UNIVERSITY *of the* WESTERN CAPE

Faculty of Naturel Sciences

MASTERS OF SCIENCE THESIS

SANBI
South African National Bioinformatics Institute

Data Science techniques for predicting plant genes involved in secondary metabolites production

Author **:**
Mr. **Ben Ilunga Muteba**

Course of studies **:**
Bioinformatics

A thesis submitted in partial fulfilment for the degree of Master of Science at the South African National Bioinformatics Institute, Faculty of Natural Sciences, University of the Western Cape

DATE : *Cape Town, 29 November 2018*

Declaration

I, Ben Ilunga Muteba, declare that this thesis titled, "*Data Science techniques for predicting plant genes involved in secondary metabolites production*" and the work presented in it is my own, that has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have indicated and acknowledged by complete references.

Signed : **Date** : 29 Nov 2018

Dedication

Dedicated to every single person who is in abject poverty as I was, and who despite the unfairness of life is willing to educate oneself.

Acknowledgement

I would like to acknowledge, Dr. Uljana Hesse for her invaluable comments, views and guidance provided during the writing up of my thesis. I would like to acknowledge, Prof Junaid Gamieldien for his invaluable comments, views and guidance provided during the writing up of my thesis. I would like to thank Dr. Dominique Anderson for her important comments, and editing of my thesis, I am very humbled by her work and her willingness to help me through despite the odds. I would like to thank Mr. Peter Van Heusden for his valuable comments and guidance provided during my writing up of my thesis. I would like to thank The Director of SANBI, Prof Alan Christoffels for his critical advice and guidance throughout the odds. I would like to thank the Deputy Dean: Research & Postgraduate Studies, Prof N. Ndiko Ludidi for his invaluable assistance and support during the perplex time of my academic journey.

I would like the thank the Dean of the Faculty of Natural Science: Prof Davies-Coleman Micheal, for his support and invaluable guidance of my research and graduate studies at the University of the Western Cape.

I would also like to thank my wife Emily Westerlund for her encouragement, insightful comments, involvement and challenging questions, for her patience, and understanding, for her extraordinary support during my educational journey from my first year all the way to my masters' year. Special thanks go to my editor Randel Zachman for her assistance. I would like to thank all the people who contributed in some way to the work described in this thesis as well as my family for their encouragement and love. Lastly, my special gratitude goes to Ada Bertie Levenstein and the National Research Fund (NRF) for financial support during this masters' research.

Finally, I would like to thank the Deputy Vice Chancellor, Prof Vivienne Lawack and the Dean of the Faculty of Natural Sciences,Prof Davies-Coleman Micheal for publishing my essay on:

A reflection on my creative and critical thinking journey as an undergraduate student at UWC and the impact of this thinking experience on my postgraduate research

https://www.uwc.ac.za/News/Pages/Academic-Week-2018-Project-Why-Celebrates-Critical-Thinking-And-Creativity-.aspx

ABSTRACT

Author: Ben Ilunga Muteba

Title: Data Science techniques for predicting plant genes involved
in secondary metabolites production

Institution: University of the Western Cape

Department: SANBI

Degree: Master of Science

Year: 2018

Plant genome analysis is currently experiencing a boost due to reduced costs associated with the development of next generation sequencing technologies. Knowledge on genetic background can be applied to guide targeted plant selection and breeding, and to facilitate natural product discovery and biological engineering. In medicinal plants, secondary metabolites are of particular interest because they often represent the main active ingredients associated with health-promoting qualities.

Plant polyphenols are a highly diverse family of aromatic secondary metabolites that act as antimicrobial agents, UV protectants, and insect or herbivore repellents. Most of the genome mining tools developed to understand genetic materials have very seldom addressed secondary metabolite genes and biosynthesis pathways. Little significant research has been conducted to study key enzyme factors that can predict a class of secondary metabolite genes from polyketide synthases.

The objectives of this study were twofold: Primarily, it aimed to identify the biological properties of secondary metabolite genes and the selection of a specific gene, naringenin-chalcone synthase or chalcone synthase (CHS). The study hypothesized that data science approaches in mining biological data, particularly secondary metabolite genes, would enable the compulsory disclosure of some aspects of secondary metabolite (SM).

Secondarily, the aim was to propose a proof of concept for classifying or predicting plant genes involved in polyphenol biosynthesis from data science techniques and convey these techniques in computational analysis through machine learning algorithms and mathematical and statistical approaches.

Three specific challenges experienced while analysing secondary metabolite datasets were: 1) class imbalance, which refers to lack of proportionality among protein sequence classes; 2) high dimensionality, which alludes to a phenomenon feature space that arises when analysing bioinformatics datasets; and 3) the difference in protein sequences lengths, which alludes to a phenomenon that protein sequences have different lengths.

Considering these inherent issues, developing precise classification models and statistical models proves a challenge. Therefore, the prerequisite for effective SM plant gene mining is dedicated data science techniques that can collect, prepare and analyse SM genes.

PCA and TSNE were implemented to visualise the behavior of the SM datasets. Three feature sets were developed: i) Amino acid frequency-based features, ii) Value-based features, and iii) frequency-based features. Eight features were engineered from ii and iii. Feature selection was then performed on these later two feature sets and it was found that all the eight features were significant, to which data visualisation was applied to visualise their significance levels. These eight features were further transformed into 8-Selected feature matrix (8SFM), and the feature set in i) was transformed into a twenty relative frequency feature matrix (20RFFM).

Both matrices were then used to conduct inferential statistical analysis with ANOVA and Chi-squared models with their post hoc tests (Tukey's HDS and Bonferroni respectively) and a boxplot, and to train eight binary classification models: Logistic Regression (LR), Decision Tree (DT), Random Forest with 100 trees (RF100), Support Vector Machine (SVM), K-Nearest Neighbor (4NN and 2NN), Naïve Bayes (NB), Single Perceptron (SLP), and Multilayer Perceptron (MLP) neural network.

The hypotheses were tested on these learned models, producing positive results, with a performance of 94.2% as the highest average accuracy of the 2NN binary classifier. Furthermore, the statistical models used inferential statistics to make judgments of the distribution of SM genes and reveal interesting inferential statistics among the three SM datasets under observation. The statistical analysis conducted for this study resulted in a 95% confidence that the labeled class of reviewed chalcone synthase (RCHS) and the labeled class of unreviewed chalcone synthase (UCHS) within each dataset held similar properties as opposed to the labeled class of Not chalcone synthase (NCHS).

In summary, the proof of concepts, and techniques developed as part of this study hold the prospective revolution of the preparation, analysis and understanding of SM genes involved in polyphenol production, but can be extended to other metabolomics, proteomics and genomics studies.

Keywords: Medicinal plants, Secondary metabolites, Polyphenols, Chalcone Synthase, feature engineering, feature selection, data visualisation, machine learning techniques, mathematic-statistical approaches.

To my twelve siblings, my parents, my extended family, despite the odds, you have produced the first-generation graduate and rightfully so. Obtaining this master's degree is an achievement in and of itself. I hope this changes the culture in the family and that extended education is not alien anymore.

Contents

Declaration ii

Dedication ii

Acknowledgement iii

ABSTRACT iv

List of Figures xi

Conferences and Presentations xiii

List of Abbreviations xiv

Preface: xv

Chapter One: Introduction 1

1.1 Introduction and Problem Statement 1

1.2 Data Science 1

1.3 Research Questions 2

1.4 Research Objectives 2

1.5 Research Aim 2

1.6 Contributions 3

1.7 Dissertation Structure 3

Chapter Two: Literature Review 5

2.1 Medicinal Plants 5

2.2 Secondary Metabolites 5

2.2.1 Terpenoids 7

2.2.2 Alkaloids 7

2.2.3 Phenolics 7

2.2.4 Glycosides 7

2.3 Polyphenols 7

2.3.1 Flavonoids 9

2.3.2 Phenolic Acids 9

2.3.3 Lignans 10

2.3.4 Stilbenes 10

2.4 Effect of Polyphenols on Human Diseases 10

2.5 Chalcone Synthase 11

2.6 Chalcone Synthase Catalytic Activity 12

2.7 The Shikimate Pathway and the Phenylpropanoid Pathway 12

2.8 Computational Biology 13

2.9 Mathematical and Statistical Approaches 14

2.9.1 Data Quality 15

2.10 Statistical Analysis on Biological Data 15

2.10.1 Chi-square Hypothesis Testing 15
2.10.2 Bonferroni Correction test 16
2.11 Statistical Model 19
2.11.1 Analysis of Variance Hypothesis Testing 19
2.11.2 Tukey's Honest Significant Difference test 20
2.12 Machine Learning 21
2.12.1 Data Cleaning 22
2.12.2 Data Pre-Processing 22
2.12.3 Feature Engineering 23
2.12.4 Feature Selection 25
2.12.5 Model building 27
2.12.6 Machine Learning Algorithms 28
2.12.6.1 Supervised Learning Algorithms 28
2.12.6.2 Unsupervised Learning Algorithms 35
2.12.6.3 Model Evaluation Performance Metric 37
2.12.6.4 Model Overfitting 39
2.12.6.5 Cross-validation 40
2.13 Summary 40
Chapter Three: Research Process 42
3.1 Contributions 42
3.2 Introduction 43
Methods Motivation for exploratory Proof of Concept on Secondary Metabolite genes 43
3.3 Data Collection 43
3.4 Data Preparation 44
3.4.1 Materials (Software and Hardware Specifications) 44
3.4.2 Data Integration 45
3.4.3 Data Cleaning 45
3.4.4 Data Transformation and Feature Engineering 45
3.4.5 Baseline Dataset 46
3.4.6 Data Standardisation 46
3.4.7 Feature Selection 46
3.4.8 Experimental Data Quality 47
3.5 Data Analysis 47
3.5.1 Data Visualisation 47
3.5.2 Supervised Machine Learning Classifiers 47
3.5.3 Statistical Analysis 47
3.6 Limitations of the study 48
Chapter Four: Feature Engineering – Implementation and Result 49
4. Introduction 49
4.1 Data Collection 49

4.1.1 Plant Protein Sequence Data Resources 49
4.2 Data Preparation Implementation and Discussion 51
4.2.1 Data Integration 51
4.2.2 Data Cleaning Implementation 52
4.2.3 Data Transformation and Feature Engineering Implementation and Discussion 52
4.2.4 Baseline Database -- Data Transformation -- Feature Engineering Result and Discussion 57
4.3 Summary 57
CHAPTER FIVE: Data Visualisation and Feature Selection Techniques 59
5. Implementation and Results 59
5.1 Introduction 59
5.2 Contribution 60
5.3 Principal Component Analysis on 20RFFM 60
5.4 t-Distributed Stochastic Neighbor Embedding on 20RFFM 62
5.5 Data Visualisation for Computational Feature Selection on 8SFM 64
5.5.1 Feature Selection using Random Forest and Forest of Trees Techniques 64
5.5.2 Spearman's Rank correlation on 8SFM 66
5.5.3 Data Visualisation ANOVA and Mutual Information feature Selection techniques 67
5.6 Summary 68
Chapter Six: Machine Learning Supervised Classifiers Implementation and Results 69
6. Introduction 69
6.1 Contribution 70
6.2 Machine learning Supervised Classification Implementation and Results 70
6.2.1 Model Building 70
6.2.2 Model Training 71
6.2.3 Performance Evaluation Metrics of Proof of Concept Binary Models 71
6.2.4 Performance Evaluation Metrics of Proof-of-Concept Multiclass Models 75
6.3 Average Performance Accuracy of Binary and Multiclass Models 77
6.4 Summary 78
Chapter Seven: Statistical Analysis and Results 79
7.1 Introduction 79
7.2 Contribution 79
7.3 Normal Distribution 79
7.4. Chi Square Hypothesis Test 80
7.4.1 Bonferroni Correction test 81
7.5 One-way Analysis of Variance 82
7.5.1 Tukey's Honest Significant Difference Post Hoc Test 83
7.6 Boxplot 84
7.5 Summary 85
Chapter Eight Conclusion 87

8.1 Conclusion 87
8.2 Summary of findings 87
8.3 Future Work 90
References 92
Appendix A.2 98
Appendix A.1 99
Appendix B 100

List of Figures

Figure 2. 1 Principle biosynthetic pathway leading to synthesis of secondary metabolites. Taken from 'Effect of CO2 enrichment on synthesis of some primary and secondary metabolites in ginger,' by A. Ghasemzadeh & H. Jaafar, 2011, International Journal of Molecular Sciences 12(2). 6

Figure 2. 2 Diet rich in polyphenols 8

Figure 2. 3 Types of Polyphenols 8

Figure 2. 4 Main food sources of polyphenols and Molecular Structure 9

Figure 2. 5 Phenylpropanoid Metabolic Pathway 11

Figure 2. 6 Shikimate Pathway 13

Figure 2. 7 Gene Clustering Analysis. In these graphs, seed genes are the genes given as input, while output genes are differentially represented according to their importance in terms of degree. 17

Figure 2. 8 Hierarchical cluster analysis dendrogram using Euclidean distance and the link between the groups by War method for agronomic traits: Plant height, stem diameter, number of primary branches, seed weight, seed width, seed length etc. 17

Figure 2. 9 Underlying mechanism of Machine Learning Model Building 27

Figure 2. 10 Three Major Groups of Machine Learning Algorithms 27

Figure 2. 11 The difference between classification and regression algorithms 29

Figure 2. 13 Logistic Model Vs. Regression Model 32

Figure 2. 14 Perceptron Neuron Network Supervised Learning 33

Figure 2. 15 Multiple Feed-forward Neural Network 34

Figure 3. 1 Data Science Computational Pipeline 42

Figure 4. 1 Data Collection-- Different curated plant protein sequences not having any chalcone synthase catalytic activity (True Negative control set), Excel compilation derived from information from the database 51

Figure 4. 2 Integration of Three classes for Model's building 51

Figure 4. 3 Data transformation and Feature engineering process 53

Figure 4. 4 A synopsis of the 2 feature (Frequency-based feature and Value-based feature) 56

Figure 4. 5 A synopsis of the 20 Amino Acid feature engineered 57

Figure 5. 1 Graphical representation of Chapter 5 contents 60

Figure 5. 2 First and Second Principal Components of PCA on RCHS, UCHS, and NCHS 61

Figure 5. 3 First and Second Principal Components of PCA on RCHS, and NCHS 62

Figure 5. 4 Principal Component's Steps, indicating that two pca component can represent more than 95% of the data 62

Figure 5. 5 TSNE with RCHS, NCHS, and UCHS coloured by target 63

Figure 5. 6 TSNE-PCA Analysis of the 3 Classes; RCHS, UCHS, UCHSC 63

Figure 5. 7 Forest of Trees Feature Selection Techniques 65

Figure 5. 9 Five Features and their F-scores and MI-scores 67

Figure 6. 1 Optimal Value of K using Cross-validation. Each point on the line represents a K value, and 4 yields a higher prediction accuracy. 72

*Figure 6. 2 Optimal Value of K using Grid-search Cross-validation. Each point on the line represents a K value, and **2** yields a higher prediction accuracy.* 73

Figure 6. 3 Validation confusion matrix for the binary 2NN proof-of-concept model. 74

Figure 6. 4 AUC average probability the SVM multiclass classifier assigns a higher probability to chalcone synthase Proof of Concept 77

Figure 7. 3 mean =0.069643 sigma = 0.004330 RCHS Distribution 80
Figure 7. 3 mean =0.068556 sigma =0.005692 UCHS Distribution 80
Figure 7. 3 mean =0.081414 sigma =0.014353 NCHS Distribution 80
Figure 7. 4 Boxplot of the 3 classes grouped by AromaFeature 84

Conferences and Presentations

SANBI

Vera-Solution

DataHack4FI Community

List of Abbreviations

AUC	**:** Area Under the receiver operating characteristics Curve
DT	**:** Decision Tree
KNN	**:** K-Nearest Neighbor
LR	**:** Logistic Regression
MLP	**:** Multilayer Perceptron Network
GNB	**:** Gaussian Naïve Bayes
8SFM	**:** 8 Selected Feature Matrix
NCHS	: Not Chalcone Synthase
RCHS	**:** Reviewed Chalcone Synthase
RF100	: Random Forest with 100 trees
20RFFM	**: 20** Relative Frequency Feature Matrix
SLP	**:** Single Perceptron Network
SM	**:** Secondary Metabolites
SVM	**:** Support Vector Machine
UCHS	**:** Unreviewed Chalcone Synthase
WHO	**:** World Health Organisation

Preface:

The focus of the thesis is on the development of exploratory proof of concept computational models which use feature engineering and predictive modelling techniques to produce outputs which mimic the results of traditional biological studies. It is important to note that computational modeling is based on mathematical and statistical estimated approaches which interrogate the properties of data and draw inferences. It is not intended to explain the biological properties of the data under review, but rather to apply computational models to draw conclusions about the data. Therefore, although the development of the model may be informed, on some level, by biological theory, it is not intended to follow the exact theory and methodology of traditional biological enzyme analysis, but to merely produce the same outcome in a faster, cost-efficient manner.

Chapter One: Introduction

1.1 Introduction and Problem Statement

World-wide, six floral kingdoms exist, and South Africa is home to one of these kingdoms. Over 9000 species belong to the Cape floral kingdom, 6200 of which are endemic and occur nowhere else in the world. More than 3000 plants in South Africa with medicinal properties are known. These plants represent a tremendous, yet untapped biological resource for bioprospecting towards development of novel drugs (Yadav, Khare, & Singhal, 2017, Lui *et al.*, 2004). Genome analysis of South African medicinal plants has been initiated through the Aspalathus linearis (rooibos) genomics programme at the University of the Western Cape, which encompasses the sequencing of the rooibos genome (150x genome coverage) as well as six diverse transcriptomes. It was funded by the NRF in 2016 (project numbers RTF150421117446 and CSUR150714125961) with the aim to:

1) Improve rooibos production through the development of genomic markers for agronomically important traits (e.g. nodulation, stress tolerance, production of medicinal compounds) to target plant selection.

2) Open the rooibos genome for biotechnological exploration.

3) Enhance our understanding of the biology of fynbos plants.

4) Develop biocomputational approaches for future medicinal plant genome analyses.

Plant genome analysis is currently increasing due to the reduced sequencing costs associated with the development of improved next generation sequencing technologies. Knowledge on the genetic background of plants can be applied to guide targeted selection and breeding, and to facilitate natural product discovery and biological engineering. Medicinal plants are of particular interest, as their genomes encode diverse metabolic pathways for pharmacologically active compounds.

The prerequisite for effective plant genome mining is dedicated biocomputational tools that can be used to identify gene pathways involved in the production of natural plant products. To date, only one tool has been developed, called PlantiSMASH. PlantiSMASH applies comparative genomics and transcriptomics analyses to pinpoint clusters of metabolic gene loci within the plant kingdom (Kautsar, Suarez Duran, Blin, Osbourn, & Medema, 2017). However, the genes for secondary metabolite production are not always co-localized across plant genomes. Furthermore, plant enzyme classification is still in progress with the current protein domain library of PlantiSMASH limited to 62 entries.

As part of this programme, this study aims to implement biocomputational approaches through data science techniques to study genes involved in the biosynthesis of secondary metabolite production. This research project lays the foundation for the development of new and innovative solutions for biocomputational mining of medicinal and crop plant genomes.

1.2 Data Science

Traditional statistics has been around for centuries, and for just as long researchers have been plugging away at trying to build models that aid us to extract information from data. Until recently, many of these studies were unobserved by the public, as their rationales seldom offered practical solutions to problems that involve the often noisy data that exists in real world. Through the data science revolution currently underway, a recent new wave of development in computer hardware and software has been the engine fueling the field of mathematical statistics to touch almost every type of dataset. For many years, different machine learning algorithms have cast a long shadow over statistical models that have been so crucial in helping statisticians in predicting and interpreting data, such as ANOVA, Chi-square and so on.

In this study, data science is introduced to infer different disciplines that are practically concerned with inference (the relationship between independent variables and dependent variables, i.e., mathematical statistics) and prediction (statement about future behaviors, i.e., machine learning). Consequently, we make use of data science (inference and prediction) approaches to utilize statistics and machine learning as two interconnected forces.

1.3 Research Questions

The following research questions arise, and constitute the main pillars of the research conducted in this study:

1. Can machine learning algorithms be trained to recognize plant secondary metabolite genes involved in the production of medicinally active compounds (e.g. polyphenols)?
2. Can mathematic statistical estimated approaches be carried out pertaining to the preparation, analysis and interpretation of secondary metabolite genes?

1.4 Research Objectives

The objectives of this study were:

(i) To develop a baseline dataset and data science computational pipeline.

(ii) To develop novel feature sets for secondary metabolite genes and protein sequences in general.

(iii) To develop inferential statistical models for secondary metabolite genes and protein sequences analysis using the novel feature set derived in (ii).

(iv) To develop machine learning supervised binary classifiers and multi classifiers with an appropriate feature set derived in (ii) for secondary metabolite gene classification using reviewed secondary metabolite genes i.e., chalcone synthase, and to test the model on available secondary metabolite data.

1.5 Research Aim

The challenge in classifying SM genes lies in profiling a set of identified plant SM genes. To accomplish the goal understanding of SM genes, this study uses data science (mathematics and statistics, and machine learning) approaches, to make new predictions, or infer new biological statistical insights. The aim of this study is therefore to apply data science techniques to gain a good understanding of plant SM genes and present a computational dynamic technique to mine plant SM genes.

1.6 Contributions

This thesis involves the exploration of different aspects related to the analysis of secondary metabolite plant genes and proposes a data science computational pipeline that introduces;

i) data preparation steps that help in dealing with bioinformatics data noise, high dimensionality, and class imbalance, and

ii) data analysis procedures for machine learning classification models and statistical models

Three feature sets were developed, and best practice guidelines are provided. These guidelines allow for development of improved statistical analysis models for the discovery of new biological insights, and, machine learning classifiers for the prediction of secondary metabolite genes involved in polyphenols biosynthesis.

The key research contributions are listed below:

1. Evaluation on three exploratory proof of concept sets of feature engineering when learning from high dimensional bioinformatics datasets with varying length sequences (heterogeneous data) is introduced.
2. The study provides the first comprehensive use of relative frequency feature matrix (RFFM) and 8-Selected feature matrix (8SFM) as a proof of concept model trained and as a data preparation tool, in the context of data quality and the alleviation of sequence unevenness. These matrices have proven to boost the predictive power and inferential analysis in a computational cost-efficient manner.
3. This study presents the effectiveness of statistical approaches on bioinformatics datasets which are empirically addressed in Chapter 7.
4. Different approaches for developing feature engineering and examining feature selection are presented through data visualisation, in the context of data quality, to determine best practices (Chapter 5).
5. The development of a baseline dataset and data science computational pipeline.
6. The building of a biocomputational program that handle all the processes of biological data preparation, machine learning classification analysis and statistical analysis.

1.7 Dissertation Structure

The Thesis is organized as follows:

Chapter 1 presents an introduction to this study and the departmental project from which this study is part of. Chapter 2 provides a deep sight of the literature tracked in experiments performed throughout different works. This chapter elaborates first on medicinal plants, biological and biochemical properties of plant genes and specifically secondary metabolite genes. Second, it discusses the computational biology approaches, and mathematical and statistical approaches in bioinformatics. Lastly, it elaborates on major aspects of machine learning techniques in bioinformatics. Chapter 3 introduces a computational pipeline, materials, method motivation needed for the implementation of the study and evaluation of computational result. This chapter presents an overview of the tools used for the computation of secondary metabolite genes. Chapter 4 provides, to our knowledge, the first development of three sets of feature extraction and feature selection techniques when learning from high dimensional bioinformatics datasets with varying lengths of data quality, through mathematical and statistical approaches. Chapter 5 provides the first comprehensive data visualisation for secondary metabolite gene features, examination of feature engineering, and feature selection, when learning from bioinformatics secondary metabolite gene datasets with varying lengths of amino acid sequences. Chapter 6 assesses the effectiveness of eight supervised machine learning classification techniques as well as classification models when learning from bioinformatics secondary metabolite gene datasets, by applying data sampling for classification problems. This chapter addresses the first question of our research and present the machine learning output on predicting SM genes. Chapter 7 Presents different statistical models, combining feature selection and data sampling in the context of inferential statistical analysis. These statistical models provide an empirical analysis to give practitioners guidance on best practices when analysing bioinformatics data to retrieve meaningful, important and reliable information from biological datasets. Chapter 8 presents conclusion of the work and suggestions for future work.

2.1 Medicinal Plants

The World Health Organization (WHO) has calculated that globally, around 60,000 plant species are utilized for their remedial, dietary or aromatic properties (Robinson & Zhang, 2011). It is estimated that over 500,000 tons of material from these plant species are exchanged per year. Worldwide trade in plants for medicinal purposes is exponentially increasing and estimated at more than 2.5 billion USD (Dushenkov, 2016). Currently, pharmacopoeia records worldwide contain medicinal drugs extracted from indigenous plants (Yadav *et al.*, 2017) and as such, medicinal plants have remained the most prominent natural source of medicines.

One of the oldest traditional techniques employed by humans for treating maladies is the use of therapeutic plants (Geethangili & Tzeng, 2011). Medicinal plants have been utilized remedially around the world, forming a critical component of different customary medication schemes Forms of phytotherapy have been used as the basis of treatment regimes in a variety of traditional medical systems, from Ayurveda to Unani, and various pharmacological drugs are derived from plant products (Van Wyk & Wink, 2017).

Plant parts contain some chemical compounds and phytochemicals that are used as active ingredients in the biosynthesis of secondary metabolites within the plant metabolism. The nature of the active ingredients may necessitate subsequent regulation of the use of phytomedicines. The parts of medicinal plants that contain natural chemical compounds are called phytochemicals. They provide the most important sources for the treatment of many diseases (Lui *et al.*, 2004). Phytochemicals present in plants are responsible for a multitude of plant properties, such as plant colour, odour and flavor these phytochemicals are responsible for different plant properties, such as the organoleptic properties of the plant and plant colour (Yadav *et al.*, 2017, Lui *et al.*, 2004). The consumption of the whole plant with phytochemicals can produce potential health benefits and could be used as dietary supplements. Phytochemicals in foods have diverse and complex chemical structures and are classified into polyphenols, terpenoids, alkaloids and other nitrogen compounds, carbohydrates and lipids (Slimestad *et al.*, 2005) (figure 1).

The presence of phytochemicals in a plant, as well as the combinations of the active compounds that yield specific physiological action on the health of humans, can have an influence on the value of a medicinal plant (Saxena, Saxena, Nema, Singh & Gupta, 2013). Some phytochemicals found in indigenous medicinal plants, such as flavonoids, phenolic compounds, alkaloids, and tannins, can be of greater importance for medicinal purposes than the other compounds (Yadav *et al.*, 2017).

2.2 Secondary Metabolites

Plants possess many phytochemicals, with approximately 10 000 being identified (Zhang *et al.*, 2015). These compounds are produced to help plants fight against predators or pathogens (Upadhyay, Upadhyaya, Kollanoor-Johny, & Venkitanarayanan, 2014). However, not all phytochemicals are beneficial to health and some are considered to be poisonous and detrimental to human health (Francisco *et al.*, 2017).

Medicinal plants are the source of secondary metabolites (SM). As opposed to the primary metabolites of plants, SM of plants are organic compounds that do not directly contribute to the reproduction, growth or development of plants (Kaul, Gupta, Sharma, & Dhar, 2017). However, these SM do indirectly impact plant health in that they can serve as a protective mechanism against herbivory, facilitate interactions in plant species, and impact plant fertility (Stevenson, Nicolson, & Wright, 2017, Clemensen *et al.*, 2017).

Secondary metabolites are used by humans for flavoring, medicinal and recreational purposes. Many are known to exhibit antioxidant, antimicrobial, anticoagulant, anti-inflammatory, antidiabetic, anthelminthic and lipid-lowering properties (Kaul *et al.,* 2017). One group of these chemical compounds are toxic to living cells (cytotoxic) which can help prevent the spread of tumors and angiogenesis by boosting the immune system to fight against cancer cells. Another group of these chemical compounds can protect nerve cells against chemicals and strokes (oxygen deprivation), used in a way that promotes nerve cell regeneration. Another group of chemical compounds is employed to protect the skin from ultraviolet damage, to protect the liver against poisons such as carbon tetrachloride, to thwart calcium loss from bone and increase fetal lung maturation (Stevenson *et al.*, 2017).

The specificity of SM has been well studied and can be characterized to individual medicinal plant species. Their specificity has been mostly restricted to a narrow set of plant species found in a phylogenetic group (Francisco *et al.,* 2017; Clemensen *et al.*, 2017). Many secondary metabolite sources that have been studied such as, flavonoids, phenolic compounds, alkaloids, tannin, etc. are commonly found in specific medicinal plants (Francisco *et al.*, 2017). The presence of these SM in plants gives a plant a high potential healthful benefit (Clemensen *et al.*, 2017).

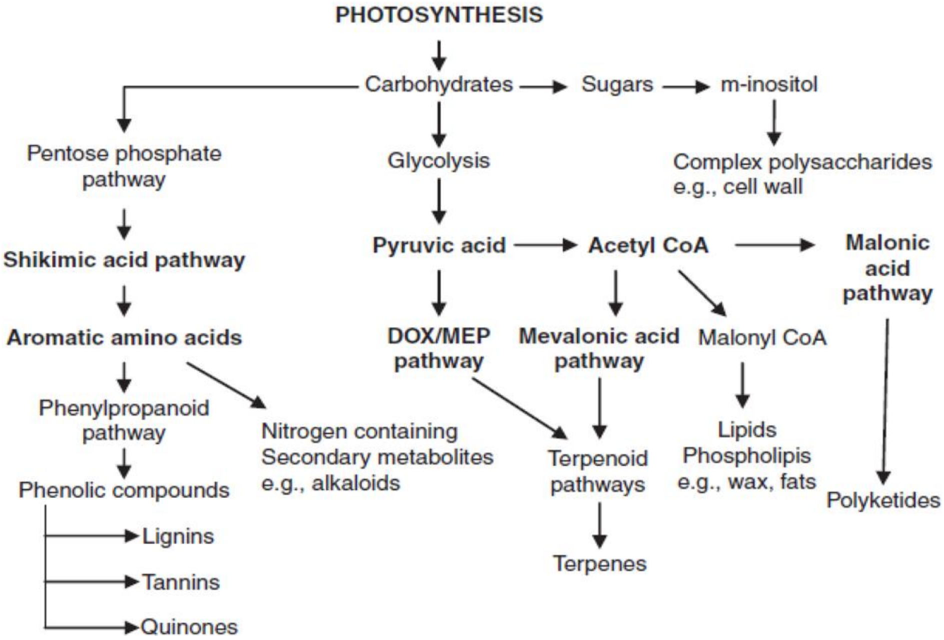


Figure 2. 1 Principle biosynthetic pathway leading to synthesis of secondary metabolites. Taken from ‘Effect of CO2 enrichment on synthesis of some primary and secondary metabolites in ginger,’ by A. Ghasemzadeh & H. Jaafar, 2011, *International Journal of Molecular Sciences 12(2).*

The healthful benefits of SM are enormous and very essential for human resistance against diseases. There is an estimate of 250,000 secondary metabolites in plants (Rehman, 2016). Their classification is determined by their biosynthetic pathways, derived from primary metabolites, where a chemical is derived. The four major classes of secondary metabolites are: Terpenoids; Alkaloids; Phenolics; and Glycosides.

2.2.1 Terpenoids

Terpenes are generally 5-carbon unit polymers of isoprene and are promoters of scent, flavor and colors. Some plant hormones (phytohormones) that influence plant physiology are produced during the terpenoid pathway (Cseke *et al.*, 2016).

2.2.2 Alkaloids

Alkaloids are nitrogenous compounds primarily found in plants and include classes such as morphine, nicotine and caffeine (Hussain *et al.*, 2018). These bitter tasting compounds are derived from amino acids such as phenylalanine, tyrosine, tryptophan, histidine, anthranilic acid, lysine and ornithine (Krechmer *et al.*, 2015; Bodi *et al.*, 2014).

2.2.3 Phenolics

Phenolic secondary metabolites are ubiquitous in plants, having an influence on plant reproduction strategy, plant defenses against biotic and abiotic stress and even plant-plant interaction (Heleno, Martins, Queiroz, & Ferreira, 2015). Phenolics contain a core formed by at least one phenol ring and are derived from aromatic amino acids such as phenylalanine, tyrosine (although generally grouped as neutral), and tryptophan (Heleno *et al.*, 2015; Działo *et al.*, 2016). Some examples of plant phenolics include coumarins (antimicrobial agents, feeding deterrents, and germination inhibitors) and lignin (abundant in secondary cell wall, and resistant to extraction or many degradation reagents, such as anthocyanins, flavones, flavanols).

2.2.4 Glycosides

Glycosides assume various vital roles in many living organisms. Glycosides molecules are formed when a sugar group (glycone) binds to a different functional group (aglycone) via a glycosidic bond. There are four main glycosidic bonds that allow glycosides to link: an O- (an O-glycoside), N- (a glycosylamine), S- (a thioglycoside), or C- (a C-glycosyl). In plants, which store glycosides in non-active form, enzyme hydrolysis is required for their activation by hydrolyzing the sugar moiety and exposing the rest of the molecule, which can be used in medicines. Some plants are also utilizing these secondary metabolite compounds as a chemical defense system against their predators. These secondary metabolite classes lead to the production of polyphenols.

2.3 Polyphenols

Polyphenols are micronutrients with antioxidant activity and are naturally occurring compounds that are usually found in vegetables, fruits, cereals, green tea, black tea, red wine, coffee, chocolate, olives, and extra virgin olive oil (Figure 2.2). Generally, plant-based foods carry a complex mixture of polyphenols. This large heterogeneous group of phytochemicals contains more than one phenolic hydroxyl group. Increasing scientific evidence is emerging on the potential healthful benefits of nutritional plant-based polyphenols. *In vitro* and *in vivo* studies have demonstrated that polyphenols possess anti-inflammatory, antioxidative, chemo preventive and neuroprotective activities and that the consumption of foods rich in polyphenols is associated, to a great extent, with lowered risk of major chronic diseases.

Figure 2. 2 Diet rich in polyphenols. Taken From (https://www.thesynergycompany.com/blog/why-a-diet-rich-in-polyphenols-is-good-for-you/)

Generally, polyphenols are divided into four diverse groups (figure 2.3):

1. Flavonoids
2. Phenolic Acids
3. Lignans
4. Stilbenes

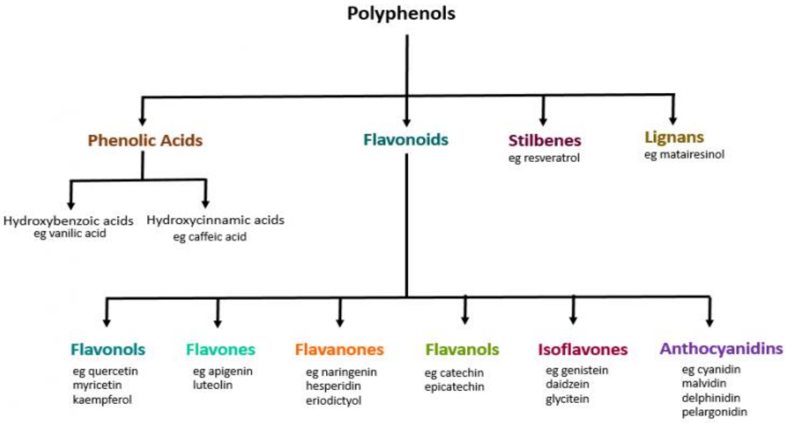


Figure 2. 3 Types of Polyphenols. Taken from (https://edurankessay.bid/?p=UG9seXBoZW5vbHM%3D)

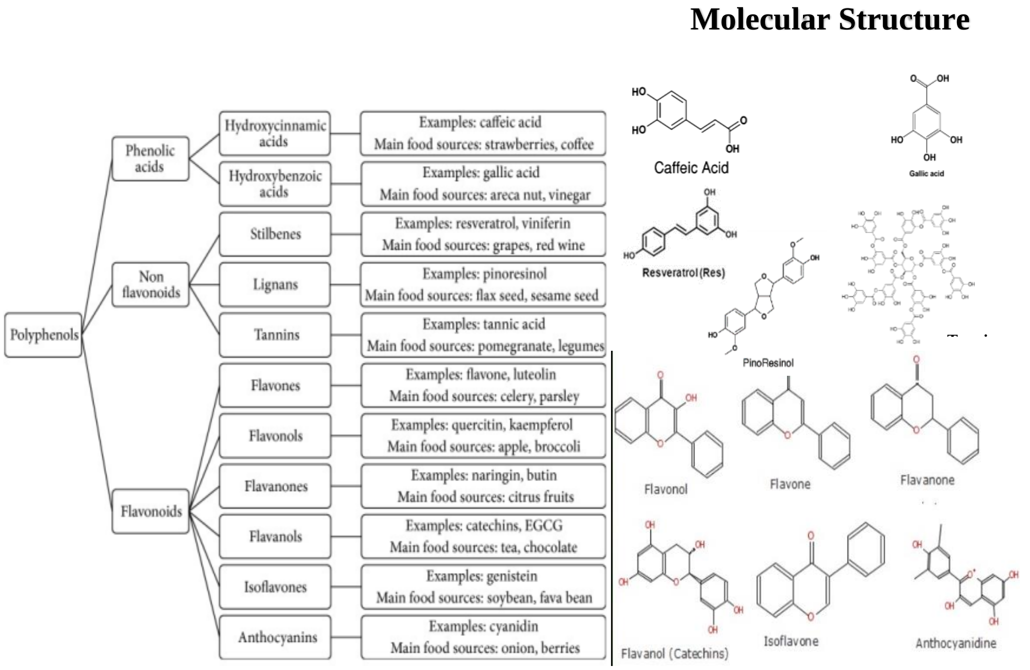

Molecular Structure
Polyphenols
Phenolic acids
Non flavonoids
Flavonoids
Hydroxycinnamic acids
Hydroxybenzoic acids
Stilbenes
Lignans
Tannins
Flavones
Flavonols
Flavanones
Flavanols
Isoflavones
Anthocyanins
Examples: caffeic acid
Main food sources: strawberries, coffee
Examples: gallic acid
Main food sources: areca nut, vinegar
Examples: resveratrol, viniferin
Main food sources: grapes, red wine
Examples: pinoresinol
Main food sources: flax seed, sesame seed
Examples: tannic acid
Main food sources: pomegranate, legumes
Examples: flavone, luteolin
Main food sources: celery, parsley
Examples: quercitin, kaempferol
Main food sources: apple, broccoli
Examples: naringin, butin
Main food sources: citrus fruits
Examples: catechins, EGCG
Main food sources: tea, chocolate
Examples: genistein
Main food sources: soybean, fava bean
Examples: cyanidin
Main food sources: onion, berries
Caffeic Acid
Gallic acid
Resveratrol (Res)
PinoResinol
Flavonol
Flavone
Flavanone
Flavanol (Catechins)
Isoflavone
Anthocyanidine

Figure 2. 4 Main food sources of polyphenols and Molecular Structure. Taken from 'Interactions between CYP3A4 and dietary polyphenols,' by L. Basheer & Z. Kerem, 2015, *Oxidative Medicine and Cellular Longevity.*

2.3.1 Flavonoids

Flavonoids are an assorted family of hydroxylated polyphenolic structures found in secondary plant metabolites (Panche, Diwan, & Chandra, 2016). Flavonoids are the largest group of phytonutrients in plant-based food products, with more than 5,000 identified compounds (Hosseini, Gholami, & Haghgu, 2016). Flavonoid compounds have been demonstrated to have antioxidant and anti-inflammatory activity (Zhang & Tsao, 2016). Furthermore, diets rich in flavonoids play a substantial role in cardiovascular health and assist in prevention of diseases such as cancer, which is caused by free-radical damage (Farzaneh & Carvalho, 2015).

However, in order to determine whether flavonoids alone are responsible for these beneficial effects on health, further studies are required (Manach *et al.*, 2017).

2.3.2 Phenolic Acids

Phenolic acids (phenolcarboxylic) are aromatic secondary plant metabolites and are found in a variety of plant-based foods (Toldra, 2017; Saltveit, 2017). They are produced via the shikimic acid through the phenylpropanoid pathway (Lynch *et al.*, 2017). Phenolic acids are a by-product of the monolignol pathway and a breakdown product of lignin and cell wall polymers in vascular (higher) plants (Saltveit, 2017). Phenolic acids in plant cell walls and lignin present a unique chemical structure of C6-C3 (phenylpropanoid type), in contrast to C6-C1 (Phenylmethyl type), which is of microbial origin (Lin *et al.*, 2016; Jang, Gang, Kim, & Choi, 2017). Hydroxybenzoic acids and Hydroxycinnamic acids (Lynch *et al.*, 2017; Demirbas, 2017) are the two significant naturally occurring types of phenolic acids (See figure 2.4).

The highest concentrations of phenols are found in plant seeds and skins of fruits and the leaves of vegetables (Lin *et al.,* 2016*;* Jang *et al.,* 2017). As seen in figure 2.4, phenolic acids are generally classified into two categories:

1. Benzoic acid with its derivatives, such as gallic acid (diet sources: tea and grape seeds)
2. Cinnamic acid and its derivatives, as well as caffeic acid (diet sources coffee, blueberries, kiwis, plums, cherries and apples) and ferulic acid (outer covering of cereal grains, corn flour, whole grain wheat, rice, and oat flours).

Phenolic acids are easily absorbed through the walls of our intestinal tract. They work as antioxidants and promote anti-inflammatory conditions and help to prevent diseases caused by oxidative damage such as coronary heart disease, stroke, and cancers (Smith, 2015).

2.3.3 Lignans

Lignans are non-flavonoid polyphenols (Panche *et al.*, 2016) and occur at highest concentration in flaxseed and bakery products containing flaxseed (secoisolariciresinol diglucoside) (Saltveit, 2017). They are widely available in drinks such as tea, coffee or wine, and in whole grains, nuts, legumes, fruits, cruciferous vegetables such as broccoli and cabbage, and seeds. In addition to these are cereals, soybeans, apricots and strawberries. Lignans are one of the largest groups of chemical compounds (polyphenols) found in plant-based foods (Lin *et al.,* 2016*;* Jang *et al.,* 2017).

Lignans, being antioxidants, support the immune system and contribute to balancing hormone levels in the body (Smith, 2015). Lignans and lignin biosynthesis source materials are byproducts of shikimic-phenylpropanoid-monolignols pathway (Calvo-Flores, Dobado, Isac-García, & Martín-Martínez, 2015). Monolignols are phytochemicals whose starting material for production is the aromatic amino acid phenylalanine (Lynch *et al.*, 2017; Calvo-Flores *et al.*, 2015). Figure 2.1 shows that the first reaction in biosynthesis is shared via the phenylpropanoid pathway. They (lignans) are classified as phytoestrogens which are estrogen-like, beneficial for the health of menopausal women and possess antioxidant activity that help protect against breast cancer (Lynch *et al.*, 2017; Saltveit, 2017).

2.3.4 Stilbenes

Stilbenes are a small family of nonflavonoid phytochemicals produced via the phenylpropanoid pathway (figure 2.1 and figure 2.4) (Calvo-Flores *et al.*, 2015, Jang *et al.*, 2017). They are polyphenolic compounds, structurally characterized by the presence of a 1,2-diphenylethylene nucleus and constitute a unique chemical scaffold in the search for bioactive molecules (Smith, 2015). Stilbenes are found in various plant families, such as Vitaceae (Lin *et al.*, 2016*;* Smith, 2015), but are less abundant in foods when compared to flavonoids, phenolic acid or Lignans (Zamora-Ros *et al.,* 2016). Food sources of Stilbene resveratrol include grape skins, red wine, peanuts, blueberries, cranberries, while stilbene pterostilbene can be found in food sources such as blueberries and grapes (Reinisalo, Kårlund, Koskela, Kaarniranta, & Karjalainen, 2015; Calabriso *et al.,* 2016). Many plant scientists have studied stilbenes to highlight their health benefits in treating chronic diseases and inflammation in aging-related diseases such as obesity, macular degeneration, Alzheimer's disease, cancer, diabetes (type 2), and heart disease.

2.4 Effect of Polyphenols on Human Diseases

Scientific biomedical research conducted in several studies has demonstrated that the consumption of polyphenols decreases the incidence of coronary heart diseases (Wang *et al.,* 2014; Clauss *et al.,* 2017; Goetz *et al.*, 2016, Reinisalo *et al.*, 2015; Lynch *et al.*, 2017; Calvo-Flores *et al.*, 2015).

Recent research by Clauss and co-authors (2017) confirmed strong anti-cancer effects of polyphenols on human cancer. Studies have reported the effects of polyphenols on human cancer cells are most often protective, and induce a reducation of the number of tumors and their growth (Clauss *et al.*, 2017; Wang *et al.*, 2014; Goetz *et al.*, 2016). The mechanisms of polyphenol underlying actions are estrogenic activity, anti-proliferation, prevention of oxidation, and anti- inflammatory activity (Wang et al., 2014).

In addition, polyphenols influence the metabolism of pro-carcinogens by simultaneously modulating the expression of cytochrome P450 enzymes involved in their activation to carcinogens (Clauss *et al.*, 2017; Wang *et al.*, 2014). Studies have also shown that onion polyphenols, especially quercetin (flavonoids-flavanols) are known to possess strong anti-diabetic activity, significantly protecting the lipid peroxidation system in diabetics (Kumar *et al.*, 2015; Cheetham & Katz, 2013). Polyphenol quercetins, particularly in red onion, have shown to be efficient against mortality from coronary thrombosis heart disease (Tresserra-Rimbau *et al.*, 2014; Tedesco, Carbone, Spagnuolo, Minasi, & Russo, 2015).

2.5 Chalcone Synthase

Chalcone synthase or naringenin-chalcone synthase (CHS) is a key enzyme in the family of type III polyketide synthase enzymes (PKS) (Shimizu, Ogata, & Goto, 2017). The specificity of type III PKS is based on its association with the production class of organic compounds, found mainly in plants as a natural defense, known as chalcones (Shimizu *et al.*, 2017; Ratnam, Choong, & Javed, 2017). CHS catalyzes the first step of flavonoid biosynthesis by directing carbon flux from general phenylpropanoid metabolism to the flavonoid pathway (Ibdah, Martens, & Gang, 2017). Naringenin-chalcone synthase produces chalcone in the phenylpropanoid pathway and flavonoid pathway by condensing one p-coumaroyl- and three malonyl-coenzyme A thioesters into a polyketide reactive intermediate that cyclizes (Yu *et al.*, 2015; Ibdah *et al.*, 2017). The CHS enzyme catalyzes the first committed step for the biosynthesis of flavonoid antimicrobial phytoalexins and anthocyanin pigments (pathway) in plants, by administering carbon flux from wide phenylpropanoid metabolism to flavonoid pathway (Ibdah et *al.*, 2017). In addition to being part of plant growth, development and adaptation, the CHS gene expression is induced in plants under stress conditions such as UV light, bacterial or fungal infection (Ibdah *et al.*, 2017; Yu *et al.*, 2015).

Chalcones, or chalconoids, are an aromatic ketone and an enone that forms the central scaffold found in a variety of important biological compounds (Abbot *et al.*, 2017). Chalcones possess a broad spectrum of interesting biological activities such as insecticidal, antioxidative, antibacterial, antiulcer, anticancer, amoebicidal, anthelmintic, antifungal, antitumor, antiprotozoal, antiviral and anti-inflammatory properties (Ibdah *et al.*, 2017; Ratnam *et al.*, 2017).

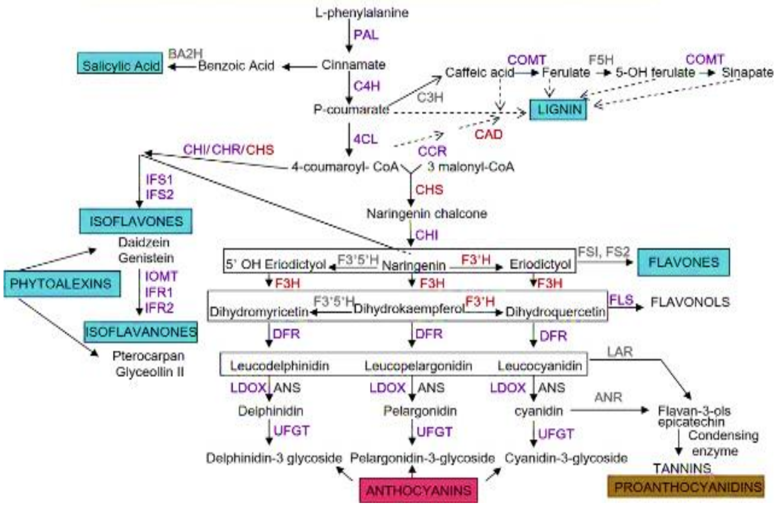


Figure 2. 5 Phenylpropanoid Metabolic Pathway. Taken from ‘Transcriptome changes in the phenylpropanoid pathway of Glycine max in response to Pseudomonas syringae infection,’ by G. Zabala et al., 2006, *BMC Plant Biology*.

In plants, chalcone synthase is a key enzyme ubiquitous to higher plants and was first observed in barley leaves (Han *et al.,* 2016). CHS proteins are found in various plant organs hence, CHS flavonoids are found in the core of diverse plant species such as *arabidopsis thaliana*, rice, grapes, medusa, *tsuga canadensis* etc.

Studies show that CHS enzyme produces flavonoids (*i.e.* lignin, suberin) and isoflavonoids (*i.e.* genistein, wighteone and lutein) which possess the power to absorb UV light radiation and hence can protect plant DNA from being damaged and from the attack ok pathogens (Shimizu *et al.*, 2017; Ibdah et al., 2017).

2.6 Chalcone Synthase Catalytic Activity

The phenylpropanoid pathway is regulated by the activity of CHS: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA +naringenin chalcone + 3 CO2 (Sun *et al.*, 2015; Gill *et al.*, 2017). CHS catalytic activity was first described in 1972 in extracts of parsley (Petroselinum crispum) (Sun et al., 2015). CHS catalytic activity is controlled through the following mechanisms (Sun *et al.*, 2015; Gill *et al.*, 2017):

- Metabolic control
- Control of CHS turnover
- Control of CHS through trans-genes

The phenylpropanoid pathway is known to be the producer of many types of secondary metabolite polyphenolic compounds such as stilbenes, phenolic compounds, lignin, flavonoids, isoflavones and flavones (see figure 2.5 and 2.4) (Yu *et al.*, 2015; Ibdah *et al.*, 2017). The flavonoids consist of various groups of plant SM such as chalcones, aurones, flavanones, isoflavonoids, flavones, flavanols, and anthocyanins (Yu *et al.*, 2015). Thus, the flavonoid pathway produces polyphenolic compounds such as naringenin, naringenin chalcone and the other end products of CoA esters that inhibit the activity of CHS in several crops (Sun *et al.*, 2015; Yu *et al.*, 2015).

2.7 The Shikimate Pathway and the Phenylpropanoid Pathway

The shikimic acid pathway (shikimate pathway) is known for its seven-step metabolic path (see appendix A.1) in the biosynthesis of folates and aromatic amino acids (phenylalanine, tyrosine, and tryptophan) used to synthesize some protozoan, bacterial, fungal, algal, and plant metabolites (Tullius, 2017; Pfister *et al.*, 2014).

The shikimate pathway initializes the phenylpropanoid biosynthesis from the shikimate aromatic compound phenylalanine (Phe), via the intermediate chorismic acid (Haslam, 2014; Tullius, 2017). The chorismic acid acts as a substrate to produce quinones and tocopherols which are important electron acceptors in photosynthesis and aerobic respiration (Gomes, Carbonari, Velini, Trindade, & Silva, 2015). The shikimate pathway as a core unit, produces SM through some of its intermediates via general phenylpropanoid metabolism (Haslam, 2014; Gomes *et al.*, 2015). The metabolic pathway of the phenylpropanoid involves several enzymes which serve as a strong foundation of plant SM (Gomes *et al.*, 2015; Kaul *et al.*, 2017).

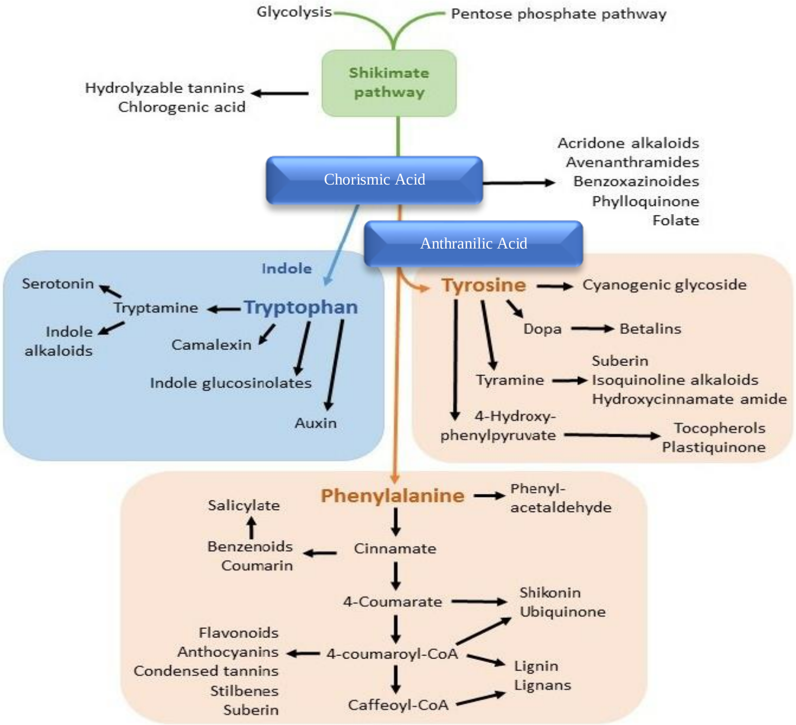


Figure 2. 6 Shikimate Pathway. Modified from 'The shikimate pathway: aromatic amino acids and phenylpropanoids,' by P.M. Dewick, 2009, *Medicinal Natural Products 137,* 86.

Plants allocate a large percentage of their fixed carbon into synthesizing phenylpropanoids. The biosynthesis of phenylpropanoids starts with the amino acids phenylalanine and tyrosine (Tullius, 2017; Haslam, 2014). The branch point enzyme; PAL (also known as phenylalanine or tyrosine ammonia-lyase) is the enzyme responsible for the biosynthesis of L-phenylalanine or tyrosine into trans-cinnamic acid or p-coumaric acid and ammonia respectively (Tullius, 2017). Hence, phenylpropanoids are a group of plant SM sourced from phenylalanine which have a large diversity of functions in terms of structural classes and signaling molecules (Gomes *et al.*, 2015). While phenylpropanoids and their byproducts have routinely been characterized as SM, some studies show their relevance to the survival of plants through different experiments in *Arabidopsis* and other plant species (Krivoruchko & Nielsen, 2015; Tullius, 2017). These studies have provided additional knowledge on various aspects of the phenylpropanoid pathway, its enzymes, molecules and the interrelationship of the pathway with the entire plant metabolism (Haslam, 2014; Pfister *et al.*, 2014).

2.8 Computational Biology

Innovative advances in the field of biology such as genomics, proteomics, imaging, biophysics, cell biology, biochemistry, and evolution have resulted in exponential increases in molecular and cell profiling information derived from substantial quantities of biological data (Boudreau & Lakhani, 2015). Developed data-analytical and theoretical methods, mathematical modeling and computational simulation techniques have been applied to the field of computational biology (Buettner *et al.*, 2015; Angermueller, Pärnamaa, Parts, & Stegle, 2016). These computational methods, when applied to the field of biology allow for the analysis of large collections of biological data in an attempt to make new predictions, or discover new biological insights (Angermueller *et al.*, 2016; Huber *et al.*, 2015).

The rapidly increasing rate of biological data generation is creating highly dimensional datasets which are challenging to analyses using conventional data analysis methodology. Current computational and machine learning techniques are demonstrating promise in leveraging the vast datasets to find concealed information within the datasets and attempt to make precise classifications or predictions. In this part of Literature review these novel types of analytical approaches for the analysis of plant SM genes involved in polyphenol biosynthesis, is explored. An overview of mathematical-statistical and machine learning approaches is presented, and the settings in which these approaches can be successfully applied to profile biological insights of SM genes are discussed.

2.9 Mathematical and Statistical Approaches

The biological data of organisms may be represented as different types of data, e.g., a protein may be represented in two dimensional images, three dimensional structures or one-dimensional sequences (Robert & Gouet, 2014). In addition, biological data is often applied in comparing the behavior of one organism, gene or sequence with the behavior of another biological unit (Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2015), resulting in even more dimensional data points to be analyzed. Mathematical and statistical approaches have helped to provide an understanding of some of the complex aspects of biological systems (Anderberg, 2014) such as comparing the behavior of known infectious disease epidemics with the behavior of an unknown, new disease epidemic (Zelditch, Swiderski, & Sheets, 2012). Different studies have conducted analyses and simulations of deterministic and stochastic models (Wilkinson, 2011; Gnauck & Straškraba, 2013), with the sole objective of establishing the epidemiological and social conditions behind the distribution and determinants of health and disease conditions in defined populations (Buettner et al., 2015). These studies have enriched the field of biomedical research, placing a major focus on the relationships between models, and disease data by putting a significant importance on the application of mathematical and statistical techniques that compute model and data veracity (Anderberg, 2014; Wilkinson, 2011).

Statistics is a broad mathematical discipline which focuses on the organisation, collection, presentation, interpretation, and analysis of data. The value of statistics lies on its ability to recognize a pattern, summarize, and draw conclusions from sample data (Wang & Peng, 2014). Statistics has become relevant to many other disciplines that are scientific, industrial, or social in nature. In relation to biology, statistics is being used to model, process and study biological problems with a view to decipher and decode the issue at hand (Wilkinson, 2011). Biological data can produce meaningful information when subjected to statistical objective concepts (Fowler, Cohen, & Jarvis, 2013; Anderberg, 2014).

Mathematical concepts, approaches, formulas, models, and techniques are used in statistical analysis to provide an explicit way of understanding a given problem, and present possible solutions. Biological data tends to be highly complex, as not all data characteristics in biology are known (Fois, Fenu, Lombrana, Cogoni, & Bacchetta, 2015; Brauer, Castillo-Chavez, & Castillo-Chavez, 2012). Fowler *et al. (*2013) have shown that statistical approaches can be used on the representative selection drawn from a given biological dataset, to uncover a number of hidden biological properties.

Complex biological processes coupled with the noisy nature of experimental data (e.g., cellular heterogeneity, microarray data, sequence data) create data uncertainty. Mathematical and statistical approaches are largely analytical methods where the general objective is to find a small number of shape functions (interpolation) or sinusoidal functions (function produced by shifting, stretching or compressing the sine function), or a small number of eigenvectors (characteristic vector), etc. that determine with sufficient accuracy the spatial and temporal properties of the biological data (Brauer *et al.*, 2012; Fowler *et al.*, 2013; Fois *et al.*, 2015). This section will discuss a number of statistical approaches that can be applied, in order to make sense of biological data.

2.9.1 Data Quality

The qualities of the results obtained in many studies rely heavily on the quality of the data being analyzed. In general terms, data quality refers to the assessment of the data's fitness for its intended uses in a given context. Data quality commonly includes the core following dimensions: accuracy, completeness, consistency, integrity, reasonability, timeliness, uniqueness, validity, and accessibility (Cai & Zhu, 2015).

Accuracy is the degree to which data correctly describes the "real world" object or event being described. Consistency is the absence of difference, when comparing two or more representations of an attribute against a definition. Integrity is the maintenance of, and the assurance of the accuracy and consistency of, data. Timeliness is the degree to which data represents reality from the required point in time. Completeness indicates that the proportion of all data fields necessary for an observation unit is captured. Uniqueness means nothing is recorded more than once based upon how a specific element is identified. Validity indicates that the data conforms to the syntax (type, format, range) of its definition. Reasonability indicates ease of understanding and concise representation of data. Accessibility represents the data source trustworthiness.

2.10 Statistical Analysis on Biological Data

A biological data experiment aims to prove, or disprove, a hypothesis. This can be answered by the significance of the results obtained. As such, "significance", has a high level of importance in biology (Fowler *et al.*, 2013; Anderberg, 2014). Statistical analysis of biological data can assign "statistical significance" to the experiment and may elaborate on the result obtained in a given study (Anderberg, 2014).

In many biological studies, a null hypothesis can be generated based on the expected result, before any experiment is undertaken (Fowler *et al.*, 2013; Fitzgerald *et al.*, 2015). The proposed null hypothesis can either be substantiated by the biological data or not, leading to the approval of statistical alternative hypothesis. For example, one of the statistical tests used most frequently to determine Mendelian ratios is the chi-square test (Burgess & Smith, 2017). Pearson's chi-square test examines if the production of deviations between observed and expected values happens by chance (null hypothesis) or by a significant factor (alternative hypothesis). In the case that the probability obtained from Pearson's chi-square happens to be high, the null hypothesis is accepted, otherwise the alternative hypothesis is accepted (Burgess & Smith, 2017; Williams, Trejo, & Schwartz, 2017).

2.10.1 Chi-square Hypothesis Testing

The chi-square formula is given by:

$$X^2 = \sum \frac{(O - E)^2}{E}$$

$$X^2 \ = The\ test\ statistics\ , O \ = \ Observed\ frequencies\ , E \ = \ Expected\ frequencies$$

The Null Hypothesis (H0) stated that there is no relationship between two variables, while the Alternative Hypothesis (H1) stated that there is a relationship between two variables.

The chi-square assumptions ensure the validity of the chi-square test results include:

- The cells of the data are counts
- The classes of the variables being tested are mutually exclusive
- Each observation contributes to only one cell within the chi-square table
- The classes tested are collected independently
- The value of the expected cells is greater than five (5). If a cell had an expected frequency less than 5, it would have used Fisher's Exact test to overcome this problem.

The chi-square tests if there is an association between target and variables. If the hypothesis results in a very small chi-square test statistic, this means that the observed data fits the expected data extremely well. In other terms, there is a relationship between the categorical variables. However, if the hypothesis results in a very large chi-square test statistic, this means that the data does not fit very well. In other terms, there is no relationship between the categorical variables (Burgess & Smith, 2017).

Chi-square hypothesis testing, tests for independence as it is for other tests like Analysis of Variance (ANOVA), where a least a test statistic is computed and compared to a critical value. The critical value for the chi-square statistic is determined by the level of the significance (usually 0.05) and the degrees of freedom. The degrees of freedom for the chi-square are calculated through the formula:

$$Df = (R\text{-}1) \text{ x } (C\text{-}1)$$

Where R is the number of rows and C is the number of columns. If the observed chi-square test statistic is greater than the critical value, the null hypothesis is rejected (Williams *et al.*, 2017).

2.10.2 Bonferroni Correction test

A post hoc test, Bonferroni correction (Armstrong, 2014), can be conducted to determine where exactly the relationship is between the different groups (Williams *et al.*, 2017). The Bonferroni correction is a method used to counteract the problem of inflated type I errors while engaging in multiple pairwise comparisons between subgroups. That is, it corrects for multiple trials by lowering the threshold of the significant p-value.

The Bonferroni correction is carried out to locate the exact association between the classes. This in practice leads to the implementation of multiple 2x2 chi-square tests using the Bonferroni-adjusted p-values. The advantage of the Bonferroni correction method lies in its capability to adjust the p-values based on planned pairwise comparisons being conducted (Armstrong, 2014). The formula is p/N,

Where:

- p = the original tests p-value and $N = (k)\ (k\text{-}1)/2$ possible pairs

Where:

- k = the number of classes

Pearson's chi-square test is largely used in genetic data, where the biological data possess enough expected values in each set (Burgess & Smith, 2017). By employing the chi-square approach, some studies have predicted plant genome size by examining whether the production of deviations between observed plant tissues and expected plant tissue values happens by chance or by a significant factor (Zhang & Finer, 2016). Other studies which have performed statistical evaluations on plant genetics, amino acids, polyphenol content and antioxidant activities of plant genes, use personal component analysis (PCA) and cluster analysis (CA) (Merel & Zwiener, 2016; Sochor *et al.*, 2011; Heller, Tripp, Turk-Kubo, & Zehr, 2014; Martinez, 2011).

A study by Sochor *et al.*, (2011) focused on detection of primary and secondary metabolites in a selection of apricot cultivars and combined the two methods of PCA and CA to evaluate the biological activity of different apricot cultivars. This study demonstrates the power of combining the two methods, where CA appears useful as a control method for PCA, in which some information goes astray.   In this study, the authors performed a normalisation of parameters based on mathematical mean to maintain the majority of the information contained in the original data. This was followed by a selection of a K-matrix of the principal components in terms of the distribution of the original data, and the selection of the parameter P expressing the degree of dispersion of the original data.

Due to the expected size of biological data sets, and the multidimensional data analysis required, these statistical approaches are largely aimed at establishing suitable bioinformatics tools for retrieving hidden biological information or predicting or classifying biological properties (Merel & Zwiener, 2016; Sochor *et al.*, 2011).

Statistical cluster analysis has been utilized in the field of genetics, and clustering approaches have been used to comprehend the diverse nature of a data set. CA has been beneficial in understanding the levels of diversity, similarity or dissimilarity in plant genes, providing a crucial understanding for the development of plant biosynthetic gene cluster (BGC) techniques used by the biocomputational tool PlantiSMASH (Kautsar *et al.*, 2017). CA can be designed to select a main group from plant metabolites, combining important biological compounds of both primary and secondary metabolites as parameters. One approach for this selection of a main group consists of splitting the plant metabolites into determined biological classes, combining them with statistical parameters to create a robust approach (Sochor *et al.*, 2011; Heller *et al.*, 2014; Kautsar *et al.*, 2017). The expected size and number of clusters is determined by the size of the data and the various biological attributes, and defined by their statistical significance (Sochor *et al.*, 2011; Fowler *et al.*, 2013; Anderberg, 2014). In this sense, CA provides an understanding of the relationships among plant metabolite classes and the biological compounds which they are made up of. The CA approach is based on the method of projection of objects to be analysed onto a multidimensional space as seen in figure 2.7. The number of dimensions is defined by the number of determined parameters, and the hierarchical interlinking of objects are structured upon their communal distance. Mathematically, objects of similar nature can be evaluated based on their similarities and placed into clusters. The standardized Euclidian distance among these distinct objects is the straight line between two points (Madzarov & Gjorgjevikj, 2010).

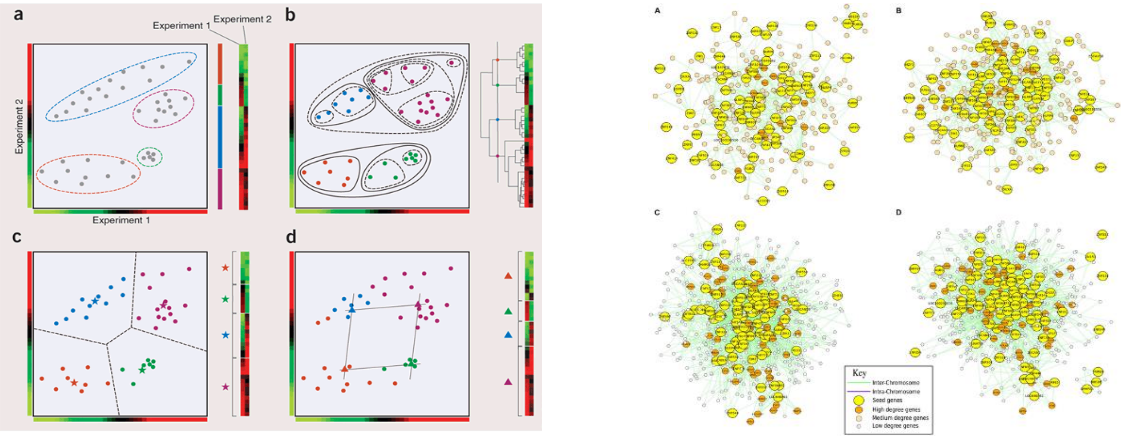


Figure 2. 7 Gene Clustering Analysis. In these graphs, seed genes are the genes given as input, while output genes are differentially represented according to their importance in terms of degree. Taken from ‘Integrating multi-omic features exploiting Chromosome Conformation Capture data,’ by I. Merelli et al., 2015, *Frontiers in Genetics 6* (40)

With this distance, Euclidean space becomes a metric space calculated as:

$$d_E(X_k\, , X_l) = \sqrt{\sum_{j=1}^{m}(X_{kj} - \quad X_{lj})^2}$$

The above Euclidean distance presents k and l as two different objects, whose distance is defined and represent the vertices of a triangle at the hypotenuse. The third vertex j defines the length of the sides, X_{kj} and X_{lj}, of a right-angled triangle, which is defined for all objects $j \epsilon \langle 1; m \rangle$. Where, m is the number of dimensions. The advised approach is to standardize the distance outputs before projecting the objects onto space. This is performed with respect to high degree of variability, for the simple reason that they have a major impact on the degree of similarity (Madzarov & Gjorgjevikj, 2010; Sochor *et al.*, 2011).

Another method of CA is hierarchical cluster (HC) dendrogram (a tree diagram showing taxonomic relationships). Hierarchical cluster illustrates the relationships among instances based on morphological traits and distance similarity. In Figure 2.7, the HC dendrogram illustrates the distance between groups as proportional to the height of the horizontal line that joins two groups. As seen in figure 2.7, the distance between groups 1 and 2 is shown to be approximately 15. The HC also orders the sub-tree in terms of cluster tightness, with the tighter clusters positioned on the left and the wider clusters positioned on the right (Van Verk Bol & Linthorst, 2011).

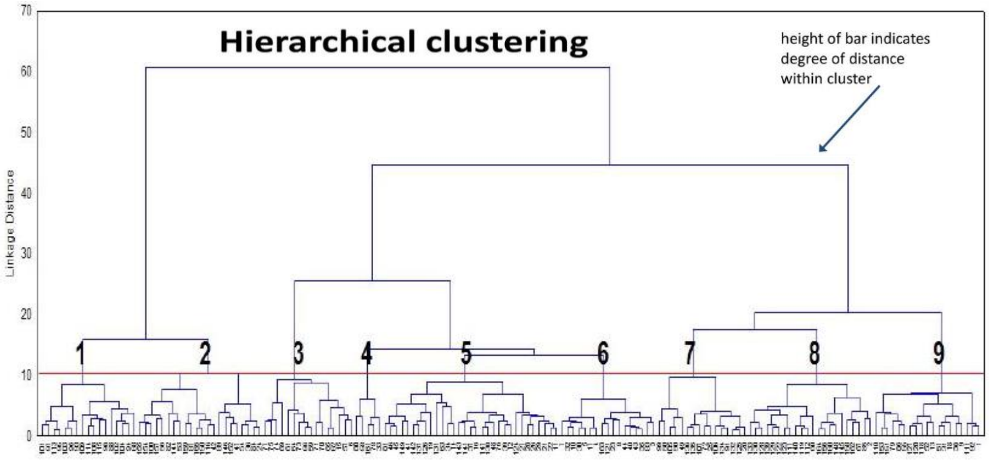


Figure 2. 8 Hierarchical cluster analysis dendrogram using Euclidean distance and the link between the groups by War method for agronomic traits: Plant height, stem diameter, number of primary branches, seed weight, seed width, seed length etc. Modified from 'Image segmentation by histogram thresholding using hierarchical cluster analysis,' by A.Z. Arifin and A. Asano, 2006, *Pattern Recognition Letters 27*(13)

Cluster analysis does however have a few limitations, as it is only suitable for examination of a data set. CA's applicability to a data set is highly dependent on the subjective choice of methods for calculating distances and cluster. For example, with proteins that are clustered based on their similarity to one another, those represented in the same cluster are potentially considered to be interacting partners (Franceschini *et al.*, 2012). However, CA may fail to properly cluster proteins that are ubiquitous, but are not necessarily functionally linked (Nepusz, Yu, & Paccanaro, 2012).

In addition, with CA one cannot tell the significance levels of instances from the observed structure (Cole *et al.*, 2013). Therefore, to understand the statistical significance of instances in clusters, additional methods such as posterior tabulation, ordination and regression, are required (Kautsar *et al.*, 2017; Cole *et al.*, 2013; Franceschini *et al.*, 2012; Nepusz *et al*., 2012).

Principal Component Analysis (PCA) is a statistical approach used to identify variations, emphasize patterns in the data set, and bring out strong expressions of the data in such a way as to highlight their differences and similarities (Sochor *et al*, 2011; Buettner *et al.*, 2015; Anders, Pyl, & Huber, 2015). This approach is based on an orthogonal transformation of original observed variables to new uncorrelated values, which are defined as principal components. PCA converts a given set of observed variables of probably correlated instances, into a set of values of linearly uncorrelated instances. Studies show that since patterns in data are not easily found in high dimensional data (Buettner *et al.*, 2015; Anders *et al.*, 2015) and graphical representation is not always possible, PCA can be used to mechanically analyse the data (Sochor *et al*, 2011).

Often data of high dimensions produce several linear cross-correlations of the observed variables, providing a more extensive understanding which is necessary for data description. Reduction in the number of linear cross-correlations is the primary benefit of PCA (Buettner *et al.*, 2015; Sochor *et al*, 2011). Studies show that the advantages of PCA are that it can be used to rank the principal components according to their decreasing distribution, and once patterns in the data have been identified, the data can then be compressed by reducing the number of dimensions, without loss of information (Candès, Li, Ma, & Wright, 2011; Sochor *et al*, 2011).

2.11 Statistical Model

Statistical modeling is a branch of mathematical modeling in which a mathematically formalized set of assumptions (a process that may have given rise to observed data) is used to approximate reality using sample data from a larger population to make predictions, classifications and interpretations (Fowler *et al.*, 2013; Fitzgerald *et al.*, 2015). Statistical modeling represents a set of probability distributions which estimate the population distribution from which the collected data is sampled. The assumptions embodied in the probability distributions of statistical models differentiate statistical models from other non-statistical and mathematical models (Wang & Peng, 2014).

In addition, statistical models are fundamental constituents of statistical inference, often embodying mathematical equations that relate random variables (and occasionally non-random variables) to drive a formal representation of a given data set (Brauer *et al.*, 2012, Fitzgerald *et al.*, 2015). This is often defined through natural transformations, functions, algebraic terms and morphisms (Fitzgerald *et al.*, 2015). Through these concepts, units, time points, instances, subjects and variables can be used to infer prediction, classification or interpretation of the data set (Wang & Peng, 2014; Fitzgerald *et al.*, 2015).

2.11.1 Analysis of Variance Hypothesis Testing

In bioinformatics, choosing the correct statistical model is not a straightforward approach. A biological dataset does not come with its own adapted model. Assumptions need to be made in relation to the desired statistical modeling (Sochor *et al.*, 2011; Fowler *et al.*, 2013; Fois *et al.*, 2015). Every statistical modelling approach is specific to the research question and to the type of data at hand (Fois *et al.*, 2015). For example, a biological study shows that glycaemia related to a distinct type of diabetes can be elucidated by a qualitative variable such as sex. Because the study was conducted on a single qualitative variable, a selection of Analysis of Variance (ANOVA) was an appropriate statistical model to analyse the biological data (Hertroijs *et al.*, 2018). However, with the same biological data, age could be selected as a quantitative variable to depict any linearly increasing or decreasing trend of glycaemia based on the age of the patients (Hertroijs *et al.*, 2018). In such a situation it is appropriate to make used of linear regression analysis to explain the data.

The analysis of variance hypothesis test compares the means of a condition between two classes. Similar to chi-square, ANOVA is an omnibus test which reviews the dataset as a whole. However, the ANOVA test does not identify where the difference is between the classes. To locate these differences in relationship between the classes, a post-hoc test can be conducted (Sochor *et al.*, 2011; Fowler *et al.*, 2013; Fois *et al.*, 2015).

For this specific ANOVA test the null hypothesis (H_0) state that there is no difference between the means of the classes, while the alternative hypothesis (H_1) state that a difference between the means exists somewhere between the classes.

The ANOVA assumptions that are applied to ensure the validity of the results of the ANOVA test include (Fowler *et al.*, 2013; Fois *et al.*, 2015):

- The variables are normally distributed in each group that is being compared in the one-way ANOVA.
- There is homogeneity of variances. This means that the population variances in each class are equal.
- There is an independence of observations.
- A caveat to these assumptions is that if the class sizes are equal, the F- statistic is robust to violations of normality and homogeneity of variance.

2.11.2 Tukey’s Honest Significant Difference test

The Tukey’s Honest Significant Difference test is a post-hoc test based on the studentised range distribution (Dominguez-Bello *et al.*, 2016). An ANOVA test can tell if the results are significant overall, but it will not tell us exactly where those differences lie. After an ANOVA test has been conducted and found significant results, then the Tukey’s HSD can be computed to find out which specific groups’ means (compared with each other) are different. The test compares all possible pairs of means (Dominguez-Bello *et al.*, 2016).

The Tukey HSD, calculates HSD for each pair of means using the following formula:

$$\text{HSD} = \frac{M_i - M_j}{\sqrt{\frac{MS_w}{n_h}}}$$

Where:

- $|M_i - M_j|$ is the (absolute) difference between the pair of means.
- MS_w is the Mean Square Within, and n is the number in the group or treatment.

The confidence coefficient for the set, when all sample sizes are equal, is exactly $1-\alpha$. For unequal sample sizes, the confidence coefficient is greater than $1-\alpha$. In other words, the Tukey method is conservative when there are unequal sample sizes.

Assumptions for the test:

- Observations are independent within and among groups.
- The groups for each mean in the test are normally distributed.

- There is equal within-group variance across the groups associated with each mean in the test (homogeneity of variance).

The ANOVA test must be performed. Assuming the F value is significant, and then the post hoc test can be computed. If the HSD statistic value for the Tukey test is greater than the critical value, it can be concluded that the two means are significantly different (Dominguez-Bello *et al.,* 2016).

Choosing a statistical model which can be applied to biological data may be inferred by the shape of the relationships between the dependent and the independent instances. This can happen through a graphical representation of these relationships, and in the situation where the shapes happen to be curved, polynomial or other nonlinear models may be more convenient over single regressions (Kleinbaum, Kupper, Nizam, & Rosenberg, 2013). Another study (Goličnik, 2011) has closely linked the choice of a statistical model to the precise question raised in the research. One of these approaches can be seen in studies which estimate the Vmax and Km (concentration of substrates when the reaction reaches half of Vmax) parameters of Michaelis-Menten enzyme kinetics. Enzyme kinetic studies which use the Michaelis - Menten equation relate reaction rate (dependent variable) to substrate concentration (independent variable) in a nonlinear fashion (Goličnik, 2011).

Similarly, when the main purpose of the study is to make predictions from a large dataset with many variables, the correct approach would be to employ a model other than a parametric model. For example, in chemometrics, where outputs are usually predicted by a wide band of wavelengths, the Partial Least Square regression (PLS) approach may be used to infer prediction of a dependent variable from multiple independent variables which may possibly be correlated (Menden *et al.,* 2013; Kelley, Snoek, & Rinn, 2016).

2.12 Machine Learning

This section of the literature review discusses some of the main computational methods for gene and genome analysis, and biological context of proteins in complete genomes, through machine learning. Applications of machine learning to analyse approaches in regulatory genomics and genetic data are outlined. Recurring challenges associated with machine learning analysis are also discussed and a practical guideline is presented for applying machine learning to extract novel and meaningful biological insights of SM genes involved in polyphenol production.

The analysis of biological data often requires rigorous and laborious experimental techniques, and there is a significant cost associated with laboratory work. In the past decade, different computational methods have been developed to analyse biological data on both small and large scale, replacing some of the traditional lab methods used. These different computational methods (Goličnik, 2011; Sochor *et al.,* 2011; Brauer *et al.,* 2012; Fowler *et al.,* 2013; Menden *et al.,* 2013; Wang & Peng, 2014; Eduati *et al.,* 2015; Kelley *et al,* 2016; Kautsar *et al.,* 2017; Hertroijs *et al.,* 2018) have been used to examine biological aspects of protein structure, phylogeny, molecular interactions, and gene expression. The field of machine learning holds promises to enable bioinformaticians and computational biologists to make sense of very large and complex datasets and identify patterns (Menden *et al.,* 2013; Kelley *et al.,* 2016). There are several machine learning applications in bioinformatics that can assist computational biologist in the construction of classification models to characterise new attributes using the previous known attributes (Menden *et al.,* 2013; Kelley *et al.,* 2016).

The advantage of machine learning is its ability to learn functional relationships from data with or without the need to define them a priori (Murphy, 2012; Michalski, Carbonell, & Mitchell, 2013). In machine learning, an algorithm that is developed improves with experience and becomes ‘smarter’ with time

(Murphy, 2012). In bioinformatics, where underlying mechanisms of an instance are inadequately defined or unknown, machine learning promises to derive predictive models without a need for strong assumptions about these underlying mechanisms (Michalski *et al.*, 2013). Several studies have demonstrated the predictive capability of machine learning that has been successfully applied in genomics research (Libbrecht & Noble, 2015; Kelley *et al.*, 2016).

In genomics, machine learning has been used through the development of algorithms that have the capability to learn how to predict the locations or the positions of transcription start sites in a genome sequence (Libbrecht & Noble, 2015; Murphy, 2012; Michalski *et al.*, 2013). The dataset, which includes a collection of true and false transcription start sites, is then fed into the algorithm to build the model (Libbrecht & Noble, 2015). Once the algorithm has learned from the original data set, new annotated sequences are passed through the algorithm and the model can predict which sequences are transcription start sites and which are not (Michalski *et al.*, 2013). In the case that the built model has learned successfully, many or most of the predicted annotated transcription start sites (TSS) for every sequence will be accurate. If not, the outputs must be tested independently in the lab (Libbrecht & Noble, 2015).

Machine learning is a very labor-intensive process, with the typical canonical machine learning workflow consisting of three phases discussed below;

2.12.1 Data Cleaning

Data cleaning, is an essential part of statistical analysis and is the process of altering data in a given storage resource by detecting and correcting corrupt, or inaccurate records from a record set, table, or database (Murphy, 2012). In practice it is often more time-consuming than the statistical analysis itself (Michalski *et al.*, 2013). In machine learning, data cleaning refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the data (Michalski *et al.*, 2013; Kelley *et al.*, 2016). This process helps a machine learning researcher to maintain consistent and accurate datasets by identifying and/or correcting data that may impact on the study results (Menden *et al.*, 2013).

A multitude of existing data cleaning techniques focus on removing data noise which is caused by data objects that are irrelevant or insignificant, but which can significantly hinder most of data type analysis (Murphy, 2012). In bioinformatics, data noise is one of the main data quality challenges that impact the analysis of bioinformatics datasets (Angermueller *et al.*, 2016) and addressing missing or insignificant values is a process that contributes to the data quality in terms of consistency, accuracy, completeness, and cleanness. A study by Angermueller *et al* (2016) on deep learning for computational biology demonstrated that inappropriately addressed data noise can result in low quality data and recommends that data noise be evaluated and corrected prior to any statistical analysis. Even though data noise remains a prevalent problem in bioinformatics datasets, there are continuous studies being undertaken in this field to explore different approaches or techniques of data cleaning that can address this issue (Wald, Khoshgoftaar, & Shanab, 2012).

2.12.2 Data Pre-Processing

Data preprocessing is an important step in the data mining process. It is commonly used as a preliminary data mining technique that involves transforming raw data into an understandable format that can be more easily and effectively processed for the purpose of machine learning. Bioinformatics data is often incomplete (sequence gaps), inconsistent (sequence length), and/or lacking in certain behaviors or trends, and is likely to contain data noise (Wald *et al.*, 2012; Michalski *et al.*, 2013). Data preprocessing is a technique proven to resolve such issues. It involves execution of five critical steps:

1. Data cleaning; (as mentioned above) the process of filling in missing values, smoothing noisy data, identifying or removing outliers and addressing inconsistencies.
2. Data integration; the process of integrating different or multiple data classes or files.
3. Data transformation; the process of data normalisation, aggregation and encoding (for example by applying one-hot coding technique to data attributes).
4. Data reduction; the process of deriving a reduced representation in dimension or volume but producing identical or similar analytical outputs.
5. Data discretisation; the process of discretising numerical data attributes

Different studies show that conducting data preprocessing techniques before analysing the data substantially improves the overall quality of the instances mined and cuts down on the time required for the actual data analysis (Wald *et al.*, 2012; Kelley *et al.*, 2016).

Furthermore, data preprocessing on bioinformatics datasets is of the utmost importance because bioinformatics datasets exhibit high dimensionality, class imbalance and heterogeneousness. High dimensionality (overabundance of attributes) contributes significantly to the challenges of data analysis, producing uncertainty in classification performance and resulting in reduced predictive accuracy of classifiers (Blagus & Lusa, 2012).

High dimensionality in bioinformatics datasets can negatively impact the computational time, as not every feature makes the same contribution to the model. However, class imbalance in machine learning can grossly impact classification performance due to the biasedness of the classes, yielding a very high rate of false negatives. These biases (unequal distribution of instances between classes) can also affect the behavior of some feature selection techniques. Lastly, the heterogeneousness of bioinformatics datasets presents a challenge in learning from the data. For example, a large group of bioinformatics data can include the amino acid sequence of a gene's protein product, which can infer some evolutionary relationships to other proteins across a wide variety of species (Blagus & Lusa, 2012; Eduati *et al.*, 2015).

Such bioinformatics datasets are difficult to mine, since most machine learning and statistical approaches for classification require that the datasets are of the same fixed-length vectors composed of real numbers (Libbrecht & Noble, 2015). This assumption cannot be met in bioinformatics datasets due to the heterogeneity of the amino acid or DNA sequences—the sequences are made up of a string of letters which vary in length. A large number of bioinformatics gene expression datasets possess the aforementioned challenges, rendering the construction of accurate classification models more difficult (Angermueller *et al.*, 2016).

2.12.3 Feature Engineering

Notwithstanding the importance of information in a data set, too much information can lessen the efficiency of data mining. Studies show that not all the attributes assembled for building and testing a model may meaningfully contribute necessary information to the model (Blagus & Lusa, 2012; Khalid, Khalil, & Nasreen, 2014). Too many attributes may indeed weaken the model's accuracy and quality.

In machine learning, feature engineering is the process of compacting attributes into features, starting from an initial set of measured data and building derived values intended to be non-redundant and informative. Feature engineering plays a very crucial role in many areas of data analysis and data processing. Prior to obtaining features, data preprocessing techniques (included in the aforesaid five steps of data preprocessing) concomitant with thresholding, resizing, binarisation, etc. are applied on the sample data, subsequently yielding feature engineering techniques that will be important for model building (Khalid *et al.*, 2014).

Feature engineering transforms the input data (attributes) into a set of features (distinctive properties of input patterns or transformed attributes) that help distinguish the types of input patterns. The important role of feature engineering in data mining relies on its ability to reduce attributes into features that have a linear combination of the original attributes. In Khalid *et al.*,'s (2014) study, models built on extracted features displayed a higher quality because the data is defined in a much smaller number of meaningful attributes.

Bioinformatics datasets are described by many attributes that generally present processing challenges for machine learning algorithms (Menden *et al.,* 2013; Kelley *et al.*, 2016). The attributes of the model represent the dimensions of the processing space used by the algorithm, resulting in higher dimensionality of the processing space. To address the challenge posed by high dimensionality, feature engineering techniques may be applied to process the datasets into a much smaller and richer set of attributes (Khalid *et al.,* 2014).

This can be useful for data visualisation, as a complex dataset can be efficiently visualized when it is reduced to two or three dimensions (Khalid *et al.,* 2014). Dimension reduction is a desirable step in data mining, helping to minimize the effects of noise and attribute correlation (Blagus & Lusa, 2012). Feature engineering techniques are often used in data visualisation, latent semantic analysis, data compression, data decomposition and pattern recognition. In addition, feature engineering enhances the speed and effectiveness of different supervised algorithms (Khalid *et al.,* 2014).

A study in regulatory genomics by Zhou and Troyanskaya (2015) considered predicting chromatin marks from DNA sequence. Features were engineered based on the size of the input sequence window, where larger windows up to one kb were used to capture sequence features at different genomic length scales. Another study used multiple output variables (so-called multitask architectures) as a feature engineering technique to predict multiple chromatin states in parallel (Russakovsky *et al.*, 2015).

Zhang *et al.,*'s (2015) study on deep model-based transfer and multi-task learning for biological image analysis performed feature engineering techniques by transferring model parameters in bioimage analysis. The authors made use of feature engineering techniques on an open corpus from ImageNet (Russakovsky *et al.,* 2015), of more than one million diverse images, to capture rich features at different scales (Xie, Xing, Kong, Su, & Yang, 2015), improving the prediction of Drosophila melanogaster (common fruit fly) developmental stages from situ hybridisation (DNA or RNA) images (Zhang *et al.,* 2015).

These feature engineering techniques allowed learning of shared features between outputs and, in doing so, improved generalisation performance, noticeably decreasing model learning computation cost (Dahl, Jaitly, & Salakhutdinov, 2014).

Best practices (Murphy, 2012) show that irrelevant attributes simply contribute to data noise which results in high computation cost and affects the accuracy of the model. The disadvantages of irrelevant attributes are not only its negative impact on the model but, on the time and resources needed for model building and scoring (Kelley *et al.,* 2016). Since bioinformatics datasets possess many attributes, there is a chance that the datasets may contain groups of attributes that are correlated, creating redundancy. In cases where attributes measure the same underlying feature, feature engineering will discard these groups of attributes to eliminate their presence in the model, preventing the model logic from skewing and influencing the accuracy of the algorithm (Khalid *et al.,* 2014).

2.12.4 Feature Selection

Feature engineering is very different from Feature selection. The former consists of combining attributes into a new reduced set of features or transforming arbitrary data, such as text or images, into numerical features usable for machine learning, while the latter is a machine learning technique applied to features to select the most relevant attributes. In other words, it ranks the existing features according to their predictive significance (Khalid *et al.,* 2014).

The feature selection phase links to the model learning phase. During the feature selection phase (Wald, Khoshgoftaar & Shanab, 2013), the main goal is to obtain measures of information theory that can be used to compute the significance of features. These measures include mutual information (MI), interaction information (II), conditional mutual information (CMI) and joint mutual information (JMI).

In machine learning, feature selection techniques under supervised learning rank the extracted features according to their relevance in predicting or classifying a target (Shanab, Khoshgoftaar, & Wald, 2012). Feature selection techniques become imperative for identifying the most significant predictors of datasets. The objective is to seek the principal features of the datasets that can best represent the datasets (Murphy, 2012). For learning models such as Naïve Bayes or Support Vector Machine, feature relevance is very useful as a preprocessing step in classification modeling. On the other hand, Decision Tree algorithm possesses mechanisms that rank features as part of the model building (Michalski *et al.,* 2013). The output of feature selection is the features of the built data ranked by their measured predictive influence.

Random Forest and Forest of Trees feature selection techniques are an instance of ensemble models. An ensemble model is a model built with some combination of different underlying models. This allows ensemble models to outperform single models because different models may distinguish diverse trends in the data (Shanab, Khoshgoftaar, Wald, & Napolitano, 2012). For this reason, ensemble models tend to minimise the biasedness that single models have to overfit the data. Random Forest and Forest of Trees models assign a significance value to each feature used in the training. Features with higher significance are more influential in building the model, expressing a stronger association with the dependent variable. Feature importance is based on a significance level of 0.05 and this is used as a threshold that can help identify useful features and eliminate features that do not contribute much to the model (Shanab *et al.,* 2012; Khalid *et al.,* 2014).

Although, in information theory, feature selection uses techniques such as, MI, II, CMI, JMI, in practice feature selection in machine learning uses two major techniques: ranker-based techniques and subset-based techniques (Shanab *et al.,* 2012; Wald *et al.,* 2013; Khalid *et al.,* 2014). The former analyses one feature at a time by means of statistical procedures, while the latter analyses complete subsets individually by means of a classifier (wrapper-based feature selection) or statistical procedures (filterbased subset selection).

The ranker-based technique usually requires very little computational power compared to other feature selection techniques, as a feature ranker focuses only on a single score for each feature, while subsets are built based on ranked feature lists (Shanab *et al.,* 2012). Subset-based selection techniques analyse groups of features (subsets) in lieu of each individual feature (Wald *et al.,* 2012). The shortcoming of subset-based selection is that it is computationally expensive, as the computational cost attains O (2^n). O (2^n) is a computational running time of often recursive algorithms that solve a dataset of size N by recursively solving two smaller problems of size N-1 (Shanab *et al.,* 2012).

On the other hand, subset-based selection techniques offer the benefit of capturing highly correlated features in a given set among features (Wald *et al.,* 2013). This ability to detect redundancy among features can help in dimensionality reduction and improve the classification model (Khalid *et al.,* 2014). As such, subset-based selection techniques are more efficient than feature rankers.

Wald *et al.,*'s, 2013 study and Khalid *et al.,*'s, 2014 study elaborate on three different feature selection techniques: filter-based feature ranking, filter-based subset selection, and wrapper-based subset selection. These studies show that filter-based feature ranking evaluates individual features, selecting the highest *N* features. The other two subset selection-based groups (filter-based subset selection, and wrapper-based subset selection), make use of a search approach that explores the space of any possible feature subsets to avoid reaching an O (2^n) computational cost.

In practice, the machine learning researcher decides which input data to provide to the algorithm. This requires prior data source knowledge. A study by Kelley *et al.,* (2016) argues that prior knowledge of the data allows the researcher to confidently decide which dataset features are likely to be significant or not significant. The study argues that the procedure for selecting significant features can be a scientific study in itself to consider.

Fakoor, Ladhak, Nazi, and Huber (2013) present the problem of building a multiclass classifier to differentiate measurements of gene expression among various kinds of cancers. The study shows that the classifier firstly helped to establish precise diagnoses in cases of atypical presentation or histopathology, and secondly, the built learning model helped to perform feature selection through the identification of subsets of genes whose expression patterns have specifically contributed to various kinds of cancers.

Best practices show that it is worthwhile to define the motivations for carrying out feature selections in a specific case. Understanding the task at hand helps to guide the machine learning researcher in selecting the most appropriate feature selection techniques (Murphy, 2012; Michalski *et al.,* 2013).

Some tasks, such as the need to produce a low-cost approach in the identification of a disease phenotype on the merit of the evaluated gene expression levels, may be merely concerned with the identification of a very small set of features that offer the best possible classifier (Fakoor *et al.,* 2013). Other tasks may require a deeper understanding of the underlying biological mechanism (Glaab, Bacardit, Garibaldi, & Krasnogor, 2012). In this case feature selection techniques can be performed with the knowledge of functional annotations or biological pathways that provide insight into the etiology of disease (Glaab *et al.,* 2012).

In even more complex cases, where one needs to train the most accurate possible classifier (Urbanowicz, Granizo-Mackenzie, & Moore, 2012), feature selection techniques can be applied which enable the classifier to identify and eliminate noise or redundancy. The machine learning researcher must be able to select the most appropriate feature selection techniques for the specific task at hand (Fakoor *et al.,* 2013).

In the case where bioinformatics datasets, which include proteomic, epigenomic, genomic or metabolomic data, suffer from high-dimensionality (Urbanowicz *et al.,* 2012; Dahl *et al.,* 2014) due to the growing number of input dimensions (number of data features input to a machine learning classifier), the latter application of feature selection will be deemed most appropriate. However, although this application will improve the data training performance, the shortcoming is the poor generalisation of the model that results, due to the training data being overfitted (Menden *et al.,* 2013). Performing a feature selection method (e.g., principal component analysis) that can project the data from higher to lower dimensions may solve this problem (Buettner *et al.,* 2015).

2.12.5 Model building

Model building in machine learning requires a lot of experimentation and discovery. Building the most relevant model is not always a straightforward task and is often defined by the researcher's understanding of the task and their prior knowledge of the datasets (Glaab *et al.*, 2012; Kelley *et al*, 2016; Hertroijs *et al.*, 2018).

The process takes into consideration different ways of collecting data, processing data, and understanding and discovering features and patterns that deserve the most attention (Sochor *et al.*, 2011). Through this process, a machine learning researcher determines the most appropriate methods for feature selection and tests multiple algorithms in an attempt to answer the questions that are being asked (Eduati *et al.*, 2015). The underlying mechanisms of machine learning model building are the disciplines of statistics, mathematics, information theory, and computer science (Sochor *et al.*, 2011; Murphy, 2012; Michalski *et al.*, 2013; Wang & Peng, 2014; Kautsar *et al.*, 2017).

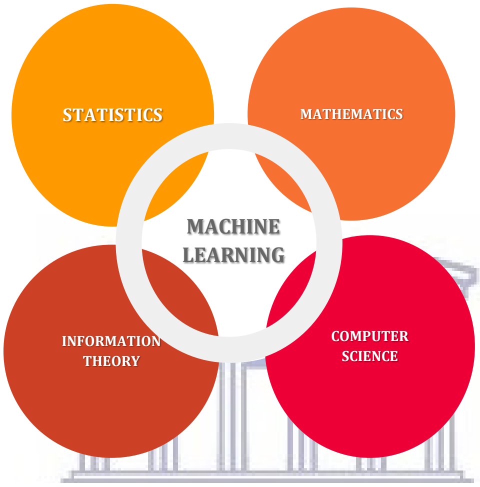


Figure 2. 9 Underlying mechanism of Machine Learning Model Building

Our approach to gaining understanding of secondary metabolite genes is to make use of statistical and mathematical approaches and draw on information theory and computer science concepts to analyse SM gene properties. These disciplines present crucial skills that are important for model building. Chapter 3 elaborates on the scientific methods used in the study, demonstrating how these various disciplines are drawn upon in the process of model building to develop predictions that hold true to test.

A study by Schmidhuber (2015) shows that the building of successful models in machine learning that have the ability to generalize future data, requires thoughtful consideration of the datasets and assumptions about existing training algorithms. This study argues that an appropriate selection and interpretation of assessment criteria are the ultimate fuel to evaluate a machine learning model's quality. Machine learning consists of various algorithms with the power to automate analytical model building (Michalski *et al.*, 2013). These algorithms can iteratively learn from a dataset and assist researchers in discovering hidden insights from a large set of data without being explicitly programmed on where to look (LeCun, Bengio, & Hinton, 2015).

The process of model building includes algorithm design, learning, and testing with the objective to test a hypothesis (Murphy, 2012; Michalski *et al.*, 2013). For example, one bioinformatics study conducted a hypothesis on a specific algorithm (Alipanahi, Delong, Weirauch, & Frey, 2015) that can learn to recognize TSSs, where the algorithm is used as a hypothesis generator. In this case, the algorithm itself hypothesizes that given a set of sequences, sequence X is a TSS. At this point, the key research question would be whether the resulting scientific theory is instantiated in the model produced by the learning algorithm (Libbrecht & Noble, 2015).

2.12.6 Machine Learning Algorithms

Algorithms are fundamental in machine learning and while many machine learning algorithms exist (Murphy, 2012; Michalski *et al.*, 2013), this section will focus on those that are of interest in the current study. Machine learning algorithms can be categorised into three major groups (figure 2.10); supervised learning, unsupervised learning, and reinforcement learning. The sections below will focus on **supervised** and **unsupervised** learning (Michalski *et al.,* 2013).

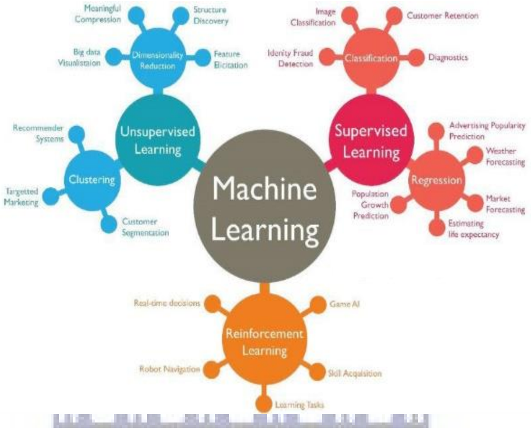


Figure 2. 10 Three Major Groups of Machine Learning Algorithms. Taken from 'Generative adversarial networks for ground penetrating radar in hand held explosive hazard detection,' by C. Veal et al., 2018, *Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XXIII* (Vol. 10628, p. 106280T).

Many studies in bioinformatics model their problems on their interaction with experience, environment or input data (Eickholt & Cheng, 2013; Dahl *et al.*, 2014; Leung, Xiong, Lee, & Frey, 2014; Sønderby & Winther, 2014). For this reason, the learning style of an algorithm is the first consideration, and may be either supervised or unsupervised. These learning styles are the taxonomy of organising machine learning algorithms to purposely think about the roles of the input data and the model preparation process, and the selection of the most relevant algorithm to get the best result for the task at hand (Murphy, 2012; Menden *et al.,* 2013; Michalski *et al.,* 2013).

2.12.6.1 Supervised Learning Algorithms

Supervised learning algorithms build prediction models by means of labeled data to predict either a categorical value or a numerical value (Michalski *et al.,* 2013). Categorical values are obtained through classification models, while numerical values are obtained through regression models (Murphy, 2012). Supervised learning can only be applied when a labeled training set exists (Menden *et al.,* 2013). In their application, supervised machine learning models intend to learn a function, from a list of training pairs for which data are recorded.

The building function of supervised models implies the following:

$$f(x) = y$$

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), \text{etc.}$$

function: = $f(x)$ where $x \in \mathbb{R}^m$, and $y \in \mathbb{R}^n$.

In the case where the desired outcome is continuous, the model will use linear or multi-linear regression and if the required outcome is discrete, the model will use classification, where the outputs are unordered categories (i.e. not numerically meaningful). In both cases (regression and classification) the input data is the training data and follows the assumption that for every $\boldsymbol{x}_i$ there exist a $\boldsymbol{y}_i$. Once the assumption is met, the model is put through a training phase in which it learns from data patterns, begins making predictions, and undergoes correction when those predictions are incorrect (Zhou & Troyanskaya, 2015). The training phase continues until the model (classification or regression) reaches a significant level of accuracy with the training data (figure 2.11).

The images below illustrate the output of a classifier model (classification-based algorithm, figure 2.11 A) which groups the data into classes, and a regression model (regression-based algorithm figure 2.11 B) which maps the data into a linear regression.

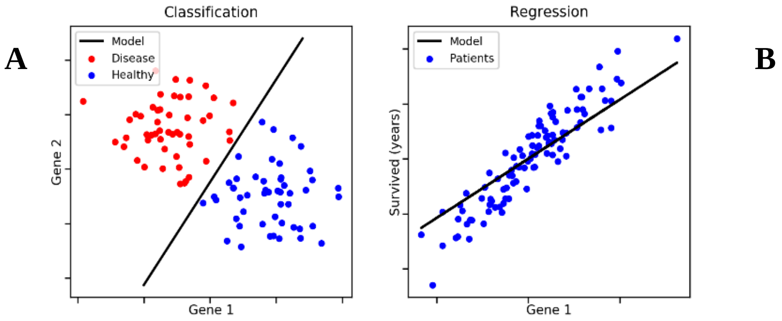


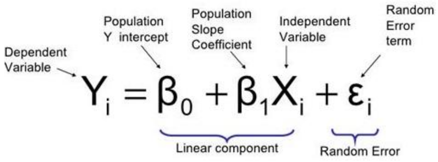
Figure 2. 11 The difference between classification and regression algorithms

Technically, a regression model predicts a numerical output value using the training data (Murphy, 2012). In this case, the data type is made of real or continuous numbers, yielding to a regression problem. In an attempt to predict or forecast a future scenario, the regression model fits a straight line based on the patterns of the data (Michalski *et al*., 2013). Statistically, a linear regression model predicts the variable of interest from single or multiple independent variables by means of a linear mathematical formula (Sochor *et al.*, 2011; Michalski *et al.*, 2013). The regression model can be used to analyse the correlation between independent variables and the dependent variable, and understand the relationship between them (Blagus & Lusa, 2012).

Machine learning can perform various parametric and non-parametric regression analysis techniques (Michalski *et al.*, 2013) where parametric regression models include methods such as linear regression, and least square regression. The regression function is formulated by a finite number of unknown parameters that are derived from the dataset. In the case of non-parametric regression models techniques are applied that permit the regression function to lie in a defined set of functions, which may have infinite dimensions (Michalski *et al.*, 2013).

Linear regression models can be in the form of single or simple linear regression, multi-linear regression, or ordinary least squares (OLS).

1. *Simple Linear Regression model is based on the formula:*



- Where ε is the error term value needed to correct for a prediction error between the observed and predicted value.
- The predictor X is simple, meaning one-dimensional ($X = X_i$). It is assumed to be linear with variance depending on X.

2. *The Multi- Linear Regression model is based on the formula:*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

$$= \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \varepsilon$$

$$= \langle \beta, X \rangle + \varepsilon$$

Where

$$\beta := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad X := \begin{pmatrix} 1 \\ X_1 \\ \vdots \\ X_p \end{pmatrix}$$

- With several predictor variables

$$(X_1, X_2, \dots, X_p:)$$

- And with p + 1 parameters

$$(\beta_0, \beta_1, \dots, \beta_p)$$

Thus, the intercept is handled like any other parameter, for the artificial constant variable$(X_0 \equiv 1)$

- Multiple linear simultaneous equations for a whole given dataset can be represented as,

$$(x_1, y_1) \quad , \dots , \quad (x_n, y_n)$$

$$Y = X\beta + \varepsilon$$

Where

$$Y := \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}, \varepsilon := \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

Multiple linear regression formula is essentially the same as a simple linear regression except that there are multiple coefficients and independent variables.

3. *The Ordinary Least Squares (OLS) Regression model is based on the formula*:

Where $\hat{\beta}$minimize $$\|Y - \hat{Y}\|^2 = \|Y - X\hat{\beta}\|^2$$

- The OLS estimates $\hat{\beta}$ are computed via

$$X^T X\hat{\beta} = X^T Y$$

Statistically, OLS is a technique for estimating the unknown parameters in a given linear regression model. The main concern is to minimize the sum of the differences between the explanatory variables in a given random dataset and the responses predicted by the linear approximation of the data (Hair Jr, Hult, Ringle, & Sarstedt, 2016). In other words, this can be seen (see figure 2.13) as the sum of the vertical distances between each data point in the dataset and the equivalent point on the regression model with the aim to achieve the smallest possible difference to fit the model to the data (Hair Jr *et al.*, 2016).

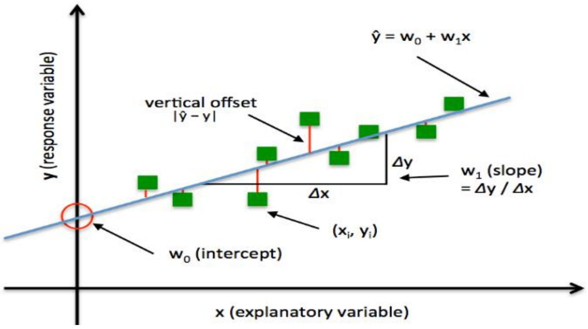


Figure 2. 12 Visually this is seen as the sum of the vertical distances between each data point in the set and the corresponding point on the regression line

For purposes of this study, supervised classification models were used as a focus in order to gain insights into secondary metabolite gene analysis.

Classification models, unlike linear regression models, group the output into classes using the training data (Michalski *et al.,* 2013). In this case the data type is made of discrete, or categorical variables, implying a classification problem. Classification models involve a two-step process (Murphy, 2012; Michalski *et al.,* 2013):

1. *Model construction:* building of a model for a defined set of pre-determined classes. Each data point is meant to belong to a determined class, predefined by the class label instances. The training dataset is used to construct the model, which is represented as classification rules, decision trees, or mathematical formula.
2. *Model usage*: classification of unknown data points until the accuracy is accepted and the model is used to classify data samples whose class labels are unknown. At this stage the accuracy of the model in predicting unknown future attributes can be determined. The known label of the test sample is compared with the classified output from the model, and the percentage of the test set samples that are correctly classified by the model yields the test accuracy rate. Although the test dataset is independent of the training dataset, it is good practice to ensure that both datasets follow the same distribution.

The above two-step process outlines the learning process of the classification model to map each independent variable *x* to one of the predefined class labels *y*. Classification models can be useful for predictive modeling or descriptive modeling (Murphy, 2012; Menden *et al.*, 2013; Kelley *et al.*, 2016). In practice, a classification model is best fit to a dataset with binary or nominal categories, as it does not take into consideration the implicit order among different instances (Menden *et al.*, 2013; Kelley *et al.*, 2016). It is a perfect model for a set of attributes for which order does not matter (e.g., nucleotides, or amino acids).

1. *A classification model is generally based on the Logistic Regression formula:*

$$p = \frac{1}{1 + e^{-y}}$$

Where *y* is equal to a linear regression (single or multiple as seen by above formula **1** and **2**).

Logistic regression prediction is based on a jointly exhaustive and mutually exclusive approach that results in the partition of a set into two classes. The probability of an outcome results in only two values (binary). Logistic regression inputs numerical and categorical variables with the aim to predict the value of a binary variable. Figure 2.13 shows a logistic regression, producing a logistic curve, and a linear regression, producing a linear regression line. There exists a similarity between the logistic regression and linear regression, but the curve of the logistic regression is shaped using the natural logarithm of the response variable in lieu of probability (Murphy, 2012), and is bound by values 0 and 1. The logistic regression does not require the assumptions of equal variances or normal distribution among group attributes (Michalski *et al.*, 2013).

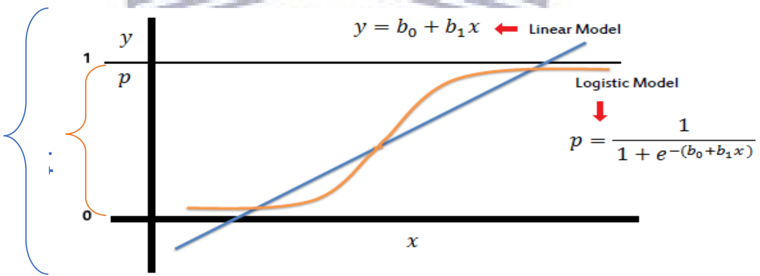


Figure 2. 13 Logistic Model Vs. Regression Model

The constant (b_0) moves the logistic regression curve left and right and the slope (b_1) determines the steepness of the curve. In the case of a logistic regression that involves any number of numerical or categorical variables, the formula can be rewritten as,

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_p x_p)}}$$

The curve function of the logistic regression for classification is also known as the sigmoid function which maps any input *X* between zero and one.

Neural Network

An artificial neural network (ANN) is a statistical learning algorithm with a group of nodes, similar to the vast network of neurons in a brain. In an ANN, each node represents an artificial neuron and an arrow represents a connection from the output of one artificial neuron to the input of another (Sønderby & Winther, 2014; Alipanahi *et al.*, 2015). There are different kinds of ANN, however the single-layer perceptron network (SLPN or SLP) and the multilayer feed-forward neural network (MFNN or MLP) are the most widely used (Zhou & Troyanskaya, 2015; Kelley *et al.*, 2016). SLPN and MFNN are initially trained as any other machine learning algorithms, with input data fed into these neural networks, followed by the learning phase, which is conducted until accurate predictions occur (Bengio, Courville, & Vincent, 2013; Schmidhuber, 2015; LeCun *et al.*, 2015).

1 Single Layer Perceptron Network

Single layer perceptron network (SNLP, often known as perceptron) is seen as analogous to a biological neuron which fires an impulse once the total sum inputs pass the threshold (Dahl *et al.*, 2014; Leung *et al.*, 2014). Perceptron emulates the thresholding behavior by acting as a switch by means of the *classification* sigmoid function (Bengio *et al.*, 2013). Different classification problems that make use of a perceptron algorithm (Logistic regression for classification) set a threshold at the output of the perceptron, which classifies the outputs into two groups (Alipanahi *et al.*, 2015). However, the supervised perceptron can represent both logistic regression (sigmoid function) and linear regression (Zhou & Troyanskaya, 2015; Kelley *et al.*, 2016). The outputs of the perceptron are obtained by a sum of the weighted inputs plus a bias term which, are the parameters that define the learned behavior (Zhou & Troyanskaya, 2015; Kelley *et al.*, 2016).

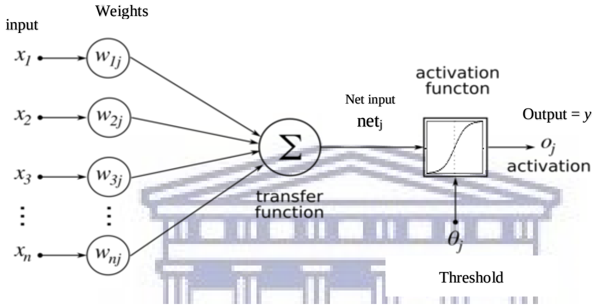


Figure 2. 14 Perceptron Neuron Network Supervised Learning. Modified from ‘A learning rule for very simple universal approximators consisting of a single layer of perceptrons,’ by P. Auer, H. Burgsteiner, & W. Maass, 2008, *Neural Networks, 21*(5).

Figure 2.14 shows the perceptron neuron binary classification with one neuron setting a threshold at the output of the perceptron. Supposing a binary class (yes, and no) problem a threshold (t) could be set to be:

i. T = 0.5

- *a. if y > T: output = yes*
- *b. else output = no*

While the above describes a case of one neuron class, there also exists classification where many neurons (nodes) are put in parallel and each node processes its binary output out of *N* possible classes (Alipanahi *et al.*, 2015; Zhou & Troyanskaya, 2015; Kelley *et al.*, 2016).

2 Multilayer Feed-Forward Neural Network

Previous research has shown that ANNs are capable of solving complex nonlinear tasks. Multilayer Feed-Forward Neural Network (MFNN) is the most applied ANN to model nonlinear systems (LeCun *et al.*, 2015; Kelley *et al.*, 2016) MFNN (figure 24), is made of nodes that are ordered into layers. The first layer (from the left) is called the input layer, the middle layers are called hidden layers, and the last layer is called the output layer (Schmidhuber, 2015). The lines that connect the layers are called weights. These weights control the transfer of signal between nodes through the activation function. During the training, MFNN seeks to determine the optimal value of the weights.

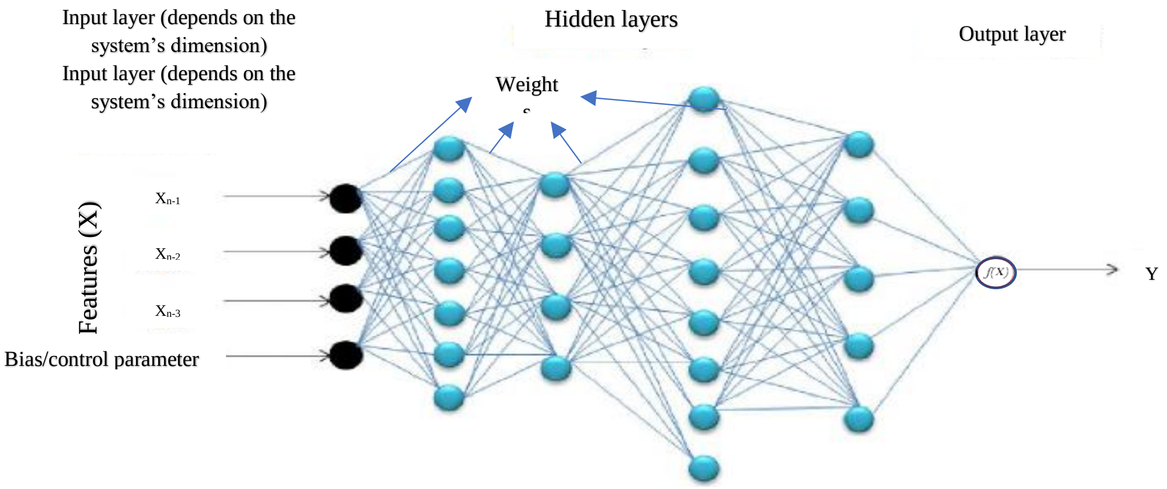

Input layer (depends on the system's dimension)
Input layer (depends on the system's dimension)
Hidden layers
Output layer
Weight s
Features (X)
X n-1
X n-2
X n-3
Bias/control parameter
f(X)
Y

Figure 2. 15 Multiple Feed-forward Neural Network. Modified from 'How transferable are features in deep neural networks?' by J. Yosinski, J. Clune, Y. Bengio, & H. Lipson, 2014, *Advances in Neural Information Processing Systems*

For the activation function *f* (*x*), the input X_{n-1} to node *n*-1 is the weighted sum of the outputs of all nodes connected to it.

A study by Kelley *et al.*, (2016), demonstrated that supervised classification models and gene selection can elicit new and meaningful knowledge from bioinformatics datasets which can be applied in the diagnosis and prognosis of a disease. Other studies demonstrate application of a supervised learning algorithm to select genes based on the nucleotide sequence of a chromosome (Zhou & Troyanskaya, 2015; Alipanahi *et al.*, 2015). The above-mentioned algorithm predicted the locations and detailed intronorexon structure of all the protein-coding genes on the chromosome. A training set (*x*) of labeled DNA sequences, consisting of all the splice sites and the locations of transcription start and termination sites of the gene (TSS and TTS), was required as input for this model. The model was trained to identify the genes based on their general properties such as the DNA sequence pattern near a donor or acceptor splice site, the occurrence of in-frame stop codons within coding exons, and the expected length distribution of 5' and 3' untranslated regions. These gene properties helped the model to identify novel genes that resemble the genes in the training set (Fakoor *et al.*, 2013; Alipanahi *et al.*, 2015; Zhou & Troyanskaya, 2015; Kelley *et al.*, 2016).

Another study (Menden *et al.*, 2013; Eduati *et al.*, 2015), aimed to predict the viability of a cancer cell line when exposed to a chosen drug. The input training set (*x*) was made of features such as somatic sequence variants of the cell line, chemical make-up of the drug and its concentration, and the measured viability (output label *y*), which were used to train classification logistic regression models (support vector machine, random forest classifier etc.). When given a new input test set (x^*), the learnt model predicted its survival (y^*) by computing the functional relationship $f\,(x^*)$.

Support Vector Machine

Support Vector Machine (SVM) classifier classifies each data attribute as a point in n-dimensional space (where n is number of features), where the value of each feature represents a specific coordinate. Classification performance is conducted by determining the hyper-plane which can differentiate two or more classes. In other words, given labeled training data, the classifier outputs an optimal hyperplane which categorizes the test file into the appropriate classes. Technically, in two-dimensional space this hyperplane separates the classes where each class lay in either side. An SVM threshold is often set to 0.5 (Zahiri *et al.*, 2013).

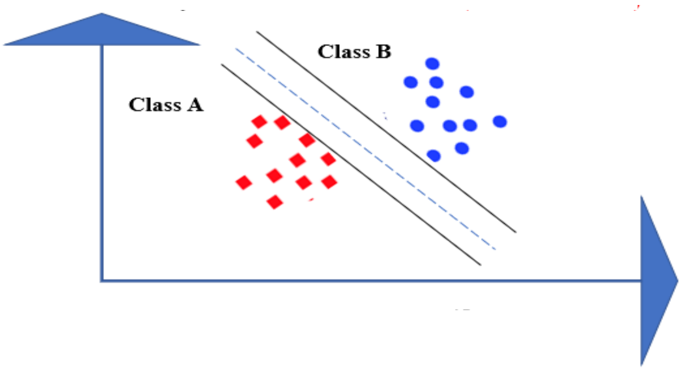


Figure 2.16: shows an SVM trained with samples from two classes with a maximum-margin hyperplane. Samples on the margin are called the support vectors. Support Vector Machine learned the representation of a hyperplane, in this figure illustrated through an enclosed rectangle that best separates the two classes (Zahiri *et al.*, 2013).

Random Forest

Random Forest (RF) is a classifier that is built based on a decision tree to boost the predictive power. This random forest operates by building a multitude of unpruned decision trees at training time and outputting the mode of the classes (classification) of the individual trees (Eduati *et al.*, 2015). Random forest 100 (RF100) has been demonstrated in previous studies as the optimum number of trees and is often used as the default recommended RF (Khoshgoftaar, Golawala, & Van Hulse, 2007). After the trees are constructed, the classifier begins testing each instance passed through each tree, and based on the majority vote of decision trees, the predicted class is thus chosen.

K-Nearest Neighbors

K-nearest neighbor (KNN) classifier is a non-parametric method used for classification and regression, where the input consists of the *k* closest training examples in the feature space built (Menden *et al.*, 2013). The KNN model stores all variable cases and predicts the numerical target based on a similarity measure. The optimal value for *k* is best calculated by first building a cross-validation algorithm to retrospectively determine the best *k* value (Eduati *et al.*, 2015). Alternatively, a grid-search cross-validation (searchgridCV) algorithm can be implemented to obtain the best *k* value (Eduati *et al.*, 2015).

Gaussian Naïve Bayes

Naïve Bayes classifier uses the Bayes theorem which is computationally efficient and easy to interpret when using binary or categorical input values. This classifier is appropriate for problems that contain a normal distribution and are assumed to be conditionally independent given the class label. A Naïve Bayes classifier can be applied to the training data for supervised learning tasks using maximum likelihood (Liu *et al.*, 2012).

2.12.6.2 Unsupervised Learning Algorithms

Unsupervised learning algorithms can be used to discover possible significant, novel, and unknown patterns or associations between covariates or sets of instances using unlabeled data

(Menden et *al.*, 2013). The goal in unsupervised learning is to identify rules (e.g., association rules and clustering) that largely link various covariate values, or cluster data attributes into a selected number of classes in such a way that each class is made of data attributes that are similar (Kelley *et al.*, 2016).

Main approaches to unsupervised learning include: Clustering Analysis (CA), K-means, Hierarchical clustering (HC), Personal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (TSNE). Unsupervised Neural Networks, Anomaly detection, Mixture models, Deep Belief Nets, Autoencoders, Hebbian Learning, and Expectation-maximisation algorithms. In the *Mathematical and Statistical Approaches* Section above CA, HA, PCA were elaborated on and are typical examples of unsupervised algorithms applied to biological data. In the sections below, we briefly focus on two unsupervised learning techniques that are important for our research (PCA and TSNE).

These techniques (PCA and TSNE) are executed through computational models computationally handled and for the sake of a concise literature review we will not dive into the mathematics behind them.

Principal Component Analysis

Before conducting data analysis and inferences, it is often beneficial to develop a visual representation of the dataset to gain a high-level view that can aid in analysis and comprehension. Principal Component Analysis (PCA) is an unsupervised machine learning classifier that takes data of high dimensions and produces several linear cross-correlations of the observed variables that assists in understanding the relationships among data points (Yu *et al*., 2012). Principal Component Analysis is also used for dimensionality reduction of linear cross-correlations among dataset attributes (Menden *et al*., 2013). Furthermore, PCA can be used to compress the data, reducing the number of dimensions without much loss of information. It will then rank the principal components according to their decreasing distributions, revealing patterns in the data (Kelley *et al*., 2016).

PCA is carried out by first calculating the set of orthogonal eigenvectors of the correlation or covariance matrix of the variable components (Vidal, Ma, & Sastry, 2016). The matrix of principal components is the product of the eigenvector matrix with the matrix of independent variables. As a result, the first principal component accounts for the largest representation of the dataset variation, and the second principal component accounts for the second largest representation of the dataset variation, and so on. The objective of principal components is to explain the maximum amount of variance based on fewest numbers of components (Vidal *et al*., 2016).

t-distributed Stochastic Neighbor Embedding:

t-Distributed Stochastic Neighbor Embedding (TSNE) is a nonlinear dimensionality reduction technique that is built to transfer a high-dimensional dataset into a low-dimensional space (2D) for visualisation. This algorithm precisely models each high-dimensional data point into a two-dimensional point in such a way that similar data points are modeled by nearby objects and unrelated data points are modeled by distant objects with high probability (Van der Maaten & Hinton, 2008).

The method of Stochastic Neighbor Embedding (SNE) converts the high-dimensional Euclidean distances between datapoints into conditional probabilities that characterize similarities (Van and Hinton, 2017). The similarity of datapoint x_j to datapoint x_i represents the conditional probability, $p_{j|i}$, in a way x_i would choose x_j as its neighbor if neighbors under a Gaussian centered at x_i were chosen proportionally to their probability density. For neighboring datapoints, $p_{j|i}$ is comparatively high, whereas for widely separated datapoints, $p_{j|i}$ will be miniscule (for reasonable values of the variance of the Gaussian, σ_i). Mathematically, the conditional probability $p_{j|i}$ is given by:

$$1. \qquad p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

The aim of TSNE technique is to optimize and produce significantly better visualizations by reducing the tendency to crowd data points together in the center of the map (Van der Maaten & Hinton, 2008; Van and Hinton, 2017).

2.12.6.3 Model Evaluation Performance Metric

Whether implementing classification or regression modeling techniques, both have the ability to make good predictions (Michalski *et al.*, 2013). Selecting a model and knowing which is the right fit for the training dataset is one thing, while knowing how to generalize the model to an unseen dataset is another (Murphy, 2012; Menden *et al.*, 2013). In circumstances where the selected model fits the training data, it is ideal to ensure that this model does not simply memorize the dataset fed into it, as this will ultimately result in a failure to predict future dataset samples (LeCun *et al.*, 2015; Schmidhuber, 2015; Kelley *et al.*, 2016).

To avoid this biased estimate of the accuracy of the learned model, it is advised to feed a labeled test dataset into the learned model, to evaluate its bias (Murphy, 2012; Menden *et al.*, 2013). In the case where the test dataset gives a less accurate output compared to the trained model, the accuracy estimate of the model is deemed to be biased (Michalski *et al.*, 2013; Schmidhuber, 2015; Kelley *et al.*, 2016).

Thus, evaluating the selection of a learning algorithm that is suitable for the application domain is an essential part of any machine learning project (Vehtari, Gelman, & Gabry, 2017). For the above classification models to give satisfying results, different metrics are used to determine the performance of the machine learning classification and compare the results. The different types of classification performance metrics which are often used include:

1. **Null Accuracy**: Also known as baseline accuracy, this is achieved when the model can consistently predict the predominant class in a dataset.

2. **Classification Accuracy**: This is the ratio of the number of correctly predicted instances to the total number of sample size.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Number of predictions made}}$$

Note: classification accuracy works well only in a case where an equal number of attributes belong to each group.

3. **Log Loss** (a.k.a. Logarithmic Loss): A performance matrix that penalizes the false classifications. In practice it works well for multi-class classification. Using Log Loss, the classifier assigns to each class an accuracy probability for all the samples.

4. **Confusion Matrix**: An evaluation metric implemented through an outputs matrix which describes the complete performance of the model. Confusion matrix has four associated terms (Batista, Prati, & Monard, 2004).

- **True Positives (TP):** The cases where the null hypothesis is an *X* observation and the actual output was also an *X* observation.
- **True Negatives (TN):** The cases where the null hypothesis is Not an *X* observation and the actual output was an *X* observation.
- **False Positives (FP):** The cases where the null hypothesis is an *X* observation and the actual output was Not an *X* observation.
- **False Negatives (FN):** The cases where the null hypothesis is Not an *X* observation and the actual output was also Not an *X* observation.

Classification Accuracy of confusion matrix is given by the relation:

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

However, there are problems with accuracy. It assumes equal costs for both kinds of errors. A 99% accuracy can be excellent, good, mediocre, poor or terrible depending upon the problem.

Recall:
Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (small number of FN).

Recall is given by the relation:

$$\textbf{\textit{Recall}} = \frac{TP}{TP + FN}$$

Precision: To get the value of precision the total number of correctly classified positive examples are divided by the total number of predicted positive examples. High Precision indicates an example labeled as positive is indeed positive (small number of FP). Precision is given by the relation:

$$\textbf{\textit{Precision}} = \frac{TP}{TP + FP}$$

High recall, low precision: means that most of the positive examples are correctly recognized (low FN) but there are a lot of false positives.

Low recall, high precision: This shows that we miss a lot of positive examples (high FN) but those we predict as positive are indeed positive (low FP)

F-measure: Since we have two measures (Precision and Recall) it helps to have a measurement that represents both of them. We calculate an F-measure which uses Harmonic Mean in place of Arithmetic Mean as it punishes the extreme values more. The F-Measure is always nearer to the smaller value of

Precision or Recall (Musicant, Kumar, Ozgur, 2003).

$$\textit{F - measure} = \frac{2\text{*Recall*Precision}}{\text{Recall + Precision}}$$

5. **Area Under the receiver operating characteristics Curve** (AUC): This is a very important and powerful metric for classifier performance, particularly for the binary classification problem. The AUC of the learned model is equal to the probability that the learned model will rank a randomly chosen X observation higher than a randomly chosen Not X observation (Murphy, 2012).
6. **ROC and AUC** are used to see how sensitivity and specificity are affected by various thresholds. ROC helped in choosing a threshold that balances sensitivity and specificity of the classifiers. AUC is a summary of the classifier performance (Vehtari *et al*., 2017).
7. **F-measure**: This calculates the mathematical mean between precision and recall. The domain interval for F-measure lies between [0, 1]. Here, F-measure explains how precise the learned model is in correctly predicting instances, as well as how robust the learned model is (i.e. it does not miss a significant number of instances). As such, the greater the F-measure, the better the performance of our model (Musicant, Kumar, Ozgur, 2003).
8. **Mean Squared Error** (MSE): There exists a similarity between MSE and MAE, with the sole difference being that MSE calculates the average of the square of the difference between the actual outputs and the predicted outputs. MSE presents an advantage in its easiness to compute the gradient, whereas MAE entails a complicated linear programming approach to compute the gradient. As the square of the error is computed, the effect of larger errors becomes more noticeable then smaller errors; hence the MSE becomes a performance metric for the model (Michalski *et al.*, 2013).

2.12.6.4 Model Overfitting

One advantage of model evaluation is its ability to detect model overfitting. During the creation of the model, the objective is to choose the model with the most appropriate hyperparameters (parameters of a prior distribution). These hyperparameters can be grouped into four parameters: regularisation, model size, number of passes and shuffle type (Michalski *et al.*, 2013; Kelley *et al.*, 2016). In certain cases, the best model parameter settings, which produce the most significant predictive accuracy on the training data, can result in overfitting (Michalski *et al.*, 2013). Overfitting gives the appearance of accuracy in the model's prediction, while in reality this is based on the model's capacity to memorize occurring patterns in the training data, meaning it will fail to predict patterns on unknown datasets (Murphy, 2012).

Overfitting can be avoided by selecting an additional dataset for validating the performance of the model. In this case, it is recommended to split the dataset into 60 percent for training, 20 percent for evaluation and 20 percent for validation (Michalski *et al.*, 2013). Once the selected model parameters have been complete and have performed well for the evaluation data, the model can then run a second evaluation on the validation dataset to see how well the learned model performs on the validation dataset. If the learned model reaches the expected threshold on the validation dataset, it can then be deemed to not overfit the data (Murphy, 2012; Michalski *et al.*, 2013).

One downside of this approach is that the splitting of dataset into three sets may result in the omission of relevant data from the training process. In an instance where the data set is small, it is advantageous to perform cross-validation (see below) and allow as much data as possible for training (Menden *et al*., 2013; Leung *et al*., 2014).

2.12.6.5 Cross-validation

Cross-validation is a machine learning technique for evaluating learning models by training the models based on input data subsets (Vehtari *et al*., 2017). The models are evaluated on the complementary subset of the data to detect overfitting. If a model suffers from overfitting, it will fail to generalize a pattern. K-fold cross-validation is a method that applies cross-validation by splitting the input data into k folds (subsets) of data. Subsequently, the model is trained on all folds excluding one-fold, and the model is evaluated on the one-fold that was not used in model training. The process goes on k times with a specific fold kept aside (not included in the training) for evaluation (Vehtari *et al.*, 2017).

2.13 Summary

The biological theory described in the literature serves to provide the context for the study. This chapter of the literature illustrates the biological background for medicinal plants and their health and economical benefits. Medicinal plants are known to be the source of secondary metabolites, and many of these secondary metabolites (SM) are used by humans for flavoring, medicinal and recreational purposes. Many SM are known to exhibit antioxidant, antimicrobial, anticoagulant, anti-inflammatory, antidiabetic, anthelminthic and lipid-lowering properties (Kaul *et al.*, 2017). Four classes (Terpenoids, Alkaloids, Phenolics, and Glycosides) of SM are studied and it was shown that these secondary metabolite classes lead to the production of polyphenols. Four classes (Flavonoids, Phenolic Acids, Lignans, and Stilbenes) of polyphenols are then explored to further understand the effect of polyphenols. Different studies have reported the effects of polyphenols on human health and their ability to treat many diseases (Clauss *et al.,* 2017; Wang *et al.,* 2014; Goetz *et al.,* 2016).

Chalcone synthase or naringenin-chalcone synthase (CHS) has been shown to be a key enzyme in the family of type III polyketide synthase enzymes (PKS) (Shimizu *et al*., 2017). CHS catalyzes the reaction 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA +naringenin chalcone + 3 CO2 and possess a broad spectrum of interesting biological activities such as insecticidal, antioxidative, antibacterial, antiulcer, anticancer, amoebicidal, anthelmintic, antifungal, antitumor, antiprotozoal, antiviral and anti-inflammatory properties (Sun *et al.*, 2015; Gill *et al.*, 2017; Ibdah *et al.*, 2017; Ratnam *et al.*, 2017).

As CHS has been well studied, we therefore use it as an exploratory proof of concept SM gene involved in polyphenol production for our computational analysis. In the case that our computational models are accurate in identifying chalcone synthase, this will have longer term benefits to society, as these models can be used to learn more about genes involved in medicinal compounds.

The mathematic statistical analysis methods are explained as they key focus of the study. Different bioinformatic studies have shown and attempted to analyse biological data and understand the inferential relationship in the biological data. Mathematic statistical approaches have been applied in many biological studies as seen in this section.

Goodness of fit or Chi-square has been used as a statistical model evaluation that can, in the context of model selection, examine the accuracy of the association between categorical variables within a dataset (D'Agostino, 2017). On the other hand, ANOVA test compared the means of a condition between two classes. Similar to chi-square, ANOVA was an omnibus test which reviews the dataset as a whole. However, the ANOVA test and chi-square do not identify where the difference is between the classes. To locate these differences in relationship between the classes, a post-hoc test can then be conducted.

The literature review also highlighted some of the studies that have used machine learning classification models to classify biological contents, and conducted an in-depth investigation of different machine learning models used in bioinformatics studies. Supervised classification models have proven to be of great efficiency in addressing different research problems in bioinformatics. Once these models have been built, model evaluation metrics can be used to determine how well a learned model classifies the output based on a new (unknown) dataset (Menden *et al*., 2013; Leung *et al.*, 2014).

This literature review describes the underlying biological theory and outlines they key computational practices that will be used to address this study's research questions. It has also served to highlight the multiple disciplines that are involved in bioinformatics generally, in particular, giving insight to which methods we can use to address our research problem.

Chapter Three: Research Process

3.1 Contributions

The main contribution of this chapter is to outline best practice and data science techniques that were used in this study. These methods are particularly effective in addressing bioinformatics dataset challenges and are appropriate in their application to machine learning and statistical analysis. In this chapter, a novel computational pipeline figure 3.1 is presented that encompasses all of the processes involved in this study, from ***collecting*** real-world bioinformatics datasets and ***preparing*** them, to ***building*** machine learning classifiers and statistical models.

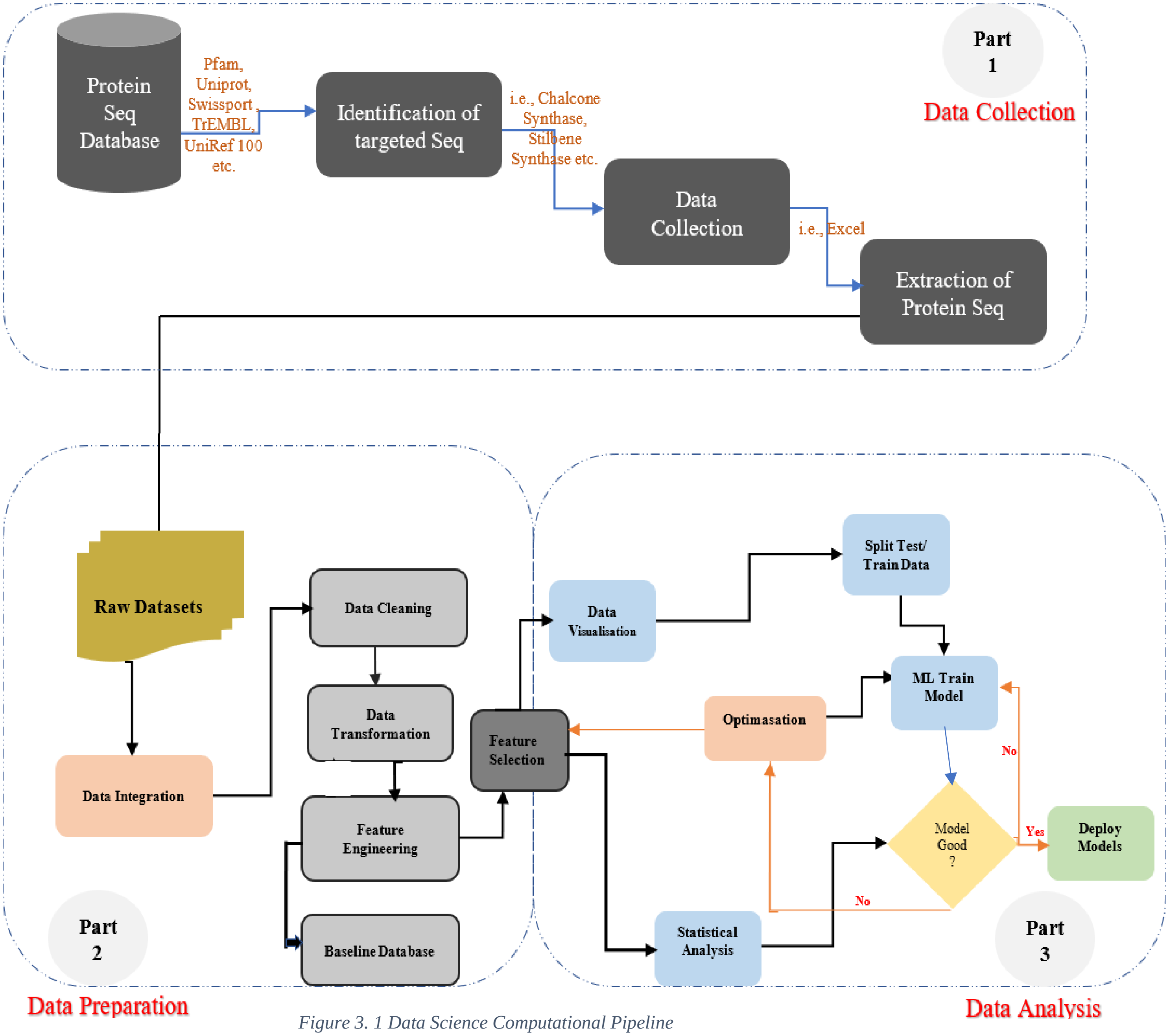


Figure 3. 1 Data Science Computational Pipeline

3.2 Introduction

The work described in the following sections of chapter 3 discusses the steps involved in each part of the computational pipeline. Section 3.3 on **Data Collection** includes a description of the data collection process. Section 3.4 on **Data Preparation** includes the steps that were conducted to ready the data for analysis and includes data integration, data cleaning, data transformation, creation of a baseline database, feature engineering, data standardisation, feature selection, experimental datasets, and data quality. Section 3.5 on **Data Analysis** includes the data visualization processes that were conducted, statistical analysis, and machine learning classification models that were used in the study.

Methods Motivation for exploratory Proof of Concept on Secondary Metabolite genes

3.3 Data Collection

One of the largest repositories of protein sequence data is *UniProtKB*, which is an open access database with large amounts of information derived from research literature. The first step was the identification of one specific enzyme known as chalcone synthase. The identification of the N-terminus and C-terminus domains (pf02797 and pf00195) of chalcone synthase was obtained from the Protein family (Pfam) website: http://pfam.xfam.org/family/. To perform supervised binary classification on machine learning classifiers, reviewed chalcone synthases (RCHS)—protein sequences with known chalcone synthase catalytic activities—were obtained by searching the Swiss-Prot section of UniProtKB using the advanced search options with the terms:

Terms	*Options*
All	*Pf02797 and Pf00195 (Chalcone synthase domains)*
Taxonomy [OC]	*Viridiplantae*
Reviewed>Unreviewed	*Reviewed>*Unreviewed

Subsequently, only the enzymes that catalyze the reaction 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA +naringenin chalcone + 3 CO2 (i.e. chalcone synthase catalytic activity) were selected. 130 RCHS protein sequences (enzymes) were collected and constituted the "true positive set" of the dataset. The "true negative set" (not chalcone synthase (NCHS)) constituted of 130 reviewed protein sequences with known catalytic activities other than RCHS's catalytic activities. These protein sequences were gathered by conducting two different searches.

The first search of the Swiss-Prot section of UniProtKB used the same advanced search options as above with the terms:

Terms	*Options*
Family and Domains > Protein family	*chalcone stilbene synthases family*
Taxonomy [OC]	*Viridiplantae*
Reviewed	*Reviewed*

This search retrieved 69 reviewed protein sequences that were confirmed to be non-chalcone synthase (NCHS).

The second search scanned the Uniref100 section of UniProtKB using the advanced search options with the terms:

Terms	*Option*
UniProt: family	*chalcone stilbene synthases family*
Taxonomy [OC]	*Viridiplantae*
Reviewed *Identity*	*Reviewed* *1.0*

61 reviewed protein sequences that were confirmed to be non-chalcone synthase (NCHS) were gathered from the 198 clusters found in Uniref100.

The 2961 unreviewed chalcone synthases (UCHS) that were used in this study were downloaded from the TREMBL section of the UniProtKB using the same advanced search terms as RCHS. For these enzymes, the catalytic activity was unknown. However, to avoid class imbalances, 130 UCHS were randomly selected to constitute the UCHS class.

These three classes—reviewed chalcone synthase (RCHS) , reviewed non-chalcone synthase (NCHS) and unreviewed chalcone synthase (UCHS)—each with 130 sequences, constituted the balanced experimental datasets from which supervised multiclassification on machine learning classifiers and statistical analysis were performed.

3.4 Data Preparation

3.4.1 Materials (Software and Hardware Specifications)

These software and hardware specifications were used for the implementation of the data preparation and data analysis.

- **Dell Computer**, Intel CORE i5 7th Gen CPU @ 2.50CHz x 4, 8 GiB, 64-bit
- **x86_64** – Machine architecture
- **x86_64** – Processor architecture
- **x86_64** – Operating system architecture
- **GNU or Linux** – Operating system
- **Linux 4.15**.0-29-generic (buildd@lcy01-amd64-024) (gcc version 5.4.0 20160609 (Ubuntu 5.4.0-6ubuntu1~16.04.10))
- **Linux Kernel** 4.15.0.29 (PC operating system)
- **Anaconda:** A data science Python distribution pre-loaded with all the most popular libraries and tools. Some of the biggest Python libraries wrapped up in Anaconda include NumPy, Pandas and Matplotlib, though the full list is over 1000 packages.
- **Jupyter Notebooks** (www.jupyter.org): Open source web application installed in Anaconda that allows for creation and sharing of live coding.

- **Jupyter kernel**: A dependency of Jupyter Notebook, which is responsible for handling various types of request (code execution, code completions, inspection), and providing a reply.
- **Python 3.6.1** | Anaconda 4.4.0 (64 bit) | Gcc 4.4.7 20120313 (Red Hat 4.4.7-1)

3.4.2 Data Integration

During data collection, the data of each class (RCHS, NCHS, UCHS) was downloaded separately in Microsoft Excel, and then integrated into one worksheet. This file was then transformed into a CSV file to form datasets for supervised binary classifications (RCHS and NCHS), supervised multiclassification and statistical models (RCHS, NCHS, and UCHS).

3.4.3 Data Cleaning

To address data noise, which may result from the presence of ambiguous amino acid letters or non-alphabetical characters ({, }, ‘,”,/, * etc.), a Python script was written to check each protein sequence and remove any unnecessary characters. To ensure correct input data logging and avoid random error and attribute noise during subsequent analyses, Python scripts were written that verified correct classification of the sequences into the three defined classes (RCHS, NCHS and UCHS). The completeness of sequence numbers and sequence labelling was also checked. The data cleaning pseudocode is explained further in Chapter 4, and the full code is available in Appendix B.

3.4.4 Data Transformation and Feature Engineering

Data transformation and feature engineering were performed simultaneously, with a view to transforming protein sequences into distinctive properties of input patterns (features) that addresses the issue of protein sequence length. The features were engineered in order to address sequence length and to convert protein sequences, which are categorical data, into numerical data. This contributed to converting the data into a format that is more conducive for feature engineering in machine learning and statistical analysis. Three novel feature sets were developed by writing Python scripts with Biopython libraries to engineer these feature sets:

1. Frequency-based features
2. Value-based features
3. Amino acid relative frequency feature

The *first* feature set (Frequency-based features) includes four features which were extracted from the protein sequences through the identification of the amino acids involved in: Aromaticity, Beta-Sheet, Alpha-Helix and Turn.

The *second* feature set (Value-based features), which was derived from specific amino acids as well, include: Entropy, Protein-Stability, Protein-GRAVY, and Protein-Isoelectric-Point.

The *third* feature set (Amino acid relative frequency feature) consists of the twenty amino acids. That is, each amino acid is defined as unique feature, and its empirical frequency is then calculated.

The rationale and explanation of these three feature sets is explained further in Chapter 4, see Section 4.2. The pseudocode for engineering these three feature sets is also presented in Chapter 4, and the full code is available in Appendix B.

3.4.5 Baseline Dataset

The three transformed feature datasets, along with the original raw datasets that were captured directly from various sections of UniProtKB, make up the baseline dataset. This dataset was compiled and saved as a Microsoft Excel workbook to ensure that all of the data collected for this study, and their derivative datasets, are preserved and made available for future study. The complete baseline dataset is available in Appendix B.

3.4.6 Data Standardisation

Data standardisation, since the frequency-based feature set and the amino acid relative frequency feature set are both proportional features, whereas the value-based feature set is value based, the features were standardised to center the data values around 0 using the standardisation mathematical formula :

$$x_{new} = \frac{x - \mu}{\sigma}$$

Where x_{new} is the standardised new dataset, x the observation in the old dataset, μ is the mean of old the dataset and *sigma* the standard deviation of the old dataset.

3.4.7 Feature Selection

To address the significance level of features among the engineered features of two sets (frequency-based feature and value-based feature), different feature selection techniques were used to rank these features based on their predictive significance. Feature selection is also implemented in this study to counter high dimensionality that biological datasets present and to find a ‘minimum relevant feature’ from these two feature sets to enter the model.

Different mathematical and statistical techniques used during the implementation of feature selection include:

- Principal Component Analysis (PCA): Used to compute the feature dataset and produce two principal components that are used to build the models.
- Scatterplot Matrix and Spearman’s Correlation Matrix: Used to pinpoint the correlation scores between features.
- Analysis of Variance (ANOVA): Computes the degree of linear dependency between two random features.
- Mutual Information (MI): Used to capture any kind of statistical dependency between features.
- Stats Test Standard Deviation (Std): Used to select feature relevance based on Std score.
- Histogram Technique: Used to select features based on their frequency and distributions.
- Chi-square: Used to select features with the highest values of the Chi-squared statistical test.
- Random Forest and Forest of Trees: Used to evaluate the significance of each of the features through a classification task.

More on the feature selection methods is found in chapters four and five.

3.4.8 Experimental Data Quality

To provide a measurement for data quality, Area Under the Curve (AUC) was calculated for each dataset using the numbers of sequences per class. To ensure a unique scalability of the engineered features in these datasets, data standardisation was conducted. Lastly, to ensure data fitness, the data collection (Section 3.3) and data preparation steps (Section 3.4) contributed to data quality in terms of accuracy, completeness, consistency, integrity, reasonability, timeliness, uniqueness, validity, and accessibility.

3.5 Data Analysis

3.5.1 Data Visualisation

Mathematical and statistical computational methods were used in this study to visualise the datasets and their features in such a way that information, analytics, patterns, trends and correlations could be clearly demonstrated, and include:

- t-distributed Stochastic Neighbor Embedding (TSNE)
- Principal Component Analysis (PCA)
- Regression Analysis (RA)
- Mutual Information (MI)
- Anova (F-test)
- Boxplot
- Histogram
- Forest of Trees
- Scatterplot and Spearman’s Correlation

3.5.2 Supervised Machine Learning Classifiers

The supervised classifiers built in this study are dependent on the same data collection and data preparation fundamentals cited in the above Sections 3.3 and 3.4. Eight supervised classification models (Binary and multiclassification algorithms) were implemented in this study, which were all referenced in the literature review in Chapter 2: Logistic Regression (LR), Decision Tree (DT), Random Forest with 100 trees (RF100), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Single Perceptron (SLP), and Multilayer Perceptron (MLP). These supervised classifiers were chosen due to their prevalence in the literature.

Further rationales and descriptions of the implementation of these above methods, procedures, and features are illustrated in more detail in Chapters 6.

3.5.3 Statistical Analysis

Data collection (Section 3.3) and data preparation (Section 3.4), also serve as fundamentals of the statistical analysis conducted in this study. Five mathematic statistical techniques: one-way analysis of variance (**ANOVA)**, **Tukey's range test, Chi-square test,** post hoc test **Bonferroni,** and **Boxplot** were implemented in this study. The assumptions and formulas of these techniques are elaborated in Chapter 7.

3.6 Limitations of the study

The main limitations of the study are laid out below.

- **Sample size** – our sample size is made up of 390 protein sequences, because only 130 Reviewed Chalcone Synthase (RCHS) and 130 reviewed Not Chalcone Synthase could be properly identified as true positive and true negative datasets respectively. To ensure that all of the datasets were balanced, 130 sequences from a batch of 2961 Unreviewed Chalcone Synthase were selected at random. The total sample size of 390 sequences may not be considered a large enough sample size to ensure a representative distribution of the plant enzyme population, given that the total population size of plant enzymes is unknown. This study therefore presents an exploratory proof of concept model. We suggest that future studies could be conducted with a larger sample size of true positive and true negative data if they become available, to explore what further significant relationships from the data could be found.
- **Lack of prior research studies on the topic** – various studies have served as the basis of the literature review and while these studies may form the foundation for framing the research problem under investigation, a study on data science techniques (machine learning and computational statistical models) mining chalcone synthase has not been previously reported (to our knowledge). Therefore, we present an *exploratory* rather than an *explanatory* proof of concept research design.

Chapter Four: Feature Engineering – Implementation and Result

4. Introduction

This chapter discusses the data collection processes implemented in the current study. A proof of concept for data transformation in combined with feature engineering techniques is presented. Data transformation, which results in engineered features, while tedious, is essential for building a statistical or machine learning model, particularly in cases of protein sequence analysis tasks (Qu, Yu, Gong, Xu, & Lee, 2017). Choosing appropriate data transformation techniques can assist in accelerate the mining and analysis of bioinformatics datasets (Qu *et al*., 2017; Angermueller *et al.*, 2016).

4.1 Data Collection

Section 3.3 in Chapter 3 showed the steps that were followed to collect all of the datasets for Reviewed Chalcone Synthase (RCHS), Not Chalcone Synthase (NCHS), and Unreviewed Chalcone Synthase (UCHS) protein sequences.

To construct proof of concept binary classification models, the true positive dataset (RCHS) consisting of 130 curated protein sequences with known chalcone synthase catalytic activities (3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO2. 130) was used in combination with the true negative dataset. The true negative dataset (NCHS) consisted of 130 different curated plant protein sequences confirmed to not exhibit any chalcone synthase catalytic activity, as can be seen in figure 4.1.

To construct proof of concept multi classification models and statistical models, the same true positive and true negative datasets were used. The third dataset was constructed from Unreviewed Chalcone Synthase (UCHS) and consisted of 130 non-curated protein sequences (with unknown chalcone synthase activities) selected at random from a set of 2961 protein sequences.

4.1.1 Plant Protein Sequence Data Resources

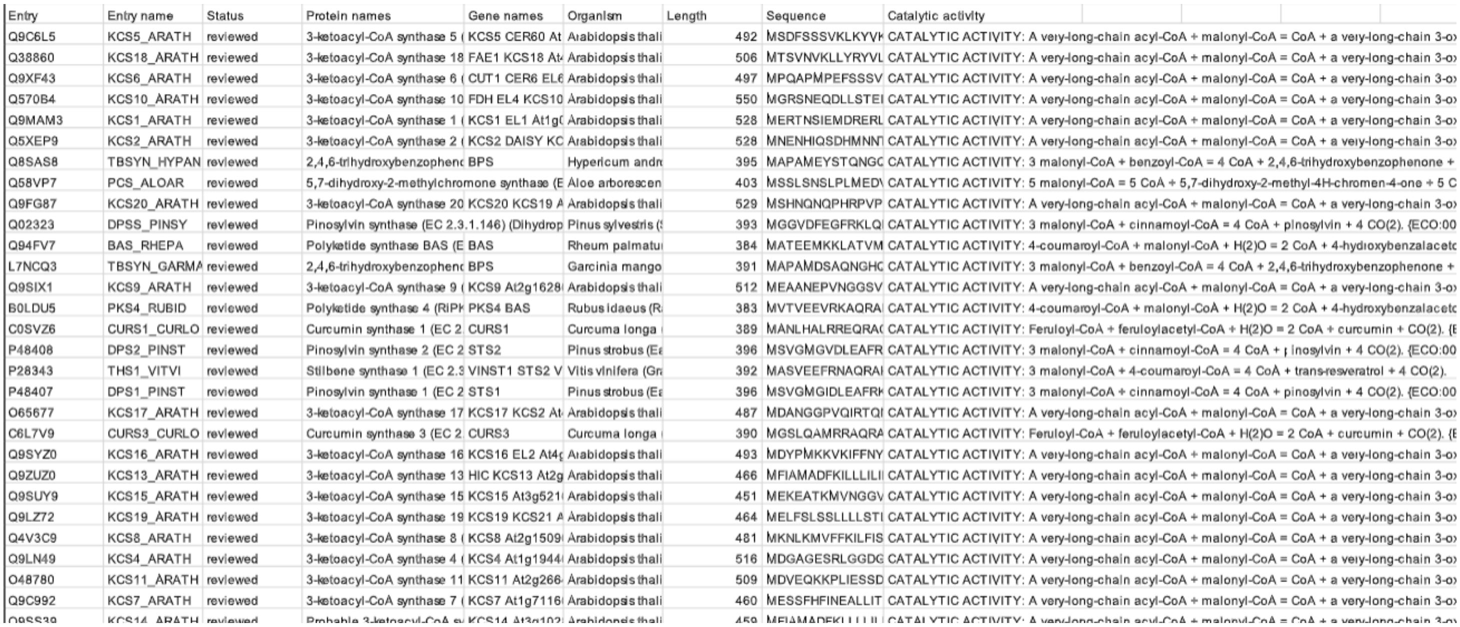
As mentioned in Chapter 3, Section 3.3, the data for this study was collected from *UniProtKB*, an open access database of protein sequence data curated from various studies. Data on plant protein sequences was collected from the following *UniProtKB* database sections:

Table 4.1 provides a summary of the various UniProtKB database sections that were used, along with direct hyperlinks.

Table 4. 1 Data Collection Resources

UniProtKB Database Section	Description
	.
Swiss-Prot	**Manually annotated.** Records with information extracted from literature and curator-evaluated computational analysis. database section stores manually curated (reviewed) protein sequences. https://www.uniprot.org/uniprot/
TREMBL	**Computationally analyzed** Records that await full manual annotation. database section stores non-curated (unreviewed) protein sequences. https://www.uniprot.org/uniprot/
UniRef100	combines identical sequences and sub-fragments with 11 or more residues from any organism into a single UniRef entry. database section stores both reviewed and unreviewed protein sequences https://www.uniprot.org/uniref

A preview of the raw datasets (taken from the true negative, Not Chalcone Synthase dataset) can be seen above, in figure 4.1. The column '*Length'* indicates the number of amino acids present in each protein sequence. Here, we note the significant differences in the length of the protein sequences, which can pose challenges for handling the data (Qu *et al*., 2017; Angermueller *et al.*, 2016). To address the challenge that presented by the difference in sequence length of each protein sequence, data transformation and feature engineering techniques were applied to resolve the discrepancies in sequence lengths (Qu *et al*., 2017; Angermueller *et al*., 2016). The column '*Status'* indicates whether the sequences were manually curated (reviewed). The column '*Sequence* shows the entries of different protein sequences that constitute the various genes from each dataset. The column '*Catalytic activity*' shows the enzyme reactions that the protein sequences catalyze. The column '*Entry'* simply presents the unique identifier of each row entry in UniProtKB. The column '*Protein names'* shows the names of each protein sequence, and the '*Gene names'* column shows the name of each gene. Finally, the column '*Organism'* shows the name of the plant organism from which the proteins originate. Each of these dataset attributes is also included in both the ***true positive dataset*** (Reviewed Chalcone Synthase) and the Unreviewed Chalcone Synthase dataset as seen in Appendix B.



Entry	Entry name	Status	Protein names	Gene names	Organism	Length	Sequence	Catalytic activity
Q9C6L5	KCS5_ARATH	reviewed	3-ketoacyl-CoA synthase 5 (	KCS5 CER60 At	Arabidopsis thali	492	MSDFSSSVKLKYVK	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
Q38860	KCS18_ARATH	reviewed	3-ketoacyl-CoA synthase 18	FAE1 KCS18 At	Arabidopsis thali	506	MTSVNVKLLYRYVL	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
Q9XF43	KCS6_ARATH	reviewed	3-ketoacyl-CoA synthase 6 (	CUT1 CER6 EL6	Arabidopsis thali	497	MPQAPMPEFSSSV	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
Q570B4	KCS10_ARATH	reviewed	3-ketoacyl-CoA synthase 10	FDH EL4 KCS10	Arabidopsis thali	550	MGRSNEQDLLSTEI	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
Q9MAM3	KCS1_ARATH	reviewed	3-ketoacyl-CoA synthase 1 (	KCS1 EL1 At1g0	Arabidopsis thali	528	MERTNSIEMDRERL	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
Q5XEP9	KCS2_ARATH	reviewed	3-ketoacyl-CoA synthase 2 (	KCS2 DAISY KC	Arabidopsis thali	528	MNENHIQSDHMNNT	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
Q8SAS8	TBSYN_HYPAN	reviewed	2,4,6-trihydroxybenzophenc	BPS	Hypericum andr	395	MAPAMEYSTQNGC	CATALYTIC ACTIVITY: 3 malonyl-CoA + benzoyl-CoA = 4 CoA + 2,4,6-trihydroxybenzophenone +
Q58VP7	PCS_ALOAR	reviewed	5,7-dihydroxy-2-methylchromone synthase (E		Aloe arborescen	403	MSSLSNSLPLMEDV	CATALYTIC ACTIVITY: 5 malonyl-CoA = 5 CoA + 5,7-dihydroxy-2-methyl-4H-chromen-4-one + 5 C
Q9FG87	KCS20_ARATH	reviewed	3-ketoacyl-CoA synthase 20	KCS20 KCS19 A	Arabidopsis thali	529	MSHNQNQPHRPVP	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
Q02323	DPSS_PINSY	reviewed	Pinosylvin synthase (EC 2.3.1.146) (Dihydrop		Pinus sylvestris (S	393	MGGVDFEGFRKLQI	CATALYTIC ACTIVITY: 3 malonyl-CoA + cinnamoyl-CoA = 4 CoA + pinosylvin + 4 CO(2). {ECO:00
Q94FV7	BAS_RHEPA	reviewed	Polyketide synthase BAS (E	BAS	Rheum palmatu	384	MATEEMKKLATVM	CATALYTIC ACTIVITY: 4-coumaroyl-CoA + malonyl-CoA + H(2)O = 2 CoA + 4-hydroxybenzalaceto
L7NCQ3	TBSYN_GARMA	reviewed	2,4,6-trihydroxybenzophenc	BPS	Garcinia mango	391	MAPAMDSAQNGHC	CATALYTIC ACTIVITY: 3 malonyl-CoA + benzoyl-CoA = 4 CoA + 2,4,6-trihydroxybenzophenone +
Q9SIX1	KCS9_ARATH	reviewed	3-ketoacyl-CoA synthase 9 (	KCS9 At2g1628	Arabidopsis thali	512	MEAANEPVNGGSV	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
B0LDU5	PKS4_RUBID	reviewed	Polyketide synthase 4 (RiPK	PKS4 BAS	Rubus idaeus (Ra	383	MVTVEEVRKAQRAI	CATALYTIC ACTIVITY: 4-coumaroyl-CoA + malonyl-CoA + H(2)O = 2 CoA + 4-hydroxybenzalaceto
C0SVZ6	CURS1_CURLO	reviewed	Curcumin synthase 1 (EC 2.	CURS1	Curcuma longa (	389	MANLHALRREQRAC	CATALYTIC ACTIVITY: Feruloyl-CoA + feruloylacetyl-CoA + H(2)O = 2 CoA + curcumin + CO(2). {E
P48408	DPS2_PINST	reviewed	Pinosylvin synthase 2 (EC 2	STS2	Pinus strobus (Ea	396	MSVGMGVDLEAFR	CATALYTIC ACTIVITY: 3 malonyl-CoA + cinnamoyl-CoA = 4 CoA + pinosylvin + 4 CO(2). {ECO:00
P28343	THS1_VITVI	reviewed	Stilbene synthase 1 (EC 2.3	VINST1 STS2 V	Vitis vinifera (Gr	392	MASVEEFRNAQRAI	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
P48407	DPS1_PINST	reviewed	Pinosylvin synthase 1 (EC 2	STS1	Pinus strobus (Ea	396	MSVGMGIDLEAFRK	CATALYTIC ACTIVITY: 3 malonyl-CoA + cinnamoyl-CoA = 4 CoA + pinosylvin + 4 CO(2). {ECO:00
O65677	KCS17_ARATH	reviewed	3-ketoacyl-CoA synthase 17	KCS17 KCS2 At	Arabidopsis thali	487	MDANGGPVQIRTQI	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
C6L7V9	CURS3_CURLO	reviewed	Curcumin synthase 3 (EC 2.	CURS3	Curcuma longa (	390	MGSLQAMRRAQRA	CATALYTIC ACTIVITY: Feruloyl-CoA + feruloylacetyl-CoA + H(2)O = 2 CoA + curcumin + CO(2). {E
Q9SYZ0	KCS16_ARATH	reviewed	3-ketoacyl-CoA synthase 16	KCS16 EL2 At4g	Arabidopsis thali	493	MDYPMKKVKIFFNY	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
Q9ZUZ0	KCS13_ARATH	reviewed	3-ketoacyl-CoA synthase 13	HIC KCS13 At2g	Arabidopsis thali	466	MFIAMADFKILLLILI	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
Q9SUY9	KCS15_ARATH	reviewed	3-ketoacyl-CoA synthase 15	KCS15 At3g521	Arabidopsis thali	451	MEKEATKMVNGGV	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
Q9LZ72	KCS19_ARATH	reviewed	3-ketoacyl-CoA synthase 19	KCS19 KCS21 A	Arabidopsis thali	464	MELFSLSSLLLLSTI	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
Q4V3C9	KCS8_ARATH	reviewed	3-ketoacyl-CoA synthase 8 (	KCS8 At2g1509	Arabidopsis thali	481	MKNLKMVFFKILFIS	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
Q9LN49	KCS4_ARATH	reviewed	3-ketoacyl-CoA synthase 4 (	KCS4 At1g1944	Arabidopsis thali	516	MDGAGESRLGGDG	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
O48780	KCS11_ARATH	reviewed	3-ketoacyl-CoA synthase 11	KCS11 At2g266	Arabidopsis thali	509	MDVEQKKPLIESSD	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
Q9C992	KCS7_ARATH	reviewed	3-ketoacyl-CoA synthase 7 (	KCS7 At1g7116	Arabidopsis thali	460	MESSFHFINEALLIT	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox
Q9SS39	KCS14_ARATH	reviewed	Probable 3-ketoacyl-CoA sy	KCS14 At3g102	Arabidopsis thali	459	MFIAMADFKLLLIL	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-ox

Figure 4. 1 Data Collection-- Different curated plant protein sequences not having any chalcone synthase catalytic activity (True Negative control set), Excel compilation derived from information from the database

All datasets used in the construction of our proof of concept models were downloaded from these sections of UniProtKB as explained in Section 3.3, in 2018. More information on all three protein sequence datasets (RCHS, NCHS, and UCHS) are found in Appendix B.

4.2 Data Preparation Implementation and Discussion

4.2.1 Data Integration

Three classes of RCHS, NCHS, UCHS, each made of 130 sequences, constituted our ***balanced*** sample size. To build binary classification models, the true positive (RCHS) and true negative (NCHS) datasets were integrated to form one dataset. To build the multiclassification and statistical models, all three datasets were integrated to form one dataset, as illustrated in figure 4.2.

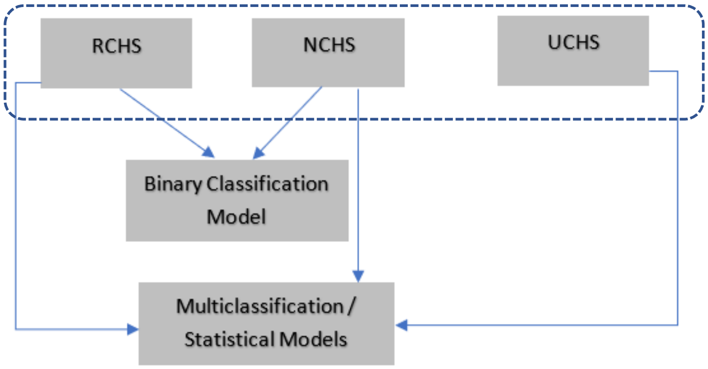

RCHS
NCHS
UCHS
Binary Classification Model
Multiclassification / Statistical Models

Figure 4. 2 Integration of Three classes for Model's building

Two datasets (binary class and multiclass) were therefore produced and used for the analysis in the current study.

4.2.2 Data Cleaning Implementation

Cleaning of datasets was performed by implementing cleaning steps on the multiclass dataset, followed by deriving the binary class from this, as all classes were included in the multiclass dataset. This was performed so that each of the two datasets would not have to be cleaned individually as, depicted in figure 4.2.

The Python pseudocode below shows the implementation of the data cleaning process as explained in Section 3.4.3. The prerequisite *Biopython libraries* are very important, as they speed up the process of writing code by providing API, libraries and functions.

```
# PREREQUISITES:
import csv
import sequtils
import sys
from Bio.SeqUtils.ProtParam import ProteinAnalysis, ProtParamData, IsoelectricPoint
from Bio.Seq import Seq,
from Bio.Alphabet import IUPAC
from Bio.Data import IUPACData

# CLEANING
Function to check (sequence, Amino_Acids = "MVATGSLDQRPNCEFHIKWY"):
    for AminoAcids in sequence:
        AminoAcids = AminoAcids.rstrip().remove('any letters and characters not in Amino_Acids')
        if AminoAcids not in AminoAcids:
            print(AminoAcids),and return False and True
# INPUT
Open the dataset as csv file
with open('path input file') as csvfile:
    Protein Sequence = csv.DictReader(csvfile)
    for row in  Protein sequence:
       dataset = Call the check function((row['Amino_Acid_Sequence']))
              Cleaned_protein_sequences = ProteinAnalysis(dataset)
 #OUTPUT
       Write the output of the program as an csv (tab delimited) via stdout

# close the current CSV file
        FILE_HANDLE.close()
```

4.2.3 Data Transformation and Feature Engineering Implementation and Discussion

This section discusses the implementation of data transformation and feature engineering, as referenced in Section 3.4.4 of Chapter 3, to address sequence length and high dimensionality of the datasets, and to engineer features by converting protein sequences into numerical data (Libbrecht & Noble, 2015).

The objective for the data transformation and feature engineering implemented in this study was to enable transformation of a set of sequences into a set of engineered features (Libbrecht & Noble, 2015). The raw datasets are used as an input into the feature engineering model with the resulting output being transformed data that is presented as the engineered features (Figure 4.4). The three feature sets that were developed in this study are explained further below.

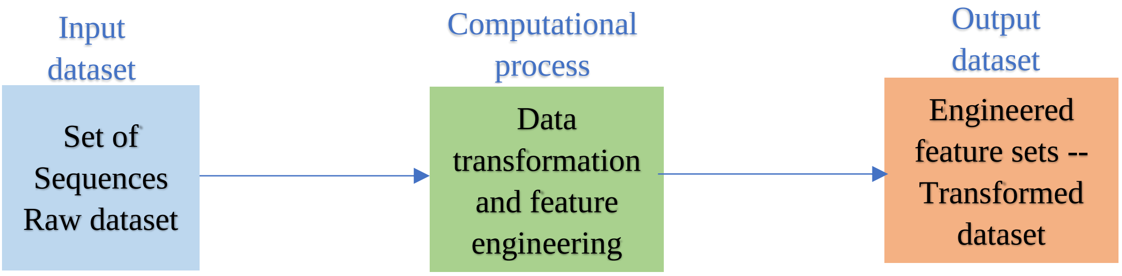

Input dataset
Set of Sequences Raw dataset
Computational process
Data transformation and feature engineering
Output dataset
Engineered feature sets -- Transformed dataset

Figure 4. 3 Data transformation and Feature engineering process

In light of the shortcomings of feature engineering methods identified from previous studies, elaborated in Sections 2.12.3 and 2.12.3.1 of the literature review: data conversion, which results in feature engineering, is very tedious work but essential for building a statistical or machine learning model in most cases of protein sequence classification tasks. Different methods, such as n-gram approaches, physiochemical properties-based extraction approaches, and homology-based approaches, have been introduced in previous studies. Even though these methods work well in many cases, their resource-intensive nature poses a practical challenge.

Three computational approaches on protein sequence data transformation are proposed as exploratory proof of concept for feature engineering:

1. Frequency-based features
2. Value-based features
3. Amino acid relative frequency feature

The features sets were computationally calculated using the Biopython libraries. The documentation with all necessary resources can be found at: https://biopython.org/wiki/Documentation

and http://biopython.org/DIST/docs/api/Bio.SeqUtils.ProtParam.ProteinAnalysis-class.html.

A brief description of pseudocode to engineer the features is presented in the box below. Feature engineering is performed with the assumption that data cleaning has been conducted.

```
# PREREQUISITES:
import csv
import sequtils
import sys
from Bio.SeqUtils.ProtParam import ProteinAnalysis, ProtParamData,
IsoelectricPoint
from Bio.Seq import Seq,
from Bio.Alphabet import IUPAC
from Bio.Data import IUPACData

# CLEANING
All cleaning steps are executed
# INPUT
Open the dataset as csv file
with open('path input file') as csvfile:
    Protein Sequence = csv.DictReader(csvfile)
    for row in  Protein sequence:
       dataset = Call the check function((row['Amino_Acid_Sequence']))
              Cleaned_protein_sequences = ProteinAnalysis(dataset)
# ENGINEERED FEATURES
       Frequency-based features : From Cleaned_protein_sequences
       compute(Aromaticity(), Beta-Sheet(), Alpha-Helix(), Turn(), using built-
       in functions from the # PREREQUISITES.
       Amino acid relative-frequency feature : From Cleaned_protein_sequences
       compute relative frequency of (M,V,A,T,G,S,L,D,Q,R,P,N,C,E,F,H,I,K,W,Y)
       by getting amino acids percent using the # PREREQUISITES
       Value-based features: From Cleaned_protein_sequences compute(Protein-
       Molecular-Weight(), Entropy(), Protein-Stability(), Protein-GRAVY(), and
       Protein-Isoelectric-Point()) using built-in functions from the #
       PREREQUISITES.
 #OUTPUT
       Write the output of the program as an csv (tab delimited) via stdout
       The columns outputted as follows:

        Frequency-based features : AromaFeature | AlHydroFeature | SulphFeature
       | AcidicFeature | BasicFeature | BetaFeature | AlphaFeature |
       TurnFeature | and AliphaticFeature
       Amino acid relative-frequency feature :
       M|V|A|T|G|S|L|D|Q|R|P|N|C|E|F|H|I|K|W|Y|
       Value-based features : | Entropy_Feature |Protein-
       Stability_Feature|Protein-GRAVY_Feature | Isoelectric-Point_Feature
# Close the current CSV file
```

The First Feature Set

The *first* feature set (Frequency-based features) includes four engineered features from the protein sequence amino acids that, according to biological literature, are known to be involved in Aromaticity, Sheet, Helix, and Turn. It is to be noted that the engineered features (i.e., regular structures; beta-sheet, alpha-helix, Turn) are not meant to reflect their exact biological realities (secondary structures). Rather, the features were engineered solely based on the presence of Amino Acids that are known to be involved in these biological properties, without considering the position or the sequencing of the Amino Acids within the protein. This approach served to segment different amino acids into computational features intended to be non-redundant and statistically informative (Zhou and Troyanskaya, 2015). It can be noted that mathematically, the relative frequency of these amino acids involved in these biological properties can be segmented from each sequence, hence, the name frequency-based features. By doing so we can prevent the model logic from skewing and influencing the accuracy of the statistical or machine learning models (Khalid *et al.*, 2014).

Table 4. 2 First Set of Features --Frequency-based features

Biological Properties	Amino Acids Present (their Alphabets)	Name of Engineered Features
Aromaticity	*FWY*	*AromaFeature Calculate the aromaticity according to Lobry, 1994.*
Alpha-Helix	*VIYFWL*	*HeliFeature*
Turn	*NPGS*	*TurFeature*
Beta-Sheet	*EMAL*	*SheeFeature*

The Second Feature Set

The *second* feature set (Value-based features), was engineered by computing the numerical value of four protein features, which are not directly derived from specific amino acids. These protein features include: Entropy, Protein-Stability, Protein-GRAVY, and Protein-Isoelectric-Point.

Entropy is associated with the number of conformations of a molecule. Protein stability refers to the thermodynamic stability of a protein, often in terms of whether the protein is in its native, folded state, or its unfolded state. Protein-GRAVY is a calculation of the grand average of hydropathy of a protein sequence, and the protein-isoelectric point is the pH at which the net electrical charge of protein is neutral.

Isoelectric point same with GRAVY is a measure of hydrophobicity in a protein and is related to the amino acids and their properties.

All of these features represent specific biological properties of the protein sequences, each of which hold a numerical value. These value-based features were computationally computed from each protein sequence using the Biopython methods.

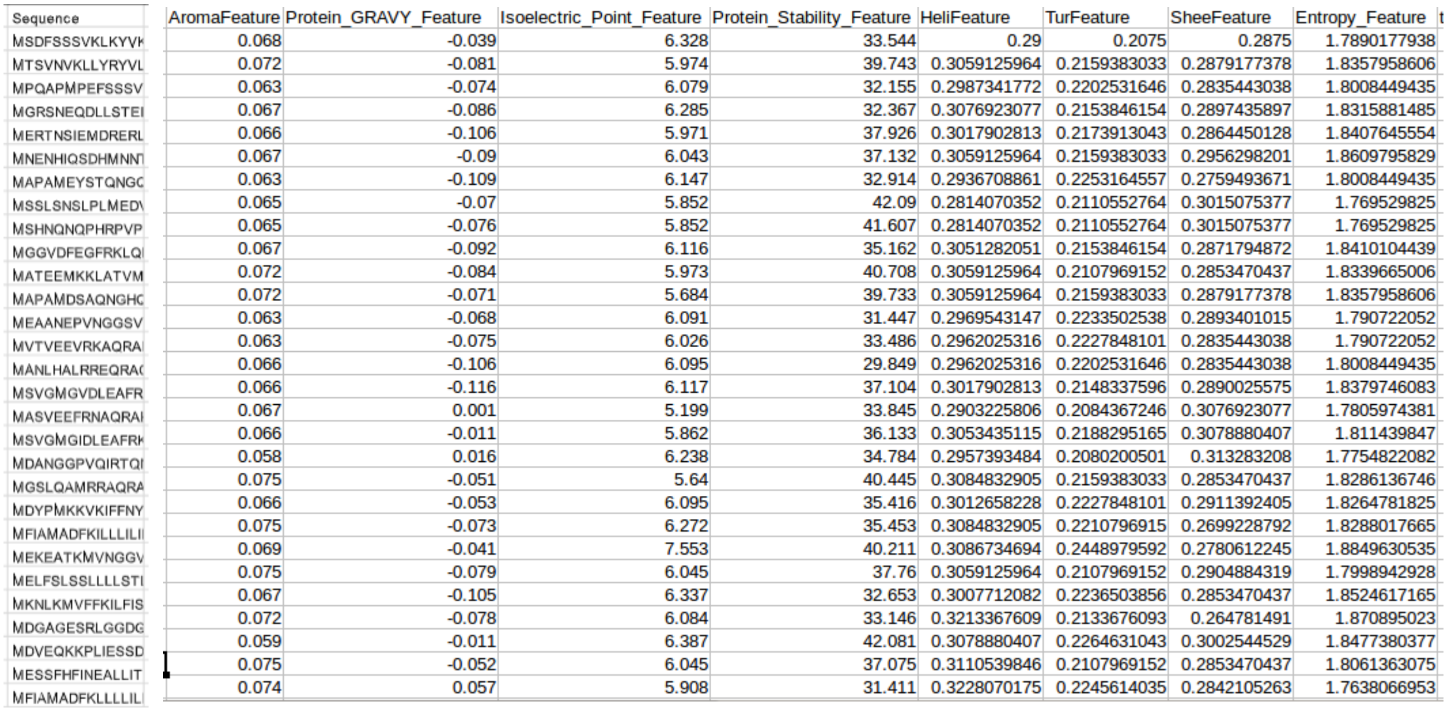
Table 4.3 below lists the four functions and the feature names that were used to label each feature.

Table 4. 3 Second Set of Features --Value-based features

Biological Properties	Name of Engineered Features
Entropy	*Entropy_Feature*
Protein Stability	*Protein Stability_Feature Calculate the instability index according to Guruprasad et al 1990.*
Protein GRAVY	*Protein GRAVY_Feature Calculate the gravy according to Kyte and Doolittle, 1982.*
Protein Isoelectric point	*Isoelectric point_Feature*

These value-based features were computationally computed on each protein sequence using the Biopython methods.

A synopsis of the two feature sets (Frequency-based features, and Value-based features), were integrated to form one dataset with all eight features (*AromaFeature*, *Portein_Gravy_Feature*, *Isoelectric_Point_Feature*, *Protein_Stability_Feature*, *HeliFeature*, *TurFeature*, *SheeFeature*, *Entropy_Feature*) which can be seen in figure 4.5.



Sequence	AromaFeature	Protein_GRAVY_Feature	Isoelectric_Point_Feature	Protein_Stability_Feature	HeliFeature	TurFeature	SheeFeature	Entropy_Feature	t
MSDFSSSVKLKYVK	0.068	-0.039	6.328	33.544	0.29	0.2075	0.2875	1.7890177938	
MTSVNVKLLYRYVL	0.072	-0.081	5.974	39.743	0.3059125964	0.2159383033	0.2879177378	1.8357958606	
MPQAPMPEFSSSV	0.063	-0.074	6.079	32.155	0.2987341772	0.2202531646	0.2835443038	1.8008449435	
MGRSNEQDLLSTEI	0.067	-0.086	6.285	32.367	0.3076923077	0.2153846154	0.2897435897	1.8315881485	
MERTNSIEMDRERL	0.066	-0.106	5.971	37.926	0.3017902813	0.2173913043	0.2864450128	1.8407645554	
MNENHIQSDHMNNT	0.067	-0.09	6.043	37.132	0.3059125964	0.2159383033	0.2956298201	1.8609795829	
MAPAMEYSTQNGC	0.063	-0.109	6.147	32.914	0.2936708861	0.2253164557	0.2759493671	1.8008449435	
MSSLSNSLPLMEDV	0.065	-0.07	5.852	42.09	0.2814070352	0.2110552764	0.3015075377	1.769529825	
MSHNQNQPHRPVP	0.065	-0.076	5.852	41.607	0.2814070352	0.2110552764	0.3015075377	1.769529825	
MGGVDFEGFRKLQI	0.067	-0.092	6.116	35.162	0.3051282051	0.2153846154	0.2871794872	1.8410104439	
MATEEMKKLATVM	0.072	-0.084	5.973	40.708	0.3059125964	0.2107969152	0.2853470437	1.8339665006	
MAPAMDSAQNGHC	0.072	-0.071	5.684	39.733	0.3059125964	0.2159383033	0.2879177378	1.8357958606	
MEAANEPVNGGSV	0.063	-0.068	6.091	31.447	0.2969543147	0.2233502538	0.2893401015	1.790722052	
MVTVEEVRKAQRAI	0.063	-0.075	6.026	33.486	0.2962025316	0.2227848101	0.2835443038	1.790722052	
MANLHALRREQRAC	0.066	-0.106	6.095	29.849	0.2962025316	0.2202531646	0.2835443038	1.8008449435	
MSVGMGVDLEAFR	0.066	-0.116	6.117	37.104	0.3017902813	0.2148337596	0.2890025575	1.8379746083	
MASVEEFRNAQRAI	0.067	0.001	5.199	33.845	0.2903225806	0.2084367246	0.3076923077	1.7805974381	
MSVGMGIDLEAFRK	0.066	-0.011	5.862	36.133	0.3053435115	0.2188295165	0.3078880407	1.811439847	
MDANGGPVQIRTQI	0.058	0.016	6.238	34.784	0.2957393484	0.2080200501	0.313283208	1.7754822082	
MGSLQAMRRAQRA	0.075	-0.051	5.64	40.445	0.3084832905	0.2159383033	0.2853470437	1.8286136746	
MDYPMKKVKIFFNY	0.066	-0.053	6.095	35.416	0.3012658228	0.2227848101	0.2911392405	1.8264781825	
MFIAMADFKILLLILI	0.075	-0.073	6.272	35.453	0.3084832905	0.2210796915	0.2699228792	1.8288017665	
MEKEATKMVNGGV	0.069	-0.041	7.553	40.211	0.3086734694	0.2448979592	0.2780612245	1.8849630535	
MELFSLSSLLLLSTI	0.075	-0.079	6.045	37.76	0.3059125964	0.2107969152	0.2904884319	1.7998942928	
MKNLKMVFFKILFIS	0.067	-0.105	6.337	32.653	0.3007712082	0.2236503856	0.2853470437	1.8524617165	
MDGAGESRLGGDG	0.072	-0.078	6.084	33.146	0.3213367609	0.2133676093	0.264781491	1.870895023	
MDVEQKKPLIESSD	0.059	-0.011	6.387	42.081	0.3078880407	0.2264631043	0.3002544529	1.8477380377	
MESSFHFINEALLIT	0.075	-0.052	6.045	37.075	0.3110539846	0.2107969152	0.2853470437	1.8061363075	
MFIAMADFKLLLLILI	0.074	0.057	5.908	31.411	0.3228070175	0.2245614035	0.2842105263	1.7638066953	

Figure 4. 4 A synopsis of the 2 feature (Frequency-based feature and Value-based feature)

The Third Feature Set

The *third* feature set (Amino acid relative frequency feature) is made up of the twenty amino acids. and was engineered by computing the relative frequency of each amino acid in each protein sequence as seen in figure 4.6.

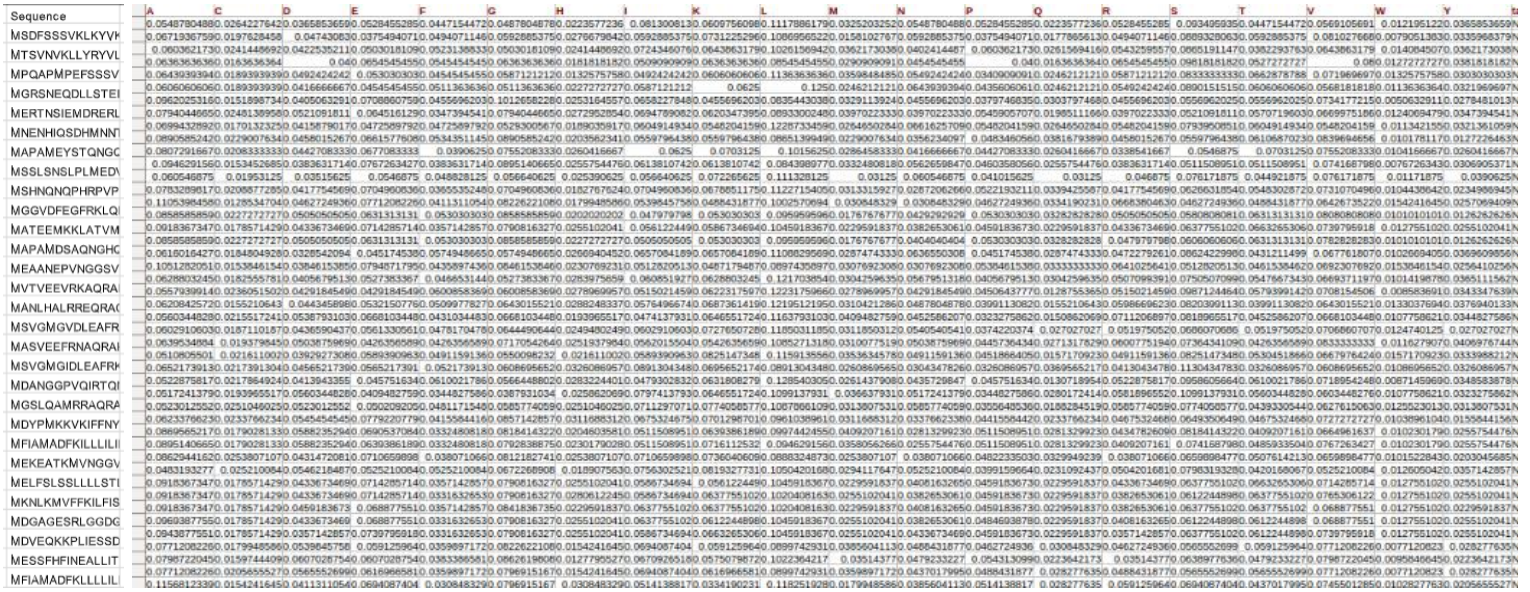
Note: there are actually 22 amino acids. While the other two they are very rare, one of them occurs in Eukaryotes. "Selenocysteine (Sec) and pyrrolysine (Pyl) are rare amino acids that are cotranslationally inserted into proteins and known as the 21st and 22nd amino acids in the genetic code". Our Python code checked for these two amino acids in the whole datasets and none of them were found. Hence we built the feature based on 20 amino acids.

In this way, each of the twenty essential amino acids are engineered as a single feature. All twenty amino acids and their one letter codes are listed in Table 4.4.

Table 4. 4 Twenty Amino Acids and their letter code

Amino Acids	One Letter Code	Amino Acids	One Letter Code
Alanine	A	glutamine	Q
Arginine	R	glutamic acid	E
Cysteine	C	glycine	G
Lysine	K	histidine	H
Methionine	M	isoleucine	I
Phenylalanine	F	leucine	L
Proline	P	Serine	S
Threonine	T	tryptophan	W
Tyrosine	Y	valine	V
aspartic acid	D	asparagine	N

Figure 4.6 shows a synopsis of the final result when a set of sequences are transformed into a set of amino acid relative frequency features. The complete dataset of the amino acid relative frequency features is included in **Appendix B.**



Sequence
MSDFSSSVKLKYVK
MTSVNVKLLYRYVL
MPQAPMPEFSSSV
MGRSNEQDLLSTEI
MERTNSIEMDRERL
MNENHIQSDHMNNT
MAPAMEYSTQNGC
MSSLSNSLPLMEDV
MSHNQNQPHRPVP
MGGVDFEGFRKLQI
MATEEMKKLATVM
MAPAMDSAQNGHC
MEAANEPVNGGSV
MVTVEEVRKAQRAI
MANLHALRREQRAC
MSVGMGVDLEAFR
MASVEEFRNAQRAI
MSVGMGIDLEAFRK
MDANGGPVQIRTQI
MGSLQAMRRAQRA
MDYPMKKVKIFFNY
MFIAMADFKILLLILI
MEKEATKMVNGGV
MELFSLSSLLLLSTI
MKNLKMVFFKILFIS
MDGAGESRLGGDG
MDVEQKKPLIESSD
MESSFHFINEALLIT
MFIAMADFKLLLLIL

A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	ta
0.0548780488	0.0264227642	0.0365853659	0.0528455285	0.0447154472	0.0487804878	0.0223577236	0.081300813	0.0609756098	0.1117886179	0.0325203252	0.0548780488	0.0528455285	0.0223577236	0.0528455285	0.093495935	0.0447154472	0.0569105691	0.012195122	0.0365853659	N
0.0671936759	0.0197628458	0.04743083	0.0375494071	0.0494071146	0.0592885375	0.0276679842	0.0592885375	0.0731225296	0.1086956522	0.0158102767	0.0592885375	0.0375494071	0.0177865613	0.0494071146	0.0889328063	0.0592885375	0.081027668	0.0079051383	0.0335968379	N
0.060362173	0.0241448692	0.0422535211	0.0503018109	0.0523138833	0.0503018109	0.0241448692	0.0724346076	0.0643863179	0.1026156942	0.0362173038	0.0402414487	0.060362173	0.0261569416	0.0543259557	0.0865191147	0.0382293763	0.0643863179	0.014084507	0.0362173038	N
0.0636363636	0.0163636364	0.04	0.0654545455	0.0545454545	0.0636363636	0.0181818182	0.0509090909	0.0636363636	0.0854545455	0.0290909091	0.0454545455	0.04	0.0163636364	0.0654545455	0.0981818182	0.0527272727	0.08	0.0127272727	0.0381818182	N
0.0643939394	0.0189393939	0.0492424242	0.053030303	0.0454545455	0.0587121212	0.0132575758	0.0492424242	0.0606060606	0.1136363636	0.0359848485	0.0549242424	0.0340909091	0.0246212121	0.0587121212	0.0833333333	0.0662878788	0.071969697	0.0132575758	0.0303030303	N
0.0606060606	0.0189393939	0.0416666667	0.0454545455	0.0511363636	0.0511363636	0.0227272727	0.0587121212	0.0625	0.125	0.0246212121	0.0643939394	0.0435606061	0.0246212121	0.0549242424	0.0890151515	0.0606060606	0.0568181818	0.0113636364	0.0321969697	N
0.0962025316	0.0151898734	0.0405063291	0.0708860759	0.0455696203	0.1012658228	0.0253164557	0.0658227848	0.0455696203	0.0835443038	0.0329113924	0.0455696203	0.0379746835	0.0303797468	0.0455696203	0.0556962025	0.0556962025	0.0734177215	0.0050632911	0.0278481013	N
0.0794044665	0.0248138958	0.0521091811	0.064516129	0.0347394541	0.0794044665	0.0272952854	0.0694789082	0.0620347395	0.0893300248	0.0397022333	0.0397022333	0.0545905707	0.0198511166	0.0397022333	0.0521091811	0.0570719603	0.0669975186	0.0124069479	0.0347394541	N
0.0699432892	0.0170132325	0.0415879017	0.0472589792	0.0472589792	0.0529300567	0.0189035917	0.0604914934	0.0548204159	0.1228733459	0.0264650284	0.0661625709	0.0548204159	0.0264650284	0.0548204159	0.0793950851	0.0604914934	0.054820416	0.011342155	0.0321361059	N
0.0890585242	0.0229007634	0.0458015267	0.0661577608	0.0534351145	0.0890585242	0.0203562341	0.0559796438	0.0559796438	0.0865139949	0.0229007634	0.0356234097	0.048346056	0.0381679389	0.0458015267	0.0559796438	0.0610687023	0.0839694656	0.010178117	0.0127226463	N
0.0807291667	0.0208333333	0.0442708333	0.0677083333	0.0390625	0.0755208333	0.0260416667	0.0625	0.0703125	0.1015625	0.0286458333	0.0416666667	0.0442708333	0.0260416667	0.0338541667	0.0546875	0.0703125	0.0755208333	0.0104166667	0.0260416667	N
0.094629156	0.0153452685	0.0383631714	0.0767263427	0.0383631714	0.0895140665	0.0255754476	0.0613810742	0.0613810742	0.084398977	0.0332480818	0.0562659847	0.0460358056	0.0255754476	0.0383631714	0.0511508951	0.0511508951	0.074168798	0.0076726343	0.0306905371	N
0.060546875	0.01953125	0.03515625	0.0546875	0.048828125	0.056640625	0.025390625	0.056640625	0.072265625	0.111328125	0.03125	0.060546875	0.041015625	0.03125	0.046875	0.076171875	0.044921875	0.076171875	0.01171875	0.0390625	N
0.0783289817	0.0208877285	0.0417754569	0.0704960836	0.0365535248	0.0704960836	0.0182767624	0.0704960836	0.0678851175	0.1122715405	0.0313315927	0.0287206266	0.0522193211	0.0339425587	0.0417754569	0.0626631854	0.0548302872	0.0731070496	0.0104438642	0.0234986945	N
0.1105398458	0.0128534704	0.0462724936	0.0771208226	0.0411311054	0.0822622108	0.0179948586	0.0539845758	0.0488431877	0.1002570694	0.030848329	0.030848329	0.0462724936	0.0334190231	0.0668380463	0.0462724936	0.0488431877	0.0642673522	0.0154241645	0.0257069409	N
0.0858585859	0.0227272727	0.0505050505	0.0631313131	0.053030303	0.0858585859	0.0202020202	0.047979798	0.053030303	0.095959596	0.0176767677	0.0429292929	0.053030303	0.0328282828	0.0505050505	0.0580808081	0.0631313131	0.0808080808	0.0101010101	0.0126262626	N
0.0918367347	0.0178571429	0.0433673469	0.0714285714	0.0357142857	0.0790816327	0.0255102041	0.056122449	0.0586734694	0.1045918367	0.0229591837	0.0382653061	0.0459183673	0.0229591837	0.0433673469	0.0637755102	0.0663265306	0.0739795918	0.012755102	0.0255102041	N
0.0858585859	0.0227272727	0.0505050505	0.0631313131	0.053030303	0.0858585859	0.0227272727	0.0505050505	0.053030303	0.095959596	0.0176767677	0.0404040404	0.053030303	0.0328282828	0.047979798	0.0606060606	0.0631313131	0.0782828283	0.0101010101	0.0126262626	N
0.0616016427	0.0184804928	0.0328542094	0.045174538	0.0574948665	0.0574948665	0.0266940452	0.0657084189	0.0657084189	0.1108829569	0.0287474333	0.0636550308	0.045174538	0.0287474333	0.0472279261	0.0862422998	0.0431211499	0.067761807	0.0102669405	0.0369609856	N
0.1051282051	0.0153846154	0.0384615385	0.0794871795	0.0435897436	0.0846153846	0.0230769231	0.0512820513	0.0487179487	0.0897435897	0.0307692308	0.0307692308	0.0538461538	0.0333333333	0.0641025641	0.0512820513	0.0461538462	0.0692307692	0.0153846154	0.0256410256	N
0.0628803245	0.0182555781	0.0405679513	0.0527383367	0.046653144	0.0527383367	0.0283975659	0.060851927	0.0628803245	0.121703854	0.0304259635	0.0567951318	0.0405679513	0.0304259635	0.0507099391	0.0750507099	0.0547667343	0.0669371197	0.0101419878	0.0365111562	N
0.0557939914	0.0236051502	0.0429184549	0.0429184549	0.0600858369	0.0600858369	0.0278969957	0.0515021459	0.0622317597	0.1223175966	0.0278969957	0.0429184549	0.0450643777	0.0128755365	0.0515021459	0.0987124464	0.0579399142	0.0708154506	0.008583691	0.0343347639	N
0.0620842572	0.0155210643	0.044345898	0.0532150776	0.0509977827	0.0643015521	0.0288248337	0.0576496674	0.0687361419	0.1219512195	0.0310421286	0.0487804878	0.0399113082	0.0155210643	0.0598669623	0.0820399113	0.0399113082	0.0643015521	0.0133037694	0.0376940133	N
0.0560344828	0.0215517241	0.0538793103	0.0668103448	0.0431034483	0.0668103448	0.0193965517	0.0474137931	0.0646551724	0.1163793103	0.0409482759	0.0452586207	0.0323275862	0.0150862069	0.0711206897	0.0818965517	0.0452586207	0.0668103448	0.0107758621	0.0344827586	N
0.0602910603	0.0187110187	0.0436590437	0.0561330561	0.0478170478	0.0644490644	0.0249480249	0.0602910603	0.0727650728	0.1185031185	0.0311850312	0.0540540541	0.0374220374	0.027027027	0.051975052	0.0686070686	0.051975052	0.0706860707	0.0124740125	0.027027027	N
0.0639534884	0.019379845	0.0503875969	0.0426356589	0.0426356589	0.0717054264	0.0251937984	0.0562015504	0.0542635659	0.1085271318	0.0310077519	0.0503875969	0.0445736434	0.0271317829	0.0600775194	0.0736434109	0.0426356589	0.0833333333	0.011627907	0.0406976744	N
0.0510805501	0.021611002	0.0392927308	0.0589390963	0.0491159136	0.0550098232	0.021611002	0.0589390963	0.0825147348	0.115913556	0.0353634578	0.0491159136	0.0451866405	0.0157170923	0.0491159136	0.0825147348	0.0530451866	0.0667976424	0.0157170923	0.0333988212	N
0.0652173913	0.0217391304	0.0456521739	0.0565217391	0.052173913	0.0608695652	0.0326086957	0.0891304348	0.0695652174	0.0891304348	0.0260869565	0.0304347826	0.0326086957	0.0369565217	0.0413043478	0.1130434783	0.0326086957	0.0608695652	0.0108695652	0.0326086957	N
0.0522875817	0.0217864924	0.0413943355	0.045751634	0.0610021786	0.0566448802	0.0283224401	0.0479302832	0.0631808279	0.128540305	0.0261437908	0.0435729847	0.045751634	0.0130718954	0.0522875817	0.0958605664	0.0610021786	0.0718954248	0.0087145969	0.0348583878	N
0.0517241379	0.0193965517	0.0560344828	0.0409482759	0.0344827586	0.0387931034	0.025862069	0.0797413793	0.0646551724	0.1099137931	0.036637931	0.0517241379	0.0344827586	0.0280172414	0.0581896552	0.1099137931	0.0560344828	0.0603448276	0.0107758621	0.0323275862	N
0.0523012552	0.0251046025	0.0523012552	0.050209205	0.0481171548	0.0585774059	0.0251046025	0.0711297071	0.0774058577	0.1087866109	0.0313807531	0.0585774059	0.0355648536	0.0188284519	0.0585774059	0.0774058577	0.0439330544	0.0627615063	0.0125523013	0.0313807531	N
0.0623376623	0.0233766234	0.0545454545	0.0779220779	0.0415584416	0.0857142857	0.0311688312	0.0675324675	0.0701298701	0.0961038961	0.0311688312	0.0337662338	0.0441558442	0.0233766234	0.0467532468	0.0649350649	0.0467532468	0.0727272727	0.0103896104	0.0155844156	N
0.0869565217	0.0179028133	0.0588235294	0.0690537084	0.0332480818	0.0818414322	0.0204603581	0.0511508951	0.0639386189	0.0997442455	0.0409207161	0.0281329923	0.0511508951	0.0281329923	0.0434782609	0.0818414322	0.0409207161	0.0664961637	0.010230179	0.0255754476	N
0.0895140665	0.0179028133	0.0588235294	0.0639386189	0.0332480818	0.0792838875	0.0230179028	0.0511508951	0.0716112532	0.094629156	0.0358056266	0.0255754476	0.0511508951	0.0281329923	0.0409207161	0.074168798	0.0485933504	0.0767263427	0.010230179	0.0255754476	N
0.0862944162	0.0253807107	0.0431472081	0.0710659898	0.038071066	0.0812182741	0.0253807107	0.0710659898	0.0736040609	0.0888324873	0.0253807107	0.038071066	0.0482233503	0.0329949239	0.038071066	0.0659898477	0.0507614213	0.0659898477	0.0101522843	0.0203045685	N
0.0483193277	0.025210084	0.0546218487	0.0525210084	0.0525210084	0.0672268908	0.018907563	0.0756302521	0.0819327731	0.1050420168	0.0294117647	0.0525210084	0.0399159664	0.0231092437	0.0504201681	0.0798319328	0.0420168067	0.0525210084	0.012605042	0.0357142857	N
0.0918367347	0.0178571429	0.0433673469	0.0714285714	0.0357142857	0.0790816327	0.0255102041	0.0586734694	0.056122449	0.1045918367	0.0229591837	0.0408163265	0.0459183673	0.0229591837	0.0433673469	0.0637755102	0.0663265306	0.0714285714	0.012755102	0.0255102041	N
0.0918367347	0.0178571429	0.0433673469	0.0714285714	0.0331632653	0.0790816327	0.0280612245	0.0586734694	0.0637755102	0.1020408163	0.0255102041	0.0382653061	0.0459183673	0.0229591837	0.0382653061	0.0612244898	0.0637755102	0.0765306122	0.012755102	0.0255102041	N
0.0918367347	0.0178571429	0.0459183673	0.068877551	0.0357142857	0.0841836735	0.0229591837	0.0637755102	0.0637755102	0.1020408163	0.0229591837	0.0408163265	0.0459183673	0.0229591837	0.0382653061	0.0637755102	0.0637755102	0.068877551	0.012755102	0.0229591837	N
0.0969387755	0.0178571429	0.0433673469	0.068877551	0.0331632653	0.0790816327	0.0255102041	0.0637755102	0.0612244898	0.1045918367	0.0255102041	0.0382653061	0.0484693878	0.0229591837	0.0408163265	0.0612244898	0.0612244898	0.068877551	0.012755102	0.0255102041	N
0.0943877551	0.0178571429	0.0357142857	0.0739795918	0.0331632653	0.0790816327	0.0255102041	0.0586734694	0.0663265306	0.1045918367	0.0255102041	0.0433673469	0.0459183673	0.0229591837	0.0357142857	0.0637755102	0.0612244898	0.0739795918	0.012755102	0.0255102041	N
0.0771208226	0.0179948586	0.0539845758	0.059125964	0.0359897172	0.0822622108	0.0154241645	0.0694087404	0.059125964	0.0899742931	0.0385604113	0.0488431877	0.0462724936	0.030848329	0.0462724936	0.0565552699	0.059125964	0.0771208226	0.0077120823	0.028277635	N
0.0798722045	0.0159744409	0.0607028754	0.0607028754	0.0383386581	0.0862619808	0.0127795527	0.0670926518	0.0575079872	0.1022364217	0.03514377	0.0479233227	0.054313099	0.0223642173	0.03514377	0.0638977636	0.0479233227	0.0798722045	0.0095846645	0.0223642173	N
0.0771208226	0.0205655527	0.0565552699	0.0616966581	0.0359897172	0.0796915167	0.0154241645	0.0694087404	0.0616966581	0.0899742931	0.0359897172	0.0437017995	0.0488431877	0.028277635	0.0488431877	0.0565552699	0.0565552699	0.0771208226	0.0077120823	0.028277635	N
0.1156812339	0.0154241645	0.0411311054	0.0694087404	0.030848329	0.0796915167	0.030848329	0.0514138817	0.0334190231	0.118251928	0.0179948586	0.0385604113	0.0514138817	0.028277635	0.059125964	0.0694087404	0.0437017995	0.0745501285	0.0102827763	0.0205655527	N

Figure 4. 5 A synopsis of the 20 Amino Acid feature engineered

Each column above shows a distinct amino acid letter code, which represents an engineered feature. The frequency of each amino acid within a specific protein sequence is listed as a decimal value in the cells below. These twenty columns, representing all twenty amino acid letter codes, become the *transformed* dataset ready to be fed into machine learning and statistical models for data analysis. Each column represents a dimension of the *transformed* dataset. These engineered features are also used in data visualisation models such as PCA, and TSNE to visualise the behavior of the dataset.

4.2.4 Baseline Database -- Data Transformation -- Feature Engineering Result and Discussion

To ensure that the raw datasets and the transformed datasets, including all three of the feature sets, are reusable, a baseline database was set up to store all of the data. This baseline dataset ensures that the transformed data is preserved and can be used for future studies. This baseline dataset can be found in Appendix B.

4.3 Summary

This chapter presented processes related to the collection and preparation of the protein sequence datasets. In order to address sequence length, which poses a very difficult problem in the handling of bioinformatics data, different studies (Khalid *et al.*, 2014; Zhou & Troyanskaya, 2015; Libbrecht & Noble, 2015) have explored using feature engineering techniques with protein and DNA sequences of different organisms and have been successful at converting these sequences into numerical data for machine learning and statistical analysis. This study took the same approach of data transformation to convert plant secondary metabolite genes (protein sequences) into numerical features. Data transformation and feature engineering methods were applied to resolve the discrepancies in sequence length and to address the high dimensionality of the protein sequence data.

Three feature sets were engineered based on the biological properties of the protein sequences. These features were computationally calculated using the Biopython libraries, which is available at: https://biopython.org/wiki/Documentation and http://biopython.org/DIST/docs/api/Bio.SeqUtils.ProtParam.ProteinAnalysis-class.html.

Feature engineering is a complicated and challenging process that requires in-depth literature review and intensive computational immersion. However, once the computational processes of engineering features are developed, these same processes can be used in future studies to engineer new features.

5. Implementation and Results

5.1 Introduction

This chapter discusses the use of data visualisation techniques use for inception of the dataset for feature selection and model building. Data visualisation tools such as PCA and TSNE were used as proof of concept to observe the behaviour of the datasets (RCHS, NCHS, and UCHS). The chapter goes on to discuss different proof of concept feature selection techniques that are applied on the eight engineered features.

Once the feature engineering phase was completed as seen in Chapter 4, the next step was to examine the eight features of the two feature sets, frequency-based and value-based features:

- AromaFeature
- Protein_Gravy_Feature
- Isoelectric_Point_Feature
- Protein_Stability_Feature
- HeliFeature
- TurFeature
- SheeFeature
- Entropy_Feature

To ensure uniform treatment among these eight features, ***data standardisation*** techniques were first applied on all eight features to center the features' values around zero, as described in Section 3.4.6. This was implemented using the Python pseudocode described below.

```
# PREREQUISITES:
From sklearn import preprocessing
import Pandas as pd
# INPUT
Open the datasetFeature as csv file
     Get column names first
ColumNames = datasetFeature.columns
      Create the Scaler object to hold the standardised
datasetFeature:
ScalerObject = preprocessing.StandardScaler()
      We then fit the data on the Scaler object
Standardised_datasetFeature = scaler.fit_tranform(datasetFeature)
Standardised_datasetFeature = pd.Dataframe(Standardised_datasetFeature,
columns =names)
 #OUTPUT
       Write the output of the program as an csv (tab delimited) via stdout
# close the current CSV file
        FILE_HANDLE.close()
```

The statistical significance level of these eight features was then measured to determine which features should be selected to enter the model. The statistical analysis revealed that all of the above eight features were significant. These eight features are henceforth referred to as the “eight significant features matrix (8SFM)” throughout the remaining sections of the document.

The third feature set, amino acid relative frequency, is referred to as twenty relative frequency feature matrix (20RFFM) in the remaining sections of the thesis. Based on the fact that each amino acid is nominal categorical data, this implied that the twenty amino acid relative frequency features are discrete categories which do not overlap. Based on this statistical understanding of this feature set’s data type, feature selection was not performed on this third feature set (Amino acid relative frequency feature).

The diagram below illustrates the various data visualization techniques that were applied to each of the feature sets. These applications are discussed in detail in the following sections.

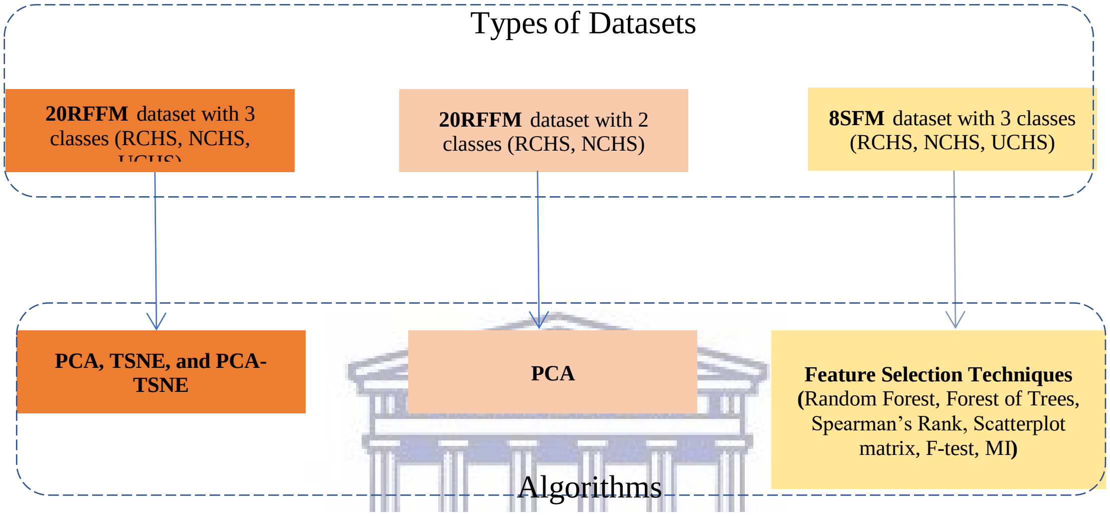

Types of Datasets
20RFFM dataset with 3 classes (RCHS, NCHS, UCHS)
20RFFM dataset with 2 classes (RCHS, NCHS)
8SFM dataset with 3 classes (RCHS, NCHS, UCHS)
PCA, TSNE, and PCA-TSNE
PCA
Feature Selection Techniques (Random Forest, Forest of Trees, Spearman’s Rank, Scatterplot matrix, F-test, MI**)**
Algorithms

Figure 5. 1 Graphical representation of Chapter 5 contents

This chapter is organised as follows: Section 5.2 outlines the contribution of this chapter, Section 5.3 presents Principal Component Analysis (PCA), Section 5.4 presents t-Distributed Stochastic Neighbour Embedding (TSNE), Section 5.5 outlines data visualisation for feature selection techniques, and lastly Section 5.6 presents the summary.

5.2 Contribution

This chapter provides an exploratory proof of concept approach to visualising secondary metabolite gene datasets and engineered features for feature selection. It provides a visual understanding of the data and the features that were engineered in Chapter 4 (Frequency-based features and Value-based features) using different feature selection techniques.

5.3 Principal Component Analysis on 20RFFM

The third feature set, referred to as twenty relative frequency feature matrix (20RFFM) dataset, was run through PCA using *sklearn.decomposition*. Principal Component Analysis (PCA) is used as a statistical approach to identify patterns in the dataset in such a way as to highlight differences and similarities in the data (Sochor *et al*, 2011; Buettner *et al.*, 2015; Anders *et al.*, 2015).

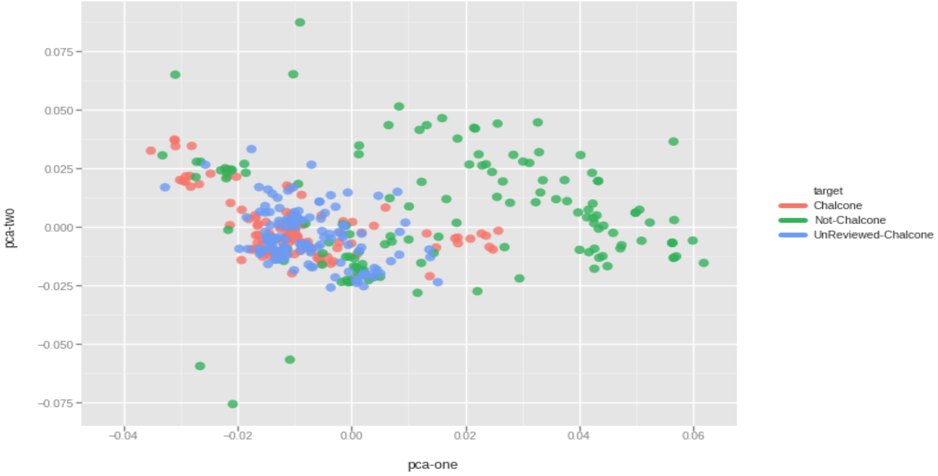


Figure 5. 2 First and Second Principal Components of PCA on RCHS, UCHS, and NCHS coloured by Target

The 20RFFM dataset consisted of the three classes, RCHS, NCHS, and UCHS, which are indicated in red, green, and blue respectively, as seen in figure 5.2. The three classes are projected onto the first two principal components (pca-one and pca-two), to visualise the behaviour of these three classes (of secondary metabolite genes) in terms of their differences and similarities. The visualisation shows that the NCHS class (in green) is scattered across the graph, whereas RCHS (in red) and UCHS (in blue) tend to cluster together around a concentrated area within the graph. This clustering of RCHS and UCHS indicates that these two classes (may) have similar properties (Sochor *et al.*, 2011). This quantitative information (finding) is important in demonstrating how RCHS and UCHS classes (may) possess very similar biological functionalities (i.e., catalytic activity). Although the RCHS class is made of experimentally reviewed chalcone synthase (meaning that their catalytic activity is known), and that the UCHS class is made of non-experimentally reviewed chalcone synthase (meaning that their catalytic activity is unknown), PCA's result suggests that these two classes' similarities are more than their differences. Whereas, the NCHS class's difference is more noticeable and shows that it has very little similarities either with RCHS or UCHS.

Furthermore, this PCA computational result suggests that the UCHS and RCHS protein sequences could potentially yield similar statistical inferences.

We then plotted the 20RFFM dataset using only two classes, RCHS and NCHS, which can be seen from figure 5.3. In this representation, the NCHS (blue data point) class presents the same behaviour as in figure 5.2, where it is scattered across the area of the graph. The RCHS (red data points), on the other hand, tends to cluster in one area of the graph. An explanation for this behaviour could lie in the fact that the RCHS class is made up of same enzyme, chalcone synthase. The RCHS clustering is therefore a representation of how similar and identical these enzyme sequences' properties are. On the other hand the NCHS class is made up of different enzymes, hence the scattered behaviour of the NCHS data points in this visualisation.

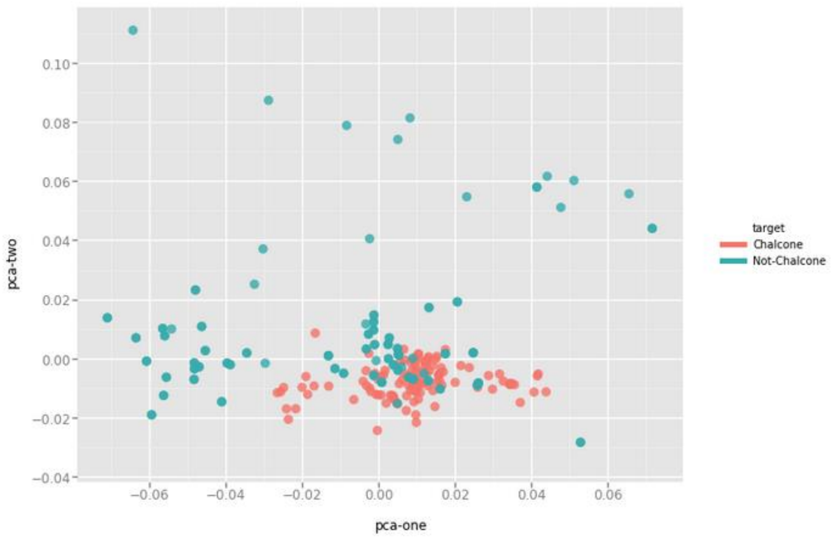


Figure 5. 3 First and Second Principal Components of PCA on RCHS, and NCHS coloured by Target

Since PCA is a dimensionality reduction tool, it is important to determine the extent to which the data is fully represented using these two components (pca-one and pca-two). A stairs plot graph was therefore plotted, based on the PCA singular values.

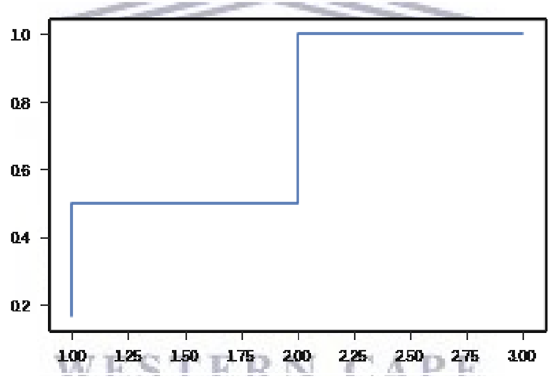


Figure 5. 4 Principal Component's Steps, indicating that two pca component can represent more than 95% of the data

The X-axis presents the singular values (principal components), and the Y-axis presents the probability of representing the data by singular values. Figure 5.4 shows that PCA has produced three singular values, and that using one singular value incorporates almost *50%* of the data, and two singular values incorporate almost *95%* of the data. This assessment proved that the data was well represented by using those two principal components.

5.4 t-Distributed Stochastic Neighbor Embedding on 20RFFM

t-Distributed Stochastic Neighbour Embedding (TSNE) was another data visualisation technique performed to analyse the behaviour of the three classes (RCHS, NCHS, and UCHS). TSNE analysis makes use of a probability distribution neighbouring embedded formula, as explained in Section 2.12.6.2 of the literature review. TSNE is a nonlinear dimensionality reduction technique that can transfer a high-dimensional dataset into a low-dimensional space (2D) for visualisation, in such a way that similar data points are modelled by nearby objects and unrelated data points are modelled by distant objects with high probability (Van der Maaten & Hinton, 2008).

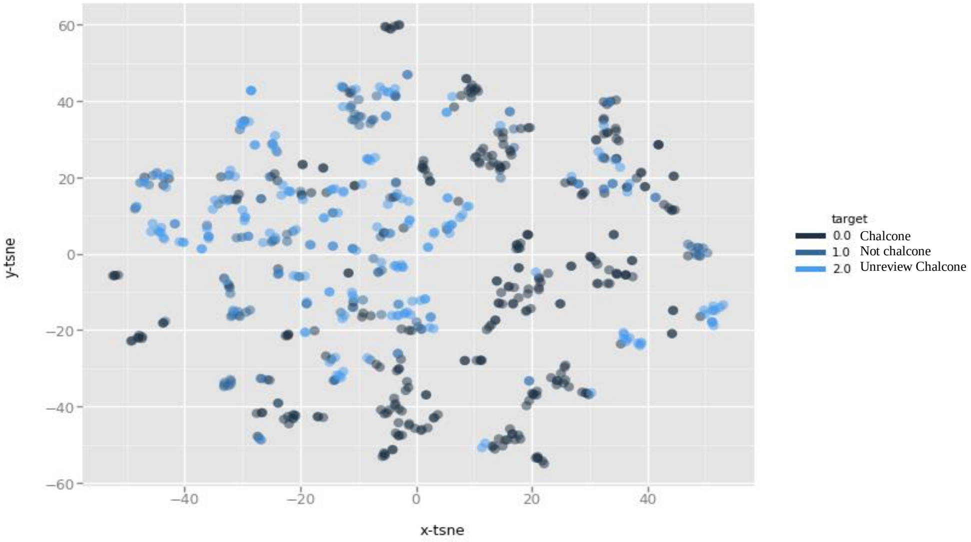


Figure 5. 5 TSNE with RCHS, NCHS, and UCHS coloured by target

TSNE was computed using Python *sklearn.manifold libraries* to analyse the RCHS (navy blue data points), NCHS (blue data points), and UCHS (light blue data points) datasets. TSNE struggled to clearly differentiate the behaviour of the three classes. One explanation may be the fact that the dataset is not sufficiently large enough for TSNE algorithm.

In an attempt to get a clearer visualisation, PCA and TSNE were combined using the PCA output as the input for TSNE's algorithm.

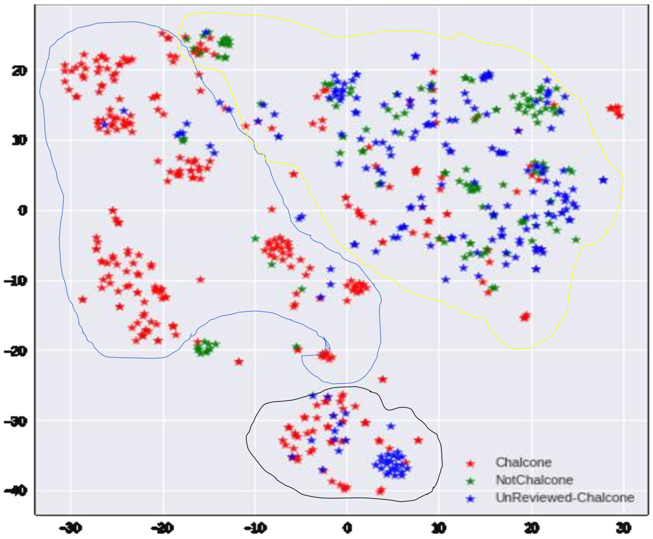


Figure 5. 6 TSNE-PCA Analysis of the 3 Classes; RCHS, UCHS, UCHSC

Figure 5.6 shows the three distributions clustered by the three outlines (blue, black, yellow). Each cluster represents the potential distribution between the three classes (RCHS, NCHS, and UCHS). In this visualisation, the PCA-TSNE appears to have minimised the divergence between two distributions: the first is the distribution that measured pairwise similarities of the classes and the second is the distribution that measured pairwise similarities of the corresponding low-dimensional data points of the three classes. However, even this approach PCA-TSNE, did not really do justice in differentiating the behaviour of the three classes of the dataset. Another drawback of TSNE and TSNE-PCA is that the algorithms take up much memory, CPU power, and time, as table 5.1 presents.

Table 5. 1 Computational cost of three dimensionality reduction algorithms

Algorithms	Elapsed Time	Dataset (RCHS, NCHS, UCHS)
PCA	0.36 seconds	20RFFM
TSNE	3.12 seconds	20RFFM
PCA-TSNE	55.59 seconds	20RFFM

TSNE and PCA-TSNE algorithms are computationally heavy and therefore pose some serious limitations to the use of these techniques. Another important limitation lies in the fact that it was not possible to feed only two classes within the dataset (RCHS and NCHS) into TSNE or PCA-TSNE, as was done with PCA, due to the fact that the two combined classes did not constitute a sufficient dataset size for TSNE and PCA-TSNE.

The PCA and TSNE were additionally run using the 8SFM, but the 20RFFM presented a much better result of these methods, and for the sake of a non-redundant study, only the performance of the 20RFFM is included here. The full code behind these algorithms is found in Appendix B.

5.5 Data Visualisation for Computational Feature Selection on 8SFM

Feature selection is of utmost importance to enhance the efficiency and improve the accuracy of supervised classifier algorithms as discussed in Section 2.12.4. Data visualisation was used in feature selection techniques to visualise the significance level of the eight features. As stated in the introduction section, and graphed in figure 5.1, all eight features from 8SFM dataset were examined.

Feature selection techniques were implemented through Python data visualisation tools such as Pandas, Seaborn, Matplotlib and Scikit-learn feature selection libraries. The full codes of all of the below techniques are found in Appendix B.

5.5.1 Feature Selection using Random Forest and Forest of Trees Techniques

As stipulated in Chapter 2 of the literature review, Section 2.12.4, a threshold of 0.05 is set to evaluate the significance of each of the eight features through a classification task, using Forest of Trees and Random Forest. Table 5.2 presents the ranking of the Random Forest feature selection technique against the Forest of Trees feature selection technique. It was observed that the four features of highest significance in each of these two techniques were similar.

Table 5. 2 Random Forest and Forest of Trees -- Features sorted by their score

Random Forest Features and Scores	*Forest of Trees Features and Scores*
1. *Isoelectric_Point_Feature : 0.158025*	1. *Isoelectric_Point_Feature : 0.165458*
2. *TurFeature : 0.150884*	2. *AromaFeature : 0.153747*
3. *AromaFeature : 0.134756*	3. *TurFeature : 0.140038*
4. *HeliFeature : 0.133901*	4. *HeliFeature : 0.129198*
5. *Entropy_Feature : 0.114431*	5. *Protein_Gravy_Feature : 0.111841*
6. *Protein_Gravy_Feature: 0.112555*	6. *Entropy_Feature : 0.104701*
7. *Protein_Stability_Feature: 0.098784*	7. *SheeFeature : 0.098235*
8. *SheeFeature : 0.096659*	8. *Protein_Stability_Feature : 0.096782*

However, neither of these feature selection techniques rank any feature to be two to three times higher than any of the other features. For instance, the highest feature on both models is *Protein Isoelectric Point*. This feature accounts for sixteen percent (*16%*), whereas the least significant feature on both models is *Protein stability* and *Sheet*, each accounting for almost ten per cent (*10%*). Furthermore, despite their rankings, all eight features are proven to be significant to the classification model, as their significance level is above *0.05*, and therefore all are important for model classification model building.

Figure 5.7 presents a visualisation of the Forest of Trees feature selection technique. The red bars indicate the significance level of the features (their inter-trees variability). As can be confirmed in the visual, the Forest of Trees ranked all eight features above *0.05 (default significance level)*.

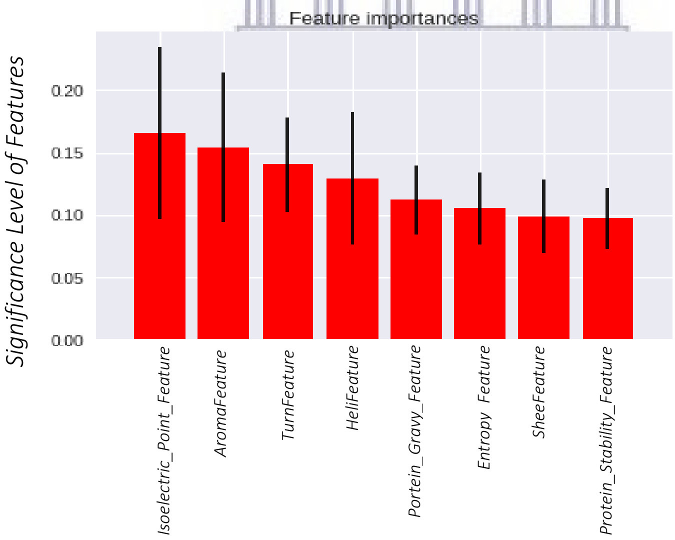


Figure 5. 7 Forest of Trees Feature Selection Techniques

It should be noted that these feature measurements are made possible only after the model has been trained and the model is dependent on all the above features. As such, if the model is trained without *Isoelectric_Point_F*eature, for example, it does not drop the model performance by *16 %*, for the simple reason that these features are independent of one another, and the 100 % significance level will be distributed among the remaining features.

Random Forest and Forest of Trees are strong methods for feature selection. However, the disadvantage of these two cited feature selection techniques often lies in their data interpretation, especially where correlated features are concerned. Important features can end up with low scores and the methods can be biased toward features with many categories. This disadvantage does not apply in our case, as the features are not highly correlated.

5.5.2 Spearman’s Rank correlation on 8SFM

Further analysis to investigate the correlation of the eight features was performed through the implementation of Spearman’s rank correlation. Spearman’s rank correlation coefficient, often denoted by the Greek letter ρ (rho), is a nonparametric measure of statistical dependence between the rankings of two features. Through the rho scores, which vary between [-1, 1], the direction and the strength of association between two ranked features was calculated.

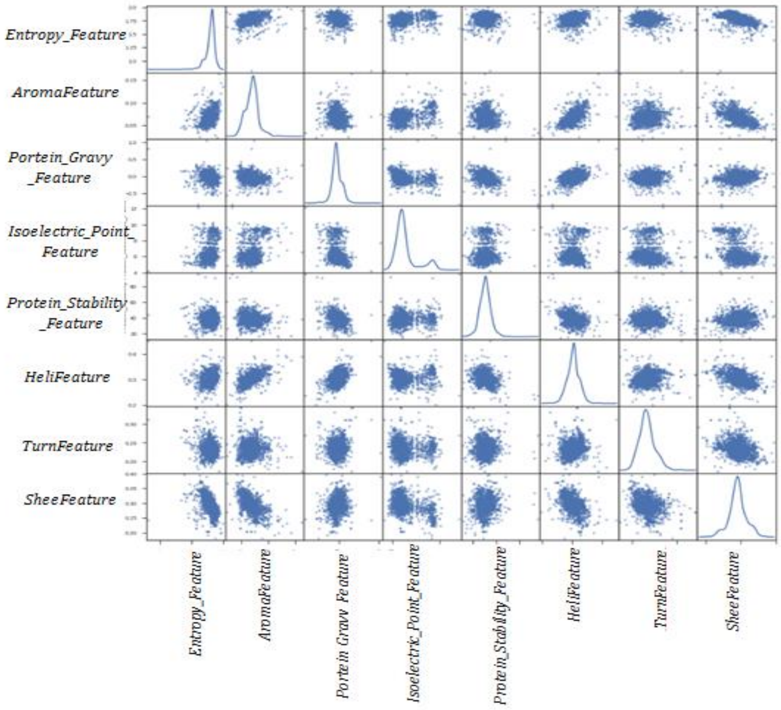


Figure 5. 9 Scatterplot matrix for features' correlation

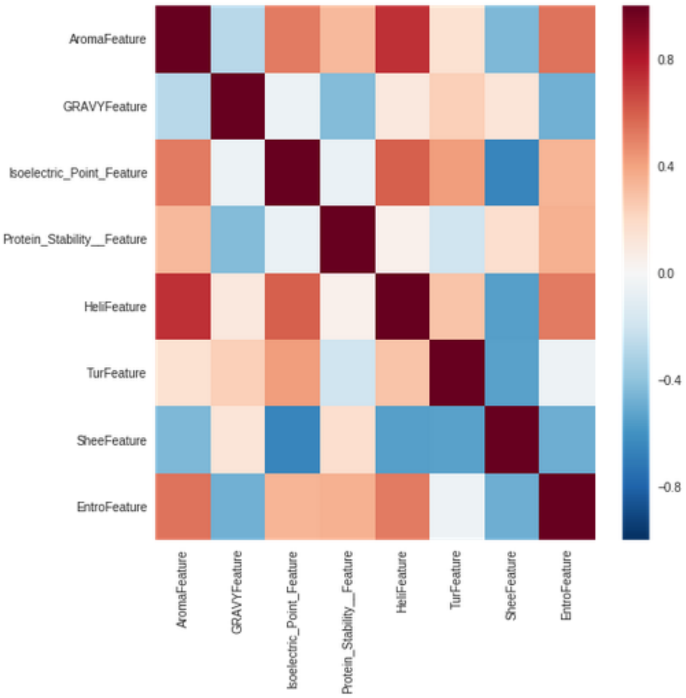


Figure 5. 8 Spearman's correlation for features' correlation

Figure 5.9 and 5.8 show a visualisation of feature correlation performed using a matrix. The scatterplot matrix (figure 5.9) has features presented in a distribution fashion (plot of linear correlation as, a line plot), while Spearman’s correlation matrix (figure 5.8) has features presented through a heat map (plot of positive and negative strongly correlated feature with a dark red and dark blue, respectively).

Both matrices present the eight features with perfect correlation from top left to bottom right in a diagonal fashion, where identical features mirror each other. This is the resulting representation when each feature is plotted against itself. For example, *AromaFeature* of figure 5.9 is plotted in the second column on the X-axis (left to right -- horizontal axis) and seventh row on the Y-axis (bottom to top – vertical axis) of both matrices, and *HeliFeature of figure 5.9* is the sixth column (left to right) on the X-axis and third row on the Y-axis (bottom to top). Both matrices have each feature on the X-axis mirroring eight features on the Y-axis, displaying their correlations.

In this regard, figure 5.8 reflects a positive correlation between *HeliFeature* and *AromaFeature*, and a very weak positive correlation between *AromaFeature* and *TurFeature*. At the same we can see a medium negative correlation between *SheeFeature* and *Isoelectric_Point_Feature*. However, both matrices indicate that many features likely have little to no correlation. Statistically, correlation does not explain cause and effect. Therefore, the conclusion that can be drawn from this visualisation is limited to the simple observation of feature correlation. The less these features are correlated, the better we can avoid model skewedness.

To pinpoint the correlation scores between these eight features numerically, a statistical analysis was performed to reveal the exact correlation scores of the eight features. These scores were taken directly from the scatterplot and Spearman's correlation matrices and are expressed as numerical values between -1 and +1 (ranges of correlation).

Table 5. 3 Numerical Correlation of the Eight Features

	AromaFeature	**Protein_Gravy_Feature**	**Isoelectric_Point_Feature**	**Protein_Stability_Feature**	**HeliFeature**	**TurFeature**	**SheeFeature**	**Entropy_Feature**
AromaFeature	**1.000000**	0.129037	0.762416	0.071704	0.732789	0.117342	-0.502525	0.392325
Protein_Gravy_Feature	0.129037	**1.000000**	0.219618	-0.343029	0.528988	0.139373	-0.127591	-0.276027
Isoelectric_Point_Feature	**0.762416**	0.219618	**1.000000**	-0.001767	0.617734	0.218557	-0.485811	0.299774
Protein_Stability_Feature	0.071704	-0.343029	-0.001767	**1.000000**	-0.224233	0.124105	0.248817	0.106563
HeliFeature	**0.732789**	0.528988	0.617734	-0.224233	**1.000000**	0.159825	-0.482219	0.228762
TurFeature	0.117342	0.139373	0.218557	0.124105	0.159825	**1.000000**	-0.333577	-0.111618
SheeFeature	-0.502525	-0.127591	-0.485811	0.248817	-0.482219	-0.33357	**1.000000**	-0.417499
Entropy_Feature	0.392325	-0.276027	0.299774	0.106563	0.228762	-0.11161	-0.417499	**1.000000**

Table 5.3 presents the exact correlation between identical features in a diagonal line (scores coloured in green), and the positive correlation (scores coloured in red) between *Aroma*Feature and both *Isoelectric_Point_Feature* and *HeliFeature*. The rest of scores indicate that the features are weak and poorly correlated.

These correlation techniques sufficiently prove that the eight features are independent features, and that removing one feature (as explained in Section 5.5.1) will not reduce model performance.

5.5.3 Data Visualisation ANOVA and Mutual Information feature Selection techniques

A visualisation of the ANOVA (F-test) and Mutual Information (MI) was also conducted to visualise the behaviour of five features within the 8SFM and their F-scores and MI-scores. F-test score tells the degree of linear dependency between random features, and MI-score captures any kind of statistical dependency.

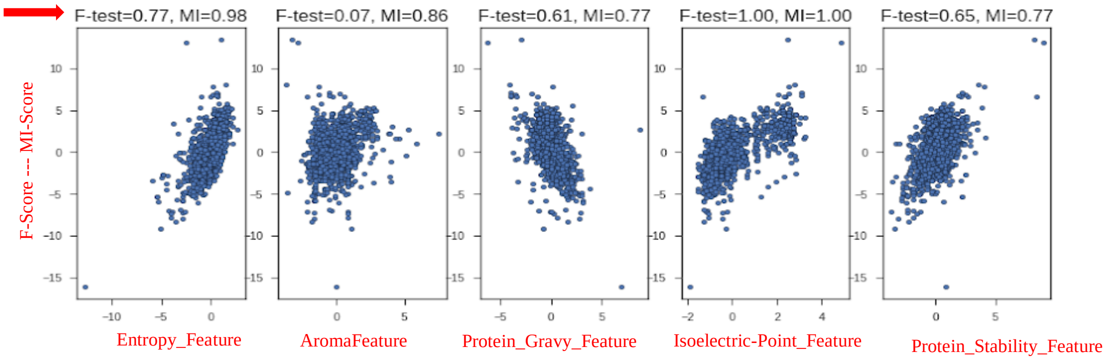


Figure 5. 10 Five Features and their F-scores and MI-scores

The red arrow in figure 5.10 points to F-test scores and MI-scores for five different features: *Entropy_Feature*, *AromaFeature*, *Protein_Gravy_Feature*, *Isoelectric_Point_Feature* and *Protein_Stability_Feature*. This method of feature selection uses a threshold of 0.05 significance level, as with the Random Forest and Forest of Trees methods. The features with F-test and MI scores of higher than 0.05 significance level would therefore be selected to enter the model. In this case, all the five features have a significance level score higher than 0.05. As with the previous feature selection techniques, this technique again confirms the independence of the features under review, signifying that they are significant for the model building.

Note: The five features were randomly chosen for this exercise and serve primarily as demonstration cases. This same technique was applied to the other three features in the 8SFM which are not mentioned here, HeliFeature, TurFeature, SheeFeature. Each had a significance score higher than 0.05.

5.6 Summary

This chapter has elaborated on proof of concept data visualisation tools such as PCA and TSNE, and their application to understanding the behaviour of secondary metabolite gene dataset. These techniques can be applied as best practice for visualising the data before analysis, and can actually provide initial insight on how to approach the data. In this case, the output of TSNE, PCA and PCA-TSNE of the 20RFFM dataset certainly leads us to the use of classification models over linear models. The data points tend to cluster for RCHS and UCHS classes, and the NCHS class is widely scattered across the graph, as opposed to a linear behaviour. Bearing this in mind, it became evident that classification models were best suited for the secondary metabolite gene datasets.

This chapter also addressed a few techniques for feature selection. The goal here was to propose an exploratory way of performing feature selection, in the instance where there are many engineered features and we want to determine their significance level. These proofs of concept feature selection techniques have revealed that all eight features in the 8SFM dataset were significant, and that classification models could be built on each of them. In next chapter, we therefore present exploratory proof of concept classification models.

Chapter Six: Machine Learning Supervised Classifiers Implementation and Results

6. Introduction

The data science computational pipeline displayed in Chapter 3 presented three major phases of this study, the first being "*data collection*"; the second, "*data preparation*"; and the third, "*data analysis*". The focus of Chapter 6 is on the third phase, data analysis, using supervised machine learning classifiers. One of the research questions for the study asks:

> *Can machine learning algorithms be trained to recognize plant secondary metabolite genes involved in the production of medicinally active compounds (e.g. polyphenols)?*

In this chapter we present exploratory proof-of-concept machine learning classifiers, to address our research question. Two types of supervised machine learning classifiers (Michalski *et al.*, 2013) are presented: binary classification and multiclass classification (Murphy, 2012). The binary classification consists of classifying (Menden *et al.*, 2013; Wang *et al.*, 2014) a protein sequence into the NCHS class or RCHS class, whereas the multiclass classification consists of classifying a protein sequence into either the RCHS class, NCHS class, or UCHS class.These two types of classifiers (binary classification and multiclass classification) are trained using both the 8SFM dataset and the 20RFFM dataset, each with two classes (RCHS and UCHS) and three classes (RCHS, NCHS, UCHS). Table 6.1 provides an overview of the different classification models applied to each of the datasets.

Table 6. 1 Representation of ML Supervised algorithms and their dataset types

Algorithms	Binary Classification		Multi Classification	
	20RFFM dataset with 2 classes (RCHS, NCHS)	**8SFM** dataset with 2 classes (RCHS, NCHS)	**20RFFM** dataset with 3 classes (RCHS, NCHS, UCHS)	**8SFM** dataset with 3 classes (RCHS, NCHS, UCHS)
Logistic Regression	✓	✓	✓	✓
Decision Tree	✓	✓	✓	✓
Random Forest	✓	✓	✓	✓
Support Vector	✓	✓	✓	✓
Gaussian Naïve Base	✓			
K-Nearest Neighbour	✓			
Single Layer Perceptron	✓			
Multilayer Perceptron	✓			

Eight supervised classification models were implemented in this study: Logistic Regression (LR); Decision Tree (DT); Random Forest with 100 trees (RF100); Support Vector Machine (SVM); K-Nearest

Neighbor (4NN and 2NN); Naïve Bayes (NB); Single Perceptron (SLP); and Multilayer Perceptron (MLP). These supervised classifiers were chosen due to their prevalence in the literature (Murphy, 2012; Menden *et al.,* 2013; Michalski *et al*., 2013; Kelley et al., 2016). While there are recommended default parameters for each of these algorithms, in order to boost the predictive power of some classifiers, some parameter changes were appropriate for improving the classification models (Alipanahi *et al.*, 2015). Once the classifier models were built, performance evaluation metrics were used, such as classification accuracy, AUC, recall, precision, F-measure, confusion matrix, and Cross-validation, to determine how well the classifier models performed. The results showed that the binary KNN model outperformed all the other binary models, whereas SVM model outperformed all the other multiclass models. This chapter details the performance of these two models. All of these supervised models were built using Anaconda Jupyter Notebook Python 3.6.1, and a variety of machine learning Scikit-learn and Scipy libraries as described in Chapter 3. The Python code behind their implementation is found in Appendix B.

This chapter is organised as follows: Section 6.1 outlines the contribution of this chapter, Section 6.2 presents Machine Learning Supervised classification implementation and result, Section 6.3 Model Accuracy of binary and multiclass models, and lastly Section 6.4 presents the summary.

6.1 Contribution

Most gene analysis tools that have been built previously to analyse gene material rarely address secondary metabolite genes. A minuscule amount of research has been dedicated to the study of key enzyme factors that can categorise a class of secondary metabolite genes. The main contribution of this chapter is the development of machine learning classification algorithms that classify (predict) chalcone synthase from a set of other plant enzymes. Binary and multi classification models are presented that are trained with two types of datasets built from two types of feature engineer techniques: 20RFFM and 8SFM. The ultimate goal of building machine learning models to predict plant enzymes, such as chalcone synthase, is to speed up the process of identifying plant secondary metabolite genes involved in the production of medicinally active compounds.

6.2 Machine learning Supervised Classification Implementation and Results

6.2.1 Model Building

The datasets used for the model building constituted the three different protein sequence classes (RCHS, NCHS, and UCHS) that were collected from UniProt database as explained in Chapters 3 and 4. For the binary classification models, 130 RCHS (experimentally verified chalcone synthase protein sequences) served as a true positive dataset, and 130 NCHS (experimentally verified protein sequences that are not chalcone synthase) served as a true negative dataset. For the multi classification models, we derived an extra class of 130 UCHS (not experimentally verified chalcone synthase protein sequences), randomly selected from a set of 2961 UCHS.

As explained in Chapter 4, data transformation and feature engineering were performed to transform the protein sequences of these three classes were converted into numerical feature datasets (20RFFM and 8SFM). Data standardization was then performed on the 8SFM, as explained in Chapter 5, and feature selection techniques were conducted to determine the significance level of the eight features of the 8SFM. All eight features were proven to be significant. Because the 20RFFM feature's values were typically zero-centered, we adapted this dataset (20RFFM) as such.

For the model building phase, the above rationales serve as the foundation of all classification models.

6.2.2 Model Training

The models were then trained with the purpose to establish parameters that minimize an objective function which measures the fit between the predicted instance and the actual instance. The parameters that were found to minimize the objective function and increase the predictive power of the models are described in table 6.2. The table presents the parameters within each of the different models that have been changed from their default values to reach a higher predictive power. It should be noted that the best parameter configuration is data driven and application-dependent (Bengio, 2012). We recommend that models with different configuration should be trained and their performance evaluated on a validation set. As the number of configurations grows exponentially with the number of parameters, testing them all would not be feasible in practice (Bengio, 2012). Therefore, we only focus on few of the parameters that were explored, while keeping all other parameters constant.

Table 6. 2 Recommended parameter settings for each model

Models	Parameters	Default	Recommended
Logistic Regression (LR)	solver	warn	lbfgs
	C	1.0	0.7
	N_jobs	none	1
	PCA	none	0.95
Decision Tree (DT)	criterion	gini	entropy
	random state	none	9 or 28
Random Forest (RF)	criterion	gini	entropy
	number of estimators	warn	100
	cross-validation	none	10
Support Vector Machine (SVM)	Anova f_regression		none
	C	1.0	1.4
	K	2	4 or 6 or 7 or 8
Gaussian Naïve Base (GNB)	random state	none	9
K-Nearest Neighbour (KNN)	CV	5	10
	random state	none	6
	number of neighbors	5	4 or 2
	weights	uniform	[uniform, distance]
Single Perceptron (SLP)	random state	none	9
Multilayer Perceptron (MLP)	solver	adam	lbfgs
	alpha	0.0001	1e-5
	random state	none	5
	Hidden_layer_sizes	(100,)	(5 , 2)

6.2.3 Performance Evaluation Metrics of Proof of Concept Binary Models

6.2.3.1 Area Under the receiver operating characteristics Curve

The area under the receiver operating characteristics curve (AUC) was used in this study to evaluate the performance of all classification models. AUC was observed by plotting the True Positive Rate (Y-axis) against the False Positive Rate (X-axis). The selection of this performance metric was motivated by its common use in the literature. The experiment was conducted using our two types of datasets (20RFFM and 8SFM), and eight binary classifiers (LR, DT, RF100, SVM, GNB, KNN, SLP, and MLP) to build predictive classifier models. Their performance was then measured using AUC, and 10-fold cross-validation to build and test the models. Table 6.3 presents the results of the average classification performance in terms of AUC. The result showed that KNN outperformed all other models across the two datasets (20RFFM and 8SFM).

Table 6. 3 Average Binary Model Classification Performance in terms of AUC

Models	Binary Classification	
	20RFFM dataset with 2 classes (RCHS, NCHS)	**8SFM** dataset with 2 classes (RCHS, NCHS)
LR	0.8531	0.8952
DT	0.8603	0.8421
RF100	0.8865	0.9101
SVM	0.8795	0.8653
GNB	0.8236	0.7862
KNN	0.9356	0.9158
SLP	0.7985	0.6958
MLP	0.8565	0.8324
Average Performance	0.8617	0.8349

The table also shows that some models performed better with the 8SFM dataset, whereas other models performed better with the 20RFFM dataset. For instance, RF100 achieved 0.91 average AUC performance with the 8SFM dataset, whereas RF100 with 20RFFM achieved 0.89 average AUC performance. DT achieved 0.86 average AUC performance with the 20RFFM, whereas it achieved 0.84 with the 8SFM. When considering SLP, the table shows that it was the worst performing model on both datasets. The last row of the table shows the average performance of each dataset. On average, the 20RFFM dataset outperformed the 8SFM dataset.

6.2.3.2 K-Nearest Neighbor Binary Models

For the sake of limiting the scope and content discussed in the study, this chapter will only explain in detail the application of the models that had the best performance for each of the binary and multiclass problems. As K-Nearest Neighbors (KNN) was discovered to have the best AUC performance of the binary classification models, this section will focus on the development of the KNN model.

K-Nearest Neighbors (KNN) model was built for binary classification, where the input consisted of the K closest training examples in the feature space. In this study, the optimal value for *K* was determined by first building a *cross-validation algorithm* to retrospectively determine the best *K* value. Next, a *grid-search cross-validation algorithm* was implemented to obtain the best *K* value. Both approaches used the 20RFFM dataset (since table 6.3 indicated the 20RFFM outperformed the 8SFM) to validate the optimal *K*. *K = 4* and *K = 2* where found to be the optimal *K* values using these two K-fold cross-validation algorithms.

Figure 6.1 shows the output of *cross-validation algorithm*. The X-axis represents a value of K, and the Y-axis represents the prediction accuracy. From this graph we can conclude that the best K value is 4, as it yielded a 92% cross-validation predictive power.

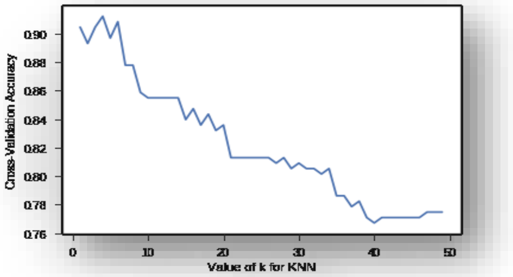


Figure 6. 1 Optimal Value of K using Cross-validation. Each point on the line represents a K value, and 4 yields a higher prediction accuracy.

Figure 6.2 shows the output of the *grid-search cross-validation algorithm*. This graph shows that the best value of K from grid-search cross-validation is 2, yielding close to a 95.5% cross-validation predictive power.

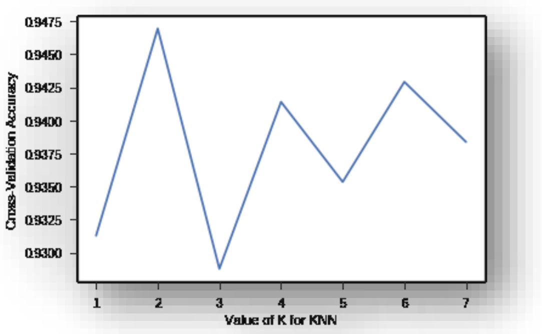


Figure 6. 2 Optimal Value of K using Grid-search Cross-validation. Each point on the line represents a K value, and **2** *yields a higher prediction accuracy.*

Both 4NN and 2NN on 10-fold cross-validation fold produced very significant accuracy rates. However, performing a Grid-search Cross-validation algorithm increased the prediction rate by *3.5%* as opposed to a Cross-validation algorithm. It can also be seen, from both of the algorithms' behavior (figure 6.1 and 6.2), that as the value of KNN increases, the prediction rate decreases. Therefore, we recommend building a KNN model on a 10-fold cross-validation, with the Grid-search Cross-validation technique, to find the optimal K for the best KNN model.

6.2.3.2 K-Nearest Neighbor Model Confusion Matrix

Since 2NN model with 20RFFM dataset proved the best classifier of our proof-of-concept model, we therefore chose to report the confusion matrix based on this model.

Note: The confusion matrix Python code was used for all other seven classifiers as well. But for the sake of a concise study, we chose to report on the best performing model.

As elaborated in Section 6.2.1 above, the dataset that was used for binary classification consisted of 130 RCHS, which served as a true positive set, and 130 NCHS, which served as a true negative set. Therefore, the final binary class contained 260 sequences. The 20RFFM dataset, which consisted of the 20 amino acid features, was used to train the 2NN proof-of-concept model and to perform 10-fold cross-validation on the final model. Training was conducted on 80% of the dataset, and the remaining 20% of the dataset was used to test the model, as is standard in machine learning best practice. The 2NN proof-of-concept model achieved an average accuracy of 94.2% on 10-fold cross-validation, 96.2% on training 80% of the 20RFFM data, and 94.4% on testing 20% of the 20RFFM data.

The train or test split is useful because of it flexibility and speed. However, one limitation of the train test split method is that it can result in a high variance estimate of out-of-sample data. To overcome this limitation, a 10-fold cross-validation is performed by repeating the train or test split multiple times and averaging the results. The pseudocode of the 2NN model can be found below.

```
# PREREQUISITES:
From sklearn import all necessary libraries i.e.,
import numpy as np, import pandas as pd
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.neighbor import KNeighborsClassifier
from sklearn.metrics import confusion_matrix
from sklearn.cross_validation import cross_val_score
from sklearn.cross_validation import KFold
etc.

# INPUT
Input the datasetFeature as csv file using pandas
Convert the pandas dataset into numpy
Split the numpy dataset with X as 20 features(Xinput) and Y as class label(Yinput)

# Model Building - Model Training -Model Evaluation
  Evaluate by training and testing using data split procedure
  Locate K by creating a for loop and record the testing accuracy

  Training the model with Xinput and Yinput
  Xtrain, Xtest, Ytrain, Ytest = train_test_split(Xinput, Yinput, test_size =
0.20,we can use our recommended parameter in table 6.2 here)

  2NN = KNeighboursClassifier (n)neighbors = 2)
  2NN = knn.fit(Xinput, Yinput)
  Ypredict = 2NN.predict(Xinput)

  Calculate accuracy on training set:
  metrics.accuracy_score(Yinput, Ypred)

  Calculate the mean of 10-fold cross-validation:
  Scores = cross_val_score(2NN, Xinput, Yinput, CV = 10 scoring='accuracy')

 Calculate the mean accuracy of 10-fold cross-validation:
  Scores.mean()

  Calculate the Confusion Matrix to describe model performance
  ConfusionMatrix = metrics.confusion_matrix(Ytest, Ypredict)
  If confusion metrics are weak, adjust the classification threshold either below
  or higher than 0.05
```

The classification accuracy (94.2%) does not explain the underlying distribution of response values, nor does it reveal the type of errors that result from the classifier. Therefore, we chose to address this with confusion matrix (figure 6.3), which best describes the performance of our classification model by calculating a range of model evaluation metrics.

N = 52	Predicted: NCHS	Predicted: RCHS	
Actual: NCHS	TN = 13	FP = 2	15
Actual: RCHS	FN = 1	TP = 36	37
	14	38	

Figure 6. 3 Validation confusion matrix for the binary 2NN proof-of-concept model.

Figure 6.3 shows the validation confusion matrix, where each row represents the actual class, while each column represents the instances in a predicted class. The advantage of using confusion matrix as a metric evaluation tool is that the data mining analyzer is capable of detecting whether or not the classifier is confusing two classes. Further definition of the confusion matrix terms have been explained in the literature review under Section 2.12.6.3.

Table 6.4 describes the results in figure 6.3 in further detail.

Table 6. 4 Validation confusion matrix explained

Metrics	Description	Rate
Accuracy	Accuracy of the model based on how often the model is correct overall	(TP+TN)/N = 94.2%
Misclassification Rate (Error Rate)	Based on how often the model is wrong overall	(FP+FN)/N = 5.77%
True Positive Rate (Sensitivity or Recall)	The rate at which the model predicts RCHS when the input is actually RCHS	TP/Actual RCHS = 97.3%
False Positive Rate	The rate at which the model predicts RCHS when the inputs is actually NCHS	FP/Actual NCHS = 13.3%
True Negative Rate (Specificity)	The rate at which the model predicts NCHS when the input is actually NCHS	TN/Actual NCHS = 86.7%
Precision	The rate at which the model is correct when predicting RCHS	TP/Predicted RCHS = 94.7%
Prevalence	The rate at which the RCHS condition occurs in the sample	Actual RCHS/N = 71.1%
Null Error Rate	The rate at which the model is wrong if it always predicts the majority class	Actual NCHS/N = 28.8%

The Null Error Rate shows that the 2NN model would be wrong only 28.8% of the time if it predicted the majority class each time. The Misclassification Rate is shown to be very low at 5.77%. We also see that the model achieved a recall (True positive Rate) of 97.3%.

The confusion matrix has clearly described the performance of the 2NN proof-of-concept model, and from this we can see that the model indicates promise for using machine learning classification models to predict plant secondary metabolite genes.

6.2.4 Performance Evaluation Metrics of Proof-of-Concept Multiclass Models

6.2.4.1 AUC of the Multi Classification Models

The multi classification experiment was conducted using our two types of datasets (20RFFM and 8SFM), on four multiclass classifiers (LR, DT, RF100, and SVM), following the same model building as explained in Section 6.2.1 above. Their performance was measured using AUC, and 10-fold cross-validation. Table 6.5 presents the results of the average classification performance in terms of AUC. The result showed that the multiclass SVM outperformed all other models across the two datasets (20RFFM and 8SFM).

Table 6. 5 Average Multiclass Classification Model Performance in terms of AUC

Models	Multiclass Classification	
	20RFFM dataset with 3 classes (RCHS, NCHS, UCHS)	**8SFM** dataset with 3 classes (RCHS, NCHS, UCHS)
LR	0.7631	0.6952
DT	0.8503	0.8621
RF100	0.8705	0.8581
SVM	0.9095	0.9013
Average Performance	0.8484	0.8120

Table 6.5 also shows that some models performed better with one dataset than the other. For instance, DT achieved 0.86 average AUC performance, whereas DT with 20RFFM achieved 0.85 average AUC performance. On the other hand, LR achieved a 0.76 average with 20RFFM, while it achieved a 0.70 average AUC performance with 8SFM. When considering LR, the table shows that it was the worst performing model for both datasets, whereas SVM outperformed all other models on both datasets and achieved almost the same accuracy rate. The last row of the table shows the average performance of each dataset, indicating that the 20RFFM dataset outperformed the 8SFM dataset.

The pseudocode of the multiclass SVM can be seen below,

```
# PREREQUISITES:
from sklearn.model_selection import train_test_split, from sklearn.preprocessing import StandardScaler
,import pandas as pd, import numpy as np, from sklearn.svm import SVC, from sklearn import svm, from
sklearn import metrics, from sklearn.pipeline import make_pipeline, from sklearn.metrics import
classification_report, etc.

# Input:
df =pd.read_csv('fielpath.csv')
#read each column of the dataset and their class label
# Model building --Model training -Model Evaluation
# test_size: what proportion of original data is used for test set
train_x, test_x, train_lbl, test_lbl = train_test_split(x, Y1, test_size=0.20, random_state=0)
scaler = StandardScaler()
#Scale the all the column values (data into one scale in case of 8SFM)
# Apply transform to both the training set and the test set.
train_x = scaler.transform(train_x)
test_x = scaler.transform(test_x)

# Step 3: Training the model on the data, storing the information learned from the data
# Model is learning the relationship between feature values and labels

# Step 4: Predict the labels of new data
# Uses the information the model learned during the model training process

# all parameters not specified are set to their defaults SEE "default parameter below
svmModel = svm.SVC(C=1.4, kernel='rbf',decision_function_shape='ovr' ,coef0=0.0 )
```

Figure 6.5 shows the SVM Area Under the ROC (AUC), determining each class's predictive power when using the 20RFFM dataset. These ROC curves (teal, blue and orange) were generated by varying the decision threshold *t* used to transform the normalized features into a predicted class. The ROC curves plot the true positive rate of the classes (RCHS, NCHS, UCHS) on the Y-axis, versus the false positive rate of these three classes on the X-axis, where the classifier decision threshold is varied from 0 to 1.

Note: *There is a controversy over AUC being equated to predictive power. In this study, we lean on the side of those scientists who equate it to predictive power.*

The area under the ROC curves is then used as a single numerical metric to indicate the performance of the SVM multi classifier: 0.91 for RCHS; 0.60 for NCHS; and 0.79 for UCHS. In the construction of the ROC, the three classes (RCHS, NCHS, and UCHS) outputs are binarized: 0 for RCHS, 1 for NCHS, and 2 for UCHS.

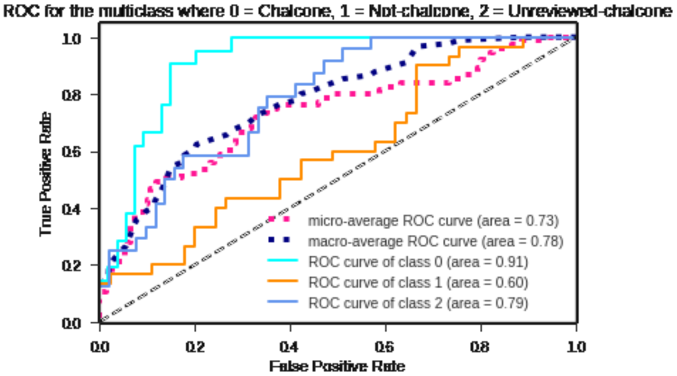


Figure 6. 4 AUC average probability the SVM multiclass classifier assigns a higher probability to chalcone synthase Proof of Concept

The RCHS class reaches the predictive power of 0.91, the NCHS class 0.60, and the UCHS class 0.79 in terms of AUC. However, it should be noted that RCHS class and UCHS class share the same micro-average and macro-average ROC area. This suggests that for future work the two classes (RCHS, UCHS) could potently be combined into one class. However, in the case where the RCHS and UCHS are combined into one, as can be seen from the graph, we would expect the number of false positives, along with true positives, to increase.

Since the exploratory proof-of-concept multiclass model that was built could differentiate the three classes, this presents promise that binary classification models, which are commonly used are not the only models suitable for predicting genes. Multi classification models could also be built to predict different secondary metabolite genes such as chalcone synthase, stelibene synthase, aloesome synthase, etc. which could be predicted or classified at the same time.

6.3 Average Performance Accuracy of Binary and Multiclass Models

Table 6. 6 Average Performance Accuracy of Binary and Multiclass Models

Algorithms	Binary Classification				Multi Classification Model Accuracy			
	20RFFM dataset with 2 classes (RCHS, NCHS)		**8SFM** dataset with 2 classes (RCHS, NCHS)		**20RFFM** dataset with 3 classes (RCHS, NCHS, UCHS)		**8SFM** dataset with 3 classes (RCHS, NCHS, UCHS)	
	Model Accuracy	**AUC**	**Model Accuracy**	**AUC**	**Model Accuracy**	**AUC**	**Model Accuracy**	**AUC**
Logistic Regression	0.881	**0.85**	0.905	**0.89**	0.794	**0.76**	0.678	**0.69**
Decision Tree	0.852	**0.86**	0.832	**0.84**	0.843	**0.85**	0.853	**0.86**
Random Forest	0.872	**0.88**	0.895	**0.91**	0.886	**0.87**	0.855	**0.86**
Support Vector	0.898	**0.88**	0.887	**0.86**	0.904	**0.91**	0.891	**0.90**
Gaussian Naive Base	0.810	**0.82**	0.823	**0.78**				
2-Nearest Neighbour	0.942	**0.94**	0.929	**0.92**				
Single Perceptron	0.812	**0.80**	0.732	**0.69**				
Multilayer Perceptron	0.881	**0.86**	0.848	**0.83**				

Table 6.6 shows the performance of all the binary classifiers and multi classifiers in terms of model accuracy and AUC with their respective datasets. It can be seen that the 2NN has overall outperformed all the other binary models, whereas SVM multiclass model has outperformed all the other multiclass models. SLP binary model has shown a lower performance among the other binary models, and LR multiclass model has shown a lower performance compared to the other multiclass models.

6.4 Summary

The goal of this chapter was twofold. First, it aimed to identify the best binary classification model for prediction of RCHS. The binary 2NN model proved to be the most accurate model for predicting RCHS in the presence of an experimentally validated true negative dataset (NCHS). The 2NN model was trained and tested (80% and 20% respectively) on a 10-fold cross-validation with two different datasets: 20RFFM and 8SFM. The 20RFFM dataset outperformed the 8SFM in terms of AUC. Therefore, the proof-of-concept 20RFFM 2NN model was further evaluated using confusion matrix and delivered good results, achieving a 94.2% average accuracy, a precision of 94.7% and a sensitivity of 97.3%.

The second aim of this chapter was to identify the best multiclass classification model for prediction of RCHS. SVM was identified as the most accurate model for predicting RCHS, and was trained and tested (80% and 20% respectively) on a 10-fold cross-validation with two different datasets: 20RFFM and 8SFM. Again, the 20RFFM dataset outperformed the 8SFM in terms of AUC, achieving 91% average accuracy model. The rationale behind using a multiclass classifier to build a proof-of-concept model for classifying chalcone-synthase in secondary metabolite genes is that there is limited RCHS and NCHS data available, and UCHS could therefore be used to increase the amount of RCHS, so that the binary model can be trained with a larger positive dataset. In addition to this, the multiclass classifier can potentially be used in a study that is trying to predict different secondary metabolite genes, so that the prediction of secondary metabolites is not limited to One vs. All (e.g. RCHS, which makes up the true positive dataset, versus all of the not chalcone synthase genes which constitute the true negative dataset). Instead, the predictions can include a pairwise comparison, identifying multiple classes of secondary metabolite genes (e.g. chalcone synthase vs. stelibene synthase vs. aloesome synthase).

This exploratory proof-of-concept experiment suggests that secondary metabolite genes can be predicted using supervised machine learning models. To our knowledge, there is no previous work that implemented this strategy for predicting RCHS using machine learning techniques. Therefore, we cannot conduct a comparison of this study against previous studies. However, our exploratory proof-of-concept model shows good results for predicting secondary metabolite genes.

Chapter Seven: Statistical Analysis and Results

7.1 Introduction

This chapter addresses one of the key research questions stipulated in Chapter 1 Section 1.3:

> *Can mathematical and statistical estimated approaches be carried out pertaining to the preparation, analysis and interpretation of secondary metabolite genes?*

In this chapter, the mathematical statistical analysis and interpretation of secondary metabolite genes will be conducted to determine the inferences between the three classes: RCHS, NCHS, and UCHS.

In Chapter 5, the application of feature selection techniques led to the conclusion that all the eight features of the 8SFM dataset—*AromaFeature*, *Protein_Gravy_Feature*, *Isoelectric_Point_Feature*, *Protein_Stability_Feature*, *HeliFeature*, *TurFeature*, *SheeFeature*, and *Entropy_Feature*—were statistically significant. Therefore, a selection of features from the 8SFM dataset is used in this chapter to conduct mathematical statistical analysis of reviewed chalcone synthase (RCHS), not chalcone synthase (NCHS), and unreviewed chalcone synthase (UCHS).

This chapter is as follow: Section 7.2 explains the main contribution, Section 7.3 presents the methods used in this work for normal distribution, Section 7.4 present the chi-square model and its post hoc test, Section 7.5 presents the ANOVA model and its post hoc test, Section 7.6 presents the boxplot model and finally, Section 7.7 presents the summary.

7.2 Contribution

This chapter presents the mathematical and statistical analysis conducted and explains the statistical tests and probability distributions that were used in this study. It elaborates on the rationale behind these statistical approaches. The main contribution of this chapter is to present an empirical analysis to provide guidance on best practices when analysing bioinformatics data statistically (Fowler *et al.*, 2013; Anderberg, 2014; Burgess & Smith, 2017). In this chapter, three classes (RCHS, NCHS and UCHS) were analysed on a balanced 8SFM dataset. We used Python *Statsmodels API*, *Pandas*, *Numpy*, and *Matplotlib* libraries to implement all the statistical models. All Python code written for the implementation of this chapter is found in Appendix B.

7.3 Normal Distribution

In this chapter, statistical assumptions was observed throughout to ensure that the statistical analysis and statistical interpretation carried out are statistically significant. Normalisation of the data was performed on each of the three classes of the 8SFM dataset to safeguard that most values of the eight features remain around the mean value. The normal distribution was computed as seen in table 7.1.

Table 7. 1 Normal Distribution of the Three Classes of 8SFM

Target	count	Mean	std	min	25%	50%	75%	max
RCHS	130.0	0.069643	0.004330	0.058	0.067	0.069	0.072	0.081
UCHSC	130.0	0.081414	0.014353	0.048	0.070	0.079	0.094	0.106
UCHS	130.0	0.068556	0.005692	0.051	0.066	0.069	0.071	0.102

Histograms were computed, over which the probability distribution of each class was plotted, as seen in figures 7.1, 7.2, 7.3. These histograms were plotted based on the *count*, *mean* and *std*. columns of table 7.1.

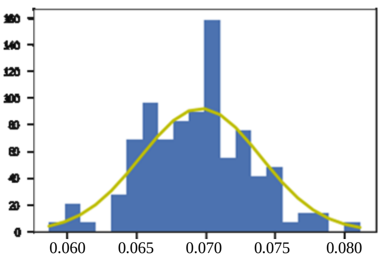


Figure 7. 5 mean =0.069643 sigma = 0.004330 RCHS Distribution

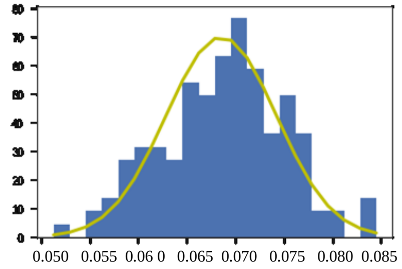


Figure 7. 5 mean =0.068556 sigma =0.005692 UCHS Distribution

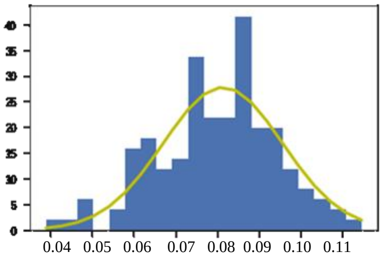


Figure 7. 5 mean =0.081414 sigma =0.014353 NCHS Distribution

These figures show the range of each distribution (values on X-axis) and their steepness (yellow line). The RCHS distribution range lies between *0.060* to *0.080*, followed by the UCHS distribution range of *0.050* to *0.085*, and finally the NCHS distribution range which is the largest from *0.040* to *0.110*. The distribution range of each class shows the spread of the deviation. The smaller the range, the smaller the deviation. Therefore, it can be statistically stated that RCHS and UCHS classes have a smaller distribution range than the NCHS class. To further infer about these three classes, hypothesis tests were conducted.

7.4. Chi Square Hypothesis Test

In order to test if there was a significant relationship between any two of the three classes (RCHS, NCHS, and UCHS), a chi-square test for independence was carried out by comparing one feature at a time from the 8SFM dataset to the rest of the three classes. The assumptions of chi-square that were followed in this analysis are explained in Chapter 2, Section 2.10.1.

The chi-square test tested if there was an association (Zhang & Finer, 2016) between the three classes based on a given feature.

The Null Hypothesis (H_0) stated that there is no relationship between two given classes, while the Alternative Hypothesis (H_1) stated that there is a relationship between two given classes. Table 7.2 lists the results of the chi-square test.

Table 7. 2 Chi-square test results of the 8SFM dataset

8SFM Features	Critical Value	Chi-square stat
AromaFeature	2845	47.54
Protein_Gravy_Feature	2845	8.59
Isoelectric_Point_Feature	2845	47.55
Protein_Stability_Feature	2845	19.40
HeliFeature	2845	8.58
TurFeature	2845	8.6
SheeFeature	2845	47.55
Entropy_Feature	2845	19.39

Since the critical value of for the chi-square statistics is determined by the level of significance (typically 0.05) and the degree of freedom, the critical value for all the features will be identical (Zhang & Finer,

2016) As it can be seen from table 7.2, the observed chi-square stats are lesser than the critical values and we therefore accept the alternative hypothesis and conclude that there is in fact a relationship between two given classes.

Statistically, we can add that since the hypothesis results in very small chi-square test statistic, this means that the observed data fits the expected data extremely well (D'Agostino, 2017).

7.4.1 Bonferroni Correction test

Although the chi-square test was significant, because the analysis was 3 x 8 (number of classes times number of features), the chi-square test could not yet illustrate where the relationship was between the classes. A post hoc test, Bonferroni correction, as discussed in Section 2.10.2, was conducted to determine where exactly the relationship was between the different classes (Armstrong, 2014).

In this study, k= 3 (number of classes: RCHS, NCHS, UCHS), so there are 3 x (3-1)/2 = 3 pairwise differences to consider.

The formula was therefore: 0.05 / 3 = 0.017.

Hence, for the planned pairwise comparisons to be significant, the p-value must be less than 0.017. To conduct multiple 2x2 chi-square tests, the classes were regrouped for each test, where one class was compared against the other two:

Table 7. 3 Bonferroni Correction test result

Classes	**Chi-square Test**	**P-value**	**Degrees of Freedom**
RCHS Vs. NCHS	325.0534144065055	0.17733575213536487	216
RCHS Vs. UCHS	235.11175243959863	**0.2320208060446495e-06**	216
NCHS Vs. UCHS	337. 9344044779791	0.14608020962424864	216

Using the Bonferroni-adjusted p-value of 0.017, one of the three planned pairwise comparisons was found to be significant—*RCHS Vs. UCHS, p-value < 0.017*. This result confirmed a relationship between the RCHS class and *the* UCHS class.

These findings are of high importance, as there is statistical evidence which confirms that the protein sequences labelled as unreviewed chalcone synthase (UCHS) have a relationship with those labelled as Reviewed chalcone synthase (RCHS). In contrast, the protein sequences of the NCHS class and these other classes differ and present no relationship. In terms of evaluating RCHS vs. NCHS, this result leads to believe that chalcone synthase can be predicted from a class of NCHS. Biologically it would infer that chalcone synthase do not have the same catalytic activity as the other genes in the NCHS class, and that the constituency of chalcone synthase is different from the rest of the secondary metabolite genes in the NCHS class.

Another hypothesis test to verify these findings was carried out with analysis of variance (ANOVA) hypothesis testing (Sochor *et al.*, 2011; Fowler *et al.*, 2013; Fois *et al.*, 2015).

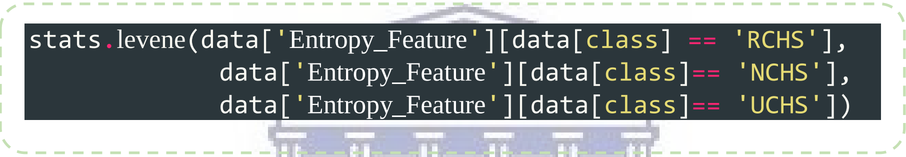
7.5 One-way Analysis of Variance

The analysis of variance (ANOVA) hypothesis test and its assumptions were carried out to compare the means of a condition between two classes (Fowler *et al.*, 2013; Fois *et al.*, 2015), as elaborated in Section 2.11.1 of Chapter 2. For this specific ANOVA test, the Null hypothesis (H_0) stated that there is no difference between the means of the classes (RCHS, NCHS, and UCHS), while the Alternative hypothesis (H_1) stated that a difference between the means exists somewhere between the classes.

The assumption of homogeneity of variance (Fowler et al., 2013; Fois et al., 2015):

- The variables are normally distributed in each group that is being compared in the one-way ANOVA.
- There is homogeneity of variances. This means that the population variances in each class are equal.
- There is an independence of observations.
- A caveat to these assumptions is that if the class sizes are equal, the F- statistic is robust to violations of normality and homogeneity of variance.

was checked with Levene's test (code below) for homogeneity of variance by implementing the stat-levene-method that is a part of the Python scipy.stats library.



```
stats.levene(data['Entropy_Feature'][data[class] == 'RCHS'],
                 data['Entropy_Feature'][data[class]== 'NCHS'],
                 data['Entropy_Feature'][data[class]== 'UCHS'])
```

Note: The reason I prefer using these methods is that the homogeneity of variance assumption can be checked for each feature of the 8SFM dataset as opposed to other stats methods, in this instance the feature being used for the ANOVA analysis is 'Entropy_Feature'. Any feature of the 8SFM dataset was checked simply by replacing, the Entropy_Feature argument with for instance, 'HeliFeature, AromaFeature etc.' I tested all the eight features and they all led to the same ANOVA conclusion as elaborated below with the ANOVA conclusion reached with the Entropy_Feature. Since this is a one way ANOVA, we therefore only need to test for one feature as the level of categorical variable.

The Levene result obtained was:

Test Statistic = 0.98 and *P-value = 0.053*.

The Levene's test of homogeneity of variance was shown to be non-significant which indicated that the classes (RCHS, NCHS, and UCHS) had related variances. Once the assumptions had been checked, the ANOVA test was then conducted with *Entropy_Feature* being the categorical variable of the three classes (RCHS, NCHS, and UCHS).

Table 7. 4 The Result of ANOVA Model-- Comparing the means between the 3 classes (RCHS, NCHS, UCHS)

	sum_sq	Df	mean_sq	F	PR (>F)	eta_sq	omega_sq
RCHS-NCHS-UCHS	0.013471	2.0	0.006736	76.95955	0.0001	0.28455	0.280394
Residual	0.033871	387	0.000086	---	---	---	---

The first row in table 7.4 (RCHS-NCHS-UCHS) presents the results of the ANOVA model, illustrating the overall experimental effect. The model explains the significance level of variance, $F_{(2,387)} = 76.96$, and p-value < *0.05 (see table PR (>F = 0*.0001))

Note: With p-value being less than 0.05 the H_o is therefore rejected, and H_1 is accepted.

The sum of square (denoted by SSM) is presented in the first column of the model: *SSM =0.013471*. This indicates the extent to which each class variance is explained by the model.

The residual row (second row) is the unsystematic variation in the data. The sum of square residual (denoted by SSR) is known as the unexplained variance (*0.033871)*. In this case, the SSR represents the statistical individual differences in the three classes.

The total variance, *SST =0.047342*, is equal to the sum of SSM + SSR.

The mean-squares (mean_sq) eliminates the bias present in the SSM and SSR, and was used to calculate the F-statistic and omega-squared. The biasedness of SSM and SSR are caused by the number of values summed to calculate them. The mean-squares (MSM and MSR) were calculated as followed:

$$MSM = SSM / df_M = 0.013471 / 2 = 0.0067355$$

$$MSR = SSR / df_R = 0.033871 / 389 = 0.00008752$$

Therefore,

$$F\text{-statistic} = MSM/MSR = 76.96$$

Eta-squared (eta_sq) and omega-squared (omega_sq) indicate the level of impact that the experiment will have in the real world. Omega-squared is considered a better measure of effect size than eta-squared because it is unbiased in its calculation. However, the results of these two measures, eta-squared and omega-squared, are almost the same. This means that both measures agree that the feature in the model (*Entropy_Feature*.) accounts for 28% of the variance in contributing to the analysis of the three classes (RCHS, NCHS, UCHS).

The overall ANOVA model proved to be significant, as the H_0 was rejected at P-value <0.05 (see table 7.4). This confirms that a difference in means exists (Fowler *et al.*, 2013; Fois *et al.*, 2015) somewhere among the three classes. To determine where the difference in means lies, a post hoc test was then conducted.

7.5.1 Tukey's Honest Significant Difference Post Hoc Test

The one one-way ANOVA test was followed up with a post hoc test known as Tukey's Honest Significant Difference test (HSD), to compare the means of each class (Dominguez-Bello et al., 2016) from Entropy_Feature as the level of categorical variable and determine exactly where those differences lie as elaborated in Section 2.11.2.

Table 7. 5 Tukey’s HSD Post-hoc Testing comparison

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1 group2	Meandiff	lower	upper	Reject H_0
RCHS Vs. NCHS	1.4871	1.1926	1.7816	True
RCHS Vs. UCHS	-0.0831	-0.3776	0.211	False
NCHS Vs. UCHS	-1.5702	-1.8625	-1.278	True

The Tukey’s HSD post-hoc test controls for type I error and maintains the family wise error rate (FWER) at 0.05 (Dominguez-Bello *et al.*, 2016) as indicated on the top right corner of Table 7.5 (*FWER = 0.05*). The *group1* and *group2* column indicates the classes that were compared. The *meandiff* column indicates the difference in means of the two groups, calculated as group2 - group1. The *lower* and *upper* columns indicate the lower and upper boundaries of the 95% confidence interval.

Lastly, the *reject* column states whether or not the null hypothesis was rejected. The null hypothesis in this case states that there is no difference in the means of the classes being compared (Fowler *et al.*, 2013; Fois *et al.*, 2015; Dominguez-Bello *et al.*, 2016). The Tukey’s HSD post-hoc test results demonstrated that the difference in the means of the RCHS class and the UCHS class is very miniscule, whereas the mean of the NCHS class differed significantly from the rest. This Tukey’s HSD post-hoc test provided therefore strong evidence at 95% confidence that the sequences labelled as UCHS inferred the same biological functionalities and properties as the protein sequences labelled RCHS, as such H_o should do not be rejected as there is no enough evidence that the means of these two classes differ.

However, on the other hand, the Tukey’s HSD post-hoc test provided strong evidence at 95% confidence that the sequences labelled NCHS do not infer the same biological properties in this case do not possess the same catalytic activities as the sequences labelled as RCHS or UCHS. In this logic, H_o should therefore be rejected and it can be concluded with 95% confidence that the mean of the NCHS do differ with the means of the RCHS and UCHS class.

To further investigate the three classes, a boxplot method was used to visualise the patterns of these three classes.

7.6 Boxplot

To graphically display the patterns (variability or spread in a data set and interquartile range) of the classes RCHS, UCHS, and NCHS, one feature, *AromaFeature*, was selected from the 8SFM dataset for the boxplot observation in figure 7.4.

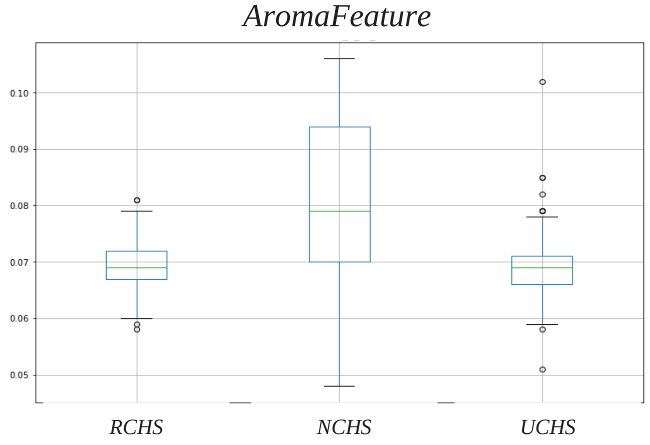


Figure 7. 7 Boxplot of the 3 classes grouped by AromaFeature

Note: Since the eight features of the 8SFM were all found significant, any feature could be used to reach inferential statistics of the two or three classes (RCHS and NCHS or UCHS).

Within the boxes a horizontal green line is drawn indicating the median. Two vertical blue lines, called whiskers, extend from the top and bottom of the box. The bottom whisker goes from quartile one (Q1) to the smallest non-outlier (separate points on the chart), and the top whisker goes from quartile three (Q3) to the largest non-outlier in the dataset. In the boxplot above, four outliers are shown on top of Q3 of the UCHS class, and two outliers are shown on the bottom of Q1. One outlier is shown on top of Q3 of the RCHS, and two outliers on the bottom of Q1. These observed outliers are due to the slight variability in the measurements. They indicate values that were more than the upper limit (Q3) or lesser than the lower limit (Q1). The NCHS class did not present any outliers.

The median of the RCHS class and UCHS class are the same (*0.069*), in contrast to NCHS's median (*0.079*). It is noteworthy that the range (spread of the data) in figure 7.4 reflects the ranges in figures 7.1, 7.2, 7.3, confirming the distribution of the classes using two different statistical methods.

The interquartile range (IQR) of RCHS (*0.072 - 0.066 =0.006*) and UCHS (*0.071-0.065 = 0.006*) is identical, whereas the NCHS class shows an IQR of *0.023.*

In fact, the boxplot has indicated, just as the other statistical analysis carried out above, that RCHS class and UCHS class have the same statistical skewedness patterns, as opposed to NCHS class.

7.5 Summary

In summary, all of the statistical experiments conducted in this chapter have revealed that mathematic statistical estimated approaches can be carried out pertaining to the analysis and interpretation of secondary metabolite genes. The RCHS class and the UCHS class have shown to have the same statistical distributions. These statistical approaches are therefore recommended for the analysis of plant genes involved in the secondary metabolites production, as they are adequate to prove the inferences among plant genes.

Through the use of these exploratory statistical methods, we have been able to see the relatedness of the sequences in the RCHS class and UCHS class. This despite the fact that the UCHS class was made of randomly selected unreviewed CHS sequences, meaning that the catalytic activity of these sequences has not been yet proven. Through this proof-of-concept statistical approach, we aimed to show that these 130 randomly chosen sequences, which constituted the UCHS class, may have the same catalytic activity as the reviewed CHS sequence of the RCHS class, and that the sequences in the NCHS class provide statistical evidence that they differ completely from the other two classes (RCHS and UCHS).

We therefore recommend that these 130 sequences that constituted the UCHS class be manually tested in the lab to see whether the lab findings reflect the findings from our exploratory proof of concept statistical approaches. In the case that 95% of these sequences of the UCHS class do present the same catalytic activities as the sequences of the RCHS class, we can then be certain that these exploratory proofs of concept statistical approaches can be adopted as a new method for classifying secondary genes without the necessity of the lab. Mathematically, we can say that the confidence interval (CI) will be 1-α, where α (alpha) is the error difference above or less than our current 95% confidence. In the case where less than 95% of the sequences in the UCHS are found to have the same catalytic activities as the sequences in the RCHS, then we will be less than 100% certain (1-alpha) that these proof of concept statistical methods do work for accurately classifying secondary metabolite genes.

To our knowledge, there is no previous work that implemented this strategy for studying RCHS using mathematical and statistical techniques. Therefore, we cannot conduct a comparison of this study against previous studies. However, our exploratory proof-of-concept model shows good results for analysing and interpreting secondary metabolite genes.

Chapter Eight Conclusion

8.1 Conclusion

Secondary metabolites are of high interest in medicinal plants because they often represent the majority of active ingredients associated with health-promoting qualities. Very little significant research has been conducted to study key enzyme factors that can categorize or predict a class of secondary metabolite genes. Identification of approaches that are essential to understanding the mechanisms by which secondary metabolite genes can be predicted and analysed have been lacking in the field of Bioinformatics.

Genome analysis of South African medicinal plants has been initiated through the Aspalathus linearis (rooibos) genomics programme at the University of the Western Cape, which encompasses the sequencing of the rooibos genome as well as six diverse transcriptomes. One of the genomic programme's aims is to develop biocomputational approaches for future medicinal plant gene analyses. Analogously, the aims of this exploratory proof-of-concept study were to develop data science techniques which present a dynamic way of mining plant SM genes that may lead to the analysis and understanding of plants SM genes through a different modus operandi.

The study presented data science techniques that can assess, inform and accelerate experimental endeavours on secondary metabolite genes, and provide practitioners guidance on best practices. In the course of this research, a number of methods were tested in preparing and analysing the protein sequence data using various machine learning and statistical techniques. The following sections present the key findings of the study and make recommendations for future work.

8.2 Summary of findings

This study attempted to address two key research questions related to protein sequence data of plant secondary metabolite genes:

1. *Can machine learning algorithms be trained to recognize plant secondary metabolite genes involved in the production of medicinally active compounds (e.g. polyphenols)?*
2. *Can mathematical and statistical estimated approaches be carried out pertaining to the preparation, analysis and interpretation of secondary metabolite genes?*

This section will present the key findings from the study in response to these research questions.

To begin with, existing literature on plant secondary metabolite genes and the identification of chalcone synthase was reviewed to provide the contextual basis for the study, and previous studies were drawn on to identify the most appropriate machine learning and statistical approaches for this study. Chapter 2 presented the background of biological properties of medicinal plants, secondary metabolite genes, polyphenols, biosynthetic pathways, and the identification of chalcone synthase which is used as true positive set in the analysis performed in this work. Different studies on computational biology were also investigated, along with statistical analysis in the field of plant genes and other organisms. Machine learning techniques were explored, such as unsupervised and supervised learning techniques, which have been used in the field of bioinformatics. A number of supervised classification models were identified as the most appropriate machine learning techniques to address the research problem. The knowledge obtained from these different studies informed the design and implementation of this exploratory proof-of-concept study.

Chapter 4 identified feature engineering techniques that were best suited to address the challenges that bioinformatics data poses in its application to computational analysis. In this case, the protein sequence data in its raw form presented a string of alphabetical letters with varying lengths. Previous studies have addressed this issue by converting the protein sequences into numerical values before performing data analysis on them. This study followed the same approach of converting amino acids into numerical values, transforming the dataset into three exploratory proof-of-concept sets of engineered features: Frequency-based features; Value-based features; and Amino acid relative frequency-based features. The frequency-based features and value-based features were combined into one dataset called 8SFM, and the Amino acid relative frequency feature dataset was labelled 20RFFM. Having converted the raw protein sequence data of the three classes (RCHS, NCHS, and UCHS) into numerical datasets (8SFM and 20RFFM), secondary metabolite (SM) data mining then became possible.

Data visualisation techniques were then discussed in Chapter 5 as a tool to gain insight into the patterns that exist within the datasets. With unsupervised techniques such as PCA, the SM dataset (20RFFM) showed that RCHS and UCHS classes were clustering in a specific location, whereas the NCHS class tended to be scattered. This foreshadowed the findings that were confirmed in the machine learning and statistical analysis phases of the study, which found that RCHS and UCHS shared similar biological properties. This finding of clustered classes informed the selection of supervised machine learning classification models as an appropriate model, given the behaviour of the data set.

Chapter 5 also explored the use of data visualisation as a tool for feature selection. By applying feature selection techniques such as Forest of Trees, ANOVA and Mutual Information, the features could be presented through a visual ranking display to indicate their significance. Findings from these feature selection techniques indicated that all eight features in the 8SFM dataset were significant. These exploratory proof-of-concept data visualisation and feature selection techniques were shown to be helpful in the interpretation of protein sequence data of SM genes.

Once the feature selection process was complete, two exploratory proof-of-concept supervised machine learning classification models were built: a binary classification model and a multi classification model. The development and performance of these models were discussed in Chapter 6.

With the binary classification model, it was shown that the 2NN showed a very high predictive power over the rest of the binary models. In training the model, the RCHS was used as the true positive set, while the NCHS was used as the control dataset (true negative set). The 2NN model was trained and tested (80% and 20% respectively) on a 10-fold cross-validation with two different datasets: 20RFFM and 8SFM. The AUC was used as a performance evaluation metric of these two 2NN models, resulting in the 20RFFM dataset indicating a higher predictive power over the 8SFM dataset. The proof-of-concept 20RFFM 2NN model was further evaluated using confusion matrix and presented good results, achieving a 94.2% average model accuracy, a precision of 94.7% and a sensitivity of 97.3% at predicting RCHS in set of diverse secondary metabolite genes (i.e. a test dataset made up of both RCHS and NCHS protein sequences).

With the multi classification model, the SVM was shown to present a very high predictive power over the rest of the multi classification models, achieving a 91% average accuracy model. Three classes were used to train the multi classification models: RCHS, NCHS, and UCHS.

The motivation behind performing a multiclass classifier to build a proof-of-concept model for classifying chalcone-synthase in secondary metabolite genes is that curated (reviewed) secondary metabolite genes are limited. As UCHS is a class of unreviewed chalcone synthase, it was important to observe the treatment of this class via the multi classification model. The AUC metric showed that the multi classifier ranked the sequences in the RCHS and UCHS classes higher than the sequences in the NCHS class. This led to the belief that either the UCHS or RCHS class could be treated as a true positive set. This outcome is of high importance, as it means that the UCHS and RCHS classes can ultimately be combined to increase the size of the true positive set, therefore addressing the problem of a limited pool of curated secondary metabolite genes. This same technique could be applied, in theory, to unreviewed non-chalcone synthase secondary metabolite genes in order to increase the size of the true negative as well. Increasing the true positive and true negative datasets would allow for training of the machine learning classifiers on much larger datasets, therefore leading to improved generalisation power of the models.

Multi classification models can also be used to classify different secondary metabolite genes where the goal is not to only predict one type of secondary metabolite gene, but to classify a range of secondary metabolite genes (enzyme classification) from a cluster of metabolic gene loci within the plant kingdom. Although this study focused strictly on chalcone synthase secondary metabolite genes, this multi classification model could be applied in future studies to classify chalcone synthase genes as well as other types of secondary metabolite genes (e.g. stelibene synthase, aloesome synthase, etc.). These outputs from Chapter 6 show that machine learning classification models can be effectively trained to predict and classify plant secondary metabolite genes involved in the production of medicinally active compounds.

Chapter 7 proposes exploratory proof-of-concept mathematical and statistical approaches that can be applied to study the statistical inferences between RCHS, NCHS, and UCHS. The chi-square hypothesis test and its post hoc test (Bonferroni Correction) indicated with 95% confidence that a relationship exists between RCHS and UCHS. Despite the fact that the UCHS class consists of protein sequences which have not yet been reviewed, the chi-square and the Bonferroni Correction tests suggest with 95% certainty that the UCHS class presents the same biological secondary metabolite properties as the RCHS class. The same chi-square and Bonferroni Correction tests have shown that the NCHS class does not present a relationship with either the RCHS or the UCHS.

Further analyses were conducted with ANOVA and its post hoc test (Tukey's HSD), which show with 95% confidence that the means of the RCHS and UCHS do not differ, whereas these means differ with that of the NCHS class. The fact that RCHS and UCHS have the same mean implies that either of these classes (RCHS, UCHS) could be used as a true positive set to be analysed against the NCHS class, and the outcome would be the same.

Chapter 7 also presented the results of the boxplot method that was applied to the datasets. The results show that UCHS and RCHS have very similar distribution, whereas the NCHS's distribution differed from RCHS and UCHS. The skewedness patterns that are revealed through the boxplot method demonstrate how similar or different these classes of secondary metabolite genes are to one another. These statistical methods (chi-square, ANOVA, boxplot) present an exploratory approach that could be more broadly used to infer the distribution that exists among a broader range of secondary metabolite genes.

The results of the statistical methods suggest with 95% confidence the RCHS and UCHS possess the same biological properties. Therefore, either of these protein sequence datasets could be used as a true positive set, or ultimately combined into one large class, increasing the size of RCHS from 130 to 260. This finding suggests that future work that makes use of the same dataset could start with a larger true positive set from the outset. Moreover, the remaining sequences from the 2961 UCHS class that were collected from UniProtKB-TREMBL could be further investigated with the same statistical approaches outlined above to confirm their similarity (or not) to the RCHS class. In doing so, the size of RCHS class could potentially be increased exponentially, enhancing the generalisation power of the models.

The outcomes of these five different statistical methods are very important as they are in accordance with the outcomes of the multi classification models and binary models seen in Chapter 6, which ranked the RCHS and UCHS classes higher than the NCHS class in multiclassification, and the RCHS higher than the NCHS in the binary classification. As the NCHS class is made up of reviewed sequences that are known to not involve the same catalytic activities as reviewed chalcone synthase (RCHS), these statistical outcomes confirm that mathematical and statistical estimated approaches can be applied to analyse and interpret secondary metabolite genes.

8.3 Future Work

This comprehensive study consisted of several components which could be expanded on in future studies conducted in bioinformatics. Some of the possibilities for future studies include:

- Studies which attempt to establish guidelines or standards on collecting secondary metabolite data.
- Studies which make use of the baseline database that was built to store the data for this particular study, to build a graph database that uses graph structures for semantic queries with nodes, edges and properties to relate the stored protein sequences and their information.
- Further laboratory analysis of the class of UCHS to determine the number of protein sequences within this class that are indeed true positive chalcone synthase, to evaluate the accuracy and real-life application of the model. If the outcome of the lab work is positive, this would boost the dataset size of RCHS and allow for building of classifier models trained on much larger datasets.
- Studies which attempt to predict and analyse other secondary metabolite enzymes using the machine learning and mathematical and statistical approaches developed in this study.
- Further study as part of the Aspalathus linearis genomics programme at the University of the Western Cape to conduct the genome analysis of Rooibos using these data science techniques.

References

- Abbot, V., Sharma, P., Dhiman, S., Noolvi, M. N., Patelc, H. M., & Varun Bhardwaj, V. (2017). Small hybrid heteroaromatics: Resourceful biological tools in cancer research. *RSC Advances*, *7*, 28313-28349.
- Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature Biotechnology*, *33* (8), 831.
- Anderberg, M. R. (2014). *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks* (Vol. 19). Academic press.
- Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31* (2), 166-169.
- Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, *12* (7), 878.
- Arifin, A. Z., & Asano, A. (2006). Image segmentation by histogram thresholding using hierarchical cluster analysis. *Pattern Recognition Letters*, *27*(13), 1515-1521.
- Armstrong, R. A. (2014). When to use the Bonferroni correction**.** *Ophthalmic and Physiological Optics, 34* (5), 502-508.
- Auer, P., Burgsteiner, H., & Maass, W. (2008). A learning rule for very simple universal approximators consisting of a single layer of perceptrons. *Neural Networks*, *21*(5), 786-795.
- Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20-29.
- Basheer, L., & Kerem, Z. (2015). Interactions between CYP3A4 and dietary polyphenols. *Oxidative Medicine and Cellular Longevity*, *2015*.
- Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35* (8), 1798-1828.
- Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In G. Montavon, G.B. Orr and K.R. Müller (Eds.), *Neural networks: Tricks of the trade* (pp. 437-478). Berlin: Heidelberg.
- Blagus, R., & Lusa, L. (2012, December). Evaluation of SMOTE for high-dimensional class-imbalanced microarray data. In *2012 11th International conference on machine learning and applications (ICLMA),* (Vol. 2, pp. 89-94). IEEE.
- Bodi, D., Ronczka, S., Gottschalk, C., Behr, N., Skibba, A., Wagner, M., ... & These, A. (2014). Determination of pyrrolizidine alkaloids in tea, herbal drugs and honey. *Food Additives & Contaminants: Part A*, *31* (11), 1886-1895.
- Boudreau, K. J., & Lakhani, K. R. (2015). "Open" disclosure of innovations, incentives and follow-on reuse: Theory on processes of cumulative innovation and a field experiment in computational biology. *Research Policy*, *44* (1), 4-19.
- Brauer, F., Castillo-Chavez, C., & Castillo-Chavez, C. (2012). *Mathematical models in population biology and epidemiology (*Vol. 1). New York: Springer.
- Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., ... & Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, *33* (2), 155.
- Burgess, S., & Smith, G. D. (2017). Mendelian randomization implicates high-density lipoprotein cholesterol–associated mechanisms in etiology of age-related macular degeneration. *Ophthalmology*, *124* (8), 1165-1174.
- Calabriso, N., Scoditti, E., Massaro, M., Pellegrino, M., Storelli, C., Ingrosso, I., ... & Carluccio, M. A. (2016). Multiple anti-inflammatory and anti-atherosclerotic properties of red wine polyphenolic extracts: differential role of hydroxycinnamic acids, flavonols and stilbenes on endothelial inflammatory gene expression. *European Journal of Nutrition*, *55* (2), 477-489.
- Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14, p.2.
- Calvo-Flores, F. G., Dobado, J. A., Isac-García, J., & Martín-Martínez, F. J. (2015). *Lignin and lignans as renewable raw materials: chemistry, technology and applications*. John Wiley & Sons.
- Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis?. *Journal of the ACM (JACM)*, *58* (3), 11.
- Cheetham, P. J., & Katz, A. E. (2013). Diet and prostate cancer: A holistic approach to management. In A.K. Tewari (Ed.), *Prostate cancer: A comprehensive perspective* (pp. 355-367). London: Springer.
- Clauss, S., Scherr, J., Hanley, A., Schneider, J., Klier, I., Lackermair, K., ... & Nickel, T. (2017). Impact of polyphenols on physiological stress and cardiac burden in marathon runners–results from a substudy of the BeMaGIC study. *Applied Physiology, Nutrition, and Metabolism*, *42* (5), 523-528.
- Clemensen, A. K., Provenza, F. D., Lee, S. T., Gardner, D. R., Rottinghaus, G. E., & Villalba, J. J. (2017). Plant secondary metabolites in alfalfa, birdsfoot trefoil, reed canarygrass, and tall fescue unaffected by two different nitrogen sources. *Crop Science,* 57 (2), 964-970.
- Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., ... & Tiedje, J. M. (2013). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42* (D1), D633-D642.
- Cseke, L. J., Kirakosyan, A., Kaufman, P. B., Warber, S., Duke, J. A., & Brielmann, H. L. (2016). *Natural products from plants*. CRC press.
- D'Agostino, R. (2017). *Goodness-of-fit-techniques*. Routledge.
- Dahl, G. E., Jaitly, N., & Salakhutdinov, R. (2014). Multi-task neural networks for QSAR predictions. *arXiv preprint arXiv:1406.1231*.

- Demirbas, A. (2017). Higher heating values of lignin types from wood and non-wood lignocellulosic biomasses. *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, *39* (6), 592-598.
- Dewick, P. M. (2009). The shikimate pathway: Aromatic amino acids and phenylpropanoids. *Medicinal Natural Products*, *137*, 86.
- Dominguez-Bello, M. G., De Jesus-Laboy, K. M., Shen, N., Cox, L. M., Amir, A., Gonzalez, A., ... & Mendez, K. (2016). Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. *Nature Medicine*, *22* (3), 250.
- Dushenkov, V. (2016). Biodiversity of medicinal plants in the highlands: problems and perspectives. In M.M. Yakubova (Ed.), *The state of biological resources in mountain regions in relation to climate change* (pp. 191-192). Khorog, Tajikistan: Donish.
- Działo, M., Mierziak, J., Korzun, U., Preisner, M., Szopa, J., & Kulma, A. (2016). The potential of plant phenolics in prevention and therapy of skin disorders. *International Journal of Molecular Sciences*, *17* (2), 160.
- Eduati, F., Mangravite, L. M., Wang, T., Tang, H., Bare, J. C., Huang, R., ... & Zhan, X. (2015). Prediction of human population responses to toxic compounds by a collaborative competition. *Nature Biotechnology*, *33* (9), 933.
- Eickholt, J., & Cheng, J. (2013). DNdisorder: predicting protein disorder using boosting and deep networks. *BMC Bioinformatics*, *14* (1), 88.
- Fakoor, R., Ladhak, F., Nazi, A., & Huber, M. (2013, June). Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the International Conference on Machine Learning* (Vol. 28).
- Farzaneh, V., & Carvalho, I. S. (2015). A review of the health benefit potentials of herbal plant infusions and their mechanism of actions. *Industrial Crops and Products*, *65*, 247-258.
- Fitzgerald, T. W., Gerety, S. S., Jones, W. D., Van Kogelenberg, M., King, D. A., McRae, J., ... & Barrett, D. M. (2015). Large-scale discovery of novel genetic causes of developmental disorders. *Nature*, *519* (7542), 223.
- Fois, M., Fenu, G., Lombrana, A. C., Cogoni, D., & Bacchetta, G. (2015). A practical method to speed up the discovery of unknown populations using Species Distribution Models. *Journal for Nature Conservation*, *24*, 42-48.
- Fowler, J., Cohen, L., & Jarvis, P. (2013). *Practical statistics for field biology*. John Wiley & Sons.
- Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., ... & Jensen, L. J. (2012). STRING v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, *41* (D1), D808-D815.
- Francisco, M., Tortosa, M., Martínez-Ballesta, M., Velasco, P., íía-Viguera, C., & Moreno, D. A. (2017). Nutritional and phytochemical value of Brassica crops from the agri-food perspective. *Annals of Applied Biology*, *170* (2), 273-285.
- Geethangili, M., & Tzeng, Y. M. (2011). Review of pharmacological effects of Antrodia camphorata and its bioactive compounds. *Evidence-Based Complementary and Alternative Medicine*, *2011*.
- Ghasemzadeh, A., & Jaafar, H. (2011). Effect of CO2 enrichment on synthesis of some primary and secondary metabolites in ginger (Zingiber officinale Roscoe). *International Journal of Molecular Sciences*, *12*(2), 1101-1114.
- Gill, U. S., Uppalapati, S. R., Gallego-Giraldo, L., Ishiga, Y., Dixon, R. A., & Mysore, K. S. (2017). Metabolic flux towards the (iso)flavonoid pathway in lignin modified alfalfa lines induces resistance against Fusarium oxysporum f. sp. medicaginis. *Plant, Cell & Environment, 41* (9), 1997-2007.
- Glaab, E., Bacardit, J., Garibaldi, J. M., & Krasnogor, N. (2012). Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. *PloS One*, *7* (7), e39932.
- Gnauck, A. H., & Straškraba, M. (2013). *Freshwater ecosystems: modelling and simulation* (Vol. 8). Elsevier.
- Goetz, M. E., Judd, S. E., Safford, M. M., Hartman, T. J., McClellan, W. M., & Vaccarino, V. (2016). Dietary flavonoid intake and incident coronary heart disease: The Reasons for Geographic and Racial Differences in Stroke (REGARDS) study. *The American Journal of Clinical Nutrition*, *104* (5), 1236-1244.
- Goličnik, M. (2011). Evaluation of enzyme kinetic parameters using explicit analytic approximations to the solution of the Michaelis–Menten equation. *Biochemical Engineering Journal*, *53* (2), 234-238.
- Gomes, G. L. G. C., Carbonari, C. A., Velini, E. D., Trindade, M. L. B., & Silva, J. R. M. (2015). Extraction and simultaneous determination of glyphosate, ampa and compounds of the shikimic acid pathway in plants. *Planta Daninha*, *33* (2), 295-304.
- Hair Jr, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage Publications.
- Han, S., Li, D., Trost, E., Mayer, K. F., Vlot, A. C., Heller, W., ... & Rothballer, M. (2016). Systemic responses of barley to the 3-hydroxy-decanoyl-homoserine lactone producing plant beneficial endophyte Acidovorax radicis N35. *Frontiers in Plant Science*, *7*.
- Haslam, E. (2014). *The shikimate pathway: biosynthesis of natural products series*. Elsevier.
- Heleno, S. A., Martins, A., Queiroz, M. J. R., & Ferreira, I. C. (2015). Bioactivity of phenolic acids: Metabolites versus parent compounds: A review. *Food Chemistry, 173,* 501-513.
- Heller, P., Tripp, H. J., Turk-Kubo, K., & Zehr, J. P. (2014). ARBitrator: a software pipeline for on-demand retrieval of auto-curated nifH sequences from GenBank. *Bioinformatics*, *30* (20), 2883-2890.
- Hertroijs, D. F., Elissen, A. M., Brouwers, M. C., Schaper, N. C., Köhler, S., Popa, M. C., ... & Ruwaard, D. (2018). A risk score including body mass index, glycated haemoglobin and triglycerides predicts future glycaemic control in people with type 2 diabetes. *Diabetes, Obesity and Metabolism*, *20* (3), 681-688.
- Hosseini, S., Gholami, M. R., & Haghgu, M. (2016). A computational study of lipophilicity of E-2-arylmethylen-1-tetralones and their heteroanalogues using QSAR and DFT based molecular surface electrostatic potential. *Journal of Physical & Theoretical Chemistry*, *13* (2), 171-177.
- Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., ... & Gottardo, R. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12* (2), 115.
- Hussain, G., Rasul, A., Anwar, H., Aziz, N., Razzaq, A., Wei, W., ... & Li, X. (2018). Role of plant derived alkaloids and their mechanism in neurodegenerative disorders. *International Journal of Biological Sciences*, *14* (3), 341.

- Ibdah, M., Martens, S., & Gang, D. R. (2017). Biosynthetic pathway and metabolic engineering of plant dihydrochalcones. *Journal of Agricultural and Food Chemistry*.
- Jang, S., Gang, H., Kim, B. G., & Choi, K. Y. (2017). FCS and ECH dependent production of phenolic aldehyde and melanin pigment from l-tyrosine in Escherichia coli. *Enzyme and Microbial Technology, 112* (May 2018), 59-64.
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2015). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44* (D1), D457-D462.
- Kaul, S., Gupta, S., Sharma, S., & Dhar, M. K. (2017). The fungal endobiome of medicinal plants: A prospective source of bioactive metabolites. In *Medicinal plants and fungi: Recent advances in research and development* (pp. 167-228). Singapore: Springer.
- Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A., & Medema, M. H. (2017). plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Research, 45* (W1), W55-W63.
- Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, *26* (7), 990-999.
- Khalid, S., Khalil, T., & Nasreen, S. (2014, August). A survey of feature selection and feature extraction techniques in machine learning. In *Science and Information Conference (SAI), 2014* (pp. 372-378). IEEE.
- Khoshgoftaar, T. M., Golawala, M., & Van Hulse, J. (2007, October). An empirical study of learning from imbalanced data using random forest. In *19th International conference on tools with artificial intelligence (ICTAI), 2007* (Vol. 2, pp. 310-317). IEEE.
- Kleinbaum, D. G., Kupper, L. L., Nizam, A., & Rosenberg, E. S. (2013). *Applied regression analysis and other multivariable methods*. Cengage Learning.
- Krechmer, J. E., Coggon, M. M., Massoli, P., Nguyen, T. B., Crounse, J. D., Hu, W., ... & Nowak, J. B. (2015). Formation of low volatility organic compounds and secondary organic aerosol from isoprene hydroxyhydroperoxide low-NO oxidation. *Environmental Science & Technology*, *49* (17), 10330-10339.
- Krivoruchko, A., & Nielsen, J. (2015). Production of natural products through metabolic engineering of Saccharomyces cerevisiae. *Current Opinion in Biotechnology*, *35*, 7-15.
- Kumar, N. B., Pow-Sang, J., Egan, K. M., Spiess, P. E., Dickinson, S., Salup, R., ... & Parnes, H. L. (2015). Randomized, placebo-controlled trial of green tea catechins for prostate cancer prevention. *Cancer Prevention Research, 8* (10), 879-887.
- Kyte, J., & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, *157*(1), 105-132.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521* (7553), 436.
- Leung, M. K., Xiong, H. Y., Lee, L. J., & Frey, B. J. (2014). Deep learning of the tissue-regulated splicing code. *Bioinformatics*, *30* (12), i121-i129.
- Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16* (6), 321.
- Lin, D., Xiao, M., Zhao, J., Li, Z., Xing, B., Li, X., ... & Chen, S. (2016). An overview of plant phenolic compounds and their importance in human nutrition and management of type 2 diabetes. *Molecules, 21*(10), 1374.
- Liu, M., & Sadovy, Y. (2004). The influence of social factors on adult sex change and juvenile sexual differentiation in a diandric, protogynous epinepheline, Cephalopholis boenak (Pisces, Serranidae). *Journal of Zoology*, *264*(3), 239-248.
- Lynch, J. H., Orlova, I., Zhao, C., Guo, L., Jaini, R., Maeda, H., ... & Pilot, G. (2017). Multifaceted plant responses to circumvent Phe hyperaccumulation by downregulation of flux through the shikimate pathway and by vacuolar Phe sequestration. *The Plant Journal*, *92* (5), 939-950.
- Madzarov, G., & Gjorgjevikj, D. (2010, June). Evaluation of distance measures for multi-class classification in binary svm decision tree. In *International Conference on Artificial Intelligence and Soft Computing* (pp. 437-444). Berlin, Heidelberg: Springer
- Manach, C., Milenkovic, D., Wiele, T., Rodriguez-Mateos, A., Roos, B., Garcia-Conesa, M. T., ... & Morand, C. (2017). Addressing the inter-individual variation in response to consumption of plant food bioactives: Towards a better understanding of their role in healthy aging and cardiometabolic risk reduction. *Molecular Nutrition & Food Research*, *61* (6).
- Martinez, M. (2011). Plant protein-coding gene families: emerging bioinformatics approaches. *Trends in Plant Science*, *16* (10), 558-567.
- Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., & Saez-Rodriguez, J. (2013). Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. *PLoS one*, *8* (4), e61318.
- Merel, S., & Zwiener, C. (2016). Accurate mass screening and data evaluation approaches for ozonation by-products in wastewater treatment plant effluents. In J.E. Drewes & T. Letzel (Eds.) *Assessing transformation products of chemicals by non-target and suspect screening− strategies and workflows Volume 2* (pp. 3-27). American Chemical Society.
- Merelli, I., Tordini, F., Drocco, M., Aldinucci, M., Liò, P., & Milanesi, L. (2015). Integrating multi-omic features exploiting Chromosome Conformation Capture data. *Frontiers in Genetics*, *6*, 40.
- Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). *Machine learning: An artificial intelligence approach*. Springer Science & Business Media.
- Murphy, K. P. (2012). *Machine learning: A probabilistic perspective.* Cambridge: The MIT Press.
- Musicant, D. R., Kumar, V., & Ozgur, A. (2003, May). Optimizing F-measure with support vector machines. In *FLAIRS conference* (pp. 356-360).
- Nepusz, T., Yu, H., & Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods*, *9* (5), 471.
- Panche, A. N., Diwan, A. D., & Chandra, S. R. (2016). Flavonoids: An overview. *Journal of Nutritional Science*, *5*.
- Pfister, B., Lu, K. J., Eicke, S., Feil, R., Lunn, J. E., Streb, S., & Zeeman, S. C. (2014). Genetic evidence that chain length and branch point distributions are linked determinants of starch granule formation in Arabidopsis. *Plant Physiology*, *165* (4), 1457-1474.

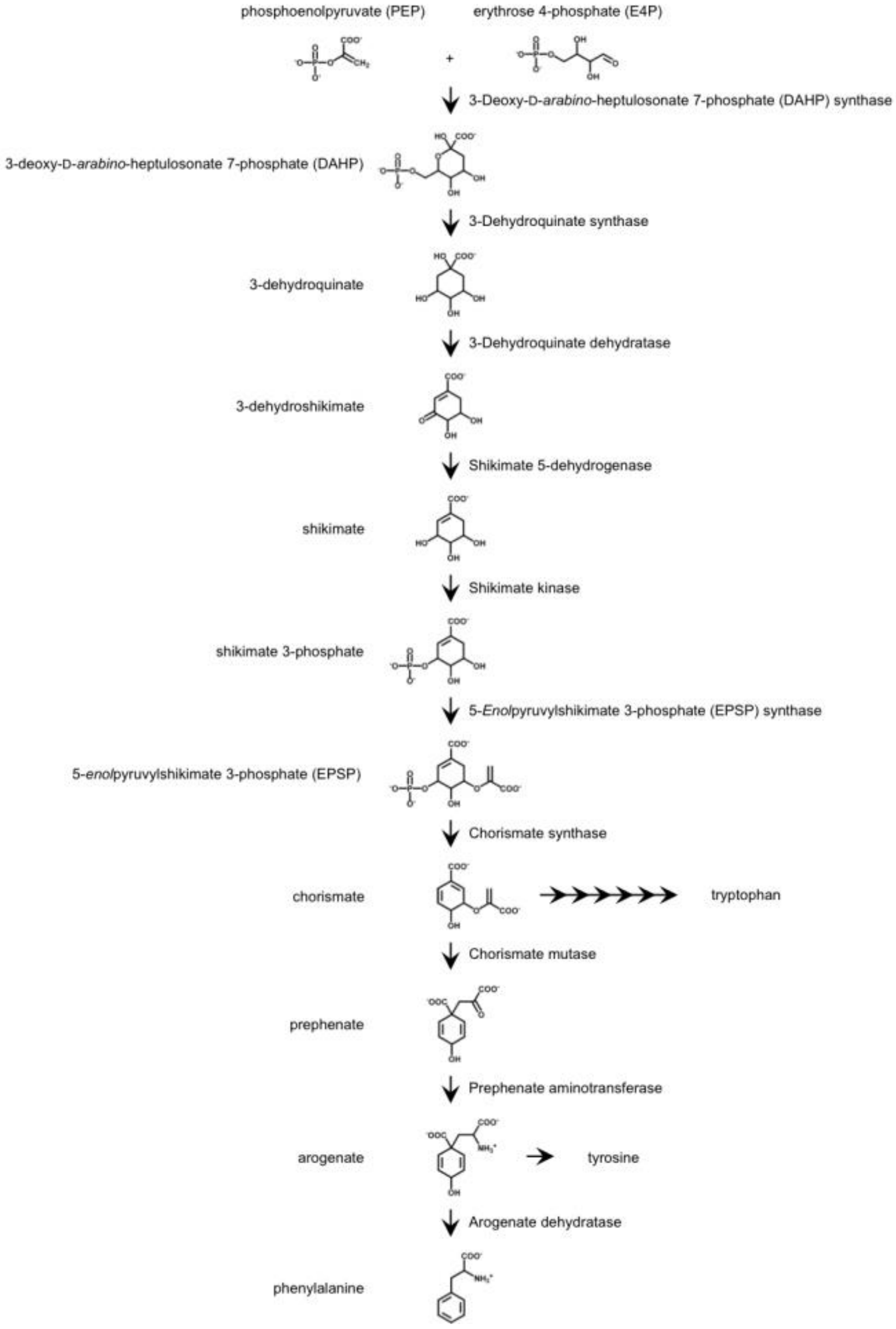
- Qu, Y. H., Yu, H., Gong, X. J., Xu, J. H., & Lee, H. S. (2017). On the prediction of DNA-binding proteins only from primary sequences: A deep learning approach. *PLoS one*, 12 (12), e0188129.
- Ratnam, W., Choong, C. Y., & Javed, M. A. (2017). Development of genomic resources and assessing their potential for accelerated Acacia breeding. In S.N.A. Abdullah, H. Chai-Ling, & C. Wagstaff (Eds.), *Crop Improvement* (pp. 117-135). Cham: Springer.
- Rehman, S. (2016). Endophytes: The producers of important functional metabolites. *International Journal of Current Microbiology and Applied Sciences*, *5* (5), 377-391.
- Reinisalo, M., Kårlund, A., Koskela, A., Kaarniranta, K., & Karjalainen, R. O. (2015). Polyphenol stilbenes: Molecular mechanisms of defence against oxidative stress and aging-related diseases. *Oxidative Medicine and Cellular Longevity*, *2015*.
- Robert, X., & Gouet, P. (2014). Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Research*, *42* (W1), W320-W324.
- Robinson, M. M., & Zhang, X. (2011). The world medicines situation 2011, traditional medicines: Global situation, issues and challenges. *Geneva: World Health Organization*, 1-4.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Berg, A. C. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115* (3), 211-252.
- Saltveit, M. E. (2017). Synthesis and metabolism of phenolic compounds. *Fruit and Vegetable Phytochemicals: Chemistry and Human Health, 2 Volumes*, 115.
- Saxena, M., Saxena, J., Nema, R., Singh, D., & Gupta, A. (2013). Phytochemistry of medicinal plants. *Journal of Pharmacognosy and Phytochemistry, 1* (6).
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85-117.
- Shanab, A. A., Khoshgoftaar, T. M., & Wald, R. (2012, May). Robustness of threshold-based feature rankers with data sampling on noisy and imbalanced data. In *FLAIRS Conference*.
- Shanab, A. A., Khoshgoftaar, T. M., Wald, R., & Napolitano, A. (2012, August). Impact of noise and data sampling on stability of feature ranking techniques for biological datasets. In *13th International conference on information reuse and integration (IRI),* (pp. 415-422). IEEE.
- Shimizu, Y., Ogata, H., & Goto, S. (2017). Type III polyketide synthases: Functional classification and phylogenomics. *ChemBioChem*, *18* (1), 50-65.
- Slimestad, R., Torskangerpoll, K., Nateland, H. S., Johannessen, T., & Giske, N. H. (2005). Flavonoids from black chokeberries, Aronia melanocarpa. *Journal of Food Composition and Analysis*, *18*(1), 61-68.
- Smith, I. (2015). *Impact of grape pomace on growth performance and blood chemistry of young rats* (Doctoral dissertation, North Carolina Agricultural and Technical State University).
- Sochor, J., Skutkova, H., Babula, P., Zitka, O., Cernei, N., Rop, O., ... & Kizek, R. (2011). Mathematical evaluation of the amino acid and polyphenol content and antioxidant activities of fruits from different apricot cultivars. *Molecules*, *16* (9), 7428-7457.
- Sønderby, S. K., & Winther, O. (2014). Protein secondary structure prediction with long short term memory networks. *arXiv preprint arXiv:1412.7828*.
- Stevenson, P. C., Nicolson, S. W., & Wright, G. A. (2017). Plant secondary metabolites in nectar: impacts on pollinators and ecological functions. *Functional Ecology, 31* (1), 65-75.
- Sun, W., Meng, X., Liang, L., Jiang, W., Huang, Y., He, J., ... & Wang, L. (2015). Molecular and biochemical analysis of chalcone synthase from Freesia hybrid in flavonoid biosynthetic pathway. *PLoS One*, *10* (3), e0119054.
- Tedesco, I., Carbone, V., Spagnuolo, C., Minasi, P., & Russo, G. L. (2015). Identification and quantification of flavonoids from two southern Italian cultivars of Allium cepa L., Tropea (Red Onion) and Montoro (Copper Onion), and their capacity to protect human erythrocytes from oxidative stress. *Journal of Agricultural and Food Chemistry*, *63* (21), 5229-5238.
- Toldra, F. (2017). *Advances in food and nutrition research (*Vol. 82). Academic Press.
- Tresserra-Rimbau, A., Rimm, E. B., Medina-Remón, A., Martínez-González, M. A., De la Torre, R., Corella, D., ... & Fiol, M. (2014). Inverse association between habitual polyphenol intake and incidence of cardiovascular events in the PREDIMED study. *Nutrition, Metabolism and Cardiovascular Diseases*, *24* (6), 639-647.
- Tullius, R. M. (2017). *High-throughput biosensing using chiral plasmonic nanostructures* (Doctoral dissertation, University of Glasgow).
- Upadhyay, A., Upadhyaya, I., Kollanoor-Johny, A., & Venkitanarayanan, K. (2014). Combating pathogenic microorganisms using plant-derived antimicrobials: a minireview of the mechanistic basis. *BioMed Research International*, *2014*.
- Urbanowicz, R. J., Granizo-Mackenzie, A., & Moore, J. H. (2012). An analysis pipeline with statistical and visualization-guided knowledge discovery for michigan-style learning classifier systems. *IEEE computational intelligence magazine*, *7* (4), 35-45.
- Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9 (Nov), 2579-2605.
- Van Verk, M. C., Bol, J. F., & Linthorst, H. J. (2011). Prospecting for genes involved in transcriptional regulation of plant defenses, a bioinformatics approach. *BMC Plant Biology*, *11* (1), 88.
- Van Wyk, B. E., & Wink, M. (2017). *Medicinal plants of the world (Ed. 2).* CABI.
- Veal, C., Dowdy, J., Brockner, B., Anderson, D. T., Ball, J. E., & Scott, G. (2018, April). Generative adversarial networks for ground penetrating radar in hand held explosive hazard detection. In S.S. Bishop & J.C. Isaacs (Eds.) *Detection and sensing of mines, explosive objects, and obscured targets XXIII* (Vol. 10628, p. 106280T). International Society for Optics and Photonics.
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27* (5), 1413-1432.
- Vidal, R., Ma, Y., & Sastry, S. S. (2016). Principal component analysis. In *Generalized principal component analysis* (pp. 25-62). New York, NY: Springer.

- Wald, R., Khoshgoftaar, T. M., & Shanab, A. A. (2012, October). The effect of measurement approach and noise level on gene selection stability. In *International conference on bioinformatics and biomedicine (BIBM),* (pp. 1-5). IEEE.
- Wald, R., Khoshgoftaar, T. M., & Shanab, A. A. (2013, November). Comparison of two frameworks for measuring the stability of gene-selection techniques on noisy class-imbalanced data. In *25th International conference on tools with artificial intelligence (ICTAI),* (pp. 881-888). IEEE.
- Wang, X., & Peng, Z. (2014). Method of moments for estimating uncertainty distributions. *Journal of Uncertainty Analysis and Applications*, *2* (1), 5.
- Wang, X., Ouyang, Y., Liu, J., Zhu, M., Zhao, G., Bao, W., & Hu, F. B. (2014). Fruit and vegetable consumption and mortality from all causes, cardiovascular disease, and cancer: Systematic review and dose-response meta-analysis of prospective cohort studies. *BMJ*, *349*, g4490.
- Wilkinson, D. J. (2011). *Stochastic modelling for systems biology*. CRC press.
- Williams, J. N., Trejo, I., & Schwartz, M. W. (2017). Commonness, rarity, and oligarchies of woody plants in the tropical dry forests of Mexico. *Biotropica*, *49* (4), 493-501.
- Xie, Y., Xing, F., Kong, X., Su, H., & Yang, L. (2015, October). Beyond classification: Structured regression for robust cell detection using convolutional neural network. In *International conference on medical image computing and computer-assisted intervention* (pp. 358-365). Cham: Springer.
- Yadav, R., Khare, R. K., & Singhal, A. (2017). Qualitative phytochemical screening of some selected medicinal plants of Shivpuri district (MP). *International Journal of Life Sciences, 3* (1), 844-847.
- Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. In M. Jordan, Y. LeCun, & S.A. Solla (Eds.) *Advances in neural information processing systems* (pp. 3320-3328).
- Yu, H. N., Wang, L., Sun, B., Gao, S., Cheng, A. X., & Lou, H. X. (2015). Functional characterization of a chalcone synthase from the liverwort Plagiochasma appendiculatum. *Plant Cell Reports*, *34* (2), 233-245.
- Yu, S., Tranchevent, L., Liu, X., Glanzel, W., Suykens, J. A., De Moor, B., & Moreau, Y. (2012). Optimized data fusion for kernel k-means clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34* (5), 1031-1039.
- Zabala, G., Zou, J., Tuteja, J., Gonzalez, D. O., Clough, S. J., & Vodkin, L. O. (2006). Transcriptome changes in the phenylpropanoid pathway of Glycine max in response to Pseudomonas syringae infection. *BMC Plant Biology*, *6*(1), 26.
- Zahiri, J., Hannon Bozorgmehr, J., & Masoudi-Nejad, A. (2013). Computational prediction of protein–protein interaction networks: algorithms and resources. *Current Genomics*, *14*(6), 397-414.
- Zamora-Ros, R., Knaze, V., Rothwell, J. A., Hémon, B., Moskal, A., Overvad, K., ... & Touillaud, M. (2016). Dietary polyphenol intake in Europe: the European Prospective Investigation into Cancer and Nutrition (EPIC) study. *European Journal of Nutrition*, *55* (4), 1359-1375.
- Zelditch, M. L., Swiderski, D. L., & Sheets, H. D. (2012). *Geometric morphometrics for biologists: a primer*. Academic Press.
- Zhang, H., & Tsao, R. (2016). Dietary polyphenols, oxidative stress and antioxidant and anti-inflammatory effects. *Current Opinion in Food Science*, *8*, 33-42.
- Zhang, Y. J., Gan, R. Y., Li, S., Zhou, Y., Li, A. N., Xu, D. P., & Li, H. B. (2015). Antioxidant phytochemicals for the prevention and treatment of chronic diseases. *Molecules*, *20* (12), 21138-21156.
- Zhang, Z., & Finer, J. J. (2016). Use of cytokinin pulse treatments and micrografting to improve sunflower (Helianthus annuus L.) plant recovery from cotyledonary tissues of mature seeds. *In Vitro Cellular & Developmental Biology-Plant*, *52* (4), 391-399.
- Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, *12* (10), 931.

Appendix A.1

Source (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3268504/figure/f01_01/)

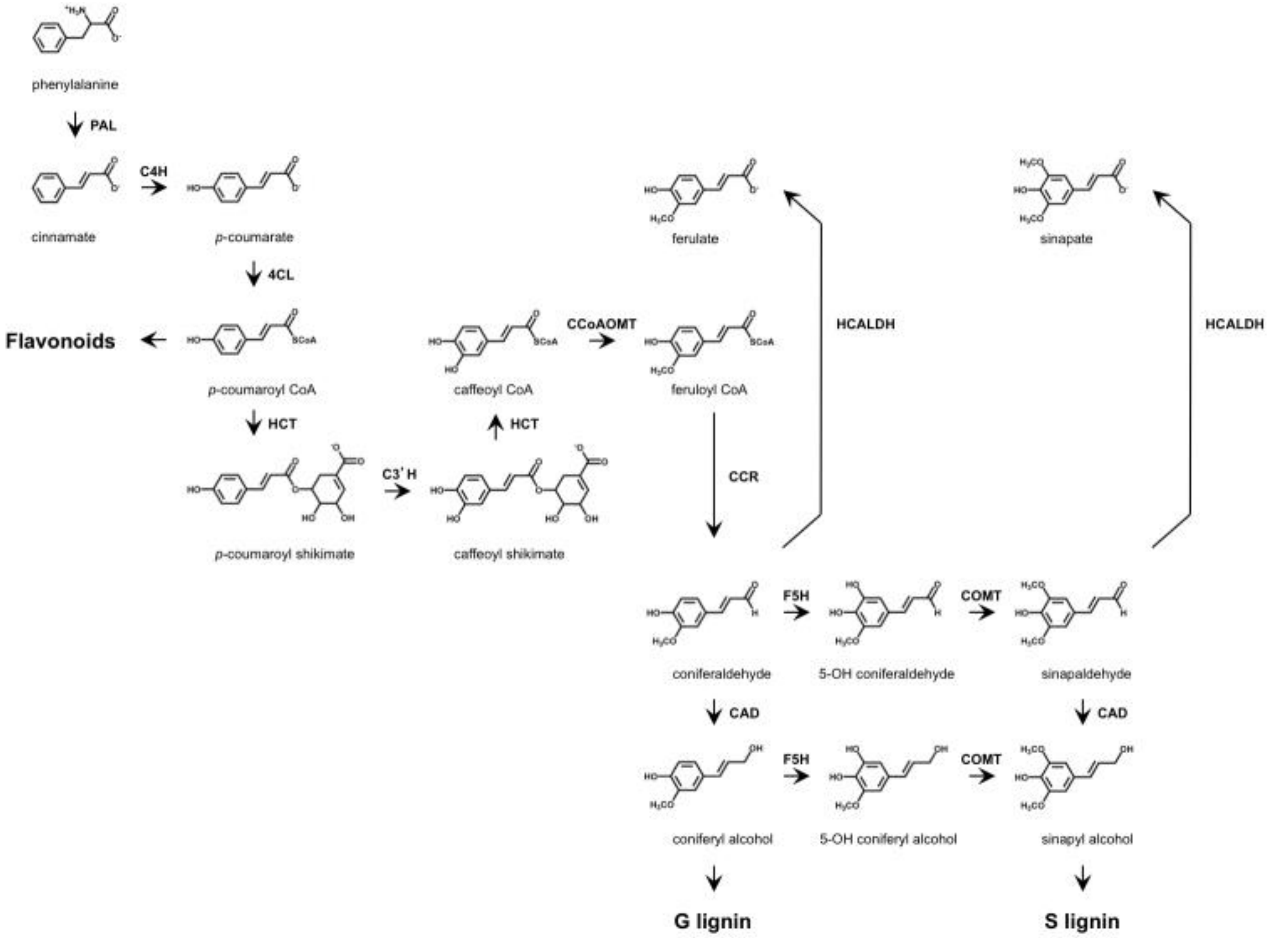
Metabolic intermediates of the shikimate pathway in higher plants. The first seven steps are common to the biosynthesis of phenylalanine, tyrosine, and tryptophan, with chorismate being the last shared precursor of the three amino acids. The phenylalanine and tyrosine pathways diverge after the biosynthesis of arogenate. Some reactants (ATP and PEP), and products (inorganic phosphate and water), have been omitted for the sake of clarity.


phosphoenolpyruvate (PEP)
erythrose 4-phosphate (E4P)
+
3-Deoxy-D-*arabino*-heptulosonate 7-phosphate (DAHP) synthase
3-deoxy-D-*arabino*-heptulosonate 7-phosphate (DAHP)
3-Dehydroquinate synthase
3-dehydroquinate
3-Dehydroquinate dehydratase
3-dehydroshikimate
Shikimate 5-dehydrogenase
shikimate
Shikimate kinase
shikimate 3-phosphate
5-*Enol*pyruvylshikimate 3-phosphate (EPSP) synthase
5-*enol*pyruvylshikimate 3-phosphate (EPSP)
Chorismate synthase
chorismate
tryptophan
Chorismate mutase
prephenate
Prephenate aminotransferase
arogenate
tyrosine
Arogenate dehydratase
phenylalanine

Appendix A.2

Source (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3268504/figure/f02_01/)

Phenylpropanoid metabolism in Arabidopsis. Horizontal reactions correspond to ring modifications, vertical reactions correspond to side-chain modifications. PAL, phenylalanine ammonia-lyase; C4H, cinnamic acid 4-hydroxylase; 4CL, 4-coumarate:CoA ligase; HCT, hydroxycinnamoyl-coenzyme A shikimate:quinate hydroxycinnamoyl-transferase; C3′H, p-coumaroyl shikimate 3′-hydroxylase; CCoAOMT, caffeoyl CoA 3-O-methyltransferase; CCR, cinnamoyl-CoA reductase; F5H, ferulate 5-hydroxylase; COMT, caffeic acid/5-hydroxyferulic acid O-methyltransferase; CAD, cinnamyl alcohol dehydrogenase; HCALDH, hydroxycinnamaldehyde dehydrogenase. The reaction catalyzed by HCALDH leads to the synthesis of ferulate and the sinapate esters.



Appendix B

All Supplementary materials, meaning; the datasets RCHS, NCHS, UCHS and their converted datasets 20RFFM and 8SFM, and all the Python code of the computational pipeline for data preparation, and data analysis were both burned into three discs that should have been brought along with the thesis. However, I was informed by our department's secretary that, UWC faculty of natural sciences does no longer accept CDs or DVDs only a thesis as a single PDF. For this reason I had to insert the data and decrease the font size for all the data and codes to fit in the appendix properly.

For a better visualization, one can simply zoom in.

True Negative Raw Dataset- NCHS

Entry	Entry name	Status	Protein names	Gene names	Organism	Length	Sequence	Catalytic activity
Q9C6L5	KCS5_ARATH	reviewed	3-ketoacyl-CoA synthase 5 (KCS-5) (EC 2.3.1.199) (Eceriferum 60) (Very long-chain fatty acid condensing enzyme 5) (VLCFA condensing enzyme 5)	KCS5 CER60 At1g25450 F2J7.9	Arabidopsis thaliana (Mouse-ear cress)	492	MSDFSSSVKLKYVKLGYQYLINNFLTLLLIPVIATVAIELLRMGPEEILSVLNSLHFELLHILCSSFLIIFVSTVYFMSKPRTVYLVDYSCYKPPVTCRVPFSSFMEHSRLILKDNPKSVEFQMRILERSGLGEETCLPPAIHYIPPTPTMESARNEAQMVIFTAMEDLFKNTGLKPKDIDILIVNCSLFSPTPSLSAMIINKYKLRSNIKSYNLSGMGCSASLISVDVARDLLQVHPNSNAIISTEIITPNYYKGNERAMLLPNCLFRMGGAAILLSNRRSDRWRAKYKLCHLVRTHRGADDKSYNCVMEQEDKNGNVGINLSKDLMTIAGEALKANITIGPLVLPASEQLLFLSSLIGRKIFNPKWKPYIPDFKQAFEHFCIHAGGRAVIDELQKNLQLSGEHVEASRMTLHRFGNTSSSSLWYELSYIEAQGRMKRNDRVWQIAFGSGFKCNSAVWKCNRTIKTPTDGAWSDCIERYPVFIPEVVKL	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q38860	KCS18_ARATH	reviewed	3-ketoacyl-CoA synthase 18 (KCS-18) (EC 2.3.1.199) (Protein FATTY ACID ELONGATION 1) (Very long-chain fatty acid condensing enzyme 18) (VLCFA condensing enzyme 18)	FAE1 KCS18 At4g34520 T4L20.100	Arabidopsis thaliana (Mouse-ear cress)	506	MTSVNVKLLYRYVLTNFFNLCLFPLTAFLAGKASRLTINDLHNFLSYLQHNLITVTLLFAFTVFGLVLYIVTRPNPVYLVDYSCYLPPPHLKVSVSKVMDIFYQIRKADTSSRNVACDDPSSLDFLRKIQERSGLGDETYSPEGLIHVPPRKTFAASREETEKVIIGALENLFENTKVNPREIGILVVNSSMFNPTPSLSAMVVNTFKLRSNIKSFNLGGMGCSAGVIAIDLAKDLLHVHKNTYALVVSTENITQGIYAGENRSMMVSNCLFRVGGAAILLSNKSGDRRRSKYKLVHTVRTHTGADDKSFRCVQQEDDESGKIGVCLSKDITNVAGTTLTKNIATLGPLILPLSEKFLFFATFVAKKLLKDKIKHYYVPDFKLAVDHFCIHAGGRAVIDELEKNLGLSPIDVEASRSTLHRFGNTSSSSIWYELAYIEAKGRMKKGNKAWQIALGSGFKCNSAVWVALRNVKASANSPWQHCIDRYPVKIDSDLSKSKTHVQNGRS	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000269|PubMed:11341960, ECO:0000269|PubMed:12135493, ECO:0000269|PubMed:16765910}.
Q9XF43	KCS6_ARATH	reviewed	3-ketoacyl-CoA synthase 6 (KCS-6) (EC 2.3.1.199) (Cuticular protein 1) (Eceriferum 6) (Very long-chain fatty acid condensing enzyme 6) (VLCFA condensing enzyme 6)	CUT1 CER6 EL6 KCS6 At1g68530 T26J14.10	Arabidopsis thaliana (Mouse-ear cress)	497	MPQAPMPEFSSSVKLKYVKLGYQYLVNHFLSFLLIPIMAIVAVELLRMGPEEILNVWNSLQFDLVQVLCSSFFVIFISTVYFMSKPRTIYLVDYSCYKPPVTCRVPFATFMEHSRLILKDKPKSVEFQMRILERSGLGEETCLPPAIHYIPPTPTMDAARSEAQMVIFEAMDDLFKKTGLKPKDVDILIVNCSLFSPTPSLSAMVINKYKLRSNIKSFNLSGMGCSAGLISVDLARDLLQVHPNSNAIIVSTEIITPNYYQGNERAMLLPNCLFRMGAAAIHMSNRRSDRWRAKYKLSHLVRTHRGADDKSFYCVYEQEDKEGHVGINLSKDLMAIAGEALKANITTIGPLVLPASEQLLFLTSLIGRKIFNPKWKPYIPDFKLAFEHFCIHAGGRAVIDELQKNLQLSGEHVEASRMTLHRFGNTSSSSLWYELSYIESKGRMRRGDRVWQIAFGSGFKCNSAVWKCNRTIKTPKDGPWSDCIDRYPVFIPEVVKL	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000269|PubMed:10330468}.
Q570B4	KCS10_ARATH	reviewed	3-ketoacyl-CoA synthase 10 (KCS-10) (EC 2.3.1.199) (Protein FIDDLEHEAD) (Very long-chain fatty acid condensing enzyme 10) (VLCFA condensing enzyme 10)	FDH EL4 KCS10 At2g26250 T1D16.11	Arabidopsis thaliana (Mouse-ear cress)	550	MGRSNEQDLLSTEIVNRGIEPSGPNAGSPTFSVRVRRRLPDFLQSVNLKYVKLGYHYLINHAVYLATIPVLVLVFSAEVGSLSREEIWKKLWDYDLATVIGFFGVFVLTACVYFMSRPRSVYLIDFACYKPSDEHKVTKEEFIELARKSGKFDEETLGFKKRILQASGIGDETYVPRSISSSENITTMKEGREEASTVIFGALDELFEKTRVKPKDVGVLVVNCSIFNPTPSLSAMVINHYKMRGNILSYNLGGMGCSAGIIAIDLARDMLQSNPNSYAVVVSTEMVGYNWYVGSDKSMVIPNCFFRMGCSAVMLSNRRRDFRHAKYRLEHIVRTHKAADDRSFRSVYQEEDEQGFKGLKISRDLMEVGGEALKTNITTLGPLVLPFSEQLLFFAALLRRTFSPAAKTSTTTSFSTSATAKTNGIKSSSSDLSKPYIPDYKLAFEHFCFHAASKVVLEELQKNLGLSEENMEASRMTLHRFGNTSSSGIWYELAYMEAKESVRRGDRVWQIAFGSGFKCNSVVWKAMRKVKKPTRNNPWVDCINRYPVPL	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9MAM3	KCS1_ARATH	reviewed	3-ketoacyl-CoA synthase 1 (KCS-1) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 1) (VLCFA condensing enzyme 1)	KCS1 EL1 At1g01120 T25K16.11	Arabidopsis thaliana (Mouse-ear cress)	528	MERTNSIEMDRERLTAEMAFRDSSSAVIRIRRRLPDLLTSVKLKYVKLGLHNSCNVTTILFFLIILPLTGTVLVQLTGLTFDTFSELWSNQAVQLDTATRLTCLVFLSFVLTLYVANRSKPVYLVDFSCYKPEDERKISVDSFLTMTEENGSFTDDTVQFQQRISNRAGLGDETYLPRGITSTPPKLNMSEARAEAEAVMFGALDSLFEKTGIKPAEVGILIVNCSLFNPTPSLSAMIVNHYKMREDIKSYNLGGMGCSAGLISIDLANNLLKANPNSYAVVVSTENITLNWYFGNDRSMLLCNCIFRMGGAAILLSNRRQDRKKSKYSLVNVVRTHKGSDDKNYNCVYQKEDERGTIGVSLARELMSVAGDALKTNITTLGPMVLPLSEQLMFLISLVKRKMFKLKVKPYIPDFKLAFEHFCIHAGGRAVLDEVQKNLDLKDWHMEPSRMTLHRFGNTSSSSLWYEMAYTEAKGRVKAGDRLWQIAFGSGFKCNSAVWKALRPVSTEEMTGNAWAGSIDQYPVKVVQ	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000269|PubMed:10074711, ECO:0000269|PubMed:16765910}.
Q5XEP9	KCS2_ARATH	reviewed	3-ketoacyl-CoA synthase 2 (KCS-2) (EC 2.3.1.199) (Docosanoic acid synthase) (Very long-chain fatty acid condensing enzyme 2) (VLCFA condensing enzyme 2)	KCS2 DAISY KCS17 At1g04220 F20D22.1	Arabidopsis thaliana (Mouse-ear cress)	528	MNENHIQSDHMNNTIHVTNKKLPNFLLSVRLKYVKLGYHYLISNAVYILILPVGLLAATSSSFSLTDLTLLYNHLLKFHFLSSTLFAALLIFLTTLYFTTRPRRIFLLDFACYKPDSSLICTRETFMDRSQRVGIFTEDNLAFQQKILERSGLGQKTYFPEALLRVPPNPCMSEARKEAETVMFGAIDAVLEKTGVNPKDIGILVVNCSLFNPTPSLSAMIVNKYKLRGNVLSYNLGGMGCSAGLISIDLAKQLLQVQPNSYALVVSTENITLNWYLGNDRSMLLSNCIFRMGGAAVLLSNRSSDRCRSKYQLIHTVRTHKGSDDNAFNCVYQREDNDDNKQIGVSLSKNLMAIAGEALKTNITTLGPLVLPMSEQLLFFATLVARKVFNVKKIKPYIPDFKLAFEHFCIHAGGRAVLDEIEKNLDLSEWHMEPSRMTLNRFGNTSSSSLWYELAYSEAKGRIKRGDRTWQIAFGSGFKCNSAVWRALRTIDPSKEKKKKTNPWIDEIHEFPVPVPRTSPVTSSSESR	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q8SAS8	TBSYN_HYPAN	reviewed	2,4,6-trihydroxybenzophenone synthase (EC 2.3.1.220) (2,3',4,6-tetrahydroxybenzophenone synthase) (EC 2.3.1.151) (Benzophenone synthase) (HaBPS)	BPS	Hypericum androsaemum (Tutsan)	395	MAPAMEYSTQNGQGEGKKRASVLAIGTTNPEHFILQEDYPDFYFRNTNSEHMTELKEKFKRICVKSHIRKRHFYLTEEILKENQGIATYGAGSLDARQRILETEVPKLGQEAALKAIAEWGQPISKITHVVFATTSGFMMPGADYVITRLLGLNRTVRRVMLYNQGCFAGGTALRVAKDLAENNEGARVLVVCAENTAMTFHAPNESHLDVIVGQAMFSDGAAALIIGACPDVASGERAVFNILSASQTIVPGSDGAITAHFYEMGMSYFLKEDVIPLFRDNIAAVMEEAFSPLGVSDWNSLFYSIHPGGRGIIDGVAGNLGIKDENLVATRHVLGEYGNMGSACVMFILDELRKSSKVNGKPTTGDGKEFGCLIGLGPGLTVEAVVLQSVPILQ	CATALYTIC ACTIVITY: 3 malonyl-CoA + benzoyl-CoA = 4 CoA + 2,4,6-trihydroxybenzophenone + 3 CO(2). {ECO:0000269|PubMed:12795704, ECO:0000269|PubMed:19710020, ECO:0000269|PubMed:9459298}.; CATALYTIC ACTIVITY: 3 malonyl-CoA + 3-hydroxybenzoyl-CoA = 4 CoA + 2,3',4,6-tetrahydroxybenzophenone + 3 CO(2). {ECO:0000269|PubMed:12795704, ECO:0000269|PubMed:19710020, ECO:0000269|PubMed:9459298}.
Q58VP7	PCS_ALOAR	reviewed	5,7-dihydroxy-2-methylchromone synthase (EC 2.3.1.216) (Pentaketide chromone synthase) (PCS)		Aloe arborescens (Kidachi aloe)	403	MSSLSNSLPLMEDVQGIRKAQKADGTATVMAIGTAHPPHIFPQDTYADVYFRATNSEHKVELKKKFDHICKKTMIGKRYFNYDEEFLKKYPNITSYDEPSLNDRQDICVPGVPALGTEAAVKAIEEWGRPKSEITHLVFCTSCGVDMPSADFQCAKLLGLHANVNKYCIYMQGCYAGGTVMRYAKDLAENNRGARVLVVCAELTIMMLRAPNETHLDNAIGISLFGDGAAALIIGSDPIIGVEKPMFEIVCTKQTVIPNTEDVIHLHLRETGMMFYLSKGSPMTISNNVEACLIDVFKSVGITPPEDWNSLFWIPHPGGRAILDQVEAKLKLRPEKFRAARTVLWDYGNMVSASVGYILDEMRRKSAAKGLETYGEGLEWGVLLGFGPGITVETILLHSLPLM	CATALYTIC ACTIVITY: 5 malonyl-CoA = 5 CoA + 5,7-dihydroxy-2-methyl-4H-chromen-4-one + 5 CO(2) + H(2)O. {ECO:0000269|PubMed:15686354}.
Q9FG87	KCS20_ARATH	reviewed	3-ketoacyl-CoA synthase 20 (KCS-20) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 20) (VLCFA condensing enzyme 20)	KCS20 KCS19 At5g43760 MQD19.11	Arabidopsis thaliana (Mouse-ear cress)	529	MSHNQNQPHRPVPVHVTNAEPNPNPNNLPNFLLSVRLKYVKLGYHYLISNALYILLLPLLAATIANLSSFTINDLSLLYNTLRFHFLSATLATALLISLSTAYFTTRPRRVFLLDFSCYKPDPSLICTRETFMDRSQRVGIFTEDNLAFQQKILERSGLGQKTYFPEALLRVPPNPCMEEARKEAETVMFGAIDAVLEKTGVKPKDIGILVVNCSLFNPTPSLSAMIVNKYKLRGNILSYNLGGMGCSAGLISIDLAKQMLQVQPNSYALVVSTENITLNWYLGNDRSMLLSNCIFRMGGAAVLLSNRSSDRSRSKYQLIHTVRTHKGADDNAFGCVYQREDNNAEETGKIGVSLSKNLMAIAGEALKTNITTLGPLVLPMSEQLLFFATLVARKVFKVKKIKPYIPDFKLAFEHFCIHAGGRAVLDEIEKNLDLSEWHMEPSRMTLNRFGNTSSSSLWYELAYSEAKGRIKRGDRTWQIAFGSGFKCNSAVWKALRTIDPMDEKTNPWIDEIDDFPVQVPRITPITSS	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q02323	DPSS_PINSY	reviewed	Pinosylvin synthase (EC 2.3.1.146) (Dihydropinosylvin synthase) (Pinosylvin-forming stilbene synthase) (Stilbene synthase) (STS)		Pinus sylvestris (Scots pine)	393	MGGVDFEGFRKLQRADGFASILAIGTANPPNAVDQSTYPDFYFRITGNEHNTELKDKFKRICERSAIKQRYMYLTEEILKKNPDVCAFVEVPSLDARQAMLAMEVPRLAKEAAEKAIQEWGQSKSGITHLIFCSTTTPDLPGADFEVAKLLGLHPSVKRVGVFQHGCFAGGTVLRMAKDLAENNRGARVLVICSETTAVTFRGPSETHLDSLVGQALFGDGASALIVGADPIPQVEKACFEIVWTAQTVVPNSEGAIGGKVREVGLTFQLKGAVPDLISANIENCMVEAFSQFKISDWNKLFWVVHPGGRAILDRVEAKLNLDPTKLIPTRHVMSEYGNMSSACVHFILDQTRKASLQNGCSTTGEGLEMGVLFGFGPGLTIETVVLKSVPIQ	CATALYTIC ACTIVITY: 3 malonyl-CoA + cinnamoyl-CoA = 4 CoA + pinosylvin + 4 CO(2). {ECO:0000269|PubMed:1426272}.; CATALYTIC ACTIVITY: 3 malonyl-CoA + dihydrocinnamoyl-CoA = 4 CoA + dihydropinosylvin + 4 CO(2). {ECO:0000269|PubMed:1426272}.
Q94FV7	BAS_RHEPA	reviewed	Polyketide synthase BAS (EC 2.3.1.212) (Benzalacetone synthase) (RpBAS)	BAS	Rheum palmatum (Chinese rhubarb)	384	MATEEMKKLATVMAIGTANPPNCYYQADFPDFYFRVTNSDHLINLKQKFKRLCENSRIEKRYLHVTEEILKENPNIAAYEATSLNVRHKMQVKGVAELGKEAALKAIKEWGQPKSKITHLIVCCLAGVDMPGADYQLTKLLDLDPSVKRFMFYHLGCYAGGTVLRLAKDIAENNKGARVLIVCSEMTTTCFRGPSETHLDSMIGQAILGDGAAAVIVGADPDLTVERPIFELVSTAQTIVPESHGAIEGHLLESGLSFHLYKTVPTLISNNIKTCLSDAFTPLNISDWNSLFWIAHPGGPAILDQVTAKVGLEKEKLKVTRQVLKDYGNMSSATVFFIMDEMRKKSLENGQATTGEGLEWGVLFGFGPGITVETVVLRSVPVIS	CATALYTIC ACTIVITY: 4-coumaroyl-CoA + malonyl-CoA + H(2)O = 2 CoA + 4-hydroxybenzalacetone + 2 CO(2). {ECO:0000269|PubMed:11389739, ECO:0000269|PubMed:17383877}.
L7NCQ3	TBSYN_GARMA	reviewed	2,4,6-trihydroxybenzophenone synthase (GmBPS) (EC 2.3.1.220)	BPS	Garcinia mangostana (Mangosteen)	391	MAPAMDSAQNGHQSRGSANVLAIGTANPPNVILQEDYPDFYFKVTNSEHLTDLKEKFKRICVKSKTRKRHFYLTEQILKENPGIATYGAGSLDSRQKILETEIPKLGKEAAMVAIQEWGQPVSKITHVVFATTSGFMMPGADYSITRLLGLNPNVRRVMIYNQGCFAGGTALRVAKDLAENNKGARVLVVCAENTAMTFHGPNENHLDVLVGQAMFSDGAAALIIGANPNLPEERPVYEMVAAHQTIVPESDGAIVAHFYEMGMSYFLKENVIPLFGNNIEACMEAAFKEYGISDWNSLFYSVHPGGRAIVDGIAEKLGLDEENLKATRHVLSEYGNMGSACVIFILDELRKKSKEEKKLTTGDGKEWGCLIGLGPGLTVETVVLRSVPIA	CATALYTIC ACTIVITY: 3 malonyl-CoA + benzoyl-CoA = 4 CoA + 2,4,6-trihydroxybenzophenone + 3 CO(2). {ECO:0000269|PubMed:22390826}.
Q9SIX1	KCS9_ARATH	reviewed	3-ketoacyl-CoA synthase 9 (KCS-9) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 9) (VLCFA condensing enzyme 9)	KCS9 At2g16280 F16F14.22	Arabidopsis thaliana (Mouse-ear cress)	512	MEAANEPVNGGSVQIRTENNERRKLPNFLQSVNMKYVKLGYHYLITHLFKLCLVPLMAVLVTEISRLTTDDLYQIWLHLQYNLVAFIFLSALAIFGSTVYIMSRPRSVYLVDYSCYLPPESLQVKYQKFMDHSKLIEDFNESSLEFQRKILERSGLGEETYLPEALHCIPPRPTMMAAREESEQVMFGALDKLFENTKINPRDIGVLVVNCSLFNPTPSLSAMIVNKYKLRGNVKSFNLGGMGCSAGVISIDLAKDMLQVHRNTYAVVVSTENITQNWYFGNKKAMLIPNCLFRVGGSAILLSNKGKDRRRSKYKLVHTVRTHKGAVEKAFNCVYQEQDDNGKTGVSLSKDLMAIAGEALKANITTLGPLVLPISEQILFFMTLVTKKLFNSKLKPYIPDFKLAFDHFCIHAGGRAVIDELEKNLQLSQTHVEASRMTLHRFGNTSSSSIWYELAYIEAKGRMKKGNRVWQIAFGSGFKCNSAVWVALNNVKPSVSSPWEHCIDRYPVKLDF	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
B0LDU5	PKS4_RUBID	reviewed	Polyketide synthase 4 (RiPKS4) (EC 2.3.1.212) (EC 2.3.1.74) (Benzalacetone synthase PKS4) (RiBAS) (Naringenin-chalcone synthase PKS4)	PKS4 BAS	Rubus idaeus (Raspberry)	383	MVTVEEVRKAQRAEGPATVLAIGTATPPNCVGQSTYPDYYFRITNSEHKIELKQKFQRMCDKSMIKKRYMYLTEEILKENPSMCEYMAPSLDARQDMVIVEIPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLIKLFGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITVVTFRGPSDTHLDCLVGQALFGDGVASIIVGADPLPEIEKPLFELVSAAQTILPDSEGAIEGHLREVGLTFHLLENVPALISKNIEKSLNETFKPLDIMDWNSLFWIAHPGGPAILDQVEAKLGLKPEKLEATGHILSEYGNMSSACVLFILDVVRRKSAANGVTTRILSIGQISKSLLILAWFLFSLV	CATALYTIC ACTIVITY: 4-coumaroyl-CoA + malonyl-CoA + H(2)O = 2 CoA + 4-hydroxybenzalacetone + 2 CO(2). {ECO:0000269|PubMed:12226219, ECO:0000269|PubMed:18068110}.; CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023, ECO:0000269|PubMed:12226219, ECO:0000269|PubMed:18068110}.
C0SVZ6	CURS1_CURLO	reviewed	Curcumin synthase 1 (EC 2.3.1.217)	CURS1	Curcuma longa (Turmeric) (Curcuma domestica)	389	MANLHALRREQRAQGPATIMAIGTATPPNLYEQSTFPDFYFRVTNSDDKQELKKKFRRMCEKTMVKKRYLHLTEEILKERPKLCSYKEASFDDRQDIVVEEIPRLAKEAAEKAIKEWGRPKSEITHLVFCSISGIDMPGADYRLATLLGLPLTVNRLMIYSQACHMGAAMLRIAKDLAENNRGARVLVVACEITVLSFRGPNEGDFEALAGQAGFGDGAGAVVVGADPLEGIEKPIYEIAAAMQETVAESQGAVGGHLRAFGWTFYFLNQLPAIIADNLGRSLERALAPLGVREWNDVFWVAHPGNWAIIDAIEAKLQLSPDKLSTARHVFTEYGNMQSATVYFVMDELRKRSAVEGRSTTGDGLQWGVLLGFGPGLSIETVVLRSMPL	CATALYTIC ACTIVITY: Feruloyl-CoA + feruloylacetyl-CoA + H(2)O = 2 CoA + curcumin + CO(2). {ECO:0000269|PubMed:19258320, ECO:0000269|PubMed:21148316}.
P48408	DPS2_PINST	reviewed	Pinosylvin synthase 2 (EC 2.3.1.146) (Dihydropinosylvin synthase 2) (Stilbene synthase 2) (STS 2)	STS2	Pinus strobus (Eastern white pine)	396	MSVGMGVDLEAFRKSQRADGFASILAIGTANPPNVVDQSTYPDYYFRNTNNEDNTDLKDKFKRICERSAIKKRHMYLTEEILKKNPELCAFLEVPSLDTRQAMLAVEVPRLGKEAAEKAIEEWGQPKSRITHLIFCTTTTPDLPGADFEVAKLLGLHPSVKRVGVFQHGCFAGGTVLRLAKDLAENNRGARVLVVCSENTAVTFRGPSETHLDGLVGLALFGDGAAALIVGADPIPQVEKPCFEIVWTAQTVVPNSDGAISGKLREVGLTFQLKGAVPDLISTNIEKCLVEAFSQFNISDWNQLFWIAHPGGRAILDQVEASLNLDPTKLRATRHVMSEYGNMSSACVHFILDETRKASRQNGCSTSGGGFQMGVLFGFGPGLTVETVVLKSIPFP	CATALYTIC ACTIVITY: 3 malonyl-CoA + cinnamoyl-CoA = 4 CoA + pinosylvin + 4 CO(2). {ECO:0000269|PubMed:7698342}.; CATALYTIC ACTIVITY: 3 malonyl-CoA + dihydrocinnamoyl-CoA = 4 CoA + dihydropinosylvin + 4 CO(2). {ECO:0000269|PubMed:7698342}.
P28343	THS1_VITVI	reviewed	Stilbene synthase 1 (EC 2.3.1.95) (PSV25) (Resveratrol synthase 1) (Trihydroxystilbene synthase 1) (StSy 1) (Vitis stilbene synthase 1)	VINST1 STS2 VST1 GSVIVT00009226001 LOC100256566 VITISV_035301	Vitis vinifera (Grape)	392	MASVEEFRNAQRAKGPATILAIGTATPDHCVYQSDYADYYFRVTKSEHMTELKKKFNRICDKSMIKKRYIHLTEEMLEEHPNIGAYMAPSLNIRQEIITAEVPRLGRDAALKALKEWGQPKSKITHLVFCTTSGVEMPGADYKLANLLGLETSVRRVMLYHQGCYAGGTVLRTAKDLAENNAGARVLVVCSEITVVTFRGPSEDALDSLVGQALFGDGSSAVIVGSDPDVSIERPLFQLVSAAQTFIPNSAGAIAGNLREVGLTFHLWPNVPTLISENIEKCLTQAFDPLGISDWNSLFWIAHPGGPAILDAVEAKLNLEKKKLEATRHVLSEYGNMSSACVLFILDEMRKKSLKGEKATTGEGLDWGVLFGFGPGLTIETVVLHSVPTVTN	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
P48407	DPS1_PINST	reviewed	Pinosylvin synthase 1 (EC 2.3.1.146) (Dihydropinosylvin synthase 1) (Stilbene synthase 1) (STS 1)	STS1	Pinus strobus (Eastern white pine)	396	MSVGMGIDLEAFRKSQRADGFASILAIGTANPPNVVDQSTYPDYYFRVTNNEDNTDLKDKFKRICERSAIKKRHMYLTEEILKKNPELCAFLEVPSLDTRQAMLAAEVPRLGKEAAEKAIEEWGQPKSRITHLIFCTTTTPDLPGADFEVAKLLGLHPSVKRVGVFQHGCFAGGTVLRLAKDLAENNRGARVLVVCSENTAVTFRGPSETHLDGLVGLALFGDGASALIVGADPIPQVEKPCFEIVWTAQTVVPNSDGAISGKLREVGLTFQLKGAVPDLISTNIEKCLVEAFSQFNISDWNQLFWIAHPGGHAILDQVEASLNLDPTKLRATRHVMSEYGNMSSACVHFILDETRKASRQNGCSTSGGGFQMGVLFGFGPGLTVETVVLKSIPFP	CATALYTIC ACTIVITY: 3 malonyl-CoA + cinnamoyl-CoA = 4 CoA + pinosylvin + 4 CO(2). {ECO:0000269|PubMed:7698342}.; CATALYTIC ACTIVITY: 3 malonyl-CoA + dihydrocinnamoyl-CoA = 4 CoA + dihydropinosylvin + 4 CO(2). {ECO:0000269|PubMed:7698342}.
O65677	KCS17_ARATH	reviewed	3-ketoacyl-CoA synthase 17 (KCS-17) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 17) (VLCFA condensing enzyme 17)	KCS17 KCS2 At4g34510 T4L20.90	Arabidopsis thaliana (Mouse-ear cress)	487	MDANGGPVQIRTQNYVKLGYHYLITHFFKLMFLPLMAVLFMNVSLLSLNHLQLYYNSTGFIFVITLAIVGSIVFFMSRPRSIYLLDYSCYLPPSSQKVSYQKFMNNSSLIQDFSETSLEFQRKILIRSGLGEETYLPDSIHSIPPRPTMAAAREEAEQVIFGALDNLFENTKINPREIGVLVVNCSLFNPTPSLSAMIVNKYKLRGNIKSFNLGGMGCSAGVIAVDLASDMLQIHRNTFALVVSTENITQNWYFGNKKAMLIPNCLFRVGGSAVLLSNKPLDRKRSKYKLVHTVRTHKGSDENAFNCVYQEQDECLKTGVSLSKDLMAIAGEALKTNITSLGPLVLPISEQILFFATFVAKRLFNDKKKKPYIPDFKLALDHFCIHAGGRAVIDELEKSLKLSPKHVEASRMTLHRFGNTSSSSIWYELAYTEAKGRMRKGNRVWQIAFGSGFKCNSAVWVALRNVEPSVNNPWEHCIHRYPVKIDL	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000269|PubMed:16765910}.
C6L7V9	CURS3_CURLO	reviewed	Curcumin synthase 3 (EC 2.3.1.217) (Demethoxycurcumin synthase) (EC 2.3.1.219)	CURS3	Curcuma longa (Turmeric) (Curcuma domestica)	390	MGSLQAMRRAQRAQGPATIMAVGTSNPPNLYEQTSYPDFYFRVTNSDHKHALKNKFRVICEKTKVKRRYLHLTEEILKQRPKLCSYMEPSFDDRQDIVVEEIPKLAKEAAEKAIKEWGRPKSEITHLVFCSISGIDMPGADYRLATLLGLPLSVNRLMLYSQACHMGAQMLRIAKDLAENNRGARVLAVSCEITVLSFRGPDAGDFEALACQAGFGDGAAAVVVGADPLPGVERPIYEIAAAMQETVPESERAVGGHLREIGWTFHFFNQLPKLIAENIEGSLARAFKPLGISEWNDVFWVAHPGNWGIMDAIETKLGLEQGKLATARHVFSEYGNMQSATVYFVMDEVRKRSAAEGRATTGEGLEWGVLFGFGPGLTIETVVLRSVPLP	CATALYTIC ACTIVITY: Feruloyl-CoA + feruloylacetyl-CoA + H(2)O = 2 CoA + curcumin + CO(2). {ECO:0000269|PubMed:19622354}.; CATALYTIC ACTIVITY: 4-coumaroyl-CoA + feruloylacetyl-CoA + H(2)O = 2 CoA + demethoxycurcumin + CO(2). {ECO:0000269|PubMed:19622354}.; CATALYTIC ACTIVITY: 4-coumaroyl-CoA + (4-coumaroyl)acetyl-CoA + H(2)O = 2 CoA + bisdemethoxycurcumin + CO(2). {ECO:0000269|PubMed:19622354}.
Q9SYZ0	KCS16_ARATH	reviewed	3-ketoacyl-CoA synthase 16 (KCS-16) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 16) (VLCFA condensing enzyme 16)	KCS16 EL2 At4g34250 F10M10.20	Arabidopsis thaliana (Mouse-ear cress)	493	MDYPMKKVKIFFNYLMAHRFKLCFLPLMVAIAVEASRLSTQDLQNFYLYLQNNHTSLTMFFLYLALGSTLYLMTRPKPVYLVDFSCYLPPSHLKASTQRIMQHVRLVREAGAWKQESDYLMDFCEKILERSGLGQETYVPEGLQTLPLQQNLAVSRIETEEVIIGAVDNLFRNTGISPSDIGILVVNSSTFNPTPSLSSILVNKFKLRDNIKSLNLGGMGCSAGVIAIDAAKSLLQVHRNTYALVVSTENITQNLYMGNNKSMLVTNCLFRIGGAAILLSNRSIDRKRAKYELVHTVRVHTGADDRSYECATQEEDEDGIVGVSLSKNLPMVAARTLKINIATLGPLVLPISEKFHFFVRFVKKKFLNPKLKHYIPDFKLAFEHFCIHAGGRALIDEMEKNLHLTPLDVEASRMTLHRFGNTSSSSIWYELAYTEAKGRMTKGDRIWQIALGSGFKCNSSVWVALRNVKPSTNNPWEQCLHKYPVEIDIDLKE	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.

Q9ZUZ0	KCS13_ARATH	reviewed	3-ketoacyl-CoA synthase 13 (KCS-13) (EC 2.3.1.199) (Protein HIGH CARBON DIOXIDE) (Very long-chain fatty acid condensing enzyme 13) (VLCFA condensing enzyme 13)	HIC KCS13 At2g46720 T3A4.10	Arabidopsis thaliana (Mouse-ear cress)	466	MFIAMADFKILLLILILISLFELDLLHFHHDFFSPFPVKIGLLLISIFFYAYSTTRSKPVYLVDFSCHQPTDSCKISSETFFNMAKGAQLYTDETIQFMTRILNRSGLGDDTYSPRCMLTSPPTPSMYEARHESELVIFGALNSLFKKTGIEPREVGIFIVNCSLFNPNPSLSSMIVNRYKLKTDVKTYNLSGMGCSAGAISVDLATNLLKANPNTYAVIVSTENMTLSMYRGNDRSMLVPNCLFRVGGAAVMLSNRSQDRVRSKYELTHIVRTHKGSSDKHYTCAEQKEDSKGIVGVALSKELTVVAGDSLKTNLTALGPLVLPLSEKLRFILFLVKSKLFRLKVSPYVPDFKLCFKHFCIHAGGRALLDAVEKGLGLSEFDLEPSRMTLHRFGNTSSSSLWYELAYVEAKCRVKRGDRVWQLAFGSGFKCNSIVWRALRTIPANESLVGNPWGDSVHKYPVHVT	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9SUY9	KCS15_ARATH	reviewed	3-ketoacyl-CoA synthase 15 (KCS-15) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 15) (VLCFA condensing enzyme 15)	KCS15 At3g52160 F4F15.270	Arabidopsis thaliana (Mouse-ear cress)	451	MEKEATKMVNGGVKSKSPKGSPDFLGYNLRYVKLGYIYLLSLSRTFCFFLPPLLLLFIFVSRFLPILAFPLSTFFILLIYHYLTPSSVFLLDFSCYRPPDHLKITKSDFIELAMKSGNFNETAIELQRKVLDQSGIGEESYMPRVVFKPGHRVNLRDGREEAAMVIFGAIDELLAATKINVKHIKILVLNCGVLNTTPSLSAMVINHYKLRHNTESYNLGGMGCSAGVIAIDLAKDLLNAHQGSYALVVSTEIVSFTWYSGNDVALLPPNCFFRMGAAAVMLSSRRIDRWRAKYQLMQLVRTHKGMEDTSYKSIELREDRDGKQGLYVSRDVMEVGRHALKANIATLGRLEPSFEHICVLASSKKVLDDIHKDLKLTEENMEASRRTLERFGNTSSSSIWYELAYLEHKAKMKRGDRVWQIGFGSGFKCNSVVWKALKNIDPPRHNNPWNL	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9LZ72	KCS19_ARATH	reviewed	3-ketoacyl-CoA synthase 19 (KCS-19) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 19) (VLCFA condensing enzyme 19)	KCS19 KCS21 At5g04530 T32M21.130	Arabidopsis thaliana (Mouse-ear cress)	464	MELFSLSSLLLLSTLFVFYIFKFVFKRRNQRNCYMLHYECYKGMEERKLDTETCAKVVQRNKNLGLEEYRFLLRTMASSGIGEETYGPRNVLEGREDSPTLLDAHSEMDEIMFDTLDKLFHKTKGSISPSDIDILVVNVSLFAPSPSLTSRVINRYKMREDIKSYNLSGLGCSASVISIDIVQRMFETRENALALVVSTETMGPHWYCGKDRSMMLSNCLFRAGGSSVLLTNAARFKNQALMKLVTVVRAHVGSDDEAYSCCIQMEDRDGHPGFLLTKYLKKAAARALTKNLQVLLPRVLPVKELIRYAIVRALKRRTSAKREPASSGIGLNLKTGLQHFCIHPGGRAIIEGVGKSLGLTEFDIEPARMALHRFGNTSSGGLWYVLGYMEAKNRLKKGEKILMMSMGAGFESNNCVWEVLKDLDDKNVWEDSVDRYPELSRIPNPFVEKYDWINDDTMSFVRVD	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q4V3C9	KCS8_ARATH	reviewed	3-ketoacyl-CoA synthase 8 (KCS-8) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 8) (VLCFA condensing enzyme 8)	KCS8 At2g15090 T15J14.13	Arabidopsis thaliana (Mouse-ear cress)	481	MKNLKMVFFKILFISLMAGLAMKGSKINVEDLQKFSLHHTQNNLQTISLLLFLVVFVWILYMLTRPKPVYLVDFSCYLPPSHLKVSIQTLMGHARRAREAGMCWKNKESDHLVDFQEKILERSGLGQETYIPEGLQCFPLQQGMGASRKETEEVIFGALDNLFRNTGVKPDDIGILVVNSSTFNPTPSLASMIVNKYKLRDNIKSLNLGGMGCSAGVIAVDVAKGLLQVHRNTYAIVVSTENITQNLYLGKNKSMLVTNCLFRVGGAAVLLSNRSRDRNRAKYELVHTVRIHTGSDDRSFECATQEEDEDGIIGVTLTKNLPMVAARTLKINIATLGPLVLPLKEKLAFFITFVKKKYFKPELRNYTPDFKLAFEHFCIHAGGRALIDELEKNLKLSPLHVEASRMTLHRFGNTSSSSIWYELAYTEAKGRMKEGDRIWQIALGSGFKCNSSVWVALRDVKPSANSPWEDCMDRYPVEIDI	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9LN49	KCS4_ARATH	reviewed	3-ketoacyl-CoA synthase 4 (KCS-4) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 4) (VLCFA condensing enzyme 4)	KCS4 At1g19440 F18O14.21	Arabidopsis thaliana (Mouse-ear cress)	516	MDGAGESRLGGDGGGDGSVGVQIRQTRMLPDFLQSVNLKYVKLGYHYLISNLLTLCLFPLAVVISVEASQMNPDDLKQLWIHLQYNLVSIIICSAILVFGLTVYVMTRPRPVYLVDFSCYLPPDHLKAPYARFMEHSRLTGDFDDSALEFQRKILERSGLGEDTYVPEAMHYVPPRISMAAAREEAEQVMFGALDNLFANTNVKPKDIGILVVNCSLFNPTPSLSAMIVNKYKLRGNIRSYNLGGMGCSAGVIAVDLAKDMLLVHRNTYAVVVSTENITQNWYFGNKKSMLIPNCLFRVGGSAVLLSNKSRDKRRSKYRLVHVVRTHRGADDKAFRCVYQEQDDTGRTGVSLSKDLMAIAGETLKTNITTLGPLVLPISEQILFFMTLVVKKLFNGKVKPYIPDFKLAFEHFCIHAGGRAVIDELEKNLQLSPVHVEASRMTLHRFGNTSSSSIWYELAYIEAKGRMRRGNRVWQIAFGSGFKCNSAIWEALRHVKPSNNSPWEDCIDKYPVTLSY	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
O48780	KCS11_ARATH	reviewed	3-ketoacyl-CoA synthase 11 (KCS-11) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 11) (VLCFA condensing enzyme 11)	KCS11 At2g26640 F18A8.1	Arabidopsis thaliana (Mouse-ear cress)	509	MDVEQKKPLIESSDRNLPDFKKSVKLKYVKLGYHYLITHGMYLFLSPLVLVIAAQISTFSVTDLRSLWEHLQYNLISVVVCSMLLVFLMTIYFMTRPRPVYLVNFSCFKPDESRKCTKKIFMDRSKLTGSFTEENLEFQRKILQRSGLGESTYLPEAVLNVPPNPCMKEARKEAETVMFGAIDELLAKTNVNPKDIGILIVNCSLFNPTPSLSAMVVNHYKLRGNILSYNLGGMGCSAGLISIDLAKHLLHSIPNTYAMVISMENITLNWYFGNDRSKLVSNCLFRMGGAAILLSNKRWDRRRSKYELVDTVRTHKGADDKCFGCITQEEDSASKIGVTLSKELMAVAGDALKTNITTLGPLVLPTSEQLLFFATLVGRKLFKMKIKPYIPDFKLAFEHFCIHAGGRAVLDELEKNLKLTEWHMEPSRMTLYRFGNTSSSSLWYELAYSEAKGRIKKGDRIWQIAFGSGFKCNSSVWRAVRSVNPKKEKNPWMDEIHEFPVEVPKVSTI	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000269|PubMed:16765910}.
Q9C992	KCS7_ARATH	reviewed	3-ketoacyl-CoA synthase 7 (KCS-7) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 7) (VLCFA condensing enzyme 7)	KCS7 At1g71160 F23N20.15	Arabidopsis thaliana (Mouse-ear cress)	460	MESSFHFINEALLITQTFITFHQFLVASACVLIAVFGYYFFKPRCIIYLIDFSCYQPPDFLRAPVSNFIEHLTISGVFDQESLDLQQKILERSGISDDASVPATVHEIPPNASISAAREETHEILFAIVQDLFSKHEIDPKSIDILVSNCSLFCPSPSITSMIINKFGMRSDIKSFSLSGMGCSAGILSVNLVKDLMKIHGDSLALVLSMEAVSPNGYRGKCKSMLIANTIFRMGGAAILLSNRKQDSHKAKYKLQHIIRTHVGSDTESYESVMQQVDEEGKVGVALSKQLVRVASKALKINVVQLGPRVLPYSEQLKYIISFIQRKWGMHKEIYTPNFKKAFEHFCIHAGGRAIIEGVEKHLKLDKEDVEASRSTLYRYGNTSSSSLWYELQYLEAKGRMKMGDKVWQIGFGSGFKANSAVWKCISEIDSRGRNAWSDRIHLYPVCGDTSSALKTELLS	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9SS39	KCS14_ARATH	reviewed	Probable 3-ketoacyl-CoA synthase 14 (KCS-14) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 14) (VLCFA condensing enzyme 14)	KCS14 At3g10280 F14P13.12	Arabidopsis thaliana (Mouse-ear cress)	459	MFIAMADFKLLLLILILLSLFELDLLHFHHDFFSPFPVKIGLLLISIFFYAYSTTRSKPVYLVDFSCHQPTDSCKISSETFFNMAKGAQLYTEETIQFMTRILNRSGLGDDTYSPRCMLTSPPTPSMYEARHESELVIFGALNSLFKKTGIEPREVGIFIVNCSLFNPNPSLSSMIVNRYKLKTDVKTYNLSGISVDLATNLLKANPNTYAVIVSTENMTLSMYRGNDRSMLVPNCLFRVGGAAVMLSNRSQDRVRSKYELTHIVRTHKGSSDKHYTCAEQKEDSKGIVGVALSKELTVVAGDTLKTNLTALGPLVLPLSEKLRFILFLVKSKLFRLKVSPYVPDFKLCFKHFCIHAGGRALLDAVEKGLGLSEFDLEPSRMTLHRFGNTSSSSLWYELAYVEAKCRVKRGDRVWQLAFGSGFKCNSIVWRALRTIPANESLVGNPWGDSVHKYPVHVT	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9FH27	KCS21_ARATH	reviewed	Probable 3-ketoacyl-CoA synthase 21 (KCS-21) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 21) (VLCFA condensing enzyme 21)	KCS21 KCS20 At5g49070 K20J1.4	Arabidopsis thaliana (Mouse-ear cress)	464	MNQTIHRVSPISMSISELTTLLSSGVSVFEIFAGLLVVHLIYQRIRTRVKVYLLDFTCYRAPDSNRVPMSTLIETIYLDDKLDQESIDFQARILERSWLSNQTSIPRSLMEIPLKKSLSSVKIETMTTIFTSVEDLLRKNKLSPRSIDILITNCSLHSPSPSLSAMVINKFHMRSNIKSFNLSGMGCAAGILSVNLANDLLQAHRGSLALIVSTEALNTHWYIGKDRSMLLTNCLFRMGAAAVLMSSNDHDRDNAKYELLHVVRKNKAKDDRAYRCIYQDIDSDEKQGVSITKDVISVAGDMLKMNLTSLGPLVLPYLEQFQYVIQHILCKKLKIYESNSSYTPNFKTAFEHFCIHTGGRAVIQAMEMNLKLTKVDIEPSKMTLHRFGNTSSSSIWYALSYLEAKRRMKKGDRVLQIAFGSGFKCNSAVWRCIRKVEPNTENKWLDFIDSYPVDVPDSTNIRPG	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9LQP8	KCS3_ARATH	reviewed	3-ketoacyl-CoA synthase 3 (KCS-3) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 3) (VLCFA condensing enzyme 3)	KCS3 At1g07720 F24B9.18	Arabidopsis thaliana (Mouse-ear cress)	478	MDLLVMLLSLLVSYLIFKIWKRIDSKRDQNCYILDYQCHKPSDDRMVNTQFSGDIILRNKHLRLNEYKFLLKAIVSSGIGEQTYAPRLFFEGREQRPTLQDGLSEMEEFYIDTIEKVLKRNKISPSEIDILVVNVSMLNSTPSLSARIINHYKMREDIKVFNLTAMGCSASVISIDIVKNIFKTYKNKLALVVTSESLSPNWYSGNNRSMILANCLFRSGGCAVLLTNKRSLSRRAMFKLRCLVRTHHGARDDSFNACVQKEDELGHIGVHLDKTLPKAATRAFIDNLKVITPKILPVTELLRFMLCLLLKKLRSSPSKGSTNVTQAAPKAGVKAGINFKTGIDHFCIHTGGKAVIDAIGYSLDLNEYDLEPARMTLHRFGNTSASSLWYVLGYMEAKKRLKRGDRVFMISFGAGFKCNSCVWEVVRDLNVGEAVGNVWNHCINQYPPKSILNPFFEKYGWIHEEEDPDTFKMPEGFM	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
B1Q2B6	OLIS_CANSA	reviewed	3,5,7-trioxododecanoyl-CoA synthase (EC 2.3.1.206) (Olivetol synthase) (Polyketide synthase-1) (Tetraketide synthase)	OLS CAN24 PKS-1 TKS	Cannabis sativa (Hemp) (Marijuana)	385	MNHLRAEGPASVLAIGTANPENILLQDEFPDYYFRVTKSEHMTQLKEKFRKICDKSMIRKRNCFLNEEHLKQNPRLVEHEMQTLDARQDMLVVEVPKLGKDACAKAIKEWGQPKSKITHLIFTSASTTDMPGADYHCAKLLGLSPSVKRVMMYQLGCYGGGTVLRIAKDIAENNKGARVLAVCCDIMACLFRGPSESDLELLVGQAIFGDGAAAVIVGAEPDESVGERPIFELVSTGQTILPNSEGTIGGHIREAGLIFDLHKDVPMLISNNIEKCLIEAFTPIGISDWNSIFWITHPGGKAILDKVEEKLHLKSDKFVDSRHVLSEHGNMSSSTVLFVMDELRKRSLEEGKSTTGDGFEWGVLFGFGPGLTVERVVVRSVPIKY	CATALYTIC ACTIVITY: 3 malonyl-CoA + hexanoyl-CoA = 3 CoA + 3,5,7-trioxododecanoyl-CoA + 3 CO(2). {ECO:0000269|PubMed:19454282, ECO:0000269|PubMed:19581347, ECO:0000269|Ref.3}.
Q9FSC2	ACS3_RUTGR	reviewed	Probable acridone synthase 3 (EC 2.3.1.159) (Acridone synthase III)	ACS3	Ruta graveolens (Common rue)	391	MESLKEMRKAQMSEGPAAILAIGTANPDNVYMQADYPDYYFKMTKSEHMTELKDKFRTLCEKSMIRKRHMCFSEEFLKANPEVCKHMGKSLNARQDIAVVETPRLGNEAAVKAIKEWGQPKSSITHLIFCSSAGVDMPGADYQLTRILGLNPSVKRMMIYQQGCYAGGTVLRLAKDLAENNKGSRVLVVCSELTAPTFRGPSPDAVDSLVGQALFADGAAALVVGADPDSSIERALYYLVSASQMLLPDSDGAIEGHIREEGLTVHLKKDVPALFSGNIDTPLVEAFKPLGISDWNSIFWIAHPGGPAILDQIEEKLGLKEDKLRASKHVMSEYGNMSSSCVLFVLDEMRSRSLQDGKSTTGEGLDWGVLFGFGPGLTVETIVLRSVPIEA	CATALYTIC ACTIVITY: 3 malonyl-CoA + N-methylanthraniloyl-CoA = 4 CoA + 1,3-dihydroxy-N-methylacridone + 3 CO(2).
Q9FSC0	ACS2_RUTGR	reviewed	Acridone synthase 2 (EC 2.3.1.159) (Acridone synthase II)	ACS2	Ruta graveolens (Common rue)	391	MESLKEMRKAQKSEGPAAILAIGTATPDNVYIQADYPDYYFKITKSEHMTELKDKFKTLCEKSMIRKRHMCFSQEFLKANPEVCKHMGKSLNARQDIAVVETPRIGKEAAVKAIKEWGHPKSSITHLIFCTSAGVDMPGADYQLTRMLGLNPSVKRMMIYQQGCYAGGTVLRLAKDLAENNKGSRVLVVCSELTAPTFRGPSPDAVDSLVGQALFADGAAALVVGADPDTSVERALYYIVSASQMLLPDSDGAIEGHIREEGLTVHLKKDVPALFSANIDTPLVEAFRPLGISDWNSIFWIAHPGGPAILDQIEVKLGLKEDKLRASKHVMSEYGNMSSSCVLFVLDEMRNKSLQDGKSTTGEGLDWGVLFGFGPGLTVETVVLRSVPVEA	CATALYTIC ACTIVITY: 3 malonyl-CoA + N-methylanthraniloyl-CoA = 4 CoA + 1,3-dihydroxy-N-methylacridone + 3 CO(2).
O80400	VPS_HUMLU	reviewed	Phloroisovalerophenone synthase (Valerophenone synthase) (EC 2.3.1.156) (3-methyl-1-(trihydroxyphenyl)butan-1-one synthase)	VPS	Humulus lupulus (European hop)	394	MASVTVEQIRKAQRAEGPATILAIGTAVPANCFNQADFPDYYFRVTKSEHMTDLKKKFQRMCEKSTIKKRYLHLTEEHLKQNPHLCEYNAPSLNTRQDMLVVEVPKLGKEAAINAIKEWGQPKSKITHLIFCTGSSIDMPGADYQCAKLLGLRPSVKRVMLYQLGCYAGGKVLRIAKDIAENNKGARVLIVCSEITACIFRGPSEKHLDCLVGQSLFGDGASSVIVGADPDASVGERPIFELVSAAQTILPNSDGAIAGHVTEAGLTFHLLRDVPGLISQNIEKSLIEAFTPIGINDWNNIFWIAHPGGPAILDEIEAKLELKKEKMKASREMLSEYGNMSCASVFFIVDEMRKQSSKEGKSTTGDGLEWGALFGFGPGLTVETVVLHSVPTNV	CATALYTIC ACTIVITY: Isovaleryl-CoA + 3 malonyl-CoA = 4 CoA + 3 CO(2) + 3-methyl-1-(2,4,6-trihydroxyphenyl)butan-1-one.
Q9SIB2	KCS12_ARATH	reviewed	3-ketoacyl-CoA synthase 12 (KCS-12) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 12) (VLCFA condensing enzyme 12)	KCS12 At2g28630 T8O18.8	Arabidopsis thaliana (Mouse-ear cress)	476	MDLLFLFFSLLLSYLFFKIWKLIDSKQDKDCYILDYQCHKPTDDRMVSTQFSGEIIYRNQNLGLTEYKFLLKAIVSSGIGEQTYAPRLVFEGREERPSLQDGISEMEEFYVDSIGKLLERNQISPKDIDILVVNVSMLSSTPSLASRIINHYKMRDDVKVFNLTGMGCSASLISVDIVKNIFKSYANKLALVATSESLSPNWYSGNNRSMILANCLFRSGGCAILLTNKRSLRKKAMFKLKCMVRTHHGAREESYNCCIQAEDEQGRVGFYLGKNLPKAATRAFVENLKVITPKILPVTELIRFMLKLLIKKIKIRQNPSKGSTNLPPGTPLKAGINFKTGIEHFCIHTGGKAVIDGIGHSLDLNEYDIEPARMTLHRFGNTSASSLWYVLAYMEAKKRLKRGDRVFMISFGAGFKCNSCVWEVVRDLTGGESKGNVWNHCIDDYPPKSILNPYLEKFGWIQDEDPDTFKVPDAFM	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
A2ICC6	THS7_VITVI	reviewed	Stilbene synthase 6 (EC 2.3.1.95) (Resveratrol synthase 6) (Trihydroxystilbene synthase 6) (StSy 6)	STS GSVIVT00009216001 LOC100242994	Vitis vinifera (Grape)	392	MASVEEFRNAQRAKGPATILAIGTATPDHCVYQSDYADYYFRVTKSEHMTELKKKFNRICDKSMIKKRYIHLTEEMLEEHPNIGAYMAPSLNIRQEIITAEVPRLGRDAALKALKEWGQPKSKITHLVFCTTSGVEMPGADYKLANLLGLETSVRRVMLYHQGCYAGGTVLRTAKDLAENNAGARVLVVCSEITVVTFRGPSEDALDSLVGQALFGDGSSAVIVGSDPDVSIERPLFQLVSAAQTFIPNSAGAIAGNLREVGLTFHLWPNVPTLISENIEKCLTQAFDPLGISDWNSLFWIAHPGGPAILDAVEAKLNLEKKKLEATRHVLSEYGNMSSACVLFILDEMRKKSLKGENATTGEGLDWGVLFGFGPGLTIETVVLHSIPTVTN	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
P51071	THS3_VITVI	reviewed	Stilbene synthase 3 (EC 2.3.1.95) (PSV368) (Resveratrol synthase 3) (Trihydroxystilbene synthase 3) (StSy 3)	VIT_16s0100g01030	Vitis vinifera (Grape)	392	MASVEEIRNAQRAKGPATILAIGTATPDHCVYQSDYADYYFRVTKSEHMTELKKKFNRICDKSMIKKRYIHLTEEMLEEHPNIGAYMAPSLNIRQEIITVEVPKLGKEAALKALKEWGQPKSKITHLVFCTTSGVEMPGADYKLANLLGLETSVRRVMLYHQGCYAGGTVLRTAKDLAENNAGARVLVVCSEITVVTFRGPSEDALDSLVGQALFGDGSAAVIVGSDPDVSIERPLFQLVSAAQTFIPNSAGAIAGNLREVGLTFHLWPNVPTLISENVEKCLTQAFDPLGISDWNSLFWIAHPGGPAILDAVEAKLNLDKKKLEATRHVLSEYGNMSSACVLFILDEMRKKSHKGEKATTGEGLDWGVLFGFGPGLTIETVVLHSIPMVTN	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
A5AEM3	THS4_VITVI	reviewed	Stilbene synthase 4 (EC 2.3.1.95) (Resveratrol synthase 4) (Trihydroxystilbene synthase 4) (StSy 4)	GSVIVT00005194001 LOC100241891 VITISV_031376	Vitis vinifera (Grape)	392	MASVEEIRNAQRAKGPATVLAIGTATPDNCLYQSDFADYYFRVTKSEHMTELKKKFNRICDKSMIKKRYIHLTEEMLEEHPNIGAYMAPSLNIRQEIITAEVPKLGKEAALKALKEWGQPKSKITHLVFCTTSGVEMPGADYKLANLLGLEPSVRRVMLYHQGCYAGGTVLRTAKDLAENNAGARVLVVCSEITVVTFRGPSEDALDSLVGQALFGDGSAAVIVGSDPDISIERPLFQLVSAAQTFIPNSAGAIAGNLREVGLTFHLWPNVPTLISENIENCLTKAFDPIGISDWNSLFWIAHPGGPAILDAVEAKVGLDKQKLKATRHILSEYGNMSSACVLFILDEMRKKSLKEGKTTTGEGLDWGVLFGFGPGLTIETVVLHSVGTDSN	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
P51070	THS2_VITVI	reviewed	Stilbene synthase 2 (EC 2.3.1.95) (PSV21) (Resveratrol synthase 2) (Trihydroxystilbene synthase 2) (StSy 2)	GSVIVT00004047001 LOC100246143 VITISV_010833; GSVIVT00008253001 LOC100259169 VITISV_024260	Vitis vinifera (Grape)	392	MASVEEIRNAQRAKGPATILAIGTATPDHCVYQSDYADYYFRVTKSEHMTALKKKFNRICDKSMIKKRYIHLTEEMLEEHPNIGAYMAPSLNIRQEIITAEVPKLGKEAALKALKEWGQPKSKITHLVFCTTSGVEMPGADYKLANLLGLEPSVRRVMLYHQGCYAGGTVLRTAKDLAENNAGARVLVVCSEITVVTFRGPSEDALDSLVGQALFGDGSAAVIVGSDPDISIERPLFQLVSAAQTFIPNSAGAIAGNLREVGLTFHLWPNVPTLISENIEKCLTQAFDPLGISDWNSLFWIAHPGGPAILDAVEAKLNLDKKKLEATRHVLSEYGNMSSACVLFILDEMRKKSLKGERATTGEGLDWGVLFGFGPGLTIETVVLHSIPMVTN	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
A5C9M2	THS5_VITVI	reviewed	Stilbene synthase 5 (EC 2.3.1.95) (Resveratrol synthase 5) (Trihydroxystilbene synthase 5) (StSy 5)	GSVIVT00007357001 LOC100250301 VITISV_036852	Vitis vinifera (Grape)	392	MASVEEIRNAQRAKGPATILAIGTATPDHCVYQSDYADYYFKVTKSEHMTELKKKFNRICDKSMIKKRYIHLTEEMLEEHPNIGAYMAPSLNIRQEIITAEVPKLGKEAALKALKEWGQPKSKITHLVFCTTSGVEMPGADYKLANLLGLETSVRRVMLYHQGCYAGGTVLRTAKDLAENNAGARVLVVCSEITVVTFRGPSEDALDSLVGQALFGDGSAAVIVGSDPNVSIERPLFQLVSAAQTFIPNSAGAIAGNLREVGLTFHLWPNVPTLISENIEKCLSQAFDPLGISNWNSLFWIAHPGGPAILDAVEAKLNLEKKKLEATRHVLSEYGNMSSACVLFILDEMRKKSLKGEKATTGEGLDWGVLFGFGPGLTIETVVLHSVPMVTN	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
P20178	THS1_ARAHY	reviewed	Stilbene synthase 1 (EC 2.3.1.95) (Resveratrol synthase 1) (RS1) (Trihydroxystilbene synthase 1)		Arachis hypogaea (Peanut)	389	MVSVSGIRKVQRAEGPATVLAIGTANPPNCVDQSTYADYYFRVTNGEHMTDLKKKFQRICERTQIKNRHMYLTEEILKENPNMCAYKAPSLDAREDMMIREVPRVGKEAATKAIKEWGQPMSKITHLIFCTTSGVALPGVDYELIVLLGLDPSVKRYMMYHQGCFAGGTVLRLAKDLAENNKDARVLIVCSENTAVTFRGPNETDMDSLVGQALFADGAAAIIIGSDPVPEVENPIFEIVSTDQQLVPNSHGAIGGLLREVGLTFYLNKSVPDIISQNINGALSKAFDPLGISDYNSIFWIAHLGGRAILDQVEQKVNLKPEKMKATRDVLSNYGNMSSACVFFIMDLMRKKSLETGLKTTGEGLDWGVLFGFGPGLTIETVVLRSMAI	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
P20077	THS2_ARAHY	reviewed	Putative stilbene synthase 2 (EC 2.3.1.95) (Resveratrol synthase 2) (RS2) (Trihydroxystilbene synthase 2) (Fragment)		Arachis hypogaea (Peanut)	313	LKENPNMCAYKAPSLDAREDMMIREVPRVGKEAATKAIKEWGQPMSKITHLIFCTTSGVALPGVDYELIVLLGLDPSVKRYMMYHQGCFAGGTVLRLAKDLAENNKDARVLIVCSENTAVTFRGPSETDMDSLVGQALFADGAAAIIIGSDPVPEVENPLFEIVSTDQKLVPNSHGAIGGLLREVGLTFYLNKSVPDIISQNINDALSKAFDPLGISDYNSIFWIAHPGGPAILDQVEQKVNLKPEKMNATRDVLSNYGNMSSACVFFIMDLMRKKSLEEGLKTTGEGLDWGVLFGFGPGLTIETVVLRSVAI	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
P51069	THS3_ARAHY	reviewed	Stilbene synthase 3 (EC 2.3.1.95) (Resveratrol synthase 3) (RS3) (Trihydroxystilbene synthase 3)		Arachis hypogaea (Peanut)	389	MVSVSGIRKVQRAEGPATVLAIGTANPPNCIDQSTYADYYFRVTNSEHMTDLKKKFQRICERTQIKNRHMYLTEEILKENPNMCAYKAPSLDAREDMMIREVPRVGKEAATKAIKEWGQPMSKITHLIFCTTSGVALPGVDYELIVLLGLDPCVKRYMMYHQGCFAGGTVLRLAKDLAENNKDARVLIVCSENTAVTFRGPSETDMDSLVGQALFADGAAAIIIGSDPVPEVEKPIFELVSTDQKLVPGSHGAIGGLLREVGLTFYLNKSVPDIISQNINDALNKAFDPLGISDYNSIFWIAHPGGRAILDQVEQKVNLKPEKMKATRDVLSNYGNMSSACVFFIMDLMRKRSLEEGLKTTGEGLDWGVLFGFGPGLTIETVVLRSVAI	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).

C0SVZ5	DCS_CURLO	reviewed	Phenylpropanoylacetyl-CoA synthase (EC 2.3.1.218) (Diketide CoA synthase)	DCS	Curcuma longa (Turmeric) (Curcuma domestica)	389	MEANGYRITHSADGPATILAIGTANPTNVVDQNAYPDFYFRVTNSEYLQELKAKFRRICEKAAIRKRHLYLTEEILRENPSLLAPMAPSFDARQAIVVEAVPKLAKEAAEKAIKEWGRPKSDITHLVFCSASGIDMPGSDLQLLKLLGLPPSVNRVMLYNVGCHAGGTALRVAKDLAENNRGARVLAVCSEVTVLSYRGPHPAHIESLFVQALFGDGAAALVVGSDPVDGVERPIFEIASASQVMLPESAEAVGGHLREIGLTFHLKSQLPSIIASNIEQSLTTACSPLGLSDWNQLFWAVHPGGRAILDQVEARLGLEKDRLAATRHVLSEYGNMQSATVLFILDEMRNRSAAEGHATTGEGLDWGVLLGFGPGLSIETVVLHSCRLN	CATALYTIC ACTIVITY: Feruloyl-CoA + malonyl-CoA = feruloylacetyl-CoA + CO(2) + CoA. {ECO:0000269|PubMed:19258320}.; CATALYTIC ACTIVITY: 4-coumaroyl-CoA + malonyl-CoA = (4-coumaroyl)acetyl-CoA + CO(2) + CoA. {ECO:0000269|PubMed:19258320}.
D2DRC5	BIPS3_SORAU	reviewed	4-hydroxycoumarin synthase 2 (EC 2.3.1.208) (Biphenyl synthase 3) (SaBIS3)	BIS3	Sorbus aucuparia (European mountain ash) (Rowan)	388	MAPVVKNEPQHAKILAIGTANPPNVFHQKDYPDFLFRVTKNEHRTDLREKFDRICEKSRTKKRYLHLTEEMLKANPNIYTYGAPSLNVRQDICNIEVPKLGQEASLKAIKEWGQPISKITHLIFCTASCVDMPGCDFQLIKLLGLDPSVTRTMIYEAGCYAGATVLRMAKDFAENNKGARVLVVCAEITTVFFHGLTDTHLDILVGQALFADGASAVIVGANPEPEIERPLFEIVACRQTILPNSEHGVVANIREMGFNYYLSGDVPKFVGGNVVDFMTKTFEKVDGKNKDWNSLFFSVHPGGPAIVDQVEEKLGLKEGKLRATRHVLSEYGNMGAPTVHFILDEMRNKSIEEGKTTTGEGLEWGVVIGIGPGLTVETAVLRSEFITY	CATALYTIC ACTIVITY: Malonyl-CoA + 2-hydroxybenzoyl-CoA = 2 CoA + 4-hydroxycoumarin + CO(2).
P53416	BBS_PHASS	reviewed	Bibenzyl synthase (EC 2.3.1.-)	BIBSY212 PCS1	Phalaenopsis sp. (Moth orchid)	390	MLSLESIKKAPRADGFASILAIGRANPDNIIEQSAYPDFYFRVTNSEHLVDLKKKFQRICEKTAIRKRHFVWNEEFLTANPCFSTFMDKSLNVRQEVAISEIPKLGAKAATKAIEDWGQPKSRITHLIFCTTSGMDLPGADYQLTQILGLNPNVERVMLYQQGCFAGGTTLRLAKCLAESRKGARVLVVCAETTTVLFRAPSEEHQDDLVTQALFADGASAVIVGADPDEAADERASFVIVSTSQVLLPDSAGAIGGHVSEGGLLATLHRDVPQIVSKNVGKCLEEAFTPFGISDWNSIFWVPHPGGRAILDQVEERVGLKPEKLSVSRHVLAEYGNMSSVCVHFALDEMRKRSANEGKATTGEGLEWGVLFGFGPGLTVETVVLRSVPL	CATALYTIC ACTIVITY: 3 malonyl-CoA + M-hydroxyphenylpropionyl-CoA = 4 CoA + 4 trihydroxybibenzyl + 3 CO(2).
Q8LIL0	CUS_ORYSJ	reviewed	Bisdemethoxycurcumin synthase (EC 2.3.1.211) (Curcuminoid synthase)	Os07g0271500 LOC_Os07g17010 OJ1001_C01.122 OSJNBb0002J01.6	Oryza sativa subsp. japonica (Rice)	402	MAPTTTMGSALYPLGEMRRSQRADGLAAVLAIGTANPPNCVTQEEFPDFYFRVTNSDHLTALKDKFKRICQEMGVQRRYLHHTEEMLSAHPEFVDRDAPSLDARLDIAADAVPELAAEAAKKAIAEWGRPAADITHLVVTTNSGAHVPGVDFRLVPLLGLRPSVRRTMLHLNGCFAGCAALRLAKDLAENSRGARVLVVAAELTLMYFTGPDEGCFRTLLVQGLFGDGAAAVIVGADADDVERPLFEIVSAAQTIIPESDHALNMRFTERRLDGVLGRQVPGLIGDNVERCLLDMFGPLLGGDGGGGWNDLFWAVHPGSSTIMDQVDAALGLEPGKLAASRRVLSDYGNMSGATVIFALDELRRQRKEAAAAGEWPELGVMMAFGPGMTVDAMLLHATSHVN	CATALYTIC ACTIVITY: 2 4-coumaroyl-CoA + malonyl-CoA + H(2)O = 3 CoA + bisdemethoxycurcumin + 2 CO(2).
C6L7V8	CURS2_CURLO	reviewed	Curcumin synthase 2 (EC 2.3.1.217)	CURS2	Curcuma longa (Turmeric) (Curcuma domestica)	391	MAMISLQAMRKAQRAQGPATILAVGTANPPNLYEQDTYPDYYFRVTNSEHKQELKNKFRLMCEKTMVKRRYLYLTPEILKERPKLCSYMEPSFDDRQDIVVEEVPKLAAEAAENAIKEWGGDKSAITHLVFCSISGIDMPGADYRLAQLLGLPLAVNRLMLYSQACHMGAAMLRIAKDIAENNRSARVLVVACEITVLSFRGPDERDFQALAGQAGFGDGAGAMIVGADPVLGVERPLYHIMSATQTTVPESEKAVGGHLREVGLTFHFFNQLPAIIADNVGNSLAEAFEPIGIKDWNNIFWVAHPGNWAIMDAIETKLGLEQSKLATARHVFSEFGNMQSATVYFVMDELRKRSAAENRATTGDGLRWGVLFGFGPGISIETVVLQSVPL	CATALYTIC ACTIVITY: Feruloyl-CoA + feruloylacetyl-CoA + H(2)O = 2 CoA + curcumin + CO(2). {ECO:0000269|PubMed:19622354}.
Q9FSC1	ACS4_RUTGR	reviewed	Probable acridone synthase 4 (EC 2.3.1.159) (Acridone synthase IV)	ACS4	Ruta graveolens (Common rue)	391	MESLKEMRKAQMSEGPAAILAIGTATPNNVYMQADYPDYYFRMTKSEHMTELKDKFRTLCEKSMIRKRHMCFSEEFLKANPEVSKHMGKSLNARQDIAVVETPRLGNEAAVKAIKEWGQPKSSITHLIFCSSAGVDMPGADYQLTRILGLNPSVKRMMVYQQGCYAGGTVLRLAKDLAENNKGSRVLVVCSELTAPTFRGPSPDAVDSLVGQALFADGAAALVVGADPDSSIERALYYLVSASQMLLPDSDGAIEGHIREEGLTVHLKKDVPALFSANIDTPLVEAFKPLGISDWNSIFWIAHPGGPAILDQIEEKLGLKEDKLRASKHVMSEYGNMSSSCVLFVLDEMRSRSLQDGKSTTGEGLDWGVLFGFGPGLTVETVVLRSVPIEA	CATALYTIC ACTIVITY: 3 malonyl-CoA + N-methylanthraniloyl-CoA = 4 CoA + 1,3-dihydroxy-N-methylacridone + 3 CO(2).
Q9SLX9	VPS_PSINU	reviewed	Phloroisovalerophenone synthase (Valerophenone synthase) (EC 2.3.1.156) (3-methyl-1-(trihydroxyphenyl)butan-1-one synthase)	VPS	Psilotum nudum (Whisk fern) (Lycopodium nudum)	406	MIQNSDSATATLLPRKKERASGPVSVLAIGSANPPNVFHQSLFPDFYFNITQSNHMAEVKAKFTRMCAKSGIKKRRMHINEDILEAHPSIRSYHDNSLDVRQDMLVEEVPKLGKVAADNAIAEWGQPKSNITHLIFCTSSGIDMPGADWALMKLLGLRPTVNRVMVYQQGCFAGCTVLRIAKDLAENNKGSRILVVCSELTLISFRGPTEDHPENLVGQALFGDGAAALIVGADPIPHAENASFEIHWARSSVVPDSDDAVTGNIKENGLVLHLSKTIPDLIGQNIHTLLKDALEEMFDACNPSSFNDLFWVIHPGGPAILDAVEEELNLKSERTHASREILSQYGNMVSPGVLFVLDYMRKRSVDERLSTTGEGLEWGVMLGFGPGLTVETLILKSVPTQAFKYF	CATALYTIC ACTIVITY: Isovaleryl-CoA + 3 malonyl-CoA = 4 CoA + 3 CO(2) + 3-methyl-1-(2,4,6-trihydroxyphenyl)butan-1-one.
C6L7V8	CURS2_CURLO	reviewed	Curcumin synthase 2 (EC 2.3.1.217)	CURS2	Curcuma longa (Turmeric) (Curcuma domestica)	391	MAMISLQAMRKAQRAQGPATILAVGTANPPNLYEQDTYPDYYFRVTNSEHKQELKNKFRLMCEKTMVKRRYLYLTPEILKERPKLCSYMEPSFDDRQDIVVEEVPKLAAEAAENAIKEWGGDKSAITHLVFCSISGIDMPGADYRLAQLLGLPLAVNRLMLYSQACHMGAAMLRIAKDIAENNRSARVLVVACEITVLSFRGPDERDFQALAGQAGFGDGAGAMIVGADPVLGVERPLYHIMSATQTTVPESEKAVGGHLREVGLTFHFFNQLPAIIADNVGNSLAEAFEPIGIKDWNNIFWVAHPGNWAIMDAIETKLGLEQSKLATARHVFSEFGNMQSATVYFVMDELRKRSAAENRATTGDGLRWGVLFGFGPGISIETVVLQSVPL	CATALYTIC ACTIVITY: Feruloyl-CoA + feruloylacetyl-CoA + H(2)O = 2 CoA + curcumin + CO(2). {ECO:0000269|PubMed:19622354}.
Q9FSC1	ACS4_RUTGR	reviewed	Probable acridone synthase 4 (EC 2.3.1.159) (Acridone synthase IV)	ACS4	Ruta graveolens (Common rue)	391	MESLKEMRKAQMSEGPAAILAIGTATPNNVYMQADYPDYYFRMTKSEHMTELKDKFRTLCEKSMIRKRHMCFSEEFLKANPEVSKHMGKSLNARQDIAVVETPRLGNEAAVKAIKEWGQPKSSITHLIFCSSAGVDMPGADYQLTRILGLNPSVKRMMVYQQGCYAGGTVLRLAKDLAENNKGSRVLVVCSELTAPTFRGPSPDAVDSLVGQALFADGAAALVVGADPDSSIERALYYLVSASQMLLPDSDGAIEGHIREEGLTVHLKKDVPALFSANIDTPLVEAFKPLGISDWNSIFWIAHPGGPAILDQIEEKLGLKEDKLRASKHVMSEYGNMSSSCVLFVLDEMRSRSLQDGKSTTGEGLDWGVLFGFGPGLTVETVVLRSVPIEA	CATALYTIC ACTIVITY: 3 malonyl-CoA + N-methylanthraniloyl-CoA = 4 CoA + 1,3-dihydroxy-N-methylacridone + 3 CO(2).
Q9SLX9	VPS_PSINU	reviewed	Phloroisovalerophenone synthase (Valerophenone synthase) (EC 2.3.1.156) (3-methyl-1-(trihydroxyphenyl)butan-1-one synthase)	VPS	Psilotum nudum (Whisk fern) (Lycopodium nudum)	406	MIQNSDSATATLLPRKKERASGPVSVLAIGSANPPNVFHQSLFPDFYFNITQSNHMAEVKAKFTRMCAKSGIKKRRMHINEDILEAHPSIRSYHDNSLDVRQDMLVEEVPKLGKVAADNAIAEWGQPKSNITHLIFCTSSGIDMPGADWALMKLLGLRPTVNRVMVYQQGCFAGCTVLRIAKDLAENNKGSRILVVCSELTLISFRGPTEDHPENLVGQALFGDGAAALIVGADPIPHAENASFEIHWARSSVVPDSDDAVTGNIKENGLVLHLSKTIPDLIGQNIHTLLKDALEEMFDACNPSSFNDLFWVIHPGGPAILDAVEEELNLKSERTHASREILSQYGNMVSPGVLFVLDYMRKRSVDERLSTTGEGLEWGVMLGFGPGLTVETLILKSVPTQAFKYF	CATALYTIC ACTIVITY: Isovaleryl-CoA + 3 malonyl-CoA = 4 CoA + 3 CO(2) + 3-methyl-1-(2,4,6-trihydroxyphenyl)butan-1-one.
A2ICC6	THS7_VITVI	reviewed	Stilbene synthase 6 (EC 2.3.1.95) (Resveratrol synthase 6) (Trihydroxystilbene synthase 6) (StSy 6)	STS GSVIVT00009216001 LOC100242994	Vitis vinifera (Grape)	392	MASVEEFRNAQRAKGPATILAIGTATPDHCVYQSDYADYYFRVTKSEHMTELKKKFNRICDKSMIKKRYIHLTEEMLEEHPNIGAYMAPSLNIRQEIITAEVPRLGRDAALKALKEWGQPKSKITHLVFCTTSGVEMPGADYKLANLLGLETSVRRVMLYHQGCYAGGTVLRTAKDLAENNAGARVLVVCSEITVVTFRGPSEDALDSLVGQALFGDGSSAVIVGSDPDVSIERPLFQLVSAAQTFIPNSAGAIAGNLREVGLTFHLWPNVPTLISENIEKCLTQAFDPLGISDWNSLFWIAHPGGPAILDAVEAKLNLEKKKLEATRHVLSEYGNMSSACVLFILDEMRKKSLKGENATTGEGLDWGVLFGFGPGLTIETVVLHSIPTVTN	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
P28343	THS1_VITVI	reviewed	Stilbene synthase 1 (EC 2.3.1.95) (PSV25) (Resveratrol synthase 1) (Trihydroxystilbene synthase 1) (StSy 1) (Vitis stilbene synthase 1)	VINST1 STS2 VST1 GSVIVT00009226001 LOC100256566 VITISV_035301	Vitis vinifera (Grape)	392	MASVEEFRNAQRAKGPATILAIGTATPDHCVYQSDYADYYFRVTKSEHMTELKKKFNRICDKSMIKKRYIHLTEEMLEEHPNIGAYMAPSLNIRQEIITAEVPRLGRDAALKALKEWGQPKSKITHLVFCTTSGVEMPGADYKLANLLGLETSVRRVMLYHQGCYAGGTVLRTAKDLAENNAGARVLVVCSEITVVTFRGPSEDALDSLVGQALFGDGSSAVIVGSDPDVSIERPLFQLVSAAQTFIPNSAGAIAGNLREVGLTFHLWPNVPTLISENIEKCLTQAFDPLGISDWNSLFWIAHPGGPAILDAVEAKLNLEKKKLEATRHVLSEYGNMSSACVLFILDEMRKKSLKGEKATTGEGLDWGVLFGFGPGLTIETVVLHSVPTVTN	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
A5AEM3	THS4_VITVI	reviewed	Stilbene synthase 4 (EC 2.3.1.95) (Resveratrol synthase 4) (Trihydroxystilbene synthase 4) (StSy 4)	GSVIVT00005194001 LOC100241891 VITISV_031376	Vitis vinifera (Grape)	392	MASVEEIRNAQRAKGPATVLAIGTATPDNCLYQSDFADYYFRVTKSEHMTELKKKFNRICDKSMIKKRYIHLTEEMLEEHPNIGAYMAPSLNIRQEIITAEVPKLGKEAALKALKEWGQPKSKITHLVFCTTSGVEMPGADYKLANLLGLEPSVRRVMLYHQGCYAGGTVLRTAKDLAENNAGARVLVVCSEITVVTFRGPSEDALDSLVGQALFGDGSAAVIVGSDPDISIERPLFQLVSAAQTFIPNSAGAIAGNLREVGLTFHLWPNVPTLISENIENCLTKAFDPIGISDWNSLFWIAHPGGPAILDAVEAKVGLDKQKLKATRHILSEYGNMSSACVLFILDEMRKKSLKEGKTTTGEGLDWGVLFGFGPGLTIETVVLHSVGTDSN	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
A5C9M2	THS5_VITVI	reviewed	Stilbene synthase 5 (EC 2.3.1.95) (Resveratrol synthase 5) (Trihydroxystilbene synthase 5) (StSy 5)	GSVIVT00007357001 LOC100250301 VITISV_036852	Vitis vinifera (Grape)	392	MASVEEIRNAQRAKGPATILAIGTATPDHCVYQSDYADYYFKVTKSEHMTELKKKFNRICDKSMIKKRYIHLTEEMLEEHPNIGAYMAPSLNIRQEIITAEVPKLGKEAALKALKEWGQPKSKITHLVFCTTSGVEMPGADYKLANLLGLETSVRRVMLYHQGCYAGGTVLRTAKDLAENNAGARVLVVCSEITVVTFRGPSEDALDSLVGQALFGDGSAAVIVGSDPNVSIERPLFQLVSAAQTFIPNSAGAIAGNLREVGLTFHLWPNVPTLISENIEKCLSQAFDPLGISNWNSLFWIAHPGGPAILDAVEAKLNLEKKKLEATRHVLSEYGNMSSACVLFILDEMRKKSLKGEKATTGEGLDWGVLFGFGPGLTIETVVLHSVPMVTN	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
B0LDU5	PKS4_RUBID	reviewed	Polyketide synthase 4 (RiPKS4) (EC 2.3.1.212) (EC 2.3.1.74) (Benzalacetone synthase PKS4) (RiBAS) (Naringenin-chalcone synthase PKS4)	PKS4 BAS	Rubus idaeus (Raspberry)	383	MVTVEEVRKAQRAEGPATVLAIGTATPPNCVGQSTYPDYYFRITNSEHKIELKQKFQRMCDKSMIKKRYMYLTEEILKENPSMCEYMAPSLDARQDMVIVEIPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLIKLFGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITVVTFRGPSDTHLDCLVGQALFGDGVASIIVGADPLPEIEKPLFELVSAAQTILPDSEGAIEGHLREVGLTFHLLENVPALISKNIEKSLNETFKPLDIMDWNSLFWIAHPGGPAILDQVEAKLGLKPEKLEATGHILSEYGNMSSACVLFILDVVRRKSAANGVTTRILSIGQISKSLLILAWFLFSLV	CATALYTIC ACTIVITY: 4-coumaroyl-CoA + malonyl-CoA + H(2)O = 2 CoA + 4-hydroxybenzalacetone + 2 CO(2). {ECO:0000269|PubMed:12226219, ECO:0000269|PubMed:18068110}.; CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023, ECO:0000269|PubMed:12226219, ECO:0000269|PubMed:18068110}.
B0LDU6	PKS5_RUBID	reviewed	Polyketide synthase 5 (RiPKS5) (EC 2.3.1.74) (Naringenin-chalcone synthase PKS5)	PKS5	Rubus idaeus (Raspberry)	391	MVTVDEVRKAQRAEGPATVLAIGTATPPNCIDQSTYPDYYFRITNSEHKTELKEKFQRMCDKSMIKKRYMYLTEEILKENPSMCEYMAPSLDARQDMVVVEIPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEIXAVTFRGPSDTHLDSLVGQALFGDGAAAIIVGADPLPKIERPLFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLNEAFKPLDITDWNSLFWIAHPGGPAILDQVETKLGLKPEKLEATRHILSEYGNMSSACVLFILDEVRRKSATNGLKTTGEGLEWGVLFGFGPGLTVETVVLHSVGVTA	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023, ECO:0000269|PubMed:18068110}.
C0SVZ5	DCS_CURLO	reviewed	Phenylpropanoylacetyl-CoA synthase (EC 2.3.1.218) (Diketide CoA synthase)	DCS	Curcuma longa (Turmeric) (Curcuma domestica)	389	MEANGYRITHSADGPATILAIGTANPTNVVDQNAYPDFYFRVTNSEYLQELKAKFRRICEKAAIRKRHLYLTEEILRENPSLLAPMAPSFDARQAIVVEAVPKLAKEAAEKAIKEWGRPKSDITHLVFCSASGIDMPGSDLQLLKLLGLPPSVNRVMLYNVGCHAGGTALRVAKDLAENNRGARVLAVCSEVTVLSYRGPHPAHIESLFVQALFGDGAAALVVGSDPVDGVERPIFEIASASQVMLPESAEAVGGHLREIGLTFHLKSQLPSIIASNIEQSLTTACSPLGLSDWNQLFWAVHPGGRAILDQVEARLGLEKDRLAATRHVLSEYGNMQSATVLFILDEMRNRSAAEGHATTGEGLDWGVLLGFGPGLSIETVVLHSCRLN	CATALYTIC ACTIVITY: Feruloyl-CoA + malonyl-CoA = feruloylacetyl-CoA + CO(2) + CoA. {ECO:0000269|PubMed:19258320}.; CATALYTIC ACTIVITY: 4-coumaroyl-CoA + malonyl-CoA = (4-coumaroyl)acetyl-CoA + CO(2) + CoA. {ECO:0000269|PubMed:19258320}.
C0SVZ6	CURS1_CURLO	reviewed	Curcumin synthase 1 (EC 2.3.1.217)	CURS1	Curcuma longa (Turmeric) (Curcuma domestica)	389	MANLHALRREQRAQGPATIMAIGTATPPNLYEQSTFPDFYFRVTNSDDKQELKKKFRRMCEKTMVKKRYLHLTEEILKERPKLCSYKEASFDDRQDIVVEEIPRLAKEAAEKAIKEWGRPKSEITHLVFCSISGIDMPGADYRLATLLGLPLTVNRLMIYSQACHMGAAMLRIAKDLAENNRGARVLVVACEITVLSFRGPNEGDFEALAGQAGFGDGAGAVVVGADPLEGIEKPIYEIAAAMQETVAESQGAVGGHLRAFGWTFYFLNQLPAIIADNLGRSLERALAPLGVREWNDVFWVAHPGNWAIIDAIEAKLQLSPDKLSTARHVFTEYGNMQSATVYFVMDELRKRSAVEGRSTTGDGLQWGVLLGFGPGLSIETVVLRSMPL	CATALYTIC ACTIVITY: Feruloyl-CoA + feruloylacetyl-CoA + H(2)O = 2 CoA + curcumin + CO(2). {ECO:0000269|PubMed:19258320, ECO:0000269|PubMed:21148316}.
C6L7V8	CURS2_CURLO	reviewed	Curcumin synthase 2 (EC 2.3.1.217)	CURS2	Curcuma longa (Turmeric) (Curcuma domestica)	391	MAMISLQAMRKAQRAQGPATILAVGTANPPNLYEQDTYPDYYFRVTNSEHKQELKNKFRLMCEKTMVKRRYLYLTPEILKERPKLCSYMEPSFDDRQDIVVEEVPKLAAEAAENAIKEWGGDKSAITHLVFCSISGIDMPGADYRLAQLLGLPLAVNRLMLYSQACHMGAAMLRIAKDIAENNRSARVLVVACEITVLSFRGPDERDFQALAGQAGFGDGAGAMIVGADPVLGVERPLYHIMSATQTTVPESEKAVGGHLREVGLTFHFFNQLPAIIADNVGNSLAEAFEPIGIKDWNNIFWVAHPGNWAIMDAIETKLGLEQSKLATARHVFSEFGNMQSATVYFVMDELRKRSAAENRATTGDGLRWGVLFGFGPGISIETVVLQSVPL	CATALYTIC ACTIVITY: Feruloyl-CoA + feruloylacetyl-CoA + H(2)O = 2 CoA + curcumin + CO(2). {ECO:0000269|PubMed:19622354}.
C6L7V9	CURS3_CURLO	reviewed	Curcumin synthase 3 (EC 2.3.1.217) (Demethoxycurcumin synthase) (EC 2.3.1.219)	CURS3	Curcuma longa (Turmeric) (Curcuma domestica)	390	MGSLQAMRRAQRAQGPATIMAVGTSNPPNLYEQTSYPDFYFRVTNSDHKHALKNKFRVICEKTKVKRRYLHLTEEILKQRPKLCSYMEPSFDDRQDIVVEEIPKLAKEAAEKAIKEWGRPKSEITHLVFCSISGIDMPGADYRLATLLGLPLSVNRLMLYSQACHMGAQMLRIAKDLAENNRGARVLAVSCEITVLSFRGPDAGDFEALACQAGFGDGAAAVVVGADPLPGVERPIYEIAAAMQETVPESERAVGGHLREIGWTFHFFNQLPKLIAENIEGSLARAFKPLGISEWNDVFWVAHPGNWGIMDAIETKLGLEQGKLATARHVFSEYGNMQSATVYFVMDEVRKRSAAEGRATTGEGLEWGVLFGFGPGLTIETVVLRSVPLP	CATALYTIC ACTIVITY: Feruloyl-CoA + feruloylacetyl-CoA + H(2)O = 2 CoA + curcumin + CO(2). {ECO:0000269|PubMed:19622354}.; CATALYTIC ACTIVITY: 4-coumaroyl-CoA + feruloylacetyl-CoA + H(2)O = 2 CoA + demethoxycurcumin + CO(2). {ECO:0000269|PubMed:19622354}.; CATALYTIC ACTIVITY: 4-coumaroyl-CoA + (4-coumaroyl)acetyl-CoA + H(2)O = 2 CoA + bisdemethoxycurcumin + CO(2). {ECO:0000269|PubMed:19622354}.
D2DRC4	BIPS2_SORAU	reviewed	4-hydroxycoumarin synthase 1 (EC 2.3.1.208) (Biphenyl synthase 2) (SaBIS2)	BIS2	Sorbus aucuparia (European mountain ash) (Rowan)	390	MAPSVKDQVEPQHAKILAIGTANPPNVYYQEDYPDFLFRVTKNEHRTDLREKFDRICEKSRTRKRYLYLTEEILKANPCIYTYGAPSLDVRQDMLNPEVPKLGQEAALKAIKEWGQPISKITHLIFCTASCVDMPGADFQLVKLLGLNPSVTRTMIYEAGCYAGATVLRLAKDFAENNEGARVLVVCAEITTVFFHGLTDTHLDILVGQALFADGASAVIVGANPEPEIESPLFEIVACRQTIIPNSEHGVVANIREMGFNYYLSGEVPKFVGGNVVDFLTKTFEKVDGKNKDWNSLFFSVHPGGPAIVDQVEEQLGLKEGKLRATRHVLSEYGNMGAPSVHFILDEMRKKSIEEGKATTGEGLEWGVVIGIGPGLTVETAVLRSEFITC	CATALYTIC ACTIVITY: Malonyl-CoA + 2-hydroxybenzoyl-CoA = 2 CoA + 4-hydroxycoumarin + CO(2).
D2DRC5	BIPS3_SORAU	reviewed	4-hydroxycoumarin synthase 2 (EC 2.3.1.208) (Biphenyl synthase 3) (SaBIS3)	BIS3	Sorbus aucuparia (European mountain ash) (Rowan)	388	MAPVVKNEPQHAKILAIGTANPPNVFHQKDYPDFLFRVTKNEHRTDLREKFDRICEKSRTKKRYLHLTEEMLKANPNIYTYGAPSLNVRQDICNIEVPKLGQEASLKAIKEWGQPISKITHLIFCTASCVDMPGCDFQLIKLLGLDPSVTRTMIYEAGCYAGATVLRMAKDFAENNKGARVLVVCAEITTVFFHGLTDTHLDILVGQALFADGASAVIVGANPEPEIERPLFEIVACRQTILPNSEHGVVANIREMGFNYYLSGDVPKFVGGNVVDFMTKTFEKVDGKNKDWNSLFFSVHPGGPAIVDQVEEKLGLKEGKLRATRHVLSEYGNMGAPTVHFILDEMRNKSIEEGKTTTGEGLEWGVVIGIGPGLTVETAVLRSEFITY	CATALYTIC ACTIVITY: Malonyl-CoA + 2-hydroxybenzoyl-CoA = 2 CoA + 4-hydroxycoumarin + CO(2).
B1Q2B6	OLIS_CANSA	reviewed	3,5,7-trioxododecanoyl-CoA synthase (EC 2.3.1.206) (Olivetol synthase) (Polyketide synthase-1) (Tetraketide synthase)	OLS CAN24 PKS-1 TKS	Cannabis sativa (Hemp) (Marijuana)	385	MNHLRAEGPASVLAIGTANPENILLQDEFPDYYFRVTKSEHMTQLKEKFRKICDKSMIRKRNCFLNEEHLKQNPRLVEHEMQTLDARQDMLVVEVPKLGKDACAKAIKEWGQPKSKITHLIFTSASTTDMPGADYHCAKLLGLSPSVKRVMMYQLGCYGGGTVLRIAKDIAENNKGARVLAVCCDIMACLFRGPSESDLELLVGQAIFGDGAAAVIVGAEPDESVGERPIFELVSTGQTILPNSEGTIGGHIREAGLIFDLHKDVPMLISNNIEKCLIEAFTPIGISDWNSIFWITHPGGKAILDKVEEKLHLKSDKFVDSRHVLSEHGNMSSSTVLFVMDELRKRSLEEGKSTTGDGFEWGVLFGFGPGLTVERVVVRSVPIKY	CATALYTIC ACTIVITY: 3 malonyl-CoA + hexanoyl-CoA = 3 CoA + 3,5,7-trioxododecanoyl-CoA + 3 CO(2). {ECO:0000269|PubMed:19454282, ECO:0000269|PubMed:19581347, ECO:0000269|Ref.3}.
L7NCQ3	TBSYN_GARMA	reviewed	2,4,6-trihydroxybenzophenone synthase (GmBPS) (EC 2.3.1.220)	BPS	Garcinia mangostana (Mangosteen)	391	MAPAMDSAQNGHQSRGSANVLAIGTANPPNVILQEDYPDFYFKVTNSEHLTDLKEKFKRICVKSKTRKRHFYLTEQILKENPGIATYGAGSLDSRQKILETEIPKLGKEAAMVAIQEWGQPVSKITHVVFATTSGFMMPGADYSITRLLGLNPNVRRVMIYNQGCFAGGTALRVAKDLAENNKGARVLVVCAENTAMTFHGPNENHLDVLVGQAMFSDGAAALIIGANPNLPEERPVYEMVAAHQTIVPESDGAIVAHFYEMGMSYFLKENVIPLFGNNIEACMEAAFKEYGISDWNSLFYSVHPGGRAIVDGIAEKLGLDEENLKATRHVLSEYGNMGSACVIFILDELRKKSKEEKKLTTGDGKEWGCLIGLGPGLTVETVVLRSVPIA	CATALYTIC ACTIVITY: 3 malonyl-CoA + benzoyl-CoA = 4 CoA + 2,4,6-trihydroxybenzophenone + 3 CO(2). {ECO:0000269|PubMed:22390826}.
O48780	KCS11_ARATH	reviewed	3-ketoacyl-CoA synthase 11 (KCS-11) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 11) (VLCFA condensing enzyme 11)	KCS11 At2g26640 F18A8.1	Arabidopsis thaliana (Mouse-ear cress)	509	MDVEQKKPLIESSDRNLDFKKSVKLKYVKLGYHYLITHGMYLFLSPLVLVIAAQISTFSVTDLRSLWEHLQYNLISVVVCSMLLVFLMTIYFMTRPRPVYLVNFSCFKPDESRKCTKKIFMDRSKLTGSFTEENLEFQRKILQRSGLGESTYLPEAVLNVPPNPCMKEARKEAETVMFGAIDELLAKTNVNPKDIGILIVNCSLFNPTPSLSAMVVNHYKLRGNILSYNLGGMGCSAGLISIDLAKHLLHSIPNTYAMVISMENITLNWYFGNDRSKLVSNCLFRMGGAAILLSNKRWDRRRSKYELVDTVRTHKGADDKCFGCITQEEDSASKIGVTLSKELMAVAGDALKTNITTLGPLVLPTSEQLLFFATLVGRKLFKMKIKPYIPDFKLAFEHFCIHAGGRAVLDELEKNLKLTEWHMEPSRMTLYRFGNTSSSSLWYELAYSEAKGRIKKGDRIWQIAFGSGFKCNSSVWRAVRSVNPKKEKNPWMDEIHEFPVEVPKVSTI	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000269|PubMed:16765910}.

O80400	VPS_HUMLU	reviewed	Phloroisovalerophenone synthase (Valerophenone synthase) (EC 2.3.1.156) (3-methyl-1-(trihydroxyphenyl)butan-1-one synthase)	VPS	Humulus lupulus (European hop)	394	MASVTVEQIRKAQRAEGPATILAIGTAVPANCFNQADFPDYYFRVTKSEHMTDLKKKFQRMCEKSTIKKRYLHLTEEHLKQNPHLCEYNAPSLNTRQDMLVVEVPKLGKEAAINAIKEWGQPKSKITHLIFCTGSSIDMPGADYQCAKLLGLRPSVKRVMLYQLGCYAGGKVLRIAKDIAENNKGARVLIVCSEITACIFRGPSEKHLDCLVGQSLFGDGASSVIVGADPDASVGERPIFELVSAAQTILPNSDGAIAGHVTEAGLTFHLLRDVPGLISQNIEKSLIEAFTPIGINDWNNIFWIAHPGGPAILDEIEAKLELKKEKMKASREMLSEYGNMSCASVFFIVDEMRKQSSKEGKSTTGDGLEWGALFGFGPGLTVETVVLHSVPTNV	CATALYTIC ACTIVITY: Isovaleryl-CoA + 3 malonyl-CoA = 4 CoA + 3 CO(2) + 3-methyl-1-(2,4,6-trihydroxyphenyl)butan-1-one.
P20077	THS2_ARAHY	reviewed	Putative stilbene synthase 2 (EC 2.3.1.95) (Resveratrol synthase 2) (RS2) (Trihydroxystilbene synthase 2) (Fragment)		Arachis hypogaea (Peanut)	313	LKENPNMCAYKAPSLDAREDMMIREVPRVGKEAATKAIKEWGQPMSKITHLIFCTTSGVALPGVDYELIVLLGLDPSVKRYMMYHQGCFAGGTVLRLAKDLAENNKDARVLIVCSENTAVTFRGPSETDMDSLVGQALFADGAAAIIIGSDPVPEVENPLFEIVSTDQKLVPNSHGAIGGLLREVGLTFYLNKSVPDIISQNINDALSKAFDPLGISDYNSIFWIAHPGGPAILDQVEQKVNLKPEKMNATRDVLSNYGNMSSACVFFIMDLMRKKSLEEGLKTTGEGLDWGVLFGFGPGLTIETVVLRSVAI	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
C0SVZ6	CURS1_CURLO	reviewed	Curcumin synthase 1 (EC 2.3.1.217)	CURS1	Curcuma longa (Turmeric) (Curcuma domestica)	389	MANLHALRREQRAQGPATIMAIGTATPPNLYEQSTFPDFYFRVTNSDDKQELKKKFRRMCEKTMVKKRYLHLTEEILKERPKLCSYKEASFDDRQDIVVEEIPRLAKEAAEKAIKEWGRPKSEITHLVFCSISGIDMPGADYRLATLLGLPLTVNRLMIYSQACHMGAAMLRIAKDLAENNRGARVLVVACEITVLSFRGPNEGDFEALAGQAGFGDGAGAVVVGADPLEGIEKPIYEIAAAMQETVAESQGAVGGHLRAFGWTFYFLNQLPAIIADNLGRSLERALAPLGVREWNDVFWVAHPGNWAIIDAIEAKLQLSPDKLSTARHVFTEYGNMQSATVYFVMDELRKRSAVEGRSTTGDGLQWGVLLGFGPGLSIETVVLRSMPL	CATALYTIC ACTIVITY: Feruloyl-CoA + feruloylacetyl-CoA + H(2)O = 2 CoA + curcumin + CO(2). {ECO:0000269|PubMed:19258320, ECO:0000269|PubMed:21148316}.
Q93X68	FABG5_BRANA	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase 5, chloroplastic (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase 5) (Fragment)	bkr1	Brassica napus (Rape)	317	TTVAATKLTSLKATAGKLGYREICQVRQWAPLKSAMPHFGMLRCATSTVVKAQAQAQATATEQTTEEAVPKVESPVVVVTGASRGIGKAIALSLGKAGCKVLVNYARSAKEAEEVSKQIEEYGGEAITFGGDVSKEADVDSMMKTAVDKWGTIDVVVNNAGITRDTLLIRMKKSQWDEVIDLNLTGVFLCTQAATKIMMKKRKGRIINIASVVGLIGNIGQANYAAAKAGVIGFSKTAAREGASRNINVNVVCPGFIASDMTAKLGEDMEKKILGTIPLGRYGQPEYVAGLVEFLALSPASSYITGHTFSIHGGFAI	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH.
Q89AG9	FABG_BUCBP	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase FabG (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase) (Beta-Ketoacyl-acyl carrier protein reductase) (Beta-ketoacyl-ACP reductase)	fabG bbp_321	Buchnera aphidicola subsp. Baizongia pistaciae (strain Bp)	245	MKTTKKIAVITGANRGLGKGIAEELSNTNNITVIGTSTSQKGCKIINKYLKNNGIGIKLDITNPNEITKTMDFVYKNFGRVDILINNAGIIRDKLLINMKTQDWNSVLNVNLNSIFYMSKSVIRNMIKNKQGKIITIGSVIAHIGNCGQTNYSAAKLGLVGFHKSLALELAPKGITVNMIAPGLIKTGMTNNLSQKQLSKYLSKIPMKRLGTIKEISKITLFLISNDANYITGQVIHVNGGMYMP	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH.
P99093	FABG_STAAN	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase FabG (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase) (Beta-Ketoacyl-acyl carrier protein reductase) (Beta-ketoacyl-ACP reductase)	fabG SA1074	Staphylococcus aureus (strain N315)	246	MKMTKSALVTGASRGIGRSIALQLAEEGYNVAVNYAGSKEKAEAVVEEIKAKGVDSFAIQANVADADEVKAMIKEVVSQFGSLDVLVNNAGITRDNLLMRMKEQEWDDVIDTNLKGVFNCIQKATPQMLRQRSGAIINLSSVVGAVGNPGQANYVATKAGVIGLTKSAARELASRGITVNAVAPGFIVSDMTDALSDELKEQMLTQIPLARFGQDTDIANTVAFLASDKAKYITGQTIHVNGGMYM	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH.
Q9PKF7	FABG_CHLMU	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase FabG (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase) (Beta-Ketoacyl-acyl carrier protein reductase) (Beta-ketoacyl-ACP reductase)	fabG TC_0508	Chlamydia muridarum (strain MoPn / Nigg)	248	MNSLLVNKAAIVTGGSRGIGFGIAKLFAEHGANVQIWGINEEAGKSAAQDLSDKTGSKVSFALVDVSKNDMVSAQVQKFLAEYGTIDVVVNNAGITRDSLLMRMSEEEWSSVIDTNLGSIYNVCSAVIRPMIKARSGAIVNISSIVGLRGSPGQTNYAAAKAGIIGFSKALSKEVGSKNIRVNCIAPGFIDTDMTKGLSDNLKNEWLKGVPLGRVGTPEEIAMAALFLASNQSSYITGQVLSVDGGMA	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH.
Q949M2	FABG4_BRANA	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase 4 (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase 4) (Fragment)	bkr4	Brassica napus (Rape)	254	TTTEEEEAVPKVESPVVVVTGASRGIGKAIALSLGKAGCKVLVNYARSAKEAEEVSKQIEEYGGQAITFGGDVSKEADVDAMMKTAVDKWGTIDVVVNNAGDTLLIRMKKSQWDEVMDLNLTGVFLCSQAATKIMMKKRKGRIINIASVVGLIGNIGQANYAAAKAGVIGFSKTAAREGASRNINVNVVCPGFIASDMTAKLGEDMEKKILGTIPLGRYGQPEDVAGLVEFLALSPAASYITGQTFTIDGGIAI	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH.
P9WGT2	FABG_MYCTO	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase FabG1 (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase) (Beta-Ketoacyl-acyl carrier protein reductase) (Beta-ketoacyl-ACP reductase)	fabG1 mabA MT1530	Mycobacterium tuberculosis (strain CDC 1551 / Oshkosh)	247	MTATATEGAKPPFVSRSVLVTGGNRGIGLAIAQRLAADGHKVAVTHRGSGAPKGLFGVECDVTDSDAVDRAFTAVEEHQGPVEVLVSNAGLSADAFLMRMTEEKFEKVINANLTGAFRVAQRASRSMQRNKFGRMIFIGSVSGSWGIGNQANYAASKAGVIGMARSIARELSKANVTANVVAPGYIDTDMTRALDERIQQGALQFIPAKRVGTPAEVAGVVSFLASEDASYISGAVIPVDGGMGMGH	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH.
Q5HGK2	FABG_STAAC	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase FabG (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase) (Beta-Ketoacyl-acyl carrier protein reductase) (Beta-ketoacyl-ACP reductase)	fabG SACOL1245	Staphylococcus aureus (strain COL)	246	MKMTKSALVTGASRGIGRSIALQLAEEGYNVAVNYAGSKEKAEAVVEEIKAKGVDSFAIQANVADADEVKAMIKEVVSQFGSLDVLVNNAGITRDNLLMRMKEQEWDDVIDTNLKGVFNCIQKATPQMLRQRSGAIINLSSVVGAVGNPGQANYVATKAGVIGLTKSAARELASRGITVNAVAPGFIVSDMTDALSDELKEQMLTQIPLARFGQDTDIANTVAFLASDKAKYITGQTIHVNGGMYM	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH.
P43713	FABG_HAEIN	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase FabG (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase) (Beta-Ketoacyl-acyl carrier protein reductase) (Beta-ketoacyl-ACP reductase)	fabG HI_0155	Haemophilus influenzae (strain ATCC 51907 / DSM 11121 / KW20 / Rd)	242	MQGKIALVTGSTRGIGRAIAEELSSKGAFVIGTATSEKGAEAISAYLGDKGKGLVLNVTDKESIETLLEQIKNDFGDIDILVNNAGITRDNLLMRMKDEEWFDIMQTNLTSVYHLSKAMLRSMMKKRFGRIINIGSVVGSTGNPGQTNYCAAKAGVVGFSKSLAKEVAARGITVNVVAPGFIATDMTEVLTDEQKAGILSNVPAGRLGEAKDIAKAVAFLASDDAGYITGTTLHVNGGLYLS	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH.
P0A5Y5	FABG_MYCBO	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase FabG (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase) (Beta-Ketoacyl-acyl carrier protein reductase) (Beta-ketoacyl-ACP reductase)	fabG fabG1 BQ2027_MB1519	Mycobacterium bovis (strain ATCC BAA-935 / AF2122/97)	247	MTATATEGAKPPFVSRSVLVTGGNRGIGLAIAQRLAADGHKVAVTHRGSGAPKGLFGVECDVTDSDAVDRAFTAVEEHQGPVEVLVSNAGLSADAFLMRMTEEKFEKVINANLTGAFRVAQRASRSMQRNKFGRMIFIGSVSGSWGIGNQANYAASKAGVIGMARSIARELSKANVTANVVAPGYIDTDMTRALDERIQQGALQFIPAKRVGTPAEVAGVVSFLASEDASYISGAVIPVDGGMGMGH	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH.
C0SVZ6	CURS1_CURLO	reviewed	Curcumin synthase 1 (EC 2.3.1.217)	CURS1	Curcuma longa (Turmeric) (Curcuma domestica)	389	MANLHALRREQRAQGPATIMAIGTATPPNLYEQSTFPDFYFRVTNSDDKQELKKKFRRMCEKTMVKKRYLHLTEEILKERPKLCSYKEASFDDRQDIVVEEIPRLAKEAAEKAIKEWGRPKSEITHLVFCSISGIDMPGADYRLATLLGLPLTVNRLMIYSQACHMGAAMLRIAKDLAENNRGARVLVVACEITVLSFRGPNEGDFEALAGQAGFGDGAGAVVVGADPLEGIEKPIYEIAAAMQETVAESQGAVGGHLRAFGWTFYFLNQLPAIIADNLGRSLERALAPLGVREWNDVFWVAHPGNWAIIDAIEAKLQLSPDKLSTARHVFTEYGNMQSATVYFVMDELRKRSAVEGRSTTGDGLQWGVLLGFGPGLSIETVVLRSMPL	CATALYTIC ACTIVITY: Feruloyl-CoA + feruloylacetyl-CoA + H(2)O = 2 CoA + curcumin + CO(2). {ECO:0000269|PubMed:19258320, ECO:0000269|PubMed:21148316}.
P20178	THS1_ARAHY	reviewed	Stilbene synthase 1 (EC 2.3.1.95) (Resveratrol synthase 1) (RS1) (Trihydroxystilbene synthase 1)		Arachis hypogaea (Peanut)	389	MVSVSGIRKVQRAEGPATVLAIGTANPPNCVDQSTYADYYFRVTNGEHMTDLKKKFQRICERTQIKNRHMYLTEEILKENPNMCAYKAPSLDAREDMMIREVPRVGKEAATKAIKEWGQPMSKITHLIFCTTSGVALPGVDYELIVLLGLDPSVKRYMMYHQGCFAGGTVLRLAKDLAENNKDARVLIVCSENTAVTFRGPNETDMDSLVGQALFADGAAAIIIGSDPVPEVENPIFEIVSTDQQLVPNSHGAIGGLLREVGLTFYLNKSVPDIISQNINGALSKAFDPLGISDYNSIFWIAHLGGRAILDQVEQKVNLKPEKMKATRDVLSNYGNMSSACVFFIMDLMRKKSLETGLKTTGEGLDWGVLFGFGPGLTIETVVLRSMAI	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
P48407	DPS1_PINST	reviewed	Pinosylvin synthase 1 (EC 2.3.1.146) (Dihydropinosylvin synthase 1) (Stilbene synthase 1) (STS 1)	STS1	Pinus strobus (Eastern white pine)	396	MSVGMGIDLEAFRKSQRADGFASILAIGTANPPNVVDQSTYPDYYFRVTNNEDNTDLKDKFKRICERSAIKKRHMYLTEEILKKNPELCAFLEVPSLDTRQAMLAAEVPRLGKEAAEKAIEEWGQPKSRITHLIFCTTTTPDLPGADFEVAKLLGLHPSVKRVGVFQHGCFAGGTVLRLAKDLAENNRGARVLVVCSENTAVTFRGPSETHLDGLVGLALFGDGASALIVGADPIPQVEKPCFEIVWTAQTVVPNSDGAISGKLREVGLTFQLKGAVPDLISTNIEKCLVEAFSQFNISDWNQLFWIAHPGGHAILDQVEASLNLDPTKLRATRHVMSEYGNMSSACVHFILDETRKASRQNGCSTSGGGFQMGVLFGFGPGLTVETVVLKSIPFP	CATALYTIC ACTIVITY: 3 malonyl-CoA + cinnamoyl-CoA = 4 CoA + pinosylvin + 4 CO(2). {ECO:0000269|PubMed:7698342}.; CATALYTIC ACTIVITY: 3 malonyl-CoA + dihydrocinnamoyl-CoA = 4 CoA + dihydropinosylvin + 4 CO(2). {ECO:0000269|PubMed:7698342}.
P48408	DPS2_PINST	reviewed	Pinosylvin synthase 2 (EC 2.3.1.146) (Dihydropinosylvin synthase 2) (Stilbene synthase 2) (STS 2)	STS2	Pinus strobus (Eastern white pine)	396	MSVGMGVDLEAFRKSQRADGFASILAIGTANPPNVVDQSTYPDYYFRNTNNEDNTDLKDKFKRICERSAIKKRHMYLTEEILKKNPELCAFLEVPSLDTRQAMLAVEVPRLGKEAAEKAIEEWGQPKSRITHLIFCTTTTPDLPGADFEVAKLLGLHPSVKRVGVFQHGCFAGGTVLRLAKDLAENNRGARVLVVCSENTAVTFRGPSETHLDGLVGLALFGDGAAALIVGADPIPQVEKPCFEIVWTAQTVVPNSDGAISGKLREVGLTFQLKGAVPDLISTNIEKCLVEAFSQFNISDWNQLFWIAHPGGRAILDQVEASLNLDPTKLRATRHVMSEYGNMSSACVHFILDETRKASRQNGCSTSGGGFQMGVLFGFGPGLTVETVVLKSIPFP	CATALYTIC ACTIVITY: 3 malonyl-CoA + cinnamoyl-CoA = 4 CoA + pinosylvin + 4 CO(2). {ECO:0000269|PubMed:7698342}.; CATALYTIC ACTIVITY: 3 malonyl-CoA + dihydrocinnamoyl-CoA = 4 CoA + dihydropinosylvin + 4 CO(2). {ECO:0000269|PubMed:7698342}.
P51069	THS3_ARAHY	reviewed	Stilbene synthase 3 (EC 2.3.1.95) (Resveratrol synthase 3) (RS3) (Trihydroxystilbene synthase 3)		Arachis hypogaea (Peanut)	389	MVSVSGIRKVQRAEGPATVLAIGTANPPNCIDQSTYADYYFRVTNSEHMTDLKKKFQRICERTQIKNRHMYLTEEILKENPNMCAYKAPSLDAREDMMIREVPRVGKEAATKAIKEWGQPMSKITHLIFCTTSGVALPGVDYELIVLLGLDPCVKRYMMYHQGCFAGGTVLRLAKDLAENNKDARVLIVCSENTAVTFRGPSETDMDSLVGQALFADGAAAIIIGSDPVPEVEKPIFELVSTDQKLVPGSHGAIGGLLREVGLTFYLNKSVPDIISQNINDALNKAFDPLGISDYNSIFWIAHPGGRAILDQVEQKVNLKPEKMKATRDVLSNYGNMSSACVFFIMDLMRKRSLEEGLKTTGEGLDWGVLFGFGPGLTIETVVLRSVAI	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
P51070	THS2_VITVI	reviewed	Stilbene synthase 2 (EC 2.3.1.95) (PSV21) (Resveratrol synthase 2) (Trihydroxystilbene synthase 2) (StSy 2)	GSVIVT00004047001 LOC100246143 VITISV_010833; GSVIVT00008253001 LOC100259169 VITISV_024260	Vitis vinifera (Grape)	392	MASVEEIRNAQRAKGPATILAIGTATPDHCVYQSDYADYYFRVTKSEHMTALKKKFNRICDKSMIKKRYIHLTEEMLEEHPNIGAYMAPSLNIRQEIITAEVPKLGKEAALKALKEWGQPKSKITHLVFCTTSGVEMPGADYKLANLLGLEPSVRRVMLYHQGCYAGGTVLRTAKDLAENNAGARVLVVCSEITVVTFRGPSEDALDSLVGQALFGDGSAAVIVGSDPDISIERPLFQLVSAAQTFIPNSAGAIAGNLREVGLTFHLWPNVPTLISENIEKCLTQAFDPLGISDWNSLFWIAHPGGPAILDAVEAKLNLDKKKLEATRHVLSEYGNMSSACVLFILDEMRKKSLKGERATTGEGLDWGVLFGFGPGLTIETVVLHSIPMVTN	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
P51071	THS3_VITVI	reviewed	Stilbene synthase 3 (EC 2.3.1.95) (PSV368) (Resveratrol synthase 3) (Trihydroxystilbene synthase 3) (StSy 3)	VIT_16s0100g01030	Vitis vinifera (Grape)	392	MASVEEIRNAQRAKGPATILAIGTATPDHCVYQSDYADYYFRVTKSEHMTELKKKFNRICDKSMIKKRYIHLTEEMLEEHPNIGAYMAPSLNIRQEIITVEVPKLGKEAALKALKEWGQPKSKITHLVFCTTSGVEMPGADYKLANLLGLETSVRRVMLYHQGCYAGGTVLRTAKDLAENNAGARVLVVCSEITVVTFRGPSEDALDSLVGQALFGDGSAAVIVGSDPDVSIERPLFQLVSAAQTFIPNSAGAIAGNLREVGLTFHLWPNVPTLISENVEKCLTQAFDPLGISDWNSLFWIAHPGGPAILDAVEAKLNLDKKKLEATRHVLSEYGNMSSACVLFILDEMRKKSHKGEKATTGEGLDWGVLFGFGPGLTIETVVLHSIPMVTN	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
P53416	BBS_PHASS	reviewed	Bibenzyl synthase (EC 2.3.1.-)	BIBSY212 PCS1	Phalaenopsis sp. (Moth orchid)	390	MLSLESIKKAPRADGFASILAIGRANPDNIIEQSAYPDFYFRVTNSEHLVDLKKKFQRICEKTAIRKRHFVWNEEFLTANPCFSTFMDKSLNVRQEVAISEIPKLGAKAATKAIEDWGQPKSRITHLIFCTTSGMDLPGADYQLTQILGLNPNVERVMLYQQGCFAGGTTLRLAKCLAESRKGARVLVVCAETTTVLFRAPSEEHQDDLVTQALFADGASAVIVGADPDEAADERASFVIVSTSQVLLPDSAGAIGGHVSEGGLLATLHRDVPQIVSKNVGKCLEEAFTPFGISDWNSIFWVPHPGGRAILDQVEERVGLKPEKLSVSRHVLAEYGNMSSVCVHFALDEMRKRSANEGKATTGEGLEWGVLFGFGPGLTVETVVLRSVPL	CATALYTIC ACTIVITY: 3 malonyl-CoA + M-hydroxyphenylpropionyl-CoA = 4 CoA + 4 trihydroxybibenzyl + 3 CO(2).
Q02323	DPSS_PINSY	reviewed	Pinosylvin synthase (EC 2.3.1.146) (Dihydropinosylvin synthase) (Pinosylvin-forming stilbene synthase) (Stilbene synthase) (STS)		Pinus sylvestris (Scots pine)	393	MGGVDFEGFRKLQRADGFASILAIGTANPPNAVDQSTYPDFYFRITGNEHNTELKDKFKRICERSAIKQRYMYLTEEILKKNPDVCAFVEVPSLDARQAMLAMEVPRLAKEAAEKAIQEWGQSKSGITHLIFCSTTTPDLPGADFEVAKLLGLHPSVKRVGVFQHGCFAGGTVLRMAKDLAENNRGARVLVICSETTAVTFRGPSETHLDSLVGQALFGDGASALIVGADPIPQVEKACFEIVWTAQTVVPNSEGAIGGKVREVGLTFQLKGAVPDLISANIENCMVEAFSQFKISDWNKLFWVVHPGGRAILDRVEAKLNLDPTKLIPTRHVMSEYGNMSSACVHFILDQTRKASLQNGCSTTGEGLEMGVLFGFGPGLTIETVVLKSVPIQ	CATALYTIC ACTIVITY: 3 malonyl-CoA + cinnamoyl-CoA = 4 CoA + pinosylvin + 4 CO(2). {ECO:0000269|PubMed:1426272}.; CATALYTIC ACTIVITY: 3 malonyl-CoA + dihydrocinnamoyl-CoA = 4 CoA + dihydropinosylvin + 4 CO(2). {ECO:0000269|PubMed:1426272}.
Q27Z07	BIPS1_SORAU	reviewed	3,5-dihydroxybiphenyl synthase (EC 2.3.1.177) (Biphenyl synthase 1) (SaBIS1)	BIS1	Sorbus aucuparia (European mountain ash) (Rowan)	390	MAPLVKNHGEPQHAKILAIGTANPPNVYYQKDYPDFLFRVTKNEHRTDLREKFDRICEKSRTRKRYLHLTEEILKANPSIYTYGAPSLDVRQDMLNSEVPKLGQQAALKAIKEWGQPISKITHLIFCTASCVDMPGADFQLVKLLGLNPSVTRTMIYEAGCYAGATVLRLAKDFAENNEGARVLVVCAEITTVFFHGLTDTHLDILVGQALFADGASAVIVGANPEPKIERPLFEIVACRQTIIPNSEHGVVANIREMGFTYYLSGEVPKFVGGNVVDFLTKTFEKVDGKNKDWNSLFFSVHPGGPAIVDQVEEQLGLKEGKLRATRHVLSEYGNMGAPSVHFILDDMRKKSIEEGKSTTGEGLEWGVVIGIGPGLTVETAVLRSESIPC	CATALYTIC ACTIVITY: 3 malonyl-CoA + benzoyl-CoA = 4 CoA + 3,5-dihydroxybiphenyl + 4 CO(2). {ECO:0000269|PubMed:14595561, ECO:0000269|PubMed:17109150}.
Q38860	KCS18_ARATH	reviewed	3-ketoacyl-CoA synthase 18 (KCS-18) (EC 2.3.1.199) (Protein FATTY ACID ELONGATION 1) (Very long-chain fatty acid condensing enzyme 18) (VLCFA condensing enzyme 18)	FAE1 KCS18 At4g34520 T4L20.100	Arabidopsis thaliana (Mouse-ear cress)	506	MTSVNVKLLYRYVLTNFFNLCLFPLTAFLAGKASRLTINDLHNFLSYLQHNLITVTLLFAFTVFGLVLYIVTRPNPVYLVDYSCYLPPPHLKVSVSKVMDIFYQIRKADTSSRNVACDDPSSLDFLRKIQERSGLGDETYSPEGLIHVPPRKTFAASREETEKVIIGALENLFENTKVNPREIGILVVNSSMFNPTPSLSAMVVNTFKLRSNIKSFNLGGMGCSAGVIAIDLAKDLLHVHKNTYALVVSTENITQGIYAGENRSMMVSNCLFRVGGAAILLSNKSGDRRRSKYKLVHTVRTHTGADDKSFRCVQQEDDESGKIGVCLSKDITNVAGTTLTKNIATLGPLILPLSEKFLFFATFVAKKLLKDKIKHYYVPDFKLAVDHFCIHAGGRAVIDELEKNLGLSPIDVEASRSTLHRFGNTSSSSIWYELAYIEAKGRMKKGNKAWQIALGSGFKCNSAVWVALRNVKASANSPWQHCIDRYPVKIDSDLSKSKTHVQNGRS	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000269|PubMed:11341960, ECO:0000269|PubMed:12135493, ECO:0000269|PubMed:16765910}.

P38004	FABG_CHLTR	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase FabG (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase) (Beta-Ketoacyl-acyl carrier protein reductase) (Beta-ketoacyl-ACP reductase)	fabG CT_237	Chlamydia trachomatis (strain D/UW-3/Cx)	248	MSGLLVNKTAIVTGGSRGIGFSIAKLFAEQGANVQIWGINGEAGQAAAQTLSEQTGRQVSFALVDVSKNDMVSAQVQNFLAEYNTIDVIVNNAGITRDALLMRMSEEEWSSVINTNLGSIYNVCSAVIRPMIKARSGAIINISSIVGLRGSPGQTNYAAAKAGIIGFSKALSKEVGSKNIRVNCIAPGFIDTDMTKSLNDNLKNEWLKGVPLGRVGMPEEIAKAALFLASDGSSYITGQVLSVDGGMA	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH.
P28643	FABG_CUPLA	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase, chloroplastic (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase)	CLKR27	Cuphea lanceolata (Cigar flower)	320	MATATAAGCSGAVALKSLGGRRLCIPQQLSPVLAGFGSHAAKSFPILSTRSIATSGIRAQVATAEKVSAGAGQSVESPVVIVTGASRGIGKAIALSLGKAGCKVLVNYARSSKEAEEVSKEIEAFGGQALTFGGDVSKEEDVEAMIKTAVDAWGTVDILVNNAGITRDGLLMRMKKSQWQEVIDLNLTGVFLCTQAAAKIMMKKKKGRIINIASVVGLVGNAGQANYSAAKAGVIGFTKTVAREYASRNINVNAVAPGFISSDMTSKLGDDINKKILETIPLGRYGQPEEVAGLVEFLAINPASSYVTGQVFTIDGGMTM	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH.
Q94FV7	BAS_RHEPA	reviewed	Polyketide synthase BAS (EC 2.3.1.212) (Benzalacetone synthase) (RpBAS)	BAS	Rheum palmatum (Chinese rhubarb)	384	MATEEMKKLATVMAIGTANPPNCYYQADFPDFYFRVTNSDHLINLKQKFKRLCENSRIEKRYLHVTEEILKENPNIAAYEATSLNVRHKMQVKGVAELGKEAALKAIKEWGQPKSKITHLIVCCLAGVDMPGADYQLTKLLDLDPSVKRFMFYHLGCYAGGTVLRLAKDIAENNKGARVLIVCSEMTTTCFRGPSETHLDSMIGQAILGDGAAAVIVGADPDLTVERPIFELVSTAQTIVPESHGAIEGHLLESGLSFHLYKTVPTLISNNIKTCLSDAFTPLNISDWNSLFWIAHPGGPAILDQVTAKVGLEKEKLKVTRQVLKDYGNMSSATVFFIMDEMRKKSLENGQATTGEGLEWGVLFGFGPGITVETVVLRSVPVIS	CATALYTIC ACTIVITY: 4-coumaroyl-CoA + malonyl-CoA + H(2)O = 2 CoA + 4-hydroxybenzalacetone + 2 CO(2). {ECO:0000269|PubMed:11389739, ECO:0000269|PubMed:17383877}.
Q9C6L5	KCS5_ARATH	reviewed	3-ketoacyl-CoA synthase 5 (KCS-5) (EC 2.3.1.199) (Eceriferum 60) (Very long-chain fatty acid condensing enzyme 5) (VLCFA condensing enzyme 5)	KCS5 CER60 At1g25450 F2J7.9	Arabidopsis thaliana (Mouse-ear cress)	492	MSDFSSSVKLKYVKLGYQYLINNFLTLLLIPVIATVAIELLRMGPEEILSVLNSLHFELLHILCSSFLIIFVSTVYFMSKPRTVYLVDYSCYKPPVTCRVPFSSFMEHSRLILKDNPKSVEFQMRILERSGLGEETCLPPAIHYIPPTPTMESARNEAQMVIFTAMEDLFKNTGLKPKDIDILIVNCSLFSPTPSLSAMIINKYKLRSNIKSYNLSGMGCSASLISVDVARDLLQVHPNSNAIIISTEIITPNYYKGNERAMLLPNCLFRMGGAAILLSNRRSDRWRAKYKLCHLVRTHRGADDKSYNCVMEQEDKNGNVGINLSKDLMTIAGEALKANITTIGPLVLPASEQLLFLSSLIGRKIFNPKWKPYIPDFKQAFEHFCIHAGGRAVIDELQKNLQLSGEHVEASRMTLHRFGNTSSSSLWYELSYIEAQGRMKRNDRVWQIAFGSGFKCNSAVWKCNRTIKTPTDGAWSDCIERYPVFIPEVVKL	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9C992	KCS7_ARATH	reviewed	3-ketoacyl-CoA synthase 7 (KCS-7) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 7) (VLCFA condensing enzyme 7)	KCS7 At1g71160 F23N20.15	Arabidopsis thaliana (Mouse-ear cress)	460	MESSFHFINEALLITQTFITFHQFLVASACVLIAVFGYYFFKPRCIIYLIDFSCYQPPDFLRAPVSNFIEHLTISGVFDQESLDLQQKILERSGISDDASVPATVHEIPPNASISAAREETHEILFAIVQDLFSKHEIDPKSIDILVSNCSLFCPSPSITSMIINKFGMRSDIKSFSLSGMGCSAGILSVNLVKDLMKIHGDSLALVLSMEAVSPNGYRGKCKSMLIANTIFRMGGAAILLSNRKQDSHKAKYKLQHIIRTHVGSDTESYESVMQQVDEEGKVGVALSKQLVRVASKALKINVVQLGPRVLPYSEQLKYIISFIQRKWGMHKEIYTPNFKKAFEHFCIHAGGRAIIEGVEKHLKLDKEDVEASRSTLYRYGNTSSSSLWYELQYLEAKGRMKMGDKVWQIGFGSGFKANSAVWKCISEIDSRGRNAWSDRIHLYPVCGDTSSALKTELLS	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9FG87	KCS20_ARATH	reviewed	3-ketoacyl-CoA synthase 20 (KCS-20) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 20) (VLCFA condensing enzyme 20)	KCS20 KCS19 At5g43760 MQD19.11	Arabidopsis thaliana (Mouse-ear cress)	529	MSHNQNQPHRPVPVHVTNAEPNPNPNNLPNFLLSVRLKYVKLGYHYLISNALYILLLPLLAATIANLSSFTINDLSLLYNTLRFHFLSATLATALLISLSTAYFTTRPRRVFLLDFSCYKPDPSLICTRETFMDRSQRVGIFTEDNLAFQQKILERSGLGQKTYFPEALLRVPPNPCMEEARKEAETVMFGAIDAVLEKTGVKPKDIGILVVNCSLFNPTPSLSAMIVNKYKLRGNILSYNLGGMGCSAGLISIDLAKQMLQVQPNSYALVVSTENITLNWYLGNDRSMLLSNCIFRMGGAAVLLSNRSSDRSRSKYQLIHTVRTHKGADDNAFGCVYQREDNNAEETGKIGVSLSKNLMAIAGEALKTNITTLGPLVLPMSEQLLFFATLVARKVFKVKKIKPYIPDFKLAFEHFCIHAGGRAVLDEIEKNLDLSEWHMEPSRMTLNRFGNTSSSSLWYELAYSEAKGRIKRGDRTWQIAFGSGFKCNSAVWKALRTIDPMDEKTNPWIDEIDDFPVQVPRITPITSS	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9FH27	KCS21_ARATH	reviewed	Probable 3-ketoacyl-CoA synthase 21 (KCS-21) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 21) (VLCFA condensing enzyme 21)	KCS21 KCS20 At5g49070 K20J1.4	Arabidopsis thaliana (Mouse-ear cress)	464	MNQTIHRVSPISMSISELTTLLSSGVSVFEIFAGLLVVHLIYQRIRTRVKVYLLDFTCYRAPDSNRVPMSTLIETIYLDDKLDQESIDFQARILERSWLSNQTSIPRSLMEIPLKKSLSSVKIETMTTIFTSVEDLLRKNKLSPRSIDILITNCSLHSPSPSLSAMVINKFHMRSNIKSFNLSGMGCAAGILSVNLANDLLQAHRGSLALIVSTEALNTHWYIGKDRSMLLTNCLFRMGAAAVLMSSNDHDRDNAKYELLHVVRKNKAKDDRAYRCIYQDIDSDEKQGVSITKDVISVAGDMLKMNLTSLGPLVLPYLEQFQYVIQHILCKKLKIYESNSSYTPNFKTAFEHFCIHTGGRAVIQAMEMNLKLTKVDIEPSKMTLHRFGNTSSSSIWYALSYLEAKRRMKKGDRVLQIAFGSGFKCNSAVWRCIRKVEPNTENKWLDFIDSYPVDVPDSTNIRPG	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9FSC0	ACS2_RUTGR	reviewed	Acridone synthase 2 (EC 2.3.1.159) (Acridone synthase II)	ACS2	Ruta graveolens (Common rue)	391	MESLKEMRKAQKSEGPAAILAIGTATPDNVYIQADYPDYYFKITKSEHMTELKDKFKTLCEKSMIRKRHMCFSQEFLKANPEVCKHMGKSLNARQDIAVVETPRIGKEAAVKAIKEWGHPKSSITHLIFCTSAGVDMPGADYQLTRMLGLNPSVKRMMIYQQGCYAGGTVLRLAKDLAENNKGSRVLVVCSELTAPTFRGPSPDAVDSLVGQALFADGAAALVVGADPDTSVERALYYIVSASQMLLPDSDGAIEGHIREEGLTVHLKKDVPALFSANIDTPLVEAFRPLGISDWNSIFWIAHPGGPAILDQIEVKLGLKEDKLRASKHVMSEYGNMSSSCVLFVLDEMRNKSLQDGKSTTGEGLDWGVLFGFGPGLTVETVVLRSVPVEA	CATALYTIC ACTIVITY: 3 malonyl-CoA + N-methylanthraniloyl-CoA = 4 CoA + 1,3-dihydroxy-N-methylacridone + 3 CO(2).
Q9FSC1	ACS4_RUTGR	reviewed	Probable acridone synthase 4 (EC 2.3.1.159) (Acridone synthase IV)	ACS4	Ruta graveolens (Common rue)	391	MESLKEMRKAQMSEGPAAILAIGTATPNNVYMQADYPDYYFRMTKSEHMTELKDKFRTLCEKSMIRKRHMCFSEEFLKANPEVSKHMGKSLNARQDIAVVETPRLGNEAAVKAIKEWGQPKSSITHLIFCSSAGVDMPGADYQLTRILGLNPSVKRMMVYQQGCYAGGTVLRLAKDLAENNKGSRVLVVCSELTAPTFRGPSPDAVDSLVGQALFADGAAALVVGADPDSSIERALYYLVSASQMLLPDSDGAIEGHIREEGLTVHLKKDVPALFSANIDTPLVEAFKPLGISDWNSIFWIAHPGGPAILDQIEEKLGLKEDKLRASKHVMSEYGNMSSSCVLFVLDEMRSRSLQDGKSTTGEGLDWGVLFGFGPGLTVETVVLRSVPIEA	CATALYTIC ACTIVITY: 3 malonyl-CoA + N-methylanthraniloyl-CoA = 4 CoA + 1,3-dihydroxy-N-methylacridone + 3 CO(2).
Q9FSC2	ACS3_RUTGR	reviewed	Probable acridone synthase 3 (EC 2.3.1.159) (Acridone synthase III)	ACS3	Ruta graveolens (Common rue)	391	MESLKEMRKAQMSEGPAAILAIGTANPDNVYMQADYPDYYFKMTKSEHMTELKDKFRTLCEKSMIRKRHMCFSEEFLKANPEVCKHMGKSLNARQDIAVVETPRLGNEAAVKAIKEWGQPKSSITHLIFCSSAGVDMPGADYQLTRILGLNPSVKRMMIYQQGCYAGGTVLRLAKDLAENNKGSRVLVVCSELTAPTFRGPSPDAVDSLVGQALFADGAAALVVGADPDSSIERALYYLVSASQMLLPDSDGAIEGHIREEGLTVHLKKDVPALFSGNIDTPLVEAFKPLGISDWNSIFWIAHPGGPAILDQIEEKLGLKEDKLRASKHVMSEYGNMSSSCVLFVLDEMRSRSLQDGKSTTGEGLDWGVLFGFGPGLTVETIVLRSVPIEA	CATALYTIC ACTIVITY: 3 malonyl-CoA + N-methylanthraniloyl-CoA = 4 CoA + 1,3-dihydroxy-N-methylacridone + 3 CO(2).
Q9LQP8	KCS3_ARATH	reviewed	3-ketoacyl-CoA synthase 3 (KCS-3) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 3) (VLCFA condensing enzyme 3)	KCS3 At1g07720 F24B9.18	Arabidopsis thaliana (Mouse-ear cress)	478	MDLLVMLLSLLVSYLIFKIWKRIDSKRDQNCYILDYQCHKPSDDRMVNTQFSGDIILRNKHLRLNEYKFLLKAIVSSGIGEQTYAPRLFFEGREQRPTLQDGLSEMEEFYIDTIEKVLKRNKISPSEIDILVVNVSMLNSTPSLSARIINHYKMREDIKVFNLTAMGCSASVISIDIVKNIFKTYKNKLALVVTSESLSPNWYSGNNRSMILANCLFRSGGCAVLLTNKRSLSRRAMFKLRCLVRTHHGARDDSFNACVQKEDELGHIGVHLDKTLPKAATRAFIDNLKVITPKILPVTELLRFMLCLLLKKLRSSPSKGSTNVTQAAPKAGVKAGINFKTGIDHFCIHTGGKAVIDAIGYSLDLNEYDLEPARMTLHRFGNTSASSLWYVLGYMEAKKRLKRGDRVFMISFGAGFKCNSCVWEVVRDLNVGEAVGNVWNHCINQYPPKSILNPFFEKYGWIHEEEDPDTFKMPEGFM	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9LZ72	KCS19_ARATH	reviewed	3-ketoacyl-CoA synthase 19 (KCS-19) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 19) (VLCFA condensing enzyme 19)	KCS19 KCS21 At5g04530 T32M21.130	Arabidopsis thaliana (Mouse-ear cress)	464	MELFSLSSLLLLSTLFVFYIFKFVFKRRNQRNCYMLHYECYKGMEERKLDTETCAKVVQRNKNLGLEEYRFLLRTMASSGIGEETYGPRNVLEGREDSPTLLDAHSEMDEIMFDTLDKLFHKTKGSISPSDIDILVVNVSLFAPSPSLTSRVINRYKMREDIKSYNLSGLGCSASVISIDIVQRMFETRENALALVVSTETMGPHWYCGKDRSMMLSNCLFRAGGSSVLLTNAARFKNQALMKLVTVVRAHVGSDDEAYSCCIQMEDRDGHPGFLLTKYLKKAAARALTKNLQVLLPRVLPVKELIRYAIVRALKRRTSAKREPASSGIGLNLKTGLQHFCIHPGGRAIIEGVGKSLGLTEFDIEPARMALHRFGNTSSGGLWYVLGYMEAKNRLKKGEKILMMSMGAGFESNNCVWEVLKDLDDKNVWEDSVDRYPELSRIPNPFVEKYDWINDDTMSFVRVD	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9MAM3	KCS1_ARATH	reviewed	3-ketoacyl-CoA synthase 1 (KCS-1) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 1) (VLCFA condensing enzyme 1)	KCS1 EL1 At1g01120 T25K16.11	Arabidopsis thaliana (Mouse-ear cress)	528	MERTNSIEMDRERLTAEMAFRDSSSAVIRIRRRLPDLLTSVKLKYVKLGLHNSCNVTTILFFLIILPLTGTVLVQLTGLTFDTFSELWSNQAVQLDTATRLTCLVFLSFVLTLYVANRSKPVYLVDFSCYKPEDERKISVDSFLTMTEENGSFTDDTVQFQQRISNRAGLGDETYLPRGITSTPPKLNMSEARAEAEAVMFGALDSLFEKTGIKPAEVGILIVNCSLFNPTPSLSAMIVNHYKMREDIKSYNLGGMGCSAGLISIDLANNLLKANPNSYAVVVSTENITLNWYFGNDRSMLLCNCIFRMGGAAILLSNRRQDRKKSKYSLVNVVRTHKGSDDKNYNCVYQKEDERGTIGVSLARELMSVAGDALKTNITTLGPMVLPLSEQLMFLISLVKRKMFKLKVKPYIPDFKLAFEHFCIHAGGRAVLDEVQKNLDLKDWHMEPSRMTLHRFGNTSSSSLWYEMAYTEAKGRVKAGDRLWQIAFGSGFKCNSAVWKALRPVSTEEMTGNAWAGSIDQYPVKVVQ	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000269|PubMed:10074711, ECO:0000269|PubMed:16765910}.
Q4V3C9	KCS8_ARATH	reviewed	3-ketoacyl-CoA synthase 8 (KCS-8) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 8) (VLCFA condensing enzyme 8)	KCS8 At2g15090 T15J14.13	Arabidopsis thaliana (Mouse-ear cress)	481	MKNLKMVFFKILFISLMAGLAMKGSKINVEDLQKFSLHHTQNNLQTISLLLFLVVFVWILYMLTRPKPVYLVDFSCYLPPSHLKVSIQTLMGHARRAREAGMCWKNKESDHLVDFQEKILERSGLGQETYIPEGLQCFPLQQGMGASRKETEEVIFGALDNLFRNTGVKPDDIGILVVNSSTFNPTPSLASMIVNKYKLRDNIKSLNLGGMGCSAGVIAVDVAKGLLQVHRNTYAIVVSTENITQNLYLGKNKSMLVTNCLFRVGGAAVLLSNRSRDRNRAKYELVHTVRIHTGSDDRSFECATQEEDEDGIIGVTLTKNLPMVAARTLKINIATLGPLVLPLKEKLAFFITFVKKKYFKPELRNYTPDFKLAFEHFCIHAGGRALIDELEKNLKLSPLHVEASRMTLHRFGNTSSSSIWYELAYTEAKGRMKEGDRIWQIALGSGFKCNSSVWVALRDVKPSANSPWEDCMDRYPVEIDI	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q570B4	KCS10_ARATH	reviewed	3-ketoacyl-CoA synthase 10 (KCS-10) (EC 2.3.1.199) (Protein FIDDLEHEAD) (Very long-chain fatty acid condensing enzyme 10) (VLCFA condensing enzyme 10)	FDH EL4 KCS10 At2g26250 T1D16.11	Arabidopsis thaliana (Mouse-ear cress)	550	MGRSNEQDLLSTEIVNRGIEPSGPNAGSPTFSVRVRRRLPDFLQSVNLKYVKLGYHYLINHAVYLATIPVLVLVFSAEVGSLSREEIWKKLWDYDLATVIGFFGVFVLTACVYFMSRPRSVYLIDFACYKPSDEHKVTKEEFIELARKSGKFDEETLGFKKRILQASGIGDETYVPRSISSSENITTMKEGREEASTVIFGALDELFEKTRVKPKDVGVLVVNCSIFNPTPSLSAMVINHYKMRGNILSYNLGGMGCSAGIIAIDLARDMLQSNPNSYAVVVSTEMVGYNWYVGSDKSMVIPNCFFRMGCSAVMLSNRRRDFRHAKYRLEHIVRTHKAADDRSFRSVYQEEDEQGFKGLKISRDLMEVGGEALKTNITTLGPLVLPFSEQLLFFAALLRRTFSPAAKTSTTTSFSTSATAKTNGIKSSSSDLSKPYIPDYKLAFEHFCFHAASKVVLEELQKNLGLSEENMEASRMTLHRFGNTSSSGIWYELAYMEAKESVRRGDRVWQIAFGSGFKCNSVVWKAMRKVKKPTRNNPWVDCINRYPVPL	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q58VP7	PCS_ALOAR	reviewed	5,7-dihydroxy-2-methylchromone synthase (EC 2.3.1.216) (Pentaketide chromone synthase) (PCS)		Aloe arborescens (Kidachi aloe)	403	MSSLSNSLPLMEDVQGIRKAQKADGTATVMAIGTAHPPHIFPQDTYADVYFRATNSEHKVELKKKFDHICKKTMIGKRYFNYDEEFLKKYPNITSYDEPSLNDRQDICVPGVPALGTEAAVKAIEEWGRPKSEITHLVFCTSCGVDMPSADFQCAKLLGLHANVNKYCIYMQGCYAGGTVMRYAKDLAENNRGARVLVVCAELTIMMLRAPNETHLDNAIGISLFGDGAAALIIGSDPIIGVEKPMFEIVCTKQTVIPNTEDVIHLHLRETGMMFYLSKGSPMTISNNVEACLIDVFKSVGITPPEDWNSLFWIPHPGGRAILDQVEAKLKLRPEKFRAARTVLWDYGNMVSASVGYILDEMRRKSAAKGLETYGEGLEWGVLLGFGPGITVETILLHSLPLM	CATALYTIC ACTIVITY: 5 malonyl-CoA = 5 CoA + 5,7-dihydroxy-2-methyl-4H-chromen-4-one + 5 CO(2) + H(2)O. {ECO:0000269|PubMed:15686354}.
Q5XEP9	KCS2_ARATH	reviewed	3-ketoacyl-CoA synthase 2 (KCS-2) (EC 2.3.1.199) (Docosanoic acid synthase) (Very long-chain fatty acid condensing enzyme 2) (VLCFA condensing enzyme 2)	KCS2 DAISY KCS17 At1g04220 F20D22.1	Arabidopsis thaliana (Mouse-ear cress)	528	MNENHIQSDHMNNTIHVTNKKLPNFLLSVRLKYVKLGYHYLISNAVYILILPVGLLAATSSSFSLTDLTLLYNHLLKFHFLSSTLFAALLIFLTTLYFTTRPRRIFLLDFACYKPDSSLICTRETFMDRSQRVGIFTEDNLAFQQKILERSGLGQKTYFPEALLRVPPNPCMSEARKEAETVMFGAIDAVLEKTGVNPKDIGILVVNCSLFNPTPSLSAMIVNKYKLRGNVLSYNLGGMGCSAGLISIDLAKQLLQVQPNSYALVVSTENITLNWYLGNDRSMLLSNCIFRMGGAAVLLSNRSSDRCRSKYQLIHTVRTHKGSDDNAFNCVYQREDNDDNKQIGVSLSKNLMAIAGEALKTNITTLGPLVLPMSEQLLFFATLVARKVFNVKKIKPYIPDFKLAFEHFCIHAGGRAVLDEIEKNLDLSEWHMEPSRMTLNRFGNTSSSSLWYELAYSEAKGRIKRGDRTWQIAFGSGFKCNSAVWRALRTIDPSKEKKKKTNPWIDEIHEFPVPVPRTSPVTSSSESR	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q8LIL0	CUS_ORYSJ	reviewed	Bisdemethoxycurcumin synthase (EC 2.3.1.211) (Curcuminoid synthase)	Os07g0271500 LOC_Os07g17010 OJ1001_C01.122 OSJNBb0002J01.6	Oryza sativa subsp. japonica (Rice)	402	MAPTTTMGSALYPLGEMRRSQRADGLAAVLAIGTANPPNCVTQEEFPDFYFRVTNSDHLTALKDKFKRICQEMGVQRRYLHHTEEMLSAHPEFVDRDAPSLDARLDIAADAVPELAAEAAKKAIAEWGRPAADITHLVVTTNSGAHVPGVDFRLVPLLGLRPSVRRTMLHLNGCFAGCAALRLAKDLAENSRGARVLVVAAELTLMYFTGPDEGCFRTLLVQGLFGDGAAVVIVGADADDVERPLFEIVSAAQTIIPESDHALNMRFTERRLDGVLGRQVPGLIGDNVERCLLDMFGPLLGGDGGGGWNDLFWAVHPGSSTIMDQVDAALGLEPGKLAASRRVLSDYGNMSGATVIFALDELRRQRKEAAAAGEWPELGVMMAFGPGMTVDAMLLHATSHVN	CATALYTIC ACTIVITY: 2 4-coumaroyl-CoA + malonyl-CoA + H(2)O = 3 CoA + bisdemethoxycurcumin + 2 CO(2).
Q8SAS8	TBSYN_HYPAN	reviewed	2,4,6-trihydroxybenzophenone synthase (EC 2.3.1.220) (2,3',4,6-tetrahydroxybenzophenone synthase) (EC 2.3.1.151) (Benzophenone synthase) (HaBPS)	BPS	Hypericum androsaemum (Tutsan)	395	MAPAMEYSTQNGQGEGKKRASVLAIGTTNPEHFILQEDYPDFYFRNTNSEHMTELKEKFKRICVKSHIRKRHFYLTEEILKENQGIATYGAGSLDARQRILETEVPKLGQEAALKAIAEWGQPISKITHVVFATTSGFMMPGADYVITRLLGLNRTVRRVMLYNQGCFAGGTALRVAKDLAENNEGARVLVVCAENTAMTFHAPNESHLDVIVGQAMFSDGAAALIIGACPDVASGERAVFNILSASQTIVPGSDGAITAHFYEMGMSYFLKEDVIPLFRDNIAAVMEEAFSPLGVSDWNSLFYSIHPGGRGIIDGVAGNLGIKDENLVATRHVLGEYGNMGSACVMFILDELRKSSKVNGKPTTGDGKEFGCLIGLGPGLTVEAVVLQSVPILQ	CATALYTIC ACTIVITY: 3 malonyl-CoA + benzoyl-CoA = 4 CoA + 2,4,6-trihydroxybenzophenone + 3 CO(2). ECO:0000269|PubMed:9459298}.
A5C9M2	THS5_VITVI	reviewed	Stilbene synthase 5 (EC 2.3.1.95) (Resveratrol synthase 5) (Trihydroxystilbene synthase 5) (StSy 5)	GSVIVT00007357001 LOC100250301 VITISV_036852	Vitis vinifera (Grape)	392	MASVEEIRNAQRAKGPATILAIGTATPDHCVYQSDYADYYFKVTKSEHMTELKKKFNRICDKSMIKKRYIHLTEEMLEEHPNIGAYMAPSLNIRQEIITAEVPKLGKEAALKALKEWGQPKSKITHLVFCTTSGVEMPGADYKLANLLGLETSVRRVMLYHQGCYAGGTVLRTAKDLAENNAGARVLVVCSEITVVTFRGPSEDALDSLVGQALFGDGSAAVIVGSDPNVSIERPLFQLVSAAQTFIPNSAGAIAGNLREVGLTFHLWPNVPTLISENIEKCLSQAFDPLGISNWNSLFWIAHPGGPAILDAVEAKLNLEKKKLEATRHVLSEYGNMSSACVLFILDEMRKKSLKGEKATTGEGLDWGVLFGFGPGLTIETVVLHSVPMVTN	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
Q68VY7	FABG_RICTY	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase FabG (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase) (Beta-Ketoacyl-acyl carrier protein reductase) (Beta-ketoacyl-ACP reductase)	fabG RT0748	Rickettsia typhi (strain ATCC VR-144 / Wilmington)	241	MIDFTGKTSLITGASGGIGSAIARLLHKLGSKVIISGSNEKKLKLLGNTLKDNYIIEVCNLANKEECNNLISKISNLDILVCNAGITSDTLAIRMKDQDFDKVIDINLKANFILNREAIKKMIQKRYGRIINISSIVGIAGNPGQANYCASKAGLIGMTKSLSYEVATRGITVNAVAPGFIKSDMTDKLNEKQREAIVQKIPLGTYGIPEDVAYAVAFLASNHASYITGQTLHVNGGMLMV	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH.
O42866	FABG_SCHPO	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase)	oar2 SPAC3G9.02	Schizosaccharomyces pombe (strain 972 / ATCC 24843) (Fission yeast)	236	MRKVLITGGSSGLGKRIAQIWSQKGHQCHIVGRNEFHLKETLQSLSVAKGQQHTLTIADVQSDMKNLKSIFESVEIDTVVHAAGVLQSSLCVRTSEKEIDSIICTNLVSAIKLSKMAILEWFRNKNSERDRLILNISSRLSTYALPGTSVYAASKAGLESFTKVLAAEVASKGIRVNAISPGYVDTPMLSSQIRAIAEKKVPIGRLASTDEIVDACTFLLDNRYTTGTILPITGGL	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH.
P73574	FABG_SYNY3	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase)	fabG slr0886	Synechocystis sp. (strain PCC 6803 / Kazusa)	247	MTALTAQVALVTGASRGIGKATALALAATGMKVVVNYAQSSTAADAVVAEIIANGGEAIAVQANVANADEVDQLIKTTLDKFSRIDVLVNNAGITRDTLLLRMKLEDWQAVIDLNLTGVFLCTKAVSKLMLKQKSGRIINITSVAGMMGNPGQANYSAAKAGVIGFTKTVAKELASRGVTVNAVAPGFIATDMTENLNAEPILQFIPLARYGQPEEVAGTIRFLATDPAAAYITGQTFNVDGGMVMF	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH. {ECO:0000305|PubMed:26358291}.
P35731	FABG_YEAST	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase)	OAR1 YKL055C	Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast)	278	MHYLPVAIVTGATRGIGKAICQKLFQKGLSCIILGSTKESIERTAIDRGQLQSGLSYQRQCAIAIDFKKWPHWLDYESYDGIEYFKDRPPLKQKYSTLFDPCNKWSNNERRYYVNLLINCAGLTQESLSVRTTASQIQDIMNVNFMSPVTMTNICIKYMMKSQRRWPELSGQSARPTIVNISSILHSGKMKVPGTSVYSASKAALSRFTEVLAAEMEPRNIRCFTISPGLVKGTDMIQNLPVEAKEMLERTIGASGTSAPAEIAEEVWSLYSRTALET	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH.
O67610	FABG_AQUAE	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase FabG (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase) (Beta-Ketoacyl-acyl carrier protein reductase) (Beta-ketoacyl-ACP reductase)	fabG aq_1716	Aquifex aeolicus (strain VF5)	248	MEIKLQGKVSLVTGSTRGIGRAIAEKLASAGSTVIITGTSGERAKAVAEEIANKYGVKAHGVEMNLLSEESINKAFEEIYNLVDGIDILVNNAGITRDKLFLRMSLLDWEEVLKVNLTGTFLVTQNSLRKMIKQRWGRIVNISSVVGFTGNVGQVNYSTTKAGLIGFTKSLAKELAPRNVLVNAVAPGFIETDMTAVLSEEIKQKYKEQIPLGRFGSPEEVANVVLFLCSELASYITGEVIHVNGGMF	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH.
Q1RKB7	FABG_RICBR	reviewed	3-oxoacyl-[acyl-carrier-protein] reductase FabG (EC 1.1.1.100) (3-ketoacyl-acyl carrier protein reductase) (Beta-Ketoacyl-acyl carrier protein reductase) (Beta-ketoacyl-ACP reductase)	fabG RBE_0116	Rickettsia bellii (strain RML369-C)	241	MIDLSGQTALITGASGGIGGAIARQLHKLGSHVIISGSNEEKLKALGNDLKDNYTIKVCNLTNTEECSNLVAQIEKLDILVCNAGITKDTLAIRMKNEDFDQVIDINLKANFILNREAIKKMMTNRYGRIINITSIVGVSGNPGQANYCASKAGLIGMTKSLAYEVATRGITVNAVAPGFIKSDMTDKLNDEQKEAITRKIPKGTYGMPEDIANAVAFLASKQSSYITGQTLHVNGGMLMV	CATALYTIC ACTIVITY: (3R)-3-hydroxyacyl-[acyl-carrier-protein] + NADP(+) = 3-oxoacyl-[acyl-carrier-protein] + NADPH.

Q9SIB2	KCS12_ARATH	reviewed	3-ketoacyl-CoA synthase 12 (KCS-12) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 12) (VLCFA condensing enzyme 12)	KCS12 At2g28630 T8O18.8	Arabidopsis thaliana (Mouse-ear cress)	476	MDLLFLFFSLLLSYLFFKIWKLIDSKQDKDCYILDYQCHKPTDDRMVSTQFSGEIIYRNQNLGLTEYKFLLKAIVSSGIGEQTYAPRLVFEGREERPSLQDGISEMEEFYVDSIGKLLERNQISPKDIDILVVNVSMLSSTPSLASRIINHYKMRDDVKVFNLTGMGCSASLISVDIVKNIFKSYANKLALVATSESLSPNWYSGNNRSMILANCLFRSGGCAILLTNKRSLRKKAMFKLKCMVRTHHGAREESYNCCIQAEDEQGRVGFYLGKNLPKAATRAFVENLKVITPKILPVTELIRFMLKLLIKKIKIRQNPSKGSTNLPPGTPLKAGINFKTGIEHFCIHTGGKAVIDGIGHSLDLNEYDIEPARMTLHRFGNTSASSLWYVLAYMEAKKRLKRGDRVFMISFGAGFKCNSCVWEVVRDLTGGESKGNVWNHCIDDYPPKSILNPYLEKFGWIQDEDPDTFKVPDAFM	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9SIX1	KCS9_ARATH	reviewed	3-ketoacyl-CoA synthase 9 (KCS-9) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 9) (VLCFA condensing enzyme 9)	KCS9 At2g16280 F16F14.22	Arabidopsis thaliana (Mouse-ear cress)	512	MEAANEPVNGGSVQIRTENNERRKLPNFLQSVNMKYVKLGYHYLITHLFKLCLVPLMAVLVTEISRLTTDDLYQIWLHLQYNLVAFIFLSALAIFGSTVYIMSRPRSVYLVDYSCYLPPESLQVKYQKFMDHSKLIEDFNESSLEFQRKILERSGLGEETYLPEALHCIPPRPTMMAAREESEQVMFGALDKLFENTKINPRDIGVLVVNCSLFNPTPSLSAMIVNKYKLRGNVKSFNLGGMGCSAGVISIDLAKDMLQVHRNTYAVVVSTENITQNWYFGNKKAMLIPNCLFRVGGSAILLSNKGKDRRRSKYKLVHTVRTHKGAVEKAFNCVYQEQDDNGKTGVSLSKDLMAIAGEALKANITTLGPLVLPISEQILFFMTLVTKKLFNSKLKPYIPDFKLAFDHFCIHAGGRAVIDELEKNLQLSQTHVEASRMTLHRFGNTSSSSIWYELAYIEAKGRMKKGNRVWQIAFGSGFKCNSAVWVALNNVKPSVSSPWEHCIDRYPVKLDF	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9SLX9	VPS_PSINU	reviewed	Phloroisovalerophenone synthase (Valerophenone synthase) (EC 2.3.1.156) (3-methyl-1-(trihydroxyphenyl)butan-1-one synthase)	VPS	Psilotum nudum (Whisk fern) (Lycopodium nudum)	406	MIQNSDSATATLLPRKKERASGPVSVLAIGSANPPNVFHQSLFPDFYFNITQSNHMAEVKAKFTRMCAKSGIKKRRMHINEDILEAHPSIRSYHDNSLDVRQDMLVEEVPKLGKVAADNAIAEWGQPKSNITHLIFCTSSGIDMPGADWALMKLLGLRPTVNRVMVYQQGCFAGCTVLRIAKDLAENNKGSRILVVCSELTLISFRGPTEDHPENLVGQALFGDGAAALIVGADPIPHAENASFEIHWARSSVVPDSDDAVTGNIKENGLVLHLSKTIPDLIGQNIHTLLKDALEEMFDACNPSSFNDLFWVIHPGGPAILDAVEEELNLKSERTHASREILSQYGNMVSPGVLFVLDYMRKRSVDERLSTTGEGLEWGVMLGFGPGLTVETLILKSVPTQAFKYF	CATALYTIC ACTIVITY: Isovaleryl-CoA + 3 malonyl-CoA = 4 CoA + 3 CO(2) + 3-methyl-1-(2,4,6-trihydroxyphenyl)butan-1-one.
Q9SS39	KCS14_ARATH	reviewed	Probable 3-ketoacyl-CoA synthase 14 (KCS-14) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 14) (VLCFA condensing enzyme 14)	KCS14 At3g10280 F14P13.12	Arabidopsis thaliana (Mouse-ear cress)	459	MFIAMADFKLLLLILILLSLFELDLLHFHHDFFSPFPVKIGLLLISIFFYAYSTTRSKPVYLVDFSCHQPTDSCKISSETFFNMAKGAQLYTEETIQFMTRILNRSGLGDDTYSPRCMLTSPPTPSMYEARHESELVIFGALNSLFKKTGIEPREVGIFIVNCSLFNPNPSLSSMIVNRYKLKTDVKTYNLSGISVDLATNLLKANPNTYAVIVSTENMTLSMYRGNDRSMLVPNCLFRVGGAAVMLSNRSQDRVRSKYELTHIVRTHKGSSDKHYTCAEQKEDSKGIVGVALSKELTVVAGDTLKTNLTALGPLVLPLSEKLRFILFLVKSKLFRLKVSPYVPDFKLCFKHFCIHAGGRALLDAVEKGLGLSEFDLEPSRMTLHRFGNTSSSSLWYELAYVEAKCRVKRGDRVWQLAFGSGFKCNSIVWRALRTIPANESLVGNPWGDSVHKYPVHVT	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9SUY9	KCS15_ARATH	reviewed	3-ketoacyl-CoA synthase 15 (KCS-15) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 15) (VLCFA condensing enzyme 15)	KCS15 At3g52160 F4F15.270	Arabidopsis thaliana (Mouse-ear cress)	451	MEKEATKMVNGGVKSKSPKGSPDFLGYNLRYVKLGYIYLLSLSRTFCFFLPPLLLLFIFVSRFLPILAFPLSTFFILLIYHYLTPSSVFLLDFSCYRPPDHLKITKSDFIELAMKSGNFNETAIELQRKVLDQSGIGEESYMPRVVFKPGHRVNLRDGREEAAMVIFGAIDELLAATKINVKHIKILVLNCGVLNTTPSLSAMVINHYKLRHNTESYNLGGMGCSAGVIAIDLAKDLLNAHQGSYALVVSTEIVSFTWYSGNDVALLPPNCFFRMGAAAVMLSSRRIDRWRAKYQLMQLVRTHKGMEDTSYKSIELREDRDGKQGLYVSRDVMEVGRHALKANIATLGRLEPSFEHICVLASSKKVLDDIHKDLKLTEENMEASRRTLERFGNTSSSSIWYELAYLEHKAKMKRGDRVWQIGFGSGFKCNSVVWKALKNIDPPRHNNPWNL	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9SYZ0	KCS16_ARATH	reviewed	3-ketoacyl-CoA synthase 16 (KCS-16) (EC 2.3.1.199) (Very long-chain fatty acid condensing enzyme 16) (VLCFA condensing enzyme 16)	KCS16 EL2 At4g34250 F10M10.20	Arabidopsis thaliana (Mouse-ear cress)	493	MDYPMKKVKIFFNYLMAHRFKLCFLPLMVAIAVEASRLSTQDLQNFYLYLQNNHTSLTMFFLYLALGSTLYLMTRPKPVYLVDFSCYLPPSHLKASTQRIMQHVRLVREAGAWKQESDYLMDFCEKILERSGLGQETYVPEGLQTLPLQQNLAVSRIETEEVIIGAVDNLFRNTGISPSDIGILVVNSSTFNPTPSLSSILVNKFKLRDNIKSLNLGGMGCSAGVIAIDAAKSLLQVHRNTYALVVSTENITQNLYMGNNKSMLVTNCLFRIGGAAILLSNRSIDRKRAKYELVHTVRVHTGADDRSYECATQEEDEDGIVGVSLSKNLPMVAARTLKINIATLGPLVLPISEKFHFFVRFVKKKFLNPKLKHYIPDFKLAFEHFCIHAGGRALIDEMEKNLHLTPLDVEASRMTLHRFGNTSSSSIWYELAYTEAKGRMTKGDRIWQIALGSGFKCNSSVWVALRNVKPSTNNPWEQCLHKYPVEIDIDLKE	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
Q9XF43	KCS6_ARATH	reviewed	3-ketoacyl-CoA synthase 6 (KCS-6) (EC 2.3.1.199) (Cuticular protein 1) (Eceriferum 6) (Very long-chain fatty acid condensing enzyme 6) (VLCFA condensing enzyme 6)	CUT1 CER6 EL6 KCS6 At1g68530 T26J14.10	Arabidopsis thaliana (Mouse-ear cress)	497	MPQAPMPEFSSSVKLKYVKLGYQYLVNHFLSFLLIPIMAIVAVELLRMGPEEILNVWNSLQFDLVQVLCSSFFVIFISTVYFMSKPRTIYLVDYSCYKPPVTCRVPFATFMEHSRLILKDKPKSVEFQMRILERSGLGEETCLPPAIHYIPPTPTMDAARSEAQMVIFEAMDDLFKKTGLKPKDVDILIVNCSLFSPTPSLSAMVINKYKLRSNIKSFNLSGMGCSAGLISVDLARDLLQVHPNSNAIIVSTEIITPNYYQGNERAMLLPNCLFRMGAAAIHMSNRRSDRWRAKYKLSHLVRTHRGADDKSFYCVYEQEDKEGHVGINLSKDLMAIAGEALKANITTIGPLVLPASEQLLFLTSLIGRKIFNPKWKPYIPDFKLAFEHFCIHAGGRAVIDELQKNLQLSGEHVEASRMTLHRFGNTSSSSLWYELSYIESKGRMRRGDRVWQIAFGSGFKCNSAVWKCNRTIKTPKDGPWSDCIDRYPVFIPEVVKL	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000269|PubMed:10330468}.
Q9ZUZ0	KCS13_ARATH	reviewed	3-ketoacyl-CoA synthase 13 (KCS-13) (EC 2.3.1.199) (Protein HIGH CARBON DIOXIDE) (Very long-chain fatty acid condensing enzyme 13) (VLCFA condensing enzyme 13)	HIC KCS13 At2g46720 T3A4.10	Arabidopsis thaliana (Mouse-ear cress)	466	MFIAMADFKILLLILILISLFELDLLHFHHDFFSPFPVKIGLLLISIFFYAYSTTRSKPVYLVDFSCHQPTDSCKISSETFFNMAKGAQLYTDETIQFMTRILNRSGLGDDTYSPRCMLTSPPTPSMYEARHESELVIFGALNSLFKKTGIEPREVGIFIVNCSLFNPNPSLSSMIVNRYKLKTDVKTYNLSGMGCSAGAISVDLATNLLKANPNTYAVIVSTENMTLSMYRGNDRSMLVPNCLFRVGGAAVMLSNRSQDRVRSKYELTHIVRTHKGSSDKHYTCAEQKEDSKGIVGVALSKELTVVAGDSLKTNLTALGPLVLPLSEKLRFILFLVKSKLFRLKVSPYVPDFKLCFKHFCIHAGGRALLDAVEKGLGLSEFDLEPSRMTLHRFGNTSSSSLWYELAYVEAKCRVKRGDRVWQLAFGSGFKCNSIVWRALRTIPANESLVGNPWGDSVHKYPVHVT	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
C4MBZ5	PKS3_ALOAR	reviewed	Aloesone synthase (EC 2.3.1.-) (Polyketide synthase 3)	PKS3	Aloe arborescens (Kidachi aloe)	403	MGSLSDSTPLMKDVQGIRKAQKADGTATVMAIGTAHPPHIISQDSYADFYFRVTNSEHKVELKKKFDRICKKTMIGKRYFNFDEEFLKKYPNITSFDKPSLNDRHDICIPGVPALGAEAAVKAIEEWGRPKSEITHLVFCTSGGVDMPSADFQCAKLLGLRTNVNKYCIYMQGCYAGGTVMRYAKDLAENNRGARVLMVCAELTIIALRGPNDSHIDNAIGNSLFGDGAAALIVGSDPIIGVEKPMFEIVCAKQTVIPNSEEVIHLHLRESGLMFYMTKDSAATISNNIEACLVDVFKSVGMTPPEDWNSLFWIPHPGGRAILDQVEAKLKLRPEKFSATRTVLWDYGNMISACVLYILDEMRRKSAAEGLETYGEGLEWGVLLGFGPGMTIETILLHSLPPV	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
C4NF90	PKS4_ALOAR	reviewed	Octaketide synthase 2 (OKS 2) (EC 2.3.1.-) (Polyketide synthase 4)	PKS4	Aloe arborescens (Kidachi aloe)	403	MGSLSNYSPVMEDVQAIRKAQKADGTATVMAIGTAHPPHIFPQDTYADFYFRATNSEHKVELKKKFDRICKKTMIGKRYFNYDEEFLKKYPNITSFDEPSLNDRQDICVPGVPALGAEAAVKAIAEWGRPKSEITHLVFCTSCGVDMPSADFQCAKLLGLRTNVNKYCVYMQGCYAGGTVMRYAKDLAENNRGARVLVVCAELTIIGLRGPNESHLDNAIGNSLFGDGAAALIVGSDPIIGVERPMFEIVCAKQTVIPNSEDVIHLHMREAGLMFYMSKDSPETISNNVEACLVDVFKSVGMTPPEDWNSLFWIPHPGGRAILDQVEARLKLRPEKFGATRTVLWDCGNMVSACVLYILDEMRRKSVADGLATYGEGLEWGVLLGFGPGMTVETILLHSLPPV	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
C4NF91	PKS5_ALOAR	reviewed	Octaketide synthase 3 (OKS 3) (EC 2.3.1.-) (Polyketide synthase 5)	PKS5	Aloe arborescens (Kidachi aloe)	405	MGSIAESSPLMSRENVEGIRKAQRAEGTATVMAIGTAHPPHIFPQDTYADFYFRATNSEHKVELKKKFDRICKKTMIGKRYFNYDEEFLKKYPNITSFDEPSLNDRQDICVPGVPALGKEAALKAIEEWGQPLSKITHLVFCTSCGVDMPSADFQLAKLLGLNTNVNKYCVYMQGCYAGGTVLRYAKDLAENNRGSRVLVVCAELTIIGLRGPNESHLDNAIGNSLFGDGAAALIVGADPIVGIEKPIFEIVCAKQTVIPDSEDVIHLHLREAGLMFYMSKDSPETISNNVEGCLVDIFKSVGMTPPADWNSLFWIPHPGGRAILDEVEARLKLRPEKFRATRHVLWEYGNMVSACVLYILDEMRNKSAADGLGTYGEGLEWGVLLGFGPGMTVETILLHSLPPV	CATALYTIC ACTIVITY: A very-long-chain acyl-CoA + malonyl-CoA = CoA + a very-long-chain 3-oxoacyl-CoA + CO(2). {ECO:0000305}.
A2ICC6	THS7_VITVI	reviewed	Stilbene synthase 6 (EC 2.3.1.95) (Resveratrol synthase 6) (Trihydroxystilbene synthase 6) (StSy 6)	STS GSVIVT00009216001 LOC100242994	Vitis vinifera (Grape)	392	MASVEEFRNAQRAKGPATILAIGTATPDHCVYQSDYADYYFRVTKSEHMTELKKKFNRICDKSMIKKRYIHLTEEMLEEHPNIGAYMAPSLNIRQEIITAEVPRLGRDAALKALKEWGQPKSKITHLVFCTTSGVEMPGADYKLANLLGLETSVRRVMLYHQGCYAGGTVLRTAKDLAENNAGARVLVVCSEITVVTFRGPSEDALDSLVGQALFGDGSSAVIVGSDPDVSIERPLFQLVSAAQTFIPNSAGAIAGNLREVGLTFHLWPNVPTLISENIEKCLTQAFDPLGISDWNSLFWIAHPGGPAILDAVEAKLNLEKKKLEATRHVLSEYGNMSSACVLFILDEMRKKSLKGENATTGEGLDWGVLFGFGPGLTIETVVLHSIPTVTN	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).
A5AEM3	THS4_VITVI	reviewed	Stilbene synthase 4 (EC 2.3.1.95) (Resveratrol synthase 4) (Trihydroxystilbene synthase 4) (StSy 4)	GSVIVT00005194001 LOC100241891 VITISV_031376	Vitis vinifera (Grape)	392	MASVEEIRNAQRAKGPATVLAIGTATPDNCLYQSDFADYYFRVTKSEHMTELKKKFNRICDKSMIKKRYIHLTEEMLEEHPNIGAYMAPSLNIRQEIITAEVPKLGKEAALKALKEWGQPKSKITHLVFCTTSGVEMPGADYKLANLLGLEPSVRRVMLYHQGCYAGGTVLRTAKDLAENNAGARVLVVCSEITVVTFRGPSEDALDSLVGQALFGDGSAAVIVGSDPDISIERPLFQLVSAAQTFIPNSAGAIAGNLREVGLTFHLWPNVPTLISENIENCLTKAFDPIGISDWNSLFWIAHPGGPAILDAVEAKVGLDKQKLKATRHILSEYGNMSSACVLFILDEMRKKSLKEGKTTTGEGLDWGVLFGFGPGLTIETVVLHSVGTDSN	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + trans-resveratrol + 4 CO(2).

True Positive Raw Dataset – RCHS

Entry	Entry_name	Status	Protein_names	Gene_names	Organism	Length	Pathway	Catalytic_activity	Amino_Acid_Sequence
P24825	CHS2_MAIZE	reviewed	Chalcone synthase C2 (EC 2.3.1.74) (Naringenin-chalcone synthase C2)	C2	Zea mays (Maize)	400	Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAGATVTVEEVRKAQRATGPATVLAIGTATPANCVYQADYPDYYFRITKSEHLTDLKEKFKRMCDKSMIRKRYMHLTEEFLAENPSMCAYMAPSLDARQDVVVVEVPKLGKAAAQKAIKEWGQPKSRITHLVFCTTSGVDMPGADYQLTKALGLRPSVNRLMMYQQGCFAGGTVLRVAKDLAENNRGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAAAVVVGADPDDRVERPLFQLVSAAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIGRALDDAFKPLGISDWNSIFWVAHPGGPAILDQVEAKVGLDKARMRATRHVLSEYGNMSSACVLFILDEMRKRSAEDGQATTGEGLDWGVLFGFGPGLTVETVVLHSVPITTGAATA
P30074	CHS2_MEDSA	reviewed	Chalcone synthase 2 (EC 2.3.1.74) (Naringenin-chalcone synthase 2)	CHS2	Medicago sativa (Alfalfa)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVSEIRKAQRAEGPATILAIGTANPANCVEQSTYPDFYFKITNSEHKTELKEKFQRMCDKSMIKRRYMYLTEEILKENPNVCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIVCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPVPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNITKALVEAFEPLGISDYNSIFWIAHPGGPAILDQVEQKLALKPEKMNATREVLSEYGNMSSACVLFILDEMRKKSTQNGLKTTGEGLEWGVLFGFGPGLTIETVVLRSVAI
P13114	CHSY_ARATH	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase) (Protein TRANSPARENT TESTA 4)	CHS TT4 At5g13930 MAC12.11 MAC12.14	Arabidopsis thaliana (Mouse-ear cress)	395	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVMAGASSLDEIRQAQRADGPAGILAIGTANPENHVLQAEYPDYYFRITNSEHMTDLKEKFKRMCDKSTIRKRHMHLTEEFLKENPHMCAYMAPSLDTRQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRIAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDTSVGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIVKSLDEAFKPLGISDWNSLFWIAHPGGPAILDQVEIKLGLKEEKMRATRHVLSEYGNMSSACVLFILDEMRRKSAKDGVATTGEGLEWGVLFGFGPGLTVETVVLHSVPL
Q9FUB7	CHSY_HYPAN	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Hypericum androsaemum (Tutsan)	390	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023, ECO:0000269|PubMed:12795704}.	MVTVEEVRKAQRAEGPATVMAIGTAVPPNCVDQATYPDYYFRITNSEHKAELKEKFQRMCDKSQIKKRYMYLNEEVLKENPNMCAYMAPSLDARQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPTDTHLDSLVGQALFGDGAAAIIIGSDPIPEVEKPLFELVSAAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNVEKSLTEAFKPLGISDWNSLFWIAHPGGPAILDQVEAKLSLKPEKLRATRHVLSEYGNMSSACVLFILDEMRRKSKEDGLKTTGEGIEWGVLFGFGPGLTVETVVLHSVAIN
Q9AU11	PKS1_RUBID	reviewed	Polyketide synthase 1 (RiPKS1) (EC 2.3.1.74) (Naringenin-chalcone synthase PKS1)	PKS1	Rubus idaeus (Raspberry)	391	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023, ECO:0000269|PubMed:11437245}.	MVTVDEVRKAQRAEGPATILAIGTATPPNCVDQSTYPDYYFRITKSEHKTELKEKFQRMCDKSMIKKRYMYLTEEILKENPSMCEYMAPSLDARQDMVVVEIPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAIIVGSDPLPDIERPLFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLNEAFKPLDITDWNSLFWIAHPGGPAILDQVEAKLGLKPEKLEATRNILSEYGNMSSACVLFILDEVRRKSVANGHKTTGEGLEWGVLFGFGPGLTVETVVLHSVAAST
Q8RVK9	CHS_CANSA	reviewed	Naringenin-chalcone synthase (EC 2.3.1.74)	CHS CAN1069	Cannabis sativa (Hemp) (Marijuana)	389		CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023, ECO:0000269|PubMed:15120113, ECO:0000269|PubMed:19581347}.	MVTVEEFRKAQRAEGPATIMAIGTATPANCVLQSEYPDYYFRITNSEHKTELKEKFKRMCDKSMIRKRYMHLTEEILKENPNLCAYEAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPNDTHLDSLVGQALFGDGSAALIVGSDPIPEVEKPIFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLNEAFKPLGISDWNSLFWIAHPGGPAILDQVESKLALKTEKLRATRHVLSEYGNMSSACVLFILDEMRRKCVEDGLNTTGEGLEWGVLFGFGPGLTVETVVLHSVAI
P13417	CHS3_SINAL	reviewed	Chalcone synthase 3 (EC 2.3.1.74) (Naringenin-chalcone synthase 3)	CHS3	Sinapis alba (White mustard) (Brassica hirta)	395	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVMGTPSSLDEIRKAQRADGPAGILAIGTANPANHVIQAEYPDYYFRITNSEHMTDLKEKFKRMCDKSTIRKRHMHLTEEFLKDNPNMCAYMAPSLDARQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAAIIVGSDPDTSVGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLDEAFKPLGISDWNSLFWIAHPGGPAILDDVEKKLGLKAEKMRATRHVLSEYGNMSSACVLFILDEMRRKSKEDGVATTGEGLEWGVLFGFGPGLTVETVVLHSVPV
A2ZEX7	CHS1_ORYSI	reviewed	Chalcone synthase 1 (OsCHS1) (EC 2.3.1.74) (Naregenin-chalcone synthase)	CHS1 CHS OsI_035120	Oryza sativa subsp. indica (Rice)	398	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAAAVTVEEVRRAQRAEGPATVLAIGTATPANCVYQADYPDYYFRITKSEHMVELKEKFKRMCDKSQIRKRYMHLTEEILQENPNMCAYMAPSLDARQDIVVVEVPKLGKAAAQKAIKEWGQPRSRITHLVFCTTSGVDMPGADYQLAKMLGLRPNVSRLMMYQQGCFAGGTVLRVAKDLAENNRGARVLAVCSEITAVTFRGPSESHLDSMVGQALFGDGAAAVIVGSDPDEAVERPLFQMVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIERALGDAFTPLGISDWNSIFWVAHPGGPAILDQVEAKVGLDKERMRATRHVLSEYGNMSSACVLFILDEMRKRSAEDGHATTGEGMDWGVLFGFGPGLTVETVVLHSVPITAGAAA
Q2R3A1	CHS1_ORYSJ	reviewed	Chalcone synthase 1 (OsCHS1) (EC 2.3.1.74) (Naregenin-chalcone synthase)	CHS1 CHS Os11g0530600 LOC_Os11g32650 OsJ_032788	Oryza sativa subsp. japonica (Rice)	398	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAAAVTVEEVRRAQRAEGPATVLAIGTATPANCVYQADYPDYYFRITKSEHMVELKEKFKRMCDKSQIRKRYMHLTEEILQENPNMCAYMAPSLDARQDIVVVEVPKLGKAAAQKAIKEWGQPRSRITHLVFCTTSGVDMPGADYQLAKMLGLRPNVNRLMMYQQGCFAGGTVLRVAKDLAENNRGARVLAVCSEITAVTFRGPSESHLDSMVGQALFGDGAAAVIVGSDPDEAVERPLFQMVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIERALGDAFTPLGISDWNSIFWVAHPGGPAILDQVEAKVGLDKERMRATRHVLSEYGNMSSACVLFILDEMRKRSAEDGHATTGEGMDWGVLFGFGPGLTVETVVLHSVPITAGAAA
B0LDU6	PKS5_RUBID	reviewed	Polyketide synthase 5 (RiPKS5) (EC 2.3.1.74) (Naringenin-chalcone synthase PKS5)	PKS5	Rubus idaeus (Raspberry)	391	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023, ECO:0000269|PubMed:18068110}.	MVTVDEVRKAQRAEGPATVLAIGTATPPNCIDQSTYPDYYFRITNSEHKTELKEKFQRMCDKSMIKKRYMYLTEEILKENPSMCEYMAPSLDARQDMVVVEIPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEIXAVTFRGPSDTHLDSLVGQALFGDGAAAIIVGADPLPKIERPLFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLNEAFKPLDITDWNSLFWIAHPGGPAILDQVETKLGLKPEKLEATRHILSEYGNMSSACVLFILDEVRRKSATNGLKTTGEGLEWGVLFGFGPGLTVETVVLHSVGVTA
P30075	CHS4_MEDSA	reviewed	Chalcone synthase 4 (EC 2.3.1.74) (CHS12-1) (Naringenin-chalcone synthase 4)	CHS4	Medicago sativa (Alfalfa)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVSEIRKAQRAEGPATILAIGTANPANCVEQSTYPDFYFKITNSEHKTELKEKFQRMCDKSMIKRRYMYLTEEILKENPSVCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIVCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPVPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNIDKALVEAFQPLGISDYNSIFWIAHPGGPAILDQVEQKLALKPEKMRATREVLSEYGNMSSACVLFILDEMRKKSTQDGLKTTGEGLEWGVLFGFGPGLTIETVVLRSVAI
P51078	CHS5_MEDSA	reviewed	Chalcone synthase 4-2 (EC 2.3.1.74) (Naringenin-chalcone synthase 4-2)	CHS4-2 CHSI	Medicago sativa (Alfalfa)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVSEIRNAQRAEGPATTLAIGTANPTNCVEQSTYPDFYFKITNSEHKTELKEKFQRMCDKSMIKRRYMYLTEEILKENPSVCEIMAPSLDAWQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIVCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPVPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNINKALVEAFEPLGISDYNSIFWIAHPGGPAILDQVEQKLALKPEKMKATREVLSEYGNMSSACVLFILDEMRKKSAQDGLKTTGEGLEWGVLFGFGPGLTIETVVLRSVAI
P17818	CHSY_MATIN	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Matthiola incana (Common stock) (Cheiranthus incanus)	394	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVMGATSLDEIRKAQRADGPAGILGIGTANPANHVIQAEYPDYYFRITNSEHMTDLKEKFQRMCDKSMIRKRHMHLTEDFLKENPNMCAYMAPSLDARQDIVVVEVPKLGKEAAVRAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDTSVGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLEEAFKPLGISDWNSLFWIAHPGGPAILDQVEIKLGLKAEKMRATRHVLSEYGNMSSACVLFILDEMRKKSAQDGVATTGEGLEWGVLFGFGPGLTVETVVLRSVPL
P13416	CHS1_SINAL	reviewed	Chalcone synthase 1 (EC 2.3.1.74) (Naringenin-chalcone synthase 1)	CHS1	Sinapis alba (White mustard) (Brassica hirta)	395	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVMGTPSSLDEIRKAQRADGPAGILAIGTANPANHVIQAEYPDYYFRITNSEHMTDLKEKFKRMCDKSTIRKRHMHLTEEFLKDNPNMCAYMAPSLDARQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDADISAGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLDEAFKPLGISDWNSLFWIAHPGGPAILDDVEKKLGLKAEKMRATRHVLSEYGNMSSACVLFILDEMRRKSVEDGVATTGEGLEWGVLFGFGPGLTVETVVLHSVPV
Q9SEP2	CHSY_CARAN	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Cardamine amara (Large bitter-cress)	395	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVMGDTPSLDEIRKAQRADGPAGILAIGTANPANYVIQAEYPDYYFRITNSEHMTDLKEKFKRMCDKSTIRKRHMHVTEEFLKENPNMCAYMAPSLDARQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDTSVGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLDEAFKPLGISDWNSLFWIAHPGGPAILDQVEIKLGLKEEKMRATRHVMREYGNMSSACVLFILDEMRKKSAKDGVATTGEGLEWGVLFGFGPGLTVETVVLHSVPL
Q9AU09	PKS3_RUBID	reviewed	Polyketide synthase 3 (RiPKS3) (EC 2.3.1.74) (Naringenin-chalcone synthase PKS3)	PKS3	Rubus idaeus (Raspberry)	391	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023, ECO:0000269|PubMed:11437245}.	MVTVDEVRKAQRAEGPATILAIGTATPPNCVDQSTYPDYYFRITKSEHRTELKEKFQRMCDKSRIKKRYMYLTEEILKENPSMCEYMAPSLDARQDMVVVEIPKLGKEAATKAIKEWGQLKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLAVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAIIVGSDPLPDIERPLFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLNEAFKPLDITDWNSLFWIAHPGGPAILDQVEAKLGLKPEKLEATRNILSEYGNMSSACVLFILDEVRRKSVANGHKTTGEGLEWGVLFGFGPGLTVETVVLHSVAAST
Q8H4L3	CHS2_ORYSJ	reviewed	Chalcone synthase 2 (OsCHS2) (EC 2.3.1.74) (Naregenin-chalcone synthase)	CHS2 Os07g0214900 LOC_Os07g11440 OJ1116_C08.125	Oryza sativa subsp. japonica (Rice)	403	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTSTVKLEEVRRMQRAEGMAAVLAIGTATPANCVYQTDYPDYYFRVTNSEHLTNLKERFQRMCESSQIRKRYTHLTEEILQENPSMCVFTAPSLDARQDMVVAEVPKLGKAAAEEAIKEWGQPMSRITHLVFCTTNGVDMPGADYQVAKMLGLPTSVKRLMMYQQGCFAGGTVLRVAKDLAENNRGARVLVVCSEIMAMAFRGPSESHLDSLVGHALFGDGAAAVIVGSDPDEAADERPLFQIVSASQTILPGTEDAIVGHLREVGLTFHLPKDVPEFISDSVEGALTDAFMPLGVHDWNSIFWVVHPGGPAILDQVEEKVALHKARMRASRNVLSEYGNMASATVLFVLDEMRKLSADDGHATTGEGMDWGVLFGFGPGLTVETIVLHSVPITAAAPLIMQ
P51090	CHSY_VITVI	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Vitis vinifera (Grape)	393	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVAEIRKAQRAEGPATVLAIGTATPANCVYQADYPDYYFRITNSEHMTELKEKFKRMCEKSMINKRYMHLTEEILKENPNVCAYMAPSLDARQDMVVVEVPKLGKEAAAKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLKPSVKRLMMYQQGCFAGGTVLRLAKDLAENNAGSRVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIIGADPDTKIELPLFELVSAAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFTPIGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKLRATRHVLSEYGNMSSACVLFILDEMRKKSIEEGKGSTGEGLEWGVLFGFGPGLTVETVVLHSVSAPAAH
Q96562	CHS2_HORVU	reviewed	Chalcone synthase 2 (EC 2.3.1.74) (Naringenin-chalcone synthase 2)	CHS2	Hordeum vulgare (Barley)	399	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAAVRLKEVRMAQRAEGLATVLAIGTAVPANCVYQATYPDYYFRVTKSEHLADLKEKFQRMCDKSMIRKRHMHLTEEILIKNPKICAHMETSLDARHAIALVEVPKLGQGAAEKAIKEWGQPLSKITHLVFCTTSGVDMPGADYQLTKLLGLSPTVKRLMMYQQGCFGGATVLRLAKDIAENNRGARVLVVCSEITAMAFRGPCKSHLDSLVGHALFGDGAAAAIIGADPDQLDEQPVFQLVSASQTILPESEGAIDGHLTEAGLTIHLLKDVPGLISENIEQALEDAFEPLGIHNWNSIFWIAHPGGPAILDRVEDRVGLDKKRMRASREVLSEYGNMSSASVLFVLDVMRKSSAKDGLATTGEGKDWGVLFGFGPGLTVETLVLHSVPVPVPTAASA
P51077	CHS3_MEDSA	reviewed	Chalcone synthase 4-1 (EC 2.3.1.74)	CHS4-1	Medicago sativa	389	PATHWAY: Secondary metabolite	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin	MVSVSEIRQAQRAEGPATIMAIGTANPSNCVEQSTYPDFYFKITNSEHKVELKEKFQRMCDKSMIKRRYMYLTEEILKENPSVCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPIPEIEKPIFEMV

P51075	CHSY_BETPN	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Betula pendula (European white birch) (Betula verrucosa)	395	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MASVEEIRKAQRAHGPATVLAIGTATPSNCITQADYPDYYFRITKSDHMTELKEKFKRMCDKSMIKKRYMYLNEEILNENPNMCAYMAPSLDARQTIVVVEVPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPTDTHLDSLVGQALFGDGAAAVIVGADPDTSVERPLFELISAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGIISKNIEKSLAEAFAPLGISDWNSLFWIAHPGGPAILDQVESKLGLKEEKLRATRHVLSEYGNMSSACVLFILDEMRRNSLEGGKVTTGEGLEWGVLFGFGPGLTVETVVLHSVPVPVEASH
P49440	CHSY_PHAVU	reviewed	Chalcone synthase 17 (EC 2.3.1.74) (Naringenin-chalcone synthase 17)	CHS17	Phaseolus vulgaris (Kidney bean) (French bean)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVSEIRQAQRAEGPATILAIGTATPSNCVDQSTYPDYYFRITNSEHMTDLKEKFQRMCDKSMIKKRYMHLNEEILKENPNMCAYMAPSLDARQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIVGSDPIPQIEKPLFELVWTAQTIAPDSDGAIDGHLREVGLTFHLLKDVPGIVSKNIGKALFEAFNPLNISDYNSIFWIAHPGGPAILDQVEQKLGLKPEKMKATRDVLSDYGNMSGACVLFILDEMRRKSAEKGLKTTGEGLEWGVLFGFGPGLTIETVVLHSVAI
P22924	CHSB_PETHY	reviewed	Chalcone synthase B (EC 2.3.1.74) (Naringenin-chalcone synthase B)	CHSB	Petunia hybrida (Petunia)	392	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MKVENGQLQGWWAQRAEGPAKILAIGTATPFHWVDQNSYPDYYFRVTNSQHLVDLKEKFRRICSKTMIKKRHMFLTEELLRKNPTLCSHNEPSLDIRQDILVSEIPKLGKEAALKAIGEWGQPKSTITHLVFCTRSGVDMPGADCQLVKLLGLSPSVQRLMMYQQGCFAGGTMLRLAKDLAENNKGARVLVVCAESSAIGFRGPSEANVDNLIAQALFGDGAAALIIGSDPKPGLERPVFEIFSAAQTFVPNGDCHLALHLREMGLTFHCTKDVPPTIAKNVESCLIKAFEPLGISDWNSLFWILHPGGNAIVDQVESTLGLGPEKLRATRNILSEYGNLSSACCLFILDEIRKKSAREGMRTSGDGLDLGVLLSFGPGLTIETVVLRSVPI
P51088	CHS6_TRISU	reviewed	Chalcone synthase 6 (EC 2.3.1.74) (Naringenin-chalcone synthase 6)	CHS6	Trifolium subterraneum (Subterranean clover)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVAEIRKAQRAEGPATILAIGTANPANKVEQATYPDFYFKITNSEHKTELKEKFQRMCDKSMIKSRYMYLTEEILKENPSLCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPVPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNIDKALVEAFQPLNISDYNSIFWIAHPGGPAILDQVEQKLALKPEKMKATRDVLSEYGNMSSACVLFILDEMRKKSAQNGLKTTGEGLDWGVLFGFGPGLTIETVVLHSVAI
P08894	CHSA_PETHY	reviewed	Chalcone synthase A (EC 2.3.1.74) (Naringenin-chalcone synthase A)	CHSA	Petunia hybrida (Petunia)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEYRKAQRAEGPATVMAIGTATPTNCVDQSTYPDYYFRITNSEHKTDLKEKFKRMCEKSMIKKRYMHLTEEILKENPSMCEYMAPSLDARQDIVVVEVPKLGKEAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGCDYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPNDTHLDSLVGQALFGDGAGAIIIGSDPIPGVERPLFELVSAAQTLLPDSHGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLEEAFKPLGISDWNSLFWIAHPGGPAILDQVEIKLGLKPEKLKATRNVLSDYGNMSSACVLFILDEMRKASAKEGLGTTGEGLEWGVLFGFGPGLTVETVVLHSVAT
P22926	CHSF_PETHY	reviewed	Chalcone synthase F (EC 2.3.1.74) (Naringenin-chalcone synthase F)	CHSF	Petunia hybrida (Petunia)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVEEVRRAQRREGPATILAIGTATPLNCVDQTTYPDYFFRVTNSDHKTELKEKFKRMFERSMIKKSYLHLTEEILKENPSICEHKAPSFDARQDIVVVEVPKLGKEAAQNAIKEWGQPKSKITHLVFCTTTGVDMHGADYQLTKLLGLSPSVKRLMMYQLGCYGGGTVLRLAKDLAENNKGARALVVCSEITAITFHAPSDTDLDVLVGQALFGDGAASVIIGSDPNLEVEKPLFELVSAAQTLVPDCGHKIYGKTSDVGLTFHLHKDVPRLVSQNIEKSLVEVFQPLGIFDWNSIFWVGHAGGREILDQVELKLGLKPEKLNVTRHVMSEYGNMASACVLFVLDEMRKTSTKEGFGTNVEGLEWGVLCSFGPGLTIETIVLHSVSI
P22927	CHSG_PETHY	reviewed	Chalcone synthase G (EC 2.3.1.74) (Naringenin-chalcone synthase G)	CHSG	Petunia hybrida (Petunia)	393	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MATVEEIRKAQRAEGPATVLAIGTANPSNCVDQSAYPDFLFRITTSDHKTELKEKFKHMCEGSMIKKRYLHLTEEILKNNPNICEHKAPSLNARQEIAVAEAPKLGKRAAQKAIEEWSQSKSKITHLVFCTTTSVELPGADYQLTKLLGLSPSVKRSMMYQQGCYGGGTALRLAKDLAENNKGARVLVVCVEITVMSFQAPSRNDTDELDVLVGQALFADGASAVIIGSDPILAIEKPLFELVFATQTLIPDSGHVICANLTEAGLIPHLLKDAPIVISQNIERRLVEVFKPLGISDWNSIFWVAHPGGPAILNQIELKLGLKPEKLRAARHVLSEYGNMSSACVLFVLDEMRKGTIEKGMGTTGEGLEWGLLFGFGPGLTIETVVLHSVSIN
P51085	CHS3_TRISU	reviewed	Chalcone synthase 3 (EC 2.3.1.74) (Naringenin-chalcone synthase 3)	CHS3	Trifolium subterraneum (Subterranean clover)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVSEIRQAQRAEGPATILAIGTANPANKVEQATYPDFYFKITNSEHKVELKEKFQRMCDKSMIKSRYMYLTEEILKENPSLCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPVPEIEKPIFVMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNIDKALVEAFQPLGISDYNSIFWIAHPGGPAILDQVEQKLALKPEKMKATREVLSEYGNMSSACVLFILDEMRKKSTKDGLKTTGEGLDWGVLFGFGPGLTIETVVLHSVAI
P51079	CHS6_MEDSA	reviewed	Chalcone synthase 6-4 (EC 2.3.1.74) (Naringenin-chalcone synthase 6-4) (Fragment)	CHS6-4	Medicago sativa (Alfalfa)	285	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	LGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPIPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNIDKALVEAFQPLNISDYNSIFWIAHPGGPAILDQVEQKLGLKPEKMKATREVLSEYGNMSSACVLFILDEMRKKSAQQGLKTTGEGLDWGVLFGFGPGLTIETVVLHSVAL
P51080	CHS7_MEDSA	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase) (Fragment)	CHSII	Medicago sativa (Alfalfa)	265	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	ITHLIVCTTSVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPVPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNINKALVEAFEPLGISDYNSIFWIAHPGGPAILDQVEQKLALKPEKMKATREVLSEYGNMSSACVLFILDEMRKKSAQDGLKTTGEGLEWGVLFGFGPGLTIETVVLRSVTI
P51087	CHS5_TRISU	reviewed	Chalcone synthase 5 (EC 2.3.1.74) (Naringenin-chalcone synthase 5)	CHS5	Trifolium subterraneum (Subterranean clover)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVAEIRQAQRAEGPATILAIGTANPANKVEQATYPDFYFKITNSEHKVELKEKFQRMCDKSMIKSRYMYLTEEILKENPSVCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPVPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNIDKALVEAFQPLNISDYNSIFWIAHPGGPAILDQVEQKLALKPEKMKATRDVLSEYGNMSSACVLFILDEMRKKSAQNGLKTTGEGLDWGVLFGFGPGLTIETVVLHSVAI
P51086	CHS4_TRISU	reviewed	Chalcone synthase 4 (EC 2.3.1.74) (Naringenin-chalcone synthase 4) (Fragment)	CHS4	Trifolium subterraneum (Subterranean clover)	311	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVSEIRKAQRAEGPATILAIGTANPANKVEQATYPDFYFKITNSEHKVELKEKFQRMCDKSMIKSRYMYLTEEILKENPSVCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPVPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNIDKALVEAFQPLGISDYNSIFWIAHPGGPAILD
P24824	CHS1_MAIZE	reviewed	Chalcone synthase WHP1 (EC 2.3.1.74) (Naringenin-chalcone synthase WHP1) (White pollen)	WHP1	Zea mays (Maize)	400	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAGATVTVDEVRKGQRATGPATVLAIGTATPANCVYQADYPDYYFRITKSDHLTDLKEKFKRMCDKSMIRKRYMHLTEEFLSENPSMCAYMAPSLDARQDVVVTEVPKLGKAAAQEAIKEWGQPKSRITHLVFCTTSGVDMPGADYQLTKALGLRVVNRLMMYQQGCFAGGTVLRVAKDVAENNRGARVMVVCSEITAVTFRGPSESHVDSLVGQALFGDGAAAGRGGADPDGRVERPLFQLVSAAQTILPDSEGAIDGHLREVGLAFHLLKDVPGLISKNIERALEDAFEPLGISDWNSIFWVAHPGGPAILDQVEAKVGLDKARMRATRHVLSEYGNMSSACVLFILDEMRKRPAEDGQSTTGEGLDWGVLFGFGPGLTVETVVLHSVPITTGAPTAA
P26018	CHS1_HORVU	reviewed	Chalcone synthase 1 (EC 2.3.1.74) (Naringenin-chalcone synthase 1)	CHS1 CHS	Hordeum vulgare (Barley)	398	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAATMTVEEVRNAQRAEGPATVLAIGTATPANCVYQADYPDYYFKITKSDHMADLKEKFKRMCDKSQIRKRYMHLTEEILEENPNMCAYMAPSLDARQDIVVVEVPKLGKAAAQKAIKEWGQPRSKITHLVFCTTSGVDMPGADYQLTKMLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPHESHLDSLVGQALFGDGAAAVIIGADPDLSVERPLFQLVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIERALEEAFKPLGIDHWNSVFWIAHQGGPAILDMVEAKVNLNKERMRATRHVLSEYGNMSSACVLFIMDEMRKRSAEDGHATTGEGMDWGVLFGFGPGLTVETVVLHSVPISAGATA
Q9ZS41	CHS1_DAUCA	reviewed	Chalcone synthase 1 (EC 2.3.1.74) (DcCHS1) (Naringenin-chalcone synthase 1)	CHS1	Daucus carota (Wild carrot)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVNEFRKAQRAEGPATVLAIGTATPPNCVDQSAYADYYFRITNSEDKPELKEKFRRMCEKSMINTRYMHLTEDLLKQNPSFCEYMASSLDARQDIVVNEVPKLGKEAALRAIKEWGQPKSKITHLIFCTTSGVDMPGADFRLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNKNARVLVVCSEITVITFRGPNDTHLDSLVGQALFGDGAGAVIVGSDPVIGIEKPLFEIVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIRKSLVEAFKPLGISDWNSIFWIAHPGGPAILDQVETELSLKPEKLKSTRQVLRDYGNMSSACVLFILDEMRKASAKDGHRTTGEGLDWGVLFGFGPGLTVETVVLHSVPP
Q9ZS40	CHS2_DAUCA	reviewed	Chalcone synthase 2 (EC 2.3.1.74) (DcCHS2) (Naringenin-chalcone synthase 2)	CHS2	Daucus carota (Wild carrot)	397	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MANHNAEIEEIRKRQRAQGPANILAIGTATPSNCVYQADYPDYYFRITNSEHMSDLKLKFKRMCEKSMIRKRYMHITEEYLKENPNVCAYEAPSLDARQDLVVVEVPRLGKEAAAKAIKEWGHPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNTGARVLVVCSEITAVTFRGPSDSHLDSLVGQALFGDGAAAVIVGSDPDLSVERPLFQLISAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLKEAFGPIGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKMRATRQVLSDYGNMSSACVLFILDEMRKKSIEEGKATTGDGLDWGVLFGFGPGLTVETVVLHSVPATITH
P30081	CHS7_SOYBN	reviewed	Chalcone synthase 7 (EC 2.3.1.74) (Naringenin-chalcone synthase 7)	CHS7	Glycine max (Soybean) (Glycine hispida)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVAEIRQAQRAEGPATILAIGTANPPNRVDQSTYPDYYFRITNSDHMTELKEKFQRMCDKSMIKTRYMYLNEEILKENPNMCAYMAPSLDARQDMVVVEVPKLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKQLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIVGSDPIPQVEKPLYELVWTAQTIAPDSEGAIDGHLREVGLTFHLLKDVPGIVSKNIDKALFEAFNPLNISDYNSIFWIAHPGGPAILDQVEQKLGLKPEKMKATRDVLSEYGNMSSACVLFILDEMRRKSAENGHKTTGEGLEWGVLFGFGPGLTIETVVLHSVAI
P30080	CHS6_SOYBN	reviewed	Chalcone synthase 6 (EC 2.3.1.74) (Naringenin-chalcone synthase 6)	CHS6	Glycine max (Soybean) (Glycine hispida)	388	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVEEIRKAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSDHMNELKEKFKRMCDKSMIKKRYMYLNEEILKENPSVCAYMEPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRYMMYQQGCFAGGTVLRLAKDLAENNTGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIVGSDPLPAEKPLFELVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIQKALVEAFQPLGIDDYNSIFWIAHPGGPAILDQVEAKLGLKPEKMEATRHVLSEYGNMSSACVLFILDQMRKKSIENGLGTTGEGLEWGVLFGFGPGLTVETVVLRSVTV
P24826	CHS1_SOYBN	reviewed	Chalcone synthase 1 (EC 2.3.1.74) (Naringenin-chalcone synthase 1)	CHS1	Glycine max (Soybean) (Glycine hispida)	388	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVEEIRKAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSEHMTELKEKFKRMCDKSMIKKRYMYLNEEILKENPSVCAYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPTDTHLDSLVGQALFGDGAAAVIVGSDPLPVEKPLFQLVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKALVEAFQPLGISDYNSIFWIAHPGGPAILDQVEAKLGLKPEKMEATRHVLSEYGNMSSACVLFILDQMRKKSIENGLGTTGEGLDWGVLFGFGPGLTVETVVLRSVTL
P48406	CHS5_SOYBN	reviewed	Chalcone synthase 5 (EC 2.3.1.74) (Naringenin-chalcone synthase 5)	CHS5	Glycine max (Soybean) (Glycine hispida)	388	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVEEIRQAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSEHMTELKEKFKRMCDKSMIKKRYMYLNEEILKENPSVCAYMAPSLDARQDMVVMEVPKLGKEAATKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPTDTHLDSLVGQALFGDGAAAVIVGSDPLPVEKPLFQLVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKALVEAFQPLGISDYNSIFWIAHPGGPAILDQVEAKLGLKPEKMEATRHVLSEYGNMSSACVLFILDQMRKKSIENGLGTTGEGLDWGVLFGFGPGLTVETVVLRSVTV
Q43188	CHS2_SOLTU	reviewed	Chalcone synthase 2 (EC 2.3.1.74) (Naringenin-chalcone synthase 2)	CHS2	Solanum tuberosum (Potato)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEVRKAQRAKGPATIMAIGTATPSNCVDQSTYPDYYFRITNSEHMTELKEKFKRMCDKSMINKRYMHLTEEILKENPNICEYMAPSLDARQDIVVVEVPKLGKEAAQKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVIRLAKDLAENNKGARVLVVCSEITAVTFSGPSDTHLDSMVGQALFGDGAAAMIIGSDPLPEVERPLFELVSAAQTLLPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLIEAFQPLGISDWNSIFWIAHPGGPAILDQVELKLGLKPEKLQATRQVLSDYGNMSSACVLFILDEMRKASSKEGLSTTGEGLDWGVLFGFGPGLTVETVVLHSVST
Q9SBL3	CHS6_SORBI	reviewed	Chalcone synthase 6 (EC 2.3.1.74) (Naringenin-chalcone synthase 6)	CHS6	Sorghum bicolor (Sorghum) (Sorghum vulgare)	401	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAGATVTVEEVRKAQRATGPATVLAIGTATPANCVHQADYPDYYFRITKSEHMTELKEKFKRMCDKSQIRKRYMHLTEEYLAENPNMCAYMAPSLDARQDIVVVEVPKLGKAAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKMLGLRPSVNRLMMYQQGCFAGGTVLRVAKDLAENNRGARVLVVCSEITAVTFRGPSESHLDSMVGQALFGDGAAAVIVGADPDERVERPLFQLVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIERSLEEAFKPLGITDYNSIFWVAHPGGPAILDQVEAKVGLKKERMRATRHVLSEYGNMSSACVLFILDEMRKRSAEDGQATTGEGFDWGVLFGFGPGLTVETVVLHSVPITTGATITA

Q9XGX2	CHS1_SORBI	reviewed	Chalcone synthase 1 (EC 2.3.1.74) (Naringenin-chalcone synthase 1)	CHS1	Sorghum bicolor (Sorghum) (Sorghum vulgare)	401	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAGATVTVEEVRKAQRATGPATVLAIGTATPANCVHQADYPDYYFRITKSEHMTELKEKFKRMCDKSQIRKRYMHLTEEYLAENPNMCAYMAPSLDARQDIVVVEVPKLGKAAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKMLGLRPSVNRLMMYQQGCFAGGTVLRVAKDLAENNRGARVLVVCSEITAVTFRGPSESHLDSMVGQALFGDGAAAVIVGADPDERVERPLFQLVSASQRILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIERALEEAFKPLGITDYNSIFWVAHPGGPAILDQVEAKVGLEKERMRATRHVLSEYGNMSSACVLFILDEMRKRSAEDGQTTTGEGFDWGVLFGFGPGLTVETVVLHSVPITTGAAITA
Q9SBL5	CHS4_SORBI	reviewed	Chalcone synthase 4 (EC 2.3.1.74) (Naringenin-chalcone synthase 4)	CHS4	Sorghum bicolor (Sorghum) (Sorghum vulgare)	401	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAAATVTVEEVRKAQRATGPATVLAIGTATPANCVHQADYPDYYFRITKSEHMTELKEKFKRMCDKSQIRKRYMHLTEEYLAENPNMCAYMAPSLDARQDIVVVEVPKLGKAAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKMLGLRPSVKRLMMYQQGCFAGGTVLRVAKDLAENNRGARVLVVCSEITAVTFRGPSESHLDSMVGQALFGDGAAAVIVGADPDERVERPLFQLVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIERSLEEAFKPLGITDYNSIFWVAHPGGPAILDQVEAKVGLKKERMRATRHVLSEYGNMSSACVLFILDEMRKRSAEDGQATTGEGLDWGVLFGFGPGLTVETVVLHSVPITTGAAITA
Q9XGX1	CHS7_SORBI	reviewed	Chalcone synthase 7 (EC 2.3.1.74) (Naringenin-chalcone synthase 7)	CHS7	Sorghum bicolor (Sorghum) (Sorghum vulgare)	400	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAGATVTVEEVRKAQRATGPATVLAIGTATPANCVHQADYPDYYFRITKSEHMTDLKEKFKRMCDKSQIRKRYMHLTEEYLAENPNMCAYMAPSLDARQDIVVVEVPKLGKAAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKMLGLRPSVNRLMMYQQGCFAGGTVLRVAKDLAENNRGARVLVVCSEITAVTFRGPSESHLDSMVGQALFGDGAAAVIVGADPDKRVECPLFQLVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIERSLEEAFKPLGITDYNSIFWVAHPGGPAILDQVEAKVGLKKERMRATRHVLSEYGNMSSACVLFILDEMRKRSAEEGQATTGEGFDWGVLFGFGPGLTVETVVLHSVPITIAAITA
Q43163	CHSB_SOLTU	reviewed	Chalcone synthase 1B (EC 2.3.1.74) (Naringenin-chalcone synthase 1B)	CHS1B	Solanum tuberosum (Potato)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEYRKAQRAEGPATILAIGTSTPSNCVDQSTYPDYYFRITNSEHKTELKEKFKRMCDKSMIKKRYMHLTEEILKENPNMCAYMAPSLDARQDIVVVEVPKLGKEAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGCDYQLAKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAAAIIMGSDPIIGVERPLFELVSAAQTLVPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLLEAFQPLGISDWNSLFWIAHPGGPAILDQVELKLGLKQEKLRATREVLSNYGNMSSACVLFILDEMRKASTKEGLGTTGEGLEWGVLFGFGPGLTVETVVLHSVAT
P17957	CHS2_SOYBN	reviewed	Chalcone synthase 2 (EC 2.3.1.74) (Naringenin-chalcone synthase 2)	CHS2	Glycine max (Soybean) (Glycine hispida)	388	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVEEIRKAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSEHMTELKEKFKRMCDKSMIKKRYMYLNEEILKENPSVCAYMAPSLDARQDMVVMEVPKLGKEAATKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPTDTHLDSLVGQALFGDGAAAVIVGSDPLPVEKPLFQLVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKALVEAFQPLGISDYNSIFRIAHPGGPAILDQVEAKLGLKPEKMEATRHVLSEYGNMSSACVLFILDQMRKKSIENGLGTTGEGLDWGVLFGFGPGLTVETVVLRSVTV
Q41436	CHSA_SOLTU	reviewed	Chalcone synthase 1A (EC 2.3.1.74) (Naringenin-chalcone synthase 1A)	CHS1A	Solanum tuberosum (Potato)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEYRKAQRAEGPATILAIGTSTPSNCVDQSTYPDYYFRITNSEHKTELKEKFKRMCDKSMIKKRYMHLTEEILKENPNMCAYMAPSLDARQDIVVVEVPKLGKEAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGCDYQLAKLLGLRPSVKRLMMYQQGCFVGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAAAIIMGSDPIIGVERPLFELVSAAQTLVPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLLEAFQPLGISDWNSLFWIAHPGGPAILDQVELKLGLKQEKLRATREVLSNYGNMSSACVLFILDEMRKASTNEGLGTTGEGLEWGVLFGFGPGLTVETVVLHSVAT
P51076	CHSY_FRAAN	reviewed	Chalcone synthase RJ5 (EC 2.3.1.74) (Naringenin-chalcone synthase) (Fragment)	CHLNCDRAFT_48950	Fragaria ananassa (Strawberry) (Fragaria chiloensis x Fragaria virginiana)	99	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	NITDWNSLFWIAHPGGPAILDQVEAKLALKPEKLEATRHILSEYGNMSSACVLFILDEVRRKSAANGHKTTGEGLEWGVLFGFGPGLTVETVVLHSVSA
Q9SBL6	CHS3_SORBI	reviewed	Chalcone synthase 3 (EC 2.3.1.74) (Naringenin-chalcone synthase 3)	CHS3	Sorghum bicolor (Sorghum) (Sorghum vulgare)	401	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAAATVTVEEVRKAQRATGPATVLAIGTATPANCVHQADYPDYYFRITKSEHMTDLKEKFKRMCDKSQIRKRYMHLTEEYLAENPNMCAYMAPSLDARQDIVVVEVPKLGKAAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKMLGLRPSVNRLMMYQQGCFAGGTVLRVAKDLAENNRGARVLVVCSEITAVTFRGPSESHLDSMVGQALFGDGAAAVIVGADPDERVERPLFQLVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIERSLEEAFKPLGITDYNSIFWVAHPGGPAILDQVEAKVGLEKERLRATRHVLSEYGNMSSACVLFILDEMRKRSAEDGQATTGEGFDWGVLFGFGPGLTVETVVLHSVPITTGAAITA
P23419	CHS2_SOLLC	reviewed	Chalcone synthase 2 (EC 2.3.1.74) (Naringenin-chalcone synthase 2)	CHS2	Solanum lycopersicum (Tomato) (Lycopersicon esculentum)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEVRRAQRAKGPATIMAIGTATPSNCVDQSTYPDYYFRITNSEHMTELKEKFKRMCDKSMINKRYMHLTEEILKENPNICEYMAPSLDARQDIVVVEVPKLGKEAAQKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVIRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSMVGQALFGDRAAAMIIGSDPLPEVERPLFELVSAAQTLLPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLIEAFQPLGISDWNSIFWIAHPGGPAILDQVELKLSLKPEKLRATRQVLSDYGNMSSACVLFILDEMRKASSKEGLSTTGEGLDWGVLFGFGPGLTVETVVLHSVST
P23418	CHS1_SOLLC	reviewed	Chalcone synthase 1 (EC 2.3.1.74) (Naringenin-chalcone synthase 1)	CHS1	Solanum lycopersicum (Tomato) (Lycopersicon esculentum)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEYRKAQRAEGPATILAIGTSTPSNCVDQSTYPDYYFRITNSEHKTELKEKFKRMCDKSMIKKRYMHLTEEILKENPNMCAYMAPSLDARQDIVVVEVPKLGKRGTQKAIKEWGQPKSKITHLVFCTTSGVDMPACDYQLAKLLPVRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAAAIIIGSDPIIGVERPLFELVSAAQTLVPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLLEAFQPLGISDWNSLFWIAHPGGPAILDQVELKLGLKPEKLRATREVLSNYGNMSSACVLFILDEMRKASTKEGLGTTGEGLEWGVLFGFGPGLTVETVVLHSVAA
P19168	CHS3_SOYBN	reviewed	Chalcone synthase 3 (EC 2.3.1.74) (Naringenin-chalcone synthase 3)	CHS3	Glycine max (Soybean) (Glycine hispida)	388	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVEEIRNAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSEHMTELKEKFKRMCDKSMIKKRYMYLNEEILKENPSVCAYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPTDTHLDSLVGQALFGDGAAAVIVGSDPLPVEKPLFQLVWTAQTILPDSEGAIDGHLGEVGLTFHLLKDVPGLISKNIEKALVEAFQPLGISDYNSIFWIAHPGGPAILDQVEAKLGLKPEKMEATRHVLSEYGNMSSACVLFILDQMRKKSIENGLGTTGEGLDWGVLFGFGPGLTVETVVLRSVTV
Q9XJ57	CHS2_CITSI	reviewed	Chalcone synthase 2 (EC 2.3.1.74) (Naringenin-chalcone synthase 2)	CHS2	Citrus sinensis (Sweet orange) (Citrus aurantium var. sinensis)	391	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MATVQEIRNAQRADGPATVLAIGTATPAHSVNQADYPDYYFRITKSEHMTELKEKFKRMCDKSMIKKRYMYLTEEILKENPNMCAYMAPSLDARQDIVVVEVPKLGKEAATKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLIGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPADTHLDSLVGQALFGDGAAAVIVGADPDTSVERPLYQLVSTSQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLSEAFAPLGISDWNSIFWIAHPGGPAILDQVESKLGLKGEKLKATRQVLSEYGNMSSACVLFILDEMRKKSVEEAKATTGEGLDWGVLFGFGPGLTVETVVLHSVPIKA
O22045	CHSD_IPONI	reviewed	Chalcone synthase D (EC 2.3.1.74) (Naringenin-chalcone synthase D) (CHS-D)	CHSD	Ipomoea nil (Japanese morning glory) (Pharbitis nil)	388	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEVRKAQRAEGPATILAIGTATPANCVDQSTYPDYYFRITNSDHMTDLKQKFQRMCDKSMITKRYMHLTEEILKENPSFCEYMAPSLDARQDIVVVEVPKLGKEAAQSAIKEWGQPKSKITHVIFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRVAKDLAENNKAARVLVVCSEITVVTFRGPNETHLDSLVGQALFGDGAAAIIVGSDPTPAEKPLFQLVSAAQTLAPDSCGAIDGHLREVGLTFHLLKDVPSIVSNNIEKCLSEAFNPLGISDWNSIFWIAHPGGPAILDQVEDKLGLKPEKLRATRHVLSEYGNMSSACVLFILDEMRKASSNDGLGTTGEGLEWGVLFGFGPGLTIETVVLHSVPA
O22046	CHSE_IPONI	reviewed	Chalcone synthase E (EC 2.3.1.74) (Naringenin-chalcone synthase E) (CHS-E)	CHSE	Ipomoea nil (Japanese morning glory) (Pharbitis nil)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEVRKAQRAQGPATIMAIGTSTPQNCVDQSTYPDYYFRITNSEHLVELKEKFKRMCEKSMIKKRYMYLTEEILKENPNICAYMAPSLDARQDIVVVEVPKLGKEAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLQPSVKRFMMYQQGCFAGGTVIRLAKDLAENNKGARVLVVCSEITAVTFRGPSDAHLDSLVGQALFGDGAAALIIGSDPDPDLERPLFQLVSAAQTILPDSGGAIDGHLREVGLTFHLLKDVPGLISKHIEKSLNEAFQPLGIHDWNSLFWIAHPGGPAILDQVEEKLELKPEKLRATRHVLSEYGNMSSACVLFILDEMRKASSKEGLNTTGEGLEWGVLFGFGPGLTVETVVLHSVSA
P30079	CHSY_PINSY	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Pinus sylvestris (Scots pine)	396	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAAGMMKDLEAFRKAQRADGPATILAIGTATPPNAVDQSSYPDYYFKITNSEHMTELKEKFRRMCDKSAIKKRYMYLTEEILKENPKVCEYMAPSLDARQDMVVVEVPRLGKEAAAKAIKEWGQPKSKITHVIFCTTSGVDMPGADYQLTKLLGLRPSVKRVMMYQQGCFAGGTVLRVAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSMVGQALFGDGAAALIVGADPVPEVEKPCFELMWTAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFQQFGISDWNQLFWIAHPGGPAILDQVEAKLNLDPKKLSATRQVLSDYGNMSSACVHFILDEMRKSSKEKGCSTTGEGLDVGVLFGFGPGLTVETVVLKSVPLLD
Q9SML4	CHS1_CICAR	reviewed	Chalcone synthase 1 (EC 2.3.1.74) (Naringenin-chalcone synthase 1)	CHS1	Cicer arietinum (Chickpea) (Garbanzo)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVSEIRKAQRAEGPATILAIGTANPSNRVEQSTYPDFYFKITNSEHKVELKQKFQRMCDKSMIKSRYMYLTEEILKENPSVCEYMAPSLDVRQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPIPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNIDKALIEAFQPLNISDYNSIFWIAHPGGPAILDQVEEKLALKPEKMRATREVLSEYGNMSSACVLFILDEMRRKSAKDGLKTTGEGLEWGVLFGFGPGLTIETVVLHSVAI
P30073	CHS1_MEDSA	reviewed	Chalcone synthase 1 (EC 2.3.1.74) (Naringenin-chalcone synthase 1)	CHS1	Medicago sativa (Alfalfa)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVAEIRQAQRAEGPATIMAIGTANPANCVEQSTYPDFYFKITNSEHKVELKEKFQRMCDKSMIKRRYMYLTEEILKDNPRVCEYMAPSLAARQDMAVVVVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEETPVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPIPEIEKPIFEMVWTAHTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNIDKALIEAFQPLNISDYNSIFWIAHPGGPAILDQVEEKLGLKPEKMKATREVLSEYGNMSSACVLFILDEMRKKSVQAGLKTTGEGLDWGVLFGFGPGLTIETVVLHSVAI
Q01287	CHS2_PEA	reviewed	Chalcone synthase 2 (EC 2.3.1.74) (Naregenin-chalcone synthase 2)	CHS2	Pisum sativum (Garden pea)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVSEIRKAQRAEGPATILAIGTANPANCVEQSTYPDFYFKITNSEHKTVLKEKFQRMCDKSMIKRRYMYLTEDILKENPSLCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVFRLAKDLAENNKNARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPVPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNIDKALVEAFKPLGISDYNSIFWIAHPGGPAILDQVEQKLALKPEKMRATREVLSEYGNMSSACVLFILDEMRKKSTQDGLNTTGEGLEWGVLFGFGPGLTIETVVLRSVAI
P51082	CHSB_PEA	reviewed	Chalcone synthase 1B (EC 2.3.1.74) (Naringenin-chalcone synthase 1B)	CHS-1B	Pisum sativum (Garden pea)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVDEIRQAQKAEGPATVLAIGTATPPNCVDQSTYPDYYFRITNSEHKTELKEKFQRMCDKSMIKKRYMHLTEEILKENPSVCEYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPHVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIVGSDPLPQVEKPLFELVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKALVEAFQPLGISDYNSLFWIAHPGGPAILDQVEAKLSLKQRKMQATRHVLSEYGNMSSACVLFILDEMRRKSKEEQLGTTGEGLEWGVLFGFGPGLTVETVVLHSVAT
Q9SBL4	CHS5_SORBI	reviewed	Chalcone synthase 5 (EC 2.3.1.74) (Naringenin-chalcone synthase 5)	CHS5	Sorghum bicolor (Sorghum) (Sorghum vulgare)	401	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAAATVTVEEVRKAQRATGPATVLAIGTATPANCVHQADYPDYYFRITKSEHMTELKEKFKRMCDKSQIRKRYMHLTEEYLAENPNMCAYMAPSLDARQDIVVVEVPKLGKAAAHKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKMLGLRPSVNRLMMYQQGCFAGGTVLRVAKDLAENNRGARVLVVCSEITAVTFRGPSESHLDSMVGQALFGDGAAAVIVGADPDERVERPLFQLVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIERSLEEAFKPLGITDYNSIFWVAHPGGPAILDQVEAKVGLEKERMRATRHVLSEYGNMSSACVLFILDEMRKRSAEDGQATTGEGFDWGVLFGFGPGLTVETVVLHSVPITTGAAITA
P51081	CHSA_PEA	reviewed	Chalcone synthase 1A (EC 2.3.1.74) (Naringenin-chalcone synthase 1A)	CHS-1A	Pisum sativum (Garden pea)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVNEIRQAQRAEGPATVFAIGTATPQNCVEQSTYPDFYFRITNSQHKTELKEKFQRMCDKSMIKKRYMHLTEEILKENPSLCEYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIVGSDPLPDVEKPLFELVWTAQTIVPDSEGAIDGHLREAGLTFHLLKDVPSLVSKNIEKALVEAFQPLNISDYNSIFWIAHPGGPAILDQVEAKLGLKQRKMQATRHVLSEYGNMSSACVLFILDEMRRKSKEDGLATTGEGLEWGVLFGFGPGLTVETVLLHSMAT
P06515	CHSY_ANTMA	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Antirrhinum majus (Garden snapdragon)	390	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEVRRAQRAEGPATVLAIGTATPANCVDQSTYPDYYFRITNSEHMTELKEKFKRMCDKSNIKRRYMHLTEEILKENPAMCEYMAPSLDARQDIVVVEVPRLGKEAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRMAKDLAENNAGARVLVVCSEITAVTFRGPADTHLDSLVGQALFGDGAAAVIVGSDPVVGVERPLFQIVTAAQTLLPDSHGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLKEAFDPLGISDWNSVFWIAHPGGPAILDQVEEKLGLKPEKLRSTRQVLSEYGNMSSACVLFILDEMRKSSAKEGMSTTGEGLDWGVLFGFGPGLTVETVVLHSVPLN

P51089	CHSY_VIGUN	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Vigna unguiculata (Cowpea)	388	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVSEIRQAQRAEGPATILAIGTATPPNCVDQSTYPDYYFRITNSEHMTDLKEKFQRMCDKSMIKKRYMHVTEEILKENPSMCAYMAPSLDARQDIVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIVGSDPIPQIEKPLFELVWTAQTIAPDSEGAIDGHLREVGLTFHLLKDVPGIVSKNIGKALSEAFDPLNISDYNSIFWIAHPGGPAILDQVAQKLGLKPEKMKATRDVLSDYGNMSSACVTFHLDEIEKSVENGLKTTGKDLEWGVLFGFGPGLSLETVVLHSVAI
Q9SBL7	CHS2_SORBI	reviewed	Chalcone synthase 2 (EC 2.3.1.74) (Naringenin-chalcone synthase 2)	CHS2	Sorghum bicolor (Sorghum) (Sorghum vulgare)	401	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAGATVTVEEVRKAQRATGPATVLAIGTATPANCVHQADYPDYYFRITKSEHMTELKEKFKRMCDKSQIRKRYMHLTEEYLAENPNMCAYMAPSLDARQDIVVVEVPKLGKAAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKMLGLRPSVNRLMMYQQGCFAGGTVLRVAKDLAENNRGARVLVVCSEITAVTFRGPSESHLDSMVGQALFGDGAAAVIVGADPDERVERPLFQLVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIERSLEEAFKPLGITDYNSIFWVAHPGGPAILDQVEAKVGLEKERMRATRHVLSEYGNMSSACVLFILDEMRKRSAEDGRATTGEGFEWGVLFGFGPGLTVETVVLHSVPITTGAAITA
P48392	CHS3_GERHY	reviewed	Chalcone synthase 3 (EC 2.3.1.74) (Naringenin-chalcone synthase 3)	CHS3	Gerbera hybrida (Daisy)	403	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MATSPAVIDVETIRKAQRAEGPATILAIGTATPANCVYQADYPDYYFRVTESEHMVDLKEKFQRMCDKSMIRKRYMHITEEFLKENPSMCKFMAPSLDARQDLVVVEVPKLGKEAATKAIKEWGFPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPNEGHLDSLVGQALFGDGAAAVIIGSDPDLSVERPLFEMVSAAQTILPDSEGAIDGHLKEVGLTFHLLKDVPALIAKNIEKALIQAFSPLNINDWNSIFWIAHPGGPAILDQVEFKLGLREEKLRASRHVLSEYGNMSSACVLFILDEMRKKSIKDGKTTTGEGLEWGVLFGFGPGLTVETVVLHSLPATISVATQN
P23569	CHSY_PUEML	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Pueraria montana var. lobata (Kudzu vine) (Pueraria lobata)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVAEIRQAQRAEGPATILAIGTANPPNCVDQSTYPDYYFRITNSEHMTELKEKFQRMCDKSMIKKRYMYLTEEILKENPNMCAYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKQLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIVGSDPIPQVEKPLYELVWTAQTIAPDSEGAIDGHLREVGLTFHLLKDVPGIVSKNIDKALFEAFNPLNISDYNSIFWIAHPGGPAILDQVEQKLGLKPEKMKATRDVLSDYGNMSSACVLFILDEMRRKSAENGLKTTGEGLEWGVLFGFGPGLTIETVVLRSVAI
P16107	CHSY_PETCR	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Petroselinum crispum (Parsley) (Petroselinum hortense)	398	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MANHHNAEIEEIRNRQRAQGPANILAIGTATPSNCVYQADYPDYYFRITNSEHMTDLKLKFKRMCEKSMIRKRYMHITEEYLKENPNVCAYEAPSLDARQDLVVVEVPRLGKEAASKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNAGARVLVVCSEITAVTFRGPSDSHLDSLVGQALFGDGAAAVILGSDPDLSVERPLFQLISAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLKEAFGPIGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKMRATRQVLSDYGNMSSACVLFILDEMRKKSIEEGKATTGEGLDWGVLFGFGPGLTVETVVLHSVPATFTH
O22652	CHSY_RAPSA	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Raphanus sativus (Radish)	394	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVGTTSSLDEIRKAQRADGPAGILAIGTANPANHVLQAEYPDYYFRITNSEHMTDLKEKFKRMCDKSTIRKRHMHLTEEFLKENPNMCAYMAPSLDARQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDVSAGEKPIFEMVSAAQTILPDSDGAIDGHLREVGITFHLLKDVPGLISKNIEKSLDEAFKPLGISDWNSLFWIAHPGGPAILDDVEKKLGLKAEKMRATRHVLSEYGNMSSACVLFILDEMRRKSLDDGVATTGEGLEWGVLFGFGPGLTVETVVLHSVPV
P48386	CHS1_CAMSI	reviewed	Chalcone synthase 1 (EC 2.3.1.74) (Naringenin-chalcone synthase 1)	CHS1	Camellia sinensis (Tea)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEDIRRAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSEHKAELKEKFKRMCDKSMIKKRYMYLTEEILKENPQVCEYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAIIVGSDPIPEVEKPLFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLAEAFQPLGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKLRATRHVLSEYGNMSSACVLFILDEMRKKSAADGLKTTGEGLEWGVLFGFGPGLTVETVVLHSLST
Q9FSB9	CHS1_RUTGR	reviewed	Chalcone synthase 1 (EC 2.3.1.74) (Naringenin-chalcone synthase 1)	CHS1	Ruta graveolens (Common rue)	393	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAAVTVEAIRKAQRADGPAAVLAIGTATPANYVTQADYPDYYFRITKSEHMTELKEKFKRMCDKSMIRKRYMHLTEDILKENPNMCAYMAPSLDARQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGCDYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPADTHLDSLVGQALFGDGAAAVIVGADPDESIERPLYQLVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLKEAFGPIGISDWNSIFWIAHPGGPAILDQVEAKLGLKEEKLRATRQVLSEYGNMSSACVLFILDEMRKNCAEEGRATTGEGLDWGVLFGFGPGLTVETVVLRSVPIKA
O23729	CHS3_BROFI	reviewed	Chalcone synthase 3 (EC 2.3.1.74) (Naringenin-chalcone synthase 3)	CHS3	Bromheadia finlaysoniana (Orchid)	394	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAPAMEEIRQAQRAEGPAAVLAIGTSTPPNALYQADYPDYYFRITKSEHLTELKEKFKRMCDKSMIKKRYMYLTEEILKENPNICAFMAPSLDARQDIVVTEVPKLAKEAAVRAIKEWGHPKSRITHLIFCTTSGIDMPGADYQLTRLLGLRPSVNRFMLYQQGCFAGGTVLRLAKDLAENNAGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAAAIIVGSDPDSATERPLFQLVSASQTILPESEGAIDGHLREIGLTFHLLKDVPGLISKNIQKCLLDAFKPLGVHDWNSIFWIAHPGGPAILDQVEIKLGLKAEKLAASRNVLAEYGNMSSACVLFILDEMRRRSAEAGQATTGEGLEWGVLFGFGPGLTVETIVLRSVPIAGAE
P48399	CHSC_IPOPU	reviewed	Chalcone synthase C (EC 2.3.1.74) (Naringenin-chalcone synthase C) (CHS-C) (Fragment)	CHSC	Ipomoea purpurea (Common morning glory) (Pharbitis purpurea)	352	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MSTAVTILTDTWTRREKRFDGHAKILAIGTAIPANWVDQTTYPDFYFRITNSEHLLEYKEKFRRICNKSKIRKRHLVITEELLKKNPNLCTYNDASLNTRQDILVSEVPKLGKEAAMKAIKEWGRPISEITHLVFCTSSGVDMPGADFQLAKLLGLSSSVNRLMMYQQGCNAGAVMLRLAKDLAENNKGGRVLVVCSEVMLSVFRGPSLQQEDNLLAQCLFGDGSAAVIVGTEPRPGLETPLFELVSAAQTTIPDTDSYLKLQLREMGLTFHCSKAVPSLITQNIEDCLVKAFEPFGISDWNSIFWILHPGGNAILDGVEEKLGLEPEKLRASRDVLSQYGNLTSACVLFKP
P22925	CHSD_PETHY	reviewed	Chalcone synthase D (EC 2.3.1.74) (Naringenin-chalcone synthase D)	CHSD	Petunia hybrida (Petunia)	419	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEVRNAQRAEGPATVLAIGTATPSNCVDQSTYPDYYFRITDSEHKTELKEKFKRICDKSMIKKRYMHLTEKILKENPNICESMAPSLDARTNIYAVEVPKLGKEAAEKAIEEWNQPKSRITHLVFCTTTGVSMPGADFQLTKLLGLGSSVKRFMMNQLGCFAGGTVLRLAKDLAENNKGARVLVVCSEITVVTFRGPNDTHFDSLVGQALFGDGAAAVIIGSDPIPNVERPLFELVSAAQTLLPDSKNSICGELREIGLTFHLLKDVAELISNNIEKSLVEVFQPLGISAWNSIFWVAHPGGPAILNQVELKLGLNPEKLGATRHVLSEYGNMSSASILFVLDEMRKSSTQKGFDTTGEGLKWGVLIGFGPGITFETIVLHSVSTQGSFGIRDWDYNFGIEFKIIFIPYLVLGIS
P48390	CHS1_GERHY	reviewed	Chalcone synthase 1 (EC 2.3.1.74) (Naringenin-chalcone synthase 1)	CHS1	Gerbera hybrida (Daisy)	398	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MASSVDMKAIRDAQRAEGPATILAIGTATPANCVYQADYPDYYFRITKSEHMVDLKEKFKRMCDKSMIRKRYMHITEEYLKQNPNMCAYMAPSLDVRQDLVVVEVPKLGKEAAMKAIKEWGHPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPNDTHLDSLVGQALFGDGAAAVIVGSDPDLTTERPLFEMVSAAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKALTTAFSPLGINDWNSIFWIAHPGGPAILDQVELKLGLKEEKLRATRHVLSEYGNMSSACVLFIIDEMRKKSSENGAGTTGEGLEWGVLFGFGPGLTVETVVLHSVPTTVTVAV
Q01286	CHS1_PEA	reviewed	Chalcone synthase 1 (EC 2.3.1.74) (Naregenin-chalcone synthase 1)	CHS1	Pisum sativum (Garden pea)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVSEIRKPQRAEGPATILAIGTANPANCVEQSTYPDFYFKITNSEHKTVLKEKFQRMCDKSMIKRRYMYLTEEILKENPSLCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPVPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREQGLTFHLLKDVPGIVSKNIDKALVEAFKPLGISDYNSIFWIAHPGGPAILDQVEQKLGLKPEKMRATREVLSEYGNMSSACVLFILDQMRKKSTQDGLNTTGEGLEWGVLFGFGPGLTIETVVLHSVAI
P48388	CHS3_CAMSI	reviewed	Chalcone synthase 3 (EC 2.3.1.74) (Naringenin-chalcone synthase 3)	CHS3	Camellia sinensis (Tea)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEDVWRAQRARGPATVLAIGTATPPNCVDQSTYPDYYFRITNSEHKVELKEKFKRMCDKSMIKKRYMYLTEEILKENPLVCEYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQSLFGDGAAAIIIGSDPIPEVEKPLFELVSAAQTILPSSDGAIDGHLREVGLTFHLLKDVPRLISMNVEKSLVEAFQPLGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKLRATRHVLSEYGNMSSACVLFILDEMRKKSAEEGLKTTGEGLEWGVLFGFGPGLTVETVVLHSLCT
O22047	CHSE_IPOPU	reviewed	Chalcone synthase E (EC 2.3.1.74) (Naringenin-chalcone synthase E) (CHS-E)	CHSE	Ipomoea purpurea (Common morning glory) (Pharbitis purpurea)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEVRKAQRAQGPATIMAIGTSTPQNCVDQSTYPDYYFRITNSEHLVELKEKFKRMCEKSMIKKRYMYLTEEILTENPNICAYMAPSLDARQDIVVVEVPKLGKEAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLQPSVKRFMMYQQGCFAGGTVIRLAKDLAENNKGARVLVVCSEITAVTFRGPSDAHLDSLVGQALFGDGAAALIIGSDPDPDLERPLFQLVSAAQTILPDSGGAIDGHLREVGLTFHLLKDVPGLISKHIEKSLNEAFQPLGIRDWNSLFWIAHPGGPAILDQVEEKLELKPEKLRATRHVLSEYGNMSSACVLFILDEMRKASSKEGLNTTGEGLEWGVLFGFGPGLTVETVVLHSVSA
Q9LKP7	CHSY_DIAMO	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Dianthus monspessulanus	391	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MASIEEIRQAQRADGPATILAIGTATPPNAIYQADYPDYYFRVTKSEHMTELKEKFRRMCDKSMIKKRYMYLTEEILKENPNLCEYMGSSLDTRQDMVVSEVPRLGKEAAVKAIKEWGQPKSKITHVIMCTTSGVDMPGADYQLTKLLGLRPSVRRFMLYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAICFRGPTEAALDSMVGQALFGDGAGALIVGSDPDLSIERPLFQMAWAGQTLLPDSEGAIDGHLREVGLTFHLLKDVPGIISKNITNALEEAFSPIGVSDWNNLFWIAHPGGPAILDQVEAKLGLKEEKLAATRNVLSDFGNMSSACVLFILDEMRKKSLRDGATTTGEGLDWGVLFGFGPGLTVETVVLHSVPLNC
O82144	CHSY_HYDMC	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Hydrangea macrophylla (Bigleaf hydrangea) (Viburnum macrophyllum)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEVRKAQRAEGPATILAIGTATPPNYVDQSTYPDFYFRVTNSEHKKELKAKFQRMCDNSQIKKRYMHLTEEILKENPNICAYMAPSLDARQDMVVVEIPKLGKEAATRAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIIGSDPMPEVEKPLFEIVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFRPLDISDWNSIFWIAHPGGPAILDQVEKKLALKPEKLRATRNVLSDYGNMSSACVLFIMDEMRKNSAEEGLMTTGEGLEWGVLFGFGPGLTVETVVLHGVST
O22586	CHSY_ONOVI	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Onobrychis viciifolia (Common sainfoin)	390	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVDVAESRRTQRAEGPRTIFAIATANPPNCVEQSTYPDFYFKITNSEHKVELKEKFQRMDDKSMIKRRYMYLTEEILKENPNVCEYMAPSLDARQDMFVVEVPRLGKEAAVKAIKEWGQPKSKITHLIVCTTSGVDMPGADYQLTKLLGLRPHVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPIPEIEKPLFELIWTAQTIAPDSEGAIDGHLREVGLTFHLLKDVPGIVSKNIDKALVEAFQPLGISDYNSIFWIAHPGGPAILDQVEQKLAFKPEKMRATREVLSEYGNMSSACVLFILDEMRKKSAQNGLKTTGEGLEWGVLFGFGPGLTIETVVLLRSVAI
Q9M5M0	CHS7_PICMA	reviewed	Chalcone synthase 7 (EC 2.3.1.74) (Naregenin-chalcone synthase 7)	CSF7	Picea mariana (Black spruce) (Abies mariana)	395	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAGGLMADLEAFRKAQRADGPATILAIGTATPPNAVDQSTYPDYYFKITNSEHMTELKEKFQRMCDKSAIKKRYMYLTDEILKENPNVCEYMAPSLDARQDMVVVEVPRLGKEAATKAIKEWGQPKSKITHVIFCTTSGVDMPGADYQLTKLLGLRPSVKRVMMYQQGCFAGGTVLRVAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSMVGQALFGDGAAALIVGADPIPQVEKPCFELMWTAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFQQFGISDWNQLFWIAHPGGPAILDQVEAKLNLDPKKLRATRQVLSEYGNMSSACVHFILDEMRKSSNEKGCSTTGEGLDVGVLFGFGPGLTVETVVLKSVPLQ
P48398	CHSB_IPOPU	reviewed	Chalcone synthase B (EC 2.3.1.74) (Naringenin-chalcone synthase B) (CHS-B)	CHSB	Ipomoea purpurea (Common morning glory) (Pharbitis purpurea)	396	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MSTTVTVLTDTWSRRAKRLEGDAKILAIGTATPASWVDQTTYPDFYFRITNSQHLLEHKEKFRRICNKSKIRKRHLVLTEELLKKNPNLCTYNETSLNTRQDILVSEVPKLGKEAAMKAIKEWGRPISEITHLVFCTTSGVDMPGADFQLTKLLGLNSSVKRLMMYQQGCNAGAAMLRLAKDLAESNKGGRVLVVCAEITINIFRGPSLEQDDNLLAQCLFGDGAAAMIVAADPRPGLETPLFELVSSAQTIVPNTDSHLKLHLREMGLTFHCSKAVPSVLAENVEDCLVKAFEPYGISDWNSIFWVFHPGGNAIVDRVEERLGLGPEKLRASRDVLSEYGNLTSACVLFILDEMRKKSKKDEQMTTGEGLEWGVVFGFGPGLTIDTIIIRSVPIN
Q9SEP4	CHSY_ARAAL	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Arabis alpina (Alpine rock-cress)	391	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAPSLEEIRKAQRADGPAGILGIGTANPPNHVLQAEYPDYYFRITNSDHMTDLKEKFKRMCDKSMIRKRHMHLTEEFLKENPKMCAYMAPSLDTRQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDTSVGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLDEAFKPLGISDWNSLFWIAHPGGPAILDQVEIKLGLKAEKMRATRHVLSEYGNMSSACVLFILDEMRKKSAKDGAATTGEGLEWGVLFGFGPGLTVETVVLHSVPL
O04220	CHSY_CHRAE	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Chrysosplenium americanum (Golden saxifrage)	396	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MSSAALMEEIRNAQRAEGPATILAIGTATPANCVIQADYPDFYFRITNSEHKTELKEKFQRMCDKSMIKKRYMHLTEDLLKENPKMCEYMAPSLDARQDMVVVEIRKLGKEAAVKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFPGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAALIVGADPDTAIERPLFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLIEAFKPIGINDWNSIFWIAHPGGPAILDQVEHKLAPGRRANSRATRHILSEYGNMSSACVLFILDEMRKKSARRREATTGDGLEWGVLFGFGPGLTVETVVLHSVPAITA
Q9MBB1	CHSY_EQUAR	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Equisetum arvense (Field horsetail) (Common horsetail)	405	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MTVLEESADASSRRLAQRANGPATVLAIGTANPANVFEQSSYPDFYFDITNSQHMTELKLKFSRMCQKSGIKKRYMHLNSEILKANPSLCAYWEKSLDVRQDIAVVEVPKLGKEASLKAIKEWGQPKSKITHLVFCTTSGVDMPGADWALTKLLGLRPSVKRLMMYQQGCFAGGTVLRVAKDVAENNKGARVLVVCSEITCVTFRGPSETHLDSLVGQALFGDGAAAVILGSDPLPEENPCFELHWSGSNILPDSDGAIDGHLREVGLTFHLMKDVPGIISKNIGKVLNDAFRSAFDESGNAEDRPASVNDIFWIAHPGGPAILDQVEEKMKLAPEKMRATRDVLSEYGNMSSACVLFIMDHMRRMSAQNKLQTTGEGLDWGVLLGFGPGLTVETVLLKSIRLAC

O04111	CHSY_PERFR	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Perilla frutescens (Beefsteak mint) (Perilla ocymoides)	391	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEDIRRAQRAEGPATVMAIGTATPENCVDQSTYPDYYFRITNSEHRTDLKEKFKRMCEKSMIRKRYMHLTEEFLKENPNMTAYMAPSLDARQDIVVVEVPKLGKEAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRMAKDLAENNAGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAAAVIVGSDPVVGVERPLFQLVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLKEAFGPLGISDWNSVFWIAHPGGPAILDQVEAKLGLKPEKLRSTRHVLGEYGNMSSACVLFILDEMRKSSAKEGMSSTGEGLDWGVLFGFGPGLTVETVVLHSVPINN
P30077	CHS9_MEDSA	reviewed	Chalcone synthase 9 (EC 2.3.1.74) (Naringenin-chalcone synthase 9)	CHS9	Medicago sativa (Alfalfa)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVSEIRQAQRAEGPATIMAIGTANPANCVEQSTYPDFYFKITNSEHKVELKEKFQRMCDKSMIKRRYMYLTEEILKENPSVCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPIPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNIDKALVEAFQPLNISDYNSIFWIAHPGGPAILDQVEQKLGLKPEKMKATREVLSEYGNMSSACVLFILDEMRKKSAQAGLKTTGEGLDWGVLFGFGPGLTIETVVLHSVAI
P22928	CHSJ_PETHY	reviewed	Chalcone synthase J (EC 2.3.1.74) (Naringenin-chalcone synthase J)	CHSJ	Petunia hybrida (Petunia)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEIRRAQRAEGPATIMAIGTATPSNCVDQSTYPDYYFRITNSEHKTELKEKFQRMCDKSMIKKRYMHLTEEILKENPNICEYMAPSLDARQDIVVVEVPKLGKEAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRSSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPNDTHLDSLVGQALFGDGAAAIIIGSDPLPGVERPLFELVSASQTLLPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIQKSLVEAFQPLGISDWNSIFWIAHPGGPAILDQVELKLGLKPEKLRATRHVLSEYGNMSSACVLFILDEMRKASSKEGLGTTGEGLEWGVLFGFGPGLTVETVVLHSVST
P48385	CHSY_CALCH	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Callistephus chinensis (China aster) (Callistemma chinense)	398	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MASTIDIAAIREAQRRQGPATILAIGTATPSNCVYQADYPDYYFRITKSEHMVDLKEKFKRMCDKSMIRKRYMHLTEEYLKENPSLCEYMAPSLDARQDVVVVEVPKLGKEAATKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPNDTHLDSLVGQALFGDGAAAVIVGADPDLTTERPLFEMISAAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKALTQAFSPLGITDWNSIFWIAHPGGPAILDQVELKLGLKEEKMRATRHVLSEYGNMSSACVLFIIDEMRKKSAEDGAATTGEGLDWGVLFGFGPGLTVETVVLHSLPTTMAIAT
P48389	CHSY_DIACA	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Dianthus caryophyllus (Carnation) (Clove pink)	391	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MASIEEIRQAPRADGPATILAIGTATPPNAIYQADYPDYYFRVTKSEHMTELKEKFRRMCDKSMIKKRYMYLTEEILKENPNLCEYMGSSLDTRQDMVVSEVPRLGKEAAVKAIKEWGQPKSKITHVIMCTTSGVDMPGADYQLTKLLGLRPSVRRFMLYQQGCFAGGTVLRLAKDLAENNKDARVLVVCSEITAICFRGPTEAALDSMVGQALFGDGAGALIVGSDPDLSIERPLFQMAWAGQTLLPDSDGAIDGHLREVGLTFHLLKDVPGIISKNITNALEDAFSPIGVSDWNNLFWIAHPGGPAILDQVEAKLGLKEEKLAATRNVLSDFGNMSSACVLFILDEMRKKSLRDGATTTGEGLDWGVLFGFGPSLTVETVVLHSVPLNC
Q9FSB7	CHS3_RUTGR	reviewed	Chalcone synthase 3 (EC 2.3.1.74) (Naringenin-chalcone synthase 3)	CHS3	Ruta graveolens (Common rue)	393	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAAVTVEEIRKAQRADGPATVLAIGTATPANYVTQADYPDYYFRITKSEHMTDLKEKFKRMCDKSMIRKRYMHLTEEILKENPNMCAYMAPSLDARQDIVVVEVPKLGKEAAAKAIKEWGQPKSKITHLIFCTTSGVDMPGCDYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPVDTHLDSLVGQALFGDGAAAVIVGADPDESIERPLYQLVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLKEAFGPIGISDWNSIFWIAHPGGPAILDQVEEKLELKEEKLRATRHVLSEYGNMSSACVLFILDEMRKKCAEEGMATTGEGLEWGVLFGFGPGLTVETVVLRSVPIKA
O23730	CHS4_BROFI	reviewed	Chalcone synthase 4 (EC 2.3.1.74) (Naringenin-chalcone synthase 4)	CHS4	Bromheadia finlaysoniana (Orchid)	394	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAPAMEEIRQAQRAEGPAAVLAIGTSTPPNALYQADYPDYYFRITKSEHLTELKEKFKRMCDKSMIKKRYMYLTEEILKENPNICAFMAPSLDARQDIVVTEVPKLAKEAAVRAIKEWGHPKSRITHLIFCTTSGIDMPGADYQLTRLLGLRPSVNRFMLYQQGCFAGGTVLRLAKDLAENNAGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAAAIIVGSDPDSATERPLFQLVSASQTILPESEGAIDGHLREIGLTFHLLKDVPGLISKNIQKCLLDAFKPLGVHDWNSIFWIAHPGGPAILDQVEIKLGLKAEKLAASRSVLAEYGNMSSACVLFILDEMRRRSAEAGQATTGEGLEWGVLFGFGPGLTVETIVLRSVPIAGAE
Q9ZRR8	CHS1_CASGL	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS1	Casuarina glauca (Swamp oak)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEVRKAQRAEGPATVLAIGTATPPNCLDQSTYPDYYFRITNSEHKTELKEKFQRMCDKSMIKKRYMYLTEEILKEHPNMCAYMAPSLDARQDMVVVEIPKLGKEAAVKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAIIVGADPLPEVEKPLFEVVSTAQTILPDSDGAIDGHLREVGVTFHLLKDVPGLISKNIEKSLVEAFQPLGISDWNSLFWIAHPGGPAILDQVEEKLALKPEKLGATRHVLSEYGNMSSACVLFILDEMRRKSAEKGLKTTGEGLDWGVLFGFGPGLTVETVVLHSLTT
P48387	CHS2_CAMSI	reviewed	Chalcone synthase 2 (EC 2.3.1.74) (Naringenin-chalcone synthase 2)	CHS2	Camellia sinensis (Tea)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEVRRAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSEHKTELKEKFQRMCDKSMIKKRYMYLTEEILKENPNVCAYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDAHLDSLVGQALFGDGAAAIIVGSDPIPEVEKPLFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLNEAFQPLNITDWNSLFWIAHPGGPAILDQVELKLALKPEKLRATRHVLSEYGNMSSACVLFILDEMRKSSAKKGLKTTGEGLDWGVLFGFGPGLTVETVVLHSVST
Q01288	CHS6_PEA	reviewed	Chalcone synthase 6 (EC 2.3.1.74) (Naregenin-chalcone synthase 6)	CHS6	Pisum sativum (Garden pea)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVSEIRKAQRAEGPATILAIGTATPANCVEQSTYPDFYFKITNSEHKTVLKEKFQRMCDKSMIKRRYMYLTEEILKENPSLCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPLPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNINKALVEAFQPLNIDDYNSIFWIAHPGGPAILDQVEEKLGLKPEKMKATREVLSEYGNMSSACVLFILDEMRKKSAQQGLKTTGEGLDWGVLFGFGPGLTIETVVLHSVAI
P30076	CHS8_MEDSA	reviewed	Chalcone synthase 8 (EC 2.3.1.74) (Naringenin-chalcone synthase 8)	CHS8	Medicago sativa (Alfalfa)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVSEIRTAQRAEGPATILAIGTANPANCVEQSTYPDFYFKITNSEHKTELKEKFQRMCDKSMIKRRYMYLTEEILKENPSVCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIVCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPVPEIEKPIFEMVWTAQTIAPDSEGGIDGHLREAGLTFHLLKDVPGIVSKNINKALVEAFEPLGISDYNSIFWIAHPGGPAILDQVEQKLALKPEKMKATREVLSEYGNMSSACVLVILDEMRKKSAQDGLKTTGEGLEFGVLFGFGPGLTIETVVLRSVAI
Q9FSB8	CHS2_RUTGR	reviewed	Chalcone synthase 2 (EC 2.3.1.74) (Naringenin-chalcone synthase 2)	CHS2	Ruta graveolens (Common rue)	393	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAAVTVEEIRKAQRADGPAAVLAIGTATPANYVTQADYPDYYFRITKSEHMTELKEKFKRMCDKSMIRKRYMYLTEDILKENPNMCAYMAPSLDARQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGCDYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPADTHLDSLVGQALFGDGAAAVIVGADPNESIERPLYQLVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLKEAFGPIGISDWNSIFWIAHPGGPAILDQVEAKLGLKEEKLRATRQVLSEYGNMSSACVLFILDEMRKKCAEEGRATTGEGLDWGVLFGFGPGLTVETVVLRSVPINA
O23882	CHS4_PEA	reviewed	Chalcone synthase 4 (EC 2.3.1.74) (Naregenin-chalcone synthase 4)	CHS4	Pisum sativum (Garden pea)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVSEIRKAQRAEGPATILAIGTANPANCVEQSTYPDFYFRITNSEHKIELKQKFQRMCDKSMINRRYMYLTEEILKENPSVCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPLPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPAIVSKNIDKALVEAFQPLGISDYNSIFWIAHPGGPAILDQVEQKLALKPEKMKATREVLSEYGNMSSACVLFILDEMRRKSIQNGLKTTGEGLEWGVLFGFGPGLTIETVVLHSVVI
O23884	CHS5_PEA	reviewed	Chalcone synthase 5 (EC 2.3.1.74) (Naregenin-chalcone synthase 5)	CHS5	Pisum sativum (Garden pea)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVSEIRKAQRAEGPATILAIGTANPANCVEQSTYPDFYFRITNSEHKTELKQKFQRMCDKSMINRRYMYLTEEILKENPSVCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPIPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPAIVSKNIDKALVEAFQPLGISDYNSIFWIAHPGGPAILDQVEQKLALKPEKMKATREVLSEYGNMSSACVLFILDEMRRKSIQNGLKTTGEGLEWGVLFGFGPGLTIETVVLHSVAI
Q9MB33	CHS1_IPOBA	reviewed	Chalcone synthase LF1 (EC 2.3.1.74) (Naringenin-chalcone synthase LF1)	CHS-LF1	Ipomoea batatas (Sweet potato) (Convolvulus batatas)	393	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEVRKAQRAEGPATILAIGTVTPANCVNQSTYPDYYFRITNSEHKTELKEKFQRMCDKSMITKRYMHLTEEILKENPSFCEYMAPSLDARQDIAVVEVPKLGKEAAQSAIKEWGQPKSKITHVVFCTTSGIDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLIVCSEITVVTFRGPSETHLDSLVGQALFGDGAAAVIVGADPTPAEKPLFQLVSAAQTLAPNSCGAIDGHLREVGLTFHLLKDVPSVVSNKIEKCLFEAFNPLGISDWNSVFWIAHPGGPAILDQVEDKLGLKPEKLRATRHVLSEYGNMSSACVLFILDEMRKASSNAGLGTTGEGLEWGVLFGFGPGLTIETVVLHSVPIKPGPH
P53414	CHS1_SECCE	reviewed	Chalcone synthase 1 (EC 2.3.1.74) (Naringenin-chalcone synthase 1)	CHS1	Secale cereale (Rye)	392	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAATMTVEEVRKAQRAEGPATVLAIGTATPANCVYQADYPDYYFKITKSDHMADLKEKFKRMCDKSQIRKRYMHLTEEILQDNPNMCAYMAPSLDARQDIVVVEVPKLGKAAAQKAIKEWGQPRSKITHLVFCTTSGVDMPGADYQLTKMLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPHEFDSLVGQALFGDGAAAVIVGADPDESVERPLFQLVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIERALEDAFKPLGIDDWNSVFWIAHPGGPAILDMVEAKVNLNKERMRATRHVLSEYGNMSSACVLFIMDEMRKRSAEDGHTTTGEGMDWGVLFGFGPGLTVETVVLHSVPVTA
P51083	CHS1_TRISU	reviewed	Chalcone synthase 1 (EC 2.3.1.74) (Naringenin-chalcone synthase 1)	CHS1	Trifolium subterraneum (Subterranean clover)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVAEIRKAQRAEGPATILAIGTANPPNRVEQATYPDFYFKITNSEHKVELKEKFQRMCDKSMIKSRYMYLTEEILKENPSVCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPVPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNIDKALVEAFQPLNISDYNSIFWIAHPGGPAILDQVEQKLSLKPEKMKATRDVLSEYGNMSSACVLFILDEMRKKSAQDGLKTTGEGLEWGVLFGFGPGLTIETVVLHSVAI
Q9MB41	CHS2_IPOBA	reviewed	Chalcone synthase LF2 (EC 2.3.1.74) (Naringenin-chalcone synthase LF2)	CHS-LF2	Ipomoea batatas (Sweet potato) (Convolvulus batatas)	393	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEVRKAQRAEGPATILAIGTVTPANCVNQSTYPDYYFRITNSEHKTELKEKFQRMCDKSMITKRYMHLTEEILKENPSFCEYMAPSLDARQDIAVVEVPKLGKEAAQSAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLSPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLIVCSEITVVTFRGPSETHLDSLVGQALFGDGAAAVIVGADPTPAEKPLFQLVSAAQTLAPDSCGAIDGHLREVGLTFHLLKDVPSVVSNNIEKCLFEAFNPLGISDWNSVFWIAHPGGPAILDQVEDKLGLKPEKLRATRHVLSEYGNMSSACVLFILDEMRKASSNAGLGTTGEGLEWGVLFGFGPGLTIETVVLHSVPIKPGPH
P53415	CHS2_SECCE	reviewed	Chalcone synthase 2 (EC 2.3.1.74) (Naringenin-chalcone synthase 2)	CHS2	Secale cereale (Rye)	394	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAATMTVEEVRKAQRAEGPATVLAIGTATPANCVYQADYPDYYFKITKSDHMADLKEKFKRMCDKSQIRKRYMHLTEEILQDNPNMCAYMAPSLDARQDIVVVEVPKLGKAAAQKAIKEWGQPRSKITHLVFCTTSGVDMPGADYQLTKMLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPHESHLDSLVGQALFGDGAAAVIIGADPDESIERPLFQLVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIERALEDAFKPLGIDDWNSVFWIAHPGGPAILDMVEAKVNLNKERMRATRHVLSEYGNMSSACVLFIMDEMRKRSAEDGHTTTGEGMDWGVLFGFGPGLTVETVVLHSVPVTA
Q9MB39	CHS4_IPOBA	reviewed	Chalcone synthase LF4 (EC 2.3.1.74) (Naringenin-chalcone synthase LF4)	CHS-LF4	Ipomoea batatas (Sweet potato) (Convolvulus batatas)	393	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEVRKAQRAEGPATILAIGTVTPANCVNQSTYPDYYFRITNSEHKTELKERFQRMCDKSMITKRYMHLTEEILKENPSFCEYMAPSLDARQDIAVVEVPKLGKEAAQSAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLIVCSEITVVTFRGPSETHLDSLVGQALFGDGAAAVIVGADPTPAEKPLFQLVSAAQNLAPDSCGAIDGHLREVGLTFHLLKDVPSVVSNNIEKCLFEAFNPLGISDWNSVFWIAHPGGPAILDQVEDKLGLKPEKLRATRHVLSEYGNMSSACVLFILDEMRKASSNAGLGTTGEGLEWGVLFGFGPGLTIETVVLHSVPIKPGPH
P48394	CHSB_IPOCO	reviewed	Chalcone synthase B (EC 2.3.1.74) (Naringenin-chalcone synthase B) (CHS-B) (Fragment)	CHSB	Ipomoea cordatotriloba (Tievine)	363	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MCTTVTVLTDTWSRREKRFEGHAKILAIGTATPANWVDQTTYPDFYFRITNSQHLLDHKEKFRRICNKSKIRKRHMILTEELLKKNPNLCTYNDASLNTRQDILVSEVPKLGKEAAMKAIKEWGRPISEITHLVFCTTSGVDMPGADFQLTKLLGLNSSVKRLMMYQQGCNAGAAMLRLAKDVAENNKGARVLVVCSEVMLSVFRGPSLQQEDNLLAQCLFGDGSAALIVGTDPRPGLETPLFELISAAQTIIPNTDSHLKLHVREMGLTFHCSKAVPTFITQNVEDCLVKAFEPYGISDWNSIFWVLHPGGNAIVDGVEETLGLAPEKLRASRDVLSGYGNLTSACVLFILDEVRKKSKKDE
P48404	CHSB_IPOTR	reviewed	Chalcone synthase B (EC 2.3.1.74) (Naringenin-chalcone synthase B) (CHS-B) (Fragment)	CHSB	Ipomoea triloba (Trilobed morning glory)	366	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MSTTVTVLTDTWGRRAKRFEIEGYAKILAIGTATPANWVDQTTYPDFYFRITNSQHLLEHKEKFRRICNKSKIRKRHLVLTEELLQKNPSLCTYNETSLNTRQDILVSEVPKLGKEAAMKAIKEWGRPISEITHLVFCTTSGVDMPGADFRLTKLLGLNSSVKRLMMYQQGCNAGAAMLRLAKDLAESNKGGRILVVCSEITINIFRGPSLEQDDNLLAQCLFGDGSAAMIVGTDPRPDLETPLFELVSSAQTIVPNTDSHLKLTLREMGLTFHCSRAVPSVLAENVEDCLVKAFEPYGISDWNSIFWVFHPGGNAIVDRVEERLGLGAQRFRASRDVLSEYGNLTSACVLFILDEVRNKSKKNEQ

Q9ZRS4	CHSY_CATRO	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Catharanthus roseus (Madagascar periwinkle) (Vinca rosea)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVNVEEIRNAQRAQGPATVLAIGTSTPSNCVDQSTYPDYYFRITNSEHKTELKEKFKRMCEKSMIKKRYMHLTEEILQENPNICAYMAPSLDARQNIVVVEVPKLGKEAAQKAIKEWGQSKSKITHLVFCTTSGVDMPGADYQLTKLLGLRSSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAAAIIVGSDPIPEIERPLFELVSAAQTLLPDSHGAIDGHLREVGLTFHLLKDVPGLISKNIGKALDEAFQPLGISDWNSIFWIAHPGGPAILDQVEEKLGLKPEKLRATRHVLSEYGNMSSACVLFILDEMRKASARDGLSTTGEGLEWGVLFGFGPGLTVETVVLHSVNV
O23731	CHS8_BROFI	reviewed	Chalcone synthase 8 (EC 2.3.1.74) (Naringenin-chalcone synthase 8)	CHS8	Bromheadia finlaysoniana (Orchid)	394	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MAPAMEEIRQAQRAEGPAAVLAIGTSTPPNALYQADYPDYYFRITKSEHLTELKEKFKRMCDKSMIKKRYMYLTEEILKENPNICAFMAPSLDARQDIVVTEVPKLAKEAAARAIKEWGHPKSRITHLIFCTTSGIDMPGADYQLTRLLGLRPSVNRFMLYQQGCFAGGTVLRLAKDLAENNAGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAAAIIVGSDPDSATERPLFQLVSASQTILPESEGAIDGHLREIGLTFHLLKDVPGLISKNIQKCLLDAFKPLGVHDWNSIFWIAHPGGPAILDQVEIKLGLKAEKLAASRNVLAEYGNMSSACVLFILDEMRRRSAEAGQATTGEGLEWGVLFGFGPGLTVETIVLRSVPIAGAE
P48402	CHSB_IPOTF	reviewed	Chalcone synthase B (EC 2.3.1.74) (Naringenin-chalcone synthase B) (CHS-B) (Fragment)	CHSB	Ipomoea trifida (Morning glory)	366	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MSTAVTMLTDTWSRRAKRFEIEGYAKILAIGTATPANWVDQTTYPDFYFRITNSQHLLEHKEKFRRICNKSNIRKRHMVLTEELLKKNPNLCTYNDASLNTRQDILVSEVPKLGKEAAMKAIKVWGRPISEITHLVFCTTSGVDMPGADFQLTKLLGLNSSVKRLMMYQQGCNAGAAMLRLAKDLAENNKGARVLVVCSEVMLSVFRGPSLQQEDNLLAQCLFGDGSAAIIIGTDPRPGLETPLFELISAAQTIIPDTDSHLKLHVREMGLTFHCSKAVPTFITQNVEDCLVKAFEPYGISDWNSIFWVLHPGGNAIVEGVEETLGLAPEKLRASRDVLSEYGNLTSACVLFILDEVRKKSKKDEQ
O65872	CHSY_PINST	reviewed	Chalcone synthase (EC 2.3.1.74) (Naringenin-chalcone synthase)	CHS	Pinus strobus (Eastern white pine)	395	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MPGGMMADLEAFRKAQRADGPATILAIGTATPPNAVDQSTYPDYYFKITNSEHMTELKEKFRRMCDKSGIKKRYMYLTEEILNENPSVCAYMAPSLDARQDMVVVEVPRLGKEAAAKAIKEWGQPKSKITHVIFCTTSGVDMPGADYQMTKLLGLRPSVKRVMMYQQGCFAGGTVLRVAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSMVGQALFGDGARALIVGADPVPEVEKPCFEMLWTAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFQQFGISDWNQLFWIAHPGGPAILDQVEAKLNLDPKKLRATRQVLSEYGNMSSACVHFILDEMRKSSQQNGCSTTGEGLDVGVLFGFGPGLTVETVVLKSVPLQ
O23883	CHS3_PEA	reviewed	Chalcone synthase 3 (EC 2.3.1.74) (Naringenin-chalcone synthase 3)	CHS3	Pisum sativum (Garden pea)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVSEIRKAQRAEGPATILAIGTANPANCVEQSTYPDFYFRITNSEHKTELKQKFQRMCDKSMINRRYMYLTEEILKENPSVCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPLPEIENPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPAIVSKNIDKALVEAFQPLGISDYNSIFWIAHPGGPAILDQVEQKLALKPEKMKATREVLSEYGNMSSACVLFILDEMRRKSIQNGLKTTGEGLEWGVLFGFGPGLTIETVVLHSVAI
Q9MB37	CHS7_IPOBA	reviewed	Chalcone synthase DIII (EC 2.3.1.74) (Naringenin-chalcone synthase DIII)	CHS-DIII	Ipomoea batatas (Sweet potato) (Convolvulus batatas)	393	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEVRKAKRAEGPATILAIGTATPANCVNQSTYPDYYFRITNSEHKTELKEKFQRMCDKSMITKRYMHLTEEILKENPSFCEYMAPSLDARQDIAVVEVPKLGKEAAQSAIKEWGQPKSKITHVFFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLIVCSEITVVTFRGPSETHLDSLVGQALFGDGAAAVIVGADPTPAEKPLFQLVSAAQTLAPDSCGAIDGHLREVGLTFHLLKDVPSVVSNNIEKCLFEAFNPLGISDWNSVFWIAHPGGPAILDQVEDKLGLKPEKLRATRHVLSEYGNMSSACVLFILDEMRKASSNAGLGTTGEGLEWGVLFGFGPGLTIETVVLHSVPIKPGPH
Q9SB26	CHS9_DAUCA	reviewed	Chalcone synthase 9 (EC 2.3.1.74) (Naringenin-chalcone synthase 9)	CHS9	Daucus carota (Wild carrot)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVNEFRKAQRAEGPATVLAIGTATPPNCVDQSAYADYYFRITNSEDKPELKEKFRRMCEKSMINTRYMHLTEDLLKQNPSFCEYMASSLDARQDIVVNEVPKLGKEAALRAIKEWGRPKSKITHLIFCTTSGVDMPGADFRLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNKNARVLVVCSEITVITFRGPNDTHLDSLVGQALFGDGAGAVIVGSDPVIGIEKPLFEIVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIRKSLVEAFKPLGISDWNSIFWIAHPGGPAILDQVETELSLKPDKLKSTRQVLRDYGNMSSACVLFILDEMRNASAKDGHRTTGEGLDWGVLFGFGPGLTVETVVLHSVPT
Q9MB38	CHS6_IPOBA	reviewed	Chalcone synthase DII (EC 2.3.1.74) (Naringenin-chalcone synthase DII)	CHS-DII	Ipomoea batatas (Sweet potato) (Convolvulus batatas)	393	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEVRKAQRAEGPATILAIGTATPANCVNQSTYPDYYFRITNSEHKTELKEKFQRMCDKSMITKRYMHLTEEILKENPSFCEYMAPSLDARQDIAVVEVPKLGKEAAQSAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLIVCSEITVVTFRGPSETHLDSLVGQALFGDGAAAVIVGADPTPAEKPLFQLVSAAQTLAPNSCGAIDGHLREVGLTFHLLKDVPSVVSNNIEKCLFEAFNPLGISDWNSVFWIAHPGGPAILDQVEDKLGLKPEKLRATRHVLSEYGNMSSACVLFILDEMRKASSNAGLGTTGEGLEWGVLFGFGPGLTIETVVLHSVLIKPGPH
Q9MB36	CHS8_IPOBA	reviewed	Chalcone synthase DIV (EC 2.3.1.74) (Naringenin-chalcone synthase DIV)	CHS-DIV	Ipomoea batatas (Sweet potato) (Convolvulus batatas)	388	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVEEVRKAQRAEGPATILAIGTVTPANCVNQSTYPDYYFRITNSEHKTELKEKFQRMCDKSMITKRYMHLTEEILKENPSFCEYMAPSLDARQDIAVVEVPKLGKEAAQSAIKEWGQPKSKITHVVFCTTSGIDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLIVCSEITVVTFRGPSETHLDSLVGQALFGDGAAAVIVGADPTPAEKPLFQLVSAAQTLAPNSCGAIDGHLREVGLTFHLLKDVPSVVSNNIEKCLFEAFNPLGISDWNSVFWIAHPGGPAILDQVEDKLGLKPEKLRATRHVLSEYGNMSSACVLFILDEMRKASSNAGLGTTGEGLEWGVLFGFGPGLTIETVVLHSVPT
P48403	CHSA_IPOTR	reviewed	Chalcone synthase A (EC 2.3.1.74) (Naringenin-chalcone synthase A) (CHS-A) (Fragment)	CHSA	Ipomoea triloba (Trilobed morning glory)	362	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MSPTVTVQLTDDTAKRFEGHAKLLAIGTATPTNWVDQATYPDFYFRITNSERLLEHKEKFRRICNKSKIRFRHLVLTEELLKKSPNLCTYNDASLNTRQDILVSEVPKLGKEAAMKAIKEWGRPISEITHLVFCTTSGVDMPGADFQLTKLLGLNSSVKRLMMYQQGCNAGAAMLRLVKDLAENNKGARVLVVCSEITINIFRGPSLEQDDNLLAQCLFGDGSAAMIVGKDPRPGLETPLFELVSSAQTIVPNTDSHLKLTLREMGLTFHCSRAVPSVLAENVEDCLVKAFEPYGISDWNSIFWVFHPGGYAIVDRVEERLGLGPERLRASRDVLSEYGNLTSACVLFILDEMRKKSKKDEQ
P48396	CHSB_IPONI	reviewed	Chalcone synthase B (EC 2.3.1.74) (Naringenin-chalcone synthase B) (CHS-B) (Fragment)	CHSB	Ipomoea nil (Japanese morning glory) (Pharbitis nil)	363	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MSTILNVLTDTWSPRAKKLEGDAKIWAIGTATPANWVDQTTYPDFYFRITNSQHLLEHKEKFRRICNKSKIRKRHLVLTEELLKKNPNLCTYNETSLNTRQDILVAEVPKLGKEAAMKAIKEWGRPISEITHLVFCTTSGVDMPGADFQLTKLLGLNSSVKRLMMYQQGCNAGAAMLRLAKDLAESNKGGRVLVVCAEITINIFRGPSLEQDDNLLAQCLFGDGAAAMIVAADPRPGLETPLFELVSSAQTIVPNTDSHLKLHLREMGLTFHCSKAVPSVLAENVEDCLVKAFEPYGISDWNSIFWVFHPGGNAIVDRVEERLGLGPEKFRASRDVLSEYGNLTSACVLFTLDEMRKKSKKDE
Q9ZU06	CHSY_PERAE	reviewed	Chalcone synthase (EC 2.3.1.74) (Naregenin-chalcone synthase)	CHS	Persea americana (Avocado)	392	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVNVEAIRKVQRAEGPATIMAIGTSTPPNAVDQSEYPDYYYFGSPTASTRPSSRRSFKRMCEKSMIKKRYMYLTETYWKRIQMFVPTWLLPLKARQDMVVVEVPKLGKEVQPKAIKGMGQPKSKINPLVFCTTSGVDMPGADYQLTKVFGLPPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAALIVGADPVPGVENPMFELVSAGQTILPDSDGAIDGHLREVGLTFHLLKVVPGLISKNIEKSLVEAFEPLGISDWNSLFWIAHPGGPAILDPGGDQTRPEARESCGNQACFSVSMATCQVFVCSSFSTRCEGSPKEEGLKTTGEGIEWGVLFGFGPGLTVETVVLHSLPTH
Q9XJ58	CHS1_CITSI	reviewed	Chalcone synthase 1 (EC 2.3.1.74) (Naringenin-chalcone synthase 1)	CHS1	Citrus sinensis (Sweet orange) (Citrus aurantium var. sinensis)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVDEVRKAQRAQGPATIMAIGTATPPNCVDQSTYPDYYFRITNSEHMTDLKEKFKRMCDKSMIKKRYMYLTEEILKENPNVCAYMAPSLDTRQDMVVVEVPRLGKAATEAIKEWGQPKSKITHLVFCTTSGVDMPGADYRLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAIIIGADPIPEIEKPMFELVSTAQTILPDSDGSIDGHLREAGLTFHLLKDVPGLISKNIQKSLTEAFKPLGISDWNSIFWIAHPGGPAILDQVEEKLGLKPEKLRATRHVLSEYGNMSSACVLFILDEMRKKSAEDGLETAGEGLEWGVLFGFGPGLTVETVVLHSVAAA
P51084	CHS2_TRISU	reviewed	Chalcone synthase 2 (EC 2.3.1.74) (Naringenin-chalcone synthase 2)	CHS2	Trifolium subterraneum (Subterranean clover)	389	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVSVSEIRKAQRAEGPATILAIGTANPANRVEQATYPDFYFKITNSEHKVELKEKFQRMCDKSMIKSRYMYLTEEILKENPSVCEYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPVPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNIDKALVEAFQPLNISDYNSIFWIAHPGGPAILDQVEQKLALKPEKMKATREVLSEYGNMSSACVLFILDEMRKKSAQNGLKTTGEGLEWGVLFGFGPGLTIETVVLHSVAI
Q9MB40	CHS3_IPOBA	reviewed	Chalcone synthase LF3 (EC 2.3.1.74) (Naringenin-chalcone synthase LF3)	CHS-LF3	Ipomoea batatas (Sweet potato) (Convolvulus batatas)	388	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MVTVGEVRKAQRAEGPATILAIGTATPANCVNQSTYPDYYFRITNSEHKTELKEKFQRMCDKSMITKRYMHLTEEILKENPSFCEYMAPSLDARQDIAVVEVPKLGKEAAQSAIKGWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLIVCSEITVVTFRGPSEAHLDSLVGQALFGDGAAAVIVGADPTPAEKPLFQLVSAAQTLAPDSCGAIDGHLREVGLTFHLLKDVPSVVSNNIEKCLFEAFNPLGISDWNSVFWIAHPGGPAILDQVEDKLGLKPEKLRATRHVLSEYGNMSSACVLFILDEMRKASSNDGLGTTGEGLEWGVLFGFGPGLTIETVVLHSVPT
P48395	CHSA_IPONI	reviewed	Chalcone synthase A (EC 2.3.1.74) (Naringenin-chalcone synthase A) (CHS-A) (Fragment)	CHSA	Ipomoea nil (Japanese morning glory) (Pharbitis nil)	344	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MSTILTNTWTRREKRIEGHAKILAIGTAIPANWVDQTTYPDFYFRITNSEHLLEHKEKFRRICNKSKIRKRHLVITEELLKKNPNLCTYNEASLNTRQDILVSEVPKLGKEAAMKAIKEWGRPISEITHLVFCTSSGVDMPGADFQLAKLLGLSSSVNRLMMYQQGCNAGAAMLRLAKDLAENNKGGRVLVVCSEVMLNVFRGPSLEQEDYLLAQCLFGDGSAAVIVGTEPRPGLETPLFELVSAAQTTIPDTDSHLKLHLREMGLTFHCSKAVPSLITQNVEDYLVKAFEPFGISDWNSIFWILHPGGIAILDRVEEKLGLEPEKLRASRDVLSESGNLTSAC
P48401	CHSA_IPOTF	reviewed	Chalcone synthase A (EC 2.3.1.74) (Naringenin-chalcone synthase A) (CHS-A) (Fragment)	CHSA	Ipomoea trifida (Morning glory)	362	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MSPTATVQLTDDTAKRFEGHAKLLAIGTATPTNWVDQATYPDFYFRITNSEHLLEHKEKFRRICNKSKIRKRHLVLTKELLKKNPNLCTYNDASLNTRQDILVSEVPKLGKEAAMKAIKEWGRPISEITHLVFCTTSGVDMPGADFQLTKLLGLNSSVKRLMMYQQGCNAGAAMLRLVKDLAENNKGARVLVVCSEITINIFRGPSLEQDDNLLAQCLFGDGSAAMIVGKDPRPGLETPLFELVSSAQTIVPNTDSHLKLHLREMGLTFHCSRAVPSVLAENVEDCLVKAFEPYGISDWNSIFWVFHPGGNAIVDRVEERLGLGPERLRASRDVLSEYGNLTSACVLFILDEMRKKSKKDEQ
P48393	CHSA_IPOCO	reviewed	Chalcone synthase A (EC 2.3.1.74) (Naringenin-chalcone synthase A) (CHS-A) (Fragment)	CHSA	Ipomoea cordatotriloba (Tievine)	361	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MTPTVTVQLTDDTAKRFEGHAKLLAIGTATPTNWVDQATYPDFYFRITNSEHLLEHKEKFRRICNKSKIRKRHLVLTEELLKENPNLCTYNDASLNTRQDILVSEVPKLGKEAAMKAIKEWGRPISEITHLVFCTTSGVDMPGADFQLTKLLGLNSSVKRLMMYQQGCNAGAAMLRLVKGLAENNKGARVLVVCSGITINIFRGPSLEQDDNLLAQCLFGDGSAAMIDGKDPRPGLETPLFELVSSAQTIVPNTDSHLKLHLREMGLTFHCSRAVPSVLAENVEDCLVKAFEPYGISDWNSIFWVFHPGGNAIVDRVEERSGLGPERLRASRDVLSEYGNLTSACVLFILDEMRKKSKKDE
P48400	CHSA_IPOPL	reviewed	Chalcone synthase A (EC 2.3.1.74) (Naringenin-chalcone synthase A) (CHS-A) (Fragment)	CHSA	Ipomoea platensis (Morning glory)	362	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MSTTVLPDTWSRRAKRFEGHAKILAVGTATPANWVDQTTYPDFYFRITNSEHLLEHKEKFRRICNKSKIRKRHLVLTEELLKKNPNLCTYNETSLNTRQDTLVSEVPKLGKEAAMKAIKEWGRPISEITHLVFCTTSGVDMPGADFQLTKLLGLNISVKRLMMYQQGCNAGAAMLRLAKDLAENNKGARVLVVCSEVTLSVFRGPSLQQEDNLLAQCLFGDGSAAVIVGTDPRPGLETPLFELVSSAQTIIPDTDSHLKLHLLEMGLTFHCSKAVPSLITQNVEDCLVKAFEPFGISDWNSIFWILHPGGNAILDRVEERLGLGPEKLRASRDVLSEYGNLTSACVLFILDLVRRKSKKQQQ
P30078	CHSY_MALDO	reviewed	Chalcone synthase (EC 2.3.1.74) (Naregenin-chalcone synthase) (Fragment)	CHS	Malus domestica (Apple) (Pyrus malus)	232	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIIGADPVPEVEKPLFELVSAAQTVLPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLNEALKPIGISDWNSLFWIAHPGGPAILDQVEAKLALKPEKLEATRQVLSDYGNMSSACVLFILDEVRRKSAEKGLETTGEGLEWGVLFGFGPGLTVETVVLHSVAA
P48391	CHS3_GERHY	reviewed	Chalcone synthase 3 (EC 2.3.1.74) (Naringenin-chalcone synthase 3)	CHS3	Gerbera hybrida (Daisy)	403	PATHWAY: Secondary metabolite biosynthesis; flavonoid biosynthesis.	CATALYTIC ACTIVITY: 3 malonyl-CoA + 4-coumaroyl-CoA = 4 CoA + naringenin chalcone + 3 CO(2). {ECO:0000255|PROSITE-ProRule:PRU10023}.	MATSPAVIDVETIRKAQRAEGPATILAIGTATPANCVYQADYPDYYFRVTESEHMVDLKEKFQRMCDKSMIRKRYMHITEEFLKENPSMCKFMAPSLDARQDLVVVEVPKLGKEAATKAIKEWGFPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPNEGHLDSLVGQALFGDGAAAVIIGSDPDLSVERPLFEMVSAAQTILPDSEGAIDGHLKEVGLTFHLLKDVPALIAKNIEKALIQAFSPLNINDWNSIFWIAHPGGPAILDQVEFKLGLREEKLRASRHVLSEYGNMSSACVLFILDEMRKKSIKDGKTTTGEGLEWGVLFGFGPGLTVETVVLHSLPATISVATQN

Unreviewed CHS Raw Dataset--UCHS

Entry	Entry name	Status	Protein names	Gene names	Organism	Length	Pathway	Catalytic activity	Sequence
G7KXB8	G7KXB8_MEDTR	unreviewed	Chalcone synthase protein	11440049 MTR_7g084300	Medicago truncatula (Barrel medic) (Medicago tribuloides)	391			MVTVEEIRKAQRSNGPATILAFGTATPSHCVTQAEYPDYYFRITNSEHMTDLKEKFKRMCEKSMIKKRYMHITEEFLKENPNMCAYMAPSLDARQDLVVVEVPKLGKDAAKKAIAEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLKPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKNARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAMIIGADPDLTVERPIFEIVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGIISKNIEKSLVEAFAPIGISDWNSIFWVAHPGGPAILDQVEEKLRLKEEKLRSTRHVLSEYGNMSSACVLFILDEMRKRSKEEGKITTGEGLEWGVLFGFGPGLTVETVVLHSVPVQG
Q64HV0	Q64HV0_ARALY	unreviewed	Chalcone synthase protein	CHS	Arabidopsis lyrata (Lyre-leaved rock-cress) (Arabis lyrata)	375			MVMAAGASSLDEIRKAQRADGPAGILAIGTANPENHVLQAEYPDYYFRITNSEHMTDLKEKFKRMCDKSMIRKRHMHLTEDFLKENPHMCAYMAPSLDTRQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRIAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDTSVGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIVKSLDEAFKPLGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKMRMTRHVLSEYGNMSSACVLFILDEMRRKSAKDGVATTGGGLEWG
D0E301	D0E301_ARALL	unreviewed	Chalcone synthase protein	CHS	Arabidopsis lyrata subsp. lyrata (Lyre-leaved rock-cress)	367			MVMAAGASSLDEIRKAQRADGPAGILAIGTANPENHVLQAEYPDYYFRITNSEHMTDLKEKFKRMCDKSMIRKRHMHLTEDFLKENPHMCAYMAPSLDTRQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRIAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDTSVGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIVKSLDEAFKPLGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKMRMTRHVLSEYGNMSSACVLFILDEMRRKSAKDGVAT
D0E303	D0E303_ARALP	unreviewed	Chalcone synthase protein	CHS	Arabidopsis lyrata subsp. petraea (Northern rock-cress) (Cardaminopsis petraea)	367			MVMAAGASSLDEIRKAQRADGPAGILAIGTANPENHVLQAEYPDYYFRITNSEHMTDLKEKFKRMCDKSMIRKRHMHLTEDFLKENPHMCAYMAPSLDTRQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRIAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDTSVGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIVKSLDEAFKPLGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKMRMTRHVLSEYGNMSSACVLFILDEMRRKSAKDGVAT
Q705Q9	Q705Q9_ARALL	unreviewed	Chalcone synthase protein	chs	Arabidopsis lyrata subsp. lyrata (Lyre-leaved rock-cress)	391			MVMAAGASSLDEIRKAQRADGPAGILAIGTANPENHVLQAEYPDYYFRITNSEHMTDLKEKFKRSDKSMIRKRHMHLTEDFLKENPHMCAYMAPSLDTRQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRIAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDTSVGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIVKSLDEAFKPLGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKMRMTRHVLSEYGNMSSACVLFILDEMRRKSAKDGVATTGEGLEWGVLFGFGPGLTVETVVLH
Q705Q6	Q705Q6_ARALP	unreviewed	Chalcone synthase protein	chs	Arabidopsis lyrata subsp. petraea (Northern rock-cress) (Cardaminopsis petraea)	391			MVMAAGASSLDEIRKAQRADGPAGILAIGTANPENHVLQAEYPDYYFRITNSEHMTDLKEKFKRSDKSMIRKRHMHLTEDFLKENPHMCAYMAPSLDTRQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRIAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDTSVGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIVKSLDEAFKPLGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKMRMTRHVLSEYGNMSSACVLFILDEMRRKSAKDGVATTGEGLEWGVLFGFGPGLTVETVVLH
Q705Q4	Q705Q4_ARALP	unreviewed	Chalcone synthase protein	chs	Arabidopsis lyrata subsp. petraea (Northern rock-cress) (Cardaminopsis petraea)	391			MVMAAGASSLDEIRKAQRADGPAGILAIGTANPENHVLQAEYPDYYFRITNSEHMTDLKEKFKRSDKSMIRKRHMHLTEDFLKENPHMCAYMAPSLDTRQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRIAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDTSIGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIVKSLDEAFKPLGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKMRMTRHVLSEYGNMSSACVLFILDEMRRKSAKDGVATTGEGLEWGVLFGFGPGLTVETVVLH
D6C6K8	D6C6K8_NELNU	unreviewed	Chalcone synthase (chalcone synthase 1) (EC 2.3.1.74)	CHSC CHS3 CHS5 CHS7 CHSB CHSD CHSE CHSF CHSG LOC104602160	Nelumbo nucifera (Sacred lotus)	389			MVTVEDIRKAQRAEGPATVMAIGTANPPNCVDQSTYPDYYFRITNSEHKTELKEKFKRMCEKSMIKKRYMHLTEEILKENPNICEYMASSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSELTAVTFRGPSDTHLDSLVGQALFGDGAAAVIVGADPVPGVEKPLFELVSAAQTILPDSHGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFQPLGISDWNSIFWIAHPGGPAILDQVEEKLALKPEKLSATRHILSEYGNMSSACVLFILDEMRKKSIEDGLKTTGEGLEWGVLFGFGPGLTVETVVLHSIAA
Q9SEN1	Q9SEN1_ARALL	unreviewed	Chalcone synthase	ARALYDRAFT_488219	Arabidopsis lyrata subsp. lyrata (Lyre-leaved rock-cress)	396			MVMAAGASSLDEIRKAQRADGPAGILAIGTANPENHVLQAEYPDYYFRITNSEHMTDLKEKFKRMCDKSMIRKRHMHLTEDFLKENPHMCAYMAPSLDTRQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRIAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDTSVGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIVKSLDEAFKPLGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKMRMTRHVLSEYGNMSSACVLFILDEMRRKSAKDGVATTGEGLEWGVLFGFGPGLTVETVVLHSVPL
Q0H703	Q0H703_ARALL	unreviewed	Chalcone synthase (Chalcone synthase family protein) protein		Arabidopsis lyrata subsp. lyrata (Lyre-leaved rock-cress)	389			MVMAAGASSLDEIRKAQRADGPAGILAIGTANPENHVLQAEYPDYYFRITNSEHMTDLKEKFKRMCDKSMIRKRHMHLTEDFLKENPHMCAYMAPSLDTRQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRIAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDTSVGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIVKSLDEAFKPLGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKMRMTRHVLSEYGNMSSACVLFILDEMRRKSAKDGVATTGEGLEWGVLFGFGPGLTVETV
Q9SBU8	Q9SBU8_ARALP	unreviewed	Chalcone synthase		Arabidopsis lyrata subsp. petraea (Northern rock-cress) (Cardaminopsis petraea)	396			MVMAAGASSLDEIRKAQRADGPAGILAIGTANPENHVLQAEYPDYYFRITNSEHMTDLKEKFKRMCDKSMIRKRHMHLTEDFLKENPHMCAYMAPSLDTRQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRIAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDTSIGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIVKSLDEAFKPLGISDWNSLFWIAHPEGPAILDQVELKLGLKEEKMRMTRHVLSEYGNMSSACVLFILDEMRRKSAKDGVATTGEGLEWGVLFGFGPGLTVETVVLHSVPL
Q9SBU7	Q9SBU7_ARALP	unreviewed	Chalcone synthase		Arabidopsis lyrata subsp. petraea (Northern rock-cress) (Cardaminopsis petraea)	396			MVMAAGASSLDEIRKAQRADGPAGILAIGTANPENHVLQAEYPDYYFRITNSEHMTDLKEKFKRMCDKSMIRKRHMHLTEDFLKENPRMCAYMAPSLDTRQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRIAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDTSVGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIVKSLDEAFKPLGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKMRMTRHVLSEYGNMSSACVLFILDEMRRKSAKDGVATTGEGLEWGVLFGFGPGLILETVVLHSVPL
A0A1Z1N350	A0A1Z1N350_PRUP	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Prunus persica (Peach) (Amygdalus persica)	343			MVLVILIWQFTGKTTTVKTRYVVMSDEILEKYPELTTEGTPTIKQRLHICNEAVTQMAIEASGACIKNWGRPISDITHLVYVSSSEARLPGGDIYLAKGLGLRPETQRVLLYFSGCSGGVAGLRVAKDIAENNPGSRVLLATSETTIIGYKPPSAHRPYDLVGVALFGDGAGAMLIGSDPDLISEKPLFELHTAIQEFLPDTEKTIDGRVTEEGISFKLGRELPQIIEDHIEGFCGRLMGVLGYDNKEYNKMFWAVHPGGPAILNRLEKRLDLFPEKLNASRRALTDYGNASSNTIVYVLEYMIEESKKIKKEQQEGDGEWGLILAFGPGITFEGILARNLAV
Q6X0M9	Q6X0M9_SOYBN	unreviewed	Chalcone synthase (EC 2.3.1.74) (Chalcone synthase CHS3)	100791524 CHS GLYMA_08G109300 GLYMA_08G110300 GLYMA_08G110900	Glycine max (Soybean) (Glycine hispida)	388			MVSVEEIRNAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSEHMTELKEKFKRMCDKSMIKKRYMYLNEEILKENPSVCAYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPTDTHLDSLVGQALFGDGAAAVIVGSDPLPVEKPLFQLVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKALVEAFQPLGISDYNSIFWIAHPGGPAILDQVEAKLGLKPEKMEATRHVLSEYGNMSSACVLFILDQMRKKSIENGLGTTGEGLDWGVLFGFGPGLTVETVVLRSVTV

D1MIB1	D1MIB1_9CARY	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Fagopyrum dibotrys	395			MAPTVQEIRKAQRAEGPATVLAIGTATPPNCVYQADYPDYYFRVTNSDHMTDLKEKFRRMCDKSQIEKRYMYLTEDILKEHPNMCEYMAPSLDSRQDMVVTEVPKLGKEAAQKAIKEWGQPKSKITHVIVCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRMAKDLAENNRGARVLVVCSEITAVCFRGPTDTHLDSMVGQALFGDGAGAVIVGADPDLSIEKPIFELVWTSQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLTEAFSPLNIADWNSLFWIAHPGGPAILDQVEAKLGLKEEKLKATRQVLNDYGNMSSACVLFILDEMRKKSLENGHATTGEGLDWGVLFGFGPGLTVETVVLHSVPTTTLAN
Q93V86	Q93V86_HUMLU	unreviewed	Chalcone synthase (EC 2.3.1.74)	chs2	Humulus lupulus (European hop)	394			MTSMTVDQIRRPLRAEGLATILAIGTANPANYITQADYPDYYFRVTKSEHMTDLKNKFQRMCDRSMIKKRHMYFTEEHLKQNPNMCDYTAPSLDARQSILVVEVPKLGKEACIKAIKEWNQPKSKITHFIFTTISGIDMPGADYQCSKLLGLNPSVKRVMLYNLGCHASGTILRMAKDIAENNKGARVLAVCSDIMTGHFRGPAESHLDSMIGQALFGDGASAIIVGAEPDESAGEQPIFELVSTAQTTLPDSDGVARGHLKEAGVVIHLHQSLPGLISTNIEKSLTEAFAPIGISDWNSIFWITHPAGRAVLEEIEAKLQLKNEKLADSRHVLSEYGNMSSACVFFIMDKLRKRSLEQRKSTTGDGLEWGVLFGFGPGLTVETVVLHSVANKF
D1MJ10	D1MJ10_FAGTA	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Fagopyrum tataricum (Tartarian buckwheat) (Polygonum tataricum)	395			MAPTVQEIRKAQRAEGPATVLAIGTATPPNCVYQADYPDYYFRVTNSDHMTDLKEKFRRMCDKSQIEKRYMYLTEEILKEHPNMCEYMAPSLDSRQDMVVTEVPKLGKEAAQKAIKEWGQPKSKITHVIVCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRMAKDLAENNRGARVLVVCSEITAVCFRGPTDTHLDSMVGQALFGDGAGAVIVGADPDLSIEKPIFELVWTSQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLTEAFSPLNIADWNSLFWIAHPGGPAILDQVEAKLGLKEEKLKATRQVLNDYGNMSSACVLFILDEMRKKSLENGHATTGEGLDWGVLFGFGPGLTVETVVLHSVPTTTLAN
A0A0A0P7Q3	A0A0A0P7Q3_RHEA	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS2	Rheum australe (Himalayan rhubarb) (Rheum emodi)	392			MAPTVQEIRKAQRAEGPATVLAIGTATPANCIYQADYPDYYFRVTNSEHMTDLKEKFRRMCDKSMIEKRYMHLTEEILKENQNMCEYMAPSLDSRQDMVVSEVPRLGKEAAQKAIKEWGQPKSKITHVIMCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENTRGARVLVVCSEITAICFRGATDTHLDSMVGQALFGDGCRSRHHWIRTRRPLRGQSSSWFGRPRPSYPTPKAPLTATYVRLGSPFISSRTFPALISKNIHKSLAGGFSLPSIFTDWNSLFWHCSGGPAILDQVEAKLGLKEEKLKATRQVLNDYGNMSSACVLFIMNEMRKKSLENGHATTGEGLDWGVLFGFGPGLTVETVVLHSVPVAH
D1MJ11	D1MJ11_FAGES	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Fagopyrum esculentum (Common buckwheat) (Polygonum fagopyrum)	392			MAPTVQEIRKAQRAEGPATVLAIGTATPPNCVYQADYPDYYFRVTNSDHMTDLKEKFRRMCDKSQIEKRYMYLTEEILKEHPNMCEYMAPSLDSRQDMVVTEVPKLGKEAAQKAIKEWGQPKSKITHVIVCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRMAKDLAENNRGARVLVVCSEITAVCFRGPTDTHLDSMVGQALFGDGAGAVIVGADPDLSVEKPIFELVWTSQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLTEAFSPLNIADWNSLFWIAHPGGPAILDQVEAKLGLKEEKLKATRQVLNDYGNMSSACVLFILDEMRKKSLENGHATTGEGLDWGVLFGFGPGLTVETVVLHSVPVAN
D7KP48	D7KP48_ARALL	unreviewed	Chalcone synthase family protein	ARALYDRAFT_470071	Arabidopsis lyrata subsp. lyrata (Lyre-leaved rock-cress)	396			MSNSRMNGVEKLSSISTRRVANPGKATLLALGKAFPSQVVPQENLVEGFLRDTKCEDAFIKEKLEHLCKTTTVKTRYTVLSREILDKYPELTTEGSPTIKQRLEIANEAVVEMALEASLGCIKEWGRPVEDITHIVYVSSSEIRLPGGDLYLSAKLGLRNDVNRVMLYFLGCYGGVTGLRVAKDIAENNPGSRVLLTTSETTILGFRPPNKARPYDLVGAALFGDGAAAVIIGADPRECEAPFMELHYAVQQFLPGTQNVIDGRLTEEGINFKLGRDLPQKIEENIEEFCKKLMGKAGGDESMEFNDMFWAVHPGGPAILNRLETKLKLEKEKLESSRRALVDYGNVSSNTILYVMEYMREELKKKGDEAQEWGLGLAFGPGITFEGLLIRSLTSP
A0A1S3EG19	A0A1S3EG19_CICA	unreviewed	LOW QUALITY PROTEIN: chalcone synthase 6-4-like	LOC105852659	Cicer arietinum (Chickpea) (Garbanzo)	374			MVSVSEIRKAQRAEGPATILAIGTANPSNRVEQSTYPDFYFKITNSEHKVELKQKFQRMCKYIIFHYLVLYEHWFLYYSKFVLLILFYYKXKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEVTAVTFRGPSDTHLDSLVGQALFGDGAAALIVGSDPIPEIEKPIFEMVWTAQTIAPDSEGAIDGHLREAGLTFHLLKDVPGIVSKNIDKALIEAFQPLNISDYNSIFWIAHPGGPAILDQVEEKLALKPEKMRATREVLSEYGNMSSACVLFILDEMRRKSAKDGLKTTGEGLEWGVLFGFGPGLTIETVVLHSVAI
A0A0S1M144	A0A0S1M144_HELA	unreviewed	Chalcone synthase 2 (Putative chalcone synthase)	CHS CHSY HannXRQ_Chr14g0441041	Helianthus annuus (Common sunflower)	398			MASSIDIAAFREAQRAQGPATILAIGTATPSNCVYQADYPDYYFRITKSEHMVDLKEKFKRMCDKSMIRKRYMHLTEEYLKENPSICEYMAPSLDARQDVVVVEVPKLGKEAAVKAIKEWGKPKSQITHLIVCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRMAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSMVGQALFGDGAAALVVGSDPDLTTERPLFQMVSAAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKALTQAFSPLGISDWNSIFWIAHPGGPAILDQVEQKLKLKEEKMRATRHVLSEYGNMSSACVLFILDEMRKKSVEDGAATTGEGLDWGVLFGFGPGLTVETVVLHSVPTTMPIAP
Q9M5B2	Q9M5B2_PETHY	unreviewed	Chalcone synthase (EC 2.3.1.74)	chs CHS	Petunia hybrida (Petunia)	389			MVTVEEYRKAQRAEGPATVMAIGTATPTNCVDQSTYPDYYFRITNSEHKTDLKEKFKRMCEKSMIKKRYMHLTEEILKENPSMCEYMAPSLDARQDIVVVEVPKLGKEAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGCDYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPNDTHLDSLVGQALFGDGAGAIIIGSDPIPGVERPLFELVSAAQTLLPDSHGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLEEAFRPLSISDWNSLFWIAHPGGPAILDQVEIKLGLKPEKLKATRNVLSNYGNMSSACVLFILDEMRKASAKEGLGTTGEGLEWGVLFGFGPGLTVETVVLHSVAT
H6UKR4	H6UKR4_VACCO	unreviewed	Chalcone synthase	chs	Vaccinium corymbosum (Highbush blueberry)	419			MVTVEEVRTAMRAEGPATVLAIGTATPAHCVEQATYPDYYFRVTNSEHKAELKEKFQRMCDKSQIKKRYMYLTEEILKENPNVCAYMAPSLDARQDMVVVEIPKLGKEAAVKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDSHLDSLVGQALFGDGAAAIIVGADPIPEVEKPLFEVVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFQPLGIPDWNSIFWIAHPGGPAILDQVEQKLALKPEKLRATRHVLSEYGNMSSACVLFTMDELRKSRPKMGSRPPVKGSTGACSLGSGLGLPLRPQCSIVCALSGLVGLWVLFMAFLLRLWSLLWTGLCLIPL
B4G105	B4G105_MAIZE	unreviewed	Chalcone synthase (Type III polyketide synthase B)	100283134 ZEAMMB73_Zm00001d019478	Zea mays (Maize)	417			MVSSSMDTTSDKRASSMLAPNPGKATILALGHAFPQQLVMQDYVVDGFMKNTNCDDPELKEKLTRLCKTTTVRTRYVVMSDEILKNYPELAQEGLPTMNQRLDISNAAVTQMATEASLSCVRSWGGALSSITHLVYVSSSEARFPGGDLHLARALGLSPDVRRVMLAFTGCSGGVAGLRVAKGLAESCPGARVLLATSETTIVGFRPPSPDRPYDLVGVALFGDGAGAAVIGTDPAPAERPLFELHSALQRFLPDTERTIEGRLTEEGIKFQLGRELPHIIEAHVEDFCQKLMKERQSGEDADGGGPEPMSYGDMFWAVHPGGPAILTKMEGRLGLGADKLRASRCALRDFGNASSNTIVYVLENMVEDTRRRRRLLAADDGGEDCEWGLILAFGPGITFEGILARNLQATARASAQL
A0A072ULL9	A0A072ULL9_MEDT	unreviewed	Chalcone synthase family protein	25492852 MTR_4g075560	Medicago truncatula (Barrel medic) (Medicago tribuloides)	390			MGDEGIVRGVTKQVTAGKATILALGKAFPHQLVMQEYLVDGYFRDTNCDNPELKQKLARLCKTTTVKTRYVVMNEEILKKYPELAVEGASTVKQRLEICNEAVTQMAIEASQVCIKNWGGSLSDITHVVYVSSSEARLPGGDLYLSKGLGLNPTIRRTMLYFSGCSGGVAGLRVAKDIAENNPGSRVLLATSETTIIGFKPPSADRPYDLVGVALFGDGAGAMIIGSDPKLETEMPLFELHTSAQEFIPDTEKKIDGRLTEEGISFTLARELPQIIEDNVEGFCDRLMDAVELENKEYNKLFWAVHPGGPAILNRVEKRLELSPEKLNASRKALMDYGNASSNTIIYVLEYMLDESKKIRKEGGYPEWGLILAFGPGITFEGILARNLCP
A2ICC5	A2ICC5_VITVI	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS VIT_14s0068g00930	Vitis vinifera (Grape)	393			MVSVAEIRKAQRAEGPATVLAIGTATPANCVYQADYPDYYFRITNSEHMTELKEKFKRMCEKSMINKRYMHLTEEILKENPNVCAYMAPSLDARQDMVVVEVPKLGKEAAAKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLKPSVKRLMMYQQGCFAGGTVLRLAKDLAENNAGSRVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIIGADPDTKIERPLFELVSAAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFTPIGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKLRATRHVLSEYGNMSSACVLFILDEMRKKSIEEGKGSTGEGLEWGVLFGFGPGLTVETVVLHSVSAPAAH
Q2ENC2	Q2ENC2_POPTR	unreviewed	Chalcone synthase (Chalcone synthase family protein)	CHS POPTR_0014s14200g POPTR_014G145100v3	Populus trichocarpa (Western balsam poplar) (Populus balsamifera subsp. trichocarpa)	396			MAPSIEEIRKAQRASGPATILAIGKATPANCVSQADYPDYYFRITNSEHMTELKEKFKRMCDKSMIKKRYMHLTEEILKENSSMCEYMAPSLDARQDMVVVEVPKLGKEAAAKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRSSVKRFMMYQQGCFAGGTVLRLAKDLAENNKGSRVLVVCSEITAVTFRGPSDTHLDSMVGQALFGDGAAAVIVGADPDTSIERPLFQIVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFAPIGINDWNSIFWIAHPGGPAILDQVEIKLDLKEEKLRATRNVLSDYGNMSSACVLFILDEMRNKSLEEGKSTTGEGLEWGVLFGFGPGLTVETVVLHSVPVEQTIYS
A3E7Z7	A3E7Z7_9TRAC	unreviewed	Chalcone synthase-like polyketide synthase	Pks1	Huperzia serrata	399			MTIKGSGSAAFEGTRLCPRVIKPDGPATILAIGTSNPTNIFEQSTYPDFFFDVTNCNDKTELKKKFQRICDKSGIKKRHFHLTDEILRKNPSICKFKEASLDPRQDIAVLEVPKLAKEAAISAIKQWGQPKSKITHLVFATTSGVDMPGADFQLAKLLGLRPTVKRVMLYQQGCYAGATVLRVAKDLAENNKGARVLVACSEVTAVTFRAPSETHLDGLVGSALFGDGAAALIVGSDPVPQEEKPLFEIHWAGEAVLPDSDGAINGHLREAGLIFHLLKDVPGLISKNIDKVLAEPLEYVHFPSYNDMFWAVHPGGPAILDQIEAKLGLSTDKMQASRDVLASYGNMSSASVLFVLDQIRKNSEELHLPTTGEGFEWGFVIGFGPGLTVETLLLRSINI
Q8W3P6	Q8W3P6_VITVI	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS3 VIT_05s0136g00260 VITISV_030544	Vitis vinifera (Grape)	389			MVTVNEVRNAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSEHKTELKEKFKRMCDKSMIKKRYMHLTEEILKENPNVCEYMAASLDARQDMVVVEVPKLGKEAAAKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIVGSDPIPGVEKPMFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLNEAFQPLGIKDWNSIFWIAHPGGPAILDQVEEKLALKPEKLRSTRHVLSEYGNMSSACVLFILDEMRRKSAEEGLKTTGEGLEWGVLFGFGPGLTVETVVLHSVST
B3F5J6	B3F5J6_SOYBN	unreviewed	Chalcone synthase 9	CHS9 100790997 GLYMA_08G109500	Glycine max (Soybean) (Glycine hispida)	388			MVSVEAIRKAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSEHMTELKEKFKRMCDKSMIKKLYMYLNEEILKENPSVCAYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPTDTHLDSLVGQALFGDGAAAVIVGSDPLPVEKPLFQLIWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKALVEAFQPLGISDYNSIFWIAHPGGPAILDQVEAKLGLKPEKMEATRHVLSEYGNMSSACVLFILDQMRKKSIENGLGTTGEGLDWGVLFGFGPGLTVETVVLRSVTV
H9DV62	H9DV62_POLCS	unreviewed	Chalcone synthase (EC 2.3.1.74)		Polygonum cuspidatum (Japanese knotweed)	393			MAPSVQEIRKAQRAEGPATVLAIGTATPPNCIYQADYPDYYFRVTNSEHMTDLKEKFRRMCDKSMIEKRYMHLTEEILKENQNMCAYMASSLDSRQDMVVSEVPRLGKEAAQKAIKEWGQPKSKITHVIMCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTALRLAKDLAENTKGARVLVVCSEITAICFRGPTDTHLDSMVGQALFGDGAGAVIIGADPDLSIERPIFELVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLTEAFSPLNISDWNSLFWIAHPGGPAILDQVEAKLGLKEEKLKATRQVLNDYGNMSSACVLFIMDEMRKKSLENGHATTGEGLDWGVLFGFGPGLTVETVVLHSVPVAHH
Q6X0M8	Q6X0M8_SOYBN	unreviewed	Chalcone synthase (EC 2.3.1.74) (Chalcone synthase CHS1)	ICHS1 GLYMA_08G109400	Glycine max (Soybean) (Glycine hispida)	388			MVSVEEIRKAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSEHMTELKEKFKRMCDKSMIKKRYMYLNEEILKENPSVCAYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPTDTHLDSLVGQALFGDGAAAVIVGSDPLPVEKPLFQLVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKALVEAFQPLGISDYNSIFWIAHPGGPAILDQVEAKLGLKPEKMEATRHVLSEYGNMSSACVLFILDQMRKKSIENGLGTTGEGLDWGVLFGFGPGLTVETVVLRSVTL
A0A1P8B585	A0A1P8B585_ARAT	unreviewed	Chalcone and stilbene synthase family protein	LAP5 LESS ADHESIVE POLLEN 5 polyketide synthase B PSKB At4g34850 F11I11.90 F11I11_90	Arabidopsis thaliana (Mouse-ear cress)	319			MSEEILKKYPELAIEGGSTVTQRLDICNDAVTEMAVEASRACIKNWGRSISDITHVVYVSSSEARLPGGDLYLAKGLGLSPDTHRVLLYFVGCSGGVAGLRVAKDIAENNPGSRVLLATSETTIIGFKPPSVDRPYDLVGVALFGDGAGAMIIGSDPDPICEKPLFELHTAIQNFLPETEKTIDGRLTEQGINFKLSRELPQIIEDNVENFCKKLIGKAGLAHKNYNQMFWAVHPGGPAILNRIEKRLNLSPEKLSPSRRALMDYGNASSNSIVYVLEYMLEESKKVRNMNEEENEWGLILAFGPGVTFEGIIARNLDV
O80407	O80407_VITVI	unreviewed	Chalcone synthase (EC 2.3.1.74)	gVvCHS1	Vitis vinifera (Grape)	393			MVSVAEIRKAQRAEGPATVLAIGTATPANCVYQADYPDYYFRITNSEHMTELKEKFKRMCEKSMINKRYMHLTEEILKENPNVCAYMAPSLDARQDMVVVEVPKLGKEAAVKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLKPSVKRLMMYQQGCFAGGTVLRLAKDLAENNAGSRVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIIGADPDTKIERPLFELVSAAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFTPIGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKLRATRHVLSEYGNMSSACVLFILDEMRKKSIEEGKASTGEGLEWGVLFGFGPGLTVETVVLHSVSAPPAH
Q8W3P5	Q8W3P5_VITVI	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS2	Vitis vinifera (Grape)	393			MVSVGEIRKSQRAEGPATVLAIGTATPANCVYQADYPDYYFRITNSEHMTELKEKFKRMCEKSMINKRYMHLTEEILKENPNVCAYMAPSLDARQDMVVVEVPKLGKEAAVKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLKPSVKRLMMYQQGCFAGGTVLRLAKDLAENNAGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAIIIGADPDTKIERPLFELVSAAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFKPIGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKLRATRHVLSEYGNMSSACVLFILDEMRKKSIEEGKGTTGEGLEWGVLFGFGPGLTVETVVLHSLATQSTH

A0A1S2XQV	A0A1S2XQV5_CICA	unreviewed	chalcone synthase	LOC101514543	Cicer arietinum (Chickpea) (Garbanzo)	391			MVTVEEIRNAQRSNGPATILAFGTATPSNCITQADYPDYYFRITNSEHMTDLKEKFKRMCEKSMIKKRYMHLTEEFLKENTNMCAYMAPSLDVRQDIVVVEVPKLGKEAAKKAISEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLKPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKNARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAMIIGADPDLTVERPIFEIVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGIISKNIEKSLVEAFSPIGITDWNSIFWVAHPGGPAILDQVEEKLRLKEDKLRSTRHVLSEYGNMSSACVLFILDELRKKSKEEGKNTTGDGFEWGVLFGFGPGLTVETVVLHSVPIQG
U5INM8	U5INM8_MALDO	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS-2	Malus domestica (Apple) (Pyrus malus)	389			MVTVEEVRKAQRAEGPATVLAIGTATPSNCVDQATYPDYYFRITNSEHKTELKEKFQRMCDKSMIKKRYMYLTEEILKENPTVCEYMAPSLDARQDMVVVEVPRLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIIGADPLPEVEKPLFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLNEAFKPIGISDWNSLFWIAHPGGPAILDQVESKLALKPEKLEATRQVLSNYGNMSSACVLFILDEVRRRKSTEKGLRTTGEGLEWGVLFGFGPGLTVETVVLHSVAA
A0A1S3BNX	A0A1S3BNX6_CUC	unreviewed	chalcone synthase 2	LOC103491706	Cucumis melo (Muskmelon)	400			MASVVSEIRKAQRADGPATVLAIGTATPPHSVLQSDYPDYYFRITKSEHMTQLKEKFSRMCEKSMIRKRHMYLTEEILRENPNMCEYMAPSLDARQDMVVVEVPKLGKEAAAKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLLKLLGLRPSVKRYMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPSETHLDSMVGQALFGDGAGAVIVGSDPDLSAERPLYELVWTGSTVLPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLKEAFTPLGISDWNSIFWIAHPGGPAILDQVEAKLSLKEEEKMRATREVLSEYGNMSSACVLFIMDQMRKNSMEEGNSTTGEGLEWGVLFGFGPGLTVETVVLHSVDIKESTVNVASY
A0A1U7Z9S2	A0A1U7Z9S2_NELN	unreviewed	chalcone synthase	LOC104591893	Nelumbo nucifera (Sacred lotus)	393			MVSVEEIRNAQRAQGPATVLAIGTATPANCVYQADYPDYYFRITNSEHKTELKEKFKRMCEKSMIKKRYMHLTEEILKENPNICAYMAPSLDARQDMVVVEVPKLGKEAASKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNAGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIVGADPDTSIERPLFQLVSAAQTILPDSHGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFTPIGVSDWNSLFWIAHPGGPAILDQVEEKLGLKEEKLRATRNVLSEYGNMSSACVLFILDEMRKKSVEEGRATTGEGLEWGVLFGFGPGLTVETVVLHSLPAVPAH
E7BLG7	E7BLG7_PAESU	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Paeonia suffruticosa (Tree peony) (Paeonia moutan)	394			MASVEEIRNAQRAQGPATILAIGTATPAHFINQAEYPDYYFRITNSEHKTELKEKFKRMCDKSMINKRYMYLTEEILKENPKMCEYMAPSLDARQDMVVVEIPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGSRVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIVGADPDVKIERPLFQIVSAGQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFKPIGINDWNSIFWIAHPGGPAILDQVELKLGLKEEKLKNTRHVLSEYGNMSSACVLFILDETRKRSLEEGKATTGEGLDWGVLFGFGPGLTVETVVLHSVPAITISE
Q45XN8	Q45XN8_9ASPA	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS5	Phalaenopsis hybrid cultivar	394			MAPAMEEIRRAQRAEGPAAVLAIGTSTPPNAVYQADYPDYYFRITNCEHLTDLKEKFKRMCEKSMIKKRYMYLTEEFLKENPNICAFMAPSLDARQDIVVAEVPKLAKEAAARAIKEWGHPKSRITHLIFCTTSGVDMPGADYQLTRLLGLRPSVNRFMLYQQGCFAGGTVLRLAKDLAENNAGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAAAIIVGSDPDLATERPLFQLVSASQTILPESEGAIDGHLREIGLTFHLLKDVPGLISKNIQKCLLEAFKPLGVLDWNSIFWIAHPGGPAILDQVETKLGLKSEKLAASRNVLAEYGNMSSACVLFILDEMRRRSAEAGQSTTGEGLEWGVLFGFGPGLTVEAVVLRSVPIGGTE
A8CLH4	A8CLH4_RHEPA	unreviewed	Chalcone synthase 2	CHS2	Rheum palmatum (Chinese rhubarb)	392			MAPTVQEIRKAQRAEGPATVLAIGTATPANCIYQADYPDYYFRVTNSEHMTDLKEKFRRMCDKSMIEKRYMHLTEEILKENQNMCEYMAPSLDSRQDMVVSEVPRLGKEAAQKAIKEWGQPKSKITHVIMCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENTRGARVLVVCSEITAICFRGPTDTHLDSMVGQALFGDGAGAVIIGSDPDLSIERPIFELVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIHKSLAEAFSPLNITDWNSLFWIAHPGGPAILDQVEAKLGLKEEKLKATRQVLNDYGNMSSACVLFIMDEMRKKSLENGHATTGEGLDWGVLFGFGPGLTVETVVLHSVPVAH
Q43041	Q43041_PETHY	unreviewed	Chalcone synthase (EC 2.3.1.74) protein	chsH	Petunia hybrida (Petunia)	330			DKSMIKKRYMHLTEEILKEDPNICECMAPSLDARQDIVVVEVPKLGKEAAEKAMEEWGQHKSRITHLVFCTTTGVGLPGADFQLTQLLGLGSSVKRFMMNQLGCFAGGTVLRLAKDLAENNKGARVLVVCSEITVVTFRAPNETHLDSLVGQALFGDGAAAIIIGSDPIPNVERPLFELISAAQTLLPNSKNAICGELREIGLTFYLLKDVPKLISNNIRKSLVDVFQPLGISDWNSIFWVAHPGRSAILNQVELKLGLKPEKLGATRHVLSEYGNMSSPSILLVLDEIRKSSVNKGFGTTGEGLEWGVLIGFGPGLTFETVVLHSVSTL
Q2HZ40	Q2HZ40_POPAL	unreviewed	Chalcone synthase	CHS	Populus alba (White poplar)	396			MAPSIEEIRKAQRASGPATILAIGKATPANCVSQADYPDYYFRITNSEHMTELKEKFKRMCDKSMIKKRYMHITEEILKENPSMCEYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGRPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGSRVLVVCSEITAVTFRGPSDTHLDSMVGQALFGDGAAAVIVGADPDTSIERPLFQLVSAAQTILPDFDGAIDGHLREVGLTFHLLKDVPGLISRNIEKSLVEAFAPIGINDWNSIFWIAHPGGPAILDQVEIKLDLKEEKLRATRNVLSDYGNMSSACVLFILDEMRNKSLEEGKSTTGEGLEWGVLFGFGPGLTVETVVLHSVPVEQTIYS
Q9SBS4	Q9SBS4_9LILI	unreviewed	Chalcone synthase	CHS2	Lilium hybrid division VII	412			MANVDEIRQSQQDRISYNLAMANVDEIRQSQRAKGPATVLAIGTATPANMIYQSEYPDYYFRITKNEHMTDLKEKFKRMCDKSMIRKRYMHLTEEILKENPNMCAYMARSLDARQDMVVVEVPKLGKEAAVKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVNRLMMYQQGCFAGGTVLRLAKDLAENNKSARVLVVCSEITAVTFRGPNDSHLDSLVGQALLGDGAAAIIGADPDLAVERPLFQLVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLTGAFQPLPISDWNSIFWIAHPGGPAILDQVEERLALKKEKMKATRHVLSEYGNMSSACVLFILDEMRKNSAAEGKATTGEGLEWGVLFGFGPGLTVETVVLHSVPIVVA
Q58G81	Q58G81_9ASPA	unreviewed	Chalcone synthase (EC 2.3.1.74)	chs	Phalaenopsis hybrid cultivar	390			MPTIESIKKAPRAHGFASILAIGKANPENFIEQCHYPDFYFRVTSSEHLVDLKEKFQRMCDRTAIRKRHFVWNEDLLTANPCLRTYMDKSLNIRQEVAIREIPKLGAEAATKAIQEWGQPKSSITHLIFCTTSGMDLPGADFQLTQILGLNPNVERVMLYQQGCFAGGTTLRLAKCLAESREGARVLVVCAETTTVVFRAPSEEHQDDLVTQALFADGASAVIVGVDPNEAAHERASFIIVSASQVLLPDSAGAIGGHVSEGGLTATLHRDVPQIVSKNVGKCLEEAFTPFGISDWNSIFWVPHAGGRAILDQVEERVGLKPEKLSVSRHVLAEYGNMSSVCVHFALDEMRKKSAKEAKATTGEGLEWGVLFGFGPGLTVETVVLHSVPI
Q27Z06	Q27Z06_SORAU	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Sorbus aucuparia (European mountain ash) (Rowan)	391			MVTVEEVRKAQRAEGPATVMAIGTATPSNCVDQATYPDYYFRITNSEHKVELKEKFQRMCDKSMIKKRYMYLTEEILKENPSVCEYMAPSIDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIIGADPVTEVEKPLFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLNEAFKPVGISDWNSLFWIAHPGGPAILDQVEAKLALKPEKLEATRQVLSDYGNMSSACVLFILDEVRRKSAEKGLKTTGEGLEWGVLFGFGPGLTVETVVLHSVGLTA
A0A1U7XEJ2	A0A1U7XEJ2_NICS	unreviewed	chalcone synthase	LOC104233646	Nicotiana sylvestris (Wood tobacco) (South American tobacco)	395			MSQNGKNINGASKYFFQPSTRLPTPGKATILAMGKAFPAQLVPQDCLVEGYIRDTNCQDLAIKEKLERLCKTTTVKTRYTVMSKEILDKYPELATEGTPTIKQRLEIANPAVVEMAKQASQACIKEWGRSAEEITHIVYVSSSEIRLPGGDLYLATELGLRNDIGRVMLYFLGCYGGVTGLRVAKDIAENNPGSRVLLTTSETTILGFRPPNNARPYDLVGAALFGDGAAAVIIGTEPIMGKESPFMELNFATQQFLPGTNNVIDGRLTEEGINFKLGRDLPEKIQDNIEEFCKKIIAKADLREAKYNDLFWAVHPGGPAILNRLENTLKLQSEKLDCSRRALMDYGNVSSNTIFYVMEYMREELKNKKNGGEEWGLALAFGPGITFEGILLRSL
A0AMG7	A0AMG7_HUMLU	unreviewed	Naringenin-chalcone synthase (EC 2.3.1.74)	chs_H1-1539	Humulus lupulus (European hop)	389			MVTVEEVRKAQRAEGPATILAIGTATPANCILQSEYPDYYFRITNSEHKTELKEKFKRMCDKSMIRKRYMHLTEEILKENPNLCAYEAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRVAKDLAENNKGARVLVVCSEITAVTFRGPNDTHLDSLVGQALFGDGSAALIIGADPIPEIEKPTFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFKPLGISDWNSLFWIAHPGGPAILDQVESKLALKPEKLRATRHVLGEYGNMSSACVLFILDEMRRKCAEDGLKTTGEGLEWGVLFGFGPGLTVETVVLHSVGI
A0AMG9	A0AMG9_HUMLU	unreviewed	Naringenin-chalcone synthase (EC 2.3.1.74)	chs_H1-211	Humulus lupulus (European hop)	389			MVTVEEVRKAQRAEGPATILAIGTATPANCILQSEYPDYYFRITNSEHKTELKEKFKRMCDKSMIRKRYMHLTEEILKENPNLCAYEAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRVAKDLAENNKGARVLVVCSEITAVTFRGPNDTHLDSLVGQALFGDGSAALIIGADPIPEIEKPIFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFKPLGISDWNSLFWIAHPGGPAILDQVESKLALKPEKLRATRHVLGEYGNMSSACVLFILDEMRRKCAEDGLKTTGEGLEWGVLFGFGPGLTVETVVLHSVGI
Q9FEY5	Q9FEY5_HUMLU	unreviewed	Chalcone synthase (EC 2.3.1.74)	chs_H1	Humulus lupulus (European hop)	389			MVTVEEVRKAQRAEGPATILAIGTATPANCILQSEYPDYYFRITNSEHKTELKEKFKRMCGKSMIRKRYMHLTEEILKENPNLCAYEAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSEITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRVAKDLAENNKGARVLVVCSEITAVTFRGPNDTHLDSLVAQALFGDGSAALIIGADPIPEIEKPIFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFKPLGISDWNSLFWITHPGGPAILDQVESKLGLKPEKLRATRHVLGEYGNMSSACVLFILDEMRRKCAEDGVKTTGEGLEWGVLFGFGPGLTVETVVLHSVGI
A0AMG8	A0AMG8_HUMLU	unreviewed	Naringenin-chalcone synthase (EC 2.3.1.74)	chs_H1-132	Humulus lupulus (European hop)	389			MVTVEEVRKAQRAEGPATILAIGTATPANCILQSEYPDYYFRITNSEHKTELKEKFKRMCDKSMIRKRYMHLTEEILKENPNLCAYEAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRVAKDLAENNKGARVLVVCSEITAVTFRGPNDTHLDSLVGQALFGDGSAALIIGADPTPEIEKPIFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFKPLGISDWNSLFWIAHPGGPAILDQVESKLALKPEKLRATRHVLGEYGNMSSACVLFILDEMRRKCAEDGLKTTGEGLEWGVLFGFGPGLTVETVVLHSVGI
Q5QPX9	Q5QPX9_ARATH	unreviewed	Chalcone synthase (EC 2.3.1.74) protein	chs	Arabidopsis thaliana (Mouse-ear cress)	376			MAGASSLDEIRQAQRADGPAGILAIGTANPENHVLQAEYPDYYFRITNSEHMTDLKEKFKRMCTCDKSTIRKRHMHLTEEFLKENPHMCAYMAPSLDTRQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRIAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDTSVGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIVKSLDEAFKPLGISDWNSLFWIAHPGGPAILDQVEIKLGLKEEKMRATRHVLSEYGNMSSACVLFILDEMRRKSAKDGVATTGEGLEWGVL
A0A1B4XSV	A0A1B4XSV6_CART	unreviewed	CHS1 (EC 2.3.1.74) (Chalcone synthase)	CtCHS1	Carthamus tinctorius (Safflower)	397			MASLTDIAEIRKAQRAEGPATILAIGTATPNNCIYQADYPDYYFRITNSEHMVELKQKFKRMCDKSMIRKRYMHITEEFLKENPNMCEYMAPSLDARQDVVVVEVPKLGKEAATKAIKEWGQPKSKITHLIVCTTSGVDMPGADYQITKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPNDTHLDSLVGQALFGDGAAAVIVGADPDLATERPLFEMVSAAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKALVQAFSPLGISDWNSLFWIAHPGGPAILDQVEQKLGLKEEKMRATRHVLSEYGNMSSACVLFIIDEMRKKSAEDGCATTGEGLDWGVLFGFGPGLTVETVVLHSVPTTTPIAA
Q8S4Y7	Q8S4Y7_9FABA	unreviewed	Root-specific chalcone synthase (EC 2.3.1.74)	chs	Senna alata	390			MVSVEEIRKAQRAQGPATVLAIGTATPPNCVDQSTYPDYYFRITNSEHKTELKEKFKRMCEKSMIKKRYMHLTEDLLKENPNMCAYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLIVCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAIIVGSDPLPEVEKPLFELVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKALVEAFQPLGISDYNSIFWIAHPGGPAILDQVEAKLELKPEKMRATRHVLSEYGNMSSACVLFILDEMRRKSIQDGLETTGEGLEWGVLFGFGPGLTVETVVLRSVAVN
Q8S4Y8	Q8S4Y8_9FABA	unreviewed	Root-specific chalcone synthase (EC 2.3.1.74)	chs	Senna alata	389			MVKVEEIRKAQRAEGAATVMAIGTATPANCVEQSTYPDYYFRVTNSEHMTELKEKFQRMCDKSMIKKRYMHLTEEILKENPNMCAYMAPSIDARQDIVVLEVPKLGKEAATKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIVGSDPIPQVETPLFELVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKALVEAFNPLGISDYNSIFWIAHPGGPAILDQVEAKLGLKPEKMQATRHVLSEYGNMSSACVLFIMDEMRRKSTKDGLGTTGEGLEWGVLFGFGPGLTVETVVLHSIAI
Q705N9	Q705N9_ARATH	unreviewed	Chalcone synthase (EC 2.3.1.74) protein	chs	Arabidopsis thaliana (Mouse-ear cress)	390			MVMAGASSLDEIRQAQRADGPAGILAIGTANPENHVLQAEYPDYYFRITNSEHMTDLKEKFKRSDKSTIRKRHMHLTEEFLKENPHMCAYMAPSLDTRQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRIAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFSDGAAALIVGSDPDTSVGEKPIFEMVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIVKSLDEAFKPLGISDWNSLFWIAHPGGPAILDQVEIKLGLKEEKMRATRHVLSEYGNMSSACVLFILDEMRRKSAKDGVATTGEGLEWGVLFGFGPGLTVETVVLH

Q9SB27	Q9SB27_DAUCA	unreviewed	Chalcone synthase (EC 2.3.1.74)	gCHS2	Daucus carota (Wild carrot)	389			MVTVNEFRKAQRAEGPATVLAIGTATPPNCVDQSAYADYYFRITNSEDKPELKEKFRRMCEKSMINTRYMHLTEDLLKQNPSFCEYMASSLDARQDIVVNEVPKLGKEAALRAIKEWGQPKSKITHLIFCTTSGVDMPGADFRLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNKNARVLVVCSEITVITFRGPNDTHLDSLVGQALFGDGAGAVIVGSDPVIGIEKPLFEIVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIRKSLVEAFKPLGISDWNSIFWIAHPGGPAILDQVETELSLKPEKLKSTRQVLRDYGNMSSACVLFILDEMRKASAKDGHRTTGEGLDWGVLFGFGPGLTVETVVLHSVPT
Q2ENC4	Q2ENC4_POPAL	unreviewed	Chalcone synthase	CHS	Populus alba (White poplar)	396			MAPSIEEIRKAQRASGPATILAIGKATPANCVSQADYPDYYFRITNSEHMTELKEKFKRMCDKSMIKKRYMHITEEILKENPSMCEYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGSRVLVVCSEITAVTFRGPSDTHLDSMVGQALFGDGAAAVIVGADPDTSIERPLFQLVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFAPIGINDWNSIFWIAHPGGPAILDQVEIKLDLKEEKLRATRNVLSDYGNMSSACVLFILDEMRNKSLEEGKSTTGEGLEWGVLFGFGPGLTVETVVLHSVPVEQTIYS
A8E1V8	A8E1V8_PINPS	unreviewed	Chalcone synthase (EC 2.3.1.74)	chs1	Pinus pinaster (Maritime pine)	395			MAGGMMADLEAFRKAQRADGPANILAIGTATPPNAVDQSTYPDYYFKITNSEHMTELKEKFRRMCDKSAIKKRYMYLTEEILQENPSVCAYMAPSLDARQDMVVVEVPRLGKEAAAKAIKEWGQPKSKITHVIFCTTSGVDMPGADYQLTKLLGLRPSVKRVMMYQQGCFAGGTVLRVAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSMVGQALFGDGAAALIVGADPVPEVEKPCFEMLWTSQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFQQFGISDWNQLFWIAHPGGPAILDQVEAKLNLDPKKLRATRQVLSDYGNMSSACVHFILDEMRKSSHQNGCSTTGEGLDVGVLFGFGPGLTVETVVLKSVPLQ
A0A0K0KBH(	A0A0K0KBH0_ARAH	unreviewed	Chalcone synthase	CHS2	Arachis hypogaea (Peanut)	393			MVSVEEIRNAARANGPATVLAIGTATPSNCVYQDTYPDYYFRITNSEHMTDLKEKFKRMCEKSMIRKRYMHLTEDFLKKNPSMCEYMAPSLDARQDIVVVEVPKLGKEAATKAIKDWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRSSVKRLMMYQQGCFAGGTVLRLAKDLAENNKNARVLVVCSEITAVTFRGPSDTHLDSMVGQALFGDGAATVIVGADPDTKVERPLFEMVSAGQTILPDSEGAIDGHLREVGLTFHLLKDVPGIISKNIEKSLVEAFTPIGISDWNSIFWIAHPGGPAILDQVEAKLQLKEEKLKATRHVLGEYGNMSSACVLFILDEMRKRSIEEGKATTGEGFDWGVLFGFGPGLTVETVVLHSLPLENLS
Q9AVC0	Q9AVC0_9LILI	unreviewed	Chalcone synthase (EC 2.3.1.74)	LhCHSB	Lilium hybrid division I	393			MSKTVEEVRKAQRAQGPATILAIGTATPSNVIYQADYPDYYFRITNSEHLTDLKQKFKRMCKKSMIKKRYIHLNEEILQENRNMCAYMAPSLDARQDIVVVEVPKLGKEAASKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVNRFMMYQQGCFAGGTVLRFAKDLAENNCDARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAAAVIVGSDPDTSVERPLFQIVSASQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLTQAFAPLGITDWNSIFWIAHPGGPAILDQVELKLALDKKKMQATRHVLSEYGNMSSACVLFILDEMRKASAEQGKATTGEGLDWGVLFGFGPGLTVETVVLHSIPITTN
Q42864	Q42864_9ROSI	unreviewed	Naringenin-chalcone synthase (EC 2.3.1.74)	CHS1	Juglans nigra x Juglans regia	389			MVTVEDVRRAQRAEGPATVMAIGTATPPNCVDQSAYPDYYFRITNSEHKTELKEKFKRMCEKSMIKKRYMHLTEEILKENPNVCAYMASSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAALIVGADPVPGVEKPLFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFQPLGITDWNSLFWIAHPGGPAILDQVESKLELKPEKLRATRHVLSEYGNMSSACVLFILDEMRKKSAEDRLKTTGEGLEWGVLFGFGPGLTVETVVLHSVSA
Q42865	Q42865_9ROSI	unreviewed	Naringenin-chalcone synthase (EC 2.3.1.74)	CHS2	Juglans nigra x Juglans regia	389			MVTVEDVRRAQRAEGPATVMAIGTATPPNCVDQSAYPDYYFRITNSEHKTELKEKFKRMCGKSMIKKRYMHLTEEILKENPNVCAYMASSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAALIVGADPVPGVEKPLFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFQPLGITDWNSLFWIAHPGGPAILDQVESKLELKPEKLRATRHVLSEYGNMSSACVLFILDEMRKKSAEDRLKTTGEGLEWGVLFGFGSGLTVETVVLHSVSA
G9F7X4	G9F7X4_CURLO	unreviewed	Chalcone synthase-like protein (EC 2.3.1.74)	ClPKS9	Curcuma longa (Turmeric) (Curcuma domestica)	396			MAKLVTEIRKSQRAEGPAAVLAIGTATPPNVVYQADYPDYYFRITRSEHLVELKEKFKRMCDKSMIRKRHMYLTEEILRENPKMCAYMEASLDARQDIVVVEVPRLGKEAAVKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVNRFMMYQQGCFAGGTVIRLAKDLAENNRGARVLVVCSEITAVTFRGPSDSHLDSMVGQALFADGAGAIIVGADPDPATERPLFELVSASQTILPDSEGAIDGHLREAGLTFHLLKDVPGLISKNIEKSLVEAFKPLGISDWNSLFWIAHPGGPAILDQVEAKLALDKDKMKATRNVLSEYGNMSSACVLFILDEMRRRSAEEGKATTGEGLEWGVLFGFGPGLTVETVVLHSVPISAAATH
A8CLG3	A8CLG3_RHEPA	unreviewed	Chalcone synthase 1	CHS1	Rheum palmatum (Chinese rhubarb)	391			MAPTVQEIRKAQRAEGPATILAIGTATPPNCVYQADYPDYYFRVTNSDHMTDLKEKFRRMCDKSMIEKRYMHLTEDLLKQNPGMCEYMASSLDARQDMVVSEVPRLGKEAAQRAIKEWGQAKSKITHVIMCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRMAKDLAENNKGARVLVVCSEITAICFRGPTDTHLDSMVGQALFGDGAGALVVGADPDLSIEKPIFELVWTAQTILPDSEGAIDGHLREVGLTFHLLKTCPGDLQTSRKSLAEAFSPLDISDWNSLFWVAHPGGPAILDQVEAKLGLKEEKLKATRQVLNDYGNMSSACVLFILDEMRKKSIKNGHATTGEGLDWGVLFGFGPGLTVETVVLHSVPTAN
I6QHR8	I6QHR8_SILMA	unreviewed	Chalcone synthase 2 (EC 2.3.1.74)	chs2	Silybum marianum (Blessed milk-thistle) (Carduus marianus)	412			MASSSIDIAEIRKAQRAQGPATILAIGTATPANCIYQADYPDYYFRITNSEHMVDLKEKFKRMCDKSMIRKRYMHITEEFLKENPNMCEYMAPSLDARQDVVVVEVPKLGKEAATKAIKEWGNPKSKITHLIVCTTSGVDMPGADYQITKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDIAENNKGSRVLVVCSEITAVTFRGPNDTHLDSLVGQALFGDGAAAIIVGADPDLKTERPLFEMVSAAQTILPDSEGAIDGHLREVGLTFHLLKDAPGLISKNIEKALIQAFSPLGITDWNSIFWIAHPGGPAILDQVEQKLGLKEEKMRATRHVLSEYGNMSSACVLFIIDEMRKKSAAEGATTTGEGLDWGVLFGFGPGLTVETVVLHSLPTATMPIAPQNLSQQVVINHE
Q94LW8	Q94LW8_HUMLU	unreviewed	Chalcone synthase (EC 2.3.1.74)	chs3	Humulus lupulus (European hop)	399			MSSSITVDQIRKAQRAEGPATILAIGTATPANFIIQADYPDYYFRVTKSEHMTNLKKRFQRICDRTMIKKRHLVLSEDHLKENPNMCEFMAPSLDVRQDILVVEVPKLGKEACMKAIKEWDQPKSKITHFIFATTSGVDMPGADYQCAKLLGLSSSSVKRVMMYQQGCFAGGTVLRIAKDIAENNKGARVLALCSEITTCMFHGPTESHLDSMVGQALFGDGASAVIVGAEPDESAGERPIYELVSAAQTILPNSEGAIDGHLMETRLTFHLLKDVPGLISNNIEKSLIEAFTPIGINDWNSIFWVTHPGGPAILDEVEAKLELKKEKLAISRHVLSEYGNMSSASVFFVMDELRKRSLEEGKSTTGDGLDWGVLFGFGPGLTVEMVVLHSVENKVKSET
A1E0A0	A1E0A0_POLCS	unreviewed	Chalcone synthase	CHS1	Polygonum cuspidatum (Japanese knotweed)	393			MAPSVQEIRKAQRAEGPATVLAIGTATPPNCIYQADYPDYYFRVTNSEHMTDLKEKFRRMCDKSMIEKRYMHLTEEILKENQNMCAYMASSLDSRQDMVVSEVPRLGKEAAQKAIKEWGQPKSKITHVIMCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTALRLAKDLAENTKGARVLVVCSEITAICFRGPTDTHLDSMVGQALFGDGAGAVIIGADPDLSIERPIFELVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLTEAFSPLNISDWNSLFWIAHPGGPAILDQVEAKLGLKEEKLKATRQVLNDYGNMSSACVLFIMDEMRKKSLENGHATTGEGLDWGVLFGFGPGLAVETVVLHSVPVAHH
B0FYP9	B0FYP9_IPOBA	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Ipomoea batatas (Sweet potato) (Convolvulus batatas)	393			MVTVEEVRKAQRAEGPATILAIGTATPANCVNQSTYPDYYFRITNSEHKTELKEKFQRMCDKSMITKRYMHLTEEILKENPSVCEYMAPSLDARQDIAVVEVPKLGKEAAQSAIKEWGQPKSKITHVVFCTTSGVDMPGADYQLTKLLGLCPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLIVCSEITVVTFRGPSETHLDSLVGQALFGDGAAAVIVGADPTPAEKPLFQLVSAAQTLAPDSCGAIDGHLREVGLTFHLLKDVPSVVSNNIEKCLFEAFNPLGISDWNSVFWIAHPGGPAILDQVEDKLGLKPEKLRATRHVLSEYGNMSSACVLFILDEMRKASSNDGLGTTGEGLEWGVLFGFGPGLTIETVVLHSVPIKPGPH
Q9AVC1	Q9AVC1_9LILI	unreviewed	Chalcone synthase (EC 2.3.1.74)	LhCHSA	Lilium hybrid division I	394			MASMTVDEVRQAQRAQGPATVLAIGTATPSNVIYQADYPDYYFRITKSEHLTGLKEKFKRMCEKSMIRKRYMHLNEEILAENHNVCAYMAPSLDVRQDMVVVEVPKLGKEAAAKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVNRFMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAAAVIVGSDPDTAVERPLFELVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLTGAFAPLGISDWNSLFWIAHPGGPAILDQVAAKLGLQKEKMRATRHVLSEYGNMSSACVLFILDEMRKTSAKMGKATTGEGLDWGVLFGFGPGLTVETVVLHSLPIAAE
G1ETS5	G1ETS5_PAELC	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Paeonia lactiflora (Chinese peony) (Paeonia albiflora)	393			MASVEEIRNAQRAQGPATILAIGTATPAHCINQAEYPDYYFRITNSEHKTELKEKFKRMCDKSMINKRYMYLTEEILKENPNMCEYMAPSLDARQDMVVVEIPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGPRVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIVGADPDVKIERPLFQIVSAGQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFKPIGINDWNSILWIAHPGGPAILDQVELKLGLKEEKLKNTRHVLSEYGNMSSACVLFILDETRKRSLEEGKATTGEGLDWGVLFGFGPGLTVETVVLHSVPAITNQ
A0A1D1XGY!	A0A1D1XGY5_9ARA	unreviewed	Chalcone synthase	CHS1_10 g.113742	Anthurium amnicola	391			MVTTSLEAIRKAQRADGPATILAIGTAVPPNAVDQSTYPDYYFRITNSEHQVELKEKFRRMCDKSMIKKRHMYLTEEILKENPSMCAYMAPSLDARQDMVVVEVPRLGKEAATRAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLAKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVICSEVTAVTFRGPSESHLDSLVGQALFGDGASALVVGADPVEGVERPIFQVVSAAQTILPDSHGAIDGHLREVGLTFHLLKDVPGLISKNIDKSLVEAFEPLGISDWNSLFWIAHPGGPAILDQVEDKLRLQSEKLAATRRVLSEYGNMSSACVHFILDEMRKRSAEEGRGTTGEGLQWGVLFGFGPGLTVETLVLHSVAI
F6M1S2	F6M1S2_FAGTA	unreviewed	Chalcone synthase (EC 2.3.1.74) protein	CHS	Fagopyrum tataricum (Tartarian buckwheat) (Polygonum tataricum)	335			DKSQIEKRCMYLTEEILKEHPNMCEYMAPSLASRQDMVVTEVPKLGKEAAQKAIKEWGQPKSKITHVIVCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRMAKDLAENNRGARVLVVCYEITAVCFRGPTDTHLDSMVGQALFGDGAGAVIVGADPDLSIEKPIFELVWTSQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLTEAFDDFSTSPTGTPSSGSPTPGAPLSSTRSRPSSDSRRRRSSRATRQVLNDYGNMSSACVLFILDEMRKKSLENGHATTGEGLDWGVLFGFGPGLTVETVVLHSVPTTTLAN
A0A0A0P544	A0A0A0P544_9ROS	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Pyrus x bretschneideri (Chinese white pear)	389			MVTVEEVRKAQRAEGPATVLAIGTSTPPNCVDQATYPDYYFRITNSEHKTELKEKFQRMCDKSMIKTRYMYLTEEILKENPTVCEYMAPSLDARQDMVVVEVPRLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIIGSDPVPEVEKPLFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLNEAFKPIGISDWNSLFWIAHPGGPAILDQVESKLALKPEKLEATRQVLSDYGNMSSACVLFILDEVRRRKSAEKGLKTTGEGLEWGVLFGFGPGLTVETVVLHSVAA
I1Y8W8	I1Y8W8_9CARY	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Persicaria minor	392			MAPSVEQIGKAQRAEGPATVLAIGTATPPNCVSQADYPDYYFRVTNSEHMTDLKEKFRRMCDKSMIEKRYMYLTEEILKENPNMCAYMEKSLDSRQDIVVTEVPRLGKEAAQKAIKEWGQAKSKITHVIMCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVCFRGPTDTHLDSMVGQALFGDGAGAVIIGADPDLSIEKPIFELVWTAQTILPDSEGAIDGHLREVGLTFHPLKDVPGLISKNIDKSLKEAFSPLNIDDWNSLFWIAHPGGPAILDQVEAKLGLKQEKLKATRQVLNDYGNMSSACVLFIMDEMRKKSVENGHATTGEGLDWGVLFGFGPGLTVETVVLHSVPVAH
A0A0A1E1L3	A0A0A1E1L3_9SOL	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Iochroma calycinum	389			MMTVEEVRRAQRAKGPATIMAIGTATPSNCVDQSTYPDYYFRITNSEHMTELKEKFKRMCDKSMIKKRYMHLTEEILKENPNICEYMAPSLDARQDIVVVEVPKLGKEAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVIRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSMVGQALFGDGAAAIIIGSDPLPGVERPLFELVSAAQTLLPDSEGAIDGHLREVGLTFHLLKDVPGLISQNIEKSLIEAFQPLGISDWNSVFWIAHPGGPAILDQVELKLGLKPEKLRATRQVLSDYGNMSSACVLFILDEIRKASAKEGLSTTGEGLDWGVLFGFGPGLTVETVVLHSVST

A0A0D3MTI6	A0A0D3MTI6_FAGT	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS2	Fagopyrum tataricum (Tartarian buckwheat) (Polygonum tataricum)	369			MAPSVEEEIRKAQRADGPATVLAIGTATPPNCIYQADYPDFYFKVTNSEHMTDLKQKFKRMCDKSMIEKRFMHLTEEEILKENQNMCAYMAPSLDSRQDMVVSEVPRLGKEAAQKAIKEWGQSKSKITHVIMCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNKGARVLVALFGDGAGAVIIGADPDLSIEKPIFELVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIHKSLDEAFSPLNISDWNSLFWIAHPGGPAILDQVEEKLGLKAEKMKATRQVLNDYGNMSSACVLFIMDEMRKKSLENGHATTGEGLEWGVLFGFGPGLTVETVVLHSVPVVAQ
U3N0Z6	U3N0Z6_9MYRT	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Melastoma malabathricum	391			MVTVEEVRRSQRAEGPATILAIGTATPPNCVDQSTYPDYYFRITNSEHKAELKEKFQRMCDKSMIKKRYMYLTEEEILKENPSVCAYEAPSLDARQDMVVVEVPKLGKEAAAKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAGAIIIGSDPIPGVEKPMFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNVEKSLVEAFQPLGISDWNSLFWIAHPGGPAILDQVEEKLGLKPEKMRATRQVLSDYGNMSSACVLFILDEMRRSSKQNGFKTTGEGLEWGVLFGFGPGLTVETVVLHSIATEA
D7R4L1	D7R4L1_FAGTA	unreviewed	Chalcone synthase protein	CHS	Fagopyrum tataricum (Tartarian buckwheat) (Polygonum tataricum)	334			DKSQIEKRYMYLTEEEILKEHPNMCEYMAPSLDSRQDMVVTEVPKLGKEAAQKAIKEWGQPKSKITHVIVCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRMAKDLAENNRGARVLVVCSEITAVCFRGPTDTHLDSMVGQALFGDGAGAVIVGADPDLSIEKPIFELVWTSQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLTEAFSPLNIADWNSLFWIAHPGGPAILDQVEAKLGLKEEKLKATRQVLNDYGNMSSACVLFILDEMRKKSLENGHATTGEGLDWGVLFGFGPGLTVETVVLHSVPTTTTLAN
X4QM74	X4QM74_FAGTA	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS1	Fagopyrum tataricum (Tartarian buckwheat) (Polygonum tataricum)	396			MAPTVQEIRKAQRAEGPATVLAIGTATPPNCVYQADYPDYYFRVTNSDHMTDLKEKFRRMCDKSQIEKRYMYLTEEEILKEHPNMCEYMAPSLDSRQDMVVTEVPKLGKEAAQKAIKEWGQPKSKITHVIVCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRMAKDLAENNRGARVLVVCSEITAVCFRGPTDTHLDSMVGQALFGDGAGAVIVGADPDLSIEKPIFELVWTSQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLTEAFSPFSTSPTGTPSSSGSPTPGAPLSSSTRSRPSSSDSRRRRSSRATRQVLNDYGNMSSACVLFILDEMRKKSLENGHATTGEGLDWGVLFGFGPGLTVETVVLHSVPTTTTLAN
F2VR45	F2VR45_PRUAV	unreviewed	Naringenin-chalcone synthase (EC 2.3.1.74)	CHS1	Prunus avium (Cherry) (Cerasus avium)	391			MVTVEEVRKAQRAEGPATVLAIGTSNPPNCVDQATYPDYYFRITNSEHKTELKEKFQRMCDKSMIKKRYMYLTEEEILKENPTMCEYMAPSLDARQDMVVVEIPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRSSSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAIVGADPIPEIEKPLFEVVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIZKSLNEAFQPLGISDWNSLFWIAHPGGPAILDQVESKLALKPEKLEATRHILSEYGNMSSACVLFILDEVRKRATKKGLKTTGDGLDWGVLFGFGPGLTVETVVLHSVGLNA
I7GY80	I7GY80_9LILI	unreviewed	Chalcone synthase C	CHSC	Lilium hybrid division I	392			MANLDEIRQSQRAEGPATVLAIGTATPANMIYQSEYPDYYFRITKSEHMTELKEKFKRMCDKSMIRKRYMHLNEEEILTENPNMCAYMAPSLDARQDMVVVEVPKLGKEAAVKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVNRLMMYQQGCFAGGTVLRLAKDLTENNKSARVLVVCSEITAVTFRGPDNAHLDSLVGQALFGDGAAAIVGADPDMTVERPLFQLVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPALISKNIEKSLTAAFQPLGISDWNSIFWIAHPGGPAILDQVEERLALKKEKMKATRHVLSEYGNMSSACVLFILDEMRKKSAEEGKATTGEGLDWGVMFGFGPGLTVETVVFHSLPIVAA
A0A1B4XSV2	A0A1B4XSV2_CART	unreviewed	Chalcone synthase	CtCHS2	Carthamus tinctorius (Safflower)	401			MASSPAIIDIDAIRKSQRAQGPATILAIGTATPSNCVLQADYPDYYFRITNSEHLVDLKEKFKRMCDKSMIRKRYMHITEEFLKENPNMCAYEAPSLDARQDLVVVEVPKLGKEAAVKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNAGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAAAVIVGSDPDPATERPLFEMVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKALTQAFSPLGISDWNSLFWIAHPGGPAILDQVEERLGLKEEKMRATRKVLSEYGNMSSACVLFILDEMRRKSAEDGRATTGEGLDWGVLFGFGPGLTVETVVLHGIPTTISVET
A1E025	A1E025_POLCS	unreviewed	Chalcone synthase	CHS1	Polygonum cuspidatum (Japanese knotweed)	393			MAPSVQEIRKAQRAEGPATVLAIGTATPPNCIYQADYPDYYFRVTNSEHMTDLKEKFRRMCDKSMIEKRYMHLTEEEILKENQNMCAYMASSLDSRQDMVVSEVPRLGKEAAQKAIKEWGQPKSKITHVIMCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTALRLAKDLAENTKGARELVVCSEITAICFRGPTDTHLDSMVGQALFGDGAGAVIIGADPDLSIERPIFELVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLTEAFSPLNISDWNSLFWIAHPGGPAILDQVEAKLGLKEEKLKATRQVLNDYGNMSSACVLFIMDEMRKKSLENGHATTGEGLDWGVLFGFGPGLAVETVVLHSVPVAHH
C4MJ52	C4MJ52_9FABA	unreviewed	Chalcone synthase (EC 2.3.1.74)	chs	Glycyrrhiza inflata	389			MVSVAEIRKAQRAEGPANILAIGTANPPNCVDQSTYPDFYFKITNSEHKTELKEKFQRMCDKSMIKKRYMYLTEEILKENPNICAYMAPSLDARQDMVVVEVPRLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPYVKRYMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPTDTHLDSLVGQALFGDGAAAVIVGSDPVPEIEKPIFELVWTAQTIAPDSEGAIDGHLREVGLTFHLLKDVPGIVSKNIDKALTGAFQPLGISDYNSIFWIAHPGGPTILDQVEQKLALKPEKMKATRDVLSDYGNMSSACVLFILDEMRKKSAQNGLKTTGEGLEWGVLFGFGPGLTIETVVLHSVAI
G9F7X3	G9F7X3_CURLO	unreviewed	Chalcone synthase-like protein (EC 2.3.1.74)	ClPKS10	Curcuma longa (Turmeric) (Curcuma domestica)	389			MANLHALRREQRTQGPATIMAIGTATPPNLYEQSTFPDFYFRVTNSDDKQELKEKFRRMCNKSMVKKRYLYLTEEILKERPKLCSYKEPSFDDRQDIVVEEIPKLAKEAAEKAISTRGRPKSEITHLVFCSISGIDMPGADYRLATLLGLPLTVNRLMIYSQACHMGAAMLRIAKDLAENNRGARVLVVACEITVLSFRGPNERDFEALAGQAGFADGAAAVVVGADPLERVEKPIYEIAAAMQETVAESQEAVGGHLRAFGWTFYFLNQLPAIISNNIGKSLERALVPLGVREWNDVFWVAHPGNWAIMDAIEAKLQLTPDKLSTARHVFSEYGNMQSATVYFVMDELRKRSAVEGRSTTGDGLQWGVLFGFGPGLSIETVVLRSMPL
A0A0A1DZ26	A0A0A1DZ26_9SOL	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Iochroma cyaneum	389			MVTVEEVRRAQRAKGPATIMAIGTATPSNCVDQSTYPDYYFRITNSEHMTELKEKFKRMCDKSMIKKRYMHLTEEEILKENPNICEYMAPSLDARQDIVVVEVPKLGKEAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVIRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSMVGQALFGDGAAAIIIGSDPLPGVERPLFELVSAAQTLLPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLIEAFQPLGISDWNSVFWIAHPGGPAILDQVELKLGLKPEKLRATRQVLSDYGNMSSACVLFILDEMRKASAKEGLSTTGEGLDWGVLFGFGPGLTVETVVLHSVST
X2D809	X2D809_MAGLI	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Magnolia liliiflora (Mulan magnolia) (Yulania liliiflora)	394			MSTVSIDAIRKAQRAGGPATVLAIGTATPSNEVIQADYPDYYFKITKSEHMTELKEKFKRMCDKSMIRKRYMHLTEDILKENPDMCAYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGHPKSKITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRYMMYQQGCFAGGTVLRLAKDLAENNAGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIVGADPNSSERPLFQLVSTAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFEPLGINDWNSIFWIAHPGGPAILDQVEAKLNLKAEKMRATRHVLSEYGNMSSACVLFILDEMRKKSAEEGKGTTGEGLDWGVLFGFGPGLTVETVVLHSLPITGAH
A0A1L5J714	A0A1L5J714_ECHPI	unreviewed	Chalcone synthase (EC 2.3.1.74)	chsB1	Echinacea pallida (Pale purple coneflower) (Rudbeckia pallida)	398			MASTIDIAAFREAQRAQGPATILAIGTATPSNCVYQADYPDYYFRITKSEHMVDLKEKFKRMCDKSMIRKRYMHLTEEFLKENPSICEYMAPSLDARQDVVVVEVPKLGKEAAVKAIKEWGKPKSQITHLIFCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSMVGQALFGDGAAALVVGSDPDLTTERPLFQMVSAAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIQKALIQAFSPLGISDWNSIFWIAHPGGPAILDQVQLKLGLREEKMRATRHVLSEYGNMSSACVLFILDEMRKKSAEDGATTTGEGLDWGVLFGFGPGLTVETVVLHSLPTTMPIAP
A0A1L5J726	A0A1L5J726_ECHPI	unreviewed	Chalcone synthase (EC 2.3.1.74)	chsA	Echinacea pallida (Pale purple coneflower) (Rudbeckia pallida)	389			MLSIQEFRKAQRAEGPATILAIGTATPPNCVVQSTYPDYYFRITKSEDKKDLKEKFTRMCEKSMIRQRYMYLTEEILKDKPNMCAYNAPSLDDRQDIVVVEVPKLGKEAATRAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRSSSVKRFMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPDDTHLDSLVGQALFGDGAAAIIVGSDPLPDVEKPIFEIISAAQTILPDSGGAIDGHLREVGLTFHLLKDVPGLISKHIEKSLVDAFQPLGINDRNSVFWIAHPGGPAILDQVEEKLALTPDKLRASRHVLSEYGNMSSACVLFILNEMRHTSATDGFSTTGEGVEWGVLFGFGPGLTVETVVLHSVSI
A0A1L5J720	A0A1L5J720_ECHPI	unreviewed	Chalcone synthase (EC 2.3.1.74)	chsB2	Echinacea pallida (Pale purple coneflower) (Rudbeckia pallida)	398			MASSIDIAAIREAQRAQGPATILAIGTATPSNCVCQADYPDYYFRITKSEHMTDLKEKFRRICNKSMIRKRYMHLTEEFLKENPSICEFMAPSLDARLDVAVVEVPKLGKEAAVKAIKEWGKPKSLITHLIFCTTSGVDMPGADYQPTKLLGLHPSVKRFMIYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEIITVVTFRGPSDTHLDSLVGQALFGDGAAAVVVGSDPELTTEQPLFEIVSAAQTILPDSEGVKGHLREVGLTFHLLKDIPVVIADNIEKALKQAFLPLGISDWNSIFWIAHPGGPAILDQVQIKLGLKEEKMRATRRVLSEYGNMSSACVLFILNEMRKKSAEEGAATAGEGLDWGVLFGFGPGLTVETVVLHSLPIIMRIAP
A0A0A1E1Z4	A0A0A1E1Z4_9SOL	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Iochroma cyaneum	389			MVTVEEVRRAQRAKGPATIMAIGTATPSNCVDQSTYPDYYFRITNSEHMTELKEKFKRMCDKSMIKKRYMHLTEEEILKENPNICEYMAPSLDARQDIVVVEVPKLGKEAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVIRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSMVGQALFGDGAAAIIIGSDPLPGVERPLFELVSTAQTLLPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLIEAFQPLGISDWNSVFWIAHPGGPAILDQVELKLGLKPEKLRATRQVLSDYGNMSSACVLFILDEMRKASAKEGLSTTGEGLDWGVLFGFGPGLTVETVVLHSVST
I4CHP2	I4CHP2_LYCBA	unreviewed	Chalcone synthase (EC 2.3.1.74) protein	CHS	Lycium barbarum (Matrimony vine)	383			MVTVEEYRKAQRAEGPATVMAIGTATPSNCVDQSTYPDYYFRITDSEHKTELKEKFKRMCDKSMIKKRYMHLTEEEILKENPNMCAYMAPSLDARQDIVVVEVPKLGKEAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGCDYQLAKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAAAIIMGSDPIPGVERPLFELVSAAQTLLPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFQPLGISDWNSLFWIAHPGGPAILDQVELKLALKPEKLRATRDVLSNYGNMSSACVLFILDEMRKASAKEGLGTTGEGLGWGVLFGFGPGLTVETVV
A0A0A0RBQ	A0A0A0RBQ5_9MA	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Plagiochasma appendiculatum	411			MAPQAVDAACAAEATPVIPTSRPRLIAQAVGPATVLAMGKAVPHNVFEQATYPDFFFNITKCNDKPTLKAKFQRICDKSGIKKRHFYLDQKILESNPSMCTYMETSLNCRQEIAIAHVPKLAKEAAQVAIKEWGRPKSEITHIVMATTSGVNMPGAELATAKLLGLRPNVRRVMMYQQGCFAGATVLRVAKDLAENNAGARVLAICSEVTAVTFRAPCETHIDGLVGSALFGDGAAAVIVGADPRPEIERPMFEMHWAGEMVLPESDGAIDGHLTEAGLVFHLLKDVPGLITKNIGGFLKDTKNLVGASTWNDLFWAVHPGGPAILDQVEAKLELEKNKFQASRDILADYGNMSSASVLFVLDRVRERSLESNKYTFGEGSEWGFLIGFGPGLTVETLLLLRALPTEQAASA
Q3YMX4	Q3YMX4_ARAHG	unreviewed	Chalcone synthase family protein protein	CHS1	Arabidopsis halleri subsp. gemmifera (Arabis gemmifera)	393			MSNSRMNGVEKLSSISTRRVANPRKATLLALGKAFPSQVVPQENLVEGFLRDTKCDDAFIKEKLEHLCKTTTVKTRYTVLSREILDKYPELTTEGSPTIKQRLEIANEAVVEMALEASLGCIKEWGRPVEDITHIVYVSSSEIRLPGGDLYLSAKLGLRNDVNRVMLYFLGCYGGVTGLRVAKDIAENNPGSRVLLTTSETTILGFRPPNKARPYDLVGAALFGDGAAAVIIGADPRECEAPFMELHYAVQQFLPGTQNVIDGRLTEEGINFKLGRDLPQKIEENIEEFCKKLMGKAGGDEPMEFNDMFWAVHPGGPAILNRLETKLKLEKEKLESSRRALVDYGNVSSNTILYVMEYMREELKKKGDAAQEWGLGLAFGPGITFEGLLIRSL
Q3YMX3	Q3YMX3_ARAHG	unreviewed	Chalcone synthase family protein protein	CHS1	Arabidopsis halleri subsp. gemmifera (Arabis gemmifera)	393			MSNSRMNGVEKLSSISTRRVANPRKATLLALGKAFPSQVVPQENLVEGFLRDTKCDDAFIKEKLEHLCKTTTVKTRYTVLSREILDKYPELTTEGSPTIKQRLEIANEAVVEMALEASLGCIKEWGRPVEDITHIVYVSSSEIRLPGGDLYLSAKLGLRNDVNRVMLYFLGCYGGVTGLRVAKDIAENNPGSRVLLTTSETTILGFRPPSKARPYGLVGAALFGDGAAAVIIGADPRECEAPFMELHYAVQQFLPGTQNVIDGRLTEEGINFKLGRDLPQKIEENIEEFCKKLMGKAGGDESMEFNDMFWAVHPGGPAILNRLETKLKLEKEKLESSRRALVDYGNVSSNTILYVMEYMREELKKKGDAAQEWGLGLAFGPGITFEGLLIRSL
Q3YMY3	Q3YMY3_ARAHG	unreviewed	Chalcone synthase family protein protein	CHS1	Arabidopsis halleri subsp. gemmifera (Arabis gemmifera)	393			MSNSRMNGVEKLSSISTRRVANPRKATLLALGKAFPSQVVPQENLVEGFLRDTKCEDAFIKEKLEHLCKTTTVKTRYTVLSREILDKYPELTTESSPTIKQRLEIANEAVVEMALEASLGCIKEWGRPVEDITHIVYVSSSEIRLPGGDLYLSAKLGLRNDVNRVMLYFLGCYGGVTGLRVAKDIAENNPGSRVLLTTSETTILGFRPPNKARPYDLVGAALFGDGAAAVIIRADPRECEAPFMELHYAVQQFLPGTQNVIDGRLTEEGINFKLGRDLPQKIEENIEEFCKKLMGKAGGDESMEFNDMFWAVHPGGPAILNRLETKLELEKEKLESSRRALVDYGNVSSNTIMYVMEYMREELKKKGDAAQEWGLGLAFGPGITFEGLLIRSL
Q3YMY4	Q3YMY4_ARAHG	unreviewed	Chalcone synthase family protein protein	CHS1	Arabidopsis halleri subsp. gemmifera (Arabis gemmifera)	393			KSNSRMNGVEKLSSISTRRVANPGKATLLALGKAFPSQVVPQENLVEGFLRDTKCEDAFIKEKLEHLCKTTTVKTRYTVLSREILDKYPELTTEGSPTIKQRLEIANEAVVEMALEASLGCIKEWGRPVEDITHIVYVSSSEIRLPGGDLYLSAKLGLRNDVNRVMLYFLGCYGGVTGLRVAKDIAENNPGSRVLLTTSETTILGFRPPNKARPYDLVGAALFGDGAAAVIIGADPRECEAPFMELHYAVQQFLPGTQNVIDGRLTEEGINFKLGRDLPQKIEENIEEFCKKLMGKAGGDESMEFNDMFWAVHPGGPAILNRLETKLKLEKEKLESSRRALVDYGNVSSNTIMYVMEYMREELKKKGDAAQEWGLGLAFGPGITFEGLLIRSL
					Arabidopsis halleri				ISNSRMNGVEKLSSISTRRVANPGKATLLALGKAFPSQVVPQENLVEGFLRDTKCEDAFIKEKLEHLCKTTTVKTRYTVLSREILDKYPELTTEGSPTIKQRLEIANEAVVEMALEASLGCIKEWGRPVEDITHIVYVSSSEIRLPGGDLYLSAKLGLRNDVNRVMLYFLGCYGGVTGLRVAKDIAENNPGSRVLLTTSETTILGFRPPNKARPYDLVGAALFGDGAAAVIIGADPRECEAPFMELHYAVQQFLPGTQNVIDGRLTEEGINFKLGRDLPQKIEENIEEFCKKLMGKAGGDESMEFNDMFWAVHPGGPAILN

Q3YMX7	Q3YMX7_ARAHG	unreviewed	Chalcone synthase family protein protein	CHS1	Arabidopsis halleri subsp. gemmifera (Arabis gemmifera)	393			MSNSRMNGVEKLSSISTRRVANPRKATLLALGKAFPSQVVPQENLVEGFLRDTKCDDAFIKEKLEHLCKTTTVKTRYTVLSREILDKYPELTTEGSPTIKQRLEIANEAVVEMALEASLGCIKEWGRPVEDITHIVYVSSSEIRLPGGDLYLSAKLGLRNDVNRVMLYFLGCYGGVTGLRVAKDIAENNPGSRVLLTTSETTILGFRPPNKARPYDLVGAALFGDGAAAVIIGADPRECEAPFMELHYAVQQFLPGTQNVIDGRLTEEGINFKLGRDLPQKIEENIEEFCKKLMGKAGGDESMEFNDMFWAVHPGGPAILNRLETKLKLEKEKLESSRRALVDYGNVSSNTILYVMEYMREELKKKGDAAQEWGLGLAFGPGITFEGLLIRSL
Q3YMX6	Q3YMX6_ARAHG	unreviewed	Chalcone synthase family protein protein	CHS1	Arabidopsis halleri subsp. gemmifera (Arabis gemmifera)	393			MSNSRMNGVEKLSSISTRRVANPRKATLLALGKAFPSQVVPQENLVEGFLRDTKCDDASIKEKLEHLCKTTTVKTRYTVLSREILDKYSELTTEGSPTIKQRLEIANEAVVEMALEASLGCIKEWGRPVEDITHIVYVSSSEIRLPGGDLYLSAKLGLRNDVNRVMLYFLGCYGGVTGLRVAKDIAENNPGSRVLLTTSETTILGFRPPNKARPYDLVGAALFGDGAAAVIIGADPRECEAPFMELHYAVQQFLPGTQNVIDGRLTEEGINFKLGRDLPQKIEENIEEFCKKLMGKAGGDESMEFNDMFWAVHPGGPAILNRLETKLKLEKEKLESSRRALVDYGNVSSNTILYVMEYMREELKKKGDAAQEWGLGLAFGPGITFEGLLIRSL
Q3YMX1	Q3YMX1_ARAHG	unreviewed	Chalcone synthase family protein protein	CHS1	Arabidopsis halleri subsp. gemmifera (Arabis gemmifera)	393			MSNSRMNGVEKLSSISTRRVADPRKATLLALGKAFPSQVVPQENLVEGFLRDTKCDDAFIKEKLEHLCKTTTVKTRYTVLSREILDKYPELTTEGSPTIKQRLEIANEAVVEMALEASLGCIKEWGRPVEDITHIVYVSSSEIRLPGGDLYLSAKLGLRNDVNRVMLYFLGCYGGVTGLRVAKDIAENNPGSRVLLTTSETTILGFRPPNKARPYDLVGAALFGDGAAAVIIGADPRECEAPFMELHYAVQQFLPGTQNVIDGRLTEEGINFKLGRDLPQKIEENIEEFCKKLMGKAGGDESMEFNDMFWAVHPGGPAILNRLETKLKLEKEKLESSRRALVDYENVSSNTILYVMEYMREELKKKGDAAQEWGLGLAFGPGITFEGLLIRSL
A0A103XI88	A0A103XI88_CYNCS	unreviewed	Chalcone/stilbene synthase, C-terminal	Ccrd_006751	Cynara cardunculus var. scolymus (Globe artichoke) (Cynara scolymus)	413			MSNTNGNGVAERRDSSATRRAPTPGKATVLAIGKAFPSQLIPQDCLVEGYFRDTNCADFAMKEKLARLFLFTSSTIFPFCLFLFLQMVGKTTTVKTRYTVMSKEILDKYPELATEGSPTITQRLDIANQAVTEMAKEASLACIKQWGRPAGDITHVIYVSSSEIRLPGGDLYLASELGLRSDVNRVMLYFLGCYGGVTGLRIAKDIAENNPGSRVLLTTSETTILGFRPPNKSRPYDLVGAALFGDGAAAAIIGADPMTKVESPFMELSFAVQQFLPGTHSVIDGRLSEEGINFKLGRDLPQKIDDNIEGFCQKLMEKAGGLEDFNDLFWAVHPGGPAILNRLETTLKLRGEKLDCSRRALMDFGNVSSNTIIYVMEYMKEELMNRENGEEWGLALAFGPGITFEGILLRNLN
Q3YMX2	Q3YMX2_ARAHG	unreviewed	Chalcone synthase family protein protein	CHS1	Arabidopsis halleri subsp. gemmifera (Arabis gemmifera)	393			MSNSRMNGVEKLSSISTRRVANPRKATLLALGKAFPSQVVPQENLVEGFLRDTKCDDAFIKEKLEHLCKTTTVKTRYTVLSREILDKYPELTTEGSPTIKQRLEIANEAVVEMALEASLGCIKEWGRPVEDITHIVYVSSSEIRLPGGDLYLSAKLGLRNDVNRVMLYFLGCYGGVTGLRVAKDIAENNPGSRVLVTTSETTILGFRPPNKARPYDLVGAALFGDGAAAVIIGADPRECEAPFMELHYAVQQFLPGTQNVIDGRLTEEGINFKLGRDLPQKIEENIEEFCKKLMGKAGGDESMEFNDMFWAVHPGGPAILNRLETKLKLEKEKLESSRRALVDYGNVSSNTILYVMEYMREELKKKGDAAQEWGLGLAFGPGITFEGLLIRSL
Q3YMX5	Q3YMX5_ARAHG	unreviewed	Chalcone synthase family protein protein	CHS1	Arabidopsis halleri subsp. gemmifera (Arabis gemmifera)	393			MSNSRMNGVEKLSSISTRRVANPRKATLLALGKAFPSQVVPQENLVEGFLRDTKCDDAFIKEKLEHLCKTTTVKTRYTVLSREILDKYPELTTEGSPTIKQRLEIANEAVVEMALEASLGCIKEWGRPVEDITHIVYVSSSEIRLPGGDLYLSAKLGLRNDVNRVMLYFLGCYGGVTGLRVAKDIAENNPGSRVLLTTSETTILGFRPPNKARPYDLVGAALFGDGAAAVIIGAAPRECEAPFMELHYAVQQFLPGTQNVIDGRLTEEGINFKLGRDLPQKIEENIEEFCKKLMGKAGGDESMEFNDMFWAVHPGGPAILNRLETKLKLEKEKLESSRRALVDYGNVSSNTILYVMEYMREELKKKGDAAQEWRLGLAFGPGITFEGLLIRSL
Q3YMY0	Q3YMY0_ARAHG	unreviewed	Chalcone synthase family protein protein	CHS1	Arabidopsis halleri subsp. gemmifera (Arabis gemmifera)	393			MSNSRMNGVEKLSSISTRRVANPGKATLLALGKAFPSQVVPQENLVEGFLRDTKCEDAFIKEKLEHLCKTTTVKTRYTVLSREILDKYPELTTEGSPTIKQRLEIANEAVVEMALEASLGCIKEWGRPVEDITHIVYVSSSEIRLPGGDLYLSAKLGLRNDVNRVMLYFLGCYGGVTGLRVAKDIAENNPGSRVLLTTSETTILGFRPPNKARPYDLVGAALFGDGAAAVIIGADPRECEAPFMELHYAVQQFLPGTQNVIDGRLTEEGINFKLGRDLPQKIEENIEEFCKKLMGKAGGDESMEFNDMFWAVHPGGPAILNRLETKLKLEKEKLESSRRALVDYGNVSSNTIMYVMEYMREELKKKGDAAQEWGLGLAFGPGITFEGLLIRSL
A0A060D9S4	A0A060D9S4_MANIN	unreviewed	Chalcone synthase 2 (EC 2.3.1.74) (Chalcone synthases-1)	chs2	Mangifera indica (Mango)	397			MATVSVEEIINAQRAKGPATILAIGTATPANCVYQADYPDYYFRITNSEHKTELKEKFQRMCDKSMIKKRYMHLTEDILKENPNMCAYMAPSLDARQDIVVVEVPKLGKEAAVKAIKEWGQPKSKITHLIFCTTSGVDMPGADYQLTKILGLRPSVKRFMMYQQGCFAGGMVLRFAKDLAENNKGARVLVVCSEITAVTFRGPSDIHLDSLVGQALFGDGAGALIVGSDPDTSIERPLYQIISAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLIAKNIEKSLGEAFTPIGINDWNSIFWVVHPGGPAILDQVEAKLGLKEEKMRATRQVLSDYGNMSSACVLFILDEMRKKSIEEGKPTTGEGLDWGVLFGFGPGLTVETVVLHSVPLAPAAAH
Q3YMX9	Q3YMX9_ARAHG	unreviewed	Chalcone synthase family protein protein	CHS1	Arabidopsis halleri subsp. gemmifera (Arabis gemmifera)	392			MSNSRMNGVEKLSSKSTRRVANAGKATLLAPGKAFPSQVVPQENLVEGFLRDTKCDDAFIKEKLEHLCKTTTVKTRYTVLTREILAKYPELTTEGSPTIKQRLEIANEAVVEMALEASLGCIKEWGRPVEDITHIVYVSSSEIRLPGGDLYLSAKLGLRNDVNRVMLYFLGCYGGVTGLRVAKDIAENNPGSRVLLTTSETTILGFRPPNKARPYDLVGAALFGDGAAAVIIGADPRECEAPFMELHYAVQQFLPGTQNVIDGRLTEEGINFKLGRDLPQKIEENIEEFCKKLMGKAGDESMEFNDMFWAVHPGGPAILNRLETKLKLEKEKLESSRRALVDYGNVSSNTILYVMEYMRDELKKKGDAAQEWGLGLAFGPGITLEGLLIRSL
A0A0K0MWZ	A0A0K0MWZ1_9LAM	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS	Phlomoides rotata	391			MVTVEDIRRAQRAEGPATVLAIGTATPSNCVDQSTYPDYYFRITNSEHMTELKEKFKRMCEKSTINKRYMHLTEDYLKENPNVCAYMAPSLDARQDLVVVEVPKLGKEAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPAVKRFMMYQQGCSAGGTVLRLAKDLAENNAGARVLAVCSEITVISFRGPSEDHLDNLVGQALFGDGAAAVIVGSDPVVGVERPLFQLVSTVQTLLPDCEASINGHLREMGIIFQILKDVPFLISEHIEKSLKEAFDPLGISDWNSIFWIAHPGGPAILNGVEAKLGLEPDKLRSSRHVLSEYGNMLSACVLFVMDEMRKVSANEGRSSTGEGLDWGVLLGFGPGLTIETVVLHSVPITN
E7DZ85	E7DZ85_FAGTA	unreviewed	Chalcone synthase	CHS	Fagopyrum tataricum (Tartarian buckwheat) (Polygonum tataricum)	395			MAPTVQEIRKAQRAEGPATVLAIGTATPPNCVYQADYPDYYFRVTNSDHMTDLKEKFRRMCDKSQIEKRYMYLTEEILKEHPNMCEYMAPSLDSRQDMVVTEVPKLGKEAAQKAIKEWGQPKSKITHVIVCTTSGVDMPGADYQLTKLLGLRPSVKRFMMYQQGCFAGGTVLRMAKDLAENNRGARVLVVCSEITAVCFRGPTDTHLDSMVGQALFGDGAGAVIVGADPDLSVEKPIFELVWTSQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLTEAFSPLNIADWNSLFWIAHPGGPAILDQVEAKLGLKEEKLKATRQVLNDYGNMSSACVLFILDEMRKKSLENGHATTGEGLDWGVLFGFGPGLTVETVVLHSVPTTTLAN
Q9AVB9	Q9AVB9_9LILI	unreviewed	Chalcone synthase (EC 2.3.1.74) protein	LhCHSC	Lilium hybrid division I	197			AGGTVLRLAKDLAENNKSARVLVVCSEITAVTFRGPNDSHLDSLVGQALFGDGAAAIIIGADPDLAVERPLFQLVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPALISKNIEKSLTGAFQPLGISDWNSIFWIAHPGGPAILDQVEERLALKKEKMKATKHVLSEYGNMSSACVLFILDEMRKSSAAEGKATTG
A0A1J3D586	A0A1J3D586_NOCC	unreviewed	Chalcone synthase 2 protein	GA_TR19319_c20_g1_i1_g.63923	Noccaea caerulescens (Alpine penny-cress) (Thlaspi caerulescens)	440			KKKKKKNRMLGVQEKQKEIGQSTRRVANQGKATVLALGKAFPSNLVSQDNLVEEYLREIKCDDLSIKDKLQHLCKLSFTTFLFHPRIRIRIRIRIPKHVFVSCVLVLVVLGKTTSVKTRYTVMTRETLQKYPELATEGSPTIKQRLEIANEAVVQMAYEASLACIKEWGRALQDITHLVYVSSSEFRLPGGDLHLSAQLGLSNEVQRVLLYFSGCYGGLSGLRVAKDIAENNRGSRVLLTTSETMVLGFRPPNKARPYDLVGATLFGDGAAALIIGADPRESESPFMELHCALQRFLPGTQGVIDGRMSEEGISFKLGRDLPQKIEDSVEDFCKKLVAKAGPAASASLELNDLFWAVHPGGPAILNGMEARLKLKPEKLESSRRALVDYGNVSSNTIFYIMDKVRDEFEKKGRGGGPEWGLGLAFGPGITFEGFLMRSLF
A0A0H3WFJ5	A0A0H3WFJ5_9MON	unreviewed	Chalcone synthase (EC 2.3.1.74)	CHS2	Dryopteris fragrans	369			MVSHGCFKASARKVERADGPATVLAIGTATPPNVFPQRDYAEFYFNITNSNHMTELKEKFHRMCQESGINKRYLYVNEELLKANPSMCAHSERSLDVRQDIVVEVPKLGKEAAMKAIKEWGQPKSKTTHLIFCTSSGLDMPGADWALTKLLGLAPTVKRVMLYSPGCYAGGKVMRIAKDLAENNKGARVLAVCSEITAMSFRGPSDKQIDNLVGQALFGDGASAAIIGADPIPQVERAWFELQFAASNILPESDGAVDGHIREVGLMIHLRKDVAGLIGKNIGSVLKDAFAKVFGANAPSFNDLFWITHPGGPAILDQVEQKLQLKPEKMAPSRHVLFEYGNTISSCVFFIMDHIRKKSVAQKCSHLW
G9F7Y1	G9F7Y1_CURLO	unreviewed	Chalcone synthase-like protein (EC 2.3.1.74) protein	ClPKS7	Curcuma longa (Turmeric) (Curcuma domestica)	189			NNRGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAGAVIVGADPDPETERPLFELVSASQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFAPLGIDDWNSIFWIAHPGGPAILDQVEAKLALEKEKMAATRQVLSEYANMSSACDLFILDEMRRRSAEEGKATTGDGFNWG
B5LXY1	B5LXY1_GOSHI	unreviewed	Chalcone synthase (chalcone synthase 1-like)	LOC107914895 CHS5	Gossypium hirsutum (Upland cotton) (Gossypium mexicanum)	389			MVTVEEVRKAQRAQGPATVLAIGTSTPPNCVDQSTYPDYYFRITNSEHKTELKEKFKRMCEKSMIKKRYMYLTEEILKENPNVCEYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRVAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIIGADPMPEIEKPMFEIVSVAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFQPLGISDWNSLFWIAHPGGPAILDQVEAKLALKPEKLRATRHVLSEYGNMSSACVLFILDEMRKKSREDGLQTTGEGLEWGVLFGFGPGLTVETVVLHSVAA

O22519	O22519_VITVI	unreviewed	Chalcone synthase	CHS	Vitis vinifera (Grape)	454			MVSVGEIRKSQRAEGPATVLAIGTATPANCVYQADYPDYYFRITNSEHMTELKEKFKRMCDKSMINKRYMHLNEEILKENPNVCAYMAPSLDARQDMVVVEVPKLGKEAAVKAIKEWGQPKSKITHLVFCTTSGCDMPGADYQLTKLLGLKPSVKRLMMYQQGCFAGGTVLRLAKDLAENNAGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAIIIGADPDTKIERPLFELVSAAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFKPIGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKLRATRHVLSEYGNMSSACVLFILDEMRKKSIEEGKGTTGEGLEWGVLFGFGPGLTVETVVVAQPCYTIDSLSHSSTYNTEGKMGMAAALGTGEDCMSSCANLRSYPSFLCYVLLYFYVLLCPCAFSPFTLK
G3G7Z0	G3G7Z0_GOSHI	unreviewed	Chalcone synthase (chalcone synthase 2-like)	CHS4 LOC107914136	Gossypium hirsutum (Upland cotton) (Gossypium mexicanum)	392			MGSLDTMNENSRGRAAVLAIGTANPPHCFNQVDYPDFYFRVTKSHHLTSLKDKFRRICEKSAIRKRYMHLTEDIINKNPNLIIYKAPSFDARQEILVTEVPKLGKDAALKAIKEWGQPISNITHLIVCTSSGIDMPAADHQLAKLIGLKSSVQRFMLYQQGCFAAGTALRLAKDLAENNPGARVLAVCSEIMVGSFQPPSETHLDVLVGSALFSDGAAAVIVGANPNATINERPLFQIVSAKQAVIPDSDDVIIAKIREMGMAYYLSKKLPNVIANNIEQCLFETLGPCGVDDWNKLFYVVHPGGPAVLKRIEEKLGLGSDKLKASWHVLSEYGNMWSPSVLFVLDEMRKRSTEEGKTAAATEGLEWGVLLAFGPGLTVETVVLRSIAADSA
D7MDI2	D7MDI2_ARALL	unreviewed	Chalcone synthase family protein	ARALYDRAFT_491172	Arabidopsis lyrata subsp. lyrata (Lyre-leaved rock-cress)	392			MGSIDAAVLGSVKKSNPGKATILALGKAFPHQLVMQEYLVDGYFKTTNCDDPELKQKLTRLCKTTTVKTRYVVMSEEILKKYPELAIEGGSTVTQRLDICNDAVTEMAVEASRACIKNWGRSISEITHLVYVSSSEARLPGGDLYLAKGLGLSPDTHRVLLYFVGCSGGVAGLRVAKDIAENNPGSRVLLATSETTIIGFKPPSVDRPYDLVGVALFGDGAGAMIVGSDPDPICEKPLFELHTAIQNFLPDTEKTIDGRLTEQGINFKLARELPQIIEDNVENFCKKLIGKAGLAHKNYNQMFWAVHPGGPAILNRMEKRLNLSPEKLSPSRRALMDYGNASSNSIVYYVLEYMLEESKKVRNMNEEEDEWGLILAFGPGVTFEGIIARNLDV
A7L2Z4	A7L2Z4_GOSHI	unreviewed	Chalcone synthase	CHS CHS7	Gossypium hirsutum (Upland cotton) (Gossypium mexicanum)	389			MVTVEEVRKAQRAQGPATVLAIGTSTPPNCVDQSTYPDYYFRITNSEHKTELKEKFKRMCEKSMIKKRYMYLTEEILKENPNVCEYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRVAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIIGADPVPEIEKPMFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFQPLGISDWNSLFWIAHPGGPAILDQVEAKLALKPEKLRATRHVLSEYGNMSSACVLFILDEMRKKSREDGLQTTGEGLEWGVLFGFGPGLTVETVVLHSVAA
A0A1U8ISL4	A0A1U8ISL4_GOSH	unreviewed	chalcone synthase 1-like	LOC107897842 LOC107897841	Gossypium hirsutum (Upland cotton) (Gossypium mexicanum)	389			MVTVEEVRKAQRAQGPATVLAIGTSTPPNCVDQSTYPDYYFRITNSEHKTELKEKFKRMCEKSMIKKRYMYLTEEILKENPNVCEYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRVAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIIGADPVPEIEKPMFEIVSVAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFQPLGISDWNSLFWIAHPGGPAILDQVEAKLALKPEKLRATRHVLSEYGNMSSACVLFILDEMRKKSREDGLQTTGEGLEWGVLFGFGPGLTVETVVLHSVAA
B6Z259	B6Z259_WHEAT	unreviewed	Chalcone synthase		Triticum aestivum (Wheat)	422			MVSARDVDTTTAANKQQQATCLAPNPGKATILALGHAFPQQLVMQDYVVEGFMRNTNCKDPELKEKLTRLCKTTTVKTRYVVMSDEILKSYPELAQEGLPTMKQRLDISNKAVTQMATEASLACIKAWGGDLSAITHLVYVSSSEARFPGGDLHLARALGLSPDVRRVMLAFTGCSGGVAGLRVAKGLAESCPGARVLLATSETTVAGFRPPSPDRPYDLVGVALFGDGAGAAVVGTDPTDQERPLFELYSALQRFLPDTEKTIDGRLTEEGIKFQLGRELPHIIEAHVESFCQKLIKEHPAAAAAEGDHGEQLTYDKMFWAVHPGGPAILTKMEGRLGLDGGKLRASRSALRDFGNASSNTIVYYVLENMVEESRQRTEAPEPEPEPEGGQECEWGLILAFGPGITFEGILARNLQARLGAN
M4VMV7	M4VMV7_CAMSI	unreviewed	Chalcone synthase	CHS1 CHSa	Camellia sinensis (Tea)	389			MVTVEDIRRAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSEHKAELKEKFKRMCDKSMIKKRYMYLTEEILKENPQVCEYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAIIVGSDPIPEVEKPLFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLAEAFQPLGISDWNSLFWIAHPGGPAILDQVELKLGLKEEKLRATRHVLSEYGNMSSACVLFILDEMRKKSAADGLKTTGEGLEWGVLFGFGPGLTVETVVLHSLST
M4VKY3	M4VKY3_CAMSI	unreviewed	Chalcone synthase	CHS2 CHSc	Camellia sinensis (Tea)	389			MVTVEEVRRAQRAEGPATVMAIGTATPPNCVDQSTYPDYYFRITNSEHKTELKEKFQRMCDKSMIKKRYMYLTEEILKENPNVCAYMAPSLDARQDMVVVEVPKLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDAHLDSLVGQALFGDGAAAIIVGSDPIPEVEKPLFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLNEAFQPLNITDWNSLFWIAHPGGPAILDQVELKLALKPEKLRATRHVLSEYGNMSSACVLFILDEMRKSSAKKGLKTTGEGLDWGVLFGFGPGLTVETVVLHSVST
K9MUA0	K9MUA0_MALDO	unreviewed	CHS2 chalcone synthase (EC 2.3.1.74)		Malus domestica (Apple) (Pyrus malus)	389			MVTVEEVRKAQRAEGPATVLAIGTATPSNCVDQATYPDYYFRITNSEHKTELKEKFQRMCDKSMIKKRYMYLTEEILKENPTVCEYMAPSLDARQDMVVVEVPRLGKEAATKAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPYVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSDTHLDSLVGQALFGDGAAAVIIGADPLPEVEKPLFELVSAAQTILPDSDGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLNEAFKPIGISDWNSLFWIAHPGGPAILDQVESKLALKPEKLEATRQVLSNYGNMSSACVLFILDEVRRKSTEKGLRTTGEGLEWGVLFGFGPGLTVETVVLHSVAA
A0A0N6XTG9	A0A0N6XTG9_LYCR	unreviewed	Chalcone synthase protein	CHS1 CHS	Lycium ruthenicum (Black goji berry)	389			MVTVEEYRKAQRAEGPATVMAIGTATPSNCVDQSTYPDYYFRITDSEHKTELKEKFKRMCDKSMIKKRYMHLTEEILKENPNMCAYMAPSLDARQDIVVVEVPKLGKEAAQKAIKEWGQPKSKITHLVFCTTSGVDMPGCDYQLAKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNKGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAAAIIMGSDPIPGVERPLFELVSAAQTLLPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKSLVEAFQPLGISDWNSLFWIAHPGGPAILDQVELKLALKPEKLRATRDVLSNYGNMSSACVLFILDEMRKASAKEGLGTTGEGLEWGVLFGFGPGLTVETVVLHSVAA
G3FJ87	G3FJ87_9ASPA	unreviewed	Chalcone synthase (EC 2.3.1.74)		Freesia hybrid cultivar	389			MVNVEEIRKAQRAEGPAAILAIGTATPPNAIEQSEYPDYYFRVTNSEDKVELKEKFKRMCEKSMIKKRYLYLTEDILKENPNVCAYMATSLDARQDMVVVEVPKLGKEAATRAIKEWGQPKSKITHLVFCTTSGVDMPGADYQLTKLLGLRPSVKRLMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAVTFRGPSESHLDSLVGQALFGDGAAALIVGSDAIEGIERPIFEMVSAAQTILPDSEGAIDGHLREVGLTFHLLKDVPGIISKNIEKSLEEAFKPLGITDYNSLFWIAHPGGPAILDQVEAKIGLKPEKLRATRHVLSEYGNMSSACVLFILEEMRKKSAEEKNGTTGEGLEWGVLFGFGPGLTVETVVLHSVEA
B9GPA0	B9GPA0_POPTR	unreviewed	Chalcone and stilbene synthase family protein	POPTR_0002s14240g POPTR_0955s00200g	Populus trichocarpa (Western balsam poplar) (Populus balsamifera subsp. trichocarpa)	388			MSESDSNGASKHCTTPSRRAPTLGKATLLAIGKAFPSQLIPQECLVEGYIRDTKCDDASIKEKLERLCKTTTVKTRYTVMSREILDKYPELATEGTPTIRQRLEIANPAVVEMALKASMACINEWGGSVEDITHIVYVSSSEVRLPGGDLYLASQLGLRNDVGRVMLYFLGCYGGVTGLRVAKDIAENNPGSRVLLTTSETTILGFRPPSKARPYDLVGAALFGDGAAAVIIGANPVIGKESPFMELNYSVQQFLPGTQNVIDGRLSEEGIHFKLGRDLPQKIEDNIEEFCNKLMSKAGLTDFNELFWAVHPGGPAILNRLESKLKLNEEKLECSRRALMDYGNVSSNTIVYVLEYMRDELKRGGGEWGLALAFGPGITFEGILLRSL
B2MWQ9	B2MWQ9_POLCS	unreviewed	CHS2 (Type III polyketide synthase)	CHS2 PKS3	Polygonum cuspidatum (Japanese knotweed)	393			MAPAVADIRKAQRAEGPATVLAIGTATPPNCVYQKDYPDYYFRVTNSDHMTDLKEKFRRMCEKSNIEKRYMYLTEEILKENPNMCSYMQTSSLDTRQDMVVSEVPRLGKEAAQKAIKEWGQPKSKITHVIMCTTSGVDMPGADYQLTKLLGLHPSVKRFMMYQQGCFAGGTVLRLAKDLAENNRGARVLVVCSEITAICFRGPTDTHPDSMVGQALFGDGSGAVIIGADPDLSIEKPIFELVWTAQTILPDSEGAIDGHLREVGLTFHLLKDVPGLISKNIEKNLTEAFSPLNVSDWNSLFWIAHPGGPAILDQVETKLGLKEEKLKATRQVLNDYGNMSSACVLFIMDEMRKKSVENGHATTGEGLEWGVLFGFGPGLTVETVVLHSVPVAN