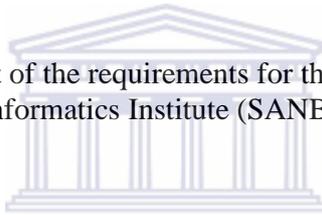A KNOWLEDGEBASE OF STRESS REPONSIVE GENE REGULATORY ELEMENTS
IN *Arabidopsis thaliana*

MUHAMMED SALEEM ADAM

Thesis submitted in fulfillment of the requirements for the degree of Magister Scientiae, at the
South African National Bioinformatics Institute (SANBI), University of the Western Cape.

May 2011

Supervisors: Prof. Vladimir Bajic and Prof. Alan Christoffels

KEYWORDS

Arabidopsis thaliana

Abiotic stress

Binding site

Biotic stress

Client-Server

Context model

Curation

Database design
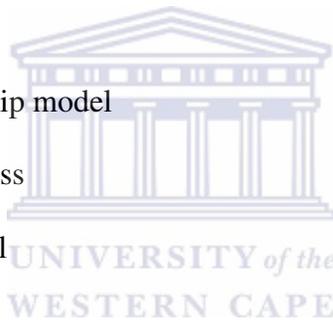
Defence

Entity relationship model

Modelling process

Relational model

Thin client

Transcription factors

Transcription factor family

Stress response

ABSTRACT

Stress responsive genes play a key role in shaping the manner in which plants process and respond to environmental stress. Their gene products are linked to DNA transcription and its consequent translation into a response product. However, whilst these genes play a significant role in manufacturing responses to stressful stimuli, transcription factors co-ordinate access to these genes, specifically by accessing a gene's promoter region which houses transcription factor binding sites. Here transcriptional elements play a key role in mediating responses to environmental stress where each transcription factor binding site may constitute a potential response to a stress signal.

Arabidopsis thaliana, a model organism, can be used to identify the mechanism of how transcription factors shape a plant's survival in a stressful environment.

Whilst there are numerous plant stress research groups, globally there is a shortage of publicly available stress responsive gene databases. In addition a number of previous databases such as the Generation Challenge Programme's comparative plant stress-responsive gene catalogue, Stresslink and DRASTIC have become defunct whilst others have stagnated.

There is currently a single *Arabidopsis thaliana* stress response database called STIFDB which was launched in 2008 and only covers abiotic stresses as handled by major abiotic stress responsive transcription factor families. Its data was sourced from microarray expression databases, contains numerous omissions as well as numerous erroneous entries and has not been updated since its inception.

The Dragon Arabidopsis Stress Transcription Factor database (DASTF) was developed in response to the current lack of stress response gene resources. A total of 2333 entries

were downloaded from SWISSPROT, manually curated and imported into DASTF. The entries represent 424 transcription factor families. Each entry has a corresponding SWISSPROT, ENTREZ GENBANK and TAIR accession number. The 5' untranslated regions (UTR) of 417 families were scanned against TRANSFAC's binding site catalogue to identify binding sites.

The relational database consists of two tables, namely a transcription factor table and a transcription factor family table called DASTF_TF and TF_Family respectively.

Using a two-tier client-server architecture, a webserver was built with PHP, APACHE and MYSQL and the data was loaded into these tables with a PYTHON script. The DASTF database contains 60 entries which correspond to biotic stress and 167 correspond to abiotic stress while 2106 respond to biotic and/or abiotic stress.

Users can search the database using text, family, chromosome and stress type search options. Online tools have been integrated into the DASTF database, such as HMMER, CLUSTALW, BLAST and HYDROCALCULATOR. User's can upload sequences to identify which transcription factor family their sequences belong to by using HMMER.

The website can be accessed at http://apps.sanbi.ac.za/dastf/ and two updates per year are envisaged.

DECLARATION

DECLARATION

I declare that "A KNWOLEDGEBASE OF STRESS REPONSIVE GENE REGULATORY ELEMENTS IN *Arabidopsis thaliana* "is my own work, that it has not been submitted for degree or examination at any other university, and that all the sources I have used or quoted, and all work which was the result of joint effort, have been indicated and acknowledged by complete references.

MUHAMMED SALEEM ADAM   May 2011

Signed: _____

UNIVERSITY *of the*
WESTERN CAPE

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

CONTENTS   **Page**

LIST OF FIGURES

UNIVERSITY *of the*
WESTERN CAPE

LIST OF TABLES

ABBREVIATIONS

5'UTR	Five prime untranslated region
ABA	Abscisic acid
AGI	Arabidopsis Genome Initiative
AP2-EREBP	APETALA 2/ Ethylene responsive element binding protein
bHLH	Basic helix loop helix
BLAST	Basic local alignment search tool
BZIP	Basic leucine zipper
ChIP	Chromatin immunoprecipitation
ChIP-Seq	Chromatin immunoprecipitation sequencing
CSS	Combined consensus scale
DAMPD	Dragon Antimicrobial Peptide Database
DASTF	Dragon Arabidopsis Stress Transcription Factor Database
DATF	Database of Arabidopsis Transcription Factors
DBD	DNA binding domain
DNA	Deoxyribonucleic acid
DRASTIC	Database for the Analysis of Signal Transduction in Cells
ET	Ethylene
ERD	Entity relationship diagram
GCP	Generation Challenge Programme
GO	Gene ontology
HMMER	Hidden markov marker and profiles
HTML	Hypertext markup language
IDA	Inferred from direct assay
IEA	Inferred from electronic annotation
JA	Jasmonic acid
KEGG	Kyoto Encyclopedia of Genes and Genomes
M:N	Many to many
MADS	Minichromosome maintenance protein1 agamous deficiens serum response factor
NAC	No apical meristem *Arabidopsis thaliana* transcription factor cup-shaped cotyledon 2

NAR	Nucleic Research Acids
PlnTFDB	Plant Transcription Factor Database
PE	Protein existence
PUBMED	Public Medline
SA	Salicylic acid
SQL	Structured query language
STIFDB	Stress Responsive Transcription Factor Database
RARTF	RIKEN Arabidopsis Transcription Factor Database
TAIR	The Arabidopsis Thaliana Information Resource
TFBS	Transcription factor binding site
TBP	TATA-binding protein
UTR	Untranslated region

UNIVERSITY *of the*
WESTERN CAPE

CHAPTER ONE

INTRODUCTION

CONTENTS PAGE

INTRODUCTION

A plant's response to the encroachment of a wide range of environmental stresses is mediated by its sessile nature which essentially shapes its ability to respond to these factors. These demands which negatively affect a plant's metabolic functioning, growth and adaptive capacity are referred to as stress. Stress may thus be viewed as a condition effected by a stressor (singularly or in combination) leading to a stress response which may result in damage such as cell death, organ incapacitation and permanent tissue and organ damage (Lichtenthaler 1995). Furthermore, a plant's response to stress-related events tacitly assumes a physiological norm. That is, a basal level implying environmental conditions supplying required quantities of water, light, temperature and nutrients under which a plant is able to thrive. An occurrence of stressors or stress events above a particular threshold or physiological norm would thus constitute a significant deviation and hence activates the need for a significant response. For example, a substantial increase in the intensity of environmental temperature requires a consequent change in the organism's perception and classification of the respective heat-related stimuli. That is, the level of stress signaling as indicated by responsive signal transduction mechanisms has exceeded a basal limit to a point where the increase in demand intensity can no longer be compensated for by conventional responses. At this point, stress is no longer perceived as something competitive but as hostile and is processed as an attack which requires appropriate defence operators and mechanisms (Lichtenthaler 1998). At this point a distinction should be made between biotic and abiotic stress. Stress events such as heat, cold, water deprivation, salinity, flooding, pesticides and soil nutrient depletion are classified as abiotic stress, that is, non-live stress

entities. Entities such as viruses, pathogens, necrophytes, biotrophs and insects are classified as biotic or live stressors that are capable of responding to a host's defence strategies.

1.1 TRANSCRIPTIONAL REGULATION IN PLANTS

Transcription is the process whereby response mechanisms coded in a native language or instruction set are operationalised in relation to environmental stress stimuli. The system operates via a network of reception, activation and amplification via signal transduction components where specific genes have been proposed to have stimuli-specific responses (Rodríguez, Canales and Borrás-Hidalgo 2005).

Many biological processes in a plant are regulated at the level of transcription, indicating the manner in which genes are expressed in relation to environmental stimuli (McGinley 2000). Changes in gene expression have been shown to underlie responses to environmental cues and stresses such as changes in light, temperature, nutrient availability and defence responses to pathogens (Aarts and Fiers 2003). The above responses are mirrored in an intricate network of components and mechanisms where signaling and transduction systems operate, activating/deactivating and shuttling response cues to various parts of the plant (Mahajan and Tuteja 2005). Furthermore, whilst the mechanisms of transcription are largely common across eukaryotes, their components vary among kingdoms (Arabidopsis Genome Initiative 2000; Riechmann et al., 2000).

Transcription factors are protein complexes that can shape the relationship between an organism and its response to environmental stress by influencing (initiate, enhance or inhibit) the transcription of specific genes. RNA polymerase is the enzyme that

transcribes genes to make messenger RNA, which is used to generate the necessary response-proteins required by the system. Transcription factors assist RNA polymerase to bind to specific segments of DNA in an area known as the promoter region. The promoter is a regulatory region of DNA which marks the target site where RNA polymerase binds and also known as the transcription start site.

Every gene has a promoter region, however in the case of eukaryotes, its location may vary. For example, some promoters are located towards the three prime (3') region of the gene. The binding of RNA polymerase to the start site initiates the transcription process where instructions are written to the coding region of a gene. In eukaryotes, the promoters of many (but not all) genes contain the sequence TATAA, also known as the "TATA box". This region is twenty-five to thirty nucleotides upstream from the transcription start site. This sequence, in turn, is recognized by the TATA-binding protein (TBP) (Riechman 2002). The TBP binds to the sequence thereby marking the start site of transcription. By controlling RNA polymerase's access to the gene, transcription factors control the rate at which a gene is transcribed, thereby effectively regulating the rate at which genes are expressed. This control of the transcription process is underscored by the relationship between transcription factors and DNA cis-regulatory elements occurring upstream in the five prime untranslated region (5' UTR) of a gene. That is, a gene's responsiveness to certain stimuli is a result of their predisposition or 'hard wiring' to these cis-regulatory elements (Zhang et al., 2005).

However, experimental data on these binding specificities are scant and with respect to *Arabidopsis thaliana,* approximately three percent of these binding sites have been determined experimentally (Schröder et al., 2010). Furthermore, wet lab based

determination of transcription factor binding sites (TFBS) using traditional DNA footprinting and chromatin immunoprecipitation (ChIP) technologies such as CHIP-sequencing (ChIP-seq) can be time consuming and expensive (Grau et al., 2005; Chan et al., 2011). Hence numerous binding site prediction tools such as MATCH (Kel et al., 2003), VOMBAT (Grau et al., 2005) and PROMOTERSWEEP (Val et al., 2009) have been developed.

1.2 TRANSCRIPTION FACTORS

Transcription factors operate in an environment sensitive fashion, contingent on the tissue or cell-type that they occur in and the environmental stimuli that trigger their activation. They possess discrete DNA binding domains which are functionally specific and are drawn from a limited set of motifs (Latchman 1997).

These domains mediate sequence specific binding features by recognizing and matching an attendant, compatible series of nucleotide bases. It is the latter selective DNA recognition process and hence response generation process which underlies the cue-specific responses of an organism (Riechmann and Ratcliffe 2000). That is, they generate sequence specific control over transcription factor bindings thereby eliciting timeous, context sensitive and response specific results. It is in this context that the transcription factor binding sites are referred to as response elements. DNA binding domains (DBD) are used to either bind directly to DNA or as part of a large protein complex. DNA binding domains possess structural motifs or recurring elements which are used to divide transcription factors into classes (also known as superclass), families and subfamilies (Riechmann et al., 2000). Furthermore, individual family or sub-family members may

play contrasting roles. For example, some members may play an activating role whilst others play repressive roles. Based on the structural motif concept, five main structural classes or superclasses can be identified namely, basic domain Helix-turn-Helix domain, zinc coordinating domain, beta scaffold with minor groove contacts domain and other domains (Table 1.1) (Stegmaier, Kel and Wingender 2004).

Transcription factor families which have been implicated in *Arabidopsis thaliana* stress response activities are AP2-EREBP, BZIP, bHLH, HSF, MYB and MYB-related, NAC and WRKY (Glazebrook 1999; Singh, Foley and Onate-Sánchez 2002; Eulgem 2005; Eulgem et al., 2005; Varshney and Koebner 2007; Bu et al., 2008; Van Verk, Gatz and Linthorst 2009).

Transcription factors also use signal transduction pathways as messaging mechanisms to respond to stress related signals, specifically, pathways related to the hormones salicylic acid (SA), jasmonic acid (JA) and ethylene (ET) (Numchuk et al., 2003; Kwon 2010). Ethylene is involved in various developmental processes, such as plant growth and fruit ripening. Besides these processes, ethylene is also involved in environmental stress signaling upon wounding or pathogen attack. Jasmonic acid is induced in defence related activities in response to necrotrophic pathogens and herbivorous insects whilst salicylic acid plays a central role in recognizing pathogen signatures and is induced after attack by biotrophic pathogens (Dong et al., 1991; Chung et al., 2008).

Furthermore, accompanying the increasing body of literature reporting on stress responses in plants, is a plethora of genes encoding biotic and abiotic stress-related genes (Hirt and Shinozaki 2004; Jenks and Hassegawa 2006; Rao, Raghavendra and Reddy

2006; Hirayama and Shinozaki 2010; Pareek et al., 2010). Hence the need for a database which collects and integrates this information.

1.2.1 TRANSCRIPTION FACTOR DATABASES

There are *A. thaliana*-specific transcription factor databases such as DATF (Database of Arabidopsis Transcription Factors) and Athamap (Steffens et al., 2005) as well as plant transcription factor databases which have an *A. thaliana* sub-category. These include the German PlnTFDB (Plant Transcription Factor Database) (Perrez-Rodrıguez et al., 2010) the Chinese PlnTFDB (Plant Transcription Factor Database) (Zhang et al., 2011) and RARTF (RIKEN Arabidopsis Transcription Factor database) (Iida et al., 2005). These databases do not focus on stress related issues but on transcription factors and their families. The Plant Stress Gene Database is a cross-species database that has an *A. thaliana* sub-section but only contains 33 entries and does not have any associated analytical tools.

There is currently one *A. thaliana* -specific stress related database namely, Stress Responsive Transcription Factor Database (STIFDB). It contains 2629 entries which have been sourced from public microarray related databases. Genes which were significantly upregulated in response to abiotic stresses such as cold, salinity, light and water were selected as candidates for their database and organized in transcription factor families (Shameer et al., 2008). STIFDB's methodology for its binding site prediction algorithm was based on identifying ten families known to be involved in abiotic stress response which in turn led to the creation of a set of 22 Hidden Markov models (HMM). The upstream sequence including the 5' UTR of each gene was scanned using these

models to predict a binding site for each transcription factor. The database features the sequence alignment tool BLAST and has not been updated since its inception in 2008.

1.3 DATABASE DESIGN MODELLING PROCESS

A database design requires a context model that represents the most salient aspects of the problem. In other words, the context model serves as the starting point for the iterative identification of objects and relationships constituting the basis of an information model. The process starts with understanding a client's problem context with a view to eliciting information which serves as input to the development of a specification. This, in turn, is interpreted using data model-based conceptual and notational techniques (Avison and Fitzgerald 1995).

Information elicitation is an iterative process which involves continuous interaction with users and their target context. Information may be captured and conveyed by various means. Human language is the most important conveyor and repository of meaning. However, meanings, derived from language-use can be vague and ambiguous. Context can play an important role, leading to the interpretation of words in varied and contradictory ways. Information, to a certain extent, is also dependant on the user and their understanding of their needs. That is, analysts will glean information filtered through the client's interpretative lenses. Furthermore, a client is not a monolithic entity, but rather refers to the various role players belonging to the single problem context. Hence, there is a variety of perspectives which may emerge (from a single context) affecting the quality of the information produced (Satzinger, Burd, and Jackson 2002). The sources of information may vary from verbal descriptions to various documentary

pieces of information describing the current context. Evolving from the process is the encapsulation of the information in terms of a model, specifically, a context model. The model serves as a 'concrete' basis for further iterative refinements but most importantly a base around which central objects and their relationships are identified (Easterbrook 1993).

In the bioinformatics community, Ontologies such as Gene Ontology (GO) have specifically been developed to minimize the problems related to communication and reference. GO terms ultimately aim at developing a standardized vocabulary whereby a common set of terms and their respective meanings can be used across the vast spectrum of biological researchers and their respective communities (Helden et al., 2000; Harris et al., 2004).

The context model (Figure 1.1) attempts to capture some of the basic aspects of transcription initiation-regulation. The stimuli processing model indicates the rudimentary path of stimuli reception via signal transduction mechanisms to the transcription factor machinery which is implicated in chromatin remodelling. This process opens a path to the promoter region of DNA which houses transcription factor binding sites. Once transcription factors bind to these sites it initiates the transcription process. Furthermore, a second path is related to certain transcription factors which, irrespective of the chromatin packaging are able to recognize and bind to their respective target binding sites. Hence the core objects under investigation can be identified, namely, transcription factors, binding sites, promoter, DNA and the relational aspect of transcription factors binding to binding sites. These objects operate in a biological context and as such are located in a network of other biological objects which assist in

developing a more complete picture of the target context, namely, the biological information model (Figure 1.2).

The development of context model also initiates the model conversion process, whereby an initial non-technical model is gradually transformed into a database design. The information model (Avison and Fitzgerald 1995) basically identifies and abstracts biological objects to the level of a sequence concept, namely, a sequence of nucleotides and amino acids. Furthermore, it assists in identifying additional objects, which, in this instance is the aggregate or family object formed on the basis of identifiable motifs/domains or shared sequence or structural signatures. The model also introduces non-biological objects such as the annotation object which is predicated on GO terms and journal publications. The basic idea behind the annotation concept is that all information in this model is ultimately referenced by publications.
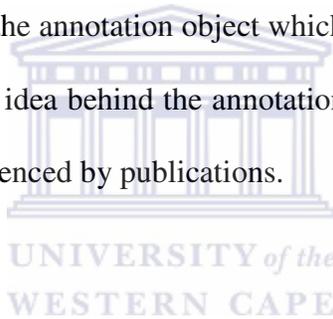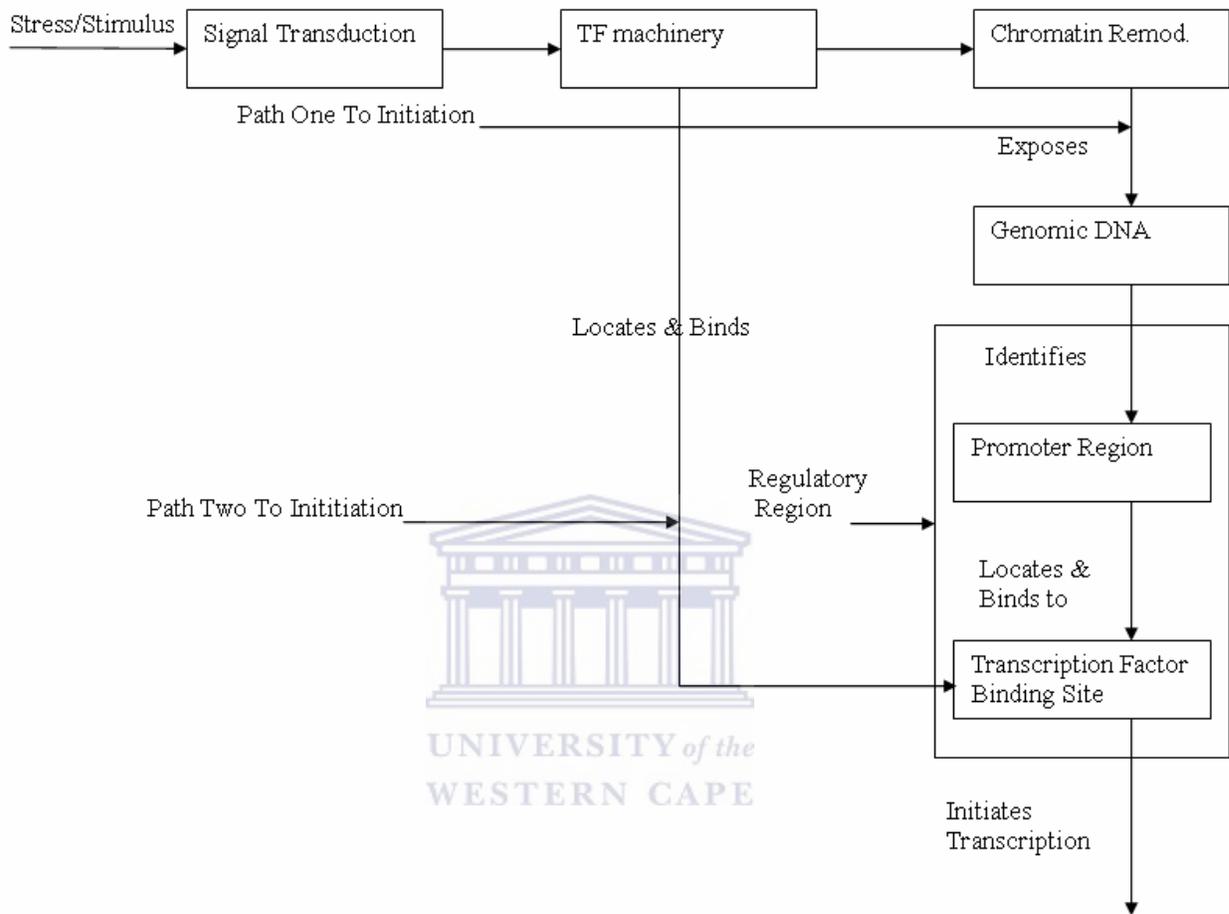
Figure 1.1: Context model for transcription factor initiation
The context model captures some basic aspects of transcription initiation regulation. Firstly, the rudimentary path of stimuli reception via signal transduction mechanisms to the transcription factor machinery which is implicated in chromatin remodelling. Secondly, chromatin remodelling opens a path to the promoter region of DNA which houses transcription factor binding sites. Transcription factors bind to these sites and initiate the transcription process.
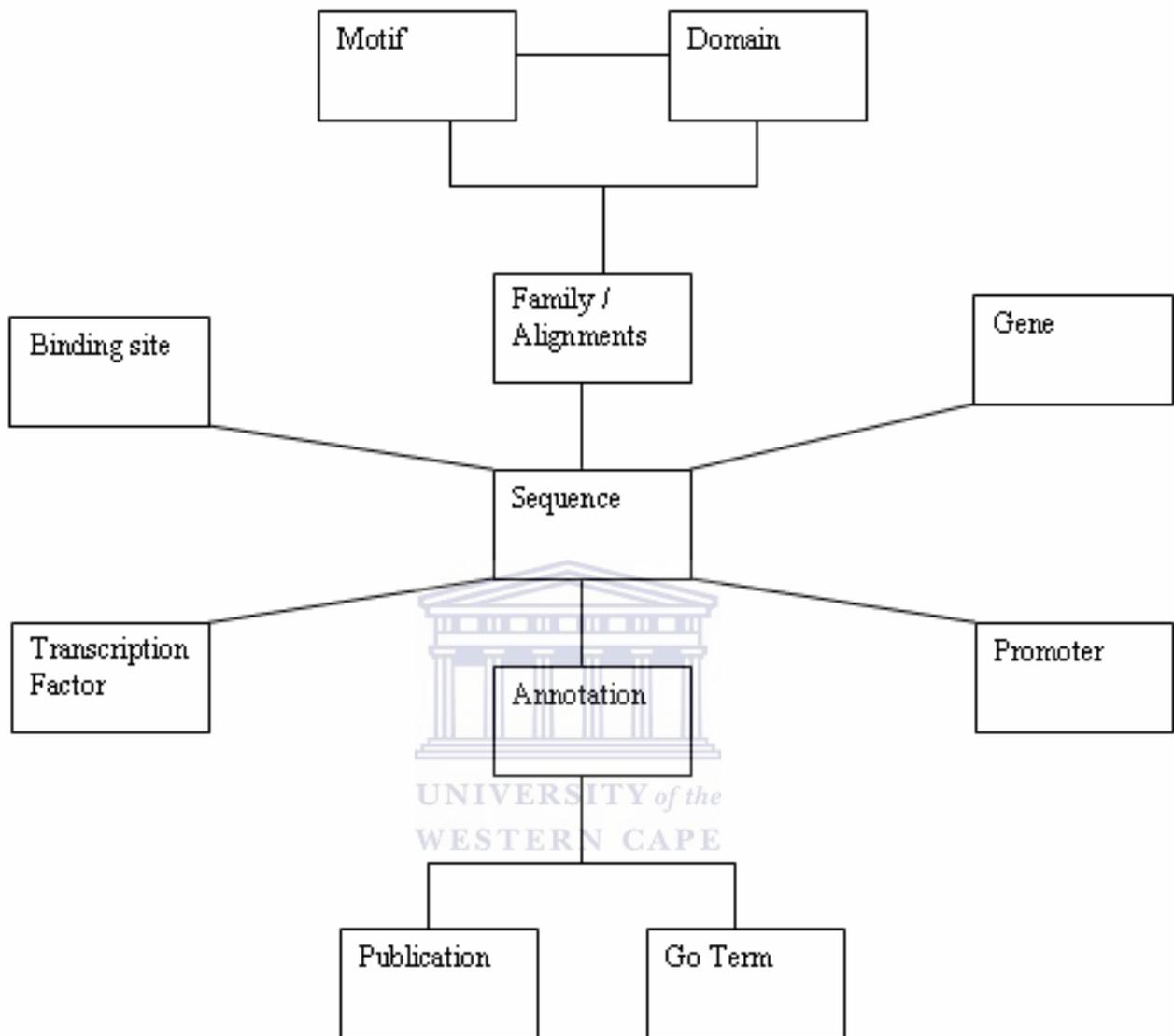
Figure 1.2: Biological information model
This model describes the basic information requirements for each relationship and its participants. For each transcription factor, a series of property and object relationship dimensions are produced.

1.4 THE CONCEPTUAL MODEL

There is a difference between the relational model and the entity-relational model. The relational model was developed by Codd in 1969 whilst the entity-relational model developed by Chen in 1976 (Silberschatz, Korth and Sudarshan 2005). The entity-relational model is a way of capturing/modeling a reality in terms of objects and their dependencies. It is amenable to a variety of contexts and as such has found particular use in systems analysis and design methodologies. This entity relational model, which uses entity relational diagrams (ERD) as a means of depiction, can then be led along different logical modeling paths including a path leading to Codd's relational model. That is, the ERD approach is used to diagrammatically map a logical path or argument to particular design structure at a particular level. In other words, it may be viewed as depicting the logical structure of the premises leading to a particular relational database design (Whitten and Bentley 2002).

The 'implementation' detail of the structure is then administered by applying the principles of the relational model. In terms of Codd's model, reality is described in terms of records expressed as a tabular structure consisting of rows and fields where fields are object properties or object relationship properties and rows are populated with instances of object values.

A conceptual model (Figure 1.3) attempts to map the entire background to the network of biologically related objects from the association between transcription factor-gene-binding site relationships to the gene's production of a response protein. Here, the conceptual model has been appropriated to represent the biological flow of information whilst located within the background of molecular biology's 'Central Dogma' (Keet

2003). The latter constitutes the business rules which determine the 'objects of interest' and the rules which bind these 'objects of interest' as relationships (Nelson, Reisinger, and Henry 2003).

ERDs use a technique where entities and their properties represent the 'objects of interest' about which information is stored and relations depict the associative links between entities (Chen and Carlis 2003). The numerical additions indicate cardinality or the level of the relationship between entities. For example, entities which start with a number greater than zero indicate that they are conceptually, mandatory participants in a relationship. Furthermore, a number greater than one indicates the 'many' side aspect of the participant. A many to many relationship is depicted as 'M:N'. Essentially, it means that the relationship between participating entities have numerous associative permutations or are conceptually non-limited in terms of their combinations. For example, any gene can theoretically, produce numerous proteins and proteins in turn may be associated with numerous participating genes, thus leading to a many to many relationship (Bornberg-Bauer and Paton 2002).
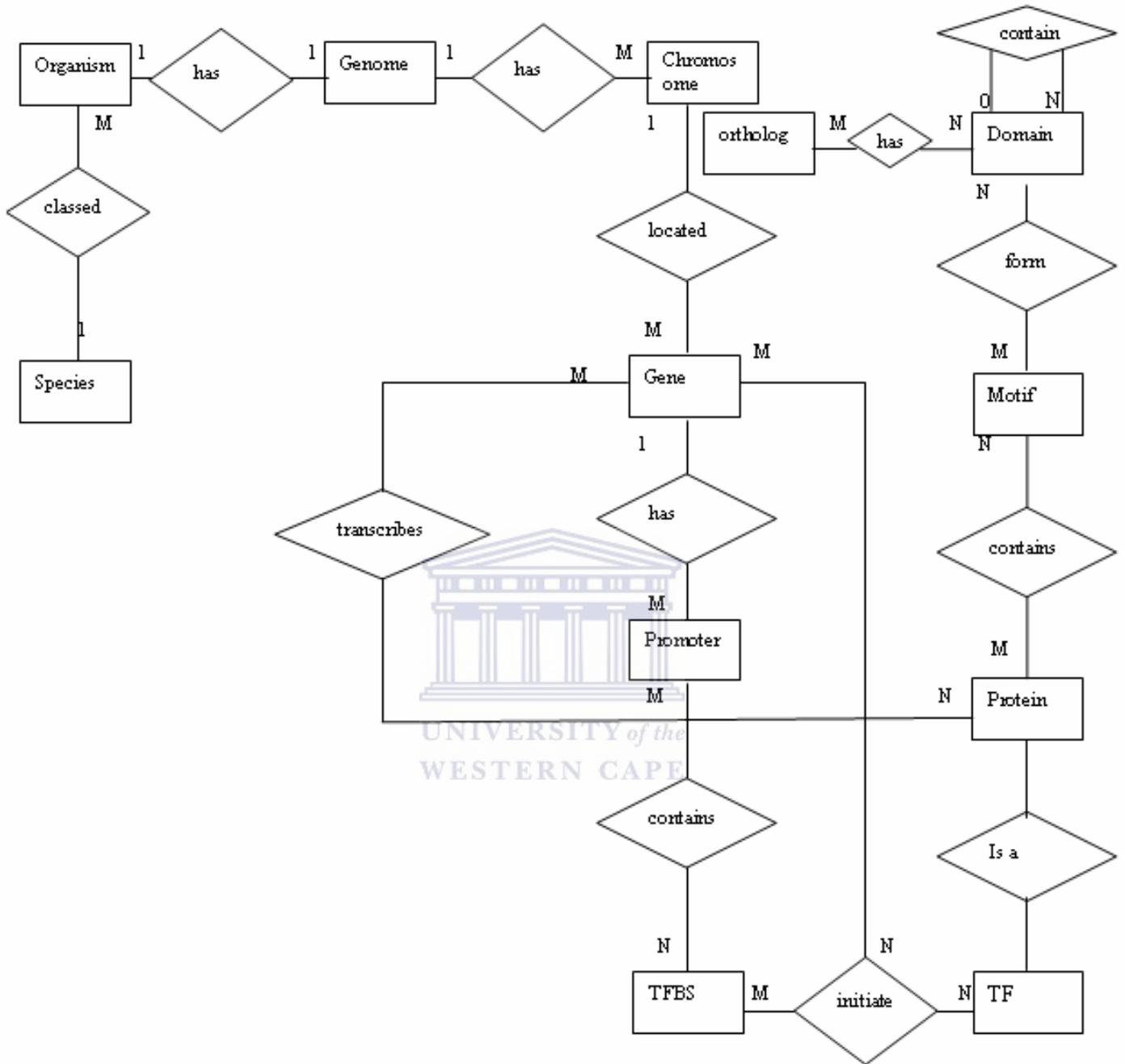
Figure 1.3: A conceptual model of biological entities
A global view of entities and their relationships. Namely the relationship between entities such as species, organism, genome, chromosome, gene, promoter, transcription factor binding sites, transcription factors, proteins, motifs, domains and orthologs.

1.5 NORMALISATION

As the modeling process moves to the logical design stage, decomposition takes place. That is, objects and relationships are decomposed to discrete self-subsisting units. Specifically, the 'many to many' relationships are all dissolved into 'one to many' relationships. Strategically, the process may be viewed as the unbundling or dissection of the original target context into discrete units which can then be reconstituted in numerous ways. This process allows the user to generate information of the target context in novel and insightful ways which is a function of the database querying process.

There is a fundamental set of guidelines to this information structuring process which is referred to as the process of normalization (Atzeni et al., 1999). It is implemented via the usage of the so-called normal forms and guides the development of a consistent database which reduces redundancy and removes anomalous behaviour. Normal forms guide the building of objects in terms of distinctive properties which uniquely identifies a row of data. This is referred to as a primary key and is used to locate data and bind relationships which are now viewed as dependencies connecting objects or tables of data. During the querying process these keys are used by the database to join tables, make calculations and locate data. The normal forms are structured hierarchically, that is, graduation to a subsequent normal form requires compliance with the previous normal form. Hence, the second normal form requires compliance with the first normal form.

The first normal form requires management of the most basic aspect of database design. As indicated earlier, this involves the decomposition of objects into discrete atomic units such that a column or field represents a single property or unit of information. This also

requires the decomposition of multi-valued attributes into single-valued attributes (Kent 1983).

The next and most important principle of the second normal requires that all properties referred to as non-key properties have to be dependent on the primary key. The primary key as indicated earlier essentially defines a row of data with a unique property or set of properties. That is, more than one column participates in key formation and as such it is called a composite key. Hence, non-complying non-key properties may point to information about a separate object which can be modeled using a separate table and in turn can then be constructed with its own primary key. In addition to compliance with the second normal form, the third normal form requires the mutual exclusivity of table non-key properties or columns. That is, whilst the second normal form requires the dependence of all non-key properties on the primary key, the third normal form requires that each non-key property should be mutually independent. Hence, any transitive dependencies between non-key properties should be identified and resolved

1.6 SWISSPROT

SWISSPROT is a curated protein database that aims to provide a high level of annotation specifically functional annotation (Figure 1.4). This data layout provides a useful starting point to retrieve quality data for plant stress-related genes. The SWISSPROT record contains amongst other fields, the "KEYWORD", "GO" and "COMMENTS" fields. Each of these fields can be used to carry out a comprehensive search of SWISSPROT. Keywords are used to provide a summary of a record's contents which is then indexed according to a set of ten categories. These include biological process, cellular component,

molecular function, coding sequence diversity, developmental stage, disease, domain, ligand and post-translation modification. GO terms and SWISSPROT keywords are stored under the same section namely, the "Ontologies" section of an entry (Figure 1.4) (Schneider et al., 2009). GO annotations are applied manually and keywords reflecting the contents of these GO annotations are also generated manually. GO annotations are classified in terms of evidence codes which are used to describe the source and strength of a particular annotation. For example, 'Inferred From Direct Assay' (IDA) indicates that the source for an annotation is an experiment and hence has the highest level of evidence classification. Whilst 'Inferred From Electronic Annotation' (IEA) is applied where an annotation has taken place though an automated computational procedure and a curator has not personally verified the annotation such as through automated importing of annotations from a related database. (Berardini et al., 2007).

Finally, the 'comments' section, which follows a similar logic to the 'keyword section', generally conveys information about a protein's function. Annotations in this area are grouped according to 'topics' such as Function, Induction, Involvement, Enzyme regulation, Pathway, Subcellular location, Tissue specificity and Developmental stage. Furthermore, as in the case with keywords, comments about a protein's function require the use of standardized terms to facilitate text searches and database interoperability (Schneider, Tognolli and Bairoch 2004; Scheider et al., 2009).

Due to its usage of manual curation SWISSPROT is regarded as a database with high quality data however its annotation cannot be accepted at face value. Even at the most basic level there is always the possibility of error. For example, the following entries**:**

http://www.uniprot.org/uniprot/Q9M9V8

contain the annotation comment "By drought and high-slat stress". The phrase 'high-slat

stress' should read 'high salt stress'.

FIGURE 1.4: The 'General annotation' section of a SWISSPROT record

The major categories are the 'Comments' and 'Ontologies' section. With regard to this entry, the GO 'Biological process' and 'Molecular function' category shows transcription factor involvement in cold acclimation via transcriptional regulation. The 'Function' category indicates the details of the stress response. The cited references confirm the annotations as correct. Hence the 'Keywords' section shows the terms 'stress response', 'transcription' and 'transcriptional regulation'.

1.7 MOTIVATION AND RATIONALE

Existing food shortages in tandem with changing weather patterns have exacerbated global food security especially in relation to staple food crops such as rice and maize.

In part, changing weather patterns provide the basis for environmental stresses as experienced by plants. An understanding of the mechanisms by which plants perceive and process responses to extremities, may provide insights which can be applied to developing stress resilient crops (Hirt and Shinozaki 2003).

The study of stress responsive genes is facilitated through using *Arabidopsis thaliana* as a point of reference due to its portability, availability, easy cultivation and comparatively small genome size. Furthermore, the availability and accessibility of information about *Arabidopsis thaliana* has also played a role in reinforcing its status as a model plant organism (Bevan and Walsh 2005).

Globally, the availability of numerous plant stress related research groups is not mirrored by a similar availability of public database resources for plant stress genes. This scarcity has been reported by the journal Nucleic Acid Research (NAR) (http://www.oxfordjournals.org/nar/database/subcat/13/39). The few databases that are still available such as the Plant environmental stress transcript database, the Plant Stress Gene database and the Stress Genomics database have not been updated since 2006 whilst others such as the Generation Challenge Programme comparative plant stress-responsive gene catalogue (GCP) have long since become defunct (Balaji et al., 2006; Wanchana et al., 2008). Furthermore, the Plant Stress Gene database and the Stress Genomics database do not have any accompanying academic publications and hence it is difficult to assess their quality.

The existing database of abiotic stress related transcription factors, STIFDB, unfortunately suffers from a number of problems. Firstly, in terms of the exclusion of stress related genes, some basic abiotic stress related transcription factors have been omitted from STIFDB (Table 1.2). Secondly, transcription factors have been included that have no correlation to abiotic stress (Table 1.3). For example, three of the entries have no abiotic stress related function. In fact, their functional annotation as indicated by the 'Stress Response' column is incomplete and the PATHOGENESIS-RELATED GENE 1 is involved in biotic stress-related defence.

Due to the key role transcription factors play in plant responses to environmental stress, the current project Dragon Arabidopsis Stress Transcription Factor database (DASTF) hopes to fill a gap in stress related transcription factor databases. The development of a manually curated database will assist in structuring the understanding of these stress responsive mechanisms. In turn, reliable data may serve as a springboard for harnessing analytical and predictive tools to generate greater insight into the environmental response mechanisms of *Arabidopsis thaliana*.

TABLE 1.1: Plant transcription factor families and their DNA binding domains (Panchon 2008)

Domain superclass	TF family
Basic domain	BES1, bHLH, bZIP, EIL, GeBP, TCP
Helix-turn-helix domain	ARR-B, E2F-DP, FHA, G2-like, HB, HSF,MYB,MYB-related, WP-RK, Sigma70-like, zf-HD
Zinc coordinating domain	Alfin-like, C2C2-CO-like, C2C2-Dof, C2C2-GATA, C2C2-YABBY, C2H2, C3H, CPP, GRF, HRT, LIM, PHD, PLATZ, SBP, SRS, TAZ, VOZ, WRKY, ZIM
Beta-scaffold with minor groove contacts domain	CCAAT, CSD, GRAS, HMG, MADS
Others	AP2-EREBP, ARF, ARID, BBR/BPC, CAMTA, DBP, DDT, Jumonji, LFY, NAC, NOZZLE, PBF-2-like, RB, S1Fa-like, Trihelix, TUB, ULT, ABI3VP1

TABLE 1.2: Abiotic stress-related transcription factors omitted from STIFDB

SWISSPROT ID	TAIRID	NAME	STRESS
Q6R0H0	AT1G01520	MYB FAMILY	SALT STRESS
Q8W4M7	AT1G03190	UVH6 (ULTRAVIOLET HYPERSENSITIVE 6)	HEAT AND UV LIGHT
Q9C8Y3	AT1G66350	RGL1 RGL1 (RGA-LIKE)	ABSCISIC ACID, SALT STRESS
Q38998	AT2G26650	AKT1 (ARABIDOPSIS K TRANSPORTER 1)	SALT STRESS
Q8GXW1	AT3G03450	RGL2 (RGA-LIKE 2)	ABSCISIC ACID, SALT STRESS, ROS REGULATION
Q9SMQ4	AT4G40010	SNRK2.7 (SNF1-RELATED PROTEIN KINASE 2.7)	SALT STRESS
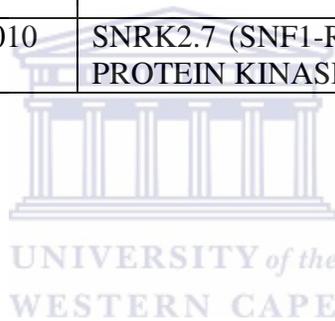
TABLE 1.3: Erroneous entries included in STIFDB

SWISSPROT ID	TAIRID	NAME	STRESS
P33154	AT2G14610	PR1 PR1 (PATHOGENESIS-RELATED GENE 1)	BIOTIC DEFENCE
O22825	AT2G43780	HYPOTHETICAL PROTEIN	FUNCTION UNKNOWN
O65547	AT4G31030	HYPOTHETICAL PROTEIN	FUNCTION UNKNOWN
O81838	AT4G27350	HYPOTHETICAL PROTEIN	FUNCTION UNKNOWN

1.8 AIMS AND OBJECTIVES

1. Produce a manually curated set of plant stress-related proteins

2. Identify transcription factor binding sites for the above proteins

3. Develop a web portal for retrieving plant stress-related regulatory information

1.9 THESIS OUTLINE

This thesis consists of four chapters:

Chapter 1:

• Brief overview of (i) stress-related genes in plants, (ii) transcription regulation, (iii) current databases for stress-related genes in plants and (iv) database design strategies

Chapter 2:

• Describes the methodology for collection and curating stress-related genes and webserver development.

Chapter 3:

• Describes the data captured in the DASTF database and the functional features.

Chapter 4:

• Discussion and Conclusion

CHAPTER 2

METHODS

CONTENTS   **PAGE**

2.1 COLLECTION OF STRESS RESPONSE TRANSCRIPTION FACTORS

Using the SWISSPROT query engine, three text-based searches were carried out to retrieve regulatory elements that respond to environmental stress as follows**:**

(a) organism:"*Arabidopsis thaliana* [3702]" AND keyword:"Stress response [KW-0346]"

(b) organism:"*Arabidopsis thaliana* [3702]" AND go:"response to stress [0006950]"

(c) organism:"*Arabidopsis thaliana* [3702]" AND annotation:(type:function stress)

The above queries generated 209, 2849 and 353 records respectively. Duplicated records were removed and resulted in final dataset of 2904 entries specific for stress response proteins.

A total of 2333 out of 2904 SWISSPROT records were mapped to their corresponding GENBANK and The Arabidopsis Information Resource (TAIR) IDs using a SWISSPROT online tool called "ID-Mapping Tool" (http://www.uniprot.org).   A PYTHON parser script was written to load the 2333 records into a local MYSQL database (see section 2.4 for database design).

2.2 CURATION

The adopted curation strategy was based on experience with a project called DAMPD (Sundarajan et al., in prep.) which is a collection of manually curated antimicrobial peptides. The curation methodology involved checking SWISSPROT keyword terms and function annotation (in the 'general annotation (comments)') against the references listed for each record. That is, if the SWISSPROT keyword section had the terms "response to stress", it was checked against the function annotation section and then cross-referenced

against the list of sources for the respective entry. Each record was curated manually using this procedure. SWISSPROT protein existence codes or evidence at protein level (PE codes) were also appropriated in the same manner namely PE level one (evidence at protein level) and PE level two. PE level one indicates that there is clear experimental evidence for the protein's existence. PE level two indicates that although there is no clear experimental evidence, there is at least gene expression data pointing to the existence of a transcript. Finally, two additional stress response databases namely STIFDB and Plant Stress Gene Database were cross-referenced to retrieve further information where applicable.

2.3 WEBSERVER DESIGN
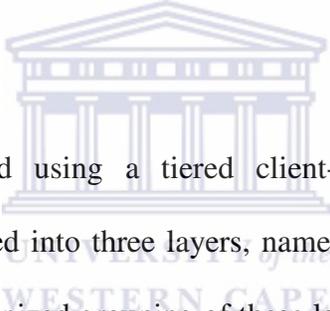
The system was implemented using a tiered client-server architecture where an application is essentially divided into three layers, namely, a presentation layer, a logic layer and a data layer. The organized grouping of these layers as in the case of a two-tier or three-tier system determines the type of client-server architecture (Ramanathan 1995). The DASTF system was implemented using a two-tier architecture which is also known as a thin client because the majority of processing is handled by the server (Figure 2.1). These include the logic and data layers whilst the client only handles the presentation and layout processing (Steiert 1998).

The presentation layer deals with the user interface of the client machine and it provides the end-user with the visual means of accessing and querying the system. This layer handles the resulting output of a request which is formatted in Hypertext Markup

Language (HTML) using cascading style sheets as a means to manage the appearance of a webpage.

The logic layer constitutes the so-called business rules of the application and is responsible for error and conformance checking. That is, making sure the user has filled in the necessary details as provided by the web interface which in turn constitutes the parameters of a request (Ward and Dafoulas 2006). As part of the logic layer, the latter process continues by translating the client request using its business logic modules into a database readable format, namely, Structured Query Language (SQL). The database was designed using MYSQL (Figure 2.1) and resides on an Apache webserver that acts as a medium routing requests from the client to the database. In other words, PHP uses its inbuilt modules to speak to the database via the Apache software.
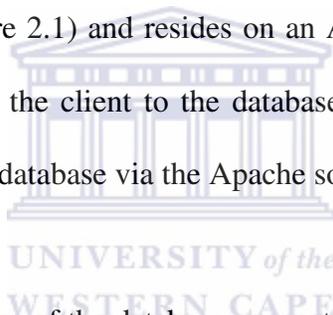
Finally, the data layer is made up of the database connection layer and the database layer. The data connection layer is responsible for connecting a user request to the actual database (Ward and Dafoulas 2006). It functions as a generic component capable of connecting to a database irrespective of the source database's architecture (network, hierarchical or relational), location path (where the database resides) or vendor (Oracle, Microsoft, Borland, MYSQL). In turn, the database layer processes (inserts, updates, deletes, queries) the request in terms of its own set of constraints, such as integrity checking for example. It then returns the result of the request along the same pathway it was received (Bahrami 1999).

2.4 DATABASE DESIGN

Two MYSQL tables were generated namely DASTF_TF and TF_Family tables (Figure 2.2). Each SWISSPROT record was parsed using a PYTHON script to extract fields needed to populate the MYSQL database (Appendix I).

2.5 STRESS RESPONSE TRANSCRIPTION FACTOR FAMILIES

Stress response proteins were mapped to ENSEMBL IDs using a SWISSPROT ID-mapping tool and each ENSEMBL ID was used to extract 5'UTRs from the ENSEMBL *Arabidopsis thaliana* core database. The 5' UTRs were scanned for transcription factor binding sites with TRANSFAC professional database version 2011.1 and the output file was parsed with a PERL script (Appendix II) (Matys et al., 2006).

Stress response genes were divided into two exclusive type groups, namely those which only respond to abiotic stress and those which only respond to biotic stress.

The terms plant defence, bacteria, fungus, virus, pathogen, nematode, oomycete, wounding, insect, chitin, jasmonic acid, ethylene and salicylic acid were used to identify biotic stress. Abiotic stress was identified by terms related to temperature (cold, heat), light (ultraviolet, red light, blue light), water (drought, flooding, deprivation), precipitations (hail, snow, frost, fire), abscisic acid, salinity, herbicide and pesticide. The exclusive categories were tested for any combination of these terms. Those which had any combination of these terms were excluded from that category and placed into a 'biotic and or abiotic' category.
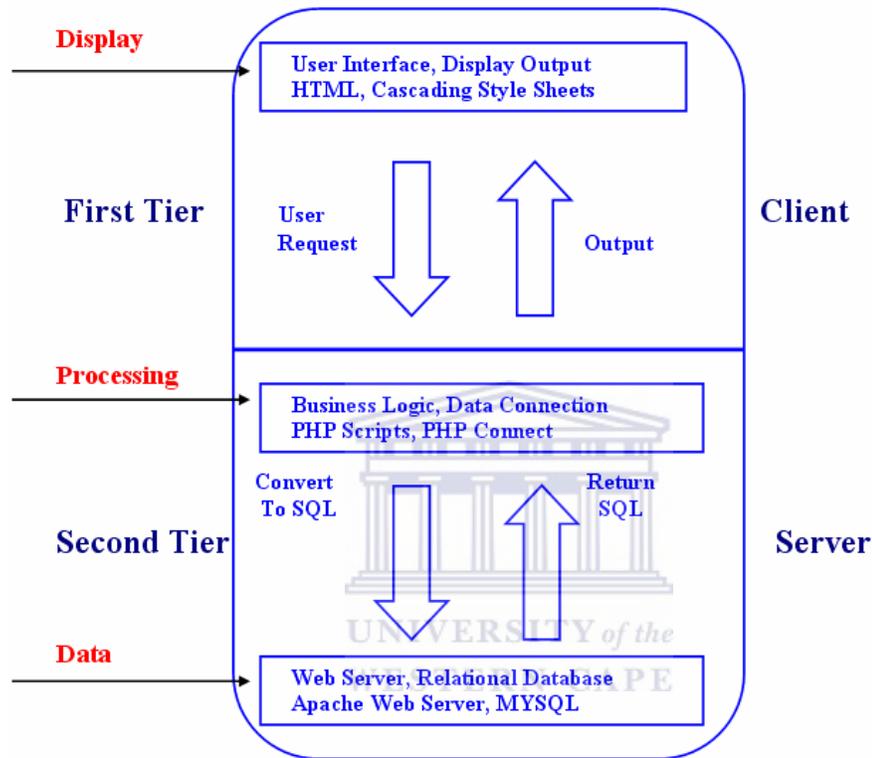
Figure 2.1: The two-tier client-server model implemented in DASTF
The majority of processing takes place on the second tier. In this implementation the second tier refers to the server side also known as 'back-end' of the system. The client has minimal processing load and is also known as the 'front-end' of the system.

DASTF_TF

| GeneID |
| Entry Name |
| Uniprot_Acc |
| GeneID |
| TAIRID |
| PubmedID |
| PFAMID |
| KEGGID |
| Gene_Ont |
| PE Level |
| Chromosome |
| Sequence |
| Stress_Type |

TF_Family

| FAMID |
| Fam_Name |
| Bind_Site |
| DASTFID |

1..N   1..1
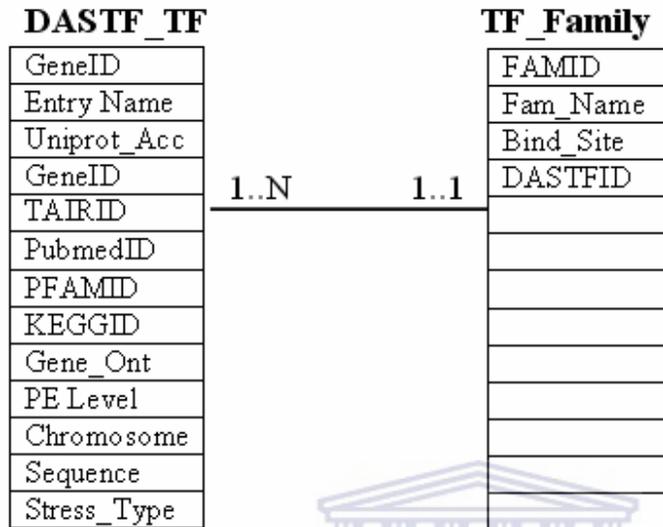
Figure 2.2: Entities with their associated properties
Transcription factor details are stored in the DAST_TF table and includes
it's name, accession number, sequence and associated GO descriptions.
Transcription factor family data is stored in the TF_Family table. Such as the
family name, transcription factor members of the family and their binding sites.

CHAPTER 3

RESULTS

UNIVERSITY of the
WESTERN CAPE

3.1 DASTF WEBSERVER

The DASTF system was developed with a two-tier client-server architecture which is better suited to smaller projects where a faster application development time is required and in an environment where user traffic is expected to be lower (Hemmer 1995). Hence it is suited to a small community-specific project such as DASTF where batch user downloads are not supported.

The DASTF database (Figure 3.1) currently holds 2333 entries. SWISSPROT's 'ID mapping' tool was used to trace these entries to their TAIR and ENTREZ GENBANK counterparts. These genes cover both biotic and abiotic stresses such as cold, heat, light, salt, water deprivation, wounding, oxidation, fungi and bacteria. There are two query categories, namely, an exclusive category and a related category. Genes that are found in the exclusive categories are either exclusively biotic or exclusively abiotic which means that there is no overlap in stress response between these genes. A total of 60 entries only correspond to biotic stress whilst 167 entries only correspond to abiotic stress.

The related type stress category refers to genes that respond to biotic and/or abiotic stress types. The overwhelming majority are found in this category namely 2106 entries.

Furthermore, there are 424 transcription factor families of which 417 contain genes for which UTRs (untranslated region) could be extracted from ENSEMBL and UTRs shorter than 50 nucleotides were excluded.

Unique transcription factor binding sites were identified in 13 stress response genes belonging to characterised protein families (Table 1.4). A total of 10 stress response genes without any protein family classification contained unique transcription factor binding sites. Among these unique binding sites was SED motif that bound to both

AT2G42910 (methyltransferase superfamily) and AT3G12810 (unclassified protein family).

Chromosome one, five, three, two and four contain 599, 574, 417, 372 and 371 genes respectively. The most common abiotic related stresses in the database are responses to light, heat, salinity, cold, water deprivation and oxidative stress respectively. The distribution of abiotic only and biotic only responsive genes mirrors the overall distribution of stress responsive genes over the five chromosomes.

Table 1.4: Unique TFBS motifs in stress-related genes families

Transcription factor binding site	Transcription factor protein family	Stress response gene	TRANSFAC motif
PCF2	Belongs_to_the_copine_family	AT5G07300	GTGGGtccca
SED	Belongs_to_the_methyltransferase_superfamily	AT2G43910	gtacCCTTTt
EmBP-1b	Contains_1_Rieske_domain	AT3G44880	tcCACGTgtt
SPF1	Belongs_to_the_DEFL_family	AT5G44420	aaATAGTaat
MYB.Ph3	Contains_4_EF-hand_domains	AT5G37770	ctAACCGtttttt
AGL2	Contains_1_F-box_domain	AT3G26810	gtttctattTATGGtttt
E2F	Belongs_to_the_DNA_mismatch_repair_mutS_family	AT4G02070	tgtTTCCCgcc
Alfin1	Contains_4_Kelch_repeats	AT1G54040	aaatagGTGGGgcag
ABF	Belongs_to_the_myo-inositol-1-phosphate_synthase	AT2G22240	aaacccgccaCGTGTcctccctcc
P	In_the_N-terminal_section;_belongs_to_the_glutamate_5-	AT2G39800	acCTACCct
ABZ1	Contains_1_CSD_cold-shock_domain	AT4G38680	gggtgACGTGgcag
HBP-1b	Belongs_to_the_sugar_epimerase_family	AT2G37660	gTGACGtggcgaaa
CPRF-2	Belongs_to_the_EIN3_family	AT3G20770	gccACGTGat
GBF	Unclassified protein family	AT3G11930	ttgCACGTggcc
AGL1	Unclassified protein family	AT4G38580	ttttcctttTCTGGaata
OSBZ8	Unclassified protein family	AT5G09230	ACGTGtcgcgtttc
CPRF-1	Unclassified protein family	AT2G47770	gcCACGTgta
PCF-2	Unclassified protein family	AT4G16990	GTGGGtccca
SED	Unclassified protein family	AT3G12810	aAAAGGgtat
TEIL	Unclassified protein family	AT3G52430	agaTACAT
E2F	Unclassified protein family	AT5G61460	tctTTCCCgcc
CDC5	Unclassified protein family	AT4G24520	aacGCTGAgcc
AGL15	Unclassified protein family	AT5G64440	tttcttcaTTTAGtaa

3.3 WEBSITE OVERVIEW

The DASTF database can be accessed via the 'search page' (Figure 3.2). At present, there are four search options, namely, a simple search, search by stress type, search by chromosome and search by transcription factor family. The stress type search allows the user to search by two categories, namely, an exclusive stress-type response and a related type stress response. The exclusive option consists of two types, that is, 'BIOTIC' or 'ABIOTIC' which is followed by the 'BIOTIC AND OR ABIOTIC' category.

The text search can be used in instances where the user has a protein name or a list of accession numbers from UNIPROT, TAIR, ENTREZ or PFAM. The chromosome search option reports all stress-related genes on a particular chromosome. Transcription factor family search allows the user to search a set of transcription factor families encoding stress-related functions.
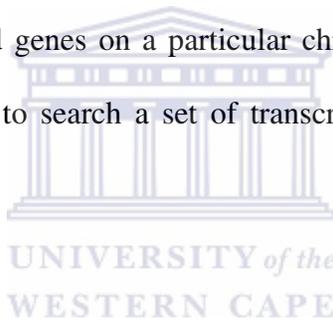
3.4 DATA ENTRY RECORD

Each entry is identified by a series of fields (Figure 3.3). The DASTF entry field is the database's unique accession number for each record whilst 'Entry name' identifies the name of the transcription factor. The 'Uniprot Accession', 'Gene ID' and 'TAIR ID' fields are the respective UNIPROT protein, ENTREZ gene and TAIR accession numbers. Links to UNIPROT, ENTREZ gene and TAIR accession numbers allow the user to cross-reference annotations as well as to access more detailed descriptions from the respective source databases. Similarly, the 'Pubmed field' provides a link to the publications which reference a particular entry. The 'Family field' identifies the transcription factor family that maps to a protein. The 'KEGG' field allows the user to access pathways that intersect

a transcription factor while the 'PFAM' field provides protein domain structures for a specific transcription factor. The 'PE' field identifies the evidence level used during curation to validate the protein. 'Evidence at protein level' or 'PE = 1' indicates that a protein was identified by experimental evidence. The Chromosome field describes the chromosomal location of a gene in base pairs. Furthermore, by clicking the chromosome hyperlink an image is generated showing the exact location of the gene as well as its proximity to other genes on the same chromosome. The 'TFBS' field is linked to a gene's transcription factor binding site and when a user clicks the 'Click here to see the TFBS' hyperlink a binding site (Figure 3.4) is reported for the particular gene.

The GO field provides the functional annotation for each record. Each annotation such as stress response, evidence and other related activities are referenced by their respective GO identifier. Finally, the sequence field shows the protein sequence.
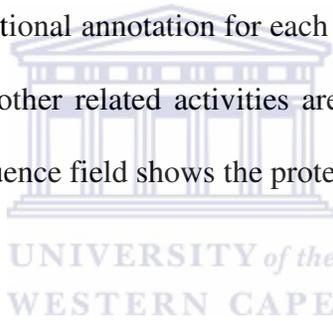
3.4 TOOLS

The database features four tools which are accessible via the tools page (Figure 3.4). These are BLAST (Altschul et al., 1990), CLUSTALW (Thompson, Higgins and Gibson 1994), HMMER (Eddy et al., 1995) and HYDROCALCULATOR (Tossi et al., 2002).

For BLAST the user uploads a sequence or list of sequences to test against the database. The sequences have to be in a specific format called the FASTA format. The utility then checks the sequences against the DASTF database and generates a score indicating which sequences are most similar to the user's input sequences. This allows users to identify whether their sequences are potential stress related genes. Another approach is using HMMER where a model is generated by inputting sequences with a high degree of

validity. That is, by using sequences where the entries have a high GO evidence type and a high SWISSPROT protein existence level. Protein sequences are subsequently searched against the model to assign them to families by using similarity as a basis for comparison. In this manner users can predict which stress related transcription factor family their sequences belong to. HYDROCALCULATOR generates protein analysis by analyzing amino acid hydrophobicity using scales such as the combined consensus scale (CSS), Kyte and Doolittle and Eisenberg.
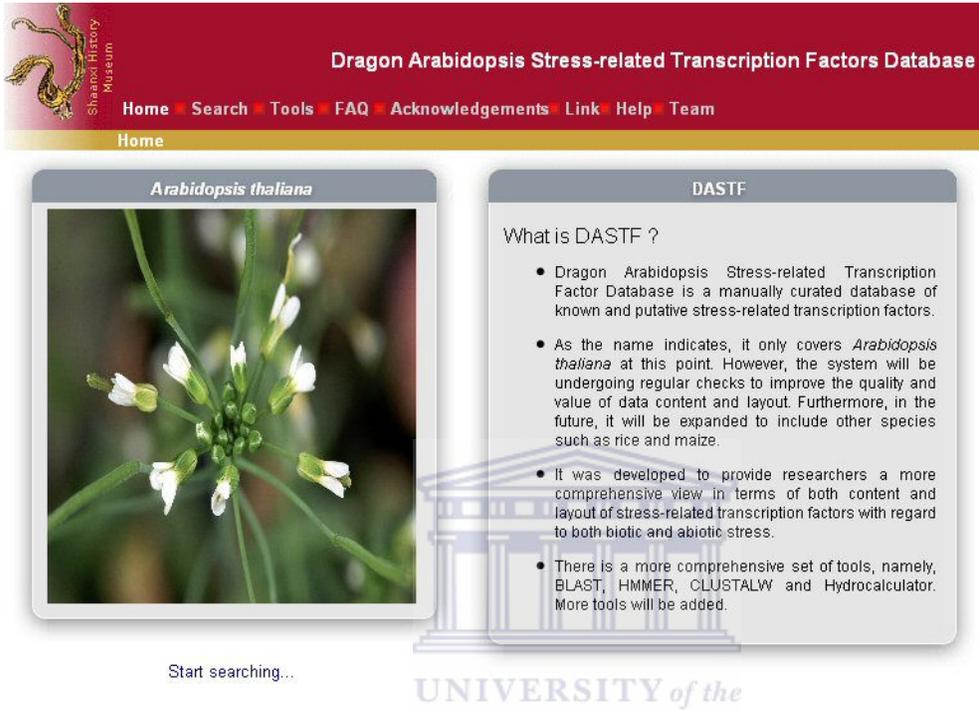
FIGURE 3.1: The DASTF Homepage
The homepage of the DASTF website. The menu section at the top allows
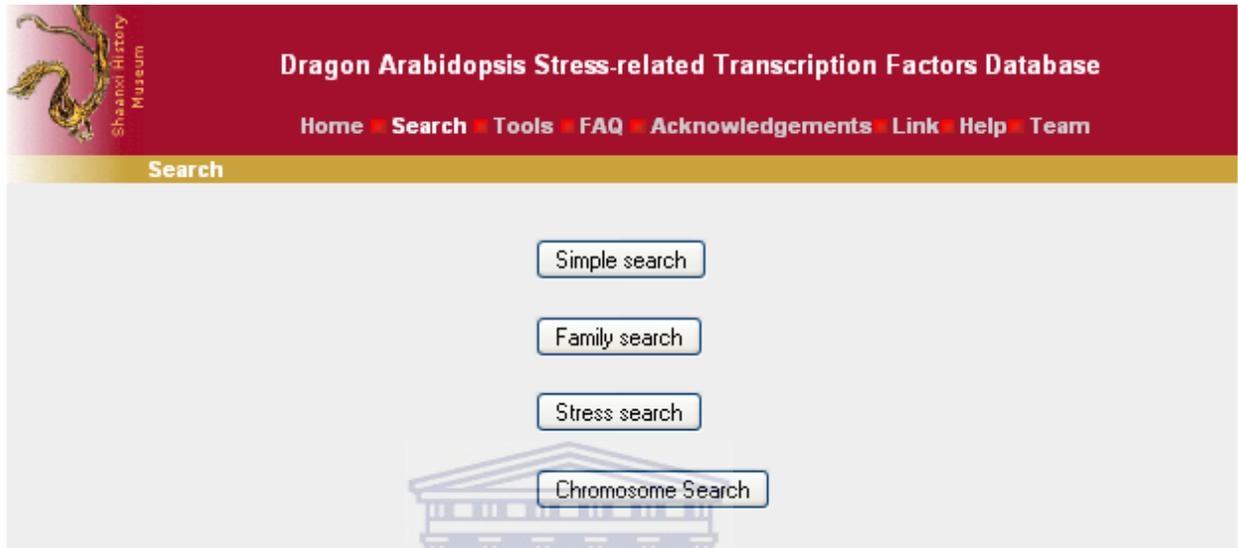access to the search engine and online tools.

FIGURE 3.2: The search page
Search options are divided into 'simple search', search by 'chromosome', 'family and search using 'stress' categories. The simple search uses entry names and accession numbers as input and the 'stress search' uses 'BIOTIC', 'ABIOTIC' and 'BIOTIC AND OR ABIOTIC' categories.

DASTF entry	DASTF_2241
Entry name	Q9SKY8_ARATH
UniProt Accession	Q9SKY8
Gene ID	817771
Pubmed	10617197
Family	Belongs to the heat shock protein 70 family
TFBS	(Click here to see the TFBS)
Tair	AT2G32120
Kegg	ath:AT2G32120
Pfam	PF00012
PE	Evidence at transcript level
Chromosome	Chromosome 2: 13651510-13653921
Stress type	ABIOTIC
GO	GO:0005524;F:ATP binding; IEA:UniProtKB-KW. GO:0009408;P:response to heat; IEP:TAIR. GO:0009644;P:response to high light intensity; IEP:TAIR. GO:0042542;P:response to hydrogen peroxide; IEP:TAIR.
Sequence	MAEAAYTVASDSENTGEEKSSSSPSLPEIALGIDIGTSQCSIAVWNGSQVHILRNTRNQKLIKSFVTFKD EVPAGGVSNQLAHEQEMLTGAAIFNMKRLVGRVDTDPVVHASKNLPFLVQTLDIGVRPFIAALVNNAWRS TTPEEVLAIFLVELRLMAEAQLKRPVRNVVLTVPVSFSRFQLTRFERACAMAGLHVLRLMPEPTAIALLY AQQQQMTTHDNMGSGSERLAVIFNMGAGYCDVAVTATAGGVSQIKALAGSPIGGEDILQNTIRHIAPPNE EASGLLRVAAQDAIHRLTDQENVQIEVDLGNGNKISKVLDRLEFEEVNQKVFEECERLVVQCLRDARVNG GDIDDLIMVGGCSYIPKVRTIIKNVCKKDEIYKGVNPLEAAVRGAALEGAVTSGIHDPFGSLDLLTIQAT PLAVGVRANGNKFIPVIPRNTMVPARKDLFFTTVQDNQKEALIIIYEGEGETVEENHLLGYFKLVGIPPA PKGVPEINVCMDIDASNALRVFAAVLMPGSSSPVVPVIEVRMPTVDDGHGWCAQALNVKYGATLDLITLQ RKM

FIGURE 3.3: List of properties for entry DASTF_2241
The TFBS detail (circled in red) is hyperlinked and explained in Figure 3.4.

Motif name	Position	Strand	Core score	Matrix score	Binding site sequence
MYBAS1	118	(+)	0.999	0.999	ccGCAACtgct
MYBAS1	119	(+)	0.999	0.999	ccGCAACtgct
PBF	139	(-)	1	1	CCTTTt
PBF	140	(-)	1	1	CCTTTt

FIGURE 3.4: TFBS prediction associated with UTR for entry DASTF_2241
'motif name' refers to the TFBS name in TRANSFAC; 'position' refers to the base position in the UTR; 'strand' indicates the positive (+) or negative (-) DNA strand; 'core score' and 'matrix score' are thresholds for binding efficiencies when using TRANSFAC; 'binding site sequence' is the motif identified by TRANSFAC.

FIGURE 3.5: Tools
A list of the online tools available BLAST, CLUSTALW, HMMER and HYDROCALCULATOR

CHAPTER FOUR

DISCUSSION

	CONTENTS	PAGE

4.1 METHODOLOGICAL ISSUES

The DASTF database contains 2333 *Arabidopsis thaliana* stress responsive genes that were collected from SWISSPROT, they were curated, organised into protein families and TFBS predictions were generated for the corresponding UTRs. A competing database, STIFDB published in 2008, relied exclusively on microarray databases to identify genes which respond to abiotic stresses such as cold, drought, salinity and light. Candidate genes had to be significantly upregulated in at least three replicate microarray experiments. Limitations of microarray experiments include repeatability of results, the influence of statistical methods when interpreting the data and the influence of environmental and background noise on results (Draghici et al., 2006). The GO consortium has advocated a set of cautionary guidelines where expression data is used to assign functions to genes and the evidence code 'Inferred from Expression Pattern' has been developed for this purpose (Evidence Codes Group 2007). Firstly, their guidelines specifically state that it may be difficult to conclusively identify the function of a gene based on the timing of its expression pattern in relation to an experimental condition such as a stress.

Secondly, microarray expression data should not be used to assign GO 'molecular function' annotation claims. For example, genes that are upregulated during a stress response should rather be classified according to a GO biological process category such as 'response to stress' and not according to the 'molecular function' category (Evidence Codes Group 2007). The wisdom of these guidelines became apparent when STIFDB's microarray sourced data was reviewed. A number of the entries were reviewed manually by comparing them with annotations in source databases such as SWISSPROT, TAIR

and GENBANK to identify any links with stress response. It was found that a number of genes (Table 1.5, Appendix III) with unknown functions were incorrectly included in STIFDB database.

The records for these genes with unknown function annotations were updated in 2010 and 2011 by GENBANK, SWISSPROT and TAIR. STIFDB was launched in 2008 which implies that inspite of a two year gap since the publication of its data, none of the updates to these gene records in GENBANK, SWISSPROT and TAIR indicate any relationship with stress response. Furthermore there is no indication that the STIFDB team had curated its data prior to its storage (Shameer et al., 2008).

The approach to gathering data for the DASTF database was influenced by the results of the STIFDB data review process and hence the need for a foundation based on well annotated and curated entries was identified as a priority.

The DASTF database covers both biotic and abiotic stresses whilst STIFDB only contains entries related to abiotic stress. Furthermore, STIFDB restricted their database to a list of 22 abiotic stress related families. However, abiotic stress responses are not limited to these 22 families and hence DASTF includes other abiotic stress related families such as glutathione peroxidase family, carotenoid oxygenase family, cytochrome P450 family, GST superfamily and a number of unclassified families as well.

Similarly, STIFDB predicted transcription factor binding sites for genes belonging to its set of 22 families whilst DASTF contains binding sites for biotic, abiotic as well as unknown families of stress response genes. Unique transcription factor binding sites were identified in 13 stress response genes belonging to characterised protein families and 10 stress response genes without any protein family classification. These motifs have been

implicated in stress regulatory environments involving responses to salt, oxidation, cold, abscisic acid and light (Singh 1998; Schwechheimer et al., 1998; Choi et al., 2000; Hasegawa et al., 2000; Jackoby et al., 2000; Grover et al., 2001; Memelink et al., 2001; Pastori and Foyer 2002; Shen, Cao and Wang 2008). Furthermore, for each gene that is identified, the database provides information identifying both the metabolic processes that a gene is involved in and as well as its location relative to other proximate genes. In addition, the database is further enhanced by the integration of tools such as HMMER, BLAST, HYDROCALCULATOR and CLUSTALW. An area which requires attention is the categorization of previously unclassified protein sequences into families. By using the HMMER tool proteins such as DNA repair protein REV1 and F16G20.140 have been identified as potentially related to the family '6-phosphogluconate dehydrogenase'.

4.2 COMPLEXITIES, LIMITATIONS AND FUTURE WORK

Arabidopsis thaliana has a smaller more compact genome in comparison with other organisms such as rice (Karlowski et al., 2003). The distribution of genes across *A.thaliana* five chromosomes reflects the need for compact organisation and economical usage of its genetic resources (Mayer et al., 1999; Arabidopsis Genome Initiative 2000; Holtorf, Guitton and Reski 2002).

It was reported (in chapter three) that 2106 genes out of 2333 stress responsive genes, were able to respond to both biotic and abiotic stress which may be interpreted in terms of a phenomenon called cross-talk. Genes use common stress related signalling pathways to generate complex cascades of gene expression in response to environmental stress.

For example, plant defence genes use signal transduction pathways such as jasmonic acid, ethylene and salicyclic acid to cope with a variety of viral, bacterial and fungal pathogens. Various permutations of pathway and transcription factor combinations impact on the activation and deactivation of specific types of biotic defence related genes (Fujital et. al., 2006; Century Reuber and Ratcliffe 2008). Cross-talk permutations have been postulated as a means whereby genes are able to fine-tune their responses to a wider range of threats from its environment. Furthermore, pathways are also able to operate independent from each other, collaboratively as well as antagonistically (Nimchuk et al., 2003; Van Verk and Gatz 2009).

The issue of context adds another layer of complexity to this study which influences the manner in which developmental processes can be viewed. In normal circumstances, various parts of an organism are supplied with nutrients on a systematic basis leading to its growth and development. However, in the event of a biotic attack as in the case of a virus, for example, a plant may voluntarily cease supplying the affected area with nutrients. That is, in order to contain and prevent the spread of the virus which needs nutrients to survive, the plant may effectively 'kill off' the affected part and starve the virus of nutrients (Lichtenthaler, 1995; Lichtenthaler, 1998). Hence annotations which identify the manner in which developmental processes such as nutrient transport and other context specific behaviour become important. From the perspective of database annotation, identifying context, the components involved and their context specific expression is one of the major limitations of the database. However, it should be noted that whilst the latter processes are inherently complex, its absence does not detract from

the value of DASTF as a resource which provides a comprehensive, curated catalogue of stress response genes in *Arabidopsis thaliana*.

Furthermore, the database needs to be evaluated and updated on a regular basis by identifying new annotations which can embellish and augment existing data thereby providing a more comprehensive view of stress related genes and proteins.

Areas that require attention are:

i) the construction of a systematic and comprehensive library of motifs and their stress responsive associations which can be used to classify new entries.

ii) there is an urgent need to classify the large number of currently unclassified transcription factor families in the database.

iii) additional levels of detail should be added to entries as in the case of stress responses to viruses, bacteria and fungi such as data detailing which types of fungal, bacterial and viral strains have been studied should be added to the database to provide a greater degree of specificity to its annotation.

Finally, the database can be extended to include other species such as rice and maize and thereby also attempt to identify orthologs between these species.

REFERENCES

Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408(6814), 796-815.

Aarts, M.G and Fiers M.W. (2003) What drives plant stress genes? *Trends Plant Sci.* 8(3), 99-102.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17), 3389-402.

Atzeni, P., Ceri, S., Paraboschi, S. and Torlone., R. (1999) Database Systems - Concepts, Languages and Architectures. New York, McGraw-Hill.

Avison, D.E. and Fitzgerald, G. (1995) Information Systems Development: Methodologies, Tools and techniques. Maidenhead, McGraw-Hill.

Bahrami, A. (1999) Object Oriented Systems Development: Using the Unified Modelling Language. Singapore, McGraw-Hill.

Bailey, T.L., Williams, N., Misleh, C., Li, W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* 34, 369-73.

Balaji, J., Crouch, J.H., Prasad, P.PV.N.S. and Hoisington, D.A. (2006) A database of annotated tentative orthologs from crop abiotic stress transcripts. *Bioinformation* 1(6), 225–227.

Bevan, M. and Walsh, S. (2005) The Arabidopsis genome: a foundation for plant research. *Genome Res*. 15(12), 1632-42.

Birney, E. and Clamp, M. (2004) Biological database design and implementation. *Brief Bioinform* 5(1), 31-38.

Bornberg-Bauer, E. and Paton, N.W. (2002) Conceptual data modeling for bioinformatics. *Brief Bioinform*. 3(2), 166-80.

Bu, Q, Jiang, H., Li, C.B., Zhai, Q., Zhang, J., Wu, X., Sun, J., Xie, Q. and Li, C. (2008) The Role of the *Arabidopsis thaliana* NAC transcription factors ANAC019 and ANAC055 in regulating jasmonic acid-signaled defence responses. *Cell Res*. 18(7), 756-67.

Bülow, L., Brill Y. and Hehl, R.. (2010) AthaMap-assisted transcription factor target gene identification in *Arabidopsis thaliana*. *Database* (Oxford).

Button, D.K., Gartland, K.M., Ball, L.D., Natanson, L., Gartland, J.S. and Lyon, G.D. (2006) DRASTIC INSIGHTS: querying information in a plant gene expression database. *Nucleic Acids Res*. 34.

Chan, T., Wong, K., Lee, K., Wong, M., Lau, C., Tsui, S.K., and Leung, K. (2011) Discovering approximate-associated sequence patterns for protein–DNA interactions *Bioinformatics* 27(4), 471–478.

Chung, K.M., Igari, K., Uchida, N., Tasaka, M. (2007) New perspectives on plant defence responses through modulation of developmental pathways. *Mol Cells*. 26(2), 107-12.
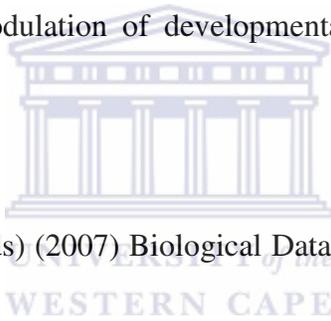
Chen, J. and Amandeep, S. (Eds) (2007) Biological Database Modeling. London, Artech House Publishers.

Chen, J.Y. and Carlis, J.V. (2003) Genomic data modeling. *Information Systems* 28, 287–310.

Chen, P. (1976) The Entity-Relationship Model - Toward a Unified View of Data. *ACM Transactions on Database Systems* 1(1), 9–36.

Codd, E.F. (1970) A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM* 13(6), 377–387.

Choi, H., Hong, J., Ha, J., Kang, J. and Kim, S.Y. (2000) ABFs, a family of ABA-responsive element binding factors. *J. Biol Chem*. 275(3), 1723-30.

Coupland, G. and Prat, M.S. (2005) Cell signalling and gene regulation signalling mechanisms in plants: examples from the present and the future. *Curr Opin Plant Biol*. 8(5), 457-61.

Dimmer, E., Berardini, T.Z., Barrell, D. and Camon, E. (2007) Methods for gene ontology annotation. *Methods Mol Biol*. (406), 495-520.

Dong, X., Mindrinos, M., Davis, K.R. and Ausubel, F.M. (1991) Induction of Arabidopsis defence genes by virulent and avirulent Pseudomonas syringae strains and by a cloned avirulence gene. *Plant Cell*. 3(1), 61-72.
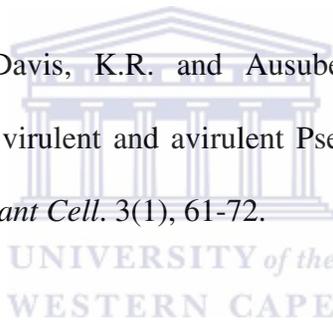
Draghici, S., Khatri, P., Eklund, A.C. and Szallasi, Z. (2006) Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet*. 22(2),101-9.

Easterbrook, S. (1993) Negotiation and the Role of the Requirements Specification.

in P. Quintas (Ed.) *Social Dimensions of Systems Engineering: People, processes, policies and software development*. London, Ellis Horwood, pp144-164.

Eddy, S., Mitchison, G. and Durbin, R. (1995) Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol*. 2, 9-23.

Elmasri, R., Navathe, S.B. (2003) Fundamentals of Database Systems. U.S.A, Addison Wesley.

Eulgem, T., Rushton, P.J., Robatzek, S. and Somssich, I.E. (2000) The WRKY superfamily of plant transcription factors. *Trends Plant Sci*. 5(5), 199-206.

Eulgem, T. (2005) Regulation of the Arabidopsis defence transcriptome. *Trends in Plant Science* 10(2).

Evidence Codes group. (2007) Evidence Code Proposals Draft. GOC meeting, Cambridge UK.

Fung, D. C. Y. (2008) Developing a DNA variant database. *Methods Mol Med*.141, 219-43.
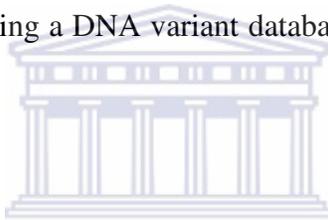
Gallaugher, J. and Ramanathan, S. (1996) The Critical Choice of Client-Server Architecture: A Comparison of Two and Three-Tier Systems. *Information Systems Management*. Spring, 7-13.

Glazebrook, J. (1999) Genes controlling expression of defence responses in Arabidopsis. *Curr Opin Plant Biol*. 2(4), 280-6**.**

Goodrich, J. and Tweedie, S. (2002) Remembrance of things past: chromatin remodeling in plant development. *Annu Rev Cell Dev Biol*. 18, 707-46.

Guo, A., He, K., Liu, D., Bai, S., Gu, X., Wei, L. and Luo, J. (2005) DATF: a database of Arabidopsis transcription factors. *Bioinformatics* 21(10).

Grau J., Ben-Gal I., Posch S., Grosse I. (2006) VOMBAT: Prediction of Transcription Factor Binding Sites using Variable Order Bayesian Trees. *Nucleic Acids Research*, 34(W529–W533).

Grover, A., Kapoor, A., Satya-Lakshmi, O., Agarwal, S., Sahi, C., Katiyar-Agarwal, S., Agarwal, A. and Dubey, H. (2001) Understanding molecular alphabets of the plant abiotic stress responses. *Curr. Sci*. 80, 206-216.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T. and White, R. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. 32, 258-61.

Hasegawa, P.M., Bressan, R.A., Zhu, J.K. and Bohnert, H.J. (2000) Plant cellular and molecular responses to high salinity. *Annu Rev Plant Physiol Plant Mol Biol*. 51, 463-499.

Havas, K., Whitehouse, I. and Owen-Hughes, T. (2001) ATP-dependent chromatin remodeling activities. *Cell Mol Life Sci*. 58(5-6), 673-82.

Helden, J.V., Naim, A., Mancuso, R., Eldridge, M., Wernisch, L., Gilbert, D. and Wodak, S.J. (2000) Representing and analysing molecular and cellular function using the computer. *Biol Chem*. 381(9-10), 921-35.

Hirayama, T., Shinozaki, K. (2010) Research on plant abiotic stress responses in the post-genome era: past, present and future. *Plant J*. 61(6), 1041-52.

Hirt, H. and Shinozaki, K. (Eds.) (2004) Plant Responses to Abiotic Stress. Springer, Humana Press.
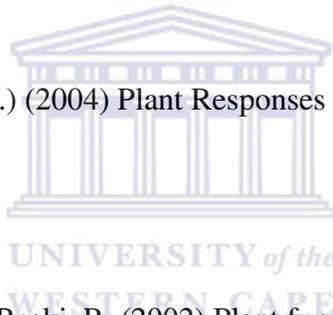
Holtorf, H., Guitton, M.C. and Reski, R. (2002) Plant functional genomics. *Naturwissenschaften* 89(6), 235-49.

Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S. and Young R.A. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*. 95(5), 717-28.

Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T., Konagaya, A. and Shinozaki, K. (2005) RARTF: database and tools for complete sets of Arabidopsis transcription factors. *DNA Res*. 12(4), 247-5

Jakoby, M., Weisshaar, B., Dröge-Laser, W., Vicente-Carbajosa, J., Tiedemann, J., Kroj, T. and Parcy, F. (2002) bZIP transcription factors in Arabidopsis. *Trends Plant Sci*. 7(3), 106-11.

Kornberg, R.D. and Klug, A. (1981) The Nucleosome. *Sci. Amer*. 244(52).

Karlowski, W.M., Schoof, H., Janakiraman, V., Stuempflen, V. and Mayer, K.F. (2003) MOsDB: an integrated information resource for rice genomics. Nucleic Acids Res. 31(1),190-2.

Karsch-Mizrachi, I. and Ouellette, B.F.F. (2001) The Genbank Sequence Database. In Baxevanis, A.D. and Ouellette, B.F.F. (Eds.) *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. New York, John Wiley and Sons.

Keet, C.M. (2003) Biological Data and Conceptual Modelling Methods. Journal of *Conceptual Modeling* 29.

Keet, C.M. (2004) Conceptual Modelling and Ontologies for Biology: experiences with the bacteriocin database. Agropecuaria Conference, San Jose de las Lajas, Cuba.

Kel, A.E., Gössling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V. and Wingender, E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*. 31(13), 3576-9.

Kent, W. (1983) A Simple Guide to Five Normal Forms in Relational Database Theory, *Communications of the ACM*. 26, 120–125.

Kwon, C., (2010) Plant Defence Responses Coming To Shape. *Plant Pathol. J*. 26(2), 115-120.

Latchman, D.S. (1997) Transcription factors: an overview. *Int. J. Biochem. Cell. Biol*. 29(12), 1305-12.

Lesk, A.M. (Ed) (2005) Database Annotation in Molecular Biology: Principles and Practice. New York, John Wiley and Sons.
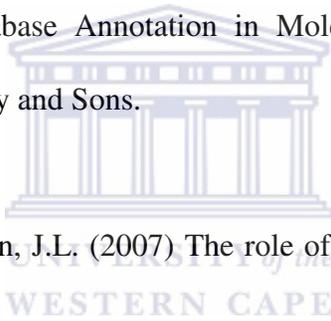
Li, B., Carey, M. and Workman, J.L. (2007) The role of chromatin during transcription. *Cell*. 128(4), 707-19.

Lichtenthaler, H.K. (1995) Vegetation stress: An Introduction to stress concept in plants. *J. Plant. Physiol*. 148, 4–14.

Lichtenthaler, H.K. (1998) The stress concept in plants: an introduction. *Ann N Y Acad Sci*. 851, 187-98.

Ma, S. and Bohnert, H.J. (2007) Integration of *Arabidopsis thaliana* stress-related transcript profiles, promoter structures, and cell-specific expression. *Genome Biol*. 8(4), R49.

Madhava Rao, K.V., Raghavendra, A.S., Janardhan Reddy, K. (2006) Physiology and Molecular Biology of Stress Tolerance in Plants. Springer, Humana Press.

Mahajan S., Tuteja, N. (2005) Cold, salinity and drought stresses: an overview. *Arch. Biochem. Biophys*. 444(2), 139-58.

Marx, B. (1999) Introduction to Database Modeling in Bioinformatics. In *The EBI Online Manual on Molecular Biology Databases*
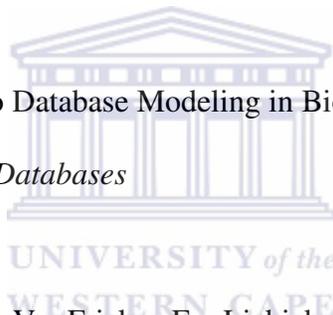
Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki0Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, *Nucleic Acids Res*. 34, D1080D110.

McGinley, S. (2000) Tracking Corn Gene Regulation - Learning More about Basic Gene Mechanisms. University of Arizona, College of Agriculture and Life Sciences.

Memelink, J., Kijne, J.W., van der Heijden, R. and Verpoorte, R. (2001) Genetic modification of plant secondary metabolite pathways using transcriptional regulators. *Adv Biochem Eng Biotechnol*. 72, 103-25.

Mitra, M., Shah, N., Mueller, L., Pin, S. and Fedoroff, N. (2002) StressDB: a locally installable web-based relational microarray database designed for small user communities. *Comp Funct Genomics*. 3(2), 91-6.

Mishra, N.S., Tuteja, R. and Tuteja, N. (2006) Signaling through MAP kinase networks in plants. *Arch. Biochem. Biophys*. 452(1), 55-68.

Mitsuda, N. and Ohme-Takagi M. (2009) Functional analysis of transcription factors in Arabidopsis. *Plant Cell Physiol*. 50(7), 1232-48.
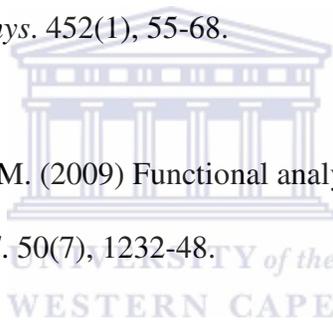
Montgomery, S.B., Griffith, O.L., Sleumer, M.C., Bergman, C.M., Bilenky, M., Pleasance, E.D., Prychyna, Y., Zhang, X. and Jones SJ (2006) ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics* 22(5), 637-40.

Nelson, M., Reisinger, S. and Henry, S. (2004) Designing databases to store biological information. *Biosilico* 1(4), 134-142.

Nimchuk, Z., Eulgem, T., Holt, B.F. and Dangl, J.L. (2003)Recognition and response in the plant immune system. *Annu Rev Genet*. 37, 579-609.

Pareek, A., Sopory, S.K., Bohnert, H. J. and Govindjee (Eds.) (2010) Abiotic Stress Adaptation in Plants: Physiological, Molecular and Genomic Foundation. Netherlands, Springer.

Pastori, G.M. and Foyer, C.H. (2002) Common components, networks, and pathways of cross-tolerance to stress. The central role of "redox" and abscisic acid-mediated controls. *Plant Physiol*. 129(2), 460-8.

Pérez-Rodríguez, P., Riaño-Pachón, D.M., Corrêa, L.G., Rensing, S.A., Kersten, B. and Mueller-Roeber, B. (2010) PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Res*. 38, 822-7.

Peterson, C.L. and Workman, J.L. (2000) Promoter targeting and chromatin remodeling by the SWI/SNF complex. *Curr. Opin. Genet. Dev*. 10(2), 87-92.

Ramakrishnan, R. and Gehrke, J. (2002) Database Management Systems. USA, McGraw-Hill

Rao, M.K.V., Raghavendra, A.S., Janardhan Reddy, K. (2006) Physiology and Molecular Biology of Stress Tolerance in Plants. Springer, Humana Press.

Rhee, S.Y. and Bill, C. (2005) Biological Databases for Plant Research. *Plant Physiology* 138, 1-3.

Riaño-Pachón, D.M. (2008) Identification of transcription factor genes in plants. Unpublished Phd thesis. Institutional Repository of the Potsdam University**.**

Riechmann, José Luis (2002) Transcriptional Regulation: a Genomic Overview. In: *The Arabidopsis Book*.pp. 1-46. Rockville, U.S.A, American Society of Plant Biologists**.**

Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R., Creelman, R., Pilgrim, M., Broun, P., Zhang, J.Z., Ghandehari, D., Sherman, B.K. and Yu G. (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 290(5499), 2105-10.

Riechmann, J.L. and Ratcliffe, OJ. (2000) A genomic perspective on plant transcription factors. *Curr Opin Plant Biol*. 3(5), 423-34.

Rob, P. and Coronel, C. (1997) Database Systems: Design, Implementation, and Management. USA, Course Technology Press

Rodríguez, M., Canales, E. and Borrás-Hidalgo, O. (2005) Molecular aspects of abiotic stress in plants. *Biotecnología Aplicada* 22, 1-10.

Satzinger, J.W., Jackson, R.B. and Burd, S.D. (2002) System Analysis and Design in a Changing World. USA, Course Technology Press.

Shameer, K., Ambika, S., Varghese, S.M., Karaba, N., Udayakumar, M., Sowdhamini, R. (2009) STIFDB: Arabidopsis Stress Responsive Transcription Factor DataBase. *International Journal of Plant Genomics.*

Schneider, M., Tognolli, M. and Bairoch, A. (2004) The SWISSPROT protein knowledgebase and ExPASy. *Plant Physiol. Biochem*. 42(12), 1013-21.

Schneider, M., Bairoch, A., Wu, C.H. and Apweiler, R. (2005) Plant Protein Annotation in the UniProt Knowledgebase. *Plant Physiol.* 138(1), 59-66.

Schneider, M., Lane, l., Boutet, E., Lieberherr, D., Tognolli, M., Bougueleret, L. and Bairoch, A. (2009) The UniProtKB/SWISSPROT knowledgebase and its Plant Proteome Annotation Program. *J. Proteomics* 72(3), 567-73.

Schröder, A., Eichner, J., Supper, J., Eichner, J., Wanke, D., Henneges, C. and Zell, A. (2010) Predicting DNA-Binding Specificities of Eukaryotic Transcription Factors. *PLoS ONE.* 5(11), 13876.

Shen, H., Cao, K. and Wang, X. (2008) AtbZIP16 and AtbZIP68, two new members of

GBFs, can interact with other G group bZIPs in Arabidopsis thaliana. *BMB Rep*. 41(2),

132-8.

Schwechheimer, C., Zourelidou, M. and Bevan, M.W. (1998) Plant transcription factor
studies. *Annu Rev Plant Physiol Plant Mol Biol.* 49, 127-150.

Silberschatz, A., Korth, H. and Sudarshan, S. (2005) Database System Concepts.

Boston, McGraw-Hill

Singh, K., Foley, R.C. and Oñate-Sánchez, L. (2002) Transcription factors in plant

defence and stress responses. *Curr Opin Plant Biol*. 5(5), 430-6.

Stegmaier, P., Kel, A.E. and Wingender, E. (2004) Systematic DNA-binding domain

classification of transcription factors. *Genome Inform*. 15(2), 276-86.

Steiert, H. (1998) Towards a Component-based n-Tier Client-Server Architecture.

Florida, USA.

Sundar, AS., Varghese, S.M., Shameer, K., Karaba, N., Udayakumar, M. and

Sowdhamini, R. (2008) STIF: Identification of stress-upregulated transcription factor

binding sites in *Arabidopsis thaliana*. *Bioinformation* 2(10), 431–437.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 22(22), 4673-80.

Tossi, A. Sandri, L. and Giangaspero, A. (2002) New consensus hydrophobicity scale extended to non-proteinogenic amino acids. *In Peptides: Proceedings of the twenty-seventh European peptide symposium. Edizioni Ziino*, 416-417.

Val, C., Pelz, O., Glatting, K., Barta, E. and Hotz-Wagenblatt, A (2010) PromoterSweep: a tool for identification of transcription factor binding sites. *Theoretical Chemistry Accounts* 125 (3-6), 583-591.

van Verk, M.C., Gatz, C. and Linthorst, H.J.M. (2009) Transcriptional Regulation of Plant Defence Responses. *Adv. Bot. Res*. 51, 397-438.

Varshney, R.K. and Koebner, RMD. (Eds.) (2007) Model Plants and Crop Improvement. USA, CRC Press.

Wanchana, S., Thongjuea, S., Ulat, V.J., Anacleto, M., Mauleon, R., Conte, M., Rouard, M., Ruiz, M., Krishnamurthy, N., Sjolander, K., van Hintum, T. and Bruskiewich, R.M.. (2008) The Generation Challenge Programme comparative plant stress-responsive gene catalogue. *Nucleic Acids Res. 36* (Database issue).

Ward, P. and Dafoulas, G. (2006) Database management systems. Belmont, Cengage Learning Business Press.

Whitten, J.L and Bentley, L. (1998) Systems Analysis and Design Methods. Boston, McGraw-Hill

Wieringa, R, (1998) A survey of structured and Object Oriented Software specification Methods and Techniques. *ACM Computer Surveys* 30 (4), 459-527.

Wingender, E. (1997) Classification of eukaryotic transcription factors. *Mol. Biol*. 31(4), 584-600.
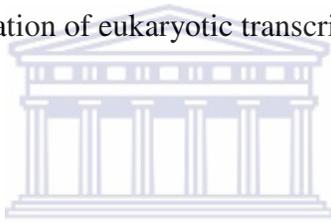
Zhang, H., Jin, J.P., Tang, L., Zhao, Y., Gu, X.C., Gao, G., Luo and J.C. (2011) PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. *Nucleic Acids Research* 39, D1114-D1117.

Zhang, W., Ruan, J., Ho, T.H,, You, Y., Yu, T. and Quatrano, R.S. (2005) Cis-regulatory element based targeted gene finding: genome-wide identification of abscisic acid- and abiotic stress-responsive genes in Arabidopsis thaliana.

Zimmermann, I.M., Heim, M.A., Weisshaar, B. and Uhrig, J.F. (2004) Comprehensive identification of *Arabidopsis thaliana* MYB transcription factors interacting with R/B-like BHLH proteins. *Plant Journal*; 40(1), 22-34.

APPENDIX I

Python SWISSPROT parser script

Parses a SWISSPROT flat file into individual records and imports the records into

a MYSQL table called DASTF

```python
#!/usr/lib/python2.4
from __future__ import division  # it ensures that  the division 1/2 --> 0.5 and not 0

"""
FILE: swiss.py
USAGE: swiss.py flatfile
"""

"""
This script is for getting certains fields from UniProt flatfiles of a given accession
"""

import random, math, sys, os, glob, time, numpy, MySQLdb

def parse(swiss_file):
        filename = open(swiss_file)

        counter_RX = 0
        counter_CC = 0
        label = []
        seq = ""
        uni_list = []

        for line in filename.readlines():
         field =  line.split()
         if field[0] == "ID":
             print "ID = ", field[1]
             uni_list.append(field[1])
             ID = field[1]
         elif field[0] == "AC":
             print "AC = ", field[1].split(";")[0]
             uni_list.append(field[1].split(";")[0])
         elif field[0] == "RX":
             counter_RX = counter_RX + 1
             if counter_RX ==  1:
```

```
      for i in range(len(field)):
        if "PubMed" in field[i]:
          pb = field[i].split("=")[1].split(";")[0]
      uni_list.append(pb)
      print "Pubmed = ", pb
      #print uni_list
  elif field[0] == "DR":
      if field[1].split(";")[0] == "GeneID":
        gene = field[2].replace(";", " " )
        uni_list.append(gene)
       print "GeneID = ", gene
        #print uni_list
  elif field[0] == "PE":
      pe = line.split(":")[-1].replace(";", "").replace("Evidence",
          "Evidence").rstrip()
        print  "PE = ", pe
        uni_list.append(pe)
      #print uni_list
  elif field[0] == "CC" and len(field) > 2:
      if field[2].split(":")[0] == "FUNCTION":
       field[3:] = [' '.join(field[3:])]
       print "Function = ", field[3].split(".")[0]
       uni_list.append(field[3].split(".")[0])
       #print uni_list
     elif field[2].split(":")[0] == "SIMILARITY":
       counter_CC = counter_CC + 1
       if counter_CC ==  1:
        #print field[3:] = [' '.join(field[3:])]
        field[3:] = [' '.join(field[3:])]
        print "Family = ", field[3].split(".")[0]
        uni_list.append(field[3].split(".")[0])
        #print uni_list
  #if field[0] not in label:
  #  label.append[field[0]]
  #if field[0] not in label:
  #  label.append(field[0])
  if field[0] not in ['ID', 'AC', 'DT', 'DE', 'GN', 'OS', 'OC', 'OX', 'RN', 'RP',
     'RC', 'RX', 'RA', 'RT', 'RL', 'CC', 'DR', 'PE', 'KW', 'FT', 'SQ','//']:
     field[0:] = [' '.join(field[0:])]
     seq =  seq + field[0].replace(" ", "")
     #print seq
print "sequence = ", seq
uni_list.append(seq)

conn = MySQLdb.connect (db = "dastf")
cursor = conn.cursor ()
```

```python
        cursor.execute('insert into uniprot values
                        ("%s","%s","%s","%s","%s","%s","%s","%s")'%
                        (uni_list[0],uni_list[1],uni_list[2],uni_list[3],uni_list[4],
                         uni_list[5],uni_list[6],uni_list[7],))

def main():
        if len(sys.argv) != 2:
                print 'Usage: python swiss.py swiss_flat_file'
                sys.exit()
        swiss_file = sys.argv[1]
        parse(swiss_file)

if __name__ == '__main__':
   main()
```

APPENDIX II

Perl TRANSFAC parser script

Parses TRANSFAC file containing binding site motifs and retrieves: TAIRID, Motif name, Motif sequence, Strand direction, Position, Core score and Matrix score

```perl
#! /usr/bin/perl
#
#
use strict;

$/="\nScanning sequence ID:";
my $f = shift;
my $rec= 0;
my %hash;
open(F, $f);
while (<F>) {
        $rec++;
        chomp;
        next unless ($rec > 1);
        next if ($_=~/No sites found for this sequence/);
        $_=~/^(\s+)(\S+)/;
        my $acc = $2;
        #$hash{$acc} = 1;
#       print "$acc\n";
        my @lines = split(/\n/);
        foreach my $line(@lines) {
                next if ($line =~/^$/);
                next unless ($line =~/^P/);
                #print "$line\n";
                my ($tfbs_id, $pos, $str, $cmatch, $mmatch, $string, $tfbs_name) = split(/\s+/, $line);
                #print "$acc $tfbs_id $pos $str $cmatch $mmatch $string $tfbs_name\n";
                my @e =  ($pos,$cmatch,$mmatch,$str, $tfbs_name,$string);
                push @{$hash{$acc}}, [@e];
        }
}
close(F);
foreach my $acc(keys %hash) {
        foreach my $e(@{$hash{$acc}}) {
                print       "$acc\t".$e->[0]."\t".$e->[1]."\t".$e->[2]."\t".$e->[3]."       ".$e->[4]."\t".$e->[5]."\n";
        }
}
```

APPENDIX III

TABLE 1.5: Extended list of erroneous entries included in STIFDB

	TAIRID	GENBANKID	SWISSPROTID	NAME/ALIAS	FUNCTION
1	AT1G11310	837673	B3H6R0	MLO2	biotic
2	AT2G02100	814741	Q39182	LCR69	biotic
3	AT2G02130	814744	Q9ZUL7	LCR 68	biotic
4	AT2G14560	815943	Q9ZQR8	LURP1	biotic
5	AT2G14610	815949	P33154	PR-1	biotic
6	AT2G19990	816518	Q39186	PR-1-LIKE	biotic
7	AT2G26560	817197	O48723	PLA2A	biotic
8	AT4G02150	827472	O04294	MOS6	biotic
9	AT4G14400	827085	Q8LPS2	ACD6 ACD6	biotic
10	AT4G31550	829282	Q9SV15	WRKY11	biotic
11	AT5G06320	830520	Q9FNH6	NHL3 NHL3	biotic
12	AT5G20900	832214	Q9C5K8	JAZ12	biotic
13	AT1G02870	839496	Q8RWK5	F22D16.13	unknown
14	AT1G03250	838557	Q8L3X7	F15K9.15	unknown
15	AT1G03610	838961	Q8LF98	F21B7.22	unknown
16	AT1G04960	839346	A8MRN9	F13M7.5	unknown
17	AT1G05340	837033	O23035	YUP8H12.4	unknown
18	AT1G07040	837215	Q9LMJ7	F10K1.25	unknown
19	AT1G10020	837537	O80593	T27I1.4	unknown
20	AT1G12080	837760	O65370	F12F1.4	unknown
21	AT1G12830	837839	Q9LPW6	F13K23.8	unknown
22	AT1G13990	837959	Q94F46	F7A19.8	unknown
23	AT1G17830	838361	Q9LMU5	F2H15.6	unknown
24	AT1G18060	838386	Q9LM40	T10F20.7	unknown
25	AT1G19400	838523	Q8VYC6	F18O14.16	unknown
26	AT1G20100	838599	Q9LNT6	T20H2.11	unknown
27	AT1G21500	838749	C0Z3H5	F24J8.11	unknown
28	AT1G22750	838881	Q949W5	T22J18.8	unknown
29	AT1G23710	838981	Q9ZUC4	F5O8.26	unknown
30	AT1G26470	839188	Q9FZD2	T1K7.16	unknown
31	AT1G26650	839205	Q94CC9	T24P13.3	unknown
32	AT1G27020	839591	O04551	T7N9.8	unknown
33	AT2G02515	814781	Q8S8R5	No Name	unknown
34	AT2G03350	814864	Q9ZQ71	T4M8.22	unknown
35	AT2G19160	816433	Q4DPX0	T2G19160	unknown
36	AT2G19180	816435	Q94F29	T20K24.20	unknown
37	AT2G19270	816444	O64560	F27F23.7	unknown
38	AT2G19390	816458	Q93YU9	F27F23.19	unknown
39	AT2G20740	816603	Q9SKU3	F5H14.29	unknown
40	AT2G21180	816653	Q9SKP5	F26H11.6	unknown

41	AT2G22660	816797	Q9ZQ47	T9I22.10	unknown
42	AT2G23120	816844	Q94K79	F21P24.18	unknown
43	AT2G24100	816944	Q9ZUI1	F27D4.1	unknown
44	AT2G28370	817385	Q9SKN3	T1B3.11	unknown
45	AT2G28400	817388	Q9SKN0	T1B3.8	unknown
46	AT2G28570	817405	Q9SK01	T17D12.13	unknown
47	AT2G28690	817418	Q9SIA6	T8O18.2	unknown
48	AT3G02420	821152	Q9M898	F16B3.5	unknown
49	AT3G02640	821289	Q9M878	F16B3.27	unknown
50	AT3G03870	821099	Q8RXM3	F20H23.8	unknown
51	AT3G04550	819611	Q9SR19	F7O18.2	unknown
52	AT3G06080	819781	Q93YQ2	F24F17.6	unknown
53	AT3G07350	819923	Q4WWW7	F21O3.6	unknown
54	AT3G07760	819967	Q93VV3	MLP3.21	unknown
55	AT3G07790	819970	Q9S7V6	MLP3.24	unknown
56	AT3G09180	820074	Q8RWM3	MZB10.22	unknown
57	AT3G10020	820163	B3H7G1	T22K18.16	unknown
58	AT3G12320	820411	Q9LHH5	T2E22.34	unknown
59	AT4G01150	828181	O04616	F2N1.18	unknown
60	AT4G03420	827928	Q9ZT70	F9H3.4	unknown
61	AT4G12340	826843	Q9STH7	T4C9.180	unknown
62	AT4G20480	827796	Q8H1G2	F9F13.130	unknown
63	AT4G21930	828282	Q8L864	F1N20.2	unknown
64	AT4G25670	828672	Q9SZZ5	L73G19.50	unknown
65	AT4G26130	828719	Q9SZI4	F20B18.240	unknown
66	AT4G27350	828843	O81838	F27G19.7	unknown
67	AT4G27450	828854	Q9SZS0	F27G19.50	unknown
68	AT4G32020	829333	O49389	F10N7.170	unknown
69	AT4G32340	829368	O49358	F8B4.40	unknown
70	AT4G33960	829542	O81764	F17I5.150	unknown
71	AT5G01350	830568	Q93W37	T10O8.60	unknown
72	AT5G03210	831903	Q949Q2	F15A17.240	unknown
73	AT5G03230	831900	Q9LYW2	F15A17.260	unknown
74	AT5G03460	831829	Q9LZE0	F12E4.230	unknown
75	AT5G04550	830334	Q9LZ71	T32M21.140	unknown
76	AT5G06980	830589	Q9FL48	MOJ9.15	unknown
77	AT5G08400	830738	B9DGL0	F8L15.130	unknown
78	AT5G11280	830998	Q94AQ7	F2I11.170	unknown
79	AT5G11420	831013	Q8H168	F15N18.10	unknown
80	AT5G11680	831040	Q4WVD2	T22P22.70	unknown
81	AT5G13970	831245	Q9FFX8	MAC12.6	unknown
82	AT5G16550	831517	Q94EY7	MQK4.30	unknown
83	AT5G18130	831931	Q3E9G1	MRG7.9	unknown
84	AT5G18420	831960	Q8L846	F20L16.140	unknown
85	AT5G19860	832107	Q7XA63	T29J13.3	unknown

86	AT5G22280	832288	Q9FMS3	T6G21.7	unknown
87	AT5G24450	832516	Q8VZB0	T31K7.3	unknown
88	AT5G24990	832569	Q94AK1	F6A4.200	unknown
89	AT5G25265	832598	Q8W4E6	unknown	unknown
90	AT5G25770	832646	Q8L7E1	F18A17.20	unknown
91	AT5G27730	832835	Q94CC1	T1G16.60	unknown
92	AT5G35320	833486	O65233	T26D22.2	unknown
93	AT5G35460	833509	Q9FJB4	MOK9.4	unknown
94	AT5G37360	833710	Q93Z11	MNJ8.18	unknown
95	AT5G38380	833821	Q8RWM9	MXI10.5	unknown
96	AT5G39570	833953	Q8L7C5	MIJ24.6	unknown
97	AT2G04460	814986	Q6NNK5	T1O3.13	unknown
98	AT2G26530	817194	Q949N6	T9J22.20	unknown
99	AT4G32190	829352	Q8H1E5	F10M6.170	unknown
100	AT4G33666	829508	Q94AJ7	unknown	unknown