Analyses of Sequence Divergence Using Completely Sequenced Genomes

Msc Bioinformatics, Department of SANBI,
University of the Western Cape

Victoria P. Nembaware

A mini thesis submitted in partial fulfilment of the requirements for the degree of Masters in Bioinformatics in the Department of SANBI, University of the Western Cape.

January 2003

Supervisor Dr Cathal Seoighe.

Keywords

Divergence
K. lactics
S. cerevesiae
BLAST

Abstract

Using the complete genome, *Saccharomyces cerevisiae*, which duplicated after its speciation from *Kluyveromyces lactics*, a dataset of 119 putative *S. cerevisiae* – *K. lactis* ortholog-pairs was constructed. *S. cerevisiae* paralogous pairs that are likely to have duplicated during the whole genome duplication of *S. cerevisiae* were obtained and the approach taken in our previous work (Nembaware *et al.*, 2002), was repeated to test whether the presence of a paralogue in *S. cerevisiae* had an effect on the rate of sequence divergence of the 119 pairs of orthologous genes. We found, however, that substitutions at synonymous sites had reached saturation and this prevented us from being able to repeat the previous finding with *S. cerevisiae* and *K. lactis* . From this study a publicly available web-server (http://hamlyn.sanbi.ac.za/~victoria) that automates the calculation of Ka:Ks values given a pairs homologous CDS sequences is presented.

The second part of this work consisted of clustering a protein set from 15 completely sequenced genomes based simply on results of BLAST homology searches. Observations from this study will contribute to future work on and the creation of a database of orphan genes.

Acknowledgements

I would like to express my gratitude to my supervisor Dr C. Seoighe, for his continuous encouragement and guidance throughout this investigation. He is acknowledged, in particular, for his patience and understanding. I gratefully acknowledge the support and presence in this study of the SANBI and EG team for the various illuminating discussions and valuable comments. I am grateful to my family, Justin and Jack for their support.

Declaration

I declare that, **Analyses of sequence divergence using completely sequenced genomes** is my own work that it has not been submitted for any degree or examination in any other university and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Full name: Victoria P. Nembaware January 2003

Signed:

Table of Contents:

Chapter 1

Literature Review

1. Literature Review

1.1 What is gene duplication?

Gene duplication is a mutation that results in identical copies of a gene thus increasing the number of copies of a DNA segment (Li, 1997). Characteristic of any other mutation, gene duplications are random, occur across phyla and the sizes of the duplicated DNA segment fluctuate. Gene duplications can occur on a small-scale, that is, gene segments, or at the level of complete genes. On the other hand the duplicated genes could be products of larger scale duplications i.e. whole chromosomes or even whole genomes can be duplicated (Wolfe, 2001). In this review the emphasis is on any form of duplication that results in whole genes increasing in number unlike gene segments i.e. domains and exons.

1.2 Evidence of Duplications

1.2.1 Whole genome duplications

Although met with a huge amount of scepticism (Hughes *et al., 2001*), the hypothesis of whole-genome duplications in a number of organisms has been backed by a significant amount of evidence (Vision and Tanksley, 2000). Whole gene duplications are rare events that are followed by massive gene loss, however evidence of their occurrence can be found in many gene clusters. For example, invertebrates have a single copy of the Hox gene cluster whereas vertebrates such as human have four copies, implying the occurrence of large (or even whole-genome) duplications after the split of vertebrates from invertebrates (Hughes, 1999).

Substantial evidence of whole genome duplications comes from whole genome sequences. The dot matrix plots of the complete Arabidopsis genome, pairs off most of the genomic segments (Vision and Tanksley, 2000), which could only be explained by the occurrence of at least one whole genome duplication. The Arabidopsis Genome Initiative, 2000, and Lynch and Conery, 2000 share the same findings of a whole Arabidopsis genome duplication having occurred. Lynch and Conery's conclusion is based on their dating of gene duplications in Arabidopsis that were estimated to have

occurred 65 million years ago. Several other organisms have shown strong evidence of whole genome duplication having occurred in the past (Wolfe and Shields 1997; Wong *et al.,* 2002).

Evidence of whole genome duplication in the human genome was not conclusive from the initial draft. This issue was recently addressed by Gu *et al., 2002,* who supported a whole genome duplication in human from the observation of a rapid increase in paralogous genes after the speciation of human from invertebrates (Martin, 1999; Wolfe, 2001). Various other independent studies and extensive discussions support whole genome duplication in human (McLysaught *et al.,* 2002; Wolfe, 2001). In view of this more organisms are expected to show similar evidence with the availability of their complete sequences. However, due to a high rate of divergence and gene loss between ancient duplicated segments most duplicates could have been lost or diverged rapidly such that they are no longer detectable.

1.2.2 Small-scale duplications

There has been substantial evidence for both recent and ancient segmental duplications (Eichler, 2001). Using a quarter of human gene families, the age distribution of the gene families highlighted an average rate of ~10 million segmental duplications per genome per million years (Gu *et al.,* 2002). Previously, using a different method, Lynch and Conery, 2000 estimated a similar average rate of the occurrence of segmental and tandem duplications.

Syntenic regions in comparative genome maps can provide evidence of segmental duplication (Blanc *et al.,* 2000). In hominids, the gene order is highly conserved in paired chromosomal regions (Wolfe, 2001). In the human genome similar regions on chromosomes 2,7,12 and 17 are evidence of ancient duplications (Venter *et al.,* 2000).

1.2.2.1 Gene families

Chromosomal mapping studies have shown that consecutive gene duplications (both large scale and small scale), give rise to gene families (Li, 1997). Gene families are a group of duplicate genes that may have similar functions. Members of a gene family

usually exist as a cluster. In spite of gene translocations, and deletions a significant number of gene clusters are still intact and they provide evidence of recent tandem gene duplications (Brooke *et al.,* 1998).

1.3 Mechanisms of duplications

1.3.1 Replication errors

Replication errors are frequent both in somatic and gametic cells. Such errors if not fatal, could be disease causing (Li, 1997). One such error is unequal crossing over during meiosis that can cause tandem gene duplications (Li, 1997). In addition to tandem gene duplications, there is a high chance of reproduction errors ultimately resulting in the doubling of genomes. In organisms with well-differentiated sex chromosomes, polypoidy is likely to be eliminated, unlike in bisexual genomes (Wolfe, 2001).

1.3.2 Jumping genes

Transposons, 'jumping genes' also ensure the continual occurrence of gene duplications. (Li *et al.*, 2001). There are two distinct types of transposons Class I and Class II. The retrotransposons (Class I) transcribe DNA into RNA. The RNA is then reverse transcribed and inserted into a new location (Clark and Kidwell, 1997). This mechanism results in duplicates being found on different chromosomes, which are almost always in opposite direction. Class II transposons simply copy DNA and translocate it into another part of the genome. Activation of a transposed gene is rare, however experimental work supports a feasible mechanism (Moran, 1999).

1.4 Fate of duplicated genes

1.4.1 Gene loss

Amino amino acid altering mutations occur in all organisms and a large fraction of them are deleterious (Fay and Wu, 2001). In hominids, ~38% of amino acid mutations are significantly deleterious (Eyre-Walker *et al.,* 1999). Due to purifying selection,

most deleterious mutations are lost from the gene pool. However, in the case of duplicates, deleterious mutations on one copy are likely to escape any form of selection, as the other gene pair will still be functioning normally. Such deleterious mutations can result in the formation of pseudogenes (Lynch and Conery, 2000; Harrison *et al.*, 2002).

Lynch and Conery, 2000 also suggested that the silencing of one gene of a duplicate gene pair is important in the speciation of organisms. This speciation model describes how the loss of different copies of a duplicated gene in geographically separated populations could genetically isolate these populations (Lynch and Conery, 2000).

1.4.2 Preservation of duplicates

1.4.2.1 Redundancy in function

Redundancy in function is common in recent duplicates but rare in older ones (Lynch and Conery, 2000). Maintenance of this redundancy in function is often advantageous depending on the functional specificity of the redundant pair (Gu *et al.*, 2002). Several cases of identical paralogs are associated with increased quantities of the gene products required for normal functioning of the organism (Nowak *et al.*, 2001). Paralogs with a transcriptional function frequently maintain redundancy in function so as to produce large amounts of the protein when needed (Li, 1997). Genes for rRNA and tRNAs are in duplicates and the duplicates maintain the same function as they are required in large quantities during the S phase of the cell cycle when DNA is replicated (Li, 1997). Yet another example of the advantage of maintaining redundancy in function in paralogous gene pairs is found in *S. cerevisiae*. When phosphate is a limiting factor to growth in *S. cerevisiae*, a duplicate of the phosphate acid carrier gene enables the gene to produce twice the amount of enzyme to utilize the surplus phosphate effectively (Li, 1997).

Maintaining redundancy in function in paralogs minimizes the phenotypic effects of null alleles and/or developmental accidents hence increase fitness of the organism (Wagner, 2000; Cooke *et al.*, 1997). In multi-domain developmental genes, point mutations in one of the redundant gene pair often have severe impact on the organism

unlike complete deletion of these genes (Gibson and Spring, 1998). This could be interpreted as a way of preventing loss in redundancy of such genes (Gibson and Spring, 1998). Its crucial not to make any conclusion on a pair of paralogs that have maintained redundancy in function before dating their duplication event, as young duplicates are most likely to have redundant functions.

1.4.2.2 Creation of novel gene functions

Soon after duplication both copies of the gene perform the same function. With two copies retaining the same function, selection pressure may be freed on one copy; the result could be the creation of a novel gene (Force *et al.*, 1999; Lynch and Conery, 2000). There has even been speculation that this creation of new function might even result in the formation of a new species (Taylor *et al.*, 2001). Differentiation in function usually requires a significant number of amino acid changing substitutions. However there are cases in which only one amino acid replacement is required for creation of differentiation of function of paralogs. For example in the case of lactate dehydrogenase and its paralog, malate dehydrogenase, only one amino acid replacement is required for switching between the two (Li, 1997).

Functional divergences of duplicates could be the underlying cause of physiological and morphological differences in ancestrally related species. Due to duplicates evolving new functions, *S. cerevisiae* has acquired novel genes that make it a far better sugar fermentor than any of the other yeasts (Wolfe and Shields, 1997). Although the new metabolic functions acquired by *S. cerevisiae* are in disagreement with Cooke *et al.*, 1997 who pointed out that developmental genes are more likely to acquire new developmental functions than metabolic functions, this finding makes the *S. cerevisiae* duplicates of great interest.

Duplication followed by divergence could be the fundamental genetic change that sparked the progression of ancient invertebrates into complex vertebrates. Duplication of the AmpiEomes gene of the invertebrate amphioxus gave rise to two genes Eomesodermin and T-brain-1 in vertebrates (Ruvinsky *et al.*, 2000). The Eomesoderm and AmpiEomes have an identical function in the mesoderm formation, however the

duplicate copy T-brain-1 has acquired a novel function essential in the development of the forebrain (Ruvinsky *et al.*, 2000).

1.4.2.3 Subfunctionalization

Subfunctionalization of duplicates has recently been proposed (Lynch and Force, 1999). Complementary mutations can occur on the duplicates especially if the original gene performed more than one function or coded for multiple domains (Force *et al.*, 1999). In such instances, the function of the parent gene could be partitioned between the duplicates (Lynch and Force, 1999). Survival of gene clusters resulting from tandem duplications can be attributed to subfunctionalization. Most gene clusters code for proteins subunits of a functional complex. For activation of such complexies, the proteins segments anneal to each other. Selection could favour the compactness of the clusters, as it would become extremely difficult to co-ordinate augmentation of the gene clusters' products if the individual genes were positioned far from each other (Skaer *et al.*, 2002).

1.5.1 Comparative genomics

Homology is based on common ancestry and thus provides information about past evolution of related species (Fitch, 2000). There are two broad categories of homologous sequences, orthologs and paralogs. Orthologs refer to genes in different genomes that have evolved vertically from a common ancestral gene (Theben, 2002; Li, 1997). Paralogs on the other hand, as discussed in the previous sections, are homologs within the same genome and result from duplication events (Jensen, 2001).

To detect homology, comparative analysis of species that originated from the same ancestor is used. Darwin's comparative analysis of morphological features has given way to sequences comparisons. Comparative genomics is the art of equating genomes at molecular level in search of their similarities and differences (Copley *et al.*, 2001; Makalowski and Boguski, 1998). The similarities and differences may reflect how the species have diverged from their ancestral species in terms of gene birth and death. Gene duplicates and their subsequent fates make the field of comparative genomics

fascinating as one to one mapping of orthologous sequences is almost always distorted.

Comparison of full genome sequences of 13 bacterial species has allowed for the first time a chance to see the addition and loss of all the genes in compared species, showing that 20-50% of the genes are gained or lost (Snel *et al.*, 1999). Comparisons of genomes can be at different levels, gene structure, protein coding gene content and location or even nucleotide base frequencies.

1.5.2 Determining Orthology

It is vital to identify accurately orthologous pairs for evolutionary studies, functional assignments along with better understanding of variation and connections between species (Mushegian *et al.*, 1998; Chervitz *et al.*, 1998). In human model organisms, studies of orthologs in both disease and normal states have profound effect on disease control and eradication. Failure to resolve orthologs can lead to misinterpretation of disease gene biochemistry and unfruitful medical results (Mushegian *et al.*, 1998). Currently orthologs for 178 of the 287 human disease genes have been identified in Drosophila (Rubin., 2000). It is predicated that the number of 'orphan' disease genes will reduce radically with the availability of complete genome sequences of other primates (Rubin, 2000).

Functionally annotating novel protein sequences based on their high percentage of similarity with an annotated ortholog is dependent on the precision of the assignment (Mushiegan *et al.*, 1998). However there are cases in which orthologs may have subtle physiological and functional differences. For example, iron transporters in yeast have their human counterparts as copper transporters (Askwith and Kaplan, 1998).

Finding unique genes is just as important as assignment of orthologous relationships. There are two major explanations to the existence of orphan genes. Many orphans genes might consist of genes whose phylogenetic distribution is restricted to certain evolutionary lineages e.g. they are specific to plants or vertebrates. Some orphan genes may diverge rapidly between closely related species because the proteins they encode are unconstrained in their sequence evolution, or under selective pressure,

whereas their structure and function might be conserved even between distantly related organisms. Unique genes in bacterial species have been shown frequently to encode pathogenic products. Drugs can often be formulated that target such species-specific genes and their pathways.

Currently there is no sure way of assigning orthology. Regardless of the method used, there is always a need to validate the putative dataset construct (Wheelan *et al.*, 1999). With the growing number of complete genomes previously ortholog databases are continually being reviewed and revised.

1.5.2.1 Similarity methods

Sequence comparisons allow one to infer orthology through similarity measures. Orthology inferences using entire sets of proteins encoded by two species have increased confidence in the similarity-based methods (Huynen and Bork,1998). The first-ever comparison of two complete protein sets of *Caenorhabditis elegans* and *Saccharomyces cerevisiae* produced relatively consistent set of orthologous proteins as most of the pairs carried out the same biological functions (Chervitz *et al.*, 1998). Although orthology (as the term is used here) does not necessary mean the genes should carry out the same function, in the study by Chervitz *et al.*, 1998 most of the orthologs sharing a function perform core biological functions. Such genes with core-biology functionality could have been conserved after the speciation of *C. elegans* and *S. cerevisiae.*

Genes may have very different evolutionary rates (Mushegian *et al.*, 1998). With different evolutionary rates, pairs of orthologs from two organisms may exhibit different degrees of similarities making it difficult to decide what cut-off to use when inferring orthology from sequence similarity.

1.5.2.2 BLAST

Currently BLAST (Altshul, 1996) is the most widely used tool for inferring orthology. Although the Smith-Waterman algorithm is more sensitive than the BLAST algorithm, its CPU intensive character hampers its popularity. There have been

several concerns concerning the performance of BLAST in similarity searches. An assessment of sequence comparison methods in the identification of distant evolutionary relationships, BLAST was the worst performer (Brenner *et al.*, 1998). BLAST only identified 15% of the evolutionary relationships in comparison to SSEARCH and FASTA, which were capable of identifying about 18% the homologous relationships (Brenner *et al.*, 1998).

1.5.5.3 Limitations of BLAST

Multi-domain Proteins

BLAST's algorithm picks up similarities by local alignments, which are likely to be motifs or domains (Tatusov *et al.*, 1997). Non-orthologous proteins with identical domains are likely to be picked up as false positives. Hence, multi-domain proteins pose a serious problem in BLAST searches. Non-orthologous proteins that simply possess orthologous domains are classified falsely as orthologs.

Convergent evolution

Convergent evolution has been known to occur. Genes that evolved in parallel exhibit a degree of similarity posing problems in orthology assignment (Haney, 1999; Huynen and Bork, 1998). Convergent evolution is common in bacterial species hence non-orthologous genes can be falsely classified as homologs (Haney, 1999).

Sequence divergence

Due to divergence, similarities in homologs can be corroded such that BLAST cannot pick them as homologs. A specific example is the Guanylyl cyclase gene from *A. thaliana*. The guanylyl cyclase gene product is required for catalysing the formation of guanosine monophosphate (cGMP) from guanosine triphosphate (GTP). When the *A. thaliana* guanylyl cyclase gene is BLAST searched against the NCBI database no significant hits are found in Arabidopsis itself nor to guanylyl cyclase genes in other organisms. (Ludidi and Gehring, 2000). The proteins that had significant hits with the guanylyl cyclase genes were human predicted proteins surprisingly the other guanylyl

cyclase homologoues had very poor scores with the query protein. A plausible reason for this observation could be that the guanylyl cyclase under-study could have divergent rapidly from the rest of the gene family. There is also a chance that the molecule may not be a homologue of previously annotated guanylyl cyclases. The same problem is experienced with the presellin enzyme gene. Sequence database searches using all available methods including different types of profile analysis, have failed to detect any appreciable sequence similarity between presenilins and any known proteases which belong to the same family as itself (Steiner *et al.,* 2000). Pattern searches of both the guanylyl cyclase and presenilin gene identified signatures present in each of their families, suggesting that they are catalytically active. In-vitro assays of the guanylyl cyclase have also confirmed catalytic activity.

Gene duplications and Horizontal gene transfer.

Gene duplications, their subsequent fates and horizontal gene transfers result in complex orthologous relationships. When using BLAST the highest hit is considered the closest homolog of the query. Even though it is accurate in some instances, detection of the highest-hit is also not capable of identifying one (many)-to-many evolutionary associations.

Furthermore, in using BLAST there is no way of distinguishing between horizontal gene transfers and gene duplications (Huynen and Bork, 1998). Such transferred genes often display nucleotide frequencies different from the rest of the genome that allow them to be identified (Huynen and Bork, 1998).

1.5.3 Protein Clusters

The fastest and easiest way to do whole genome comparisons is through comparison of their open reading frames (ORFs) comparisons. Protein sequences allow detection of distant ancestral relationships to be established as compared to nucleotide sequences (Brenner *et al.,* 1998).

Post-editing of BLAST results, through clustering creates more informative networks of protein relationships. Due to the complex evolutionary relationships that exist

between proteins, clustering has been carried out to group homologous proteins. This involves making graphs. Each sequence is a vertex, an edge joins it to another protein. The edge can be weighted according to statistical significance of the alignment score between two proteins. This weighting allows one to remove and add edges to see how the number of vertices fluctuate, reflecting the strength and weaknesses of the evolutionary relationships between the proteins.

1.5.3.2 COGs

Clustering has paved way to databases such as COGs (Clusters of Orthologous groups) Tatusov *et al,* 2001). Similarity searches were performed on complete genomes. Orthologous proteins that are identified through reciprocal best hits are classified as a COG. The clusters would include orthologs from different organisms and sometimes a cluster may contain paralogs due to gene duplications. Functional classification of proteins from newly sequenced genomes using the COGs database has resulted in 17 broad functional groups.

1.6 Evolutionary distances

Ancestrally related DNA sequences commonly have a significant amount of differences. The observable differences in sequences alignments are indels and substitutions. The number of base substitutions per site can estimate evolutionary distances. Base substitutions that occur in protein-coding sequences can be classified as synonymous and non-synonymous substitutions because the degeneracy of the genetic code (Hughes, 1999). Synonymous mutations do not alter the amino acid sequences while non-synonymous substitutions do cause a change in the amino acid sequence. Non-synonymous substitutions have a high probability of being deleterious although they may have no effect or even improve the protein.

1.6.1 Ka/Ks calculations

The ratio of the number of nonsynonymous substitutions per nonsynonymous site (Ka) and the number of synonymous substitutions per synonymous site (Ks) is a powerful measure of the selective pressure acting on a pair of homologous sequences.

According to the Kimura's neutral theory of evolution, Ks is proportional to the mutation rate of the gene (Li, 1997). Commonly purifying selection is detected where Ka/Ks << 1, whereas in positively selected genes Ka/Ks is greater than 1. If amino acid changes are neutral owing to neutral substitutions, Ka/Ks is close to 1. However, Ka:Ks is not totally reliable, in cases where one part of the gene experiences positive selection and the other part neutral selection, Ka/Ks detects false purifying selection (Li, 1997; Hurst, 2002).

Besides evolutionary analysis, Ka:Ks ratio test has been used in increasing the accuracy of gene predictions (Nekrutenko *et al.,* 2001). The Ka:Ks ratio is used to identify/confirm protein coding exons using human/mouse orthologous sequences, as Ks is normally larger than Ka in coding regions (Nekrutenko *et al.,* 2001). Where the Ka/Ks ratio is significantly less than one for a human/mouse orthologous genomic sequence that contains a reading frame, such a region is more likely to be a protein-coding region (Nekrutenko *et al.,* 2001).
Although the Ka;Ks test cannot on its own be used for gene prediction as it is incapable of identifying exon/intron boundaries, promoters and poly-adenylation sites it can be incorporated into other gene prediction methods making them more robust (Nekrutenko *et al.,* 2001).

For Ka and Ks estimations to be feasible, sequences should be divergent enough to observe the substitutions that have occurred (Hughes, 1999). On the other-end of the spectra very divergent sequences can reach saturation in their Ks values, see **Figure 1.** The number of substitutions observed become too high to make an estimate of the number of substitutions that might have occurred.

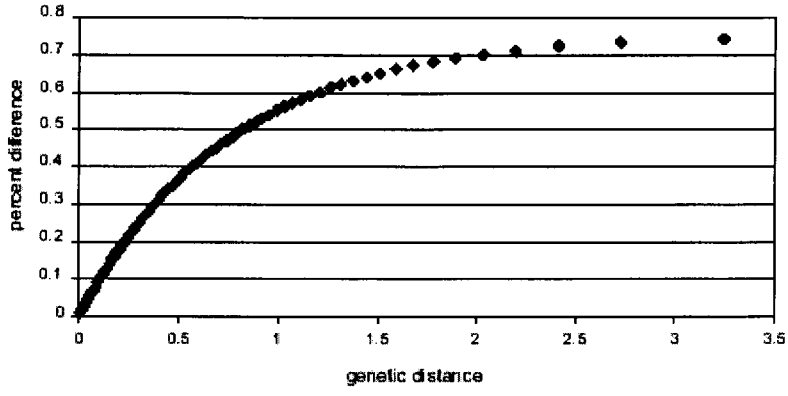

Figure 1: Effects of saturation on synonymous substitutions.

When differences between the sequences are above 75%, estimations of genetic distances are unreliable. This is from a simple Juke-Cantor model.

(Graph from http://hiv-web.lanl.gov/content/hiv-db/TREE_TUTORIAL/Tree-tutorial.html).

1.6.2 Conservative and Radical non-synonymous substitutions

The use of CDSs for Ka/Ks calculations may become unfeasible as synonymous substitutions have a tendency to become saturated. In addition to saturation, base frequencies may vary between sequences making alignments unfeasible. In such instances, selection studies on CDS sequences have been abandoned for protein sequences (Yang *et al.,* 1998; Zhang, 2000; Hughes, 2001). The ratio of radical (RA) and conservative amino acid (CO) replacements can be used to detect the type of selection that favours change of certain properties of amino acids (Hughes, 1999). Amino acids have distinct physiochemical properties that can be utilized for CO and RA for example charge and polarity. A radical substitution in a radical non-synonymous site (RA) can lead to changes in the amino acid hence in the protein as a whole (Hughes *et al.,* 2000; Hughes, 2000). Although there are cases in which positive selection in genes has been detected when RA > CO, there has been a strong critique of this method (Smith and Hurst, 1998). The method has no way of distinguishing between other factors (unrelated to selection) that might result in an increase in RA as compared to genuine amino acid changes in favour of adaptation (Smith and Hurst, 1998).

Chapter 2

Effect of paralogs on the rate of divergence of *Kluvyromyces lactics* and *Saccharomyces cerevisiae* orthologs

2.1 Introduction

Gene duplications have unquestionably played a significant role in the innovation of gene functions (Li, 1997; Lynch and Force, 2000; Martin, 1999) and developmental pathways (Li, 1997). Furthermore gene duplication events have recently been hypothesised to contribute in the creation of new species (Lynch, 1999; Nowak *et al.*, 1997). Although neutral evolution promotes loss of a significant portion of duplicates the few that are retained are of great value in studying sequence divergence patterns (Hughes, 1993).

In an attempt to elucidate the evolutionary forces acting on pairs of paralogs, Kondrashov *et al.*, 2002 reported that young duplicated genes evolve faster than orthologs that share the same magnitude of divergence. Using a genome wide analysis of a nearly complete genome Lynch and Conery, 2000 suggested neutral evolution immediately after duplication followed by purifying pressure as the paralogs acquire different roles. However, Lynch and Conery, 2000 did not take into account the effect of averaging Ka:Ks of gene pairs that diverged at different times. To illustrate the limitation of Lynch and Conery's study, in a previous study we estimated the Ka:Ks ratios between 119 pairs of chimpanzee-human orthologs with diverse divergence times (Nembaware *et al.*, 2002). The average Ka:Ks obtained was far too high, reflecting a figure that would be appropriate only for recently diverged sequences (Nembaware *et al.*, 2002).

To measure the effect of the presence of a human duplicate on the rate of sequence divergence more accurately, two databases of human-mouse orthologs were created based on the synonymous distance of the human gene to its paralog (Nembaware *et al.*, 2002). From an initial human-mouse orthologous dataset of 5341 gene pairs, 180 human-mouse orthologs with close paralogs were identifed followed by a set of 70 gene pairs with intermediate human paralogs. By using only gene pairs that diverged during a common period our results provided greater insight on the effect a paralog can have on sequence diverge of orthologs (Nembaware *et al.*, 2002). For a set of paralogs that appeared at about the same time as the human and mouse speciation, acceleration in the rate of non-synonymous substitution in a set of mouse and human

orthologs was observed compared to the human-mouse orthologs with a very close paralog (Nembaware *et al.*, 2002).

Aims of study

The aim of the study presented here was to replicate a previous study done on human-mouse orthologs (Nembaware et al., 2002) on *S. cerevisae* and *K. lactics*. According to Wolfe and Shields, 1997, the *S. cerevisiae* whole genome duplication occurred after the speciation of S. cerevisae from *K. lactics,* making these two genomes ideal for assembling a dataset for use in this current study. As mentioned previously, it is essential to compare Ka:Ks ratios of gene pairs that share a common divergence time. The *S. cerevisiae* paralogs from Wolfe amd Shields' dataset are ideal for this as they diverged at the same time.

Although use of Ka:Ks for detection of positively selected genes has its shortfalls (Hughes, 1999), the Ka:Ks ratios are useful for characterization of the evolutionary forces acting on homologs (Zhang, *et al.*, 2000; Nembaware *et al.*, 2002), therefore this method was implemented in this study.

Another aim of the study was to create web-server that automates estimation of Ka:Ks values for sets of protein-coding nucleotide sequences.

2.2 Data and Methods

2.2.1 *K. lactis*

K. lactis 144 proteins were downloaded from the SWISSPROT database (http://srs.ebi.ac.uk/srs6bin/cgi-bin/wgetz) on the 15/10/2001. Proteins from the SWISSPROT database were chosen over other databases as it is a manually curated database with minimal redundancy. All the available K. lactis nucleotide sequences, which amounted to 7106, were also downloaded from the NCBI taxonomy browser (http://www.ncbi.nlm.nih.gov).

2.2.2 S. cerevisiae

The set of sequences used by Wolfe and Shields (1997) to investigate whole genome duplication in *S. cerevisiae* was obtained from their website. This set consisted of 5790 open reading frames as well as the nucleotide sequences of the 16 *S. cerevisiae* chromosomes. 406 pairs of paralogs retained after the *S. cerevisiae* genome duplication were also part of the *S. cerevisiae* data. The presence of these 406 gene paralogous pairs thus served as a point of entry into this current study

2.2.3 Identification of *K. lactis* and *S. cerevisiae* Orthologues

144 *K. lactis* proteins were used as queries in a BLASTP search against the whole S. cerevisiae genome ORFs to identify putative orthologs. A threshold E-value cut-off of 0.0001 was used after examining the score distribution. Similarity measures may not be the best method for assigning orthologous relationships but if one or both genomes are completely sequenced it has been shown to produce plausible results (Rubin, 2000). Orthologous relationships between *K lactis* and *S. cerevisiae* are expected to be 1:1 or 1:2 mappings as speciation from *K lactis* predates the whole genome duplication of *S. cerevisiae* which occurred ~100 million years ago. The set of 406 *S. cerevisiae* paralogs aids in the task of distinguishing between ortholog-pairs that have a surviving paralog (1:2 orthologous relationship) from the S. cerevisiae whole genome duplication and those that do not. The number of genes in *S. cerevisiae* and *K. lactis* is very similar as well, which should make the orthologous relationships simpler to map (Ozier-kalogeropoulos *et al.*, 1998 and Wong *et al.*, 2001). In addition, research that leads to the understanding of both organisms is of great benefit

as both yeast species are of interest to industrial fermentation (Ozier-Kalogeropooulos *et al.,* 1998). The genetic code is degenerate i.e. two or more codons code for the same amino acid. As a result protein sequences are more conserved and more appropriate for defining orthologs between highly divergent organisms (Wheelan *et al.*, 1999; Mushegian *et al.,* 1998; Chervitz *et al.,* 1998). A PERL script was designed and used to extract the query and best hits with a score better or equal to the above threshold , see **APPENDIX.**

2.2.4 Analysis of the orthologs

Needle from the EMBOSS package, was used to calculate the percentage identity of the orthologous pairs (Wheelan *et al.,* 1999). Needle is a global alignment tool. For the gap opening and gap closing penalties the default values were used.

2.2.5 Estimation of evolutionary rates

Calculation of the Ka (rate of non-synonymous substitution) and Ks (rate of synonymous substitution) values was done using a PERL script; calc_ka_ks.pl from a previous study (Nembaware *et al.*, 2002) Modifications have since been made to make the scripts more efficient and to make use of a likelihood model. A brief overview of the procedure carried out by the calc_Ka_Ks.pl script is in **Figure 2.** See **APPENDIX** for PERL script.

2.2.5.1 Input

Obtaining the CDSs that are required as input into the calc_Ka_Ks.pl was a critical part of the project; particularly for *K. lactis,* which is still undergoing sequencing. The *K. lactis* proteins were BLAST searched against the *K. lactis* DNA sequences. The frame and the coordinates of the *K. lactis* DNA alignment to the *K. lactis* protein were parsed out and these were used to extract the CDS for *K. lactis* proteins using Perl scripts written for this project.

For *S. cerevisiae* the ORF dataset used had the CDS coordinates relative to the chromosome. A PERL script was designed to take the co-ordinates from the S. cerevisiae ORF and splice out the corresponding CDS from the chromosome.

2.2.5.2 Aligning the protein sequences

The first stage in the calc_Ka_Ks.pl script is translation of the CDSs into their corresponding protein sequences. For this step, transeq, an EMBOSS tool is used. The script had to be altered, to reverse complement some DNA sequences before translating them to their corresponding protein sequences. Revseq, and EMBOSS program was used for altering the DNA orientation when necessary. ClustalW was used to align the two pairs of protein sequences obtained from this stage. This protein alignment is used later as a template for the CDS alignments as well as for protein distances, which are parsed out by a sub-routine in the calc_Ka_Ks.pl script.

2.2.5.3 Nucleotide (CDS) sequence alignment

The output from the proteins alignment in the previous stage is needed as template for the CDS alignments. TRANALIGN, an EMBOSS program has been developed to align nucleotide sequences using the protein alignments as a guide. It requires two files on the command line, with the protein sequences in one file and the CDS in another.

2.2.5.4 Codeml Program

The codeml program is part of the PAML package which is freely available from the following website http://abacus.gene.ucl.ac.uk/software/paml.html. Codeml implements a Maximum Likelihood estimate of pairwise non-synonymous and synonymous substitutions. Codeml accounts for transition/transversion rate bias as well as codon usage bias more efficiently than other methods (Goldman and Yang, 1994; Yang and Nielsen, 1998).

Basic Model for Likelihood analysis

The basic model for likelihood analysis is the Goldman and Yang, 1994 codon substitution model. The substitution rate from codon i to j, where i and j are not identical is

0 if the codon pair have one or more differences in their positions
πj for synonymous transversion,
κπj for synonymous transitions
πωj for nonsynonymous transversion
ωκπi for nonsynonymous transtion
key
κπ = transtion/transversion rate ratio
ω= Ka/Ks
Πj =equilibrium frequency of codon j

Output for the calc_Ka_Ks.pl

The output from the codeml program is written to a file. The ouput includes the sequence of both homologous pairs, protein distances, number of synonymous and non-synonymous sites, Ks, Ka and the Ka:Ks ratios. The results are written in a tabular form, which makes post-processing of results efficient.

2.2.6 Web-server for Ks and Ka estimations

Calculation of Ka/Ks ratios can be very tedious especially for those unfamiliar with molecular evolution packages such as EMBOSS, ClustalW and PAML and the required sequence format changes. The majority of Ka:Ks calculation software is only compatable with UNIX based systems, e.g. GenomeHistory (Conant and Wagner, 2002) and this can be a major inconvenience to researchers not familiar with this operating system. It is thus useful to have a web-based and easy-to-use tool that can

calculate Ka/Ks ratios for sets of coding sequences easily. This tool would be especially attractive for researchers without access to UNIX based systems.

Codeml from the PAML package is not user friendly especially to researchers with minimal UNIX skills. Through this project, the easy-to-use calc_Ka_Ks.pl pipeline for Ks and Ka estimations was made available from the SANBI website. In addition to this, for researchers only interested in calculations Ka and Ks for one pair of sequences, a web-server has been created which utilises a CGI version of the calc_Ka_Ks.pl script.

2.3 Results

2.3.1 Database of K. lactis and S. cerevisiae putative Orthologs

121 pairs of *K. lactis S. cerevisiae orthologs* were retained but two cases were detected where two K. lactis proteins had an identical *S. cerevisiae* protein. A list of the generated 119 putative K. lactis and S. cerevisiae orthologs can be viewed at http://hamlyn.sanbi.ac.za/~victoria/database.html.

2.3.1.2 Percent identities of the dataset of 119 *S. cerevisiae* and *K. lactis* orthologous proteins

Global alignments of the S. cerevisiae and K. lactis orthologous amino acid sequences resulted in the percentage identity distribution illustrated in Figure 3. Average identity = 67.84% and the standard deviation = 16.70% with a range from 28.12%-97.07%. The presence of low percentage identities in some of the orthologs is expected, the cut-off similarity is 26% identity.

2.3.2 Estimation of Ka and Ks for *K. lactis* and *S. cerevisiae* orthologs

A total of 107 orthologs were used for further analysis. For the remaining 12 pairs, CDS could not be successful obtained due to frame shifts in the *K. lactics* DNA. The Ks distribution of the 107 putative orthologs (Figure 4) showed less than 40 pairs with Ks values less than 3.7. The rest of the orthologs had Ks well above 3.4 with some

outliers exhibiting Ks well above 70. Only six Ks values were below 0.75. The average Ks value for the whole dataset is above 20, indicating saturation of substitutions at the synonymous sites (see **Chapter 1**). With saturated Ks values, the Ka: Ks ratios can no longer be used as a measure of selective pressure.

The distribution of Ka (**Figure 5**) illustrates active selection against most non-synonymous mutations as expected as the distribution is skewed to the left.

2.3.3 Protein distances and CDS distances

Output from the calc_Ka-Ks.pl script included amino acid distances of the orthologs. These amino acid distances, shown in a scatter plot (**Figure 6**) were calculated using Clustalw. The average protein distance was 0.63 with a standard deviation of 0.63 for the 107 dataset.

Investigating the orthologs with a paralog present and those without

The average amino acid divergence of the whole ortholog set was compared to the average value of amino acid divergence of ortholog pairs for which a paralogous gene was present in *S. cerevisiae.* The Ks values could not be considered as they were saturated. The Ka values were not significantly ifferent for the two sets contrary to the results for the previous study. Although there is a 0.10 difference in the Ka values of the whole orthologs dataset as compared to the smaller dataset, it should be noted that the standard deviations are very large.

Summary of the evolutionary distance calculations

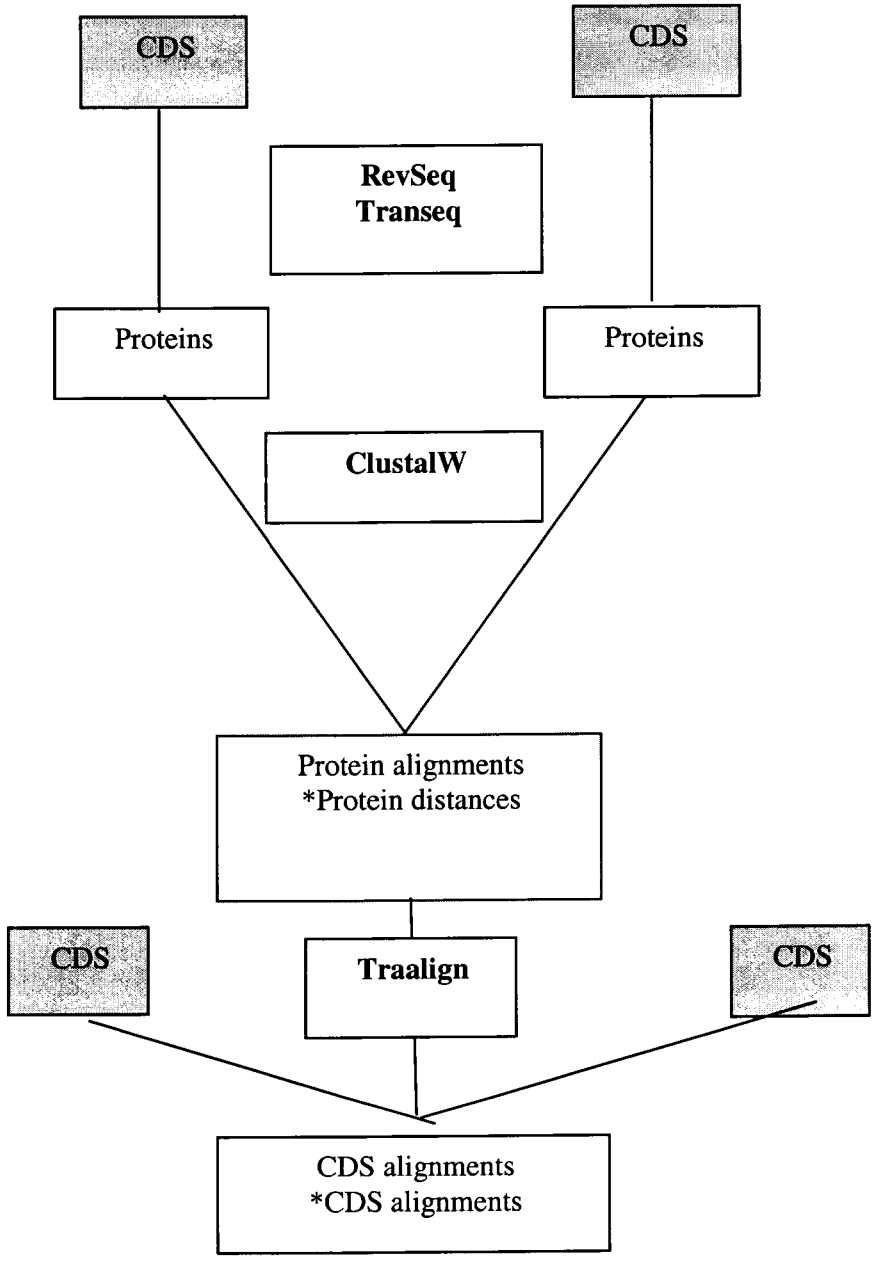

Figure 2 Summary of the intermediate stages implemented in the cals_Ka_Ks.pl script in calculation of Ka:Ks ratios. A stage that has been omitted to minimize clutter in the script is that calculation of nucleic distances using ClustalW. The Output format is described in detail in the text.

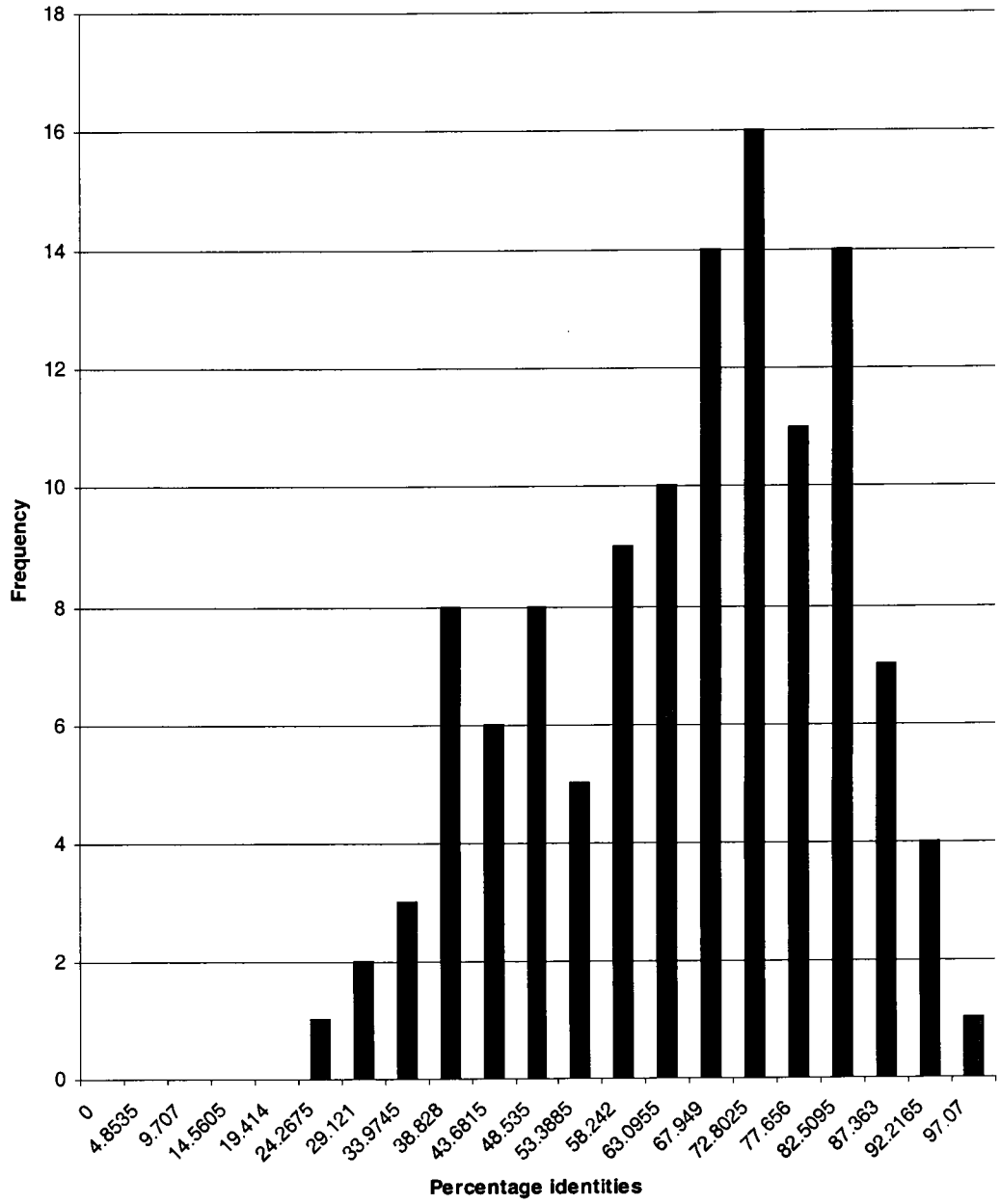

Figure 3 Percentage identity frequency distribution of the 119 *K. lactis* and *S. cerevisiae* orthologs. The alignments were obtained from using a Needle, a global alignment algorithm

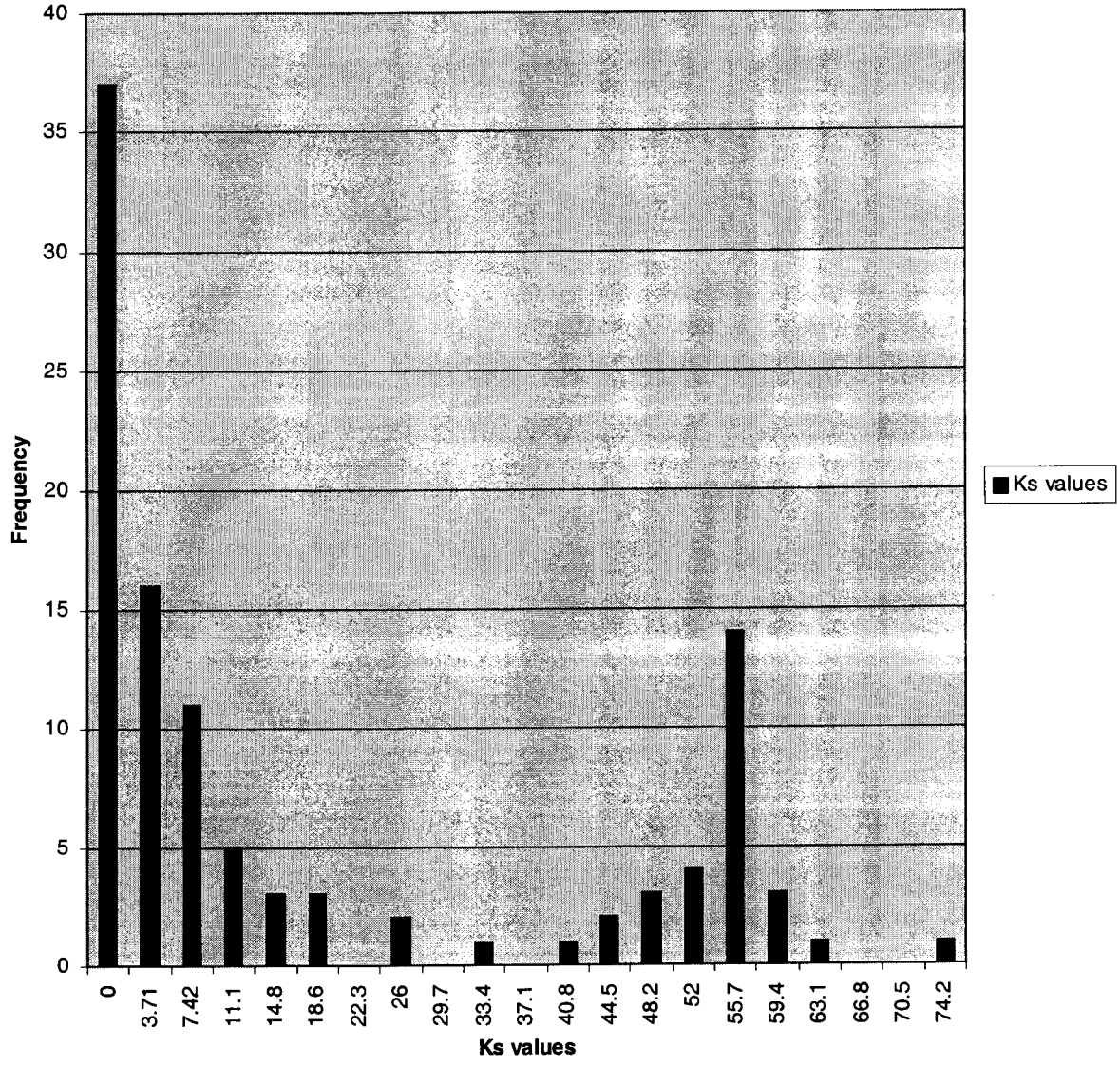

Figure 4 Frequency distribution graph depicting the Ks values of the 107 K. lactics and S. cerevisiae orthologs

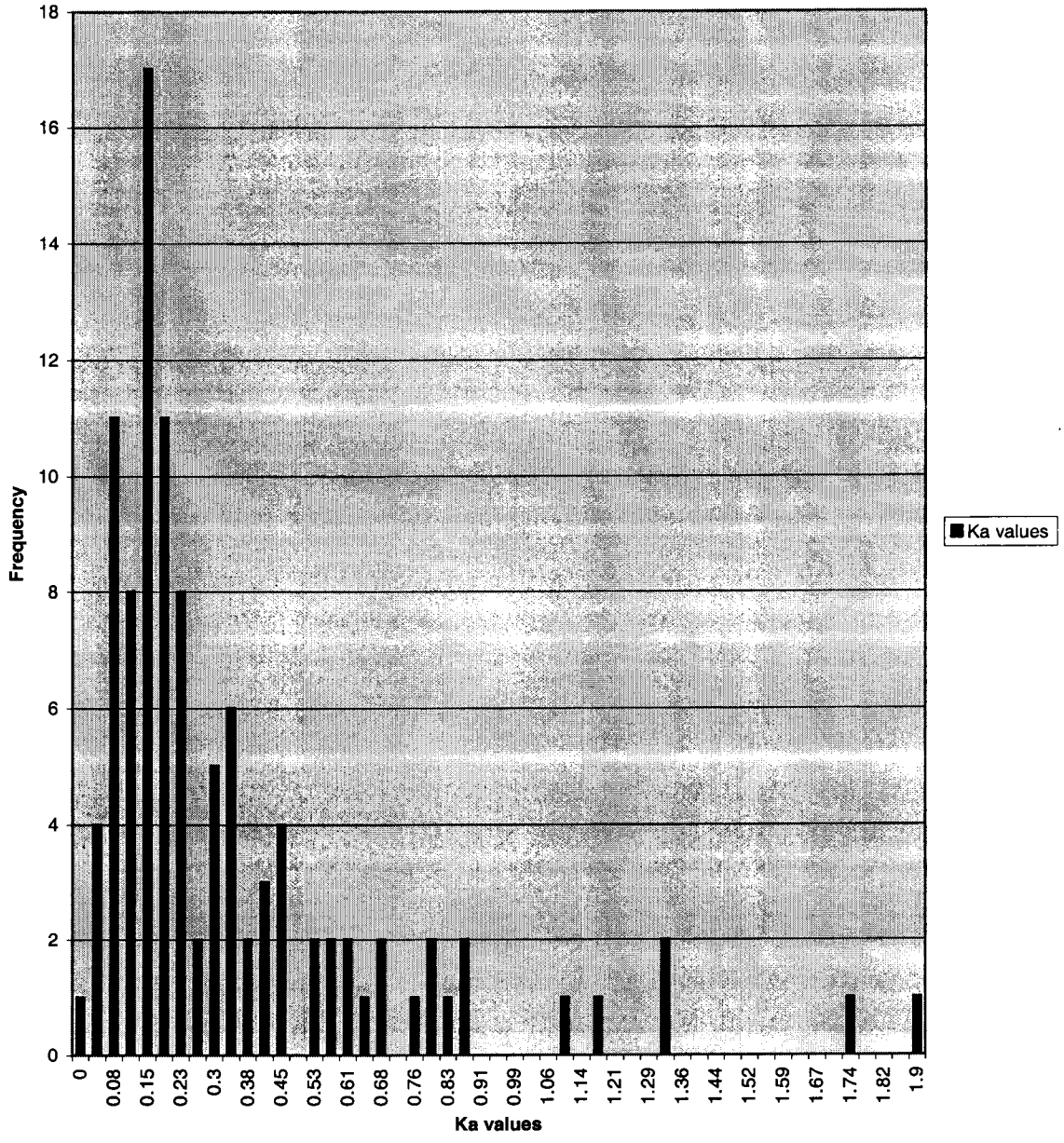

Figure 5 Frequency distribution of Ka values calculated from 107 orthologous pairs.

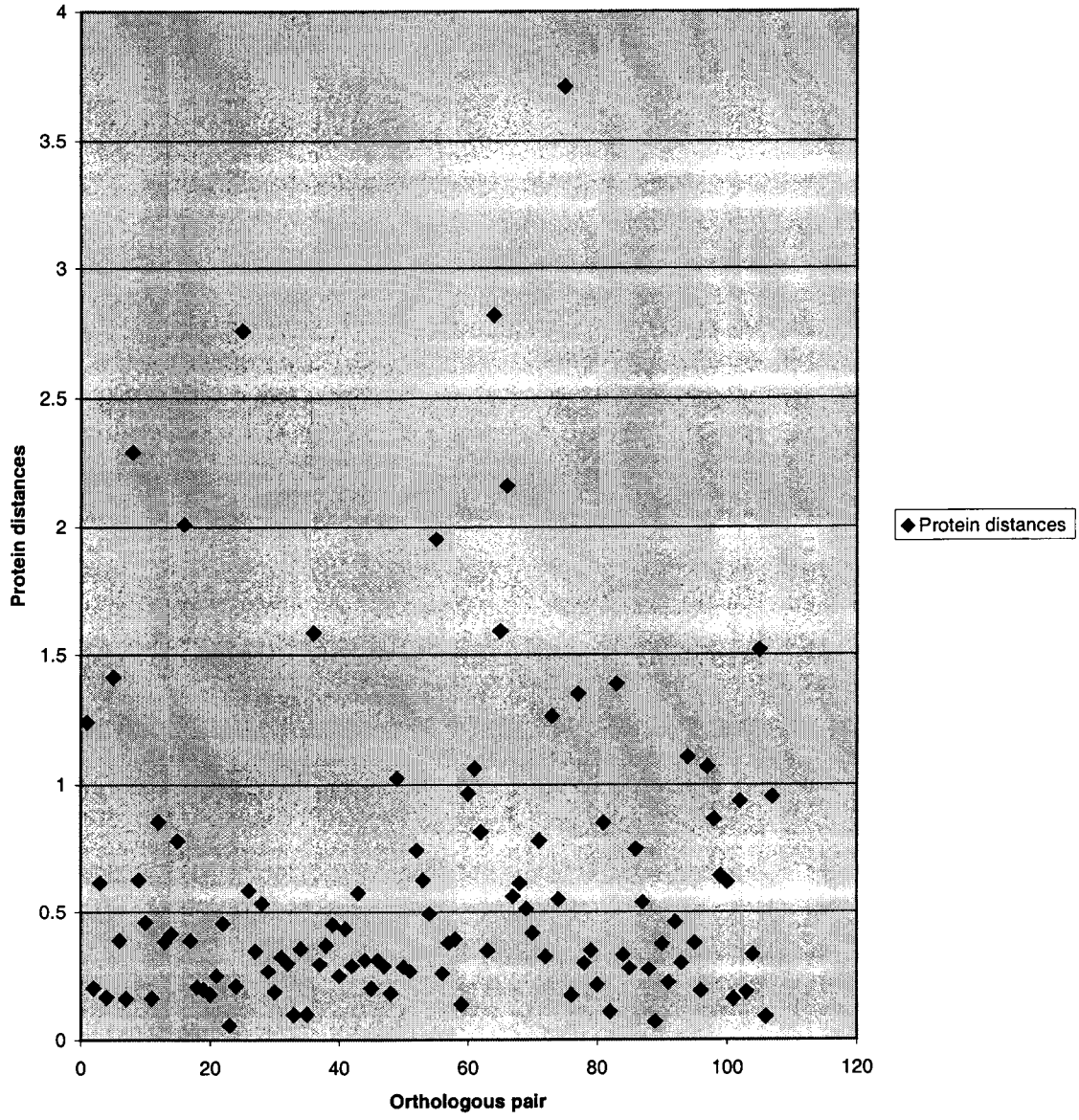

Figure 6 A scatter plot of the protein distances of 107 orthologous pairs.

Table 1: Comparison of the whole orthologous set to the smaller set of orthologs that have paralogs present

Datasets	Whole orthologous dataset (107)	Duplicated orthologs (13)
Ks: ave (sd)	20.76 (23)	17.10 (21)
Ka: ave (sd)	0.36 (0.31)	0.33 (0.25)
Amino acid distances	0.634 (0.63)	0.53 (0.46)

2.4 Discussion and Conclusion

2.4.1 Orthologous dataset

Accuracy in identification of orthologous relationships using similarity measures increases with the compeleteness of the genomes in question (Huynen and Bork, 1998). Although the *K. lactis* genome is far from complete and the protein set used in this study represents a very small percentage of the K. lactis genome, S. cerevisiae (the first eukaryote to be sequenced) is complete and well characterised. The observation that two pairs of *K. lactis* genes share one *S. cerevisiae* protein as their best BLAST hit provides evidence of duplication events that have occurred in K. lactis. This is the only case of possible gene duplications in *K. lactis* since the divergence of *S. cerevisiae* and *K. lactis* that has been reported (Wolfe and Shields, 1997; Wong *et al.*, 2001).

2.4.2 Rate of evolution of K. lactis and S. cerevisiae orthologs

The measures of the Ka: Ks ratios were computed using codeml program from PAML, which makes use of a maximum likelihood approach (Yang and Nielsen, 1997). Despite using a maximum likelihood approach, which has realistic models on mutations and substitution processes (Yang and Nielsen, 1997), saturation in Ks was observed for almost all the sequence pairs under study. Selection is measured in terms of the ratio of Ka: Ks, such an analysis is only applicable to cases in which $Ks < 3$ (Smith and Eyre-Walker, 2001). Previous groups that have experienced the saturation of Ks values have resorted to calculating the radical and conservative non-synonymous changes in the amino acid sequences (Zhang, 2000; Ford, 2001). However, the possibility that factors unrelated to selection, influence the ratio of radical to conservative changes was pointed out by Dagan *et al,* 2002, and as a result we did not compare radical and non-radical amino acid replacements in this project.

Previous work, including our own work on mouse and human orthologs, showed evidence of the affect of duplication on orthologue divergence (Nembaware *et al.*, 2002; Malawoski and Boguski 1998). The speciation of human and mouse has been estimated to have occurred ~100 million years ago, there was little saturation in synonymous substitutions (Nembaware *et al.*, 2002; Malawoski and Boguski, 1998). Based on that observation, it seemed reasonable to expect that the synonymous

substitutions between *K. lactics* and *S. cerevisiae* orthologs that diverged ~150 million years ago might not be saturated.

Factors Contributing to saturation of Ks

Several molecular characteristics have been observed to correlate with synonomous substitution rates. GC content, codon bias and different mutational rates among species and even among genomic regions in the same species contribute to differences in the Ks values (Matassi *et al.,* 1999). All the factors that could contribute to saturation in the Ks values could all be playing a role as the inter-play between the factors is not always clear.

Effect of the presence of a paralog on an ortholog

To understand the mechanisms that govern evolution of duplicated genes, there has been a number of studies that have attempted to correlate various factors (gene family size of paralog, functional group etc) to the rate of evolution of duplicated genes (Conant and Wagner, 2002; Lynch and Conery, 2000). Lynch and Conery, 2000 concluded from their analysis of Ka:Ks values of duplicates that was an increase in the efficiency of purifying selection acting on duplicates with time. However Lynch and Conery's findings are compromised as the authors disregarded the effect the age of a paralog can have on Ka values. Using chimpanzee and human orthologs, our previous study took into account the age of duplicates for each analysis, only duplicates that share a divergence time were used. This method employed by Nembaware *et al.,* 2002, quantified the effect of a retaining a duplicate on the rate of evolution of an ortholog (Nembaware *et al.,* 2002). The presence of an ancient paralog was shown to increase the rate of non-synonymous substitution in ortholog-pairs. Although the aim of this study was to employ the same, approach that we have used previously (Nembaware *et al.,* 2002), the research was hindered by saturation of Ks values. Comparison of protein distances was of little use as the standard deviations were too large.

2.4.4 Web-server for Ks and Ka calculations

A web-server that calculates Ka:Ks ratios was set up. A similar website has been set-up SNAP which is at (http://hivweb.lanl.gov/content/hivdb/SNAP/WEBSNAP/SNAP.html) (Ota and Nei, 1994) but this website implements an approximate method and requires an alignment as input. The approximate method implemented by SNAP, does not model rate variation across sites effectively as researchers have since shown their preference to maximum likelihood methods such as the one implemented on our website. The script, calc_Ka_Ks.pl was modified into a CGI perl program. The parameter settings in the codeml.ctl file are in the READ_ME file available from the website (http://hamlyn.sanbi.ac.za/~victoria). A screen shot is illustrated in **Figure 7** below. For datasets that are too big to down load, there is also an option of downloading the original calc_Ka_Ks.pl perl script. A sample of the results from the web-server is shown in **Figure 8.** Future improvements to the web-server will include an e-mail of the in-frame alignments used as input for codeml to the user. Its essential for the user to have accessto the alignments used as a check on the quality of the calculation.

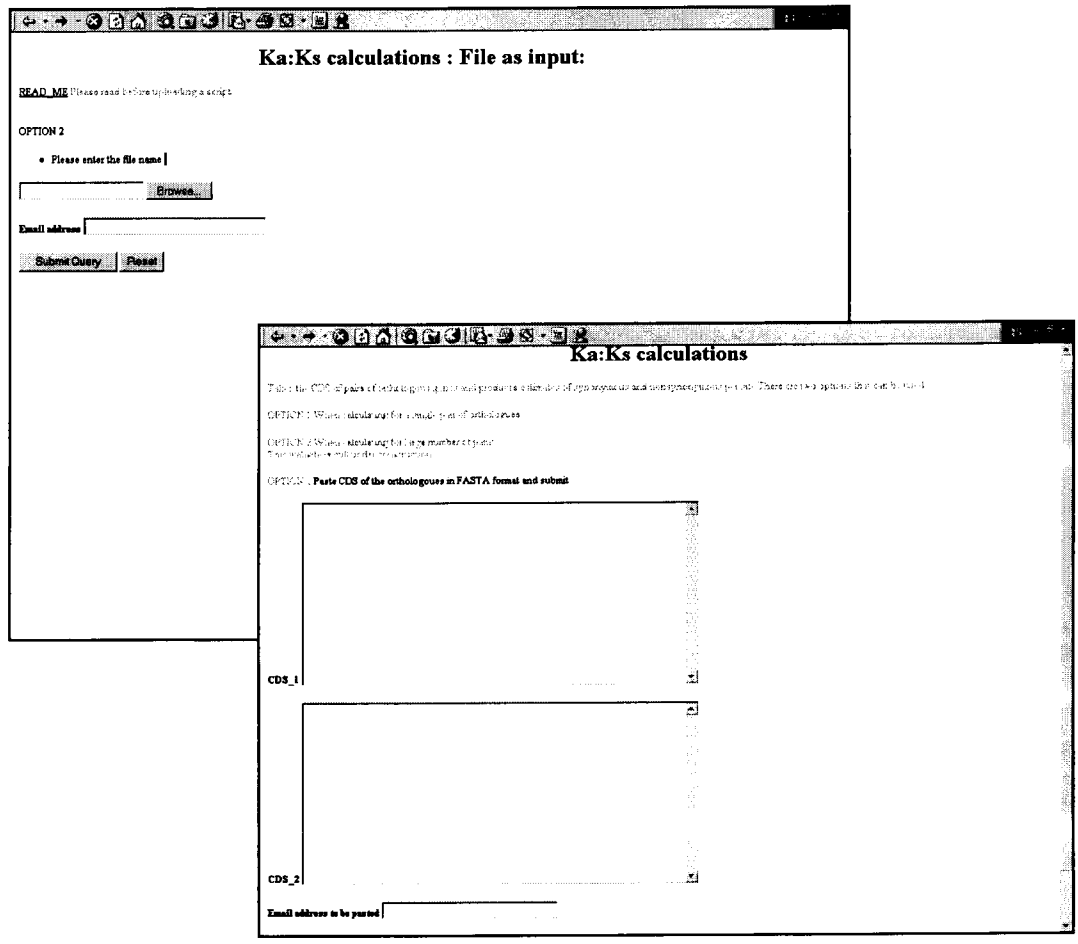

Figure 8 A screen shot of the web-server for calculation of Ka:Ks using Codeml. There are three options available of which two are shown in the figure. Option 1 is to paste two pairs of sequences for the Ka:Ks calculations. Option three is uploading a file of sequences onto the server .The last option is a free download of the script used.

Date: Fri, 15 Nov 2002 12:22:35 +0200

From: Victoria Nembaware <victoria@sanbi.ac.za>

To: jaechild@hotmail.com

Subject : Results_Ka_Ks

Sequence names		nucleic acid distances	protein distances	S	N	Ks	Ka	Ks/Ka
SEQ_1	SEQ_2	0.307	0.200	377.1	1248.9	2.9896	0.1194	0.0399

ERROR LOG FILE: web_infile

```
This part contains a list of error messages for why divergence
calculations could not be carried out for certain pairs of
orthologous sequences which have been entered into the program
================================================================
```

Figure 9: An example of an email with results from the web-server calculating Ka:Ks ratios. If for some reason the Ka:Ks could not be calculated, the heading error_log_web_infile could have a paragraph briefly describing why the calculation could not be carried out by the perl script calc_Ks_Ka.pl.

Chapter 3

Exploration of graphs obtained from BLAST results

3.1 Introduction

The primary step after sequencing of a gene is identification of homologs, largely through BLAST searches. The inadequacies of the BLAST tool have been highlighted in its failure to detect close homologs of proteins such the guanlyly cyclase (discussed in Introduction). Such limitations of the sequence similarity programs in identifying homologues, has led to clustering methodology gaining in popularity (Perriere *et al.,* 2000; Yona *et al.,* 2000; Sasson *et al.,* 1998). By clustering sequences into groups, based on the E-score values associated with them, one can discover relationships that direct sequence comparisons fail to uncover (Yona *et al.,* 2000).

Motivation for this investigation stems from the observation that sequence similarity searches done on complete genomes increase the chances of finding accurate homologous relationships (Bork and Huynen, 1999). In addition, this could also provide information about proteins missing from the genomes. This investigation is largely an exploratory study, and is centred on analysing graphs constructed from a BLAST output from a set of proteins from 15 completely sequenced genomes queried against itself. This investigation does not attempt to produce a clustering tool for effective clustering therefore makes no use of complex algorithms. We implement a very basic PERL script to create graphical representation of the BLAST output. Each protein is taken as a node and pairs of nodes are connected if they have a BLAST match with an E-value less than or equal to a cut-off value.

From this study we have emphasized the number of connected clusters that visit each genome at a particular cut-off as well as complete clusters. These results were aimed at estimating the homologous networks that exist among the genomes.

Unlike the COGS and Clustr and HOBACGEEN databases (Perriere *et al.,* 2000), that aim to define groups of related proteins, the - study is intended at providing the base for potential future work on defining and analysis of "orphan genes".

3.2. Data acquisition and methods

3.2.1 Complete Genomes

Completely sequenced genomes have increased drastically as the sequencing technologies are constantly improving. Presently, there is a total of 118 published completely sequenced genomes are more than 300 still undergoing sequencing. Most of these complete genomic sequences have been made available to the public through various genome initiatives as listed at the GOLD website (http://wit.integratedgenomics.com/GOLD/completegenomes.html). For the current study, 15 genomes that represent the almost all the major phyla of organisms were selected. **Table 1** gives an outline of the organisms used as well as the corresponding websites from which they can be downloaded.

Table 1: Genome used for the BLAST analysis

Organism	Website
Aquifex aeolicus	http://www.bio.nite.go.jp/
Arabidopsis thaliana	http://www.tigr.org/
Borrelia burgdorferi	http://www.tigr.org/
Caenorhabditis elegans	http://www.sanger.ac.uk/
Drosophila melanogaster	http://flybase.harvard.edu:7081/
Haemophilus influenzae	http://www.tigr.org/
Helicobacter pylori	http://www.tigr.org/
Homo sapiens	http://www.ensembl.org
Listeria monocytogenes	http://www.ncbi.nlm.nih.gov
Mycobacterium tuberculosis	http://www.tigr.org/
Oryza sativa	http://www.ncbi.nlm.hih.gov
Saccharomyces cerevisiae	http://www.sanger.ac.uk/
Streptococcus pneumoniae	http://www.tigr.org/
Synechocystis PCC6803	http://www.ncbi.nlm.nih.gov
Vibrio cholerae	http://www.tigr.org/

3.2.2 Pre-clustering phase

Open reading frames (ORFs) of 15 complete genomes were downloaded the web sites shown in **Table 1** above. ORFS of the downloaded complete genomes were concatenated and stored in one file **(Figure 9).** A comprehensive all-against-all sequence comparison, using BLASTp based on the matrix BLOSUM62 was performed. The option of filtering low complexity sequences was switched on for the BLASTp search to reduce noise. Default values of BLAST parameters were used. To reduce the time required for the clustering process, blast results were parsed, for all the query-hit pairs. The accessions for each query and hit were transformed so that

there would be uniformity for each specific organism. Standardisation of each record was essential for the clustering stage.

3.2.3 Exploration of BLAST results

The E-scores values from the BLAST output were used as edges joining the nodes (proteins). Only queries and hits that have E-scores better or equal to the stipulated threshold were used in the clustering procedure. This means that short proteins with E-values larger than the cut-off will be omitted from the analysis instead of contributing to the number of orphans. With the algorithm described below, a graph was constructed and analysed at the specified E-scores cut-offs.

Algorithm for Graph theory implementation:

```
Algorithm
        if new Query
                current Query= query;
        unless Query is contained in a cluster
                make new cluster with 1 element
                        = current_query;
Case 1: Hit is not already contained in a cluster {

                add hit to cc of current_query

                }
Case 2: Hit is contained in cluster(Z ){

                merge cluster(Z) && cluster of current_query

                }
```

The Perl script used in this parsing is in **APPENDIX**

Clusters

As illustrated in **Figure 10,** a cluster is defined if it contains at least one protein. Clusters containing only one protein are termed orphans. Orphans are a result of self-hits at a specified cut-off. A perl script was designed to implement the algorithm shown above. Clusters were created at various cut-offs E-score cut-off.

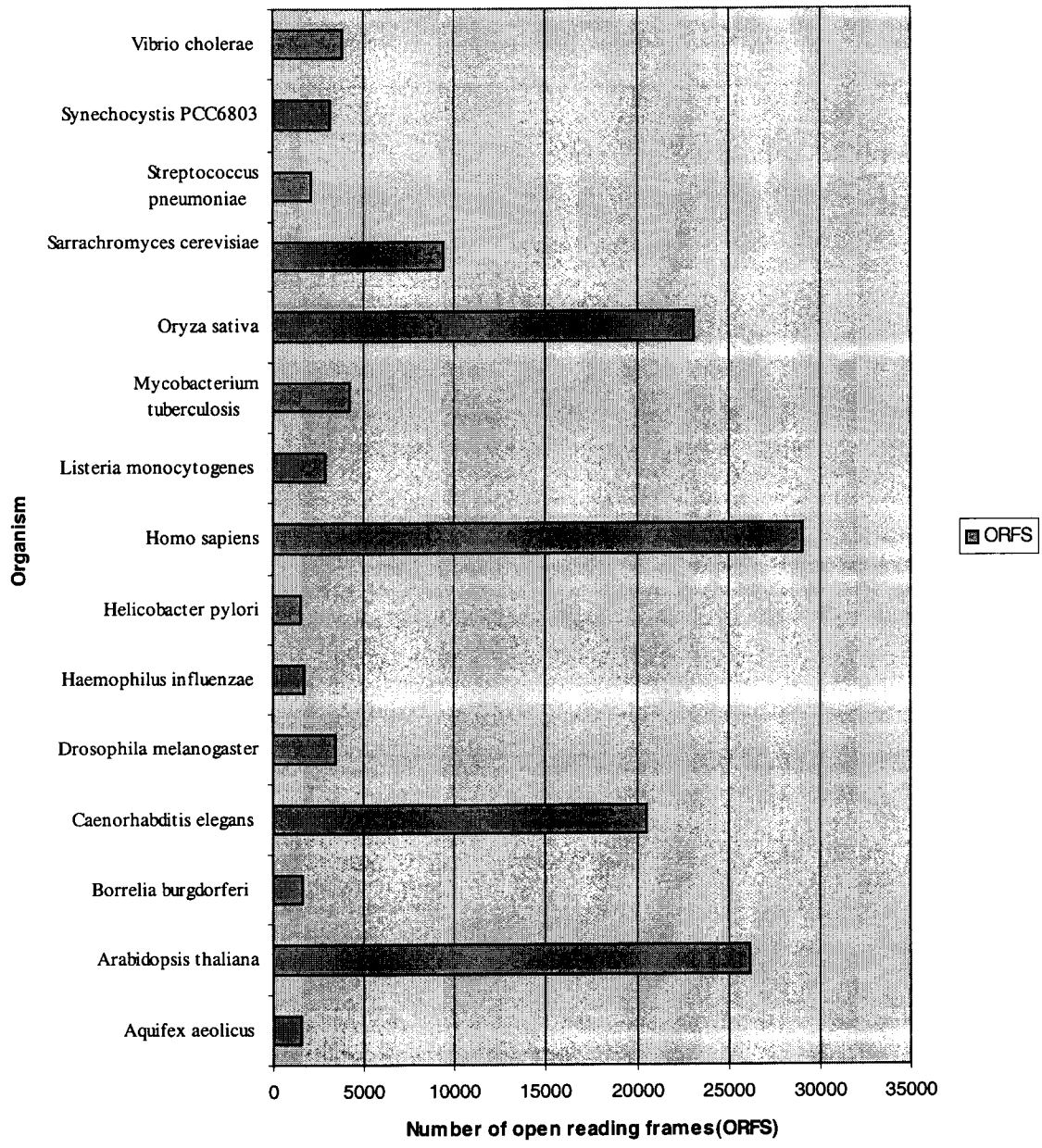

Figure 9: Number of ORFs available per genome

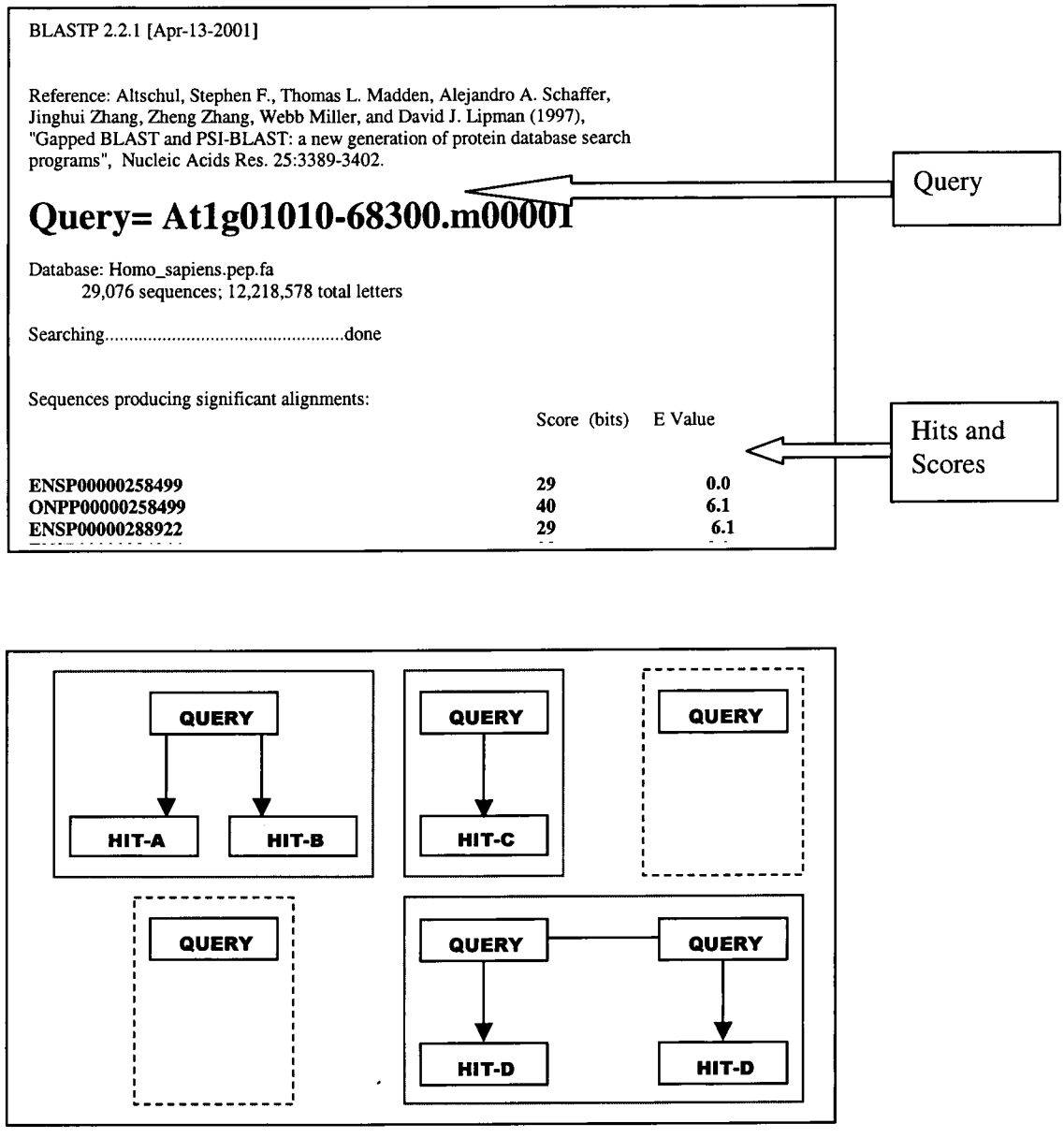

Figure 10: An illustration of the clustering procedure for a graphical representation of BLAST results. 3a shows a sample outline of BLAST. In 3b, Each box enclosing either a query or a query and its hit represents a cluster. The dotted lines represent orphans while the solid lines represent connected compounds.

3.2.3 Blast search of Arabidopsis genome against the human genome

The human genome is a point of reference for most of the studies that are carried out in various other genomes. A significant number of genes in human that are disease causing have orthologs in Arabidopsis (Arabidopsis Genome Initiative, 2000). The importance of establishing homologous relationships between the distantly related species therefore cannot be ruled out. BLAST can give a general indication of the homology that exists between two organisms when the distribution of the scores is analyses.

3.3 Results

3.3.1 Arabidopsis versus Human genome

The Arabidopsis genome has quite a significant number of hits to the human genome (results not shown). Since the two genomes are complete this is expected. The guanylyl cyclase gene has its best non-self-hit at a cut-off of 1e-23 which is a reasonable score however its second best hit is a predicated human gene hence is of minimal value in characterisation of this putative gene.

3.3.2 Cluster analyses

3.3.2.1 The guanylyl cyclase gene

In the unclustered BLAST output the Arabidopsis guanylyl cyclase has its closest hit as a human entry. The human entry is a predicted proteins hence this would not be very helpful for characterisation of the proposed guanylyl cyclase gene. The guanylyl cyclase gene has its best non-self hit at an E-score of 5e-23 while the remaining hits have highly insignificant hits (0.70). However when the E-score cut-off is relaxed, the guanylyl cyclase clustered with various other proteins, which are bound to be some of its closest homologs.

3.3.2.3 Distribution of Clusters

At stringent E-value cut-off very few proteins clustered, as expected. At stringent E-values such as 0 and 1e-180, most of the proteins only have themselves as hits. The number of clusters increases gradually as shown in **Figure 11** as queries have more non-self hits at such E-score cut-offs. As the E-score stringency is relaxed at about 1e-20, the number of clusters decreases drastically as most of the clusters begin to merge forming bigger clusters.

The orphan gene clusters are lower than expected as some of the proteins hit their homologs as their best hit instead of themselves due to shortcomings of the BLAST tool as well as annotation inaccuracies. When a query has its best E-scores and P-scores being shared by two differently named hits, the BLAST tool outputs the results in alphabetical order, this seems to have affected the orphan's distribution to some extent. The graph tails off as stringency becomes more relaxed mainly because at these values many proteins exist in connected clusters thus drastically decreasing the overall number of clusters as well as the number of orphans.

Complete clusters

A complete cluster, as discussed here, refers to clusters that contain at least one protein from each of the 15 organisms. The complete clusters are distributed as shown in **Figure 12.** At the tail of the curve a significant portion of the BLAST output is clustered into one huge connected cluster with very few proteins that are self hits still existing as unconnected clusters (orphans).

For each genome, a cluster was counted only if it contained a protein sequence from that specific genome. **Figure 13** shows the number of such clusters per genome. Arabidopsis has the most number of clusters associated with it at stringent E-score values. This is unexpected as the human genome has more proteins than any of the other genomes, this is observation requires further study. The proportion of short proteins in all the genomes could be investigated as well as the distribution of domains. Both these factors could be influencing the distribution of the number of clusters associated to each genome.

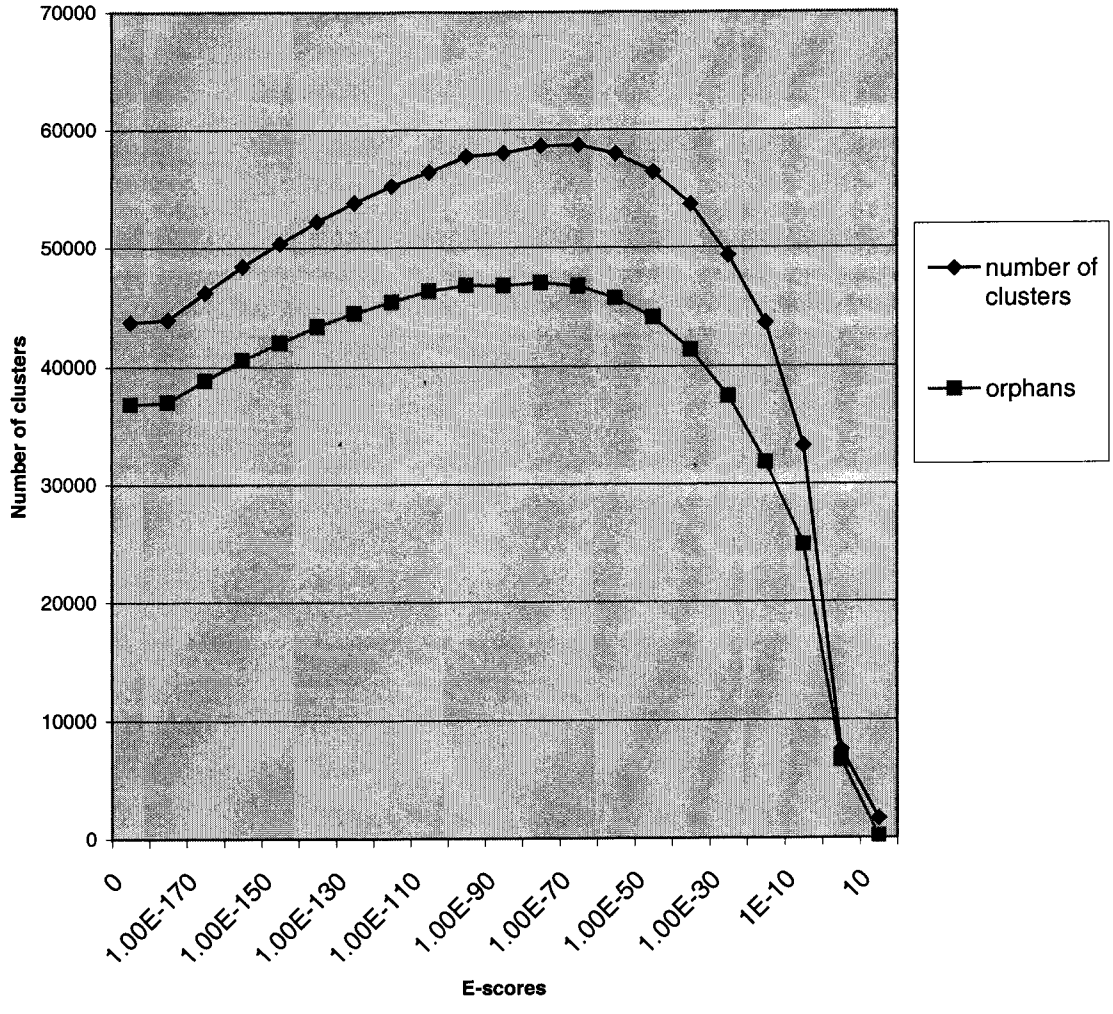

Figure 11: The number of clusters and the orphan genes plotted against E-value cut-offs.

7

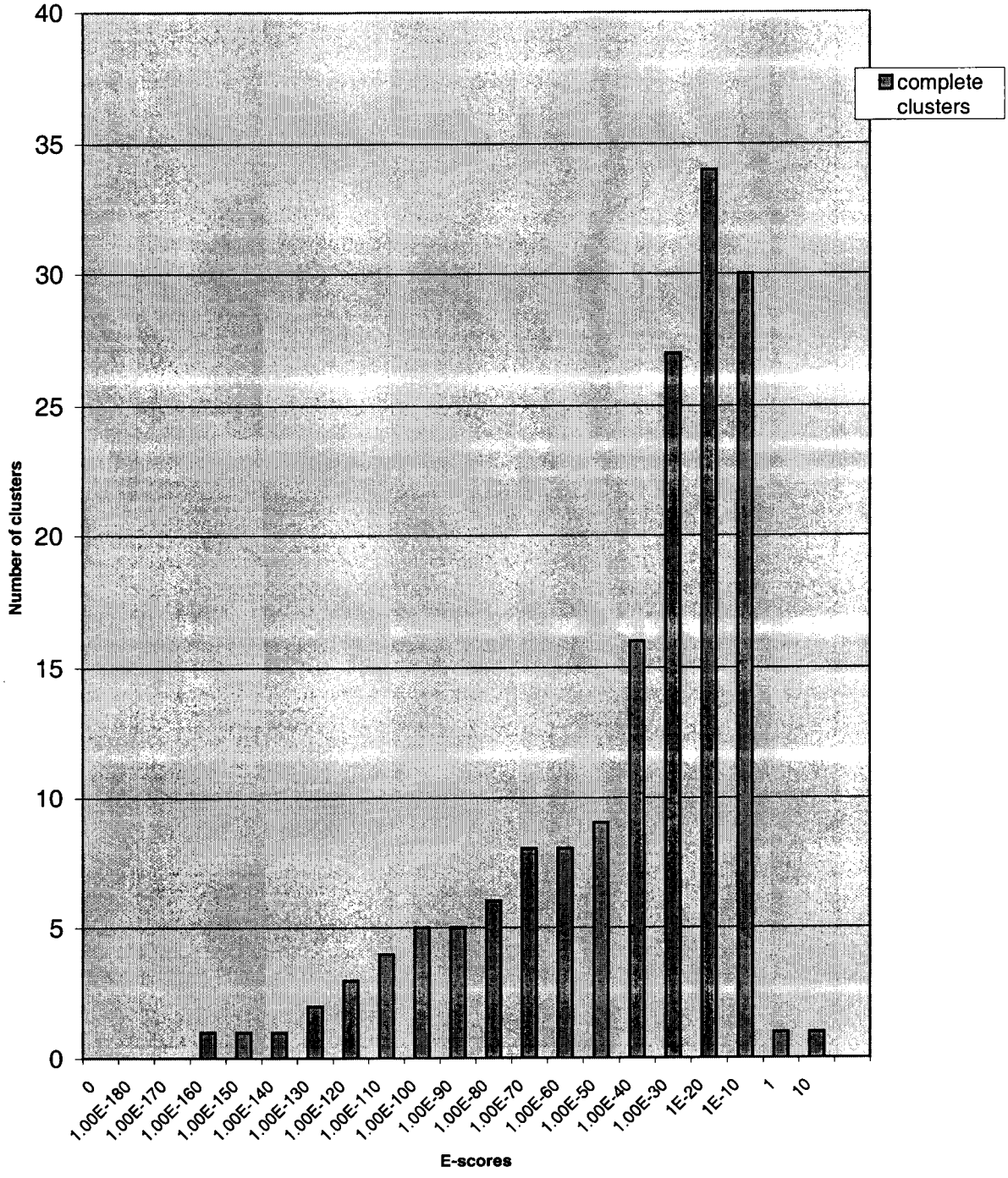

Figure 12 The number of complete clusters at various E-score cut-offs.

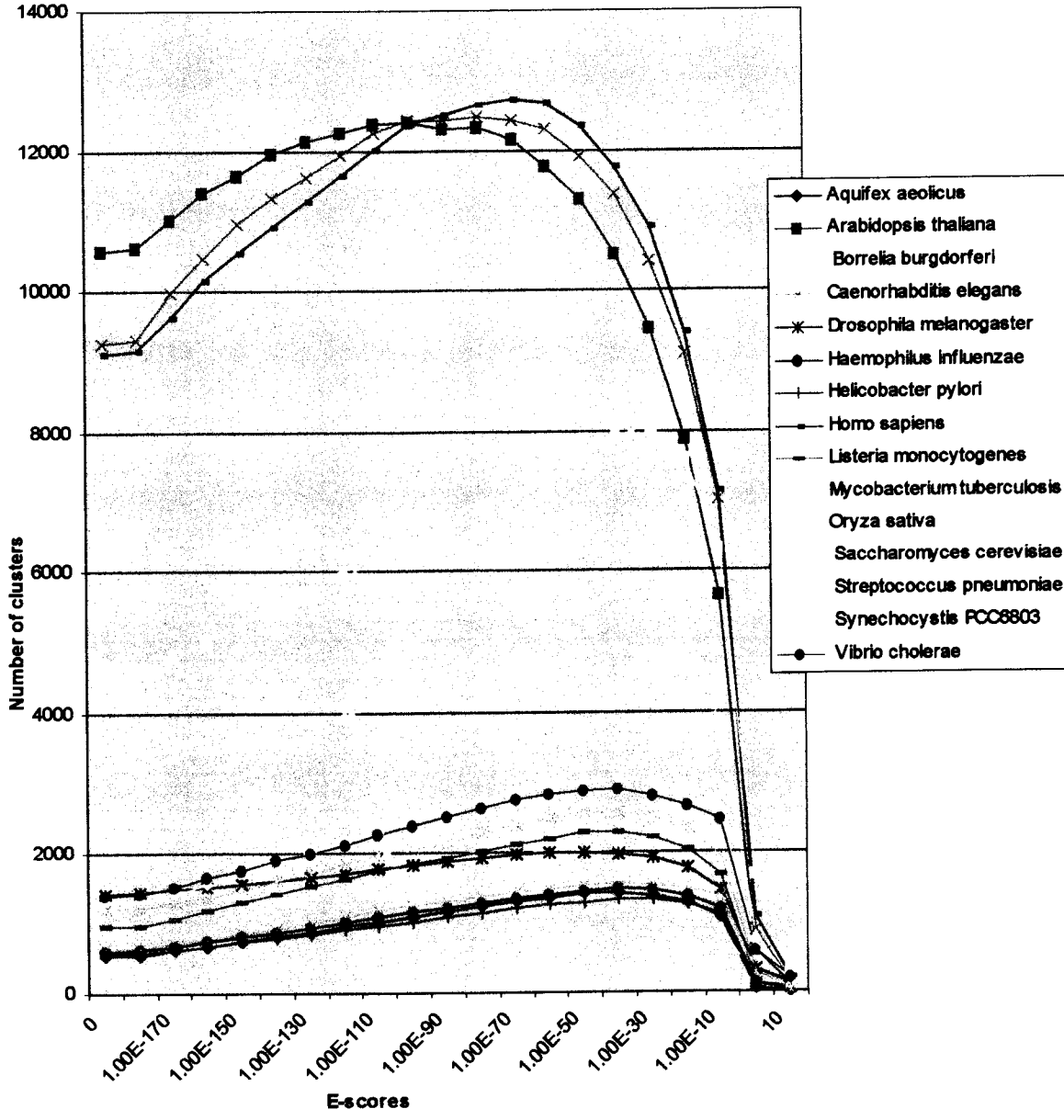

Figure 13: For each genome the number of clusters that contain a protein from the genome plotted against the E-value cut-off

3.4 Discussions and Conclusion

3.4.1 Database of complete genomes

The informativeness of a homology search increases when complete genomes are used (Chervitz *et al.,* 1998; Huynen and Bork, 1998). Complete genomes provide a higher probability in the identification of homology. Chervitz *et al.,* 1998, reported an informative set of orthologs after reciprocal BLAST searches using complete eukaryotic genomes. To increase the validity of the COGs database only complete organisms were used in the analysis (Tatusov, 2001; Perrire *et al.*, 200). Therefore in the present study, 13 complete genomes that represent as many organisms as possible from all the three kingdoms of life were used for the all-against –all BLAST search.

3.4.2 Exploration of BLAST results

Proteins that consist of multiple structural and functional domains are more common in eukaryotes than any of the other simpler life forms. The eukaryotes have the largest number of clusters in which their proteins are found at all the E-value cut-offs. Sasson *et al.*, 1998 to validate their clustering algorithm, used InterPro (database of protein domains) to check the percentage of proteins in a cluster that shared a common domain.

3.4.3 Orphans

Orphan genes are resources in providing insights in the uniqueness of species (Salama *et al*, 2000). Several orphans have been implicated in virulence of pathogens. BLAST searches of complete genomes can give a preliminary indication of orphan gene candidates. For genes such as guanylyl cyclase, sequence similarity searching was incapable of detecting the homologs at low E-value cut-offs (Ludidi and Gehring, 2001). Less stringent cut-off in clustering analysis has been shown to detect homologs in other clustering procedures (Portugaly and Linaial, 2000). In this study the guanylyl cyclase was examined and confirmed this clustering evidence. When a less

stringent cut-off was used guanylyl cyclase clustered off with a significant number of its own homologs unlike at low cut-offs where it clustered with only one human homolog.

3.4.4 Advantages of complete genomes for BLAST searches

Inadequacies of sequence algorithms paved way to the classification of homologous proteins sequences through clustering Perriere *et al.,* 2000. The clustering described allows genes to be exposed to various other evolutionary associations. Clustering may aid in revealing concealed homologous relationships among protein sequences that pair-wise BLAST is incapable of detecting. With the use of complete genomes, the advantages of clustering observed in the past, using incomplete genomes, are likely to be enhanced.

3.4.4 Future Work on Orphans

Previous groups have not attempted to answer the question raised by genes such as the guanylyl cyclase gene. It is in a class of genes that fails to detect homologs in completely sequenced genomes. The clustering algorithm implemented in this project does not involve normalisation of the database hence this aspect of the program can be improved during future work. There is a need to provide a web-based tool that navigates a BLAST graph space and provides homologs for such genes, to aid researchers speed up sequence characterisations. The design concept of the web-page interface is illustrated in **Figure 14.**

Results for option 1 of the **Figure 14** would include the number of orphan clusters at specific cut-offs. And as for option 2, the results would include the number of proteins in the cluster as well the proteins sequences that it clusters with. Such a procedure can be the first point of reference for characterisation of the genes where BLAST has failed. The user has the option of defining the e-score cut-off they would like to use for the clustering process

OPTION 1:

1. E-score cut-off

2. Complete genomes to be used in clustering (use Key)

OPTION 2:

3. Paste sequence

KEY

Organism	**Number**
Aquifex aeolicus	*1*
Arabidopsis thaliana	*2*
Borrelia burgdorferi	*3*
Caenorhabditis elegans	*4*
Drosophila melanogaster	*5*
Haemophilus influenzae	*6*
Helicobacter pylori	*7*
Homo sapiens	*8*
Listeria monocytogenes	*9*
Mycobacterium tuberculosis	*10*
Oryza sativa	*11*
Saccharomyces cerevisiae	*12*
Streptococcus pneumoniae	*13*
Synechocystis PCC6803	14
Vibrio cholerae	15

Figure 14: For option 1, the user can use the web-page to access the distribution of orphans from a BLAST file of selected complete genomes. Option 2 would allow the user to follow clustering of a sequence at various E-value cut-offs.

Chapter 4

Concluding Comments

Questions concerning sequence divergence have been primarily approached using single sequence comparative studies heralded by the cytochrome C studies in the 1960s (Li, 1997). The results of such studies may give a reflection on the evolution of single genes but have very limited information on the evolutionary forces acting on the genome as a whole. The increasing number of complete genomes aids in understanding the rate of sequence divergences, as well as uncovering the factors governing the rate of divergence (Lynch and Conery, 2000).

Comparative genomics has made it apparent that such small-scale sequence divergence studies may mask other greater forces controlling homologous sequences (Lynch and Conery, 2000). Homology describes a relationship between genes and is based on quantitative similarity measures (Theben, 2001). Homology implies that the sequences under comparison diverged in evolution from a common origin. The accumulation of base substitutions is a continual process and is crucial instrumental in shaping the molecular sequence. Recently a significant amount of research has been aimed at elucidating the rate at which sequences diverge and the factors that influence their divergence (Conant and Wagner, 2002; Spring, 2002; Chervitz *et al.*, 1998).

The unifying concept in this thesis is that of analysing sequence divergence making reference to complete sets of protein coding sequences. In chapter 2 the rate of divergence of orthologs was studied in a way analogous to our previous study (Nembaware *et al.*, 2002). Although the results were not conclusive the study resulted in an application that can facilitate future sequence divergence studies as the Ka:Ks calculations have been greatly simplified through the introduction of a web-server.

In Chapter 3, a widely used similarity measure tool, BLAST is analysed. Genes such as the guanylyl cyclase inspired this study. Genes in this class diverge to such an extent that they are no longer detectable using simple sequence similarity measures. Using the clustering method described, detection of distant homologs can be enhanced without the need for profile searching. This expoloratory study has become preliminary work in the creation of an orphan webserver. The operation of the "orphan server" described in the discussion section of Chapter 3 will eventually utilise an increased number of completely sequenced genomes to enhance the number homologs that can be clustered hence reducing the number of orphans. Yet another

functionality that can be added is that of association of the clusters to gene ontology (http://www.geneontology.org) (Kyrpides, 1999).

References:

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller WW, Lipman DJ. **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Research* 1996. **25:**3389-3402.

Askwith C, Kaplan J. **Iron and copper transport in yeast and its relevance to human disease.** *Trends Biochem Sci* 1998. **23(4):**135-8.

Blanc G, Barakat A. **Extensive Duplication and Reshuffling in the Arabidopsis Genome.** *The Plant cell* 2000. **12:**1093-1101.

Brenner SE, Chothia C, Hubbard TJP. **Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships.** *Proc. Natl. acad. Sci. USA* 1998. **95**: 6073- 6078.

Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T, Weng S, Cherry JM, Botstein D. **Comparison of the complete protein sets of worm and yeast: orthology and divergence.** *Science* 1998. **282:**2022-2028

Clark JB, Kidwell MG. **A phylogenetic perspective on P transposable element evolution in Drosophila.** *Proc Natl Acad Sci USA* 1997. **94**: 11428-11433.

Conant GC, Wagner A. **GenomeHistory: A software tool and its application to fully sequenced genomes.** *Nucleic Acids Research* 2002. **30(15):**3378-86.

Cooke J, Nowak MA, Boerlijst M, Maynard Smith J. **Evolutionary origins and maintenace of redundant gene expression during metazoan development.** *C TIG* 1997. **13(9)**:360-364

Copley RR, Letunic I, Bork P. **Genome and protein evolution in eukaryotes.** *Curr Opin Chem Biol* 2002. **6(1):**39-45.

Duret L, Mouchiroud D. **Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis.** *Proc Natl Acad Sci USA* 1999. **96**:4482-7

Eichler EE. **Segmental duplications: what's missing, misassigned, and misassembled--and should we care?** *Genome Res* 2001. **11**:653-6.

Fay JC, Wu CI. **The neutral theory in the genomic era.** *Curr Opin Genet Dev* 2001. **11(6)**:642-6.

Fitch WM. **Homology a personal view on some of the problems.** *Trends in Genetics* 2000. **16**:227-231.

Friedman R, Hughes AL.**Gene duplication and the structure of eukaryotic genomes.** *Genome Res* 2001. **11(3)**:373-81.

Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlethwait C. **Preservation of Duplicate Genes by Complementary, Degenerative Mutations.** *Genetics* 1999. **151**:1531-1545.

Goldman N, Yang Z. **A codon-based model of nucleotide substitution for protein-coding DNA sequences.** *Mol Biol Evol* 1994. **3**:230-239.

Gu Xun, Yufeng wang, Jianying Gu. **Age distribution of human gene families shows significant roles of both large- and small scale duplications in vertebrate evolution.** *Nature Genetics* 2002. **31**:206-209.

Haney P.J, Badger JH, Buldak GL, Reich CI, Woese CR, Olsen GJ. **Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species.** *Proc Natl Acad Sci USA* 1999. **96**:3578-3583.

Hughes AL, da Silva J, Friedman R. **Ancient genome duplications did not structure the human Hox-bearing chromosomes.** *Genome Res* 2001. **11(5):**771-80.

Hughes AL. **Adaptive evolution of Genes and Genomes.** Oxford 1999.

Huynen MA, Bork P. **Measuring genome evolution.** *Proc Natl Acad Sci U S A* 1998. **2695(11):**5849-56.

Jensen RA. **Orthologs and paralogs- we need to get it right.** *Genome Biology* 2001. **2(8)**:1002.1-1002.3.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. **Selection in the evolution of gene duplications.***Genome Biology* 2002. **3:**research0008.1-0008.9

Kyrpides N. **Genomes OnLine Database (GOLD): a monitor of complete and ongoing genome projects world wide.** *Bioinformatics* 1999. **15:**773-774.

Li WH, Gu Z, Wang H, Nekrutenko A. **Evolutionary analyses of the human genome.** *Nature* 2001. **409**: 847-849.

Li WH. **Molecular Evolution.** Sinauer Associate, Sunderland, MA, 1997.

Ludidi N, Gehring C. **Identification of a novel protein with guanylyl cyclase activity in Arabidopsis thaliana.** *J Biol Chem* 2002 (submitted).

Lynch M, Connery JS. **The evolutionary fate and consequences of duplicate genes.** *Science* 2000. 290:1151-1155.

Lynch M, Force A. **The probability of duplicate-gene preservation by subfunctionalization** *Genetics* 2000. **154:** 459-473.

Makalowski W, Boguski MS. **Evolutionary parameters of the transcribed mammalian genome: An analysis of 2,820 orthologous rodent and human sequences.** *Proc Natl Acad Sci USA* 1998. **95**:9407-9412.

Martin AP. **Increasing genomic complexity by gene duplication and the origin of vertebrates.** *American Naturalist* 1999. **154:**111-128.

Matassi G, Sharp PM, Gautier C. **Chromosomal location effects on gene sequence evolution in mammals** 1999. *Current Biology* **9:**786-791.

McLysaught AK, Hokamp KH, Wolfe K. **Extensive genomic duplication during early chordate evolution.** *Nature Genetics* 2002. **31**:200-204.

Moran JV. **Human L1 retrotransposition: insights and peculiarities learned from a cultured cell retrotransposition assay**. *Genetica* 1999. **107(1-3):**39-51.

Mushegian AR, Garey JR, Martin J, Liu L. **Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes.** *Genome Research* 1998. **8:**590-598.

Needleman SB, & Wunsch CD. **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *J Mol Biol 1970* **48**:443-453.

Nekrutenko A, Kateryna KD, Wen-Hsiung Li. **The Ka/Ks Test for Assessing the Protein-Coding Potential of Genomic Regions: An Empirical and Simulation Study.** *Genome research* 2001. **12:**198-202.

Nembaware V, Crum K, Kelso J, Seoighe C. **Impact of the Presence of Paralogs on Sequence Divergence in a Set of Mouse-Human Orthologs** *Genome Research* 2002. **12:** 1370-1376.

Nielsen R, Yang Z. **Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.** *Genetics* **148:**929-936.

Ota T, Nei M. **Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family.** *Mol Bio Evol* 1994. **11:**469-482.

Ozier-Kalogeropoulos O, Malpertuy A, Boyer J, Tekaia F, Dujon B. **Random exploration of the Kluyveromyces lactis genome and comparison with that of Saccharomyces cerevisiae.** *Nucleic Acids research* 1998. **26(23):**5511-5524.

Perriere G, Duret l, Goug M. **HOBACGEN: Database System for Comparative Genomics in Bacteria.** *Genome Research* 2000 **10:**379-385.

Portugaly E, Linial M. **Estimating the probability for a protein to have a new folder: A statistical computational mode.** *Proc Natl Acad Sci U S A* 1999. **96(7):**3578-83.

Rubin C. **Comparative genomics of the eukaryotes.** *Science* 2000. **287**:2204-2215.

Ruvinsky I, Silver LM, Gibson-Brown JJ. **Phylogenetic analysis of T-Box genes demonstrates the importance of amphioxus for understanding evolution of the vertebrate genome.** *Genetics* 2000. **156(3):**1249-57.

Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L, Falkow S. **A whole-genome microarray reveals genetic diversity among Helicobacter pylori strains.** *Proc Natl Acad Sci U S A* 2000. **97(26):**14668-73.

Spring J. **Genome duplication strikes back.** *Nat Genet* 2002. **31(2):**128-9.

Sasson O, Linial N, Linial M. **The metric space of proteins- comparative study of clustering algorithms.** *Bioinformatics* 2002. Science 1998. **282(5396):**2022-8.

Skaer N, Pistillo D, Gibert J, Lio P, Wülbeck C and P Simpson P. **Gene duplication at the achaete-scute complex and morphological complexity of the peripheral nervous system in Diptera.** *Trends in Genetics* 2002. **18(8):**399-405.

Smith GC, Eyre-Walker A. **Nucleotide Substitutions rate Estimstion in Enteroacteria: Approximate and Maximum-Likelihood Methods Lead to Similar Conclusions.** *Mol Biol Evol* 2001. **18(11)**:2124-2126.

Smith GC, Hurst LD. **Sensitivity of Patterns of Molecular Evolution to Alterrations in Methodology: A Critique of Hughes and Yeager.** J Mol Biol 1998. 47:493-500.

Smith TF, Waterman MS. **Identification of common molecular subsequences.** *J Mol Biol* 1981. **147**:195-197.

Snel B, Bork P, **Huynen MA. Genomes in flux: the evolution of archaeal and proteobacterial gene content.** *Genome Research* 2002. **12:**17-25.

Steiner H, Haass C. **Intramembrane proteolysis by presenilins.** *Nat Rev Mol Cell Biol* 2000. **3:**217-24.

Tatusov RL, Koonin EV, Lipman DJ. **A genomic perspective on protein families.** *Science* 1997. **278(5338):**631-7.

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. **The COG database: new developments in phylogenetic classification of proteins from complete genomes.** *Nucleic Acids Research* 2001. **29**:22-28.

Taylor JS, Van de Peer Y, Meyer A. **Genome duplication, divergent resolution and speciation.** *Trends Genet* 2001. **17**:299-301.

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

Theben G. **Secret life of Genes.** Nature 2002. 415:741

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH,

Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner

R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. **The Sequence of the Human Genome**. *Science* 2001. **291:**1304-1351.

Vision TJ, Brown DG, and Tanksley SD. **The Origins of Genomic Duplications in Arabidopsis.** *Trends Genet* 2000. **16(5):**227-31.

Wagner A. **Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate.** *Proc Natl Acad Sci U S A* 2000. **97(12):**6579-84.

Wheelan SJ, Boguski MS, Duret L, Makalowski W. **Human and nematode orthologs--lessons from the analysis of 1800 human genes and the proteome of Caenorhabditis elegans.** *Gene* 1999. **238(1):**163-70.

Wolfe HK. **Yesterdays Polyploids and The mystery of Diploidization**. *Nature Reviews* 2001. **2:**333-341.

Wolfe KH, Sharp PM, Li WH. **Mutation rates differ among regions of the mammalian genome**. *Nature* 1989. **337:**283-285.

Wolfe KH, Shields DC. **Molecular evidence for an ancient duplication of the entire yeast genome.** *Nature* 1997. **387**:708-713.

Wong S, Butler G, Wolfe KH. **Gene order evolution and paleopolyploidy in hemiascomycete yeasts.** *PNAS* 2002. **99:**9272-9277.

Yang Ji, Huang J, Gu H, Zhong Y, Yang Z. **Duplication and Adaptive Evolution of the Chalcone Sytdeshase Genes of Dedranthema (Asteracease).** *Mol Biol Evol* 2002. **19(10):**1752-1759.

Yang Z, Nielsen R. **Synonymous and Nonsynonymous Rate Variation in Nuclear Genes of Mammals.** *J Mol Evol* 1998. **46:**409-418.

Yang Ziheng, Rasmus Nielsen, et al: **Models of Amino Acids Substitution and Applications to Mitochondrial Protein Evolution.** *Mol Biol Evol* 1998. **15(12):**1600-1611.

Yona G, Linial N, Linial M. **ProtoMap: automatic classification of protein sequences and hierarchy of protein families.** *Nucleic Acids Research* 2000. **28(1):**49-55.

Zhang J, Rosenberg HF, Nei M. **Positive Darwinian selection after gene duplication in primate ribonuclease genes.** *Proc Natl Acad Sci U S A* 1998. **95(7):**3708-13.

Zhang J. **Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes**. *J Mol Evol* 2000. **50(1):**56-68.

Appendix

1. Perl script for parsing blast results:

```
#! /usr/bin/perl -w

print STDERR "Blast output\n";
$file = <STDIN>;;


open(IN,"$file");
$count = 0;
while (<IN>) {


      if ($_=~/^Query=\s(\S+)\s+(.*)\s$/) {
      $queryname = $1;
    }
          if ($_ =~/^Sequences producing significant/) {
                              <IN>;
                             # $_ = <IN>;
                             $_=<IN>;
                  /(\S+)\s+(.*)\s+(\S+)\s+(\S+)/;
                               $name = $1;
                               $P_score =$3;
                               $e_score = $4;
                      $one = "1e";
                       if ($e_score =~/^e\-\S+/) {
                        $e_score =~s/e/$one/;

                       }
                       print
"$queryname#$name#$e_score#$P_score\n";
                       $count++;

          }
}
print "Best hits with specified cut-offs points or better =
$count\n";
```

2. calc_ka_ks.pl : perl script for calculation synonymous and non-synonymous substitutions

```
#!/usr/bin/perl -w

##################################################################
#################

# NAME: cs_Ks_Ka.pl
# LANGUAGE: perl
#
# DEPENDENCIES: Uses EMBOSS, clustalw

# DESCRIPTION: takes the accession numbers of pairs of
orthologous genes and
```

```
# produces a distance database containing estimates of the
number of pairwise
# estimates of synonymous and nonsynonymous per site between
two aligned
# nucleic acids which code for proteins. Utilises Codeml to do
so.

# USAGE: input should be a file containing the accession
numbers of pairs of
# orthologous sequences separated by either a space or tab
deliminator.
#
# OUTPUT: 1: name1 2: name2 3: nucleotide distance 4: protein
distance
# 5: total number of sites 6: total number of non-degenerate
sites
# 7: total number of two-fold degenerate sites
# 8: total number of four-fold degenerate sites 9: Ks 10: Ks
standard deviation
# 11: Ka 12: Ka standard deviation 13: Ka/Ks
#
#
##################################################################
################

$| = 1;
$ortho_file = "web_infile";
$path = "/cip0/research/victoria/2002/RESULTS/RUNPIPE";
open (INFILE, "$path/$ortho_file") || die "Can't open
$ortho_file: $!";
open (OUTPUT, "+>>$path/dist_db_$ortho_file") || die "Can't
open dist_db_$ortho_file\n";
open (LOG, "+>>$path/error_log_$ortho_file");

$date = `date`;


print LOG
"================================================================
====\n";
print LOG "ERROR LOG PART: $ortho_file\n";
print LOG "This part of the output contains a list of error
messgaes for why divergence \n";
print LOG "calculations could not be carried out for certain
pairs of\n";
print LOG "orthologous sequences which have been entered into
the program\n";
print LOG
"================================================================
====\n\n";
print LOG "$date";

while (<INFILE>)
{
      @fields = split (/\s+/, $_);
      $ortho1_nuc = $fields[0];
      $ortho1_nuc =~s/\s+//g;
      $ortho2_nuc = $fields[1];
      $ortho2_nuc =~s/\s+//g;
      system("cp $path/$ortho1_nuc $path/nuc1");
      system("cp $path/$ortho2_nuc $path/nuc2");
```

```
##########next few lines are for the yeast data only

            if ($ortho1_nuc =~/\.R/) {
            system ("revseq -sequence $path/$ortho1_nuc -outseq
$path/nuc1");
            } else { system("cp $path/$ortho1_nuc $path/nuc1");
            }
            if ($ortho2_nuc =~/\.R/) {
            system ("revseq -sequence $path/$ortho2_nuc -outseq
$path/nuc2");
            } else { system("cp $path/$ortho2_nuc $path/nuc2");
            }
##################

system("transeq $path/nuc1 -outseq $path/prot1");
system("transeq $path/nuc2 -outseq $path/prot2");
&remove_star("$path/nuc1");
system("cp $path/yes_stop $path/nuc1");
&remove_star("$path/nuc2");
system("cp $path/yes_stop $path/nuc2");
system("cat $path/prot1 $path/prot2 > $path/protein_file_1");
system("clustalw -infile=$path/protein_file_1 -output=gde -
outfile=$path/prot_align_1.gde");
system("clustalw $path/prot_align_1.gde -tree -kimura -
outputtree=dist");
$prot_dist = &get_dist("$path/prot_align_1.dst");

                        open (GDE, "$path/prot_align_1.gde") ||
print LOG "Can't open prot_align_1.gde\n";
                        open (FASTA, ">$path/prot_align_1.fasta")
|| print LOG "Can't open prot_align_1.fasta\n";
                        while (<GDE>)

                        {
                              s/%/>/;
                              print FASTA $_;
                        }
                        close (FASTA);
                        close(GDE);

                        system("cat $path/prot_align_1.fasta
$path/nuc1 $path/nuc2 > $path/all_seq_1.fasta");
                        open (ALL_SEQ, "$path/all_seq_1.fasta")
|| print LOG "Can't open all_seq_1.fasta\n";
                        open (NAME_SEQ,
">$path/nuc_prot_1.fasta") || print LOG "Can't open
nuc_prot_1.fasta in order to change seq names\n";


                        $ctcs =0;
            while (<ALL_SEQ>)
                        {
                              s/_1//;
                              s/_4//;
                                    if(/^>/) {
                                    $ctcs++;
                                          if($ctcs == 1) {
                                          $name1 = $_;
                                          }
                                          if($ctcs == 2) {
                                          $name2 = $_;
                                          }
                                          if($ctcs == 3) {
```

```
                                        $_ = $name1;
                                        }
                                        if($ctcs == 4) {
                                        $_ = $name2;
                                        }
                                  }
                                  unless(/>/) {
                              tr/A-Z/a-z/;
                                  }
                              print NAME_SEQ $_;
                        }
                        close(NAME_SEQ);
                        close(ALL_SEQ);
                        #system("perl align3aa.pl
nuc_prot_1.fasta");
                        system("cat $path/nuc1 $path/nuc2 >
$path/nuc_all");
                        system("tranalign  $path/nuc_all
$path/nuc_prot_1.fasta  -outseq $path/tranalign.out");
                        open (CONVERT, "$path/tranalign.out") ||
print LOG "Can't open tranalign.out in order to convert its
sequence format\n";
print "HERE\n";
                        open (NEWFORMAT,
">$path/nuc_prot_1.aligned.gde") || print LOG "Can't open
nuc_prot_1.aligned.gde in order to print new sequence format
to it\n";
                        $cs_len = 0;
                        <CONVERT>;
                        WHILE: while (<CONVERT>) {
                              if(/>/) {
                              last WHILE;
                              }
                        $cs_len = $cs_len + length($_) - 1;
                        }

                        close(CONVERT);
                        open (CONVERT, "$path/tranalign.out") ||
print LOG "Can't open tranalign.out in order to convert its
sequence format\n";

                        while (<CONVERT>)
                        {
                              s/>/#/;
                              print NEWFORMAT $_;
                        }
                        close(NEWFORMAT);
                        close(CONVERT);
&pad_nucleotide_sequence("$path/tranalign.out");
#print "HERE calling clustal withinfile =
nuc_prot_1.aligned.gde and output should be a tree\n";
                        system("clustalw -
infile=$path/nuc_prot_1.aligned.gde  -tree -kimura -
outputtree=dist");
                        $nuc_dist =
&get_dist("$path/nuc_prot_1.aligned.dst");
                        system("clustalw -
infile=$path/nuc_prot_1.aligned.gde -convert -output=gcg -
outfile=$path/nuc_prot_1.aligned.gcg");
                        open (CODEMLFORMAT,
">$path/nuc_prot_1.aligned.cdml") || print LOG "Can't open
nuc_prot_1.aligned.cdml in order to print new sequence format
to it\n";
```

```
                        open (CONVERT, "$path/tranalign.out") ||
print LOG "Can't open nuc_prot_1.fasta.aligned2aa in order to
convert its sequence format\n";
                  print CODEMLFORMAT "2 $cs_len\n";
                        while (<CONVERT>) {
                              s/>//;
                              print CODEMLFORMAT $_;
                        }
                        close(CODEMLFORMAT);
                        close(CONVERT);
            #system("rm $path/rst");
            system("$path/codeml");
                        &parse_codeml();
                        print OUTPUT "$ortho1_nuc\t";
                        print OUTPUT "$ortho2_nuc\t";
                        print OUTPUT "$S\t";
                        print OUTPUT "$N\t";
                        print OUTPUT "$ks_value\t";
                        print OUTPUT "$ka_value\t";
                        print OUTPUT "$ka_ks_value\n";
system("rm $path/nuc_prot_1.aligned.cdml");
system("rm $path/protein_file_1");
system("rm $path/prot_align_1.gde");
system("rm $path/prot_align_1.fasta");
#system("rm $path/prot_align_1.dst");
system("rm $path/nuc1 $path/nuc2 $path/prot1 $path/prot2");
system("rm $path/all_seq_1.fasta");
system("rm $path/nuc_prot_1.fasta");
system("rm $path/tranalign.out");
system("rm $path/nuc_prot_1.aligned.dst");
system("rm $path/nuc_prot_1.aligned.gcg");
system("rm $path/nuc_prot_1.aligned.ph");
system("rm $path/prot_align_1.ph");
system("rm $path/protein_file_1.dnd");
system("rm $path/nuc_prot_1.aligned.gde");
system("chmod 0777 $path/error_log_web_infile");
system("chmod 0777 $path/dist_db_web_infile");

print
"\n\n\n\n\n\n\n\n\n\n\n\n\nXXXXXXXXXXXXXX\n\n\n\n\n\n\n\n";
}
close (OUTPUT);
close (LOG);
system("chmod 0777 $path/error_log_web_infile");
system("chmod 0777 $path/dist_db_web_infile");
system ("cat $path/dist_db_web_infile
$path/error_log_web_infile > $path/results_web_infile");
system("chmod 0777 $path/results_web_infile");




##################################################################
####


sub pad_nucleotide_sequence {
$infile = $_[0];
$outfile = $_[0];  #same file as infile
$ct = 0;
@seq = ();
```

```
print "padding sequence=$infile\n";
open(IN,"$infile");

      while(<IN>) {
            if(/>/) {
            $ct++;
            $name[$ct] = $_;
            }
            else{
            chomp($_);
            $seq[$ct] = $seq[$ct] . $_;
            }
      }
      if(length($seq[1]) > length($seq[2]) ) {
      $seq[2] = $seq[2] . "-" x (length($seq[1]) -
length($seq[2]));
      }

      elsif(length($seq[2]) > length($seq[1]) ) {
      $seq[2] = $seq[1] . "-" x (length($seq[2]) -
length($seq[1]));
      }

close(IN);
open (OUT,">$outfile");
$ct = 0;
print OUT "$name[1]";

      while($ct * 50 < length($seq[1]) - 1 )    {
      $sub = substr($seq[1],$ct*50,50);
      print OUT "$sub\n";
      $ct++;
      }

$ct = 0;
print  OUT "$name[2]";

      while($ct * 50 < length($seq[1]) - 1 )    {
      $sub = substr($seq[2],$ct*50,50);
      print OUT "$sub\n";
      $ct++;
      }

close(OUT);
}


sub parse_codeml {
$ks_value = "notfound";
$ka_ks_value = "notfound";
$ka_value = "notfound";
$S = "notfound";
$N = "notfound";

open (RST,"$path/rst") || die "can't open rst file\n";

while(<RST>) {
#print STDERR "here\n";

      if(/2\s+1\s+(\S+)\s+(\S+)\s+(\S+)\s+(\S+)\s+(\S+)/) {
      $ks_value = $4;
      $ka_ks_value = $5;
      $ka_value = $3;
```