Exploitation of Genomic Data in the Prediction of Gene Function.

Paul Erskine Boardman

May 2003

A thesis submitted to the University of Manchester Institute of Science and Technology for the degree of Doctor of Philosophy

Declaration: No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university, or other institute of learning

Acknowledgements

I feel indebted to many people who have helped me in my work and to stay sane. Here's a list of most of them. I'm bound to miss one or two people out. If that's you, I owe you a pint. Even if you're on the list, I probably owe you a pint anyway.

- My supervisors, Dr. Simon Hubbard and Prof. Steve Oliver who appointed me to this position and then guided me through the trials and tribulations of a PhD.
- Claire Wilson, who has been a close friend and colleague throughout an M.Sc. and a Ph.D., has shown me the value of true friendship and most of all has made me cakes.
- My parents, without whom I wouldn't even exist, who have supported me, both financially and emotionally, throughout my life. Thanks folks! I think I owe you a whole brewery (or would you rather a distillery?)!
- Simon Oliver, Chris Heeley and Sid for providing excellent computer support.
- Dr. Ben Stapley for persuading me that the PSB 2002 conference was worth going to; the Bioinformatics group and many people from Biomolecular Sciences at UMIST for their friendship and advice.
- The MRC for funding my PhD and the BBSRC for providing the money for the Chicken EST project.
- Noel, James, Pete, Mark, Alison, Neil, Rob and Martin for all the climbing and chilling.
- Paul D'Ambra and Mike Sutherland for the hedonism, music and headspace.
- Stuart Wilson and Simon Hubbard for employing me and allowing me to continue to develop the Chicken EST resource.
- And last but not least, Lindsey Jones for her creativity, caring, support, acceptance, cooking and friendship. It's been fun!

Abstract

Large-scale genome sequencing projects have made their way into the forefront of both academic and general news in recent years with the completion of milestones such as the first eukaryotic genome in 1996 and a draft of the human genome in 2001. These represent immensely important and useful resources in the quest to understand how life works. One of the important stages in the transformation of this sequence information into useful data is the elucidation of coding sequences within the genomes and the subsequent assignment of function to these sequences. The most widely used tools for the latter task assign function through the identification of similarity between the newly discovered genes and previously characterised sequences. Although these techniques are very powerful for functional assignment, they fail to identify relationships for remote homologues and species-specific genes. At first this may sound like a trivial problem, but every eukaryotic sequencing project to date has failed to assign functions to between 20 and 70% of all newly discovered genes. Obviously, methodologies that are not dependent on the detection of homology relationships offer vitally important avenues for the elucidation of function for these unannotated sequences. In this thesis, we first describe the generation and annotation of 330,000 chicken ESTs, which represents a classic demonstration of the use of homology-based techniques in a large-scale sequencing project. Approximately 50% of the newly generated EST sequences can be assigned a function using homology-dependent techniques, which leaves a large proportion for which similarity searches have failed to find homologues in the public databases. The second section of this thesis describes an investigation into a non-homology based technique for associating genes. This approach attempts to associate genes through similarities in the presence, absence and patterns of transcription factor binding sites in their upstream regions. This analysis results in the definition of methodologies for generating lexicons enriched in binding sites for functionally related genes, suggesting a novel approach for gene function prediction.

Table of contents

Abbreviations

BLAST	Basic Local Alignment Search Tool
CGI	Common Gateway Interface
Contig	Contiguous sequence
DNA	Deoxyribonucleic Acid
EMBL	European Molecular Biology Laboratory
EST	Expressed Sequence Tag
GTF	General Transcription Factors
HSP	High Scoring Pair
InterPro	Integrated resource of Protein Families, Domains and Sites
MIPS	Munich Information center for Protein Sequences
mRNA	Messenger RNA
mtDNA	Mitochondrial DNA
OMIM	Online Mendelian Inheritance in Man
ORF	Open Reading Frame
RNA	Ribonucleic Acid
RNAi	RNA interference
rRNA	Ribosomal RNA
RST	Random Sequence Tag
SCPD	Saccharomyces Cerevisiae Promoter Database
SDS-PAGE	Sodium Dodecyl Sulphate Ployacrylamide Gel Electrophoresis
TBP	TATA Binding Protein
TCP/IP	Transmission Control Protocol/Internet Protocol
TF	Transcription Factor
TM	Trans membrane
TRANSFAC	Transcription Factor Database
TrEMBL	Translated EMBL
URS	Upstream Regulatory Sequence
UTR	UnTranslated Region

Introduction

1.1 Functional Genomics & Bioinformatics

With the recent trend in genome sequencing, biologists have found themselves overwhelmed with vast quantities of genomic sequence data. Biology is moving from a data-poor science to becoming a data-rich science (Vukmirovic 2000).

In April 1994, work began at TIGR on a project to sequence the entire genome of *Haemophilus influenzae*. TIGR researchers applied a technique known as whole genome shotgun sequencing, which involves the sequencing of randomly generated pieces of genomic DNA. These fragments are then joined together through a process known as assembly to generate the final, complete, genome sequence. To date nine eukaryotic, 16 archeal, 96 bacterial and over 1000 viral genomes have been completely sequenced, with many more in the process of being sequenced. Table 1.1 shows a cross-section of the more important genomes that have been sequenced so far (data from web ref 1).

Table 1.1 Landmarks in the efforts of the genome sequencing projects.

Name	Size(Mb)	Year of completion	Predicted number of Genes
Bacteriophage ΦX174	0.005	1977	10
***Haemophilus influenzae* Rd**	1.8	1995	1730
Saccharomyces cerevisiae	12	1996	6,200
***Escherichia coli* (strain K12)**	4.6	1997	4289
Caenorhabditis elegans	97	1998	19,000
Drosophila melanogaster	180	2000	13,600
Arabidopsis thaliana	125	2000	25,500
Homo sapiens	3150	2001	30 → 40,000
Plasmodium falciparum	23	2002	5,300

These complete genome sequences represent an immensely valuable resource for biologists. In the case of *Plasmodium falciparum* for example, the generation of the

complete genome sequence has the potential to open up new areas of research for the development of new drugs and vaccines, improved diagnostics and effective vector control techniques (Gardner 2002). Unfortunately, the billions of bases of DNA sequence do not tell us what all the genes do, how cells work, how cells form organisms, what goes wrong in disease, how we age or how to develop a drug (Lockhart 2000). To find answers to these questions, a range of post-genomic technologies have been developed.

These next stages of analysis involve the area of functional genomics as well as other 'omics such as transcriptomics (DeRisi 1997), proteomics (Fey 1997) and metabolomics (Tweedale 1998). This is therefore more than just the determination of coding sequences and their putative functional assignments; it is also the elucidation of the organisation and control of genes in genetic pathways that operate to create the physiology of an organism. In order to put the work contained in this thesis into perspective, the science of genomics and bioinformatics will be covered in this introduction, along with the other post-genomic techniques involved in the 'omics field.

Although genomics has only relatively recently been at the forefront of biological research, the term "genome" itself is more than 75 years old and refers to the complete set of genes and chromosomes for an organism. The term "genomics" was coined in 1986 by Thomas Roderick to describe the discipline of mapping, sequencing and analysing genomes. As the emphasis of genomics shifted from the mapping and sequencing of genomes to the mining of this data, the field split into two areas. These are usually referred to as structural genomics, which is involved in the mapping and sequencing of genomes, and functional genomics, which refers to the development and application of genome-wide experimental approaches to predict gene function. This involves large-scale, high-throughput methodologies combined with statistical and computational analysis of the results (Hieter 1997).

Bioinformatics has become integral to all of these post genomic sciences. The term itself was coined in the late 80's by Hwa Lim to describe the computational biology involved in analysing, storing and searching the data that was being produced by structural genomics at the time. It now commonly refers to computational work in genomics and the other 'omics' that have developed. The precise boundaries of

bioinformatics are elusive and to muddy the water further some people distinguish bioinformatics from computational biology, using the latter to denote computational work whose agenda is clearly biological (Goodman 2002).

A large component of bioinformatics is the creation and maintenance of computational resources such as databases (e.g. EMBL (Stoesser 2003), GenBank (Benson 2003), Swiss-Prot (Boeckmann 2003) etc.) which store the vast numbers of gene and protein sequences along with associated functional annotation and meta-data, and tools used to search and mine these resources to find relationships between their component sequences (e.g. BLAST, HMMER, ClustalW). However bioinformatics is not only concerned with these areas, and is currently a broad field that has a central role in multiple biological research areas. Along with its role in genomic sequencing, mapping, genome annotation and comparison of genomes, it is also essential in transcriptomics (the study of transcribed sequences, both full length cDNAs and expressed sequence tags) and the analysis of gene expression data from DNA microarrays. It also has a crucial role in proteomics for the study of protein abundance, the analysis of protein sequences and the determination of protein structure. Bioinformatics is valuable in the study of gene regulation (the 'regulome') and in the analysis of molecular pathways and protein-protein interactions (the 'interactome'). Further to this, bioinformatics has proven to be essential in studies of evolution and phylogeny, as well as playing a significant role in structural biology. It is beyond the scope of this introduction to describe the whole field of bioinformatics in any detail. Instead, a broad overview of the relevant areas of bioinformatics to this work will be discussed; focussing on the sequence analysis techniques used to assign evolutionary relationships and biological function to gene and protein sequences.

Of course, the field of bioinformatics would not even exist without the masses of data produced from the wide variety of experimental systems for large-scale analyses of DNA, RNA and protein within an organism. In order to place these data in biological context, they will be briefly described in section 1.2.

1.1.1. Aims in brief

There are two main areas to the project described in this thesis. Firstly, the design, implementation and analysis of a large-scale chicken EST resource is undertaken. This involves a large number of classical bioinformatic techniques, which are described later on in this and further chapters. Secondly, we examine potential uses of transcription factor binding sites in the prediction of gene function. In this analysis, many, publicly available, bioinformatic programs are used in conjunction with pipelines and programs developed specifically for this project. These two analyses use a number of programs, ideas and raw data generated by the post-genomic technologies described later on in this chapter. Both are focussed on the exploitation of genomic data for the elucidation of biological function, although they represent two very different approaches.

1.2 Experimental techniques

There are many experimental techniques available for mapping interactions between genes, interactions between proteins and the interactions of proteins with DNA. Many of these associations are used to assign function to unclassified genes/proteins using the principle of "Guilt by Association". By finding an association between a gene of known function and a gene of unknown function, you gain insight into the role of the unknown gene as it is most likely to share the function of the known gene on some level. This section will cover the main areas of functional genomics that are currently being used to investigate gene/protein function.

1.2.1 Microarray experiments

DNA microarrays (cDNA arrays, GeneChips or just 'chips') are made up of thousands of oligonucleotides immobilised on a surface (e.g. glass, plastic or silicon). Traditionally these oligonucleotides were of unknown sequence but recently, thanks to the knowledge of complete genome sequences and large cDNA collections, it has become common to design arrays based on specific sequence information (Lockhart 2000). Thus, a microarray chip can be designed that contains every gene in a genome by immobilising (or 'spotting') unique oligonucleotides, representing each gene sequence, onto a surface. This can then be exploited using the standard complementary base pairing observed in biology to allow the expression of RNA transcripts to be

monitored. Labelled cDNA from cell extracts can then hybridise to these oligonucleotides allowing a quantitative analysis of gene expression levels, giving a value for every single gene. In this way 'chips' can be designed that allow the monitoring of whole genome expression levels, or of a more specific sub-set of genes, under different physiological conditions (DeRisi 1997, Zhang 1999).

By amassing a large enough set of data under varied conditions, this approach should allow the elucidation of the physiological role of many, if not all, of the genes in an organism. The most obvious first step in the analysis of the microarray data is to examine the extremes, e.g. genes that show significant difference in expression levels between different physiological or pathological conditions, or in mutant strains compared to wild type organism. Although this determination of differentially expressed genes can be useful for many areas of biological understanding including the identification of putative drug targets, this approach does not exploit the full potential of genome-wide expression analysis. This is where the "Guilt by Association" approach is applied, where the expression values of genes observed over a range of time points or experimental conditions are compared. So far, the most powerful approaches form groups (or clusters) of genes that display mathematically similar expression patterns. These can then be displayed in an intuitive manner easing the process of data mining (Eisen 1998, web ref 2). The major premise here is that genes with similar expression profiles are functionally related. This has proven to be well founded, with many analyses generating clusters which contain functionally related transcripts (e.g. Eisen 1998, Zhang 1999). An analysis on *Saccharomyces cerevisiae* by Eisen *et al.* (1998) successfully generated a number of clusters containing functionally related genes. For example, cluster H is composed entirely of genes involved in chromatin remodelling, and cluster E is composed predominantly of genes involved in glycolysis.

A further use of this technology, which has proven successful, is in the area of diagnostics. Samples from patients suffering from various tumours have been taken, RNA extractions and microarray experiments performed, and the gene expression patterns determined, allowing the diagnosis of tumour type and prescription of appropriate treatment (Hampton 2003). One exciting result form these analyses is that they can distinguish between tumours of different anatomical origin, and define new subgroups of cancer with similar histological appearance.

The fact that functionally related genes often have similar expression patterns can be taken to infer that these genes are controlled by a common mechanism. This idea has been investigated by many groups who have searched for common transcription factor binding sites in the upstream regulatory sequences (URS) of co-expressed genes (Chu 1998, Schuldiner 1998, Wolfsberg 1999, Zhang 1999, Fujibuchi 2001). These sites can be previously determined sites (Schuldiner 1998) or sequences that are predicted to be TF binding sites due to the significance of their occurrence within the URSs of the co-expressed genes (Van Helden 1998, Wolfsberg 1999). In this way potential common control sequences have been identified for many functionally related genes. Given enough data about the binding nature of a transcription factor (e.g. its binding sequence and target genes) it can be assigned a physiological or phenotypic role. This can then be used to infer physiological function to other genes found to be under the influence of this transcription factor (Fondrat 1994).

1.2.2 Yeast two-hybrid experiments

The two-hybrid system is based on the discovery that some eukaryotic transcription activators are modular. It was found that the activation domain of the yeast transcription factor GAL4 could be fused to the DNA binding domain of *E. coli* LexA to create a functional transcription activator in yeast (Brent 1985). Fields and Song (1989) presented the possibility of detecting protein-protein interactions by expressing the genes as chimeras. In their experiment, they fused SNF1 and SNF4 (yeast proteins that have been demonstrated to interact in vitro) to the DNA binding domain of GAL4 and the activation domain of GAL4 respectively. The DNA binding domain of GAL4 binds to specific DNA sequences (UASG). Fields and Song were able to successfully detect interactions of proteins fused independently to the activator and DNA binding domain by observing the activated expression of a reporter gene regulated by the UASG sequence. Hence, this provides a system for studying any potential protein-protein interaction, so long as the independent, folded and active proteins can be fused to the components of the 2-hybrid system and subsequently expressed.

This technique is well suited to large-scale analyses and it was soon adapted for this purpose. The first genomic analysis using the two-hybrid system centred on the T7

bacteriophage (Bartel 1996). The first eukaryotic whole-genome analysis was carried out on *Saccharomyces cerevisiae* by Uetz *et al.* (2000). Two, separate, approaches were taken in this investigation. One approach was to screen a protein array of 6,000 yeast transformants, with each transformant expressing one of the open reading frames as a fusion to an activation domain, with 192 yeast proteins. The other mated a library of transformants containing the 6,000 potential ORFs fused to the GAL4 activation domain and a library of transformants fused to the GAL4 DNA-binding domain. These approaches identified 957 putative interactions involving over 1,000 proteins. The interaction data placed functionally unclassified proteins into a biological context giving a clue as to their functional role.

1.2.3 Further techniques

Of the remaining technologies that have been adapted to genomic scale analyses of gene function, co-immunoprecipitation and gene disruption are two of the most prominent.

Co-immunoprecipitation

Immunoprecipitation is a technique that allows the purification of specific proteins for which an antibody has been raised. This primary antibody is either bound to agarose or can be bound to the protein/agarose beads during the procedure to physically separate the antibody-antigen complex from the remaining sample. Co-immunoprecipitation allows the identification of proteins that physically interact with the primary protein through virtue of their co-precipitation with the antibody-antigen complex.

For example, Ho *et al.* (2002) used 725 'bait' proteins from a variety of functional classes in a one-step immuno-affinity purification. Proteins from over 1,500 immunoprecipitations were resolved using SDS-PAGE which gave rise to the identification of over 8,000 potential protein-protein interactions.

Gene Disruption

The European Functional Analysis Network (EUROFAN) was established at the beginning of 1996 with the primary aim of determining the function of the ~ 4,000 novel proteins in the yeast genome with no assigned function (Oliver 1998, Dujon 1998). This was approached by creating a set of single-gene deletion mutants covering nearly all ORFs from *S. cerevisiae.* The mutants are submitted to many specialised functional assays, and studies are performed at the transcriptome and proteome levels.

A similar project has been carried out by Ross-Macdonald (1999) and co-workers who used transposon mediated gene disruption to explore the function of nearly 2,000 yeast genes under 20 different growth conditions. This also led to the identification of 300 previously non-annotated open reading frames.

Recent insights into the sequence-specific, gene-silencing, action of double stranded RNA (dsRNA) has led to the development of RNA interference (RNAi) as an experimental tool for the analysis of gene disruption (Guo 1995, Fire 1998). Kamath *et al* (2003) used RNAi to inhibit the function of ~86% of the 19,427 predicted genes of *C. elegans.* This facilitated the identification of mutant phenotypes for 1,722 genes two thirds of which were not previously associated with a phenotype. This analysis also provided further insights into genome organisation by showing that genes of similar function are often clustered in distinct regions of individual chromosomes and that these genes tend to share transcriptional profiles. In a similar, genome-wide, gene-disruption investigation, Ashrafi *et al.* (2003) identified 417 genes necessary for the normal storage of fat in *C. elegans.* Many of these genes have human homologues, which highlights targets for treating obesity and its associated diseases.

1.3 Predicting Gene Function

1.3.1 What is function?

Before the process of functional prediction can be addressed, the concept of function needs to be defined. Protein function has many levels of meaning (Skolnick 2000): -

- Biochemical function – the substrate specificity of an enzyme, the chemical reaction it catalyses e.g. serine protease, protein kinase. The protein-protein interactions involved in the binding of cofactors and regulatory molecules are also levels of biochemical function.
- Cellular function – at this level the function of a protein reflects its cellular localisation and its interactions with other macromolecules.
- Physiological - the physiological role of the protein (growth hormone, adrenaline) or which metabolic pathway the protein is involved.
- Phenotypical – the role of the protein in the whole organism. Gene knockout experiments are designed to elucidate the phenotypical function of a protein.

1.3.2 Historical perspective

Sequence determination and analysis began on proteins in the 1950s, with RNA starting about a decade later and DNA a decade later still. Hence, many of the concepts for function annotation and prediction were first developed by looking at amino acid sequences.

Analysis of proteins came first due to their greater chemical and structural stability and ease of purification in large quantities. Several approaches were introduced, separating and purifying the subunits of a protein, purifying the polypeptide fragments created via proteolysis with proteases, using specialised protein chemistry techniques to determine the residue order of the subfragments, finding the overlaps and then assembling the contiguous subsequences into the complete sequence (Hodgman 2000).

Nucleotide sequences were much more difficult to analyse due to their greater length and lower stability. It was in the 1970s that saw the birth of DNA sequencing with the

advent of *in vitro* DNA cloning, interrupted replication synthesis (Sanger 1975), modified chain-terminating nucleotides (Sanger 1977) and thin gels which allowed longer reads (Sanger 1978). These innovations facilitated the large-scale analysis of DNA sequences.

Many different areas of biological research rely on the comparison of one or more biological objects with a reference set of others (e.g. species identification for ecology, tissue status determination for pathological specimens). The assignment of function to genes and proteins uses the same principle in a process known as pairwise sequence alignment. Again this approach follows the "Guilt by Association" principle, where a function for an unknown entity is inferred, in some part, from a known entity that has some defined relationship with the unknown one. In the case of gene and protein function, this is usually achieved by establishing that the two genes or proteins share come common evolutionary ancestor, and are therefore homologous.

1.3.3 Homology-based methods

1.3.3.1 Pairwise sequence alignment

Needleman and Wunch developed the first widely regarded sequence alignment algorithm in 1970. Sequence comparison and multiple sequence alignment, of proteins with known structure, helped to determine the parameters for judging the validity of alignments. From these analyses it became apparent that percentage identity alone was inadequate in detecting evolutionary relationships.

"Direct comparison of two sequences, based on the presence in both of corresponding amino acids in an identical array, is insufficient to establish the full genetic relationships between the two proteins" (Needleman and Wunsch. 1970).

Margaret Dayhoff developed a model of protein evolution that resulted in the creation of a set of substitution matrices, called Dayhoff or PAM (Percentage Accepted Mutation) matrices (Dayhoff. 1978), which became the parameters of choice for sequence alignment. The PAM matrix is a set of weights that is derived from how often different amino acids replace other amino acids in evolution. This was based on a

dataset of 1,572 accepted mutations between 34 superfamilies of closely related sequences.

The PAM matrices derive their substitution frequencies from global alignments of very similar sequences. An alternative approach was developed by Henikoff and Henikoff (1992) using multiple alignments of more distantly related sequences. This resulted in the set of BLOSUM (BLOcks Substitution Matrix) matrices, which came into popular use.

Sequence databases have been undergoing exponential growth since the early 1980s. This necessitated the use of computers for storage, administration and querying of the sequence data. Smith and Waterman (1981) produced one of the first, sensitive database search algorithms, by extending the Needleman and Wunsch approach to detect local rather than global alignments. Like the Needleman and Wunsch algorithm, this is still an optimal solution to the alignment problem, but it is perhaps more realistic as it considers sub-sequence matches which are more typical of biological sequence "units" such as domains. Although this remains one of the more sensitive pairwise algorithms, it is limited by its speed and was superseded by the FASTA series (Lipman 1985, Pearson 1988) and later the BLAST series (Altschul 1990, 1997) of database search programs.

FASTA

Wilbur and Lipman (1983) developed the FASTA program at a time when personal computers had insufficient speed and memory to scan a database using unassisted dynamic programming. This program utilises a fast procedure for DNA scans that, in concept, searches for the most significant diagonals in a dot-plot of the two sequences being compared. Those regions that could form a longer alignment are joined and, finally, dynamic programming is used over a narrow region of the high scoring diagonal to produce a final alignment. FASTA only discovers the top scoring alignment; it does not compute all high scoring alignments between two sequences. As a result, FASTA may not identify multiple shared domains or direct repeats.

BLAST

BLAST employs a heuristic method to find the highest scoring locally optimal alignments between a query sequence and a database. Unlike FASTA, the original BLAST algorithm (Altschul 1990) did not allow gaps, though a modified version for generating gapped alignments has now been developed (Altschul 1997). The BLAST algorithm is based on Karlin and Altschul's (1990) work on the statistics of ungapped sequence alignments. These statistics allow the estimation of the probability of obtaining an ungapped alignment with a particular score when searching against a database of a given size. This means that the BLAST algorithm can efficiently identify nearly all high scoring pairs (HSPs) above a cut-off in a database, and hence is likely to find all the related protein or gene sequences to a query using a particular set of parameters and mutation data exchange matrix. However, these probability scores can be confusing when searching a sequence against multiple databases; an alignment of the same two sequences will have different probability scores in different sized databases.

BLAST has become very popular largely because implementations of it have been very efficient, and it has been optimised to work with parallel UNIX architectures from an early stage.

Profile methods

One of the most important single algorithmic developments in sequence searching methodology has been the development of position-specific iterated BLAST (PSI-BLAST, Altschul 1997). This combines the speed of the BLAST algorithm with the advantages of searching with a sequence profile. On the first iteration of the PSI-BLAST program, a normal BLAST search is carried out. Matching sequences (above a predefined threshold) from this search are used to build a profile. The profile represents an empirical description of the amino acid residues found in homologues at each point of the query sequence. This profile is then used on subsequent iterations, of the PSI-BLAST program, to search the database. The profile is further refined based on any new matches found. This methodology has the advantage of being able to detect more distant sequence relationships than the standard pairwise alignment approaches.

Another advance is in the use of hidden Markov models (HMMs) to produce a profile of protein family alignments. The use of HMM representations of protein families captures additional information beyond amino acid preference; in particular, position specific gap penalties (Michalovich 2002). HMMs have found many uses. A library of HMMs for all the major protein families has been established under the name of PFAM (Bateman 2002). Other databases created using HMMs include TIGRFAMs and SMART, which are both produced using the HMMER package. Profile HMMs also have several uses for DNA. A common example is the detection of repeat family members in large-scale genomic sequence.

1.3.4 Limitations and dangers of homology based techniques

Since the recent explosion in sequence data emanating from gene and genome sequencing projects worldwide, biologists are unable to individually study and annotate the function of newly discovered genes on a case-by-case basis using wet-lab based experimental techniques alone. Bioinformatics is the route of choice for the large-scale functional annotation of genes using homology-based approaches. Despite the ubiquitous nature of tools like BLAST and the increasing size of the databases, the limitations of assigning function using sequence similarity methods have become increasingly apparent. For example, Figure 1.1 shows the percentage of predicted ORFs that could be assigned a function for the yeast *Saccharomyces cerevisiae* on completion of its genome sequence, and the current state of functional assignment. *S. cerevisiae* is a model organism and is considered to be well characterised, but on completion of its genome, one third of its ~6000 genes could not be assigned a putative function using homology based techniques.

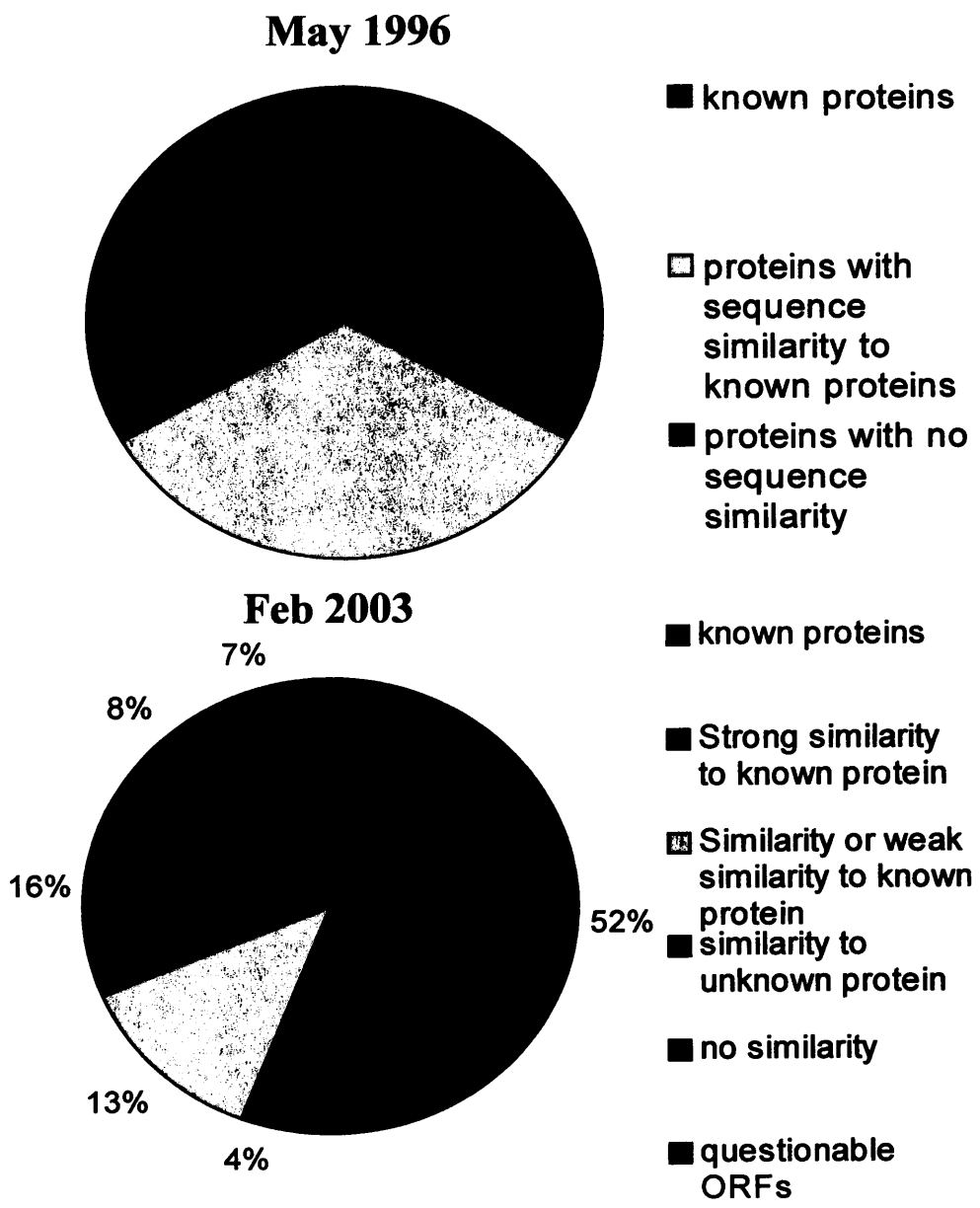


Figure 1.1. Present state of functional classification of the yeast genome.
Data taken from the MIPS website (web ref 3).

Although Figure 1.1 demonstrates an increase in the available expert knowledge for the yeast genome, it should be noted that this increased knowledge has come from characterising those genes with similarity to known proteins not from genes with no detectable homologues in the public databases. These unknown sequences (those with no detectable homology) account for almost one-third of all predicted genes in the genome. The figures for other completed (and near-complete) genomes can be even more pronounced. The bread mould *Neurospora crassa*, for example, has a predicted complement of some 10,000 genes. Of these, only 1,400 have counterparts in fruitflies,

worms and other complex eukaryotes and 4,140 (41%) lack any significant match to known proteins from public databases (Arnold 2003, Galagan 2003). One of the most pronounced examples of the failing of homology based techniques comes from an analysis of the genome of the human malaria parasite *Plasmodium falciparum.* This genome contains 5,268 predicted genes, 60% of which have no significant similarity matches in the public databases (Gardner 2002). The completion of the *falciparum* genome was heralded as a milestone in the fight against malaria. In order to design effective treatments and understand infection and cause of disease, we need to know the genome of the pathogen and have a handle on the function of the individual genes. This is obviously going to be a difficult task with complex genomes such as that of *Plasmodium falciparum.*

Even when homologues exist to our unknown genes and function assignment is possible following a database search, it is often unclear how much functional information can be legitimately transferred to the query sequence from a matched homologue. In the rush to automate the route from raw genomic data to biological and biomedical insight, we are most likely generating and propagating innumerable errors in our databases and in the literature (Brenner 1999, Attwood 2001). Brenner (1999) investigated the reliability of functional annotation by studying the inconsistencies in three different groups' functional annotations of the *Mycoplasma genitalium* genome. This organism has just 468 genes, many of which are fundamental for all life and should therefore be easily identifiable. Of the 340 genes annotated by all three groups, there was an apparent error rate of 8%. This investigation ignored minor disagreements in annotation and did not attempt to identify situations where multiple groups arrived at consistent but incorrect annotations. Therefore, the true error rate is likely to be higher than the reported figure. A similar study by Devos and Valencia (2001) using enzyme commission numbers was much more pessimistic, with estimations for the errors in annotation reaching up to 40%!

The annotation problem escalates when incorrectly annotated genes/proteins are entered into public databases (Bork 1999). These incorrect assignments can be propagated onto sequences used in subsequent searches. The repetition of this missannotation could lead to the rapid pollution of the databases. This highlights the necessity to maintain high standards of curation by the scientific community (Smith 1998) and to annotate the

source (experimental or computational) of functional assignment for a given sequence. A study by Karp *et al.* (2001) into the quality of metadata annotation (in this case the source of function annotation) in SwissProt was disappointing. Information about functional characterisation is apparently encoded in the SwissProt data bank (Junker 1999) though the study by Karp found that this mechanism is far from complete. Only 0.5% of SwissProt entries are marked as having their functions determined experimentally, which contradicts the description of the metadata system by the SwissProt authors. In defence of SwissProt, the study by Karp *et al.* was based on outdated GenBank entries and the SwissProt team are introducing evidence tags to allow users to identify the source of data items (Apweiler 2001a). Nevertheless, the functional situation remains problematic and alternative and/or complementary strategies to functional annotation by homology are needed. Some recent work has seen progress in this area.

1.4 Recent advances using non-homology-based methods

There has been a great drive to develop non-homology dependent techniques to predict gene/protein function in the hope of gaining further insight into the workings of organisms. These techniques link genes using metrics such as correlated expression profiles, correlated evolution profiles, and through the detection of putative protein-protein interactions.

1.4.1 Microarray experiments

Microarray analyses were introduced previously in section 1.2.1. It has been shown that genes with correlated expression profiles often share similar physiological functions. They have also been used to verify and assist standard bioinformatics annotations, using custom built chips that can cover large sections of genomic sequence (Shoemaker 2002).

1.4.2 Phylogenetic profiling

This method is designed to detect proteins that participate in a common structural complex or metabolic pathway. The assumption being that functionally linked proteins are likely to evolve in a correlated fashion being either preserved or eliminated in a new species (Pellegrini 1999). In general, these pairs of functionally linked proteins have no detectable sequence similarity with each other and, therefore, cannot be linked by conventional homology based techniques.

To carry out this analysis a phylogenetic profile is constructed for each protein. This consists of a string with *n* entities, each one a single bit, where *n* corresponds to the number of genomes considered in the analysis. Proteins are then clustered according to the similarity of their phylogenetic profiles. This groups together proteins with a correlated pattern of inheritance and implies functional linkage. Functions of uncharacterised proteins are likely to be similar to those of characterised proteins within a cluster.

Pellegrini and co workers validated this technique by creating phylogenetic profiles for all of the 4,290 proteins encoded by the *E. coli* genome by aligning each protein sequence with the proteins from 16 other fully sequenced genomes. They then examined the phylogenetic profiles of three well-characterised proteins, RL7 (which participates in the ribosome complex), FlgL (which is part of the flagellar structure) and HIS5 (a member of the histidine biosynthetic pathway). For both RL7 and HIS5 it was found that more than half of proteins with similar profiles (those profiles that were identical or differed by one bit) had a related function to the search protein. With FlgL all 10 proteins exhibiting the same profile were flagellar proteins. None of the three proteins share sequence similarity with their 'profile neighbours' and so these relationships couldn't have been detected using sequence similarity searches.

1.4.3 Rosetta stone analysis

This analysis is based on the observation that certain protein families in a given species consist of fused domains that usually correspond to single, full-length proteins in other species (Enright 1999). For example, the Gyr A and Gyr B subunits of *Escherichia coli* DNA gyrase are fused into a single chain in the topoisomerase II of yeast (Berger 1996).

Complete genome sequences are required for the identification of fusion events as this allows the detection of orthologous proteins across species. The underlying assumption is that if a composite protein (or fusion protein) is uniquely similar to two component proteins in another species, the component proteins are most likely to interact in some way. The domain fusion analysis makes two distinct predictions: it predicts protein pairs that have related biological functions (proteins that participate in a common metabolic pathway or structural complex) and it predicts potential protein-protein interactions (Marcotte 1999). Naturally, these two scenarios are not mutually exclusive. In an analysis on the *E. coli* genome Marcotte *et al.* found that over half of the testable predictions shared functional similarity.

1.4.4 Function from structure

The prediction of protein function through virtue of structural similarity is based on the finding that proteins adopting similar folds often have similar functions (Orengo 1999, Hegyi 1999, Koppensteiner 2000). An analysis, by Orengo *et al.* (1999), of the proteins in the CATH database suggest that structural data can allow recognition of more distant homologues compared with sequence data. In this analysis over 80% of structures with novel sequences could be assigned a putative function through 'guilt by association'. Orengo also suggests that structure could be used to predict function *ab initio* using structural features to define classes of protein. Enzymes, for example, can often be identified by the presence of a major cleft, which is frequently the location of the active site.

One such analysis by Stawiski *et al.* (2000) found that the proteases, as a group, have consistent structural similarities. These include smaller than average surface areas and smaller radii of gyration. These features were used to train an artificial neural network, which demonstrated a predictive accuracy of over 86%. It is feasible that similar structural analyses could characterise other classes of proteins and may be of use in categorising the flood of structures soon to emerge from structural genomics initiatives.

It is also possible to predict the functions of enzymes based on the spatial arrangement of catalytic residues. Wallace *et al.* (1998) developed an algorithm called TESS that automatically derives 3D motifs from PDB structures. These motifs are analogous to

those present in the PROSITE (Hofmann 1999) and PRINTS (Attwood 2000) databases. Newly elucidated structures can be scanned against these templates to identify functional sites.

A similar analysis involves mapping the conservation of amino acids on to protein surfaces in order to predict likely functional sites. In an analysis by Lichtarge *et* al. (2002), this technique was seen to identify statistically significant and functionally relevant regions in more than 80% of all proteins tested.

1.4.5 Combinatorial algorithms

In general, these approaches are most effective when used in combination. Marcotte *et al.* (1999) devised a protocol that forms links between proteins using correlated evolution, correlated mRNA expression patterns and patterns of domain fusion (Rosetta Stone analysis) to determine functional relationships among the entire protein complement of the yeast genome. These links were combined with experimentally derived protein-protein interactions from the MIPS yeast genome database and from the database of interacting proteins. Again, function is assigned through 'guilt by association'. If a protein, A, of unknown function is linked to a group of functionally related proteins, then the shared functions of these proteins provide a clue to the general function of A. This method was successful in assigning a general function, with high confidence, to 15% (374) of the 2, 557 uncharacterised proteins in the yeast *Saccharomyces cerevisiae.* They also claim that they are able to assign function, with lower confidence, to 62% (1,589) of the uncharacterised proteins. The reliability of these predictions were tested by evaluating the recovery of known protein functions by prediction. For example, the SwissProt keywords for the yeast enzyme ADE1 are 'ligase' and 'purine biosynthesis'. Based on the keyword frequencies of the 18 proteins linked to this enzyme, the general function of ADE1 was predicted to be purine biosynthesis (13.6%), lyase (13.6%) transferase (11.4%) and ligase (6.8%). ADE1 shares no sequence similarity with any of the 18 proteins with which it is linked, which demonstrates the utility of non-homology based techniques in the prediction of gene/protein function.

It is interesting to note that the ability of this technique to correctly assign function is on a par with experimental techniques when the prediction is carried out using data from multiple predictions.

1.5 Project aims

The aims of this project were two fold; Firstly, we aim to investigate the utility and limitations of current homology based techniques for the functional annotation of nucleotide sequences through the generation and analysis of a large-scale chicken EST database. Secondly, we seek to develop a non-homology based technique for the classification of gene function using transcription factor binding sites.

The initial aims of this project centred on the investigation of non-homology based techniques in the prediction of gene function. During the course of this investigation, funding was acquired for a large-scale chicken EST project and I was appointed as the primary bioinformatician. Very little genomic information was available for *Gallus gallus* at the inception of this project, and the generation of this data represented a significant platform for advancing the understanding of a novel vertebrate genome. It also represented an opportunity to demonstrate the application of classic bioinformatic techniques and to further investigate the limitations of homology based gene function annotation. Since this is an EST based project we would expect to find only coding sequence. These sequences have a lower fidelity than say a 10x genome sequence, which has the potential to cause annotation problems requiring the intelligent use of the full range of BLAST tools.

The non-homology based protein function prediction investigation represents a counterpoint to this study for classic homology-based approaches. We concentrate on the yeast *Saccharomyces cerevisiae*, which was the first eukaryote to have its genome completely sequenced. This is a very well characterised organism, and is perhaps the best characterised in terms of regulatory sites despite being relatively primitive (in comparison to higher eukaryotes such as *Homo sapiens*). Hence, given the large range of functional data available, it makes an obvious choice for this study.

Predicting gene function using homology - an informatic analysis of Chicken ESTs.

2.1 Introduction to EST libraries

Shortly after the inception of the Human Genome Project in 1990, Craig Venter developed a novel method for gene discovery. Molecules called Expressed Sequence Tags (ESTs) offered an efficient way to discover genes and explore their function.

When a gene is expressed, the bases encoding that gene are transcribed into messenger RNA (mRNA). These molecules serve as templates for protein synthesis. Messenger RNA molecules are both fragile and transient in the cell, but they can be translated into more robust complementary DNA (cDNA). By 1990, researchers had developed the techniques necessary to create libraries of cDNA molecules. Expressed sequence tags (ESTs) are short DNA sequences (usually between 200 and 600 bases in length) generated by sequencing from the 3' and 5' ends of randomly selected cDNA clones. Figure 2.1 gives an overview of EST library construction.

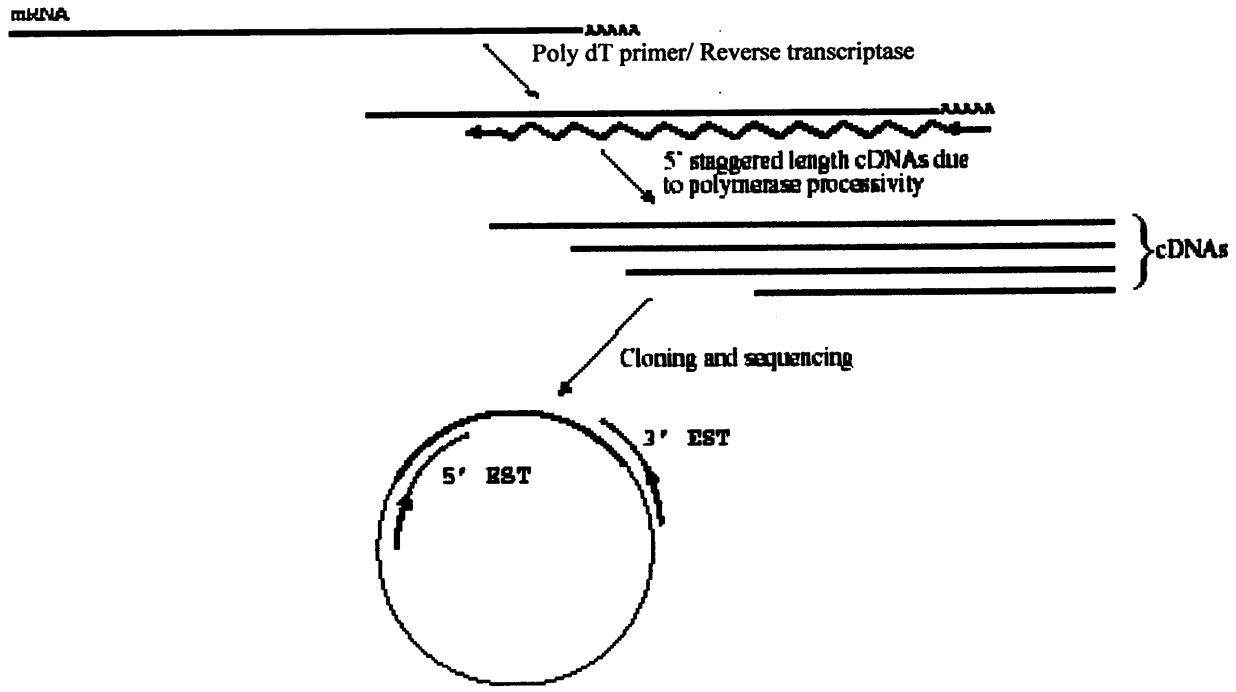


Figure 2.1. Overview of EST library construction

A gene can often encode the information necessary to produce multiple proteins. The genomic DNA is processed by the cell into a primary messenger RNA (mRNA) transcript, which undergoes splicing. This functional mRNA is exported to the cytosol and in turn translated into a protein sequence. The primary mRNA can be processed in multiple ways to produce a number of different transcripts. Figure 2.2 shows a number of different mRNA transcripts that have been characterised from the rat α-tropomyosin gene. Different exons (shown as different colour boxes) can be either included or excluded in the final, spliced form of the processed mRNA.

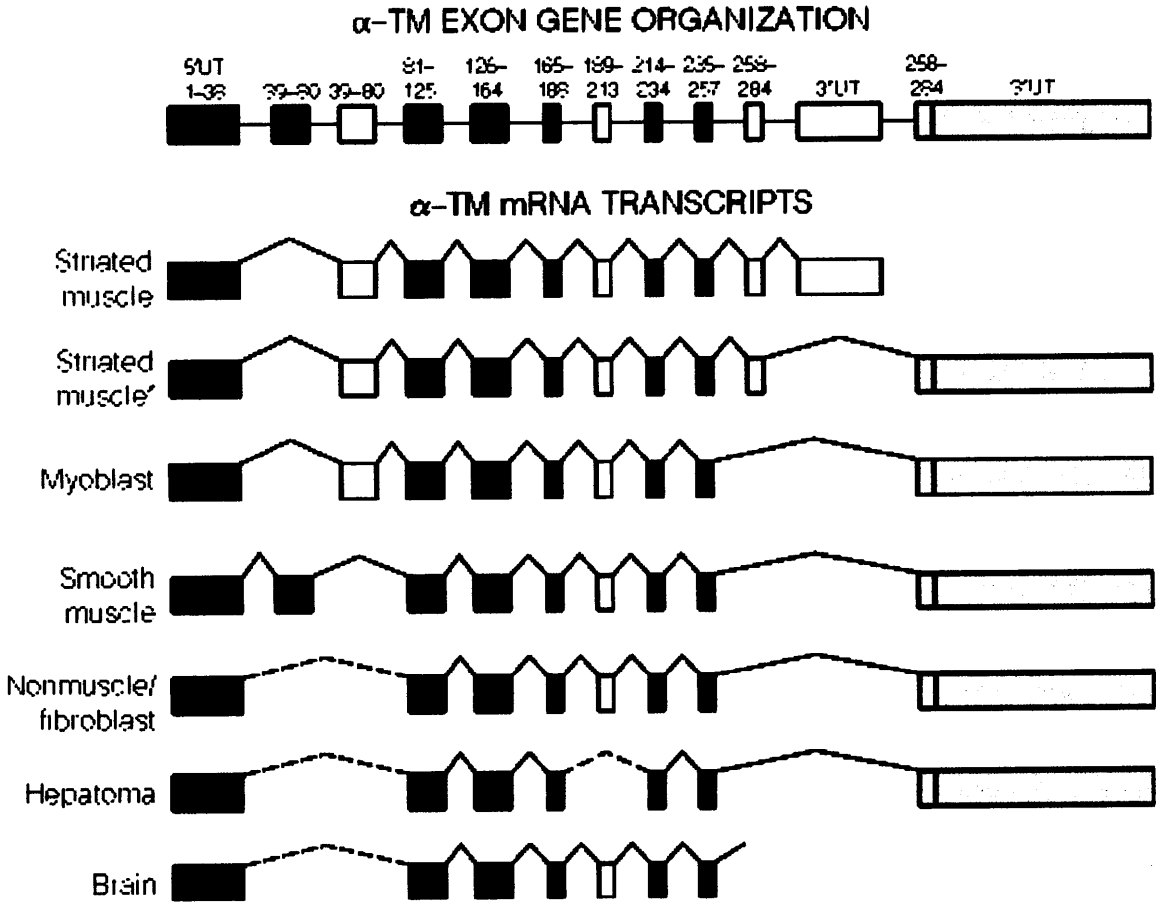


Figure 2.2. Alternate splicing of rat α-tropomyosin (taken from web ref 4).
Different coloured boxes represent different classes of exon. Red = constitutively expressed, green = smooth muscle specific, yellow = striated muscle specific and white = variable. Solid interconnecting lines represent experimentally determined splicing pathways, dotted lines represent pathways inferred from nuclease protection mapping. UT signifies untranslated regions.

It was proposed that by sequencing a cross section of mRNA molecules from a tissue or cell type, the complement of active genes in that sample could be determined. This approach of producing ESTs as a large-scale analysis tool was first used in the early 90's to characterise gene expression in the human brain (Adams 1991). Over 600 ESTs

were sequenced and compared with the complement of known genes at the time. Of these, 230 ESTs represented previously undiscovered and uncharacterised genes. This showed that EST sequencing could be used to rapidly identify new genes. It was also shown that, by mapping ESTs to chromosomes, ESTs could be used to search for families of genes and genes implicated in heritable disorders.

Large-scale EST sequencing was readily taken up by the private sector. This led to the rapid growth of proprietary data collections that surpassed, many fold, the data available in the public databases (Boguski 1995). The funding of multiple large-scale EST projects by the public sector and the deposition of this data into the public databases has addressed this imbalance. As of May 2003 there are 16,547,527 EST sequences deposited in dbEST (Boguski 1993) from 446 species. Table 2.1 gives an overview of the spread of ESTs in dbEST.

Table 2.1. Cross section of organisms represented in dbEST.

Organism	Number of ESTs in dbEST
Homo sapiens (human)	5,094,900
Mus musculus + *domesticus* (mouse)	3,721,428
Rattus sp. (rat)	525,545
Gallus gallus (chicken)	418,093
Drosophila melanogaster (fruit fly)	261,271
Caenorhabditis elegans (nematode)	192,132
Arabidopsis thaliana (thale cress)	178,538
Plasmodium falciparum (malaria parasite)	20,176
Saccharomyces cerevisiae (baker's yeast)	3,041

Data ranked by the number of EST sequences deposited (as of May 2nd 2003).

The rate of novel sequence discovery and the quality of this data depends heavily on the cDNA libraries used. The mRNA of the most abundant and medium abundance classes comprise as much as 50-65% of the total mRNA of a cell but only represent a small fraction of the total number of different mRNAs present in that cell. It is apparent that a random gene discovery strategy (as are all EST programs) will soon be overwhelmed by redundant sequences, which seriously compromises cost effectiveness. To overcome this limitation a number of normalisation procedures have been developed which aim to reduce the numbers of abundant transcripts in cDNA libraries (Ko 1990, Patanjali 1991, Bonaldo 1996). The most common method of normalisation depends on the

observation that if cDNA reannealing follows second-order kinetics, rarer species anneal less rapidly and the single-stranded fraction of cDNA becomes progressively more normalised during the course of hybridisation (Galau 1977). Although the kinetics of the reassociation reactions are more complex (Britten 1974), this technique allows the construction of libraries in which a rare mRNA becomes nearly as abundantly represented as is the most abundant mRNA. Another method for increasing the number of novel sequences gained from a cDNA library is that of subtractive hybridisation (or just subtraction). Subtraction enriches for gene transcripts present in one population of mRNAs and absent from another. It also removes all transcripts expressed equally in both populations. This has proven to be a powerful method to isolate differentially expressed genes (Klar 1992, Rissoan 2002, Tominaga 2002).

As well as being a powerful tool for the rapid discovery of genes, EST databases can be used for expression profiling (using non-normalised libraries), *in silico* cross-tissue differential expression analyses (e.g. Rajkovic 2002), electronic northern analysis (web ref 5, Schmitt 1999), gene number estimates in a genome (for vastly varying estimates see Fields 1994, Liang 2000a and Ewing 2000) and are a valuable resource for annotating genes during a genome project (Lander 2001, Gardner 2002).

2.2 Informatics for EST analysis

Any large-scale EST sequencing project involves a number of informatic stages.

1. Base calling from sequence traces
2. Removal of vector from EST sequence
3. Identification and removal of contaminants (rRNA, mtRNA, mtDNA, host sequence contamination)
4. Clustering and Assembly of ESTs into contigs (contiguous sequences)
5. Annotation of sequences (both contigs and individual ESTs)
6. Data storage and distribution

These stages are considered in more detail here, with emphasis on the more common tools used to carry them out. The next section will include a more in-depth analysis on the use of a number of these tools.

2.2.1 Sequence generation

In a large-scale EST project, sequences are generally produced using automated DNA sequencers. These generate sequence trace files or chromatograms (Figure 2.3) that must be interpreted by a base-calling program in order to generate sequence data.

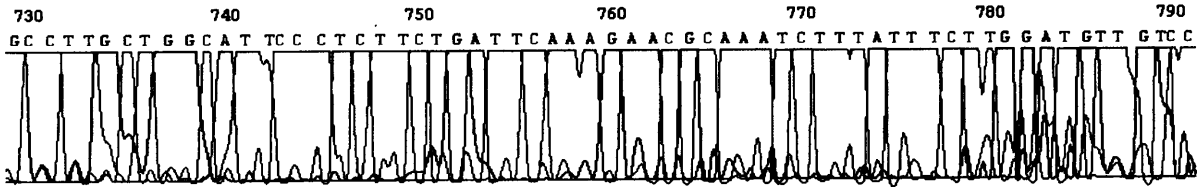


Figure 2.3. A portion of the sequence trace file for EST 602818629 from the BBSRC Chicken EST project (Boardman 2002).

The most popular base-calling program in common use is phred (Ewing 1998a, Ewing 1998b). Phred reads DNA sequencer trace data, calls bases, assigns quality values to the bases and writes the base calls and quality values to separate output files. Phred can read trace files generated by the majority of automated DNA sequencers and produce output files in multiple formats (FASTA format being the most popular example).

Phred uses simple Fourier methods to examine the four base traces in the surrounding region for each point in the dataset. In this way a series of evenly spaced locations are predicted in a way that corrects for factors that shift the peaks from their "true" locations (e.g. compressions, dropouts). Next phred examines the centres of the actual peaks and the area of these peaks relative to their neighbours. Phred then evaluates the trace surrounding each called base using multiple quality value parameters. The result is a quality value that is related to the base call error probability by the following formula (where P_e is the probability that the base call is an error): -

$$Qv = -10 \log_{10} (P_e)$$

Equation 2.1. Phred quality score calculation

Quality scores range from 4 to about 60, with higher values corresponding to higher quality. Table 2.2 shows how these scores are linked logarithmically to error probabilities.

Table 2.2. Correlation between phred quality scores and the associated base call probability/accuracy.

Phred quality score	Probability that the base is called incorrectly	Accuracy of the base call
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

These highly accurate, base-specific quality scores can be used to assess sequence quality, to compare different sequencing methods (Richterich 1998) and to generate better assemblies when used in conjunction with Phrap (web ref 6, Green 1996), which is discussed later (Section 2.2.3).

2.2.2 Vector clipping and decontamination

Both vector clipping and decontamination rely on identification of pairwise similarity with known sequences. The most widely used programs for sequence similarity searches are the BLAST (Altschul 1990, Altschul 1997) and FASTA (Pearson 1988, Pearson 2000) suite of programs (described previously in section 1.3).

2.2.2.1 Vector clipping

Due to the nature of sequencing, the 3' and 5' regions of an EST sequence may be vector sequence rather than insert sequence (Figure 2.4). Indeed, in a 5' sequencing project the presence of 5' vector sequence is good evidence that correct cloning has occurred and is an additional factor in quality control. However, the presence of vector sequence presents problems during the processes of assembly and annotation. This can be rectified by comparing the EST sequence with the vector sequence and subsequently removing any matching regions. There are a number of publicly available programs that carry out vector clipping, though many groups decide to create their own. For example, the Staden suite of programs (Staden 2000, web ref 7) contains the vector-clipping program ***Vector_clip***, which can compare batches of sequences against a vector to locate and mark contamination sites. The standard installation of phrap comes with a program, called ***cross_match***, which generates a set of vector-masked versions of the input sequences (Green 1996).

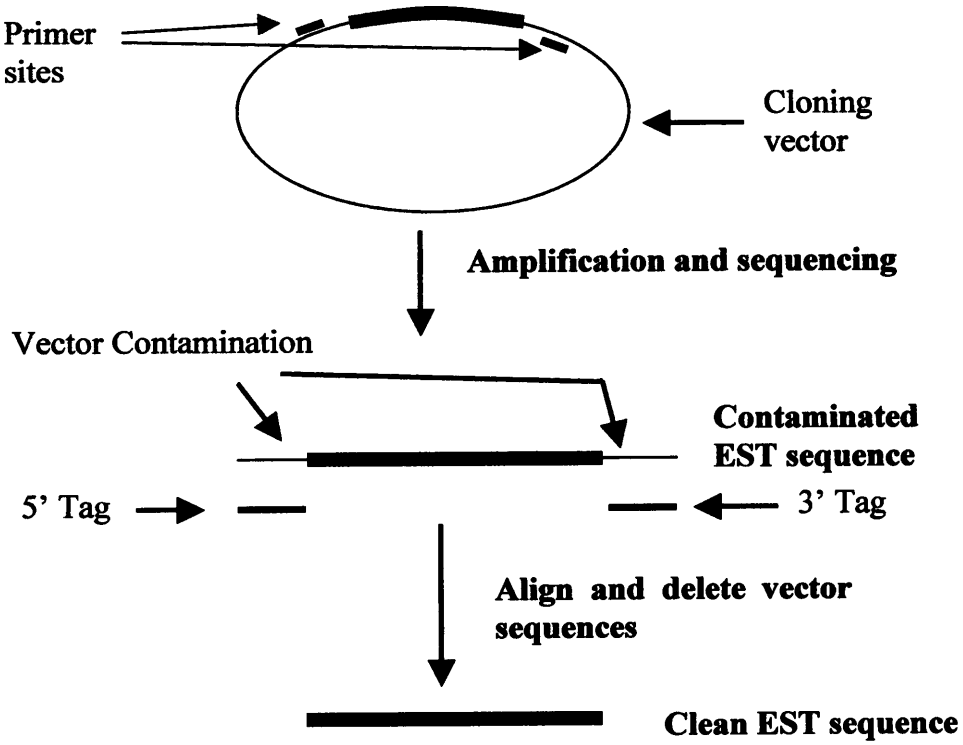


Figure 2.4. Vector clipping schematic.
Orange bars represent cDNA insert sequence.

2.2.2.2. Decontamination

As the EST data is produced, contaminants need to be removed from the sequences. These contaminants include, vector sequence, mitochondrial sequences (both DNA and RNA), ribosomal RNA (rRNA) and DNA/RNA sequences derived from the host in which the vector was cloned. As with the vector clipping stage, contaminants are identified by comparing the EST sequence with databases of known mitochondrial and rRNA sequences from the originating organism as well as databases containing all known sequences from the cloning host. Those sequences that are found to match in their entirety, or over a large portion of the EST, are flagged as contaminants and are usually removed from the EST database.

2.2.3 Clustering and Assembly

Though a database of ESTs alone is a valuable resource for comparative and functional genomics, it is possible to produce data with a higher information content and quality by grouping ESTs that originate from the same gene/mRNA transcript. There are two major, complementary, protocols used to do this; clustering and assembly. The intelligent use of these techniques allows reduction in redundancy, detection of splice

variants, creation of longer and higher quality sequences (in the case of assembly) and can reduce the number of sequences required to be analysed later on.

2.2.3.1 Clustering

Clustering is the process of partitioning a set of elements into meaningful groups so that members of each group are more similar to each other than to members of any other group.

The fastest and easiest way to cluster a set of ESTs is through virtue of shared sequence identity. There are many programs available that cluster together sequences by identification of similar, overlapping regions (e.g. uicluster2 (web ref 8), BlastClust (Altschul 1990), d2_cluster (Burke 1999)), but many groups create their own clustering pipelines based on publicly available alignment tools (e.g. the BLAST or FASTA algorithms).

For example, the UniGene database (Boguski 1995, web ref 9) partitions sequences into a non-redundant set of gene-orientated clusters using GenBank coding sequences and mRNAs as "seed" sequences for the clusters. Each UniGene cluster contains sequences that represent a unique gene. In this way closely related transcripts and alternatively spliced transcripts are partitioned into the same set. The UniGene procedure uses the megablast program (a member of the BLAST suite) to add sequences to a cluster.

2.2.3.2 Assembly

Clustering ESTs brings together those sequences that originate from the same gene or share a significant amount of similar, overlapping, sequence. Assembly goes one stage further by producing consensus sequences (termed contigs, template sequences or tentative consensus sequences) that may represent the gene or mRNA transcript of origin (Figure 2.5). This has several advantages: it separates splice variants, it separates closely related genes (that couldn't be separated by clustering) into distinct consensus sequences and it produces longer representations of the underlying gene sequences (Quackenbush 2001).

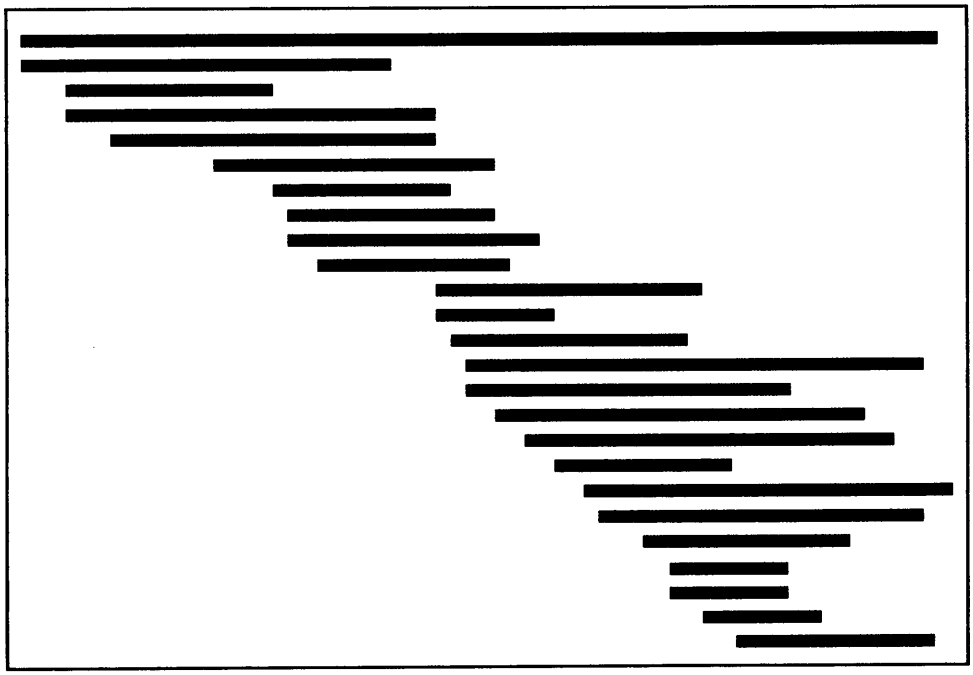


Figure 2.5. Assembling a consensus sequence.
The blue bar represents the assembled contig sequence. The red bars represent the component EST sequences used to generate the contig.

Of the numerous assembly programs available, the most widely used are CAP3 (Huang 1999), Phrap (Green 1996) and the TIGR Assembler (Sutton 1995). The quality and utility of the assembled sequences relies on the ability of these programs to effectively generate high quality consensus sequences from the EST data. An analysis by Liang *et al.* (2000) into the fidelity of the assemblies of these programs showed that CAP3 consistently provided the highest fidelity assemblies when analysing ESTs data with no quality scores. Phrap came in a close second on the majority of tests performed though Phrap tends to insert bases when there is a discrepancy in the component sequences. The construction of a consensus sequence by Phrap relies heavily on the availability of quality scores of the sequences being assembled, derived by programs such as the partner software tool, phred. When these scores are included in the assembly Phrap may outperform other assembly programs. Indeed, in our initial assembly assessment we note that Phrap appears to perform as well as (if not better than) the other assembly programs.

An example of a resource that uses an assembly strategy is the TIGR gene indices (Quackenbush 2001). These are a collection of species-specific databases, produced by analysing EST sequences in an attempt to identify all unique genes in the dataset. Gene

Indices are constructed by first clustering and then assembling EST and annotated gene sequences from GenBank for each targeted species. The clustering stage groups sequences sharing a minimum of 95% identity over a 40 nt region with less than 20 bases of mismatched sequence at either end. This stage uses the FLAST sequence comparison program, which is based on DDS (Huang 1997). Each cluster is then assembled separately using CAP3 to produce consensus sequences. Assembly produces one or more consensus sequences for each cluster and rejects any chimeric, low-quality and non-overlapping sequences. These assembled sequences can then be annotated to provide a provisional functional assignment.

2.2.4 Functional annotation

After the generation of 'clean' EST sequences and consensus sequences, the next stage is often to assign putative functions by comparing each sequence against a selection of the publicly available sequence databases (e.g. SwissProt, TrEMBL, NRDB, PIR, EMBL). If the sequence (EST/contig) shares sufficient similarity with a member of the database then it is assigned the same function as the matching sequence. By comparing the nucleotide sequences against a protein rather than a nucleotide database there is a higher likelihood of finding a biologically significant match. Protein database searching is between two and five times more sensitive than DNA database searching. This is due to a number of reasons:

1. The higher degree of conservation of protein sequence in comparison to nucleotide sequence.

2. The larger alphabet of protein sequences - since the DNA alphabet is only composed of four characters, you would expect a random pairwise alignment to show 25% sequence identity. The sensitivity of comparison is greatly improved for proteins, which have an alphabet of 20 characters.

3. The scoring matrices used – DNA comparisons are usually carried out using identity matrices. More sensitive matrices such as the PAM and BLOSUM matrices are available for protein comparisons.

2.2.5 Data storage

It is essential that any large-scale sequencing and analysis project have the capacity for intelligent data storage. This data storage system should allow for effective modes of access for data querying. Often sequences are stored and distributed as flat files but as the amount of data increases it soon becomes necessary to move onto more easily queryable systems of storage such as relational databases, object-orientated databases or a combination of these systems.

A major choice to be made is that of public availability of the data. For the most effective distribution of data, sequences are deposited in the most relevant public database or databases. For large amounts of EST data the database of choice is dbEST (Boguski 1993, described in section 2.1), which is a division of the GenBank sequence database (Benson 2002). The submission systems for other databases are not as capable of handling large numbers of sequences. GenBank maintains regular data exchange with the EMBL and DDBJ databases to ensure comprehensive worldwide coverage and consequently, data deposited in GenBank is rapidly represented in the other major databases

Submission of sequences into the public databases ensures the widest dissemination of sequence data. It can also be desirable to provide additional "value added" information by producing a specialist distribution or a website allowing users to specifically mine the EST sequences and any additional data (e.g. assembled sequences). For example, the FANTOM website (web ref 10) allows users to mine the functional annotations and other functional information for the RIKEN full-length cDNA clones (Hayashizaki 2001).

2.3 Design and implementation of informatic pipeline for the Chicken EST project

2.3.1 The Chicken EST project

The chicken (*Gallus gallus)* represents an important experimental system that is used by a diverse group of scientists, ranging from developmental biologists, immunologists, through to molecular biologists and geneticists. Two special features of chicken biology offer outstanding opportunities for investigating gene function. Firstly, chicken embryos are well characterised and their developmental mechanisms are easy to access and analyse. Secondly, there exists a chicken cell line, DT40, whose genome can be genetically modified as efficiently as that of yeast. This is due to its high rate of homologous recombination (Buerstedde 1991) and a targeted to random DNA integration ratio of more than 1:2, which exceeds that of any mammalian cell line (Winding 2001).

The major obstacle for groups working with chicken systems has been the limited knowledge of the chicken genome and the time and effort that has to be expended in isolating chicken orthologues identified in other species. Funding was granted by the BBSRC for the production of a comprehensive chicken EST database in order to promote the use of *Gallus gallus* as a model organism for functional genomics.

2.3.1.1 Overview

The focus of this sequencing project was gene discovery. This dictated the choices in sequencing strategy and tissue selection for cDNA library construction. Ongoing research and existing EST data generated from other projects (Abdrakhmanov 2000, Tirunaguru 2000) were also taken into account. Figure 2.6 shows the range of adult and embryonic tissues that were selected for sequencing.

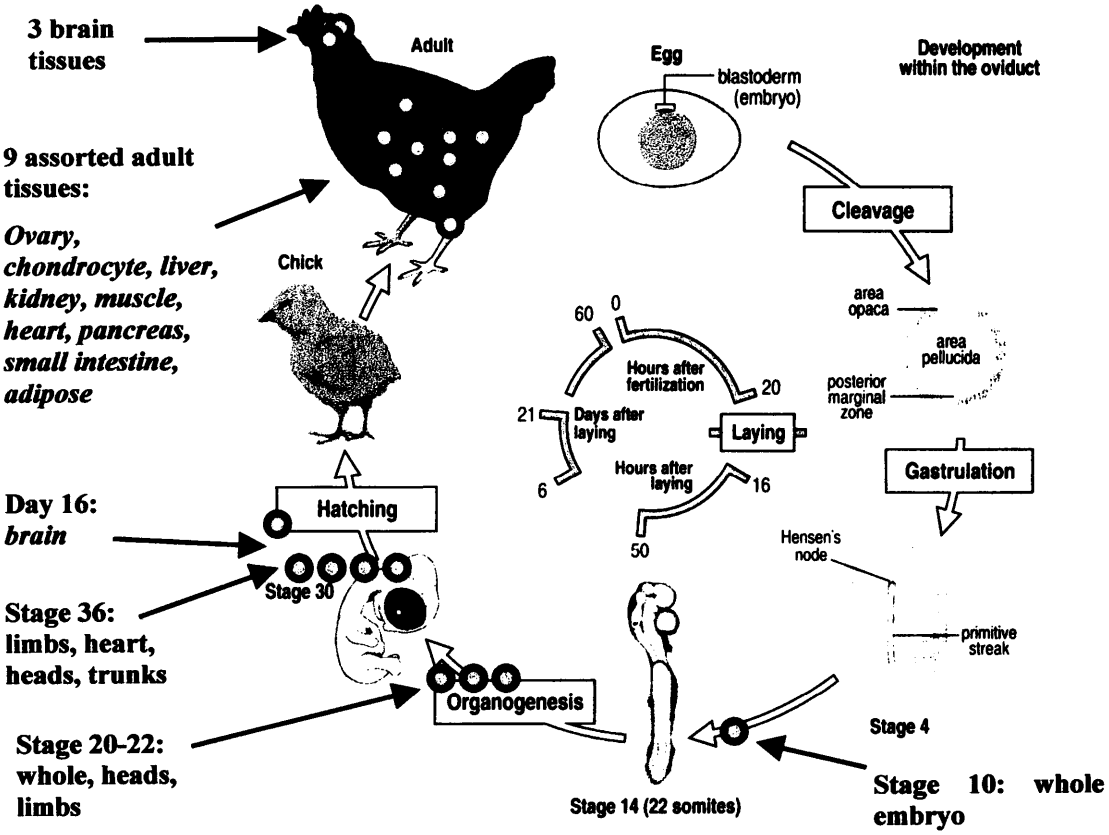


Figure 2.6. Overview of tissues selected for sequencing.
Each selected tissue is represented by an orange circle placed at the appropriate stage in the chick life cycle.

Funding was granted to sequence 350,000 ESTs from 21 different tissues. Library construction and sequencing was outsourced to Incyte Genomics (web ref 11) who provided raw sequences and chromatograms via regular ftp updates and CD shipments throughout the duration of the project.

For each tissue, one standard (Fu 2002) and at least two normalised libraries were constructed. We used Incyte's proprietary "rare-clone biased" normalisation protocol which is a refinement of the original Soares normalisation procedure (Bonaldo 1996). This new normalisation protocol is capable of producing libraries in which 75-95% of the clones are derived from the rare transcript population (Figure 2.7), which is ideal for a gene discovery motivated project.

Normally, Incyte provide a post sequencing bioinformatics service and supply the final data to the customer after all sequencing has been completed. In order to ensure cost effectiveness and maximise rates of gene discovery, it was decided that sequences

would be analysed as they were produced and further sequencing reactions would be allocated to libraries exhibiting high levels of novel sequences.

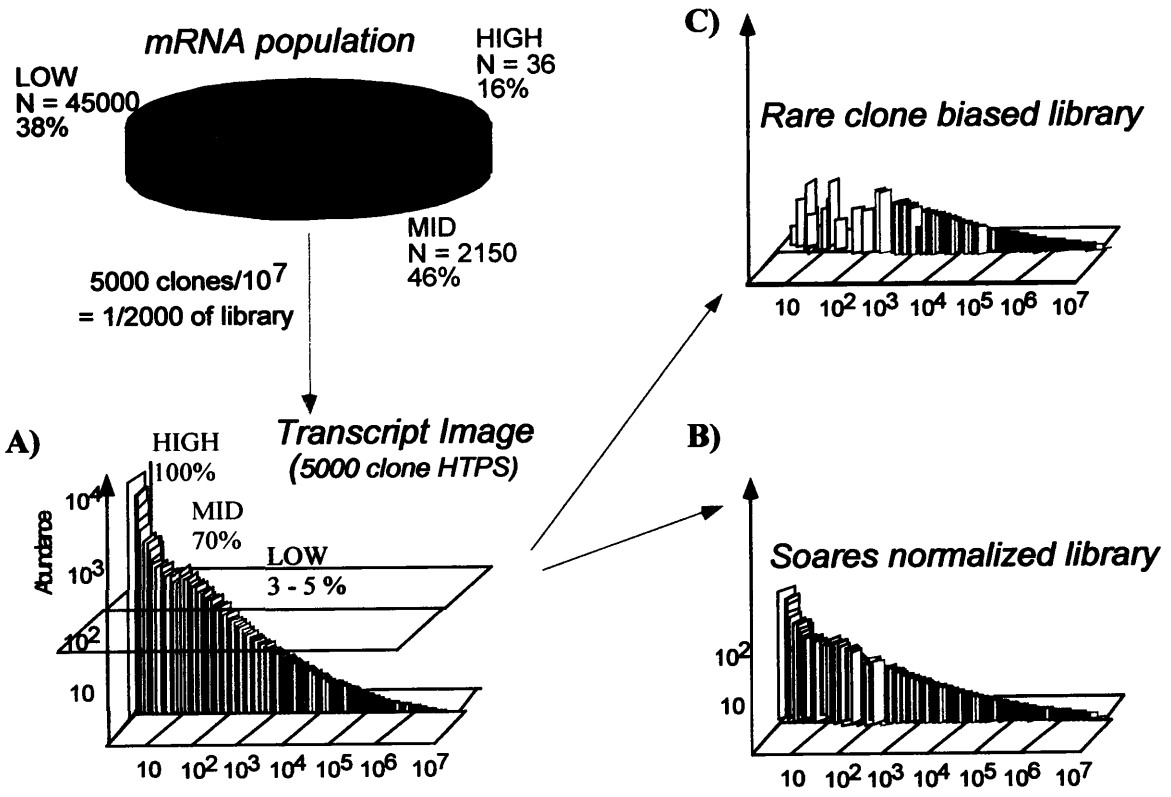


Figure 2.7. Comparison of Incyte's "rare clone biased" normalisation with Soares normalisation.

A) A typical sequence of 5,000 clones from a standard library would cover ~100% of the abundant transcripts, 70% of the mid abundant transcripts and 3-5% of the low abundant transcripts.

B) The Soares normalisation method compresses the abundance distribution of the three transcript categories preserving the original transcript distribution.

C) The Rare Clone Biased procedure produces a library in which the rare clones are enriched at the expense of mid and high abundance transcripts.

2.3.1.2 Hardware/software requirements

A 20-node linux cluster and a 300Gb RAID device were purchased to provide the processing power and storage capacity necessary to effectively analyse the large numbers of ESTs being produced and to ensure this was possible within a reasonable time-scale. The following programs were installed on the linux cluster (Rocky) and used during this project: BLAST, FASTA, ClustalW, uicluster2, Perl, C, Apache, Phred, Phrap, TIGR Assembler, MySQL, InterPro and numerous smaller scripts (described later). All programs written during the course of this project are available on the CD bound with this manuscript, or on request from the author.

2.3.2 Informatic pipeline

2.3.2.1 Overview

The general informatics necessary for an EST project were covered previously in section 2.2. It was essential to develop an automated pipeline for vector clipping and decontamination of the ESTs as they were deposited by Incyte onto our ftp site. We also added a basic annotation to the decontaminated sequences. These were then deposited into our database. Figure 2.8 shows a schematic of the basic informatic pipeline used in this EST project.

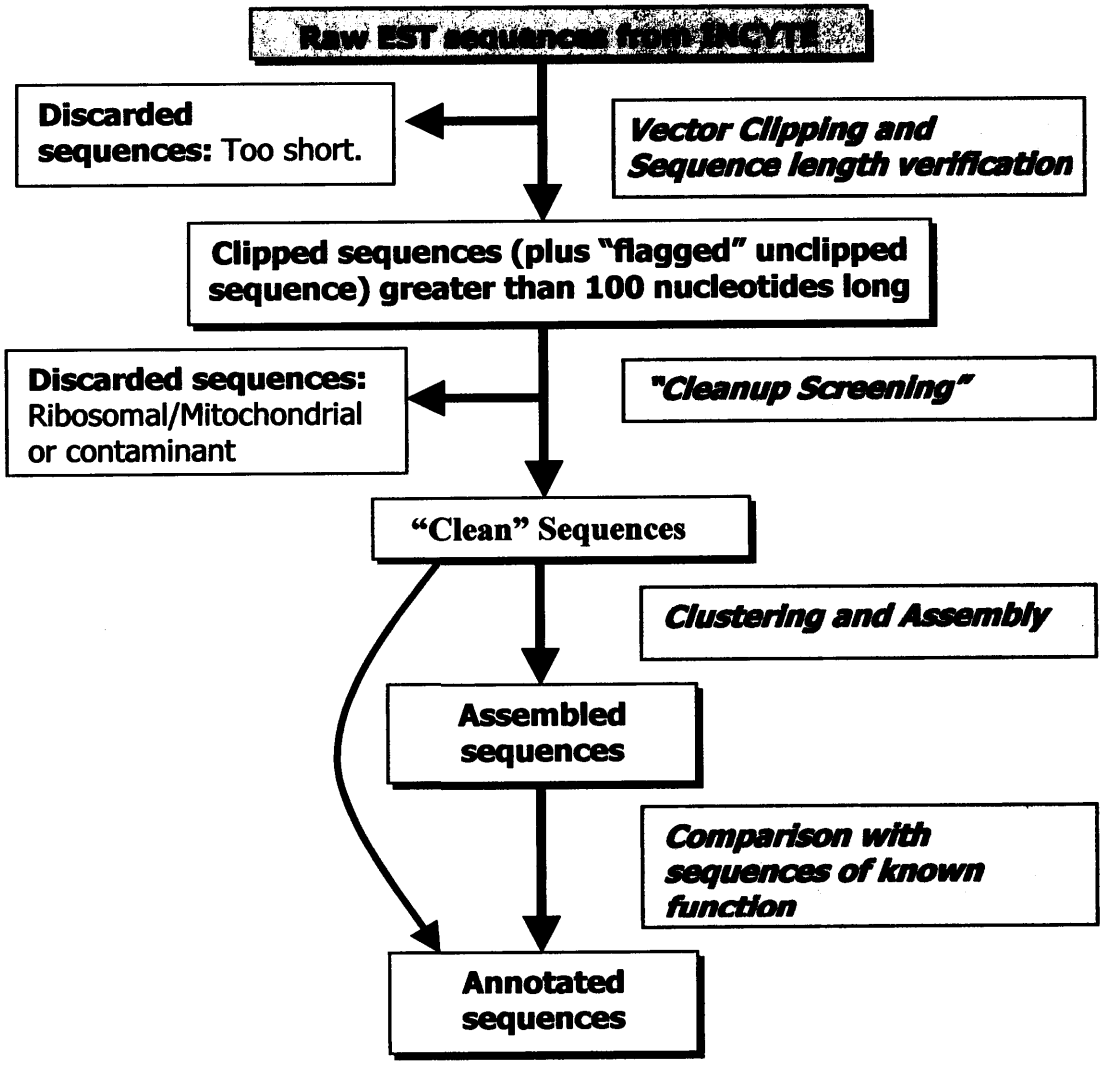


Figure 2.8. Primary informatics pipeline.

2.3.2.2 Distributed programming

It was necessary to develop a number of programs to exploit the computational power provided by the linux cluster. The most desirable function was the ability to distribute processes evenly between the separate nodes. It is essential that this process does not overload an individual node and that waiting jobs are immediately 'farmed out' as soon as nodes become available. Satellite programs can then be developed which exploit this architecture.

The client-server paradigm seemed the most fitting for this purpose. Although some publicly available software for distributing jobs on the linux cluster was initially tested, we opted to develop our own tools, limited to applications for this project. In this client-server setup a server program is set running constantly on a host machine (in our case each node of the cluster) awaiting commands. The client program connects to a server and issues requests/commands. Two Perl scripts were written to achieve this, ***client.pl*** and ***job_server***.

Client.pl

This program was designed to process either a single command or a list of commands and execute them by farming them out to available nodes in the linux cluster. The program flow for ***client.pl*** is detailed in Figure 2.9. Since the user could potentially execute 40 processes or more simultaneously on the cluster it was necessary to redirect any errors generated by these processes to a log file. This log file is identifiable by the username of the invoking user and the time of execution. A further feature of this program is the generation of a list of nodes that are accepting/refusing jobs by supplying the single command "query_servers".

Job_server

This program continually runs on all of the nodes of the linux cluster awaiting a connection from the ***client.pl*** program. Since each node houses two CPUs, the ***job_server*** will only accept two processes at any time. On accepting a command, the ***job_server*** spawns a child process to manage the process and to maintain communication with the client. This frees the parent process for further communication and allows multiple, simultaneous, connections.

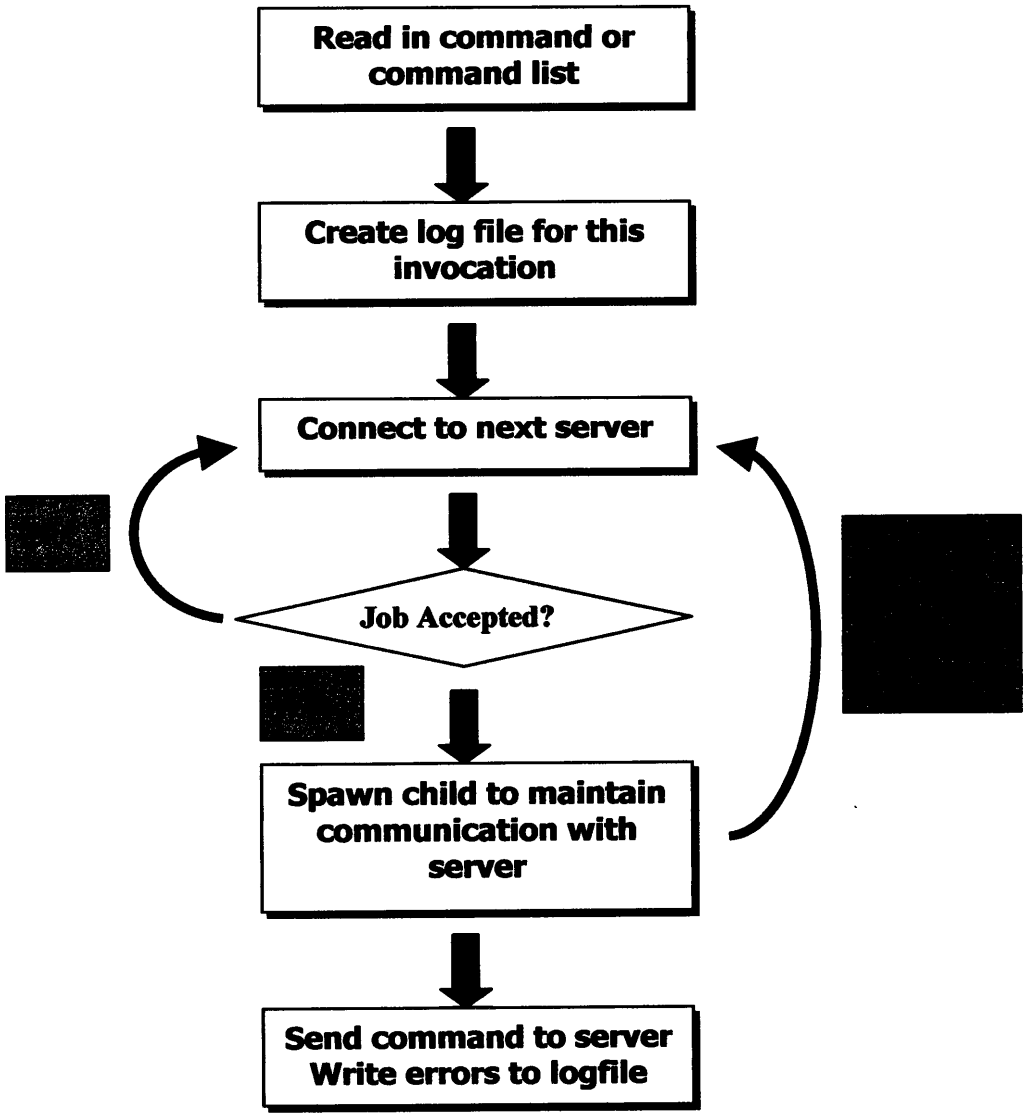


Figure 2.9. Program flow of *client.pl*.

Rocky_top

A multi-threaded program to display the current CPU load and memory usage of each node was created. This was to enable real-time monitoring of the status of the linux cluster. Figure 2.10 shows screenshots from this program.

A number of other programs were created to further exploit the functionality of this architecture. These will be described in more detail later.

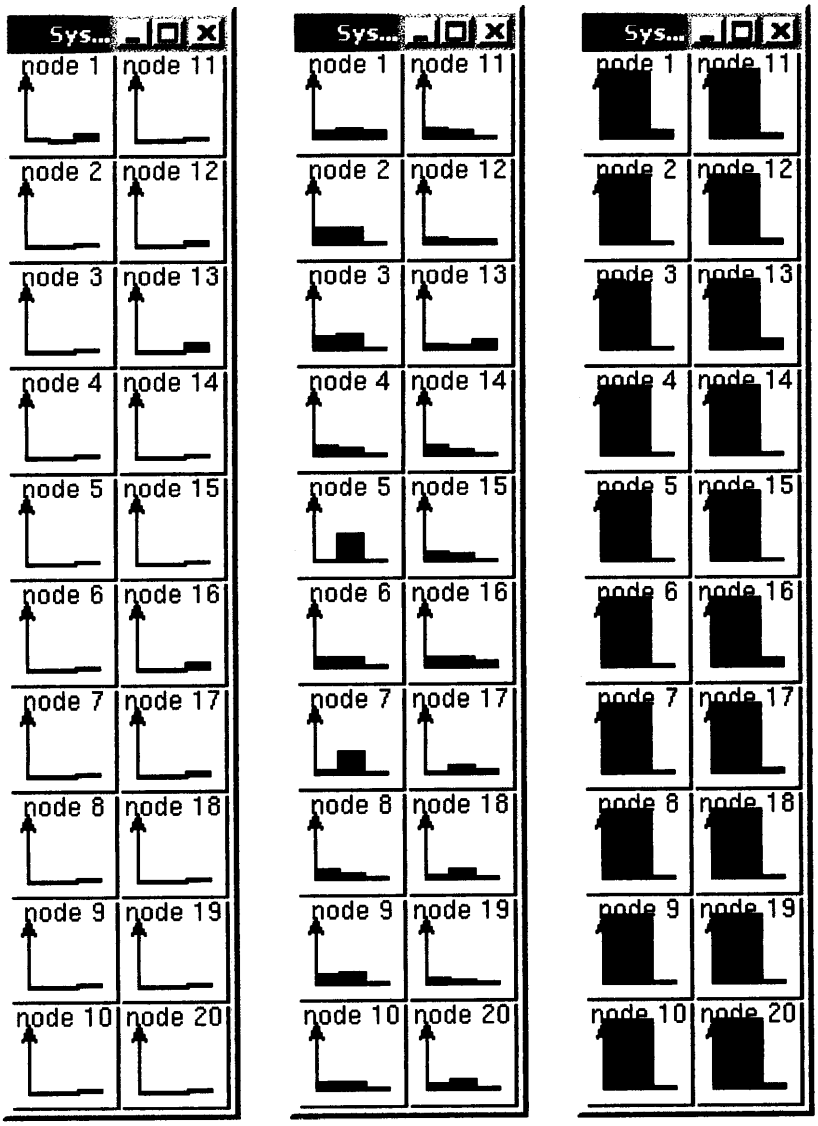


Figure 2.10. Screenshots of the *rocky_top* program.

These show rocky (the linux cluster) during times of low usage, moderate usage and heavy usage (from left to right). Blue and green bars represent individual CPU usage. Red bars represent memory usage.

2.3.2.3 Vector clipping

The vector clipping protocol was designed using the ssearch3 program, which is a member of the FASTA suite. This is an implementation of the Smith-Waterman local alignment algorithm and is more sensitive than BLASTN for nucleotide alignments. Any vector clipping protocol must account for the regions of low fidelity found at the beginning and end of sequence reads. The vector sequence to be clipped is usually located in these regions (Figure 2.4) and consequently, exact matches between the vector and EST sequences are unlikely. Two 40 nt tags (one for the 5' flanking region and one for the 3' flanking region) are aligned with the EST sequence. If either tag

matches with a sequence identity of 70% or more (or an alignment score of 150 or more) within 200 bases of their respective ends, the position is noted and the sequence is trimmed. Since all ESTs were sequenced from the 5' end, we expect to find vector sequence at the start of the sequence. If the 40 nt tag does not match a new search is instigated with a shorter tag of 20 nt. If this shorter sequence is not found the EST sequence is flagged as potentially erroneous. Clipped sequences continue through the pipeline if they have a length of 100bp or greater.

Table 1.3. Summary of vector clipping logic.

Expected Vector Position	Max Distance from 5' end	% identity	Alignment Score	Clip sequence?
5 Prime	80 nucleotides	>= 70	Not considered	YES
5 Prime	200 nucleotides	>= 80	Not considered	YES
5 Prime	80 nucleotides	<= 70	>149.9	YES
5 Prime	200 nucleotides	<= 80	>149.9	YES
5 Prime	80 nucleotides	<= 70	<150	NO
5 Prime	200 nucleotides	<= 80	<150	NO
3 Prime	Not considered	>= 80	Not considered	YES
3 Prime	Not considered	>= 70	>149.99	YES
3 Prime	Not considered	>= 70	<150	NO

Where a clipping event for an alignment is indicated, the EST sequence encompassing the alignment and all sequence 3' (for a 3' tag alignment) or 5' (for a 5' tag alignment) of the alignment is discarded.

2.3.2.4 Decontamination

The decontamination stage is carried out by searching vector-clipped ESTs against chicken mitochondrial DNA sequence, all known chicken ribosomal RNA (rRNA) sequences, and vector sequence. For mtDNA and rRNA searches, sequences are flagged as contaminants if a match is found with a threshold e-value of $1E^{-10}$ over a minimum match length of 50 bases for rRNA hits and 200 bases for mtDNA matches. A comparison with the complete vector sequence is used to remove erroneous ESTs which are composed solely of vector sequence.

2.3.2.5 Clustering and Assembly

ESTs that are potentially derived from the same gene were clustered together into gene-bins prior to assembly. A gene-bin represents all sequences that may have been derived from either the same mRNA transcript or from the same gene. Contigs are produced through assembly of the sequences in individual gene-bins, which may result in the generation of multiple contigs due to differences caused by SNPs, alternate splicing and sequencing errors. An initial clustering and assembly of 300,000 sequences was carried out by Incyte Genomics. The final 30,000 sequences were later added to this assembly at UMIST.

This final stage was carried out by assigning ESTs to existing, Incyte-defined, gene-bins through virtue of sequence identity with Incyte generated contigs. Assembly was carried out on all gene-bins with new members. ESTs with no match to Incyte contigs were clustered using a separate protocol. A BlastN comparison of the outlying ESTs is carried out to identify sequence relationships. The program, ***assign_gene_bins_from_blastm9.pl,*** was created to assign ESTs to gene-bins based on these blast results. Figure 2.11 shows a schematic of the recursive algorithm used for gene-bin assignment. Two EST sequences are placed in the same gene-bin if they share 97 percent sequence identity over a minimum of 40 bases and they share no more than 20 bases of unaligned, overlapping sequence. The inclusion of a "maximum unaligned overlapping sequence" parameter results in the assignment of the majority of splice variants (from a single gene) into separate gene-bins. This parameter can be relaxed to assign more or all of the splice variants from a gene into the same gene-bin.

The gene-bins (Incyte and UMIST) were then assembled using Phrap. A number of test runs using different parameters showed greatest convergence with the Incyte assembly when using default parameters.

The program ***phrap_for_remote_run.pl*** was created to run Phrap on the linux cluster. This program was optimised for use with the ***client.pl*** program (described in section 2.3.2.2). All files required for assembling a single gene-bin are automatically copied to the local node carrying out the assembly. This step is designed to reduce the IO burden on the head node of the cluster. Phrap is an IO intensive program and it was observed

that without moving the data to local nodes, 40 instances of Phrap running in parallel greatly reduce the efficiency of the cluster.

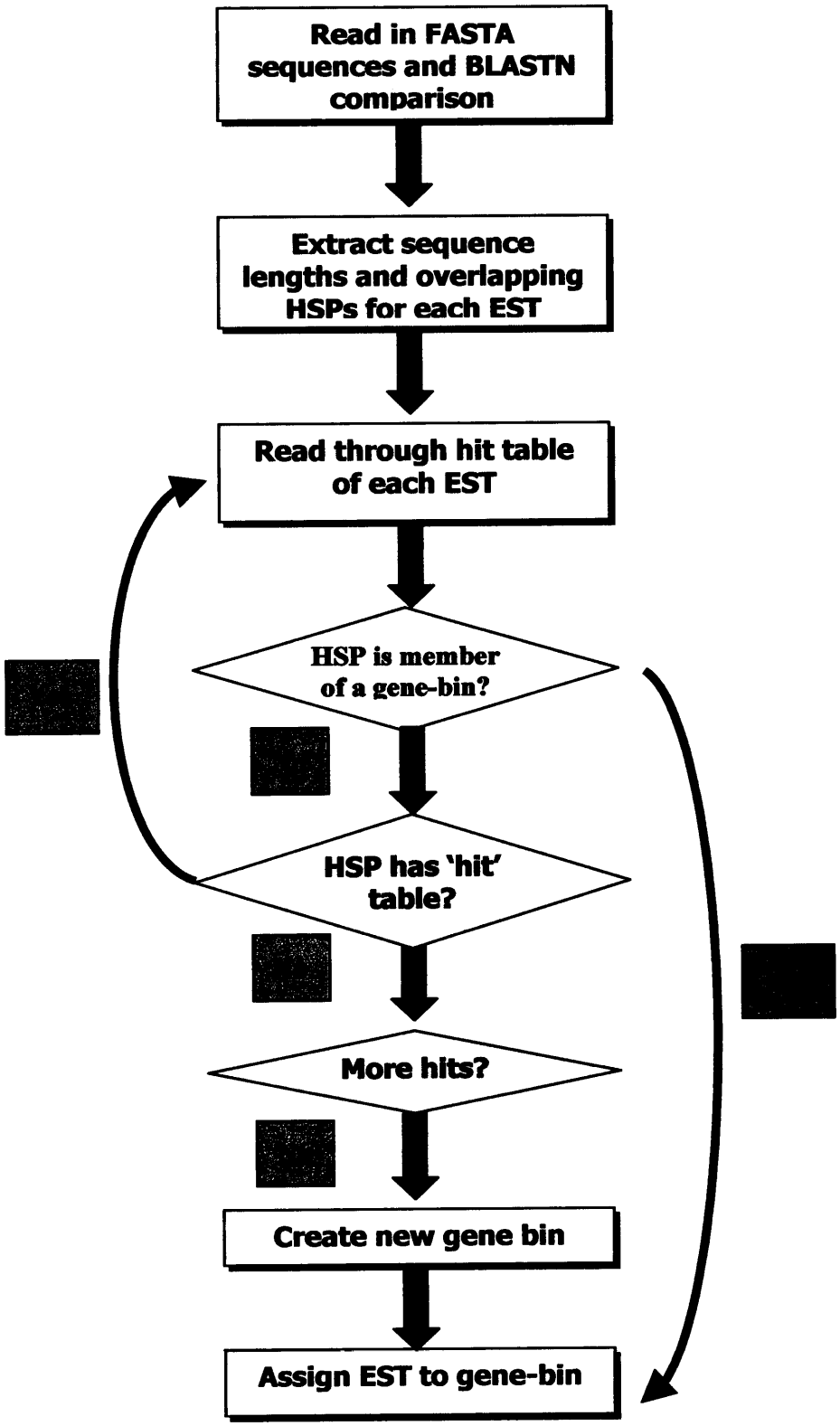


Figure 2.11. Algorithm used for clustering sequences into gene-bins.
BLAST hit tables for each EST are interrogated recursively. If one member of the search tree has been assigned to a gene-bin, this is propagated to the other members of the tree. Otherwise a new bin is created and assigned.

2.3.2.6 Data storage/distribution

The main medium for storage of our sequences is as a flat file in FASTA format. Each project is stored individually in its unclipped (raw) state as well as in its clipped and decontaminated state. Information pertaining to the clipped sequences, consensus sequences and their annotations (described later) are stored in a relational database (Figure A2.1, Appendix 2). A website was designed and linked with this database to provide access for the scientific community (Section 2.7, web ref 12).

2.4 Singleton/Redundancy analysis

In order to maximise the frequency of gene discovery, within individual tissues and across the project as a whole, sequencing was carried out in multiple stages. At each stage those libraries containing the highest proportions of novel transcripts were determined. Sequencing allocations were then biased towards these libraries. Although further minor decisions regarding the sequencing allocation were made throughout the project, the vast majority of the allocations were made after 200,000 clones had been sequenced.

Three main analyses were undertaken to determine the information content of each library: -

1. Clustering, using the uicluster2 program
2. Inter-library comparisons, using BLASTN
3. Swiss-Prot/TrEMBL comparison, using BLASTX.

2.4.1. Clustering

The clustering philosophy used here is subtly different from the clustering described in section 2.3.2.5. The difference is that, for the assembly clustering protocol, two sequences need not share any sequence overlap to be placed in the same cluster; they need only be associated through overlap with shared sequences (Figure 2.12). For the redundancy analysis the ***uicluster2*** program was used, which is fast and optimised for EST clustering. The ***uicluster2*** program is speedy because when a new sequence is being considered for membership in a cluster it is only compared to the cluster's

primary sequence (the primary sequence is the longest member of that cluster, which is continually updated). This results in clusters of sequences that all exhibit sequence identity with the primary sequence of a cluster. This frequently results in clusters where all members have similar sequences.

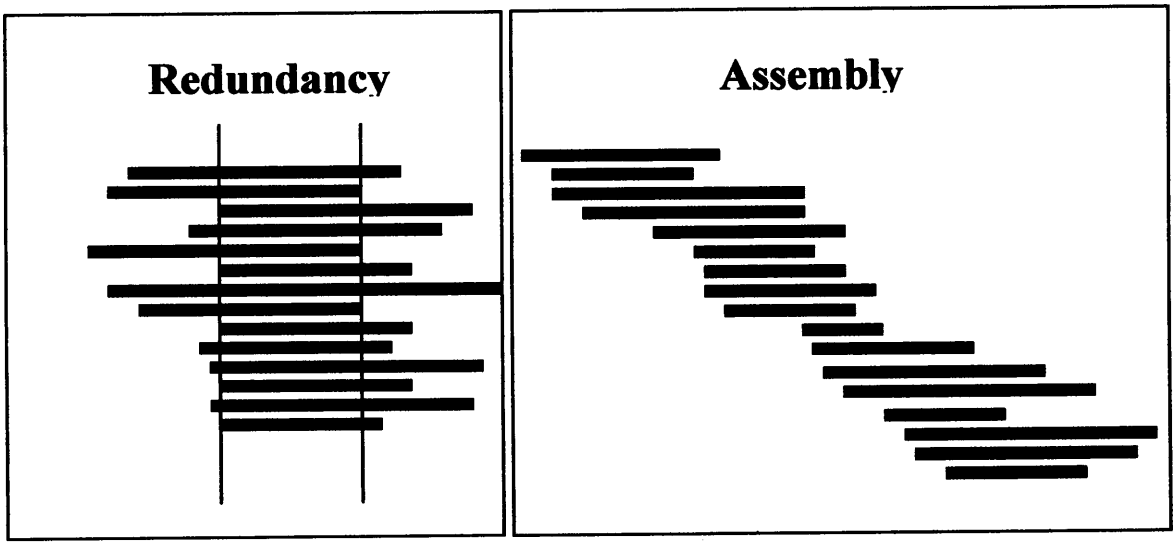


Figure 2.12. Comparison of clustering philosophies for redundancy analysis and prior to assembly.
With the uicluster2 program (redundancy analysis), all sequences must share identity with the primary sequence to be placed in the same cluster. For the assembly clustering algorithm, two sequences are grouped together even if they have no sequence overlap so long as they can be connected by virtue of other sequences with which they share similarity.

The output of ***uicluster2*** is such that the primary sequence of a cluster is easily identifiable. This sequence can be extracted and used as the representative of all sequences in the cluster (the program ***extract_primaries.pl*** was created for this purpose). It is also easy to recognise and extract those clusters that are composed of a single sequence (the program ***extract_singleton_gene_clusters.pl*** was written to extract these singleton gene clusters). In order to assess the levels of novel sequences being produced by a library, the singleton gene clusters were extracted and compared with the SwissProt/TrEMBL database. Sequencing allocation was also biased towards libraries producing novel sequences not found in the public sequence databases.

2.4.2 Inter-library and Intra-tissue comparisons

Clipped ESTs from each project were compared with ESTs within the same project, with all other ESTs generated from the same tissue and with the entire set of ESTs. This was performed using BLASTN and a cut-off e-value of $1E^{-80}$. Sequence files for each library from a tissue were concatenated together to create tissue datasets. For reasons of speed, these were compared with each other using the ***megablast*** program (a member of the BLAST suite) and the same cut-offs as with the BLASTN search above.

2.4.3 Swiss-Prot/TrEMBL comparison

Clipped ESTs were compared with a non-redundant Swiss-Prot/TrEMBL database (Swiss-Prot release 40.0, TrEMBL release 18). A threshold value of $1E^{-30}$ was used to identify those ESTs that are already represented in the public databases.

2.4.4 Results

2.4.4.1 Clustering

An overview of the results from the clustering analysis is given in Appendix 1, Table A1.2. An excerpt from this table is given below (Table 2.4). This shows examples form a highly redundant library (from the "adult pancreas" tissue), and a library considered information rich (from the "adult small intestine" tissue). The pancreatic library has a very low percentage of singleton gene clusters and a high mean cluster size, which is indicative of a very redundant set of sequences. One of the most promising libraries shown here is CSEQCHN56. This has a low level of redundancy, a low mean cluster size and a high percentage of singletons without homologues in the SwissProt/TrEMBL database. Comparing CSEQCHL18 (an un-normalised library) with CSEQCHN56 (normalised) demonstrates the benefits of normalisation. The level of redundancy has been reduced from 19% to 3% through the application of Incyte's 'rare clone biased' procedure.

Table 2.4. Excerpt from Table A1.2 (Appendix 1).
The Norm column refers to the level of normalisation that this library underwent (S = no normalisation. The letters A to F refer to different normalisation protocols). The "level of redundancy" refers to the number of sequences with a stringent match to another EST within the same library. Mean cluster size is the average number of ESTs per cluster after clustering the library with the ***uicluster2*** program (see Section 2.4.1).

Tissue	Library	Norm	Level of redundancy	Mean cluster size	% Singleton gene clusters	% Singletons w/o BLAST of total lib
adult pancreas	CSEQRBL07_JMS	S	72.5%	3.64	23.99%	11.25%
adult small intestine	CSEQCHL18_AIN	S	19.2%	1.24	75.16%	40.84%
	CSEQCHN56_VEI	C	2.7%	1.03	94.88%	**90.53%**
	CSEQCHN58_AWI	E	3.2%	1.03	94.07%	66.39%

As expected, all normalised libraries were much less redundant than the standard libraries generated from the same tissue. There are a number of measures available from the clustering analysis to take into account when appraising the "uniqueness" of a library. The most obvious is the level of redundancy in the library. A highly redundant library is less likely to provide novel sequences than a less redundant one (one of the arguments for normalisation). A second measure is the number of singleton gene clusters in a library. This is a measure of the percentage of clusters from the library that only contain a single sequence (novel sequences). The higher the percentage of singleton gene clusters, the higher the number of novel sequences present in this library. The third method used here is derived from a BLASTX search of the singletons against SwissProt/TrEMBL. The number of singleton gene clusters with no detectable homologue is reported as a percentage of the total number of sequences in the library. This provides a measure of the "uniqueness" of the library. A high percentage in this measure is indicative of a library that contains a large measure of both non-redundant and novel transcripts.

Sequencing allocations were biased towards the following libraries, which were deemed information rich: CSEQCHN59 (stage 36 limbs), CSEQRBN10 (chondrocytes), CSEQRBN13, CSEQRBN14 (ovary), CSEQCHN56 (adult small intestine), CSEQCHN03, CSEQCHN04 (stage 20-21).

2.4.4.2 Global and intra-tissue BLAST analyses

The complete results for the inter-library, global and SwissProt/TrEMBL analyses are shown in various tables in Appendix 1. Table A1.4 contains the intra-tissue analysis results. These were performed to locate libraries providing the highest proportion of unique sequences for its source tissue. Table 2.5 shows an excerpt from Table A1.4. It shows the intra-tissue results for three tissues: adult pancreas, adult small intestine and chondrocytes. The level of redundancy in the pancreatic library is very high (as mentioned previously in section 2.4.4.1). Only 10.5% of the sequences have no similarity to other ESTs in the same library. For the adult small intestine libraries, it is apparent that CSEQCHN56 is the most dissimilar. It has the highest level of novel sequences in comparison to the other libraries and the lowest level of redundancy. Apart from the standard (non-normalised) library, all of the chondrocyte libraries appear to be information rich, with CSEQRBN10 having the lowest levels of similarity to other libraries within the same tissue.

Table 2.5. Excerpt from Table A1.4 (Appendix 1).

		% Unique sequences when compared with					
Tissue	**Library name**	**lib A**	**lib B**	**lib C**	**lib D**	**lib E**	**lib F**
adult pancreas	CSEQRBL07_JMS(A)	10.5					
adult small intestine	CSEQCHL18_AIN(A)	50.35	73.49	76.37			
	CSEQCHN58_AWI(B)	84.13	70.93	71.76			
	CSEQCHN56_VEI(C)	89.9	81.39	73.83			
chondrocytes	CSEQRBL03_MCJ(A)	52.11	67.23	76.44	86.28	74.89	66.94
	CSEQRBN09_RWG(B)	89.13	78.79	87.14	92.21	86.66	83.09
	CSEQRBN10_JFU(C)	91.01	84.13	87.37	93.56	89.96	85.95
	CSEQRBN10_EPW(D)	91.73	85.09	90.17	90.17	89.03	86.98
	CSEQRBN20_AID(E)	88.02	82.77	89.76	93.64	69.72	77.3
	CSEQRBN22_BTM(F)	86.75	81.22	87.49	91.95	82.05	72.48

Table A1.3 (Appendix 1) contains the inter-library redundancy analyses and the comparisons with SwissProt/TrEMBL. Table 2.6 is an excerpt from this table showing the results for the same libraries shown in Table 2.5. The final column in the table gives the total percentage of "unique singletons" contained in a specific library. This figure represents the number of ESTs with no match in the entire EST dataset and no detectible homologue in SwissProt/TrEMBL. This data supports that shown in table 2.5. The pancreatic library has very few unique sequences when compared to the entire

EST dataset (3%). Over 80% of the sequences in the pancreatic library have homologues in the SwissProt/TrEMBL database. In diametric opposition to this is the chondrocyte library, CSEQRBN10. Over 30% of the transcripts from this library have no detectible homologue in SwissProt/TrEMBL or any similarity to other ESTs from the entire chicken EST dataset.

Table 2.6. Excerpt from Table A1.3 (Appendix 1).
Sptr refers to the SwissProt/TrEMBL protein database.

	% unique sequences when compared with					
library name	self & sptr	sptr	tissue	tissue & sptr	all ESTs	all ESTs & sptr
CSEQRBL07_JMS	6	17.34	10.5	6	3	2.62
CSEQCHL18_AIN	30.04	47.02	35.75	21.95	7	5.71
CSEQCHN58_AWI	51.45	69.82	52.91	39.47	12.41	11.15
CSEQCHN56_VEI	70.67	95.3	62.29	59.67	20.73	20.19
CSEQRBL03_MCJ	36.35	57.44	33.4	23.65	9.31	8.68
CSEQRBN09_RWG	62.64	77.18	57.17	46.83	24.96	22.69
CSEQRBN10_JFU	70.16	80.04	62.39	52.06	32.35	30.2
CSEQRBN10_EPW	72.24	78.79	63.31	52.74	33.66	30.79
CSEQRBN20_AID	54.57	77.26	43.62	35.64	17.3	16.08
CSEQRBN22_BTM	59.84	79.34	52.34	44.55	24.41	22.87

2.4.4.3 Inter-tissue analysis

The final question to address is that of the similarity between each tissue. Other analyses have discovered which library within a tissue yields the greatest proportion of unique sequences. Allocating sequencing on this data alone may unnecessarily increase the number of redundant sequences due to similarities of expression between different tissue types. Table A1.5 (Appendix 1) shows the percentage of similar sequences found between each tissue. Figure 2.13 is a graphical overview of this data produced using the ***draw_jock_plot*** program. This figure can be slightly confusing at first. It seems unintuitive that a comparison of x against y does not yield the same results as a comparison of y against x. The disparity is due to the differences in the total number of ESTs contained within each tissue. It is likely that a higher proportion of sequences from a tissue will find matches if a small tissue (e.g. the "adult adipose" tissue, 1964 sequences) is compared with a large tissue (e.g. the "chicken ovary" tissue, 19831 sequences) than in the reverse comparison. This is seen when looking at the results of comparisons of the smallest tissue, "adult adipose", with the other tissues. Looking

along the x-axis of the adipose tissue comparisons (the first row of Figure 2.13), it is apparent that many of its transcripts have similar sequences in other tissues. In contrast, the comparisons of other tissues with the adipose tissue (the first column of Figure 2.13) show very little similarity. From this one can conclude that if a tissue shows similarity with other tissues in both *x* and *y* comparisons then future sequencing of this tissue will potentially provide more redundant information than sequences from other tissues. The most striking example of this is the stage 10 embryonic tissue, which contains many transcripts also present in all of the stage 36 tissues. This came as a surprise as it was expected that early embryonic tissues would contain many novel transcripts that are only expressed in early development.

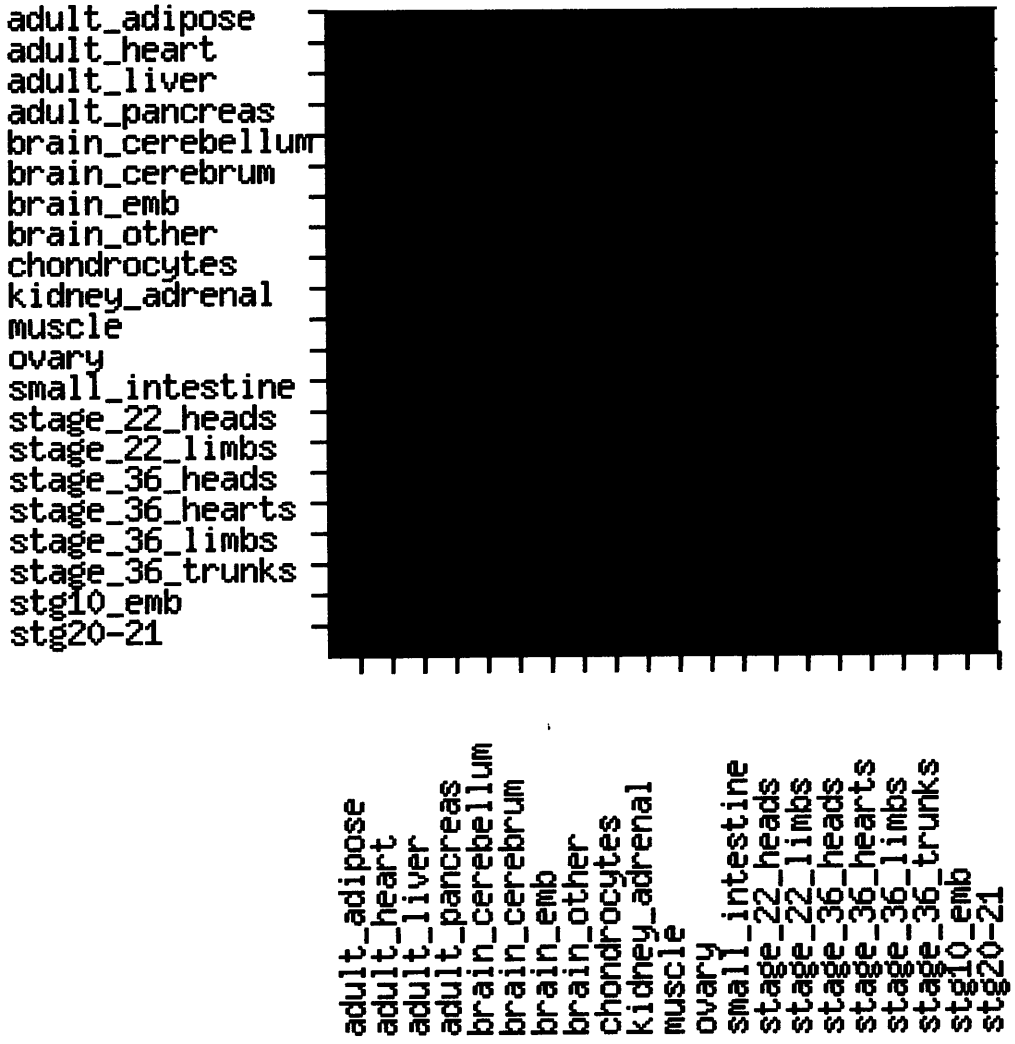


Figure 2.13. Graphical overview of the tissue vs. tissue similarity analysis.
This represents a ***megablast*** comparison of all tissues. Each square in the grid represents a comparison between two tissues. Self/self comparisons were not performed. The result of the comparison is depicted as an intensity of colour ranging from black (no similarity) to bright red (high similarity). The highest similarity shown here is a comparison of stage 36 hearts with stage 10 embryos where 38.1% of all sequences from the heart tissue find a partner in the stage 10 tissue.

2.4.4.4. Conclusions and further sequencing decisions

Figure 2.14 shows the total number of sequences obtained from round 1 and round 2 of the tissue sequencing. This data is also shown in tabular form in Table A1.6 (Appendix 1). As can be seen from the figure, no further sequences were taken from the following tissues in round 2 of the sequencing: adult adipose, adult heart and stage 36 hearts. Attempts at creating a normalised library for the adult adipose tissue failed. Further sequences were requested from the stage 10 embryonic tissue because, although the "uniqueness" statistics were not that strong, it was believed that there are some transcripts available exclusively at this stage and due to the effort required to harvest the embryos.

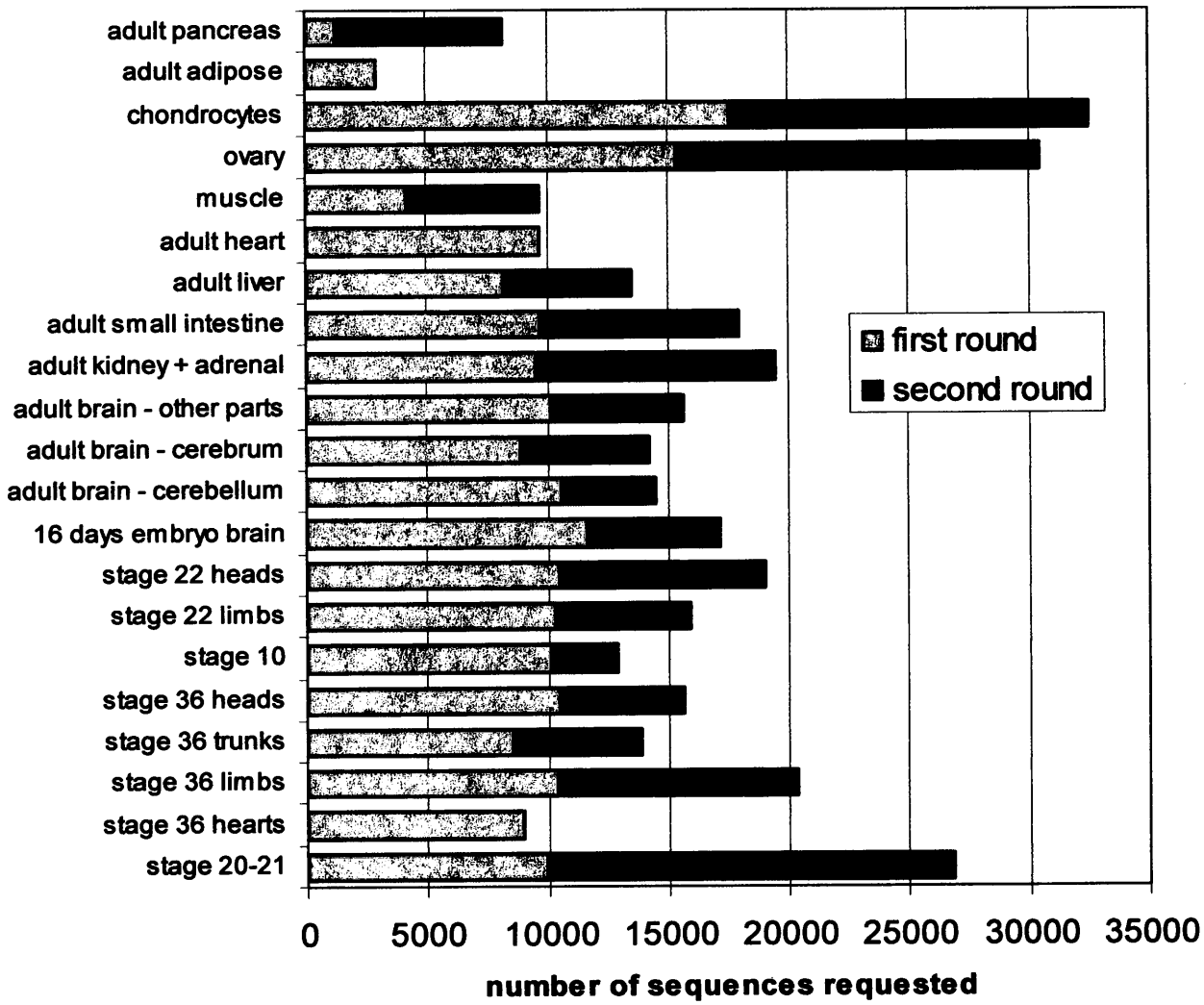


Figure 2.14. Sequence allocations from round 1 and round 2.

Two normalised pancreatic libraries were in the process of being constructed at the time of the final sequencing allocations. Although the standard pancreatic library was highly redundant and was deemed to have low information content, it would have been foolish to ignore the normalised libraries completely. The higher stringency normalised library was allocated 4,500 reactions and the lower stringency library was allocated 3,000 reactions. The libraries producing the highest proportions of novel sequences were allocated the largest numbers of sequencing reactions: stage 20-21 embryos (~17,000 reactions), ovary (16,000) and chondrocytes (16,000). Four tissues were allocated an intermediate number of sequencing reactions: stage 36 limbs (10,000), adult kidney and adrenal (10,000), stage 22 heads (~9,000) and adult small intestine (8,500). In order to maintain a broad coverage, between 3,000 and 5,000 sequencing reactions were allocated to information rich libraries selected from the remaining tissues.

2.4.4.5. Final sequencing and assembly statistics

In total, 339,314 ESTs were sequenced. Of these, 323,670 passed through the vector clipping and decontamination stages and underwent assembly and annotation. The average read length for the clipped ESTs is 745 bp.

The decontaminated ESTs were partitioned into 64,760 gene-bins, which in turn were assembled to produce 85,486 contigs. Of these, 46,674 (55%) were singleton contigs and 38,812 (45%) were multi-component contigs resulting in an average of 3.8 ESTs per contig and an average read length of 874 bp (1158 bp for multi-component contigs). Strikingly, a number of significantly larger contigs have been assembled; for example, contig 354630.6 is over 7 Kb long, corresponding to the complete chicken non-muscle myosin heavy chain open reading frame and including more than 1.5 Kb of 3' UTR (Figure 2.16). Figure 2.15 shows the distribution of EST and contig read-lengths from this analysis.

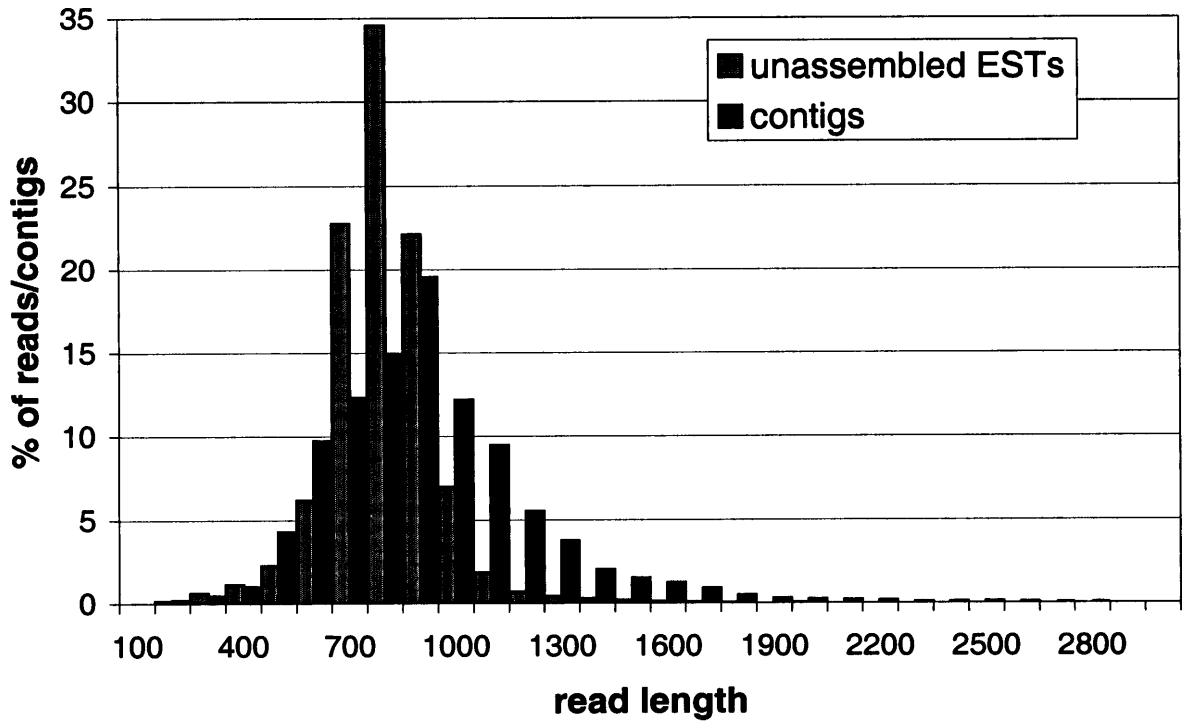


Figure 2.15. Histogram of read lengths for unassembled ESTs and contig sequences.
Read length is given in base pairs.

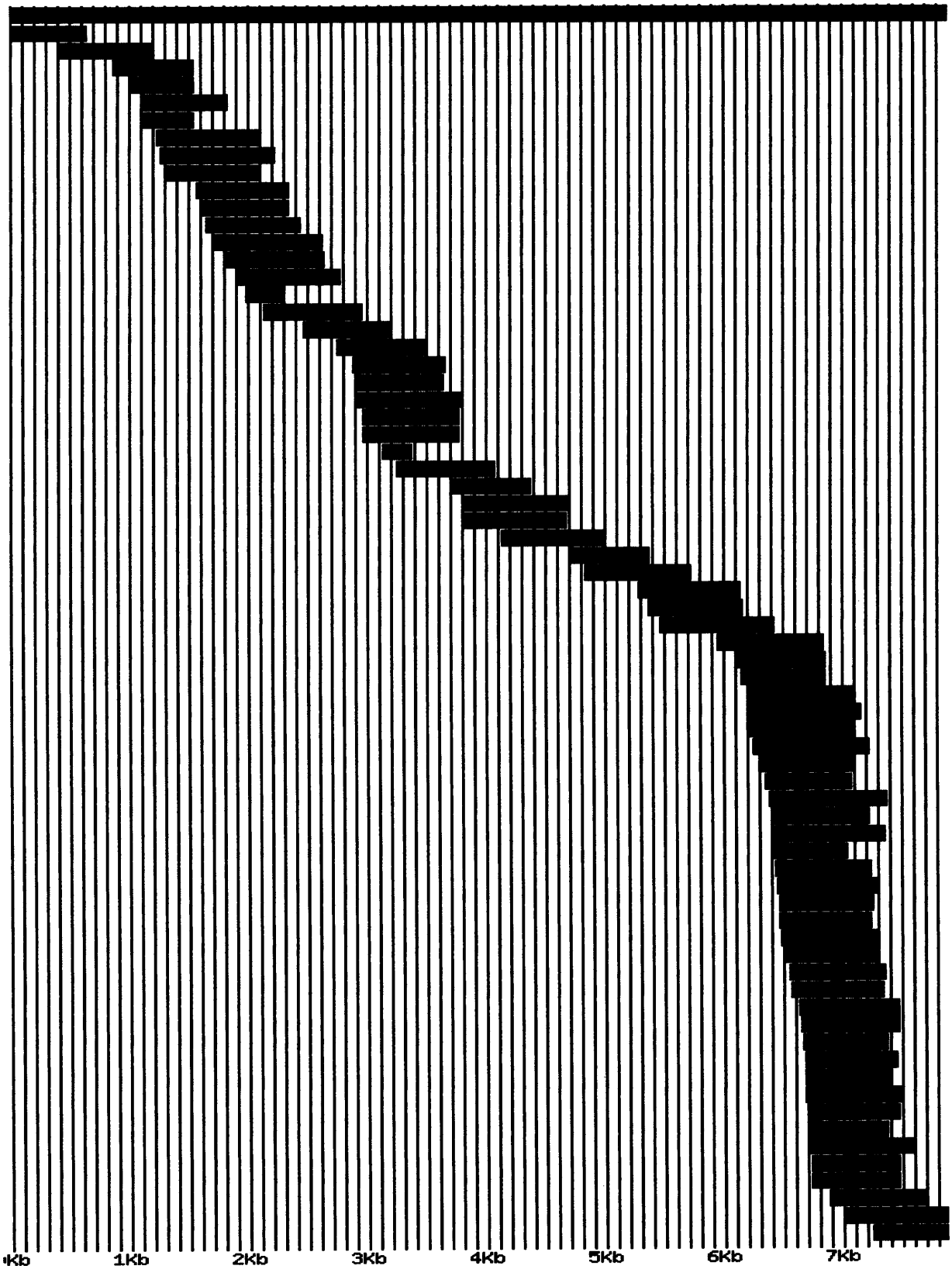


Figure 2.16. Contig sequence and component EST sequences representing chicken non-muscle myosin heavy chain (Q02015).
This is a graphical representation of the composition of contig sequence 354630.6. The blue bar at the top represents the contig sequence. The red bars represent alignment positions of the component EST sequences. The spacing of vertical lines is equivalent to 100 bp in the gapped alignment.

2.5 Annotation of gene function in Chicken ESTs

2.5.1 BLAST annotation

The main route for functional annotation of a biological sequence is through homology detection with proteins of known function. This method has been used to annotate tens of thousands of sequences and has been used in the annotation of both EST and genomic data. Primary annotation was procured by searching the six frame protein translations of every EST and consensus sequence against SwissProt/TrEMBL using the BLASTX program. We also extracted all chicken proteins from SwissProt/TrEMBL and all publicly available chicken ESTs from EMBL. A comparison of our data with these sequences provides an estimate for the depth of coverage of this resource.

2.5.1.1 Programs developed for annotation

The ***blast_large_seq_file*** program was created to speed up the annotation process of the large consensus sequence dataset (and other BLAST searches using large numbers of query sequences). This is achieved by distributing the load over multiple nodes on the linux cluster. The ***blast_large_seq_file*** program has the following logic: -

1. Split the FASTA formatted input file into multiple "chunks" with a user defined number of sequences per chunk.
2. Create and execute a command list (for use with ***client.pl***) to run ***blast_on_remote_node*** (or ***megablast_on_remote_node***) for each chunk.
3. Wait for the BLAST searches to complete and collate the blast results into a single file.
4. Delete all temporary files.

The ***blast_on_remote_node*** program copies a chunk of data onto its local node, runs the requested BLAST search and then copies the results back to the specified location. A similar program, ***megablast_on_remote_node***, was developed to run megablast searches. Both are run *via* the ***blast_large_seq_file*** program.

2.5.1.2 Parameters for annotation

ESTs are annotated with the top BLASTX hit matching with a threshold e-value of $1E^{-3}$ or better. Consensus sequences are annotated if their top hit has an e-value of $1E^{-6}$ or better. Originally annotation was much more stringent, with an e-value threshold of $1E^{-30}$. The use of a stringent threshold results in a low level of annotation and many relationships are missed. A lower threshold was introduced in order to capture more distant relationships. The e-value of the BLAST hit used to annotate a sequence is displayed alongside any annotation when viewed by users of the database. This allows an assessment of the quality and reliability of any individual annotation.

For the comparison with extant chicken data, we used a threshold e-value of $1E^{-6}$ to identify known proteins and an e-value threshold of $1E^{-30}$ for previously characterised EST sequences.

2.5.1.3. Results

Similarity searches against the SwissProt/TrEMBL database were successful in providing annotation for 173,351 (50%) ESTs and 31,005 (38.5%) consensus sequences (Figure 2.17).

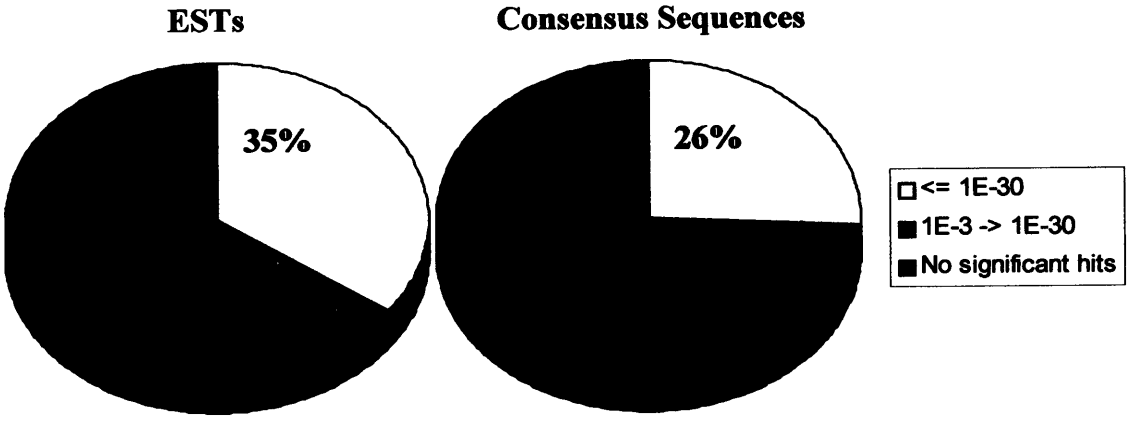


Figure 2.17. Results of BLASTX searches against SwissProt/TrEMBL.

The yellow segment shows all BLASTX hits with an e-value of $1E^{-30}$ or better. The orange segment represents BLASTX hits with an e-value of $1E^{-3}$ to $1E^{-30}$. The blue segment represents all sequences with no significant BLASTX match in the database.

These figures represent a much lower ratio of annotated to unannotated sequences than those found in previous studies (Abdrakhmanov 2000, Tirunaguru 2000). For

example, ~75% of the Buerstedde bursal EST library (Abdrakhmanov 2000) find a homologue in SwissProt/TrEMBL with an e-value of $1E^{-3}$ or better. This is very likely due to the successful application of a more advanced normalisation procedure in our libraries, coupled with the sequencing bias we placed on our libraries producing novel transcripts. The fall in the percentage of annotated sequences when considering consensus sequences rather than ESTs reflects a bias in the public databases for more abundant transcripts. In other words, as our EST resources are more redundant, they have multiple copies of abundant, well-characterised genes that will have BLASTX hits. Another possibility is that a proportion of our ESTs are composed of genomic DNA, UTR sequences or some other form of contamination or artefact. These possibilities are addressed in section 2.5.4.

Figure 2.18 shows the results for the comparisons with the extant chicken data. At the time of this analysis there were 3,042 chicken proteins and 43,900 chicken ESTs in the public databases. Of these 2,764 proteins and 34,302 EMBL ESTs find a partner in our database. The majority of ESTs currently available in EMBL are from a bursal cDNA library. This immune tissue represents a tissue type not selected for sequencing in this project. Indeed, the presence of these ESTs in EMBL and the continuing efforts of the originating laboratory to sequence this library were considerations when selecting tissues for sequencing.

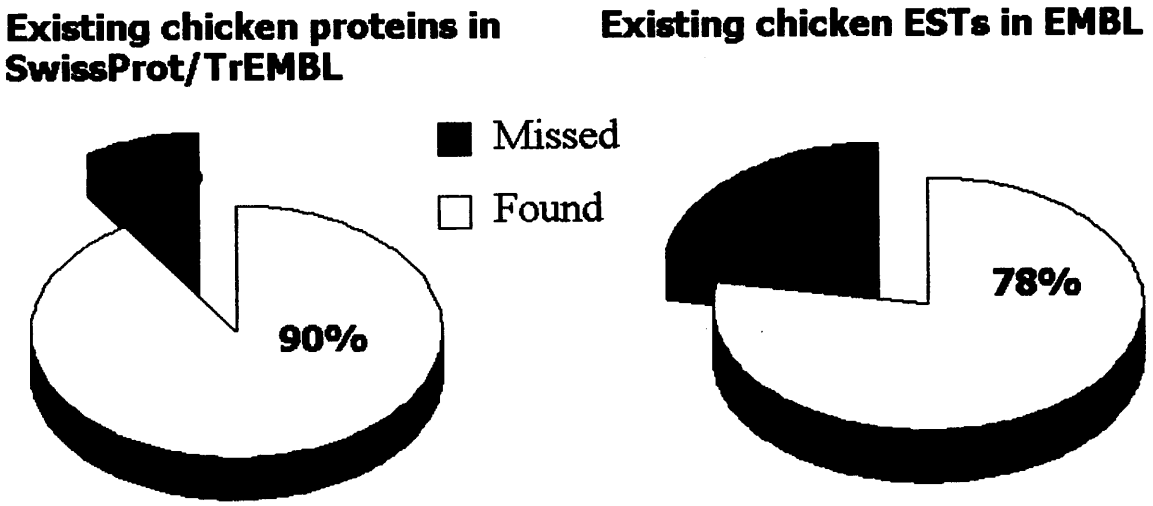


Figure 2.18. Overview of comparisons of BBSRC chicken EST database with extant chicken data.

2.5.2 InterPro annotation

2.5.2.1 InterPro and InterProScan

Secondary protein databases on functional sites and domains such as PROSITE (Hofmann 1999), PRINTS (Attwood 2000), SMART (Schultz 2000), Pfam (Bateman 2000), ProDom (Corpet 1999), are vital resources for predicting the function of novel sequences. These signature databases have very different formats and nomenclature which makes unifying the results both difficult and time consuming. To address this issue the Integrated resource of Protein Families, Domains and Sites (InterPro) was created (Apweiler 2001b). InterPro provides an integrated view of the commonly used signature databases, which lends it well to the annotation and functional classification of uncharacterised sequences. The EBI, for example, use InterPro for enhancing the automated annotation of TrEMBL (Fleischmann 1999). Using InterPro is more efficient and reliable than using each individual database because InterPro provides internal consistency checks and deeper coverage (Apweiler 2001).

InterProScan is a tool that scans input protein sequences against the protein signatures of the InterPro member databases (Zdobnov 2001). The Perl-based implementation is freely available from the EBI ftp server (web ref 13).

2.5.2.2 Annotation of contig data

Each contig sequence must be translated into protein sequence before being subjected to a scan against the InterPro member databases. Since our ESTs were produced using directional cloning, it is in principle only necessary to translate each sequence in the three forward reading frames (the ***3frametrans.pl*** program was written for this purpose). InterPro annotation was only carried out on consensus sequences due to the large amount of time and resources required for a single search against the member databases. This reduced the number of scans required from 1 million (three frame translation of ESTs) to 260,000 (three frame translation of contig sequences). Fifteen nodes (30 CPUs) of the linux cluster were in constant use over a period of two weeks to complete the InterPro scans.

Test runs of ***InterProScan*** revealed difficulties when scanning with large input datasets. On numerous occasions jobs terminated prematurely and required restarting. Fortuitously, the design of ***InterProScan*** is such that when a job is restarted it picks up where it left off. Due to these limitations, it was not possible to run InterProScan with the current distributed programming system developed for the linux cluster. Instead, it was necessary to individually run jobs on separate nodes. The ***do_batch_interpro.pl*** program was written to execute a series of ***InterProScan*** runs on an individual node. This program takes a FASTA formatted input file, splits the file into more manageable chunks (default of 500 sequences per chunk). Each chunk is then processed in turn as follows: -

1. Produce three forward frame protein translations for all sequences
2. Run InterProScan.pl on the input protein sequence file - this creates all the necessary files and directories for the InterPro search
3. Alter configuration files so that both of the nodes CPUs are used whenever possible.
4. Execute the InterProScan.

The results from each chunk on each node are then collated into a single results file representing the results for the entire dataset.

2.5.2.3 Results

Around 25% (21,222) of the consensus sequences can be annotated with hits to entries in one or more of the protein signature databases. A further 6% (5,027) have predicted coiled-coil motifs but no matches with the protein signature databases, giving a total of 26,249 annotated contigs. Table 2.7 shows the results for each individual database searched.

Table 2.7. Individual database results for InterPro scans.

These results were generated using default parameters.

Database	Number of contig hits
ProfileScan	15300
BlastProDom	3296
Coil	8538
HMMPfam	12625
HMMTigr	208
HMMSmart	6522

2.5.3 Genomic contamination assessment

Although ESTs are primarily produced from processed mRNA present in the cytoplasm of a cell, it is possible that some unprocessed mRNA will be present in the cDNA libraries. There may also be a small amount of non-coding genomic DNA. Unfortunately, there is insufficient intergenic DNA present in the public databases to examine the contamination level of this sequence type.

Two methods were employed to assess the level of unprocessed mRNA: A search of the chicken EST dataset with all known intronic sequences extracted from the limited number of complete chicken genes in the EMBL databank, and an investigation utilising the well characterised chicken beta-actin gene.

2.5.3.1 Chicken introns

All known *Gallus gallus* sequences were extracted from the EMBL nucleotide database (6016 sequences). On examination of the EMBL entries it became apparent that UTR sequence was often labelled as intronic in some databank entries. To counter this, our extraction was altered to only output introns that are found between two exons. A total of 878 introns were extracted from the 248 EMBL entries containing exon bounded introns for *Gallus gallus*.

A preliminary BLAST search of the consensus sequences against the introns was carried out. Just over 0.2% of the contigs matched to the intron database (with a minimum percentage identity of 98% over a length of 25 nucleotides or more).

To calculate a more accurate estimate it is necessary to perform the analysis with ESTs representing the intron containing sequences from EMBL. These were extracted by searching the EST dataset against the intron containing set of EMBL *Gallus gallus* sequences. Those ESTs that matched with a threshold e-value of $1E^{-30}$ were extracted and then searched against the intron database. The number of ESTs matching one or more introns from their corresponding EMBL partner was calculated. The number of ESTs containing introns in the entire dataset was then estimated by calculating the ratio of intron containing to intron free ESTs with matches to *Gallus gallus* EMBL sequences.

2.5.3.2 Beta-actin

The nucleotide sequence for the *Gallus gallus* cytoplasmic beta-actin gene was extracted from the EMBL database. This gene has both 3' (~600 bp) and 5' (~1kbp) UTR sequence, five exons and four introns (Figure 2.19). The UTR sequences are labelled as introns in the EMBL feature table but are not considered as such in this analysis (UTR sequences are present in fully processed mRNAs).

The ***megablast*** program was used to extract all ESTs matching to the EMBL actin sequence (with a threshold e-value of $1E^{-30}$).

```
FT   source          1..5046
FT   CAAT_signal     455..459
FT   TATA_signal     517..524
FT   intron          544..1542
FT   misc_RNA        544..544
FT                   /note="cap site"
FT   exon            1543..1665
FT                   /number=1
FT   intron          1666..1985
FT                   /number=1
FT   exon            1986..2225
FT                   /number=2
FT   intron          2226..2749
FT                   /number=2
FT   exon            2750..3188
FT                   /number=3
FT   intron          3189..3494
FT                   /number=3
FT   exon            3495..3676
FT                   /number=4
FT   intron          3677..4031
FT                   /number=4
FT   exon            4032..4172
FT                   /number=5
FT   intron          4173..4766
FT                   /number=5
FT   polyA_signal    4744..4751
FT   polyA_site      4764..4764
FT   polyA_site      4766..4766.
```

Figure 2.19. Feature table from the EMBL entry for *Gallus gallus* cytoplasmic beta-actin gene.

2.5.3.3 Results

Preliminary intron analysis

A total of 6865 ESTs matched to an intron containing *Gallus gallus* sequence from the EMBL database. Of these, 235 ESTs were found to contain an intronic sequence from their corresponding EMBL hit. This represents a potential intron contamination level of 3.4%.

Beta actin

Out of the 326 ESTs matching to the beta acting sequence, only two were found that contained intronic contamination. This represents a potential intron contamination level of 0.6%.

2.5.4 Estimation of full-length clones

A collection of sequenced full-length cDNAs is an important resource both for functional genomics studies and for the determination of the intron-exon structure of genes (Stapleton 2002). The number of our cDNA clones that extend to the start codon for known chicken genes was estimated by comparing the contigs with a set of 1,845 non-redundant, full-length *Gallus gallus* cDNA sequences taken from the EMBL databank. In total, 3,181 of the 85,486 contigs had stringent BLASTN matches (greater than 98% identity over 100 bases), of which 996 (31%) extend completely to the start codon. In addition, 631 (20%) of these sequence matches extend from the start codon to the stop codon, suggesting that the complete coding sequence has been obtained for these genes. Extrapolating these figures to the complete dataset suggests that a large number of contigs (17,000 → 25,000) contain component clones that encompass the start codon and may represent full-length cDNAs. Figure 2.20 illustrates the lengths of chicken gene coding sequences that the contigs cover completely; these sequences include many that extend up to and greater than 2 Kb in length. Similar results were obtained using a set of 1,676 non-redundant, full-length chicken sequences extracted from SwissProt/TrEMBL, where 916 (37%) of the 2,442 matching contigs extend to within 5 amino acids of the N-terminus of the matching chicken protein (using a match criteria of 96% identity over 30 amino acids).

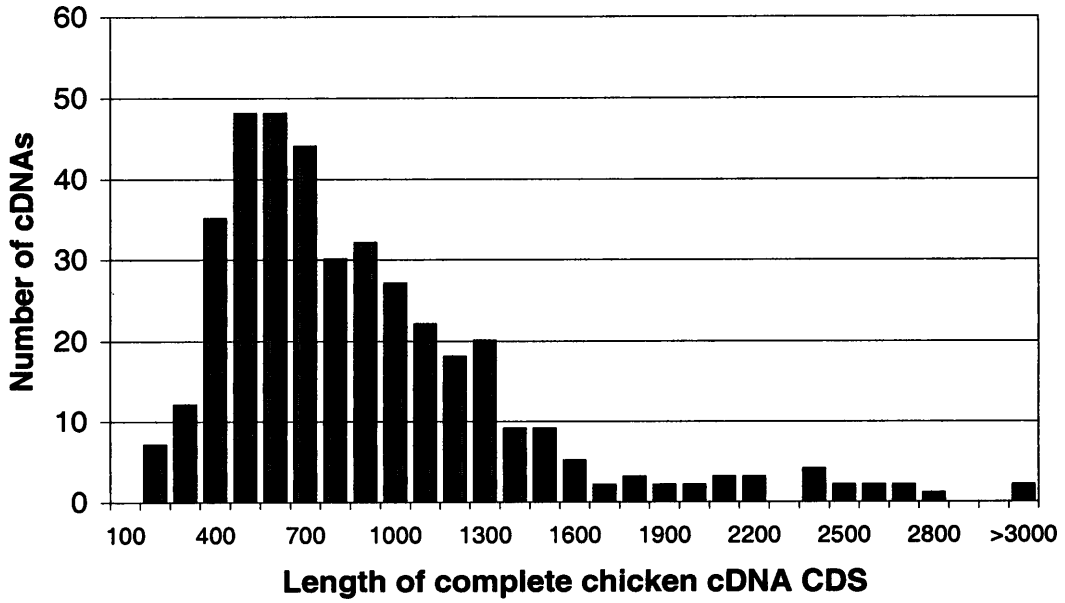


Figure 2.20. Distribution of full-length chicken cDNA coding regions (from EMBL) completely matched by individual contigs.

2.5.5 Gene number estimation

The number of protein-coding genes in an organism provides a useful first measure of its molecular complexity. Prokaryotes and single celled eukaryotes typically have a few thousand genes (e.g. *E. coli* = 4,300, *S. cerevisiae* = 6,000) whereas multicellular eukaryotes have many more (*D. melanogaster* = 13,600, *H. sapiens* = ~ 35,000). It is possible to estimate the number of genes in *Gallus gallus* using our EST data and the method developed by Brent Ewing and Phil Green (2000). In this methodology the number of genes, G, is given by:

$$G = n_1 n_2 / m_2$$

Where n_1 is the number of sequences in a representative set of full-length, unbiased, genes or proteins from the genome, n_2 is the number of sequences in the second sequence set being compared and m_2 is the number of sequences from set n_1 that match to a sequence in set n_2.

We carried out this analysis using the set of 38,812 contigs that contain at least two component ESTs and two reference sets; 1,845 clustered, full-length chicken cDNAs (taken from EMBL) and 1,676 non-redundant complete protein sequences (taken from SwissProt/TrEMBL). A match was assigned to hits possessing 98% or better sequence identity over 100 bases or more for the cDNA set. A less stringent cut-off of 96% identity and an overlap of 30 amino acids were used for the protein set because of the greater impact that minor sequencing and frameshift errors have on the translated nucleotide BLAST statistics.

The estimated total number of genes predicted using this approach were 33,228 and 35,682 for the non-redundant cDNA and protein reference sets, respectively. This is in reasonable agreement with gene number estimates for the human (Ewing 2000, Hogenesch 2001) and *Fugu rubripes* genomes (Aparicio 2002), suggesting a common baseline of around 35,000 genes for vertebrates.

2.6 Comparative genomics

2.6.1 Comparison with the Human Proteome

A comparison against the human genome gives an indication of the coverage of the EST resource and its utility for the study of human genetics. Although *Gallus gallus* and *Homo sapiens* diverged approximately 300 million years ago (Nei 2001, Shaul 2002), it can be expected that many genes will still have detectable homology. Estimates place the human/yeast divergence at over 1.3 billion years ago and yet 50% of the proteins in the *Saccharomyces cerevisiae* genome have detectible homology (at a threshold e-value of $1E^{-6}$) with the human Ensemble confirmed protein dataset.

2.6.1.1 Ensembl comparison

The confirmed human protein dataset was acquired from the Ensemble ftp site (web ref 14). This consists of 27,628 protein sequences covering 22,877 unique genomic loci. A TBLASTN search of the human proteins against the chicken consensus sequences was carried out to determine the number of human proteins with a homologue in the chicken sequence database. Over 80% of the human proteins have a sequence match to a consensus sequence using an e-value threshold of $1E^{-6}$ (Figure 2.6.1) and 64% had a match using a threshold of $1E^{-30}$. These figures are dramatically reduced when searching against the GENSCAN predicted protein set. This contains 73,128 proteins (predicted by the GENSCAN program (Burge 1997) from the raw human genome sequence) of which 40% have a match with at least one consensus sequence (with an e-value threshold of $1E^{-6}$). This predicted set includes a large proportion of erroneously assigned proteins, originating from non-coding sequences and pseudo-genes. Given that these genetic elements have a high degree of sequence variability and are free to mutate rapidly over time it is not surprising that a reduced number find matches within our consensus sequence set.

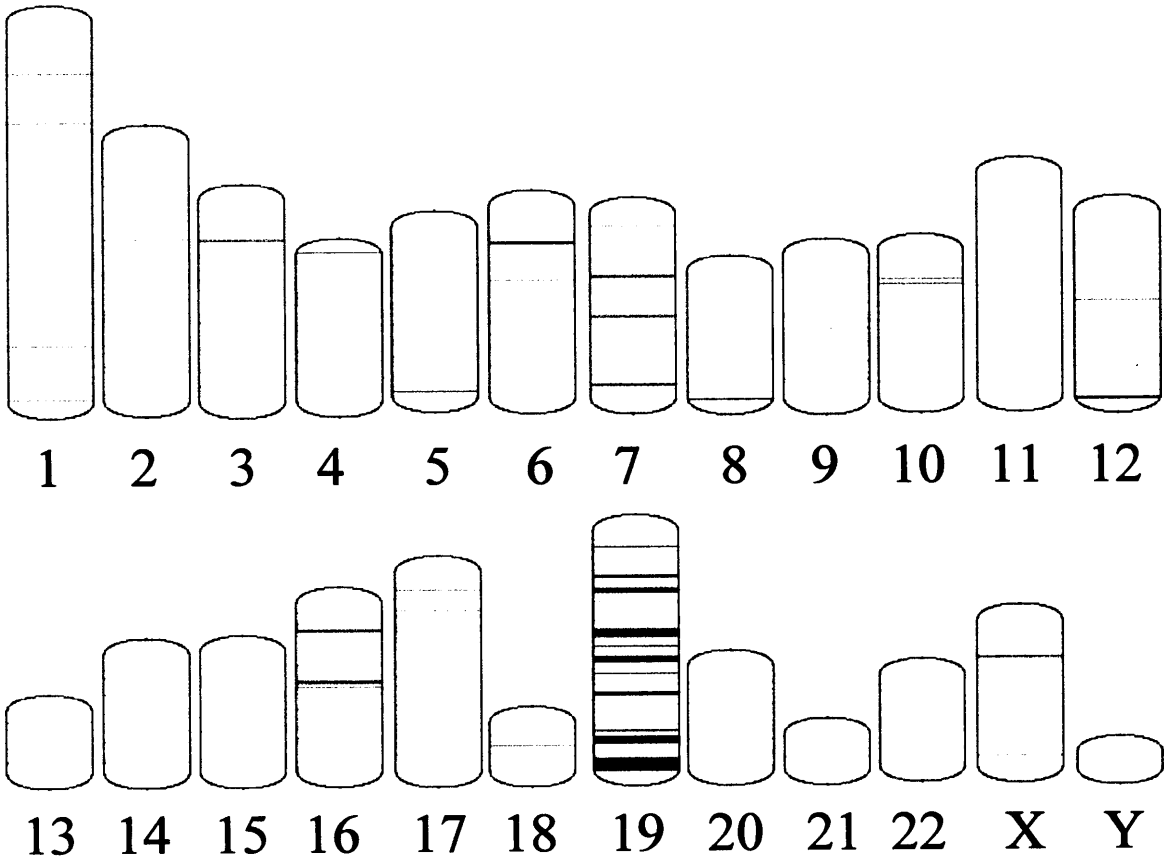


Figure 2.21. Comparison of chicken ESTs against the human proteome.

Confirmed human proteins from each chromosome were searched against the chicken EST database. Blue bars represent proteins that have a homologue in the database (using an e-value threshold of $1E^{-6}$). White bars represent proteins with no homologue in the chicken EST database. An increased intensity of colour represents a higher coverage for that protein. A more detailed graphic for each chromosome is available in Figure A2.1 (Appendix 2).

2.6.1.2 OMIM comparison

The Online Mendelian Inheritance in Man (OMIM) database contains information on human genes and genetic disorders (web ref. 15). The subset of this database known as the 'morbid map' contains details on genes linked with known hereditary genetic diseases. Accession numbers were extracted from the 'morbid map' database and used to extract the corresponding protein sequences from Genbank. A TBLASTN search of these sequences against the consensus sequences found stringent matches (threshold e-value of $1E^{-30}$) for 86% and less stringent matches (threshold e-value of $1E^{-6}$) to 95% of the 931 disease genes extracted from Genbank (Figure 2.22).

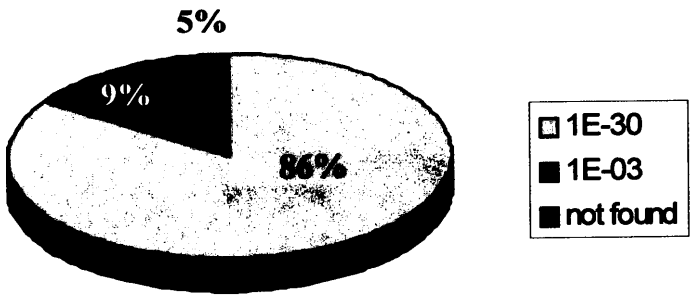


Figure 2.22. Number of OMIM morbid map genes with homologues in the consensus sequence data.

These results clearly show that the chicken EST database contains a large proportion of genes known to be involved in human disease. This high degree of disease gene representation demonstrates the utility of chickens as a model organism for the study of human disease.

2.6.2 GO comparison with other eukaryotes

To produce a more comprehensive comparison, the BBSRC chicken EST data was combined with extant chicken EST data extracted from Genbank. At this time Genbank contained around 60,000 *Gallus gallus* ESTs. The protocol described in section 2.3.2.5 was employed to generate an assembly containing over 72,000 gene bins and 97,000 consensus sequences. A GO Slim (Ashburner 2000, web ref. 16) annotation of this contig set was produced using stringent database matches (e-value threshold of $1E^{-6}$), assigning GO terms by inference from database matches with known GO assignments. Figure 2.23 shows a comparison of this GO annotation with that for five other model eukaryotes (*Homo sapiens*, *Caenorhabditis elegans*, *Drosophila melagonaster*, *Mus musculus* and *Saccharomyces cerevisiae*). This appears to show that these selected proteomes share very similar relative fractions of proteins assigned to the same high-level term from the Gene Ontology. One feature that stands out is that mice, humans and chickens contain relatively smaller fractions of enzymes and larger fractions of genes involved in cell-cell signalling in comparison to yeast and fly.

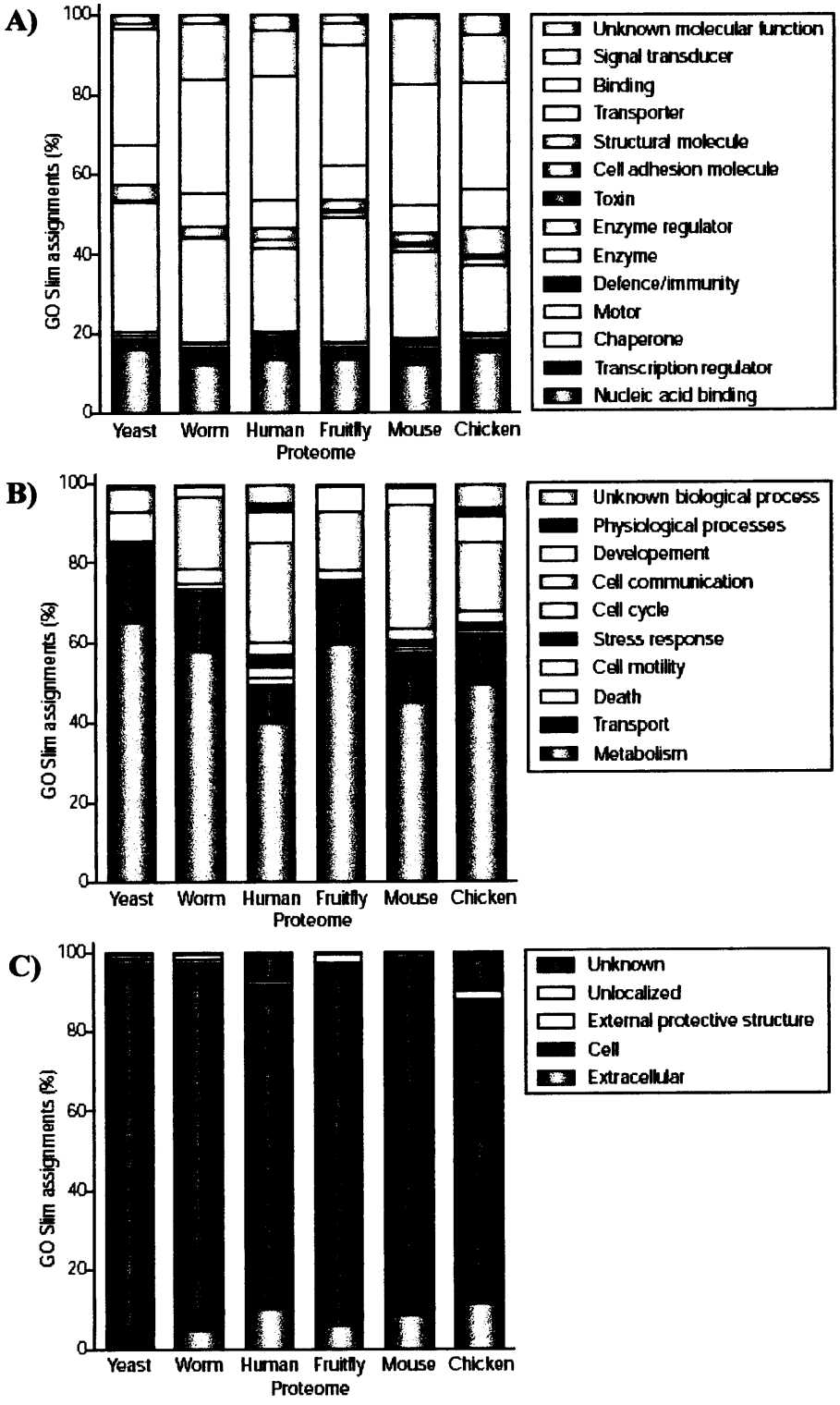


Figure 2.23. Gene Ontology functional assignments to eukaryotic proteomes that have been completely sequenced (adapted from Brown *et al.* 2003).

The relative fractions of assigned protein functions using the Gene Ontology classification for *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans* (worm), *Homo sapiens* (human), *Drosophila melagonaster* (fruitfly) and *Mus musculus* (mouse) taken from the EBI proteome web site (web ref 17) compared with the GO assignment of assembled contig sequences. A) Molecular function, B) Biological processes and C) Cellular component.

2.6.3 *In-silico* subtraction

The EST database described here represents a potentially useful resource for the analysis of differential gene expression. Typically, this kind of analysis is carried out through the creation and sequencing of a subtracted cDNA library, followed by *in-situ* hybridisation of a selection of the resulting clones. Subtraction enriches the levels of novel sequences within a library by removing sequences found in common with another library. For example, Christiansen *et al.* (2001) created a subtracted library enriched for transcripts specific to the chicken embryonic hindbrain through subtraction of an embryonic hindbrain library with a pre-streak stage chicken embryo library. This was believed to significantly deplete ubiquitously expressed genes from the subtracted library. Christiansen and colleagues then tested 445 clones from this library for tissue specific expression profiles. This was carried out through *in-situ* hybridisation of labelled clones in chicken embryos. Thirty-six of the 445 clones (8%) displayed restricted expression patterns within the hindbrain, midbrain or cranial neural crest. Twenty-two of these are novel and eleven encode peptides with homology to proteins with previously uncharacterised roles during early neural development.

We have sequenced mRNAs from 21 adult and embryonic tissues; some to an extensive level. An *in-silico* subtraction seems plausible with the extent of this data. By creating a list of clones that are present in one library but are not represented in, one or more, other libraries it may be possible to identify tissue specific transcripts.

2.6.3.1. Methods

We created a 'virtual library' enriched for transcripts present in stage 22 limbs through computational subtraction with the following libraries: "stage 20-21 whole chick embryos", "stage 22 heads" and "stage 36 trunks". These three tissues were compared with the "stage 22 limb" tissue using BLASTN and a threshold e-value of $1E^{-80}$. A strict threshold value such as this is used to subtract only unique genes rather than paralogous genes. Sequences with no match from the "stage 22 limb" tissue were extracted. A BLASTX search of these against SwissProt/TrEMBL was carried out and ESTs were grouped together through virtue of their top hits. This step groups

ESTs that potentially represent the same transcript and adds a limited amount of annotation to the data.

A selection of these clones were then used as probes for a set of *in-situ* hybridisation experiments in stage 22 chicken embryos, undertaken by our collaborators in Dundee. In this initial analysis, clones with no functional annotation were not considered for *in situ* hybridisation.

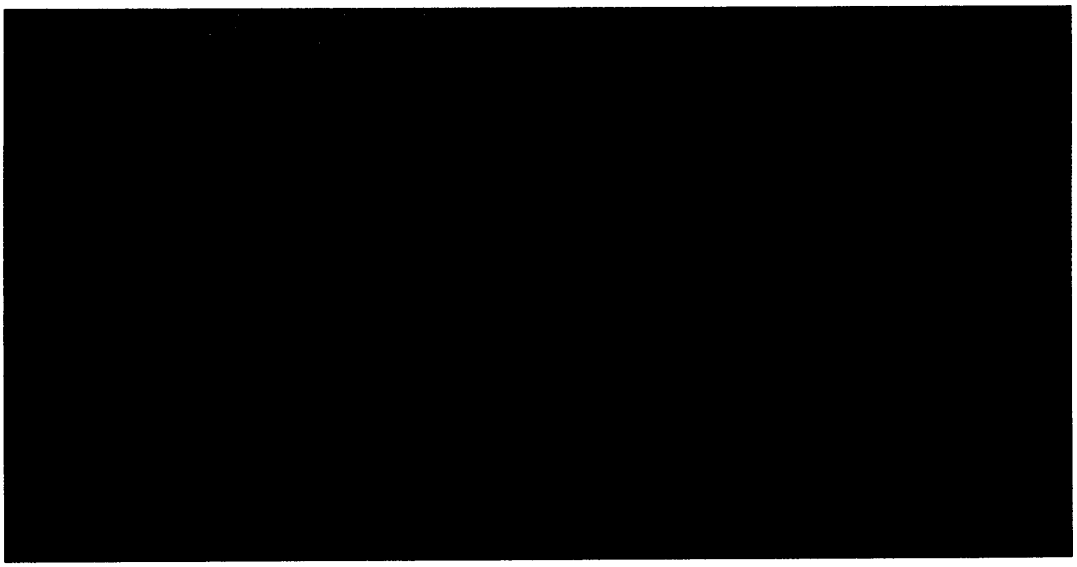
2.6.3.2. Results.

The *in-silico* subtraction resulted in a 'virtual library' of 405 clones (from the original set of 15,748). These formed 325 groups when annotated with SwissProt/TrEMBL top hits. A single clone from the 13 largest groups (i.e. those groups containing the most abundant sequences) were ordered from the MRC geneservice (web ref 18) and used as probes for *in-situ* hybridisation experiments (carried out by Eva Tiecke, Division of Cell and Developmental Biology, School of Life Sciences Research Centre, University of Dundee).

Eleven clones (~85%) were expressed in the limb and in other places, one was expressed in tissues other than the limb and expression of the final clone was undetectable. Low expression levels may be the reason why staining was undetectable within the limb for two clones. Figure 2.24 gives two example results from the *in-situ* hybridisations. This was just a preliminary examination to assess the viability of *in-silico* subtraction techniques. We also intend to perform a subtraction for the "stage 22 limbs" tissue against all other tissues in our database in order to discover transcripts unique to stage 22 limbs and to consider clones with no homologue in the SwissProt/TrEMBL database.

These preliminary results are promising, with a success rate of over eighty percent for the clones examined so far. It should be noted that the expression patterns of the clones were not limited to the limb bud, but were often quite widespread over the embryo. A whole database subtraction has the potential to provide clones that are more specific to the limb bud.

A)



B)

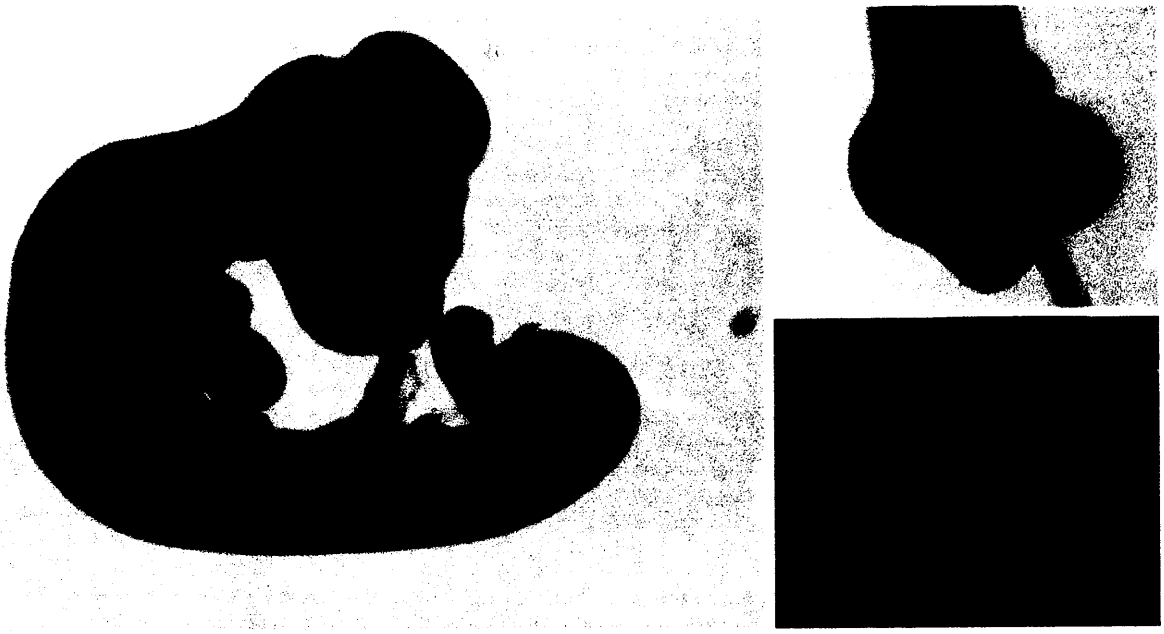


Figure 2.24. *In-situ* hybridisations to stage 22 chicken embryos.

A) Staining with clone ChEST429K9. Expression can be seen in the ventral side of the limb bud.
B) Staining with clone ChEST434B4. Expression can be seen in the apical ectodermal ridge (AER) or the limb buds, in the somites and the neural tube.

2.7 Web site & ftp site

To allow public access to the chicken EST data, an ftp site was set up to distribute the clipped EST sequences and assembled contigs. In addition, a web portal was created (web ref 12). Initially, this was designed for the dissemination of preliminary data and analyses to members of the BBSRC chicken EST consortium. Later this site was restructured and access was granted to the public. The main features of this site are: a BLAST server, a keyword search and an ID search.

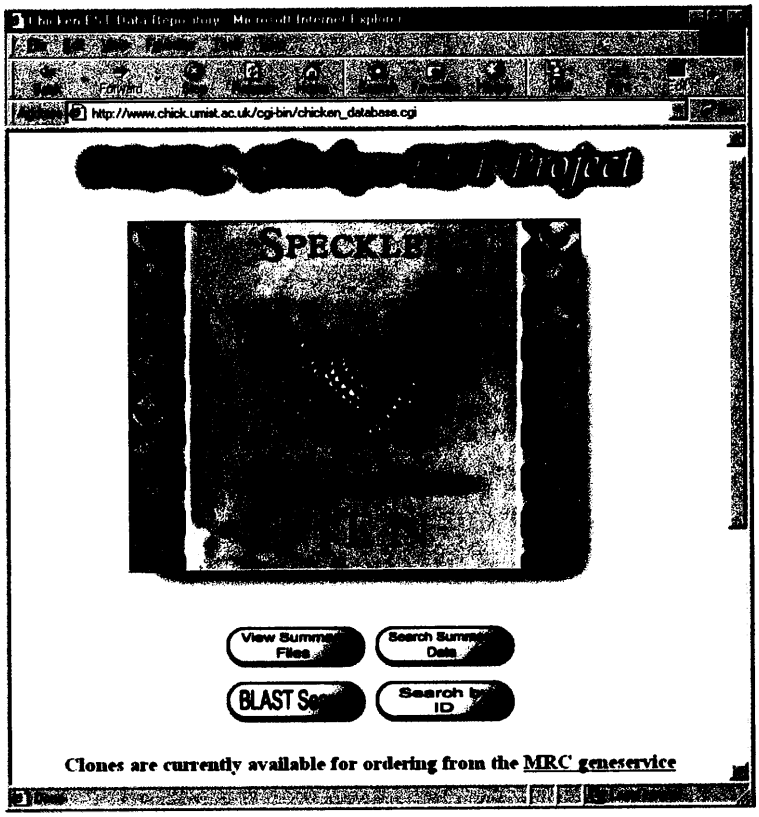


Figure 2.25. The BBSRC Chicken EST project main page.

2.7.1 BLAST facility

A local BLAST server was created to allow public queries against the chicken EST database (Figure 2.26). This supports the following BLAST programs: BLASTN (nucleotide query vs. nucleotide database), TBLASTN (protein sequence vs. translated nucleotide database) and TBLASTX (translated nucleotide query sequence vs. translated nucleotide database). Searches can be carried out against the entire EST dataset, ESTs separated by tissue or against the assembled consensus sequences.

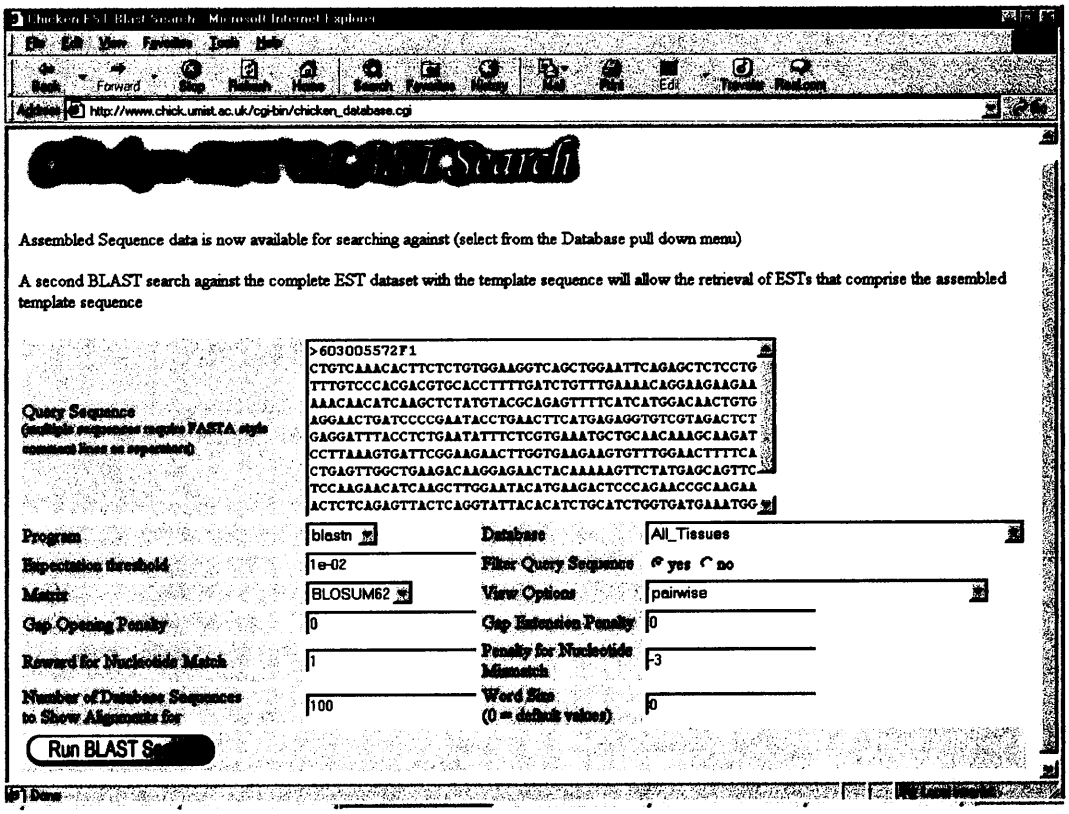


Figure 2.26. BLAST search page.

It was necessary to create a server program to process BLAST requests, control the number of BLAST jobs running on the server and to provide an intuitive interface.

2.7.1.1 blast_server

The ***blast_server*** program runs on the head node of the linux cluster (this is the only node visible to the internet). BLAST jobs are farmed out to this server from the main database CGI script (***chicken_database.cgi***) via TCP/IP. When BLAST requests are received by the server, one of two actions may occur:

1. If the maximum number of jobs are running, the server sends the user a 'job queued' web page and places the request at the bottom of a stack of pending jobs. The 'job queued' page is refreshed every five seconds. The web page is replaced when the BLAST job has been accepted (see below).

2. If the number of jobs running has not exceeded the maximum limit then a child process is spawned to process this BLAST request. The parent server is then free to accept/queue further requests. The child server generates an "in-progress" web page to notify the user that their job has been accepted and is currently running. This page refreshes every five seconds.

On completion, the BLAST output is parsed and edited to include links through to data on the EST or consensus sequences found in the search. The BLAST results file is moved/renamed and replaces the 'in progress' page viewed by the user. The child process then terminates. This event is detected by the server program, which prompts the execution of the next pending job (if any).

2.7.2 Keyword search

All ESTs and consensus sequences are annotated where possible (section 2.5). This annotation is stored in a relational database (section 2.3.2.6). A keyword search of the annotation was implemented and incorporated into the main web page. This provides an easy way to mine the EST data. The search can be refined in a number of ways; by tissue (whole EST dataset, consensus sequences or individual tissues); by the score of the BLAST hit used to annotate the sequence; and by either of the AND/OR logical operators used to combine multiple keywords (Figure 2.27).

The results of the keyword search are linked through to the following: data on the EST or consensus sequence matching the keywords, the original BLAST report used to annotate the sequence, the top hit SwissProt/TrEMBL entry. The output also contains the name of the sequences originating tissue (or "Assembled Sequence" in the case of a consensus sequence match), the number of blast matches to the SwissProt/TrEMBL database, the description of the top hit, the organism of the top hit sequence and the top hit's SwissProt/TrEMBL keywords (Figure 2.28). Clicking on an EST ID opens a sequence view for this EST (Figure 2.29) clicking on a contig ID brings up the consensus sequence view for the selected contig (Figure 2.30).

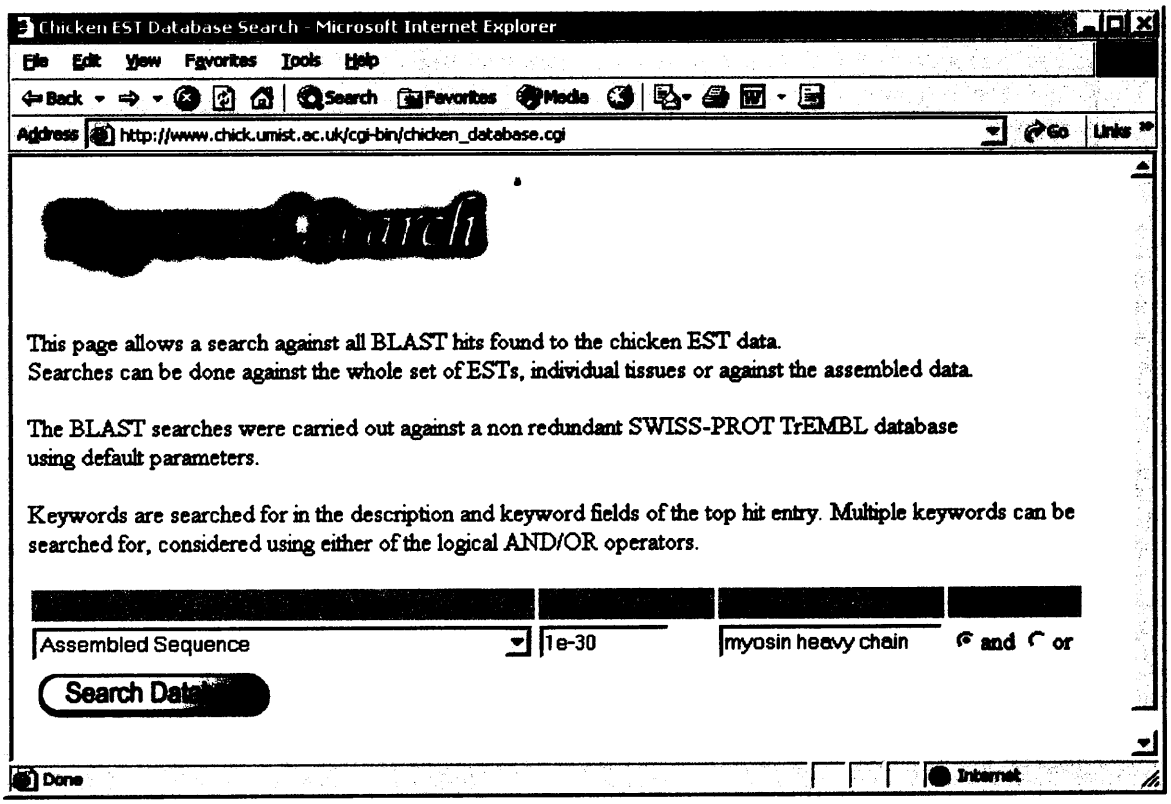


Figure 2.27. Keyword search page.

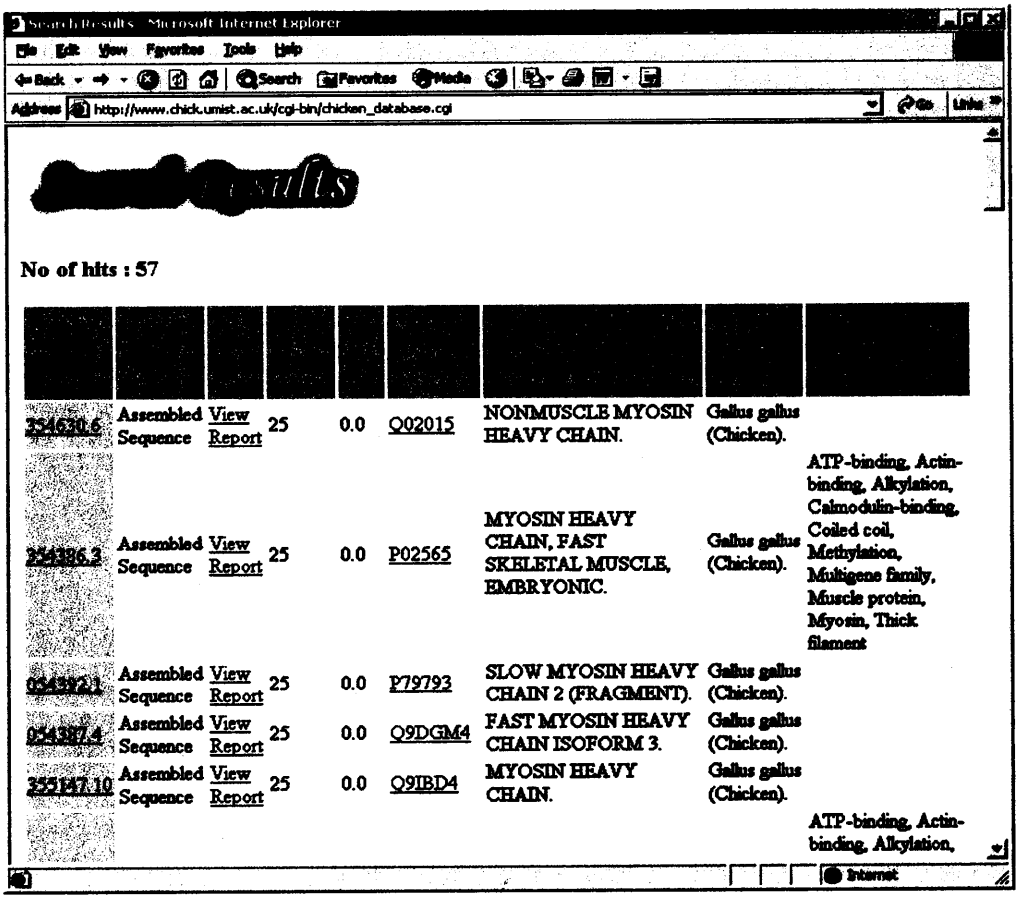


Figure 2.28. An example keyword search results page.

2.7.3 ID search and sequence view

The ID search allows users to retrieve data on ESTs or consensus sequences from their IDs. Users can search with the UMIST EST IDs, cloneIDs or with contig IDs (for the consensus sequences). EST sequences are displayed in FASTA format with any available annotation displayed underneath. The EST sequence view also contains a link through to the contig containing this sequence (Figure 2.29).

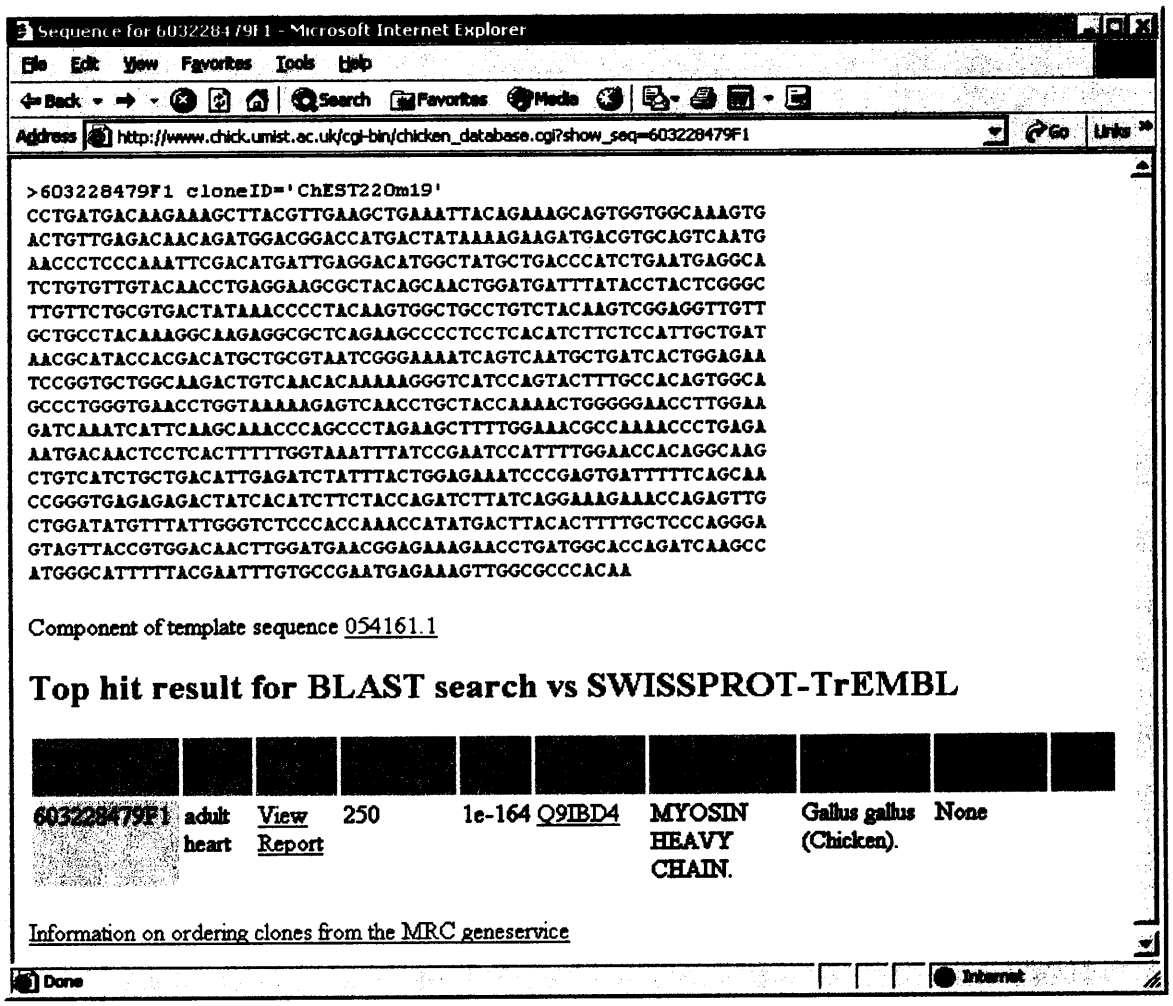


Figure 2.29. Sequence view page.

Consensus sequence data is presented as an image map showing the composition of the assembled contig (Figure 2.30). Clicking on an area of this image map takes the user to the sequence view page for the selected EST or consensus sequence (Figure 2.29).

The annotation information is currently restricted to the top-scoring hit (if any are found) from our BLASTX searches against the SwissProt/TrEMBL non-redundant databank. However, more detailed BLAST information can be returned by clicking on the "View Report" link which brings up the full BLAST report. Future additions to the website and database will allow users to view matches to both the Human and Mouse genomes.

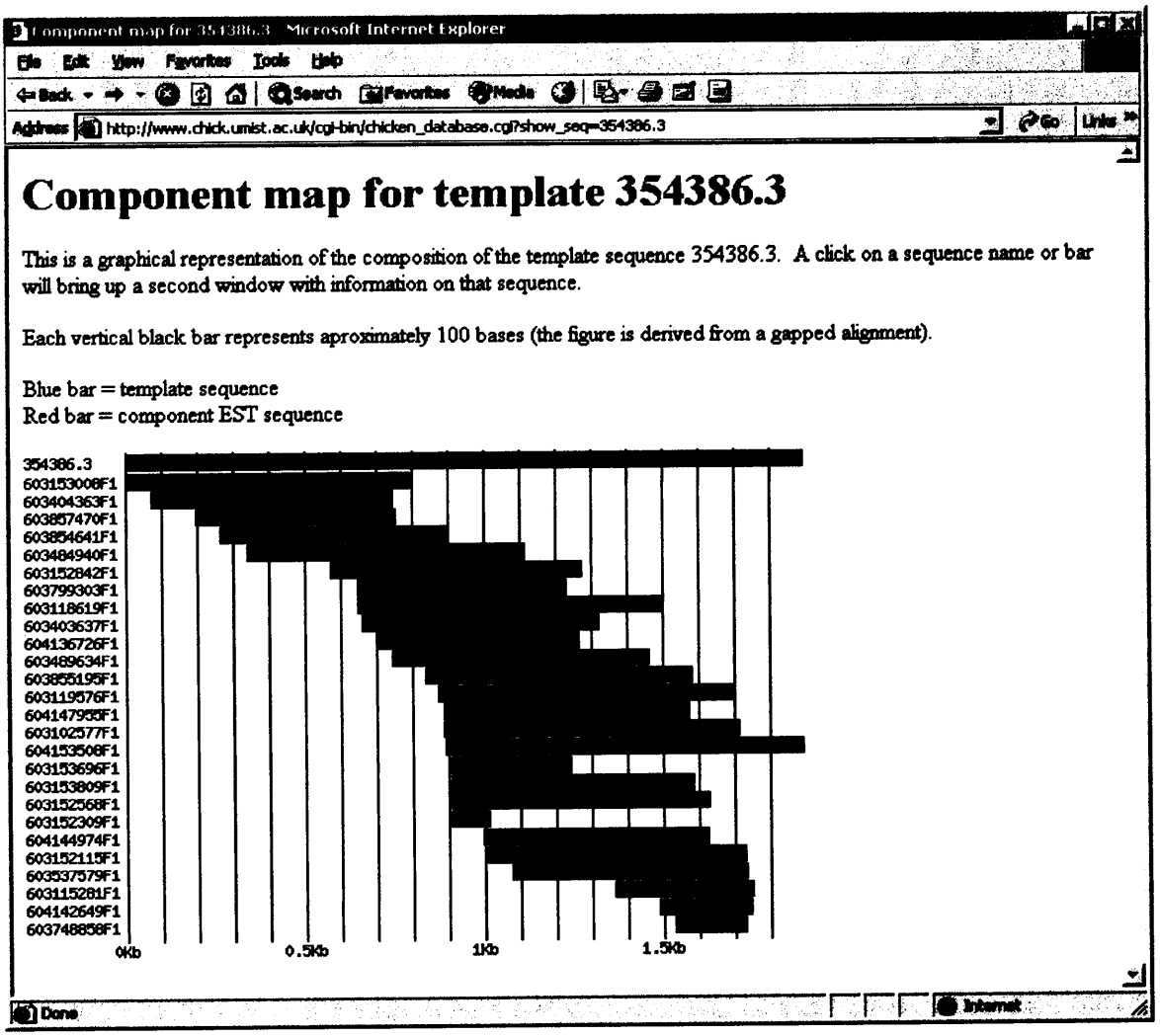


Figure 2.30. Consensus sequence view.

The blue bar represents the consensus sequence. Red bars represent the alignment positions of the component EST sequences. Vertical bars are spaced at 100 nt in the gapped alignment.

2.7.4 ftp site

An ftp site was created to allow public download of the data (web ref 19). The entire EST collection, all consensus sequences and ESTs subdivided into tissues are available for download from this site. We also included the assembly of all Genbank chicken ESTs in conjunction with the BBSRC ESTs on the ftp site.

2.7.5 Usage

The BBSRC Chicken EST web page became accessible to the public during December 2001. Since this time, the main page has been accessed over 38,000 times and over 43,000 BLAST searches have been run. These searches have been performed from many different labs all over the world (table 2.8).

Table 2.8. List of the top twenty top-level domains from IP addresses which have carried out BLAST searches via the chicken EST web site (as of 03/05/2003).

Top level domain	Country of origin	Number of BLAST searches
uk	United Kingdom	9155
edu	USA (Educational)	8690
jp	Japan	3675
il	Israel	1787
fr	France	1757
nl	Netherlands	1383
se	Sweden	1335
at	Austria	1105
com	Commercial	935
de	Germany	922
gov	USA (Government)	748
pt	Portugal	445
ca	Canada	390
cz	Czech Republic	382
org	Non-Profit Organisations	394
net	Network	366
es	Spain	332
au	Australia	259
ch	Switzerland	252
cl	Chile	95

2.8 Discussion

We have described the design, implementation, evaluation and publication of a large set of chicken ESTs and associated resources. With the publication of the 339.000 ESTs generated here we have dramatically increased the amount of genomic information available for scientists studying chickens and those wishing to use the DT40 cell line as an experimental system. Of the 85,486 contigs generated from this EST database, only 31,005 (36%) could be assigned a putative function through homology detection with the known databases. This leaves 54,481contigs with no form of functional annotation. Since our cDNA sequences were generated through poly-dT priming of the polyA tail of mRNA sequences, one would expect a large proportion of these unannotated contigs should represent protein coding mRNA sequences. With a smaller proportion being composed of non-coding sequences such as UTR sequence, genomic DNA, miRNAs and other forms of RNA sequence. We are currently developing an open reading frame prediction program to estimate the protein coding potential of the non-annotated contigs. In an analysis on 60,770 full-length mouse cDNAs (Okazaki 2002) the FANTOM consortium found 33,409 unique clusters (which they termed transcriptional units). Of these, 15,815 (47%) were found to be functional non-coding RNAs (RNAs with no apparent protein-coding region). This class of RNA may also account for a large proportion of our EST dataset, which would offer some explanation for the low percentage of BLASTX hits.

This figure is in keeping with those seen for other large-scale analyses of completed genomes. When the genome sequence for the yeast *Saccharomyces* cerevisiae was completed, functions could be assigned to approximately 43% of the predicted proteome (Mewes 1997). Okazaki *et al.* (2002) were able to assign putative functions to approximately 48% of their 60,770 full-length mouse cDNAs through similarity searches with the public databases. The genome of the worm *Caenorhabditis elegans* is predicted to contain a similar number of genes to both mouse and human, but only 42% of these could be assigned a putative function through homology based techniques (The *C. elegans* Sequencing Consortium 1998).

It is interesting to note that our chicken gene number estimation (Section 2.5.5) reveals a figure in the region of 30 → 35K genes. This is in agreement with gene

number estimations for other vertebrates such as *Fugu rubripes* and *Homo sapiens*, which suggests a common baseline of around 35,000 genes for all vertebrates. Of these ~30K genes, we estimate that our EST dataset contains a minimum of 12K individual chicken genes. This was calculated by counting unique accession numbers from the top hit of each EST's BLASTX search against SwissProt/TrEMBL. There may be many more genes present in the database than this figure suggests. Since any given completed genome fails to annotate ~50% of the predicted genes, it seems feasible that our chicken EST dataset potentially contains up to 24K genes.

The chicken genome is currently being sequenced and the draft version should be completed by the end of the year. A full-length cDNA project is also being carried out here at UMIST, where we are working with the Sanger centre to define 10,000 full-length sequences from our cDNA clone set. These projects will provide a wealth of invaluable information and resources to the scientific community.

Understandably, non-homology based techniques for the prediction of gene/protein function has been a 'hot topic' recently. An overview of a selection of these techniques was given in Section 1.4. In the following chapters, we describe an exploration of the use of other forms of genomic data for the prediction of gene function. More specifically, we are interested in transcription factor binding sites and the possibility of exploiting conserved patterns of these sites to predict the function of the downstream gene. This form of genomic information was chosen in light of the results from large-scale microarray experiments. In many of these, it would appear that genes exhibiting correlated expression patterns also have related functions. Many of these correlated genes are also seen to be under the influence of common binding proteins, identified through virtue of conserved sequences upstream of the genes in question. We use the yeast *Saccharomyces cerevisiae* as a model organism for this analysis. Yeast has a small genome and a high level of expert knowledge available on much of its genome and proteome. In addition, many yeast systems generally represent simplified versions of their higher eukaryotic counterparts enabling investigators to discover rules and relationships in a simple organism that are often applicable in higher organisms. Much of our knowledge on chromatin remodelling, transcriptional regulation and transcription initiation comes from experiments in yeast, which makes yeast an ideal choice for this analysis.

Analyses of Promoter Regions in the Yeast Genome

3.1 Regulation of transcription

The method of transcriptional control in prokaryotes was elucidated some 30 years ago. From that point attention turned to transcriptional control in eukaryotes where there is a much greater level of complexity due to the necessity for cell-type specific and developmental regulation of a much larger complement of genes. Some of the basic features of the prokaryotic system have been conserved with the addition of new layers of complexity to the transcription apparatus (Kornberg 1999a).

The logic of gene regulation in prokaryotes and eukaryotes is fundamentally different. In prokaryotic organisms, control of gene expression is primarily through repression where as in eukaryotes transcriptional control is often through activation of transcription. This fundamental difference is primarily because eukaryotic DNA is packaged into chromatin, which prevents binding of TBP (TATA Binding Protein) and other factors required for transcription.

Prokaryotic RNA polymerases recognise promoters via specific sequences immediately upstream of the initiation site. Efficient transcription is seen, *in vitro*, on purified DNA templates, with the rate and level of transcription determined solely by the quality of the promoter sequences (Struhl 1999). This property is mirrored *in vivo* showing that there is no inherent restriction on the ability of prokaryotic RNA polymerase to bind the DNA template and initiate transcription. Prokaryotes repress transcription through occlusion of RNA polymerase binding sites or by generating repressosome structures (Geanacopoulos 1999). Regulation through activation is only required where the promoters are inherently weak or repressed.

3.1.1 Nucleosomal repression

In eukaryotes, histone-dependent packaging of genomic DNA into chromatin is a central mechanism for gene regulation. At the heart of the chromatin structure is the nucleosome. This comprises 146 base pairs of DNA wrapped round a histone octamer.

This is composed of two copies each of the highly conserved histone proteins: H1, H2A, H2B, H3 and H4. These function as building blocks to package eukaryotic DNA into repeating nucleosomal units that are folded into higher-order chromatin fibres (Luger 1998, Kornberg 1999b, Strahl 2000).

The structural basis for repression by nucleosomes is due to the histone configuration. Each histone is organised in two domains, a characteristic 'histone fold' and an unstructured N-terminal tail (Kornberg 1999b). The histone-fold domains constrain the DNA in a central core particle limiting access by transcription factors and thereby repressing gene expression. The tails extend outside the core particle, providing points of interaction for gene activation, higher-order coiling and condensation in heterochromatin. Residues in these tails are subject to a wide range of post-translational modifications, including phosphorylation, acetylation, methylation and ubiquitination. These modifications have been shown to affect chromosome function through two distinct mechanisms. First, most of these modifications affect the electrostatic charge of the histone and this could change the structural properties of the histone and its binding to DNA. Secondly, these modifications could create binding surfaces to recruit specific functional complexes (Dhalluin 1999, Jacobs 2002). Some groups believe that these form a 'histone code' that governs gene expression by controlling the higher-order structure of chromatin (Strahl 2000, McKinsey 2002). Indeed, there is now evidence for complex interactions among these different modification activities (Rea 2000) and examples of a modification on one tail governing a different modification on another tail *in trans* (Sun 2002, Dover 2002).

3.1.2 Activators and enhancers

Analysis of eukaryotic transcriptional regulation has concentrated mainly on activator proteins that have a positive effect on transcription when bound to DNA control elements called enhancers. Many activators have been identified that are specific for genes or gene-families and, typically, couple transcription to the physiological needs of the cell (Kornberg 1999a). Activators have distinct DNA binding and activation domains. The structures of many different types of binding domains have been solved and these usually take one of a few well-characterised motifs (e.g. leucine zipper, zinc finger, helix-turn-helix). The activation domains on the other hand are less well

understood. It is known that these domains participate in protein-protein interactions with other parts of the transcription machinery and this action influences gene expression.

It is thought that activators exert a positive effect on transcription by recruiting chromatin-modifying complexes that relieve nucleosomal repression *via* histone acetylation. The reverse mechanism (deacetylation) has also been observed for a set of proteins called repressors that have a negative effect on transcription by re-establishing nucleosome repression (Kadonaga 1998). Activators and repressors, referred to collectively as transcription factors, can also have a more direct effect on the RNA polymerase through interactions with a complex called Mediator (Malik 2000).

3.1.3 Mediator

Mediator is a multi-protein complex widely recognised as an important interface between activators and RNA polymerase II. The first indications of a common target for transcriptional activators came in 1988 when interference was detected between different gene activator proteins in an *in vitro* transcription system (Gill 1988). The inhibition could not be relieved through the presence of excess of general initiation factors, indicating that the common target is distinct from the basal transcription machinery. Mediator was first isolated in yeast by searching for a factor that relieved the interference of activators and thus could be a common target (Kelleher 1990). Experiments on the purified Mediator fraction showed that it was capable of reversing the interference seen previously. When purified to homogeneity it was seen that this fraction was a holoenzyme form of RNA polymerase II, made up of core polymerase and a Mediator complex (Kim 1994). Mediator was subsequently purified and found to be a multiprotein complex consisting of 20 individual polypeptides (Myers 1998).

The mechanism of Mediator-dependent transcriptional activation is sill unknown. Interactions have been detected between multiple activator proteins and Mediator subunits (Park 2000). It is most likely that Mediator plays an important role in the recruitment of RNA pol II to the preinitiation complex. This is supported by the observation that RNA pol II engaged in active transcription lacks associated Mediator but the formation of the preinitiation complex is dependent on the holoenzyme form.

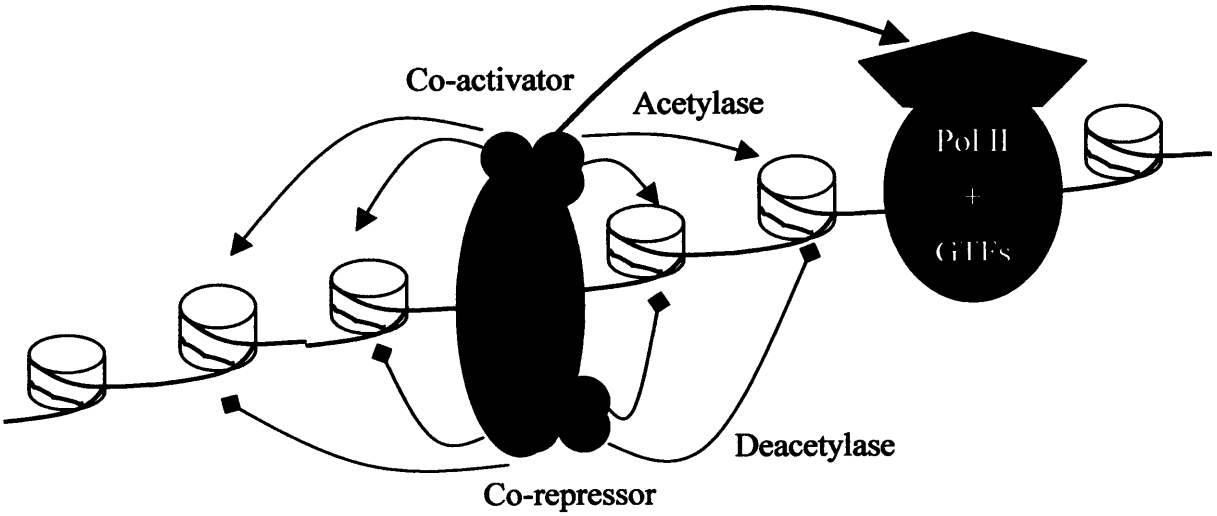


Figure 3.1. A general view of transcriptional regulation.

A transcription factor (TF), either an activator or repressor, binds to a regulatory element (enhancer or silencer respectively) and effects transcription by affecting the acetylation state of the surrounding histone octamers (white cylinders). Activators also act by direct interaction with Mediator (orange).

A somewhat surprising discovery was that only six of the 20 yeast Mediator subunits have a detectable homologue in mouse and human (Gustafsson 2001). This has led to the suggestion that metazoan Mediator is significantly different in both structure and function from the yeast Mediator. There is an alternative to this hypothesis; an important function of Mediator may be to present a dynamic interface between rapidly evolving gene-specific regulatory proteins and the highly conserved basal transcription machinery (Gustafsson 2001). A conserved Mediator core of 6-10 proteins is responsible for contacts formed with RNA pol II and TFIIH. The remaining species-specific subunits are responsible for interactions with gene-specific repressors and activators. This hypothesis is supported by a comparison of *S. cerevisiae* and *S. pombe* Mediator complexes (Spahr 2000). This analysis revealed that only the *S. cerevisiae* Mediator subunits encoded by essential genes have a detectable homologue in *S. pombe*. Further evidence is revealed by a comparison with metazoan Mediator. Six out of the 10 essential subunits have conserved homologues in the metazoan Mediator (Boube 2000, Malik 2000)

3.1.3 Non-coding RNAs

Approximately 98% of all transcriptional output in humans is non-coding RNA (Mattick 2001a). Introns account for 95% of the pre-mRNA transcripts of protein coding genes and other non-coding RNAs represent half to three quarters of all transcription from the genomes of higher organisms (Davidson 1977, Matick 2001b). It has been proposed that these intronic and other non-coding RNAs have evolved to provide a second level of transcriptional control in eukaryotes. Some evidence to support this is provided by an investigation into the bithorax-abdominal A/B complex of *Drosophila melanogaster.* This gene covers 200 kb and produces seven major transcripts. Only three of these transcripts contain protein-coding sequences but all are spatially and temporally regulated. The interruption or deletion of the DNA encoding the non-protein transcripts has known phenotypic consequences (Lipshitz 1987, Sanches-Herrero 1989).

Many proteins that are considered to be transcription factors, such as the zinc finger proteins Sp1 and WT1, winged helix-turn-helix proteins and Y-box (cold shock) proteins, appear to bind RNA or RNA-DNA hybrids. This implies that these proteins may be interacting with higher order structures formed by RNA rather than recognising the primary DNA sequence (Shi 1995, Fierro-Monti 2000).

There is also evidence that RNA regulates chromatin structure. Transfection, co-suppression and transgene silencing have all been shown to require a group of proteins known as the Polycomb proteins. These proteins are involved in chromatin remodelling via histone deacetylation. Transfection has also been shown to involve *trans*-acting RNA signals, which has led to the suggestion that RNA may be involved in chromatin remodelling by the recruitment of Polycomb complexes in a gene specific manner. It has also been shown that the Polycomb proteins (as well as other chromatin remodelling proteins) contain a conserved RNA binding domain, called a chromodomain (Akhtar 2000), which controls sequence and target specificity (Jones 2000). The transcriptional control of certain steroid receptors have been shown to require chromatin remodelling and the recruitment of histone acetyltransferases. Transcriptional co-activation of these genes involves the action of a non-coding RNA (Lanz 1999).

3.2 Simple vector-based comparison methods for gene association

One of the most obvious ways to attempt the association of functionally related genes using binding site data is through a simple search for the presence or absence of known binding sites in their upstream regulatory sequences (URSs). Indeed, microarray analyses have found conserved sites in the URSs of clusters of co-regulated genes using statistical measures to judge the over-representation of certain words in these sequences (van Helden 1998, Hampson 2002, Sinah 2002). These usually find single sites, which are present in the majority of the clustered sequences.

3.2.1 Methods

There are multiple public and private databases that specialise in providing information on transcription factors and their DNA binding sites (TRANSFAC, SCPD, TFD, TRRD, COMPEL, to name but a few). Two of the most popular databases at the time of this analysis were TRANSFAC (Wingender 1997, web ref 20) and SCPD (Zhu 1999, web ref 21). Both of these databases contain information on mapped binding sites, consensus sequences and matrices (Table 3.1). The DNA binding site data for sites from *S. cerevisiae* were obtained from these two databases.

Table 3.1. Comparison of TRANSFAC and SCPD databases.

	TRANSFAC	*SCPD*
Sites	312 (include artificial sites)	580
Factors	159 (include TFs with no mapped sites)	103
Consensus sequences	21	48
Matrices	21	24

Yeast regulatory systems are typically located within 800 – 1000 base pairs upstream of the translation start site of the genes they control (Figure 3.2, Zhu 1999, Kruglyak 2000). The set of sequences covering 800 bases upstream of all genes from *S. cerevisiae* were obtained from the SGD web-site (Weng 2003, web refs 22 and 23).

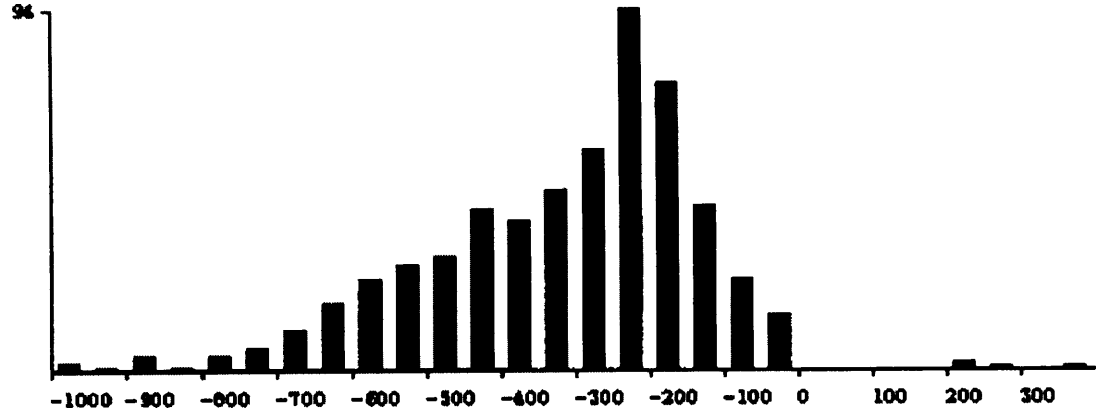


Figure 3.2. Distribution of all mapped sites present in the SCPD (Figure taken from Zhu 1999).

The Y-axis represents the total number of experimentally mapped sites determined at any given position in an URS in the yeast genome. The translational start site is at position 0.

The GCG program, ***findpatterns***, was used to search for binding sites in the URSs of all yeast genes. These pattern searches were initially carried out with three sets of binding site data: SCPD consensus sequences, TRANSFAC mapped sites and a redundant dataset combining the two.

Consensus sequences are derived from alignments of mapped sites for individual transcription factors. These were used in preference to mapped sites, as they are believed to better reflect the action of binding at regulatory sites. Indeed, the sequence variability of binding sites affects the binding affinity of transcription factors and this can therefore provide an additional, finer level of control over gene expression (Stormo 2000).

Preliminary studies suggested that the simple presence or absence of a single mapped site is insufficient for identifying genes with a common regulatory mechanism. For example, in an analysis by DeRisi *et al.* (1997), 17 genes were found to have their expression levels increased by over twofold when the transcriptional activator Yap1 was overexpressed. Only two-thirds of these genes contain a Yap1 binding site in the 600 bases upstream of their start codon. Over 1,500 URSs in the yeast genome contain at least one Yap1 binding site and 202 URSs contain two or more Yap1 sites (data not shown), the majority of which do not exhibit an increase in gene expression levels on

overexpression of Yap1. This highlights the ambiguities inherent in attempting to describe trends in transcriptional activation/repression in terms of single transcription factors and their associated binding sites. In an attempt to address this issue, we consider the presence/absence of all known binding sites when comparing the URSs of genes.

The binding profile of an URS can be represented simply as a vector with each position in the vector defining the presence of absence of a binding site (Table 3.2). The program, ***vectorsFROMfp***, was developed to produce a database of vectors from the results of a findpatterns search.

Table 3.2. Example binding profiles for the URSs of the first three genes in the *cerevisiae* genome.

URS	Transcription Factor Binding Site				
	ABF1	*ACE2*	*ORC*	*PHO4*	...
YAL001C	0	1	1	0	...
YAL002W	0	0	0	1	...
YAL003W	0	0	0	0	...

The representation of binding data as a vector allows the pairwise comparison of URSs using metrics such as Euclidean distance (Equation 3.1) and the correlation coefficient (Equation 3.2). The clustering programs used in microarray analyses take vectors as input, allowing the use of extant clustering programs (such as the program described in Eisen 1998) to cluster the binding site vectors for *S. cerevisiae*.

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots (x_i - y_i)^2}$$

Equation 3.1. Euclidean Distance

$$c = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Equation 3.2. The Correlation Coefficient.

Initial analyses noted only the presence or absence of a given binding site in the URSs of genes. Multiple copies of a single site were essentially ignored. The copy number of a site is potentially important information. Later versions of the ***vectorsFROMfp*** program were modified to include the copy number of each site in the vector outputs.

The following datasets were clustered using the Eisen's cluster program and visualised with TreeView: SCPD consensus sequences, SCPD mapped sites, TRANSFAC mapped sites and a dataset combining the SCPD and TRANSFAC mapped sites. Each SCPD consensus sequence is found in a large proportion of the URSs of *S.* cerevisiae, this represents a large amount of noise and suggests that consensus sequences do not provide any discriminatory capacity for this analysis, because of this these sites were only used for the initial binary analyses (presence/absence of site noted).

3.2.2 Results

Very few tight clusters were formed from either the simple presence/absence analysis or the copy number analysis. Figure 3.3 shows a screenshot of a visualisation of ~100 URSs from a clustering run using the TRANSFAC mapped sites. There is a single, obvious, cluster containing 14 URSs in the figure. The associated genes for these URSs are all unclassified. The main results from this analysis are:-

1. Only a small minority of URSs form tight clusters. This is indicative of a very low discriminatory capacity.
2. The few clusters that are apparent consist of genes of unknown function or genes in close physical proximity to each other.
3. Some binding sites have a very high frequency of occurrence. For example, the GCN4 consensus sequence (TGANTN) or its reverse compliment (NANTCA) occur in 6211 of all yeast URSs and at least one of the 20 GCN4 mapped sites are found in 5775 URSs.

Appendix 3 contains details on 14 clusters that were present in results for all binding site datasets. There are 62 genes (out of 6,217) that show correlated patterns in the presence or absence of binding sites in their URSs. These were analysed using the yeast gene duplications web site (web ref 24), which provides data on all potential duplications within the genome of *S. cerevisiae*, and by generating multiple sequence alignments of the URSs and ORFs from each cluster (see web ref 25 for supplemental data). This revealed an extremely high degree of sequence similarity within each cluster, which is indicative of duplication events. This disappointing result indicates that the clustering analysis is only capable of associating sequences which have a high degree of sequence similarity and as such offers no improvement on current homology based techniques.

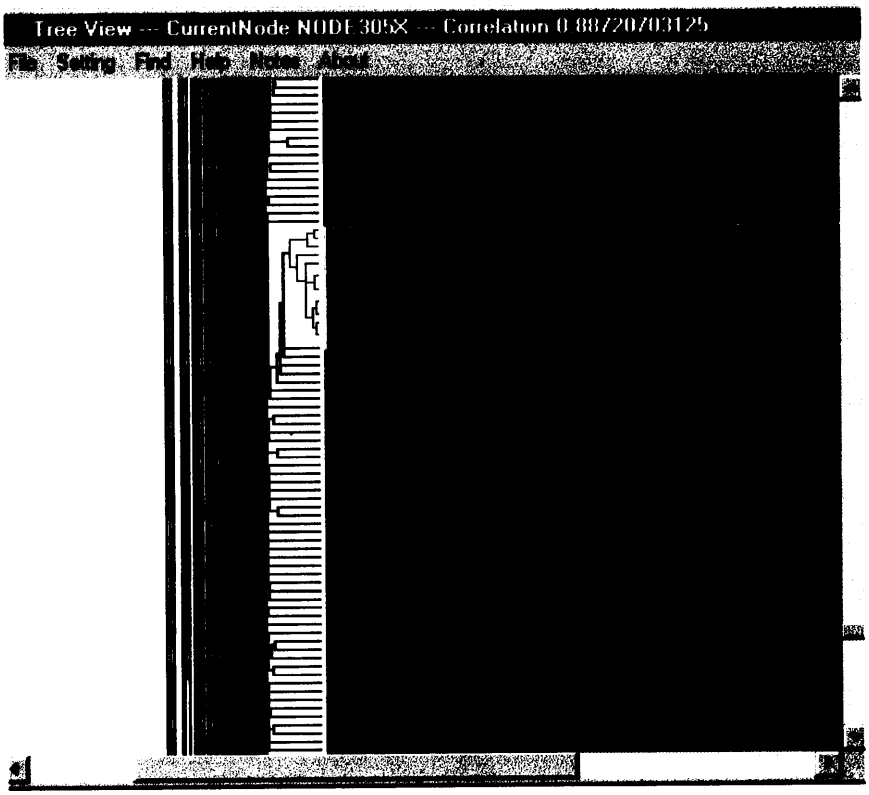


Figure 3.3. TreeView visualisation depicting the results of clustering URSs using TRANSFAC mapped sites.

The highlighted cluster is composed solely of unclassified genes. The sequences of the genes and their associated URSs are all highly similar and are thought to be related through duplication events within the genome (see Appendix 3 for multiple sequence alignments of the proteins and associated promoter regions of the genes in this cluster).

3.3 Binding site frequencies

3.3.1 Introduction

Many sequences present in the transcription factor binding site databases are either over or under represented in the yeast genome. Some consensus sequences in the SCPD dataset exhibit such a lax binding specificity that at least a single representative (and often multiple copies) of the site can be found in nearly every promoter region in the genome. On the other hand, many of the mapped sites have only a single occurrence in the entire genome. These high and low copy number sites may have a detrimental effect on the vector clustering results. The low count sites contribute to the creation of sparse vectors and the high-count sites may introduce noise to the system. The frequency of occurrence of binding sites within the complete set of *cerevisiae* URSs was calculated. This allows the assessment of the information content of the binding site datasets and allows the discrimination between low and high information sites (low information sites being those sites that exhibit one of the two extreme binding characteristics).

Random sequence datasets were produced in order to compare the frequency of occurrence of binding sites in actual genomic URSs with that expected by random. The program, ***randomDNAseqgen***, was written to create random nucleotide datasets with user defined GC content (For example, *Saccharomyces cerevisiae* has a GC content of 34.9% in non-coding DNA and 39.7% in coding DNA). A GC content of 34.9% (that of yeast URSs) was used to create the plots shown in section 3.3.2.

3.3.2 Results

Figures 3.4 and 3.5 show the frequency of occurrence for the SCPD consensus sequences and the TRANSFAC mapped site datasets within the complete set of yeast URSs. These plots show both the total number of occurrences of a site and the total number of URSs found to contain a particular site. It would appear, from these results, that the frequency of occurrence of the sites in the yeast URSs is not readily distinguishable from that for a randomly generated set of sequences. This reinforces the study by Audic and Claverie (1998), which found that, in the absence of a method to prelocate the core promoter regions, the use of consensus sequences and weight matrices to locate binding sites in promoter regions is essentially useless.

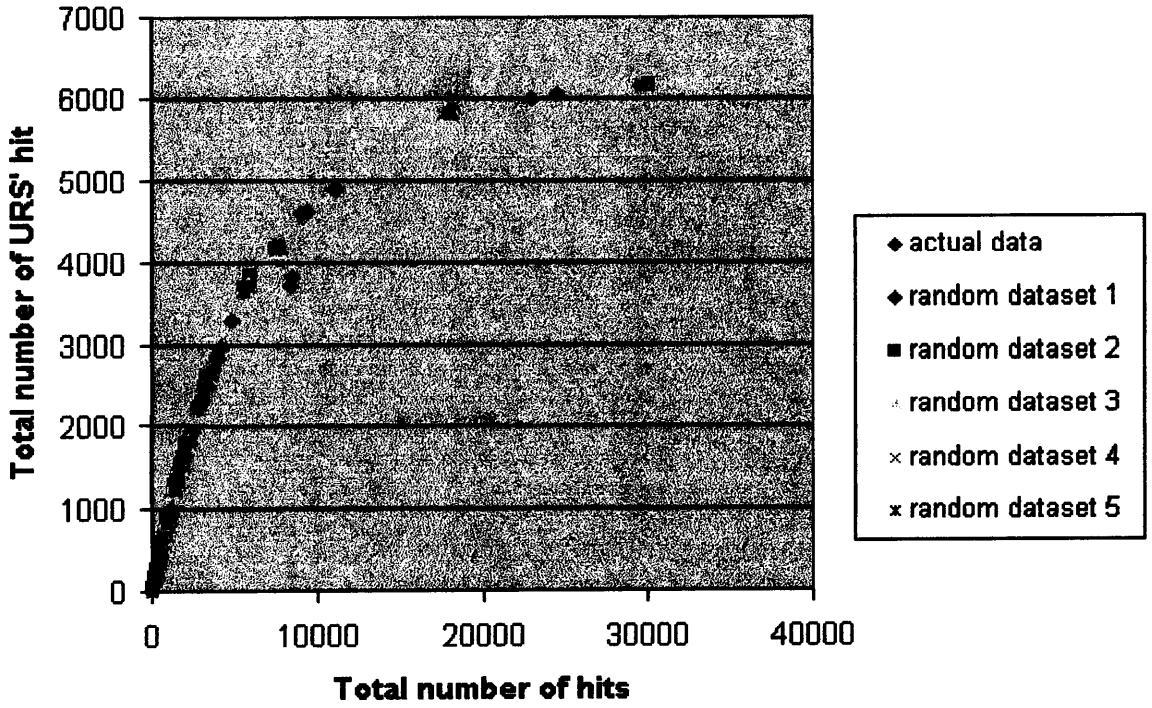


Figure 3.4. Binding statistics of SCPD consensus sequences for five random datasets and for all promoter regions in the yeast genome.

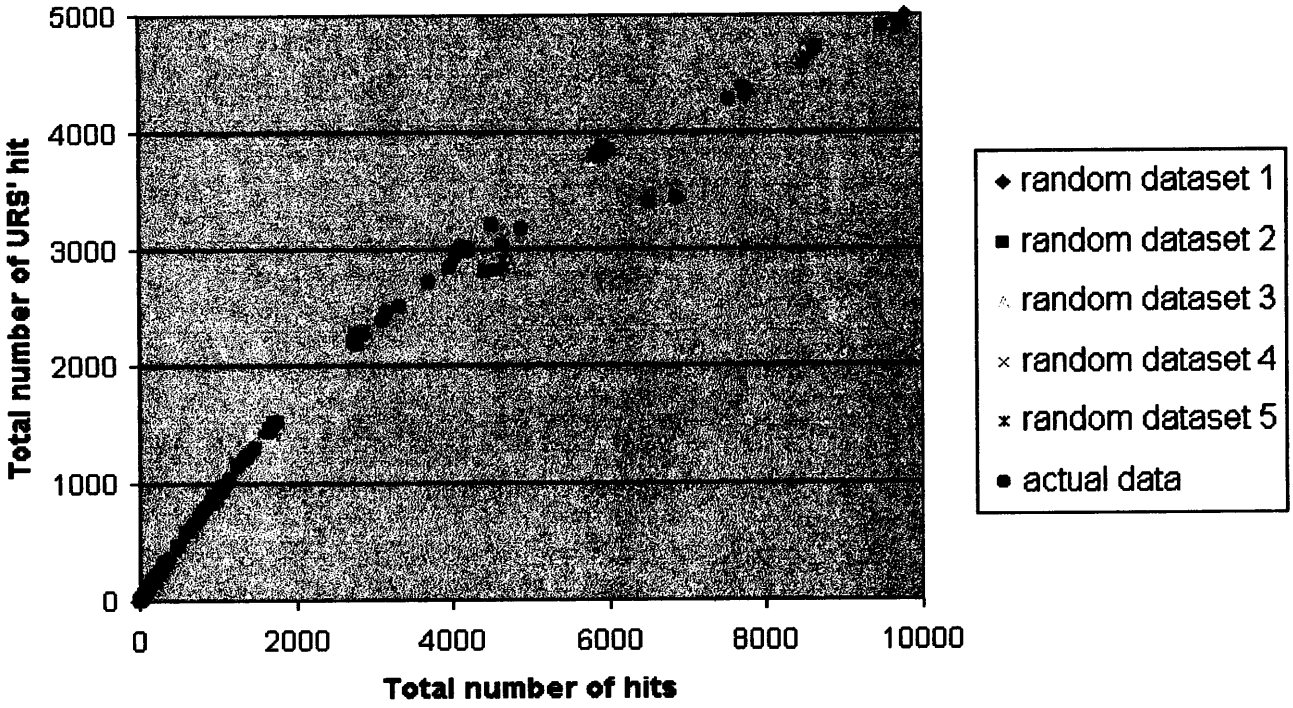


Figure 3.5. Binding statistics of TRANSFAC mapped sites for five random datasets and for all promoter regions in the yeast genome.

Transcription factors face the same problem in the cell. Experiments have shown that this problem is overcome through interactions with other factors bound to neighbouring sites (Bucher 1999). Transcription factor sites require a specific context (usually other TF sites) to elicit a biological response. The context required to initiate transcription typically contains a whole set of binding sites and is called a promoter (Werner 2003). Many predictive algorithms now use sequence contextual features, such as predicted neighbouring elements, to distinguish between functional and biologically irrelevant binding sites (Prestridge 1995). The COMPEL database (Kel-Margoulis 2000) represents a systematic attempt to collect all synergistic or antagonistic pairs of TF binding sites in all organisms.

3.3.3 Clustering of reduced binding site dataset

Functional categories in MIPS and KEGG rarely have more than a few hundred members. From this it follows that sites found in thousands of URSs within the genome are unlikely to be informative, at least on their own. Similarly, low copy number sites (those that are represented 5 times or less) are also non-informative. Therefore, for further analyses, the data from section 3.3.2 was used to create a "reduced" set of binding data with the low information sites subtracted.

Clustering these 'refined' datasets did not have the desired effect however (quite the opposite, in fact). Even fewer URSs form tight clusters (data not shown), and those that do have a very high level of sequence similarity. This implies that the clustered genes are related through duplication (as previously seen in section 3.2.2). A further interesting result is that when the number of sites used in the vector comparison is reduced, there is an increase in the level of negatively correlated pairwise comparisons. The reasons behind this are revealed in the next set of analyses.

3.4 Complementary analysis of coding and regulatory regions

Before considering a method to incorporate contextual relationships, we thought it prudent to evaluate the orthogonal data content of the URSs. That is to say that we wanted to see if there really is more information available in the regulatory sequence of a gene than that detectable by sequence comparison alone. Furthermore, this would define some quantitative measure of how many URS pairs were apparently generated from recent gene duplications and shared very high sequence similarity. The experimental procedure for this analysis is detailed in Figure 3.6.

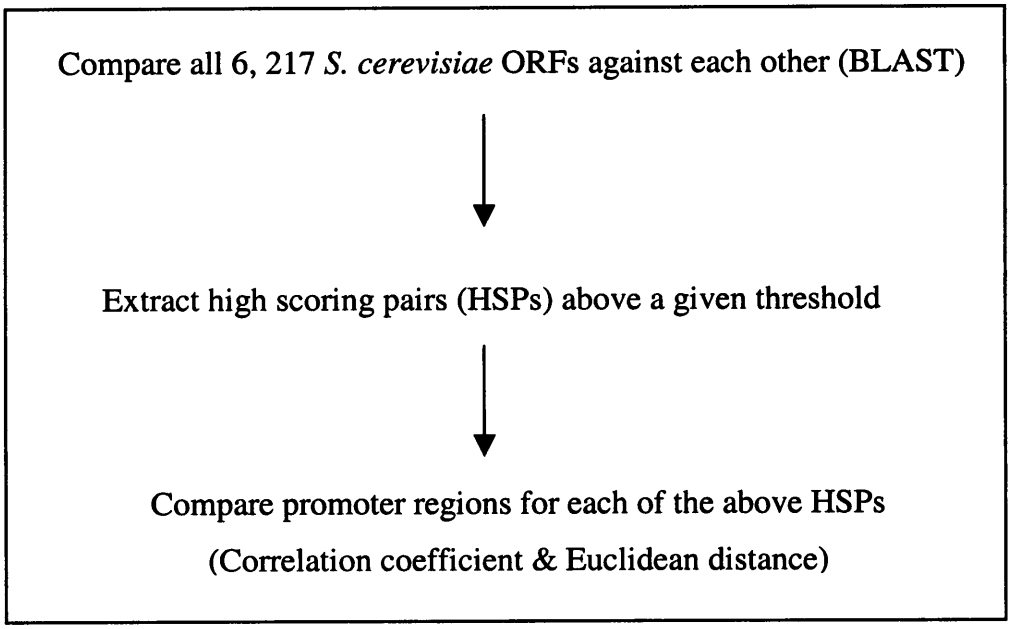


Figure 3.6. Methodology for comparison of URSs and ORFs.

If there is no extra data to be gained from an analysis of the URSs then, as the similarity of the ORFs increases, so too will the similarity of their corresponding URSs. On the other hand, if there is orthogonal information in the URSs then there should be no correlation between the similarities of the ORFs and that of the URSs.

3.4.1. Comparison of ORFs

The entire set of protein sequences for *S. cerevisiae* was downloaded from the MIPS ftp site (web ref 26). This dataset was compared with itself using the BLASTP program run with default parameters. The program, ***getGoodBLASThits***, was created to extract HSPs complying with user-defined thresholds for a combination of alignment length, percentage identity and e-value.

3.4.2 Comparison of URSs

The vectors representing the URSs of each HSP were extracted. The similarity between the two URSs was estimated using two distance metrics: the correlation coefficient and Euclidean distance. The program, ***goodHitsPlusVectorData***, was created for this task. This program takes as input a results file from ***getGoodBLASThits*** and a vector database created by ***vectorsFROMfp*** (Figure 3.7).

3.4.3 Random comparisons

The program, ***xlin***, was created to randomly select and compare vectors from a database. This allows a comparison of the true results with random.

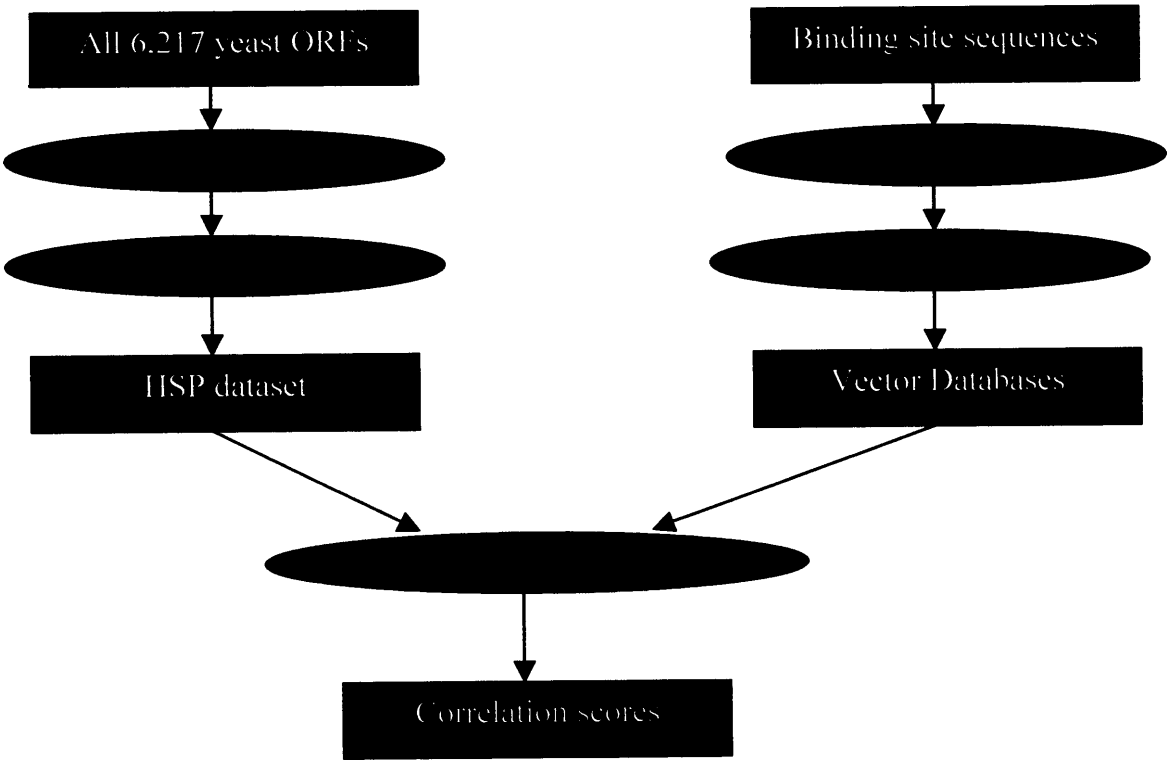


Figure 3.7. Methodology for the comparison of URS and ORF similarities.

Blue boxes represent datasets, light orange ovals represent publicly available programs and dark orange ovals represent programs developed specifically for this analysis.

3.4.3 Results

This method shows a large bias towards positive correlation scores for the pairwise comparisons of URS vectors (Figure 3.8). Each URS contains only a small number of the total possible compliment of TF binding sites. This results in a sparse vector where the majority of values are zero. For example, the vector representation for the promoter region of YAL001C from the SCPD consensus sequence dataset is an array with 92 entries. Of these, 79 have a zero value. This is much more pronounced for the mapped site datasets. The TRANSFAC mapped site vector for YAL001C has 518 zero values out of 534. This bias towards sparse vectors results in a false positive correlation as the absence of multiple sites is taken as a positive relationship between two URSs. In an attempt to remedy this, our application of the correlation coefficient was modified to only compare regions where at least one of the two URSs contain a positive number of sites. The results from this are much more evenly spread showing a reduced positive bias and some values well into the negative range (Figure 3.9).

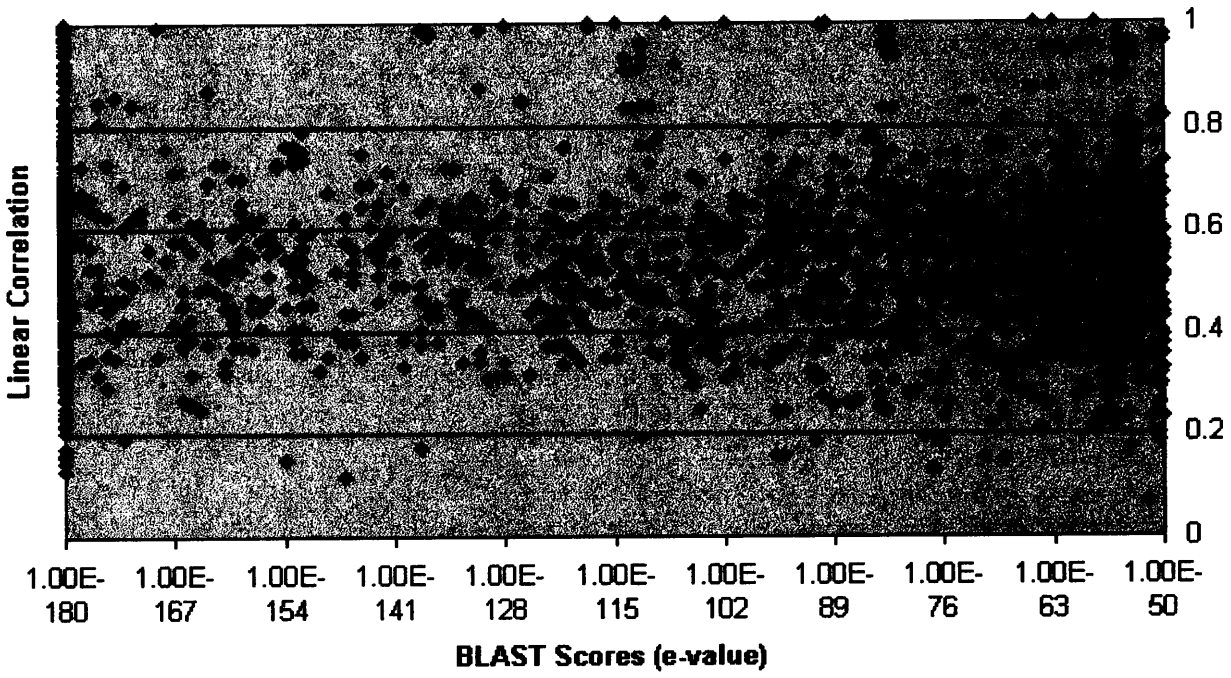


Figure 3.8. Initial results from URS and ORF comparisons.

This data was produced using a combined dataset (SCPD consensus and TRANSFAC mapped sites) and a BLAST cut-off value of $1E^{-50}$. HSPs with e-values of 0 have been placed in the $1E^{-180}$ category to allow logarithmic scaling.

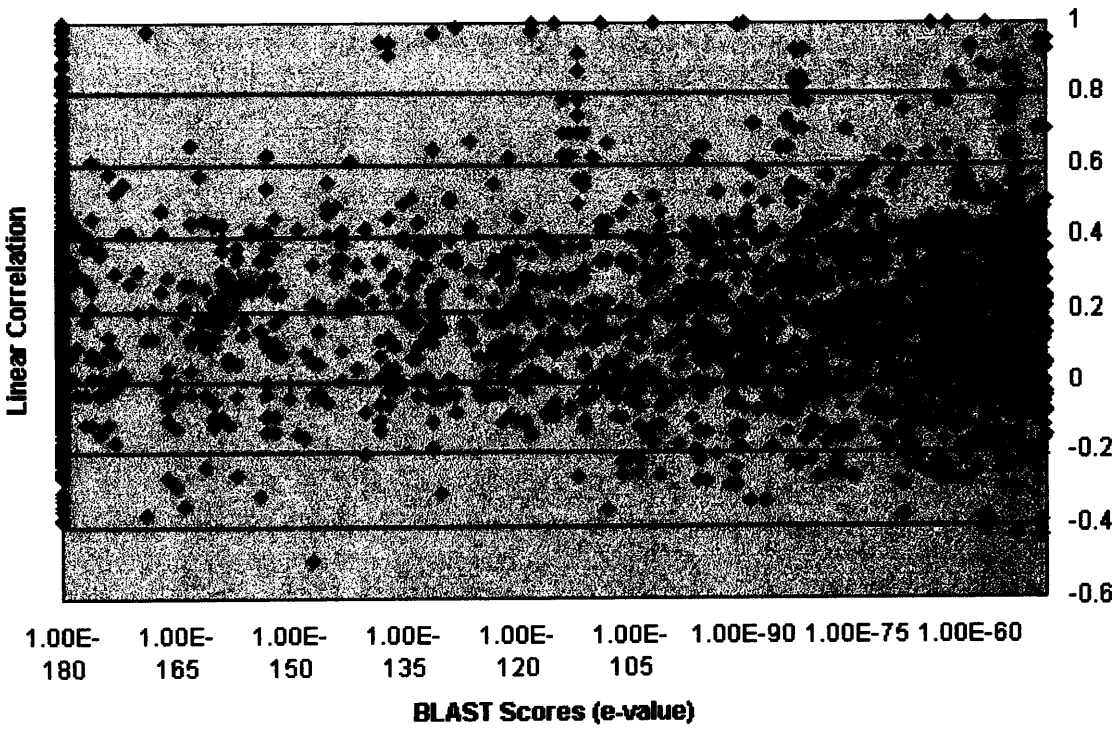


Figure 3.9. Results from URS and ORF comparisons using customised application of the correlation coefficient.

This data was produced using a combined dataset (SCPD consensus and TRANSFAC mapped sites) and a BLAST cut-off value of $1E^{-50}$. HSPs with e-values of 0 have been placed in the $1E^{-180}$ category to allow logarithmic scaling.

It is apparent that the correlation coefficient may not be the most suitable distance measure for this analysis. The, ***goodHitsPlusVectrData***, program was modified to allow the calculation of the Euclidean distance between two vectors (Equation 3.1, Figure 3.10).

No matter which metric is used, it is obvious - from all three graphs - that there is no correlation between similarity at the protein level and similarity at the TF binding profile level. One could argue that a comparison of the blast scores of the URSs and ORFs would be a more honest metric for this comparison. Figure A3.3 (Appendix 3) shows the results of such a comparison. The results are in agreement with the vector comparisons, with no correlation between URS similarity and ORF similarity in the yeast genome. The same result is obtained no matter what dataset or sub-set of binding sites are analysed.

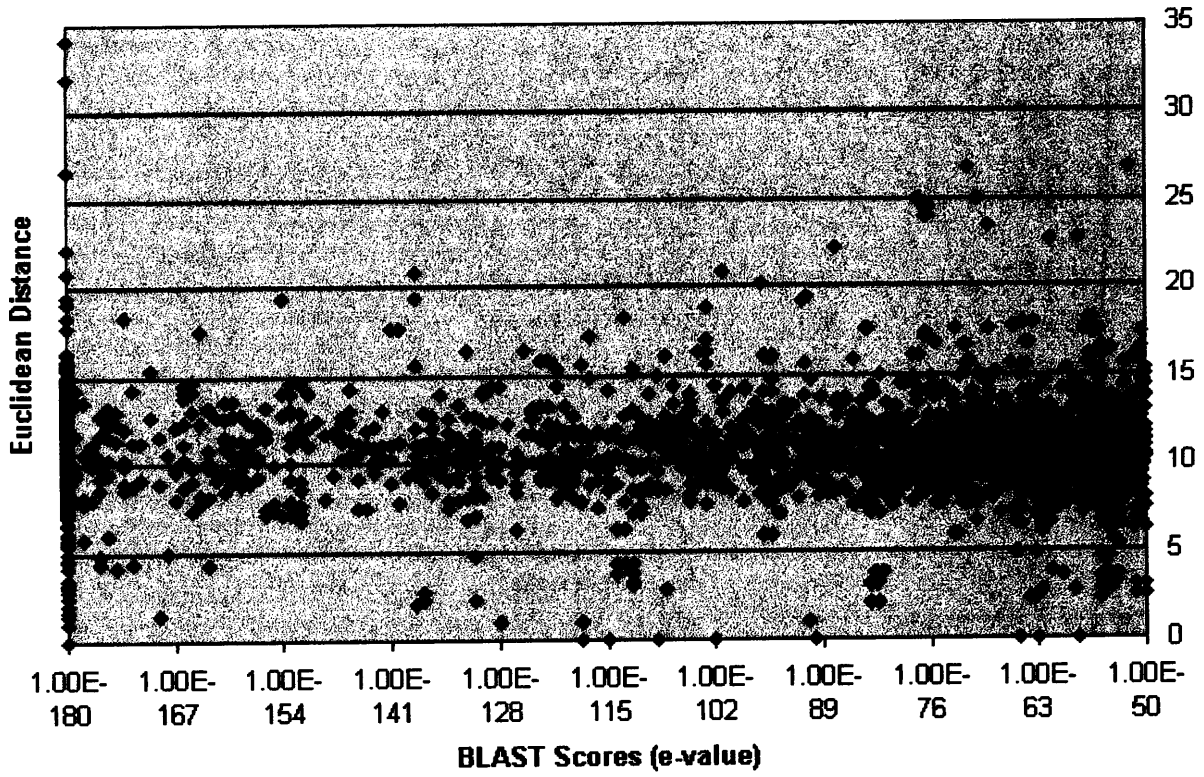


Figure 3.10. Results from URS and ORF comparisons using the Euclidean distance metric.

This data was produced using a combined dataset (SCPD consensus and TRANSFAC mapped sites) and a BLAST cut-off value of $1E^{-50}$. HSPs with e-values of 0 have been placed in the $1E^{-180}$ category to allow logarithmic scaling.

Figure 3.11 shows the results of a comparison of random and real vectors from the combined dataset (SCPD consensus sequences and TRANSFAC mapped sites). This shows that for the majority of HSPs their URSs are no more similar than what would be expected by random. There are some obvious exceptions in the 0 to 6 distance range for the lower e-values. The ORFs and URSs corresponding to these scores were extracted and analysed. The same genes detected through the previous clustering analyses are contained within this set, and the remainder have a very high degree of sequence similarity (i.e. are likely to be related through duplication).

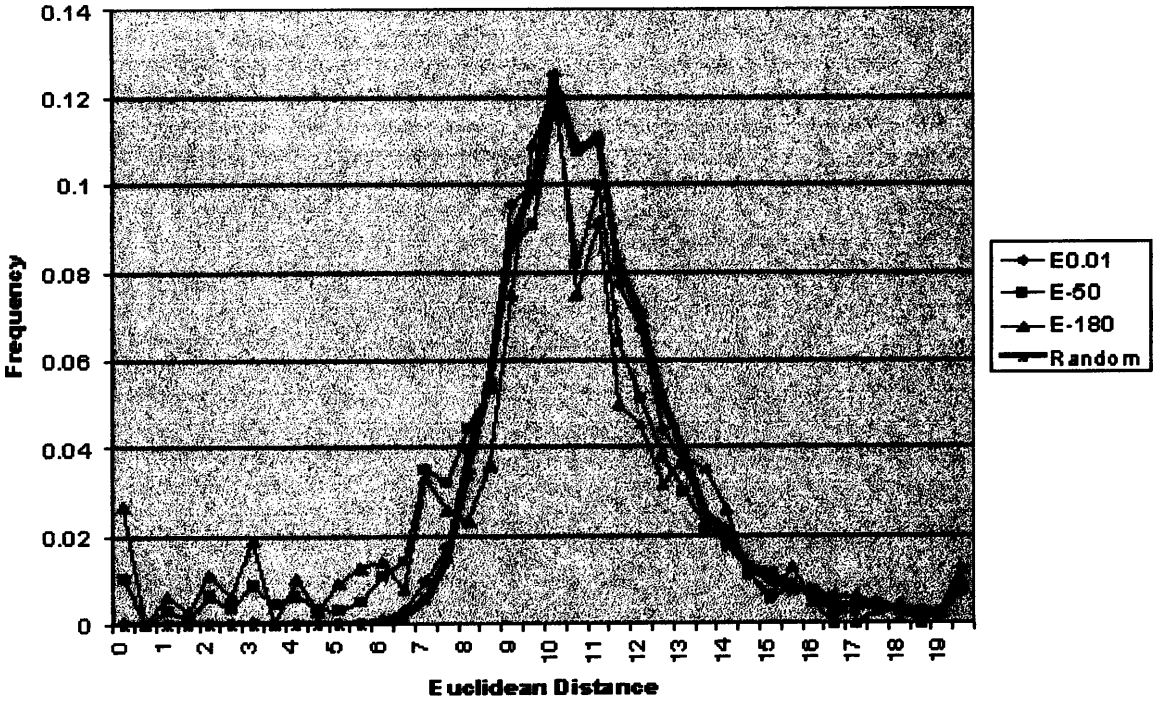


Figure 3.11. Histogram showing the frequency of Euclidean distances between URS vectors for different BLAST cut-off values for the equivalent ORFs.

The random data shown here is for 10,000 comparisons of randomly picked pairs of vectors from the combined SCPD consensus sequence and TRANSFAC mapped site vector database.

3.4.4 Conclusions & Discussion

The clustering of genes through the presence, absence or number of individual TF sites in their URSs has not provided a valid method for associating functionally related genes. In fact, it would appear that this analysis performs no better than random on all except the most similar of URSs. This is perhaps not surprising judging by the complexity represented by the transcription machinery and the plethora of methods with which the cell may control the transcription levels of a gene. Perhaps a more honest representation is that of multiple binding sites, which act synergistically to affect transcription. Many of these "promoter modules" have been identified and it would seem that both spacing and/or sequential order of the sites within these modules are important (Werner 1999). For example, the insertion of a few nucleotides between the TATA box and an upstream TF binding site (MyoD) in the desmin promoter dramatically reduces the levels of expression for this gene (Li 1994). Another example is the AP-1 binding site; Functional AP-1 sites have been found upstream and downstream of the transcriptional start site of the gene they influence. The position, or more accurately the context, in which these sites are found appears to affect the function of the transcription factor. An AP-1 site located close to the transcriptional start is important for the expression of Moloney murine leukemia virus (Sap 1989). In the rat bone sialoprotein gene, a repressing AP-1 site is found 900 bases upstream of the site of transcription start (Yamauchi 1996). This site overlaps with a set of GRE sites (glucocorticoid response elements) and is thought to cause repression through competitive binding. In the next chapter, we explore the use of distance relationships between TF binding sites for the prediction of gene function.

Positional analysis of transcription factor binding sites

Our previous studies described in Chapter 3 have strongly suggested that there is insufficient data available to functionally classify genes using solely the presence or absence of binding sites as indicators. In this chapter, we expand on the vector analysis by considering the positional information encoded by the locations of the binding site sequences. We also describe the development of a computational tool for the visualisation of binding sites in groups of nucleotide sequences.

In the last few years it has become increasingly apparent that differential gene expression is often achieved through the combinatorial regulation of transcription by specific combinations of transcription factors binding to different sites in the URSs of genes under their control (Klingenhoff 1999, Kel-Margoulis 2000). Indeed, a specialist database, COMPEL (Kel-Margoulis, 2000), has even been set up to provide data on composite regulatory elements. However, visual inspection (using BLAST and clustalW) of selected URSs in our yeast data did not point to a large number of pairwise TF binding sites that were consistently conserved in terms of relative or absolute positioning. Because it was clearly not possible to define a reasonably sized set of positionally conserved pairs of sites, it was decided to undertake a large-scale comparative analysis in an attempt to detect these relationships. Furthermore, rather than using a comprehensive treatment involving weight matrices for each site, we elected to use several more simplistic approaches which model either the presence or absence of ***pairs*** of sites. By analysing pairs of sites we hoped to discover synergistic relationships between transcription factors that affect the transcriptional activity of the downstream gene. We hoped to discover functional relationships between those genes that display an overrepresentation of these relationships.

We investigated two potential models for positional information exploitation: Absolute position and relative position. With these two approaches we hoped to address the possibility that functionally dependent sites may be under some form of distance constraint. These constraints may be expressed as a gap of a specific number of bases between two sites, which relates to the specific interactions between the two

transcription factors, or simply the same face of the DNA duplex. Alternatively, in all co-regulated URSs it may be that the distance between the sites is irrelevant and that regulation is largely dependent on the order of the binding sites. It is unclear which of these models may be correct and so a number of different analyses were undertaken to investigate all the aforementioned possibilities. We use the relative position model to investigate the possibility that the order of binding sites is more important than the distances. The absolute position model is used for examining distance constraints in preference to site order.

4.1 Absolute position

4.1.1 Methods

The comparison we term "absolute position" compares two upstream sequences by looking at the distances of binding sites within the URSs. A simple score is allocated to every pairwise comparison by counting the number of equidistant sites within the two sequences. For example, the pairwise comparison of the two URSs in Figure 4.1 yields a score of 3. This is because the following distances are equal in both sequences: green → red, red → yellow, green → yellow. Distances between multiple instances of the same site are also considered.

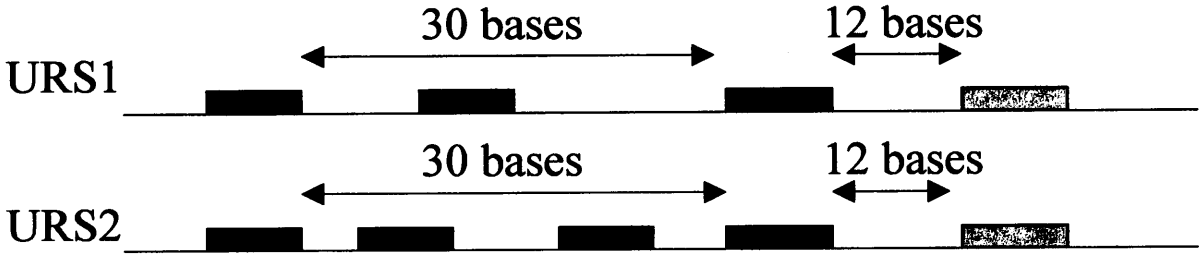


Figure 4.1. Example upstream regions for two URSs.

We created the program, ***TFPositionAnalyser***, to perform this analysis. This program takes a set of URS vectors (as in our previous analysis, Section 3.2) as input. The ***vectorsFROMfp*** program was modified to include positional information in the output vectors (Table 4.1). This investigation was carried out with multiple sets of binding site data. Initially we utilised the complete combined set of TRANSFAC and SCPD sites. A second, more refined, dataset was created by removing all duplicated sequences (contained in different database entries) and those sequences that were found to be sub-

sequences contained within other entries in the dataset. Sequences that had partners detectable through reverse complementation were also removed. A third dataset was produced by removing all TATA and TATA like sequences, as these sequences add a significant amount of noise to the analysis (this is seen later on).

Table 4.1. Example *vectorsFROMfp* output vectors for the URSs of the first three genes in the *cerevisiae* genome.

Multiple instances of a specific site are separated by the '&' character. Empty columns signify the absence of a particular site.

URS	Transcription Factor Binding Site				
	ABF1	*ACE2*	*ORC*	*PHO4*	...
YAL001C		35&128&634	12		...
YAL002W				56	...
YAL003W					...

All pairwise distances between each binding site in an URS are searched against those of all the other URSs in the dataset. If another URS is found with the same distance between the same sites then the score for these two URSs is incremented. This calculation was observed to be much quicker and more efficient than an "all against all" search, whilst producing the same results.

To account for the possibility of transcription factors acting as functional units of three or more proteins, we also implemented searches that consider URSs with identical distances between sets of three and four binding sites. Initially we only considered situations where sites are detected at exactly the same distance apart. It is likely that this is a somewhat unrealistic biological model and that these distances are more variable in practice. In an attempt to account for this, we subsequently allowed a small amount of variability in these distances. This was applied for both the pairwise, and three and four site calculations. We also included a "phase variance" term, with which we hoped to capture sites that were separated by different distances but demonstrate a separation which places the binding protein in the same orientation as the original. A single turn of B-DNA contains 10 base pairs. The "phase variance" term is a threshold limit for the difference in distance two sites may exhibit whilst still demonstrating the

correct phase. For example, if the GAL4 and ACE2 sites are found at a distance of 35 bases in sequence A and at a distance of 55 bases in sequence B, using a "phase variance" threshold of 2 (± 20 bases) will positively score this relationship because the sites are still in the same physical orientation on the DNA duplex in both URSs. If the sites were 50 bases apart in sequence B (or within any of the following distances: 0 → 14, 16 → 24, 26 → 34, 36 → 44 and 46 → 54 bases), this would not be considered a positive relationship.

An evaluation of the effectiveness of these approaches is undertaken by considering the frequency of association of functionally related genes in comparison to that expected by chance. In order to test this, gene sets were obtained where the genes were known, *a priori*, to be either co-expressed or functionally related in some way. We used microarray generated gene clusters defined by Eisen *et al.* (1998), northern blot data generated in the analysis of Brown *et al.* (2001), and genes occurring in the same section of the KEGG functional classification scheme (Kanehisa 2002, web ref 27) and the MIPS functional classifications (Mewes 2002, web ref 28) to define functionally related genes. Random values were calculated by analysing clusters of URSs picked randomly from the *Saccharomyces cerevisiae* genome. For each cluster under consideration, an identical number of URSs were selected at random and analysed in parallel to the biologically related datasets. Random selections were repeated ten times and the scores averaged to give the final randomised score for comparison.

The method described above was used to evaluate the clusters from Eisen *et al.* (1998) and selected clusters from Brown *et al.* (2001). A second methodology was used to analyse the Eisen clusters, the entire set of clusters from Brown *et al.*, the MIPS functional categories and the KEGG functional categories. In this method, the average number of links (number of distance relationships in common) between URSs in a specific category/cluster is compared with the average number of links formed between the URSs from this category to those in different categories. Given the hypothesis that the URSs of a group of functionally related genes contain a common, detectable, signal, then the average number of links formed between URSs in a cluster should be higher than the average number of links formed with non-cluster members.

To ease the interpretation of the results a simple graphical tool, ***draw_jock_plot***, was developed which takes a tab-delimited dataset of real numbers (with columns and row headings) and outputs a 'microarray style' display. It is possible to view a complete set of results for all functional groups in a single image using this style of visualisation.

4.1.2 Binding site database refinement

Initial results from the pairwise analysis using the complete binding site data were unexpected. A very large proportion (99%) of the 19,318,088 pairwise comparisons have a positive score. Indeed, some comparisons were seen to have very large scores representing a huge number of identical distances between binding sites in the URS (Table 4.2). The highest score was 3036 for the comparison of YMR050C and YMR051C. On inspection, the URSs for these genes (and those of many other high scoring comparisons) correspond to the same genomic region. These duplicated URSs were removed from further analysis. An examination of some of the remaining, high scoring, comparisons reveals a further possible cause for the unexpectedly large scores. Many of the binding site sequences in our TRANSFAC and SCPD datasets are identical or near identical. This resulted in positive scores for the majority of comparisons made using the paired site approach. Two identical (and near identical) sites will always have identical distances, because they are located at the same position (with a distance of zero). Since a zero distance was still assigned a positive score, these had a major effect on the final result. The transcription factor binding site database was therefore updated to remove redundant sites (as described in Section 4.1.1). We then performed the same analysis with the new binding site data.

Table 4.2. Distribution of scores for initial pairwise analysis.

Score range	Number of comparisons	Number of ORFs in comparisons
1000+	54	102
100 → 1000	23,628	1594
10 → 100	6,063,472	6216
1 → 10	13,010,349	6217
0	224,933	6217

Scores for this 'refined' dataset ranged from a minimum of 0 (no distances in common) to a maximum of 1327 for the comparison of YDL131W and YNL014W (not considering completely duplicated URSs). This score is still far in excess of any expected maximum. It represents the presence of 1327 identical distance relationships within the 800 bases of the two URSs. The percentage of positive scoring pairwise comparisons is still high, though markedly reduced from the previous analysis to just over 72% (Table 4.3).

Table 4.3. Distribution of scores after removal of duplicated binding sites.

Score range	Number of comparisons	Number of ORFs in comparisons
1000+	26	24
100 → 1000	10,262	541
10 → 100	604,730	5139
1 → 10	13,345,389	6216
0	5,362,029	6217

The reason for the continuing presence of high scores becomes apparent when the nucleotide sequences of the relevant URSs are examined. All 24 URSs that have scores of 1000 or more contain long AT-repeats (Figure 4.2). Within these repeat regions there are a very large number of distance relationships present for a matching binding site. These distances will all be identical to those found in other repeat regions resulting in the large scores seen in our results.

It is possible to account for these AT rich repeats in two ways. Firstly, we can remove TATA sites (and TATA related sites) from the binding site data. A second method for masking out these repeats is by using a low-complexity filter such as ***dust*** (available with the blast suite, Altschul *et al.* 1997). We initially opted to remove just the TATA sites from the database, because ***dust*** may also mask out other regions of the URS that potentially contain informative sites. Later we also performed limited experiments on URSs with low-complexity regions masked out using the ***dust*** algorithm.

```
>YDL131W 5' untranslated region, chrIV 226393 - 227192, 800bp
AATCTGCCACATACAACATTGGTCAACGAATGCATTGCGCAATCACACCAAAGGTTCAAT
GCAAAGGTTTCTATGGTCAAGAGAGCCATCGATAGCTTAATACAAAAGGGATACCTACAG
AGGGGAGACGATGGTGAATCGTATGCTTACCTTGCTTAATCATCTTTGAAGGCTTGTGCT
GATCGAACGAAGCAAATCCTACGAGTAAATACATAAGCGTATACATATATATATATATAT
ATATATATATATATATATGTATATATATATATGTGTGTGTGTGTAATTGTGTGTATTCAA
CTGAACTATGAAGAGTCTTTGACCTCTTGAGAATCTCATAGTATGAAGATATGGCACTTC
TCTTTCCGTTGTAACATCCTTTACCGGGCGGCTTTTCGGCCTGCTTGAGAAGAGATCAGG
CTGAAATGATGAGACTGACAACAAGGATCAGTCAGTGGCAGAGTTGAAATTCCGCTGGAA
ATCGTCACCAGCAGCTACCAGCTCCAAGAAATCCGCGTAGAAATCACGTCGCGCAATGTG
CGATAGTGGTGGAACAGCGAAAATTAGTCACTTTGCAAAGATGCCAAGCTGATCGTTCTC
GTTGTGTGTGATTCTACGATATTGCGCATGAGTGGTGTTGAGCATGCGTTTGTTCGCATT
TTCATGCAGTCTCTTATATACTAATATATATATATGCAACAGATTTCAAAATTTATCTTTCT
TCTTGTTGCGTTTAGGCCTTTATTTGCTACACATTTAAAAGTGCAACGACAACCCAAGTA
ATTGTATACTTTAACAAACC

>YNL014W 5' untranslated region, chrXIV 605315 - 606114, 800bp
GTACCAGATGTTCTCCACTTGAACAAGTTGAAGGAAAAGCACAATGATGTCATCGAGAAC
GTCGAAGAAGACAAAGAAGTTCATACAAATTGATTCCTGAGATAAACGCAAACACTTTTT
TACTCGGAAATTTTCTCTCCCCGATTGTAAATTTATTTCTCGTCTACTATAATAATTCAC
CAAAAAACAGCAGAGCACGTCCTATATATATATATATATATATATATATATATACATATA
TATATACGAATATATTTATGACAGTCTAAATCGTTGCTCCTGTTCAATTTTAACGCTCTT
TTATAACTGCTGTGGCGCCATTTCTTTGGCATCTAAGATGTGTCTCGGGACTTTCTTTTT
CTGGTCATCTTGCTTTCCTTGCCTTGTCATTAACAGAGCTAAACTGTGTTGAGGAAGTGG
TCTTTTAAATTTTCAAGGTCGCGGCCCGTGCATTGCGCGGCTGGATGTCTTTTAGGGGAG
CGATGCGCTTGCACAAAAGAATAGCCTTGATTTCCGCTATTAATGGGTATATTCTTCTAA
CAATTTGAGTTTTTTCTCCACTTGTTTCTTTCTTCTTTCTCTTTTCCATTCACATTTTGC
TTTACTTTTCTGGATAATATATATATGTGCATATATATATATCCATTTAAATATATATATAT
ATATATCTTTTGTCCTCCTCTCCAGTTACTTTTGCGCCAGAAATTCCTTCTTCTTTTCAA
TCTAATAGAGAAGGGTAGATAGTGTGGTGCAAAAAAAGAGCCTTTTAAACCAAAAATAAA
TAAAAAAAAGGAATCACAAA
```

Figure 4.2. Upstream sequences of YDL131W and YNL014W.

These are two of the 24 URSs sequences that have scores in excess of 1000. AT repeat sequences are highlighted in yellow. These sequences were extracted from the SGD database.

The score distribution for the TATA reduced dataset appears to be much more reasonable (Table 4.4). The vast majority of comparisons (80%) have a score of 0, with only a tiny fraction (0.0002%) having more than 10 distance relationships in common.

Table 4.4. Distribution of scores after removal of duplicate sites and TATA related sequences.

Score range	Number of comparisons	Number of ORFs in comparisons
100 → 1000	254	159
10 → 100	4,149	1368
1 → 10	3,894,009	6216
0	15,424,024	6217

There are many entries in SCPD and TRANSFAC that describe multiple binding sites for the same transcription factor. With the current methodology, these sites are considered as completely independent. This has consequences for our analysis as potentially, functional relationships between two URSs may be missed if in one URS a binding site is replaced with another for the same transcription factor. We produced a dataset where multiple binding sites for the same protein are merged and considered as one. We termed this the "simplified" set. All experiments were carried out with both full and simplified binding site datasets.

There are also many sites in the SCPD and TRANSFAC datasets that have overlapping sequences. The proteins that bind these sites may influence transcription by competitively binding to the DNA. Potentially, these overlapping sequences have no common functional relationship and should not be considered. For example, many of the documented TATA sites overlap but only a single TATA binding protein (TBP) is present in the TIIFD complex (required for transcription at the PolII promoter). Since the TBP is only capable of binding to a single site, relationships between these overlapping sites are unlikely to be biologically significant and are likely to produce spurious connections between two, otherwise unrelated, URSs. In order to take this into account two further analyses were carried out where a minimum distance of 3 and 6 bases is required between two binding sites before they are taken into consideration as a potentially conserved pair. These distance-constrained analyses were only carried out with the cluster/non-cluster methodology described at the end of section 4.1.1.

4.1.3 Results

We used four, separate, groups of genes in order to assess the ability of this methodology to associate functionally related ORFs. We extracted the microarray clusters from Eisen *et al.* (1998) and those generated in the northern analysis by Brown *et al.* (2001), genes occurring in the same section of the KEGG functional classification scheme and the MIPS functional classifications.

A combination of codes is used to identify the 'experimental conditions' under which the different analyses were performed. These are given in Table 4.5.

Table 4.5. Codes for analysis identification.

Section ID refers to the section of the analysis code where this item is located. For example, the code dntaf2s1v refers to an experiment using the full binding site dataset with TATA boxes removed. Only pairwise distances are considered, and these distances may vary from one URS to the next by up to one base pair.

Section ID	Code	Meaning
1	df	Full dataset
1	ds	Simplified dataset
1	dntaf	Full dataset with TATA sites removed
1	dntas	Simplified dataset with TATA boxes removed
2	2s	Distances considered are for two sites only
2	3s	Distances considered are for those between three sites
2	4s	Distances considered are for those between four sites
3	Blank	No phase variance or distance variance allowed
3	1v	Positive score given to sites that have the same distance with up to a 1 bp difference
3	2v	Positive score given to sites that have the same distance with up to a 2 bp difference
3	1p	Sites must have the same distance in both URSs being considered or have a difference of exactly ten bases
3	2p	Sites must have the same distance in both URSs being considered or have a difference of exactly ten or twenty bases

4.1.3.1 Microarray clusters

The clusters considered here were taken from Eisen *et al.* (1998). In this paper, ten clusters of genes with highly correlated expression profiles were identified. Each of these clusters were thought to be "high quality" as they each contained many genes with related functions (Table 4.6). These were labeled cluster B to cluster K in the paper and we maintain this nomenclature here. Unfortunately the list of genes composing cluster I is unobtainable from the paper or from the supplementary material. Because of this, cluster I cannot be considered in the evaluation.

Table 4.6. Overview of gene clusters taken from Eisen *et al.* (1998).

Cluster	Number of Genes	Primary functional classification
B	11	Spindle pole body formation
C	27	The proteasome
D	14	MRNA splicing
E	17	Glycolysis
F	20	The mitochondrial ribosome
G	14	ATP synthesis
H	8	Chromatin structure
J	5	DNA replication
K	15	The tricarboxylic acid cycle and respiration

The complete set of results for the random comparison is given in Figures A4.1 and A4.2 (Appendix 4). The example shown below is for cluster B carried out using both the simplified and the full binding site datasets (Figure 4.3). A breakdown of the abbreviations used to identify different analyses in the figure is given in Table 4.5.

In Figure 4.3, it is apparent that the scores for comparisons of genes in cluster B are lower than the scores for randomly selected URSs. This is the case for all analyses on the URSs of genes from cluster B considering all experimental variations. Analyses considering distances between 3 or 4 binding sites failed to find any relationships using the full binding site dataset and very few relationships with the simplified set. Indeed, these results are representative of all those seen in the analyses of clusters C, D, G, J and K (Figures A4.1 and A4.2, Appendix 4).

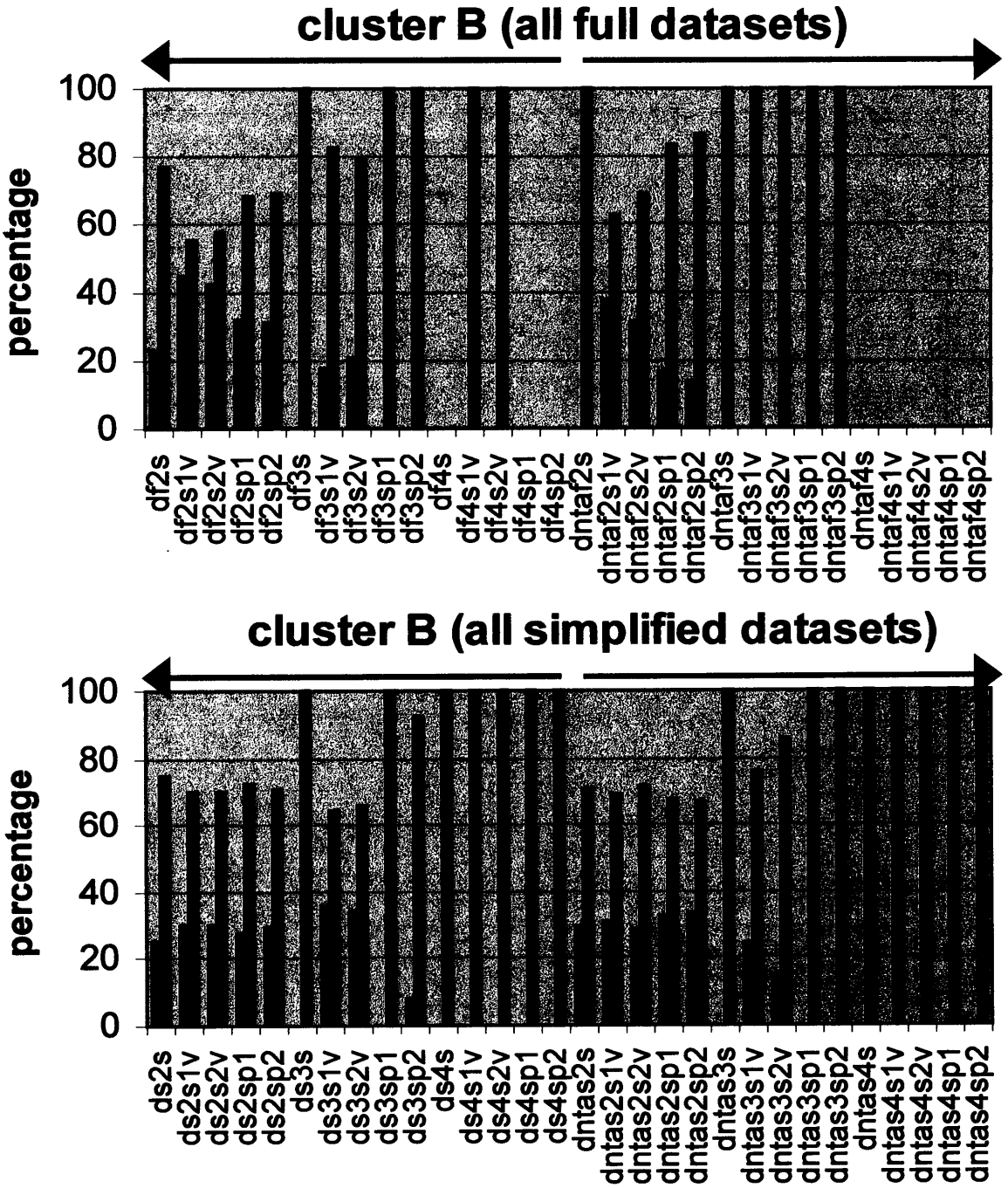


Figure 4.3. Results of both full and simplified binding site analyses for cluster B.

The final score for an analysis (x-axis) is displayed as a percentage (y-axis) of the total for both actual and random scores. Blue bars represent the scores for the analysis on the real dataset. Red bars represent average scores for ten random analyses. See Table 4.5 for an explanation of the dataset codes seen on the x-axis. The green arrow highlights those analyses carried out with TATA boxes retained in the binding site data whereas the orange arrow highlights those analyses carried out with TATA boxes removed from the binding site data.

Clusters E, F and H all show consistently higher scores than expected by random for all experiments that include TATA sites. On removal of the TATA sites from the binding site data, the scores for clusters E and F fall dramatically in comparison to random. Cluster H retains higher than random scores with all datasets apart from the 'full' binding site data where relationships between three or more sites are considered. The results for cluster H are easily explained when its component URSs are examined. This cluster consist of four pairs of genes which are located close to each other on opposite strands of the DNA. This means that each gene pair shares a portion of the 800 bp URS region between them. For all the genes in this cluster, the overlap is very large, ranging from 620 bp to 780 bp.

No single analysis gives a better than random performance for these clusters of genes. The presence of large numbers of TATA repeats in a few of the clusters biases the analyses. On removal of these sites from consideration, the analyses still fail to find more relationships than expected by chance. In cluster E, for example, there are 14 conserved distance relationships found with the df2s analysis (full binding site database including TATA boxes, comparing distances between two sites at a time). An analysis with randomly picked URSs finds an average of 13.1 conserved distances (over 10 searches). When the TATA sites were removed from consideration (dntaf2s analysis), there were no conserved distance relationships found between the URSs of this cluster. However, the random analysis found an average of 4.9 relationships (over 10 searches).

The comparison of the number of links formed between URSs in the same cluster to URSs in other clusters gives similar results (Figure 4.4). No single analysis performs uniformly better than random. The ds2s and ds2s1v results contain the highest number of positive scoring clusters (numerical results for these analyses are given in Table 4.7). It has already been shown that a high abundance of AT rich sequences in the URSs of these high scoring clusters is the cause of these results.

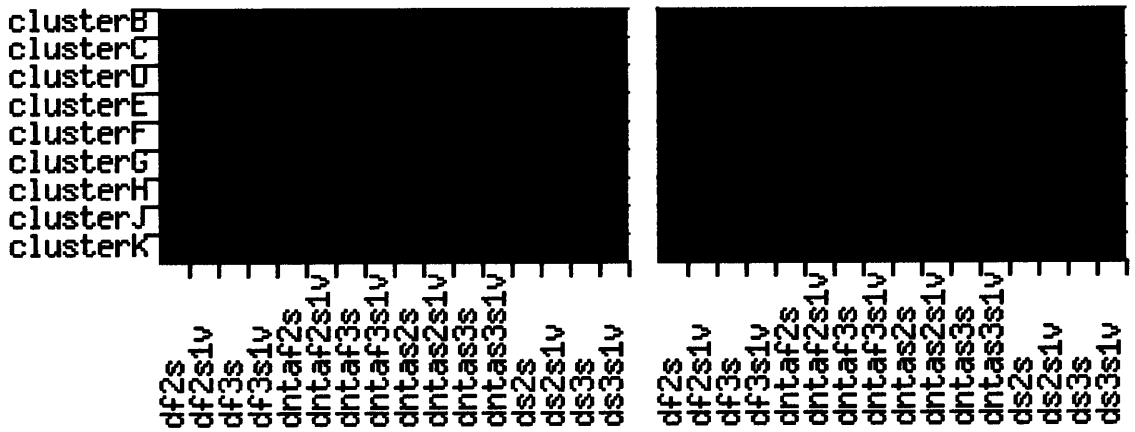


Figure 4.4. Graphical overview of the absolute position analysis of clusters taken from Eisen *et al.* (1998).

Dataset codes are given along the x-axis (see Table 4.5 for description of these codes). Each square represents the results for a single analysis of a specific cluster (y-axis). The performance of a dataset is represented as increasing intensities of red (positive score) or green (negative score). In the left hand image, the results for each cluster are scaled independently, with the maximum colour intensity set by the largest score (either positive or negative) for that cluster. This highlights the analyses that perform the best and worst for each cluster. In the right hand image the scaling is global, with the maximum colour intensity set by the largest score out of all comparisons for all clusters. This highlights those clusters with the highest number of intra-cluster links in comparison to inter-cluster links.

Table 4.7. Results for the ds2s and ds2s1v analyses on clusters taken from Eisen *et al.* (1998).

Results are given as the percentage difference between the average number of links found between URSs in the same cluster in comparison to the average number of links found between URSs in a cluster and those not in the cluster.

Cluster	Score for analysis (%)	
	ds2s	ds2s1v
B	-2.5	-3.1
C	0.6	1.5
D	-2.6	-2.7
E	12.2	16.5
F	1.3	2.0
G	-1.1	-1.3
H	7.5	9.6
J	-3.3	-4.8
K	2.0	4.0

No major effect on the discriminatory power of this analysis was observed when distance constraints were incorporated (Figure A4.3 b) and c), Appendix 4). However, one notable change is that the scores for clusters with AT rich sequences are reduced dramatically when a minimum distance of six bases is used. A further, somewhat

disappointing, result was achieved for analyses carried out on URSs with low-complexity sequences masked (Figure A4.3 d), Appendix 4). Overall, the best performing analysis is ds2s1v with an enrichment of 2.1% over random for cluster H.

4.1.3.2 Northern clusters

Brown *et al.* (2001) used a classical northern approach to analyse the expression of 1008 ORFs from *Saccharomyces cerevisiae*. Eight different physiological conditions were examined: glucose depression, glucose upshift, stationary phase, control at 30°C, ammonium starvation, hyperosmotic shock, control at 23°C and heat shock. In this paper, hierarchical clustering was applied to the expression profiles for 635 ORFs whose levels were detected in all of the physiological conditions examined. Gene clusters containing the housekeeping genes CAR1 (cluster 8), HSP12 (cluster 1) and RPL25 (cluster 29) were analysed alongside two other 'high-quality' clusters (clusters 11 and 12). All five of these clusters were selected by one of the authors (Fajar Restuhadi, personal communication). Further to this primary analysis we then examined all 28 clusters of genes provided by Fajar Restuhadi.

The number of genes in each cluster is given in Table 4.8. Figures A4.4 and A4.5 (Appendix 4) give a graphical representation of the results from the initial analysis of clusters 1, 8, 11, 12 and 29.

Table 4.8. Overview of gene clusters taken from Brown *et al.* (2001).

Cluster	Number of genes
1 (contains HSP12)	21
2	19
3	13
4	17
5	13
6	10
7	12
8 (contains CAR1)	10
9	14
10	21
11	24
12	26
13	62
14	5
15	17
16	10
17	31
19	36
20	10
21	9
22	17
23	9
24	17
25	8
26	10
27	8
28 (contains RPL25)	8
29 (contains RPL25)	100

The initial results are similar to those seen for the microarray clusters. The majority of analyses for each cluster perform worse than (or similar to) the analyses run on randomly picked sets of genes. The only exception is the CAR1-containing cluster 8. Two URSs from this cluster contain long AT rich repeats. When the TATA sites are removed from the analysis the number of relationships found in common for the cluster is reduced to a value below that found in the random analysis. Overall, the analysis on all 28 clusters gives mixed results (Figure 4.5). No single method gives consistently positive discrimination for genes within all clusters. Overall, the analyses that most often perform worse than random are those that include TATA sites (e.g. analyses ds2s and ds2s1v).

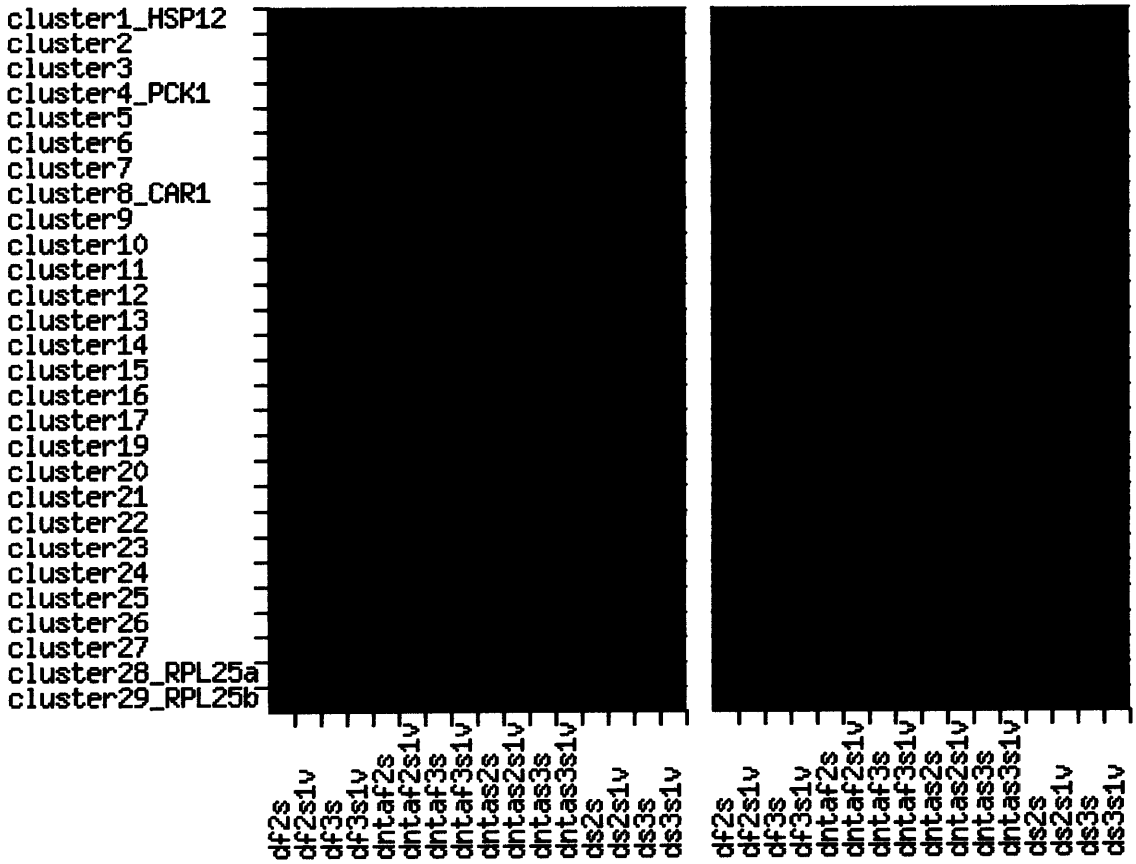


Figure 4.5. Graphical overview of the absolute position results for clusters from the Northern data (Brown *et al.* 2001).

Dataset codes are given along the x-axis (see Table 4.5 for description of these codes). Each square represents the results for a single analysis of a specific cluster (y-axis). The performance of a dataset is represented as increasing intensities of red (positive score) or green (negative score). In the left hand image the results for each cluster are scaled independently. This highlights the analyses that perform the best and worst for each cluster. In the right hand image the scaling is global, which highlights those clusters with the highest number of intra-cluster links in comparison to inter-cluster links.

The inclusion of distance constraints apparently has no major effect on the results (Figure A4.6 b) and c), Appendix 4). Cluster 8 has consistently high scores in all analyses except ones using low-complexity masked URSs. This is because two of the URSs in this cluster contain long AT rich sequences. The fact that the scores for the analyses with this cluster are reduced when using masked URSs demonstrates the utility of low complexity masking in removing noise from sequence comparisons.

4.1.3.3 MIPS classifications

At the time of this analysis, the MIPS classification hierarchy consisted of sixteen major categories (Table 4.9). Each of these categories is further subdivided into more specific classifications (except categories 15 and 16). We investigated the possibility of signal detection with all major categories except those pertaining to unclassified genes (categories 15 and 16).

Table 4.9. MIPS functional classification.

Category	Functional classification	Number of genes
1	Metabolism	1025
2	Energy	232
3	Cell growth, cell division and DNA synthesis	766
4	Transcription	708
5	Protein synthesis	348
6	Protein destination	505
7	Transport facilitation	303
8	Intracellar transport	429
9	Cellular biogenesis	167
10	Signal transduction	122
11	Cell rescue	338
12	Ionic homeostasis	120
13	Cellular organization	2124
14	Retrotransposons and plasmid proteins	8
15	Classification not yet clear-cut	149
16	Unclassified proteins	2657

We also analysed the subclasses of categories 9, 10 and 11 (Table 4.10) to investigate the possibility that the more refined groupings of genes may contain signals that are drowned out in the more general functional classifications. Due to time and computational constraints, only a limited number of analyses were carried out on the MIPS groupings. We examined the full datasets, simplified datasets and the same datasets with TATA sites removed.

Table 4.10. Sub-classifications for MIPS categories 9, 10 and 11.

Category	Functional classification	Number of genes
9.1	Biogenesis of cell wall	115
9.2	Biogenesis of plasma membrane	1
9.3	Biogenesis of cytoskeleton	17
9.4	Biogenesis of endoplasmatic reticulum	3
9.5	Biogenesis of Golgi	2
9.6	Biogenesis of intracellular transport vesicles	1
9.7	Nuclear biogenesis	5
9.8	Biogenesis of chromosome structure	18
9.9	Mitochondrial biogenesis	9
9.10	Peroxisomal biogenesis	2
9.11	Vacuolar and lysosomal biogenesis	13
10.1	Unspecified signal transduction	1
10.2	Morphogenesis	27
10.3	Osmosensing pathway	17
10.4	Nutritional response pathway	22
10.5	Pheromone response generation	34
10.6	Other signal transduction proteins	37
11.1	Stress response	154
11.2	DNA repair (direct, base excision and nucleotide excision repairs)	79
11.3	Detoxification involving cytochrome P450	97
11.4	Cell death	11
11.5	Ageing	1
11.6	Degradation of exogenous polynucleotides	1
11.7	Other cell rescue activities	9

The results for the broad functional categories show a slight enrichment in comparison to non-group members (Figure 4.6 and Figure A4.7 a), Appendix 4), although this is usually for those lexicons that contain TATA sites. Categories 9 and 14 show a significant excess of intra-cluster links in comparison to inter-cluster links. Category 14 is defined as "Retrotransposons and plasmid proteins"; only eight genes in this category are present in the nucleus, which greatly reduces its worth for the elucidation of transcriptional control mechanisms present in the nucleus.

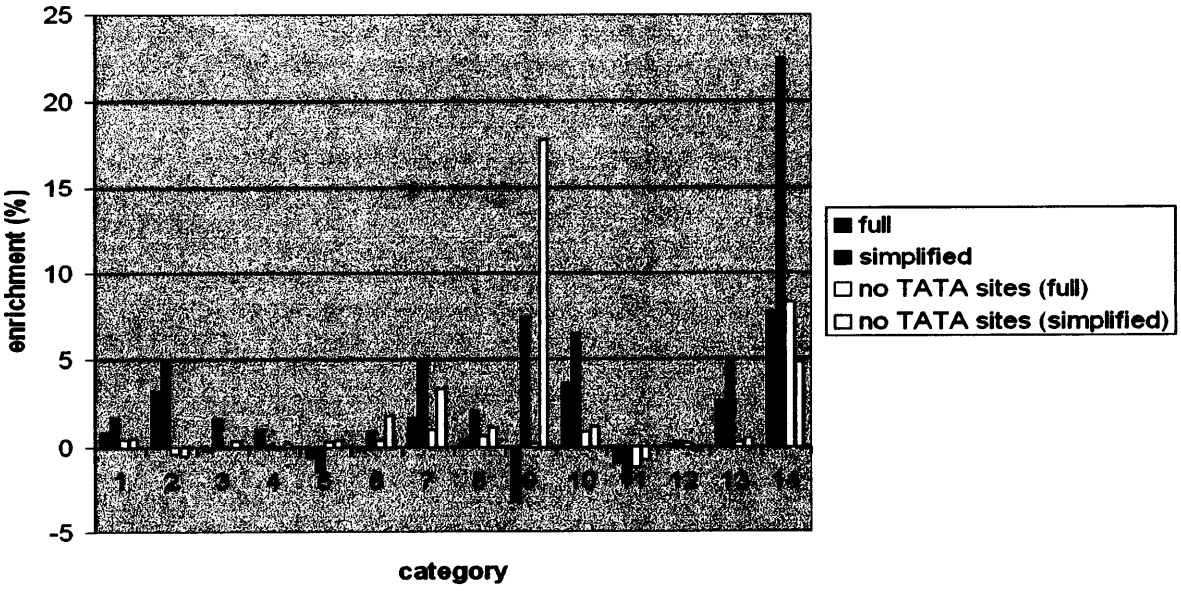


Figure 4.6. Absolute position results for broad MIPS functional categories.

Enrichment is calculated as the percentage difference of the average number of links formed between URSs clustered by functional classification with the average number of links formed to URSs outside of the cluster.

At first glance the results for the more specific MIPS categories seem contradictory to the previous results (Figure 4.7 and Figure A4.7 b), appendix 4). The majority of the subcategories of category 9 have negative scores, which seems contrary to the positive figures seen for the broader categories. The reason behind this is that the URSs with binding site relationships in the broad category have been partitioned into separate categories in the more specific analysis.

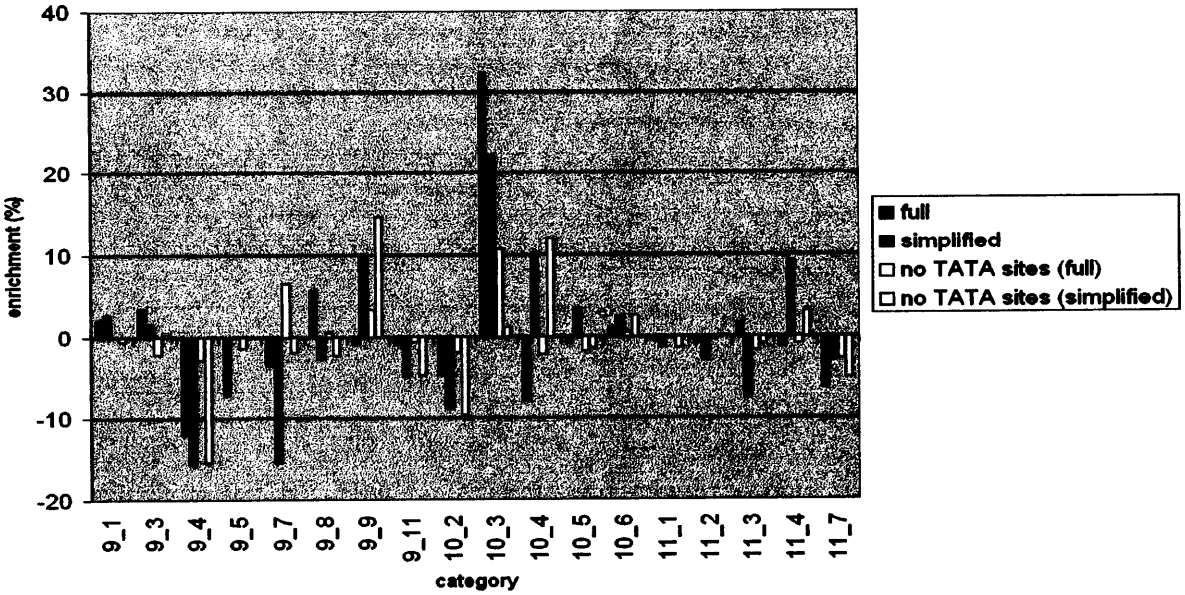


Figure 4.7 Absolute position results for a selection of specific MIPS functional categories.

See Figure 4.6 for details.

Analyses of those categories displaying positive enrichment revealed an abundance of short and simple sites such as GAL1-10, BAS2 and GATA (Table 4.11). In fact, all scores for MIPS category 9-9 were due to relationships arising through combinations of these sites. The small and simple nature of these sites means that there is a high probability that their presence is a purely random occurrence and there is no biological significance to the discovery of conserved distance relationships between them.

Table 4.11. Sequences for abundant sites in categories displaying positive enrichment.

Sequences for the GAL1-10 sites are from TRANSFAC. GATA and BAS2 sites are from the SCPD database.

Site name	Sequence
GAL1-10	ACTTATAT
GAL1-10	CGGATTAGAAGCCGCCG
GAL1-10	CGGGTGACAGCCCTCCGA
GAL1-10	AGGAAGACTGCTCCGAACAAT
GAL1-10	GAGGA
GAL1-10	GATAA
GAL1-10	AGCCT
GAL1-10	GGGG
GAL1-10	ATATAA

Site name	Sequence
GATA	GATAAG
GATA	CTTATC
GATA	GATAAC
GATA	GATAAT
GATA	GATAGA
GATA	GGTAAG
BAS2	TAATGA
BAS2	TAATAA

4.1.3.4 KEGG classifications

The KEGG functional classification system is geared towards classifying genes by metabolic pathways and interaction. This is reflected by the major divisions of the database: SSDB, BRITE, GENES, PATHWAY, LIGAND and EXPRESSION. Every gene in the database has a functional classification, which is usually provided through homology searches with the known sequence databases. The classification system for *Saccharomyces cerevisiae* is split into seventeen major categories (Table 4.12), although four of these categories are lacking any entries for *cerevisiae* genes.

Table 4.12. KEGG functional classifications (as of 1999).

Category	Description	Number of genes
1	Carbohydrate Metabolism	150
2	Energy Metabolism	122
3	Lipid Metabolism	45
4	Nucleotide Metabolism	116
5	Amino Acid Metabolism	217
6	Metabolism of Other Amino Acids	67
7	Metabolism of Complex Carbohydrates	208
8	Metabolism of Complex Lipids	159
9	Metabolism of Cofactors, Vitamins, and Other Substances	208
10	Metabolism of Macromolecules	37
11	Membrane Transport	0
12	Signal Transduction	61
13	Ligand-Receptor Interaction	0
14	Cell Cycle	0
15	Cell Death	0
16	Molecular Assembly	139
17	Unassigned	5277

As with the MIPS analysis we only undertook a limited number of analyses and we also considered selected sub-categories from the KEGG classification system (Table 4.13).

Table 4.13. Sub-categories for categories 1, 2, 9 and 12 of the KEGG functional classifications.

Category	Description	Number of genes
1.1	Glycolosys / Gluconeogenesis	38
1.2	Citrate cycle (TCA cycle)	22
1.3	Pentose Phosphate cycle	21
1.4	Pentose and glucuronate interconversions	8
1.5	Fructose and mannose metabolism	31
1.6	Galactose metabolism	24
1.7	Ascorbate and aldarate metabolism	8
1.8	Pyruvate metabolism	34
1.9	Glyoxylate and dicarboxylate metabolism	16
1.10	Propanoate metabolism	10
1.11	Butanoate metabolism	24
2.1	Oxidative phosphorylation	62
2.2	Methane metabolism	8
2.3	Carbon fixation	18
2.4	Reductive carboxylate cycle	11
2.5	Nitrogen metabolism	17
2.6	Sulfur metabolism	10
9.1	Thiamine metabolism	4
9.2	Riboflavin metabolism	12
9.3	Vitamin B6 metabolism	10
9.4	Nicotinate and nicotinamide metabolism	101
9.5	Pantothenate and CoA biosynthesis	9
9.6	Biotin metabolism	9
9.7	Folate biosynthesis	17
9.8	One carbon pool by folate	13
9.10	Porphyrin and chlorophyll metabolism	122
9.11	Terpenoid biosynthesis	4
9.12	Xenobiotics metabolism	18
9.13	Ubiquinone biosynthesis	25
9.14	Flavonoids, stilbene and lignin biosynthesis	2
12.1	Two-component system	3
12.3	MAPK signaling pathway	44
12.4	Second messenger signaling pathway	19

Figures 4.8 and 4.9 show the results from these analyses (A graphical overview created using the ***draw_jock_plot*** program is shown in Figure A4.8, Appendix 4). As with the MIPS analyses, the score for the majority of categories is generally no greater than expected by chance, unless TATA sites are included in the analyses. As with previous analyses the presence of AT rich sequences in the URSs of two or more genes in a single category can dramatically boost its score. For example, there are six URSs (out of 66) in category 12 that possess long tracts of AT repeats. This explains the high

scores seen for the full and simplified datasets. Those binding sites that have similar distances in the different URSs in the high scoring functional categories are again made up from simple sites as seen in the MIPS analysis. The most common relationships are found between the GAL1-10, BAS2 and GATA sites.

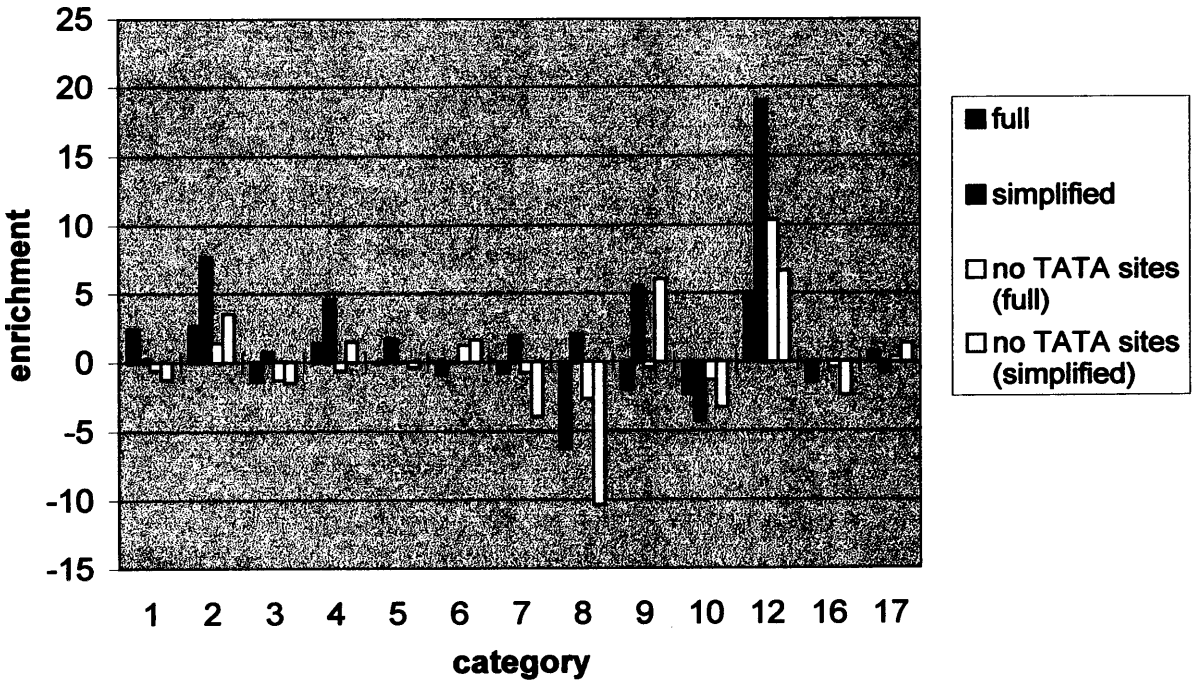


Figure 4.8. Absolute position results for broad KEGG categories.
See Figure 4.6 for details.

Results for the more refined KEGG categories are mixed. The overall score is positive but demonstrates only a slight enrichment over that expected by chance. It is obvious that the presence of TATA sites in the URSs allows a high score to be generated for a category and that the removal of these sites from consideration invariably has a detrimental effect on the final score for that group. The only category that this is not the case for is category 12.1. This is a very small group which consists of only three ORFs: YDL235C, YLR006C and YIL147C. There are only two conserved distance relationships within the URSs of YIL147C and YLR006C (BAS2 → GATA 217 bp and BAS2 → GAL1-10 218bp). The fact that there are only a small number of genes in this category affects the enrichment score.

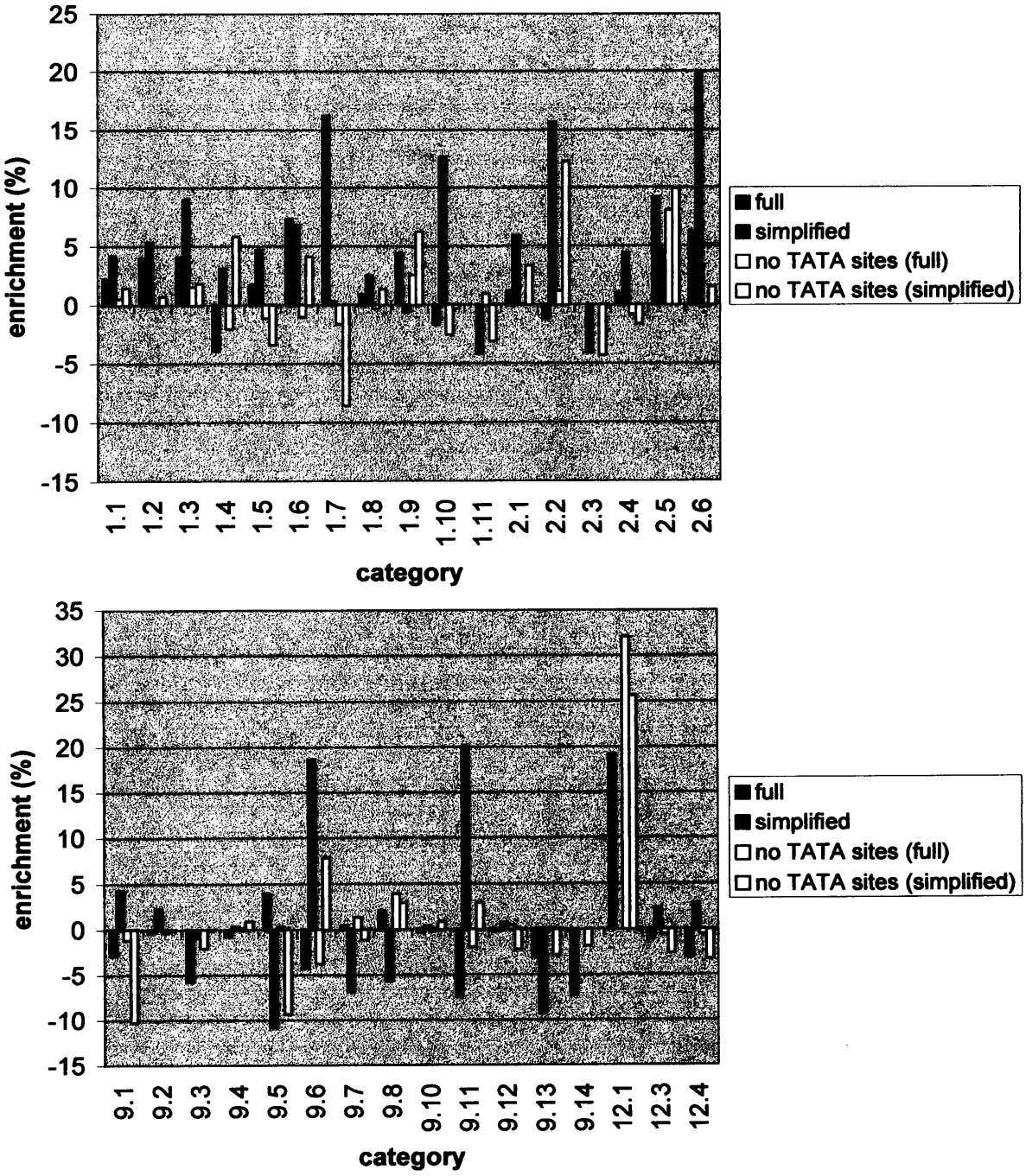


Figure 4.9. Absolute position results for a selection of specific KEGG categories.

See Figure 4.6 for details.

4.2 Relative position

4.2.1. Methods

The "relative position" analysis differs from the absolute position analysis by positively scoring URSs that contain the same patterns of binding sites regardless of absolute separation. There are many possibly ways of searching for conserved patterns; we opted for a simple model were URSs are scored by the number of identical neighbouring binding sites. For example, Figure 4.10 shows the positions of binding sites for two URSs. In this case, the comparison would score 1 as only the red and yellow sites are located next to each other in the two sequences.

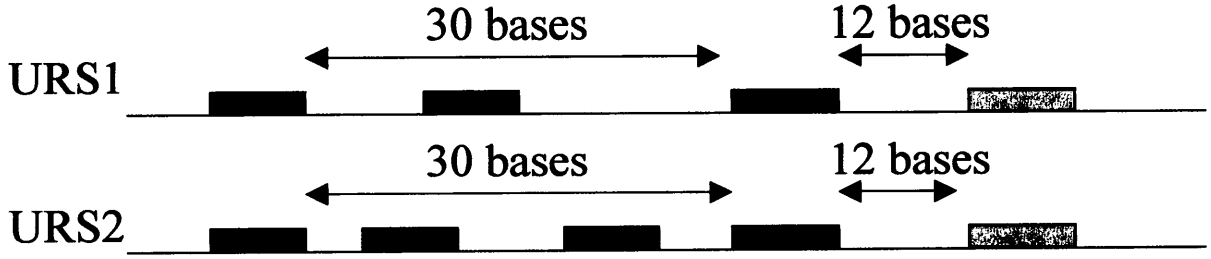


Figure 4.10. Example upstream regions for two URSs.

The program, ***TFPatternAnalyser***, was created to perform this analysis. As with the ***TFPositionAnalyser*** program, this takes a set of URS vectors as input. The ***vectorsFROMfp*** program was further modified to output binding site data in a format more suited to this analysis (Figure 4.11). The same binding site dataset was used as in section 4.1. We only carried out a primary investigation using the MIPS and KEGG functional categories.

YAL001C [10]TATA,TBP(2)&TATA,TBP(7) [203]TATA,TBP(18) [206]BAS2(4) ...
YAL002W [111]SCB [120]BAS2(4)-Rev [123]AP-1-Rev [143]TATA,TBP(5) ...
YAL003W [122]GATA(1) [123]GAL1-10-I [217]TATA,TBP(2) [219]TATA,TBP(5) ...
YAL004W [108]GATA(2)-Rev [161]GCN4(12)-Rev [180]GATA(5)-Rev

Figure 4.11. Example relative position vectors generated by the *vectorsFROMfp* program.

Each row contains information on the binding sites present in the URS of a single ORF. The binding sites are displayed in order of their appearance in the URS. The values in square brackets represent the start position of the site in the URS.

4.2.2. MIPS results

Figures 4.12 and 4.13 show the results of the relative position analysis using broad MIPS functional categories and the selected refined categories. Figure A4.9 represents these results as generated by the ***draw_jock_plot*** program as an alternative display.

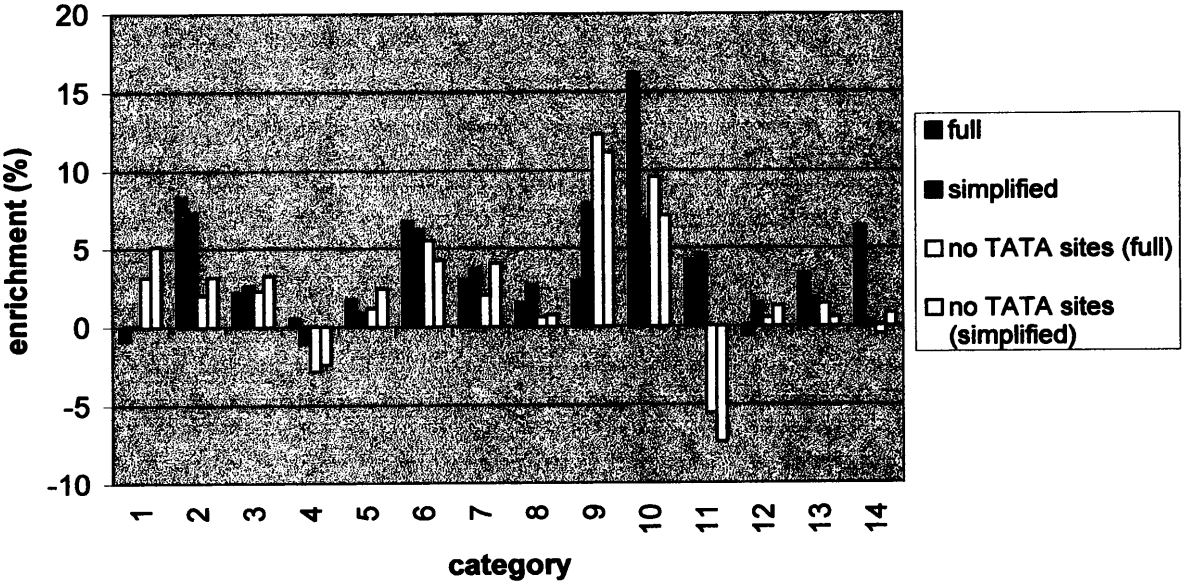


Figure 4.12. Relative position results for MIPS functional categories.

Enrichment is calculated as the percentage difference of the average number of links formed between URSs clustered by functional classification with the average number of links formed to URSs outside of the cluster.

At first glance, the results for the relative position analysis appear promising. There is a small positive enrichment of conserved relationships above random in all categories bar two. On inspection of high scoring analyses however, we see a large number of conserved relationships between the low-complexity, high abundance sites seen in the absolute position analysis. The ten most abundant conserved binding site relationships for the four analyses in MIPS category 10 is shown in Tables A4.11 to A4.14. These show that the vast majority of relationships are found between short/simple binding sites that are ubiquitous throughout the genome and thus appear to contain very little specific functional information. Even so, it appears that this methodology confers a limited form of predictive ability for all but two of the MIPS functional categories, even when TATA sites are removed from consideration.

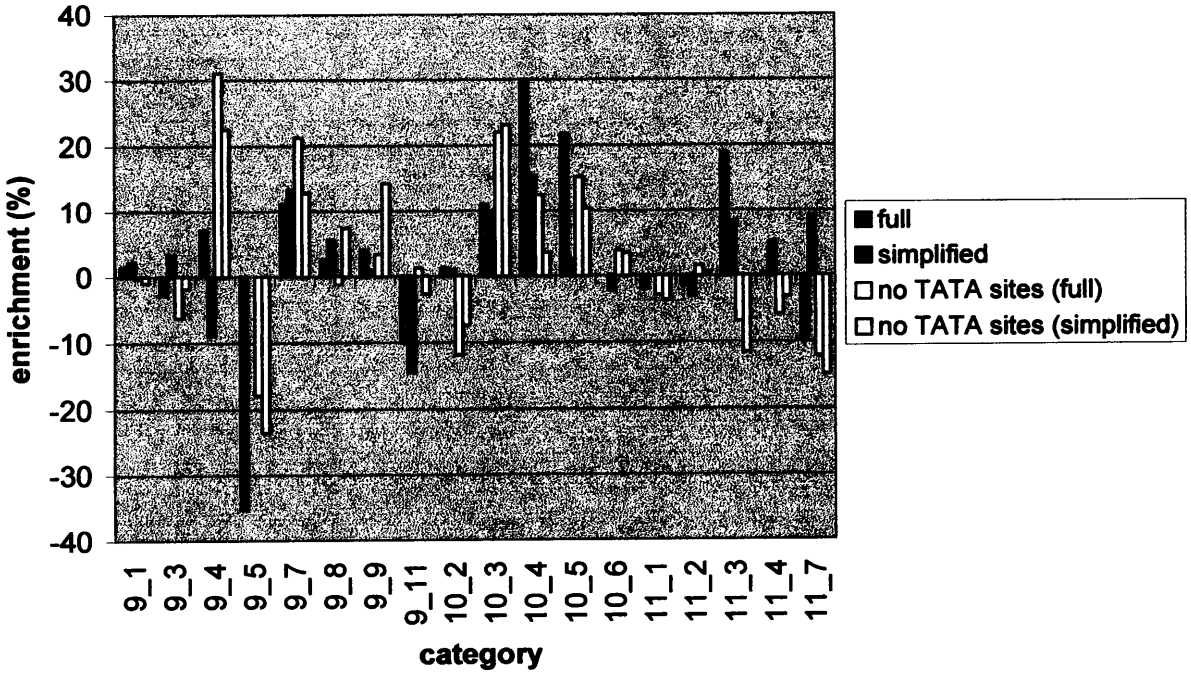


Figure 4.13. Relative position results for a selection of specific MIPS categories.

See Figure 4.12 for details. The following broad MIPS categories are represented here: category 9 – "Metabolism of Cofactors, Vitamins and Other Substances", category 10 – "Metabolism of Macromolecules" and category 11 – "Membrane Transport".

As with the more general categories the vast majority of patterns discovered in these analyses were between small/simple sites such as the GATA, GAL1-10 and BAS2 sites (as well as TATA boxes in the TATA inclusive analyses). In fact, nearly all relationships found involve one of these sites. In one extreme example, every conserved relationship discovered in every analysis for the "osmosensing pathway" category (category 10-03 above) involves one of these sites.

4.2.3. KEGG results

Figures 4.14 and 4.15 show the results of the relative position analysis using broad KEGG functional categories and the selected example refined categories. The refined categories shown here represent typical examples from the dataset. Figure A4.10 represents these results as generated by the ***draw_jock_plot*** program as an alternative display.

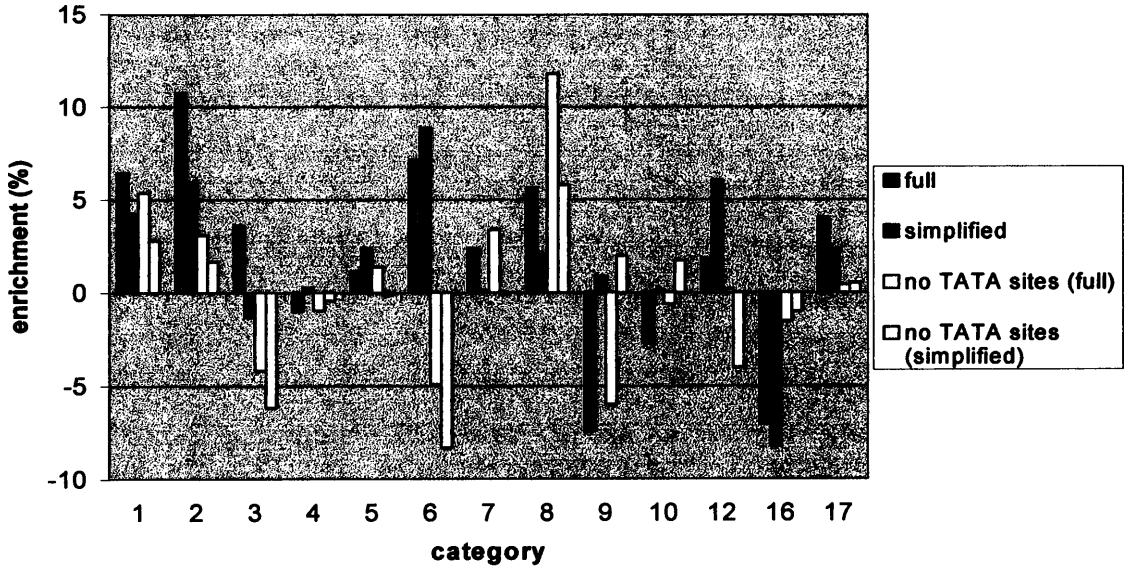


Figure 4.14. Relative position results for KEGG functional categories.

See Figure 4.12 for details.

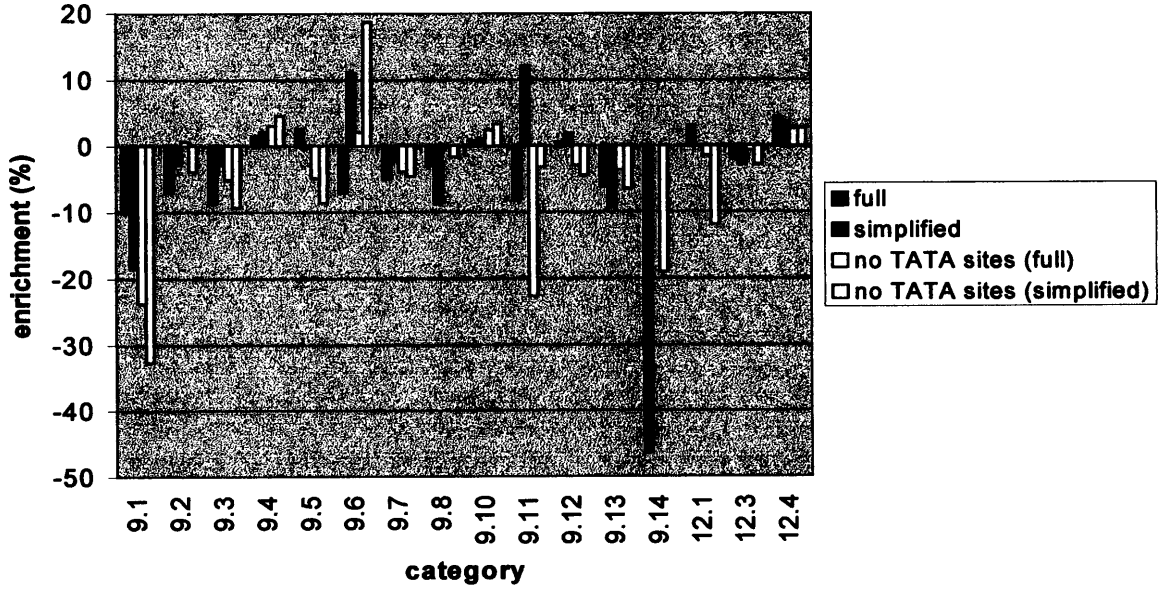


Figure 4.15. Relative position results for a selection of specific KEGG categories.

See Figure 4.12 for details.

Using the KEGG classification system to group genes has no appreciable difference over using the MIPS classification system. An analysis of broad and specific categories show that all conserved relationships discovered are again between the simple, ubiquitous, binding sites in the vast majority of cases.

4.2.4 Discussion

The results from the absolute and relative position analyses have demonstrated that the classification of genes using transcription factor binding sites is a non-trivial task. We have seen that in nearly all cases, conserved relationships between URSs are due to the presence of non-informative, low-complexity binding sites. On removal of these sites, virtually no relationships are found between URSs in functionally grouped genes. Neither of the approaches show significant ability for finding over represented relationships in functionally linked genes after removal of the TATA and other, low-complexity, sites. It should be pointed out that both the MIPS and KEGG classification systems group together genes with broad and overlapping functions and as such the URSs of genes in these categories may not be expected to contain conserved binding site relationships. This is not the case for microarray and northern data. Clustering genes through correlated expression profiles implies that these genes are under the influence of a common control mechanism and as such, we may expect to find similarities in the URSs of these genes. Our inability to find these may be due to a lack of informative binding sites in the lexicons used and/or to an inadequacy in the similarity method we have investigated. In this study, we have only investigated the use of mapped binding sites, although it is well known that many transcription factors do not bind to an exact sequence but to a variable sequence. These variable sequences can be represented using a matrix or as a static consensus sequence with IUPAC ambiguity codes. However, the number of these consensus sequences available for *Saccharomyces cerevisiae* is very small (21 in TRANSFAC, 24 in SCPD) which is why they were not utilised here.

Having presented a negative perspective on these results, it should also be pointed out that the metrics used in these methodologies account for the presence of ubiquitous sites through comparisons with random expectation. Our initial expectation was for the discovery of conserved sites or patterns in the URSs of genes clustered through virtue of correlated expression profiles. We were not overly enthusiastic about the possibility of discovering these links in more general groupings of genes (represented here by the broad MIPS and KEGG categories). However, the observed results appear to be in contradiction to these expectations. This method appears to be better at detecting conserved patterns in the general functional categories than in categories that are more

specific or in the microarray and northern clusters. However, this behaviour is not consistent enough to be used in a predictive fashion.

In the next section we describe the development of a visualisation tool that was implemented in order to visually search for conserved patterns in the URSs of functionally related/grouped genes. We believed that, if conserved motifs could be recognised by eye then an algorithm could be defined to efficiently capture this information.

4.3 SiteSeer – Visualisation and analysis of transcription factor binding sites in nucleotide sequences

Visualisation of the binding sites in the URSs of functionally related genes may help us to better understand the poor predictive capacity of the methods explored so far. This tool was originally designed for visualising the 800 bp upstream regions of genes from the yeast *Saccharomyces cerevisiae* but was later extended to allow the analysis of sequences from any organism. The tool, SiteSeer, was produced as a CGI based web-tool to allow public access and has been made publicly available (Boardman *et al.* 2003, web ref 29).

4.3.1 Overview

The software, SiteSeer, searches input nucleotide sequences for transcription factor binding sites and creates a graphical representation of the results. Binding site sequences were extracted from TRANSFAC public release version 6.0 and the *Saccharomyces cerevisiae* Promoter Database, SCPD. The user may select to search with sites from a single organism, from a taxonomic group (yeasts, plants, viruses, animals or insects) or with user-defined sites. An expectation value and an expectation ratio are calculated for each site. These two metrics provide an estimate of the significance of occurrence of the site and can be used as a filter to remove low-significance sites from the final output.

4.3.2. Program usage

4.3.2.1. Sequence input

Figure 4.16 shows a screenshot of the SiteSeer input page with an example set of input parameters. Sets of FASTA-formatted nucleotide sequences can be 'cut-and-pasted' into the sequence input box (top left of the figure). This tool was originally designed for the analysis of *S. cerevisiae* sequences and, as a reflection of this, there is an input box for *S. cerevisiae* systematic ORF names (e.g. YCR035C, YDL266W). If these are supplied, sequences encompassing 800 bp upstream of their translational start sites are

retrieved from a MySQL database. These input methods may be used in combination to provide the input sequence set.

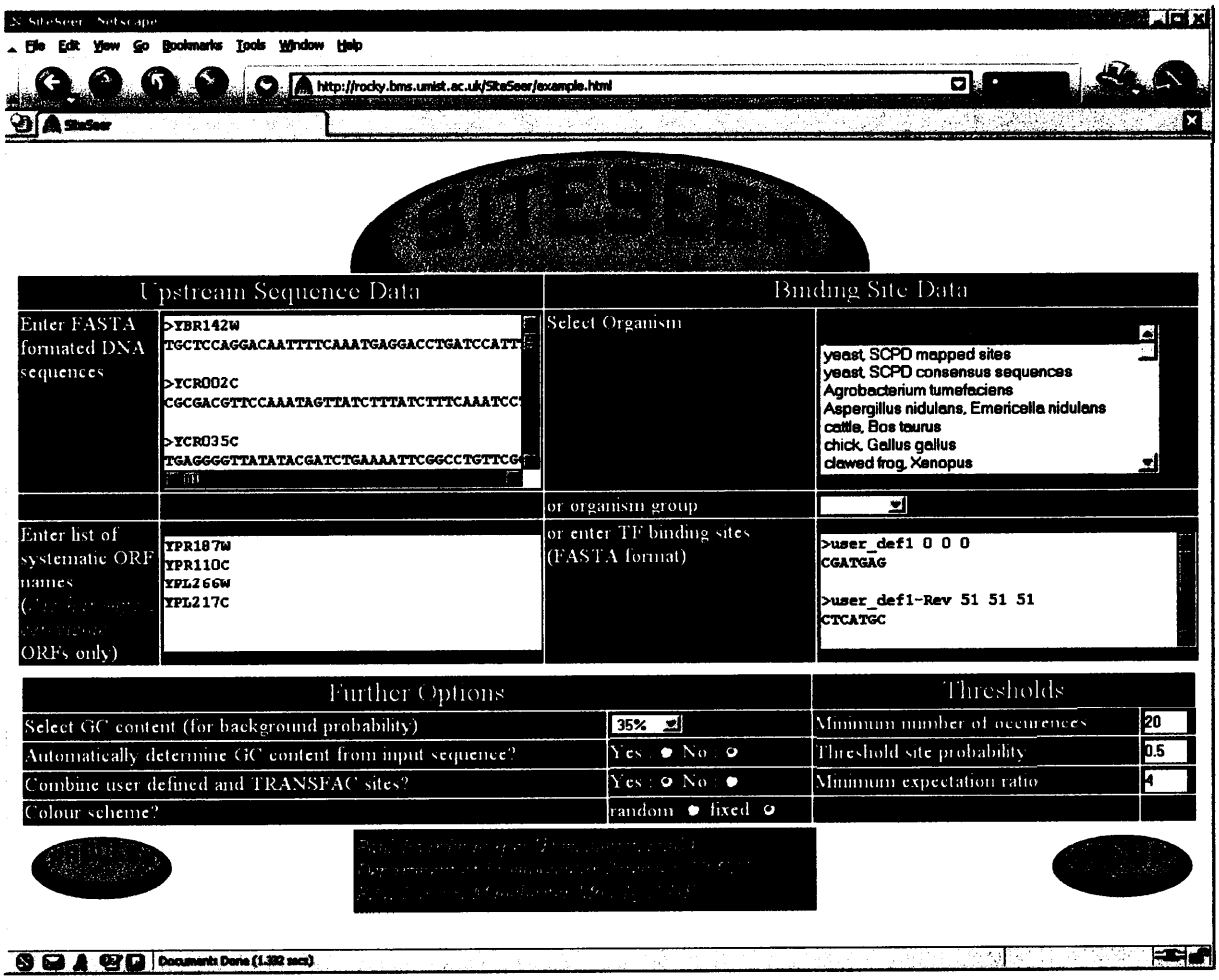


Figure 4.16. SiteSeer front page (http://rocky.bms.umist.ac.uk/SiteSeer/).

4.3.2.2. Binding-site selection

There are three ways of defining the set(s) of binding-site sequences to investigate. Sites from the TRANSFAC public release (version 6.0) were separated by organism and by taxonomic group and stored in a MySQL database for quick retrieval. Additional, user-defined, binding-site data may also be entered in the input box provided (sequences must be in FASTA format). These user-defined sites may either be entered as exact sequences (e.g. 'CAGATA') or via a simple regular expression grammar, using standard IUPAC codes (e.g. the pattern A-[any nucleotide]-[A or T]-G is represented as the string 'ANWG').

4.3.2.3 Thresholds

It was not our intention to develop a novel over-representation statistic for promoter discovery, rather we planned to allow known or user-defined sites to be easily visualised. However, we have included a simple statistical measure for over-representation derived from the input sequences themselves, and based on GC content and binding site probabilities (using a simple zero-order Markov Model). We define two metrics: the expectation value and the expectation ratio. The expectation value of a site in a given sequence is a measure of the likelihood of the site occurring in the upstream region. Let l_q be the length of the query sequence, l_s be the length of the binding-site sequence and $p(x_i)$ be the probability of finding nucleotide x at position i in the binding site. Then, we define the expectation value of a site as:

$$(l_q - l_s + 1)\prod_{i=1}^{l_s} p(x_i)$$

Equation 4.1. Expectation value.

The expectation value can be thought of as the number of occurrences (of a specific binding site) you would expect to find by chance in a given sequence. For example the ADE4 binding site 'ATTGCCGTA' has a zero order Markov chain probability of $0.325^5 \times 0.175^4 = 3.4E^{-6}$, given a background GC content of 35%. In our analyses, binding site sequences are searched against URSs of 800 bp. In one of these searches the total number of possible matches for the above site would be 800 – 9 + 1 = 792 (where $l_q = 800$ and $l_s = 9$). The expectation value for this site is then calculated as $792 \times 3.4E^{-6} = 2.69E^{-3}$.

The expectation ratio builds on this measure by taking into account the actual number of sites detected in a sequence. This is calculated by dividing the number of occurrences of a specific binding site in an upstream sequence by the expectation value of the site. If we found a single occurrence of the above site in an URS then the expectation ratio would be $1 \div 2.69E^{-3} = 371.36$. For many simple, low-complexity sites, the expectation value is higher and consequently the expectation ratio is often greatly reduced unless an

unusually large number of sites are found in a single URS. For example, when searched against an 800 bp sequence the GATA binding site 'GGGG' has an expectation value of $792 \times 0.175^4 = 0.74$ and an expectation ratio of 1.35 (if only a single occurrence is found). For this GATA site to have the same expectation ratio as the ADE4 site above there have to be over 275 occurrences found in a single URS.

The chief reason for the inclusion of these thresholds is to mask low-complexity "uninteresting" sites, such as the many small sites present in TRANSFAC and SCPD and ubiquitous TATA sites. Thus, users can remove potential noise from their sequence representations by choosing an appropriate threshold value, on any of these parameters, below which binding-sites are not drawn. In the example shown in Figure 4.17, the significance of the user-defined site STRE (the stress response element, "AGGGG") is masked by many low-complexity sites. By applying threshold limits, the less significant sites are removed from the display, leaving only the most significant sites (in this case, the STRE elements). We believe that the expectation ratio, in particular, provides a simple, intuitive, threshold measure for users to reduce the complexity of the final display by masking low significance sites.

4.3.2.4. Colour scheme

There are two options for colouring the binding sites. By default, binding sites are randomly allocated one of the 215, non-white, web-safe colours (web ref 30). The user may also choose a fixed colour scheme where the colour allocated is dependent on the name of the binding site. This ensures consistency between images, but can result in an image that is difficult to interpret as sites with similar names often have similar colours allocated to them. Finally, colours can be individually assigned for user-defined sites by including an RGB value after the site name. For example, inputting the definition line ">STRE 0 0 0" would ensure the colour black is used when drawing all STRE sites.

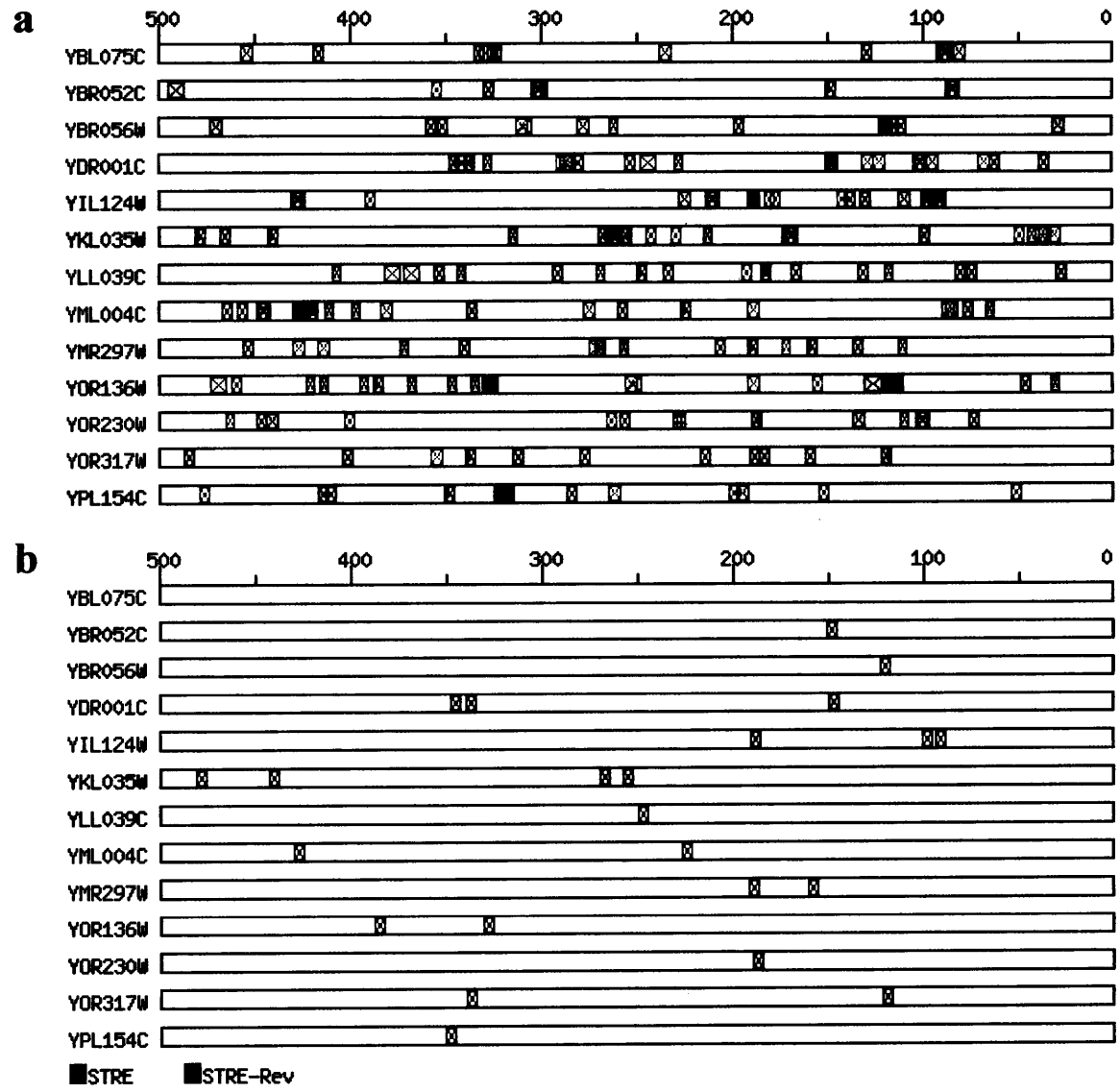


Figure 4.17. Visualisation of the upstream regions of clustered yeast genes from Gasch *et al.* (2000).

a) Data shown with default thresholds and a user-defined binding site (STRE).
b) The same data, but using threshold values to reduce the complexity of the display and to highlight potentially important sites (thresholds used were: Minimum occurrence = 8, maximum expectation value = 0.45 and minimum expectation ratio = 4).

4.3.3. Output

SiteSeer creates an image in PNG (Portable Network Graphics) format to represent the results. Each input sequence is displayed as an empty rectangle whose width (in pixels) is identical to the length of the original sequence (in nucleotides). A scale is provided at the top of the image to aid in interpretation. Individual sites are represented as a coloured cross, bounded by a rectangle of the same colour. The width of this rectangle is proportional to the length of the binding site. This image is part of a client-side

image map, which allows further exploration of the data through interaction with the map. Placing the mouse over any mapped site provides brief details on the site and the number of occurrences, whilst clicking on a binding-site brings up further, more detailed, information about the site in the lower frame of the browser window. This includes expectation values, accession number, database and sequence data. A second program, ***display_tf_data***, is sent the accession number and scores for the selected binding site. The accession number is used to retrieve further details pertaining to the site from a MySQL database. The ***display_tf_data*** program then outputs the results as HTML to the browser.

4.3.4. Application to yeast URS analysis

We used SiteSeer to visualise the URSs of the functionally grouped genes from the previous sections. A number of different threshold values for each visualisation were applied with the hope of finding conserved signals/patterns in the URSs of these genes. The complete results of these scans are available on the attached CD and via web references 31 to 36. Below we discuss some of the groups of genes that score highly in the absolute and relative position analyses.

4.3.4.1 Clusters from Eisen *et al.* (1998)

The clusters that have the most striking results for the absolute and relative position analyses were cluster E and cluster H. Cluster H was previously shown to be composed of four pairs of proximal genes that share large portions of their URSs. This causes many relationships to be discovered, especially when there are multiple instances of the same site present in a single URS. Figure 4.18 shows two SiteSeer visualizations for cluster E using two different sets of parameters. Using reasonably typical parameters (Figure 4.18, A) produces an image that is rich in binding sites and many AT rich sites. Indeed, even with the strictest thresholds used here (Figure 4.18, B), the AT rich repeats in the URSs of ORFs YKL152C and YLR153C are still apparent (solid pink boxes). The TRANSFAC site, CHA1_03, appears within the 350 bp proximal to the site of transcription start in the majority of genes. The sequence for this binding site is 'TATATAAA', which is identical to the SCPD TATA box sequence TATA,TBP_18. The only over-represented, non-TATA related, site in these sequences is CLN2_01

(which has the sequence 'CGCGAAA') although this site does not appear to demonstrate any conservation in distances or patterns with other sites in these URSs. Other clusters with notable over-represented sites detected using visual inspection with SiteSeer are listed in Table 4.14.

Table 4.14. Clusters from Eisen *et al.* (1998) that contain common over-represented binding sites.

Cluster	Description
C	Ten out of 27 URSs contain a REB1_17 site. Seven of these are within 200 bases of translational start.
G	Eight out of 14 URSs contain the HAP4_01 site 'CCAATCA'. Six of these are within 400 bases of translational start.
J	Four out of the five URSs contain an SCB_04 site within 350 bases of translational start.
K	Eight out of 15 URSs contain the HAP4_01 site, all of which are within 350 bases of transcriptional start.

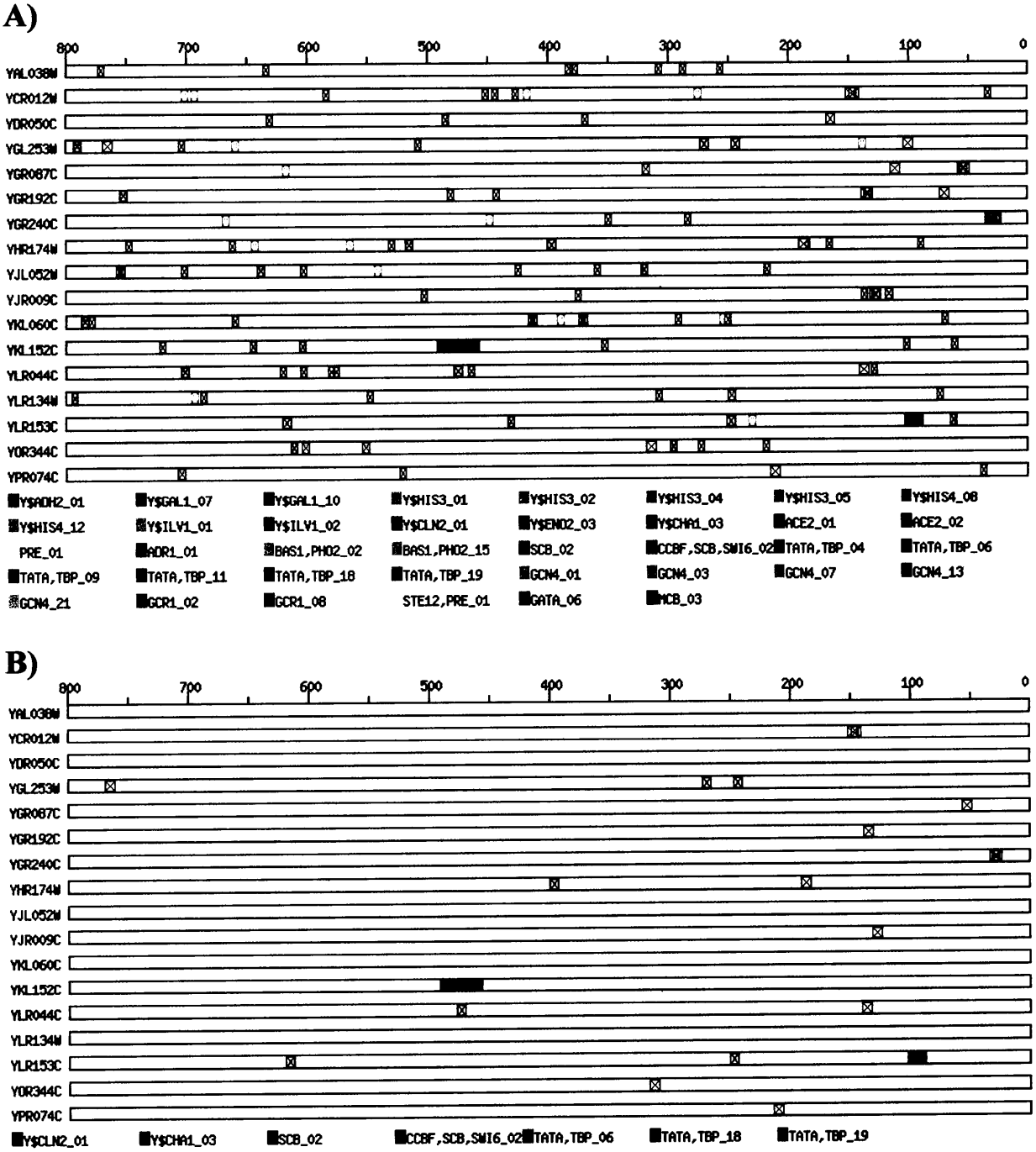


Figure 4.18. SiteSeer visualisations for cluster E from Eisen *et al.* (1998)

A) Visualisation using minimum expectation ratio = 4, maximum expectation value = 0.5, minimum number of occurrences = 3.
B) Visualisation using minimum expectation ratio = 10, maximum expectation value = 0.3, minimum number of occurrences = 5.

4.3.4.2. Northern analysis clusters from Brown *et al.* (2001).

Cluster 8 is the only cluster from this set that has consistently large positive scores for both relative and absolute position analyses. It was obvious that these scores were due to AT rich sequences from the dramatic fall in scores when TATA boxes were removed from consideration in the absolute position experiments. Figure 4.19 shows the SiteSeer visualisation for cluster 8 and show that these scores were due to two stretches of AT rich sequences (solid pink boxes) in the URSs of YDL046W and YNR035C.

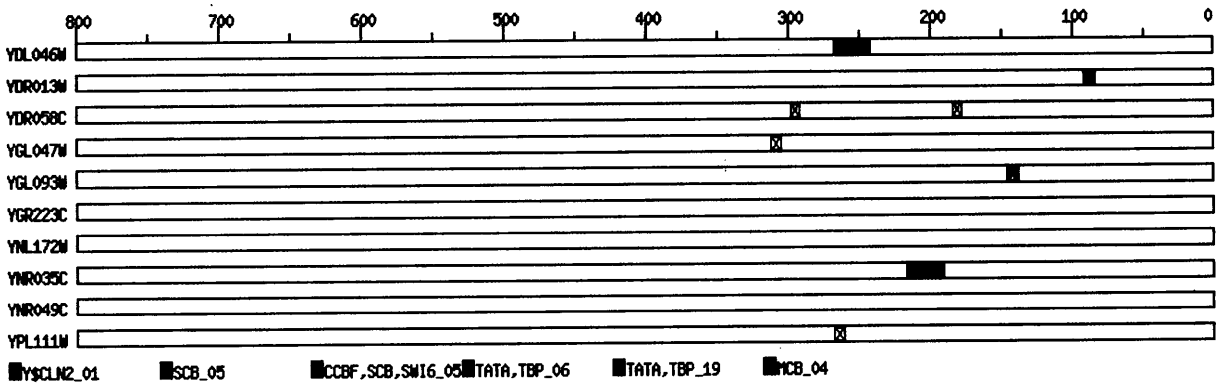


Figure 4.19. SiteSeer visualisation of cluster 8 generated from northern data (Brown *et al.* 2001).

Parameters used to generate this image were: minimum expectation ratio = 10, maximum expectation value = 0.3, minimum number of occurrences = 3

In the analysis using URSs with low complexity sequences masked out, the PCK1 containing cluster (cluster 4) has the highest score of the few positively scoring clusters. The visualisation of the URSs in this cluster reveals no consistent binding site relationships (Figure A4.15, Appendix 4). Similarly, a visual inspection of all the other clusters revealed an absence of conserved binding site relationships or any notable over-represented sites.

4.3.4.3. MIPS clusters

There are a number of specific MIPS clusters that consistently score highly in both the comparisons with randomly picked URSs and in the inter/intra cluster comparisons in the absolute and relative position analyses. These are MIPS categories 10-03, 10-04 and 09-09. Figure 4.20 shows two SiteSeer visualisations for the "Osmosensing pathway" category (10-03). Figures A4.16 and A4.17 (Appendix 4) show visualisations

for the "Nutritional response pathway" (10-04) and "Mitochondrial biogenesis" (9-9) categories.

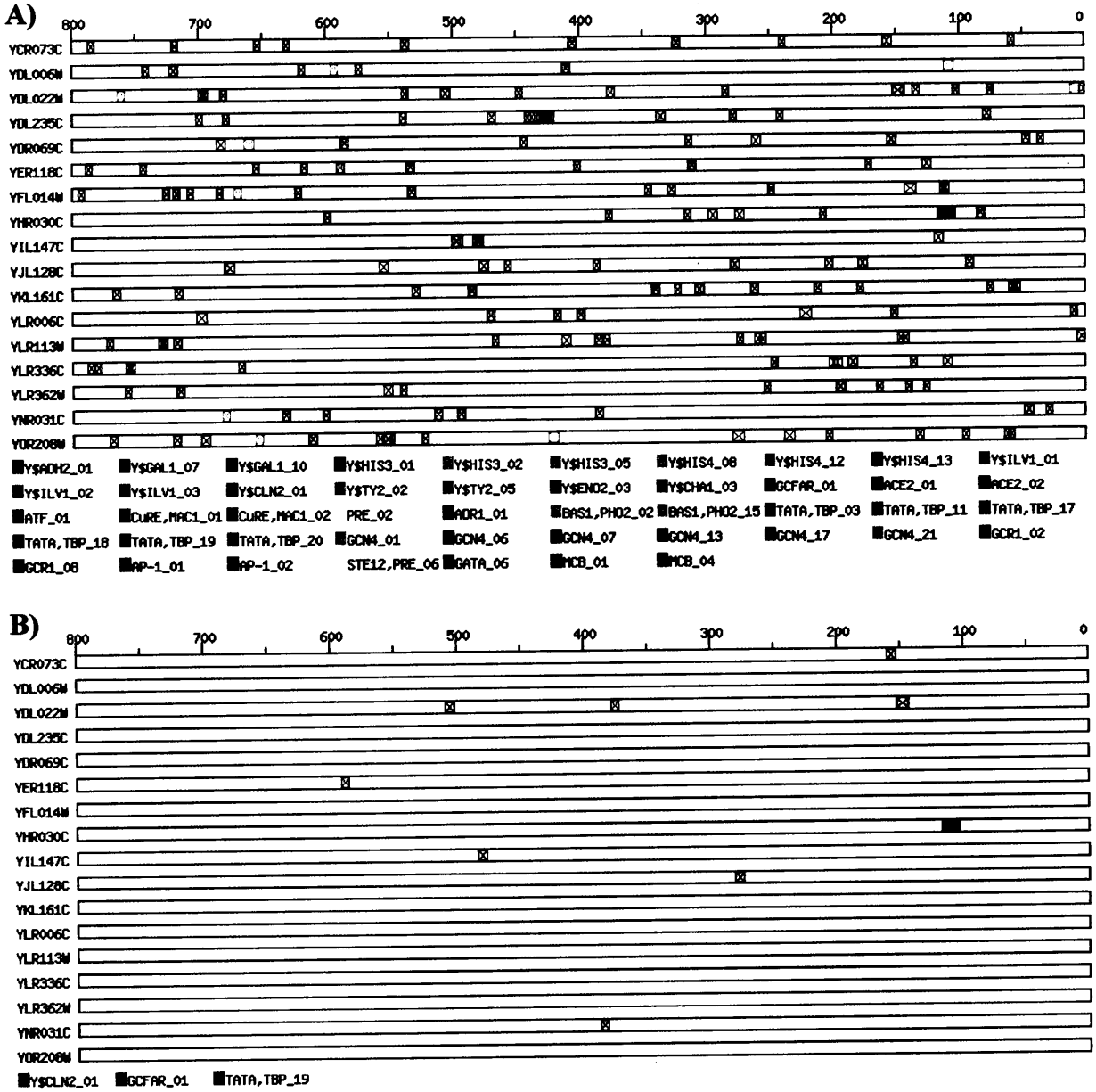


Figure 4.20. SiteSeer visualisations of URSs for genes in MIPS category 10-03 ("Osmosensing pathway").

A) Visualisation using minimum expectation ratio = 4, maximum expectation value = 0.5, minimum number of occurrences = 3.
B) Visualisation using minimum expectation ratio = 10, maximum expectation value = 0.3, minimum number of occurrences = 5.

Again, the visualisations for these categories are less than astounding. There are no obvious conserved binding site relationships within the categories. Nor are there any over-represented sites that are non-TATA sequences. Category 10-04 has a high

number of GATA sites in close proximity in a number of the URSs (Figure A4.17 A)), which may explain the positive scores in the relative position analysis. Similarly, category 9-9 contains a number of proximal GAL1 sites and GATA sites. Other MIPS categories possessing a notable over-representation of non-TATA sites are listed in Table 4.15.

Table 4.15. MIPS categories containing conserved, over-represented, sites.

Category	Description
02-05	28 out of 75 URSs contain the HAP4_01 site. The majority of these fall within 450 bp of translational start.
03-06	40 out of 83 URSs contain either an MCB_02 ('ACGCGT') or an MCB_04 ('ACGCGA') site.
10-02	15 out of 27 URSs contain either anMCB_01 ('TCGCGA') or an MCB_02 site.
13-05	9 out of 28 contain an ACE2 site and 11 contain an SCB_05 site. Where both sites are present in an URS (3 occurrences) the ACE2 site is always upstream of the SCB_05 site.
13-13	Three of the 8 URSs contain a TY2 element.

4.3.4.4 KEGG clusters

The following KEGG categories have consistently positive scores in two or more of the previous analyses: "Methane metabolism" (2.2), "Nitrogen metabolism" (2.5) and "Two-component system" (12.1). Figure 4.21 shows the visualisation for KEGG category 2.5. From this Figure, it is apparent that the consistent positive scores, in the previous analyses, are due to the high similarity between the URSs of ORFs YLR155C, YLR157C, YLR158C and YLR160C. These genes are a set of tandem repeats on chromosome 12, which explains their high scores in this analysis. Neither category 2.2 nor 12.1 contain any apparent conserved binding patterns. Other KEGG categories containing notable over-represented, non-TATA, sites are listed in Table 4.16.

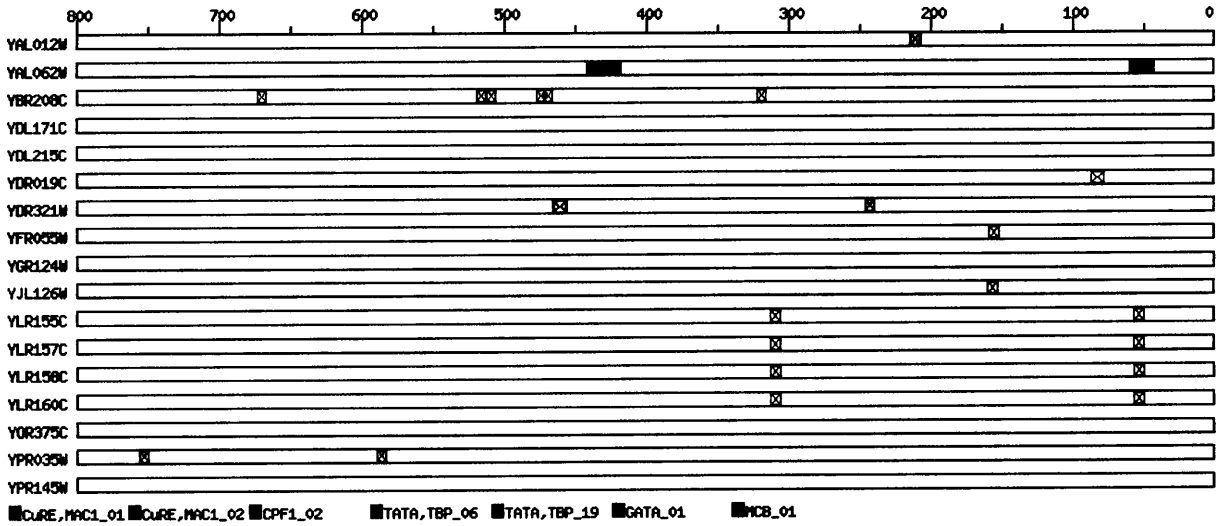


Figure 4.21. SiteSeer visualisations of URSs for genes in the KEGG functional category "Nitrogen metabolism" (2.5).

Parameters used to generate this image were: minimum expectation ratio = 10, maximum expectation value = 0.3, minimum number of occurrences = 5.

Table 4.16. KEGG categories containing conserved, over-represented, sites.

Category	Description
2.1	Overrepresentation of HAP4_01 and CPF1_02 sites. Bias towards first 400 bases upstream from translational start. No other apparent patterns.
5.16	11 out of 15 URSs contain GCN4 sites. Six of the URSs contain multiple instances of the GCN4 site though the spacing is not conserved.
5.7	8 out of 16 URSs contain GCN4 sites within 300 bases of translational start.
6.4	8 out of 16 URSs contain CPF1 sites within 400 bases of translational start.
6.5	Contains the tandemly repeated ORFs seen in Figure 4.21.

4.4 Summary and Discussion

The results obtained so far are generally unsatisfactory. The main goal of this analysis was to produce a method to link functionally related genes using patterns of transcription factor binding sites. It is evident from these analyses that this is a non-trivial task. Some enrichment is seen for a few functional groupings, but overall the effect is very small.

One of the main issues in this analysis is noise. We repeatedly find long AT rich sequences that have adverse effects using our approaches. In order to combat this we tried investigations using binding site data bereft of TATA sites and URSs with low complexity sequences masked out. Although these AT rich regions (and as a consequence TATA sites in the binding site databases) have a negative effect on the analyses, we cannot refute the possibility that they are biologically important. Indeed, TATA boxes are required elements at many polII promoters and the removal of these sites from the analyses is removing essential information. This represents the somewhat paradoxical problem of comparing yeast regulatory sequences; they are AT rich sequences that contain important sites, which themselves are frequently AT rich. A classic "wood from the trees" separation problem and noise represents a key challenge to overcome.

All of the analyses in this chapter and in the previous chapter were carried out on 800 bp stretches upstream of the genes. The size of this region may have added unnecessary noise through the inclusion of sites that are too distal from the transcription initiation site to have an effect. The distribution of mapped sites in the SCPD (Figure 3.2) shows that the majority of sites appear to occur in the first four hundred bases upstream of the genes they affect. Although this is a consideration, many important sites also occur further than 400 bp from the start codon and other investigators (e.g. van Helden 1998) had used the 800 bp definition. Hence, we did not investigate the use of shortened sequences any further and mention it now as a possible course for future investigation.

Another consideration for the poor results is the quality and quantity of sites in the lexicons (SCPD and TRANSFAC). There is a lot of redundancy both internally and between the two databases. Although we have attempted to address this through

iterative refinement, any such refined dataset is unavoidably a compromise between a desire to retain as much potentially significant information as possible and the desire for a truly unique, noise free, set. It is important to note that the sites in TRANSFAC have been mapped from the URSs of only 78 genes and those in SCPD are from a total of 199 genes. This represents quite a shallow coverage of functional classifications (Table 4.17) and a very small number of the total number of sites present in the *cerevisiae* genome.

Table 4.17. Number of genes in MIPS categories with a mapped site.

The TRANSFAC/SCPD coverage (columns 3 and 4) is defined as the number of genes from a particular category that have a mapped site entry in the respective database. For example, category 12 contains 120 genes; none of these genes has an entry in the TRANSFAC database and seven have an entry in the SCPD database.

Category	Number of genes	TRANSFAC coverage	SCPD coverage
1	1025	30	73
2	232	19	32
3	766	29	34
4	708	13	15
5	348	4	10
6	505	3	8
7	303	4	21
8	429	7	5
9	167	2	6
10	122	3	6
11	338	17	9
12	120	0	7
13	2124	32	56
14	8	0	0
15	149	1	1
16	2657	0	1

This implies that there are still a large number of binding sites yet to be characterised and this has been the focus of investigation for many groups. Large numbers of papers have been published describing techniques for the identification of potential transcription factor binding sites (e.g. van Helden 1998, Bussemaker 2001, Hampson 2002, Sinha 2002, Aerts 2003, Zheng 2003). These analyses often discover sites that are over-represented in a particular group of genes and are therefore proposed to act as binding sites for transcription factors. These newly discovered sites are sometimes already represented in the current databases (which is used as a measure to confirm the utility of the techniques). Novel sites possessing the hallmarks of TF binding sites are

also often discovered with these techniques. For example, in the study by Sinha *et al.* (2003) a novel site, CGATGAG, was discovered upstream of 45 genes in the MIPS functional category "rRNA transcription". This site demonstrates a definite positional bias towards the translational start site (Figure 4.22).

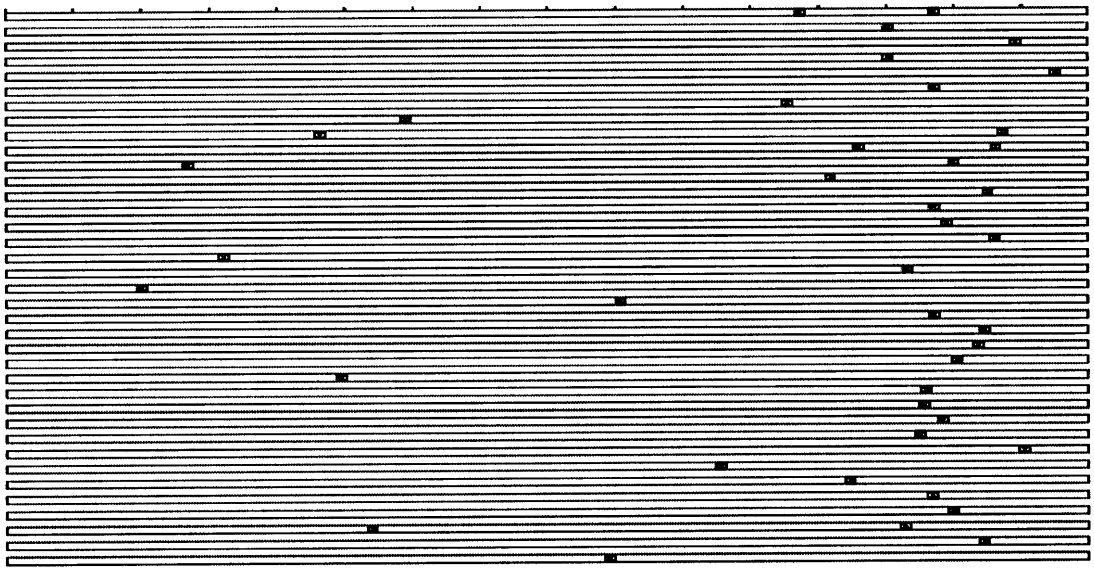


Figure 4.22. Compressed SiteSeer visualisation of the CGATGAG sequence in 37 URSs from the MIPS functional category "rRNA transcription".

Both the MIPS and KEGG functional classifications contain genes that share very broad functions. It is not necessarily a sensible assumption that URSs of the genes in these broad categories should contain some form of conserved signal, as it is currently unclear exactly how an organism could switch them all on or off, or more importantly, why this would be advantageous. Another consideration is that many ORFs have multiple classifications. The ORF YMR043W, for example, is placed in twelve MIPS categories and the ORF YER073W is placed in 14 different KEGG categories. To search for conserved signals in the URSs of these genes implies the assumption that there is some form of conserved signal in the URSs of all genes in all the functional categories to which these ORFs belong.

These analyses have been performed with the assumption that regulation is performed predominantly on the "micromanagement" level. That is to say, that sites in the region directly upstream of ORFs are the major factors in determining the expression profiles of those ORFs. It is becoming increasingly apparent that this assumption is incorrect, or at least somewhat naïve. Regulation of gene expression also occurs on the chromatin

level. The expression of multiple genes may be simultaneously induced through the remodeling of chromatin. The transformation of heterochromatin to euchromatin opens up areas of the chromosome and allows basal transcription of all genes within this domain. Further control of the transcription levels of these genes may be influenced through the action of individual transcription factors that either induce or repress expression. The genes within these domains may not be functionally related but will still exhibit correlated expression patterns (Cohen 2000, Spellman 2002). We investigated this by producing sets of chromosomal correlation maps using microarray data from Gasch *et al.* (2000). These correlation maps were generated by calculating the correlation coefficient of the expression levels of all genes in each chromosome across a range of 155 experimental expression points. This data was visualised using the ***draw_jock_plot*** program. Figure 4.23 shows a correlation map for the right arm of chromosome 1. A large portion of this arm is coexpressed. The majority of genes in this arm are unclassified and there appears to be no association in the functions of those few that are functionally classified.

To conclude, the studies so far have revealed a number of deficiencies in standard approaches to the analyses of regulatory regions and transcription factor binding sites, particularly in the definition of important binding sites and the quality and quantity of sites in the current databases. In order to address this, further studies were required to define a set of words (or binding site sequences), determined in an unbiased way, that were capable of distinguishing functionally related genes from random groupings.

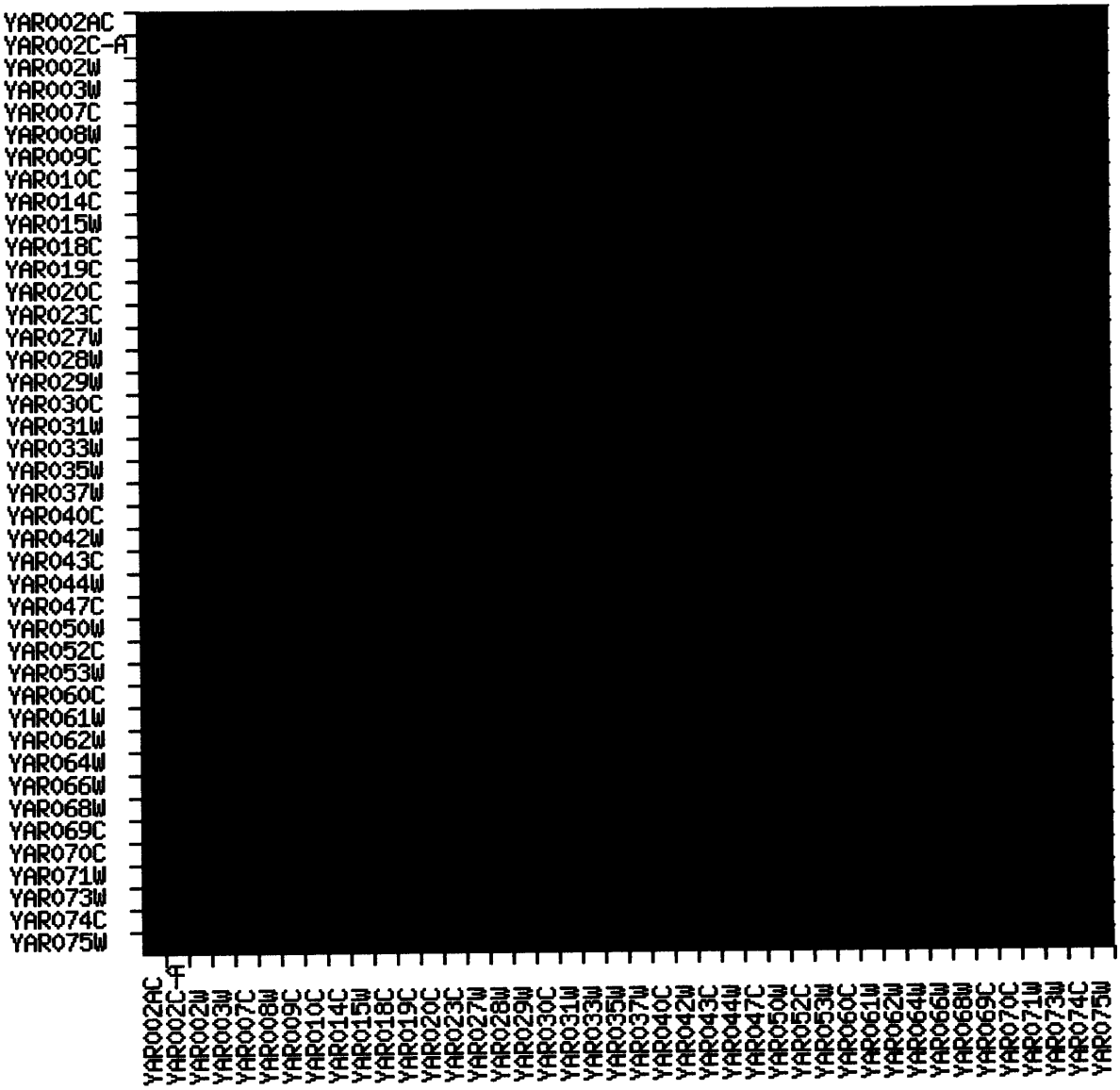


Figure 4.23. Portion of Chromosomal correlation map for *Saccharomyces cerevisiae* chromosome A showing coordinated expression patterns for large sets of proximal genes.

This data was generated using microarray expression data from Gasch *et al.* (2000). Red represents correlated expression. Green represents anti-correlated expression. This plot was scaled so that a correlation of ± 0.7 is represented by the maximum possible intensity of colour. Correlation values were not calculated for self-self comparisons (diagonal).

Conservation of regulatory sequences in yeast species

5.1 Available data

This analysis was made feasible with the completion of the Génolevures project. This involved large-scale comparative genomics between *Saccharomyces cerevisiae* and 13 other yeast species representative of the various branches of the Hemiascomycetous class (Souciet 2000, web ref 38, Table 5.1). The first stage of this project required the generation of a large set of novel sequences derived from the 13 yeast species. Importantly, this sequencing was not biased towards coding regions but was produced by sequencing random genomic libraries. The term random sequence tag (RST) is used to describe sequences generated in this fashion. This approach not only produces sequence data containing genes with detectable homology to *S. cerevisiae* sequences, but also sequences which continue into the URS of their *cerevisiae* homologues. This allows a comparative analysis of the upstream regions of homologous genes from the different yeast species.

Table 5.1. Sequence data obtained from the Génolevures project.

This shows a breakdown of the Génolevures project sequence data (downloaded from the genoscope web site, web ref 37). The number of sequences and total number of bases sequenced for each species is detailed as well as basic information from a BLASTX comparison against the *Saccharomyces cerevisiae* proteome. Percentage values in the "number of hits to *S. cerevisiae* proteome" column represent the percentage of the total number of sequences from this yeast that find a match in the *cerevisiae* proteome.

Species and species code	Number of sequences	Total number of bases (Mb)	Number of hits to *S. cerevisiae* proteome	Ave. match length (bp)	Ave. e-value
Candida tropicalis (Ct)	2541	2.38	1249 (49%)	182	$1.8E^{-5}$
Debaromyces hansenii (Dh)	2830	2.70	1747 (62%)	170	$1.6\ E^{-5}$
Hansenula polymorpha (Hp) *(P. angusta)*	5082	4.85	3748 (74%)	180	$1.1\ E^{-5}$
Kluyveromyces lactis (Kl)	6080	5.27	4268 (70%)	179	$8.5\ E^{-6}$
Kluyveromyces marxianus var. *marxuanus* (Km)	2496	2.36	1819 (73%)	181	$6.5\ E^{-6}$
Kluyveromyces thermotolerans (Kt)	2653	2.44	1846 (70%)	175	$1.3\ E^{-6}$
Pichia sorbitophila (Ps)	4829	4.03	2951 (61%)	170	$1.4\ E^{-5}$
Saccharomyces exiguus (Se)	2578	2.38	1781 (69%)	189	$8.3\ E^{-6}$
Saccharomyces kluyveri (Sk)	2528	2.48	1957 (77%)	193	$4.4\ E^{-6}$
Saccharomyces sevazzi (Ss)	2570	2.31	1920 (75%)	184	$6.3\ E^{-6}$
Saccharomyces bayanus var. *uvarum* (Su)	5140	4.84	4560 (89%)	198	$1.4\ E^{-6}$
Yarrowia lipolytica (Yl)	4940	4.92	1821 (37%)	162	$1.6\ E^{-5}$
Zygosaccharomyces rouxii (Zr)	4936	4.37	3562 (72%)	184	$6.1\ E^{-6}$

5.1.1 Extracting promoter regions of homologous sequences

Homologues to *S. cerevisiae* from the Génolevures data were obtained by comparing the RSTs with the *S. cerevisiae* proteome using BLASTX. These homologues were refined into a useful dataset of sequences containing promoter regions using the program, ***analyse_STS_vs_Sc_blasts.pl***. RST sequences were extracted if they aligned to within five amino acids of the start position of their homologous *cerevisiae* sequence, had an e-value of $1E^{-10}$ or better and contained a putative upstream region spanning 50 bases or more. We define the upstream region of an RST as the sequence that, in a genomic alignment, is located before the start site of the *cerevisiae* gene. The URS containing sequences were separated into two datasets; one containing upstream sequences and the other containing open reading frames.

To maximise the number of orthologues and minimise the number of paralogues extracted, only a single sequence (that with the best expectation value) from each yeast species was extracted per *cerevisiae* protein. It is hoped that the match with the best score with a *cerevisiae* protein is the most likely to be a true orthologue rather than a paralogue. This is an important step if one wishes to examine URSs for conserved binding sites due to the potential for significant divergence in the regulatory mechanisms of orthologous and parologous genes. A study by Gu *et al.* (2002) found that a large proportion of duplicated genes in yeast have diverged quickly in expression and that the vast majority of gene pairs eventually develop divergent expression profiles. This implies a corresponding change in the URSs of these genes to facilitate these changes. Therefore, it is important to remove paralogues from this analysis.

A total of 4,544 Génolevure RSTs matched to 2,719 *S. cerevisiae* proteins, giving an average of just under two homologous genes per *cerevisiae* protein.

A "gold standard" dataset was created from this homologue set for further analysis; *S. cerevisiae* genes with homologues in at least three other yeasts were used to create this dataset. This resulted in a reduced set of 2,036 sequences covering 450 *cerevisiae* genes. A series of BLASTN and ClustalW comparisons of these sequences with their *cerevisiae* orthologue are shown in Table 5.2. The general outcome of this analysis is

in agreement with the currently accepted evolutionary relationships of these yeasts (Wong 2002). Those yeasts that are thought to have most recently diverged from *cerevisiae* have the highest percentage identities when comparing ORFs and URSs.

Table 5.2. Number of hits and the average values of those hits from yeasts in the "gold standard" dataset to *Saccharomyces cerevisiae*.

Sequences from the "gold standard" dataset (those cerevisiae genes with homologues in 3 or more of the Génolevure yeasts) were compared on a pairwise basis to their *cerevisiae* homologue. Percentage identity was calculated using ClustalW (Thompson 1994). Expectation values were taken from the highest scoring BLASTN HSP (run with the bl2seq program).

Species	Number of hits	Average			
		URS e-values	ORF e-values	URS %Ids	ORF %Ids
Saccharomyces exiguus	90	$3.47E^{-02}$	$1.62E^{-02}$	21.29	63.49
Saccharomyces bayanus var. uvarum	207	$8.46E^{-03}$	$2.33E^{-04}$	41.61	77.56
Yarrowia lipolytica	55	$7.65E^{-02}$	$2.65E^{-02}$	10.75	48.35
Debaromyces hansenii	62	$4.97E^{-02}$	$2.06E^{-02}$	16.48	52.36
Hansenula polymorpha (P. angusta)	143	$6.26E^{-02}$	$2.09E^{-02}$	16.7	48.17
Saccharomyces kluyveri	117	$5.09E^{-02}$	$1.46E^{-02}$	19.07	58.05
Kluyveromyces lactis	198	$4.95E^{-02}$	$1.29E^{-02}$	19.84	55.77
Kluyveromyces marxianus* var. *marxianus	134	$5.23E^{-02}$	$1.64E^{-02}$	16.92	55.96
Pichia sorbitophila	111	$5.21E^{-02}$	$1.58E^{-02}$	13.1	52.59
Zygosaccharomyces rouxii	185	$1.10E^{-01}$	$1.31E^{-02}$	20.91	58.51
Kluyveromyces thermotolerans	107	$4.68E^{-02}$	$1.18E^{-02}$	15.92	54.58
Saccharomyces sevazzii	97	$2.74E^{-02}$	$1.00E^{-02}$	22.47	61.09
Candida tropicalis	63	$6.74E^{-02}$	$1.20E^{-02}$	13.57	55.37

5.2 Conservation of sequence in orthologous upstream sequences

Before attempting to extract potential binding sites from the upstream region of homologous genes in the Génolevures data, it is necessary to examine these regions and determine levels of similarity. BLAST and ClustalW analyses have been described previously (Table 5.2), here we describe an analysis that utilises entropy to investigate sequence conservation.

5.2.1 Entropy measures

Entropy is a measure of the average uncertainty of an outcome. Two entropy measures are employed here; Shannon (or standard) entropy and relative entropy.

5.2.1.1 Shannon entropy.

The Shannon entropy is used here to provide a measure for the degree of conservation between a set of sequences and is defined by:

$$H(X) = -\sum_{i} P(x_i) \log P(x_i)$$

Equation 5.1. Shannon entropy.

Where $P(x_i)$ represents the probability of finding nucleotide x at position i in a set of sequences. In the following analysis, the logarithm base 2 is used. In this case, the unit of entropy is termed a 'bit'. With a DNA alphabet of four characters, if we are maximally uncertain about the outcome of a sample (i.e. no conservation in the nucleotide sequences), the entropy value will be 2. If we are certain of the outcome of a sample (i.e. complete conservation in the nucleotide sequences) then the entropy will be zero.

5.2.1.2 Relative entropy

Relative entropy is often thought of as a measure of the distance between two probability distributions and is useful for finding unusual patterns in biological sequences. Here, we use the relative entropy measure to score groups of sequences with the premise that sequences containing conserved patterns will possess greater scores than seen in a set of random sequences. This term is defined as:

$$H(P \| Q) = \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)}$$

Equation 5.2. Relative entropy.

Where, $P(x_i)$ is the probability of finding nucleotide x at position i in the sequence set and $Q(x_i)$ is the probability of finding nucleotide x at position i in the background model. In this analysis, the background model is defined as the GC content of non-coding regions in the *S. cerevisiae* genome (~ 35%). Scores for relative entropy are opposite from those calculated using the Shannon entropy. A score of zero represents no difference of the sample from the background distribution (i.e. no conservation in the nucleotide sequences).

5.2.2 Analyses

Two analyses were carried out using both standard and relative entropy measures. In the first, the average entropy value over a window of 100 bases was calculated for homologous URSs. In the second, the average entropy of a sliding window of 10 bases was calculated. Conserved sequence features found in the homologous sequences should affect their entropy values. This affect should be visible when comparing these values with those calculated using random sequences of comparable composition and length.

For both analyses, sequences were aligned to the putative translation start site of their corresponding ORF and by ClustalW. A multiple sequence alignment has the potential to affect the entropy values by aligning conserved sequence features that are not present in a strictly conserved position (relative to the site of translation start) in all sequences.

Random sequences were generated for each analysis by applying a Fisher-Yates shuffle to the input sequences (Fisher 1938, Knuth 1997). Three sets of randomised sequences were generated for each set of input sequences. Entropy values are first calculated for a set of 3' aligned URSs sequences (aligned to the translation start site of the *cerevisiae* orthologue). ClustalW alignments were generated for both the real and randomised sequences. Entropy values were then calculated for these alignments.

Preliminary results suggested that the distribution of sequence lengths has a dramatic effect on the entropy values. To account for this the sequences were split into three 'length filtered' datasets with minimum lengths of 400, 600 and 800 bases (Table 5.3). Those *cerevisiae* genes with no homologues after filtering were removed from the analyses.

Table 5.3. Statistics of 'length filtered' datasets.

Minimum length	Number of sequences	Number of *cerevisiae* sequences in dataset	Average number of sequences per *cerevisiae* sequence
400	1550	441	3.5
600	1233	408	3
800	863	328	2.6

5.2.3 Results

5.2.3.1 Average entropy values

Figure 5.1 shows the results of the 100 base average entropy analysis with translation start aligned sequences. Entropy values for the homologous sequences do not differ significantly from random. The difference in entropy value between the 'length filtered' datasets is due to variations in numbers of component sequences. As the number of sequences being compared increases, the overall level of similarity decreases and so we see an increase in the Shannon entropy and a decrease in the relative entropy.

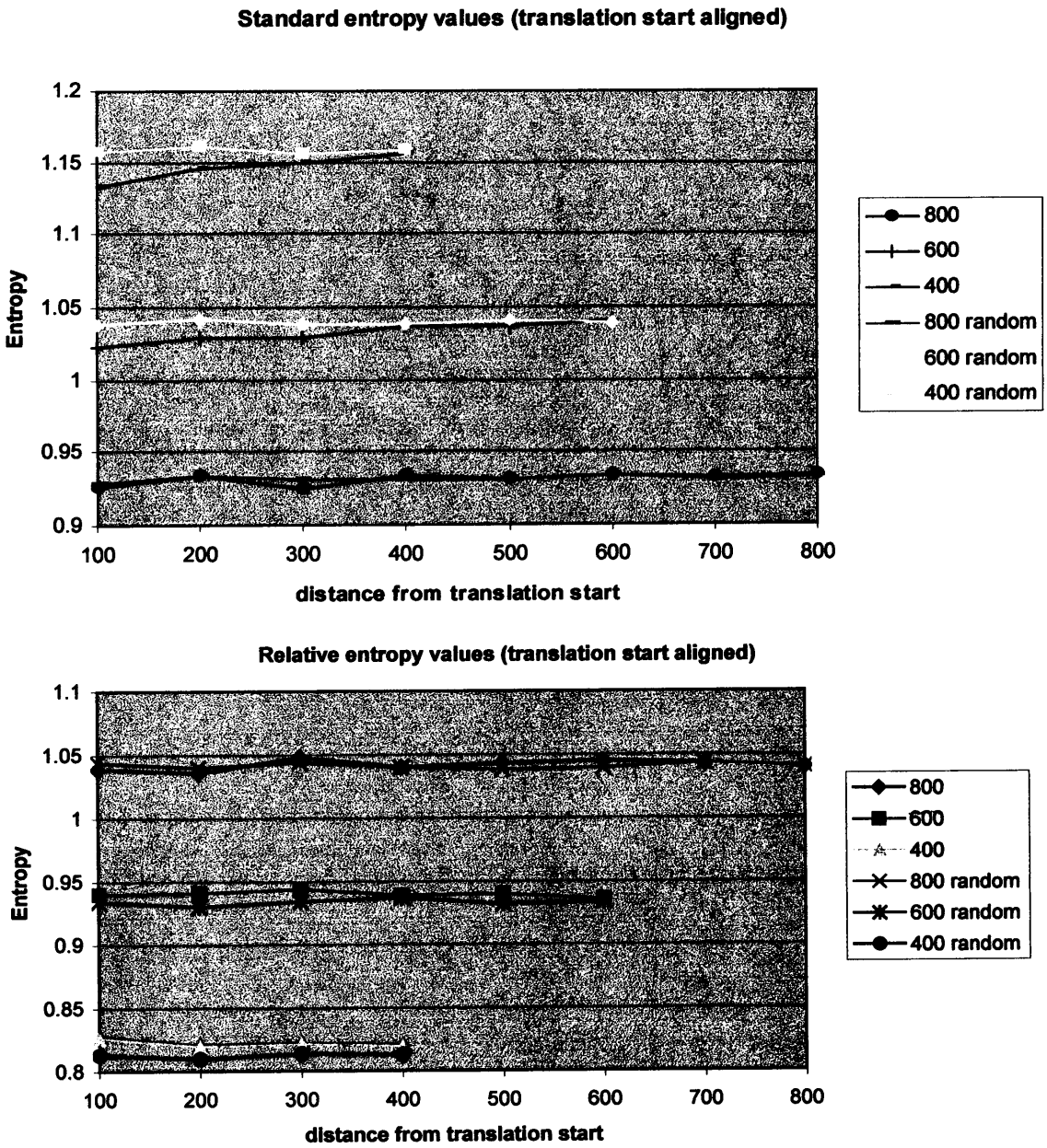


Figure 5.1. Average standard and relative entropy values averaged over 100 nucleotides from translational start for the translation start aligned sequences.

The results of the 100 base averages using ClustalW aligned sequences shows a similar trend as for the translation start aligned sequences (Figure 5.2). From this data, it would appear that ClustalW has a preference to align shorter DNA sequences internally to longer ones. This is seen as lower Shannon entropy in proximity to the site of translation start, which in turn implies that the number of sequences being compared is reduced. This is probably caused by the behaviour of the alignment algorithm coupled with the poor conservation of nucleotide sequence in the homologous non-coding regions.

Entropy values of real data are consistently lower than those of random sequences, though the trend in entropy over distance is very similar to that of the random sequences. This suggests that conserved sequence features are being aligned in the real dataset and that these features are not seen in the random sequences. The presence of a few, highly conserved, sequences in the dataset will also have a similar effect.

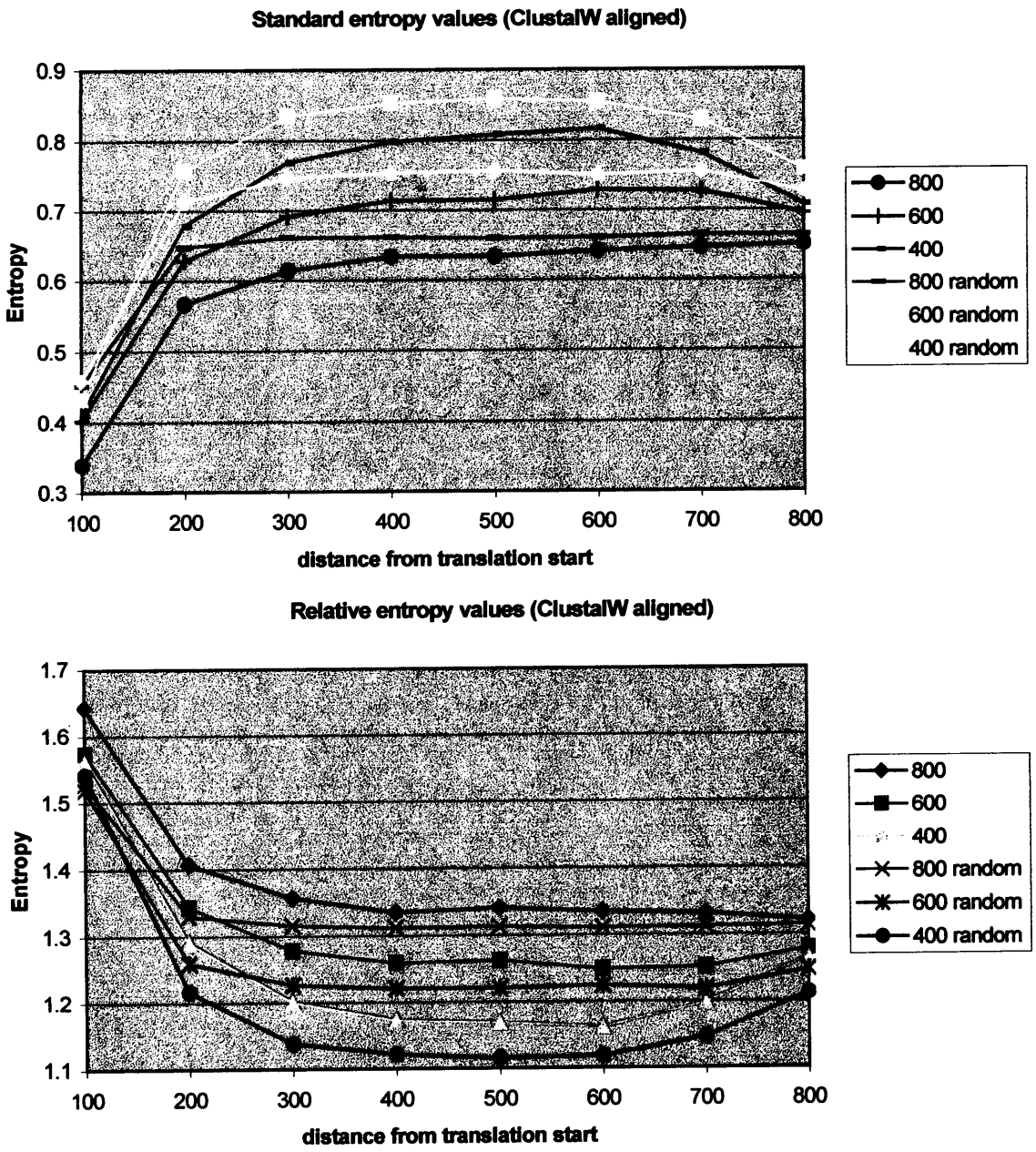


Figure 5.2. Average standard and relative entropy values averaged over 100 nucleotides from translational start for the ClustalW aligned sequences.

5.2.3.2 Sliding window analysis

The results in figures 5.3 and 5.4 show the same trends as for the previous analysis. One feature of note is the reduced standard entropy (and increased relative entropy) close to the putative translational start position in the translation start (3') aligned sequences. A probable cause of this is an AT bias in the non-randomised sequences at this position. This was confirmed by an analysis of the AT content of the homologous sequences (see section 5.2.3.3). This is common in the promoter regions of genes. It

is thought that an increase in AT content decreases the energy required to "melt" the DNA, which is a required before initiation of transcription can occur.

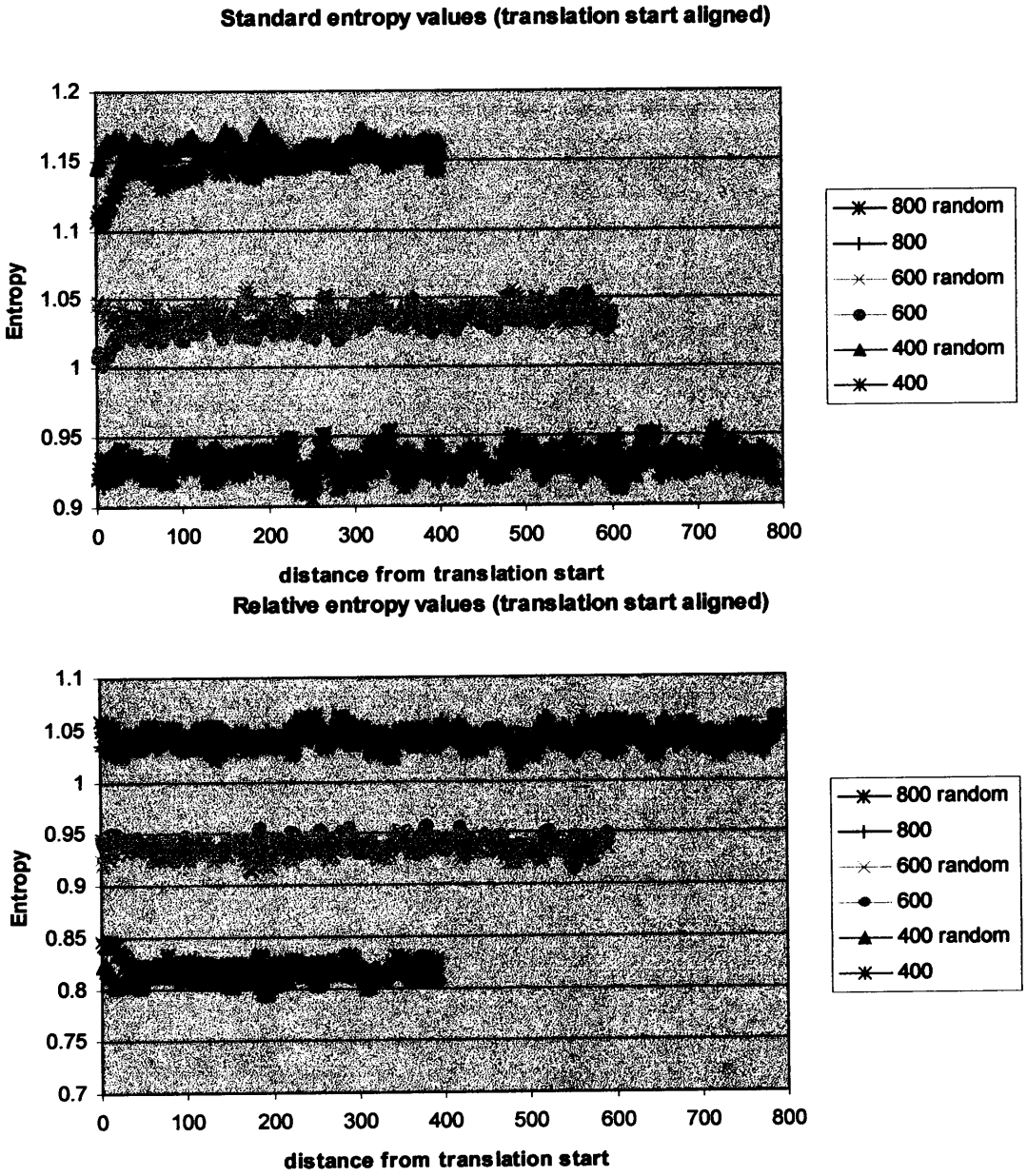


Figure 5.3. Average entropy values using a sliding window of 10 bases and aligning the sequences to their putative translation start sites.

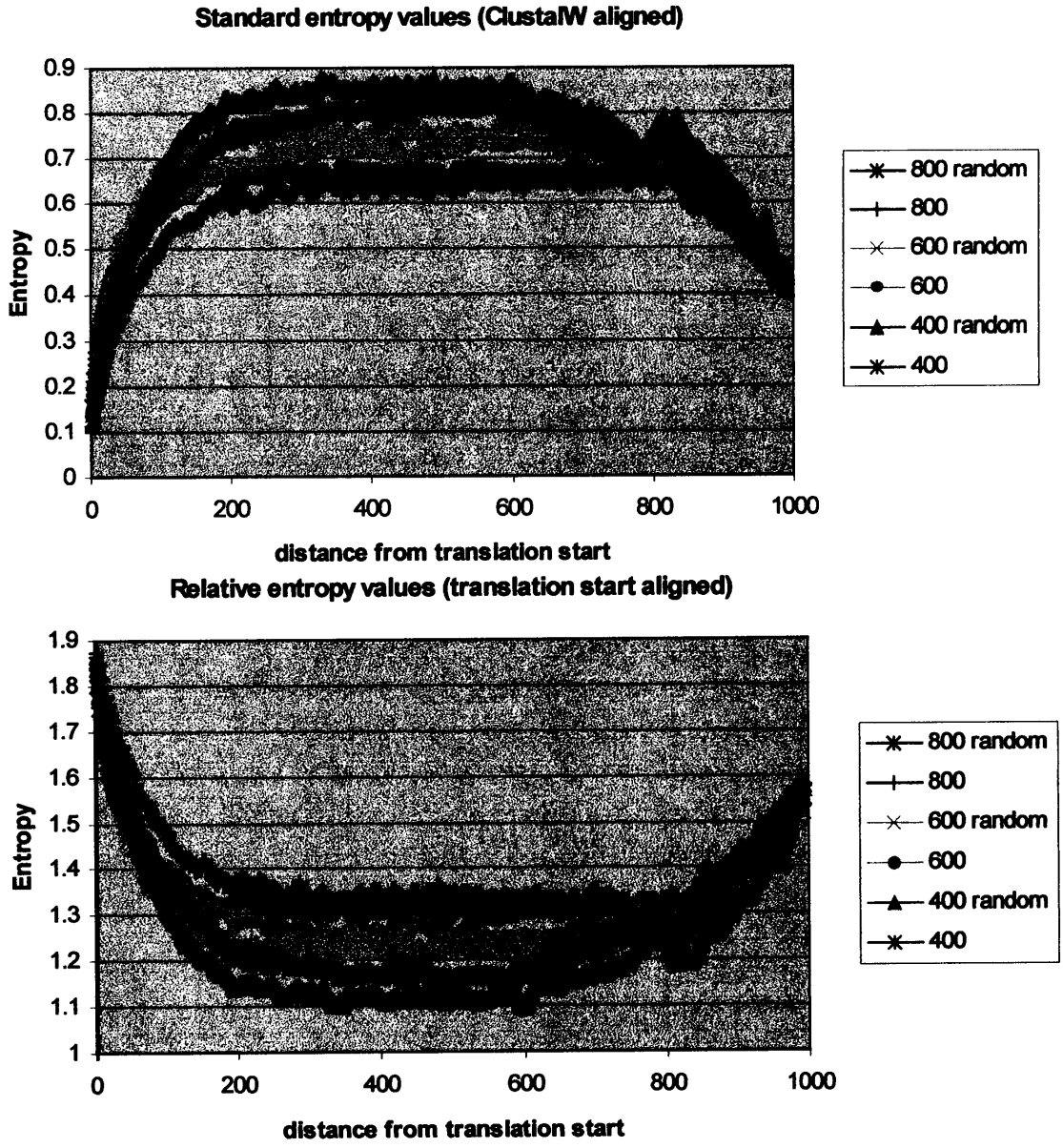


Figure 5.4. Average entropy values using a sliding window of 10 bases , aligned using ClustalW.

5.2.3.3 AT content

The program, ***calculate_positional_at_content.pl***, analyses a set of input sequences (in FASTA format) and outputs the AT content of the sequences over a user defined window size. This program was designed to handle both coding and non-coding sequences, which must be aligned at their 5' and 3' ends respectively.

The ORF and URS sequences for all of the Génolevures data combined with their *cerevisiae* homologues were analysed using this program. A chart of the AT content

for homologous URSs shows an increasing AT content over the 300 bases prior to the translation start site (Figure 5.5). This would appear to be the cause of the decrease in overall sequence complexity observed in the previous section as a fall in Shannon entropy.

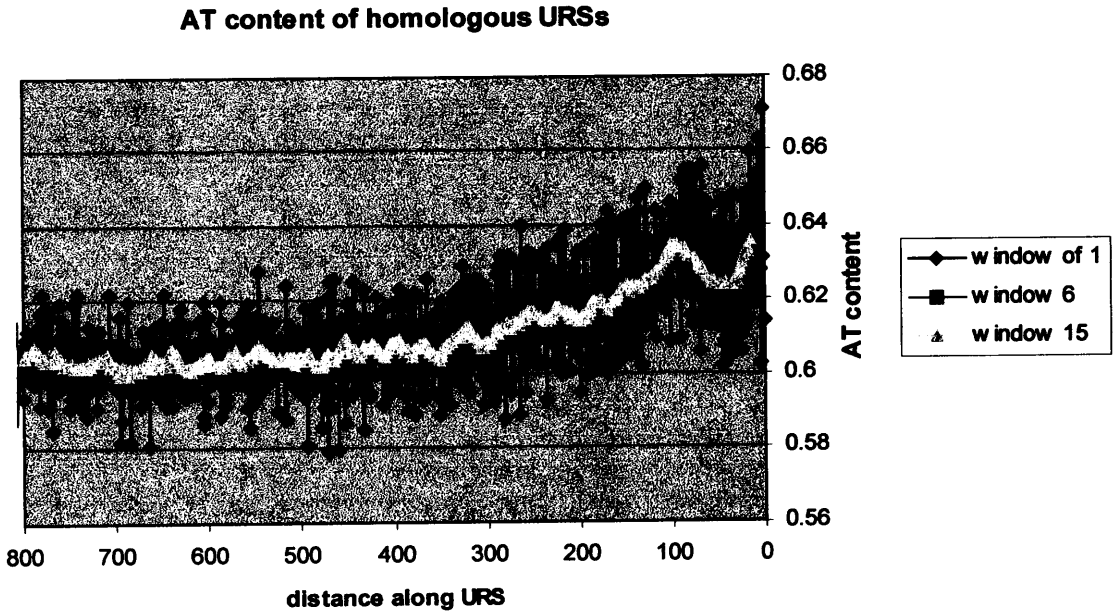


Figure 5.5. AT content of the aligned homologous upstream sequences of the Génolevures data and their corresponding *Saccharomyces cerevisiae* homologues.

As an interesting side note, an AT analysis of the homologous ORFs shows a 3 nt periodicity even when the *cerevisiae* sequences are removed (Figure A5.1, Appendix 5). The nucleotide bias seen in the periodicity is in agreement with the internal codon nucleotide bias seen in *Saccharomyces cerevisiae* (web ref 38). This is confirmation that homologous genes are being selected and aligned in the correct frame. The AT bias seen towards the translation start and the GC bias observed immediately afterwards is further confirmation that homologous sequences are being correctly detected and aligned.

5.3 Mapped sites in the upstream regions of orthologues

Although the orthologous URSs show no significant entropic relationships beyond a general AT bias within the first ~100 bases prior to the site of translation start, it may still be the case that these sequences contain conserved binding sites, but the position of these sites (relative to the putative translation start site) varies between the different yeast species. In this case, alignments of URSs are unlikely to be relevant in all but the most closely related species. Previous studies have shown that the distance of a binding site, from translational start, is not tightly constrained. Distance constraints are more likely to be related to the site of transcriptional start, or to other TF sites within functional "promoter modules" in the URSs. These features are largely uncharacterised for the vast majority of genes. It may be important, in some cases, for a site to be located within a particular region. For example, Sinha and Tompa (2002) found a motif demonstrating an obvious preference for the first 200 bases from the site of translational start in the upstream sequences of 45 genes classified as 'rRNA transcription' in the MIPS functional classification system (Figure 5.6).

To investigate the possibility that the Génolevures orthologue groups contain conserved sites (but that these sites are not in strictly conserved positions) we visualised these groups using the SiteSeer tool described in Section 4.3.

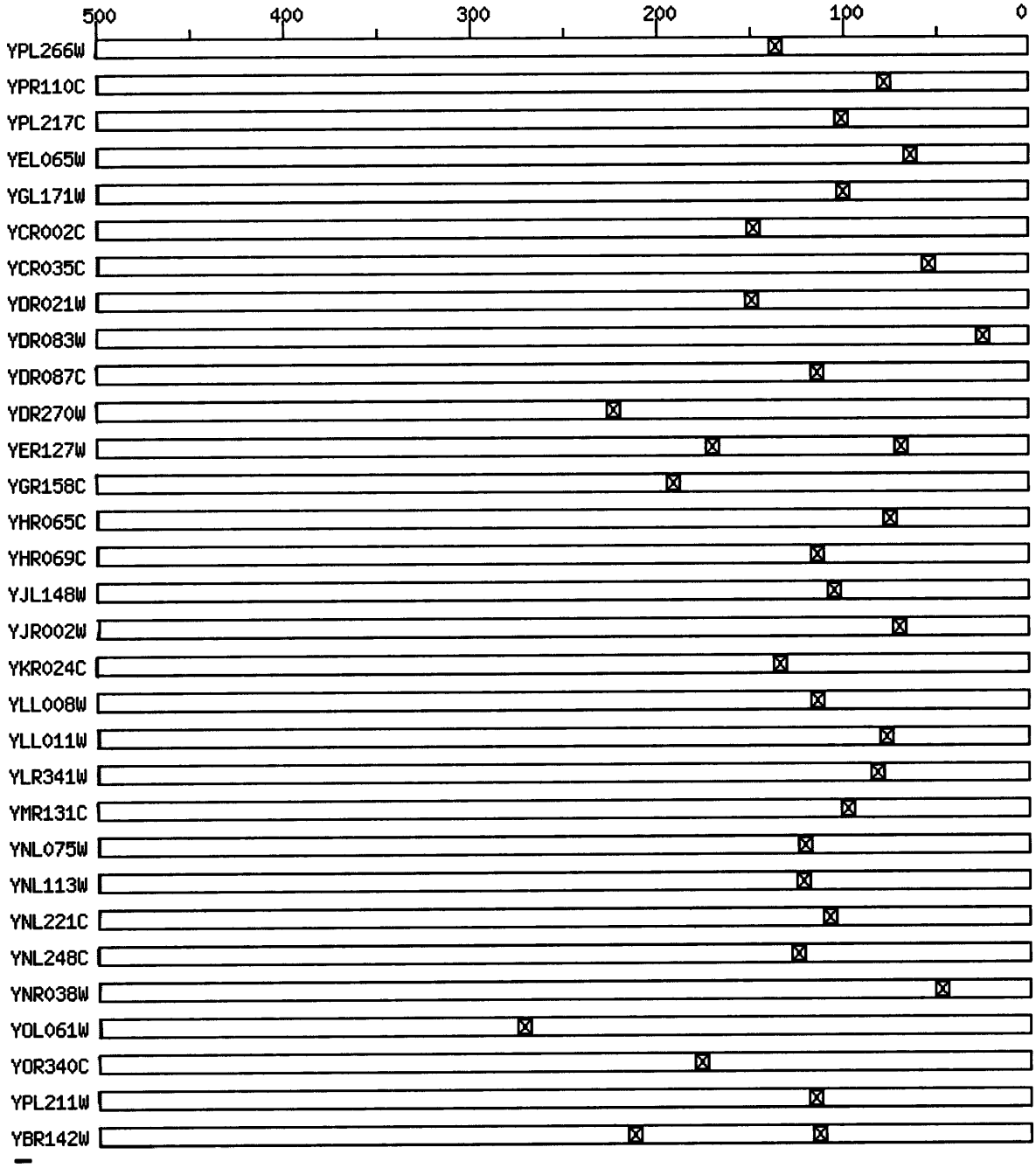


Figure 5.6. Visualisation of the 500 base URSs of 31 genes classified as 'rRNA transcription' in the MIPS functional classification.

This graphic was generated using the SiteSeer program (Section 4.3) and the user-defined motif 'CGATGAG' (represented as a black box) as discovered by Sinha and Tompa (2002).

5.3.1 Methods

Since our previous computational attempts to discover links between genes using conserved locations of binding sites was unsuccessful, a visual inspection of the orthologue groups was deemed necessary. The human eye is far superior to current computational approaches for detecting complex and variable patterns in visual data. Unfortunately, there are over 2,000 groups of orthologous genes in this dataset, which is far too many to visually inspect in the time available. To make the analysis more tractable we reduced the search space to orthologue groups containing six or more sequences. Only a single sequence from each organism is selected when creating these groups, which means that each orthologue group represents data from six or more species. These groups should be representative for the whole dataset. Fifty groups were found to contain six or more sequences. These were visually inspected for conservation in binding site preference and position, using the SiteSeer tool described in Section 4.3.

Four different visualisations, with differing parameter settings, were carried out for each orthologue set (Table 5.4). The different parameter settings represent a gradual increase in stringency for binding site visualisation. This improves the ease and speed of identification of sites with potential biological significance by eliminating smaller, less significant, sites from the visualisation.

Table 5.4. Visualisation parameters for orthologue scans.

Scan	Minimum number of site occurrences	Minimum Expectation ratio	Threshold site probability
1	1	1	1
2	No. of seqs in group ÷ 3	2	0.5
3	No. of seqs in group ÷ 2	4	0.5
4	No. of seqs in group ÷ 1.25	4	0.3

Rather than visualise each group on an individual basis through the web front end of SiteSeer, a command line version of this tool was developed (***SiteSeer_command_line.pl***). The wrapper program, ***batch_siteseer.pl***, allows the input of multiple files to ***SiteSeer_command_line.pl*** and eases the analysis of the output by creating a web page that contains thumbnail images of the results (web ref

39 and 40, Figure 5.7). Clicking on a thumbnail displays the full-size output for the desired analysis.

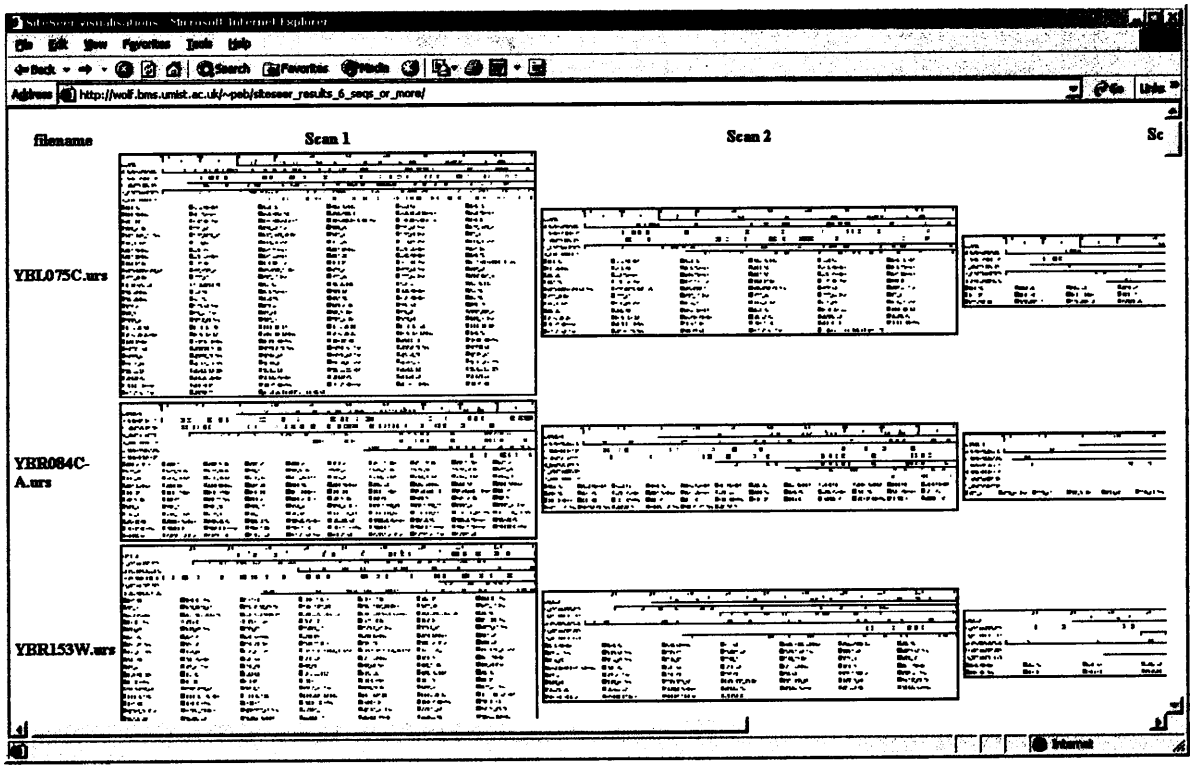


Figure 5.7. Thumbnail image page of SiteSeer visualisations.

Created by analysing 50 orthologous groups containing sequences from six or more yeast species, with the ***batch_site_seer.pl*** program.

5.3.2 Results

Although some groups contain sequences with sites in conserved positions/patterns, the majority of sequences do not have obvious similarities in site preference or order. The sites that occur most frequently in any given group of genes are those that have low-sequence complexity or are AT rich. This implies that these sites are likely to have been detected by chance and are unlikely to be biologically significant. Of the fifty orthologous groups of URSs, 27 show no apparent conservation. However, ten groups do demonstrate a distinct preference for a specific binding site, with a few of these exhibiting a weakly conserved distance constraint with respect to the translational start site (Table 5.5). For example, the most stringent scan for the YNL064C orthologue set (Figure 5.8) reveals two binding sites that appear to have distance preferences in the URSs: the ECB site appears within 200 bases of

translational start, and the PHO4 sites are found mainly between 200 → 500 bases from the translational start site.

Table 5.5. Orthologue groups exhibiting binding site composition preferences.

Groups are identified by the systematic ORF name of the *cerevisiae* protein with which all members share homology.

Orthologue Group	Observations
YNL064C	Most sequences contain an ECB site within 200 bases of translational start and a PHO4 site between 200 → 500 bases from the translational start site
YJL115W	Most sequences contain an MCB site within 300 bases from the translational start site
YHR181W	Majority of sequences contain an ECB site between 200 → 500 bases from translational start
YPR080W	High number of ACE2 and UASPHR sites. UASPHR sites all lie within 240 → 500 bases of translational start. ACE2 sites all occur over 300 bases from translational start.
YJR045C	Numerous ABF1 sites all within 450 bases of translational start.
YJR009C	ABF1 sites present in 4 (of 6) sequences between 350 → 550 bases from translational start
YDR155C	Prevalence of ECB and STRE sties.
YBL075C	Slight preference for ABF1 sites (3 out of 6 sequences).
YDL055C	ACE2 site present in all sequences
YGL148W	Preference for PHO4 (4 out of 6 sequences)

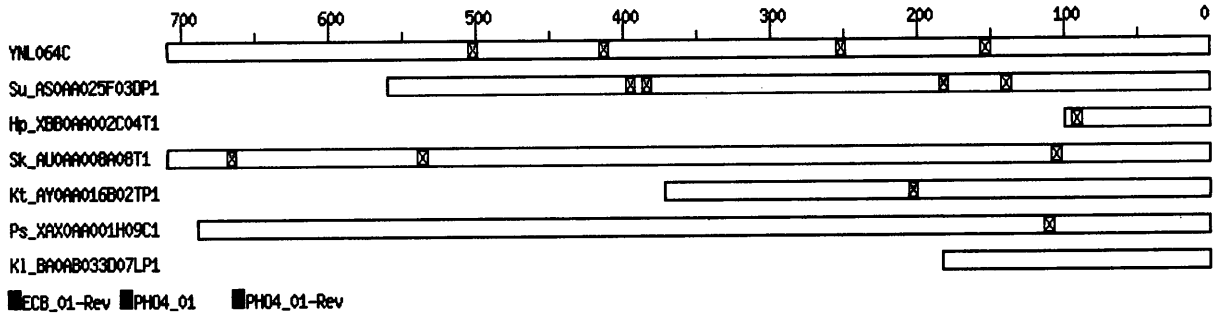


Figure 5.8. SiteSeer visualisation of the YNL064C orthologue set.

This image was produced using the following parameters: Expectation ratio = 4, minimum number of occurrences = 6, probability of occurrence = 0.3. The first two characters of a Génolevures sequence identifier represent its species code (see Table 5.1).

Finally, thirteen groups contain sequences exhibiting stronger conservation in binding site sequences and in the positioning of those sites (Table 5.6). In many of these groups, however, only two sequences are found to share common sites. For example, in the YLR175W sequence set only the *Saccharomyces cerevisiae* URS and the

Saccharomyces bayanus var. *uvarum* sequences show any similarities in binding site composition and order (Figure 5.9). The following groups show conservation between three or more sequences: YBR048C-A, YGR020C, YHR112C, YLR009W and YPR074C (Figures A5.3.1 to A5.3.5, Appendix 5).

Table 5.6. Orthologue groups exhibiting conservation of binding site positions.

Similarities between sequences are defined using the species code (defined in Table 5.1) for the sequences and arrows joining those sequences with notable conservation in mapped site positioning. The code Sc represents the yeast *Saccharomyces cerevisiae*.

Group	Species showing conservation
YBR084C-A	Sk → Kl and Zr → Ss
YDR155C	Su → Sc (also note the abundance of ECB and STRE sites)
YGL037C	Sc → Su
YGR020C	Sc → Su and Hp → Su
YHR112C	Sk → Zr and Su → Sc
YIL069C	Kl → Ps
YKL003C	Sc → Su (mainly low probability sites)
YKL199C	Ss → Su (weak similarity)
YLL039C	Hp → Yl shows strong conservation. Possible similarities between Km → Yl → Hp → Kt
YLR009W	Zr → Ps → Se
YLR175W	Sc → Su
YOR128C	Sk → Sc
YPR074C	Sc → Su

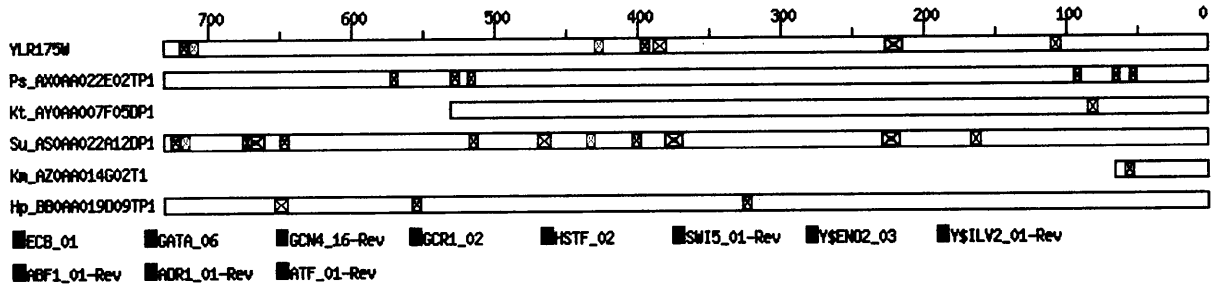


Figure 5.9. SiteSeer visualisation of the YLR175W orthologue set.

This image was produced using the following parameters: Expectation ratio = 4, minimum number of occurrences = 3, probability of occurrence = 0.5. The first two characters of a Génolevures sequence identifier represent its species code (see Table 5.1).

5.3.3 Conclusions

The majority of the URSs in the orthologous groups exhibit no apparent similarities in binding site preferences or in the order of their sites. This observation is based on the distributions of mapped sites currently detailed in public promoter databases, and it could be argued that this dataset is not sufficiently representative for this analysis. It is also possible that transcription factor binding sites are under quite different mutational pressures than coding regions. Evolution of these sites and their associated binding proteins can occur without loss of equivalent regulatory control or function. However, the available resources do contain a large number of mapped *cerevisiae* sites and it might be expected that some motifs would be conserved in a relatively close set of yeast species.

The visualisation techniques applied here are clearly useful and highlight several conserved motifs, although some sequence groups show little or no conservation. It is also possible that there is a weak conservation of over-represented motifs not clearly visible by eye. Later in this study, we employ a random background model to determine whether there is a higher conservation of sites than expected by chance.

It is worth noting that the small proportion of sequences that contain conserved sites usually originate from closely related species. For example, *Saccharomyces cerevisiae* and *Saccharomyces bayanus* var. *uvarum* are the two most commonly associated species and are the most closely related species in the dataset. It could be argued that there has been insufficient time for the divergence of these URSs and that the commonalities are due to a lack of evolutionary distance rather than the presence of biologically important sites. There has been relatively little work on the conservation of regulatory regions across species to date, and it is difficult to draw more general conclusions in comparison to other studies. However, some evidence is available concerning the close conservation of the regulatory regions of sea urchins (Prof. H. Boulouri, personal communication), which supports these conclusions. In addition, a study into somitogenesis related genes revealed sequence conservation between a diverse set of species: *Danio rerio*, *Fugu rubripes*, *Tetraodon nigroviridis*, *Xenopus laevis* and *Homo sapiens* (Gajewski 2002). The URSs for the c-hairy 1/2 class genes *her6* and *her9* from *Danio, Fugu* and *Tetraodon* were aligned with the

promoters of *x-hairy2a* and human *hes*1. The conserved regulatory sequences found consist of CCAAT boxes, TATA boxes and an SPS motif (Figure A5.3.6, Appendix 5).

5. 4 Data-mining for conserved motifs

It is apparent that the majority of orthologous URSs do not contain conserved mapped binding sites. It is still possible that the orthologue groups contain conserved sites that are not currently represented in the databases. Many different, published, techniques attempt to detect over represented words in a set of related sequences. Initially we attempted to mine for conserved words in the Génolevures orthologue groups using the Improbizer program (developed by Dr Jim Kent, web ref 41). We then implemented the over representation statistic described by Hampson *et al.* (2002) in an attempt to define over-represented words from the entire set of URSs in the *Saccharomyces cerevisiae* genome. Over-represented sites represent potentially biologically important sequences. This approach was also augmented with a positional conservation statistic to find words that were not only over-represented, but were also positionally biased with respect to the translational start of the URS in which they are found. Finally, we attempt to identify those lexicons that are over-represented in sets of functionally grouped genes (e.g. microarray clusters, MIPS categories, KEGG categories, Orthologues from the Génolevures project) with the aim of identifying one or more lexicons that contain an abundance of biologically significant sequences.

5.4.1 Improbizer analysis

5.4.1.1 Improbizer

Improbizer uses a variation of the expectation maximisation algorithm to search for motifs (in nucleotide sequences) that occur more often than expected by chance. The background model used by Improbizer is more robust and accurate than those of other available motif finding programs (Dr. William Noble, personal communication). The background model is calculated from the input sequences. This background model can be either a simple probability matrix or a higher order Markov model (up to

second order). Two further parameters have a significant effect on the results are the specification of the number of sites to report and the number of expected instances of each site per sequence.

A down side to using Improbizer is that the output log odds scores generated for each motif are unintuitive. It is not possible to compare scores between motif searches on different sets of sequences. Instead, it is necessary to generate randomised versions of the same input sequences and re-search with these. The significance of a motif in the real dataset is estimated by comparing its log odds score with those generated in the random analysis.

5.4.1.2 Methods

A set of motif searches were run on the orthologue groups that contain sequences from four or more yeast species (Table 5.7). The generation of these groups was described in Section 5.1.

Table 5.7. Improbizer motif searches.

Number of motifs to report	Number of instances expected per sequence	Background model (Markov order)
6	1	0
6	1	1
6	2	0
6	2	1

The, ***run_ameme***, program was created as a wrapper to Improbizer. This program creates a set of randomised sequences from the input file(s) and runs Improbizer on the original file and on the randomised files (a default of 3 random files per input file are produced). Motifs are separated from the HTML output and placed in their own directory. A second program, ***extract_high_scoring_motifs.pl***, is used to analyse the results and extract those motifs that have a higher log odds score than expected by random (as defined by the random Improbizer runs).

5.4.1.3 Results

Figures 5.10 and 5.11 show two sets of contrasting results from the Improbizer scans of orthologues from the Génolevures data. These examples represent two extremes. It is clear that three of the four sequences in the orthologue group YDL020C contain strongly conserved motifs (Figure 5.10). These motifs appear in all four search conditions. In contrast, the YDL100C group contains no obvious conserved motifs that are present in all four search conditions. The colour allocated, by Improbizer, to a motif reflects only the order in which the motifs were found, which can confuse initial comparisons between different search conditions. The intensity of the colour is relational to the score of the motif.

It is also worth noting that the motifs detected by Improbizer are heavily dependent on the search parameters. Using a first order Markov model (b) and d) in the figures) to represent background probabilities results in smaller motifs than a search using a zero order Markov model (a) and c) in the figures). Smaller motifs are also discovered when two sites (c) and d) in the figures) are expected per sequence in comparison to a single site (a) and b) in the figures).

In total, 3738 motifs were found to have higher scores than expected by random. The consensus sequences from 584 of these motifs find one or more match within the database of all URSs from *Saccharomyces cerevisiae*. Later on in this study we employ a random background model to determine whether this matching set of words is found more often than expected by chance in the URSs of groups of functionally related genes.

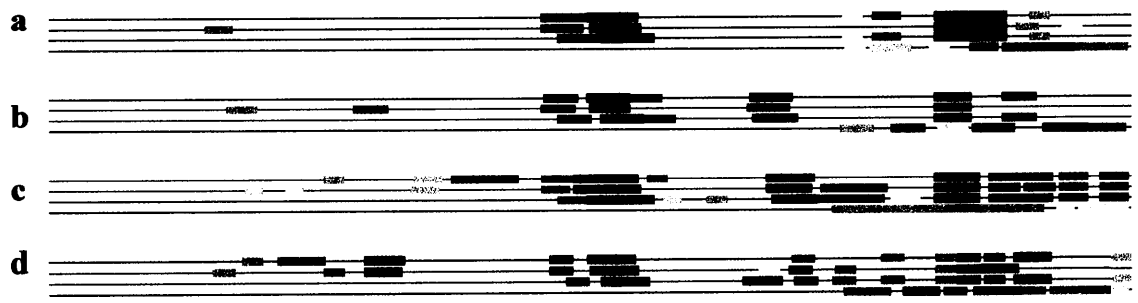


Figure 5.10. Improbizer results for the URS of YDL020C and its orthologues.

Each black line corresponds an URS. Coloured boxes represent matches to a specific motif (the colours are not consistent between the different images). The species depicted here are *Saccharomyces cerevisiae, Saccharomyces bayanus* var. *uvarum, Saccharomyces paradoxus* and *Zygosaccharomyces rouxii* respectively. The four images depict the results from searches using different parameters (see Table 5.7) **a)** one instance per sequence and a zero order Markov model, **b)** one instance per sequence and a first order Markov model, **c)** two instances per sequence and a zero order Markov model, **d)** two instances per sequence and a first order Markov model.

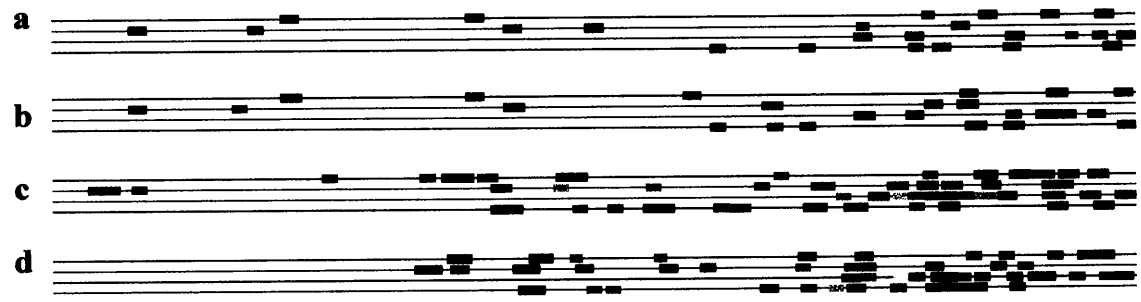


Figure 5.11. Improbizer results for the URS of YDL100C and its orthologues.

Each black line corresponds an URS. Coloured boxes represent matches to a specific motif (the colours are not consistent between the different images). The species depicted here are *Saccharomyces cerevisiae, Hansenula polymorpha, Saccharomyces kluyveri* and *Saccharomyces exiguus* respectively. The four images depict the results from searches using different parameters (see Table 5.7) **a)** one instance per sequence and a zero order Markov model, **b)** one instance per sequence and a first order Markov model, **c)** two instances per sequence and a zero order Markov model, **d)** two instances per sequence and a first order Markov model.

5.4.2 Over-representation analysis

Analyses have shown that over-represented sequences found in co-regulated families of genes are often associated with the DNA binding sites of transcription factors (van Helden 1998, Brazma 1998). Hampson *et al.* (2002) devised an effective statistic where over-representation is measured by comparing the frequency of sequences (or *k*-mers) in co-regulated genes to their frequency over all genes. We implemented the algorithm described by Hampson *et al.* and used this to extract over-represented *k*-mers in the URSs of *Saccharomyces cerevisiae*.

5.4.2.1 Algorithm

This algorithm is for the analysis of exact *k*-mers (words with a length *k*) only. The number of occurrences, *C*0, and positions of each *k*-mer can be collected in a single pass through a set of sequences. Only non-overlapping occurrences of each word are considered. This means that multiple instances of the same word found in a single sequence are only considered if they lie outside of other occurrences of the same word. This reduces the effect of low-complexity regions such as AT repeats.

Two background models were used for normalisation in this analysis. Firstly, the expectation probability of a *k*-mer is calculated using a zero order Markov model with a constant background probability of A = T = 0.31 and G = C = 0.19. So, the sequence TTACCCG has an expectation probability of $0.31^3 * 0.19^4 = 3.9E^{-5}$. This is then used to calculate the expected number of occurrences of the *k*-mer in the dataset. In our case, we have a dataset of 6217 URSs of 800 bases in length (there are a small number of shorter sequences in the dataset). The expected number of occurrences is given by the expectation probability multiplied by the total number of possible occurrences of the site (800 * 6217 – length of the *k*-mer + 1). For our previous example the expected number of occurrences would be $3.9E^{-5} * (800 * 6217 - 6) = 193$. The second background model estimates the counts of each *k*-mer as the average of *C*1, the summed count of all *k*-mers that differ from the subject in a single position.

The program, ***hampson***, calculates the following for each *k*-mer: -

1. Number of occurrences on the forward strand
2. Number of occurrences on the reverse strand
3. Zero order Markov chain probability (or expectation probability) of the *k*-mer
4. Expected number of occurrences in the dataset
5. Positional preference in the sequences

The positional preference was introduced to allow the detection of sites that have a non-random distribution in the dataset. This is calculated by sectioning the URSs into bins of 100 bases and counting the total number of occurrences of the *k*-mer in each bin. The difference in frequency of occurrence per bin in comparison to that expected by chance is then calculated, where random is defined as the total number of occurrences divided by the number of bins. The average of these differences is then used as a final "skewed" value for the *k*-mer. This "skewed" value represents the non-uniformity in the distribution. If the sites are evenly spread throughout the URS sequences, then this value will be small. If they are biased to one end, or even several positions, in the URS sequences, then the "skewed" value will be large.

We also calculate a set of normalised figures for the forward, reverse and aggregated counts of each *k*-mer. These are normalised using both the expected number of occurrences and the average of the *C*1 counts (i.e. the two separate background models).

5.4.2.2 Results

Calculations were carried out using all 7 mers, 8 mers and 9 mers. A number of lexicons were produced from each set of *k*-mers by selecting the top 50 sequences ranked by the values described in section 5.4.2.1 (Table 5.8).

Table 5.8. Ranking statistics used for lexicon production.

Normalisation values: 1 = expected number of occurrences, 2 = *C*1 counts.

Ranking Statistic	Occurrence Measure	Non-random spatial distribution	Normalisation
C0f	Forward strand only	No	1
C0a	Forward and reverse strands	No	1
C1f	Forward strand only	No	2
C1a	Forward and reverse strands	No	2
C0f_skewed	Forward strand only	Yes	1
C0a_skewed	Forward and reverse strands	Yes	1
C1f_skewed	Forward strand only	Yes	2
C1a_skewed	Forward and reverse strands	Yes	2

5.4.3 Lexicon evaluation

We have now defined a number of lexicons that may describe biologically functional sequences; SCPD mapped sites, TRANSFAC mapped sites, sequences detected by the Hampson methodology, and sequences defined by the Improbizer analysis (Section 5.4). The next logical step is to identify those lexicons that contain functionally important sites that may be used to group together genes with common function.

For this evaluation, we used the following to define groups of genes with common functions:

1. Gene clusters from Eisen *et al.* (1998)
2. Microarray data from Gasch *et al.* (2000) clustered using a self-organising map.
3. MIPS categories (top level only).
4. KEGG categories (top level and second level).
5. Orthologous genes from the Génolevures data.

Each of these groups contains a different number of sequences (Table 5.9). A single gene may be assigned to multiple functional categories within the MIPS or KEGG

classification systems, whereas the data from the Eisen and Gasch analyses do not encompass the entire genome.

Table 5.9. Total number of sequences in each functional grouping.

Functional Group	Number of Sequences
Eisen	131
KEGG (1^{st} level)	1365
Gasch	1943
KEGG (2^{nd} level)	1954
Génolevures orthologues	2036
MIPS (1^{st} level)	10274

The lexicons also contain differing numbers of sequences (Table 5.10). It is possible that many of these lexicons contain similar sequences. We performed a pairwise comparison of the lexicons, noting when a sequence in one lexicon was contained within another (Figure 5.12). Sequences that were contained as a subsequence of an entry in another lexicon were also noted. These comparisons are non-symmetrical, this means that a comparison of lexicon *a* versus lexicon *b* does not yield the same result as a comparison of lexicon *b* versus *a*. For example, a comparison of the TRANSFAC mapped sites versus the SCPD mapped sites shows that 91% of TRANSFAC sites are represented within the SCPD database but only 46% of the SCPD mapped sites are represented within the TRANSFAC database. The difference in size of these databases is a large contributing factor for this apparent anomaly (Table 5.10).

Table 5.10. Details of lexicons used in analysis.

Lexicon	Number of sites	Average site length (nt)	Percentage AT content
Combined sites	809	15.2	55
Improbizer sites	548	12.2	62
SCPD sites	542	14	55
TRANSFAC sites	267	17.6	56
7mer C0a	50	7	71
7mer C0a skewed 3	50	7	14
7mer C0f	48	7	69
7mer C0f skewed 3	50	7	14
7mer C1a	50	7	74
7mer C1a skewed 3	50	7	17
7mer C1f	50	7	75
7mer C1f skewed 3	50	7	16
7mer e	48	7	0
7mer e skewed 3	50	7	16
8mer C0a	50	8	64
8mer C0a skewed 3	50	8	19
8mer C0f	48	8	64
8mer C0f skewed 3	50	8	20
8mer C1a	50	8	55
8mer C1a skewed 3	50	8	31
8mer C1f	50	8	53
8mer C1f skewed 3	50	8	32
8mer e	48	8	100
8mer e skewed 3	50	8	39
9mer C0a	50	9	61
9mer C0a skewed 3	50	9	27
9mer C0f	48	9	56
9mer C0f skewed 3	50	9	24
9mer C1a	50	9	40
9mer C1a skewed 3	50	9	36
9mer C1f	50	9	36
9mer C1f skewed 3	50	9	32
9mer e	48	9	78
9mer e skewed 3	50	9	78

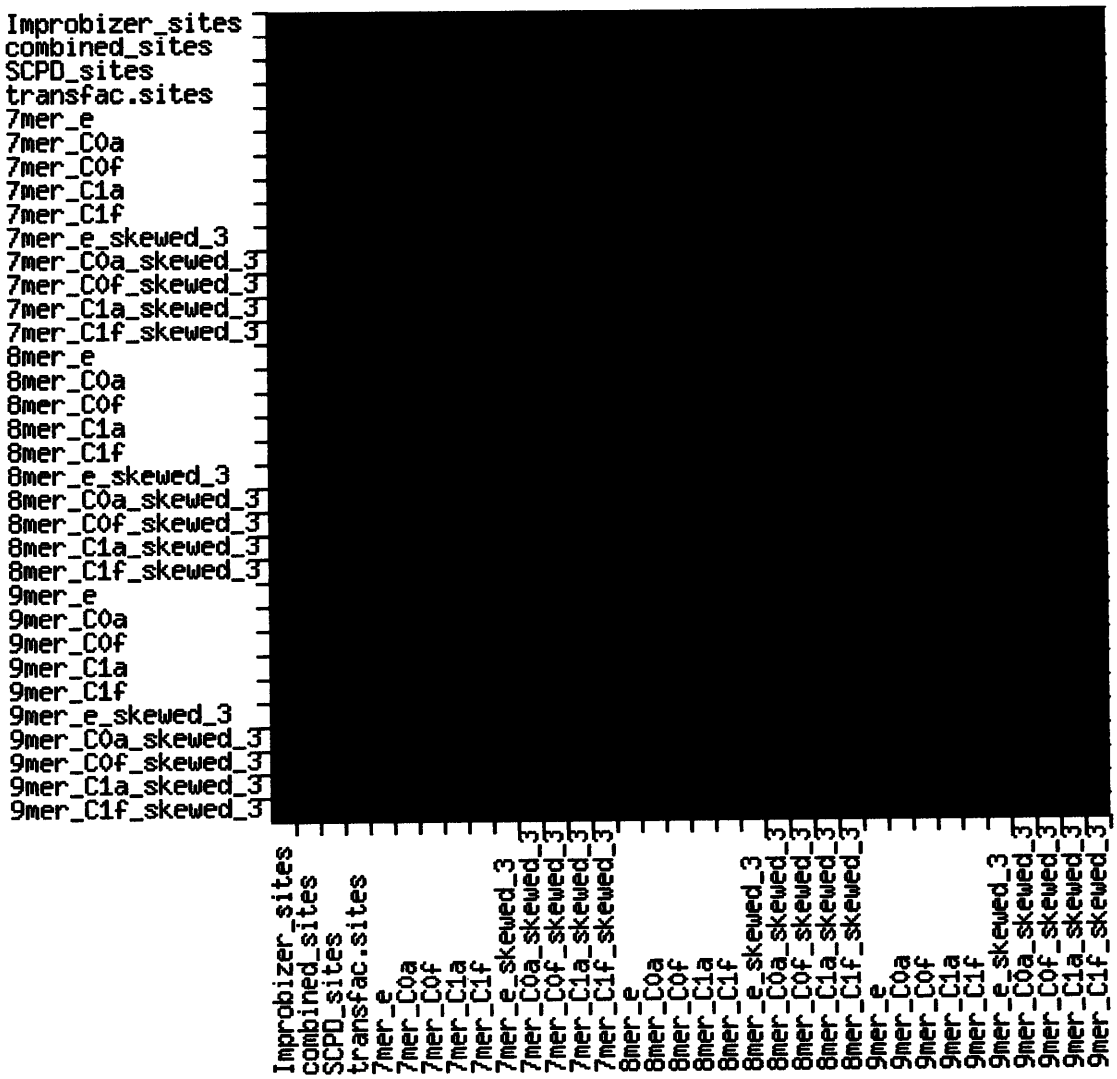


Figure 5.12. Pairwise lexicon comparisons.

Each square in the grid represents a single pairwise comparison of two lexicons. The comparison is carried out in both directions, which results in a non-symmetrical grid. Different levels of the colour red are used to represent different degrees of lexicon similarity. Black = no similarity, bright red = very similar.

5.4.3.1. Basic algorithm

The efficacy of a lexicon is evaluated by searching each group within a functional grouping of URSs (e.g., MIPS categories) with each word from the lexicon. The score for the lexicon is increased each time two sequences, within a clustered set of genes, are found to have a word in common.

The program, ***lexicon_distributions.pl***, was created to carry out these evaluations and subsequent normalisation steps.

Figure 5.13 shows the initial results of this analysis. The graph is ordered by score for the lexicons of the MIPS groupings (the other functional groupings are seen to follow a similar trend). It is unsurprising that the highest scores are seen for those analyses that compare the functional groupings containing the largest number of sequences using lexicons either with the simplest sequences or with the greatest number of sequences. The highest scoring search, for example, is an analysis of the MIPS functional groupings (which contains the largest numbers of sequences to be compared) with the 7mer_C1a lexicon (which has a high AT content). An inspection of the 7mer_C1a lexicon, revealed a high proportion of low complexity words such as AT repeats and words made up almost purely of A's or T's. Due to the nature of the yeast genome, these words have a high probability of occurrence and are unlikely to enable the discrimination of functional groups.

To counter this, a number of different normalisations were introduced in order to identify those lexicons containing the most informative sites.

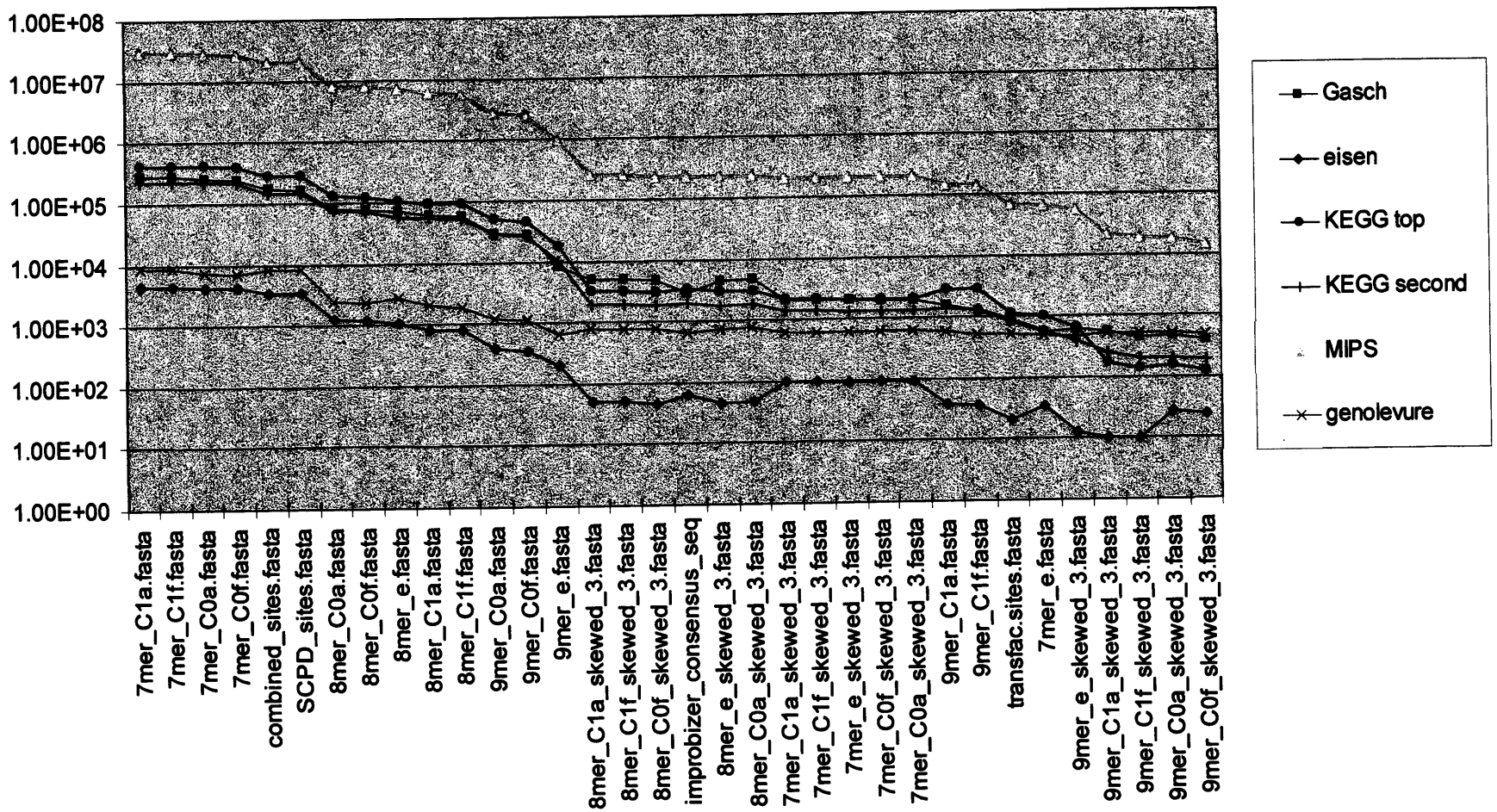


Figure 5.13. Total scores for each lexicon within each functional grouping.

The total score for each lexicon (defined in Section 5.4.3.1) is plotted on the y-axis. The Hampson lexicons are defined in Table 5.8. The 'combined_sites' lexicon was produced by combining the TRANSFAC and SCPD mapped sites.

5.4.3.2 Lexicon size

As the number of sequences in a lexicon increases the probability of finding positive matches from this lexicon also increases. To counter this we produced a set of scores that were normalised by lexicon size (Figure A5.2.1, Appendix 5). This results in a marked reduction of the scores for the larger lexicons but does not provide any discrimination between the Hampson derived lexicons as these all contain 50 sequences.

5.4.3.3 Randomised URSs

A more rigorous approach to normalisation is to compare a result to that which you would expect to find by chance. In this normalisation, random is defined as the average score from searches performed on randomised URSs. To produce randomised URSs, the input sequences are subjected to a Fisher-Yates shuffle. Input sequences were randomised and searched ten times. The average score from these searches were used for normalisation.

Figure A5.2.2 (Appendix 5) shows the results using this normalisation. Unlike the results from the previous two scoring systems, the scores between the different functional groupings do not exhibit the same patterns. This is most prominent when comparing the Génolevures results with the other functional groupings (but is apparent when comparing all functional groupings). The highest scoring lexicons for the Génolevures data are 7mer_C0f, 7mer_C0a, 7mer_C1a and 7mer_C1f. Whereas the most consistent high scoring lexicons for all the other functional groups are 9mer_C0f, 9mer_C0a, 9mer_e and 8mer_C0f. This difference may be due to higher sequence similarity within the Génolevures orthologue groups, which could potentially cause a bias towards lexicons containing small sequences with low complexity.

It is interesting to note that the mapped sites from both SCPD and TRANSFAC have consistently low scores for all functional groupings in comparison to the other lexicons, as do the Improbizer consensus sequences. It is not surprising that the highest ranking obtained by the Improbizer consensus sequences are for the Génolevures orthologous groups since these groups were used to generate the Improbizer lexicon.

Importantly, all the lexicons have an overall positive score for all functional groupings. This score is an average of that calculated for all clusters within a functional grouping; it may be the case that one or two high scoring clusters are biasing the results. Another possibility is that the majority of positive scores seen are for low complexity sites and comparisons of URSs with regions of AT repeats or poly A/T tracts. The randomisation of these sequences would remove the repeat regions, which in turn would accentuate this bias. If this is true, the bias should be most obvious for those lexicons containing words that consist solely of AT repeats or poly A/T. Indeed, a brief analysis of the lexicons shows that those mentioned previously as the highest scoring for this analysis all contain AT repeat sequences and sequences consisting solely of either poly A or poly T whereas the lower scoring lexicons do not contain these sequences. This bias is addressed to some extent with the next normalisation procedure.

5.4.3.4 Randomly picked URSs

Another method for randomisation is to pick out random selections of real URSs to search. For each set of input sequences from a functional grouping, an equivalent number of real URSs are randomly picked. For the Génolevures analysis, a random sequence is selected from each organism represented in the original input set. For the other functional groupings, sequences are randomly selected from a set of 6,217 *Saccharomyces cerevisiae* URSs. This random selection and analysis is carried out 10 times for each cluster within the functional grouping. The average of these results is subsequently used for normalisation.

Figure A5.2.3 (Appendix 5) shows the results using this normalisation step. As with the previous randomisation method for normalisation, the different functional groupings show different lexicon preferences. As before 7mer_C1a and 7mer_C1f score highly for the Génolevures groups, but we also see the SCPD mapped sites scoring highly. This may be because the SCPD sites contain between 48 and 56% of the, non-skewed, 7mer lexicons. The other functional groupings show a preference for the skewed lexicons (predominantly 8mer_C0a_skewed, 8mer_C0f_skewed, 7mer_C1f_skewed and 7mer_C0a_skewed). The skewed lexicons are those that exhibit a spatial bias in their positional distributions, which is thought to be indicative of biologically functional sites (Hampson 2002). The scores for all functional groupings, except the Génolevures

groups, have been reduced in comparison to the previous normalisation. This implies that the bias noted in the previous normalisation has been either removed or reduced. Scores for the Génolevures orthologous groups remain largely unchanged from the previous set of results. The large scores seen here may be caused by a few groups of orthologous genes that contain very similar sequences. Unfortunately, we did not have time to address this issue.

5.4.3.5 Lexicon size and sequence number normalisation

Finally, it is important to consider the results in light of the total number of sequence comparisons carried out within each search. A normalisation procedure was implemented that accounts for this.

The total number of sequence comparisons is dependent on both the size of the lexicon and the total number of sequences in the functional group being analysed. The total number of sequence comparisons is given by:

$$\frac{n^2 - n}{2} \times L$$

Where, n, is the number of sequences in the functional group being considered and, L, is the number of words in the lexicon. If, T, is the total number of matches found, then the normalised value for a search is given by:

$$\frac{2T}{(n^2 - n)L}$$

Figure A5.2.4 (Appendix 5) shows the results using this normalisation procedure. All scores are greatly reduced with this normalisation step. In addition, the trend for lexicon scores is similar between all of the functional classification systems. The highest scoring lexicons are 7mer_C1a, 7mer_C1f, 7mer_C0a and 7mer_C0f (thought the order differs slightly from one classification to the next). As discussed before, these lexicons have a high number of low-complexity and AT rich sites. The Improbizer consensus sequences perform poorly for all functional groups, the worst being the KEGG (second level) and the Génolevures orthologue analyses where the Improbizer

set is the lowest scoring lexicon. The SCPD sites perform better than the TRANSFAC sites in all cases.

5.4.4 Conclusions

In this section, we have described a number of lexicons that may be enriched for transcription factor binding sites and other biologically significant sequences. We have investigated the information content of these lexicons and applied a number of normalisation procedures to account for biases present in the data.

A rather surprising theme from the results of the above analyses is the poor performance of the SCPD mapped sites, the TRANSFAC mapped sites and the Improbizer generated sites in comparison to the sequences generated using the Hampson methodology. This may partly be due to the proportion of large sites in these three databases. Of the 266 *Saccharomyces cerevisiae* mapped sites in TRANSFAC, 211 have a length of 10 bases or greater. The case is similar for SCPD where 394 out of 542 have a length of 10 bases or more and for the Improbizer sites where 547 out of 584 have a length of 10 bases or more. Transcription factors generally bind to an area encompassing six base pairs on the DNA duplex. The larger sites in these databases contain the core binding-site sequence for a transcription factor as well as sequence not involved in transcription factor association. This, non-specific, sequence increases the stringency of the mapped site beyond that exhibited by its associated transcription factor. This causes an increase in the number of false negative site detections and reduces the number of overall site assignments.

There are obviously biases introduced by very similar sequences present in the functional groupings. This is particularly evident for the Génolevures orthologous groups when comparing the scores for the "randomised URS" and the "randomly picked URS" normalisations. Removing sequences with high identity to one another other could improve the analyses discussed in this chapter. Unfortunately, this has only become apparent after the completion of this work. It would still appear that this is a valid methodology for the identification of lexicons enriched with binding site sequences. In particular, the "randomly picked URS" normalisation appears to favour lexicons with sites exhibiting features indicative of functional sites.

This analysis also highlights the differences between methods for grouping functionally related genes. With different groupings having different perspectives on the meaning of function (discussed in Section 1.3.1). With some groupings, we would expect to see a preference for conserved binding sites in the URSs of individual groups. However, in others, perhaps we would not. For example, the most general levels of the MIPS and KEGG categories contain large varieties of genes within each group. One would not expect all of the genes within a group to be expressed at the same time or at the same rate. Given this, it is quite probable that a search for lexicons enriched with words contained within these groupings will only discover those lexicons with the smallest and simplest words. Conversely, we also examine groups of genes clustered through virtue of correlated patterns of expression. For these genes, we might expect to find common signals, or words, in their URSs. Indeed, this appears to be the case. Both the Eisen clusters and the clusters produced from data generated by Gasch *et al.* favour the "skewed" lexicons, which contain sequences with biases in their position in the URSs of *Saccharomyces cerevisiae*.

Conclusions & Discussion

During the course of this investigation, we have detailed the creation and annotation of an EST resource containing over 330,000 ESTs, which have been assembled into 85,000, high fidelity, contig sequences. This required the design and implementation of a series of bioinformatic protocols and the development of a relational database system for data storage, querying and retrieval. Our web-site, which provides public access to these resources, has been widely used by the scientific community, with over 35,000 'hits' to the front page and over 45,000 BLAST searches carried out (from labs all over the world) within the first year and a half of operation. The initial evaluation of our *in silico* subtraction technique has shown that this has the potential to be a powerful procedure, allowing the isolation of differentially expressed transcripts from EST libraries. A facility for *in silico* subtractions for all tissues represented in our dataset is being developed for the chicken EST website (web ref 12). During the course of the transcription factor binding site analysis, the visualisation tool, SiteSeer was developed. A paper describing the function and use of this tool is currently in press (Boardman 2003) and SiteSeer itself is publicly accessible over the web (web ref 29). This tool has proved to be invaluable in the analysis of upstream regions for conserved motifs among functionally related URSs, as well as for URSs of homologous genes from different yeast species. The TF binding site analysis itself, although producing few positive results, has demonstrated the complexity of eukaryotic transcriptional control and the difficulties in discovering and exploiting patterns of protein binding signatures in the prediction of gene function. We also discovered methods for the production and identification of lexicons enriched in words over-represented in the URSs of functionally related genes.

We've seen from the analyses with chicken ESTs (Chapter 2) that traditional homology-based approaches represent quick and effective ways for the annotation of function to unknown sequences. We have also seen that these techniques fail to find reliable annotations (through the detection of similarity) for many sequences from a set purporting to cover most of the genome of a given organism. In the investigation detailed in Chapter 2, more than 60% of the consensus (or contig) sequences and around

50% of the ESTs failed to find a significant match in the public databases as judged by homology-based algorithms such as BLAST. This result appears to be quite surprising at first glance since comparable vertebrate genomes have a higher level of known, previously characterised, genes. In addition, other chicken EST resources appeared to contain a much higher fraction of sequences with homologues in the known databases. A number of possible explanations for the shortfall in the BBSRC collection are apparent. For example, many of these unannotated sequences may be composed entirely of 3' or 5' UTR, which would explain the inability to find a match in the protein databases. A cursory search of the contigs versus a database of 3' and 5' UTR sequences reveals that only 407 (~0.5%) contigs have 50% or more of their sequence matching to an entry in the eukaryotic UTR databases. However, this in itself is not entirely conclusive since chicken UTR sequences may very well differ quite markedly from previously sequenced genomes in these regions. This is probably quite likely, since UTR sequences are under quite different, less stringent, mutational pressures than coding sequences and no avian genomes are available as close comparators. It is also possible that these undefined chicken sequences represent either protein coding mRNAs with no homologue characterised at present or functional non-coding mRNAs. The latter possibility is currently being addressed through comparisons with RNA databases (e.g. Rfam) and the former through the development of open reading frame prediction software tailored for EST data. Naturally, the ongoing sequencing of the chicken genome will provide many of the final answers and is eagerly anticipated.

It is quite feasible that a large proportion of the remaining, unannotated, sequences represent protein-coding mRNAs, which currently have no homologue in the sequence databases. Indeed, some of these may well represent avian specific genes, which have been discovered as part of this project. This is in keeping with the outcome of other genome sequencing projects where the complete predicted protein set of any given organism has an average automated annotation rate of around 60-70%, even for well-characterised organisms such as *Saccharomyces cerevisiae*. This represents a large number of sequences for which annotation of function is a challenge.

We conducted a basic scan of the unannotated contigs to discover potential open reading frames. A consensus chicken Kozak sequence was defined using the available full-length sequences from the EMBL databank. Coupled with codon usage bias, this

provides an intuitive way to score potential open reading frames in the unannotated contigs. 3,983 unannotated contigs contain an uninterrupted ORF of >=100 amino acids and 214 have an uninterrupted ORF of >=200 amino acids. Despite this apparent successful determination of many putative chicken coding-sequences, it is likely that many more of the contigs contain open reading frames. There are two main reasons why this approach will miss true ORFs: Firstly, this approach requires the start of an ORF to be present, and those contigs that do not cover the start codon of the associated gene will not have a detectable Kozak sequence. Secondly, a putative ORF may be prematurely "terminated" in a given frame through errors introduced into the contig sequence during sequencing. The program, ***EORF***, is currently being developed at UMIST in order to address these issues. We hope to develop an algorithm to detect ORFs in ESTs that accounts for potential sequencing errors (such as miscalled bases and insertions/deletions) through virtue of the codon-usage bias of the underlying ORF.

We look forward to the completion of the chicken genome at the end of the year. The complete genome sequence should allow us to answer many of the above questions. With this resource we will also be able to work back through the protocols and tools we developed in order to define better, more accurate and robust ways to annotate ESTs, predict ORFs, find miRNAs and to assemble high fidelity contig sequences.

As a stark contrast, the investigations into the utility of non-homology based approaches to predict gene function have shown that there is still a lot to learn about how even one of the humblest eukaryotes regulates gene function at the transcriptional level. Our ability to exploit these regulatory sequences to predict protein function is still at an early stage.

Our initial investigations into non-homology based techniques for the prediction of gene function have shown that the mere presence of a transcription factor binding-site upstream of a gene is not sufficient to infer the function of that gene. Further studies, using distance constraints and patterns of binding sites, were, on average, unable to associate functionally related genes with a frequency over random expectation. This implies either that conservation in promoter "modules" is insufficient to be detected using this approach or that the current lexicon of binding site sequences lacks many of the functional binding-sites present in the yeast genome. Indeed, the analyses detailed

in Chapter 5 show promise in the detection of new and biologically important sites that are not present in SCPD or TRANSFAC and these sites do appear to be represented above average in URS regions of functionally related genes to some extent. This suggests a way forward to expand the experimentally verified lexicons in the current databases.

There are an increasing number of groups attempting to use microarray-based approaches to detect and define putative TF binding sites from over-represented words in clustered gene sets. Many of these approaches are now being extended in attempts to predict and model regulatory networks (Pilpel 2001, Banerjee 2002, Toyoda 2003). The results detailed in this thesis suggest that these approaches may not be sufficient to define functional sites, and that many, possibly unrelated, mechanisms may be responsible for producing common expression patterns for a set of genes. Indeed, this is supported through the observation that not all co-regulated genes in a cluster share a common binding site or motif, which, in turn, implies that they may not all be the direct targets of the same TF. In addition, the same binding site can be functional in genes separated into different clusters, which emphasises the point that TFs often function in combination with other TFs and that an individual TF can act at different times or under different physiological conditions.

Studies on mammalian promoters have shown that individual TF sites are unable to induce transcription and that the specific context of the TF site is important in the elucidation of its biological function (Rao 1997, de Martin 1999, Werner 2003). Often a specific promoter function (such as a tissue-specific silencer) will require more than two sites and the organisation of these sites in relation to each other is often much more restricted than the apparent mixture of TF sites in the whole promoter suggests (Klingenhoff 1999, Werner 2003). The results detailed in Chapters 3 and 4 are in general agreement with these observations. In an analysis by Klingenhoff *et al.* (2000) a model to describe the mammalian muscle-specific actin promoter was defined. A general actin model (generated by Frech *et* al. 1998) and this muscle-specific model were evaluated by searching against actin and non-acting promoters. The results from these searches were very promising and show a specificity of 99 – 100% and a sensitivity of 77 – 84%. Figure 6.1 shows the structure of the two actin models used in their analysis. The distance constraints between the individual binding sites in the actin

models appear to be extremely lax, which provides some insight into the failures of the "absolute position" analysis detailed in Chapter 4. In addition, some of the failures of the "relative position" analysis can be deduced from these models. The gaps between individual sites are large, which leads to a high probability of detecting other binding sites within these gaps. The signal represented by the functional sites from the model will not be detected using our "relative position" analysis if these gaps contain other (potentially non-functional) sites.

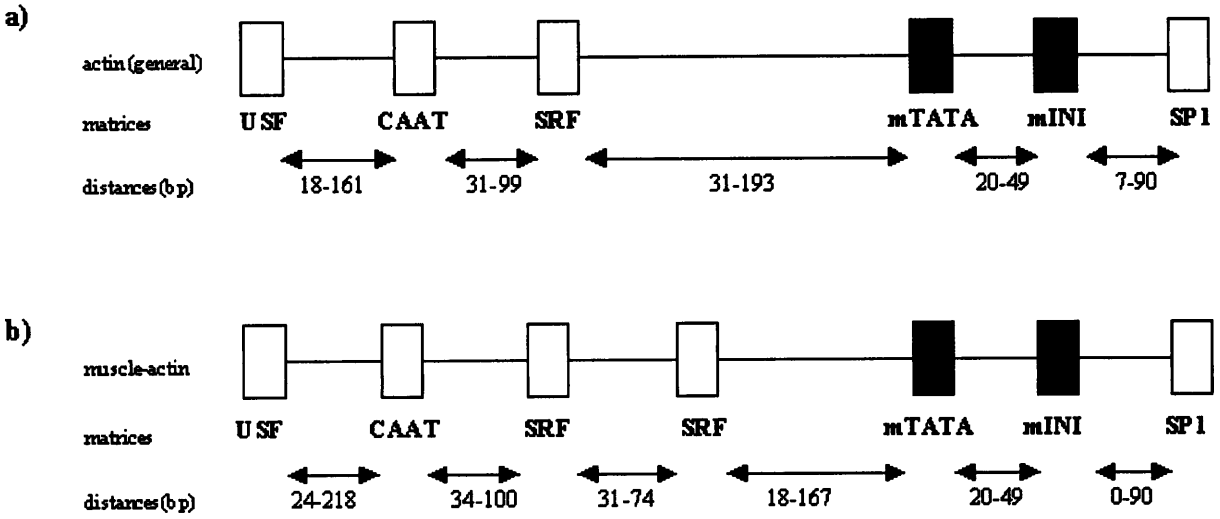


Figure 6.1. Promoter models for actin genes (taken from Klingenhoff 2000).

a) general actin model representing a regulatory module common to all mammalian actin promoters.
b) muscle-specific actin model.

We are currently working on a new method for the comparison of URSs that improves on the absolute and relative position analyses. A dynamic programming algorithm is employed to compare URS representations in a similar way that the BLAST algorithm compares two sequences. The difference lies in the fact that the representation of each URS is created using a lexicon of TF binding sites. Currently, each site is allocated a hex value and the final representation of the URS is a string of these values in the order with which the corresponding binding sites are found (Figure 6.3). Initial results using this approach with lexicons defined in Chapter 5 are promising. Figure 6.2 shows some preliminary results comparing clusters taken from Eisen *et* al. (1998) and using a lexicon of the 100 top scoring Hampson 6-mers (see Chapter 5). The distribution of scores for clusters D, E and H appear to have a more positive distribution when compared to the distribution of scores for comparisons of all URSs. Further analyses on URSs containing experimentally defined promoter modules (such as the actin "promoter

modules" described in Klingenhoff *et al.* (2000)) should help to determine the efficacy of this approach.

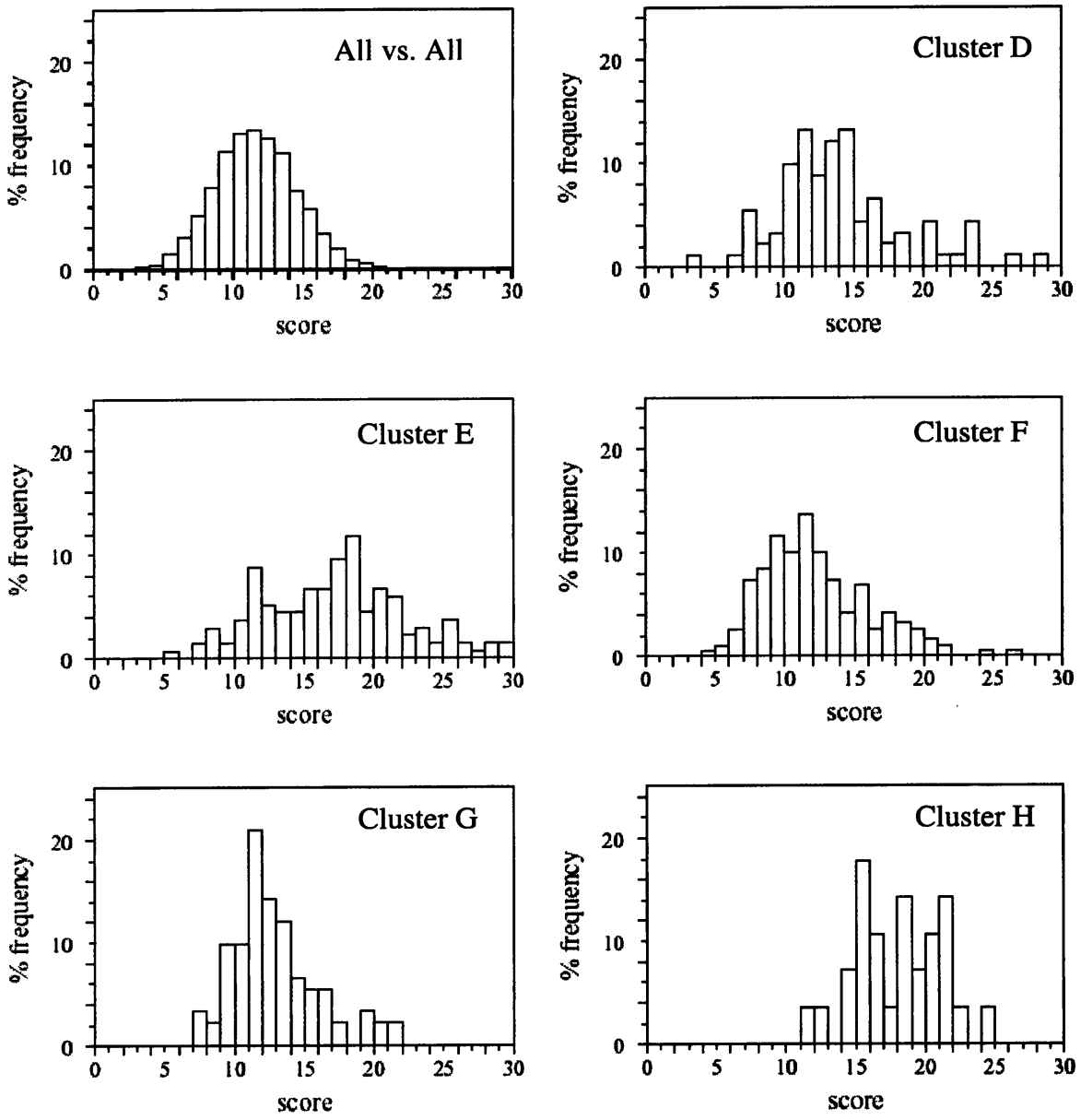


Figure 6.2. Preliminary results for URSscan.

These were performed using the 800 bp upstream regions for genes clustered together in the analysis by Eisen *et al.* (1998) and a lexicon of the top 100 Hampson 6-mers. The first graph "all vs. all" shows the score distribution for an analysis comparing all URSs in the yeast genome. This provides a random or background model with which to compare intra-cluster scores.

```
YEL054C-URS    YER137C-URS    0001 0002   11   22.45   0.24    1.38    2.890   45  49  49

Alignment type: local (type=1)
Matrix: identity (type=3)

Query: 1 17-------------58-64-39-45-44-42-48-04-46-3F-2C-03-37-29-36-61-20----58-5D-55-42-5A-0B-3B- 25
Ident: 1 17-                          44-42-                              61-20-         55-          30
Subjt: 6 17-2F-1E-31-02-0E-52-02-0E-44-42----------------------------61-20-13-4E-5B-55----50-4D-0F- 25

Query: 26 52-2A-4B-1F-0A-10-4B-14-13----------------08- 35
Ident: 31                 10-4B-14-13-               08- 45
Subjt: 26 09-16-49-3A-57-10-4B-14-13-38-4D-07-0B-3B-08- 40


YEL054C-URS    YOR166C-URS    0001 0003   15   25.86   0.15    1.76    3.257  100  87  87

Alignment type: local (type=1)
Matrix: identity (type=3)

Query:    8 48-04-46-3F-2C-03-37-29-36-61-20----------------58-5D-55---------------------------------- 21
Ident:    1 48-04-46-                   61-20-                   55-                                   30
Subjt:    2 48-04-46-------------------61-20-3F-43-0D-01-05-0F-26-55-4A-13-12-0D-05-07-5C-11-01-30-13- 25

Query:   22 42-5A-0B-3B-52-2A-4B-1F-0A-10-4B-14-13-08-06-02-0E-14-36-1E-12-13-5A---------------------- 44
Ident:   31                          0A-10-4B-                                 13-5A-                   60
Subjt:   26 12-0D-01-05-0F-09-2E-17-0A-10-4B-------------------------52-56-13-5A-1D-34-31-10-33-1F-2E- 47

Query:   45 ----------------------------------------0B-53-23-16----------------------------5D-41-59- 51
Ident:   61                                          23-16-                                          90
Subjt:   48 10-50-60-3E-5F-2C-03-04-46-3F-26-09-2C-03-04-34-23-16-24-39-24-62-04-46-36-60-06-29-37-5C- 77

Query:   52 09-16-18-5B-2F----------5D-37- 58
Ident:   91             2F-           37- 100
Subjt:   78 11-35-46-1D-2F-12-55-2C-03-37- 87
```

Figure 6.3. Example output from URSscan.

Each binding site is represented by a two digit hexadecimal number in the alignment.

References

Abdrakhmanov I, Lodygin D, Geroth P, Arakawa H, Law A, Plachy J, Korn B and Buerstedde JM. **A large database of chicken bursal ESTs as a resource for the analysis of vertebrate gene function.** *Genome Res.* 2000. 10:2062-2069.

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WE and Venter JC. **Complementary DNA sequencing: expressed sequence tags and human genome project.** *Science* 1991. 252:1651-1656.

Adams MD *et al.* (54 authors). **Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence.** *Nature* 1995. 377(Suppl.):3-174.

Aerts S, Thijs G, Coessens B, Staes M, Moreau Y and De Moor B. **Toucan: deciphering the *cis*-regulatory logic of coregulated genes.** *Nucleic Acids Res.* 2003. 31:1753-1764.

Akhtar A, Zink D and Becker PB. **Chromodomains are protein-RNA interaction modules.** *Nature* 2000. 407:405-409.

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ. **Basic local alignment search tool.** *J. Mol. Biol.* 1990. 215:403-410.

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W and Lipman DJ. **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nuc. Acid Res.* 1997. 25;3389-3402.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A *et al.* **Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*.** *Science* 2002. 297:1301-1310.

Apweiler R, Kersey P, Junker V and Bairoch A. **Technical comment to "Database verification studies of SWISS-PROT and GenBank" by Karp et al.** *Bioinformatics* 2001a. 17:533-534.

Apweiler R *et al.* (28 authors). **The InterPro database, an integrated documentation resource for protein families, domains and functional sites.** *Nuc. Acides Res.* 2001b. 29:37-40.

Arnold J and Hilton N. **Revelations from a bread mould.** *Nature* 2003. 422:821-822.

Ashburner M *et al.* (21 authors). **Gene ontology: tool for the unification of biology.** *Nature Genet.* 2000. 25:25-29.

Ashrafi K, Chang FY, Watts JL, Fraser AG, Kamath RS, Ahringer J & Ruvkun G. **Genome-wide RNAi analysis of *Caenorhabditis elegans* fat regulatory genes.** *Nature* 2003. 421:268-272.

Attwood TK and Miller CJ. **Which craft is best in bioinformatics?** *Computers & Chemistry* 2001. 25:329-339.

Attwood TK, Croning MD, Flower DR, Lewis AP, Mabey JE, Scordis P, Selley JN and Wright W. **PRINTS-S: the database formerly known as PRINTS.** *Nucleic Acids Res.* 2000. 28:225-227.

Audic S and Claverie JM. **Visualizing the competitive recognition of TATA-boxes in vertebrate promoters.** *Trends Genet.* 1998. 14:10-11.

Banerjee N and Zhang MQ. **Functional genomics as applied to mapping transcription regulatory networks.** *Curr. Opin. Microbiol.* 2002. 5:313-317.

Bartel PL, Roecklein JA, SenGupta D & Fields S. **A protein linkage map of *Escherichia coli* bacteriophage T7.** *Nature Genetics* 1996. 12:72-77.

Bateman A, Birney E, Durbin R, Eddy SR, Howe KL and Sonnhammer EL. **The Pfam protein families database.** *Nucleic Acids Res.* 2000. 28:263-266.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. **The Pfam protein families database.** *Nucleic Acids Res.* 2002. 30:276-280.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA and Wheeler DJ. **GenBank.** *Nuc. Acides. Res.* 2003. 31:23-27.

Berger JM, Gamblin SJ, Harrison SC, Wang JC. **Structure and mechanism of DNA topoisomerase II.** *Nature* 1996. 379:225-232.

Boardman PE, Sanz-Ezquerro J, Overton IM, Burt DW, Bosch E, Fong WT, Tickle C, Brown WRA, Wilson SA and Hubbard SJ. **A Comprehensive Collection of Chicken cDNAs.** *Current Biology* 2002. 12:1965-1969.

Boardman PE, Oliver SG, Hubbard SJ. **SiteSeer: visualisation and analysis of transcription factor binding sites in nucleotide sequences.** *Nucleic Acids Res.* 2003. 31, In press.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S and Schneider M. **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res.* 2003. 31:365-370.

Boguski MS, Lowe TM and Tolstoshev CM. **DbEST – database for "expressed sequence tags".** *Nature Genetics* 1993. 4:332-333.

Boguski MS, Tolstoshev CM, Bassett DE. **Gene discovery in dbEST.** *Science* 1994. 265:1993-1994.

Boguski MS & Schuler GD. **ESTablishing a Human Transcript Map.** *Nature Genetics* 1995. 10:369-371.

Bonaldo MF, Lennon G, Soares MB. **Normalization and subtraction: two approaches to facilitate gene discovery.** *Genomre Res.* 1996. 6:791-806.

Boube M, Faucher C, Joulia L, Cribbs DL and Bourbon HM. **Drosophila homologs of transcriptional mediator complex subunits are required for adult cell and segment identity specification.** *Genes Dev.* 2000. 14:2906-2917.

Boutanaev AM, Kalmykova AI, Shevelyov YY & Nurminsky DI. **Large clusters of co-expressed genes in the *Drosophila* genome.** *Nature* 2002. 420:666-669.

Brazma A, Jonassen I, Vilo J and Ukkonen E. **Predicting gene regulatory elements in silico on a genomic scale.** *Genome Res.* 1998. 8:1202-1215.

Brenner SE. **Errors in genome annotation .** *Trends in Genetics* 1999. 15:132-133.

Brent R and Ptashne M. **A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor.** *Cell* 1985. 43:729-736.

Britten RJ, Graham DE, Neufeld BR. **Analysis of repeating DNA sequences by reassociation.** *Mothods Enzymol.* 1974. 29:363-418.

Brown AJ *et al.* (21 authors). **Transcript analysis of 1003 novel yeast genes using high-throughput northern analysis.** *EMBO Journal.* 2001. 20:3177-3186.

Brown WRA, Hubbard SJ, Tickle C and Wilson SA. **The chicken as a model for large-scale analysis of vertebrate gene function.** *Nat.Rev. Genet.* 2003. 4:87-98.

Bucher P. **Regulatory elements and expression profiles.** *Curr. OPin. Struc. Biol.* 1999. 9:400-407.

Buerstedde JM and Takeda S. **Increased Ration of Targeted to Random Integration after Transfection of Chicken Cell Lines.** *Cell* 1991. 67:179-188.

Burge C and Karlin S. **Prediction of complete gene structures in human genomic DNA.** *J. Mol. Biol.* 1997. 268:78-94.

Bussemaker HJ, Li H & Siggia ED. **Regulatory element detection using correlation with expression.** *Nature Genet.* 2001. 27:167-171

Christiansen JH, Coles EG, Robinson V, Pasini A, Wilkinson DG. **Screening from a subtracted embryonic chick hindbrain cDNA library: identification of genes expressed during hindbrain, midbrain and cranial neural crest development.** *Mech. Devel.* 2001. 102:119-133.

Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO and Herskowitz I. **The transcriptional program of sporulation in budding yeast.** *Science* 1998. 282:699-705.

Cohen BA, Mitra RD, Hughes JD & Church GM. **A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression.** *Nature Genet.* 2000.26:183-186.

Corpet F, Grouzy J and Khan D. **Recent improvements of the ProDom database of protein domain families.** *Nucleic Acids Res.* 1999. 27:263-267.

Davidson EH, Klein WH and Britten RJ. **Sequence organization in animal DNA and a speculation on hnRNA as a coordinate regulatory transcript.** *Dev. Biol.* 1977. 55:69-84.

Dayhoff MO, Schwartz RM and Orcutt BC. **A model of evolutionary change in proteins.** *Atlas of Protein Sequence and Structure.* 1978. 5:345-352.

de Martin R, Schmid JA & Hofer-Warbinek R. **The NF-kappaB/Rel family of transcription factors in oncogenic transformation and apoptosis.** *Mutat. Res.* 1999. 437:231-243.

Defossez P and Gilson E. **The vertebrate protein CTCF functions as an insulator in *Saccharomyces cerevisiae*.** *Nuc. Acids Res.* 2002. 30:5136-5141.

DeRisi JL, Vishwanath RI & Brown PO. **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997. 278: 680-686.

Devos D and Valencia A. **Intrinsic errors in genome annotation.** *Trends in Genetics* 2001. 17:429-431.

Dhalluin C, Carlson JE, Zheng L, He C, Aggarwal AK, Zhou MM. **Structure and Ligand of a histone acetyltransferase bromodomain.** *Nature* 1999. 399:491-496.

Dover J, Schneider J, Tawaiah-Boateng MA, Wood A, Dean K, Johnston M & Shilatifard A. **Methylation of histone H3 by COMPASS requires Ubiquitination of histone H2B by Rad6.** *J. Biol. Chem.* 2002. 277:28368-28371.

Dujon B. **European Functional Analysis Network (EUROFAN) and the functional analysis of the *Saccharomyces cerevisiae* genome.** *Electrophoresis* 1998. 19:617-624.

Eisen MB, Spellman PT, Brown PO & Botstein D. **Cluster analysis and display of genome-wide expression patterns.** *Proc. Natl. Acad. Sci. USA* 1998. 95(25):14863-14868.

Enright AJ, Iliopoulos I, Kyrpides NC & Ouzounis CA. **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999. 402:86-90.

Ewing B, Hillier L, Wendl MC, and Green P. **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Research* 1998a. *:175-185.

Ewing B & Green P. **Base calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Research* 1998b. 8:186-194.

Ewing B & Green P. **Analysis of expressed sequence tags indicates 35,000 human genes.** *Nature Genetics* 2000. 25:232-234.

Fey S, Nawrocki A, Larsen MR, Gorg A, Roepstorff P, Skews GN, Williams R, Mose Larsen P. **Proteome analysis of *Saccharomyces cerevisiae*: A methodological outline.** *Electrophoresis* 1997. 18:1361-1372.

Fields C, Adams MD, White W & Venter JC. **How many genes in the human genome?** *Nature Genetics* 1994. 7:345-346.

Fields S and Song O. **A novel genetic system to detect protein-protein interactions.** *Nature* 1989. 340:245-246.

Fierro-Monti I and Mathews MB. **Proteins binding to duplexed RNA: one motif, multiple functions.** *Trends. Biochem. Sci.* 2000. 25:241-246.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. **Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*.** *Nature* 1998. 391:806-811.

Fisher RA and Yates F. **Example 12.** *Statistical Tables.* London, 1938.

Fleischmann W, Moller S, Gateau A and Apweiler R **A novel method for automatic functional annotation of proteins.** *Bioinformatics* 1999. 15:228-233.

Fondrat C & Kalogeropoulos A. **Approaching the function of new genes by detection of their potential upstream activation sequence in *Saccharomyces cerevisiae*: application to chromosome III.** *Current Genetics* 1994. 25:396-406.

Frech K, Quandt D and Werner T. **Muscle actin genes: A first step towards computational classification of tissue specific promoters.** *In Silico Biol.* 1998. 1, 0005. http://www.bioinfo.de/isb/1998/01/0005/

Fu GK, Stames S and Stuve L. **Construction of unidirectionally cloned cDNA libraries from messenger RNA for improved 3' end DNA sequencing.** *United States Patent. 2002.* Number 6,387,624.

Fujibuchi W, Anderson JSJ and Landsman D. **PROSPECT improves *cis*-acting regulatory element prediction by integrating expression profile data with consensus pattern searches.** *Nuc. Acids Res.* 2001. 29(19):3988-3996.

Gajewski M and Voolstra C. **Comparative analysis of somitogenesis related genes of the *hairy/Enhancer* of *split* class in *Fugu* and zebrafish.** *BMC Genomics.* 2002. 3:21. http://www.biomedcentral.com/1471-2164/3/21

Galagan JE *et al.* (78 authors). **The genome of the filamentous fungus *Neurospora crassa*.** *Nature* 2003. 422:859-869.

Galau GA, Klcin WH, Britten RJ & Davidson EH. **Significance of rare mRNA sequences in liver.** *Arch. Biochem. Biophys.* 1977. 179:584-599.

Gardner MJ *et al.* (40 authors). **Genome sequence of the human malaria parasite *Plasmodium falciparum*.** *Nature* 2002. 419:498-511.

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D and Brown PO. **Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes.** *Mol. Biol. Cell* 2000. 11:4241-4257.

Geanacopoulos M, Vasmatzis G, Lewis DEA, Roy S, Lee B and Adhya S. **GalR mutants defective in repressosome formation.** *Genes & Dev.* 1999. 13:1251-1262.

Gill G and Ptashne M. **negative effect of the transcriptional activator GAL4.** *Nature* 1988. 334:721-724.

Goodman N. **Biological data becomes computer literate: new advances in bioinformatics.** *Current Opinion in Biotechnology* 2002. 13:68-71.

Green, P. 1996. http://bozeman.mbt.washington.edu/phrap.docs/phrap.html

Gu A, Nicolae D, Lu H H-S and Li W-H. **Rapid divergence in expression between duplicate genes inferred from microarray data.** *TRENDS in Genet.* 2002. 18:609-613.

Guo S and Kemphues KJ. **par-1, a gene required for establishing polarity in C. elegans embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed.** *Cell* 1995. 81:611-620.

Gustafsson CM and Samuelsson T. **Mediator – a universal complex in transcriptional regulation.** *Mol. Microbiol.* 2001. 41:1-8.

Hampson S, Kibler D and Baldi P. **Distribution patterns of over-represented *k*-mers in non-coding yeast DNA.** *Bioinformatics* 2002. 18:513-528.

Hampton GM and Frierson Jr HF. **Classifying human cancer by analysis of gene expression.** *Trends in Genetics* 2003. 9:5-10.

Hayashizaki Y *et al.* (96 authors). **Functional annotation of a full-length mouse cDNA collection.** *Nature* 2001. 409:685-690.

Hegyi H and Gerstein M. **The Relationship between Protein Structure and Function: a Comprehensive Survey with Application to the Yeast Genome.** *J. Mol. Biol.* 1999. 288:147-164.

Henikoff S and Henikoff JG. **Amino acid substitution matrices from protein blocks.** *PNAS USA.* 1992. 89(22):10915-10919.

Hieter P and Boguski M. **Functional Genomics: It's all how you read it.** *Science* 1997 278: 601-602.

Ho Y *et al.* (46 authors). **Systemaatic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry.** *Nature* 2002. 415:180-183.

Hodgman T. **A historical perspective on gene/protein functional assignment.** *Bioinformatics* 2000 16:10-15.

Hofmann D, Bucher P, Falquet L and Bairoch A. **The PROSITE database, its status in 1999.** *Nucleic Acids Res.* 1999. 27:215-219.

Hogenesch JB, Ching KA, Batalov S, Su AI, Walker JR, Zhou Y, Kay SA, Schultz PG and Cooke MP. **A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes.** *Cell* 2002. 106:413-415.

Huang X, Adams AD, Zhou H and Kerlavage AR. **A Tool for Analyzing and Annotating Genomic Sequences.** *Genomics* 1997. 46:37-45.

Jacobs SA and Khorasanizadeh S. **Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail.** *Science* 2002. 295:2080-2083.

Jones DO, Cowell IG and Singh PB. **Mammalian chromodomain proteins: their role in genome organisation and expression.** *Bioessays* 2000. 22:124-137.

Junker VL, Apweiler R and Bairoch A. **Representation of functional information in the SWISS-PROT Data Bank.** *Bioinformatics* 1999. 15:1066-1067.

Kadonaga JT. **Eukaryotic transcription: An interlaced network of transcription factors and chromatin-modifying machines.** *Cell* 1998. 92: 307-317.

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P & Ahringer J. **Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi.** *Nature* 2003. 421:231-237.

Kanehisa M, Goto S, Kawashima S and Nakaya A. **The KEGG databases at GenomeNet.** *Nucleic Acids Res.* 2002. 30:42-46.

Karlin S and Altschul SF. **Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.** *Proc. Natl. Acad. Sci. USA* 1990. 87:2264-2268.

Kel-Margoulis OV, Romashchenko AG, Kolchanov NA, Wingender E and Kel AE. **COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation.** *Nuc. Acids Res.* 2000. 28:311-315.

Kim YJ, Bjorklund S, Li U, Sayre MH and Kornberg RD. **A mulriprotein mediator of transcriptional activation and its interaction with the C-terminal repreat domain of RNA polymerase II.** *Cell* 1994. 77:599-608.

Klar A, Baldassare M, Jessell TM. **F-spondin: a gene expressed at high levels in the floor plate encodes a secreted protein that promotes neural cell adhesion and neurite extension.** *Cell* 1992. 69:95-110.

Klingenhoff A, Frech K, Quandt K and Werner T. **Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity.** *Bioinformatics* 1999. 15:180-186.

Klingenhoff A, Frech K and Werner T. **Regulatory modules shared within gene classes as well as across gene classes can be detected by the same in silico approach.** *In Silico Biol.* 2000. 1, 0020. http://www.bioinfo.de/isb/2000010020/

Knuth DE. **The Art of Computer Programming.** *Volume 2, Third edition.* Section 3.4.2, Algorithm P, pp 145. Reading: Addison-Wesley. 1997. ISBN: 0-201-89684-2.

Ko MSH. **An "equalized cDNA library" by the reassociation of short double-stranded cDNAs.** *Nuc. Acids Res.* 1990. 18:5705-5711.

Koppensteiner WA, Lackner P, Wiederstein M and Sippl MJ. **Characterization of Novel Proteins Based on Known Protein Structures.** *J. Mol. Biol.* 2000. 296:1139-1152.

Kornberg RD. **Eukaryotic transcription control.** *Trends in Genetics* 1999a. 15(12):46-49.

Kornberg RD & Lorch Y. **Twenty-five years of the nucleosome, fundamental particle of the eukaryotic chromosome.** *Cell* 1999b. 98:285-294.

Kruglyak S and Tang H. **Regulation of adjacent yeast genes.** *Trends. Genet.* 2000. 16:109-111.

LanderES *et al.* (over 250 authors). **Initial sequencing and analysis of the human genome.** *Nature* 2001. 409:860-921.

Lanz RB, McKenna NJ, Onate SA, Albrecht U, Wong J, Tsai SY, Tsai MJ and O'Malley BW. **A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex.** *Cell* 1999. 97:17-27.

Li H and Capatanaki Y. **An E box in the desmin promoter cooperates with the E-box and MEF-2 sites of a distal enhancer to direct muscle-specific transcription.** *EMBO J.* 1994. 13:3580-3589.

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL & Quackenbush J. **Gene index analysis of the human genome estimates approximately 120,000 genes.** *Nature Genetics* 2000a. 25:239-240.

Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL & Quackenbush J. **An optimized protocol for analysis of EST sequences.** *Nuc. Acids Res.* 2000b. 28:3657-3665.

Lichtarge O and Sowa ME. **Evolutionary predictions of binding surfaces and interactions.** *Curr. Opin. Struct. Biol.* 2002. 12:21-27.

Liphitz HD, Peattie DA and Hogness DS. **Novel transcripts from the Ultrabithorax domain of the bithorax complex.** *Genes Dev.* 1987. 1:307-322.

Lipman DJ and Pearson WR. **Rapid and sensitive protein similarity searches.** *Science.* 1985. 227:1435-1441.

Lockhart DJ & Winzeler EA. **Genomics, gene expression and DNA arrays.** *Nature* 2000 405: 827-836.

Luger K & Richmond TJ. **The histone tails of the nucleosome.** *Curr. Opin. Genet. Devel.* 1998. 8:140-146.

Malik S & Roeder RG. **Transcriptional regulation through Mediator-like coactivators in yeast and metazoan cells.** *TIBS.* 2000. 25:277-283.

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO & Eisenberg D. **Detecting Protein Functino and Protein-Protein Interactions from Genome Sequences.** *Science* 1999. 285:751-753.

Mattick JS. **Non-coding RNAs: the architects of eukaryotic complexity.** *EMBO report.* 2001a. 2:986-991.

Mattick JS and Gagen MJ. **The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms.** *Mol. Biol. Evol.* 2001b. 18:1611-1630.

Matys V *et al.* (21 authors). **TRANSFAC®: transcriptional regulation, from patterns to profiles.** *Nuc. Acid. Res.* 2003. 31:374-378.

McKinsey TA, Zhang CL and Olson EN. **Signaling chromatin to make muscle.** *Curr. Opin. Cell Biol.* 2002. 14:763-772.

Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S and Weil B. **MIPS: a database for genomes and protein sequences.** *Nucleic Acids Res.* 2002. 30:31-34.

Michalovich D, Overington J and Fagan R. **Protein sequence analysis *in silico*: application of structure-based bioinformatics to genomics initiatives.** *Curr. Opin. Pharmacol.* 2002. 2:574-580.

Myers LC, Gustafsson CM, Bushnell DA, Lui M, Erdjument-Bromage H, Tempst P and Kornberg RD. **The Med proteins of yeast and their function through the RNA polymerase II carboxy-terminal domain.** *Genes Devel.* 1998. 12:45-54.

Needleman SB and 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.

Nei M, Xu P and Glazko G. **Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms.** *Proc. Natl. Acad. Sci. USA* 2001. 98:2497-2502.

Okazaki Y *et al.*(138 authors). **Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.** *Nature* 2002. 420:563-573.

Oliver SG, Winson MK, Kell DB and Baganz F. **Systematic functional analysis of the yeast genome.** *Trends in Biotechnology* 1998. 16:373-378.

Orengo CA, Pearl FMG, Bray JE, Todd AE, Martin AC, Conte LL and Thornton JM. **The CATH Database provides insights into protein structure/function relationships.** *Nuc. Acid. Res.* 1999. 27:275-279.

Patanjali SR, Parimoo S, Weissman SM. **Construction of a uniform-abundance (normalized) cDNA library.** *Proc. Natl. Acad. Sci. USA* 1991. 88:1943-1947.

Pearson WR, Lipman DJ. **Improved tools for biological sequence comparison.** *Proc. Natl. Acad. Sci. USA* 1988. 85:2444-2448.

Pearson WR. **Flexible sequence similarity searching with the FASTA2 program package.** *Methods Mol. Biol.* 2000. 137:185-219.

Pellegrini M, Marcotte EM, Thompson MJ, Eisneberg D & Yeats TO. **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc. Natl. Acad. Sci. USA* 1999. 96:4285-4288.

Pilpel Y, Sudarasanam P and Church GM. **Identifying regulatory networks by combinatorial analysis of promoter elements.** *Nat. Genet.* 2001. 29:153-159.

Prestridge DS. **Predicting Pol II promoter sequences using transcription factor binding site.** *J. Mol. Biol.* 1995. 249:923-932.

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R and White J. **The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species.** *Nuc. Acids Res.* 2001. 29:159-164

Rajkovic A, Tan C, Tan W, Klysik M, and Matzuk MM. ***Obox,* a Family of Homeobox Genes Preferentially Expressed in Germ Cells.** *Genomics* 2002. 79:711-717.

Rao A, Luo C and Hogan PG. **Transcription factors of the NFAT family: regulation and function.** *Annu. Rev. Immunol.* 1997. 15:707-747.

Rea S, Eisenhaber F, O'Carroll D, Strahl BD, Sun ZW, Schmid M, Opravil S, Mechtler K, Ponting CP, Allis CD and Jenuwein T. **Regulation of chromatin structure by site-specific histone H3 methyltransferases.** *Nature* 2000. 406:593-599.

Richterich P. **Estimation of Errors in "Raw" DNA Sequences: A Validation Study.** *Genome Research* 1998. 8:251-259.

Rissoan MC, Duhen T, Bridon JM, Bendriss-Vermare N, Peronne C, Vis Bde S, Briere F, Bates EE. **Subtractive hybridization reveals the expression of immunoglobulinlike transcript 7, Eph-B1, granzyme B, and 3 novel transcripts in human plasmacytoid dendritic cells.** *Blood* 2002. 100:3295-3303.

Ross-Macdonald P, Coelho PSR, Roemer T, Agarwai S, Kumar A, Jansen R, Cheung KH, Sheehan A, Symonlatis D, Umansky L, Heldtman M, Nelson FK, Iwasaki H, Hager K, Gerstein M, Miller P, Roeder GS & Snyder M. **Large-scale analysis of the yeast genome by transposon tagging and gene disruption.** *Nature* 1999. 402:413-418.

Sanchez-Herrero E and Akam M. **Spatially ordered transcription of regulatory DNA in the bithorax complex of *Drosophila*.** *Development* 1989. 107:321-329.

Sanger F and Coulson AR. **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.** *J.Mol.Biol.* 1975. 94:441-448.

Sanger F and Coulson AR. **The use of thin acrylamide gels for DNA sequencing.** *FEBS Letters* 1978. 87:107-110.

Sanger F, Nicklen S and Coulson AR. **DNA sequencing with chain-terminating inhibitors.** *Proc. Natl. Acad. Sci. USA* 1977. 74:5463-5467.

Sap J, Munoz A, Schmitt J, Stunnenberg H, Vennstrom B. **Repression of transcription mediated at a thyroid hormone response element by the v-erb-A oncogene product.** *Nature* 1989. 340:242-244.

Schmitt AO, Specht T, Beckmann G, Dahl E, Pilarsky CP, Hinzmann B and Rosenthal A. **Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues.** *Nuc. Acids Res.* 1999. 27:4251-4260.

Schuldiner O, Yanover C & Benvenisty N. **Computer analysis of the entire budding yeast genome for putative targets of the GCN4 transcription factor.** *Current Genetics* 1998 33:16-20.

Schultz J, Copley RR, Coerks T, Ponting CP and Bork P. **SMART: a web-based tool for the study of genetically mobile domains.** *Nucleic Acids Res.* 28:231-234.

Shaul S and Graur D. **Playing chicken (*Gallus gallus*): methodological inconsistencies of molecular divergence data estimates due to secondary calibration points.** *Gene* 2002. 300:59-61.

Shi Y and Berg JM. **Specific DNA-RNA hybrid binding by zinc finger proteins.** *Science* 1995. 268:282-284.

Shoemaker DD and Linsley PS. **Recent developments in DNA microarrays.** *Curr. Opin. Microbiol.* 2002. 5:334-337.

Sinha S and Tompa M. **Discovery of novel transcription factor binding sites by statistical overrepresentation.** *Nuc. Acids Res.* 2002 30:5549-5560

Skolnick J & Fetrow JS. **From genes to protein structure and function: novel application of computational approaches in the genomic era.** *TIBTECH* 2000 18:34-39.

Smith TF. **Functional genomics – bioinformatics is ready for the challenge.** *Trends in Genetics* 1998. 14:291-293.

Souciet JL *et al.* (24 authors). **Genomic Exploration of the Hemiascomycetous Yeasts: 1. A set of yeast species for molecular evolution studies.** *FEBS Letters* 2000. 487:3-12.

Spahr H, Beve J, Larsson T, Bergstrom J, Karlsson KA and Gustafsson CM. **Purification and characterization of RNA polymerase II holoenzume from *Schizosaccharomyces pombe*.** *J. Biol. Chem.* 2000. 275:1351-1356.

Spellman PT and Rubin GM. **Evidence for large domains of similarly expressed genes in the *Drosophila* genome.** *Journal of Biology* 2002. 1:5. http://jbiol.com/content/1/1/5

Staden R, Beal KF and Bonfield JK. **The Staden package, 1998.** *Methods Mol. Biol.* 2000. 132:115-130.

Stapleton M, Carlson J, Brokstein P, Yu C, Champe M, George R, Guarin H, Kronmiller B, Pacleb J, Park S, Wan K, Rubin GM and Celniker SE. **A *Drosophila* full-length cDNA resource.** *Genome Biology* 2002. 3:1-8. http://genomebiology.com/2002/3/12/research/0080

Stawiski EW, Baucom AE, Lohr SC and Gregoret LM. **Predicting protein function from structure: Unique structural features of proteases.** *Proc. Natl. Acad. Sci. USA* 2000. 97:3954-3958.

Stoesser G, Baker W, van den Broek A, Garcia-Pastor M, Kulikova T, Leinonen R, Lin Q, Lombard V, Lopez R, Mancuso R, Nardone F, Stoehr P, Tuli MA, Tzouvara K, Vaughan R. **The EMBL Nucleotide Sequence Database: major new developments.** *Nucleic Acids Res.* 2003. 31:17-22.

Stormo GD. **DNA binding sites: representation and discovery.** *Bioinformatics* 2000. 16:16-23.

Strahl BD & Allis D. **The language of covalent histone modifications.** *Nature* 2000. 403:41-45

Struhl K. **Fundamentally Different Logic of Gene Regulation in Eukaryotes and Prokaryotes.** *Cell* 1999. 98:1-4.

Sun ZW and Allis CD. **Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast.** *Nature* 2002. 418:104-108.

Sutton G, White O, Adams M and Kerlavage A. **TIGR Assembler: A new tool for assembling large shotgun sequencing projects.** *Genome Science & Technology* 1995. 1:9-19.

The *C. elegans* Sequencing Consortium. **Genome Sequence of the Nematode *Caenorhabditis elegans*. A Platform for Investigating Biology.** *Science* 1998. 282:2012-2018.

Thompson JD, Higgins DG and Gibson TJ. **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weigh matrix choice.** *Nucleic Acids Res.* 1994. 22:4673-4680.

Tirunaguru VG, Sofer L, Cui J and Burnside J. **An expressed sequence tag database of T-cell enriched activated chicken splenocytes: sequence analysis of 5251 clones.** *Genomics* 2000. 66:144-151.

Tominaga M, Tomooka Y. **Novel genes cloned from a neuronal cell line newly established from a cerebellum of an adult p53(-/-) mouse.** *Biochem. Biophys. Res. Commun.* 2002. 297:473-479.

Toyoda T and Konagaya A. **KnowledgeEditor: a new tool for interactive modelling and analysing biological pathways based on microarray data.** *Bioinformatics* 2003. 19:433-434.

Tweedale H, Notley-McRobb L, Ferenci T. **Effect of slow growth on metablism of *Escherichia coli*, as revealed by global metabolite pool ("Metabolome") analysis.** *Journal of Bacteriology* 1998. 180:5109-5116.

Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Quershi-Emili A, Li Y, Godwin B, Conover D, Kalbfeisch T, Vijayadamodar, G, Yang M, Johnston M, Fields S and Rothberg JM. **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*.** *Nature* 2000. 403:623-627.

Van Helden J, Andre B and Collado-Vides J. **Extracting regulatory sites from the upstream regions of yeast genes by computational analysis of oligonucleotide frequencies.** *J. Mol. Biol.* 1998 281:827-842.

Vukmirovic OG and Tilghman SM. **Exploring genome space.** *Nature* 2000. 405: 820-822.

Wallace AC, Borkakori N and Thornton JM. **TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases.** *Protein Science* 1997. 6:2308-2323.

Weng S, Dong Q, Balakrishnan R, Christie K, Costanzo M, Dolinski K, Dwight SS, Engel S, Fisk DG, Hong E, Essel-Tarver L, Sethuraman A, Theesfeld C, Andrada R, Binkley G, Lane C, Schroeder M, Botstein D and Cherry JM. ***Saccharomyces* Genome Database (SGD) provides biochemical and structural information for budding yeast proteins.** *Nuc. Acid. Res.* 2003. 31:216-218.

Werner T. **Models for prediction and recognition of eukaryotic promoters.** *Mammakuan Genome* 1999. 10:168-175.

Werner T. **Promoters can contribute to the elucidation of protein function.** *Trends Biotechnol.* 2003. 21:9-13.

Wilbur WJ and Lipman DJ. **Rapid similarity searches of nucleic acid and protein data banks.** *Proc. Natl. Acad. Sci. USA* 1983. 80:726-730.

Winding P and Berchtold MW. **The chicken B cell line DT40: a novel tool for gene disruption experiments.** *J. Immunol. Methods* 2001. 249:1-16.

Wingender E, Kel AE, Kel OV, Karas H, Heineeyer T, Dietze P, Knuppel R, Romaschenko AG, Kolchanov NA. **TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation.**

Wolfsberg TG, Gabrielian AE, Campbell MJ, Cho RJ, Spouge JL & Landsman D. **Candidate regulatory sequence elements for cell cycle-dependent transcription in *Saccharomyces cerevisiae.*** *Genome Research* 1999. 9:775-792.

Wong S, Butler G, Wolfe KH. **Gene order evolution and paleopolyploidy in hemiascomycete yeasts.** *Proc. Natl. Acad. Sci. U.S.A.* 2002. 99:9272-9277.

Yamauchi M, Ogata Y, Kim RH, Li JJ, Freedman LP and Sodek J. **AP-1 regulation of the rat bone sialoprotein gene transcription is mediated through a TPA response element within a glucocorticoid response unit in the gene promoter.** *Matrix Biol.* 1996. 15:119-130.

Zdobnov EM and Apweiler R. **InterProScan – an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001. 17:847-848.

Zhang MQ. **Large-scale gene expression data analysis: A new challenge to computational biologists.** *Genome Research* 1999. 9:681-688.

Zheng J, Wu J and Sun Z. **An approach to identify over-represented *cis*-elements in related sequences.** *Nucleic Acids Res.* 2003. 31:1995-2005.

Zhu J and Zhang MQ. **SCPD: a promoter database of the yeast *Saccharomyces cerevisiae.*** *Bioinformatics* 1999. 15:607-611.

Web references

1. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome - NCBI completed genome list
2. http://rana.stanford.edu/software/ - Microarray software
3. http://mips.gsf.de/proj/yeast/CYGD/db/index.html - Functional classification of the yeast genome
4. http://oregonstate.edu/instruction/bb492/fignumbers/Fig28.34.html – on-line figure showing alternate splicing of rat tropomyosin
5. http://www.zmdb.iastate.edu/zmdb/ZMDB_HOWTO.html
6. http://www.phrap.com – Phrap web site
7. http://www.mrc-lmb.cam.ac.uk/pubseq/ - Staden Package web-site
8. http://eyeball.eng.uiowa.edu/clustering/
9. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene – UNIGENE web-site
10. http://www.gsc.riken.go.jp/e/FANTOM/ - FANTON web-site
11. http://www.incyte.com/index.shtml – Incyte Genomics
12. http://www.chick.umist.ac.uk/ - BBSRC Chicken EST web-page
13. ftp://ftp.ebi.ac.uk/pub/sogtware/unix/iprscan/ -InterPro iprscan software
14. ftp://ftp.ensembl.org/pub/current_human/data/ - Ensembl sequence data
15. http://www.ncbi.nlm.nih.gov/omim/ - Online Mendelian Inheritance in Man database
16. http://www.geneontology.org – gene ontology web-site
17. http://www.ebi.ac.uk/proteome/
18. http://www.hgmp.mrc.ac.uk/geneservice/reagents/products/cdna_resources/index.shtml – MRC geneservice
19. ftp://rocky.bms.umist.ac.uk – BBSRC Chicken EST ftp server
20. http://www.gene-regulation.com - TRANSFAC public release
21. http://cgsigma.cshl.org/jian/ - SCPD
22. http://genome-www.stanford.edu/Saccharomyces/ - SGD main page
23. ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/genomic_sequence/orf_dna/ - SGD ftp site for sequence data
24. http://acer.gen.tcd.ie/~khwolfe/yeast/nova/index.html - Ken Wolfe, yeast gene duplications web site.
25. http://wolf.bms.umist.ac.uk/~peb/interestingClusters.html – details on clusters found from initial vector analyses.
26. ftp://ftp.mips.embnet.org/pub/yeast/ORF_SEQ/ - MIPS sequence data ftp site
27. http://www.genome.ad.jp/kegg/ - KEGG database

28. http://mips.gsf.de/proj/yeast/CYGD/db/index.html - MIPS functional classification
29. http://rocky.bms.umist.ac.uk/SiteSeer/ -SiteSeer promoter visualisation tool
30. http://www.imswebtips.com/issue56top1.htm - web safe colours
31. http://wolf.bms.umist.ac.uk/~peb/eisen_clusters/ - SiteSeer visualisations of microarray clusters from Eisen *et al.* (1998)
32. http://wolf.bms.umist.ac.uk/~peb/northern/ - SiteSeer visualisations of gene clusters generated from Brown *et al.* (2001)
33. http://wolf.bms.umist.ac.uk/~peb/mips/broad/ - SiteSeer visualisations of broad MIPS functional categories
34. http://wolf.bms.umist.ac.uk/~peb/mips/refined/ - SiteSeer visualisations of specific MIPS functional categories
35. http://wolf.bms.umist.ac.uk/~peb/kegg/broad/ - SiteSeer visualisations of broad KEGG functional categories
36. http://wolf.bms.umist.ac.uk/~peb/kegg/refined/ - SiteSeer visualisations of specific KEGG categories
37. http://cbi.labri.fr/Genolevures/index.php – Génolevures web-site
38. http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=Saccharomyces+cerevisiae+[gbpln] - codon bias and internal nucleotide frequency of codons for cerevisiae
39. http://wolf.bms.umist.ac.uk/~peb/siteseer_results_6_seqs_or_more/ - SiteSeer visualisations of the groups of orthologous genes containing 6 or more sequences
40. http://wolf.bms.umist.ac.uk/~peb/siteseer_all/ - SiteSeer visualisations of all Génolevures orthologous groups
41. http://www.soe.ucsc.edu/~kent/ - Dr Jim Kent's homepage

Appendix 1

Table A1.1. Tissues, library names, library sizes and their level of mitochondrial and rRNA contamination.

tissue	library name	Number of sequences	rRNA (%)	Mito (%)
16 days embryo brain	CSEQCHL24_VKU	2397	0	2
	CSEQCHN57_RIX	2856	0	0.14
	CSEQCHN66_IZG	2808	0	0.14
	CSEQCHN67_WGU	2855	0.04	0.11
adult adipose	CSEQRBL06_IGK	2705	0.07	0.89
adult brain - cerebellum	CSEQCHL25_YXB	2781	0	1.01
	CSEQCHN68_FOG	2850	0	0.04
	CSEQCHN69_IJA	4377	0.05	0.07
adult brain - cerebrum	CSEQCHL15_YYO	2726	0.04	0.48
	CSEQCHN53_LTX	2842	0.07	0
	CSEQCHN72_FXF	2781	0.07	0.07
adult brain - other part	CSEQCHL16_XSM	2731	0.15	1.5
	CSEQCHN73_UOQ	2900	0	0
	CSEQCHN54_QNR	4114	0	0
adult heart	CSEQCHL26_TZW	2073	0	5.31
	CSEQCHN60_OAP	2903	0	0.1
	CSEQCHN61_UVA	1940	0.05	0
adult kidney + adrenal	CSEQCHL17_TOJ	1910	0	1.47
	CSEQCHN74_OHW	2777	0	0.04
	CSEQCHN55_NCT	4257	0	0.05
adult liver	CSEQCHL19_GWB	2014	0.25	0.79
	CSEQCHN33_OJM	1453	0	0
	CSEQCHN33_LFQ	1387	0	0
	CSEQCHN34_FDV	2744	0	0
adult small intestine	CSEQCHL18_AIN	2014	0.15	0.94
	CSEQCHN58_AWI	2797	0.14	0
	CSEQCHN56_VEI	4299	0	0
adult pancreas	CSEQRBL07_JMS	1037	0.94	0
chondrocytes	CSEQRBL03_MCJ	2063	0	0.53
	CSEQRBN09_RWG	3927	0	0.03
	CSEQRBN10_JFU	2470	0	0
	CSEQRBN10_EPW	1221	0	0.08
	CSEQRBN20_AID	2705	0	0.04
	CSEQRBN22_BTM	4211	0	0.02

tissue	library name	Number of sequences	rRNA (%)	Mito (%)
muscle	CSEQRBL04_OLC	1380	0.22	2.61
	CSEQRBN11_INT	2473	0	0.24
	CSEQRBN15_KAK	4077	0.05	0.02
ovary	CSEQRBL05_CPD	1362	0.15	0.29
	CSEQRBN13_ZQZ	2533	0	0
	CSEQRBN14_IPE	3776	0	0
	CSEQRBN21_BJG	2627	0	0
	CSEQRBN19_CFF	4085	0	0.1
stage 10	CSEQCHL23_LHX	2833	0.11	1.16
	CSEQCHN64_ZJU	2687	0.07	0.04
	CSEQCHN65_XYK	4042	0.05	0.05
stage 20-21	CSEQCHL01_JIU	2684	0.3	1.83
	CSEQCHN03_PWB	2709	0.07	0.11
	CSEQCHN04_BWQ	4030	0.15	0.15
stage 22 heads	CSEQCHL14_RYH	2767	0.18	0.65
	CSEQCHN23_FZD	2857	0.07	0.32
	CSEQCHN24_JFD	4033	0	0.07
stage 22 limbs	CSEQCHL13_UZO	2792	0.04	1.07
	CSEQCHN51_GSM	2798	0	0
	CSEQCHN52_OJD	4189	0	0.07
stage 36 heads	CSEQCHL22_UCY	2701	0.07	0.41
	CSEQCHN62_NTG	2860	0	0.03
	CSEQCHN63_OTT	4350	0	0
stage 36 hearts	CSEQCHL12_NGJ	2502	0.24	1.12
	CSEQCHN70_CYQ	2872	0.03	0.1
	CSEQCHN71_AIJ	2831	0	0.07
stage 36 limbs	CSEQCHL20_YDI	2736	0.29	1.28
	CSEQCHN38_SQJ	2714	0	0
	CSEQCHN59_BKX	4365	0.02	0
stage 36 trunks	CSEQCHL21_XMQ	2664	0.3	0.9
	CSEQCHN35_TKI	2560	0.16	0.16
	CSEQCHN75_UDJ	2675	0	0

Table A1.2. Clustering analysis of 200,000 chicken ESTs.

This table gives an overview of the preliminary clustering results prior to the final 150,000 sequencing allocations. The Norm column refers to the level of normalisation that this library underwent (S = no normalisation. The letters A to F refer to different normalisation protocols). The "level of redundancy" refers to the number of sequences with a stringent match to another EST within the same library. Mean cluster size is the average number of ESTs per cluster after clustering the library with the ***uicluster2*** program (see Section 2.4.1).

Tissue	Library	Norm	Level of redundancy	Mean cluster size	% Singleton gene clusters	% Singletons w/o BLAST of total lib
16 days embryo brain	CSEQCHL24_VKU	S	11.0%	1.12	84.07%	56.43%
	CSEQCHN57_RIX	A	6.2%	1.07	89.11%	71.57%
	CSEQCHN66_IZG	F	5.3%	1.06	91.70%	69.98%
	CSEQCHN67_WGU	E	2.9%	1.03	94.85%	70.40%
stage 20-21	CSEQCHL01_JIU	S	16.2%	1.19	79.48%	47.64%
	CSEQCHN03_PWB	A	1.9%	1.02	96.72%	75.32%
	CSEQCHN04_BWQ	B	2.8%	1.03	95.00%	75.23%
adult adipose	CSEQRBL06_IGK	S	15.9%	1.19	80.15%	54.64%
adult brain - cerebellum	CSEQCHL25_YXB	S	9.0%	1.10	86.89%	62.57%
	CSEQCHN68_FOG	F	2.9%	1.03	94.95%	71.79%
	CSEQCHN69_IJA	E	3.5%	1.04	94.45%	71.92%
adult brain - cerebrum	CSEQCHL15_YYO	S	10.9%	1.12	83.73%	52.47%
	CSEQCHN53_LTX	A	8.4%	1.09	88.42%	68.05%
	CSEQCHN72_FXF	D	3.0%	1.03	94.79%	70.33%
adult brain - other parts	CSEQCHL16_XSM	S	14.7%	1.17	79.39%	51.11%
	CSEQCHN54_QNR	A	2.3%	1.02	95.75%	76.69%
	CSEQCHN73_UOQ	D	1.9%	1.02	96.62%	70.48%
adult heart	CSEQCHL26_TZW	S	27.0%	1.37	67.89%	40.67%
	CSEQCHN60_OAP	F	8.2%	1.09	87.53%	63.04%
	CSEQCHN61_UVA	E	6.3%	1.07	91.03%	63.35%
adult kidney + adrenal	CSEQCHL17_TOJ	S	14.1%	1.16	80.67%	46.51%
	CSEQCHN55_NCT	A	4.9%	1.05	91.59%	64.69%
	CSEQCHN74_OHW	D	2.4%	1.02	95.61%	67.19%
adult liver	CSEQCHL19_GWB	S	27.5%	1.38	67.36%	36.20%
	CSEQCHN33_LFQ	A	2.4%	1.02	96.13%	68.22%
	CSEQCHN33_OJM	A	2.1%	1.02	96.84%	69.55%
	CSEQCHN34_FDV	B	3.5%	1.04	94.86%	66.44%

Tissue	Library	Norm	Level of redundancy	Mean cluster size	% Singleton gene clusters	% Singletons w/o BLAST of total lib
adult pancreas	CSEQRBL07_JMS	S	72.5%	3.64	23.99%	11.25%
adult small intestine	CSEQCHL18_AIN	S	19.2%	1.24	75.16%	40.84%
	CSEQCHN56_VEI	C	2.7%	1.03	94.88%	**90.53%**
	CSEQCHN58_AWI	E	3.2%	1.03	94.07%	66.39%
chondrocytes	CSEQRBL03_MCJ	S	16.9%	1.20	77.31%	50.80%
	CSEQRBN09_RWG	A	3.7%	1.04	93.89%	73.88%
	CSEQRBN10_EPW	B	1.3%	1.01	97.66%	77.93%
	CSEQRBN10_JFU	B	1.6%	1.02	97.18%	78.32%
	CSEQRBN20_AID	C	8.3%	1.09	84.92%	66.73%
	CSEQRBN22_BTM	D	4.4%	1.05	92.45%	74.78%
muscle	CSEQRBL04_OLC	S	19.2%	1.24	75.20%	40.06%
	CSEQRBN11_INT	A	3.0%	1.03	94.72%	71.30%
	CSEQRBN15_KAK	B	4.8%	1.05	90.97%	70.44%
ovary	CSEQRBL05_CPD	S	6.1%	1.06	90.85%	57.03%
	CSEQRBN13_ZQZ	A	2.0%	1.02	96.57%	74.85%
	CSEQRBN14_IPE	B	1.4%	1.01	97.62%	**80.35%**
	CSEQRBN19_CFF	D	3.4%	1.04	94.47%	69.91%
	CSEQRBN21_BJG	C	14.0%	1.16	74.99%	55.81%
stage 10	CSEQCHL23_LHX	S	16.8%	1.20	78.36%	42.55%
	CSEQCHN64_ZJU	F	4.5%	1.05	92.07%	62.08%
	CSEQCHN65_XYK	E	6.8%	1.07	89.34%	61.97%
stage 22 heads	CSEQCHL14_RYH	S	8.4%	1.09	88.27%	53.95%
	CSEQCHN23_FZD	A	5.9%	1.06	91.35%	63.63%
	CSEQCHN24_JFD	B	10.9%	1.12	86.91%	69.09%
stage 22 limbs	CSEQCHL13_UZO	S	16.6%	1.20	77.76%	42.85%
	CSEQCHN51_GSM	C	8.3%	1.09	86.20%	64.40%
	CSEQCHN52_OJD	D	8.2%	1.09	87.04%	65.34%
stage 36 heads	CSEQCHL22_UCY	S	12.7%	1.15	82.36%	46.43%
	CSEQCHN62_NTG	F	3.2%	1.03	95.00%	63.32%
	CSEQCHN63_OTT	E	6.1%	1.07	90.62%	60.67%
stage 36 hearts	CSEQCHL12_NGJ	S	26.1%	1.35	65.06%	33.29%
	CSEQCHN70_CYQ	F	4.7%	1.05	91.99%	55.92%
	CSEQCHN71_AIJ	E	5.8%	1.06	90.64%	54.79%

Tissue	Library	Norm	Level of redundancy	Mean cluster size	% Singleton gene clusters	% Singletons w/o BLAST of total lib
stage 36 limbs	CSEQCHL20_YDI	S	15.6%	1.18	79.70%	43.22%
	CSEQCHN38_SQJ	E	8.4%	1.09	90.48%	60.21%
	CSEQCHN59_BKX	F	5.0%	1.05	91.98%	
stage 36 trunks	CSEQCHL21_XMQ	S	19.6%	1.24	73.75%	36.65%
	CSEQCHN35_TKI	A	15.3%	1.18	81.11%	48.57%
	CSEQCHN75_UDJ	D	2.4%	1.02	95.25%	65.76%

Table A1.3. BLAST based redundancy statistics.

This table shows the redundancy statistics for libraries when compared against the following sequence databases: themselves and SwissProt/TrEMBL (self & sptr), SwissProt/TrEMBL (sptr), all ESTs sequenced from the same originating tissue (tissue), all ESTs from the same tissue and SwissProt/TrEMBL (tissue & sptr), all ESTs sequenced in the project (all ESTs), all sequenced ESTs and SwissProt/TrEMBL (all ESTs & sptr).

	% unique sequences when compared with					
library name	self & sptr	sptr	tissue	tissue & sptr	all ESTs	all ESTs & sptr
CSEQCHL24_VKU	42.89	62.87	44.97	32.46	9.72	8.72
CSEQCHN57_RIX	50.25	78.99	49.93	40.51	18.52	16.42
CSEQCHN66_IZG	50.89	74.18	45.62	36.79	10.22	9.51
CSEQCHN67_WGU	55.73	73.1	55.73	42.87	14.85	13.35
CSEQRBL06_IGK	40.81	62.4	57.86	40.81	13.64	12.09
CSEQCHL25_YXB	47.79	69	48.33	37.65	11.97	10.97
CSEQCHN68_FOG	57.47	74.84	54.11	43.61	14.32	13.3
CSEQCHN69_IJA	46.47	74.66	48.46	38.61	11.74	10.56
CSEQCHL15_YYO	40.83	59.17	47.58	32.02	10.09	8.51
CSEQCHN53_LTX	47.96	74.88	47.96	38.85	12.21	11.33
CSEQCHN72_FXF	61.88	73.07	66.41	51.96	19.99	17.84
CSEQCHL16_XSM	36.87	59.39	41.08	28.49	8.02	6.52
CSEQCHN73_UOQ	60.52	72.45	63.41	48.72	17.03	14.72
CSEQCHN54_QNR	64.03	79.75	66.63	54.59	20.25	18.5
CSEQCHL26_TZW	30.34	48.38	32.71	22.09	5.93	4.97
CSEQCHN60_OAP	48.85	67.9	51.15	39.79	10.82	10.23
CSEQCHN61_UVA	52.89	66.55	55.36	40.46	10.31	9.18
CSEQCHL17_TOJ	36.02	51.94	41.41	26.7	9.69	8.38
CSEQCHN74_OHW	58.08	69.54	62.87	46.24	18.15	15.74
CSEQCHN55_NCT	54.59	69.77	62.86	46.23	17.38	15.43
CSEQCHL19_GWB	25.42	43.74	31.33	19.36	7	5.71
CSEQCHN33_OJM	63.32	70.41	67.31	49.14	17.07	14.93
CSEQCHN33_LFQ	61.43	70.73	64.67	46.36	16.29	13.99
CSEQCHN34_FDV	59	68.88	67.09	49.82	19.06	17.02

library name	% unique sequences when compared with					
	self & sptr	sptr	tissue	tissue & sptr	all ESTs	all ESTs & sptr
CSEQRBL07_JMS	6	17.34	10.5	6	3	2.62
CSEQCHL18_AIN	30.04	47.02	35.75	21.95	7	5.71
CSEQCHN58_AWI	51.45	69.82	52.91	39.47	12.41	11.15
CSEQCHN56_VEI	70.67	95.3	62.29	59.67	20.73	20.19
CSEQRBL03_MCJ	36.35	57.44	33.4	23.65	9.31	8.68
CSEQRBN09_RWG	62.64	77.18	57.17	46.83	24.96	22.69
CSEQRBN10_JFU	70.16	80.04	62.39	52.06	32.35	30.2
CSEQRBN10_EPW	72.24	78.79	63.31	52.74	33.66	30.79
CSEQRBN20_AID	54.57	77.26	43.62	35.64	17.3	16.08
CSEQRBN22_BTM	59.84	79.34	52.34	44.55	24.41	22.87
CSEQRBL04_OLC	32.25	45.22	48.12	29.64	6.88	5.8
CSEQRBN11_INT	62.92	73.84	78.08	60.86	20.26	18.52
CSEQRBN15_KAK	59.73	77.68	76.31	59.6	77.56	61.81
CSEQRBL05_CPD	47.8	59.99	51.98	35.32	13.88	12.26
CSEQRBN13_ZQZ	69.21	77.1	70.04	55.35	32.37	28.54
CSEQRBN14_IPE	72.46	81.99	71	59.48	35.96	32.94
CSEQRBN21_BJG	45.49	73.51	42.82	32.89	18.08	16.22
CSEQRBN19_CFF	61.35	72.71	64.28	49.47	26.61	23.62
CSEQCHL23_LHX	33.67	47.3	39.22	25.03	6.32	5.4
CSEQCHN64_ZJU	47.9	65.35	47.26	33.57	8.6	7.48
CSEQCHN65_XYK	48.52	66.55	51.63	39.31	12.35	11.58
CSEQCHL01_JIU	36.44	56.3	49.33	32.45	11.66	9.76
CSEQCHN03_PWB	69.44	77.22	75.38	61.31	30.68	27.61
CSEQCHN04_BWQ	66.9	78.39	75.48	61.74	35.71	32.66
CSEQCHL14_RYH	41.99	56.7	53.16	36.03	9.29	8.06
CSEQCHN23_FZD	54.99	66.19	59.96	45.64	15.75	14.32
CSEQCHN24_JFD	63.82	79.17	72.7	59.31	31.66	28.89
CSEQCHL13_UZO	32.59	46.45	43.95	26.68	5.95	4.66
CSEQCHN51_GSM	50.64	72.34	51.39	38.92	11.19	10.08
CSEQCHN52_OJD	54.62	73.26	60.59	46.93	19.98	17.04
CSEQCHL22_UCY	38.28	49.83	45.98	29.51	11.59	9.74
CSEQCHN62_NTG	54.51	65.14	58.46	42.94	17.62	15.87
CSEQCHN63_OTT	48.44	63.7	55.72	40.97	13.1	11.24

library name	% unique sequences when compared with					
	self & sptr	sptr	tissue	tissue & sptr	all ESTs	all ESTs & sptr
CSEQCHL12_NGJ	23.34	40.53	30.1	17.91	6.83	5.72
CSEQCHN70_CYQ	42.37	58.67	49.13	33.98	10.9	9.58
CSEQCHN71_AIJ	44.01	56.94	51.36	35.18	12.33	11.06
CSEQCHL20_YDI	35.53	47.99	44.59	28.14	10.16	8.3
CSEQCHN38_SQJ	56.08	67.61	61.39	45.03	21.81	19.86
CSEQCHN59_BKX	63.8	99.08	52.12	51.68	11.59	11.52
CSEQCHL21_XMQ	29.24	41.67	44.22	24.66	9.8	7.17
CSEQCHN35_TKI	44.41	55.27	55.59	38.87	16.68	14.45
CSEQCHN75_UDJ	56.11	68.93	70.24	50.21	17.76	15.1

Table A1.4. Intra tissue redundancy analysis.

All projects from a tissue were compared with each other. Each library name is associated with a letter in parentheses. This refers to which library it represents in the comparison table. For example, library CSEQCHL24_VKU (16 days embryo brain) is library A. When compared with itself 61.95% of the sequences are redundant (find another similar sequence within the library).

		% unique sequences when compared with					
tissue	library name	lib A	lib B	lib C	lib D	lib E	lib F
16 days embryo brain	CSEQCHL24_VKU(A)	61.95	83.98	73.01	74.89		
	CSEQCHN57_RIX(B)	76.44	62.96	69.82	72.16		
	CSEQCHN66_IZG(C)	72.04	71.72	64.46	68.41		
	CSEQCHN67_WGU(D)	81.54	83.01	77.34	74.29		
adult adipose	CSEQRBL06_IGK(A)	57.86					
adult brain - cerebellum	CSEQCHL25_YXB(A)	64.4	75.08	71.77			
	CSEQCHN68_FOG(B)	79.33	73.82	71.86			
	CSEQCHN69_IJA(C)	75.65	68.97	59.79			
adult brain - cerebrum	CSEQCHL15_YYO(A)	61.74	75.5	73.7			
	CSEQCHN53_LTX(B)	70.09	60.87	75.33			
	CSEQCHN72_FXF(C)	84.21	84.83	81.12			
adult brain - other part	CSEQCHL16_XSM(A)	53.75	75.58	77.37			
	CSEQCHN73_UOQ(B)	86.59	81.24	82.03			
	CSEQCHN54_QNR(C)	90.76	86.51	79.78			
adult heart	CSEQCHL26_TZW(A)	46.65	55.23	63.58			
	CSEQCHN60_OAP(B)	76.44	64.42	72.89			
	CSEQCHN61_UVA(C)	78.97	68.92	75.21			
adult kidney + adrenal	CSEQCHL17_TOJ(A)	56.81	79.06	70.05			
	CSEQCHN74_OHW(B)	91.57	81.49	78.18			
	CSEQCHN55_NCT(C)	90.77	84.17	75.95			
adult liver	CSEQCHL19_GWB(A)	42.15	73.68	73.98	65.99		
	CSEQCHN33_OJM(B)	90.92	87.75	87.2	83		
	CSEQCHN33_LFQ(C)	90.41	87.17	87.31	83.27		
	CSEQCHN34_FDV(D)	88.78	88.08	89.21	81.71		

tissue	library name	% unique sequences when compared with lib A	lib B	lib C	lib D	lib E	lib F
adult pancreas	CSEQRBL07_JMS(A)	10.5					
adult small intestine	CSEQCHL18_AIN(A)	50.35	73.49	76.37			
	CSEQCHN58_AWI(B)	84.13	70.93	71.76			
	CSEQCHN56_VEI(C)	89.9	81.39	73.83			
chondrocytes	CSEQRBL03_MCJ(A)	52.11	67.23	76.44	86.28	74.89	66.94
	CSEQRBN09_RWG(B)	89.13	78.79	87.14	92.21	86.66	83.09
	CSEQRBN10_JFU(C)	91.01	84.13	87.37	93.56	89.96	85.95
	CSEQRBN10_EPW(D)	91.73	85.09	90.17	90.17	89.03	86.98
	CSEQRBN20_AID(E)	88.02	82.77	89.76	93.64	69.72	77.3
	CSEQRBN22_BTM(F)	86.75	81.22	87.49	91.95	82.05	72.48
muscle	CSEQRBL04_OLC(A)	54.2	69.06	98.26			
	CSEQRBN11_INT(B)	89.89	82.33	97.94			
	CSEQRBN15_KAK(C)	98.92	98.99	76.6			
ovary	CSEQRBL05_CPD(A)	73.49	86.27	87.44	85.24	74.82	
	CSEQRBN13_ZQZ(B)	94.24	89.42	89.54	91.39	87.25	
	CSEQRBN14_IPE(C)	95.1	91.79	88	91.95	87.61	
	CSEQRBN21_BJG(D)	93	90.86	87.97	61.25	83.14	
	CSEQRBN19_CFF(E)	90.62	90.21	86.98	88.45	81.96	
stage 10	CSEQCHL23_LHX(A)	55.17	70.81	64.31			
	CSEQCHN64_ZJU(B)	80.91	69.71	67.7			
	CSEQCHN65_XYK(C)	81.37	74.54	66.95			
stage 20-21	CSEQCHL01_JIU(A)	56.63	77.72	75.6			
	CSEQCHN03_PWB(B)	90.22	87.52	86.42			
	CSEQCHN04_BWQ(C)	92.01	89.78	83.47			
stage 22 heads	CSEQCHL14_RYH(A)	63.61	70.47	86.38			
	CSEQCHN23_FZD(B)	76.3	74.24	85.4			
	CSEQCHN24_JFD(C)	91.67	89.07	79.87			
stage 22 limbs	CSEQCHL13_UZO(A)	53.33	85.32	87.32			
	CSEQCHN51_GSM(B)	89.49	67.51	75.16			
	CSEQCHN52_OJD(C)	94.1	82.36	71.09			
stage 36 heads	CSEQCHL22_UCY(A)	61.98	76.71	69.86			
	CSEQCHN62_NTG(B)	83.29	78.71	71.85			
	CSEQCHN63_OTT(C)	82.53	77.43	69.22			

tissue	library name	% unique sequences when compared with lib A	lib B	lib C	lib D	lib E	lib F
	CSEQCHL12_NGJ(A)	43.61	64.47	65.87			
stage 36 hearts	CSEQCHN70_CYQ(B)	78.52	65.25	70.68			
	CSEQCHN71_AIJ(C)	81.35	71.46	68.77			
	CSEQCHL20_YDI(A)	61.07	79.02	66.48			
stage 36 limbs	CSEQCHN38_SQJ(B)	83.57	81.72	75.87			
	CSEQCHN59_BKX(C)	78.08	78.65	64.38			
	CSEQCHL21_XMQ(A)	54.77	65.5	88.96			
stage 36 trunks	CSEQCHN35_TKI(B)	72.27	68.67	85.86			
	CSEQCHN75_UDJ(C)	92.22	90.54	80.75			

Table A1.5. Inter-tissue redundancy analysis.

Results of the ***megablast*** comparisons of all ESTs from each tissue against all other individual tissues. Tissues with more than 30% of their sequences in common are highlighted in blue. Tissues with between 25->30% of their sequences in common are highlighted in yellow.

Tissue	adult_adipose	adult_heart	adult_liver	adult_pancreas	brain_cerebellum	brain_cerebrum	brain_emb	brain_other	chicken_muscle	chicken_ovary	chondrocytes	kidney_adrenal	small_intestine	stage_22_heads	stage_22_limbs	stage_36_heads	stage_36_hearts	stage_36_limbs	stage_36_trunks	stg10_emb	stg20-21
adult_adipose	100.	25.4	**23.3**	2.9	19.5	**22.8**	16.2	19.4	18.8	28.6	29.9	17.8	19.0	19.9	19.1	**22.5**	25.9	**22.3**	**24.5**	19.6	**21.4**
adult_heart	12.7	100.	17.1	1.8	15.2	15.7	14.0	15.6	18.0	29.5	28.2	15.2	13.9	14.5	15.4	17.0	**20.9**	17.3	16.5	17.6	**22.5**
adult_liver	8.2	14.1	100.	1.3	9.8	13.5	9.3	13.6	10.8	**23.0**	**22.9**	16.2	10.4	13.6	14.7	15.2	15.9	14.9	16.7	15.2	16.9
adult_pancreas	13.3	14.0	13.7	100.	19.5	11.3	16.4	18.4	11.9	14.7	18.1	13.3	14.0	15.4	11.9	17.4	19.1	19.5	17.7	17.1	16.0
brain_cerebellu	8.1	13.8	11.2	1.4	100.	28.9	13.6	18.1	13.3	19.1	17.9	11.8	13.1	11.8	11.6	13.4	12.2	12.2	11.6	12.9	**20.6**
brain_cerebrum	8.7	13.9	14.0	1.3	33.8	100.	32.9	**21.8**	11.2	**22.0**	**22.4**	13.2	12.2	**23.4**	**23.7**	28.6	**21.6**	**24.2**	**21.1**	**24.6**	18.0
brain_emb	6.3	11.1	9.2	1.4	16.6	**23.3**	100.	15.1	12.1	16.1	16.6	9.9	11.5	8.9	8.8	9.5	8.5	8.9	8.0	9.6	19.0
brain_other	7.0	11.8	12.9	1.0	13.3	13.8	12.5	100.	11.5	**24.1**	**22.6**	12.1	12.2	10.9	10.6	12.5	11.4	11.4	10.6	11.8	18.4
chicken_muscle	5.2	11.3	9.4	0.9	11.9	9.5	12.4	11.4	100.	13.2	13.1	10.7	11.2	11.5	11.0	13.4	12.3	14.1	11.4	12.8	9.8
chicken_ovary	5.8	13.3	12.6	0.9	10.2	12.7	9.5	16.4	8.7	100.	13.2	16.7	14.8	15.4	16.7	16.9	14.9	17.5	13.0	18.1	13.4
chondrocytes	5.6	11.1	12.3	0.8	9.0	11.4	8.9	13.3	8.8	10.3	100.	13.7	11.7	12.9	14.4	15.2	13.5	15.1	12.6	15.7	11.7
kidney_adrenal	6.7	11.9	16.3	1.4	9.1	9.7	8.9	10.5	11.2	25.6	**23.5**	100.	15.1	8.9	9.3	10.2	10.5	9.3	10.4	10.5	19.1
small_intestine	6.5	11.3	10.9	1.2	10.1	9.5	9.5	10.8	11.9	**24.5**	**21.5**	12.8	100.	9.8	11.1	11.4	12.2	11.6	11.1	11.5	18.5
stage_22_heads	8.0	12.7	13.7	2.2	14.9	19.8	14.1	13.4	12.5	**23.7**	**22.5**	11.9	12.2	100.	15.3	16.8	15.9	15.8	14.8	17.8	**20.9**
stage_22_limbs	6.5	11.8	13.3	1.1	13.8	19.5	12.6	12.9	11.4	**24.9**	**23.9**	10.9	12.8	15.9	100.	15.7	14.7	15.9	14.0	17.2	**20.7**
stage_36_heads	8.3	14.0	15.4	1.7	18.0	28.6	14.9	16.3	14.0	28.1	27.6	12.7	14.4	**20.2**	19.0	100.	15.3	15.1	15.2	32.6	**21.2**
stage_36_hearts	11.5	**22.2**	19.9	1.9	**21.7**	28.6	16.9	19.5	17.9	31.2	30.8	17.7	19.6	**24.2**	**23.8**	28.6	100.	28.6	28.6	38.1	28.6
stage_36_limbs	8.6	14.5	15.1	1.8	16.3	**21.7**	13.5	14.7	14.7	27.6	28.1	12.0	14.7	19.1	19.2	**20.1**	14.4	100.	19.3	33.3	**22.9**
stage_36_trunks	10.3	15.8	18.3	2.2	17.1	**22.0**	14.6	16.0	14.5	28.6	28.6	15.4	16.5	19.5	19.7	**22.5**	**20.7**	19.9	100.	28.6	**22.2**
stg10_emb	8.3	16.2	17.3	1.7	18.9	**24.6**	16.4	17.6	15.8	31.5	30.5	15.1	16.0	**23.9**	**23.9**	36.3	32.5	37.5	28.6	100.	28.6
stg20-21	7.6	14.8	14.4	1.6	18.8	14.9	19.8	18.3	10.5	19.5	19.2	17.9	17.6	20.0	20.0	20.0	18.3	**20.9**	16.7	**21.2**	100.

Table A1.6. Total number of sequences gained from round 1 and round 2 sequencing.

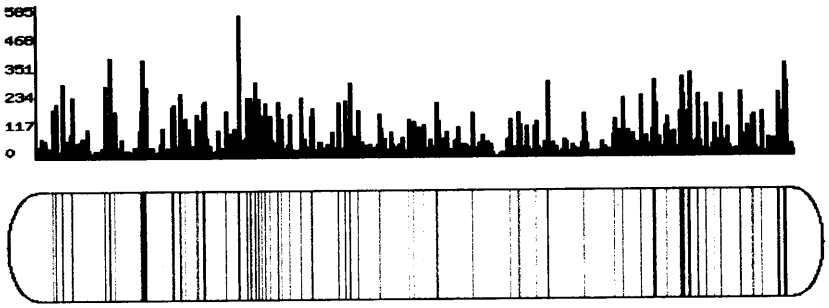
Tissue	*First phase*	*Second phase*	*Total sequencing*
stage 20-21	9941	16856	26797
stage 36 hearts	8940	0	8940
stage 36 limbs	10405	9967	20372
stage 36 trunks	8513	5338	13851
stage 36 heads	10476	5196	15672
stage 10	10162	2703	12865
stage 22 limbs	10314	5591	15905
stage 22 heads	10470	8576	19046
16 days embryo brain	11648	5453	17101
adult brain - cerebellum	10541	3948	14489
adult brain - cerebrum	8907	5265	14172
adult brain - other parts	10111	5524	15635
adult kidney + adrenal	9512	9969	19481
adult small intestine	9657	8305	17962
adult liver	8170	5358	13528
adult heart	9705	0	9705
muscle	4219	5483	9702
ovary	15410	15091	30501
chondrocytes	17632	14889	32521
adult adipose	2898	0	2898
adult pancreas	1269	6902	8171
Total	**198900**	**140414**	**339,314**

Appendix 2

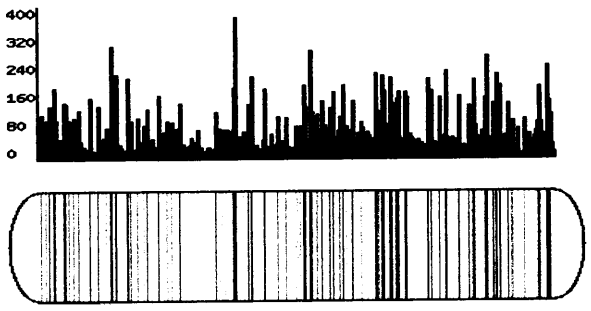
Figure A2.1. Graphical overview of the coverage of the human genome in the chicken EST database.

These 'sausage plots' show the result of a TBLASTN (protein query versus nucleotide database dynamically translated in all 6 reading frames) search of the Ensembl confirmed human protein set against the chicken consensus sequences. The data is represented in two ways: as a histogram and as a 'sausage plot' with the intensity of colour representing the number of hits. For both plots, each bar represents the total number of hits to five neighboring proteins along the chromosome. The two representations are aligned for ease of comparison.

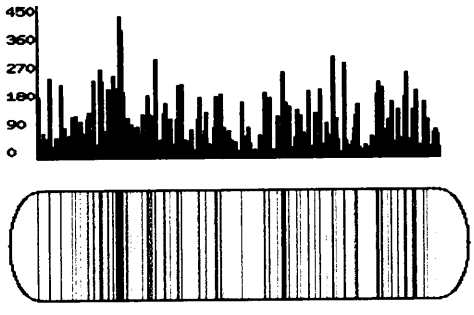
Chromosome 1



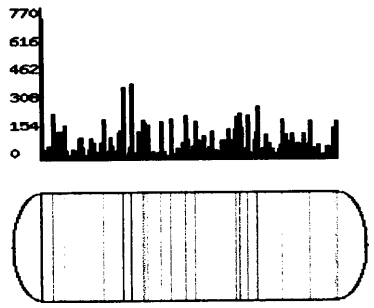
Chromosome 2



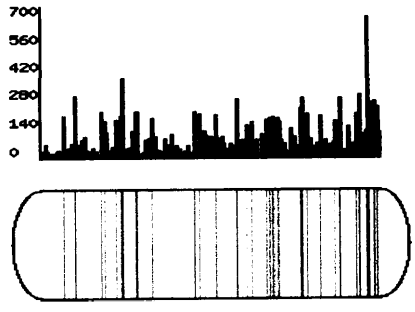
Chromosome 3



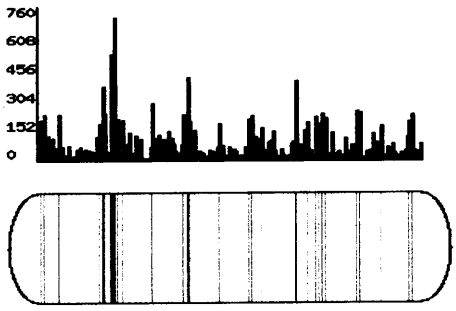
Chromosome 4



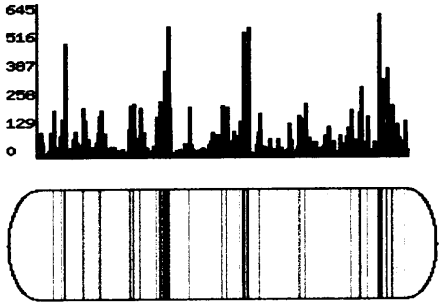
Chromosome 5



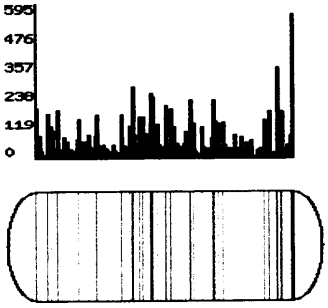
Chromosome 6



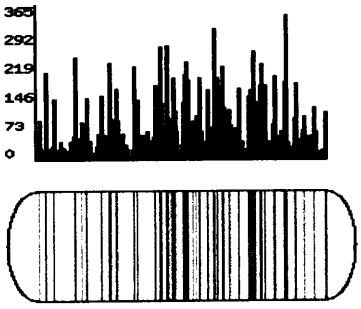
Chromosome 7



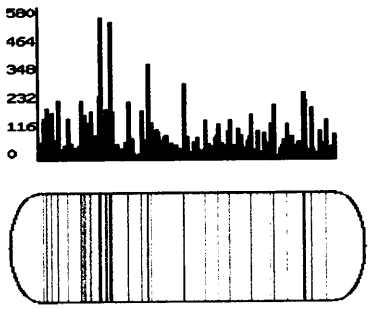
Chromosome 8



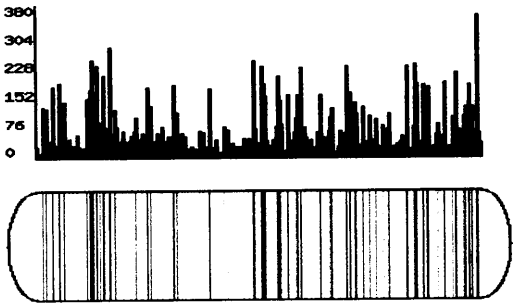
Chromosome 9



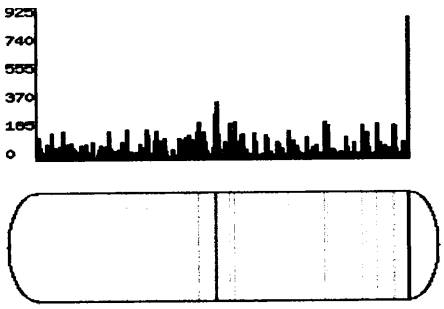
Chromosome 10



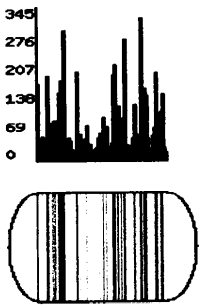
Chromosome 11



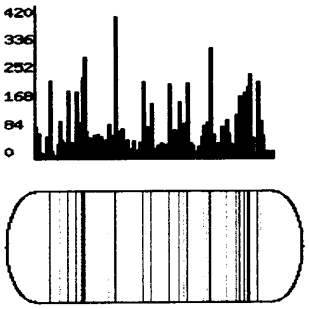
Chromosome 12



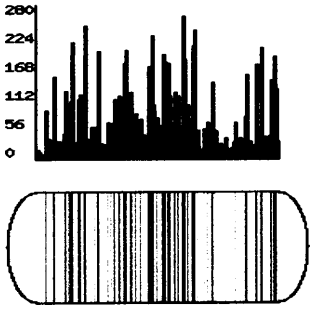
Chromosome 13



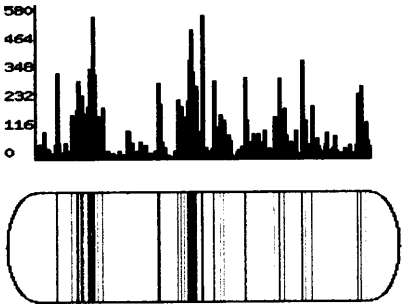
Chromosome 14



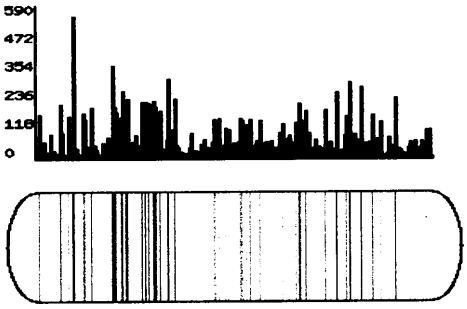
Chromosome 15



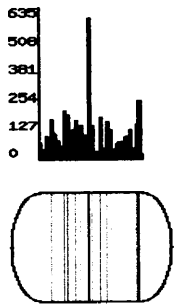
Chromosome 16



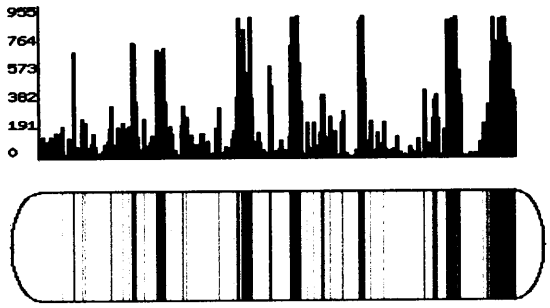
Chromosome 17



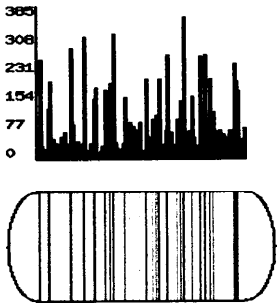
Chromosome 18



Chromosome 19

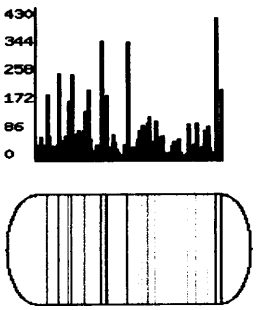


Chromosome 20

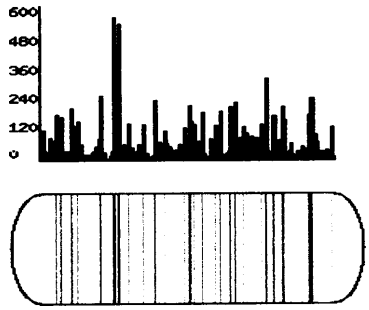


Chromosome 21

Chromosome 22



Chromosome X



Chromosome Y

Appendix 3: Simple vector analysis of yeast URSs.

Figure A3.1. ClustalW alignment of 14 URSs described in section 3.2.2.

The alignment has been trimmed. Empty sequence entries have been removed from the beginning and end of the alignment.

```
CLUSTAL W (1.82) multiple sequence alignment

YGR296W         -GGGTAAGTATGTGTGTATTATTTACGATCATTTGTTAACATTTCAATATGTTGGGTAGA
YPL283C         AGGGTAAGTATATGTGTATTATTTACGATCATTTGTTAACGTTTCAACATGGTGGGTAGA
YNL339C         -GGGTAATGGTAGTAGAGTTGGATTTGGGTAATTGGAGGGTAACGGTTATGATGGGCGGT

YGR296W         ACAACAGTATAGTGAGTAACAAGATGGGGCATGGTAGGGTAATGGCAGGGTAAGTGGTAG
YPL283C         ACAACAGTATGGTGAGTAGCG-GATGATGGATGGTAGGGTAATAGTAGGGTAAGTGGTGG
YNL339C         GGATGGTAGTAGTAAGTAGAGTGATGGATGGTGGTTGGG-AGTGGTATGGTTGAGTGGGG

YGR296W         TGGAGTTGAATATGGGCAATTGGAGGGTAACAG--GTGGTGGATGTGGGTGAGTGGTAGT
YPL283C         TGGAGTTGGATATGGGTAATTGGAGGGTAACGGTTATGATGGGCGGTGGATGGTAGTAGT
YNL339C         CAGGGTAACGAGTGGGGAGGT--AGGGTAATGG--AGGGTAAGTTGAGAGACACGTTCAT

YGR296W         AAGTAGAGAGATGGATGGTGGTTGGGGT-GTGGTATAGTTGAATGAGACAGGGTAACTTG
YPL283C         AAGTAGAGAGATGGATGGTGGTTGGGA--GTGGTATGGTTGAGTGAGACAGGGTAACGAG
YNL339C         CAGGGTTAGAATAGGGTAGGGTTAGGGTTGTGATGGGTGTGGGTGTGGGTGTGTGGGTGT

YLR462W         ------------------------------------------------TATATCTATGTCA
YGR296W         TGGGGAGGTAGGGTAATGGAGG-GTAAGTTGAGAG-ACAGGTTAAATCATATATATGTCA
YPL283C         TGGAGAGGTAGGGTAATGGAGG-GTAAGTTGAGAG-ACAGGTTCA-TCATATATATGTCA
YNL339C         GGGTGTGGTGTGGGTGTGTGGGTGTGGTGTGTGGGTGTGGGTGTGGGTATATATATGTCA

YEL075C         ----------TTTTAAGTTAATGACGCCATGGTAGTATTCATACTTCAAGTCAAAGTGTT
YBL111C         ------------------------AGACAAGGCCGTAGGGACATATAGCATCTAGGAAGT
YIL177C         -------------------------------------GGGACATATAGCATCTAGGAAGT
YJL225C         -------------------------------------GGGACATATAGCATCTAGGAAGT
YDR545W         ----------------------------------------CATATAGCATCTAGGAAGT
YLR467W         ----------------------------------------CATATAGCATCTAGGAAGT
YML133C         ---------------------------------------ACATATAGCATCTAGGAAGT
YFL064C         ------------------------GACAAGGCCGTAGGGACATATAGCATCTAGGAAGT
YER189W         -------------TGGATGGTGTTAGACAAGGCCGTAGGGACATATAGCATCTAGGAAGT
YLR462W         CCTTATTGCATGCTGGATGGTGTTAGACAAGGCCGTAGGGACATATAGCATCTAGGAAGT
YGR296W         CTGTATTGCATGCTGGATGGTGTTAGACAAGGCCGTAGGGACATATAGCATCTAGGAAGT
YPL283C         CTGTATTGCATGCTGGATGGTGTTAGACAAGGCCGTAGGGACATATAGCATCTAGGAAGT
YNL339C         CTGTATTGCATGCTGGATGGTGTTAGACAAGGCCGTAGGGACATATAGCATCTAGGAAGT

YEL075C         TATTTCGTGAAGTTGAAAGGTAGAATATTTTTATGTTTAGGTGATTTTGATGGTGATTTT
YBL111C         AACCTTGTA----CGAAAATAGGCAATATTTCCTGTTTAGGCGATTGTGACGCAGATTTT
YIL177C         AACCTTGTA----CGAAAATAGGCAATATTTCCTGTTT----------GACGCAGATTTT
YJL225C         AACCTTGTA----CGAAAATAGGCAATATTTCCTGTTT----------GACGCAGATTTT
YDR545W         AACCTTGTA----CGAAAATAGGCAATATTTCCTGTTTAGGCGATTGTGACGCAGATTTT
```

```
YLR467W         AACCTTGTA----CGAAAATAGGCAATATTTCCTGTTTAGGCGATTGTGACGCAGATTTT
YML133C         AACCTTGTA----CGAAAATAGGCAATATTTCCTGTTTAGGCGATTGTGACGCAGATTTT
YFL064C         AACCTTGTA----CGAAAATAGGCAATATTTCCTGTTTAGGCGATTGGGACGCAGATTTT
YER189W         AACCTTGTA----CGAAAATAGGCAATATTTCCTGTTT----------GACGCAGATTTT
YLR462W         AACCTTGTA----CGAAAATAGGCAATATTTCCTGTTT----------GACGCAGATTTT
YGR296W         AACCTTGTA----CGAAAATAGGCAATATTTCCTGTTTAGGCGATTGTGACGCAGATTTT
YPL283C         AACCTTGTA----CGAAAATAGGCAATATTTCCTGTTTAGGCGATTGTGACGCAGATTTT
YNL339C         AACCTTGTA----CGAAAATAGGCAATATTTCCTGTTTAGGCGATTGTGACGCAGATTTT

YEL075C         TCTGTAGTATTGACATAAGTGTATACAAGTGAAGGTGAGCATGGTGTGTGGGTGTGGGTG
YBL111C         AGTCCAACGATCTAGC------------GTCAAGGAATTTTTTTATAGTGGGACATTGCA
YIL177C         AGCCCAAAGATCTAGCG--------TTAAGGAATTTTTTTATAGTGGGACATTGCAAA--
YJL225C         AGCCCAAAGATCTAGCG--------TTAAGGAATTTTTTTATAGTGGGACATTGCAAA--
YDR545W         AGTCCAACGATCTAGCG--------TCAAGGAATTTTTTTATAGTGGGACATTGCA----
YLR467W         AGTCCAACGATCTAGCG--------TCAAGGAATTTTTTTATAGTGGGACATTGCA----
YML133C         AGTCCAACGATCTAGCG--------TCAAGGAATTTTTTTATAGTGGGACATTGCA----
YFL064C         AGTCCAACGATCTAGCG--------TCAAGGAATTTTTTTATAGTGGGACATTGCA----
YER189W         AGCCCAAAGATCTAGCG--------TTAAGGAATTTTTTTATAGTGGGACATTGCAAA--
YLR462W         AGCCCAAAGATCTAGCG--------TCAAGGAATTTTTTTATAGTGGGACATTGCAAA--
YGR296W         AGTCCAACGATCTAGCG--------TCAAGGAATTTTTTTATAGTGGGACATTGCA----
YPL283C         AGTCCAACGATCTAGCG--------TCAAGGAATTTTTTTATAGTGGGACATTGCA----
YNL339C         AGTCCAACGATCTAGCG--------TCAAGGAATTTTTTTATAGTGGGACATTGCA----

YEL075C         TGGGTGAAGTAACTTGATACGTCGTTGGTGAATGGGTCTGCTTTCTTATTCGGCGGGGTA
YHR218W         ----------------------------------------------------GGCGGGGTA
YBL111C         CCAAGGAAGTAACTTGATACGTCGTGGGTGAATGGGTCTGTTTTCTTATTCGGCGGGGTA
YIL177C         TCAAGGAAGTAACTTGATACGTCGTTGGTGAATGGGTCTGTTTTCTTATTCGGCGGGGTA
YJL225C         TCAAGGAAGTAACTTGATACGTCGTTGGTGAATGGGTCTGTTTTCTTATTCGGCGGGGTA
YDR545W         CCAAGGAAGTAACTTGATACGTCGTGGGTGAATGGGTCTGTTTTCTTATTCGGCGGGGTA
YLR467W         CCAAGGAAGTAACTTGATACGTCGTGGGTGAATGGGTCTGTTTTCTTATTCGGCGGGGTA
YML133C         CCAAGGAAGTAACTTGATACGTCGTGGGTGAATGGGTCTGTTTTCTTATTCGGCAGGGTA
YFL064C         CCAAGGAAGTAACTTGATACGTCGTGGGTGAATGGGTCTGTTTTCTTATTCGGCGGGGTA
YER189W         CCAAGGAAGTAACTTGATACGTCGTTGGTGAATGGGTCTGTTTTCTTATTCGGCGGGGTA
YLR462W         CCAAGGAAGTAACTTGATACGTCGTTGGTGAATGGGTCTGTTTTCTTATTCGGCGGGGTA
YGR296W         CCAAGGAAGTAACTTGATACGTCGTGGGTGAATGGGTCTGTTTTCTTATTCGGCGGGGTA
YPL283C         CCAAGGAAGTAACTTGATACGTCGTGGGTGAATGGGTCTGTTTTCTTATTCGGCGGGGTA
YNL339C         CCAAGGAAGTAACTTGATACGTCGTGGGTGAATGGGTCTGTTTTCTTATTCGGCGGGGTA
                                                                *** *****

YEL075C         ATACATTTT-GGGAGAAGTTTGTCTTTCTGACGCGCCATATGTAGGTACGCCAAAAAGGG
YHR218W         ATACATTTTTGGGGGAAGTTTGTCTGTCTGACGCGCCATATGTAGGTACGCCAAAAAGGG
YBL111C         ATACATTTTTGGGGGAAGTTTGTCTGTCTGACGCGCCATATGTAGGTACGCCAAAAAGGG
YIL177C         ATACATTTT-GAGGGAAGGTTGTCTGTCTGACGCGCCATATGTAGGTACGCCAAAAAGGG
YJL225C         ATACATTTT-GAGGGAAGGTTGTCTGTCTGACGCGCCATATGTAGGTACGCCAAAAAGGG
YDR545W         ATACATTTTTGGGGGAAGTTTGTCTGTCTGACGCGCCATATGTAGGTACGCCAAAAAGGG
YLR467W         ATACATTTTTGGGGGAAGTTTGTCTGTCTGACGCGCCATATGTAGGTACGCCAAAAAGGG
YML133C         ATACATTTTTGGGGGAAGTTTGTCTGTCTGACGCGCCATATGTAGGTACGCCAAAAAGGG
YFL064C         ATACATTTTTGGGGGAAGTTTGTCTGTCTGACGCGCCATATGTAGGTACGCCAAAAAGGG
YER189W         ATACATTTT-GAGGGAAGGTTGTCTGTCTGACGCGCCATATGTAGGTACGCCAAAAAGGG
YLR462W         ATACATTTTTGGGGGAAGTTTGTCTGTCTGACGCGCCATATGTAGGTACGCCAAAAAGGG
```

```
YGR296W         ATACATTTTTGGGGGAAGTTTGTCTGTCTGACGCGCCATATGTAGGTACGCCAAAAAGGG
YPL283C         ATACATTTTTGGGGGAAGTTTGTCTGTCTGACGCGCCATATGTAGGTACGCCAAAAAGGG
YNL339C         ATACATTTTTGGGGGAAGTTTGTCTGTCTGACGCGCCATATGTAGGTACGCCAAAAAGGG
                ********* * * **** ****** **********************************

YEL075C         CTCCTTTACTTCGAAGCGCGAGGTCGTATACCTAATAAGGAAATGTAATTTATAACTTTT
YHR218W         CTCCTCTACTTCGAAGCGCGAGGTCGTATACCTAATAAGGAAATGTAATTTATAACTTTT
YBL111C         CTCCTCTACTTCGAAGCGCGAGGTCGTATACCTAATAAGGAAATGTAATTTATAACTTTT
YIL177C         CTCCTTTACTTCGAAGCGCGAGGTCGTATACCTAATAAGGAAATGTAATTTATAACTTTT
YJL225C         CTCCTTTACTTCGAAGCGCGAGGTCGTATACCTAATAAGGAAATGTAATTTATAACTTTT
YDR545W         CTCCTCTACTTCGAAGCGCGAGGTCGTATACCTAATAAGGAAATGTAATTTATAACTTTC
YLR467W         CTCCTCTACTTCGAAGCGCGAGGTCGTATACCTAATAAGGAAATGTAATTTATAACTTTC
YML133C         CTCCTCTACTTCGAAGCGCGAGGTCGTATACCTAATAAGGAAATGTAATTTATAACTTTC
YFL064C         CTCCTCTACTTCGAAGCGCGAGGTCGTATACCTAATAAGGAAATGTAATTTATAACTTTC
YER189W         CTCCTTTACTTCGAAGCGCGAGGTCGTATACCTAATAAGGAAATGTAATTTATAACTTTT
YLR462W         CTCCTCTACTTCGAAGCGCGAGGTCGTATACCTAATAAGGAAATGTAATTTATAACTTTT
YGR296W         CTCCTCTACTTCGAAGCGCGAGGTCGTATACCTAATAAGGAAATGTAATTTATAACTTTC
YPL283C         CTCCTCTACTTCGAAGCGCGAGGTCGTATACCTAATAAGGAAATGTAATTTATAACTTTC
YNL339C         CTCCTCTACTTCGAAGCGCGAGGTCGTATACCTAATAAGGAAATGTAATTTATAACTTTC
                ***** ****************************************************

YEL075C         TATTATATTGGTCTTTTCGAGAGCGGAA---------------------------CGTAG
YHR218W         TATTATATTGGTCTTTTCGAGAGCGGAA---------------------------CGTAG
YBL111C         TATTATATTGGTCTTTTCGAGAGCGGAA---------------------------CGTAG
YIL177C         TATTATATTGGTCTTTTCGATAGCGGAAGAAGTTGTAGGCTAAGCGCAGGCTAAGCGTAG
YJL225C         TATTATATTGGTCTTTTCGATAGCGGAAGAAGTTGTAGGCTAAGCGCAGGCTAAGCGTAG
YDR545W         TATTATATTGGTCTTTTCGAGAGCGGAAGAAGTTGTAGGCTAAGCGCAGGCTAAGCGTAG
YLR467W         TATTATATTGGTCTTTTCGAGAGCGGAAGAAGTTGTAGGCTAAGCGCAGGCTAAGCGTAG
YML133C         TATTATATTGGTCTTTTCGAGAGCGGAAGAAGTTGTAGGCTAAGCGCAGGCTAAGCGTAG
YFL064C         TATTATATTGGTCTTTTCGAGAGCGGAAGAAGTTGTAGGCTAAGCGCAGGCTAAGCGTAG
YER189W         TATTATATTGGTCTTTTCGAGAGCGGAAGAAGTTGTAGGCTAAGCGCAGGCTAAGCGTAG
YLR462W         TATTATATTGGTCTTTTCGAGAGCGGAA---------------------------CGTAG
YGR296W         TATTATATTGGTCTTTTCGAGAGCGGAAGAAGTTGTAGGCTAAGCGCAGGCTAAGCGTAG
YPL283C         TATTATATTGGTCTTTTCGAGAGCGGAAGAAGTTGTAGGCTAAGCGCAGGCTAAGCGTAG
YNL339C         TATTATATTGGTCTTTTCGAGAGCGGAAGAAGTTGTAGGCTAAGCGCAGGCTAAGCGTAG
                ******************** *******                          *****

YEL075C         GTCCATGTTTAAAGTATCCAACAGAATATCCACGAAGCGGCTGAGCAACGAACAGAATCC
YHR218W         GTCCATGTTTAAAGTATCCAAGAGAATATCCACGAAGCGGCTGAGCAACGAACAGAATCC
YBL111C         GTCCATGTTTAAAGTATCCAAGAGAATATCCACGAAGCGGCTGAGCAACGAACAGAATCC
YIL177C         GTCCATGTTTAAAGTATCCAAGAGAATATCCACGAAGCGGCTGAGCAACGAACAGAATCC
YJL225C         GTCCATGTTTAAAGTATCCAAGAGAATATCCACGAAGCGGCTGAGCAACGAACAGAATCC
YDR545W         GTCCATATTTAAAGTATCCAAGAGAATATCCACGAAGCGGCTGAGCAACGAACAGAATCC
YLR467W         GTCCATATTTAAAGTATCCAAGAGAATATCCACGAAGCGGCTGAGCAACGAACAGAATCC
YML133C         GTCCATATTTAAAGTATCCAAGAGAATATCCACGAAGCGGCTGAGCAACGAACAGAATCC
YFL064C         GTCCATGTTTGAAGTATCCAAGAGAATATCCACGAA---------------------TCC
YER189W         GTCCATGTTTAAAGTATCCAAGAGAATATCCACGAAGCGGCTGAGCAACGAACAGAATCC
YLR462W         GTCCATGTTTAAAGTATCCAAGAGAATATCCACGAAGCGGCTGAGCAACGAACAGAATCC
YGR296W         GTCCATATTTAAAGTATCCAAGAGAATATCCACGAAGCGGCTGAGCAACGAACAGAATCC
YPL283C         GTCCATATTTAAAGTATCCAAGAGAATATCCACGAAGCGGCTGAGCAACGAACAGAATCC
```

```
YNL339C         GTCCATATTTAAAGTATCCAAGAGAATATCCACGAAGCGGCTGAGCAACGAACAGAATCC
                ****** *** ********** **************                    ***

YEL075C         TGGTTCTCCTCGACTAAGCAGATAGTTAAGATACTGTGCACCATGGAAATTGAAAACGAA
YHR218W         TGGTTCTCCTCGACTAAGCAGATAGTTAAGATACTGTGCACCATGGAAATTGAAAACGAA
YBL111C         TGGTTCTCCTCGACTAAGCAGATAGTTAAGATACTGTGCACCATGGAAATTGAAAACGAA
YIL177C         TGGTTCTCCTCGACTAAGCAGATAGTTAAGATACTGTGCACCATGGAAATTGAAAACGAA
YJL225C         TGGTTCTCCTCGACTAAGCAGATAGTTAAGATACTGTGCACCATGGAAATTGAAAACGAA
YDR545W         TGGTTCTCCTCGACTAAGCAGATAGTTAAGATACTGTGCACCATGGAAATTGAAAACGAA
YLR467W         TGGTTCTCCTCGACTAAGCAGATAGTTAAGATACTGTGCACCATGGAAATTGAAAACGAA
YML133C         TGGTTCTCCTCGACTAAGCAGATAGTTAAGATACTGTGCACCATGGAAATTGAAAACGAA
YFL064C         TGGTTCTCCTCGACTAAGCAGATAGTTAAGATACTGTGCACCATGGAAATTGAAAACGAA
YER189W         TGGTTCTCCTCGACTAAGCAGATAGTTAAGATACTGTGCACCATGGAAATTGAAAACGAA
YLR462W         TGGTTCTCCTCGACTAAGCAGATAGTTAAGATACTGTGCACCATGGAAATTGAAAACGAA
YGR296W         TGGTTCTCCTCGACTAAGCAGATAGTTAAGATACTGTGCACC------------------
YPL283C         TGGTTCTCCTCGACTAAGCAGATAGTTAAGATACTGTGCACC------------------
YNL339C         TGGTTCTCCTCGACTAAGCAGATAGTTAAGATACTGTGCACC------------------
                ****************************************

YEL075C         CGTACGTACCGACTACTTTATTTTTGCAGGCCGGAAATCAAGCGATGAATGAGACATCCT
YHR218W         AGTACGTACCGACTACTTTATTTTTGCAGGCCGGAAATCAAGCGATGAATGAGACATCCT
YBL111C         AGTACGTACCGACTACTTTATTTTTGCAGGCCGGAAATCAAGCGATGAATGAGACATCCT
YIL177C         CGTACGTACCGACTACTTTATTTTTGCAGGCCGGAAATCAAGCGATGAATGAGACATCCT
YJL225C         CGTACGTACCGACTACTTTATTTTTGCAGGCCGGAAATCAAGCGATGAATGAGACATCCT
YDR545W         AGTACGTACCGACTACTTTATTTTTGCAGGCCGGAAATCAAGCGATGAATGAGACATCCT
YLR467W         AGTACGTACCGACTACTTTATTTTTGCAGGCCGGAAATCAAGCGATGAATGAGACATCCT
YML133C         CGTACGTACCGACTACTTTATTTTTGCAGGCCGGAAATCAAG-GATGAATGAGACATCCT
YFL064C         AGTGCGTACCGACTACTTTATTTTTGCAGGCCGGAAATCAAGCGATGAATGAGACATCCT
YER189W         CGTACGTACCGACTACTTTATTTTTGCAGGCCGGAAATCAAGCGATGAATGAGACATCCT
YLR462W         AGTACGTACCGACTACTTTATTTTTGCAGGCCGGAAATCAAGCGATGAATGAGACATCCT

YEL075C         TCTGTTTTCTATGTTGTGCTTGAAGGGGACAGACAGTCGCTTATCTTAGTGAGATTT---
YHR218W         TCTGTTTTCTATGTTGTGCTTGAAGGGGACAGACAGTCGCTTATCTTAGTGAGATTTCTT
YBL111C         TCTGTTTTCTATGTTGTGCTTGAAGGGGACAGACAGTCGCTTATCTTAGTGAGATTTCTT
YIL177C         TCTGTTTTCTATGTTGTGCTTGAAGGGGACAGACAGTCGCTTATCTTAGTGAGATTTCTT
YJL225C         TCTGTTTTCTATGTTGTGCTTGAAGGGGACAGACAGTCGCTTATCTTAGTGAGATTTCTT
YDR545W         TCTGTTTTCTATGTTG----------GGACAGACAGTCGCTTATCTTAGTGAGATTTCTT
YLR467W         TCTGTTTTCTATGTTG----------GGACAGACAGTCGCTTATCTTAGTGAGATTTCTT
YML133C         TCTGTTTTCTATGTTG----------GGACAGACAGTCGCTTATCTTAGTGAGATTTCTT
YFL064C         TCTGTTTTCTATGTTGTGCTTGAAGGGGACAGACAGTCGCTTATCTTAGTGAGATTTCTT
YER189W         TCTGTTTTCTATGTTGTGCTTGAAGGGGACAGACAGTCGCTTATCTTAGTGAGATTT---
YLR462W         TCTGTTTTCTATGTTGTGCTTGAAGGGGACAGACAGTCGCTTATCTTAGTGAGATTT---

YEL075C         -------------TGTTTGCTTTTGCTG--------CACCTGCATAGCGCAGATTCTGCA
YHR218W         ATTAACTGAATTTTCTTTGCTGCTGCTGGAGATTTGCACCTGCATAGCGCAGATTCTGCT
YBL111C         ATTAACTGAATTTTCTTTGCTGCTGCTGGAGATTTGCACCTGCATAGCGCAGATTCTGCT
YIL177C         ATTAACTGAATTTTCTTTGCTGCTGCTAGAGATTTGCACCTGCATAGCGCAGATTCTGCA
YJL225C         ATTAACTGAATTTTCTTTGCTGCTGCTAGAGATTTGCACCTGCATAGCGCAGATTCTGCA
YDR545W         ATTAACTGAATTTTCTTTGCTGCTGCTGGAGATTTGCACCTGCATAGCGCAGATTCTGCT
YLR467W         ATTAACTGAATTTTCTTTGCTGCTGCTGGAGATTTGCACCTGCATAGCGCAGATTCTGCT
YML133C         ATTAACTGAATTTTCTTTGCTGCTGCTGGAGATTTGCACCTGCATAGCGCAGATTCTGCT
```

```
YFL064C         ATTAACTGAATTTTCTTTGCTGCTGCTGGAGATTTGCACCTGCATAGCGCAGATTCTGCA
YER189W         -------------TGTTTGCTTTTGCTG--------CACCTGCATAGCGCAGATTCTGCA
YLR462W         -------------TGTTTGCTTTTGCTG--------CACCTGCATAGCGCAGATTCTGCA

YEL075C         TCTTCTCAATAG-----CTTAATTATTACATTCTTAGATGATGATAAGACGGAAACTGGA
YHR218W         TCTTCTCAATAGAGTAGCTTAATTATTACATTCTTAGATGATGATAAGACGGAAACTGGA
YBL111C         TCTTCTCAATAGAGTAGCTTAATTATTACATTCTTAGATGATGATAAGACGGAAACTGGA
YIL177C         TCTTCTCAA-----TAGCTTAATTATTACATTCTTAGATGATGATAAGACGGAAACTGGA
YJL225C         TCTTCTCAA-----TAGCTTAATTATTACATTCTTAGATGATGATAAGACGGAAACTGGA
YDR545W         TCTTCTCAATAGAGTAGCTTAATTATTACATTCTTAGATGATGATAAGACGGAAACTGGA
YLR467W         TCTTCTCAATAGAGTAGCTTAATTATTACATTCTTAGATGATGATAAGACGGAAACTGGA
YML133C         TCTTCTCAATAGAGTAGCTTAATTATTACATTCTTAGATGATGATAAGACGGAAACTGGA
YFL064C         TCTTCTCAA-----TAGCTTAATTATTACATTCTTAGATGATGATAAGACGGAAACTGGA
YER189W         TCTTCTCAA-----TAGCTTAATTATTACATTCTTAGATGATGATAAGACGGAAACTGGA
YLR462W         TCTTCTCAA-----TAGCTTAATTATTACATTCTTAGATGATGATAAGACGAAAACTGGA

YEL075C         CAATCTTTTGTTTATATTGATGGATTTCTTGTCAAAAAGCATAACAATCAACATACTATT
YHR218W         CAATCTTTTGTTTATATTGATGGATTTCTTGTCAAAAAGCATAGCAATCAACATACTATT
YBL111C         CAATCTTTTGTTTATATTGATGGATTTCTTGTCAAAAAGCATAGCAATCAACATACTATT
YIL177C         CAATCTTTTGTTTATATTGATGGATTTCTTGTCAAAAAGCATAACAATCAACATACTATT
YJL225C         CAATCTTTTGTTTATATTGATGGATTTCTTGTCAAAAAGCATAACAATCAACATACTATT
YDR545W         CAATCTTTTGTTTATATTGATGGATTTCTTGTCAAAAAGCATAACAATCAACATACTATT
YLR467W         CAATCTTTTGTTTATATTGATGGATTTCTTGTCAAAAAGCATAACAATCAACATACTATT
YML133C         CAATCTTTTGTTTATATTGATGGATTTCTTGTCAAAAAGCATAACAATCAACATACTATT
YFL064C         CAATCTTTTGTTTATATTGATGGATTTCTTGTCAAAAAGCATAACAATCAACATACTATT
YER189W         GAATCTTTTGTTTATATTGATGGATTTCTTGTCAAAAAGCATAACAATCAACATACTATT
YLR462W         CAATCTTTTGTTTATATTGATGGATTTCTTGTCAAAAAGCATAAAAATCAACATACTATT

YEL075C         GTTAATTTCGAAACTTACAAAAATAAA--------------------------------
YHR218W         GTTAATTTCGAAACTTACAAAAATAAAATGAAAGTTTCCGATAGCGTAAGTTTGAAAAAG
YBL111C         GTTAATTTCGAAACTTACAAAAATAAA--------------------------------
YIL177C         GTTAATTTCGAAACTTACAAAAATAAA--------------------------------
YJL225C         GTTAATTTCGAAACTTACAAAAATAAA--------------------------------
YDR545W         GTTAATTTCGAAACTTACAAAAATAAA--------------------------------
YLR467W         GTTAATTTCGAAACTTACAAAAATAAA--------------------------------
YML133C         GTTAATTTCGAAACTTACAAAAATAAA--------------------------------
YFL064C         GTTAATTTCGAAACTTACAAAAATAAA--------------------------------
YER189W         GTTAATTTCGAAACTTACAAAAATAAA--------------------------------
YLR462W         GTTAATTTCGAAACTTACAAAAATAAA--------------------------------
```

Figure A3.2. ClustalW alignment of the associated protein sequences (translated ORFs) of the 14 URSs described in section 3.2.2.

The alignment has been trimmed. Empty sequence entries have been removed from the beginning and end of the alignment.

```
CLUSTAL W (1.82) multiple sequence alignment

YNL339C         MEIENEQICTCIAQILHLLNSLIITFLDDDKTETGQSFVYIDGFLVKKHNNQHTIVNFET
YGR296W         MEIENEQICTCIAQILHLLNSLIITFLDDDKTETGQSFVYIDGFLVKKHNNQHTIVNFET
YPL283C         MEIENEQICTCIAQILHLLNSLIITFLDDDKTETGQSFVYIDGFLVKKHNNQHTIVNFET

YBL111C         ----MKVSDRRKFEKANFDEFESALNNKNDLVHCPSITLFESIPTEVRSFYEDEKSGLIK
YML133C         ----MKVSDRRKFEKANFDEFESALNNKNDLVHCPSITLFESIPTEVRSFYEDEKSGLIK
YNL339C         YKNKMKVSDRRKFEKANFDEFESALNNKNDLVHCPSITLFESIPTEVRSFYEDEKSGLIK
YGR296W         YKNKMKVSDRRKFEKANFDEFESALNNKNDLVHCPSITLFESIPTEVRSFYEDEKSGLIK
YPL283C         YKNKMKVSDRRKFEKANFDEFESALNNKNDLVHCPSITLFESIPTEVRSFYEDEKSGLIK
YLR467W         ----MKVSDRRKFEKANFDEFESALNNKNDLVHCPSITLFESIPTEVRSFYEDEKSGLIK
YDR545W         ----MKVSDRRKFEKANFDEFESALNNKNDLVHCPSITLFESIPTEVRSFYEDEKSGLIK
YIL177C         ----MKVSDRRKFEKANFDEFESALNNKNDLVHCPSITLFESIPTEVRSFYEDEKSGLIK
YJL225C         ----MKVSDRRKFEKANFDEFESALNNKNDLVHCPSITLFESIPTEVRSFYEDEKSGLIK
YLR462W         ----MKVSDRRKFEKANFDEFESALNNKNDLVHCPSITLFESIPTEVRSFYEDEKSGLIK
YEL075C         ----MKVSDRRKFEKANFDEFESALNNKNDLVHCPSITLFESIPTEVRSFYEDEKSGLIK
YER189W         ----MKVSDRRKFEKANFDEFESALNNKNDLVHCPSITLFESIPTEVRSFYEDEKSGLIK
YFL064C         ----MKVSDRRKFEKANFDEFESALNNKNDLVHCPSITLFESIPTEVRSFYEDEKSGLIK

YBL111C         VVKFRTGAMDRKRSFEKIVVSVMVGKNVQKFLTFVEDEPDFQGGPIPSKYLIPKKINLMV
YHR218W         --------MDRKRSFEKIVVSVMVGKNVQKFLTFVEDEPDFQGGPIPSKYLIPKKINLMV
YML133C         VVKFRTGAMDRKRSFEKIVVSVMVGKNVQKFLTFVEDEPDFQGGPIPSKYLIPKKINLMV
YNL339C         VVKFRTGAMDRKRSFEKIVVSVMVGKNVQKFLTFVEDEPDFQGGPIPSKYLIPKKINLMV
YGR296W         VVKFRTGAMDRKRSFEKIVVSVMVGKNVQKFLTFVEDEPDFQGGPIPSKYLIPKKINLMV
YPL283C         VVKFRTGAMDRKRSFEKIVVSVMVGKNVQKFLTFVEDEPDFQGGPIPSKYLIPKKINLMV
YLR467W         VVKFRTGAMDRKRSFEKIVISVMVGKNVQKFLTFVEDEPDFQGGPIPSKYLIPKKINLMV
YDR545W         VVKFRTGAMDRKRSFEKIVISVMVGKNVQKFLTFVEDEPDFQGGPIPSKYLIPKKINLMV
YIL177C         VVKFRTGAMDRKRSFEKVVISVMVGKNVKKFLTFVEDEPDFQGGPIPSKYLIPKKINLMV
YJL225C         VVKFRTGAMDRKRSFEKVVISVMVGKNVKKFLTFVEDEPDFQGGPIPSKYLIPKKINLMV
YLR462W         VVKFRTGAMDRKRSFEKVVISVMVGKNVKKFLTFVEDEPDFQGGPIPSKYLVPKKINLMV
YEL075C         VVKFRTGAMDRKRSFEKVVISVMVGKNVKKFLTFVEDEPDFQGGPIPSN-----------
YER189W         VVKFRTGAMDRKRSFEKIVISVMVGKNVQKFLTFVEDEPDFQGGPIPSN-----------
YFL064C         VVKFRTGAMNRKRSFEKIVISVMVGKNVQKFLTFVEDEPDFQGGPIPSKYLIPKKINLMV
                        *:********:*:********:********************:

YBL111C         YTLFQVHTLKFNRKDYDTLSLFYLNRGYYNELSFR---VLERCYEIASARPNDSSTMRTF
YHR218W         YTLFQVHTLKFNRKDYDTLSLFYLNRGYYNELSFR---VLERCYEIASARPNDSSTMRTF
YML133C         YTLFQVHTLKFNRKDYDTLSLFYLNRGYYNELSFR---VLERCHEIASARPNDSSTMRTF
YNL339C         YTLFQVHTLKFNRKDYDTLSLFYLNRGYYNELSFR---VLERCYEIASARPNDSSTMRTF
YGR296W         YTLFQVHTLKFNRKDYDTLSLFYLNRGYYNELSFR---VLERCYEIASARPNDSSTMRTF
YPL283C         YTLFQVHTLKFNRKDYDTLSLFYLNRGYYNELSFR---VLERCYEIASARPNDSSTMRTF
YLR467W         YTLFQVHTLKFNRKDYDTLSLFYLNRGYYNELSFR---VLERCHEIASARPNDSSTMRTF
YDR545W         YTLFQVHTLKFNRKDYDTLSLFYLNRGYYNELSFR---VLERCHEIASARPNDSSTMRTF
```

```
YIL177C         YTLFQVHTLKFNRKDYDTLSLFYLNRGYYNELSFR---VLERCHEIASARPNDSSTMRTF
YJL225C         YTLFQVHTLKFNRKDYDTLSLFYLNRGYYNELSFR---VLERCHEIASARPNDSSTMRTF
YLR462W         YTLFQVHTLKFNRKDYDTLSLFYLNRGYYNELSFR---VLERCHEIASARPNDSSTMRTF
YEL075C         -------------KPRD--GLHVVSSAYF---------------EIQ-------------
YER189W         -------------KPRD--GLHVVSSAYF---------------EIQ-------------
YFL064C         YTLFQVHTLKFNRKDYDTLSLFYLNRGYYNELSFPCPGTLSRNSECQAERQLYDAYFH--
                             *  *  .*. :. .*:               *

YBL111C         TDFVSGTPIVRSLQKSTIRKYGYNLAPYMFLLLHVDELSIFSAYQASLPGEKKVDTERLK
YHR218W         TDFVSGTPIVRSLQKSTIRKYGYNLAPYMFLLLHVDELSIFSAYQASLPGEKKVDTERLK
YML133C         TDFVSGAPIVRSLQKSTIRRYGYNLAPHMFLLLHVDELSIFSAYQASLPGEKKVDTERLK
YNL339C         TDFVSGTPIVRGLQKSTIRKYGYNLAPYMFLLLHVDELSIFSAYQASLPGEKKVDTERLK
YGR296W         TDFVSGTPIVRGLQKSTIRKYGYNLAPYMFLLLHVDELSIFSAYQASLPGEKKVDTERLK
YPL283C         TDFVSGTPIVRGLQKSTIRKYGYNLAPYMFLLLHVDELSIFSAYQASLPGEKKVDTERLK
YLR467W         TDFVSGAPIVRSLQKSTIRKYGYNLAPYMFLLLHVDELSIFSAYQASLPGEKKVDTERLK
YDR545W         TDFVSGAPIVRSLQKSTIRKYGYNLAPYMFLLLHVDELSIFSAYQASLPGEKKVDTERLK
YIL177C         TDFVSGAPIVRSLQKSTIRKYGYNLAPYMFLLLHVDELSIFSAYQASLPGEKKVDTERLK
YJL225C         TDFVSGAPIVRSLQKSTIRKYGYNLAPYMFLLLHVDELSIFSAYQASLPGEKKVDTERLK
YLR462W         TDFVSGAPIVRSLQKSTIRKYGYNLAAYT-------------------------------

YBL111C         RDLCPRKPTEIKYFSQICNDMMNKKDRLGDVLATAQR-----------------------
YHR218W         RDLCPRKPTEIKYFSQICNDMMNKKDRLGDVLATAQR-----------------------
YML133C         RDLCPRKPIEIKYFSQICNDMMNKKDRLGDVLATAQR-----------------------
YNL339C         RDLCPRKPTEIKYFSQICNDMMNKKDRLGDILHIILRACALNFGAGPRGGAGDEEDRSIT
YGR296W         RDLCPRKPTEIKYFSQICNDMMNKKDRLGDILHIILRACALNFGAGPRGGAGDEEDRSIT
YPL283C         RDLCPRKPTEIKYFSQICNDMMNKKDRLGDILHIILRACALNFGAGPRGGAGDEEDRSIT
YLR467W         RDLCPRKPIEIKYFSQICNDMMNKKDRLGDILHIILRACALNFGAGPRGGAGDEEDRSIT
YDR545W         RDLCPRKPIEIKYFSQICNDMMNKKDRLGDILHIILRACALNFGAGPRGGAGDEEDRSIT
YIL177C         RDLCPRKPIEIKYFSQICNDMMNKKDRLGDILHIILRACALNFGAGPRGGAGDEEDRSIT
YJL225C         RDLCPRKPIEIKYFSQICNDMMNKKDRLGDILHIILRACALNFGAGPRGGAGDEEDRSIT

YBL111C         ------------------------------------------------------------
YHR218W         ------------------------------------------------------------
YML133C         ------------------------------------------------------------
YNL339C         NEEPIIPSVDEHGLKVCKLRSPNTPRRLRKTLDAVKALLVSSCACTARDLDIFDDNNGVA
YGR296W         NEEPIIPSVDEHGLKVCKLRSPNTPRRLRKTLDAVKALLVSSCACTARDLDIFDDNNGVA
YPL283C         NEEPIIPSVDEHGLKVCKLRSPNTPRRLRKTLDAVKALLVSSCACTARDLDIFDDNNGVA
YLR467W         NEEPIIPSVDEHGLKVCKLRSPNTPRRLRKTLDAVKALLVSSCACTARDLDIFDDTNGVA
YDR545W         NEEPIIPSVDEHGLKVCKLRSPNTPRRLRKTLDAVKALLVSSCACTARDLDIFDDTNGVA
YIL177C         NEEPIIPSVDEHGLKVCKLRSPNTPRRLRKTLDAVKALLVSSCACTARDLDIFDDNNGVA
YJL225C         NEEPIIPSVDEHGLKVCKLRSPNTPRRLRKTLDAVKALLVSSCACTARDLDIFDDNNGVA

YBL111C         ------------------------------------------------------------
YHR218W         ------------------------------------------------------------
YML133C         ------------------------------------------------------------
YNL339C         MWKWIKILYHEVAQETALKDSYRITLVPSSDGVSVCGKLFNREYVRGFYFACKAQFDNLW
YGR296W         MWKWIKILYHEVAQETALKDSYRITLVPSSDGVSVCGKLFNREYVRGFYFACKAQFDNLW
YPL283C         MWKWIKILYHEVAQETALKDSYRITLVPSSDGVSVCGKLFNREYVRGFYFACKAQFDNLW
YLR467W         MWKWIKILYHEVAQETTLKDSYRITLVPSSDGISVCGKLFNREYVRGFYFACKAQFDNLW
YDR545W         MWKWIKILYHEVAQETTLKDSYRITLVPSSDGISVCGKLFNREYVRGFYFACKAQFDNLW
YIL177C         MWKWIKILYHEVAQETTLKDSYRITLVPSSDGIS-------------------------
```

```
YJL225C         MWKWIKILYHEVAQETTLKDSYRITLVPSSDGIS--------------------------

YBL111C         -------------------IRRR---------------------------------------
YHR218W         -------------------IRRR---------------------------------------
YML133C         -------------------IRRR---------------------------------------
YNL339C         EELNDCFYMPTVVDIASLILRNREVLFREPKRGIDEYLENDSFLQMIPVKYREIVLPKLR
YGR296W         EELNDCFYMPTVVDIASLILRNREVLFREPKRGIDEYLENDSFLQMIPVKYREIVLPKLR
YPL283C         EELNDCFYMPTVVDIASLILRNREVLFREPKRGIDEYLENDSFLQMIPVKYREIVLPKLR
YLR467W         GELNNCFYMPTVVDIASLILRNREVLFREPKRGIDEYLENDSFLQMIPVKYREIVLPKLR
YDR545W         GELNNCFYMPTVVDIASLILRNREVLFREPKRGIDEYLENDSFLQMIPVKYREIVLPKLR
YIL177C         ----LLAFAGPQRNVYVDDTTRR-----------------------------------
YJL225C         ----LLAFAGPQRNVYVDDTTRR-----------------------------------

YBL111C         ------------------------------------------------------------
YHR218W         ------------------------------------------------------------
YML133C         ------------------------------------------------------------
YNL339C         RDTNKMTAALKNKVTVAIDELTVPLMWMIHFAVGYPYRYPELQLLAFAGPQRNVYVDDTT
YGR296W         RDTNKMTAALKNKVTVAIDELTVPLMWMIHFAVGYPYRYPELQLLAFAGPQRNVYVDDTT
YPL283C         RDTNKMTAALKNKVTVAIDELTVPLMWMIHFAVGYPYRYPELQLLAFAGPQRNVYVDDTT
YLR467W         RDTNKMTAALKNKVTVAIDELTVPLMWMVHFAVGYPYRYPELQLLAFAGPQRNVYVDDTT
YDR545W         RDTNKMTAALKNKVTVAIDELTVPLMWMVHFAVGYPYRYPELQLLAFAGPQRNVYVDDTT
YIL177C         ------------------------------------------------------------
YJL225C         ------------------------------------------------------------

YBL111C         --------YNKNGSSEPRLKTLDGLT----------------------------------
YHR218W         --------YNKNGSSEPRLKTLDGLT----------------------------------
YML133C         --------YNKNSSSEPRLKTLDGLT----------------------------------
YNL339C         RRIQLYTDYNKNGSSEPRLKTLDGLTSDYVFYFVTVLRQMQICALGNSYDAFNHDPWMDV
YGR296W         RRIQLYTDYNKNGSSEPRLKTLDGLTSDYVFYFVTVLRQMQICALGNSYDAFNHDPWMDV
YPL283C         RRIQLYTDYNKNGSSEPRLKTLDGLTSDYVFYFVTVLRQMQICALGNSYDAFNHDPWMDV
YLR467W         RRIQLYTDYNKNGSSEPRLKTLDGLTSDYVFYFVTVLRQMQICALGNSYDAFNHDPWMDV
YDR545W         RRIQLYTDYNKNGSSEPRLKTLDGLTSDYVFYFVTVLRQMQICALGNSYDAFNHDPWMDV
YIL177C         --IQLYTDYNKNGSSEPRLKTLDGLTSDYVFYFVTVLRQMQICALGNSYDAFNHDPWMDV
YJL225C         --IQLYTDYNKNGSSEPRLKTLDGLTSDYVFYFVTVLRQMQICALGNSYDAFNHDPWMDV

YBL111C         ------------------------------------------------------------
YHR218W         ------------------------------------------------------------
YML133C         ------------------------------------------------------------
YNL339C         VGFEDPDQVTNRDISRIVLYSYMFLNTAKGCLVEYATFRQYMRELPKNAPQKLNFREMRQ
YGR296W         VGFEDPDQVTNRDISRIVLYSYMFLNTAKGCLVEYATFRQYMRELPKNAPQKLNFREMRQ
YPL283C         VGFEDPDQVTNRDISRIVLYSYMFLNTAKGCLVEYATFRQYMRELPKNAPQKLNFREMRQ
YLR467W         VGFEDPDQVTNRDISRIVLYSYMFLNTAKGCLVEYATFRQYMRELPKNAPQKLNFREMRQ
YDR545W         VGFEDPDQVTNRDISRIVLYSYMFLNTAKGCLVEYATFRQYMRELPKNAPQKLNFREMRQ
YIL177C         VGFEDPNQVTNRDISRIVLYSYMFLNTAKGCLVEYATFRQYMRELPKNAPQKLNFREMRQ
YJL225C         VGFEDPNQVTNRDISRIVLYSYMFLNTAKGCLVEYATFRQYMRELPKNAPQKLNFREMRQ

YBL111C         ------------------------------------------------------------
YHR218W         ------------------------------------------------------------
YML133C         ------------------------------------------------------------
YNL339C         GLIALGRHCVGSRFETDLYESATSELMANHSVQTGRNIYGVDSFSLTSVSGTTATLLQER
YGR296W         GLIALGRHCVGSRFETDLYESATSELMANHSVQTGRNIYGVDSFSLTSVSGTTATLLQER
```

```
YPL283C         GLIALGRHCVGSRFETDLYESATSELMANHSVQTGRNIYGVDSFSLTSVSGTTATLLQER
YLR467W         GLIALGRHCVGSRFETDLYESATSELMANHSVQTGRNIYGVDSFSLTSVSGTTATLLQER
YDR545W         GLIALGRHCVGSRFETDLYESATSELMANHSVQTGRNIYGVDSFSLTSVSGTTATLLQER
YIL177C         GLIALGRHCVGSRFETDLYESATSELMANHSVQTGRNIYGVDSFSLTSVSGTTATLLQER
YJL225C         GLIALGRHCVGSRFETDLYESATSELMANHSVQTGRNIYGVDSFSLTSVSGTTATLLQER

YBL111C         -SERWIQWLGLESDYHCSFSSTRNAEDVVAGEAASSDHDQKISRVTRKRPREPKSTNDIL
YHR218W         -SERWIQWLGLESDYHCSFSSTRNAEDVVAGEAASSDHDQKISRVTRKRPREPKSTNDIL
YML133C         -SERWIQWLGLESDYHCSFSSTRNAEDVVAGEAASSDHHQKISRVTRKRPREPKSTNDIL
YNL339C         ASERWIQWLGLESDYHCSFSSTRNAEDVVAGEAASSDHHQKISRVTRKRPREPKSTNDIL
YGR296W         ASERWIQWLGLESDYHCSFSSTRNAEDVVAGEAASSDHHQKISRVTRKRPREPKSTNDIL
YPL283C         ASERWIQWLGLESDYHCSFSSTRNAEDVVAGEAASSDHHQKISRVTRKRPREPKSTNDIL
YLR467W         ASERWIQWLGLESDYHCSFSSTRNAEDVVAGEAASSDHDQKISRVTRKRPREPKSTNDIL
YDR545W         ASERWIQWLGLESDYHCSFSSTRNAEDVVAGEAASSDHDQKISRVTRKRPREPKSTNDIL
YIL177C         ASERWIQWLGLESDYHCSFSSTRNAEDVVAGEAASSNHHQKISRVTRKRPREPKSTNDIL
YJL225C         ASERWIQWLGLESDYHCSFSSTRNAEDVVAGEAASSNHHQKISRVTRKRPREPKSTNDIL

YBL111C         VAGRKLFGSSFEFRDLHQLRLCHEIYMADTPSVAVQAPPGYGKTELFHLPLIALASKGDV
YHR218W         VAGRKLFGSSFEFRDLHQLRLCHEIYMADTPSVAVQAPPGYGKTELFHLPLIALASKGDV
YML133C         VAGQKLFGSSFEFRDLHQLRLCHEIYMADTPSVAVQAPPGYGKTELFHLPLIALASKGDV
YNL339C         VAGQKLFGSSFEFRDLHQLRLCHEIYMADTPSVAVQAPPGYGKTELFHLPLIALASKGDV
YGR296W         VAGQKLFGSSFEFRDLHQLRLCHEIYMADTPSVAVQAPPGYGKTELFHLPLIALASKGDV
YPL283C         VAGQKLFGSSFEFRDLHQLRLCHEIYMADTPSVAVQAPPGYGKTELFHLPLIALASKGDV
YLR467W         VAGQKLFGSSFEFRDLHQLRLCHEIYMADTPSVAVQAPPGYGKTELFHLPLIALASKGDV
YDR545W         VAGQKLFGSSFEFRDLHQLRLCHEIYMADTPSVAVQAPPGYGKTELFHLPLIALASKGDV
YIL177C         VAGQKLFGSSFEFRDLHQLRLCYEIYMADTPSVAVQAPPGYGKTELFHLPLIALASKGDV
YJL225C         VAGQKLFGSSFEFRDLHQLRLCYEIYMADTPSVAVQAPPGYGKTELFHLPLIALASKGDV

YBL111C         KYVSFLFVPYTVLLANCMIRLGRRGCLNVAPVRNFIEEGCDGVTDLYVGIYDDLASTNFT
YHR218W         KYVSFLFVPYTVLLANCMIRLGRRGCLNVAPVRNFIEEGCDGVTDLYVGIYDDLASTNFT
YML133C         KYVSFLFVPYTVLLANCMIRLSRCGCLNVAPVRNFIEEGCDGVTDLYVGIYDDLASTNFT
YNL339C         KYVSFLFVPYTVLLANCMIRLSRCGCLNVAPVRNFIEEGCDGVTDLYVGIYDDLASTNFT
YGR296W         KYVSFLFVPYTVLLANCMIRLSRCGCLNVAPVRNFIEEGCDGVTDLYVGIYDDLASTNFT
YPL283C         KYVSFLFVPYTVLLANCMIRLSRCGCLNVAPVRNFIEEGCDGVTDLYVGIYDDLASTNFT
YLR467W         KYVSFLFVPYTVLLANCMIRLSRCGCLNVAPVRNFIEEGCDGVTDLYVGIYDDLASTNFT
YDR545W         KYVSFLFVPYTVLLANCMIRLSRCGCLNVAPVRNFIEEGCDGVTDLYVGIYDDLASTNFT
YIL177C         EYVSFLFVPYTVLLANCMIRLGRCGCLNVAPVRNFIEEGYDGVTDLYVGIYDDLASTNFT
YJL225C         EYVSFLFVPYTVLLANCMIRLGRRGCLNVAPVRNFIEEGYDGVTDLYVGIYDDLASTNFT

YBL111C         DRIAAWENIVECTFRTNNVKLGYLIVDEFHNFETEVYRQSQFGGITNLDFDAFEKAIFLS
YHR218W         DRIAAWENIVECTFRTNNVKLGYLIVDEFHNFETEVYRQSQFGGITNLDFDAFEKAIFLS
YML133C         DRIAAWENIVECTFRTNNVKLGYLIVDEFHNFETEVYRQSQFGGITNLDFDAFEKAIFLS
YNL339C         DRIAAWENIVECTFRTNNVKLGYLIVDEFHNFETEVYRQSQFGGITNLDFDAFEKAIFLS
YGR296W         DRIAAWENIVECTFRTNNVKLGYLIVDEFHNFETEVYRQSQFGGITNLDFDAFEKAIFLS
YPL283C         DRIAAWENIVECTFRTNNVKLGYLIVDEFHNFETEVYRQSQFGGITNLDFDAFEKAIFLS
YLR467W         DRIAAWENIVECTFRTNNVKLGYLIVDEFHNFETEVYRQSQFGGITNLDFDAFEKAIFLS
YDR545W         DRIAAWENIVECTFRTNNVKLGYLIVDEFHNFETEVYRQSQFGGITNLDFDAFEKAIFLS
YIL177C         DRIAAWENIVECTFRTNNVKLGYLIVDEFHNFETEVYRQSQFGGITNLDFDAFEKAIFLS
YJL225C         DRIAAWENIVECTFRTNNVKLGYLIVDEFHNFETEVYRQSQFGGITNLDFDAFEKAIFLS

YBL111C         GTAPEAVADAALQRIGLTGLAKKSMDINELKRSEDLSRGLSSYPTRMFNLIKEKSEVPLG
```

```
YHR218W         GTAPEAVADAALQRIGLTGLAKKSMDINELKRSEDLSRGLSSYPTRMFNLIKEKSEVPLG
YML133C         GTAPEAVADAALQRIGLTGLAKKSMDINELKRSEDLSRGLSSYPTRMFNLIKEKSEVPLG
YNL339C         GTAPEAVADAALQRIGLTGLAKKSMDINELKRSEDLSRGLSSYPTRMFNLIKEKSEVPLG
YGR296W         GTAPEAVADAALQRIGLTGLAKKSMDINELKRSEDLSRGLSSYPTRMFNLIKEKSEVPLG
YPL283C         GTAPEAVADAALQRIGLTGLAKKSMDINELKRSEDLSRGLSSYPTRMFNLIKEKSEVPLG
YLR467W         GTAPEAVADAALQRIGLTGLAKKSMDINELKRSEDLSRGLSSYPTRMFNLIKEKSEVPLG
YDR545W         GTAPEAVADAALQRIGLTGLAKKSMDINELKRSEDLSRGLSSYPTRMFNLIKEKSEVPLG
YIL177C         GTAPEAVADAALQRIGLTGLAKKSMDINELKRSEDLSRGLSSYPTRMFNLIKEKSEVPLG
YJL225C         GTAPEAVADAALQRIGLTGLAKKSMDINELKRSEDLSRGLSSYPTRMFNLIKEKSEVPLG

YBL111C         HVHKIWKKVESQPEEALKLLLALFEIEPESKAIVVASTTNEVEELACSWRKYFRVVWIHG
YHR218W         HVHKIWKKVESQPEEALKLLLALFEIEPESKAIVVASTTNEVEELACSWRKYFRVVWIHG
YML133C         HVHKIWKKVESQPEEALKLLLALFEIEPESKAIVVASTTNEVEELACSWRKYFRVVWIHG
YNL339C         HVHKIWKKVESQPEEALKLLLALFEIEPESKAIVVASTTNEVEELACSWRKYFRVVWIHG
YGR296W         HVHKIWKKVESQPEEALKLLLALFEIEPESKAIVVASTTNEVEELACSWRKYFRVVWIHG
YPL283C         HVHKIWKKVESQPEEALKLLLALFEIEPESKAIVVASTTNEVEELACSWRKYFRVVWIHG
YLR467W         HVHKIWKKVESQPEEALKLLLALFEIEPESKAIVVASTTNEVEELACSWRKYFRVVWIHG
YDR545W         HVHKIWKKVESQPEEALKLLLALFEIEPESKAIVVASTTNEVEELACSWRKYFRVVWIHG
YIL177C         HVHKIRKKVESQPEEALKLLLALFESEPESKAIVVASTTNEVEELACSWRKYFRVVWIHG
YJL225C         HVHKIRKKVESQPEEALKLLLALFESEPESKAIVVASTTNEVEELACSWRKYFRVVWIHG

YBL111C         KLGCCRKG----------------------------------------------------V
YHR218W         KLGCCRKG----------------------------------------------------V
YML133C         KLGAAEKVSRTKEFVTDGSMRVLIGTKLVTEGIDIKQLMMVIMLDNRLNIIELIQGVGRL
YNL339C         KLGAAEKVSRTKEFVTDGSMRVLIGTKLVTEGIDIKQLMMVIMLDNRLNIIELIQGVGRL
YGR296W         KLGAAEKVSRTKEFVTDGSMRVLIGTKLVTEGIDIKQLMMVIMLDNRLNIIELIQGVGRL
YPL283C         KLGAAEKVSRTKEFVTDGSMRVLIGTKLVTEGIDIKQLMMVIMLDNRLNIIELIQGVGRL
YLR467W         KLGAAEKVSRTKEFVTDGSMRVLIGTKLVTEGIDIKQLMMVIMLDNRLNIIELIQGVGRL
YDR545W         KLGAAEKVSRTKEFVTDGSMRVLIGTKLVTEGIDIKQLMMVIMLDNRLNIIELIQGVGRL
YIL177C         KLGAAEKVSRTKEFVTDGSMQVLIGTKLVTEGIDIKQLMMVIMLDNRLNIIELIQGVGRL
YJL225C         KLGAAEKVSRTKEFVTDGSMQVLIGTKLVTEGIDIKQLMMVIMLDNRLNIIELIQGVGRL

YBL111C         SHKGVCH-----------------------------------------------------
YHR218W         SHKGVCH-----------------------------------------------------
YML133C         RDGGLCYLLSRKNSWAARNRKGELPPIKEGCITEQVREFYGLESKKGKKGQHVGCCGSRT
YNL339C         RDGGLCYLLSRKNSWAARNRKGELPPIKEGCITEQVREFYGLESKKGKKGQHVGCCGSRT
YGR296W         RDGGLCYLLSRKNSWAARNRKGELPPIKEGCITEQVREFYGLESKKGKKGQHVGCCGSRT
YPL283C         RDGGLCYLLSRKNSWAARNRKGELPPIKEGCITEQVREFYGLESKKGKKGQHVGCCGSRT
YLR467W         RDGGLCYLLSRKNSWAARNRKGELPPIKEGCITEQVREFYGLESKKGKKGQHVGCCGSRT
YDR545W         RDGGLCYLLSRKNSWAARNRKGELPPIKEGCITEQVREFYGLESKKGKKGQHVGCCGSRT
YIL177C         RDGGLCYLLSRKNSWAARNRKGELPPIKEGCITEQVREFYGLESKKGKKGQHVGCCGSRT
YJL225C         RDGGLCYLLSRKNSWAARNRKGELPPIKEGCITEQVREFYGLESKKGKKGQHVGCCGSRT

YML133C         DLSADTVELIERMDRLAEKQATASMSIVALPSSFQESNSSDRCRKYCSSDEDSDTCIHG-
YNL339C         DLSADTVELIERMDRLAEKQATASMSIVALPSSFQESNSSDRCRKYCSSDEDSDTCIHG-
YGR296W         DLSADTVELIERMDRLAEKQATASMSIVALPSSFQESNSSDRCRKYCSSDEDSDTCIHG-
YPL283C         DLSADTVELIERMDRLAEKQATASMSIVALPSSFQESNSSDRCRKYCSSDEDSDTCIHG-
YLR467W         DLSADTVELIERMDRLAEKQATASMSIIALPSSFQESNSSDRCRKYCSSDEDSDTCIHG-
YDR545W         DLSADTVELIERMDRLAEKQATASMSIIALPSSFQESNSSDRCRKYCSSDEDSDTCIHG-
YIL177C         DLSADTVELIERMDRLAEKQATASMSIVALPSSFQESNSSDRYRKYCSSDEDSNTCIHGS
YJL225C         DLSADTVELIERMDRLAEKQATASMSIVALPSSFQESNSSDRYRKYCSSDEDSNTCIHGS
```

```
YML133C         ------------------------------------------------------------
YNL339C         ------------------------------------------------------------
YGR296W         ------------------------------------------------------------
YPL283C         ------------------------------------------------------------
YLR467W         ------------------------------------------------------------
YDR545W         ------------------------------------------------------------
YIL177C         ANASTNASTNAITTASTNVRTNATTNASTNATTNASTNASTNATTNASTNATTNSSTNAT
YJL225C         ANASTNASTNAITTASTNVRTNATTNASTNATTNASTNASTNATTNASTNATTNSSTNAT

YML133C         --------------------------------SANASTNATTNSSTNATTTASTNVRTSAT
YNL339C         --------------------------------SANASTNATTNSSTNATTTASTNVRTSAT
YGR296W         --------------------------------SANASTNATTNSSTNATTTASTNVRTSAT
YPL283C         --------------------------------SANASTNATTNSSTNATTTASTNVRTSAT
YLR467W         --------------------------------SANASTNATTNSSTNATTTASTNVRTSAT
YDR545W         --------------------------------SANASTNATTNSSTNATTTASTNVRTSAT
YIL177C         TTASTNVRTSATTTASINVRTSATTTESTNSSTNATTTESTNSSTNATTTESTNSNTSAT
YJL225C         TTASTNVRTSATTTASINVRTSATTTESTNSSTNATTTESTNSSTNATTTESTNSNTSAT

YML133C         TTASINVRTSATTTESTNSSTNATTTASTNVRTSATTTASINVRTSATTTESTNSNTSAT
YNL339C         TTASINVRTSATTTESTNSSTNATTTASTNVRTSATTTASINVRTSATTTESTNSNTSAT
YGR296W         TTASINVRTSATTTESTNSSTNATTTASTNVRTSATTTASINVRTSATTTESTNSNTSAT
YPL283C         TTASINVRTSATTTESTNSSTNATTTASTNVRTSATTTASINVRTSATTTESTNSNTSAT
YLR467W         TTASINVRTSAITTESTNSSTNATTTASTNVRTSATTTASINVRTSATTTESTNSNTSAT
YDR545W         TTASINVRTSAITTESTNSSTNATTTASTNVRTSATTTASINVRTSATTTESTNSNTSAT
YIL177C         TTASINVRTSATTTESTNSSTSATTTASINVRTSATTTKSINSSTNATTTESTNSNTNAT
YJL225C         TTASINVRTSATTTESTNSSTSATTTASINVRTSATTTKSINSSTNATTTESTNSNTNAT

YML133C         TTESTDSNTSATTTESTDSNTSATTTASTNSSTNATTTASTNSSTNATTTESTNASAKED
YNL339C         TTESTDSNTSATTTESTDSNTSATTTASTNSSTNATTTASTNSSTNATTTESTNASAKED
YGR296W         TTESTDSNTSATTTESTDSNTSATTTASTNSSTNATTTASTNSSTNATTTESTNASAKED
YPL283C         TTESTDSNTSATTTESTDSNTSATTTASTNSSTNATTTASTNSSTNATTTESTNASAKED
YLR467W         TTESTDSNTSATTTESTDSNTSATTTASTNSSTNATTTASTNSSTNATTTESTNASAKED
YDR545W         TTESTDSNTSATTTESTDSNTSATTTASTNSSTNATTTASTNSSTNATTTESTNASAKED
YIL177C         TTESTNSSTNATTTESTNSSTNATTTESTNSNTSAATTESTNSNTSATTTESTNASAKED
YJL225C         TTESTNSSTNATTTESTNSSTNATTTESTNSNTSAATTESTNSNTSATTTESTNASAKED

YML133C         ANKDGNAEDNRFHPVTDINKESYKRKGSQMVLLERKKLKAQFPNTSENMNVLQFLGFRSD
YNL339C         ANKDGNAEDNRFHPVTDINKESYKRKGSQMVLLERKKLKAQFPNTSENMNVLQFLGFRSD
YGR296W         ANKDGNAEDNRFHPVTDINKESYKRKGSQMVLLERKKLKAQFPNTSENMNVLQFLGFRSD
YPL283C         ANKDGNAEDNRFHPVTDINKESYKRKGSQMVLLERKKLKAQFPNTSENMNVLQFLGFRSD
YLR467W         ANKDGNAEDNRFHPVTDINKESYKRKGSQMVLLERKKLKAQFPNTSENMNVLQFLGFRSD
YDR545W         ANKDGNAEDNRFHPVTDINKESYKRKGSQMVLLERKKLKAQFPNTSENMNVLQFLGFRSD
YIL177C         ANKDGNAEDNRFHPVTDINKESYKRKGSQMVLLERKKLKAQFPNTSENMNVLQFLGFRSD
YJL225C         ANKDGNAEDNRFHPVTDINKESYKRKGSQMVLLERKKLKAQFPNTSENMNVLQFLGFRSD

YML133C         EIKHLFLYGIDIYFCPEGVFTQYGLCKGCQKMFELCVCWAGQKVSYRRMAWEALAVERML
YNL339C         EIKHLFLYGIDVYFCPEGVFTQYGLCKGCQKMFELCVCWAGQKVSYRRMAWEALAVERML
YGR296W         EIKHLFLYGIDVYFCPEGVFTQYGLCKGCQKMFELCVCWAGQKVSYRRMAWEALAVERML
YPL283C         EIKHLFLYGIDVYFCPEGVFTQYGLCKGCQKMFELCVCWAGQKVSYRRMAWEALAVERML
YLR467W         EIKHLFLYGIDVYFCPEGVFTQYGLCKGCQKMFELCVCWAGQKVSYRRMAWEALAVERML
```

```
YDR545W         EIKHLFLYGIDVYFCPEGVFTQYGLCKGCQKMFELCVCWAGQKVSYRRMAWEALAVERML
YIL177C         EIKHLFLYGIDIYFCPEGVFTQYGLCKGCQKMFELCVCWAGQKVSYRRIAWEALAVERML
YJL225C         EIKHLFLYGIDIYFCPEGVFTQYGLCKGCQKMFELCVCWAGQKVSYRRIAWEALAVERML

YML133C         RNDEEYKEYLEDIEPYHGDPVGYLKFFSVKRGEIYSQIQRNYAWYLAITRRRETISVLDS
YNL339C         RNDEEYKEYLEDIEPYHGDPVGYLKFFSVKRGEIYSQIQRNYAWYLAITRRRETISVLDS
YGR296W         RNDEEYKEYLEDIEPYHGDPVGYLKYFSVKRGEIYSQIQRNYAWYLAITRRRETISVLDS
YPL283C         RNDEEYKEYLEDIEPYHGDPVGYLKYFSVKRGEIYSQIQRNYAWYLAITRRRETISVLDS
YLR467W         RNDEEYKEYLEDIEPYHGDPVGYLKYFSVKRGEIYSQIQRNYAWYLAITRRRETISVLDS
YDR545W         RNDEEYKEYLEDIEPYHGDPVGYLKYFSVKRGEIYSQIQRNYAWYLAITRRRETISVLDS
YIL177C         RNDEEYKEYLEDIEPYHGDPVGYLKYFSVKRREIYSQIQRNYAWYLAITRRRETISVLDS
YJL225C         RNDEEYKEYLEDIEPYHGDPVGYLKYFSVKRREIYSQIQRNYAWYLAITRRRETISVLDS

YML133C         TRGKQGSQVFRMSGRQIKELYYKVWSNLRESKTEVLQYFLNWDEKKCREEWEAKDDTVFV
YNL339C         TRGKQGSQVFRMSGRQIKELYYKVWSNLRESKTEVLQYFLNWDEKKCREEWEAKDDTVFV
YGR296W         TRGKQGSQVFRMSGRQIKELYYKVWSNLRESKTEVLQYFLNWDEKKCREEWEAKDDTVFV
YPL283C         TRGKQGSQVFRMSGRQIKELYYKVWSNLRESKTEVLQYFLNWDEKKCREEWEAKDDTVFV
YLR467W         TRGKQGSQVFRMSGRQIKELYYKVWSNLRESKTEVLQYFLNWDEKKCREEWEAKDDTVFV
YDR545W         TRGKQGSQVFRMSGRQIKELYYKVWSNLRESKTEVLQYFLNWDEKKCREEWEAKDDTVFV
YIL177C         TRGKQGSQVFRMSGRQIKELYFKVWSNLRESKTEVLQYFLNWDEKKCQEEWEAKDDTVVV
YJL225C         TRGKQGSQVFRMSGRQIKELYFKVWSNLRESKTEVLQYFLNWDEKKCQEEWEAKDDTVVV

YML133C         EALEKVGVFQRLRSMTSAGLQGPQYVKLQFSRHHRQLRSRYELSLGMHLRDQLALGVTPS
YNL339C         EALEKVGVFQRLRSMTSAGLQGPQYVKLQFSRHHRQLRSRYELSLGMHLRDQLALGVTPS
YGR296W         EALEKVGVFQRLRSMTSAGLQGPQYVKLQFSRHHRQLRSRYELSLGMHLRDQLALGVTPS
YPL283C         EALEKVGVFQRLRSMTSAGLQGPQYVKLQFSRHHRQLRSRYELSLGMHLRDQLALGVTPS
YLR467W         EALEKVGVFQRLRSMTSAGLQGPQYVKLQFSRHHRQLRSRYELSLGMHLRDQLALGVTPS
YDR545W         EALEKVGVFQRLRSMTSAGLQGPQYVKLQFSRHHRQLRSRYELSLGMHLRDQLALGVTPS
YIL177C         EALEKGGVFQRLRSMTSAGLQGPQYVKLQFSRHHRQLRSRYELSLGMHLRDQIALGVTPS
YJL225C         EALEKGGVFQRLRSMTSAGLQGPQYVKLQFSRHHRQLRSRYELSLGMHLRDQIALGVTPS

YML133C         KVPHWTAFLSMLIGLFYNKTFRQKLEYLLEQISEVWLLPHWVDLANVEVLAADNTRVPLY
YNL339C         KVPHWTAFLSMLIGLFCNKTFRQKLEYLLEQISEVWLLPHWLDLANVEVLAADNTRVPLY
YGR296W         KVPHWTAFLSMLIGLFCNKTFRQKLEYLLEQISEVWLLPHWLDLANVEVLAADNTRVPLY
YPL283C         KVPHWTAFLSMLIGLFCNKTFRQKLEYLLEQISEVWLLPHWLDLANVEVLAADNTRVPLY
YLR467W         KVPHWTAFLSMLIGLFYNKTFRQKLEYLLEQISEVWLLPHWLDLANVEVLAADNTRVPLY
YDR545W         KVPHWTAFLSMLIGLFYNKTFRQKLEYLLEQISEVWLLPHWLDLANVEVLAADNTRVPLY
YIL177C         KVPHWTAFLSMLIGLFYNKTFRQKLEYLLEQISEVWLLPHWLDLANVEVLAADDTRVPLY
YJL225C         KVPHWTAFLSMLIGLFYNKTFRQKLEYLLEQISEVWLLPHWLDLANVEVLAADDTRVPLY

YML133C         MLMVAVHKELDSDDVPDGRFDIILLCRDSSREVGE
YNL339C         MLMVAVHKELDSDDVPDGRFDILLCRDSSREVGE-
YGR296W         MLMVAVHKELDSDDVPDGRFDILLCRDSSREVGE-
YPL283C         MLMVAVHKELDSDDVPDGRFDILLCRDSSREVGE-
YLR467W         MLMVAVHKELDSDDVPDGRFDIILLCRDSSREVGE
YDR545W         MLMVAVHKELDSDDVPDGRFDIILLCRDSSREVGE
YIL177C         MLMVAVHKELDSDDVPDGRFDILLCRDSSREVGE-
YJL225C         MLMVAVHKELDSDDVPDGRFDILLCRDSSREVGE-
```

Figure A3.3. BLAST scores for the comparison of HSPs and their associated URSs.

HSPs with e-values of 0 have been placed in the $1E^{-180}$ category to allow logarithmic scaling.

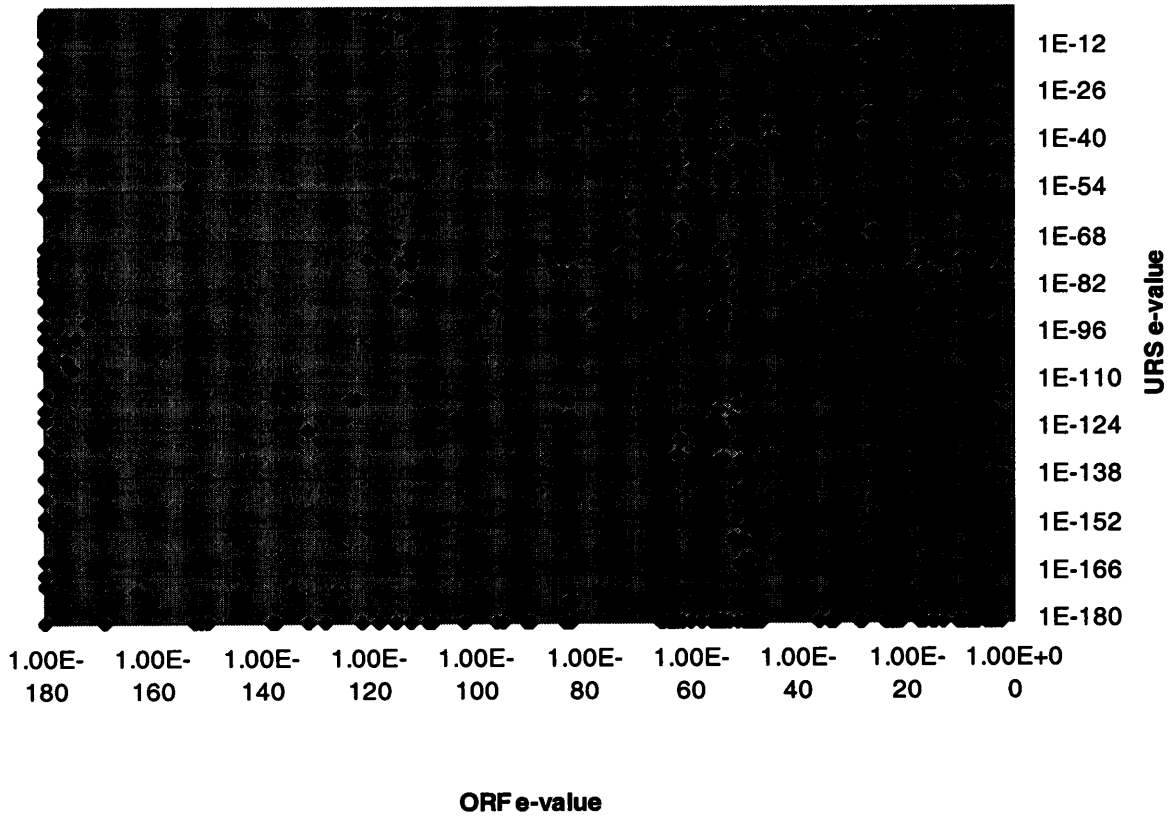


Table A3.1. Clusters present in all vector based binding site analyses.

Clusters are given an arbitrary number for identification. The systematic ORF name for each member of a cluster is listed followed by any available functional annotation (taken from MIPS (1999 schema), KEGG and SwissProt).

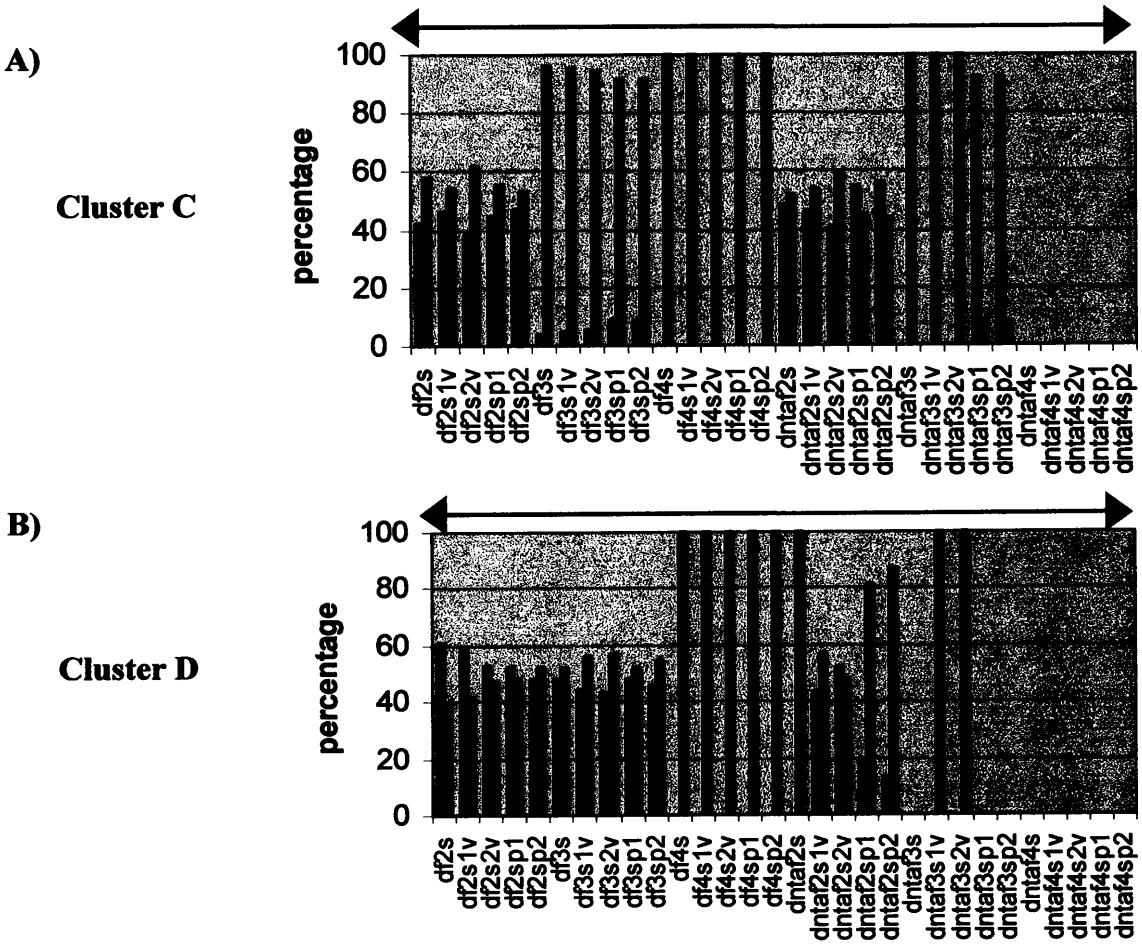
Cluster	ORF ID	Functional information
1	YDL248W YML132W YBR302C YNL336W YJR161C	Similar to subtelomeric encoded proteins, predicted transmembrane domain.
2	YEL076C-A YEL076C	Unknown
	YLL067C YLL066C	ATP/GTP binding site, mitochondrial energy transfer protein signature, predicted transmembrane domain, similar to subtelomeric encoded proteins
	YLR464W YPR203W YFL065C	Similar to subtelomeric encoded proteins
3	YFL064C YER189W	Similar to subtelomeric encoded proteins
	YOR545W YJL225C YIL177C YML133C YLR467W YBL111C	ATP/GTP binding site, mitochondrial energy transfer protein signature, predicted transmembrane domain, similar to subtelomeric encoded proteins
	YLR482W	Unknown
4	YPR204W YHL050C YFL066C	ATP/GTP binding site, mitochondrial energy transfer protein signature, predicted transmembrane domain, similar to subtelomeric encoded proteins
5	YHR217C YBL109W YNL338W YDR544C	Similar to subtelomeric encoded proteins
6	YKL223W YIR040C YGL260W YBL108W YIL175W YIL174W YCR103C	Similar to subtelomeric encoded proteins
7	YLR155C YLR157C YLR158C YLR160C	MIPS : 1-1-0, 1-1-4, 13-1-0 KEGG : 2.5, 5.2, 6.5
8	YLR161W YLR159W YLR156W	Identical to each other

9	YDL248W YML132W YBR302C YNL336W YJR161C	Similar to subtelomeric encoded proteins, predicted transmembrane regions
10	YIR041W YKL224C YGL261C	Similar to subtelomeric encoded proteins, predicted transmembrane region, stress induced protein signature
11	YAL064C-A YHR213W YAR062W	Putative pseudogene Putative transcriptional activator Putative pseudogene
12	YJR029W YJR028W YLR411W	TY1B protein TY1A protein Copper transport protein, 3 TM domains. Only expressed in strains lacking the adjacent 5' TY element
13	YMR323W YPL281C YOR393W	MIPS: 1-5-1 KEGG: 1.1 (YMR323W)
14	YMR322C YPL280W YOR391C	Predicted transmembrane domain

Appendix 4

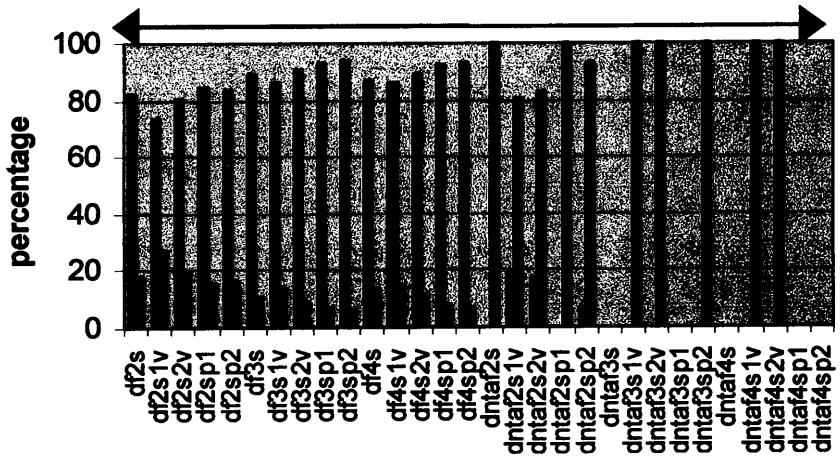
Figure A4.1. Results for absolute position analysis of Eisen clusters (Eisen *et al.* 1998) using the full binding site dataset.

The final score for an analysis (x-axis) is displayed as a percentage (y-axis) of the total for both actual and random scores. Blue bars represent the scores for the analysis on the real dataset. Red bars represent average scores for ten random analyses. See Table 4.5 (Chapter 4) for an explanation of the dataset codes seen on the x-axis. The green arrow highlights those analyses carried out with TATA boxes retained in the binding site data whereas the orange arrow highlights those analyses carried out with TATA boxes removed from the binding site data.



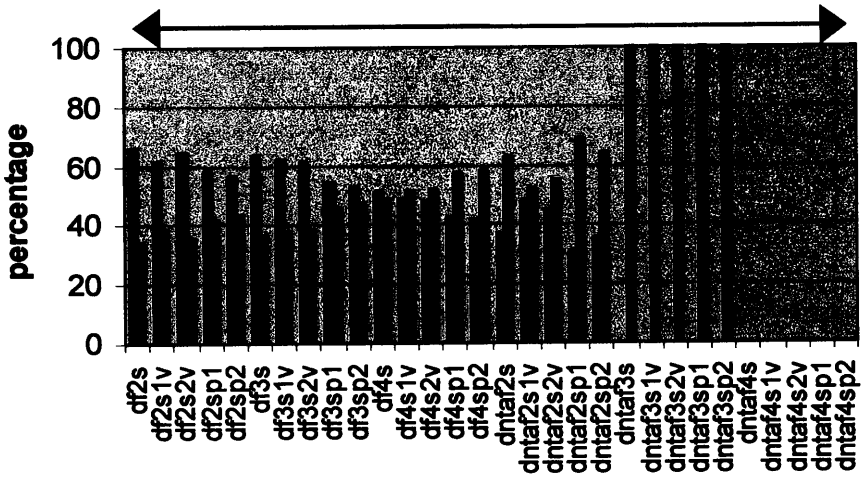
C)

Cluster E



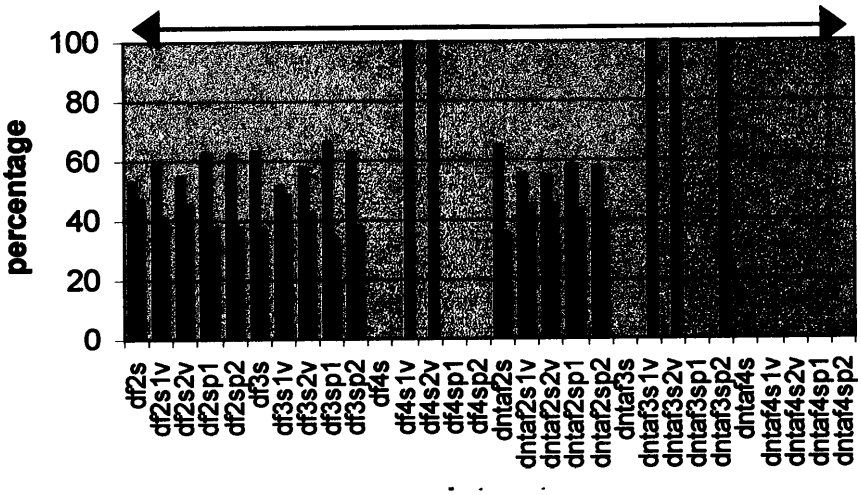
D)

Cluster F



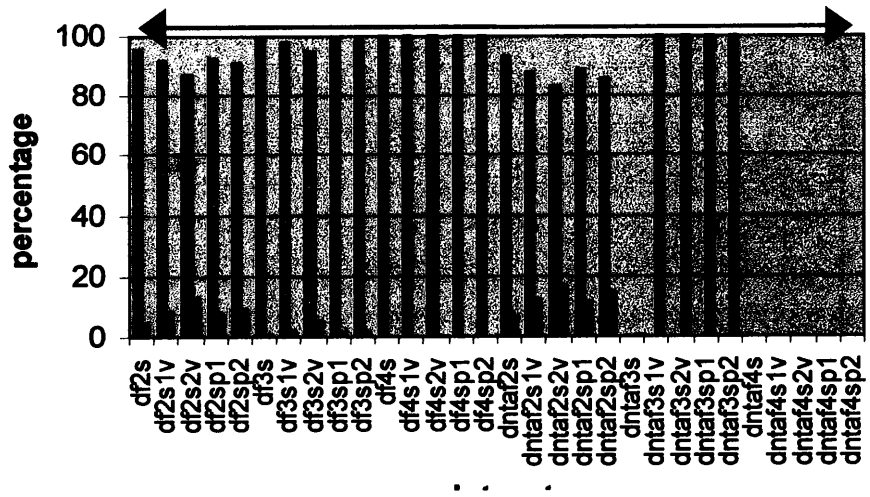
E)

Cluster G



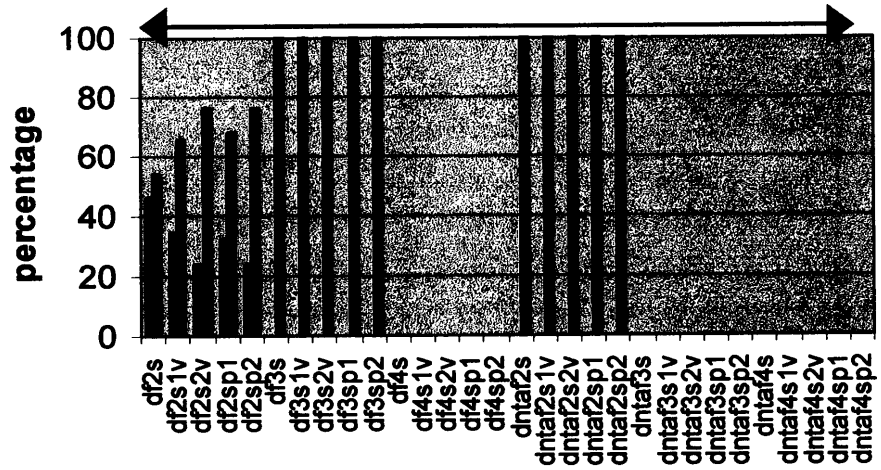
F)

Cluster H



G)

Cluster J



H)

Cluster K

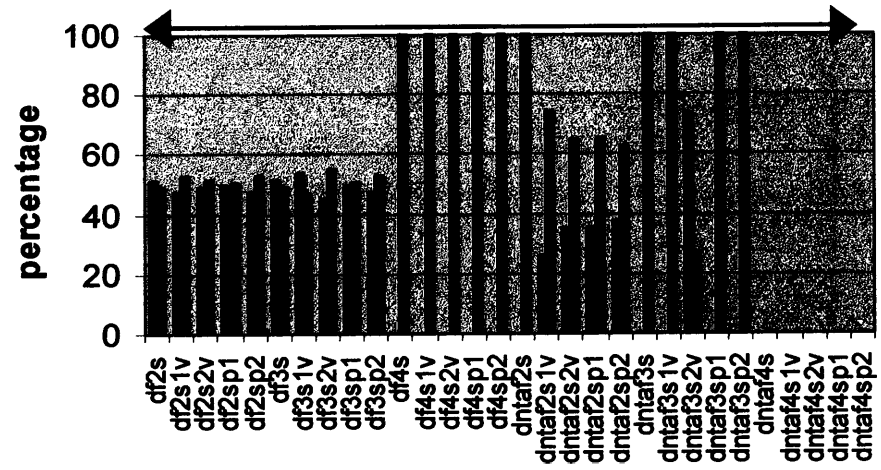
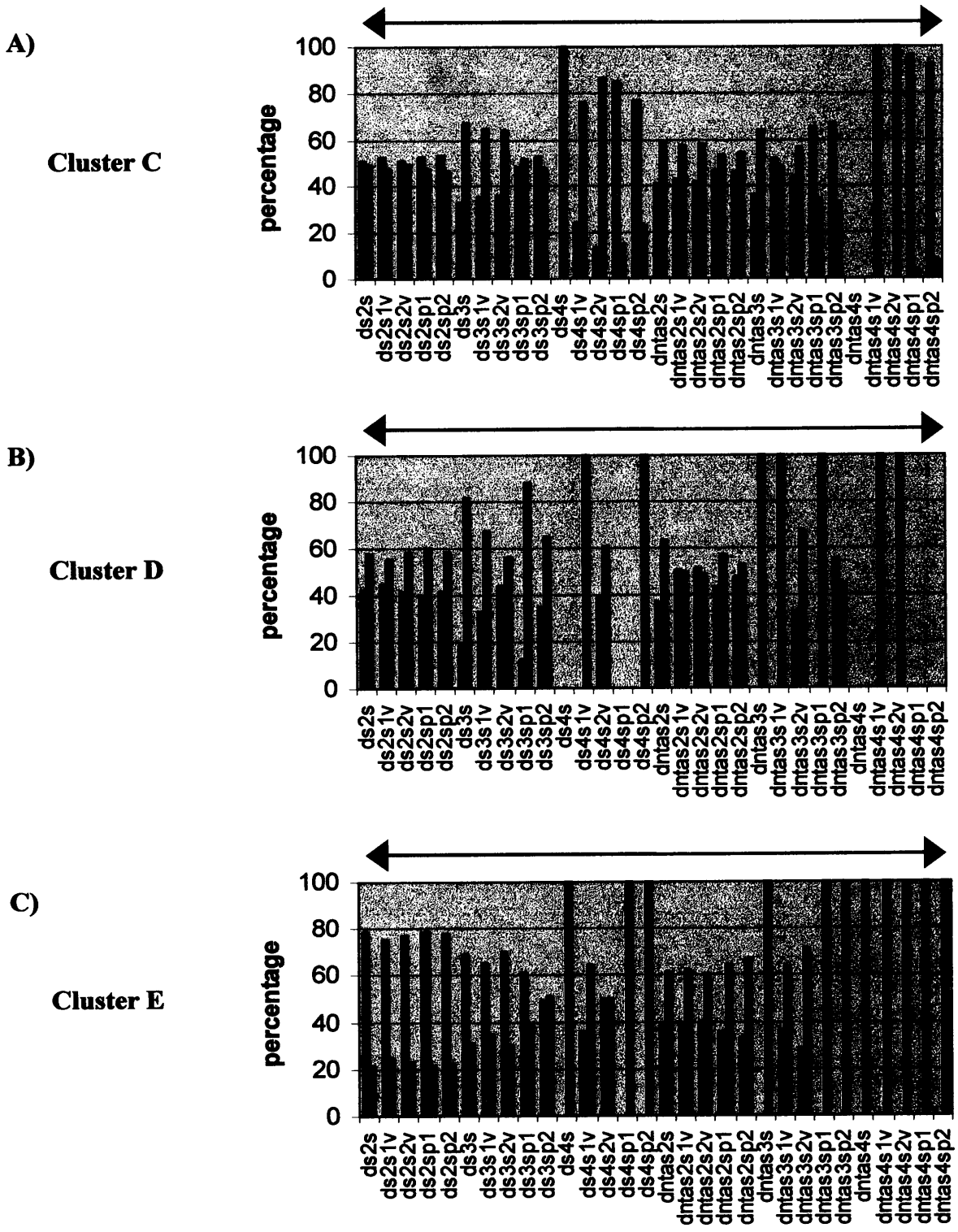


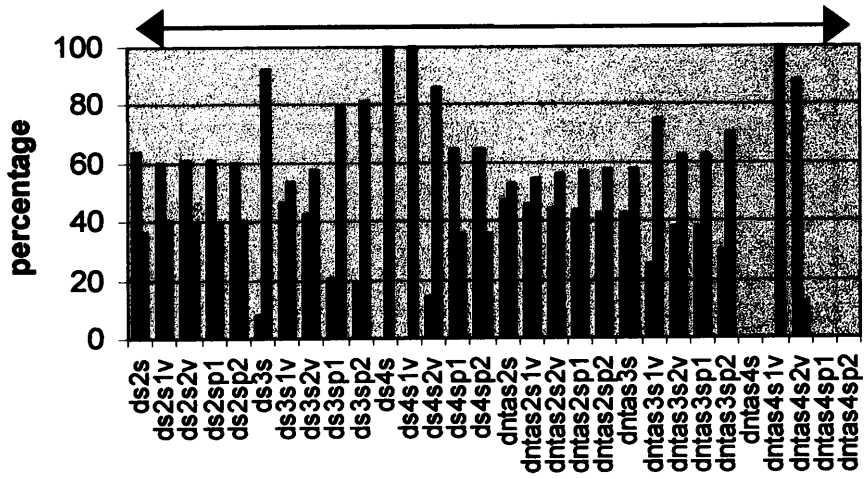
Figure A4.2. Results for absolute position analysis of Eisen clusters (Eisen *et al.* 1998) using the simplified binding site dataset.

See Figure A4.1 for details.



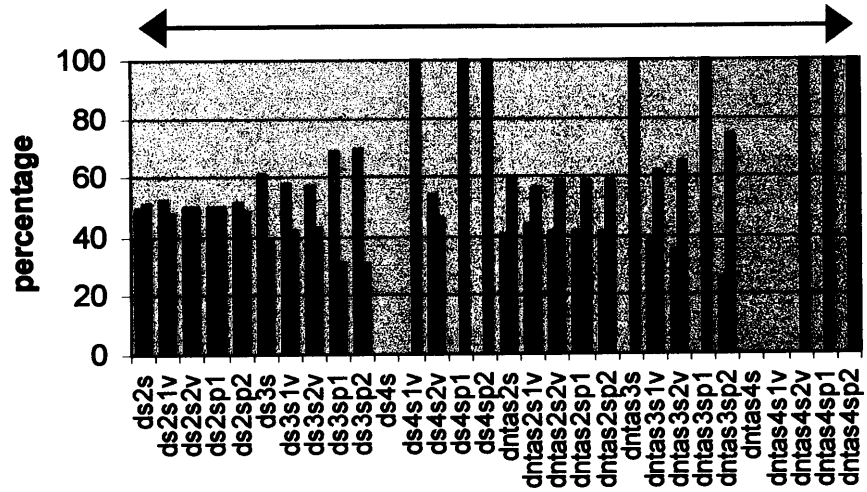
D)

Cluster F



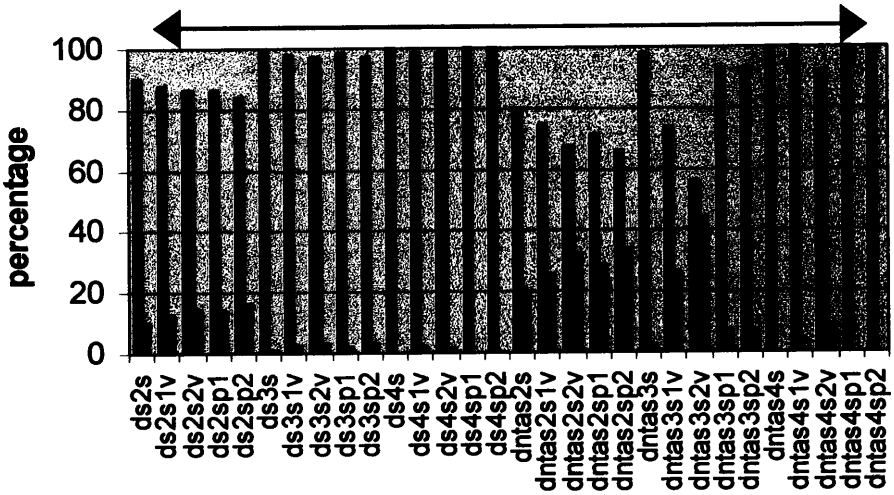
E)

Cluster G



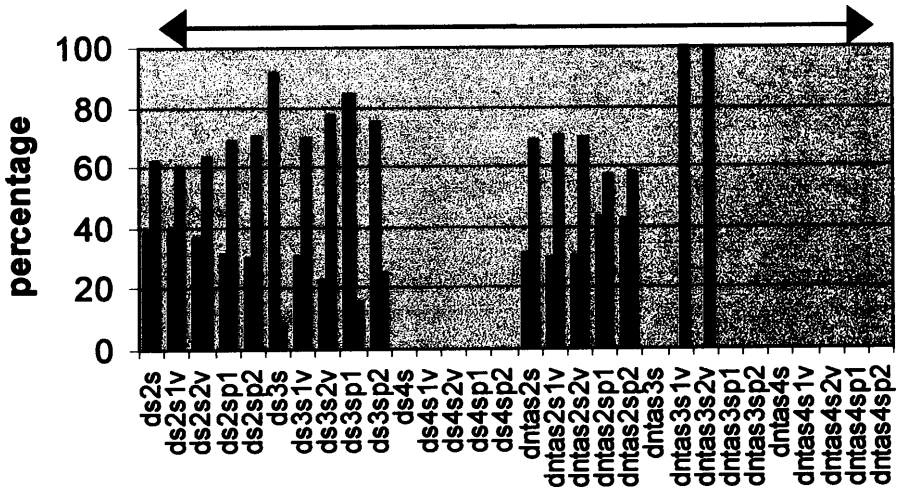
F)

Cluster H



G)

Cluster J



H)

Cluster K

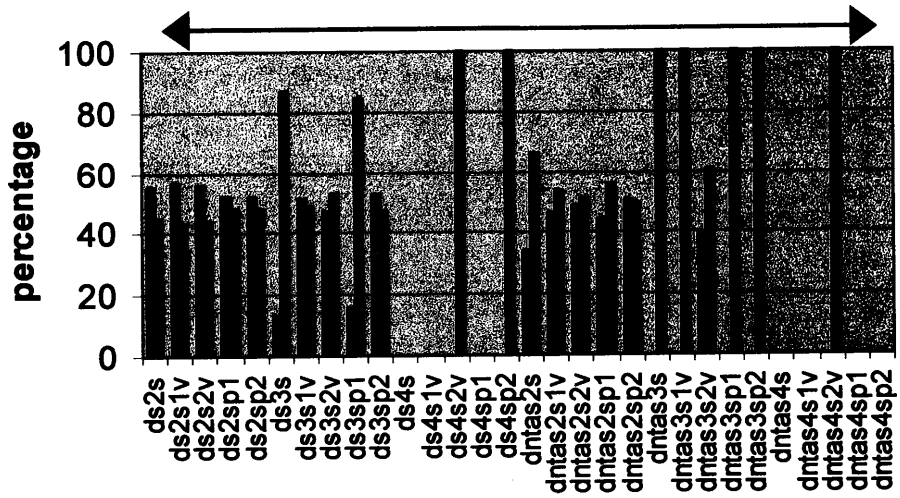
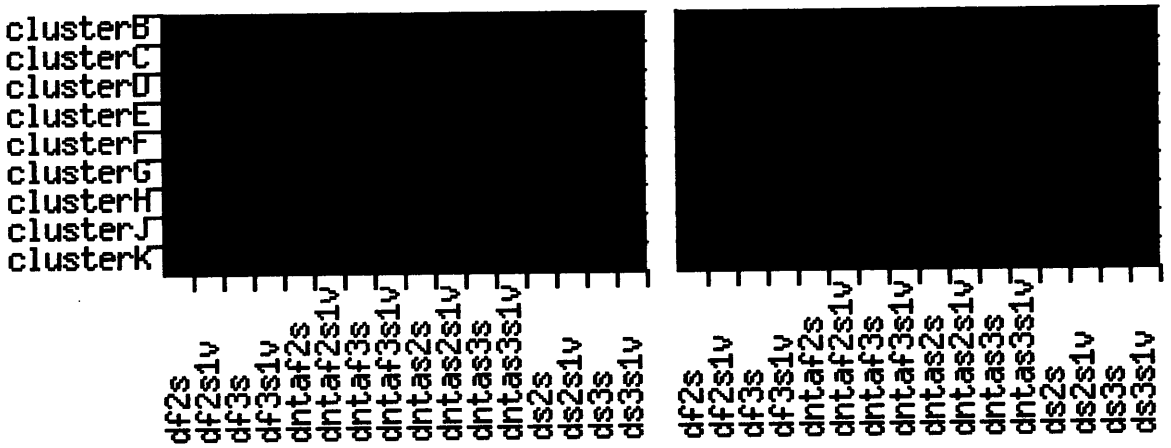


Figure A4.3. Results of Inter to Intra cluster comparison for absolute position analysis of clusters taken from Eisen *et al.* (1998).

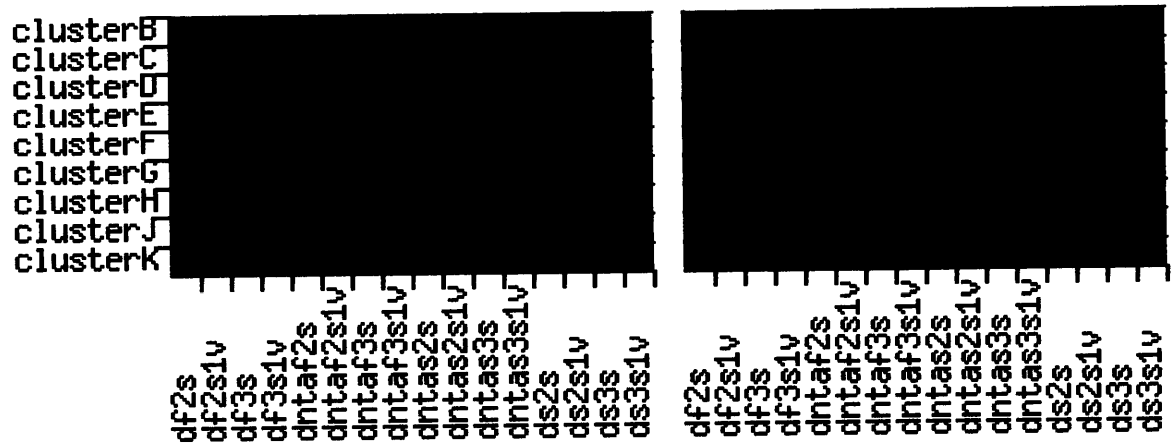
A) No distance constraints B) Distance constraint of 3 bases C) Distance constraint of 6 bases D) Low-complexity sites masked, no distance constraints

In the left hand image of each pair, the results for each cluster (row) are scaled independently. This highlights the analyses that perform the best and worst for each cluster. In the right hand image the scaling is global, which highlights those clusters with the highest number of intra-cluster links in comparison to inter-cluster links.

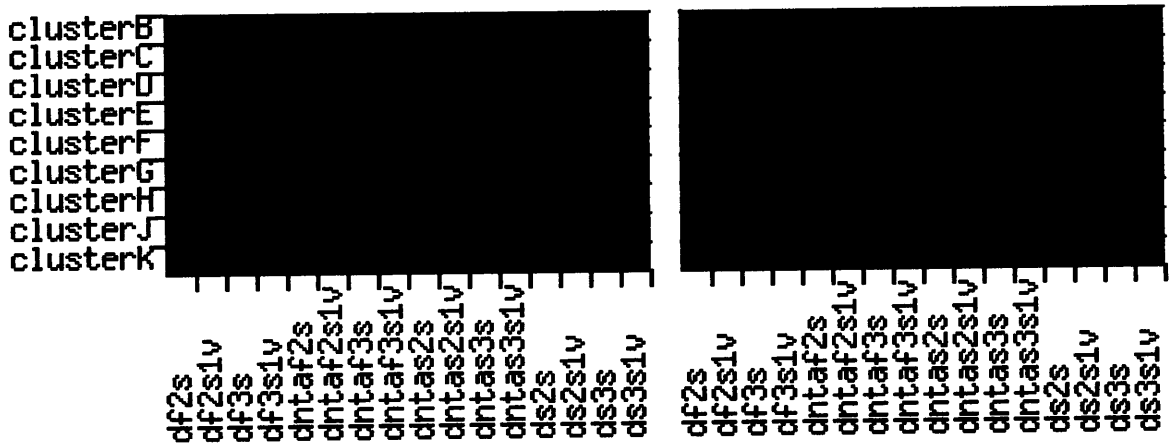
A)



B)



C)


clusterB
clusterC
clusterD
clusterE
clusterF
clusterG
clusterH
clusterJ
clusterK
df2s
df2s1v
df3s
df3s1v
dntaf2s
dntaf2s1v
dntaf3s
dntaf3s1v
dntas2s
dntas2s1v
dntas3s
dntas3s1v
ds2s
ds2s1v
ds3s
ds3s1v

D)

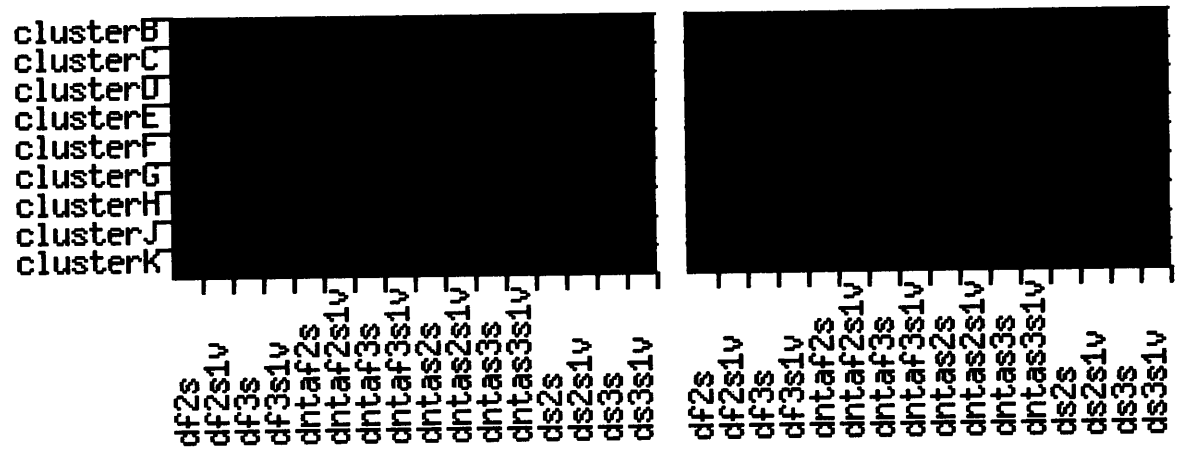

clusterB
clusterC
clusterD
clusterE
clusterF
clusterG
clusterH
clusterJ
clusterK
df2s
df2s1v
df3s
df3s1v
dntaf2s
dntaf2s1v
dntaf3s
dntaf3s1v
dntas2s
dntas2s1v
dntas3s
dntas3s1v
ds2s
ds2s1v
ds3s
ds3s1v

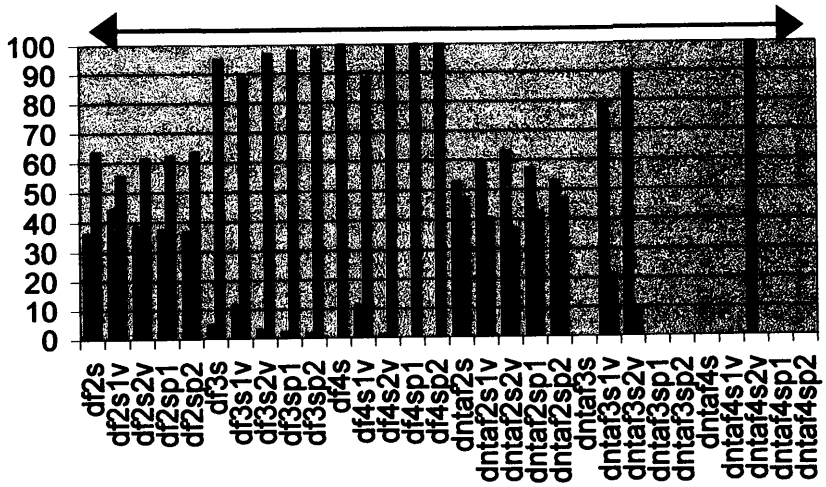
Figure A4.4. Results for absolute position analysis of Northern clusters (Brown *et al.* 2001) using the full binding site dataset.

See Figure A4.1 for details.

A)

HSP12

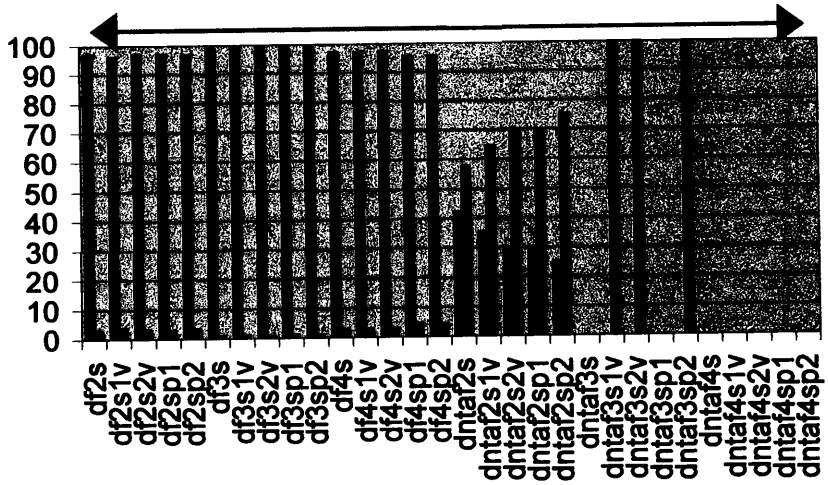
Cluster



B)

CAR1

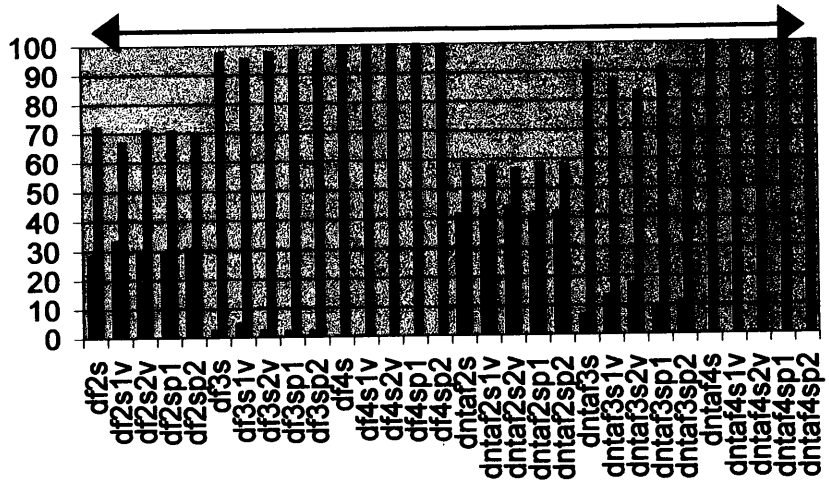
Cluster



C)

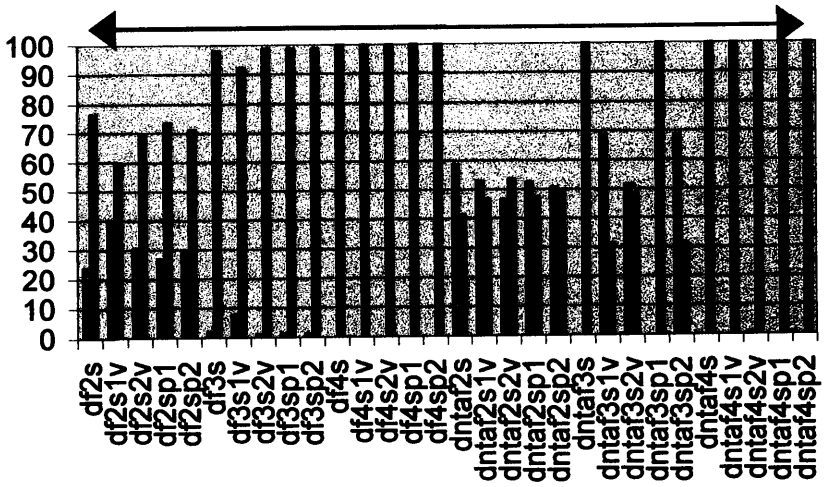
RPL25

Cluster



D)

Cluster 11



E)

Cluster 12

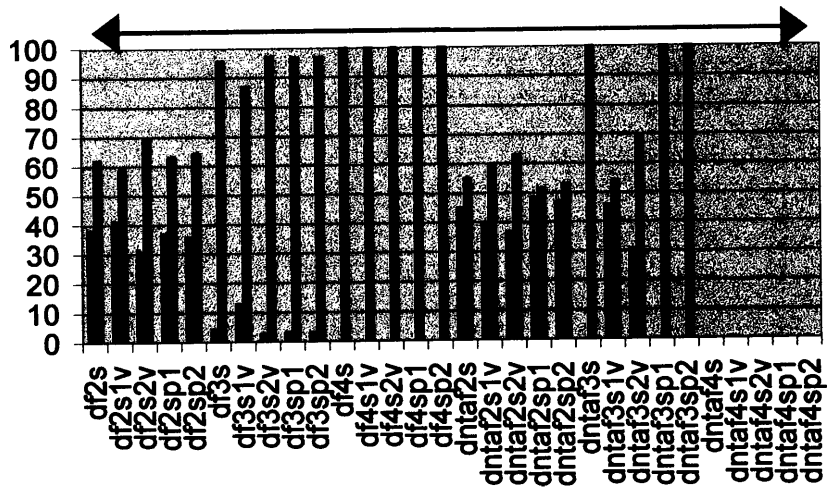


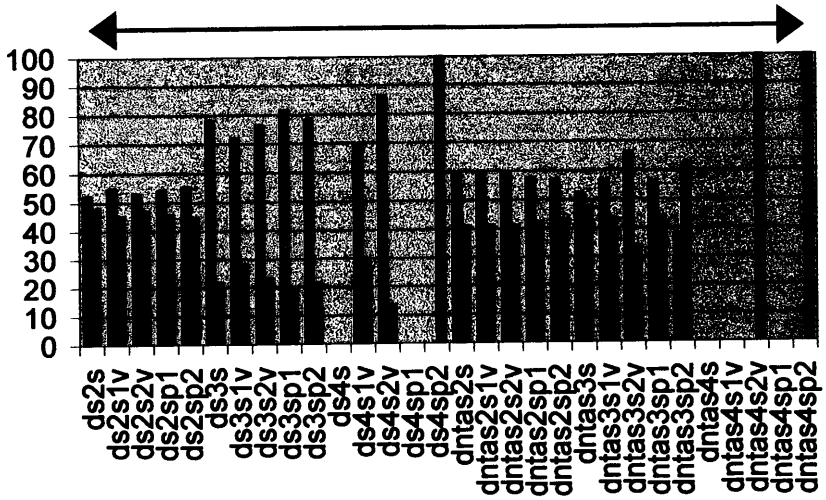
Figure A4.5. Results for absolute position analysis of Northern clusters (Brown *et al.* 2001) using the simplified binding site dataset.

See Figure A4.1 for details.

A)

HSP12

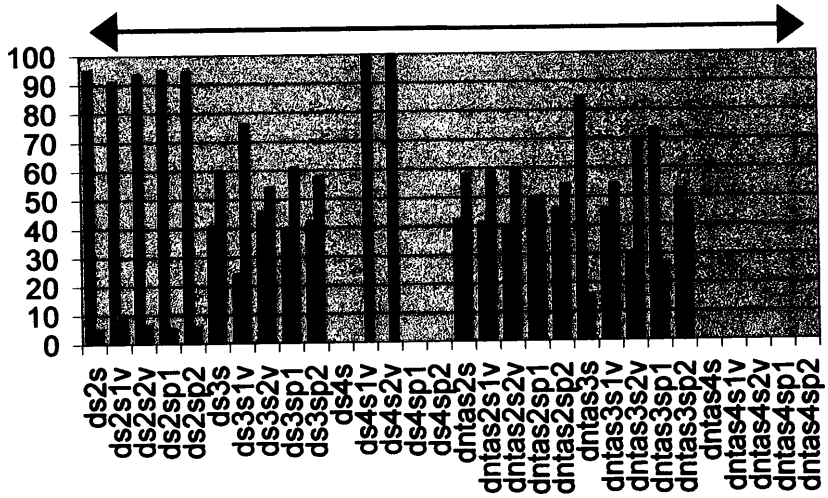
Cluster



B)

CAR1

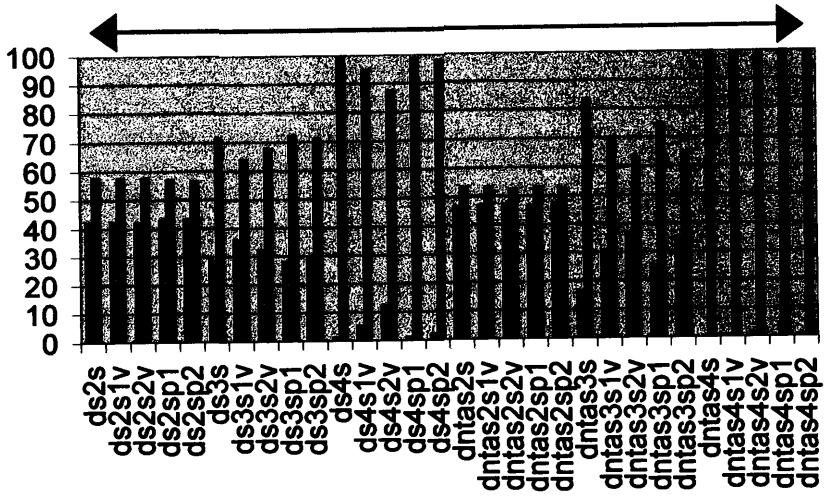
Cluster



C)

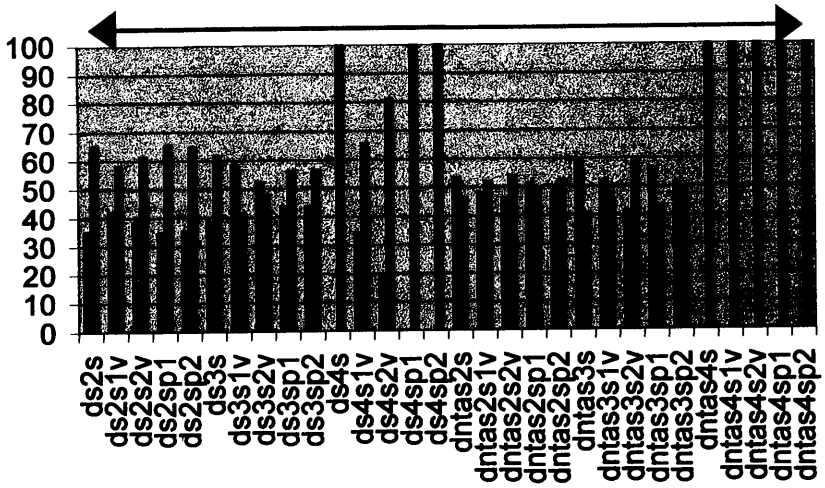
RPL25

Cluster



D)

Cluster 11



E)

Cluster 12

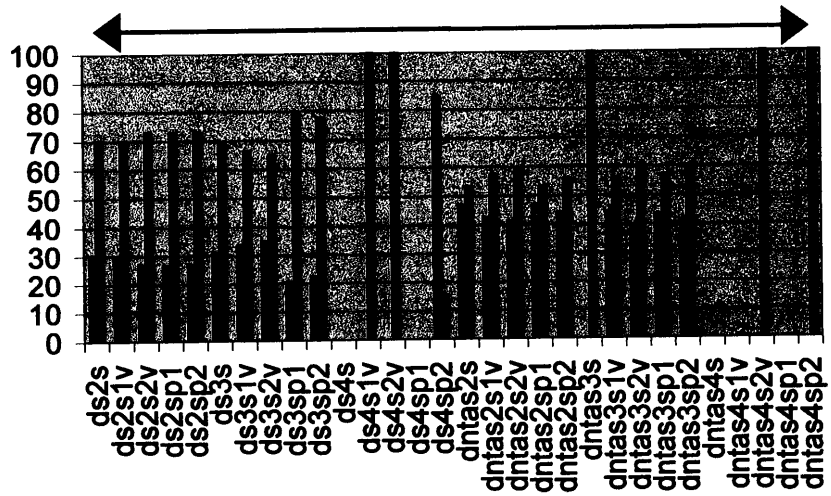
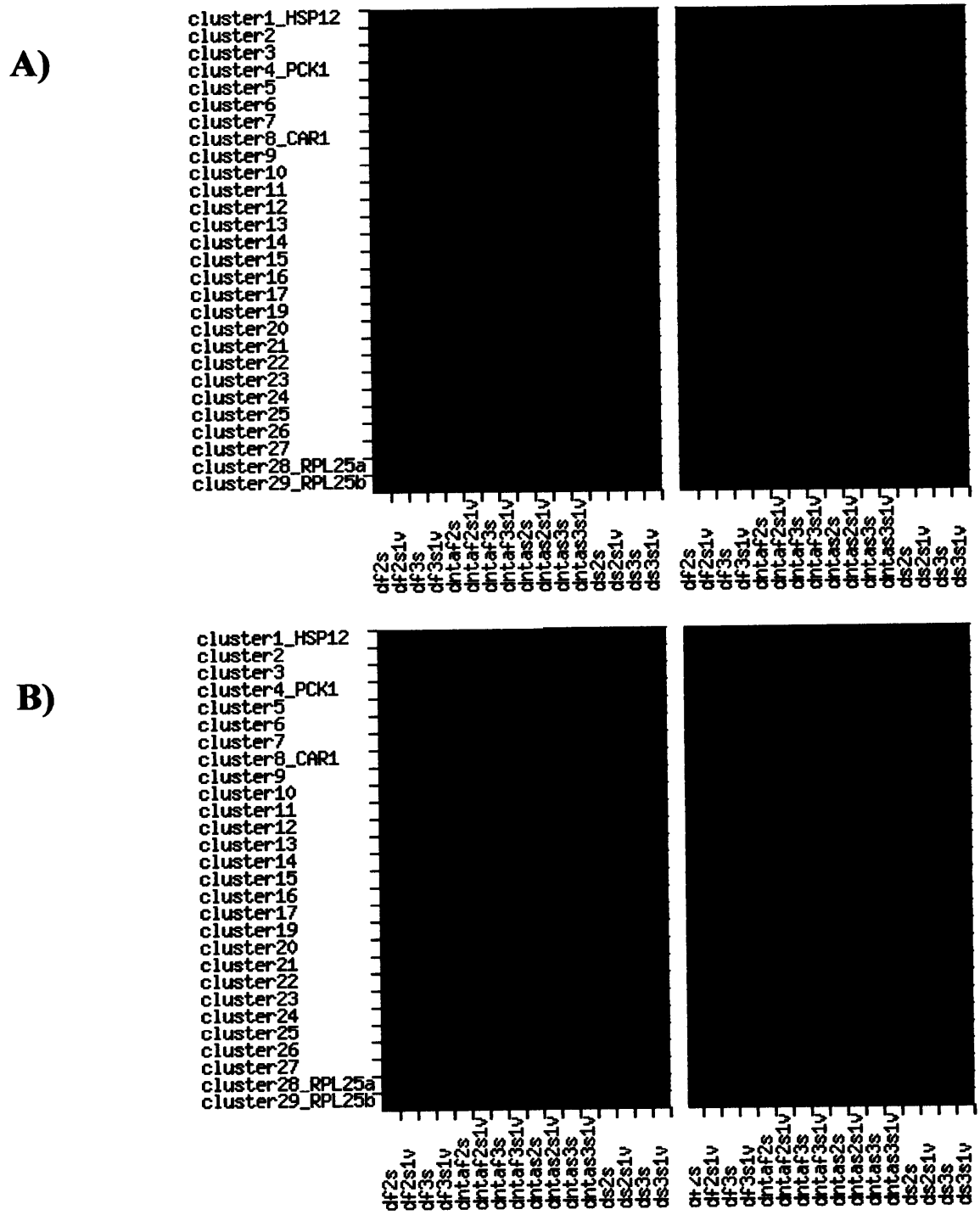


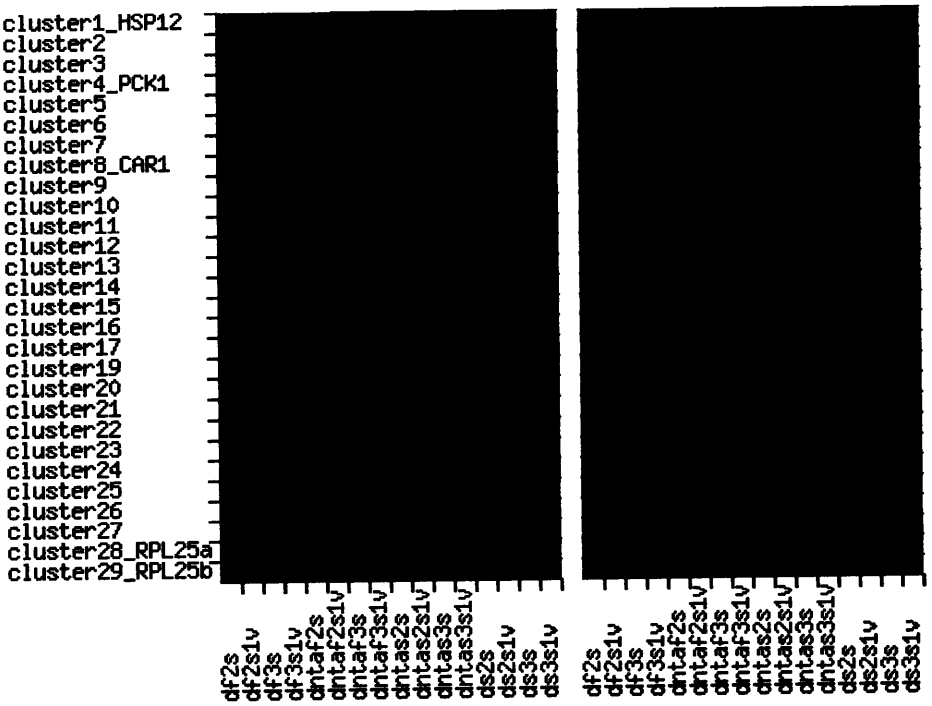
Figure A4.6. Results of Inter to Intra cluster comparison for absolute position analysis of Northern clusters (Brown *et al.*2001).

A) No distance constraints B) Distance constraint of 3 bases C) Distance constraint of 6 bases D) Low-complexity sites masked, no distance constraints

In the left hand image of each pair, the results for each cluster (row) are scaled independently. This highlights the analyses that perform the best and worst for each cluster. In the right hand image the scaling is global, which highlights those clusters with the highest number of intra-cluster links in comparison to inter-cluster links.



C)


cluster1_HSP12
cluster2
cluster3
cluster4_PCK1
cluster5
cluster6
cluster7
cluster8_CAR1
cluster9
cluster10
cluster11
cluster12
cluster13
cluster14
cluster15
cluster16
cluster17
cluster19
cluster20
cluster21
cluster22
cluster23
cluster24
cluster25
cluster26
cluster27
cluster28_RPL25a
cluster29_RPL25b
df2s
df2s1v
df3s
df3s1v
dntaf2s
dntaf2s1v
dntaf3s
dntaf3s1v
dntas2s
dntas2s1v
dntas3s
dntas3s1v
ds2s
ds2s1v
ds3s
ds3s1v

D)

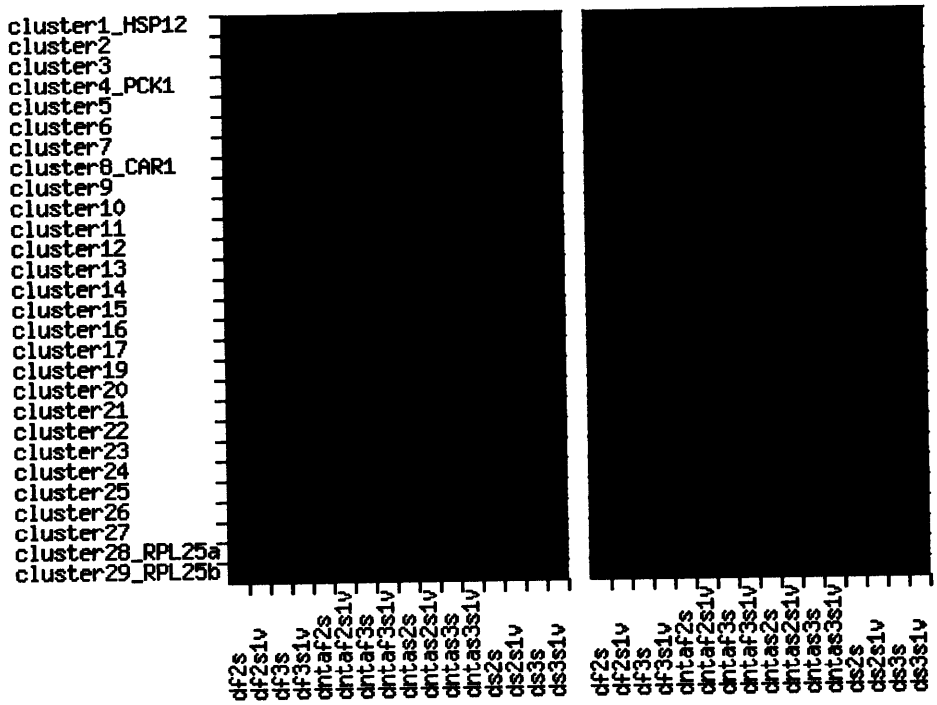

cluster1_HSP12
cluster2
cluster3
cluster4_PCK1
cluster5
cluster6
cluster7
cluster8_CAR1
cluster9
cluster10
cluster11
cluster12
cluster13
cluster14
cluster15
cluster16
cluster17
cluster19
cluster20
cluster21
cluster22
cluster23
cluster24
cluster25
cluster26
cluster27
cluster28_RPL25a
cluster29_RPL25b
df2s
df2s1v
df3s
df3s1v
dntaf2s
dntaf2s1v
dntaf3s
dntaf3s1v
dntas2s
dntas2s1v
dntas3s
dntas3s1v
ds2s
ds2s1v
ds3s
ds3s1v

Figure A4.7. Absolute position results for MIPS functional categories.

A) Broad MIPS categories
B) Selected specific MIPS categories.
Columns headings dntancf and dntancs refer to experiments where non-categorised genes were removed from the analysis.

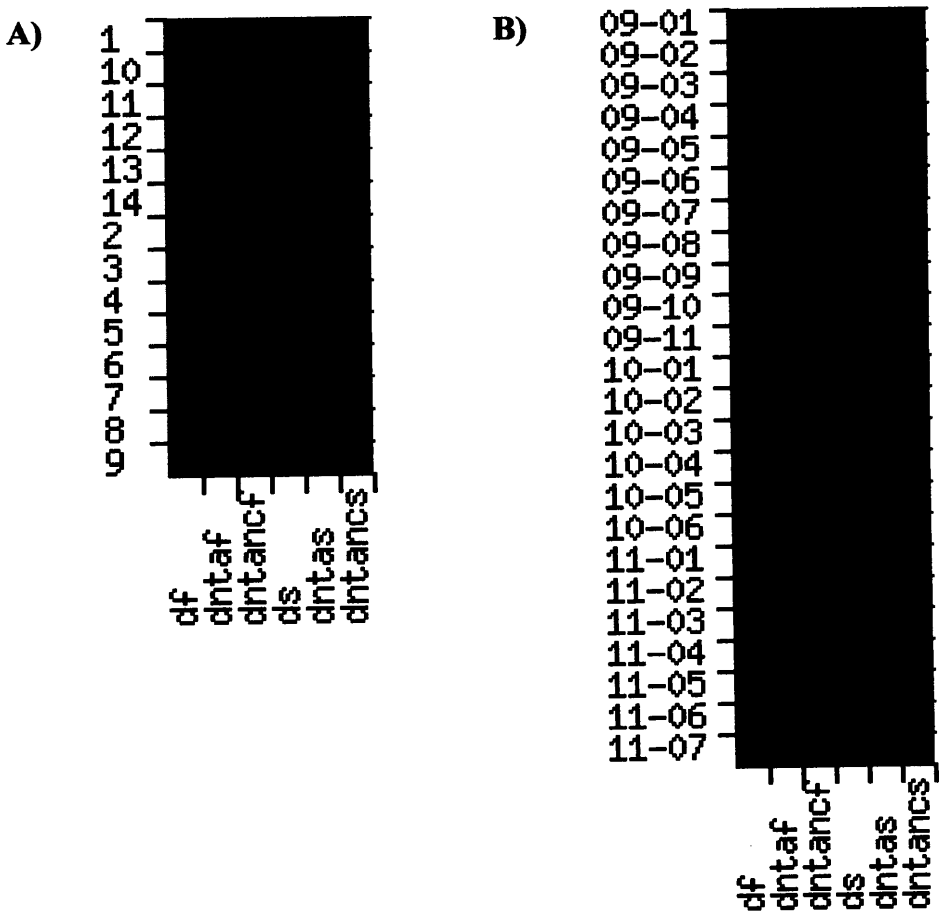


Figure A4.8. Absolute position results for KEGG functional categories.

A) Broad KEGG categories
B) Selected specific KEGG categories.
Columns headings dntancf and dntancs refer to experiments where non-categorised genes were removed from the analysis.

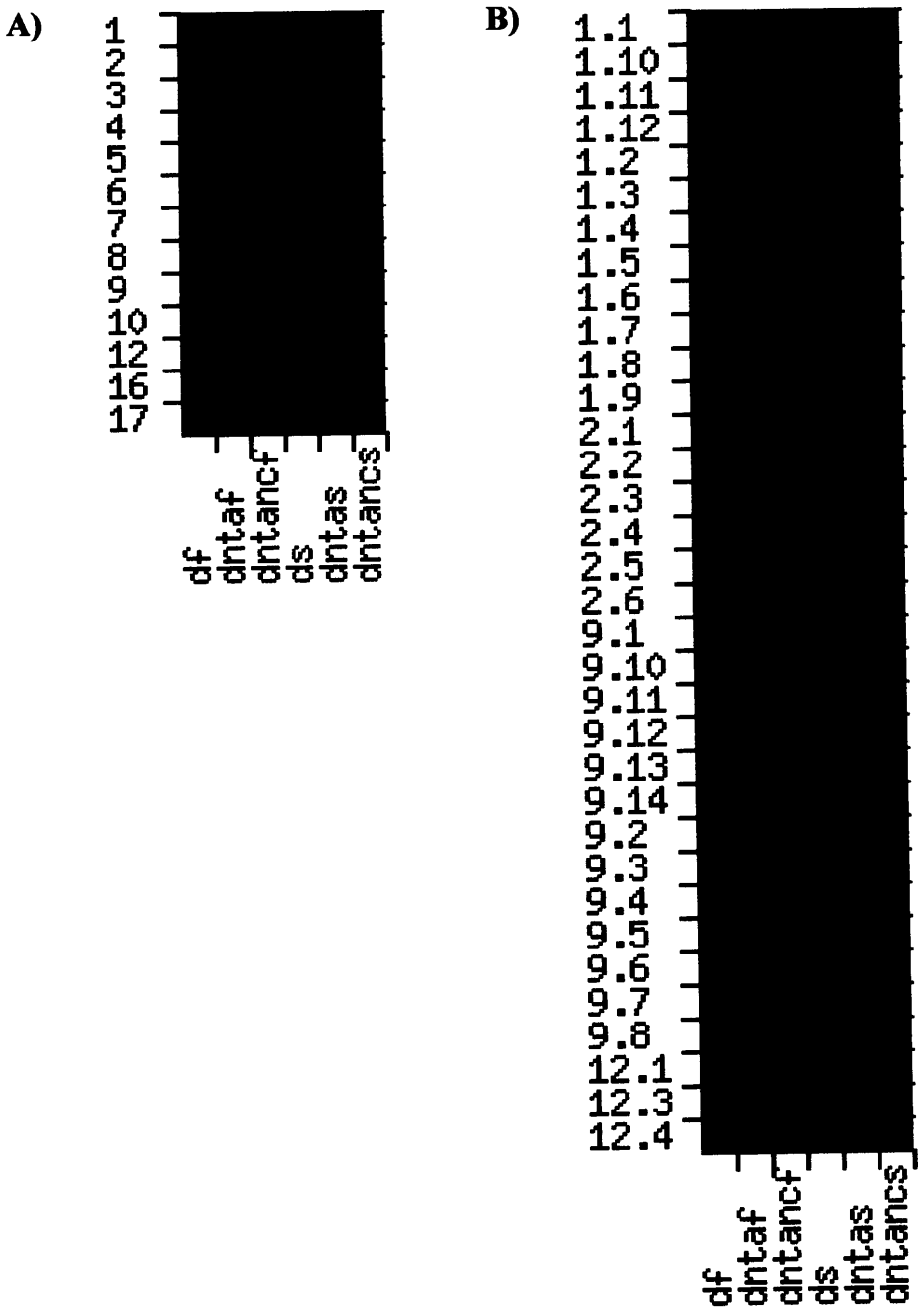


Figure A4.9. Relative position results for MIPS functional categories.

A) Broad MIPS categories
B) Selected specific MIPS categories.
Columns headings dntancf and dntancs refer to experiments where non-categorised genes were removed from the analysis.

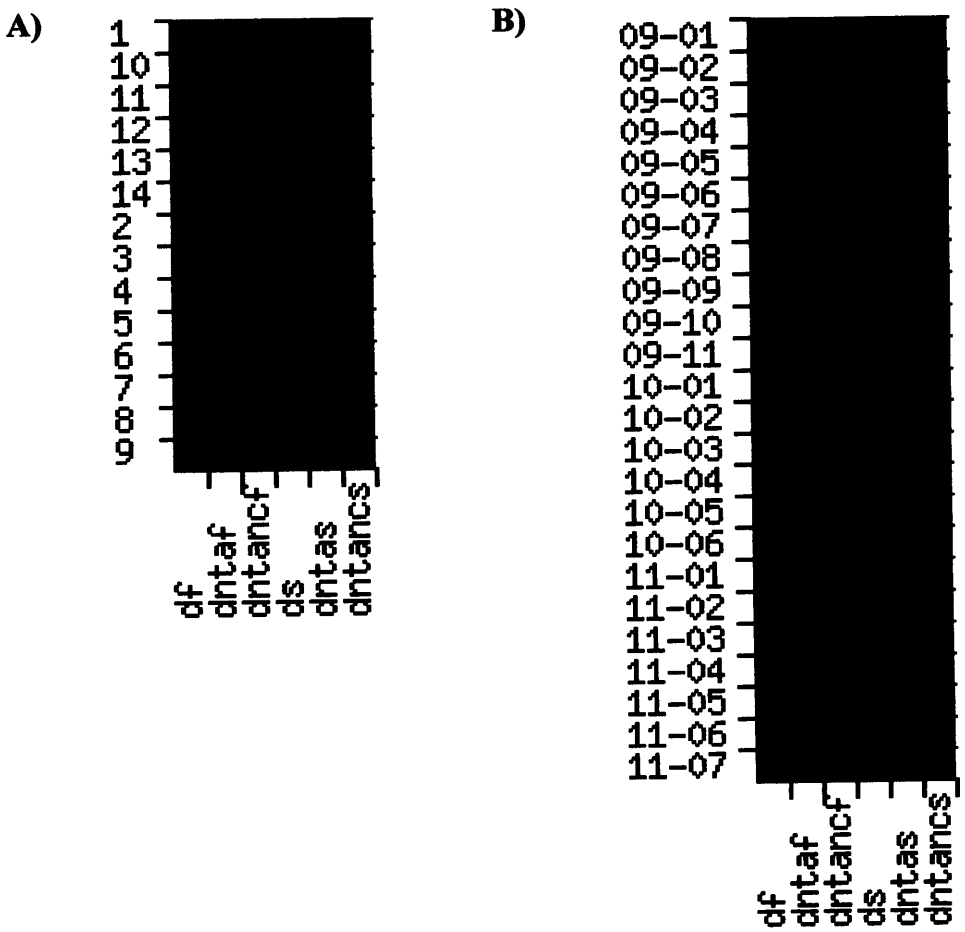


Figure A4.10. Relative position results for KEGG functional categories.

A) Broad KEGG categories
B) Selected specific KEGG categories.
Columns headings dntancf and dntancs refer to experiments where non-categorised genes were removed from the analysis.

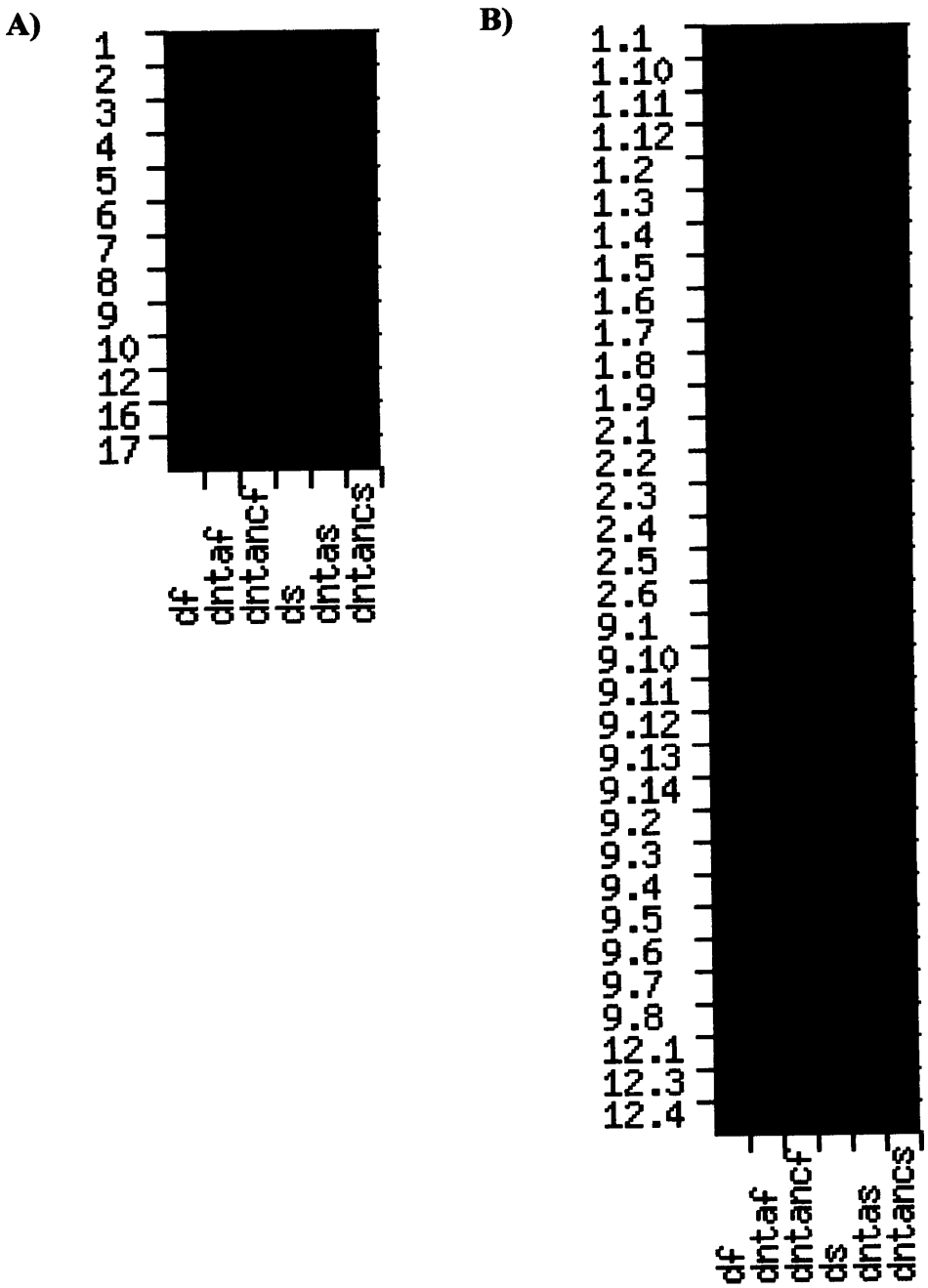


Table A4.11. Top ten binding site relationships found for MIPS functional category 10 using the full binding site dataset.

Binding site relationship	Number found
GAL1-10-c->TATA,TBP(7)	946
GATA(1)->GAL1-10-i-Rev	496
GATA(3)-Rev->GAL1-10-i-Rev	435
TATA,TBP(7)-Rev->TATA,TBP(7)	406
TATA,TBP(5)->GAL1-10-c	253
TATA,TBP(2)-Rev->TATA,TBP(7)-Rev	253
GATA(5)-Rev->GAL1-10-i-Rev	231
TATA,TBP(4)-Rev->TATA,TBP(18)-Rev	136
CHA1-d,TATA,TBP-Rev->GAL1-10-c-Rev&TATA,TBP(5)-Rev&TATA,TBP(7)-Rev	91
TATA,TBP(15)-Rev->TATA,TBP(7)-Rev	78

Table A4.12. Top ten binding site relationships found for MIPS functional category 10 using the full binding site dataset with TATA sites removed.

Binding site relationship	Number found
GATA(1)->GAL1-10-i-Rev	496
GATA(3)-Rev->GAL1-10-i-Rev	435
GATA(5)-Rev->GAL1-10-i-Rev	231
GAL1-10-i-Rev->GAL1-10-h-Rev	66
GATA(3)-Rev&Ty2-917-c->GAL1-10-i-Rev	45
GAL1-10-i-Rev->GAL1-10-c	36
BAS2(4)-Rev->GAL1-10-h	28
GAL1-10-i-Rev->BAS2(3)-Rev	28
BAS2(4)-Rev->GAL1-10-c	21
GAL1-10-h-Rev->GAL1-10-h-Rev	21

Table A4.13. Top ten binding site relationships found for MIPS functional category 10 using the simplified binding site dataset.

Binding site relationship	Number found
TATA,TBP-Rev->TATA,TBP-Rev	1485
TATA,TBP-Rev->TATA,TBP	990
GAL1-10-c->TATA,TBP	946
GATA-Rev->GAL1-10-i-Rev	946
TATA,TBP->TATA,TBP-Rev	595
GATA->GAL1-10-i-Rev	496
TATA,TBP->GAL1-10-c	300
TATA,TBP-Rev->GAL1-10-c-Rev&TATA,TBP-Rev	276
TATA,TBP-Rev->TATA,TBP&TATA,TBP	253
GAL1-10-i-Rev->TATA,TBP-Rev	171

Table A4.14. Top ten binding site relationships found for MIPS functional category 10 using the simplified binding site dataset with TATA sites removed.

Binding site relationship	Number found
GATA-Rev->GAL1-10-i-Rev	946
GATA->GAL1-10-i-Rev	528
BAS2-Rev->GATA-Rev	105
GAL1-10-i-Rev->BAS2-Rev	91
GAL1-10-i-Rev->GAL1-10-h-Rev	66
GATA-Rev&Ty2-917-c->GAL1-10-i-Rev	45
BAS2->GAL1-10-i&GATA	36
GAL1-10-i-Rev->GAL1-10-c	36
GAL1-10-h-Rev->BAS2-Rev	36
BAS2-Rev->GAL1-10-h	36

Figure A4.15. SiteSeer visualisation of cluster 4 generated from the northern data produced by Brown *et al.* (2001).

A) Visualisation using minimum expectation ratio = 4, maximum expectation value = 0.5, minimum number of occurrences = 3.
B) Visualisation using minimum expectation ratio = 10, maximum expectation value = 0.3, minimum number of occurrences = 5.

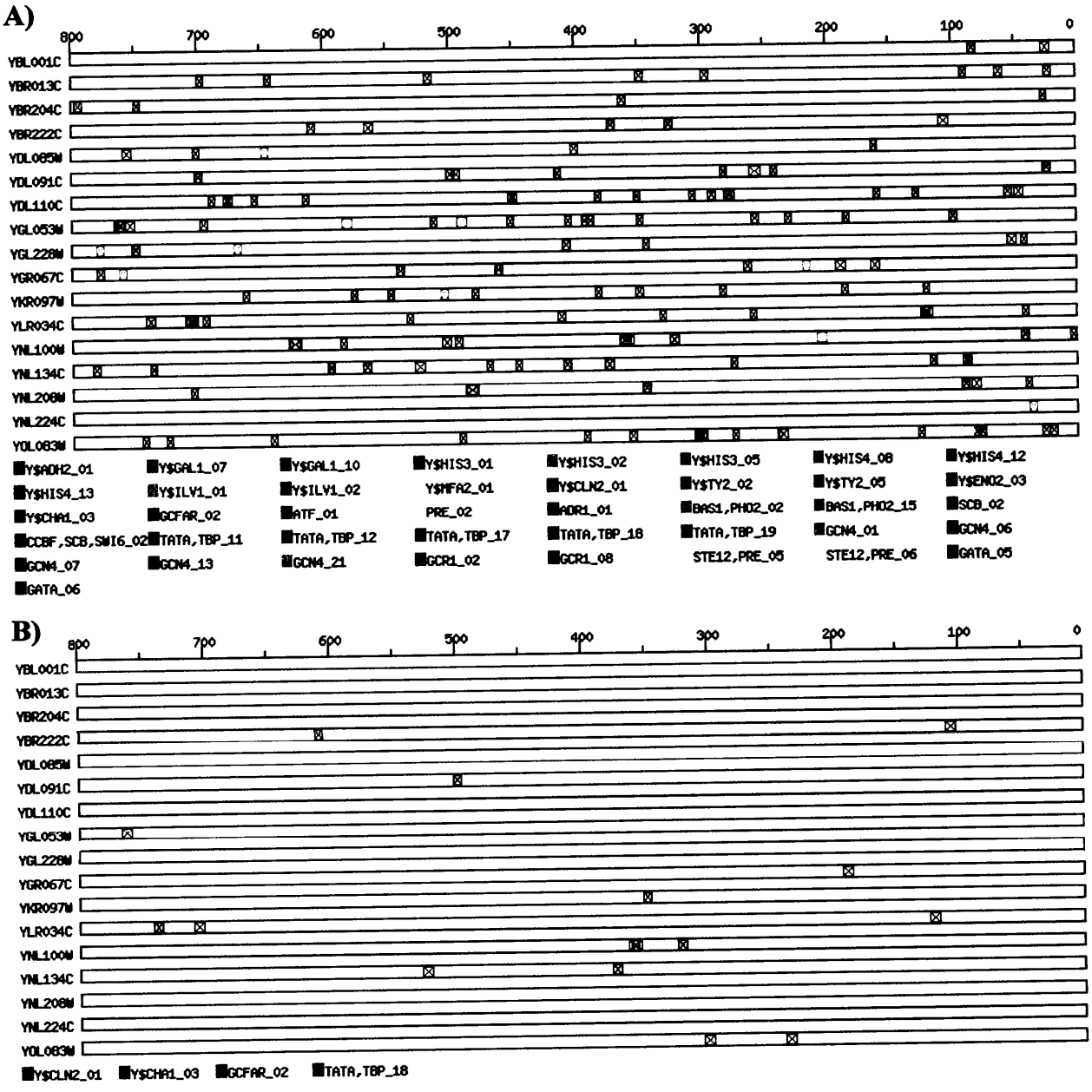
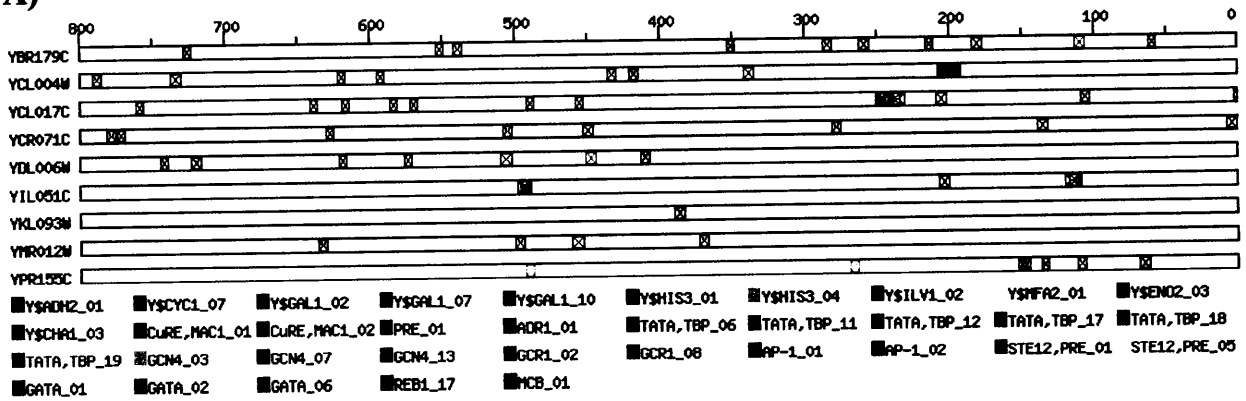


Figure A4.16. SiteSeer visualisation of the URSs of genes from the MIPS functional category "Mitochondrial biogenesis" (9-9).

A) Visualisation using minimum expectation ratio = 4, maximum expectation value = 0.5, minimum number of occurrences = 1.
B) Visualisation using minimum expectation ratio = 10, maximum expectation value = 0.3, minimum number of occurrences = 3.

A)



B)

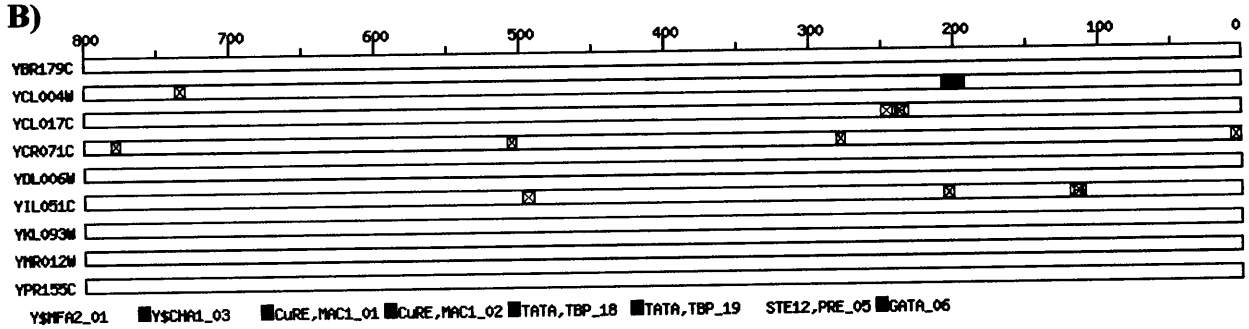


Figure A4.17. SiteSeer visualisation of the URSs of genes from the MIPS functional category "Nutritional response pathway" (10-04).

A) Visualisation using minimum expectation ratio = 4, maximum expectation value = 0.5, minimum number of occurrences = 4.
B) Visualisation using minimum expectation ratio = 10, maximum expectation value = 0.3, minimum number of occurrences = 7.

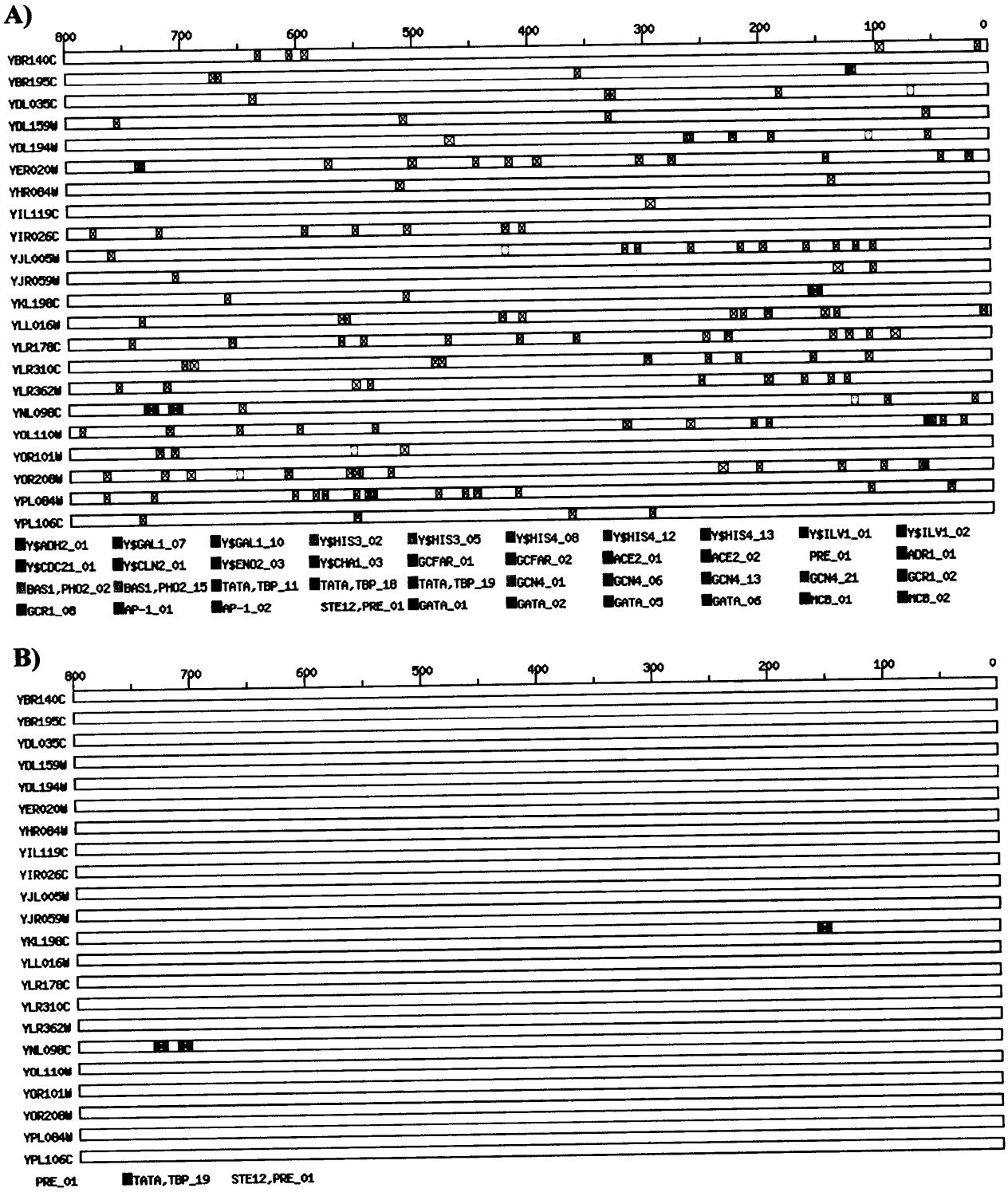


Figure A4.18. SiteSeer visualisation of the URSs of genes from the KEGG functional category "Methane metabolism" (2.2).

A) Visualisation using minimum expectation ratio = 4, maximum expectation value = 0.5, minimum number of occurrences = 1.
B) Visualisation using minimum expectation ratio = 10, maximum expectation value = 0.3, minimum number of occurrences = 3.

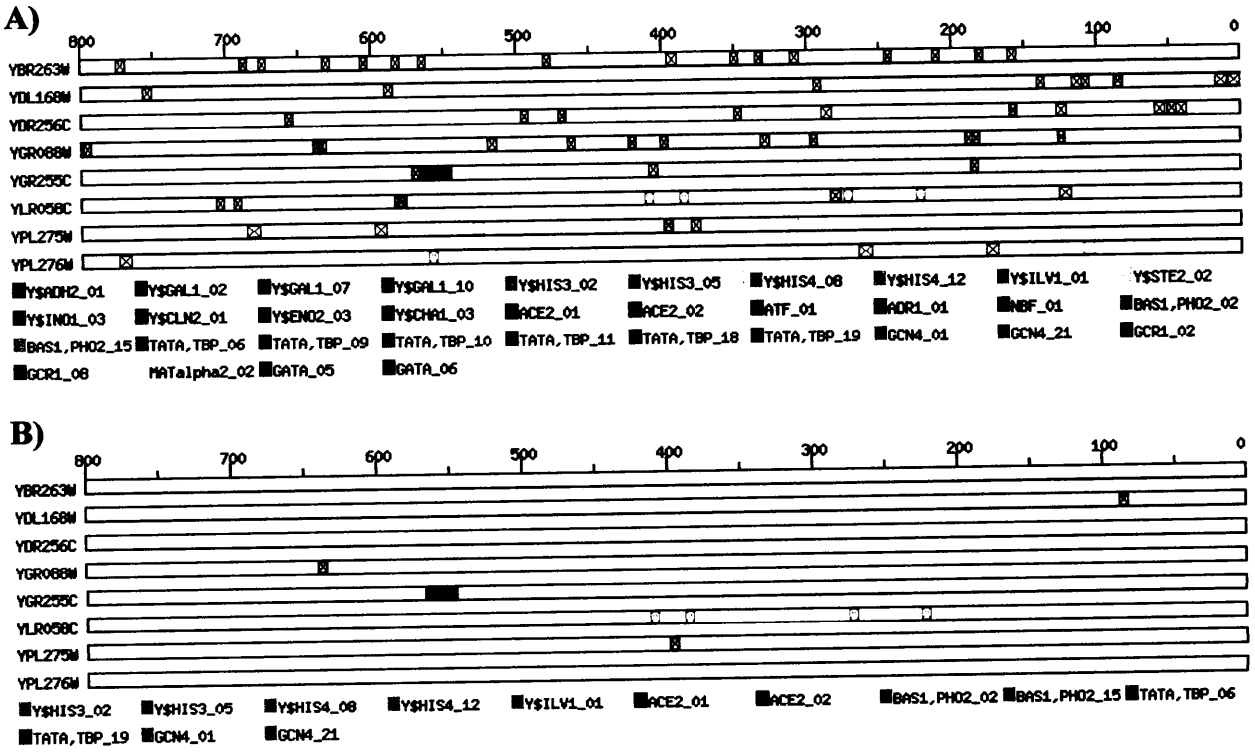
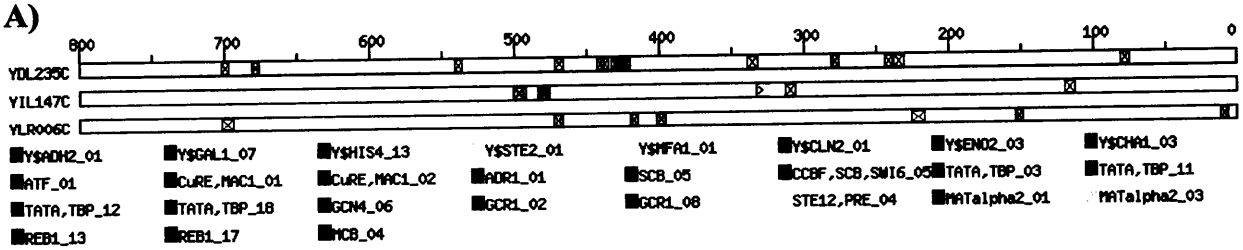


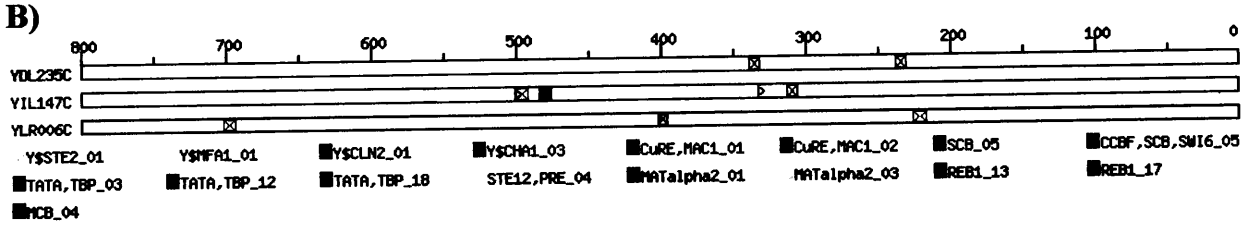
Figure A4.19. SiteSeer visualisation of the URSs of genes from the KEGG functional category "Two-component system" (12.1).

A) Visualisation using minimum expectation ratio = 4, maximum expectation value = 0.5, minimum number of occurrences = 1.
B) Visualisation using minimum expectation ratio = 10, maximum expectation value = 0.3, minimum number of occurrences = 1.

A)



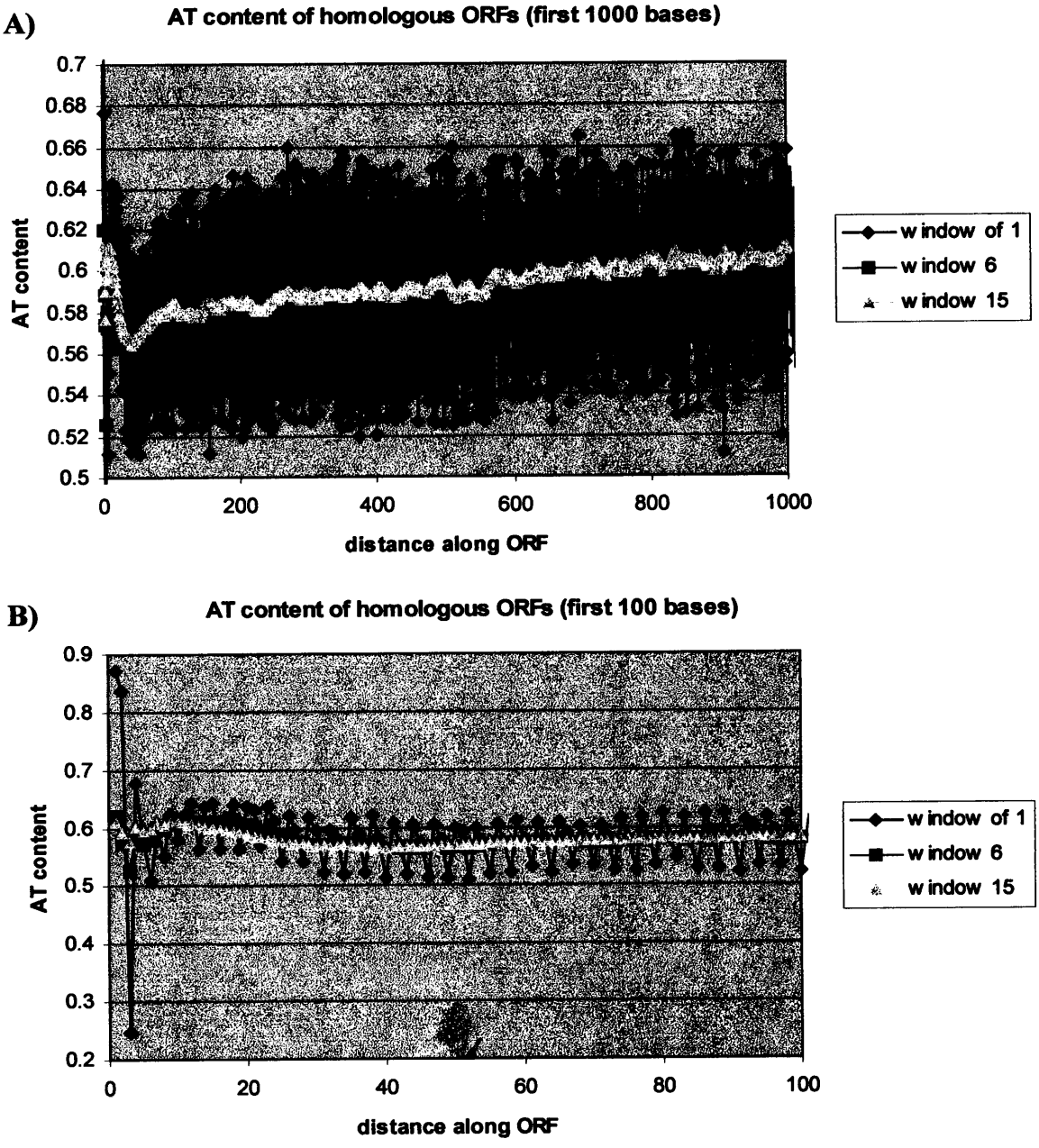
B)



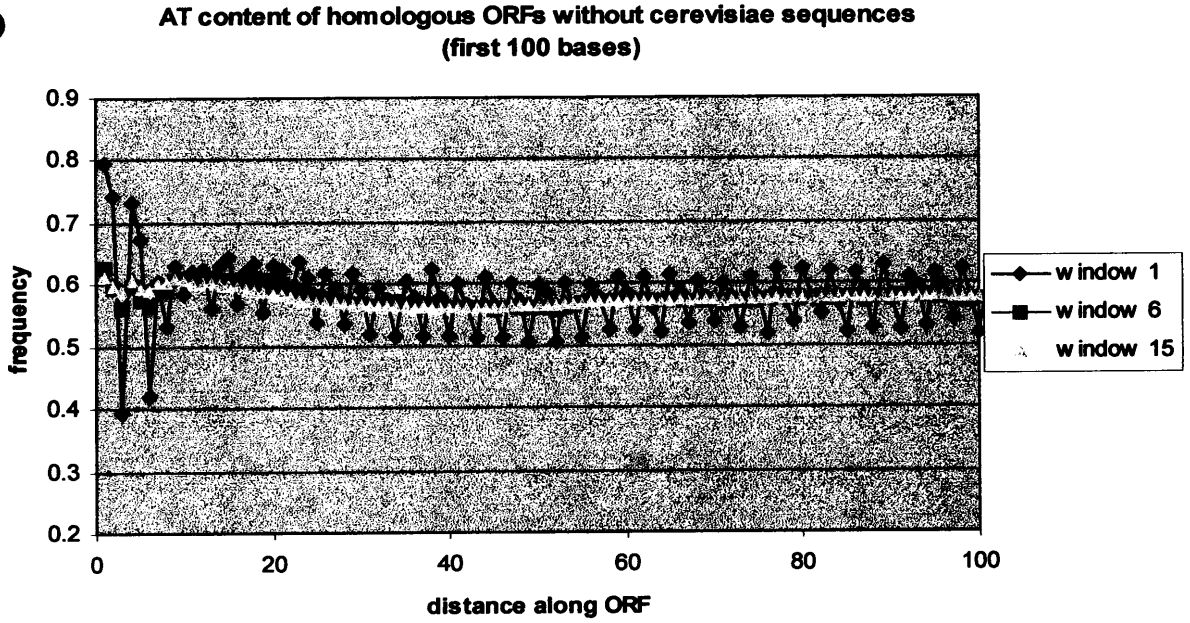
Appendix 5

Figure A5.1. AT-Content of aligned homologous open reading frames from *Saccharomyces cerevisiae* and the Génolevures data.

These plots were generated using the ***calculate_positoinal_at_content.pl*** program. Plot A) shows the AT content for the first 1000 bases, plot B) is a magnified view of the AT content over the first 100 bases of the open reading frames and plot C) is identical to B) but with *S. cerevisiae* sequences removed. The 3 nt periodicity seen here is in agreement with the internal codon nucleotide bias of *Saccharomyces cerevisiae* (1st letter GC 44.6%, 2nd letter GC 36.57%, 3rd letter GC 37.9%, web ref 37).



C)


AT content of homologous ORFs without cerevisiae sequences
(first 100 bases)
frequency
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
20
40
60
80
100
distance along ORF
window 1
window 6
window 15

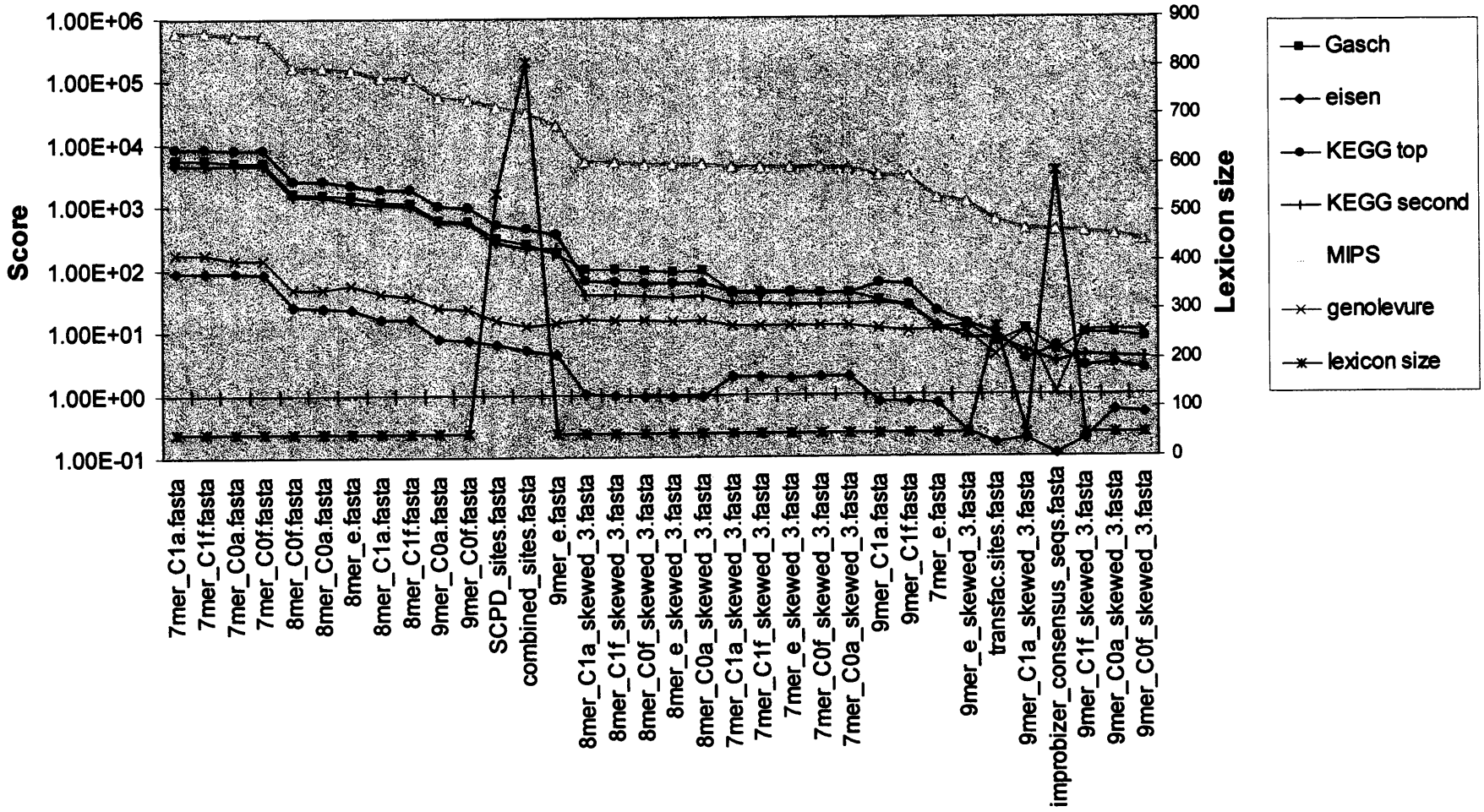


Figure A5.2.1. Lexicon size normalised scores.

This graph is ordered by lexicon score for the MIPS functional groupings. The lexicon 'consensus_sequences' refers to the set of Improbizer consensus sequences.

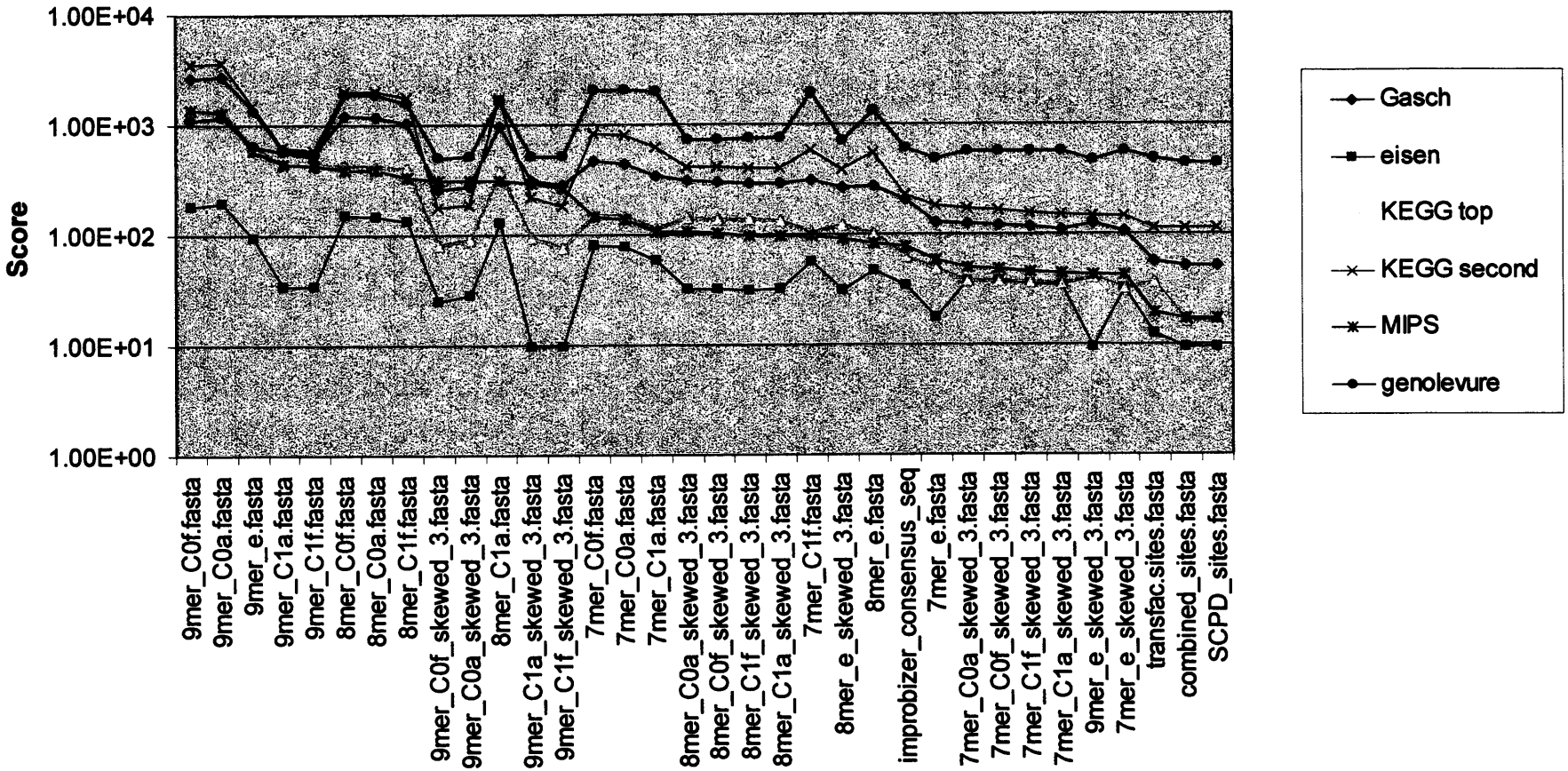


Figure A5.2.2. Randomised URS normalised scores.

This graph is ordered by lexicon score for the MIPS functional groupings. The lexicon 'consensus_sequences' refers to the set of Improbizer consensus sequences.

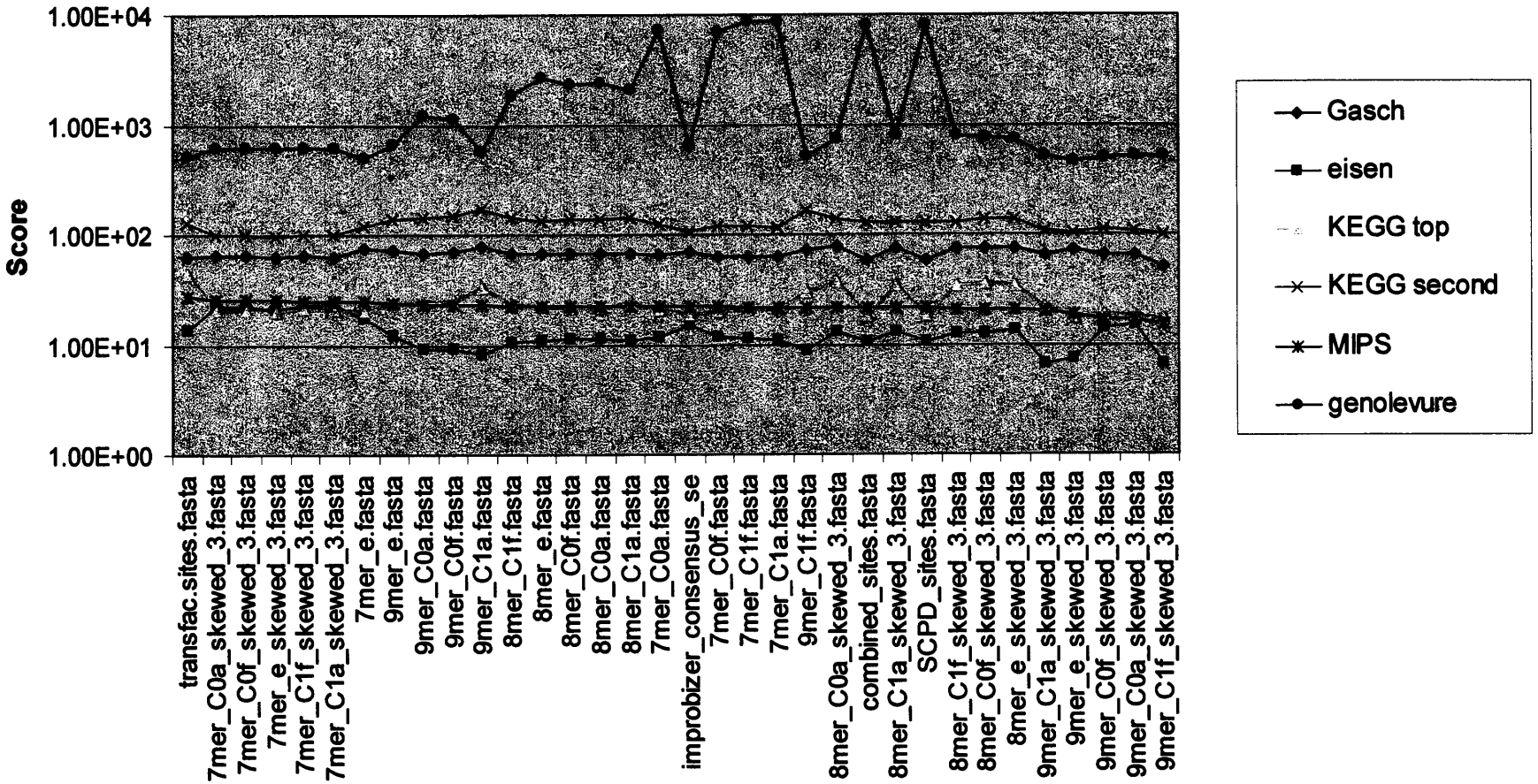


Figure A5.2.3. Randomly picked URS normalised scores.

This graph is ordered by lexicon score for the MIPS functional groupings. The lexicon 'consensus_sequences' refers to the set of Improbizer consensus sequences.

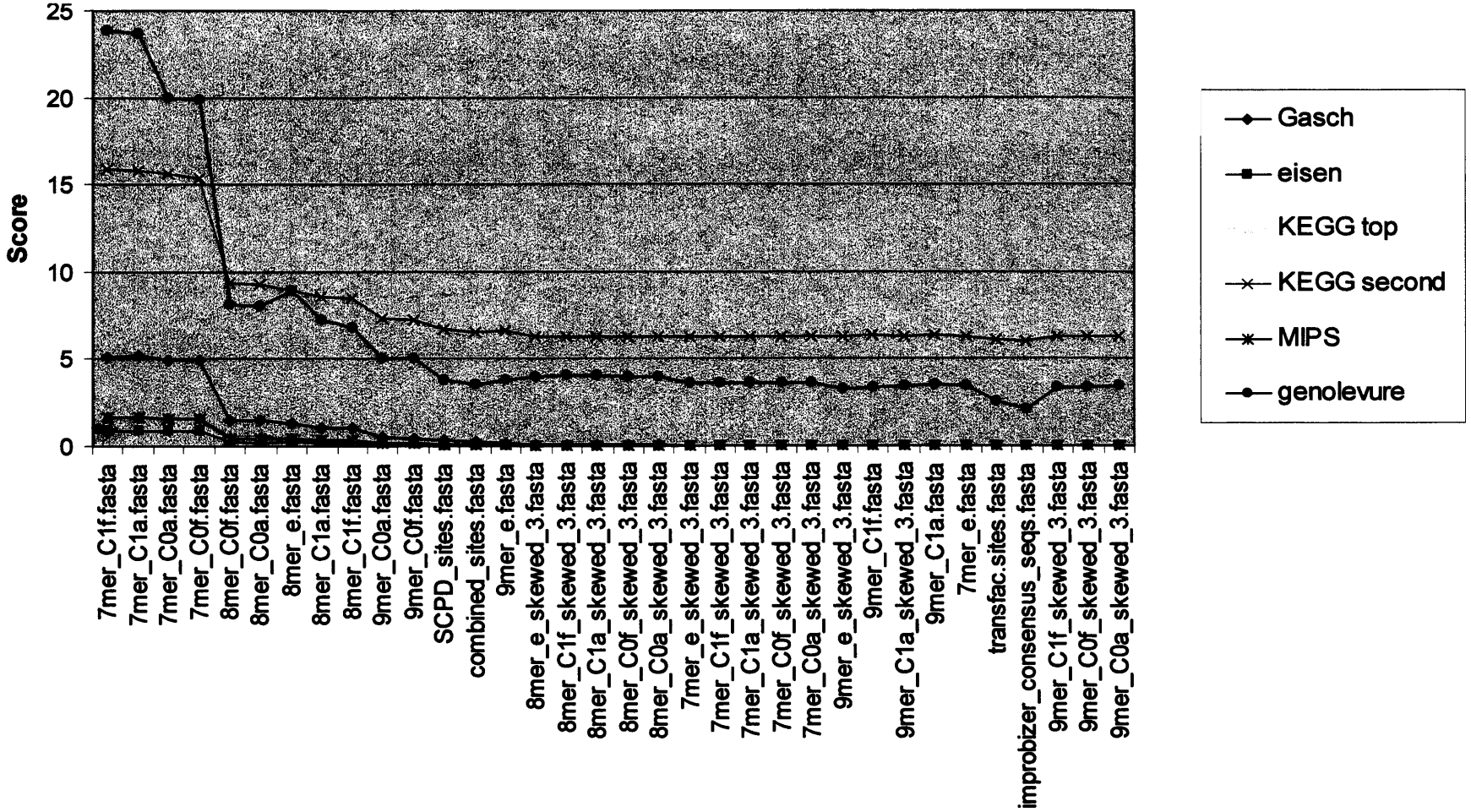


Figure A5.2.4. Search-space normalised scores.

This graph is ordered by lexicon score for the MIPS functional groupings. The lexicon 'consensus_sequences' refers to the set of Improbizer consensus sequences.

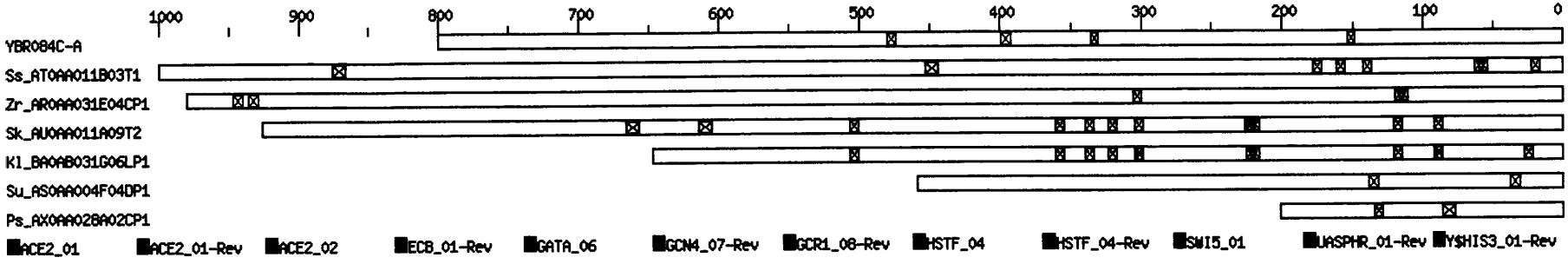


Figure A5.3.1. SiteSeer visualisation of the YBR084C-A orthologue set.

Strong conservation of binding site order can be seen between *Saccharomyces servazzi*, *Saccharomyces kluyveri* and *Kluyveromyces marxianus* var. *marxianus*. There is also a weaker conservation between *Zygosaccharomyces rouxii* and the aforementioned species.

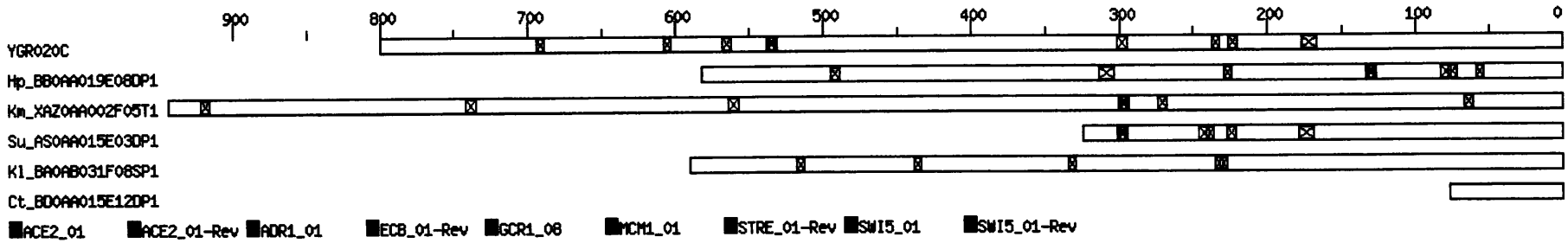


Figure A5.3.2. SiteSeer visualisation of the YGR020C orthologue set.

Conservation of binding site order can be seen between *Saccharomyces cerevisiae*, *Saccharomyces uvarum* and *Hansenula polymorpha*.

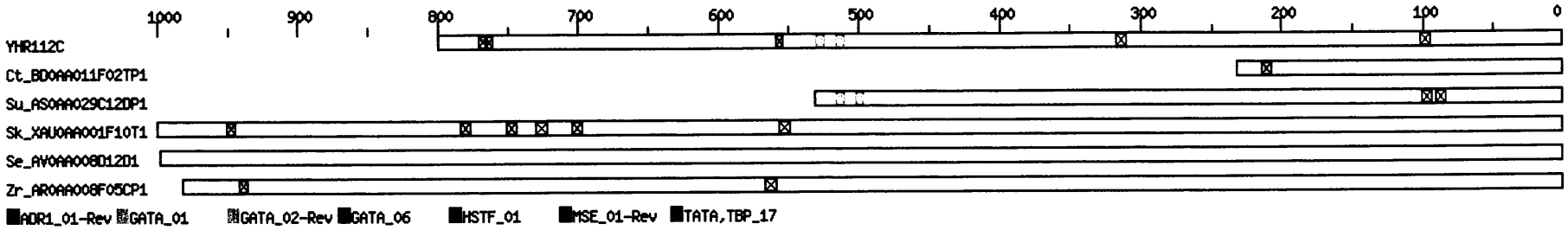


Figure A5.3.3. SiteSeer visualisation of the YHR112C orthologue set.

Conservation of binding site order can be seen between *Saccharomyces kluyveri* and *Zygosaccharomyces rouxii*. There is also conservation apparent between *Saccharomyces cerevisiae* and *Saccharomyces uvarum*.

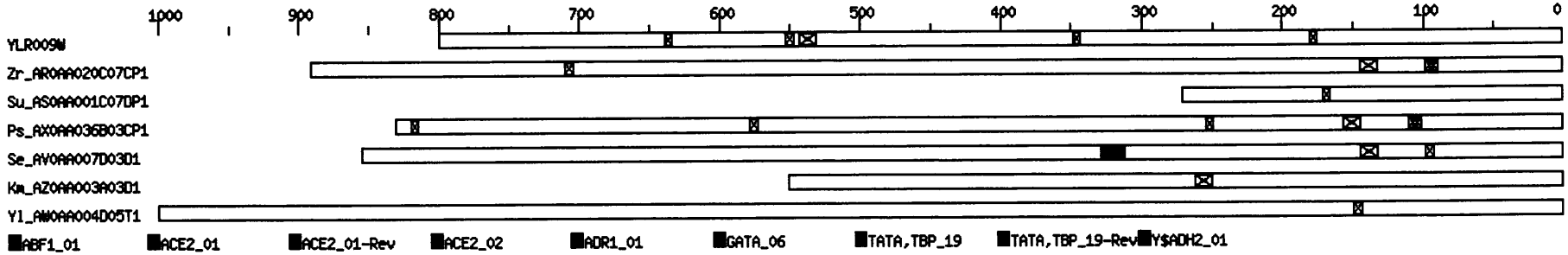


Figure A5.3.4. SiteSeer visualisation of the YLR009W orthologue set.

Conservation of binding site order can be seen between *Zygosaccharomyces rouxii*, *Pichia sorbitophila* and *Saccharomyces exiguus*.

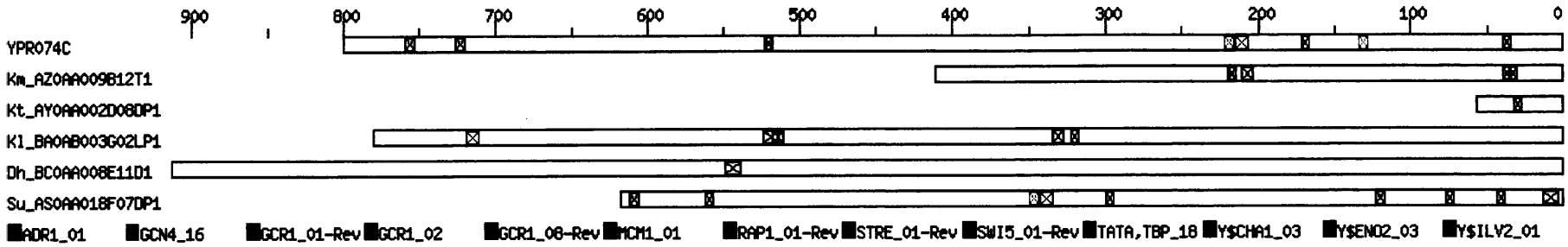


Figure A5.3.5. SiteSeer visualisation of the YPR074C orthologue set.

Conservation of binding site order can be seen between *Saccharomyces cerevisiae* and *Saccharomyces uvarum*.

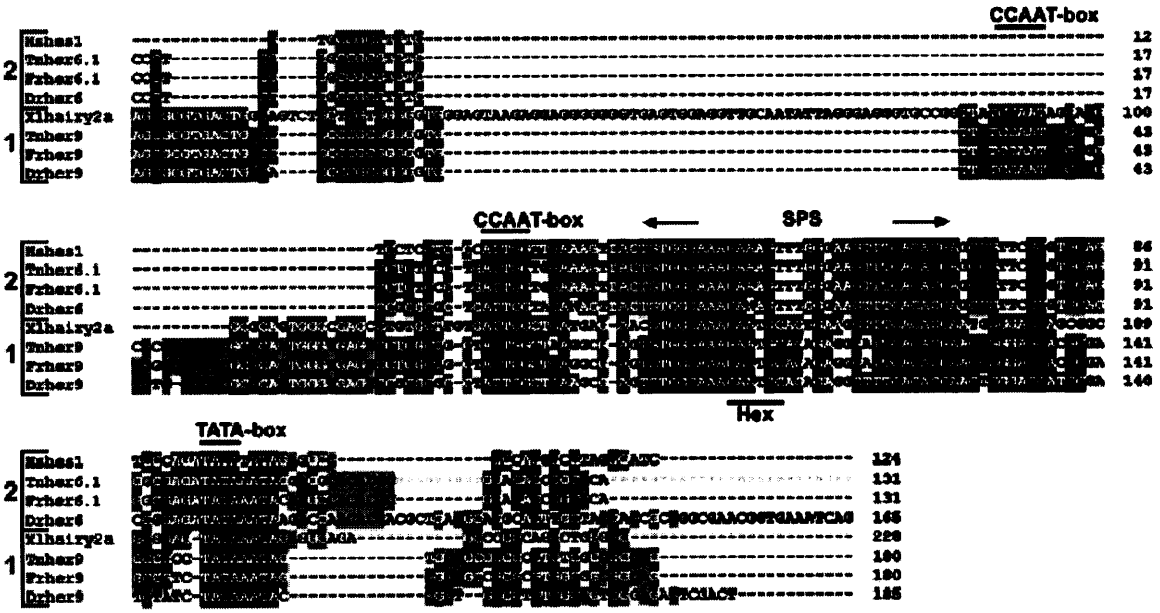


Figure A5.3.6. Alignment of the upstream sequences of *c-hairy* class genes (Taken from Gajewski 2002).

Conservation levels: 75% or more conserved nucleotides are marked in black, 50% or more conserved nucleotides are marked in dark grey and nucleotides conserved to more than 25% are indicated in light grey.