Visualising and exploring linked functional genomic data sets in the Yeast Exploration Tool Integrator (YETI)

Richard J. Orton

Biocomputing Research Unit
Institute of Cell and Molecular Biology
University of Edinburgh

Thesis presented for the degree of Doctor of Philosophy
2006

Declaration

I declare that this thesis was composed by myself and the work presented herein is my own other than when referenced to others. The work presented herein has not been submitted for any other degree or professional qualification except as specified; some of the work has already been published.

Richard J. Orton

Date: 17th May 2006

To My Family

And

To Morag

Acknowledgements

I firstly have to thank my supervisors Dietlind Gerloff, Bill Sellers and Jean Beggs for all of their guidance and advice over the years as well as the Medical Research Council for funding. I thank everyone on Swann Level 3 and 65/3 West Mains Road for making my time in Edinburgh so enjoyable; particularly Morag Bilsland, Russell Hamilton, Ralf Schmid, Dinesh Soares & Joel Tyndall from Swann and John Ellis, John Hamilton, Hichem Mortad & Stuart Ord from West Main Road as well as Graham Reid. I would also like to thank David Gilbert for supporting me through the long process of writing up as well as the other members of the Bioinformatics Research Centre at the University of Glasgow. Most importantly, I would like to thank all my family (Mum, Dad, Alan, Caroline, Graeme, Abi & Sam) and especially Morag for supporting me throughout, for keeping me sane while our flat was falling apart, for entertaining Jess whilst I was working at home and on holiday, and for keeping me smiling the whole way.

Abstract

Over the past few years there has been a relative explosion of data in the biological sciences. At the heart of this data explosion is the budding yeast *Saccharomyces cerevisiae* (*S. cerevisiae*) which is one of the most widely studied eukaryotes due to its value as a model organism in biological research; it has a fully sequenced genome that is well annotated and a variety of publicly available functional genomic data sets. Analysis of this vast amount of data is a key challenge and computers in conjunction with effective software tools are an essential part of this process. There has been a rapid increase in the number of software tools available for the visualisation and analysis of individual types of functional genomic data sets. However, there are relatively few tools available that are capable of bringing together a number of different types of data sets for integrated visualisation and analysis. As many new biological insights are likely to emerge from the combined use of data from different functional genomic strategies, there is a need for a new generation of software tools that are capable of effectively utilising the wealth of data available for *S. cerevisiae* enabling users to perform integrative analyses.

The Yeast Exploration Tool Integrator (YETI) is a novel bioinformatics tool for the integrated visualisation and analysis of *S. cerevisiae* functional genomic data sets. The YETI system consists of a database for the storage and management of data and a Java program for the integrated visualisation and analysis of data. YETI utilises publicly available data sets from a number of different functional genomic strategies, such as gene expression microarrays and yeast two-hybrid screens, and provides an

effective means for their integrated visualisation and analysis. YETI consists of a number of individual sections for the visualisation and analysis of functional genomic data sets which are closely inter-linked enabling users to swiftly move between them and investigate all aspects of any genes or proteins of interest as well as providing access to textual information, including Gene Ontology (GO) annotations, at any point. YETI enables users to easily explore the data in an integrated modular fashion, investigate the intricacies of broad biological processes and test specific hypotheses.

In this thesis, we detail the design and development of YETI and also report a number of case studies which clearly demonstrate its potential and utility as an analysis and exploration tool. Furthermore, the results of a number of correlation analyses performed between the stored functional genomic data sets are also reported.

Table of Contents

Table of Figures

Table of Tables

Chapter 1

Background

1.1: The Budding Yeast *Saccharomyces cerevisiae*

Yeasts are fungi that grow as single cells. They are simple unicellular eukaryotes that multiply by budding or direct division (fission). They typically grow in moist environments where there is a plentiful supply of simple, soluble nutrients such as sugars and amino acids. For this reason they are commonly found on fruits, leaves, flowers, roots and in various types of food. The precise classification of yeasts is accomplished using the characteristics of the cell, ascospores and colonies. Physiological characteristics are also used to identify species, with one of the more well known characteristics being the ability to ferment sugars for the production of ethanol. Budding yeasts are true fungi of the phylum *Ascomycetes*, class *Hemiascomycetes* and the true yeasts are separated into one main order, *Saccharomycetales*.

The best known and commercially significant yeasts are the related species and strains of the budding yeast *Saccharomyces cerevisiae* (*S. cerevisiae*; Figure 1.1), also known as baker's or brewer's yeast. *S. cerevisiae* has played an important part in human history for a long time through the production food, beverages and a variety of fermentation products for industry. It also has a great scientific importance through its use in biological research where it has been the subject of extensive study for the past few decades.

Figure 1.1: *Saccharomyces cerevisiae*
This is an image of the budding yeast *S. cerevisiae* in the process of budding. This image was taken from the Munich Information Centre for Protein Sequences (MIPS) Comprehensive Yeast Genome Database (CYGD; Mewes *et al.*, 1998; http://mips.gsf.de/genre/proj/yeast/index.jsp).

1.2: The *S. cerevisiae* Genome

The genome contains all the biological information needed to build and maintain a living organism and can be defined as the complete set of genes of an organism or its organelles (Oliver, 2000). The biological information contained in a genome is encoded in its deoxyribonucleic acid (DNA) base pair (bp) sequence which is typically determined by systematic DNA sequencing techniques. *S. cerevisiae* was the first eukaryotic organism to have its genome sequenced and it was chosen to be so for a number of reasons: (1) *S. cerevisiae* is one of the most widely studied eukaryotic organisms due to its value as a model organism in biological research; (2) *S. cerevisiae* is a powerful eukaryotic model system because the basic cellular mechanics of replication, recombination, cell division and metabolism are generally conserved between the yeasts and higher eukaryotes such as *Homo sapiens*; (3) *S. cerevisiae* is cheap, easy to cultivate, has short generation times and has a relatively small genome which can be manipulated and analysed readily. It can be grown on defined media giving the experimenter complete control over its chemical and

physical environment; and (4) *S. cerevisiae* is easy to manipulate by molecular techniques and its genetics and biochemistry have been well characterised. It is a unicellular eukaryote and an ideal organism for geneticists as it allows genes to be replaced, mutated or deleted by homologous recombination.

S. cerevisiae has 16 nuclear chromosomes of varying lengths and a circular mitochondrial chromosome of 86 kilo bases (kb). The mitochondrial chromosome was initially sequenced in segments during the 1980s but was subsequently re-sequenced in the 1990s (Foury *et al.*, 1998). The *S. cerevisiae* genome sequencing project began in January 1989 when a consortium of 35 European laboratories began the sequencing of *S. cerevisiae* chromosome III (Vassarotti *et al.*, 1992). In 1992 this project resulted in the release of the complete DNA sequence of chromosome III which was presented to be 315 kb in length (Oliver *et al.*, 1992). This was a scientific landmark because it was the first eukaryotic chromosome to be sequenced. However, it also revealed the extent of what remained to be understood in the genome of an otherwise extensively studied organism.

A total of 182 open reading frames (ORFs) encoding putative proteins longer than or equal to 100 codons were identified from the DNA sequence of chromosome III (Oliver *et al.*, 1992). The size limit of 100 codons was chosen because ORFs of this length have less than 0.2 % probability of occurring by chance (Sharp *et al.*, 1991), it was however recognised that a few shorter genes were likely to exist. Of the 182 genes identified, only 34 appeared on the existing *S. cerevisiae* genetic map (Mortimer *et al.*, 1989; Oliver *et al.*, 1992). This showed that even in the genome of

an organism as small and intensively studied as *S. cerevisiae*, only a minor proportion of the genes had been identified by classical means. Analyses of the newly discovered ORFs revealed how much was still left to learn about this organism. Only 10 % of ORFs showed significant sequence similarity to other genes from *S. cerevisiae*, 10 % were similar to genes from other organisms and 80 % showed no significant sequence similarity to any previously sequenced genes in any organism (Oliver *et al.*, 1992). The majority of genes on chromosome III were completely novel and to many, completely unexpected.

In April 1996, *S. cerevisiae* became the first eukaryotic organism for which a complete genome sequence was publicly available (Goffeau *et al.*, 1996); *S. cerevisiae* was shown to have a relatively small and compact genome of 12,068 kb (Goffeau *et al.*, 1996). At the beginning of the sequencing project ~ 1,000 genes encoding either protein products or ribonucleic acids (RNA) had been identified on the *S. cerevisiae* genome by genetic analyses (Mortimer *et al.*, 1992; Goffeau *et al.*, 1996). However, initial analysis of the *S. cerevisiae* genome sequence revealed the presence of 6,275 ORFs, 5,885 of which were believed to represent protein encoding genes (Goffeau *et al.*, 1996). The presence of an ORF in a genome sequence does not necessarily imply the existence of a functional gene and despite advances in bioinformatics it is still difficult to predict genes, especially small ones, accurately from genomic data (Eisenberg *et al.*, 2000; Mathe *et al.*, 2002). For example, due to discrepancies in gene numbers indicated by previous analyses, the *S. cerevisiae* genome underwent a complete re-annotation in 2001 (Wood *et al.*, 2001). In this analysis, 3 new ORFs were identified, 46 ORF coordinates were altered, 370 ORFs

were defined as totally spurious and a further 193 ORFs were defined as very hypothetical. Overall, the *S. cerevisiae* gene number estimate was revised to a new upper limit of 5,570. Although this number is likely to be closer to the true upper limit, it is still predicted to be an overestimate of the real gene number (Wood *et al.*, 2001).

The longest known ORF is YLR106C located on chromosome XII with a length of 14,733 bp (4,910 codons). However, very few ORFs are longer than 1,500 codons. The lower size limit is less clear cut because without direct information on function, real short genes cannot be easily distinguished from random occurrences of apparent short ORFs. Short genes can be identified from the genome by the presence of introns, biased codon usage or the existence of corresponding transcripts. On average a protein encoding gene is found every 2 kb of the *S. cerevisiae* genome with the typical *S. cerevisiae* gene being 1,450 bp (483 codons) in length preceded by an upstream region of 309 bp and followed by a downstream region of 163 bp making a total of only 1,922 bp (Dujon, 1996). ORFs occupy approximately 70 % of the *S. cerevisiae* genome (Dujon, 1996) which leaves little space for all other structural and functional elements as well as non-coding DNA.

One of the major findings of the initial genome sequence analysis was the presence of 'orphan' genes (Dujon, 1996). The orphan genes are a large set of previously undiscovered genes of unknown function with no sequence homologues of known function. Although gene numbers are undergoing continuous revision by the yeast community, it is currently reasonable to estimate that ~ 30 % of *S. cerevisiae* genes

are orphans. It is widely believed that these genes do make a contribution to the upkeep of the organism and there is little doubt that the majority of the sequenced ORFs are actual genes that are expressed under certain conditions. *S. cerevisiae* deletion mutants have been generated by homologous recombination for ~96% of the predicted ORFs (Winzeler *et al.*, 1999) and ~1500 genes were identified as essential for viability (Giaever *et al.*, 2002); numerous nonessential genes have been found to be required for various biological processes (Ooi *et al.*, 2001; Begley *et al.*, 2002; Deutschbauer *et al.*, 2002). Ultimately, the validity and function of each ORF can only be proven by experiments in the laboratory but given the number of orphans in the *S. cerevisiae* genome this could take some time. Therefore, there is a clear need for new experimental and computational methods to aid in the assignment of biochemical functionality.

Analysis of sequences also revealed that many genes were part of families with two or more members whose predicted protein products were at least 50 % identical (Mewes *et al.*, 1997). This apparent genetic redundancy can be partly explained by the presence of gene sets with overlapping functions (Goffeau *et al.*, 1996); most of the duplicated genes are members of families with just two or three members but some gene families are significantly larger. In addition, blocks of duplicated ORFs called cluster homology regions (CHR) were found in both the telomeric regions and at internal sites within chromosome arms (Goffeau *et al.*, 1996). Genetic redundancy appears to be common at chromosome ends and many duplicate genes seem to be phenotypically redundant. However, single gene duplication mechanisms are insufficient to account for the full extent of redundancy in the *S. cerevisiae* genome.

An alternative explanation is that the *S. cerevisiae* genome underwent a complete duplication at some stage in its evolutionary history and has subsequently been reduced to its present size via a series of deletions (Wolfe *et al.*, 1997). A recent study demonstrated that the *S. cerevisiae* genome could indeed have arisen from an ancient whole genome duplication (Kellis *et al.*, 2004). In this study, the genome of a related yeast species called *Kluyveromyces waltii* (*K. waltii*), which diverged from *S. cerevisiae* before the duplication event, was sequenced and analysed. The two genomes are related by a 1:2 mapping, with each region of *K. waltii* corresponding to two regions of *S. cerevisiae*, as expected for a whole genome duplication.

1.3: Gene Ontology

The Gene Ontology (GO) project (Ashburner *et al.*, 2000; http://www.geneontology.org/) is a collaborative effort to address the need for consistent descriptions of gene products in different genomic databases. The project began in 1998 as a collaboration between three model organism databases: FlyBase (Gelbart *et al.*, 1996; http://flybase.org/), the *Saccharomyces* Genome Database (Cherry *et al.*, 1998; http://www.yeastgenome.org/) and the Mouse Genome Database (Blake *et al.*, 2000; http://www.informatics.jax.org/). Since then, the GO Consortium (Ashburner *et al.*, 2001) has grown to include many databases including several of the world's major repositories for plant, animal and microbial genomes. The GO annotation system is split into three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated molecular

functions, biological processes and cellular components in a species-independent manner. The three ontologies are defined by the GO Consortium as follows:

1) **Molecular Function**: "A molecular function describes activities, such as catalytic or binding activities, at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products. Examples of broad functional terms are catalytic activity, transporter activity or binding; examples of narrower functional terms are adenylate cyclase activity or Toll receptor binding", (http://www.geneontology.org/).
2) **Biological Process:** "A biological process is accomplished by one or more ordered assemblies of molecular functions. Examples of broad biological process terms are cell growth and maintenance or signal transduction; examples of more specific terms are pyrimidine metabolism or alpha-glucoside transport. It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps", (http://www.geneontology.org/).
3) **Cellular Component**: "A cellular component is simply a component of a cell but with the proviso that it is part of some larger object, which may be an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene

product group (e.g. ribosome, proteasome or a protein dimer)", (http://www.geneontology.org/).

The ontologies are organised into structures called 'directed acyclic graphs' which differ from hierarchies in that a 'child' can have many 'parents'. This structure also enables queries to be performed at different levels: for example, one can use the GO system to find all the gene products in the *S. cerevisiae* genome that are involved in signal transduction, or you can zoom in on all the receptor tyrosine kinases. Furthermore, annotators are able to assign properties to gene products at different levels, depending on how much is known about a gene product. It is also important to note that a gene product can have multiple GO annotations; a gene has one or more molecular functions, is used in one or more biological processes and can be associated with one or more cellular components.

GO slims (http://www.geneontology.org/) are cut-down versions of the GO ontologies that contain a subset of the terms from the complete GO. They give a broad overview of the ontology content without the detail of the specific fine grained terms. GO slims are particularly useful for giving a summary of the results of GO annotations of a genome when broad classifications of gene product function are required. The GO Consortium provides a generic GO slim which, like the GO itself, is not species specific. However, many organism specific databases such as the *Saccharomyces* Genome Database (SGD; Cherry *et al.*, 1998; http://www.yeastgenome.org/) have created their own specie specific GO slim.

1.4: Functional Genomics

The sequencing project has essentially provided biologists with a complete catalogue of all the genes present in *S. cerevisiae*. The goal now is to understand the interactions of all gene products and ultimately their function in creating this simple eukaryotic organism. However, a large proportion of the genes in *S. cerevisiae* are still classified as proteins of unknown function and additional information is needed to place them within a biological context. Functional genomics strategies are becoming increasingly important in characterising novel proteins discovered by genome sequencing projects. Many such strategies use the principle of 'guilt by association' (Oliver, 2000) as the means of elucidating function, i.e. genes that are coexpressed or proteins that interact with one another are likely to be involved in the same or related cellular process.

1.5: The *S. cerevisiae* Transcriptome

The transcriptome can be defined as the complete set of RNA molecules present in a cell, tissue or organ at a certain point in time (Oliver, 2000). Unlike the genome, the transcriptome is highly dynamic and changes rapidly and dramatically in response to changes in the environment and during cellular events. In terms of understanding the function of a gene, knowing when and to what extent it is expressed can be crucial to understanding the activity and biological role of its encoded protein. Gene expression studies have previously relied on techniques such as northern blot analysis which measure the expression of only a single or small set of genes at one time. Newer

technologies including Serial Analysis of Gene Expression (SAGE; Velculescu *et al.*, 1997), high throughput northern analysis (Planta *et al.*, 1999) and gene expression microarrays (Schena *et al.*, 1995; Lockhart *et al.*, 1996) enable thousands of genes to be analysed at once.

1.5.1: Microarrays

Microarrays are microscopic arrays of large sets of nucleic acids immobilised on solid substrates such as glass, they are used for a wide range of analytical methods based around the detection of sequence specific nucleic acid hybridisation. Microarrays can monitor, rapidly and efficiently, the messenger RNA (mRNA) abundance of all an organism's genes, allowing massive parallel data acquisition and analysis; they provide a sensitive, global readout of the physiological state of the cell. It is important to note that the relationship between the quantity of mRNA and the abundance of the corresponding protein in the cell is not trivial due to the fact that the speed of production varies for different proteins as does the half-life of both the protein and mRNA. However, it is widely accepted that measuring the level of mRNA gives us a reasonable insight into the abundance of the corresponding protein and it is this that can be measured on a genomic scale using microarrays.

Currently, there are two general types of microarrays widely used in biological research, spotted microarrays (Schena *et al.*, 1995) and Affymetrix chips (Lockhart *et al.*, 1996), both of which rely on the same binding property of DNA. DNA and RNA are examples of nucleic acids, one characteristic of which is their tendency to

form double stranded molecules through complementary base pairing. This tendency of nucleic acids to form double stranded molecules is known as hybridisation and plays an important role in the measurement of mRNA abundance. For example, consider a specific gene and its mRNA product; given a sample of this mRNA, it is possible to reverse transcribe it to single stranded complimentary DNA (cDNA) which will hybridise to a single strand of the gene's original DNA. It is this hybridisation that underlies the operation of microarrays.

Spotted microarrays (Schena *et al.*, 1995; Figure 1.2) typically consist of a small glass slide onto which the DNA sequences of the genes to be analysed are printed at pre-defined locations to create an array of tiny spots; each spot contains many copies of the sequence of one gene. A basic spotted microarray experiment proceeds as follows (Figure 1.3), mRNA is extracted from the cell sample of interest and also from a separate control cell sample; the two samples are kept separate at this point. Reverse transcription is used to transform all the mRNA molecules into cDNA molecules labelled with distinct fluorescent dyes; typically Cy5 (red) for the experimental sample and Cy3 (green) for the control sample. The two samples are then pooled and washed over the slide and left to hybridise for a set period of time. Once this time has elapsed, the slides are rinsed and are ready to be analysed. The microarray is then scanned using a laser to excite the dyes and independent images for the green (control) and red (experimental) channels are generated. These images must then be analysed to identify all the arrayed spots and to measure their fluorescence intensities. Currently, image analysis requires significant human intervention to ensure that grids are properly aligned and artefacts are flagged and

excluded from subsequent analysis. After image processing, it is necessary to normalise the relative fluorescence intensities in each of the two scanned channels. Normalisation adjusts for differences in labelling and detection efficiencies for the fluorescent labels and for differences in the quantity of initial RNA from the two samples examined in the assay. There are three widely used techniques that can be used to normalise gene expression data (Quackenbush, 2001): (1) Total intensity normalisation; (2) Normalisation using regression analysis; and (3) Normalisation using ratio statistics.

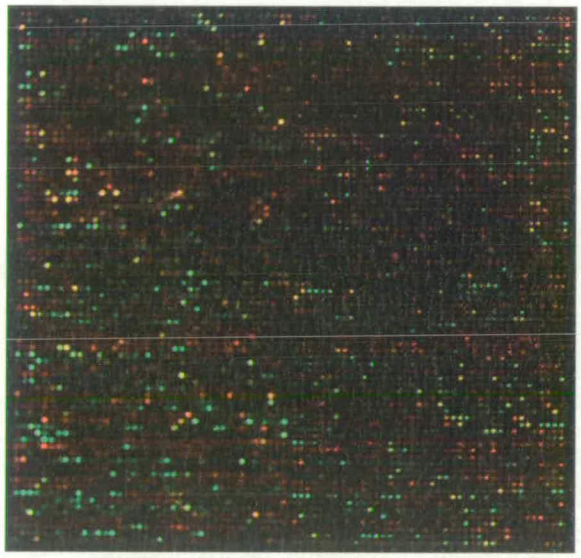


Figure 1.2: *S. cerevisiae* spotted microarray
This is an image of a spotted microarray with all the ~ 6,000 *S. cerevisiae* ORFs spotted onto it. Each spot on the microarray represents a separate ORF that has been individually synthesised and mechanically spotted onto the microarray. The colour and intensity of each spot can be used to calculate the relative expression level of the corresponding ORF in the *S. cerevisiae* genome under the experimental conditions used. This image was taken from the Stanford Microarray Database (SMD; Sherlock *et al.*, 2001; http://genome-www5.stanford.edu/).

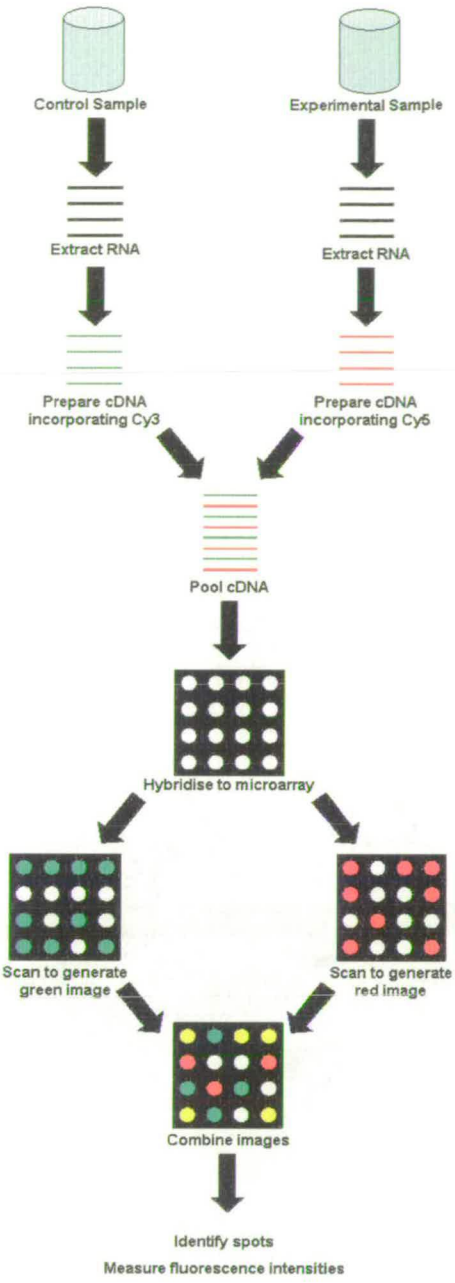


Figure 1.3: Spotted microarray experimental procedure
This image displays the experimental procedure for a typical microarray experiment from RNA extraction to image analysis: (1) Extract the RNA from both the control and experimental cell samples; (2) Prepare cDNA probes by incorporating either Cy3 (green; control) or Cy5 (red; experimental) using a single round of reverse transcription; (3) Pool the two cDNA samples; (4) Hybridise the pooled sample to a single microarray slide; (5) Scan the microarray slide in the green and red channels to create a green and red image, respectively; (6) Combine the two images to create a single image of the microarray, identify the spots and measure the fluorescence intensities in each channel for each spot.

Ultimately, the result of a spotted microarray experiment is two fluorescence values (experimental and control) for each gene spot on the microarray. The ratio of these readings provides us with a relative level of expression for the experimental sample with respect to the control. For example, if for a particular gene there is much more mRNA in the experimental sample relative to the control, the dye corresponding to the experimental sample (typically red Cy5) will fluoresce much more than the dye for the control (typically green Cy3) and we will have a high ratio. These ratios are normally logged (base 2) to preserve the symmetry between over and under expression (Eisen *et al.*, 1999).

Affymetrix chips (Lockhart *et al.*, 1996) are high density arrays of oligonucleotides synthesised *in situ* using light directed chemistry. They combine photolithography technology with DNA synthetic chemistry to enable high density oligonucleotide manufacture (Schena *et al.*, 1998). Affymetrix chips use a slightly more complicated procedure when compared to spotted microarrays, but do not need a separate control sample and hence provide absolute rather than relative expression values. For each gene that is being analysed, a number of small sections of the gene's DNA are printed at various locations around the array; these are referred to as perfect match (PM) probes. Next to each of these, the same sequence is printed but with the middle base switched; these are referred to as mismatch (MM) probes. The mRNA from an experimental sample is reverse transcribed to cDNA, labelled with a fluorescent dye, washed over the array and then excited with a laser to generate an image. Various algorithms exist to combine all these probe values into one expression value (e.g.

http://www.affymetrix.com/support/technical/technotes/statistical_reference_guide.pdf) for each gene analysed.

1.5.2: Cluster Analysis

One of the biggest challenges in applying gene expression microarray technology lies in data analysis. Currently, there are a wide variety of methods referred to as 'Cluster Analysis' that attempt to organise genes with similar expression patterns into related groups or clusters; a gene's expression pattern over a number of microarray experiments is also known as it's expression profile. The basic assumption underlying these approaches is that genes with similar expression patterns are likely to be related functionally. In this way, genes without functional assignments can be given tentative assignments based on the functions of known genes in the same expression cluster; the concept of 'guilt by association'. However, a tentative functional assignment may not be much more than a vague description or general classification.

1.5.3: Hierarchical Clustering

Hierarchical clustering has the advantage that it is simple and that the result can be easily visualised. As a result it has become one of the most widely used techniques for the analysis of gene expression data; a seminal paper in the use of hierarchical clustering for gene expression analysis was published by Eisen *et al.* (1998). Hierarchical clustering is an agglomerative approach in which single gene expression

profiles are joined together to form clusters of genes which are further joined together until the process is completed; thus forming a single hierarchical tree with a corresponding clustered gene expression data table.

The hierarchical clustering process through a number of distinct steps (Table 1.1; Quackenbush, 2001). The first step is to create a pairwise gene expression matrix. The matrix is generated by mathematically comparing every gene expression profile to every other gene expression profile in a pairwise fashion to create a distance (or similarity) score; the matrix is therefore comprised of all the pairwise distance scores between all the profiles. It is important to note that the way in which distance is measured between gene expression profiles can have a profound effect on the clusters that are produced and there are a number of different distance metrics that can be used. Perhaps the simplest method used to do this is the Euclidean distance metric which is a generalisation of the Pythagorean Theorem. However, the Pearson correlation coefficient is perhaps the most widely used measurement of distance between two expression profiles and the averaged dot (or inner) product is also commonly used; a good review of distance measures is presented in Sturn, 2001.

Step	Description
1	The pairwise distance matrix is calculated for all genes to be clustered.
2	The pairwise distance matrix is searched for the two most similar clusters (initially all clusters consist of a single gene). If more than one pair of clusters has the same similarity measure, a predetermined rule is used to decide between them.
3	The two selected clusters are merged to produce a single new cluster.
4	The distances are calculated between the new cluster and all the other clusters in the matrix. There is no need to calculate all the distances in the matrix as only those involving the new cluster have changed.
5	Steps 2-4 are repeated until all the clusters have been joined to form a single hierarchical tree.

Table 1.1: Steps of the hierarchical clustering processing
This table contains a step wise description of the hierarchical clustering process.

Pairwise linkage is a form of hierarchical clustering that has been successfully applied to sequence and phylogenetic analysis and has now been applied to clustering gene expression data. There are several variations of pairwise linkage clustering that differ in the way distances are measured between clusters as they are constructed (Table 1.2; Quackenbush, 2001), each of which will produce slightly different results. Typically for gene expression data 'pairwise average linkage' clustering gives acceptable results (Quackenbush, 2001). However, one potential problem with many hierarchical clustering methods is that as clusters grow in size the expression profile that represents the cluster might no longer represent any of the genes in the cluster. Consequently, as clustering progresses the actual expression patterns of the genes themselves become less relevant. Furthermore, if a poor assignment is made early in the process it cannot be corrected.

Method	Description
Pairwise single linkage	The distance between two clusters is calculated as the minimum distance between a member of the first cluster and a member of the second cluster.
Pairwise complete linkage	The distance between two clusters is calculated as the maximum distance between a member of the first cluster and a member of the second cluster.
Pairwise average linkage	The distance between two clusters is calculated as the average distance between all members of the first cluster and all members of the second cluster.

Table 1.2: Pairwise linkage clustering techniques
This table contains the names and descriptions of the three main types of pairwise linkage hierarchical clustering techniques.

Hierarchical clustering methods group genes with similar expression profiles together. The computed hierarchical tree can then be used to reorder the genes in the original expression data table so that genes with similar expression profiles are juxtaposed. However, the resulting ordered but still massive collection of numbers can remain difficult to visualise and comprehend. Therefore, it is essential to include a graphical representation of the data table by representing each gene expression data point with a colour that reflects its value; the hierarchical tree is then typically displayed alongside this table. The most commonly used method colours each data point on the basis of its $\log_2$ ratio, with those close to zero coloured black, those greater than zero coloured red and those with negative values coloured green. The end product is a graphical representation of complex gene expression data that, through statistical organisation and graphical display, allows biologists to understand and explore the data in a natural intuitive manner (Figure 1.4).

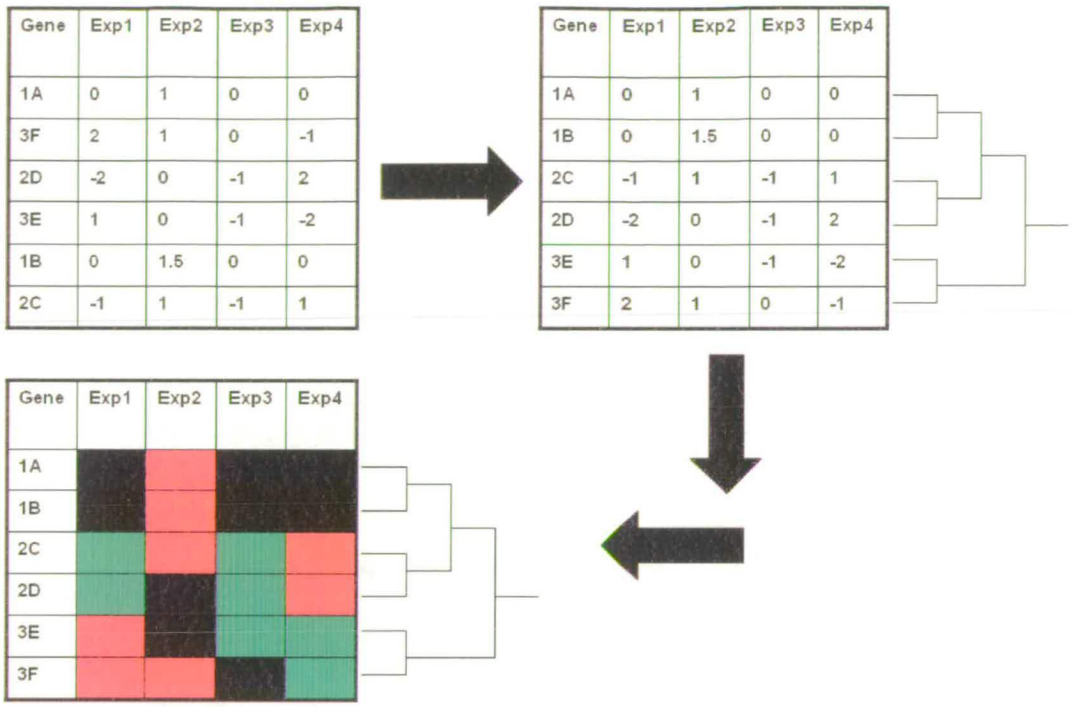


Figure 1.4: Hierarchically clustered gene expression data table
This figure shows the main steps involved in the hierarchical clustering of a microarray gene expression data set. The first step involves hierarchically clustering the gene expression data table to produce a hierarchical tree and a corresponding ordered data table. The second step involves visually representing each gene expression data point with a colour that represents its value, thus creating a clustered graphical representation of the gene expression data set. The extension of this example to include many more genes and microarray experiments is simple.

Although cluster analysis techniques are extremely powerful, great care must be taken in applying this family of techniques. The algorithms used are well defined and reproducible but selecting different algorithms, normalisations or distance metrics will place different genes into different clusters; thus giving different results depending on the route taken. Furthermore, clustering unrelated data will still produce clusters although they might not be biologically meaningful. It is therefore essential to select relevant data and apply algorithms appropriately so that data is clustered sensibly.

1.5.4: Other Clustering Methods

There are a variety of other statistical methods that can be used to analyse gene expression data and cluster genes into similar groups. Three of the major unsupervised methods for clustering gene expression data are (Quackenbush, 2001; Sturn, 2001):

1) ***k*-means clustering** (Tavazoie *et al.*, 1999) can be used as an alternative to hierarchical methods if there is advanced knowledge about the numbers of clusters that should be represented in the data. In *k*-means clustering, objects are partitioned into a fixed number (*k*) of clusters such that the clusters are internally similar but externally dissimilar; no dendrograms are produced.
2) **Self Organising Maps** (SOM; Tamayo *et al.*, 1999) are an unsupervised neural network based divisive clustering approach. A SOM assigns genes into a series of partitions on the basis of the similarity of their expression vectors to reference vectors that are defined for each partition. It is the process of defining these reference vectors that distinguishes SOMs from *k*-mean clustering.
3) **Principal Component Analysis** (PCA; Raychaudhuri *et al.*, 2000), also known as Singular Value Decomposition (SVD) is a mathematical technique that reduces the effective dimensionality of gene expression space without significant loss of information. PCA provides a 'projection' of complex data sets onto a reduced, easily visualised space.

In addition to the unsupervised methods discussed above, there are a variety of supervised methods that can be used in the analysis of gene expression data. Supervised methods represent a powerful alternative that can be applied if one has previous information about which genes are expected to cluster together. One widely used supervised approach is the Support Vector Machine (SVM; Brown *et al.*, 2000).

1.5.5: *S. cerevisiae* Microarray Experiments

Over recent years, microarrays have been used widely in biological research to effectively measure the relative mRNA abundance of all the genes in *S. cerevisiae* under a variety of experimental conditions. For example, the Stanford Microarray Database (SMD; Sherlock *et al.*, 2001; http://genome-www5.stanford.edu/) alone currently contains 40 *S. cerevisiae* microarray studies. Contained within the mass of numbers produced by this technology is an immense amount of biological information. Furthermore, microarray results can represent the first indication to the function of many *S. cerevisiae* genes and with each new microarray experiment additional information is added.

Microarrays are well suited for the analysis of temporal changes in gene expression during cellular events such as the cell cycle. Cell populations are synchronised by arresting them at a homogeneous cell cycle state then released from the arrested state and sampled at subsequent time intervals. Cho *et al.* (1998) were the first to analyse cell cycle periodic transcription patterns using microarrays. This study was quickly followed by additional studies of the mitotic (Spellman *et al.*, 1998) and meiotic

(Chu *et al.*, 1998) cell cycles in the budding yeast. Cho *et al.* (1998) used visual examination of time series plots to identify a set of 416 periodic transcripts. Spellman *et al.* (1998) used Fourier analysis of both their own data and the data from Cho *et al.* (1998) to compute a periodicity score for each gene in the array. Using this approach they scored 800 yeast genes as cell cycle periodic. Chu *et al.* (1998) evaluated the transcript profile of synchronously sporulating yeast cells in comparison with an asynchronous vegetative culture. They distinguished seven temporal classes of sporulation specific genes using cluster analysis and other methods. Other studies revealed that in rich medium, 87 % of all putative *S. cerevisiae* genes had a detectable level of expression, approximately 7 % of which were shown to have cell cycle dependent periodicity (Zweiger *et al.*, 1999).

It is well known that yeast cells change their patterns of gene expression in response to environmental stresses and microarrays can be used to measure these changes. To this end, Gasch *et al.* (2000) measured the genomic expression patterns of *S. cerevisiae* in response to environmental changes such as heat and cold shock, amino acid starvation, nitrogen depletion and steady state growth on alternative carbon sources. Microarrays have also been used to evaluate transcripts differentially expressed in yeast cells treated with DNA damaging agents (Jelinsky *et al.*, 1999; Gasch *et al.*, 2001) and for evolutionary studies of *S. cerevisiae* (Ferea *et al.*, 1999). Combining the data from several unrelated expression profiling experiments can result in more detailed and informative clustering; this was first demonstrated in *S. cerevisiae* when ~300 different experimental and genetic conditions were combined to create a so-called transcriptome compendium (Hughes *et al.*, 2000).

1.6: The *S. cerevisiae* Proteome

The proteome can be defined as the complete set of protein molecules present in a cell, tissue or organ at a certain point in time (Oliver, 2000). Messenger RNA transcripts are the transmitters of genetic information; they are not functional cellular entities. Proteins by contrast are the main catalysts, structural elements, signalling messengers and molecular machines of living cells. Proteomics is the large scale study of proteins usually by experimental biochemical means. The main methods used in proteomic research are large scale identification and localisation studies (Burns *et al.*, 1994) and protein-protein interaction studies (Fields *et al.*, 1989).

The study of protein-protein interactions is currently an important area of functional genomics. It is well recognised that protein-protein interactions play a key role in the structural and functional organisation of the cell; most proteins require physical interaction with other proteins to fulfil their biological goal. If two proteins interact with one another they often participate in the same or related cellular functions; the concept of 'guilt by association'. A detected protein-protein interaction has the potential to yield a wide array of information which can generally be classified into one of four categories (Oliver, 2000): (1) An interaction between a protein of known and a protein of unknown function may allow the role of the latter to be inferred; placing functionally unclassified proteins into a biological context; (2) An interaction between proteins involved in the same biological process can provide information on how functionally related proteins are working together in order to fulfil biological goals; (3) An interaction between proteins involved in different biological processes

can provide clues as to how processes are combining together to create larger cellular processes; and (4) An interaction between two *S. cerevisiae* proteins can imply an interaction between the orthologous proteins in another organism.

1.6.1: The Yeast Two-Hybrid System

The yeast two-hybrid system (Fields *et al.*, 1989) can be used to identify pairs of proteins that physically interact with one another (Figure 1.5). It works by separating the coding sequences of the DNA binding and activation domains of a transcriptional activator, which are then cloned into different vector molecules. The coding sequence of the protein whose partners are sought (the 'bait') is fused with the DNA binding domain. Typically, a library of coding sequences for proteins that might interact with the bait (the 'prey') is fused with the activation domain. As *S. cerevisiae* has two sexes (a and α) baits and preys can easily be introduced into the same *S. cerevisiae* cell by mating. If the two proteins physically interact, the DNA binding and activation domains are closely juxtaposed and the reconstituted transcriptional activator can mediate the switching on of a reporter gene that typically brings about a colour change to the host *S. cerevisiae* cell. As a result, the yeast two-hybrid system is simple, sensitive and amenable to high throughput methods.

One disadvantage of this approach is that it typically uses the entire protein sequence derived from the DNA sequence and so does not account for the different splice variants or post-translational modifications of the protein which could interact

differently. In addition, the two-hybrid system reveals potential protein interactions but not the biological context in which they happen. Some may occur only when *S. cerevisiae* is in a particular physiological state (i.e. when both proteins are expressed and translated from their corresponding genes), whereas others may never occur because in real life the proteins are located in separate cellular compartments.

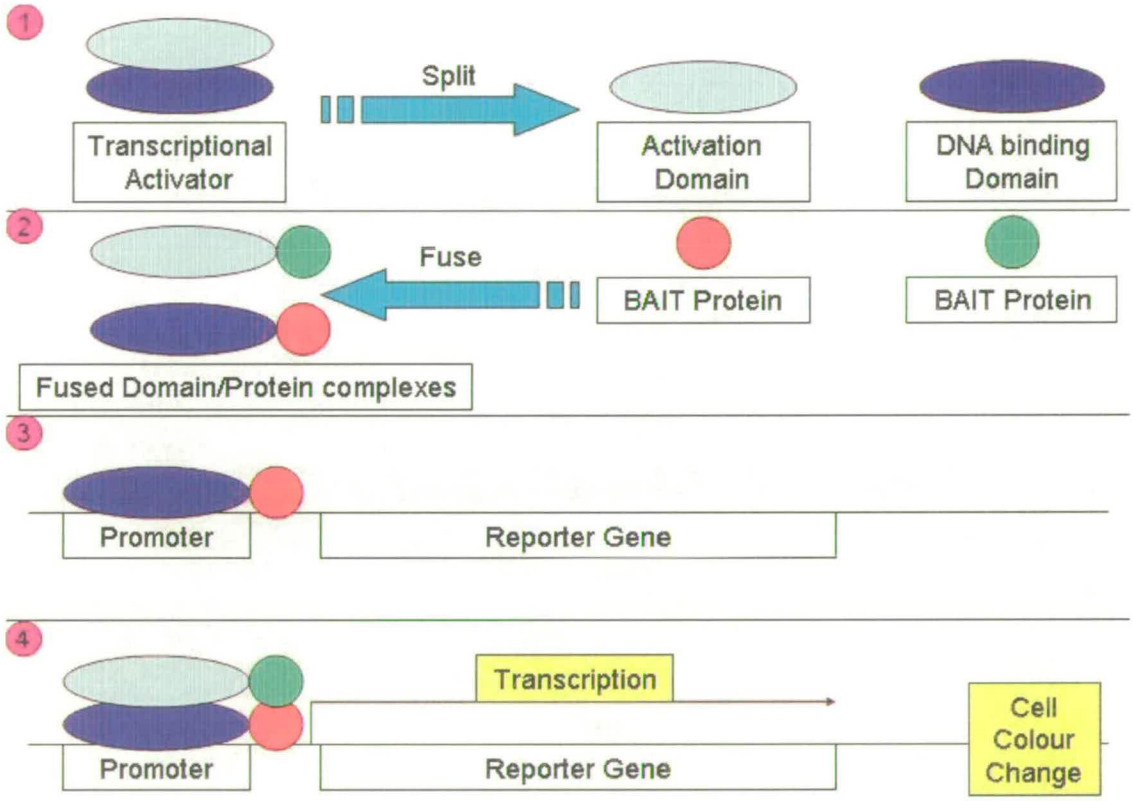


Figure 1.5: The yeast two-hybrid system
This is an image depicting the main steps of the yeast two-hybrid system. The DNA binding and activation domains of a transcription activator are split (1) and fused to a bait and prey protein, respectively (2). The DNA binding domain fused to the bait protein is still able to bind the reporter genes' promoter (3). If the two proteins interact together, the two domains are juxtaposed and the transcriptional activator is reconstituted, thus switching on the reporter gene which brings about a colour change to the hosting yeast cell (4).

The two-hybrid system combined with the complete genome sequence of *S. cerevisiae* has given biologists the opportunity to identify all possible pairwise interactions between the ~ 6,000 proteins of *S. cerevisiae*. A collaborative group

from the University of Washington and the biotechnology company CuraGen used the two-hybrid system on a large scale to identify 957 putative interactions involving 1,004 proteins (Uetz *et al.*, 2000); this group subsequently reported an additional 553 interactions, available at http://depts.washington.edu/sfields/yplm/data/index.html. A different collaborative group from Japan also used the two-hybrid system to begin the construction of a comprehensive protein-protein interaction map of *S. cerevisiae*. This group followed up their initial pilot study (Ito *et al.*, 2000) with a comprehensive two-hybrid analysis of the yeast interactome (Ito *et al.*, 2001). This study resulted in the identification of 4,549 interactions among 3,728 proteins; a core data set from within the main data set was also identified consisting of 841 interactions that were reported more than three times and involved 797 proteins. Surprisingly, there was only a small overlap between the data generated from the Uetz *et al.* (2000) and Ito *et al.* (2000 & 2001) studies (Hazbun *et al.*, 2001). Furthermore, neither of the two studies reproduced more than ~ 13 % of the published interactions previously detected by the scientific community using conventional interaction analyses (Hazbun *et al.*, 2001). Smaller scale yeast two-hybrid screens have also been performed in *S. cerevisiae* to investigate the specific interactions of splicing factors, RNA polymerase III and Sm-like proteins (Fromont-Racine *et al.*, 1997; Flores *et al.*, 1999; Fromont-Racine *et al.*, 2000). Furthermore, large scale yeast two-hybrid screens have also been performed in other organisms such as *Drosphilia melanogaster* (*D. melanogaster*; Stanyon *et al.*, 2004; Giot *et al.*, 2003), *Caenorhabditis elegans* (*C. elegans*; Li *et al.*, 2004), bacteria and phage (Rain *et al.*, 2001; Bartel *et al.*, 1996) and viruses (Uetz *et al.*, 2004).

The simplest way to display a data set of protein-protein interactions is in a simple linear list or table containing the names of all the interacting protein pairs. However, this is impractical when the data sets are large due to the sheer amount of interactions being displayed. A much more intuitive way of representing protein-protein interactions is to use a visual graphical format (Figure 1.6). Although graphical representations do in essence just repeat the information shown in textual lists and tables, the graphical representation has fundamental advantages with respect to human perception (Uetz *et al.*, 2002). Firstly, humans are better able to understand and remember a graphical representation. Secondly, in a textual representation the interactions involving a particular protein are usually spread out over different positions in the list; this requires an exhaustive search through the whole list to find all the relevant interactions. However, in a graphical layout each protein only occurs once and its interacting partners and their relationships can be easily identified and examined.

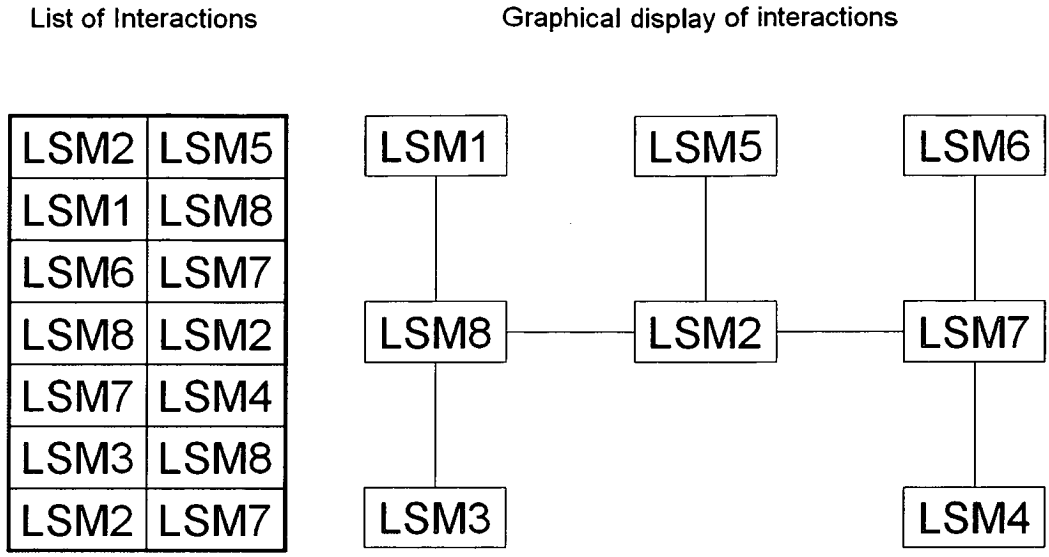


Figure 1.6: Visualisation of protein-protein interactions
This is an image of the two main ways of displaying a set of protein-protein interactions. The first way (shown on the left) is to display the interactions as a simple table where each row contains the names of the two interacting proteins. However, this method requires the user to search through the whole table to find interactions involving a protein of interest. As can be seen, a much more intuitive method of displaying interactions is by using a graphical representation (shown on the right). The user is easily able to see all the interactions, pick out proteins of interest and also get an impression of the overall connectivity between the proteins.

Protein-protein interactions can be effectively visualised using a range of computational approaches known as 'graph drawing' (Battista *et al.*, 1999). A graph consists of nodes (proteins) and edges (interactions) linking pairs of nodes together. In order to draw the graph, coordinates in either two or three dimensional space need to be associated with each node. One of the most important factors in drawing a graph is minimising the number of edge intersections and evenly spacing out nodes in the drawing space. Currently, one of the most widely used algorithms for protein-protein interaction graphs is the 'spring embedder' or 'springs and rings' algorithm (Eades, 1984). This algorithm is relatively simple and works by representing edges as springs and nodes as rings. The springs create an attracting force between the rings when they are far apart and a repulsive force when they are close together. The

algorithm searches for a placement of rings that minimises the total energy present in the system; this is commonly achieved by simulating the behaviour of the system over a certain period of time. However, these algorithms struggle to cope when the number of nodes reaches the hundreds and when there is a high connectivity between the nodes. This is because current computer technology struggles to cope with the processor time required to calculate the minimum energy in the system and sometimes to even draw an aesthetically pleasing and understandable graph. An additional problem when viewing graphs displaying a large number of nodes is the sheer size; it becomes virtually impossible to display the graph at a readable size on an object such as a computer screen. Other strategies for visualising protein-protein interaction networks include zoom and pan, focus and context (also known as fish-eye or the magnifying glass), and collapsing protein classes (Uetz *et al*, 2002).

1.6.2: Protein Interaction Complexes

Most proteins function within cellular pathways where they interact with other proteins either in pairs or as components of larger complexes. Two groups (Gavin *et al.*, 2002; Ho *et al.*, 2002) have characterised hundreds of distinct multi-protein complexes in *S. cerevisiae* using approaches in which individual bait proteins are tagged and used to catch associated proteins which are then analysed by mass spectroscopy. The approaches used by Gavin *et al.* (2002) and Ho *et al.* (2002) are similar and proceed through a number of distinct steps (Kumar & Snyder, 2002): (1) A tag is attached to the DNA coding sequence of a bait protein; (2) The DNA encoding the tagged bait protein is introduced into a yeast cell. The host cell

expresses the tagged protein allowing it to form complexes with other proteins which are naturally present in the cell at that time; (3) The bait protein is extracted using the tag which often results in the entire protein complex involving the bait protein being extracted as well; and (4) The proteins extracted with the tagged bait are identified using standard mass spectrometry methods.

Gavin *et al.* (2002) used tandem-affinity purification (TAP) and mass spectrometry in a large scale approach to characterise multi-protein complexes in *S. cerevisiae*. In this study 1,739 genes were processed and 589 protein assemblies were purified. Subsequent analysis of these assemblies identified 1,440 distinct proteins within 232 multi-protein complexes. More importantly, it proposed new cellular roles for 344 proteins including 231 with no previous functional annotation. Their analysis showed the *S. cerevisiae* proteome as a network of protein complexes at a level of organisation above pairwise interactions.

Ho *et al.*, (2002) used a technique termed high throughput mass spectrometric protein complex identification (HMS-PCI) to identify protein complexes. Numerous protein complexes were identified from the initial construction of 725 bait proteins; 3,617 associated proteins were detected involving 1,578 different proteins. The bait proteins were representative of a number of different functional classes including protein kinases, phosphatases, regulatory subunits and proteins involved in DNA damage response.

One interesting issue is how to represent the potential protein-protein interactions reported from this type of technique. Technically, these techniques only provide the identities of all the proteins in a particular complex, they do not tell use which proteins interact with which other proteins. Therefore, there are two general ways to represent the potential protein-protein interactions from these techniques: (1) The Spoke model represents a complex as a set of interactions where every protein only interacts with the tagged bait protein; and (2) The Matrix model represents a complex as a set of interactions where every protein interacts with every other protein. Furthermore, the potential protein-protein interactions detected from this technique are not really physical interactions; they are technically functional interactions as they detect groups of proteins in stable complexes, implying that they function together (Uetz *et al.*, 2005). However, functional interactions could be characterised as physical interactions in the future if additional data becomes available. It is also important to note that this type of technique has an additional difference to the yeast two-hybrid system: only proteins that are naturally present in the cell at the time of experimentation can interact with the bait protein.

Another example of a functional interaction is a genetic interaction. Genetic interactions are where the combination of alleles of two different genes has specific phenotypic consequences which is often taken to suggest that the two genes function in the same or parallel pathways affecting a particular biological process. Ongoing large-scale screens in *S. cerevisiae* have mapped thousands of genetic interactions derived from synthetic lethal mutations (Tong *et al.*, 2004). Other related functional genomic data sets include protein-DNA interaction data sets (Ren *et al.*, 2000), large

scale yeast protein localization data using GFP tagged yeast proteins (Huh *et al.*, 2003; Kumar *et al.*, 2002) and the quantification of the expression levels of approximately 4500 affinity tagged yeast proteins through western blot analysis (Ghaemmaghami *et al.*, 2003).

1.6.3: False Positives and False Negatives

The occurrence of both 'false positive' and 'false negative' interactions is perhaps the major disadvantage of the high throughput protein-protein interaction detection techniques described above. False positives wrongly indicate that two proteins interact with one another; they are generally caused by experimental errors and can be commonplace in large scale screens. On the other hand, false negative interactions wrongly indicate that two proteins do not interact with one another. Various studies have estimated that the ~6,000 *S. cerevisiae* proteins are connected by ~ 12,000 - 40,000 interactions (Wallhout *et al.*, 2000; Tucker *et al.*, 2001; Grigoriev *et al.*, 2003; Uetz *et al.*, 2005). However, the high throughput protein-protein interaction data sets described above have only detected a fraction of these. Furthermore, there is a lack of overlap between the different datasets themselves and also with published low-throughput studies which are generally considered to be less prone to false positives and false negatives (Ito *et al.*, 2001; Grunenfelder *et al.*, 2002; Cornell *et al.*, 2004; Uetz *et al.*, 2005). Taken together, this not only suggests that new or improved technologies are needed (especially for interactions involving membrane proteins) but also that more interactions could be detected by more exhaustive

application of these current techniques and that confidence scores for all detected interactions are of great importance.

There are now a number of different strategies for evaluating the reliability of large-scale protein interaction data sets (Bork *et al.*, 2004). In a recent study, various interaction data sets were tested for accuracy on confident sets of interactions and the rate of false positives for the various large-scale experimental approaches was found to vary widely, but was always larger than that for confident small scale experiments (von Mering *et al.*, 2002). However, high quality subsets could often be selected on the basis of additional criteria such as the degree to which mRNAs of interacting proteins are co-expressed in microarray experiments (Ge *et al.*, 2001; Deane *et al.*, 2002; Kemmeren *et al.*, 2002), topological properties of the resulting network (Goldberg *et al.*, 2003; Saito *et al.*, 2003), shared pathways or sub-cellular localisation (Date *et al.*, 2003; Sprinzak *et al.*, 2003) or combinations of these approaches (Bader *et al.*, 2004; Bork *et al.* 2004). Furthermore, several studies have suggested that interactions detected in multiple data sets and by different techniques or in different species are more likely to be true positives than those only found once (von Mering *et al.*, 2002; Uetz *et al.*, 2005). However, due to the high rates of false negatives in high throughput screens, there has been very little overlap between different datasets, thus limiting the opportunities for such experimental cross-validation (Uetz *et al.*, 2005). Therefore, computational tools that are able to effectively integrate the different interaction data sets together and then integrate them further with other functional genomic data sets would be extremely useful developments.

Since the first large scale data sets were published, the topological properties of protein interaction networks themselves have also been intensively studied. These networks have been shown to be both small world and scale free (Barabasi *et al.*, 2004). Interaction networks contain highly connected hub proteins which have been shown to correlate with evolutionary conserved proteins and in *S. cerevisiae* with proteins encoded by essential genes (Jeong *et al.*, 2001; Han *et al.*, 2004; Said *et al.*, 2004); therefore, a proteins relative position in a network can have implications for its function and importance. Analysis of topology also reveals clusters of highly interconnected proteins that correlate with conserved functional modules (Spirin *et al.*, 2003; von Mering, Zdobnov *et al.*, 2003; Poyatos *et al.*, 2004). This highlights the fact that even the current error prone networks can still be used to explore the hierarchical organisation of biological networks and to reveal interconnected modules that control specific biological properties (Uetz *et al.*, 2005). In addition to the study of the global topology of interacting networks, the existence of recurring local topological features, known as network motifs, has also been shown in protein-protein interaction networks (Wutchy *et al.*, 2003).

Over recent years computational methods have been increasingly used to predict protein-protein interactions; some prediction tools are now conveniently available as online services (e.g. von Mering, Huynen *et al.*, 2003). Gene expression profiles have been used to infer functional interactions among gene products based on the assumption that proteins that function together should be frequently expressed together (Jansen *et al.*, 2002; Jansen *et al.*, 2003). Genetic interactions have been predicted based on physical interactions, gene expression, protein localisation and

other experimental data (Marcotte *et al.*, 1999; Wong *et al.*, 2004). In addition, numerous methods for predicting physical protein-protein interaction have also been developed (Enright *et al.*, 1999; Aloy *et al.*, 2002; Jansen *et al.*, 2003; Lu *et al.*, 2003; Aloy *et al.*, 2004; Reiss *et al.*, 2004; Zhang *et al.*, 2004). One commonly used approach predicts that two proteins will interact if their orthologs have been shown to interact; such conserved interactions have been referred to as interlogs (Matthews *et al.*, 2001; Lehner *et al.*, 2004). Interactions have also been predicted between pairs of proteins with domains that are often observed in interacting proteins (Ng *et al.*, 2003).

1.7: Computational Resources

Currently, there is a wide variety of functional genomic data sets publicly available for the budding yeast *S. cerevisiae* which are described above; additional data sets are constantly being produced by existing and new high-throughput technologies. These data sets are often both large and complex and the analysis of this vast amount of data is now the key problem and computers in conjunction with effective software tools are an essential part of this process. Over the past few years there has been a rapid increase in the number of software tools available for the storage, visualisation and analysis of these data sets; a selection of the major resources available are described below.

1.7.1: Genome Resources

There is now a large amount of genome related data associated with *S. cerevisiae* that is being continuously generated by laboratories across the globe. This data ranges from the genome sequence and gene coordinates to descriptions and functional annotations of protein products. This vast amount of data requires efficient database systems to store and manage it as well as effective web interfaces to make it readily available to the scientific community. Currently, there are three main *S. cerevisiae* database resources available over the World Wide Web (Table 1.3).

***Saccharomyces* Genome Database (SGD)** **Cherry *et al.*, 1998** **http://www.yeastgenome.org/**
The SGD was established to provide a fast, easy and reliable method for yeast researchers to obtain information about the *S. cerevisiae* genome, the genes it contains and their possible interactions. The genome information in the SGD is organised around a 'locus' page for each ORF containing a summary of the gene, its protein product and any mutant phenotypes. The SGD contains an enormous amount of data on every ORF in *S. cerevisiae* and also provides a vast array of links to a number of relevant scientific web sites. In addition, the SGD makes a large proportion of its data publicly available for download and use.
Munich Information Centre for Protein Sequences (MIPS) **Comprehensive Yeast Genome Database (CYGD)** **Mewes *et al.*, 1998** **http://mips.gsf.de/genre/proj/yeast/index.jsp**
MIPS coordinated the collaborative effort of European groups during the *S. cerevisiae* genome sequencing project and now manages a web site that provides the yeast community with access to several genome databases. The information in MIPS is also organised around a web page for each ORF which contains a brief summary of the gene and a number of links to relevant data sources and web sites. In addition, MIPS makes a proportion of its data publicly available for download and use.
Yeast Proteome Database (YPD) **Garrels *et al.*, 1996** **http://www.incyte.com/control/researchproducts/insilico/proteome**
YPD began as a protein database rather than a genome database as emphasis was placed on providing detailed information about the *S. cerevisiae* proteins. Although much of YPDs data is included in MIPS and SGD, YPD excels at presenting its information in a very readable, compact form. It is important to note that the YPD recently became a commercial database that charges users a fee for access and use.

Table 1.3: *S. cerevisiae* online databases
This table contains the names and descriptions of the three main *S. cerevisiae* specific online database resources available to the yeast researcher.

These resources are primarily data warehouses, the main function of which is the dissemination of as much information as possible. Although, these resources do contain large amounts of information on all the genes in *S. cerevisiae* they have limited search and navigation mechanisms, basic visualisation tools and generally centre around displaying information on a single gene at a time as opposed to displaying information on entire groups of related genes at once to enable rapid comparison and analysis.

In addition to the *S. cerevisiae* specific resources described above there are also a number of more general resources that provide access to the fully sequenced genomes of other organisms, for example Schuler *et al.* (1996), Kyrpides (1999) and Peterson *et al.* (2001). Perhaps the most comprehensive of these resources is the National Centre for Biotechnology Information (NCBI) Entrez Genome database (Schuler *et al.*, 1996; http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome). The Entrez Genome database is publicly available and contains the whole genomes of a large number of viruses and over 100 other organisms. However, these resources are primarily focussed on providing information on the genome of specific organisms and do not utilise the wealth of functional genomic data available such as gene expression and protein-protein interaction data.

1.7.2: Gene Ontology Resources

Over the past few years, the Gene Ontology (GO) annotation system has been adopted by the majority of the world's major database repositories for plant, animal and microbial genomes. Furthermore, a wide variety of computational tools have now been developed that enable users to browse and search the GO annotation system itself as well as searching for the annotations of specific genes (Table 1.4).

AmiGO **Developed and maintained within the GO Consortium (Ashburner *et al.*, 2001)** **http://www.godatabase.org/**
AmiGO is an HTML based application that allows the user to browse, query and visualize data from the Gene Ontology. It allows the user to search for a GO term and view all gene products annotated to it, or search for a gene product and view all its associations. Users can also browse the ontologies to view relationships between terms as well as the number of gene products annotated to a given term.
GeneInfoViz **Zhou *et al.*, 2004** **http://genenet.org/geneinfoviz/search.php**
GeneInfoViz is a web based tool for batch retrieval of gene function information, visualization of GO structure and construction of gene relation networks. It takes an input list of genes and returns their functional annotation information. Based on the GO annotations of the given genes, GeneInfoViz allows users to visualize these genes in the DAG structure of GO, and construct a gene relation network at a selected level of the DAG.
GoFish **Berriz et al., 2003** **http://llama.med.harvard.edu/~berriz/GoFishWelcome.html**
GoFish is a Java application that allows users to search for gene products with particular gene ontology (GO) attributes, or combinations of attributes. GoFish ranks gene products by the degree to which they satisfy the search query.
GoMiner **Zeeberg *et al.*, 2003** **http://discover.nci.nih.gov/gominer/**
GoMiner is a Java-based program package that displays groups of 'interesting' genes within the framework of the GO hierarchy, both as a DAG and as the equivalent tree structure.
Onto-Express **Khatri et al., 2002** **http://vortex.cs.wayne.edu/projects.htm#Onto-Express**
Onto-Express (OE) is a novel tool to automatically translate lists of differentially regulated genes from microarray experiments into functional profiles characterizing the impact of the condition studied. OE constructs functional profiles (using GO terms) for the following categories: biochemical function, biological process, cellular role, cellular component, molecular function and chromosome location. Statistical significance values are calculated for each category.

Table 1.4: Gene ontology related computational resources
This table contains the names and descriptions of a few of the major Gene Ontology (GO) related computational resources. A full list of GO related computational tools is available at http://www.geneontology.org/GO.tools.shtml.

However, these tools tend to only be concerned with investigating the GO annotation system itself. They provide good mechanisms to visualise and browse the GO system

and search for specific terms and some tools permit the input of a group of gene names (such as the names of all genes within an expression cluster of interest) which can then collectively visualised and analysed. However, the user has to manually input gene names as these tools are not themselves integrated with other functional genomic data sources such as gene expression data.

1.7.3: Transcriptome Resources

The use of microarray technologies for the analysis of gene expression has increased dramatically over the past few years. As a result, there has been a relative explosion in the number of computational tools and resources available for the storage, visualisation and analysis of the data generated; a few of the major resources are described in Table 1.5. However, these resources tend to be solely aimed at the analysis of gene expression data, only a few have features to integrate other forms of data such as chromosome maps in Genesis (Sturn *et al.*, 2002), protein-protein interaction data in Expression Profiler (Brazma *et al.*, 2003) and cellular pathways in GeneSpring. In addition, only a few resources such as the yeast microarray global viewer (yMGV; Marc *et al.*, 2001) are aimed specifically at *S. cerevisiae*, which means that most resources are not utilising the vast array of additional information available on the genes being analysed such as GO annotations.

Cluster and TreeView **Eisen *et al.*, 1998** **http://rana.lbl.gov/EisenSoftware.htm**
Cluster and TreeView are an integrated pair of computer programs for visualising and analysing the results of complex microarray experiments. Cluster is a freely available Windows based computer program that is widely used for the analysis of gene expression data from microarray experiments; it performs a variety of data normalisation and cluster analysis techniques including hierarchical clustering, *k*-means clustering, Self-Organising Maps (SOMs) and Principal Component Analysis (PCA). TreeView is a freely available Windows based computer program that can be used to graphically browse the results of a hierarchical cluster analysis performed by Cluster; it supports tree and image based browsing of hierarchical trees and provides a number of output options for the generation of images.
GeneSpring **Silicon Genetics** **http://www.silicongenetics.com/cgi/SiG.cgi/Products/GeneSpring/index.smf**
GeneSpring is a commercial standalone program that is widely regarded as one of the leading tools for gene expression data analysis. It has a number of advanced features including: scripting, data normalisation, data clustering, 3D data visualisation, pathway views, expression profile comparison and statistical tools.
yeast Microarray Global Viewer (yMGV) **Marc *et al.*, 2001** **http://www.transcriptome.ens.fr/ymgv/**
yMGV is an online database providing a synthetic view of the transcriptional expression profiles of *S. cerevisiae* genes in a number of published expression data sets. yMGV displays a one-screen graphical representation of gene expression variations for each published genome-wide experiment, allowing a quick retrieval of experimental conditions having an effect upon expression of a selected gene. yMGV also provides tools to isolate groups of genes sharing similar transcription profiles in a defined subset of experiments.
Stanford Microarray Database (SMD) **Sherlock *et al.*, 2001** **http://genome-www5.stanford.edu/**
SMD stores raw and normalised data from microarray experiments from ongoing research projects at Stanford University and provides a web interface for the public to retrieve, analyse and visualise the data.
Genesis **Sturn *et al.*, 2002** **http://genome.tugraz.at/Software/Genesis/Description.html**
Genesis is a versatile, platform independent and easy to use Java suite for large-scale gene expression analysis. Genesis integrates various tools for microarray data analysis such as filters, normalization and visualization tools, distance measures as well as common clustering algorithms including hierarchical clustering, self-organizing maps, k-means, principal component analysis, and support vector machines. The results of the clustering are transparent across all implemented methods and enable the analysis of the outcome of different algorithms and parameters. Additionally, mapping of gene expression data onto chromosomal sequences has been implemented to enhance promoter analysis and investigation of transcriptional control mechanisms.

Table 1.5: Continued overleaf

Array Express **Brazma *et al.*, 2003** http://www.ebi.ac.uk/arrayexpress/
ArrayExpress is a public database of microarray gene expression data at the European Bioinformatics Institute (EBI), it is a generic gene expression database designed to hold data from all microarray platforms. ArrayExpress uses the annotation standard Minimum Information About a Microarray Experiment (MIAME) and the associated XML data exchange format Microarray Gene Expression Markup Language (MAGE-ML) and it is designed to store well annotated data in a structured way. The ArrayExpress infrastructure consists of the database itself, data submissions in MAGE-ML format or via an online submission tool MIAMExpress, an online database query interface and the Expression Profiler online analysis tool.

Table 1.5: Microarray related computational resources
This table contains the names and descriptions of a few of the major computational tools and resources available for the analysis and interpretation of gene expression data generated from microarray experiments.

The establishment of standards for microarray data annotation and exchange is a key issue currently being addressed by the Microarray Gene Expression Data society (MGED; http://www.mged.org). MGED is an international organisation of biologists, computer scientists and data analysts that aims to facilitate the sharing of microarray data. The current focus of MGED is on establishing standards for microarray data annotation and exchange, facilitating the creation of microarray databases and related software implementing these standards and promoting the sharing of high quality, well annotated data. The Minimum Information About a Microarray Experiment initiative (MIAME; Brazma *et al.*, 2001; http://www.mged.org/Workgroups/MIAME/miame.html) aims to outline the minimum information required to unambiguously interpret microarray data and to subsequently allow independent verification of this data at a later stage if required. MIAME is a set of guidelines that will assist with the development of microarray repositories and data analysis tools.

1.7.4: Proteome Resources

The use of high-throughput techniques in the detection of protein-protein interactions has increased rapidly over the past few years. As a result, there has also been an explosion in the number of computational tools and resources available for the storage, visualisation and analysis of protein-protein interactions. The majority of these resources are online database repositories for interaction data which have a simple graphical display tool (typically using a 'springs and rings' type algorithm); the major resources available are described in Table 1.6. However, most these resources are only concerned with protein-protein interactions and therefore do not incorporate other data such as the genomic location, GO annotations or gene expression profiles of the interacting proteins. In addition, relatively few are specifically aimed at the budding yeast *S. cerevisiae* and so do not utilise the wealth of functional genomic data available for this organism.

A Java applet for visualizing protein–protein interactions **Mrowka, 2001** **http://www.charite.de/bioinformatics/**
This is a web applet for browsing protein–protein interactions. It enables the display of interaction relationships, based upon neighbouring distance and biological function. This applet was one of the first protein-protein interaction visualisation tools to use a 'springs and rings' type algorithm.
Biomolecular Interaction Network Database (BIND) **Bader *et al.*, 2001** **http://www.bind.ca/**
BIND is an expanding database of biomolecular interaction, pathway and complex information. All information stored in BIND is freely available through a web interface that allows users to query, view and submit records. The interactions come from scientific literature, public submitters and other interaction databases.
Database of Interacting Proteins (DIP) **Xenarios *et al.*, 2000** **http://dip.doe-mbi.ucla.edu/**
DIP catalogues experimentally determined interactions between proteins. It combines information from a variety of sources to create a single, consistent set of protein-protein interactions. The data stored within the DIP database were curated manually and also automatically using computational approaches. The database is publicly available on the web and is intended to aid those studying protein-protein interactions, signalling pathways, multiple interactions and complex systems.
General Repository for Interaction Datasets (GRID) **Breitkreutz *et al.*, 2003** **http://biodata.mshri.on.ca/grid/servlet/Index**
GRID is a database of genetic and physical interactions. It contains interaction data from many sources, including several proteome wide studies and other interaction databases. GRID also has a software platform for the visualization of complex interaction networks called Osprey. Recently, the GRID database split into three organism specific databases called YeastGRID, FlyGRID and WormGRID. The YeastGRID database is now strongly linked to the SGD and incorporates the GO annotations of interacting proteins.
IntAct **Hermjakob *et al.*, (2004)** **http://www.ebi.ac.uk/intact**
IntAct provides an open source database and toolkit for the storage, presentation and analysis of protein interactions. It has a web interface that provides both textual and graphical representations of protein interactions and allows the exploration of interaction networks in the context of the GO annotations of the interacting proteins. A web service allows direct computational access to retrieve interaction networks in XML format.

Table 1.6: Continued overleaf

Molecular Interactions Database (MINT) **Zanzoni *et al.*, 2002** **http://cbm.bio.uniroma2.it/mint/**
MINT is a database designed to store functional interactions between biological molecules (proteins, RNA, DNA). Beyond cataloguing the formation of binary complexes, MINT was conceived to store other types of functional interactions namely enzymatic modifications of one of the partners. The interaction data can be easily extracted and viewed graphically with 'MINT Viewer'.
PathCalling Yeast Interaction Database **Uetz *et al.*, 2000** **http://portal.curagen.com/cgi-bin/com.curagen.portal.servlet.PortalYeastList**
PathCalling is a yeast specific interaction database that was initially designed to store the data generated from the Uetz *et al.* (2000) yeast two-hybrid study. It allows users to search for information on putative protein interactions, perform sequence analyses and view the results, extend interactions to construct pathways and to view homologues of the yeast genes. PathCalling has a basic visualisation tool that displays a static diagram of a protein and all the interactions it is involved in.
PIMRider **Hybrigenics** **http://pim.hybrigenics.com/pimrider/pimriderlobby/PimRiderLobby.jsp**
PIMRider is a commercial functional proteomics software platform for the exploration of reliable protein-protein interaction data and protein pathways.
Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) **Von Mering *et al.*, 2003** **http://www.bork.embl-heidelberg.de/STRING/**
STRING is a database of known and predicted protein-protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources: (1) Genomic Context; (2) High-throughput Experiments; (3) Co-expression; and (4) Previous Knowledge. STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable.

Table 1.6: Protein-Protein interaction related computational resources
This table contains the names and descriptions of some of the major computational resources available for the visualisation and analysis of protein-protein interaction data.

Currently, there are several well established databases for protein-protein interaction data. However, these databases provide their data in many different formats and are not synchronised with each other. Therefore, the task of combining interaction data from different sources is a common and tedious problem. The Proteomics Standards Initiative (PSI; Hermjakob *et al.*, 2004; http://psidev.sourceforge.net/) aims to define

community standards for data representation in proteomics to facilitate data comparison, exchange and verification. PSI is developing a common data standard for protein-protein interactions that will allow users to retrieve all relevant data from different sites and perform comparative analyses of different data sets much more easily than is currently possible. This standard will allow a synchronisation of the core data between public protein interaction database providers.

1.8: Integrated Analysis

The availability of complete genome sequences along with gene predictions has resulted in the development of new technologies such as microarrays and the yeast two-hybrid system enabling the analysis of gene expression and protein interactions on a genomic scale. These techniques have been used to sort genes and proteins into related groups based on shared expression profiles or interactions; the concept of guilt by association. However, these high-throughput techniques all have their own disadvantages and therefore the data obtained from any single approach should be interpreted cautiously. Furthermore, as the data from any single approach can only provide a tentative indication of a gene or protein function, it has been proposed that these limitations can be overcome by integrating data obtained from two or more distinct approaches (Walhout *et al.*, 1998; Vidal, 2001; Ge *et al.*, 2003). For example, a cluster of interacting proteins whose corresponding genes are similarly expressed under various experimental conditions and have similar GO annotations is likely to be more relevant than any other cluster for which additional information is not available. In addition, the expression profiles and GO annotations might indicate

dynamic and functional aspects of the cluster. Therefore, new biological insights are likely to emerge from the integration of data from different functional analyses and computers in conjunction with effective software tools are an essential part of this process.

Several groups have investigated the potential relationship between gene expression and protein interaction data sets (Ge *et al.*, 2001; Grigoriev, 2001; Mrowka *et al.*, 2001; Jansen *et al.*, 2002; Kemmeren *et al.*, 2002). Ge *et al.* (2001) combined a variety of high throughput and low throughput interaction data sets with expression data from cell cycle, sporulation and environmental stress experiments. A Protein Interaction Density (PID) value was calculated as the ratio of the number of observed interactions over the total number of possible interactions for a given set of proteins. PIDs were then compared between sets of protein pairs encoded by genes belonging to the same expression cluster (or intracluster pairs) and sets of protein pairs encoded by genes belonging to different clusters (or intercluster pairs). In general, average intracluster PIDs were found to be significantly greater than intercluster PIDs for interactome data sets, whereas the average intracluster and intercluster PIDs were similar for random data sets. Furthermore, low throughput data sets gave larger PIDs than high throughput data sets. This was interpreted as evidence that genes with similar expression profiles are more likely to encode interacting proteins and indicated that there was a global correlation between gene expression and protein interaction data. However, although the actual approach used seems to be sound (Mrowka *et al.*, 2003; Ge, Liu *et al.*, 2003), self-interacting protein interactions were not filtered out of the experimental data sets which would obviously bias the results.

Removal of these self-interactions was found to give similar results for the experimental and random data sets (Mrowka et al., 2003).

Grigoriev (2001) investigated the relationship between the similarity of expression patterns for a pair of genes and interaction of the proteins they encoded for both *S. cerevisiae* and the bacteriophage T7. Grigoriev (2001) found that, on average, the Pearson correlation coefficients of transcript abundance corresponding to interacting protein pairs were significantly higher (indicating a better correlation) for interactome data sets than for sets of random protein pairs. This led to the suggestion that protein pairs encoded by co-expressed genes interact with each other more frequently than with random pairs. Mrowka *et al.* (2001) compared a number of high and low throughput interaction data sets and found that interacting proteins from the low throughput data sets were much more closely related to each other with respect to transcription profiles when compared to the high throughput data sets. One explanation for this difference was the high false positives rates in the high throughput data sets. Jansen *et al.* (2002) integrated a variety of data sources for yeast to investigate the relationship of protein-protein interactions with mRNA expression levels. By focusing on known protein complexes with high confidence interactions they found that subunits of the same protein complex show significant coexpression. However, they also investigated the interactions in genome-wide data sets and found them to have only a weak relationship with gene expression. Kemmeren *et al.* (2002) showed how integration improves the utility of different types of functional genomic data by using collections of microarray expression data to assess the quality of different high-throughput protein interaction data sets and

provide functional annotation for a large number of previously uncharacterised genes. They found that, on average, the cosine correlation distances of transcript abundance corresponding to proteins pairs are significantly lower (indicating a better correlation) for interactome data sets than for random protein pairs. Werner-Washburne *et al.* (2002) created a novel tool for the visualisation and comparison of *S. cerevisiae* gene expression and protein-protein interaction data sets; visual analysis of the data using this tool showed no clear overall correlation between co-expression of genes and protein interactions. However, interesting insights were generated by focusing in on ribosomal proteins as opposed to analysing whole data sets.

Global relationships have also been examined for other pairwise combinations of functional genomic data sets. Cohen *et al.* (2000) investigated correlations between the expression patterns of genes on the same chromosome and found that in many cases adjacent pairs of genes, as well as nearby non-adjacent pairs of genes, showed correlated expression. Furthermore, they showed that genes with similar functions tended to occur in adjacent positions along the chromosome. Drawid *et al.* (2000) investigated the relationship between protein subcellular localisation and gene expression for a variety of *S. cerevisiae* whole genome expression data sets. They found high expression levels for cytoplasmic proteins, low levels for nuclear and membrane proteins and large fluctuating levels for excreted proteins. Fellenberg *et al.* (2000) developed a method for the integrative analysis of protein-protein interaction and functional classification data from *S. cerevisiae* to deduce hypotheses about the functional role of uncharacterised proteins. Ogata *et al.* (2000) investigated, for a number of different organisms, if enzymes located near each other

in the KEGG metabolic pathways (http://www.genome.jp/kegg/kegg2.html) were located near each other on the genome, forming Functionally Related Enzyme Clusters (FRECs). They found that the relative number of enzymes in FRECs was close to 50 % for *Bacillus subtilis* and *Escherichia coli* but was less than 10 % for *S. cerevisiae*. Ideker *et al.* (2001) developed an approach to integrate gene expression, protein expression and protein interaction data sets and assimilate them into biological models to predict cellular behaviour; they used this approach to investigate the properties and behaviour of the galactose-utilisation pathway. Jeong *et al.* (2001) and Oltvai *et al.* (2002) investigated correlations between high throughput protein-protein interaction and phenotype data sets in *S. cerevisiae* and found that proteins with large numbers of potential interaction partners (hubs) were often found to be essential.

As discussed above, there have now been a number of studies that have combined different functional genomic data sets together for integrated analysis which have led to some interesting insights; these studies most commonly integrate two different types of data sets in a pairwise fashion. However, there are currently very few computational resources available that enable users to perform analyses on the functional genomic data sets in an integrated fashion themselves (Table 1.7); in addition, these resources are only recent developments. Therefore, there is a now a clear need for a new generation of software tools that are capable of effectively integrating the wealth of data available for *S. cerevisiae* enabling users to readily utilise all of this data in their analyses and investigations.

Cytoscape **Shannon *et al.*, 2003** **http://www.cytoscape.org/**
Cytoscape is an open source software project for integrating biomolecular interaction networks with high-throughput expression data and other molecular states into a unified conceptual framework. Although applicable to any system of molecular components and interactions, Cytoscape is most powerful when used in conjunction with large databases of protein–protein, protein–DNA, and genetic interactions that are increasingly available for humans and model organisms. Cytoscape's software Core provides basic functionality to layout and query the network; to visually integrate the network with expression profiles, phenotypes, and other molecular states; and to link the network to databases of functional annotations. The Core is extensible through a straightforward plug-in architecture, allowing rapid development of additional computational analyses and features.
Genome Information Management System (GIMS) **Cornell *et al.*, 2003** **http://www.cs.man.ac.uk/img/gims/**
GIMS is an object database that integrates genomic data with data on the transcriptome, protein-protein interactions, metabolic pathways and annotations, such as gene ontology terms and identifiers. GIMS supports the running of integrated analyses over database and provides comprehensive facilities for handling and inter-relating the results of these analyses.
Database for Annotation, Visualisation and Integrated Discovery (David) **Dennis *et al.*, 2003** **http://www.david.niaid.nih.gov**
DAVID is a web-based tool that provides integrated solutions for the annotation and analysis of genome-scale datasets derived from high-throughput technologies such as microarray and proteomic platforms. Analysis results and graphical displays remain dynamically linked to primary data and external data repositories, thereby furnishing in-depth as well as broad-based data coverage. The functionality provided by DAVID accelerates the analysis of genome-scale datasets by facilitating the transition from data collection to biological meaning.
Genostar **http://www.genostar.org/**
Genostar is a bioinformatics platform for exploratory genomics offering a unified way of representing and managing data of various types and origins (high throughput sequencing, micro-arrays, proteomics, etc) through a set of software modules which can exchange information. The first version of Genostar consisted of three modules: (1) GenoAnnot provides an innovative solution to the annotation of genomic sequences; (2) GenoLink enables the exploration of relationships between data sets; and (3) GenoBool helps to identify correlations between data sets.

Table 1.7: Integrated computational resources
This table contains the names and descriptions of some of the major integrated computational resources.

1.9: Thesis Outline

This chapter has essentially given a broad overview of the subject areas relating to this PhD project. In Chapter 2, the specific aims and motivations behind this project are detailed and discussed. In Chapters 3 and 4, the features and functionality of the software tool developed through this project are described along with an overview of the functional genomic data sets used. In Chapter 5, a number of case studies are presented that demonstrate the utility of the developed tool to investigate the function of unknown genes. In Chapters 6 and 7, the utility of the tool in the analysis of correlations between functional genomic data sets is detailed and discussed along with the results from a number of correlation analyses; in addition, a number of case studies are presented that investigate specific genes and biological processes highlighted through the correlation analysis results. In Chapter 8, an overall discussion of the tool and the analysis results is presented along with concluding remarks and future directions.

Chapter 2

Aims

2.1: Concept

Although the budding yeast *Saccharomyces cerevisiae* (*S. cerevisiae*) is one of the most intensively studied eukaryotic organisms (due to its value as a model organism in biological research) there is still a great deal left to learn about this organism and the biological processes that maintain it. The genome sequencing project has essentially provided a complete catalogue of all the genes present in *S. cerevisiae* and the goal now is to understand the function of all the gene products and ultimately how they interact to create this simple eukaryotic organism. However, a large proportion of the genes in *S. cerevisiae* are still classified as genes of unknown function and additional information is needed to place them within a biological context. Ultimately, the validity and function of each gene can only be proven by experiments in the laboratory but given the number of unknown genes in the *S. cerevisiae* genome this could take some time. Therefore, there is a clear need for new experimental and computational methods to aid in the assignment of biochemical functionality; these methods could suggest possible biological roles for genes of unknown function which could then be validated by experiments in the laboratory.

Functional genomic strategies have become increasingly valuable in characterising novel genes discovered by genome sequencing projects. Many such strategies use the principle of 'guilt by association' as the means of elucidating function, i.e. genes that are coexpressed or proteins that interact with one another are likely to be involved in the same or related biological processes. Over recent years there has been a relative explosion of functional genomic data available for *S. cerevisiae* such as gene expression and protein-protein interaction data sets. As these data sets can be both

large and complex, the intelligent exploitation of them is dependent upon the provision of effective software tools. Software tools facilitate the exploration and analysis of these data sets by providing effective search, visualisation and analysis mechanisms. The overall aim of such tools is to aid in improving our biological understanding of *S. cerevisiae* by helping to functionally characterise individual genes and proteins, and to decipher how they work together to fulfil broader biological goals.

Over recent years, there has been a rapid increase in the number of software tools available for the visualisation and analysis of individual types of functional genomic data sets; for example, there are now many tools for the visualisation of protein-protein interactions (e.g. Mrowka, 2001) and many tools for the analysis of gene expression data (e.g. Eisen *et al.*, 1998). However, the majority of functional genomic strategies have weaknesses and disadvantages that can make the data sets produced incomplete and error prone. Combining data sets from the same strategy can reduce these disadvantages and therefore give greater confidence in any biological interpretations made from analyses of them. More importantly, many new biological insights are likely to emerge from the combined use of data from different functional genomic strategies. For example, there have now been a number of individual scientific studies that have integrated functional genomic data sets together for analysis which have led to some interesting biological insights (see section 1.8 of this thesis for more details). However, there are still relatively few software tools available that can effectively combine functional genomic data sets together and present them to the user for integrated visualisation and analysis.

Therefore, there is a clear need for a new generation of software tools that are capable of effectively integrating the wealth of functional genomic data available for *S. cerevisiae* enabling users to readily utilise all of this data in their analyses and investigations of specific genes and broader biological processes.

To this end, the first aim of this project was to design and develop a novel bioinformatics tool for the integrated visualisation and analysis of functional genomic data sets from the budding yeast *S. cerevisiae*. The initial data sets considered were gene expression data from microarrays, protein-protein interaction data from yeast two-hybrid screens as well as functional annotation data on the genes and proteins of *S. cerevisiae*; these data sets were selected as they were generated from exciting modern technologies and the combination of them had the potential to yield interesting associations. This tool was planned to be a user friendly workbench that would enable both wet and dry laboratory scientists to easily explore any and all aspects of the data in an integrated modular fashion. The second aim of this project was to use the developed tool to try and assign biochemical functionality to genes of unknown function, investigate specific biological processes, analyse the stored functional genomic data sets individually and investigate possible correlations between them.

2.2: Software Life Cycle

As one of the primary aims of this project involves the design and development of a software product, it is important to give an overview of the software life cycle at this

point. The software life cycle can be defined as the period of time beginning when a software product is conceived and ending when the product is no longer available for use. The software life cycle is typically broken into phases denoting activities such as requirements, design, programming, testing, installation, and operation and maintenance. There are many different software life cycle models such as the waterfall, prototyping, incremental, rapid application development, transformation and spiral models; for more information on the different software life cycles models see, for example, Jacobson *et al.*, 1999. The software life cycle model that best describes the design and development of the software product in this project is shown in Figure 2.1. Briefly, after the initial concept for the project was devised, potential users were consulted and an initial system design was drawn up. The development of the system then went through a number of cycles of coding and testing with a new version of the system released at the end of each development cycle, ultimately resulting in the release of the final version of the system at the end of the project. Overall, the system went through four broad cycles of development which are described in detail in section 2.6, "System Development", below.

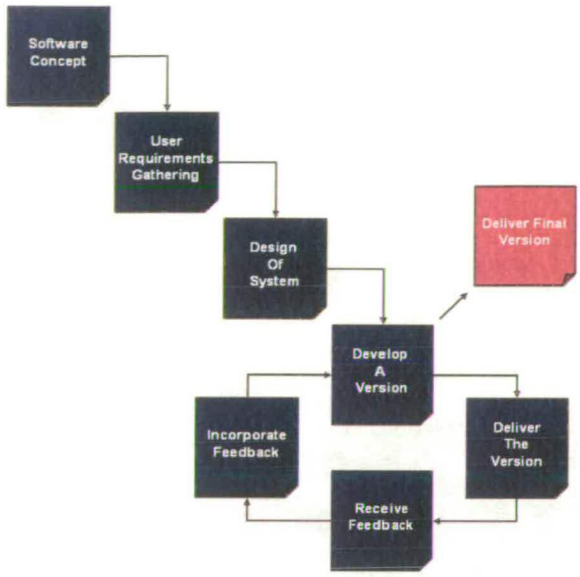


Figure 2.1: The Software Life Cycle
This is a diagram of the software life cycle model that best describes the design and development of the software product in this project.

2.3: User Requirements

A 'user requirement' can be defined as a condition or capability needed by a user to solve a problem, achieve an objective or increase productivity and the 'requirements gathering phase' can be defined as the period of time in the software life cycle during which the user requirements, such as functional and performance capabilities, are identified and documented. The requirements gathering phase is therefore one of the most important phases as it forms the basis for the design and implementation phases that follow. In this project, the initial target users were members of Professor Jean Beggs's laboratory (http://homepages.ed.ac.uk/jeanb/) in the Institute of Cell and Molecular Biology, University of Edinburgh. However, it is important to note that the other primary user in this project was also myself as I would be using the

developed software product to explore the stored functional genomic data sets and investigate possible correlations between them.

As described above, the initial concept of the project was to develop a bioinformatics tool for the integrated visualisation and analysis of functional genomic data sets from the budding yeast *S. cerevisiae*. Therefore, preliminary meetings were organised with the initial target users to discuss the potential usefulness of such a tool, what essential features would be needed and what novel features would be useful; essentially, this was the requirements gathering phase of the project. The concept for the tool received good feedback from the target users and was further backed up by observations of their current working practices. The target users would typically use multiple computational resources to find information on a specific gene of interest. For example, the *Saccharomyces* Genome Database (SGD; Cherry *et al.*, 1998) would be used to view textual information such as descriptions and annotations on a specific gene but an alternative resource would need to be used to view the gene's corresponding protein-protein interactions (e.g. PathCalling; Uetz *et al.*, 2000); furthermore, another resource would need to be used to view the expression data on the gene (e.g. Cluster; Eisen *et al.*, 1998). This lack of integration between these resources and their corresponding data sets was evidently a problem as if any genes were found to be of interest in one resource their names would have to be manually noted and subsequently entered into the other resources to be investigated further. Therefore, this would often result in users manually noting down gene names and constantly shifting between different resources to examine relevant data in the process of their investigation. In addition, users were often interested in investigating

the properties of multiple genes of interest. However, as the existing resources revolved around a single gene approach, users would have to investigate each of the genes individually, as opposed to collectively, making comparisons of their properties tedious. Whereas a group approach in conjunction with integrated functional genomic data sets would enable all the genes involved in an entire biological process to be collectively examined as a whole to investigate the dynamics of how they are working together to achieve their biological goal and to also examine what other genes they may be working with. Furthermore, this approach would enable any features of interest from one functional genomic data set to be selected and collectively investigated in further detail in the other data sets; for example, investigating if all the genes located in a specific expression cluster share similar functions and encode proteins that interact with one another.

After meeting with the target users, the essential features for the planned software tool were identified as easy to use navigation, search and display mechanisms combined with clear graphical representations of the data. While the novel features for the tool were identified as: (1) A modular or group approach enabling the collective investigation of all the properties of an entire group of genes at once; and (2) Effective integration of the data enabling users to select a feature of interest from one data set to collectively investigate further as a whole in the other data sets. In conclusion, the target users were in favour of the development of an easy to use but advanced tool for the visualisation and analysis of *S. cerevisiae* functional genomic data sets in an integrated modular fashion.

2.4: Existing Tools

This project began in October 2000 and at this time there were relatively few computational tools available for the visualisation and analysis of *S. cerevisiae* functional genomic data sets compared to the large variety available today. The available tools tended to be either data warehouses centred on displaying a large amount of textual information on a single gene of interest or tools for the visualisation and analysis of only a specific type of functional genomic data set. There were no established tools available that could effectively integrate the wealth of functional genomic data available for *S. cerevisiae* and none that could utilise a group approach in the analysis of the data.

The major computational resources available to *S. cerevisiae* researchers were the SGD, the Munich Information Centre for Protein Sequence (MIPS; Mewes *et al.*, 1998) and the Yeast Proteome Database (YPD; Garrels *et al.*, 1996). However, these resources were primarily data warehouses, the main function of which was the dissemination of as much information as possible on the genes of *S. cerevisiae.* These resources revolved around a single gene approach and were essentially designed to search for and subsequently display a datasheet on a single gene of interest. They had fairly limited and rigid search and navigation systems to find information where the main and sometimes only way of searching for information was by entering a single gene name which typically led to a datasheet on that gene. Although this is an essential feature, more flexible search mechanisms allowing keyword searches of descriptions were seldom provided; those that were would

simply lead to a list of all the genes associated with the keyword, each of which would have to be examined individually to see what they were and what their function was. Whereas, more flexible search mechanisms combined with a group approach for analysis would allow the data on an entire group of genes to be easily searched for and then collectively displayed enabling users to investigate entire biological processes as a whole. Furthermore, the above resources generally displayed data in a textual format; although some graphical representations of data were provided, such as an image of the chromosomal region surrounding a gene of interest, these displays tended to be relatively basic. Whereas more intuitive and dynamic graphical representations of the data would enable users to easily and rapidly explore the data and then select any features of interest to investigate further collectively.

There were also a number of computational tools available for the visualisation and analysis of specific types of functional genomic data sets. However, these tools tended to be focussed on a single data type and none of these were specifically aimed at *S. cerevisiae* and so did not utilise the wealth of other functional genomic data available. For example, Cluster (Eisen *et al.*, 1998) was a widely used computational tool for the analysis of gene expression data from microarray experiments. It could perform a variety of data normalisation and cluster analysis techniques including hierarchical clustering, the results of which could be graphically viewed and browsed in its associated computational tool Treeview (Eisen *et al.*, 1998). However, although Cluster and Treeview were good tools for the analysis and subsequent visualisation of gene expression data, they were only concerned with gene expression data and

therefore did not utilise the wealth of other functional genomic data available for *S. cerevisiae*. Furthermore, although annotations of the genes analysed could be incorporated into the input files, this data needed to be incorporated manually by the users themselves.

There were also a number of computational tools available for the visualisation and analysis of protein-protein interaction data. For example, PathCalling (Uetz *et al.*, 2000) was a computational tool specifically designed for the protein-protein interaction data generated from the Uetz *et al.* (2000) yeast two-hybrid study. However, this tool had limited search mechanisms, basic graphical displays and although it did include brief descriptions of the interacting proteins it did not utilise the wealth of other functional genomic data available for *S. cerevisiae*. The Database of Interacting Proteins (DIP; Xenarios *et al.*, 2000) contained interactions manually curated from the scientific literature. Although it provided a number of effective search mechanisms, these searches would simply return textual lists of interactions, as opposed to graphical displays, each of which would need to be examined individually. Furthermore, although it did contain a brief amount of information on the interacting proteins, it too did not utilise the wealth of other functional genomic data available for *S. cerevisiae*.

Therefore, there was a clear need to design and develop a new tool that would combine the advantages of the existing data warehouses, by containing a large variety of information on every gene in *S. cerevisiae*, with the advantages of the existing visualisation and analysis tools, by enabling users to explore the stored

functional genomic data sets. However, the tool would also need to utilise the wealth of functional genomic data available for *S. cerevisiae* and be able to effectively integrate the different types of data sets together as well as utilising a group approach that would enable users to collectively investigate all the properties of an entire group of genes at once

2.5: System Design

After the initial discussions with the target users, the next step in the development process was to sketch out an initial system design of the planned software tool (Figure 2.2). The initial design split the system into two parts: (1) A database for the storage and management of the data; and (2) An associated software tool for the integrated visualisation and analysis of the data. The database was planned to store a variety of information on all the genes of *S. cerevisiae* in conjunction with a variety of functional genomic data sets. The source of the information on the genes of *S. cerevisiae* was initially identified as the SGD and the initial functional genomic data sets that were considered were protein-protein interactions from large scale yeast two-hybrid screens (Uetz *et al.*, 2000 & Ito *et al.*, 2000) and genome scale gene expression microarray experiments (e.g. Eisen *et al.*, 1999 & Gasch *et al.*, 2000). The initial system design split the software tool itself into a number of inter-linked sections, namely: (1) A Genome Section for the visualisation and analysis of the *S. cerevisiae* genome; (2) A Transcriptome Section for the visualisation and analysis of gene expression data; (3) A Proteome Section for the visualisation and analysis of protein-protein interactions; (4) A Cell Section for the visualisation and analysis of a

typical *S. cerevisiae* cell; and (5) An Analysis Section for searching for information and collectively visualising data on all the search results.

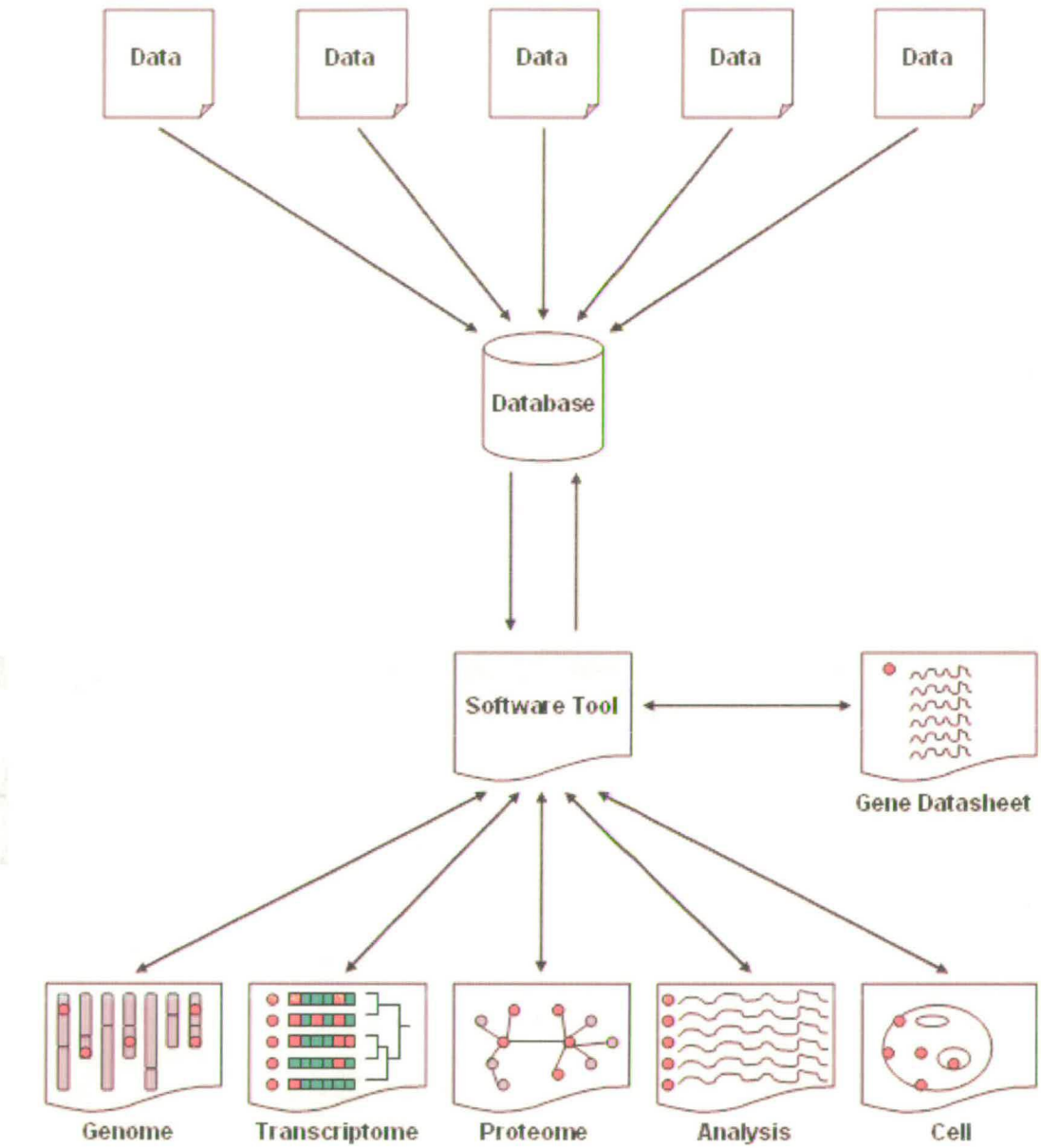


Figure 2.2: Initial system design of the software tool
This is a schema of the initial system design of the planned software tool. Briefly, the system is comprised of a database for the storage and management of the data and a software tool for the visualisation and analysis of the data. The software tool is split into a number of inter-linked sections and utilises a group approach enabling all the properties of an entire group to be analysed collectively. For example, data on the 5 genes highlighted in red on the Genome Section can be collectively viewed in the Transcriptome, Proteome, Analysis and Cell Sections, and vice versa.

The software tool was planned to utilise a group approach that, combined with the inter-linked sections, would enable users to easily select a feature of interest from one section and then swiftly move to any of the other sections where the corresponding data related to their selection would be automatically displayed and highlighted. Therefore, the tool would enable all the properties of an entire group of genes to be collectively investigated. For example, the chromosomal region surrounding a gene of interest could be selected in the Genome Section and then all the genes in this region could be collectively investigated in the other sections to examine if they are coexpressed, if their encoded products interact, if they share similar functions and if they are located in the same cellular location. Furthermore, the tool was planned to be easy to use with simple navigation and functional features, have flexible search mechanisms and provide clear graphical representations of the data enabling users to easily and rapidly find the data they want, investigate the intricacies of broad biological processes and test specific hypotheses. In addition to the initial target users, the software tool was also aimed at both wet and dry laboratory scientists with an interest in *S. cerevisiae* who would use the tool as a workbench to investigate specific genes and biological processes and to easily explore any and all aspects of the functional genomic data in an integrated modular fashion. The typical questions that this software tool aimed to help users answer are detailed in Table 2.1 below; these questions were identified through discussions with the target users. Furthermore, although the tool was specifically aimed at the budding yeast *S. cerevisiae*, it was designed with flexibility in mind so that it could be applied to other organisms with relative ease in the future.

I am interested in a particular gene of unknown function; What can this tool tell me about it and can it help me to assign biochemical functionality to it?
I am interested in a particular gene of known function; What can this tool tell me about it and what other genes it is working with to achieve its biological goals?
I am interested in a particular gene which I believe is involved in a particular biological process; Can this tool help me to investigate this?
I am interested in a specific biological process; What can this tool tell me about it, what proteins are involved and how are they working together? Can this tool help identify any new proteins of unknown or known function involved in this process?
I am interested in a specific chromosomal region; What can this tool tell me about it, what genes are located within it, what are their functions and do they work together to achieve common biological goals? Can this tool help characterise any proteins of unknown function in this region?
I am interested in a particular hierarchically clustered gene expression data set; How can this tool help me explore this data set? I am interested in a particular gene expression cluster from this data set; What can this tool tell me about it, what genes are located within it, what are their functions and do they work together to achieve common biological goals? Can this tool help characterise any genes of unknown function in this cluster?
I am interested in a particular protein-protein interaction data set; How can this tool help me explore this data set? I am interested in a particular protein interaction cluster from this data set; What can this tool tell me about it, what proteins are located within it, what are their functions and do they work together to achieve common biological goals? Can this tool help characterise any proteins of unknown function in this cluster?
I am interested in two groups of proteins that I believe are evolutionary or functionally related; Can this tool help me investigate this?

Table 2.1: Typical user questions the software tool aims to address
This table contains the typical questions the software tool aims to help users answer.

The initial system design also identified the computational technologies that would be used to actually build the system itself; the database and software tool would be built using MySQL (http://www.mysql.com) and Java (http://java.sun.com/), respectively. MySQL is an open source database management system that is fast, compact, stable and is available for most of the major computer platforms. The Java programming language is a state-of-the-art, object-oriented language with a syntax similar to the C++ programming language and is also available for most of the major

computer platforms. Furthermore, Java has a rich set of routines to support Graphical User Interface (GUI) creation, communication with databases and web based applications. Taken together, MySQL and Java would therefore enable the creation of a fast, stable, user-friendly, platform independent and web-enabled system. Before construction of the system began, the initial system design was discussed with and approved by the initial users.

2.6: System Development

The planned system was fairly large but could be effectively split up into two parts, namely the database and the software tool; furthermore, the software tool itself could be effectively split up into a number of inter-linked sections. Therefore, the system was developed in a number of stages starting from the development of the database and the core architecture of the software tool followed by the development and subsequent integration into the system of each individual section of the software tool. Each stage resulted in the release of a new version of the system which was delivered to the users for testing and feedback.

The first stage of development involved identifying and obtaining data on all the genes in the *S. cerevisiae* genome, building the database to effectively store this data and building the Genome and Analysis Sections of the software tool which could utilise this data. The SGD was identified as the initial source of this data as it contains a large amount of information on all the genes of *S. cerevisiae* and makes a large proportion of this publicly available. The initial data obtained and processed

included the name, size, location, description, GO annotations and phenotype of every gene in the genome. The MySQL database to store this data was then designed, built and subsequently loaded with the data obtained from the SGD. The core section of the software tool concerned with initialisation and communicating with the database was then developed. This was quickly followed by the development and integration of the Genome and Analysis Section. The result of this stage was the internal release of a software tool called the Virtual Yeast Cell (Version 1) which was a standalone system that could be used to visualise and analyse the *S. cerevisiae* genome as well as for searching for information and collectively visualising data on all the search results. This version highlighted the main principles underlying the whole system as it utilised a group approach for analysis, had two inter-linked sections and offered clear graphical representations of the data. As a result, this version received good feedback from the initial target users and the further development of the system was approved.

The second stage of development involved expanding the system to incorporate the Proteome Section and its associated data. The initial data identified for the Proteome Section were the two large scale yeast two-hybrid screens of the time (Uetz *et al.*, 2000; Ito *et al.*, 2001). This data was processed and subsequently integrated into the database. The Proteome Section itself was then developed and inter-linked with the other sections of the software tool. During this stage of development, the entire system was also made available for use over the World Wide Web enabling users to use the system without having to install the database or program locally. As described above, the initial name given to the system was the Virtual Yeast Cell.

However, this name frequently gave the impression that the main aim of the system was to recreate a living *S. cerevisiae* cell *in silico* as opposed to being a workbench for the integrated visualisation and analysis of *S. cerevisiae* functional genomic data sets. Therefore, the name of the system was changed to the Yeast Exploration Tool Integrator (YETI) during this stage. The result of this stage was the public release of YETI Version 1 (Orton *et al.*, 2004) which could be used as a standalone or web based system. YETI Version 1 had all the features and functionality of the Virtual Yeast Cell Version 1 but also included the Proteome Section for the visualisation and analysis of protein-protein interactions. This Proteome Section was effectively inter-linked with both the Genome and Analysis Sections and could also utilise a group approach for analysis enabling users to explore the stored data sets in an integrated modular fashion.

The third stage of development involved expanding the system to incorporate the Transcriptome Section and its associated data. The initial data identified for the Transcriptome Section were two large microarray studies that monitored the expression of all the genes in *S. cerevisiae* under a number of environmental conditions (Gasch *et al.*, 2000; Gasch *et al.*, 2001). This data was processed and subsequently integrated into the database. The Transcriptome Section itself was then developed and inter-linked with the other sections of the software tool. During this stage, additional features were also added to the existing sections that enabled them to utilise the recently incorporated gene expression data. The result of this stage was the public release of YETI Version 2 which had all the features and functionality of YETI Version 1 but also included the Transcriptome Section for the visualisation and

analysis of gene expression data. This Transcriptome Section was effectively inter-linked with the Genome, Analysis and Proteome Sections and could also utilise a group approach for analysis enabling users to further explore the stored data sets in an integrated modular fashion.

As mentioned previously, at the end of each stage of development the latest version of the system was delivered to the users for testing and feedback. The latest version of the system would be installed on the user's computer and the user would be given a demonstration of how the system works and how to use all of the features and functions. After a few weeks, the users would be met with to discuss any problems, bugs and suggestions for improvement. At this point, it is important to note that the user base was always expanding, especially after the system was made publicly available for use over the World Wide Web, and feedback from these additional users was always invited via email.

In addition, there was also an a final stage of development which involved expanding the system further to incorporate a number of additional sections that enable users to directly investigate possible global correlations between the stored functional genomic data sets, specifically: (1) A Genome vs Transcriptome Section to investigate possible correlations between gene location and gene expression; (2) A Genome vs Proteome Section to investigate possible correlations between gene location and protein interaction; and (3) A Proteome vs Transcriptome Section to investigate possible correlations between protein interaction and gene expression. During this stage of development no new data needed to be incorporated into the

database, however, some new data was generated through analysis of the existing data sets. Furthermore, these additional correlation sections were effectively inter-linked in YETI through the Analysis Section. The result of this stage was Version 3 of YETI which had all the features and functionality of YETI Version 2 but with the additional correlation analysis sections integrated into the system.

Further details on the features and functionality of the YETI system are now discussed in the forthcoming chapters of this thesis along with a number of case studies and analyses which demonstrate the utility of the tool.

Chapter 3

YETI Data & Database

3.1: Introduction

The Yeast Exploration Tool Integrator (YETI) is a novel bioinformatics tool for the integrated visualisation and analysis of *Saccharomyces cerevisiae* (*S. cerevisiae*) functional genomic data sets. Essentially, YETI consists of two parts: (1) A database for the storage and management of data; and (2) A Java program for the integrated visualisation and analysis of data. The YETI database is populated with publicly available data from both online databases and published scientific studies. However, this data needs to be checked and processed into the necessary formats before it can be imported into the YETI database. Therefore, a number of computer programs were written to extract the relevant data, check and process it into the necessary formats and then automatically update the YETI database. The data used to populate the YETI database can be split into three categories:

1) **Genome:** genomic data from the *Saccharomyces* Genome Database (SGD; Cherry *et al.*, 1998; http://www.yeastgenome.org).
2) **Transcriptome:** gene expression data from the Stanford Microarray Database (SMD; Sherlock *et al.*, 2001; http://genome-www5.stanford.edu/).
3) **Proteome:** protein-protein interaction data from the General Repository for Interaction Datasets (GRID; Breitkreutz *et al.*, 2003; http://biodata.mshri.on.ca/grid/servlet/Index).

Each of these three categories of data has had a separate computer program written for data processing. These programs have been designed for use by advanced users

only as they involve modifying large amounts of essential data in the YETI database and have limited error handing capabilities. However, the average user does not need to use these programs as updated versions of the YETI database itself are regularly available from the YETI website (http://www.bru.ed.ac.uk/~orton/yeti.html). In addition, the web based version of YETI (Web YETI) automatically connects to the latest database at the University of Edinburgh and the standalone version of YETI (Standalone YETI) also has an option to connect to this database.

3.2: Genome Data

The SGD is perhaps the largest information resource available for *S. cerevisiae*; it contains a wealth of genomic and biological information on the genes of *S. cerevisiae*, is constantly updated by a number of database curators and is a central resource for the yeast community. Therefore, the SGD is widely used and respected by yeast researchers. A large amount of its data is publicly available from the SGD data download site (ftp://genome-ftp.stanford.edu/pub/yeast/) in the form of text files. YETI currently uses six of these files to populate its own database (Table 3.1) with the extracted data ranging from gene names and chromosomal locations to descriptions of gene products and GO annotations.

File Name & Location	Description
data_download/ chromosomal_feature/ SGD_features.tab	This file contains information on all the current chromosomal features in the SGD. It also contains the coordinates of introns, exons and other subfeatures that are located within a chromosomal feature.
data_download/ literature_curation/ orf_geneontology.tab	This file contains the primary set of GO annotations for every ORF in the *S. cerevisiae* genome.
data_download/ literature_curation/ gene_association.sgd.gz	This file contains all the GO annotations for all *S. cerevisiae* gene products (protein and RNA).
data_download/ literature_curation/ go_slim_mapping.tab	This file contains the mapping of all *S. cerevisiae* gene products (protein and RNA) to a GO Slim annotation term.
data_download/ literature_curation/ go_terms.tab	This file contains detailed definitions of all the GO annotations used to characterise all the *S. cerevisiae* gene products.
data_download/ literature_curation/ phenotypes.tab	This file contains phenotype data for *S. cerevisiae* gene products; the majority of this data is from the systematic deletion project (Winzeler *et al.*, 1999).

Table 3.1: SGD data files used to populate the YETI database
This table contains the names, locations and descriptions of the six files available from the SGD data download site that are currently used to populate the YETI database.

Essentially, the data from the SGD is the core data of the YETI database because it defines the number of ORFs in the *S. cerevisiae* genome along with the name, type and location of each ORF. Recently, the SGD began characterising all ORFs as either verified, uncharacterised or dubious; these categories are defined by the SGD as follows:

1) **Verified:** ORFs for which experimental evidence exists that a gene product is produced in *S. cerevisiae*. Generally these have obvious orthologs in one or more other *Saccharomyces* species. Most named genes are in this class.

2) **Uncharacterized:** ORFs that are likely to be real due to the existence of orthologs in one or more other species, but which are not supported with specific experimental data demonstrating that a gene product is produced in *S. cerevisiae*. A few named genes may be in this class if there is no experimental evidence that they are produced. Evidence from large-scale analyses that indicates an ORF may be biologically relevant is sometimes but not always enough to upgrade an ORF from "Uncharacterized" to "Verified", depending on the individual case.

3) **Dubious:** ORFs which are not conserved in other *Saccharomyces* species and for which there is no experimental evidence that a gene product is produced in *S. cerevisiae*. Many ORFs classified as "Dubious" are small and overlap a larger ORF of the class "Verified" or "Uncharacterized"; however, overlap with another ORF does not mandate that an ORF be classified as "Dubious."

This new characterisation system can be used to eliminate ORFs that are highly unlikely to be real genes from analyses and investigations as well as highlighting those that may not be real. However, knowing the name and location of each ORF in the *S. cerevisiae* genome means little in the absence of what the function of each ORF is. Therefore, the SGD data is even more important because it describes the function of each ORF through textual descriptions of gene products and GO annotations. GO annotations provide a standard for characterising gene products and can readily be used to examine all the genes located in a specific cellular component or involved in a specific biological process. Furthermore, as standard GO annotations can be very specific, the SGD also characterises each ORF with a set of GO Slim

annotations. GO Slims are a cut-down version of the complete GO ontology and give a broad overview of the ontology content without the detail of the specific fine grained terms. Both the standard and slim GO annotations for each ORF in the *S. cerevisiae* genome are utilised in YETI.

A single combined Java program called YETI_SGD was written to collectively process all the required data from the six SGD data files described in Table 3.1. This program extracts all the required data from each of the files, checks the data, combines portions of it, assigns the appropriate ID numbers and then automatically updates the YETI database. The YETI_SGD program plays an essential role in keeping YETI an up-to-date resource as it updates the database with all the relevant data from the SGD; it can be run manually at any time or set up to run at regular intervals by the host operating system. However, it is important to note that if the data files available from the SGD change format, which has happened numerous times over the past few years, the YETI_SGD program will have to be modified in order to cope with the changes and still perform its function; in extreme cases of change, the YETI database and program will also have to be modified. This highlights one of the problems with using third party data sources in that you do not have control over the format or assurances on its continued availability.

A brief overview of the SGD data set currently stored in the YETI database is presented in Table 3.2; as can be seen a large proportion of the ORFs in the *S. cerevisiae* genome are still classified as genes of unknown function highlighting how much is still left to learn about this organism. It is important to note that the function

of a verified ORF is not necessarily known (899 verified ORFs are characterised as unknown GO molecular function) and the function of an uncharacterised ORF is not necessarily unknown (197 uncharacterised ORFs are classified as known GO molecular function).

Category	Total Number
Genomic Features	7783
ORFs	6591
Verified ORFs	4303
Uncharacterised ORFs	1470
Dubious ORFs	818
Unknown GO Function	2172
Unknown GO Process	1562
Unknown GO Component	868

Table 3.2: Overview of the current SGD data set
This table contains an overview of the SGD data set currently stored in the YETI database which was used for all the analyses and case studies presented in this thesis. Genomic Features represents the total number of genomic features in the *S. cerevisiae* genome; ORFs represents the total number of ORFs in the *S. cerevisiae* genome; Verified ORFs, Uncharacterised ORFs and Dubious ORFs represents the total number of verified, uncharacterised and dubious ORFs in the *S. cerevisiae* genome, respectively; Unknown GO Function, Unknown GO Process and Unknown GO Component represents the total number of non-dubious ORFs characterised with unknown GO molecular function, biological process and cellular component annotations, respectively.

3.3: Transcriptome Data

The SMD stores a large amount of raw and normalised data from microarray experiments from ongoing research projects at Stanford University. This data is available in a variety of formats ranging from raw microarray image files to normalised and clustered gene expression ratio tables. At present, the data from two large *S. cerevisiae* genome wide microarray studies available from the SMD are used to populate the YETI database:

1) **Genomic expression programs in the response of yeast cells to environmental changes (Gasch *et al.*, 2000):** In this study, spotted (two-colour) DNA microarrays were used to measure changes in transcript levels over time for almost every *S. cerevisiae* gene as cells responded to temperature shocks, hydrogen peroxide, the superoxide generating drug menadione, the sulfydryl-oxidising agent diamide, the disulfide reducing agent dithiothreitol, hyper- and hypo-osmotic shock, amino acid starvation, nitrogen source depletion and progression into stationary phase. A total of 93 individual microarray experiments, grouped into 13 related categories, were used to monitor how *S. cerevisiae* cells responded (via gene expression) to changes in a number of environmental conditions.
2) **Genomic expression responses to DNA damaging agents and the regulatory role of the yeast ATR homolog Mec1p (Gasch *et al.*, 2001):** In this study, spotted (two-colour) DNA microarrays were used to observe genomic expression of wild-type and mutant *S. cerevisiae* cells responding to the methylating agent methylmethane sulfonate (MMS) and ionising radiation. A total of 40 individual microarray experiments, grouped into 7 related categories, were used to monitor how different *S. cerevisiae* cell types responded (via gene expression) to a number of DNA damaging agents.

These two studies were chosen as the initial microarray data sets to populate the YETI database with; they are large genome wide data sets that complement each other well, monitoring how *S. cerevisiae* cells respond to a wide variety of environmental conditions and DNA damaging agents. The Gasch *et al.* (2000) study

is especially well respected and as a result is commonly used as a test data set for gene expression analysis programs; for example, the Gasch *et al.* (2000) data set was selected at the Yeast 2003 Conference (http://www.yeastgenome.org/community/meetings/yeast03/) as a test data set to enable biologists to easily compare the functions and performance of microarray analysis programs. The data generated from these two studies are publicly available for download from the SMD in the a variety of formats including output files generated by the Cluster program (Eisen *et al.*, 1998) after a hierarchical cluster analysis has been performed.

Cluster is a freely available Windows based computer program that is widely used for the analysis of gene expression data from microarray experiments; it performs a variety of data normalisation and cluster analysis techniques including hierarchical clustering, *k*-means clustering, Self-Organising Maps (SOMs) and Principal Component Analysis (PCA). Cluster is perhaps most commonly used for hierarchical clustering which is a conceptually simple yet very effective method of clustering gene expression data. The results of such an analysis can be represented in a visual manner that is easily comprehensible to the human mind even when hundreds of experiments are analysed on a genomic scale.

After a hierarchical cluster analysis has been performed, Cluster generates three output files containing the clustering results (Table 3.3). The clustering results can then be visualised using an associated program called TreeView (Eisen *et al.*, 1998) by importing the three Cluster output files. TreeView is a freely available Windows

based computer program that can be used to graphically browse the results of a hierarchical cluster analysis performed by Cluster; it supports tree and image based browsing of hierarchical trees and provides a number of output options for the generation of images.

File Name	Description
.cdt (clustered data table)	This file contains the original or normalised (depending on the selection) gene expression ratio data table with the rows and columns reordered based on the hierarchical clustering result. It also contains unique identifiers for each gene and microarray experiment that relate to the .gtr and .atr files.
.gtr (gene tree) **.atr (array tree)**	These two files contain the history of node joining events from the gene (.gtr) and array (.atr) clustering processes; the history of node joining events is used to recreate the resultant hierarchical tree. When clustering begins each item to be clustered is assigned a unique identifier and it is these identifiers that relate to the .cdt file. As each node is generated it is also assigned a unique identifier and each joining event is stored as a row with the node identifier, the identifier of the two joined elements and a similarity score between the two joined elements.

Table 3.3: Output files generated by the Cluster program
This table contains the names and descriptions of the three output files generated by Cluster after a hierarchical cluster analysis has been performed.

Cluster is a fairly advanced program for gene expression analysis but TreeView is a fairly basic visualisation program. TreeView was designed for the sole purpose of visualising the results of a hierarchical cluster analysis performed by Cluster. It has limited search functions to find and subsequently view relevant data, it does not utilise the underlying gene expression data tables and it is not integrated with any other data sources or resources. In addition, both Cluster and TreeView can only be used on the Windows platform which limits their usability.

A Java program called YETI_Cluster was written to process the output files generated by Cluster after a hierarchical cluster analysis has been performed and then import the results into the YETI database. The YETI_Cluster program checks all gene names, assigns the appropriate ID numbers to link the data into the YETI database, calculates precise coordinates for drawing the resultant hierarchical tree and then updates the YETI database. The YETI_Cluster program was used to process and subsequently import the Gasch *et al.* (2000) and Gasch *et al.* (2001) Cluster output files downloaded from the SMD.

Essentially, this means that the original Cluster program can be used to normalise and hierarchically cluster any *S. cerevisiae* spotted (two-colour) gene expression microarray data set and then the associated output files can be processed and imported into the YETI database using the YETI_Cluster program. YETI is then able to access the database to retrieve the expression data for visualisation and analysis via the YETI Transcriptome Section. In essence, the Transcriptome Section is a much more sophisticated version of TreeView that is fully integrated with all the other YETI Sections, has advanced search and display features and is not limited to the Windows platform.

The YETI_Cluster program described above enables hierarchically clustered gene expression data to be incorporated and integrated into the YETI system. Hierarchical clustering is a simple but effective technique for clustering gene expression data into related groups (or clusters); for example, it enables one to easily examine if a pair of genes of interest are located in the same expression cluster of the hierarchical tree.

However, hierarchical clustering lacks a true quantitative measurement of how similar two gene expression profiles are to each other. A good quantitative measurement of the similarity of two gene expression profiles is needed for correlation analyses comparing the expression profiles of, for example, neighbouring genes or interacting proteins. Therefore, an additional Java program called YETI_Pearson was written to calculate the Pearson correlation coefficient between all genes with expression profiles in the Gasch *et al.* (2000) data set. The Pearson correlation coefficient (http://mathworld.wolfram.com/CorrelationCoefficient.html; Figure 3.1) is perhaps the most widely used measure of the similarity between two expression profiles in gene expression analyses. The Pearson correlation coefficient (R) lies between –1 and 1 (inclusive) with 1 meaning that the two profiles are identical, 0 meaning they are completely independent, and -1 meaning they are perfect opposites.

$$R = \frac{N\sum xy - \sum x \sum y}{\sqrt{(N\sum x^2 - (\sum x)^2)(N\sum y^2 - (\sum y)^2)}}$$

Figure 3.1: The Pearson correlation coefficient equation
This figure shows the Pearson correlation coefficient equation used to calculate the similarity between the two gene expression profiles *x* and *y* with N data points.

3.4: Proteome Data

Many protein-protein interaction data sets are available as a simple list of interactions where each interaction is represented by the names of the two interacting proteins; for example, 'LSM1-LSM2' represents an interaction between the LSM1

protein and the LSM2 protein. A Java program called YETI_PPINTS was written to process this type of protein-protein interaction data set and integrate it into the YETI database. The YETI_PPINTS program assigns the relevant ID numbers to the interactions and the proteins themselves, checks protein names and adds a source field. It also checks for and merges any duplicate entries, counts the total number of unique interactions each protein is involved in and also identifies interactions that consist of a protein interacting with itself. Identifying self-interacting proteins is important because these interactions can bias correlation analyses investigating trends in the function, location and expression of interacting proteins. Identifying and merging duplicate protein-protein interactions serves an additional purpose as protein-protein interactions that are reported in multiple data sets are more likely to be real protein-protein interactions. Keeping track of the source of each protein-protein interaction enables users to judge for themselves if they trust the source. After the processing is complete, the program automatically updates the YETI database.

One of the largest protein-protein interaction data sets available for *S. cerevisiae* can be downloaded from the GRID database. GRID is a database of genetic and physical interactions covering many organisms including *S. cerevisiae*, it contains a large number of protein-protein interactions from a variety of sources including Mewes et al. (1998), Uetz *et al.* (2000), Bader *et al.* (2001), Ito *et al.* (2001), Gavin *et al.* (2002) and Ho *et al.* (2002). The GRID database currently contains ~ 20,000 *S. cerevisiae* protein-protein interactions; these interactions were processed by the YETI_PPINTS program and are currently used to populate the YETI database.

The YETI_PPINTS program described above essentially ensures that the YETI database contains a set of unique protein-protein interactions and also highlights the interactions that were reported multiple times as these interactions have a higher confidence of being real. However, there are also a number of other confidence measures that can be applied to protein-protein interactions. Therefore, an additional Java program called YETI_PPCON was written to apply confidence scores to all the protein-protein interactions stored in the YETI database. The YETI_PPCON program first checks whether interacting proteins are located in the same cellular component as defined by their GO annotations and additionally their GO Slim annotations; this is because two proteins can not physically interact with each other if they are not located in the same cellular compartment. The program then checks whether interacting proteins are coexpressed, as defined by the Pearson correlation coefficient of the their corresponding genes; this is because two proteins can not physically interact with each other if they are not present in the cell at the same time. The program also identifies interactions involving dubious ORFs and also highlights interactions involving uncharacterised ORFs; this is because dubious ORFs and therefore the interactions involving its encoded protein product are highly unlikely to be real. Additionally, the program also identifies the protein-protein interactions where both interacting partners share the same GO Molecular Function, Biological Process or Cellular Compartment annotation and also identifies the interactions where both the interacting protein's corresponding genes are located on the same chromosome; these checks facilitate additional analyses comparing the properties of interacting proteins.

3.5: YETI Database

Essentially, the YETI database has two main functions: (1) To store and manage all of the data outputted from the YETI data processing programs described above; and (2) To communicate with the YETI Java program by receiving and running search queries and passing back the search results. The YETI database was designed with the YETI Java program specifically in mind and the architecture of the database reflects the architecture of the program. The database itself is a relational database consisting of a number of data tables linked together through key fields (Figure 3.2); a brief description of each database table and the data it contains is presented in Table 3.4. The ORF_DATA table is the core table of the YETI database as it contains a wide variety of information on all the features in the *S. cerevisiae* genome and therefore also defines the number of current features in the genome. Each genomic feature in the ORF_DATA table is assigned a unique YETIID number and it is this number that is the main way of linking the database tables (and therefore the data within them) together. Briefly, each ORF in the ORF_DATA table can be involved in multiple protein-protein interactions in the HYBRID table, can have multiple GO annotations in the ONT_DATA table, can have multiple GO Slim annotations in the GO_SLIM table, can have an expression profile in _DATA table of each microarray study and can be involved in multiple Pearson correlation coefficients in the PEARSON table.

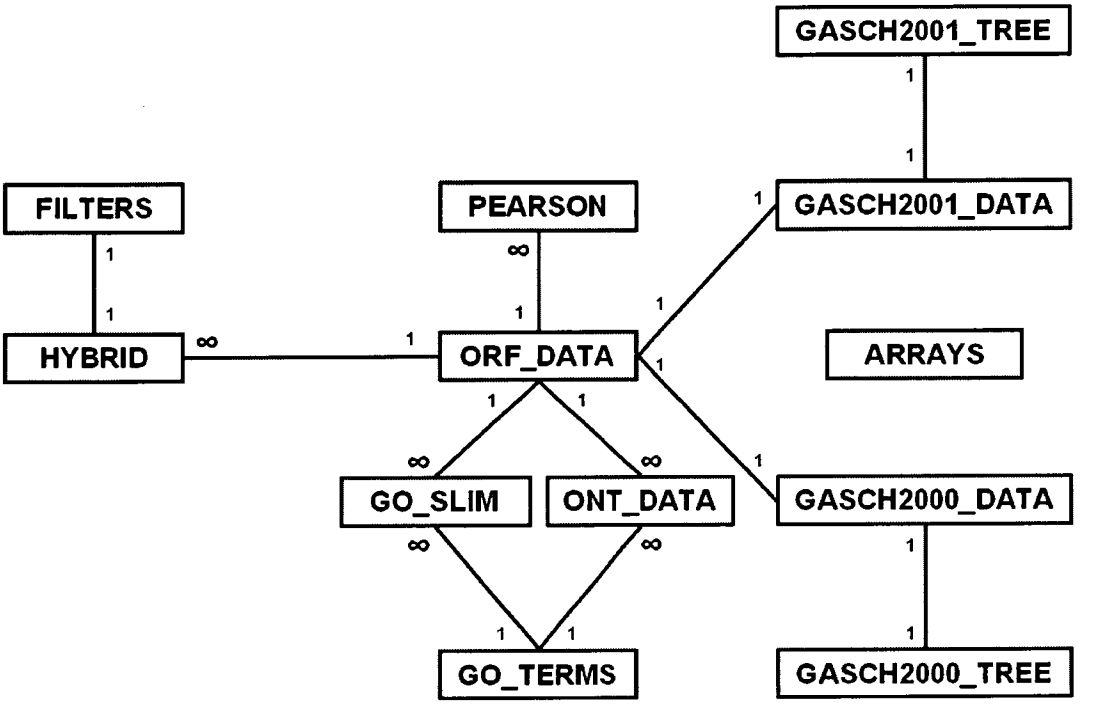


Figure 3.2: Schematic of the YETI database
This is a schematic of the YETI database showing the names of all the database tables as well as the relationships between tables; brief descriptions of each table and the data it contains can be found in Table 3.4.

Table Name	Description
ORF_DATA	This is the core table of the YETI database as it contains a wide range of information on all the features in the *S. cerevisiae* genome; this information includes the name and location of each genomic feature as well as textual descriptions and phenotypic data.
ONT_DATA	This table contains all the GO annotations of all the features in the *S. cerevisiae* genome.
GO_SLIM	This table contains all the GO Slim annotations of all the features in the *S. cerevisiae* genome.
GO_TERMS	This table contains detailed definitions of all the GO annotations used to characterise all the features in the *S. cerevisiae* genome.
HYBRID	This table contains all of the protein-protein interactions.
FILTERS	This table contains a number of different confidence scores for all of the protein-protein interactions stored in the HYBRID table.
ARRAYS	This table contains information on all the gene expression microarray data sets currently stored in the database.
GASCH2000_DATA	This table contains the hierarchically clustered gene expression ratio data from the Gasch *et al.* (2000) study.
GASCH2000_TREE	This table contains the node joining history and information for drawing the hierarchical tree for the Gasch *et al.* (2000) study.
GASCH2001_DATA	This table contains the hierarchically clustered gene expression ratio data from the Gasch *et al.* (2001) study.
GASCH2001_TREE	This table contains the node joining history and information for drawing the hierarchical tree for the Gasch *et al.* (2001) study.
PEARSON	This table contains all of the Pearson correlation coefficients between all of the ORFs with expression data in the Gasch *et al.* (2000) study.

Table 3.4: YETI database tables
This table contains the names of the tables in the YETI database and general descriptions of the data they contain.

The YETI database stores and manages all of the data that is generated from the YETI data processing programs in the format required by the YETI program for visualisation and analysis. The database was developed in tandem with the YETI program and as new sections and features were added to the program, new data and tables were incorporated in the database. The database does have some data

duplication but this duplication allows the YETI Java program to perform at faster speeds whilst having relatively little effect on the performance of the database. Furthermore, this duplication enables specific sections of the YETI program to be detached and used as standalone applications with their corresponding sections of the YETI database.

The YETI database is primarily available in MySQL format which is an open source database management system that is fast, compact, stable and is available for most of the major computer platforms. One disadvantage of MySQL is that it is fairly complicated to install from the point of view of the standard wet laboratory scientist. However, users can avoid downloading and installing the database by either using Web YETI or by connecting to the YETI database housed at the University of Edinburgh from Standalone YETI. In addition, the YETI database has also been ported across to Microsoft Access format which is much simpler to install; however, this version of the database can only be used on the Windows platform.

The YETI database has always been relatively small in size but steadily increasing as more protein-protein interactions and microarray data sets were added; this steady increase in the size of the database has not significantly affected the speed or performance of the YETI program. However, the addition of the Pearson correlation coefficients between all the ORFs with expression data in the Gasch *et al.* (2000) study increased the size of the database from ~30 MB to ~800 MB. This is because there are ~18,000,000 unique Pearson correlation coefficients stored along with the two YETIID numbers of the ORFs that each coefficient corresponds to as well as

data indexes to enable efficient database searching. This does affect the usability of YETI from the point of view of disk space and longer download and installation times. However, disk space is not the major problem it once was as modern computers currently come with extremely large hard disks and download time can be improved by compression and high speed networks. Furthermore, YETI can be used over the internet avoiding the need to download and install the database locally; however, this option will also need a high speed network to be effective as large amounts of data often need to be transferred between the YETI database and program.

3.6: Discussion

The core data in the YETI database consists of the name, location and function of every feature in the *S. cerevisiae* genome; the source of this data is the SGD which is a well respected resource of the yeast community. The SGD recently began defining each ORF as either verified, uncharacterised or dubious which can essentially be used as a confidence measure as to the validity of each ORF with verified ORFs having very high confidence, dubious ORFs having very low confidence, and uncharacterised ORFs having medium confidence of being real genes. The source of the gene expression data in the YETI database are the Gasch *et al.* (2000) and Gasch *et al.* (2001) studies which are well respected and often used as a test data set for gene expression analysis programs; these expression data sets have already been normalised and hierarchically clustered by the Cluster program. Furthermore, the Gasch *et al.* (2000) is also used to directly calculate the Pearson correlation

coefficient between all gene expression profiles. The source of the protein-protein interaction data in the YETI database is the GRID database which is a large and widely used resource. However, this protein-protein interaction data contains many interactions detected from techniques such as the yeast two-hybrid system which can be error-prone. Therefore, the protein-protein interactions in this data set are thoroughly evaluated with a number of confidence scores assigned to each interaction.

The YETI database effectively stores and links all of the data described above in a relational way. The YETI database was specifically designed for use by the YETI Java program as both were developed in tandem. The features and functions of the YETI program itself and now discussed in further detail in the next chapter of this thesis.

Chapter 4

YETI Program

4.1: Introduction

The Yeast Exploration Tool Integrator (YETI) is a novel bioinformatics tool for the integrated visualisation and analysis of *Saccharomyces cerevisiae* (*S. cerevisiae*) functional genomic data sets. Essentially, YETI consists of two parts: (1) A database for the storage and management of data; and (2) A Java program for the integrated visualisation and analysis of data. The YETI Java program itself consists of a MainFrame and a number of core inter-linked sections. The YETI MainFrame is concerned with establishing a connection with the YETI database, handling database searches as well as launching and monitoring all the other YETI sections. All the core YETI sections are closely inter-linked enabling users to swiftly move between them and investigate all aspects of any genes or proteins of interest as well as providing access to textual information, including Gene Ontology (GO) annotations, at any point. Furthermore, there are also a number of additional YETI correlation sections that enable users to investigate possible correlations between the stored functional genomic data sets. An overview of the structure of the YETI program is presented in Figure 4.1 and an overview of the main function of each of the YETI sections is presented in Table 4.1.

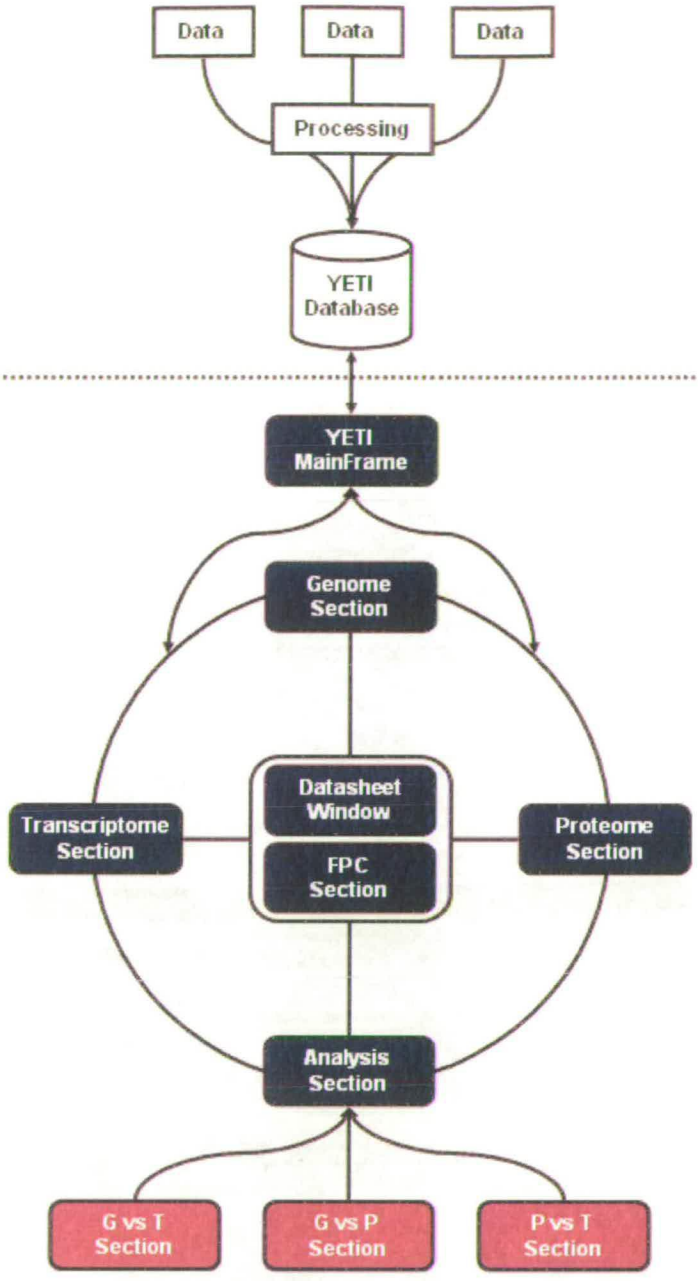


Figure 4.1: Schematic of the YETI program
This is a schematic of the overall structure of YETI. Essentially, YETI consists of two parts: (1) A database for the storage and management of data; and (2) A Java program for the integrated visualisation and analysis of data. The YETI Java program consists of a MainFrame which communicates with the database and a number of closely inter-linked core sections where data can be visualised and analysed. Sections highlighted in blue are the core sections of the YETI program whereas sections highlighted in red are additional correlation analysis sections. An overview of the main function of each section is presented in Table 4.1.

Name	Description
Genome Section	The Genome Section is concerned with the informative display of the *S. cerevisiae* genome, its chromosomes, and known and predicted genes.
Transcriptome Section	The Transcriptome Section is concerned with the visualisation and integration of gene expression data from microarray experiments.
Proteome Section	The Proteome Section is concerned with the effective visualisation of protein-protein interactions on a dynamic graphical display panel.
Analysis Section	The Analysis Section is concerned with providing a graphical interface to the YETI database and has a number of easy to use search mechanisms for both simple and complex queries.
FPC Section	The Function, Process and Component (FPC) Section is concerned with enabling users to browse GO annotations and define specific groups of genes which can then be investigated in further detail in the other YETI sections.
Datasheet Window	The Datasheet Window is concerned with displaying a wide range of information on a specific gene of interest and contains a number of direct links to the YETI Sections.
G vs T Section	The Genome vs Transcriptome (G vs T) Section is concerned with enabling users to find and investigate chromosomal regions of coexpression.
G vs P Section	The Genome vs Proteome (G vs P) Section is concerned with enabling users to find and investigate chromosomal regions containing genes whose corresponding proteins interact with one-another.
P vs T Section	The Proteome vs Transcriptome (P vs T) Section is concerned with enabling users to find and investigate interacting proteins whose corresponding genes are coexpressed.

Table 4.1: Overview of the main functions of the YETI Sections
This table contains the names and descriptions of the sections of the YETI Java program.

YETI is written in the Java programming language which is a state-of-the-art, object-oriented language with a syntax similar to the C++ programming language. Furthermore, the Java programming language is platform independent with Java Virtual Machines (JVM) available for all of the major operating systems. Therefore, this means that YETI can work on any operating system that has access to a JVM of

version 1.4 or above making it highly portable. YETI has been thoroughly tested and performs very well on the Windows operating system and also performs well on the Linux and Mac OS-X operating systems. YETI can be used online via a simple Java applet (Web YETI) or can be downloaded and installed locally onto the users own computer (Standalone YETI).

4.2: Analysis Section

Current computational resources tend to utilise a single gene approach where users simply view a datasheet on a single gene of interest at once. Although this approach is an essential feature, it does not enable users to collectively view and compare the data on a number of genes to investigate possible shared functionality, for example. In addition, current computational resources tend to have limited search capabilities where the main and sometimes only way of searching for data is by entering a single gene name. In contrast, the Analysis Section of YETI provides a sophisticated graphical interface to the YETI database with a number of different search functions to find data and an interactive data table to collectively visualise and analyse all the search results together.

At the heart of the Analysis Section is the interactive data table which displays all the results from a database search (Figure 4.2). This table can collectively display a wide range of data on a large number of genes at once enabling easy visual examination and comparison of all the properties of all the genes; each row in the table corresponds to a distinct gene and contains a wide range of information on that gene.

By collectively displaying all the results of a search together in one table users are easily able to scroll along the table to view all the data on an individual gene of interest. But more importantly, users can also scroll up and down the table to compare the properties of all the genes found in the search; this feature is extremely useful when investigating possible shared functionality among a group of genes. Furthermore, any of the genes displayed in the data table can be individually selected and investigated in further detail individually in YETI.

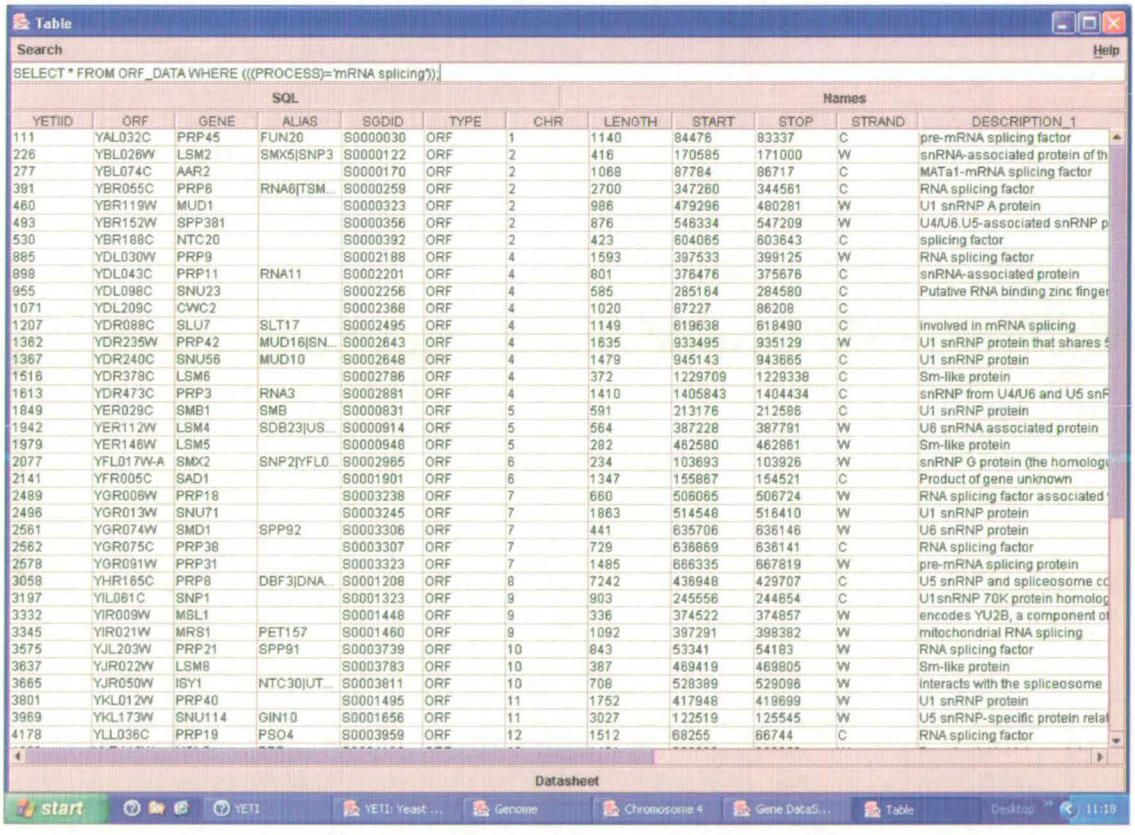


Figure 4.2: Screenshot of the Analysis Section
This is a screenshot of the Analysis Section which displays all the results of a database search collectively in an interactive data table. Each row of the table corresponds to a specific gene and contains a large amount of data relating to that gene. Search queries can be entered into the textfield at the top of the window or can be generated by using the QueryBuilder function (see below).

The Analysis Section has a flexible and powerful QueryBuilder function that enables users to search on any aspect of the available data and perform both simple and complex database searches. The QueryBuilder function enables users to easily construct large and complex queries to search the database with by simply entering their desired search criteria into a variety of labelled textfields. One of the main reasons for constructing the QueryBuilder function was to enable users to perform keyword searches on gene descriptions and GO annotations to rapidly find and then collectively examine related genes. In this case, the user simply needs to enter their desired keyword(s) into the appropriate textfield(s) and YETI will automatically search the database and collectively display the search results in the interactive data table. For example, searching for all the genes with the keyword 'spliceosome' in their description or GO annotations would return all the spliceosome and spliceosome related genes for examination and further investigation. However, the QueryBuilder function can be used to perform a wide variety of other searches both simple and complex; for example, users can search for all the genes with the text 'LSM' in their name, all the genes that contain introns, all the genes with an inviable phenotype, all the genes located on chromosome 1 or any combinations of the above. Alternatively, users can simply enter the names of multiple genes of interest and YETI will collectively display information on all of them in the data table. Furthermore, an advanced search option is also provided that enables users to directly enter a Structured Query Language (SQL) statement which YETI then uses to search the database with and collectively displays any results found in the data table; however, although this powerful option gives the user complete control over the search criteria and display settings it obviously requires knowledge of both SQL

and the YETI database structure.

The Analysis Section is effectively inter-linked with the other YETI sections enabling users to swiftly move directly into another YETI section where information related to all the genes currently displayed in the data table will be automatically displayed and highlighted. Alternatively, users can move swiftly into the Analysis Section from the other YETI sections where a range of information on all their selected genes would be automatically displayed in the data table.

4.3: Genome Section

Current computational resources such as the *Saccharomyces* Genome Database (SGD; Cherry *et al.*, 1998) tend to only have a basic graphical representation of the chromosomal area surrounding a specific gene of interest; there are few resources available that enable users to view the location of genes from a genomic perspective or enable users to easily and rapidly scroll along detailed visual representations of the chromosomes. In contrast, the Genome Section of YETI is concerned with the informative display of the *S. cerevisiae* genome, its chromosomes, and known and predicted genes. The Genome Section enables users to examine and compare the genomic location of multiple genes or multiple groups of genes on a schematic of the entire *S. cerevisiae* genome. It also enables users to scroll along detailed visual representations of the chromosomes themselves and select regions of interest to investigate in further detail in the other YETI sections.

At the heart of the Genome Section is the genome schematic which is a scaled graphical representation of the 16 nuclear chromosomes of *S. cerevisiae* (Figure 4.3). This schematic firstly provides a visual overview of the genome which enables users to make quick comparisons of chromosome sizes and centromere positions. The Genome Section does not currently take account of the mitochondrial chromosome as it fits into a different model; it is circular whereas the nuclear chromosomes are linear. In addition, there is generally less of a scientific interest in the mitochondrial genes as they do not tend to be investigated in functional genomic analyses. However, data on all the mitochondrial genes can still be accessed and examined through the Analysis Section.

The genomic location of any genes of interest can be collectively displayed on the genome schematic by simply entering their names and YETI will then highlight their location on the genome schematic; genes are highlighted with a red line at their corresponding start position on the scaled representation of their chromosome. This feature provides a quick and simple means to examine and compare the genomic location of multiple genes; a group of genes of interest could well be located near each other on a particular chromosome or be located on different chromosomes but at similar positions such as the centromere. Alternatively, as the YETI sections are closely inter-linked, the Analysis Section could be used to search for a specific group of genes to highlight on the genome schematic. For example, in Figure 4.3 YETI was used to search for all genes with an inviable phenotype and to subsequently highlight their location on the genome schematic enabling all their genomic locations to be collectively examined and compared; as can clearly be seen, there a very few

inviable and therefore essential genes located in the telomeric regions of the 16 nuclear chromosomes. This feature also enables the investigation of possible functional hotspots in the *S. cerevisiae* genome, for example, examining if genes characterised to the same or related GO biological process annotations are located in the same genomic region.

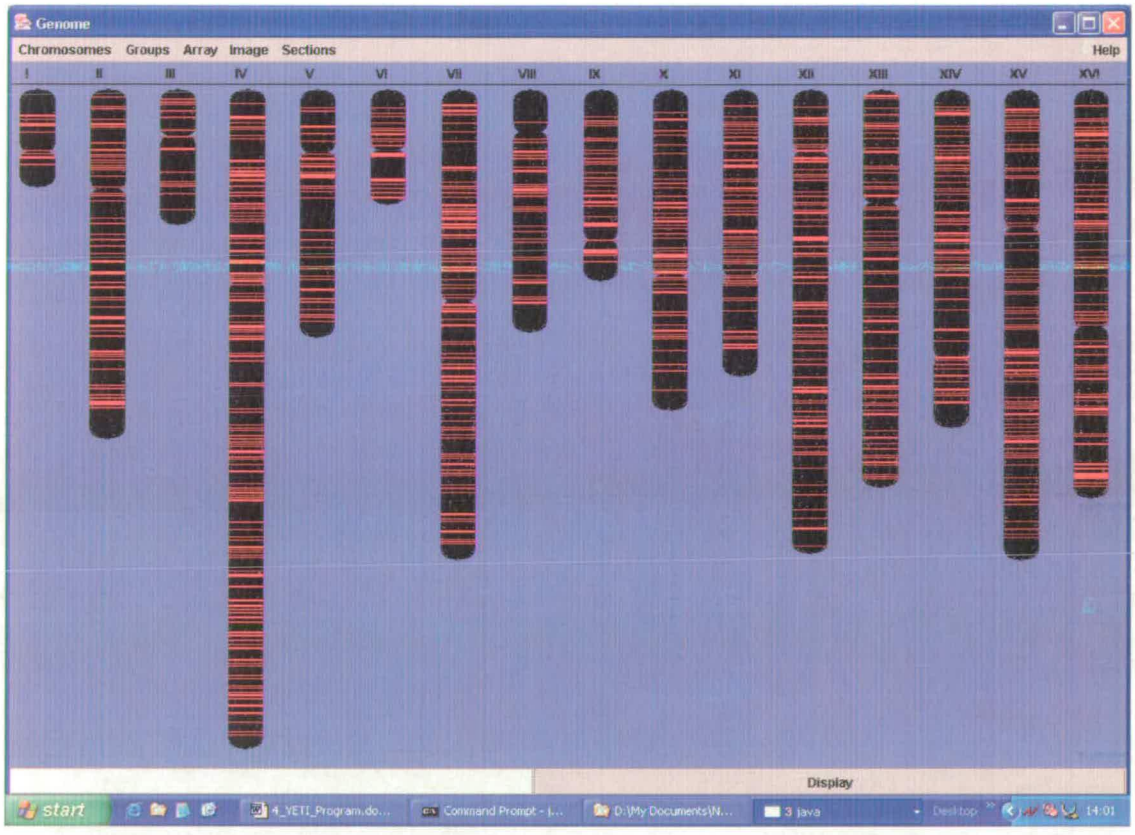


Figure 4.3: Screenshot of the Genome Section
This is a screenshot of the Genome Section displaying the genome schematic. The genome schematic displays a scaled representation of all 16 nuclear chromosomes of *S. cerevisiae*; chromosome 4 is the longest at 1,532,000 base pairs (bp) and chromosome 1 is the smallest at 230,000 bp. In this case, the genomic location of all genes with an inviable phenotype have been highlighted on the genome schematic with red lines enabling the genomic location of the entire group to be collectively examined. As can be seen, very few inviable and therefore essential genes are located in the telomeric regions of the chromosomes. Furthermore, numerous high density red regions are observed which consist of a number of inviable genes located next to each on the chromosome; these chromosomal regions could be easily selected and investigated in further detail in YETI.

The genomic location of multiple groups of genes can also be examined and compared on the genome schematic. In YETI, two different groups of genes can be defined, the so called Red and Green groups which are highlighted on the genome schematic with red and green lines, respectively. A broad example is shown in Figure 4.4 where the genomic location of all genes whose protein products are located in the cytoplasm and nucleus are highlighted with red and green lines, respectively; this example shows that even the genomic location of very large groups of genes can still be collectively examined and compared on the genome schematic. This feature is useful for investigating possible evolutionary relationships between two groups of genes through the collective comparison of their genomic locations; for example, members of the two groups could be colocated across the genome.

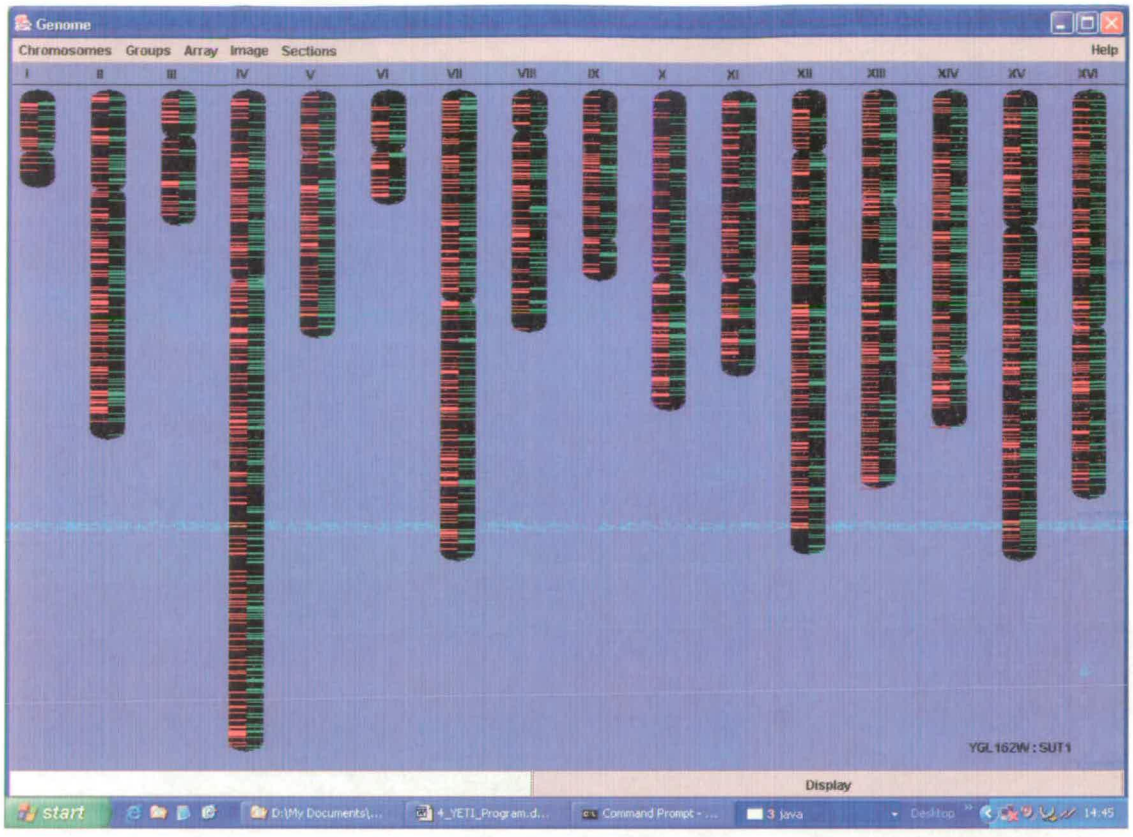


Figure 4.4: Screenshot of the Genome Section with multiple groups highlighted
This is a screenshot of the Genome Section with the location of multiple groups of genes highlighted on the genome schematic. The Red group consists of all the gene's whose protein products are located in the cytoplasm and are highlighted with red lines on the genome schematic. The Green group consists of all the gene's whose protein products are located in the nucleus and are highlighted with green lines on the genome schematic. As can be seen, even the genomic locations of very large groups can still be collectively compared and examined on the genome schematic with ease.

A unique feature of the Genome Section is the ability to overlay gene expression data onto the genome schematic to display the relative expression of every gene in the genome; every gene is highlighted on the genome schematic with a line that is coloured to reflect its relative gene expression ratio value from the selected microarray experiment. This fairly unique feature essentially enables users to view the gene expression profile of the entire *S. cerevisiae* genome; it enables users to examine the expression state of particular areas of the genome and find areas that have similar relative changes in gene expression.

4.3.1: Chromosome Window

The Chromosome Window of YETI displays a detailed visual representation of one of the 16 nuclear chromosome of *S. cerevisiae* (Figure 4.5). The chromosome is visually represented by two scaled black bars which correspond to the two strands of chromosomal DNA (Watson strand at the top and Crick strand at the bottom). Genes are represented by white rectangles within the chromosomal strands extending from their corresponding start to stop positions along with the name of the gene. The chromosome is contained within a scrollpane that enables users to easily scroll along the chromosome to view the location and distribution of genes across the whole chromosome and rapidly find areas of interest. As some genes can be located quite close together making them hard to distinguish from one another, there is a zoom function to magnify the chromosomal display and clarify the situation. By default only verified and uncharacterised ORFs are displayed on the chromosome; however, dubious ORFs and any other genomic feature types can also be selected and subsequently displayed. In addition, there is also a simple Find function that can be used to search for and subsequently highlight the location of a specific gene of interest on the chromosome.

A single datasheet on any of the genes displayed on the chromosome can be viewed simply by mouse clicking on them. The datasheet contains a wide range of information on the selected gene and has a number of direct links to the other YETI sections; more information on the features of the YETI Datasheet Window can be found below in section 4.6 of this chapter. Furthermore, entire regions of the

chromosome can be selected simply by dragging the mouse to create a selection box (Figure 4.5); all the genes located within the selection box are automatically selected and highlighted in red and can then be collectively investigated in further detail in the other sections of YETI. The Chromosome Window is effectively linked to the other YETI sections enabling users to swiftly move directly into another YETI section where information related to all the genes currently selected on the chromosome will be automatically displayed and highlighted. For example, the Transcriptome Section would automatically display and highlight the gene expression profiles of all the genes currently selected on the chromosome; this enables users to investigate if all the genes located in a particular chromosomal region, such as a telomeric region or the region surrounding a gene of interest, are coexpressed with one another.

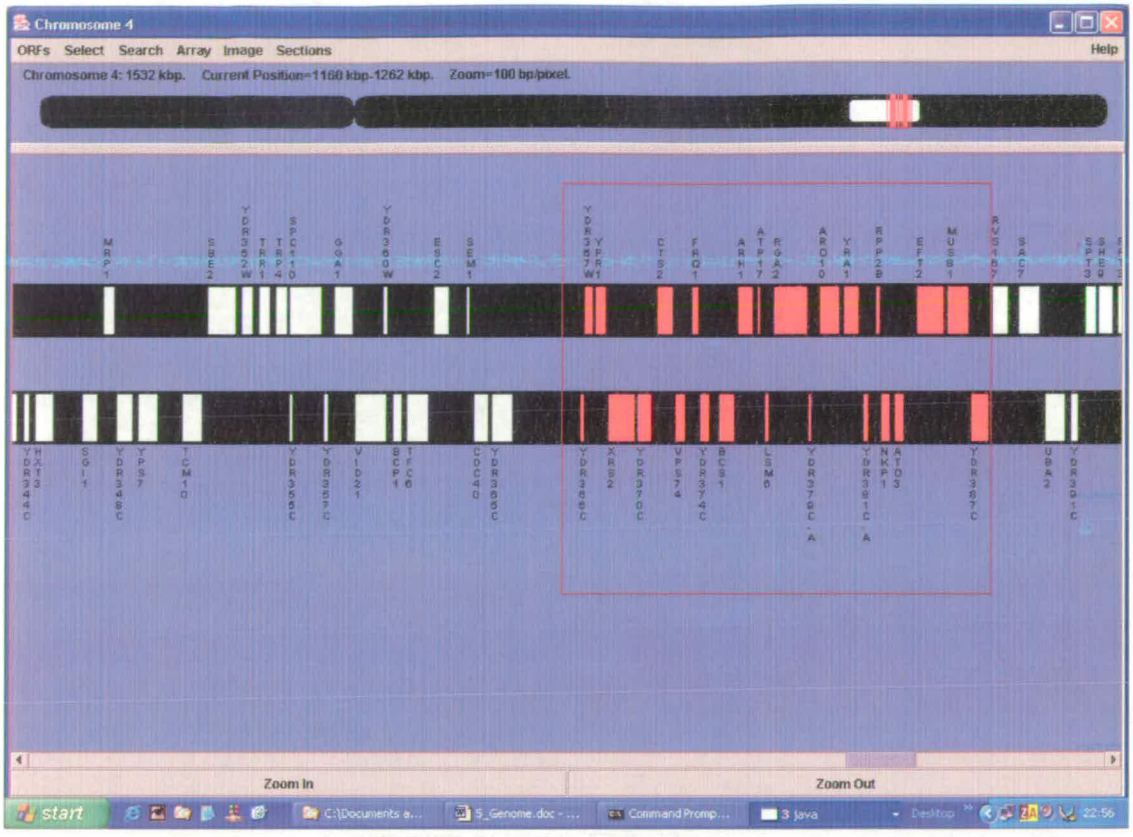


Figure 4.5: Screenshot of the Chromosome Window
This is a screenshot of the Chromosome Window with chromosome 4 displayed and the chromosomal region around the LSM6 gene selected. The top panel (above the horizontal grey bar) displays a graphical overview of the whole chromosome with the area currently being viewed in the bottom panel represented by the white box. The bottom panel displays a detailed graphical representation of the selected chromosome in a scrollpane. Selected genes are highlighted in red in the bottom panel and their location is also highlighted with red lines in the top panel.

One of the unique features of the Chromosome Window is that gene expression data from any of the microarray experiments stored in the YETI database can be overlaid onto the chromosome to display the relative expression of every gene on the chromosome; gene expression data is overlaid onto the chromosome by colouring each gene with a colour that reflects its relative expression ratio value from the selected microarray experiment. This essentially enables users to view the gene expression profile of the entire chromosome and enables them to easily find chromosomal regions with similar relative changes in expression. Furthermore, users

are able to save an image of the entire chromosome that is currently displayed complete with any gene selections or overlaid expression data (Figure 4.6). This feature is useful because it creates a detailed image of the entire chromosome that allows easy visual examination of gene locations and distributions as well as the rapid examination of the relative expression of every gene on the chromosome.

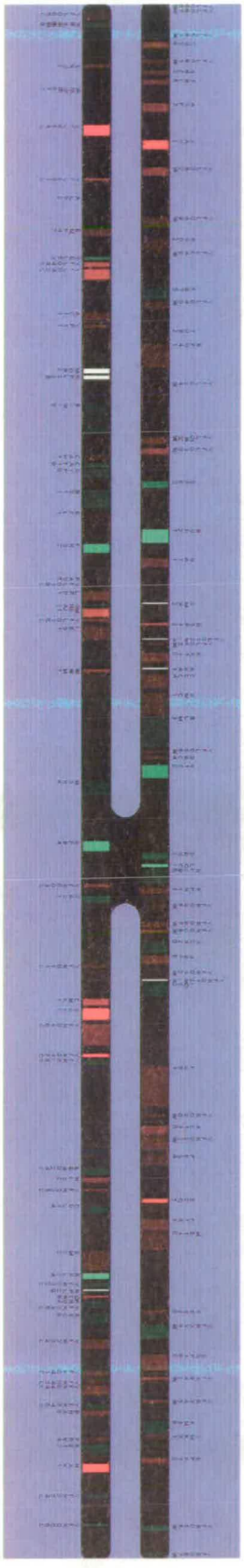


Figure 4.6: YETI generated image of chromosome 6
This is an image, created by YETI, of the whole of chromosome 6 with a microarray experiment overlaid. The image created is a large, detailed view of the entire chromosome that enables users to view gene locations and distributions along the entire length of the chromosome. In this case, a gene expression microarray experiment has been overlaid onto the chromosome which colours all the gene boxes corresponding to their relative gene expression ratio values from the selected experiment. This image allows the user to easily and rapidly examine the expression of all the genes on the chromosome and find regions of interest.

4.4: Transcriptome Section

The Transcriptome Section of YETI provides an effective means for the visualisation and analysis of gene expression data generated from microarray experiments. The YETI database contains processed gene expression data sets that have already been hierarchically clustered using the Cluster computer program (Eisen *et al.*, 1998). These hierarchically clustered gene expression data sets can be loaded into the Transcriptome Section for visualisation and analysis as well as integration with the other YETI sections and their corresponding functional genomic data sets. This highlights one of the advantages of YETI as there are few computational resources available that can effectively integrate gene expression data with other functional genomic data sets for visualisation and analysis.

At the heart of the Transcriptome Section is the graphical panel which can display any one of the hierarchically clustered gene expression microarray data sets stored in the YETI database (Figure 4.7). The graphical panel displays the gene expression data set visually by representing each relative gene expression ratio data point with a colour that reflects its value; values greater than zero are coloured with progressively brighter shades of red and values less than zero are coloured with progressively brighter shades of green. Therefore, each gene's expression profile in the data set is represented on the graphical panel by a row of data points that are all individually coloured to reflect their value. The gene rows are ordered with respect to the data set's hierarchical tree which is also displayed on the graphical panel so that the relationship between genes can be easily examined. Furthermore, the graphical panel

is contained within a scrollpane which enables users to easily scroll up and down to examine the entire hierarchically clustered gene expression data set and rapidly find regions of interest such as a particular cluster. Any regions of interest from the displayed data set can be selected simply by dragging the mouse to create a white selection box; all the genes contained within the selection box are then automatically selected and highlighted with red lines to their left (Figure 4.7). Furthermore, multiple regions of interest can simply be selected by creating multiple selection boxes.

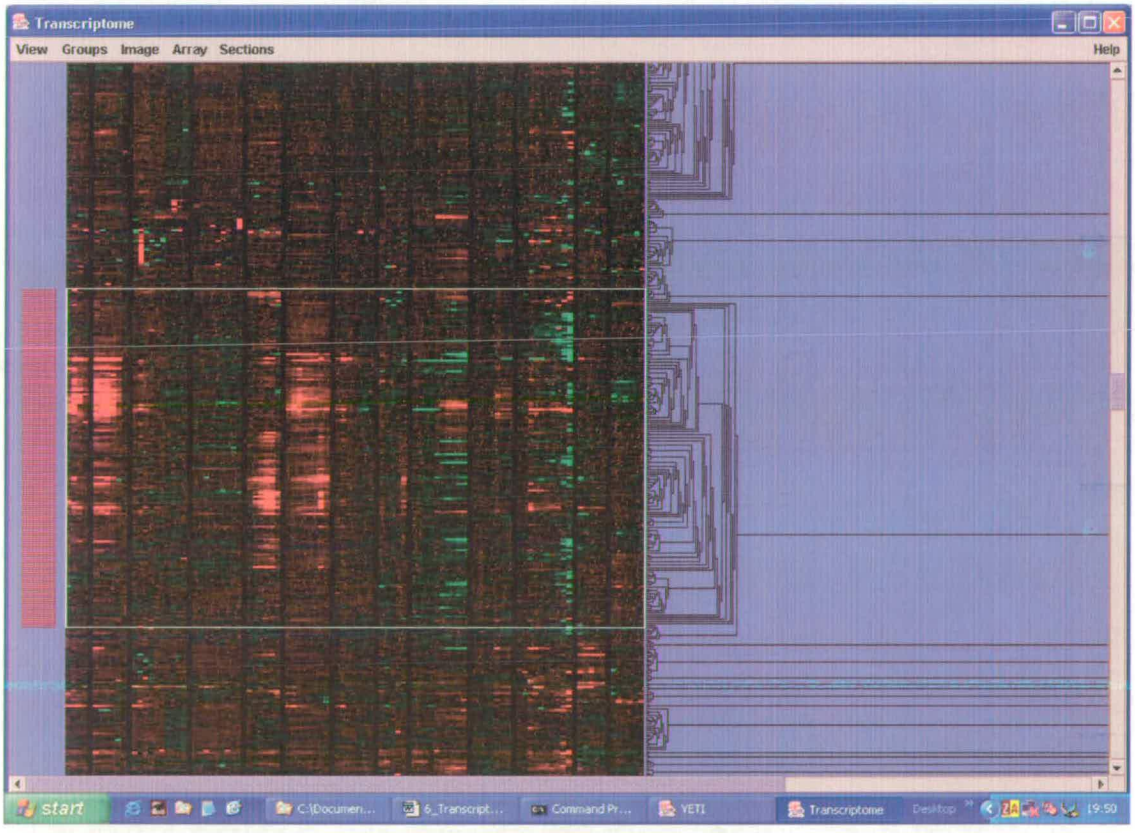


Figure 4.7: Screenshot of the Transcriptome Section
This is a screenshot of the Transcriptome Section with a region of interest selected from the hierarchically clustered gene expression data set. The graphical panel is shown in light blue and displays a visual representation of the clustered gene expression data set with the corresponding hierarchical tree. Each row in the data set corresponds to the gene expression profile of a particular gene. In this case, a region has been selected by dragging the mouse vertically to create a white selection box. All genes contained within the selection box are then automatically selected and highlighted with red lines to their left.

Once a region or regions of the gene expression data set have been selected they can be examined in more detail in the Transcriptome Section itself. The Data option of the Transcriptome Section can be used to display an expanded view of the selected regions of the data set along with the name and description of all the genes within these regions (Figure 4.8). This option enables users to rapidly examine what each selected gene is as well as giving a much clearer view of all the selected gene's expression profiles allowing easy visual examination and comparison of possible shared properties between the selected genes.

Figure 4.8: Screenshot of the Transcriptome Section with the expanded data view
This is a screenshot of the Transcriptome Section displaying an expanded view of the region of the data set selected in Figure 4.7. Each individual gene expression data point is increased in size giving an expanded view and the names and descriptions of every selected gene is displayed to its right.

The Transcriptome Section is effectively inter-linked with the other YETI sections enabling users to swiftly move directly into another YETI section where information related to all the genes currently selected from the gene expression data set will be automatically displayed and highlighted. For example, the Genome Section would automatically display and highlight the location of all the selected genes on the genome schematic; this enables users to examine if the selected coexpressed genes are also located in the same or similar regions of the genome. Alternatively, users can move swiftly into the Transcriptome Section from the other YETI sections where the expression profiles of all their selected genes would be automatically highlighted in the gene expression data set. For example, the Analysis Section could be used to search for a group of related genes, such as all the genes that contain introns, and the Transcriptome Section would highlight all their expression profiles in the data set enabling users to examine if they are coexpressed. For example, Figure 4.9 shows that a large number of the genes containing introns are coexpressed forming a large cluster of genes in the gene expression data set. This feature enables the investigation of possible functional hotspots in the gene expression data sets by examining if genes characterised with the same or related GO biological process annotations have been clustered with or near each other in the hierarchical tree. Furthermore, the Transcriptome Section can highlight the location of multiple groups of genes in the gene expression data set enabling the properties of the groups as a whole to be examined and compared; in YETI two groups can be defined, the so-called Red and Green groups, which are highlighted in the gene expression data set with red and green lines respectively.

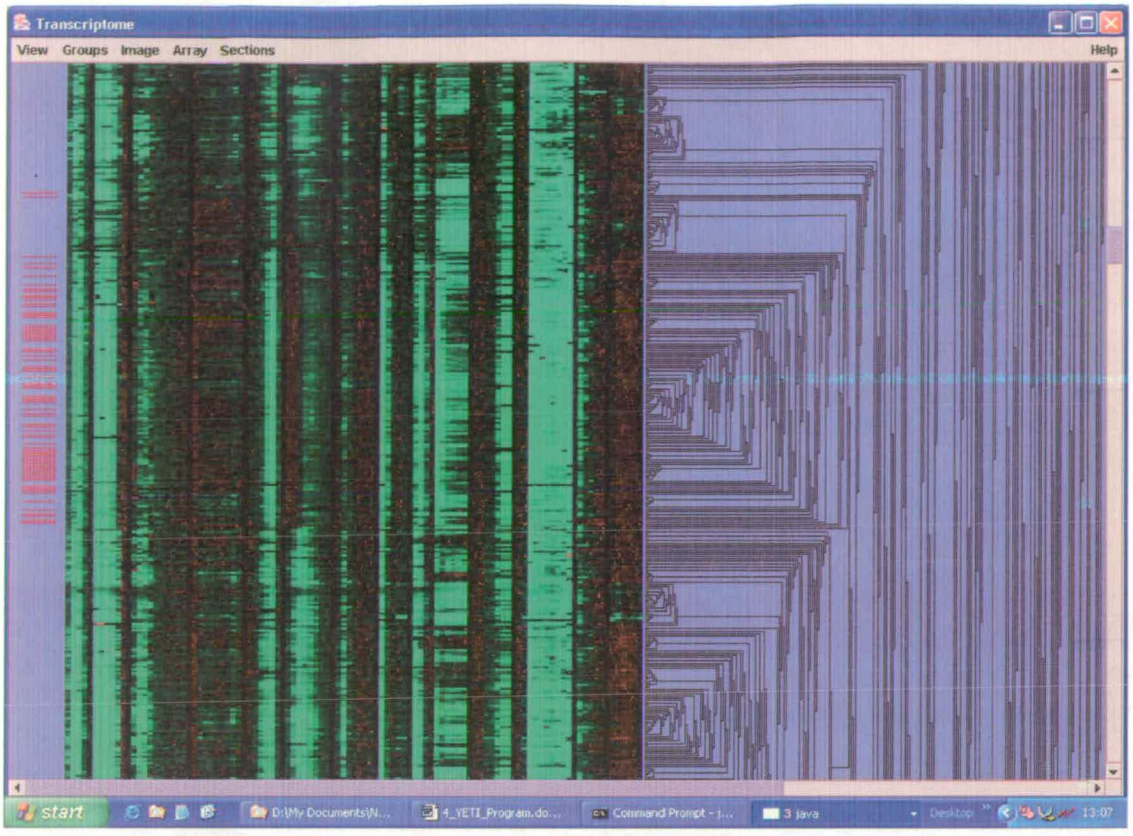


Figure 4.9: Screenshot of the Transcriptome Section highlighting the location of intron containing genes
This is a screenshot of the Transcriptome Section highlighting the expression profiles of all the genes that contain introns. As can be seen, a large cluster of genes containing introns can be observed in the data set.

An additional feature of the Transcriptome Section is the ability to save an image of the entire clustered gene expression microarray data set currently displayed in the graphical panel (Figure 4.10). This option is useful because it enables users to create a detailed image of the entire data set allowing easy visual examination of the overall properties of the entire hierarchically clustered gene expression data.

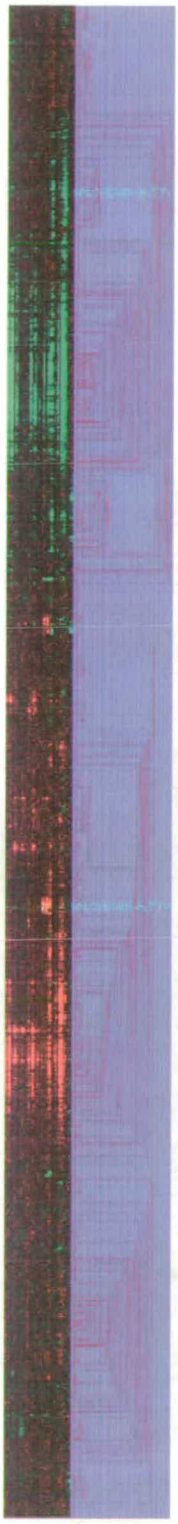


Figure 4.10: YETI generated image of the Transcriptome Section's graphical panel
This is an image of the Transcriptome Section's graphical panel displaying a hierarchically clustered gene expression data set created by YETI. The image created is a large, detailed view of the entire data set that contains both the hierarchical tree and the clustered visual representation of the gene expression data itself. This enables users to easily examine the entire data set and find regions of potential interest.

4.5: Proteome Section

Current proteomic computational resources tend to only be concerned with the visualisation of protein-protein interactions, provide only basic information on the interacting proteins and tend to utilise a single protein approach with limited search functions. In contrast, the Proteome Section of YETI is concerned with the effective visualisation of protein-protein interactions, utilises both a single protein approach and a group approach, has a number of advanced features and flexible search functions, and is fully inter-linked with the other sections of YETI. This again highlights one of the advantages of YETI as there are few computational resources available that can effectively integrate protein interaction data with other functional genomic data sets.

At the heart of this section is the dynamical graphical panel which displays all the relevant protein-protein interactions (Figure 4.11). Proteins are represented on the panel by labelled yellow boxes with a black bond linking two protein boxes together representing a protein-protein interaction between those two proteins. The graphical panel uses a 'springs and rings' type relaxation algorithm to automatically arrange all the displayed proteins in an optimal way; this algorithm is based on the publicly available relaxation algorithm from the Sun Network Mapping Java applet (http://java.sun.com/products/plugin/1.4.1/demos/applets/GraphLayout/example1.html). The relaxation algorithm is conceptually simple and essentially attempts to find space on the panel for all the displayed proteins and their interactions. It treats the bonds linking proteins together as springs which pull the two proteins together when they

are far away and which pushes them apart when they are too close together. Initially, the algorithm assigns random x and y coordinates to all the proteins displayed on the panel and then starts to calculate new positions for the proteins and begins to move them on the panel; this has the affect of bringing proteins that interact with one another closer together and moving proteins that do not interact with one another further apart.

All the protein-protein interactions of a specific protein of interest can be visualised simply by entering the protein's name; YETI then searches the database for all interactions involving the selected protein and displays any interactions found dynamically on the graphical panel (Figure 4.11). This enables users to easily and rapidly examine what proteins a specific protein of interest interacts with. Furthermore, a range of confidence scores for all the displayed interactions can be accessed via the Analysis Section enabling users to judge the relevance of each interaction themselves; these confidence scores show users if an interaction has been reported in multiple studies, whether the two interacting proteins are located in the same cellular compartment, and whether the two interacting protein's corresponding genes are coexpressed.

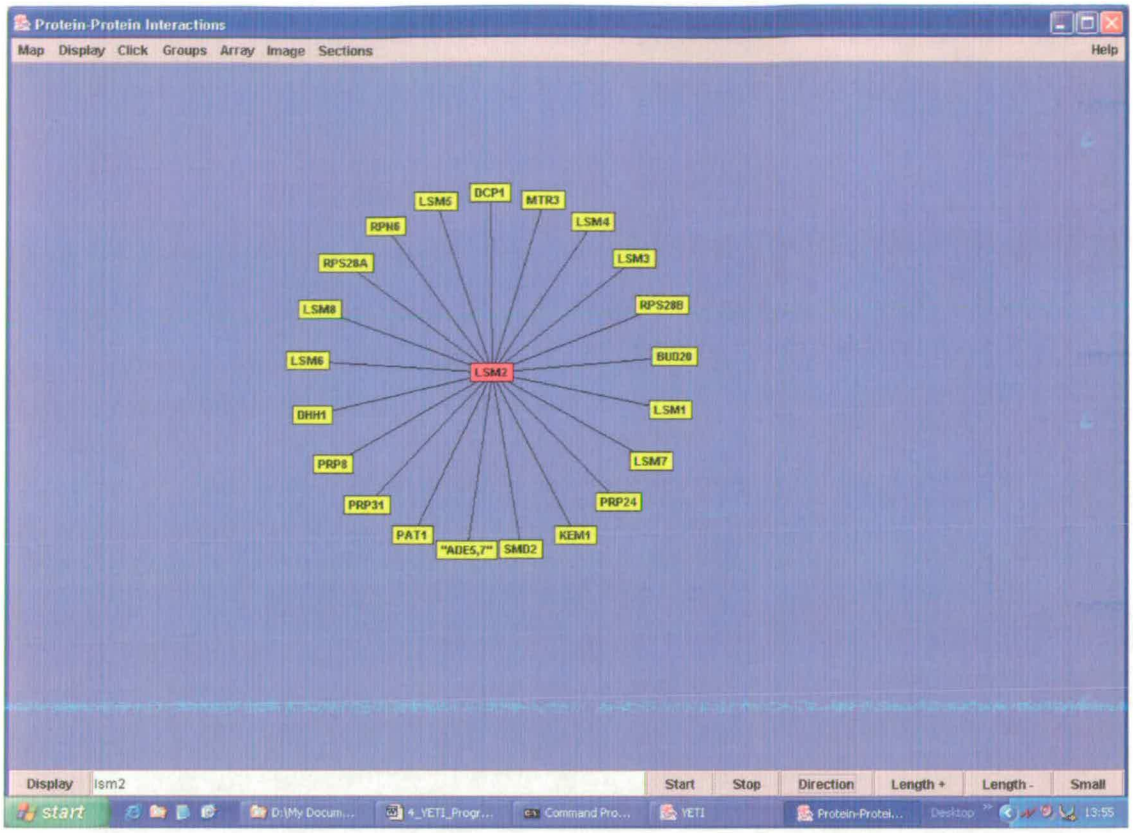


Figure 4.11: Screenshot of the Proteome Section
This is a screenshot of the Proteome Section displaying all of the protein-protein interactions involving the LSM2 protein. After LSM2 was entered into the white textfield YETI searched for all the protein-protein interactions involving LSM2 and automatically displayed all the interactions found on the dynamic graphical panel. LSM2 is highlighted in red and located in the centre of the map whereas all the proteins it interacts with are highlighted in yellow and have been automatically positioned around LSM2 in a circular fashion; the black bonds linking the protein boxes together represent the interactions between proteins.

All the interactions of multiple proteins of interest can be collectively visualised simply by entering all of their names; this enables users to examine if the proteins of interest interact directly with one another, or interact indirectly via common proteins, and to also examine what other proteins they interact with. The Proteome Section also enables paths of interactions between two proteins of interest to be visualised simply be entering their names; a path length of five was selected to be the default and maximum value as paths longer than this were not viewed as significant. This feature enables users to investigate if two proteins that may not interact directly with

each other, interact via one or more intermediate proteins.

Simply displaying all of the proteins a protein of interest interacts with means little if you do not know what all of the interacting proteins are and what their functions are. Therefore, any and all of the proteins currently displayed on the graphical panel can be selected and investigated in further detail individually or as a whole. Proteins can be investigated individually by simply mouse clicking on them and YETI will launch a Datasheet enabling users to examine a wide range of information on the selected protein. An alternative option enables multiple proteins to be selected simply by mouse clicking on them which highlights them in red; furthermore, there is also an option to select entire clusters of interacting proteins simply by mouse clicking one of the proteins in the cluster and all the other proteins in the cluster are also automatically selected. As the Proteome Section is effectively inter-linked with the other YETI sections, after a number of proteins have been selected they can be collectively investigated in further detail by swiftly moving into another YETI section where data related to the selected proteins will be automatically displayed and highlighted. For example, all of the proteins that interact with a specific protein of interest can be selected in the Proteome Section and the Analysis Section would collectively display a wide range of information on all of the selected proteins enabling users to rapidly examine and compare the functions of all the proteins. This feature is especially useful when investigating a protein of unknown function as users can examine if the unknown protein is interacting with a number of proteins of the same function.

Alternatively, users can move swiftly into the Proteome Section from another YETI section where all the protein-protein interactions involving any of their selected proteins will be automatically displayed and highlighted. For example, the Analysis Section could be used to search for a group of related proteins, such as all the proteins that contain the keyword 'cytochrome' in their description, and the Proteome Section would display all of their interactions and also highlight all of the cytochrome proteins automatically (Figure 4.12). This enables users to rapidly examine if and how a group of related proteins are interacting with one another to achieve their biological goals and to also examine what other proteins the group as a whole are interacting with. For example, in the bottom left corner of Figure 4.12 there are four cytochrome proteins (COX9, COX6, COX4 and COX5A) interacting with one another and further investigation reveals that these are all subunits of cytochrome c oxidase; furthermore, in the top left corner of Figure 4.12 there are three cytochrome proteins (COX17, SCO1 and SCO2) interacting with one another and further investigation reveals that these are all involved in the delivery of copper to cytochrome c oxidase. Overall, this feature enables users to collectively examine all the interactions of a functionally related group of proteins; it enables users to examine if a group of proteins are interacting directly with each other perhaps in order to fulfil their biological role or if they are using other key proteins as mediators to link functionally related proteins together. Furthermore, this feature has the potential to aid in the characterisation of unknown proteins; for example, when a protein of unknown function interacts with a number of proteins in the same functional group it could well have a similar function.

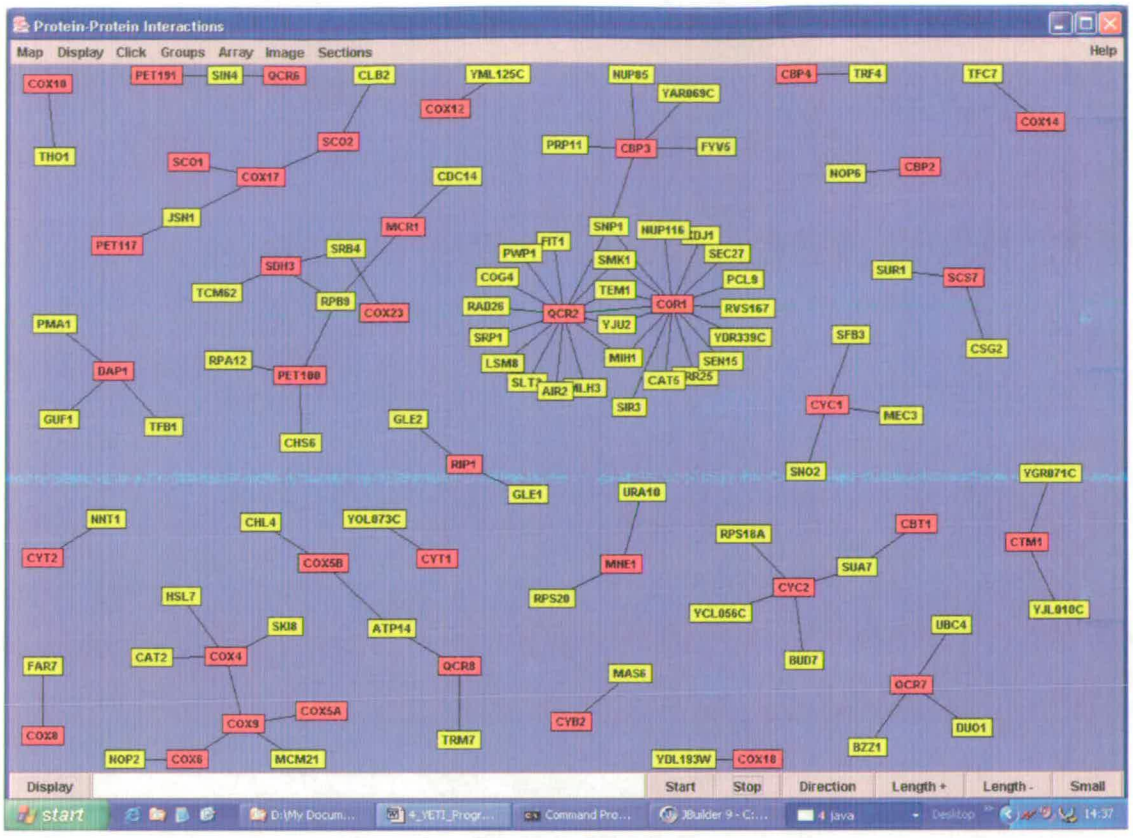


Figure 4.12: Screenshot of the Proteome Section displaying all the interactions of cytochrome proteins
This is a screenshot of the Proteome Section displaying all of the interactions involving cytochrome proteins. In this case, the Analysis Section was used to search for all proteins with the keyword 'cytochrome' in their description and the Proteome Section has displayed all of the interactions involving these proteins. Furthermore, all of the cytochrome proteins have automatically been highlighted in red. This enables users to easily and rapidly examine all of the interactions of a functionally related group of proteins to see if and how they are interacting to fulfil their biological goal.

One of the unique features of the Proteome Section is the ability to overlay gene expression data onto the graphical panel to display the relative gene expression of every protein displayed on the graphical panel (Figure 4.13). It is important to note that it is genes that are expressed not the proteins themselves; however, gene expression data can be overlaid onto the graphical panel by linking proteins to their corresponding gene. Any of the microarray experiments stored in the YETI database can be selected and overlaid onto the graphical panel; the expression of every protein on the panel is represented by colouring the protein boxes with a colour that reflects

its relative gene expression ratio value from the selected experiment. Therefore, this feature enables users to easily examine if particular clusters of interacting proteins have similar relative changes in gene expression. This is useful because a cluster of interacting proteins that are also coexpressed is more likely to be functionally relevant.

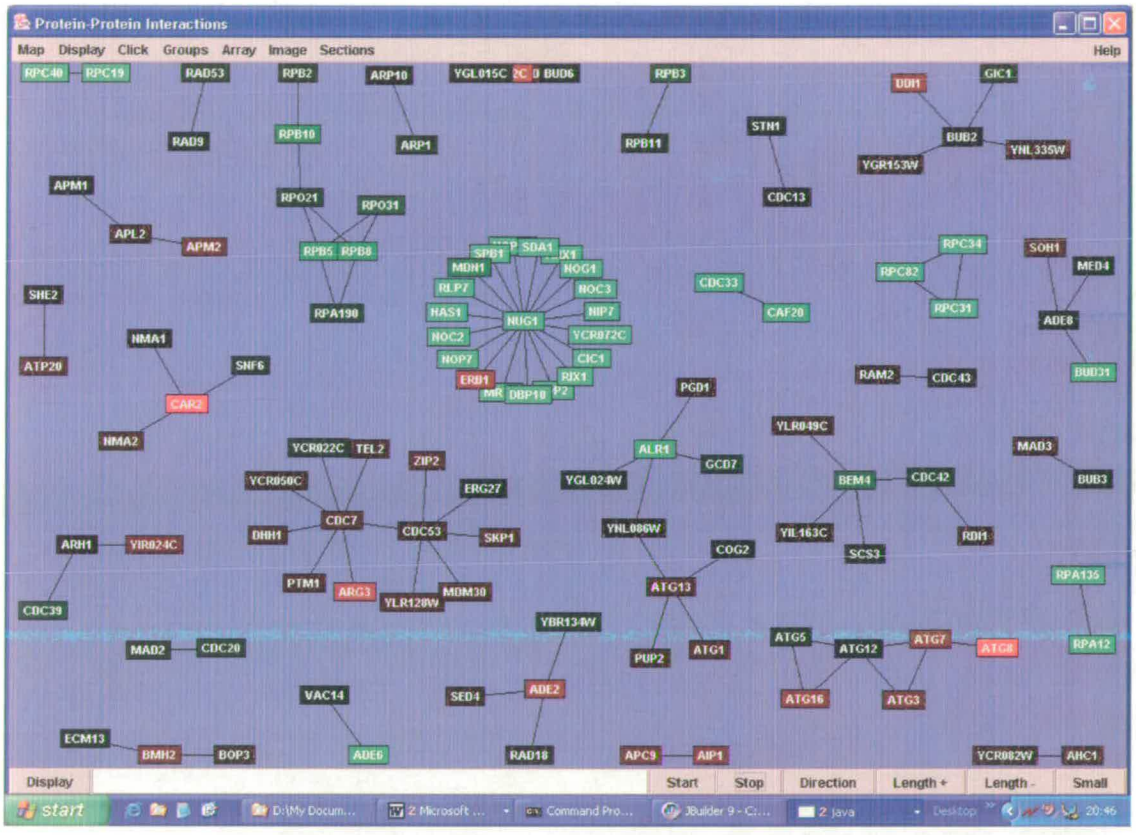


Figure 4.13: Screenshot of the Proteome Section with expression data overlaid
This is a screenshot of the Proteome Section displaying a number of protein-protein interactions with a gene expression microarray experiment overlaid onto the graphical panel. The relative expression of every protein on the panel is represented by colouring the protein boxes; the colour of each protein box is calculated from the relative gene expression ratio of the protein's corresponding gene in the selected microarray experiment. This fairly unique feature enables the user to easily see if particular clusters of interacting proteins have similar relative changes in gene expression.

One of the advanced features of the Proteome Section is the Extend function which enables users to extend or expand out of the interactions currently displayed on the

graphical panel. To do this, users simply need to click on a displayed protein and YETI searches for all interactions involving the selected protein and dynamically adds any interactions that are not already displayed onto the graphical panel. This feature enables the current investigation to be extended further through the proteins that are currently at the periphery; peripheral proteins do not tend to have all of their interactions currently displayed so they could well be linked to other proteins on the panel either directly or indirectly through intermediate proteins. An additional advanced feature is the ability to enter SQL statements directly into the Proteome Section; YETI then uses the entered SQL statement to search the database with and displays any interactions found on the graphical panel. This powerful function gives users complete control over the search criteria enabling them to perform large and complex searches; however, this function obviously requires knowledge of both SQL and the YETI database structure.

Another advanced feature is the Small function which replaces the large labelled protein boxes with small unlabelled boxes (Figure 4.14). This feature enables users to examine entire data sets of protein-protein interactions on the graphical panel to give a good overall impression of the size and connectivity of the data set. Furthermore, the Proteome Section has a Find function that can be used to highlight the location of multiple proteins of interest on the graphical panel; this feature enables users to see where a specific protein of interest is located and to examine if a specific group of proteins are located in the same region of an entire data set (Figure 4.14).

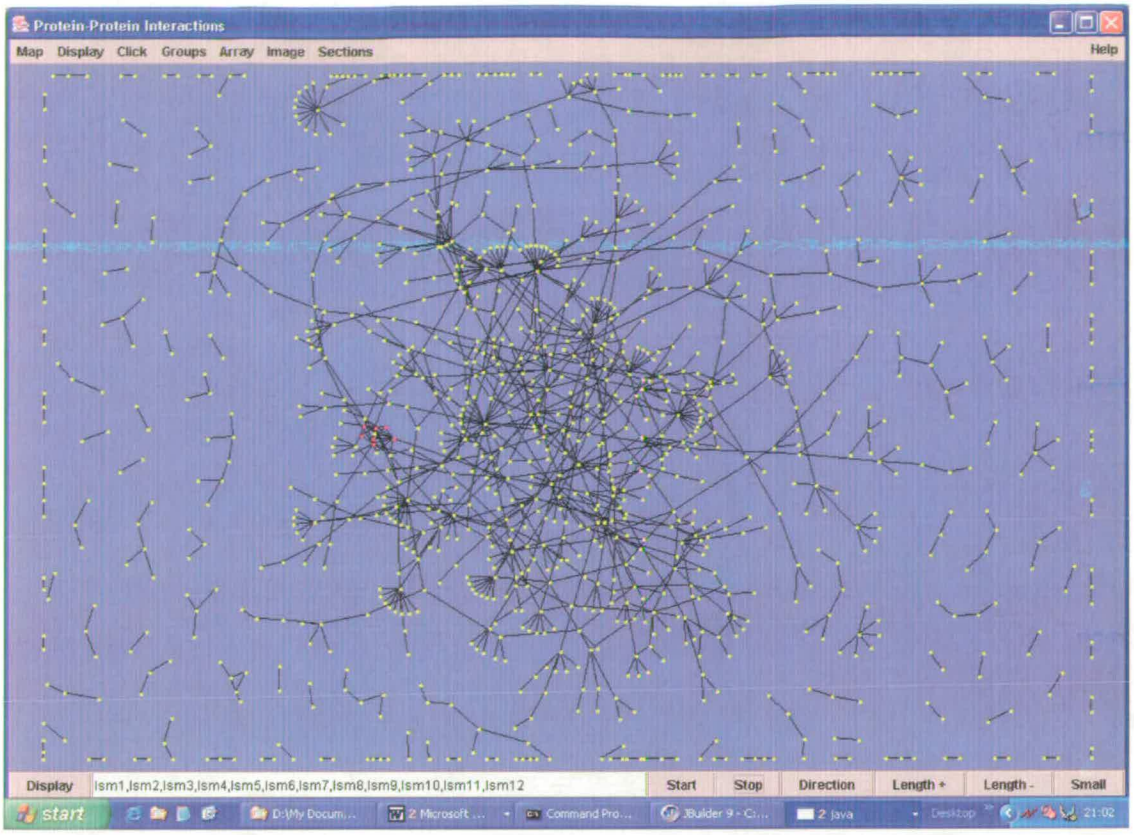


Figure 4.14: Screenshot of the Proteome Section displaying an entire data set
This is a screenshot of the Proteome Section displaying the entire Uetz *et al.* (2000) data set. When the number of interactions displayed on the graphical panel is very large the Small function is automatically activated. The small function replaces the large labelled yellow protein boxes with small unlabelled yellow protein boxes. This means that entire data sets of protein interactions can be displayed on the graphical panel at once; this gives a good overall impression of the size and connectivity of the data set. In this case, a large cluster of proteins is observed in the centre of the panel with numerous smaller clusters located around the periphery. Furthermore, the location of proteins of interest can still be highlighted on the graphical panel. In this case, the location of all the LSM proteins have been highlighted on the panel in red; the LSM proteins form a small cluster within the main large cluster located to the left of the panel centre.

There are also a number of additional features available to users in the Proteome Section: (1) The Direction function can be used to show the direction of all protein interactions displayed on the graphical panel by colouring the bonds linking interacting proteins together; (2) The Start, Stop and Reset buttons can be used to manually start, stop and rest the relaxation algorithm, respectively; (3) The Move option can be used to manually arrange the proteins on the graphical panel by simply mouse clicking on them and dragging them to a new location; (4) The Length option

can be used to increase the default length for all protein interaction bonds on the graphical panel; and (5) The Save option can be used to save the layout of the current graphical panel complete with any protein selections and the Open option can be used to open a previously save graphical panel file.

4.6: Datasheet Window

The Datasheet Window of YETI (Figure 4.15) displays a wide range of information on a single gene of interest and can be launched from numerous points in the YETI program; for example, by mouse clicking on a displayed chromosomal gene in the Chromosome Window as described above. Alternatively, the datasheet of any gene of interest can be viewed simply by entering its name into the Quick Search textfield of the YETI MainFrame. YETI then searches the database for the entered gene name and, if found, subsequently launches a datasheet for the selected gene; if many genes share the entered name, YETI launches a small window displaying the full names and chromosomal locations of all the genes found enabling users to identify the gene they wish to investigate further. The Datasheet Window itself displays a wide range of textual information on the selected gene such as its name(s), length, number of introns, chromosomal location, phenotype, description and GO annotations. Overall, this feature enables users to easily and rapidly view a wide range of information on a specific gene of interest; this is a core feature of YETI which is also utilised by most computational resources revolving around a single gene approach.

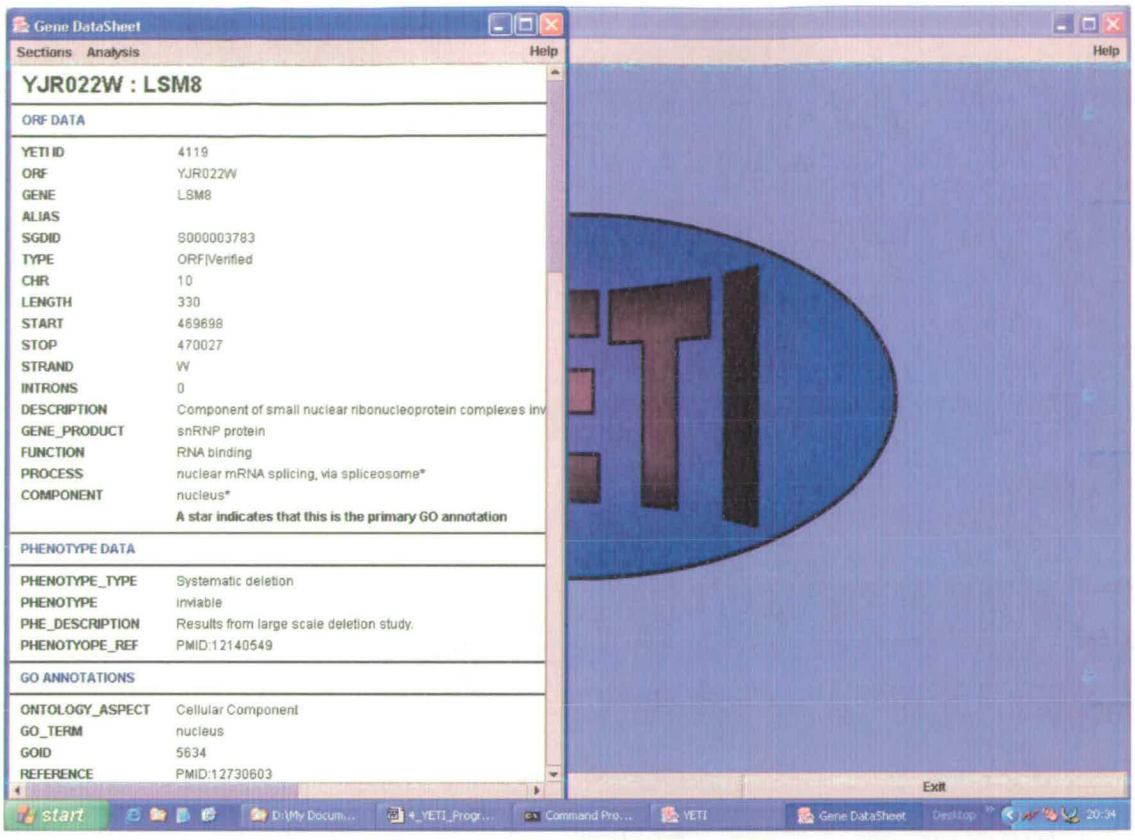


Figure 4.15: Screenshot of the Datasheet Window
This is a screenshot of the YETI Datasheet Window which contains a wide range of information for a selected gene of interest. In this case, the Quick Search function of the YETI MainFrame was used to launch the datasheet of LSM8.

However, the Datasheet Window also contains a number of advanced options that provide direct links to the YETI Sections where data relating to the selected gene is automatically displayed and highlighted (Table 4.2). These links enable users to collectively examine and subsequently investigate all the genes located in the same chromosomal region as the selected gene, all the genes the selected gene is coexpressed with, all the proteins the selected gene's corresponding protein interacts with, and all the genes that share the same GO annotations. These links are especially useful when investigating a potential function for a gene of unknown function and when investigating what other genes a gene of known function may be working with in order to achieve its biological goal. Furthermore, these links again highlight the

main advantages of YETI: (1) The inter-linked sections enable users to investigate all the aspects of a specific gene of interest; and (2) The group approach enables all the genes related to a specific gene of interest to be collectively examined and investigated.

Option	Description
Genome	This option launches the Genome Section and highlights the location of the selected gene on the genome schematic in red.
Chromosome	This option launches the Chromosome Window displaying the relevant chromosome with the selected gene highlighted in red.
Transcriptome	This option launches the Transcriptome Section displaying a hierarchically clustered genome wide gene expression microarray data set and highlights the selected genes location in red.
Proteome	This option launches the Proteome Section displaying all the protein-protein interactions that the selected gene's protein product is involved in and highlights the selected gene's protein product in red.
Function	This option launches the Analysis Section displaying a data table containing a wide range of information on all the genes that have been characterised with the same GO Molecular Function annotation as the selected gene.
Process	This option launches the Analysis Section displaying a data table containing a wide range of information on all the genes that have been characterised with the same GO Biological Process annotation as the selected gene.
Component	This option launches the Analysis Section displaying a data table containing a wide range of information on all the genes that have been characterised with the same GO Cellular Component annotation as the selected gene.
Hybrid	This option launches the Analysis Section displaying a data table containing a wide range of information on all the genes whose protein product interacts with the selected gene's protein product.
Pearson	This option launches the Analysis Section displaying a data table containing a wide range of information on all the genes that are coexpressed with the selected gene.

Table 4.2: Links available from the YETI Datasheet Window
This table contains the names and descriptions of the links available from the Datasheet Window that move users directly into one of the YETI sections where data relating to the selected gene is automatically displayed and highlighted.

4.7: FPC Section

Most current computational resources do not enable users to collectively investigate the properties of entire groups of genes at once. In contrast, the FPC Section of YETI enables users to define specific groups of genes which can then be investigated in further detail in the other sections of YETI; FPC stands for Function, Process and Component which are the three organising principles of the GO annotation system. The GO annotation system has been used over recent years to functionally characterise a large proportion of the genes in *S. cerevisiae*. This characterisation system means that functionally related groups of genes can easily be created consisting of genes that share the same or similar GO annotations. In the FPC Section, there are two groups that can be defined, the so called Red and Green groups, either or both of the groups can be defined and subsequently investigated in further detail in the other sections of YETI.

At the heart of the FPC Section is the GO annotation list which contains all the GO annotations that have been used to characterise *S. cerevisiae* genes (Figure 4.16). By default the list contains all the annotations from all three GO organising principles (Function, Process and Component) which are sorted in alphabetical order; however, principles can easily be removed and added to control what annotations are displayed in the list. Although simple, this comprehensive list enables users to easily browse all the GO annotations used to characterise *S. cerevisiae* genes and rapidly find any annotations of interest. Single or multiple annotations of interest in the list can simply be selected by mouse clicking on them and then assigning them to either the

Red or Green group; this results in all the genes characterised with the selected annotations being assigned to the selected group. Therefore, the FPC Section enables users to easily and rapidly construct specific or broad groups of functionally related genes to investigate in further detail in the other sections of YETI. Furthermore, as two different groups of genes can be defined this enables the properties of both groups to be collectively examined and compared. Alternatively, users can manually define groups themselves by simply entering the names of multiple genes of interest and assigning them to one of the two groups.

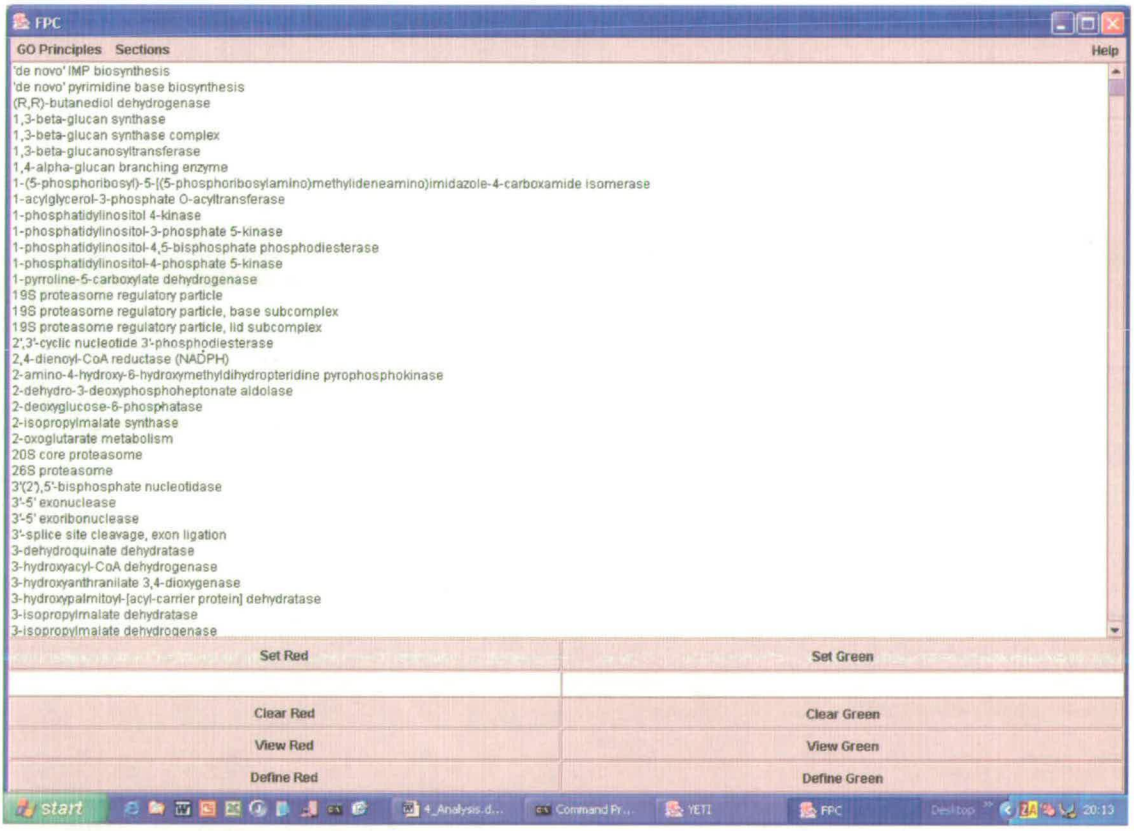


Figure 4.16: Screenshot of the FPC Section
This is a screenshot of the Function, Process and Component (FPC) Section. The FPC Section enables users to define up to two groups of genes (Red & Green) to investigate in further detail. The groups can be defined by entering the names of multiple genes into the corresponding textfield or by selecting GO annotations from the comprehensive list. Once a group has been defined users can move directly into the other YETI Sections to collectively view relevant data on all the members of the group.

The FPC Section is effectively inter-linked with the other YETI sections enabling users to swiftly move directly into another YETI section where information related to all the selected genes will be automatically displayed and highlighted. This enables all of the properties of a group of genes defined in the FPC Section to be collectively investigated in the other YETI Sections. For example, all of the genes characterised with the GO biological process annotation of 'nuclear mRNA splicing, via spliceosome' could be selected in the FPC Section and then collectively investigated in all the other YETI sections. The Analysis Section would display a wide range of information on all the spliceosome genes enabling users to examine how many genes are currently characterised as being involved in the spliceosome, what each gene is as well descriptions as to what each gene does. The Genome Section would highlight the location of all the spliceosome genes on the genome schematic enabling users to investigate if any of the genes are colocated in the same or similar genomic regions and to also examine what other genes are located in these regions. The Transcriptome Section would highlight the gene expression profiles of all the spliceosome genes in the hierarchically clustered gene expression data set enabling users to investigate if any of the genes are located in the same expression cluster and to also examine what other genes are located in these regions. The Proteome Section would display all of the protein interactions involving the spliceosome proteins enabling users to investigate if the spliceosome proteins are interacting directly or indirectly with one another and to also examine what other proteins they are interacting with. Overall, this enables all the properties of the whole group to be collectively investigated to examine if and how they are working together to achieve their biological goal and to also investigate what other genes/proteins they may be

working with; this could enable functions to be inferred on any unknown gene consistently associated with the group.

In addition, YETI also includes a Slim FPC Section which is essentially identical to the FPC Section except that it concerned with GO Slim annotations as opposed to the complete GO annotations. GO Slims are a cut-down version of the complete GO ontology and give a broad overview of the ontology content without the detail of the specific fine grained terms. The YETI Slim FPC Section displays a list of all the GO Slim annotations used to characterise the genes of *S. cerevisiae* and therefore enables users to select annotations to construct much broader groups of functionally related genes which can then be collectively investigated in the other YETI sections.

4.8: Discussion

One of the main advantages of YETI is its ease of use. YETI was designed with simplicity in mind with simple navigation mechanisms to move through the program, flexible search mechanisms and clear graphical representations of the data in unison with a number of advanced features and functionality. YETI aims to be a user friendly workbench that enables both wet and dry laboratory scientists to easily and rapidly explore all the aspects of the stored functional genomics data in an integrated modular fashion; it enables users to easily and rapidly find the data they want, investigate the intricacies of broad biological processes and test specific hypotheses.

One of the unique features of YETI is the fact that it can utilise both a single gene approach and a group approach. The single gene approach enables users to examine all of the properties of a specific gene of interest. For example, YETI can display a wide range of textual information on the gene, display what genes are located in the same chromosomal region, display what genes it is coexpressed with, display what proteins the gene's corresponding protein interacts it, and display what genes share the same GO annotations. This single gene approach is especially useful for helping to investigate the function of an unknown gene in a 'guilt by association' approach as it enables users to examine what other genes the selected gene is associated with and what their functions are. On the other hand, the group approach enables all the properties of an entire group of genes to be collectively investigated. For example, YETI can collectively display a wide range of textual information on all the genes to examine if they share the same GO annotations and are involved in the same biological process, collectively display where they are all located in the genome to examine if they are colocated and what other genes are colocated with them, collectively display all their expression profiles to examine if they are coexpressed and what other genes are coexpressed with them, and collectively display what proteins they interact with to examine if they interact with one-another and what other proteins they interact with. This group approach enables all the genes/proteins involved in an entire biological process to be collectively examined as a whole to investigate the dynamics of how they are working together to achieve their biological goal and to also examine what other proteins they may be working with; this could enable functional roles for any common proteins of unknown function to be inferred.

Furthermore, YETI enables the properties of multiple groups to be collectively examined at the same time enabling comparisons to be made between the groups.

Another unique feature of YETI is its inter-linked sections which enable users to select any feature of interest from one section and then swiftly move to another section where data relating to their selection is automatically displayed and highlighted. A feature of interest could be a specific chromosomal region from the Genome Section, a gene expression cluster from the Transcriptome Section, a protein interaction cluster from the Proteome Section, all the genes sharing a specific GO annotation from the FPC Section or all the genes returned from a specific data search from the Analysis Section. For example, when examining the hierarchically clustered gene expression data in the Transcriptome Section a specific cluster of interest can be readily selected to examine what genes are located in the cluster and what their functions are in the Analysis Section (to investigate possible shared functionality), examine where all the genes are located in the Genome Section (to investigate possible colocation), and examine what their corresponding proteins are interacting with in the Proteome Section (to investigate possible inter-connectivity). Therefore, these inter-linked sections (in combination with the group approach) enable all the properties of a specific feature of interest to be collectively investigated and also enable expansion of the investigation at any point as additional genes can be selected and added to the search group.

In this chapter, the features and functions of the core YETI program were detailed and discussed along with a number of examples to illustrate their potential use.

However, in order to demonstrate the potential and utility of YETI as an analysis tool, a number of case studies are presented in the forthcoming chapters of this thesis. In the next chapter, a number of single gene case studies are presented which demonstrate the utility of YETI in investigating potential functions for specific genes of unknown function. In later chapters, the additional correlation sections of YETI are discussed along with the results of various correlation analyses performed between the stored functional genomic data sets. Furthermore, these chapters also include a number of much broader case studies which demonstrate the utility of YETI in investigating the dynamics of how groups of functionally related proteins are working together to achieve their biological goal.

Chapter 5

Single Gene Case Studies

5.1: Introduction

As described previously, the YETI Datasheet Window enables users to easily and rapidly view a wide range of information on a specific gene of interest. However, the Datasheet Window also provides a number of direct links to the core YETI Sections where data relating to the selected gene is automatically displayed and highlighted. These links are especially useful when investigating a potential function for a gene of unknown function in a 'guilt-by-association' approach as they enable users to rapidly examine what genes/proteins are colocated, coexpressed and interact with the selected gene and also enable users to compare all of their functions to investigate possible shared functionality.

To illustrate the potential of YETI to aid in the assignment of biochemical functionality in a 'guilt by association' approach, a simple computer program was written to suggest a potential biological role for every gene of unknown function in the *S. cerevisiae* genome. Initially, this program simply finds every gene in the genome that currently has both a 'molecular function unknown' and 'biological process unknown' GO annotation. For each gene, the program retrieves all the GO biological process annotations of all the genes that it is coexpressed with (Pearson >= 0.7) and also all the GO biological process annotations of all the proteins that interact with its protein product; biological process annotations were chosen because they tend to be much broader than molecular function annotations. The program then searches for the most common GO biological process annotation associated with each gene of unknown function and outputs this annotation along with an occurrence score. Any genes that are associated with a large number of genes involved in the

same biological process could then be investigated in further detail in YETI. To illustrate this, a number of case studies are presented below that investigate possible biological roles for genes of unknown function; the title of each case study reflects the name of the unknown gene along with the GO biological process that it is potentially involved in, as suggested by the computer program.

5.2: MOH1 - Negative regulation of gluconeogenesis

The YETI Datasheet Window for MOH1 (YBL049W) shows that it is currently an 'uncharacterised ORF' of unknown function; its three GO annotations are 'molecular function unknown', 'biological process unknown' and 'cellular component unknown'. However, one clue as to the function of MOH1 is provided in its textual description which states that MOH1 is 'not required for growth on non-fermentable carbon sources'. To investigate possible functional roles for MOH1 further, the links from the Datasheet Window of MOH1 to the core sections of YETI were utilised to examine what genes were coexpressed with MOH1 and what proteins interacted with its corresponding protein product.

Firstly, the Hybrid link from the MOH1 Datasheet Window was used to move directly into the Analysis Section to collectively examine what proteins interact with MOH1 and what their functions are. YETI shows that MOH1 interacts with a total of 12 other proteins, 6 of which are characterised with the GO biological process of 'negative regulation of gluconeogenesis'; specifically: GID7, GID8, FYV10, RMD5, VID28 and VID30. Secondly, the Pearson link from the MOH1 Datasheet Window

was used to move directly into the Analysis Section to collectively examine what genes MOH1 was coexpressed with and what their functions are. YETI shows that MOH1 is most highly coexpressed with SPG4 (Pearson = 0.92) and also coexpressed SPG1 (Pearson = 0.89), SPG5 (Pearson = 0.77) which are all of unknown function. Interestingly, the descriptions of SPG4, SPG1 and SPG5 all state that, like MOH1, they are 'not required for growth on non-fermentable carbon sources'. Furthermore, YETI also shows that MOH1 is coexpressed with GID8 (Pearson = 0.70) and FYV10 (Pearson = 0.80) both of which were previously shown to interact directly with MOH1 and are both characterised with the 'negative regulation of gluconeogenesis' annotation.

As the biological process 'negative regulation of gluconeogenesis' was consistently associated with MOH1, this biological process was itself investigated in further detail in YETI. To this end, the 'negative regulation of gluconeogenesis' annotation was selected in the FPC Section; selecting an annotation in the FPC Section has the affect of selecting all the proteins currently characterised with the selected annotation and therefore enables all these proteins to be collectively investigated in the other sections of YETI. The GO biological process of 'negative regulation of gluconeogenesis' is defined as 'any process that stops, prevents or reduces the rate of gluconeogenesis'; the GO biological process of 'gluconeogenesis' is itself defined as 'the formation of glucose from non-carbohydrate precursors, such as pyruvate, amino acids and glycerol'. The Analysis Section shows that there are currently nine proteins characterised with the 'negative regulation of gluconeogenesis' annotation; specifically: GID7, GID8, FYV10, RMD5, UBC8, UBP14, VID24, VID28 and

VID30. The Proteome Section shows that a large number of these proteins interact highly with one another forming a tight interaction cluster (Figure 5.1). An integral part of this cluster is MOH1 which (as described above) directly interacts with a number of the 'negative regulation of gluconeogenesis' proteins.

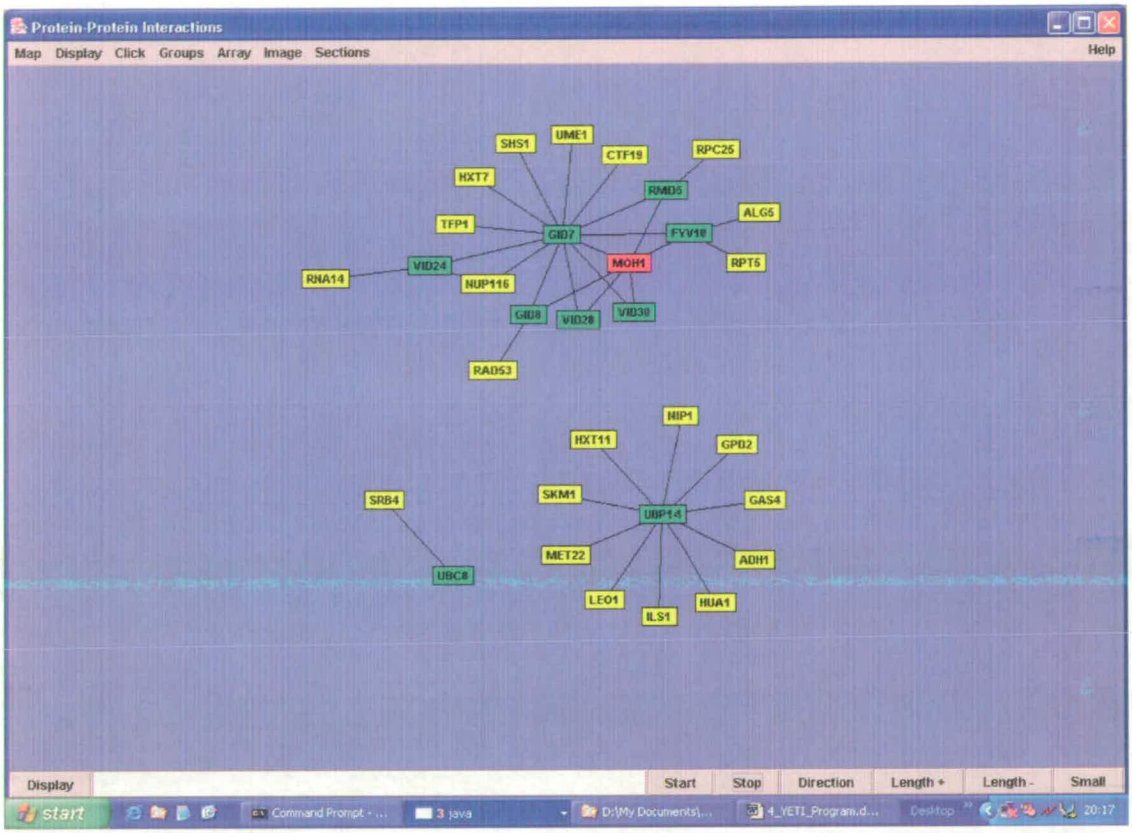


Figure 5.1: Screenshot of the Proteome Section displaying all the interactions involving 'negative regulation of gluconeogenesis' proteins
This is a screenshot of the Proteome Section displaying all the protein-protein interactions involving any of 'negative regulation of gluconeogenesis' proteins. The proteins involved in the 'negative regulation of gluconeogenesis' are highlighted in green on the graphical panel. As can be seen, there is a tight cluster consisting of a number of the 'negative regulation of gluconeogenesis' proteins. Furthermore, an integral member of this cluster is MOH1 (highlighted in red) which interacts with a number of 'negative regulation of gluconeogenesis' proteins.

Overall, the observations described above suggest that the biological process of 'negative regulation of gluconeogenesis' involves a small specialised group of proteins and given the high interactivity (and coexpression) of MOH1 with these

proteins naturally leads one to suggest that this protein is also involved in this biological process. Interestingly, there are a number of additional observations that support this. Firstly, the fact that MOH1 is not required for growth on non-fermentable carbon sources supports its role as a negative regulator of gluconeogenesis; examples of non-fermentable carbon sources are glycerol, lactate, ethanol and acetate whereas examples of fermentable carbon sources are glucose and fructose. Non-fermentable carbon sources such as ethanol are metabolised in the Krebs cycle, with ATP being obtained from respiration (Ronne, 1995). However, the cell also needs hexose phosphates for biosynthetic reactions and in the absence of glucose these must be produced by gluconeogenesis. Most gluconeogenic steps are catalysed by glycolytic enzymes but two steps are irreversible and therefore have unique gluconeogenic enzymes; specifically: fructose bisphophatase (FBP1) and PEP carboxykinase (PCK1). FBP1 and PCK1 are repressed by glucose to prevent glycolysis and gluconeogenesis from taking place simultaneously which would rapidly deplete ATP levels. Therefore, if MOH1 is involved in the negative regulation of gluconeogenesis it will be required for effective growth on fermentable carbon sources where gluconeogenesis is repressed but not required on non-fermentable carbon sources where gluconeogenesis is de-repressed. Indeed, this seems to be the case as the description of MOH1 states that it is not required for growth on non-fermentable carbon sources. Secondly, collectively examining the descriptions of all the genes currently characterised with the 'negative regulation of gluconeogenesis' annotation in the Analysis Section reveals further proof to suggest that MOH1 is also involved in this biological process. Specifically, the description of GID7 (also known as MOH2) states that 'computational analysis suggests that GID7

and MOH1 have similar functions' which further links MOH1 to this biological process.

5.3: YKL056C - Protein Biosynthesis

The YETI Datasheet Window for YKL056C shows that it is currently an 'uncharacterised ORF' of unknown function; its three GO annotations are 'molecular function unknown', 'biological process unknown' and 'cytoplasm'. YETI shows that YKL056C is coexpressed (Pearson cutoff of 0.7) with a staggering 121 genes that are characterised with both the 'structural constituent of ribosome' and 'protein biosynthesis' GO molecular function and biological process annotations, respectively; furthermore, YKL056C is coexpressed with 97 of these genes using a Pearson cutoff of 0.8, 65 at a Pearson cutoff of 0.85, and 9 at a Pearson cutoff of 0.9. Virtually all of these 'protein biosynthesis' genes are characterised equally with either the 'cytosolic small ribosomal subunit (sensu Eukaryota)' or 'cytosolic large ribosomal subunit (sensu Eukaryota)' GO cellular component annotations; as opposed to the 'mitochondrial small ribosomal subunit' or 'mitochondrial large ribosomal subunit'. Therefore, this strongly suggest that YKL056C is also a 'structural constituent of ribosome' involved in 'protein biosynthesis' and part of either the 'cytosolic small ribosomal subunit (sensu Eukaryota)' or 'cytosolic large ribosomal subunit (sensu Eukaryota)'; this is further supported by the fact that YKL056C is already characterised as being located in the cytoplasm.

There are a number of other genes of unknown function that are also highly

coexpressed with a large number of 'protein biosynthesis' genes. YMR116C is a 'verified ORF' with 'molecular function unknown', 'biological process unknown' and 'cytoplasm' as its three GO annotations. YETI shows that YMR116C is coexpressed with 112 'protein biosynthesis' genes at a Pearson cutoff of 0.7 and 74 at a Pearson cutoff of 0.8. However, the description of YMR116C already states that it is a 'core component of the ribosome' and the observations made through YETI further support this. YMR321C is currently an 'uncharacterised ORF' of unknown function; its three GO annotations are 'molecular function unknown', 'biological process unknown' and 'cellular component unknown'. YETI shows that YMR321C is coexpressed with 108 'protein biosynthesis' genes at a Pearson cutoff of 0.7 and 49 at a Pearson cutoff of 0.8. Similarly, YJR124C, YJL193W and YBR025C are all genes of unknown function that are highly coexpressed with a large number of 'protein biosynthesis' genes.

YETI has therefore been used to suggest possible functional roles for all the genes of unknown function discussed above (YKL056C, YMR116C, YMR321C, YJR124C, YJL193W and YBR025C) through examination of the GO annotations of their coexpressed genes. Specifically, they are all potentially a 'structural constituent of ribosome' involved in 'protein biosynthesis' and part of either the 'cytosolic small ribosomal subunit (sensu Eukaryota)' or 'cytosolic large ribosomal subunit (sensu Eukaryota)'. Interestingly, a number of additional facts support these predictions. Firstly, although the *Saccharomyces* Genome Database (SGD; Cherry *et al.*, 1998; http://www.yeastgenome.org/) characterises YKL056C as a protein of unknown function, the Munich Information Centre for Protein Sequences (MIPS; Mewes *et al.*,

1998; http://mips.gsf.de/genre/proj/yeast/index.jsp) characterises it as a 'protein putative involved in cytoplasmic ribosome function'. Furthermore, a recent study by Barriot *et al.* (2004) also associated YKL056C and YMR116C with numerous ribosomal genes. Barriot *et al.* (2004) proposed a new strategy for the integration of sequence data with other functional genomic data such as gene expression profiles. They developed an associated tool (BlastSets) which was used to automatically retrieve the members of the ribosome complex based on the mining of expression profiles; this enabled functional roles for genes of unknown function associated with this complex to be inferred.

5.4: YMR148W – Tricarboxylic Acid Cycle

The YETI Datasheet Window for YMR148W shows that it is currently an 'uncharacterised ORF' of unknown function; its three GO annotations are 'molecular function unknown', 'biological process unknown' and 'integral to the membrane'. YETI shows that although the protein product of YMR148W does not interact with any other proteins, YMR148W itself is coexpressed with nine other genes; four of these genes are also of unknown function. YMR148W is coexpressed with SDH4 (Pearson = 0.85), SDH1 (Pearson = 0.72) and SDH2 (Pearson = 0.71) all of which are subunits of succinate dehydrogenase characterised with both the 'tricarboxlylic acid cycle' and 'mitochondrial electron transport, succinate to ubiquinone' GO biological process annotations. YMR148W is also coexpressed with CYB2 (Pearson = 0.79) which is a cytochrome involved in 'electron transport', and MBR1 which is a mitochondrial stress response protein involved in 'aerobic respiration'. Furthermore,

all of these genes have cellular component annotations linking them to the mitochondria.

YETI has firmly linked, through its coexpression, YMR148W with aerobic respiration and the mitochondrial electron transport chain; therefore, YMR148W could well have a functional role involved in this or a related biological process as well. This hypothesis is further supported by the fact that YMR148W is already characterised with the 'integral to membrane' GO cellular component annotation; many of the proteins involved in the mitochondrial electron transport chain are located in the inner mitochondrial membrane. In addition, although YETI showed that YMR148W was only coexpressed (Pearson >= 0.7) with 9 other genes, the Transcriptome Section of YETI shows that YMR148W is located in a small cluster of genes in the gene expression hierarchical tree (Figure 5.2). Further examination of this cluster in the Analysis Section reveals that virtually all of these genes are associated with the mitochondria and are characterised with either 'aerobic respiration', 'tricarboxylic acid cycle' or 'ATP synthesis coupled proton transport' GO biological process annotations. Therefore, this further links YMR148W to the biological process of aerobic respiration and the mitochondrial electron transport chain. Although this is a fairly broad functional assignment it is a good starting point for further investigation and characterisation of this gene and its encoded protein product.

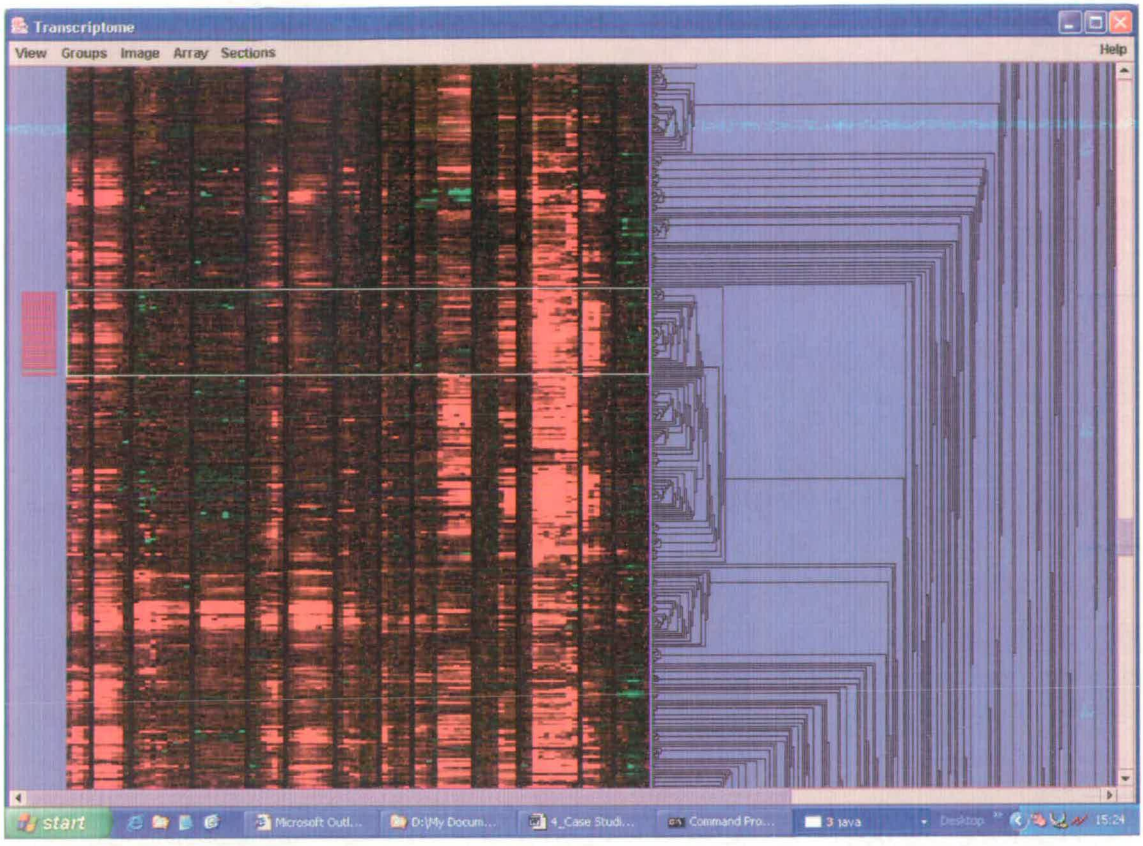


Figure 5.2: Screenshot of the Transcriptome Section highlighting the location of YMR148W
This is a screenshot of the Transcriptome Section with the location of YMR148W in the gene expression hierarchical tree highlighted with a green line to the left. Furthermore, the entire cluster that YMR148W is located in has subsequently been selected for further investigation and highlighted with red lines to the left.

5.5: YLR364W - Sulphate Assimilation

The YETI Datasheet Window for YLR364W shows that it is currently an 'uncharacterised ORF' of unknown function; its three GO annotations are 'molecular function unknown', 'biological process unknown' and 'cytoplasm'. YETI shows that although the protein product of YLR364W does not interact with any other proteins, YLR364W itself is coexpressed with five other genes. Four of the genes YLR364W is coexpressed with are MET3 (Pearson = 0.73), MET10 (Pearson = 0.72), MET1 (Pearson = 0.72) and MET16 (Pearson = 0.71); all four of these MET genes are

characterised with the 'sulphate assimilation' GO biological process annotation. Furthermore, three of these MET genes have protein products located in the cytoplasm which is also where the protein product of YLR364W is located; the cellular location of the protein product of MET1 is currently unknown.

The Transcriptome Section of YETI shows that YLR364W is located in a small cluster of genes in the gene expression hierarchical tree (Figure 5.3). Further examination of this cluster in the Analysis Section reveals that there are a large number of genes involved in the metabolism of sulphur compounds with GO biological process annotations such as 'sulphur amino acid metabolism', 'methionine metabolism', 'sulphate assimilation', 'sulphur metabolism' and 'sulphate transport'. However, there are also a large number of genes involved in the metabolism of nitrogen compounds located in this cluster with GO biological process annotations such as 'nitrogen compound metabolism', 'allantoin catabolism', 'asparagine metabolism' and 'serine family amino acid biosynthesis'.

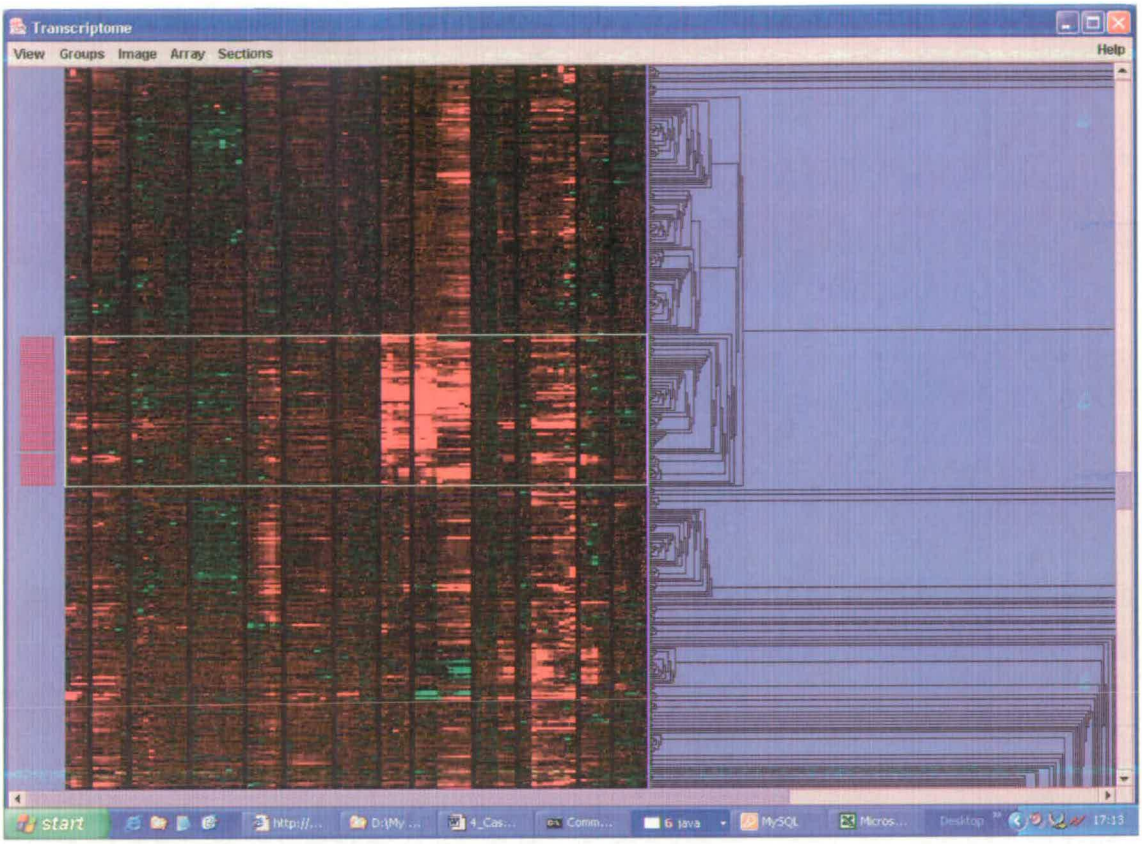


Figure 5.3: Screenshot of the Transcriptome Section highlighting the location of YLR364W
This is a screenshot of the Transcriptome Section with the location of YLR364W in the gene expression hierarchical tree highlighted with a green line to the left. Furthermore, the entire cluster that YLR364W is located in has subsequently been selected for further investigation and highlighted with red lines to the left.

Therefore, as YLR364W is located in the same region of the hierarchical tree as many genes involved in sulphur and nitrogen compound metabolism it could well be involved in one of these biological processes. However, given the fact that YLR364W is coexpressed (Pearson >= 0.7) and colocated in the cell with 4 genes involved in 'sulphate assimilation' it can be argued that it is more likely to be involved in sulphur compound rather than nitrogen compound metabolism. Interestingly, this functional prediction is supported by a recent study which used microarrays to characterise the transcriptional response of *S. cerevisiae* to growth limitation by carbon, nitrogen, phosphorus or sulphur (Boer *et al.*, 2003). In this

study, YLR364W was characterised as 'specifically higher expression under sulfur limitation' and was therefore hypothesised to be involved in sulphur compound metabolism.

5.6: IES5 – Chromatin Remodelling

The YETI Datasheet Window for IES5 (YER092W) shows that it is currently a 'verified ORF' of unknown function; its three GO annotations are 'molecular function unknown', 'biological process unknown' and 'nucleus'. However, IES5's description states that it is a 'protein that associates with the INO80 chromatin remodelling complex under low salt-conditions'; this description is based on a study by Shen *et al.* (2003) who found, through complex purification and peptide sequencing techniques, IES5 to be associated with the INO80 complex under low salt conditions. Therefore, YETI was used to see if it could further associate IES5 with the INO80 chromatin remodelling complex and also clarify its molecular function.

The GO cellular component INO80 complex is defined as a 'multisubunit protein complex that contains the Ino80p ATPase; exhibits chromatin remodelling activity and 3' to 5' DNA helicase activity'. As described above, IES5 is already characterised as being located in the nucleus which places it in the correct cellular location to potentially be involved in chromatin remodelling. The Proteome Section of YETI shows that IES5 interacts with four proteins; specifically: NHP10, DID4, ISE3 and ATG17. NHP10 are ISE3 are both characterised with the 'chromatin remodelling' biological process annotation and the 'INO80 complex' cellular

component annotation. YETI shows that these interactions were not derived from the Shen *et al.* (2003) study described above, rather, they were derived from high throughput protein-protein interactions studies; the NHP10-IES5 interaction was reported in both the Gavin *et al.* (2002) and Uetz *et al.* (2000) studies whereas the IES3-IES5 interaction was reported in the Ito *et al.* (2001) study. Therefore, this directly links IES5 to the INO80 complex and also links it with a certain degree of confidence due to the interactions being reported in multiple studies. YETI shows that IES5 is only coexpressed with one gene at a Pearson cutoff of 0.7; specifically, YKL069W which is of unknown function. Furthermore, the Transcriptome Section shows that IES5 is not located in a distinct cluster in the gene expression hierarchical tree and that the surrounding genes have a range of functions, none of which are related to chromatin remodelling. Therefore, IES5 could not be linked to the INO80 Complex through its expression pattern; however, lowering the Pearson cutoff reveals that IES5 is coexpressed with SLD3 (Pearson = 0.64) which is involved in the initiation of DNA replication and has chromatin binding activity.

By using the FPC Section to select the 'INO80 complex' annotation from the GO list, all of the proteins that are currently assigned to this complex could be collectively investigated in the other sections of YETI to examine how they are working together in order to achieve their biological goal. The Proteome Section shows that a single cluster of interacting proteins is formed that contains all of the INO80 complex proteins (Figure 5.4); this is to be expected as they are members of the same complex. At the centre of this cluster is the protein NHP10 which directly interacts with all but one of the other INO80 complex proteins; in addition, NHP10

also interacts with IES5 (as described above). Although none of the other INO80 complex proteins interact directly with one another, there are a number of additional 'bridging' proteins that link proteins of the INO80 complex together; specifically: ESA1 (histone deacetyltransferase activity), RVS167 (actin associated protein), SAP105 (protein phosphatase activity) and our protein of interest IES5. The Transcriptome Section shows that members of the INO80 complex are not colocated and are dispersed fairly evenly through the gene expression hierarchical tree. Furthermore, YETI shows that none of the INO80 complex genes are coexpressed with each other at a Pearson cutoff of 0.7. It is quite surprising that none of the INO80 complex genes are coexpressed together given that they are all members of the same functional complex. One explanation for this observation could be that the microarray data set currently stored in the YETI database (Gasch *et al.*, 2000) may not be suitable for highlighting the relationships between the expression of these genes and perhaps other microarray data sets would yield better results in this case.

Overall, YETI further supports the hypothesis that IES5 is part of the 'INO80 complex'. IES5 is located in the nucleus, directly interacts with two other members of the complex (including the apparent core member) and although it is not coexpressed with any of the other members, none of the members of this complex appear to be coexpressed with one another. However, YETI can not shed any light on the functional role of IES5 within the INO80 complex in this instance.

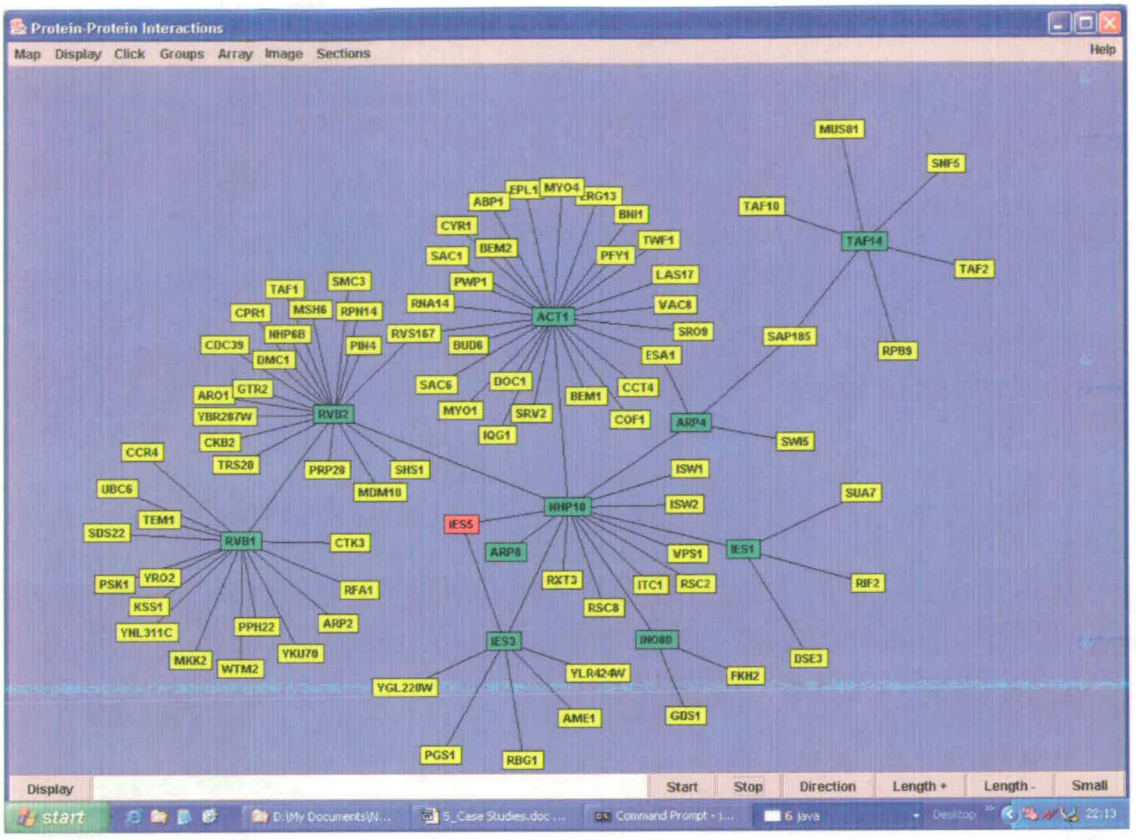


Figure 5.4: Screenshot of the Proteome Section displaying all the interactions involving 'INO80 complex' proteins
This is a screenshot of the Proteome Section displaying all the protein-protein interactions involving any of 'INO80 complex' proteins. The proteins involved in the 'INO80 complex' are highlighted in green on the graphical panel. As can be seen, one large cluster is formed consisting of all the 'INO80 complex' proteins. At the centre of this cluster is NHP10 which also interacts with IES5 (highlighted in red).

5.7: Discussion

The case studies above illustrate the potential of YETI to aid in the assignment of biochemical functionality to a specific gene of interest in a 'guilt-by-association' approach. The Datasheet Window of YETI enables users to view a wide a range of information relating to what is currently known about a specific gene in a single gene approach. However, the links available from the Datasheet Window to the core YETI sections enable users to collectively examine and compare information on all the genes associated with a specific gene in a group approach; associated can mean

coexpressed, interacted, colocated in the cell or colocated in the genome. If a gene of unknown function is associated with a large number of genes involved in the same biological process then this could enable a possible functional role to be inferred (the concept of guilt by association). Overall, the Datasheet Window and its associated links enable users to investigate the potential function of a specific gene of interest, to test whether it is involved in a specific biological process, and to investigate what others genes it may be working with in order to achieve its biological goal. However, it is important to note although the guilt by association approach can readily be used to suggest possible functional roles for genes of unknown function, these suggestions need to be confirmed by experiments in the laboratory.

The case studies presented above also highlight that the textual descriptions of genes can contain a wealth of useful information but unlike the GO annotations this information is not structured or linked in any way. For example, the description of GID7 states that it has a similar function to MOH1, however this information is not present in the description of MOH1 where it is perhaps of more use. The lack of linkage and structure of this information also highlights the usefulness of the YETI QueryBuilder function which enables keyword searches of descriptions and annotations to find potentially related groups of genes. Furthermore, these case studies also highlight that the different *S. cerevisiae* computational resources can contain different information and that the scientific literature contains a wealth of predicted biological roles for the unknown genes of *S. cerevisiae*. Therefore, the integration of effective text mining techniques that can automatically extract functional associations of unknown genes from the scientific literature with the major

S. cerevisiae computational resources would be useful developments.

Chapter 6

Genome vs Proteome Correlation Analysis

6.1: Introduction

A Genome vs Proteome correlation analysis was performed using YETI to investigate if there was a tendency for proteins that interact with one another to be located near each other on the genome. As interacting proteins are likely to be related functionally, this analysis could reveal a high level organisation of the genome where interacting proteins of similar function are colocated. For this analysis, it is important to note that every interacting protein corresponds to a specific ORF in the *S. cerevisiae* genome. The first step is to identify the number of protein-protein interactions where both interacting proteins are located on the same chromosome and test if this number is statistically relevant by comparing it to the number expected if it is assumed the genomic location of interacting proteins is random. The second step is to calculate the average distance between all interacting proteins located on the same chromosome to see if there is a tendency for them to be located near each other. In addition, whether there is an overall correlation or not, the closest interacting proteins and the chromosomal regions they are located in can be investigated in further detail using YETI.

6.2: Correlation Matrix

To investigate a potential correlation between the genomic locations of interacting proteins the approach developed by Ge *et al.* (2001) was applied. Ge *et al.* (2001) originally investigated a potential correlation between expression clusters and interaction clusters. In this analysis, expression clusters are replaced with chromosome clusters where each nuclear chromosome is considered to be a cluster

comprised of all the ORFs located on it; therefore, the 16 nuclear chromosomes of *S. cerevisiae* correspond to 16 chromosome clusters. A two-dimensional interaction matrix is generated by organising the chromosome clusters into two identical axes; for the 16 chromosome clusters, the matrix arrangement results in 16^2 squares. Each square in the matrix represents all the pairwise interactions of ORFs within a single chromosome cluster (diagonal or intracluster squares) or between different chromosome clusters (nondiagonal or intercluster squares). Therefore, pairs of ORFs whose products interact can be assigned to their corresponding intracluster or intercluster squares (Figure 6.1). For each square, an index of protein interaction density (PID) is calculated as the ratio of the number of observed protein-protein interactions (IP) to the total number of possible protein-protein interactions (PP); this IP/PP ratio is scaled by a factor of 100,000 to give PID values typically in the range of 0 to 100. It can be reasoned that for a given protein-protein interaction data set, significantly higher PIDs for intracluster (diagonal) versus intercluster (non-diagonal) squares would be the first step in revealing a possible correlation between genome location and protein interaction.

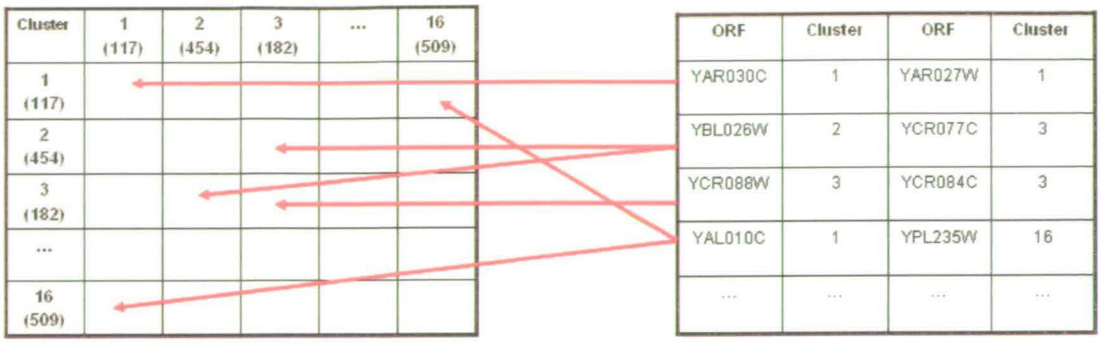


Figure 6.1: Strategy for genome-proteome correlation mapping
The two-dimensional matrix on the left shows the pairwise combinations between the 16 chromosome clusters; the chromosome cluster numbers are indicated on the corresponding rows and columns of the matrix along with the number of ORFs each chromosome cluster contains (in brackets). The table on the right shows protein interaction pairs together with the chromosome cluster to which the corresponding ORFs belong. For each interaction pair, arrows point to its corresponding squares in the two-dimensional chromosome matrix. For example, the first interaction is between two ORFs which are both located on chromosome 1; therefore, this interaction is assigned to chromosome 1's intracluster (diagonal) square which represents pairwise interactions within chromosome 1. Whereas, the second interaction is between an ORF located on chromosome 2 and an ORF located on chromosome 3; therefore, this interaction is assigned to both the chromosome 2/chromosome 3 and the chromosome 3/chromosome 2 intercluster (non-diagonal) squares which both represent pairwise interactions between these two chromosomes. In actual fact, as the matrix is duplicated on either side of the diagonal, only the squares along and below the diagonal need to be displayed. This figure is based on Figure 1a from Ge *et al.* (2001).

6.3: YETI Genome vs Proteome Section

The Genome vs Proteome Section of YETI was used to perform the genome vs proteome correlation analysis. In this section, YETI displays a genome-proteome correlation map where the PID for each square in the two-dimensional interaction matrix is calculated, as described above, and represented with a colour gradient (Figure 6.2); bright greens represent higher PIDs whereas bright reds represent lower PIDs. This visual representation of the genome-proteome correlation map enables users to easily and rapidly compare the PIDs for all intracluster squares with intercluster squares to investigate a potential correlation between genome location and protein interaction. Furthermore, it also enables users to examine the PIDs of all

the squares in the map individually to investigate if there are any specific intracluster or intercluster squares that have substantially higher PIDs than the other squares in the map.

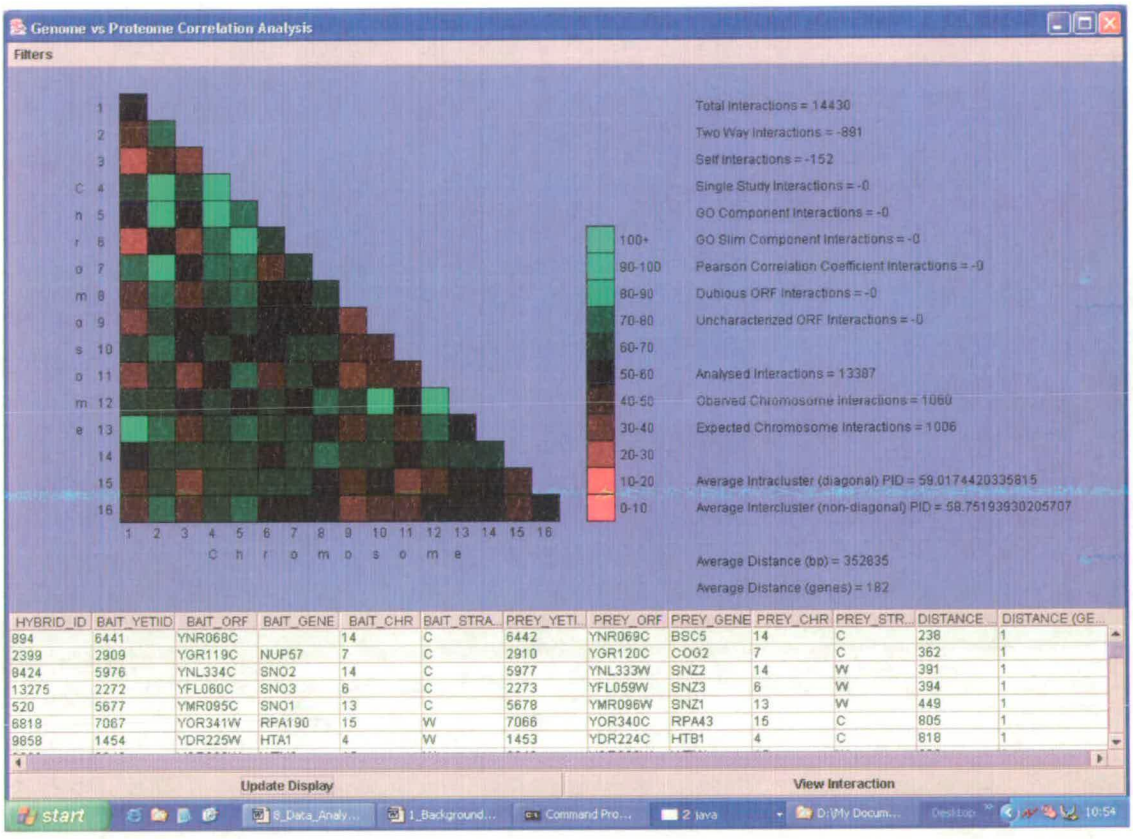


Figure 6.2: Screenshot of the Genome vs Proteome Section of YETI
This is a screenshot of the Genome vs Proteome Section of YETI. This section displays a genome-proteome correlation map where the PID for each square in the two-dimensional matrix is calculated and represented by a colour gradient. Furthermore, a variety of textual information and a data table is displayed along with the map (see text below for more details).

The Genome vs Proteome Section also has eight filters that can be used to filter the proteome dataset to remove specific types of interactions and therefore give a higher quality dataset; any combinations of the following eight filters can be used:

1) **Two-Way Interactions:** this filter can be used to remove two-way interactions from the proteome dataset. Two-way interactions are where two interactions are not technically duplicates but are essentially the same interaction. For example, consider the two reactions: A-B & B-A. Although, they are essentially the same interaction they are different because in the first interaction protein A was used as the 'bait' whilst in the second interaction protein B was used as the 'bait'. It was decided to leave these duplicate interactions in the YETI database as some researchers are interested in the 'direction' of interactions. This filter removes one of each two-way interaction as they are duplicated from an analytical viewpoint and therefore would bias the genome-proteome correlation results. 891 of the 14,430 protein-protein interactions stored in the YETI database are removed by this filter.

2) **Self Interactions:** this filter can be used to remove self-interactions from the proteome data set. Self-interactions are where an interaction is comprised of a protein interacting with another molecule of itself (it is both the 'bait' and 'prey' protein). This filter removes all the self-interactions as they would bias the genome-proteome correlation results. 152 of the 14,430 protein-protein interactions stored in the YETI database are removed by this filter.

3) **Single Study Interactions:** this filter can be used to remove all interactions that have only been reported in a single experimental study. Interactions that have been reported in more than one experimental study can more confidently be assumed to be true interactions than those reported from only a single study (von Mering *et al.*, 2002; Uetz *et al.*, 2005). 8,605 of the 14,430

protein-protein interactions stored in the YETI database are removed by this filter.

4) **GO Component**: this filter can be used to remove all interactions where the interacting proteins are not located in the same cellular compartment as defined by their GO component annotations (all the GO component annotations of the interacting proteins are compared not just the primary annotations). Interactions where the interacting proteins are not located in the same cellular compartment are less likely to be true interactions as in real life the proteins may never actually meet to interact. Furthermore, this filter also removes all protein-protein interactions involving any protein whose GO component annotation is currently unknown. 9,961 of the 14,430 protein-protein interactions stored in the YETI database are removed by this filter.
5) **GO Slim Component**: this filter can be used to remove all interactions where the interacting proteins are not located in the same cellular compartment as defined by their GO Slim component annotations and also interactions involving proteins whose GO Slim component annotation is currently unknown. GO Slim annotations are a cut-down version of the standard GO annotations meaning that proteins are assigned to broader high level terms rather than specific fine grained terms. 8,643 of the 14,430 protein-protein interactions stored in the YETI database are removed by this filter.
6) **Pearson Correlation Coefficient:** this filter can be used to remove all interactions where the corresponding ORFs of the interacting proteins are not coexpressed. Proteins that interact with one-another will not physically be able to do so if they are not both present in the cell at the same time. In this

filter, whether or not two proteins are coexpressed is defined by the Pearson correlation coefficient of the two corresponding ORFs as calculated from their corresponding expression data from the Gasch *et al.* (2000) study. For this filter, the user enters a minimum Pearson correlation coefficient value and all interactions below this cutoff value are removed. A standard cutoff used in microarray experiments is a Pearson correlation coefficient of 0.7 and 13,581 of the 14,430 protein-protein interactions stored in the YETI database are removed by this filter at this cutoff value.

7) **Dubious ORFs:** this filter can be used to remove all interactions involving proteins whose corresponding ORFs are 'dubious' and are therefore unlikely to be real ORFs. All ORFs are now defined by the *Saccharomyces* Genome Database (SGD; Cherry *et al.*, 1998) as dubious, uncharacterised or verified. 563 of the 14,430 protein-protein interactions stored in the YETI database are removed by this filter.
8) **Uncharacterised ORFs:** this filter can be used to remove all interactions involving proteins whose corresponding ORFs are 'uncharacterised', as defined by the SGD. 2,405 of the 14,430 protein-protein interactions stored in the YETI database are removed by this filter.

In addition to the actual genome-proteome correlation map, YETI also calculates and displays the average intracluster and intercluster PID, the total number of interactions analysed, the number of interactions removed by each filter, the number of expected and observed interactions where the corresponding ORFs of the interacting proteins are located on the same chromosome as well as the average distance in both base

pairs and genes between these ORFs. This YETI section also displays a data table containing information about all the protein-protein interactions found where the corresponding ORFs of the interacting proteins are located on the same chromosome (after filtering) is displayed and linked to the Analysis Section. This table enables all the identified interactions to be rapidly examined and also enables any interactions of interest to be selected and investigated in further detail in the other sections of YETI. Furthermore, any of the squares in the matrix can be individually selected to view information on all the interactions currently assigned to that square in the Analysis Section. This enables users to investigate any specific square that may be of interest in the matrix such as an intracluster or intercluster square that has a very high PID value when compared to the rest of the matrix.

6.4: Correlation Analysis Results

The genome dataset used in this analysis consisted of the 6,563 ORFs stored in the YETI database that are located on the 16 nuclear chromosomes of *S. cerevisiae*; 809 of these ORFs are characterised by the SGD as dubious and 1,468 as uncharacterised. The real proteome dataset consisted of the 14,430 protein-protein interactions stored in the YETI database that the above 6,563 ORFs are involved in. As a negative control, a random proteome dataset was generated through the creation of 14,430 random protein-protein interactions between the 6,563 ORFs of the genome dataset.

The Genome vs Proteome Section of YETI was used to analyse both the real and random proteome datasets described above against the genome dataset. The genome-

proteome correlation map for the unfiltered real proteome dataset is shown in Figure 6.3A. As can clearly be seen, there is a high-density region along the diagonal intracluster squares illustrated by the large number of bright green squares indicating high PIDs; although, there are also a few bright green non-diagonal squares in the map. Furthermore, the average intracluster PID is substantially above the average intercluster PID (79.92 vs 62.62). Taken together, this could lead one to suggest a possible global correlation between genome location and protein interaction. However, as mentioned above this is the unfiltered dataset which therefore still has self and two-way interactions which bias the correlation results. Applying the filters to remove these interactions gives completely different results as shown in Figure 6.3B. After filtering, the high density region along the diagonal is no longer apparent and the average intracluster PID is now only very slightly above the average intercluster PID (59.02 vs 58.76). Therefore, these initial results suggest that there is no global correlation between genome location and protein interaction.

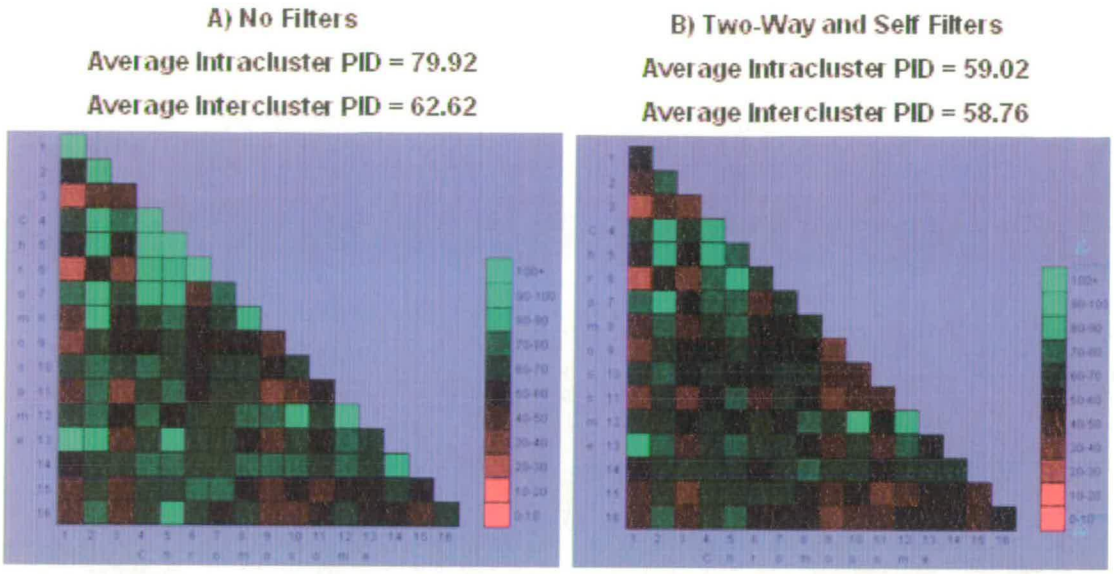


Figure 6.3: Genome-Proteome Correlation Maps
This figure contains the YETI generated genome-proteome correlation maps for the unfiltered real proteome dataset (A) and the real proteome dataset filtered for self and two-way interactions (B).

In addition to the genome-proteome correlation maps described above, YETI can be used to examine a potential correlation between the genome and proteome in more detail. YETI displays the expected and observed numbers of protein-protein interactions where both interacting proteins are located on the same chromosome. The expected number of interactions is calculated by multiplying the number of analysed interactions (after filtering) by the probability that two interacting proteins will be located on the same chromosome (Figure 6.4). Furthermore, YETI calculates the average distance between all interacting proteins located on the same chromosome in both base pairs and genes.

$$P = \sum_{c=1}^{C} \left[\frac{n_c}{t} x \frac{n_c - 1}{t - 1} \right] = 0.075306604$$

Figure 6.4: Probability that any two interacting proteins are located on the same chromosome
This figure shows the equation used to calculate the probability that any two interacting proteins will be located on the same nuclear chromosome of *S. cerevisiae*; c = the total number of chromosome clusters; n_c = the number of ORFs in chromosome cluster c; t = the total number of ORFs in all chromosome clusters. In this case: C = 16; n_1 = 117, n_2 = 454, n_3 = 182 ... n_{16} = 509; t = 6563. It is important to note that the probability changes depending on what proteome filters are selected. For example, if the self interactions filter is not selected then the '-1' components are removed from the above equation or if the dubious filter is selected all the n_c and t values are modified accordingly.

Both the real and random proteome datasets were analysed with various combinations of the YETI filters and a comprehensive account of the results is presented in Table 6.1 for the real dataset and Table 6.2 for the random dataset. Amongst others, this table contains the total number of interactions analysed after any filtering and the number of observed interactions where the interacting proteins are located on the same chromosome. To test whether the observed number of interactions located on the same chromosome is statistically significant the probability for obtaining at least the observed number of interactions by chance was calculated using the standard cumulative binomial distribution (http://mathworld.wolfram.com/BinomialDistribution.html; Figure 6.5).

$$P(i >= i_0) = \sum_{i=i_0}^{I} p^i (1-p)^{I-i} \left[\frac{I!}{I!(I-i)!} \right]$$

Figure 6.5: Cumulative binomial distribution
This figure shows the cumulative binomial distribution equation used to calculate the probability of obtaining at least the observed number of interactions where both interacting proteins are located on the same chromosome by chance. In this case: I = the total number of interactions analysed; i_0 = the observed number of interactions where both interacting proteins are located on the same chromosome; and p = the probability of two interacting proteins being located on the same chromosome (Figure 6.4).

Real Data Set										
Filters	**Ints**	**Exp**	**Obs**	**P-Value**	**Intra**	**Inter**	**Dist (bp)**	**StDev (bp)**	**Dist (genes)**	**StDev (genes)**
None	14430	1086	1289	2.77E-10	79.92	62.62	312034	290927	161	148
Two-Way & Self	13387	1006	1060	0.041233	59.02	58.75	352835	283883	182	144
Two-Way, Self & Dubious	12866	972	1016	0.076722	73.89	74.53	349156	282982	180	144
Two-Way, Self, Dubious & Single Study	5055	382	394	0.270717	23.85	28.24	359508	285594	185	145
Two-Way, Self, Dubious & GO Slim Component	5238	395	405	0.325266	26.78	29.69	356750	288773	183	147
Two-Way, Self, Dubious & GO Component	3917	296	298	0.462671	21.58	22.72	342358	280195	176	142
Two-Way, Self, Dubious & Pearson = 0.7	651	49	36	0.982538	4.16	4.22	263676	218901	137	113
Two-Way, Self, Dubious & Uncharacterised	10621	813	834	0.229779	112.20	113.58	349636	283199	180	144
Two-Way, Self & Dubious	12866	972	1016	0.076722	73.89	74.53	349156	282982	180	144
(+) Single Study	5055	382	394	0.270717	23.85	28.24	359508	285594	185	145
(+) GO Slim Component	1646	124	127	0.418258	6.51	8.85	372197	299269	191	151
(+) GO Component	1124	84	96	0.118409	5.19	6.16	369432	299681	189	152
(+) Pearson Correlation Coefficient = 0.7	85	6	4	0.892333	0.32	0.42	130058	92759	71	51

Table 6.1: Genome vs Proteome Correlation Analysis Results for the Real Proteome Dataset
This table contains the results of the Genome vs Proteome correlation analysis for the real proteome dataset performed using YETI. Ints represents the total number of protein-protein interactions analysed after any filtering; Exp and Obs represent the number of expected and observed protein-protein interactions where the interacting proteins corresponding ORFs are located on the same chromosome, respectively; P-Value represents the probability of getting at least the observed number of interactions by chance calculated using the cumulative binomial distribution; Intra represents the average intracluster PID; Inter represents the average intercluster PID; Dist (bp) represents the average distance in base pairs between interacting proteins located on the same chromosome; StDev (bp) represents the standard deviation of the average distances in base pairs; Dis (genes) represents the average distance in genes between interacting proteins located on the same chromosome; and StDev (genes) represents the standard deviation of the average distances in genes. Details on the calculation of the P-Value and a discussion of the results can be found in the text above.

Random Data Set										
Filters	Ints	Exp	Obs	P-Value	Intra	Inter	Dist (bp)	StDev (bp)	Dist (genes)	StDev (genes)
None	14430	1086	1114	0.198348	66.55	67.59	348957	283078	180	144
Two-Way & Self	14428	1084	1113	0.187967	66.91	67.58	349271	283011	181	144
Two-Way, Self & Dubious	11097	838	870	0.135613	67.93	67.88	341504	280364	177	143
Two-Way, Self, Dubious & GO Slim Component	1932	146	168	0.034415	12.42	11.06	309168	296562	160	151
Two-Way, Self, Dubious & GO Component	1151	87	93	0.266798	7.17	6.54	315580	290812	164	148
Two-Way, Self, Dubious & Pearson = 0.7	76	5	7	0.351082	0.46	0.43	334887	326838	171	163
Two-Way, Self, Dubious & Uncharacterised	6238	477	492	0.253905	65.81	69.34	343606	287904	178	147
Two-Way, Self & Dubious	11097	838	870	0.135613	67.93	67.88	341504	280364	177	143
(+) GO Slim Component	1932	146	168	0.034415	12.42	11.06	309168	296562	160	151
(+) GO Component	1133	85	92	0.252358	7.09	6.45	316371	292288	165	148
(+) Pearson Correlation Coefficient = 0.7	13	0	3	0.06969	0.12	0.05	561569	384804	282	194

Table 6.2: Genome vs Proteome Correlation Analysis Results for the Random Proteome Dataset
This table contains the results of the Genome vs Proteome correlation analysis for the random proteome dataset performed using YETI. Ints represents the total number of protein-protein interactions analysed after any filtering; Exp and Obs represent the number of expected and observed protein-protein interactions where the interacting proteins corresponding ORFs are located on the same chromosome, respectively; P-Value represents the probability of getting at least the observed number of interactions by chance calculated using the cumulative binomial distribution; Intra represents the average intracluster PID; Inter represents the average intercluster PID; Dist (bp) represents the average distance in base pairs between interacting proteins located on the same chromosome; StDev (bp) represents the standard deviation of the average distances in base pairs; Dis (genes) represents the average distance in genes between interacting proteins located on the same chromosome; and StDev (genes) represents the standard deviation of the average distances in genes. Details on the calculation of the P-Value and a discussion of the results can be found in the text above. The random dataset could not be subjected to the Single Study filter because it was randomly not experimentally generated.

The results of the analysis of the real proteome dataset are contained in Table 6.1; the unfiltered dataset should be discounted as it contains two-way and self interactions which bias the results and datasets containing dubious ORFs should also be discounted as these ORFs and therefore their interactions are very unlikely to be real. As can be seen in Table 6.1, the observed number of interactions where both interacting proteins are located on the same chromosome is nearly always above the expected number for all the filters. However, in each case the observed number is only slight above the expected number and the P-value is always above the standard cut-off of 0.05 suggesting that the observed numbers are not statistically significant. Furthermore, in each case the average intracluster and intercluster PIDs are always similar and there is no apparent trend for one being consistently higher than the other. It is important to note that although the average intracluster and intercluster PID values from the same filter can be readily compared to each other, the average PIDs obtained from different filters can not really be compared to one another. This is because the PID is calculated as the observed number of interactions for a cluster divided by the total number of possible interactions for a cluster. However, the self, dubious and uncharacterised filters change the number of possible interactions for a cluster which therefore means that the average PIDs can only really be compared within as opposed to across filters.

Identifying the interactions where both interacting proteins are located on the same chromosome is only the first step in this analysis. The fact that two interacting proteins are located on the same chromosome could mean little if they are at opposite ends. As can be seen in Table 6.1, the average distance between interacting proteins

located on the same chromosome in both base pairs and genes is very large in every case; the average distances are typically above 300,000 bp and 175 genes. Generally, the random dataset (Table 6.2) gave similar results to the real dataset with similar intracluster vs intercluster PID values, large average distances and statistically insignificant numbers of observed interactions. Therefore, altogether, these results suggest that there is no global correlation between genome location and protein interaction in *S. cerevisiae*.

However, there are still a number of interesting observations that can be made from the analysis results. Firstly, the GO component filter removes approximately 90 % of the interactions from the random dataset whereas this filter only removes approximately 70 % of the interactions from the real data set. The fact that approximately 30 % of the real protein-protein interactions share the same known GO Component annotation compared with only 10 % of the random interactions suggests that these interactions have a higher confidence of being true interactions. Furthermore, this also suggests that the GO component filter is a good filter to achieve a higher quality dataset; this level of filtration has also been suggested in Sprinzak *et al.* (2003), for example. Secondly, although overall there does not appear to be a correlation, the observed number of interactions were nearly always above the expected number for the real dataset which could suggest that there are a small number of relevant individual cases of interacting proteins being colocated.

6.5: Closest Interacting Proteins

As described above, the Genome vs Proteome Section of YETI includes a data table containing information on all the protein-protein interactions found where the corresponding genes of interacting proteins are located on the same chromosome (after any filtering). Furthermore, the interactions are ordered by the distance between interacting proteins and the table is directly linked to the Analysis Section. Therefore, this table enables users to rapidly examine and compare all of the interactions found and select any interactions of interest to investigate further in the other sections of YETI. The closest interactions found are presented in Table 6.3 which contains information on all the protein-protein interactions whose corresponding genes are located on the same chromosome and within 10,000 bp of each other.

	BAIT			PREY			DISTANCE		
No	ORF	GENE	CHR	ORF	GENE	CHR	BP	GENES	PCC
1	YNR068C		14	YNR069C	BSC5	14	238	1	0.65
2	YGR119C	NUP57	7	YGR120C	COG2	7	362	1	0.67
3	YNL333W	SNZ2	14	YNL334C	SNO2	14	391	1	0.37
4	YFL060C	SNO3	6	YFL059W	SNZ3	6	394	1	0.70
5	YMR095C	SNO1	13	YMR096W	SNZ1	13	449	1	0.64
6	YOR341W	RPA190	15	YOR340C	RPA43	15	805	1	0.58
7	YDR225W	HTA1	4	YDR224C	HTB1	4	818	1	0.89
8	YOR229W	WTM2	15	YOR230W	WTM1	15	988	1	0.01
9	YPL026C	SKS1	16	YPL028W	ERG10	16	1383	2	-0.43
10	YIL035C	CKA1	9	YIL033C	BCY1	9	1511	2	-0.16
11	YLR288C	MEC3	12	YLR290C		12	2241	2	0.35
12	YCR088W	ABP1	3	YCR084C	TUP1	3	2616	4	-0.21
13	YMR308C	PSE1	13	YMR310C		13	3495	2	0.60
14	YMR106C	YKU80	13	YMR108W	ILV2	13	3894	2	-0.14
15	YER081W	SER3	5	YER078C		5	4344	3	-0.04
16	YLR328W	NMA1	12	YLR332W	MID2	12	4558	4	0.23
17	YER022W	SRB4	5	YER019W	ISC1	5	4582	3	-0.24
18	YGR177C	ATF2	7	YGR172C	YIP1	7	5238	5	0.39
19	YNR046W	TRM112	14	YNR050C	LYS9	14	5853	4	0.12
20	YPR182W	SMX3	16	YPR178W	PRP4	16	6465	4	0.20
21	YLR319C	BUD6	12	YLR313C	SPH1	12	6976	6	0.14
22	YDR386W	MUS81	4	YDR381W	YRA1	4	8080	5	0.00
23	YMR308C	PSE1	13	YMR314W	PRE5	13	9488	6	0.05
24	YOR239W	ABP140	15	YOR232W	MGE1	15	9598	7	0.50
25	YGR095C	RRP46	7	YGR090W	UTP22	7	9600	5	0.26
26	YLR453C	RIF2	12	YLR449W	FPR4	12	9789	4	-0.37
27	YNL090W	RHO2	14	YNL085W	MKT1	14	9988	5	0.45

Table 6.3: The Closest Interacting Proteins
This table contains information on all the protein-protein interactions where the interacting proteins corresponding ORFs are located on the same chromosome and within 10,000 bp; the interactions have been filtered for two-way interactions, self interactions and dubious ORFs. The ORF name (ORF), gene name (GENE) and chromosome (CHR) is displayed for both the BAIT and PREY proteins of the interaction and the distance between them is shown in both base pairs (BP) and genes (GENES); in this analysis, two neighbouring genes have a gene distance of 1 not 0. Furthermore, the Pearson correlation coefficient (PCC) of the two interacting protein's corresponding genes is also displayed.

As discussed above, no overall correlation was observed between genome location and protein interaction. However, as can be seen in Table 6.3 there are a small number of cases of interacting proteins being located right next to each other on their corresponding chromosome; neighbouring genes have a gene distance of 1 in Table 6.3. Furthermore, these neighbouring genes are typically involved in the same

specific biological process suggesting that there is a functional reason for their colocation such as co-regulation through shared promoter regions. For example, interaction 8 involves WTM2 and WTM1 which are both involved in the GO biological process of 'regulation of meiosis', interaction 7 involves HTA1 and HTB1 which are both histones involved in 'chromatin assembly or disassembly', and interaction 6 involves RPA190 and RPA43 which are both RNA polymerase I subunits. In addition, there are three interactions (3, 4 and 5) involving neighbouring genes of the SNZ and SNO gene families which are all involved in 'thiamin biosynthesis; these interactions and proteins are discussed in further detail below in section 6.6: Thiamin Biosynthesis.

There are also a number of cases of interacting proteins located near each other on a chromosome and also involved in the same specific biological process. For example, interaction 20 involves SMX3 and PRP4 which are both involved in the GO biological process of 'nuclear mRNA splicing, via spliceosome', interaction 21 involves BUD6 and SPH1 which are both involved in 'actin filament organization', and interaction 25 involves RRP46 and UTP22 which are both involved in '35S primary transcript processing'. However, this does lead to the inevitable question of how 'near' do two interacting genes have to be for their colocation to be significant. Although there is no clear answer to this question, one consideration would be what the functions and expressions of the separating genes are. For example, if a pair of genes whose products interact with one another are separated by three other genes and all five genes are involved in the same or related biological processes and were coexpressed, then this would suggest that this colocation is relevant. In the each of

the three examples described above the separating genes were involved in a range of biological processes and the interacting genes themselves were not significantly coexpressed which suggests that these observed interactions could just be random occurrences of close genes whose products interacts.

6.6: Thiamin Biosynthesis

As can be seen in Table 6.3, three of the closest interactions involve members of the SNZ and SNO gene families; specifically, interactions (3) SNZ2-SNO2, (4) SNO3-SNZ3 and (5) SNO1-SNZ1. As the members of these two families appear to be colocated across the genome and directly interact with one another, they were investigated in further detail using YETI. YETI shows that there are three members of the SNZ gene family (SNZ1, SNZ2 and SNZ3) each of which has an SNO gene (SNO1, SNO2 and SNO3, respectively) next to it (Figure 6.6). In each case, the SNO gene is located directly upstream on the opposite strand of DNA to the SNZ gene; therefore, each SNZ/SNO gene pair is divergent which suggests they could well share the same promoter region and could be regulated by the same factors. Furthermore, given the conserved colocation of the members of these two gene families, all the SNZ/SNO gene pairs could well be collectively regulated by the same factors.

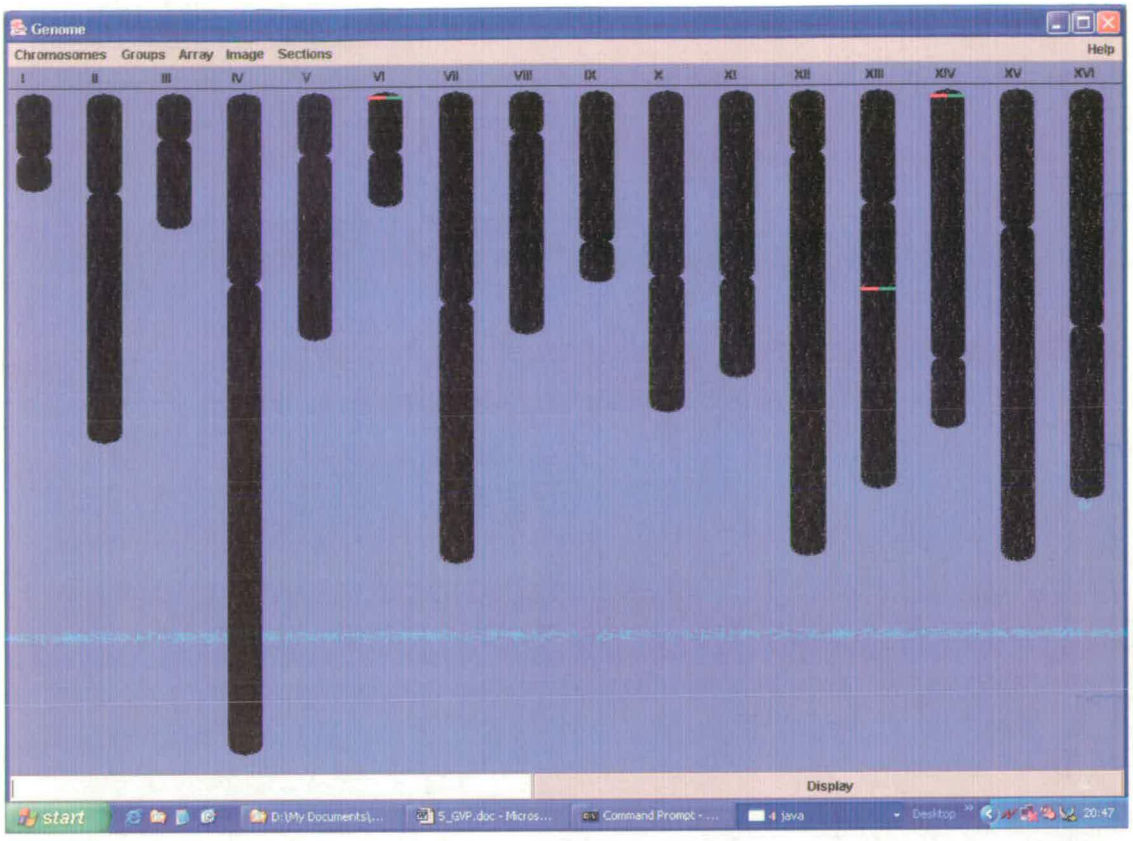


Figure 6.6: Screenshot of the Genome Section highlighting the genomic location of the SNZ/SNO gene pairs
This is a screenshot of the Genome Section of YETI with the genomic location of the three SNZ/SNO gene pairs highlighted on the genome schematic. The three SNZ genes are highlighted on the genome schematic with red lines and the three SNO genes with green lines. As can clearly be seen each SNZ gene is colocated on the genome with an SNO gene. This example highlights the potential of the genome schematic to investigate possible evolutionary relationships between two groups of genes.

Using the Chromosome Window of YETI to examine the chromosomal regions of the three SNZ/SNO gene pairs suggests that the chromosomal regions containing the SNZ2/SNO2 and SNZ3/SNO3 gene pairs are duplicated (Figure 6.7). Both pairs are located at the left arm telomere of their respective chromosomes with a THI gene followed by an AAD gene downstream of the SNZ gene and a gene of unknown function followed by a COS gene upstream of the SNO gene. Further examination of these two regions reveals that all the genes (except the AAD genes) are exactly the same length which further suggests that these two regions are duplicated. This DNA

duplication implies that the SNZ2/SNO2 and SNZ3/SNO3 gene pairs are coregulated and that they encode the same protein products. However, the chromosomal region of the SNZ1/SNO1 gene pair does not show any similarity to the other two SNZ/SNO regions; it is located in the middle of the right arm of chromosome 13 and does not contain any members of the THI, AAD or COS gene families. Furthermore, the SNZ1/SNO1 genes are slightly different in length to the SNZ2/SNO2 and SNZ3/SNO3 genes. Therefore, whether or not the SNZ1/SNO1 genes are coregulated with the SNZ2/SNO2 and SNZ3/SNO3 genes and also encode the same protein products is unclear (at this point).

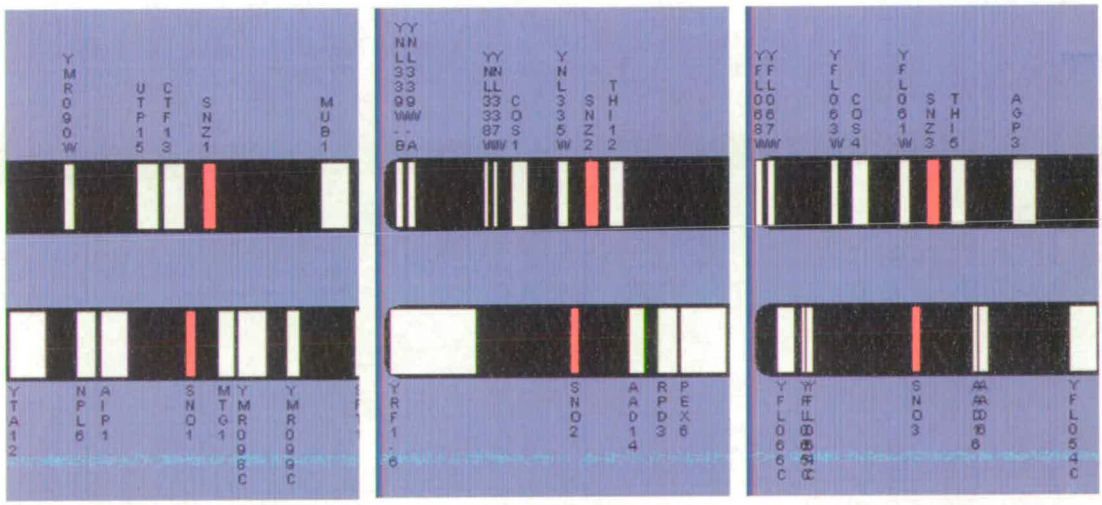


Figure 6.7: Chromosomal regions of the three SNZ/SNO gene pairs
This figure shows the chromosomal regions of the three SNZ/SNO gene pairs generated from the Chromosome Window of YETI. In each case, the SNZ and SNO genes are highlighted in red. As can clearly be seen, the chromosomal regions surrounding SNZ2/SNO2 and SNZ3/SNO3 are strikingly similar.

The three SNZ and SNO genes are all characterised as being involved in the GO biological processes of 'pyridoxine metabolism' and 'thiamin biosynthesis', therefore, YETI was used to collectively investigate all of the proteins involved in these biological processes further; pyridoxine (vitamin B_6) is a coenzyme for enzymes involved in amino acid metabolism whereas thiamin (vitamin B_1) functions as the co-enzyme thiamin pyrophosphate (TPP) in the metabolism of carbohydrates and branched-chain amino acids. YETI shows that there is an additional SNO gene in the *S. cerevisiae* genome, namely SNO4 located in the right arm telomere on chromosome 13, which is also involved in pyridoxine metabolism; however, this gene is not a true SNO gene as it is not located upstream of an SNZ gene. YETI shows that a number of other genes are characterised as being involved in thiamin biosynthesis; specifically: THI2, THI3, THI4, THI5, THI6, THI11, THI12, THI13, THI20, THI21, THI22, PDC2 and RPI1. Interestingly, THI5 and THI12 are also colocated with SNZ3 and SNZ2, respectively, as shown previously in Figure 6.7. Highlighting the genomic location of all the thiamin biosynthesis genes on the genome schematic of YETI reveals that two additional THI genes (namely, THI13 and THI11) are also located in telomeric regions. Analysing these two regions further in the Chromosome Window reveals that they also appear to be duplicated with each region consisting of a COS (unknown function), MPH (α-glucoside permease), SOR (sorbitol dehydrogenase), HXT (hexose transporter), THI (thiamin biosynthesis) and an AAD (aryl-alcohol dehydrogenase) gene (Figure 6.8). Furthermore, these two regions are also similar to the SNZ2/SNO2 and SNZ3/SNO3 telomeric regions discussed above which also span from a COS gene to a THI and AAD gene.

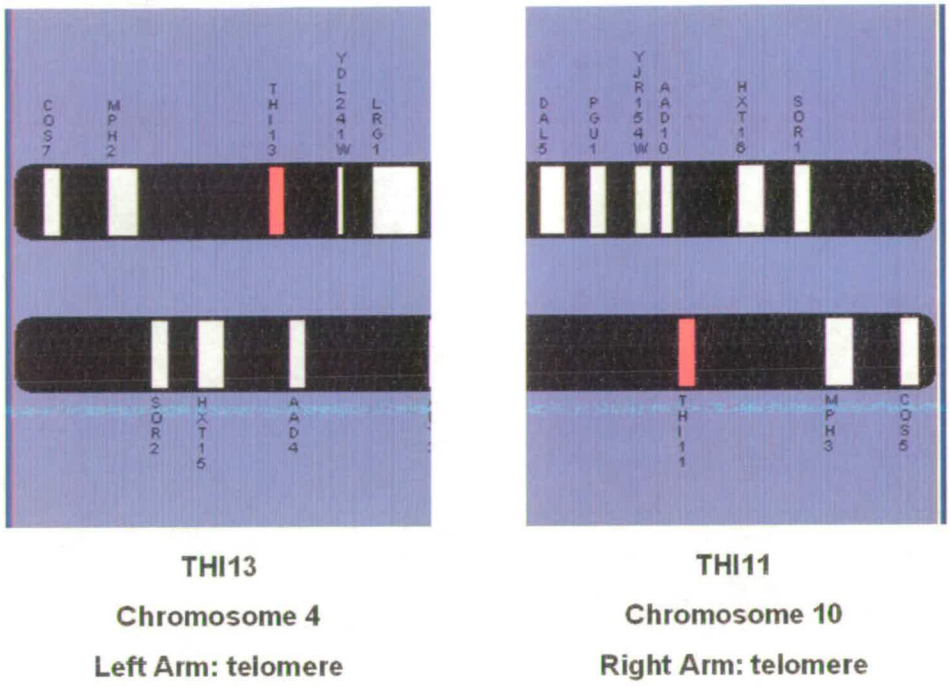


Figure 6.8: Chromosomal regions of THI13 and THI11
This figure shows the chromosomal regions of the THI13 and THI11 genes generated from the Chromosome Window of YETI; in each case, the THI gene is highlighted in red. As can clearly be seen, the two regions are strikingly similar as both consist of a COS, MPH, SOR, HXT, THI and AAD genes. One slight difference is that THI13 is located at a left arm telomere whereas THI11 is located at a right arm telomere; this is why the THI11 region is 'upside down' when compared to the THI13 region.

Overall, YETI has highlighted four similar telomeric regions in the *S. cerevisiae* genome containing genes involved in thiamin biosynthesis; each region consists of seven genes, starting with a COS gene, ending with a THI and AAD gene with three genes in between. These four regions can be split into two equal groups: (1) The two members of the first group have a gene of unknown function followed by an SNO and SNZ gene located in between the COS and THI genes; and (2) The two members of the second group have an MPH, SOR and HXT gene located in between the COS and THI gene. However, it is not yet clear how these duplicated regions arose and what the functional relevance of these duplicated regions is; for example, a whole

genome duplication event could account for the duplication of each group individually but would not account for the similarity between the two groups.

The Transcriptome Section of YETI shows that a number of the genes involved in thiamin biosynthesis are colocated in the hierarchical tree (Figure 6.9). Seven of the thiamin biosynthesis genes are located right next to each other in the tree forming a tight cluster with three additional thiamin biosynthesis genes located in the vicinity. The seven genes in the tight cluster are THI12, THI5, THI3, THI11, SNO3, SNZ2 and SNZ3 and the three additional genes are RPI1, SNO2 and THI6. Interestingly, all of the thiamin biosynthesis genes that are located in the duplicated chromosomal regions discussed above are located in this region of the gene expression hierarchical tree suggesting they are all coregulated; specifically: SNZ3, SNO3 and THI5 from chromosome 6; SNZ2, SNO2 and THI12 from chromosome 14; THI13 from chromosome 4; and THI11 from chromosome 10. The other thiamin biosynthesis genes, including SNZ1 and SNO1 are dispersed through the hierarchical tree. These observations suggest that the SNZ2/SNO2 and SNZ3/SNO3 gene pairs are indeed coregulated with each other, along with a number of other genes involved in thiamin biosynthesis, but not with the SNZ1/SNO1 gene pair. This could suggest that the SNZ1/SNO1 gene pair is not actually involved in thiamin biosynthesis or that this gene pair is regulated by different factors to the other two SNZ/SNO regions.

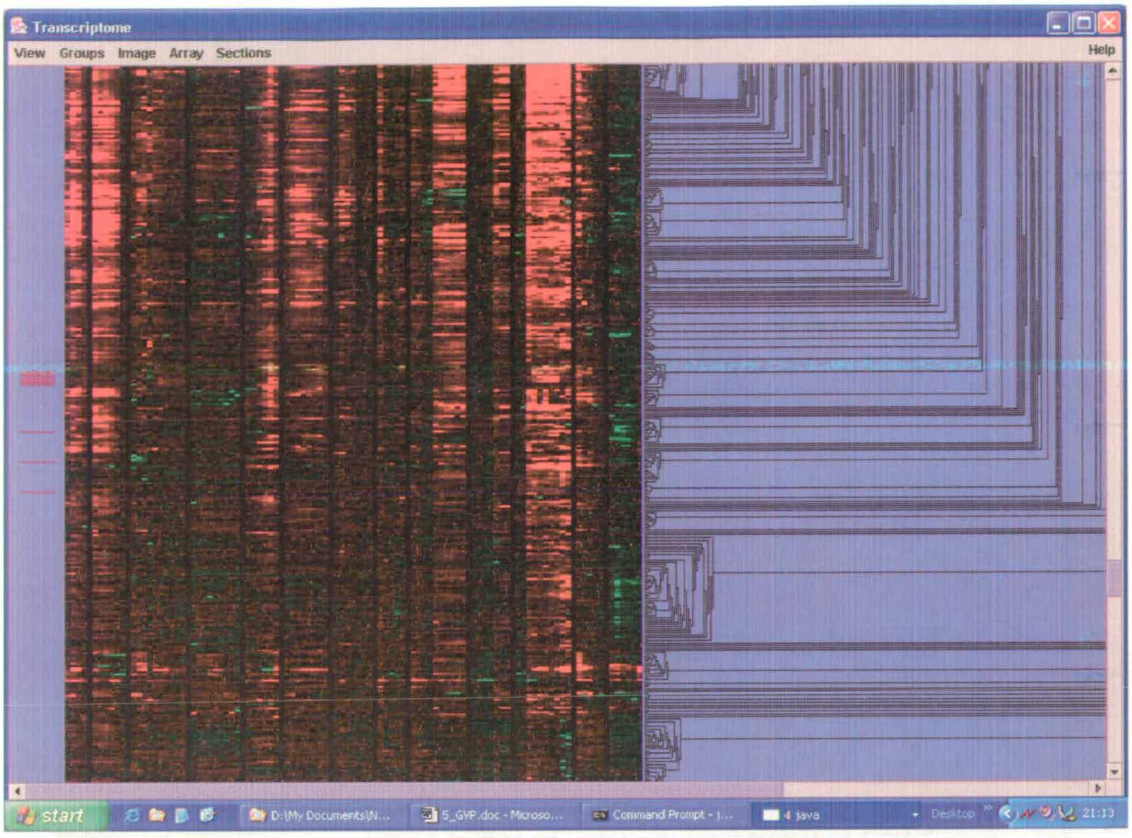


Figure 6.9: Gene expression cluster of thiamin biosynthesis genes
This is a screenshot of the Transcriptome Section with the location of all the genes involved in the GO biological process of thiamin biosynthesis highlighted on the gene expression hierarchical tree. As can be seen, a number of genes involved in thiamin biosynthesis are colocated forming a tight expression cluster.

The Proteome Section of YETI shows that all the SNZ and SNO proteins interact highly with one another (Figure 6.10). This suggests that all the SNZ and SNO proteins could work together in order to achieve their biological goal. However, if the proteins are not expressed at the same time or are located in different cellular compartments then some of these interactions could well be irrelevant. Indeed, the results from the Transcriptome Section suggest that the SNZ1/SNO1 gene pair may well be expressed at different times to the SNZ2/SNO2 and SNZ3/SNO3 gene pairs. The Proteome Section also shows that none of the SNZ/SNO proteins interact directly with any of the other proteins involved in thiamin biosynthesis; in actual fact, none of the other proteins involved in thiamin biosynthesis appear to interact

directly with one another. This is slightly surprising given their colocation, coexpression and the fact that they are all involved in the same biological process. However, this observation could be explained by an incomplete protein interaction data set or by the fact that some proteins do not need to interact with other proteins in order to fulfil their biological goal.

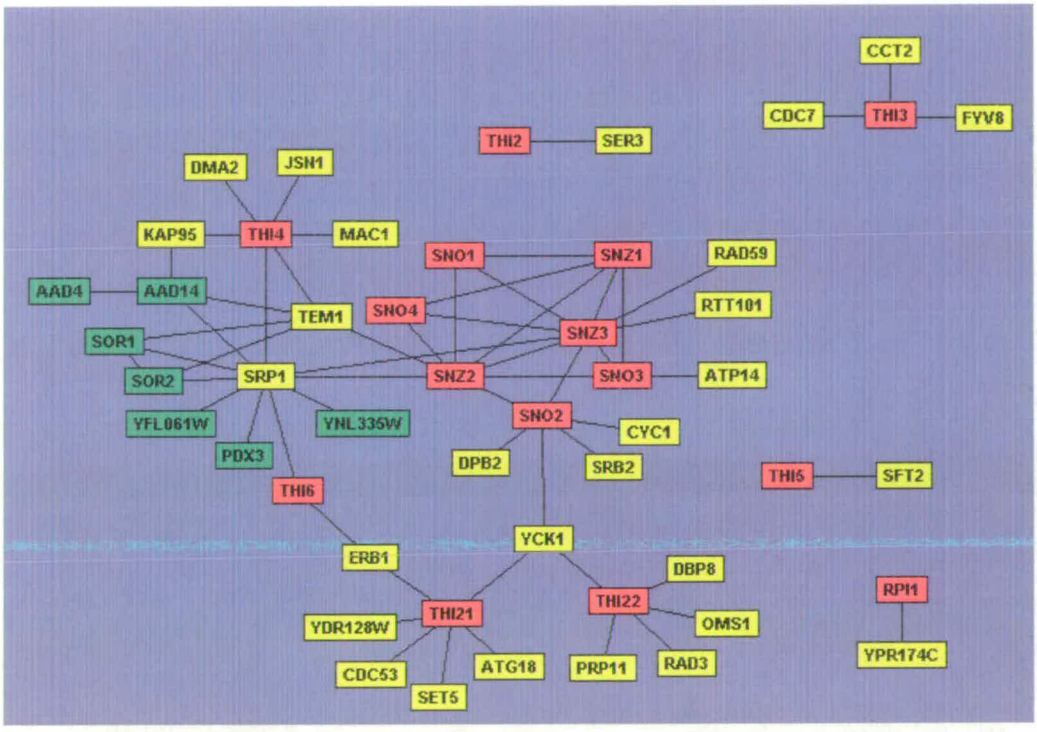


Figure 6.10: Protein interactions involving thiamin biosynthesis proteins
This is an image of all the protein interactions involving any of the proteins involved in thiamin biosynthesis created through the Proteome Section of YETI. All of the proteins involved in thiamin biosynthesis are highlighted in red; as can be seen, all of the SNO and SNZ proteins interact highly with one another. The non-thiamin biosynthesis proteins whose corresponding genes are located in the duplicated chromosomal regions discussed above are highlighted in green; as can be seen, practically all of these proteins interact directly with SRP1. An additional related gene called PDX3 which interacts with SRP1 is also highlighted in green.

Interestingly, both SNZ2 and SNZ3 interact with a protein called SRP1 which is involved in the import of nuclear proteins. SRP1 interacts with a large number of proteins including: YFL061W which is colocated on chromosome 6 with

SNZ3/SNO3; YNL335W which is colocated on chromosome 14 with SNZ2/SNO2; PDX3 which is a pyridoxine phosphate oxidase; THI4 and THI6 which are both involved in thiamin biosynthesis, AAD14 which is colocated on chromosome 14 with THI12; SOR1 which is colocated on chromosome 10 with THI11; and SOR2 which is colocated on chromosome 4 with THI13. It is interesting that a number of the proteins involved in thiamin biosynthesis, along with a number of proteins whose corresponding genes are located in the duplicated chromosomal regions discussed above, interact directly with SRP1. Although this does represent a common link between all of these proteins it does not necessarily imply that they are all involved in the same or related biological processes, especially as few of them interact directly with one another. However, as SRP1 is involved in the import of nuclear proteins, this could suggest possible cellular locations for these proteins i.e. the nucleus; interestingly, the majority of the cellular locations of the thiamin biosynthesis proteins are currently unknown. Furthermore, it is also interesting that whilst SNZ2 and SNZ3 do interact with SRP1, SNZ1 does not. This is interesting as it again suggests a distinction between the SNZ1/SNO1 gene pair and the SNZ2/SNO2 and SNZ3/SNO3 gene pairs; in the case by suggesting that SNZ1 may have a different cellular location to SNZ2 and SNZ3.

As described above, located directly upstream of the SNZ2/SNO2 and SNZ3/SNO3 gene pairs are two genes of unknown function (YFL061W and YNL335W); as these two genes are located in apparently duplicated blocks of DNA they should encode the same product and also be regulated by the same factors. Each of these genes has been duplicated along with an SNO, SNZ and a THI gene, all of which are involved

in thiamin biosynthesis. Therefore, this in itself strongly suggests that these two genes are also involved in thiamin biosynthesis or a related biological process. Furthermore, YETI shows that these two genes are coexpressed (pearson > 0.7) with SNZ2, SNZ3 and SNO4; in addition, both their corresponding proteins interact with SRP1 along with many of the proteins involved in thiamin biosynthesis and pyridoxine metabolism. However, whether or not these two genes of unknown function are directly involved in thiamin biosynthesis or pyridoxine metabolism can only be proven by experiments in the laboratory but the observations presented here suggest that they could well be.

At this point it is worth comparing the information found using YETI alone to what is currently known about the SNZ/SNO genes. SNZ1 was originally identified through studies of proteins synthesised in stationary phase *S. cerevisiae* cells, (Braun *et al.*, 1996). SNZ1 was found to be the most highly conserved protein present in all three domains, exhibiting 60 % identity with SNZ proteins in archea and bacteria (Braun *et al.*, 1996). Padilla *et al.* (1998) first identified the highly conserved SNZ gene family in *S. cerevisiae* and subsequently studied their sequence similarity, expression and phenotypes. Sequence analysis showed that SNZ2 was ~ 99 % identical to that of SNZ3 and ~ 80 % identical to that of SNZ1. Sequence analysis also showed that SNZ2 and SNZ3 were located within 7 kb telomeric regions that were nearly identical. Analysis of the sequence adjacent to the SNZ genes revealed an additional conserved, duplicated gene upstream of each SNZ gene which was subsequently called SNO (SNZ proximal ORF). Like their SNZ counterparts, SNO2 and SNO3 were found to be almost 100 % identical to each other and ~ 72 %

identical to SNO1. Using expression analysis Padilla *et al.* (1998) showed that adjacent SNZ/SNO genes were coregulated and that the SNZ1/SNO1 gene pair was induced at alternate times to the SNZ2/SNO2 and SNZ3/SNO3 gene pairs. Phenotypic analyses showed that SNZ1 was induced in an SNZ2/SNZ3 mutant at the times when SNZ2 and SNZ3 were normally induced which suggested that SNZ1 was repressed by expression of SNZ2 and SNZ3.

In order to clarify their physiological functions, Rodriguez-Navarro *et al.* (2002) further characterised the SNZ and SNO gene families. In this study, they demonstrated that SNZ1 and SNO1 were required for growth of *S. cerevisiae* in the presence of low levels of pyridoxine but that SNZ2, SNO2, SNZ3 and SNO3 were not. However, overexpression of SNZ2 or SNZ3 in SNZ1 mutants compensated for the observed growth defects suggesting that all the SNZ genes encode proteins with similar activities. Rodriguez-Navarro *et al.* (2002) also showed that the transcripts of SNZ2, SNO2, SNZ3 and SNO3 (but not SNZ1 and SNO1) accumulated in the absence of thiamin, along with THI5 and THI11 transcripts, which were known to be involved in thiamin biosynthesis. Furthermore, using the two-hybrid technique, SNZ2 and SNZ3 were found to directly interact with THI11 further associating them to thiamin biosynthesis. Overall, these results suggested that although all three SNZ genes encoded proteins with similar activities involved in the biosynthesis of pyridoxine, SNZ2 and SNZ3 were regulated by the same factors as thiamin biosynthesis genes directly linking them to this biological process as well.

The four duplicated telomeric regions containing the two SNZ/SNO gene pairs and the four THI genes (THI5, THI11, THI12 and THI13) highlighted through YETI were also highlighted by Wightman *et al.* (2003) who studied the function and redundancy of the THI5 gene family. The THI5 gene family of *S. cerevisiae* comprises four highly conserved members named THI5, THI11, THI12 and THI13 which are all homologues of the *Schizosaccharomyces pombe* nmt1 gene which functions in the biosynthesis of hydroxymethylpyrimidine (HMP). Interestingly, HMP is itself derived from pyridoxine which directly links the SNZ/SNO genes involved in pyridoxine metabolism to the THI genes involved in the biosynthesis of HMP; overall, this means that all the SNZ, SNO and THI genes are involved in the biosynthesis of thiamin. Phenotypic analyses of mutant strains showed that the four genes were functionally redundant in terms of HMP formation for thiamin biosynthesis; each gene product was found to be involved in the production of HMP from pyridoxine. However, comparative analysis of mRNA levels revealed subtle differences in the regulation of the four genes, suggesting that they respond differently to nutrient limitation. Wightman *et al.* (2003) proposed that the duplication of the SNZ and SNO genes may have been caused by a need to increase the production of pyridoxine for HMP production. Furthermore, the co-duplication of a member of the THI5 gene family with the SNZ/SNO genes and their coregulation ensured that this extra pyridoxine was channelled into thiamin biosyhthesis. However, the precise molecular functions of the SNZ and SNO genes is still not known but both Wightman *et al.* (2003) and Padilla *et al.* (1998) proposed that the SNZ and SNO are possible glutamine amidotransferases that produce phosphoribosylamine for pyridoxine and thiamin biosyntheis.

In addition, the SNZ/SNO gene pairs have previously been highlighted in other analyses investigating correlations between different functional genomic data sets (Cohen *et al.*, 2000; Ge *et al.*, 2001; Cornell *et al.*, 2001). By integrating transcriptome and interactome data, Ge *et al.* (2001) showed that although the SNZ and SNO proteins all interact highly with one another, their expression patterns suggested that they function in two distinct groups. By relating regulatory sequences to protein-protein interactions, Cornell *et al.* (2001) also identified the three SNZ/SNO pairs as neighbouring genes regulated by the same transcription factor whose corresponding proteins interact. However, neither of these analyses investigated the chromosomal locations and functions of the SNZ and SNO gene families in further detail. By correlating gene expression with gene location, Cohen *et al.* (2000) identified a large group of correlated adjacent genes on chromosome 6 which included SNO3, SNZ3 and THI5. However, although Cohen *et al.* (2000) searched for common promoter elements and upstream activating sequences (UAS) they did not investigate the functions and properties of the genes contained within this region in further detail.

The observations of the SNZ and SNO gene families made through using YETI alone conform well to what it currently known about them. Initially, the Genome vs Proteome Section of YETI highlighted that there were three cases of SNZ and SNO genes that were located next to each other in the genome and whose corresponding proteins interacted; YETI found these automatically based only on the gene location and protein interaction data and subsequently highlighted them enabling them to be

easily selected and collectively investigated in the other sections. As well as showing the similarity between all three SNZ/SNO gene pairs, all the sections of YETI consistently suggested a possible division within the SNZ and SNO gene families. The SNZ2/SNO2 and SNZ3/SNO3 gene pairs were consistently associated with one another along with many other genes involved in thiamin biosynthesis whereas the SNZ1/SNO1 gene pair, despite being related, was shown to be distinct from the other two pairs and not directly involved in thiamin biosynthesis.

Although much of what YETI highlighted about the SNZ and SNO genes was previously known before, the fact that YETI did highlight these facts based on the available data alone could be seen as a confirmation that the system and strategy works well. This case study is also a good illustration of how YETI can easily and rapidly be used to collectively investigate all the properties of a group of genes to investigate if and how they are working together to achieve their biological goals and to also examine what other genes and proteins they may be working with. Furthermore, YETI itself can also highlight potential features of interest to investigate further; in this case, neighbouring genes whose corresponding proteins interact with each other. In addition to highlighting some of the main advantages of YETI, such as the group approach and inter-linked sections, this case study also highlights the usefulness of specific features of YETI, such as the Genome Section for investigating possible evolutionary relationships between groups of genes and the Chromosome Window for providing good clear visual representations of gene locations.

However, this case study also highlights some of the disadvantages of YETI, namely the lack of sequence data and an incomplete protein interaction data set. Sequence data would enable users to directly examine the similarity of specific genes or chromosomal regions and to also examine if specific genes share similar regulatory regions. The incomplete protein interaction data is highlighted by the fact that Rodriguez-Navarro *et al.* (2002) reported interactions that are not currently present in the YETI database; specifically SNZ2-THI11 and SNZ3-THI11. This highlights one of the disadvantages with many protein-protein interaction resources as they tend to be populated with data mainly from high-throughput studies. The majority of scientific studies investigate the properties of a small number of specific proteins and subsequently report a small number of interactions between them; therefore contained within the scientific literature is a mass of important interaction data. However, to manually examine all of the published scientific literature for interactions is a major undertaking. Therefore, good text mining techniques that can automatically find and extract interactions from the literature would be extremely useful developments. Indeed, one successful protein interaction resource that currently includes text mining techniques is the STRING database (Von Mering *et al.*, 2003).

This case study also illustrates the benefits of filtering the protein-protein interactions as although all the SNO and SNZ proteins can interact highly with one another they are expressed at different times and could well be located in different cellular locations making some of these interactions irrelevant. Furthermore, it would also be of use to know if two proteins have been tested for an interaction and failed; for

example, knowing categorically whether SNZ1 does not interact with SRP1 would further suggest different cellular compartments for the SNZ proteins.

6.7: Discussion

As described above, the Genome vs Proteome Section of YETI provides a number of filters to filter the proteome dataset and remove certain types of interactions. These include filters that remove interactions that would bias the correlation results (such as self and two-way interactions), filters that remove possible false positives (such as interactions only reported once and interactions involving proteins not contained within the same cellular component) and filters that can also improve the quality of both the proteome and the genome dataset (such as removing interactions involving dubious ORFs). These filters are an essential feature as analysing the datasets without them can lead to incorrect conclusions. For example, the unfiltered proteome datasets showed a statistically significant number of observed interactions where both interacting proteins were located on the same chromosome and the average intracluster PID was substantially higher than the average intercluster PID. However, this correlation was caused by self and two-way interactions biasing the correlation results and the apparent correlation disappeared when the appropriate filters were applied. This was a problem that a recent study by Ge *et al.* (2001) experienced where a strong correlation between gene expression and protein interaction was observed. However, this study was recently discredited by Mrowka *et al.* (2003) who showed that the apparent correlation was in fact caused by the presence of self-interactions which were not removed in the original study. Furthermore, there is no

mention in either study of filtering the interactions for true duplicates and two-way interactions which could further bias the correlation results.

Overall, the results presented above suggest that there is no global correlation between genome location and protein interaction. There does not appear to be a tendency for proteins that interact with each other to be located near each other in the genome or for genes located near each other in the genome to interact with one another; although, there are a number of isolated cases. For two proteins to be located near each other in the genome they first have to be located on the same chromosome. Therefore, the first indication of a correlation would be significantly more observed protein-protein interactions where both proteins are located on the same chromosome than would be expected if it is assumed the genomic location of interacting proteins is random. However, the observed number of interactions was never found to be significant no matter what filters were applied to the datasets. The second indication of a correlation would be a low average distance between interacting proteins located on the same chromosome, especially when compared to the random dataset. However, the average distances observed were very large and similar to the average distances from the random dataset. Even though no overall correlation was observed there were a few isolated cases of interacting proteins being located next to each other on a chromosome; these were often involved in the same specific biological process suggesting that there is a functional reason for this co-location, such as co-regulation.

It could be argued that the above results are expected when one considers that the genes of eukaryotes are generally considered to be monocistronic, each with its own promoter at the 5' end and a transcription terminator at the 3' end (Blumentahl, 2004); however, it has recently become clear that not all eukaryotic genes are transcribed monocistronically (Blumenthal, 2004). To the best of our knowledge, this is the first analysis to investigate a potential correlation between protein interaction and genome location in *S. cerevisiae*. There is one related study by Ogata *et al.* (2000) who investigated, for a number of different organisms, if enzymes located near each other in the KEGG metabolic pathways were located near each other on the genome, forming Functionally Related Enzyme Clusters (FRECs). They found that the relative number of enzymes in FRECs was close to 50 % for *Bacillus subtilis* and *Escherichia coli* but was less than 10 % for *S. cerevisiae*. This ties in with the results presented here which suggest relatively few interacting, and therefore possibly functionally related, proteins are located near each other in the genome.

One improvement that could be made to the genome vs proteome analysis presented above would be higher quality datasets. Our biological understanding of *S. cerevisiae* is constantly improving and evolving with more genes being functionally characterised and more erroneous ORFs ruled out; therefore the quality of the genome dataset used in YETI is constantly improving with time. Although the protein-protein interaction dataset used in YETI is one of the largest available it is still incomplete (Wallhout *et al.*, 2000; Tucker *et al.*, 2001; Grigoriev *et al.*, 2003; Uetz *et al.*, 2005) and can be error-prone due to false-positives and false-negatives generated through techniques such as the yeast two-hybrid approach. However, new

datasets are constantly being produced, new and improved technologies are constantly being developed and filters can be applied to improve the quality of the existing dataset (Bader *et al.*, 2004; Bork *et al.* 2004). Therefore, over time the proteome data should also increase in both size and quality.

Chapter 7

Genome vs Transcriptome Correlation Analysis

7.1: Introduction

A Genome vs Transcriptome correlation analysis was performed using YETI to investigate if there was a tendency for genes located next to each other in the genome to be coexpressed. Genes that are coexpressed are likely to be related functionally (the concept of guilt by association). Therefore, it could be argued that genes that are coexpressed and colocated are even more likely to be related functionally. To examine a possible global correlation between gene location and gene expression the first step is to identify the number of physically adjacent genes in the genome that are coexpressed and test if this number is statistically significant by comparing it to the expected number derived from a control set. In addition, whether there is an overall correlation or not, the chromosomal regions displaying coexpression can be investigated in more detail using YETI.

7.2: Chromosome Correlation Maps

To investigate a potential correlation between gene location and gene expression the approach developed by Cohen *et al.* (2000) was applied. Cohen *et al.* (2000) developed a visualisation technique called chromosome correlation maps to display correlations between the expression patterns of genes on the same chromosome. A chromosome correlation map is essentially a two dimensional matrix generated by organising all the ORFs on a specific chromosome into two identical axes; ORFs are arranged by the sequential order they appear on the chromosome. If the number of ORFs on a chromosome is equal to N, then the number of squares in the matrix equals N^2. Each square in the matrix represents the Pearson correlation coefficient of

the two ORFs the square corresponds to. The Pearson correlation coefficient for each square in the matrix is represented with a colour gradient to give a visual representation of the coexpression of genes along the chromosome; bright greens represent high Pearson correlation coefficients (positive correlation) whereas bright reds represent low Pearson correlation coefficients (anti-correlation). An example chromosome correlation map for a small hypothetical chromosomal region is displayed in Figure 7.1. The bright green diagonal line from the top left corner of the map to the bottom right corner corresponds to the Pearson correlation coefficients of each ORF with itself; as each ORF has an identical pattern of expression with itself the Pearson correlation coefficient is always equal to 1 in these cases. ORFs that have correlated expression and are physically close together form green regions around the diagonal; an example region is highlighted in blue in Figure 7.1.

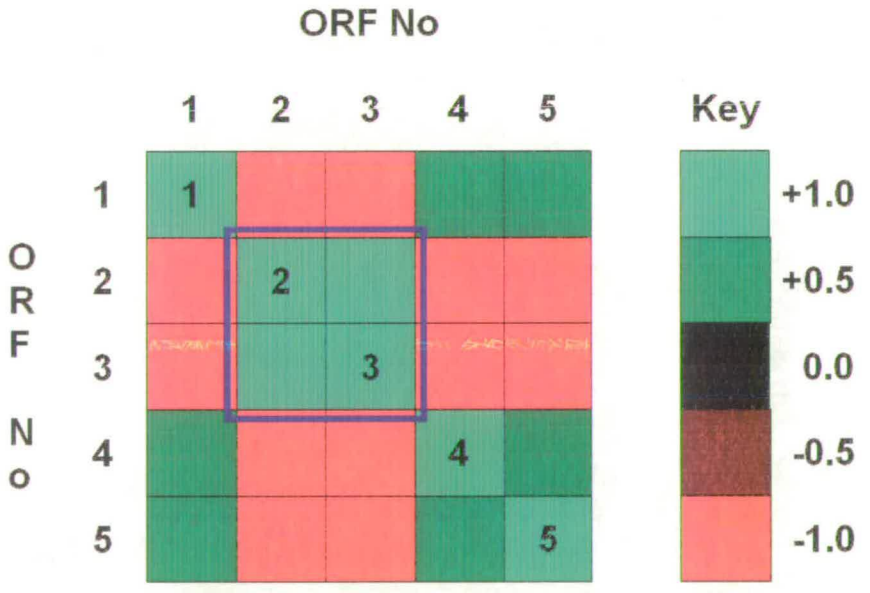


Figure 7.1: Chromosome Correlation Map
This is a figure of an example chromosome correlation map for a small hypothetical chromosomal region containing 5 ORFs. The map is essentially a two-dimensional matrix that displays the Pearson correlation coefficient of every ORF with every other ORF. The bright green diagonal from the top left corner to the bottom right corresponds to Pearson correlation coefficient of each ORF with itself which is always equal to 1. ORFs that have correlated expression and are physically close together form green regions around the diagonal as highlighted by the blue box.

7.3: YETI Genome vs Transcriptome Section

The Genome vs Transcriptome Section of YETI was used to perform the genome vs transcriptome correlation analysis, to find and investigate chromosomal regions exhibiting coexpression and to investigate if there was an overall tendency for genes located next to each other in the genome to be coexpressed. In this section, YETI displays the chromosome correlation map for a selected chromosome (Figure 7.2); the expression data used in every chromosome correlation map is currently from the Gasch *et al.* (2000) data set. All the genomic features on the selected chromosome (i.e. ORFs as well as [amongst others] tRNAs and rRNAs) are represented on the map to give a realistic impression of whether ORFs are physically adjacent or not; however, dubious ORFs are not displayed on the map as these are highly unlikely to be real genes. Any genomic feature that does not have expression data available is still represented on the map with the missing expression data displayed with black squares. The correlation maps YETI displays enable regions of coexpression on the chromosome to be rapidly found. These regions could involve ORFs that are physically close forming bright green regions around the diagonal or involve ORFs that are physically distant forming bright green regions elsewhere in the map. Any region of interest on the map can easily be selected enabling all the ORFs contained within this region to be collectively investigated in further detail in the other sections of YETI. In addition, there is also a Find function to highlight the location of any specific ORF of interest on the map.

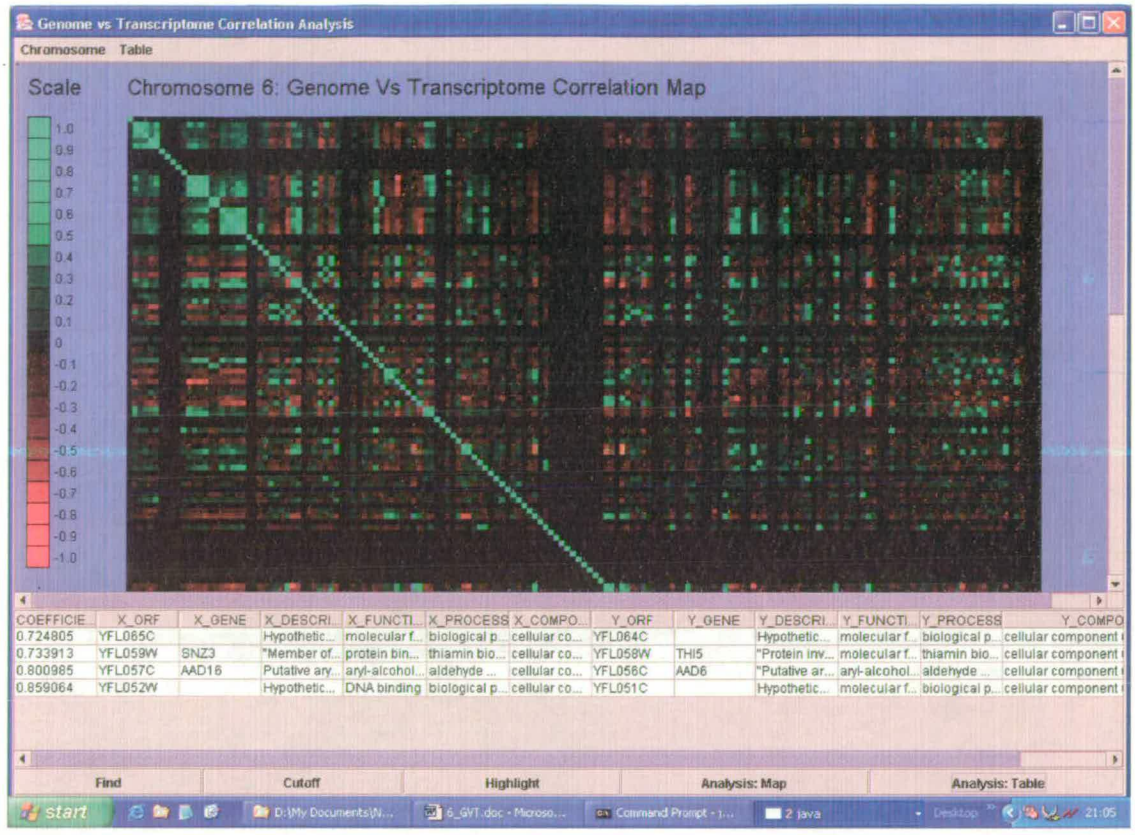


Figure 7.2: Screenshot of the Genome vs Transcriptome Section of YETI
This is a screenshot of the Genome vs Transcriptome Section of YETI which displays the chromosome correlation map of a selected chromosome. Furthermore, a data table containing information on all the adjacent ORFs that are significantly coexpressed is also displayed. In this figure, the correlation map for chromosome 6 is displayed.

In addition to the actual chromosomal correlation map, this section of YETI also includes a data table containing information on all the adjacent ORFs on the selected chromosome that are coexpressed; this table therefore gives an immediate overview of all the regions of coexpression on the chromosome which can then be investigated further. In YETI, adjacent ORFs are defined as two ORFs located on the same chromosome with no other genomic features between them and the default definition of coexpressed is two adjacent ORFs with a Pearson correlation coefficient equal to or above the standard cutoff of 0.7. In the data table, coexpressed adjacent ORFs are sorted by the order they appear on the chromosome enabling the user to easily and rapidly find chromosomal regions exhibiting coexpression. For example, in addition

to clearly showing all the coexpressed adjacent ORFs, the table could show that multiple coexpressed adjacent ORFs form larger coexpressed regions such as triplets or quadruplets. Furthermore, the data table contains the primary GO annotations of every ORF enabling the user to rapidly see if the ORFs in coexpressed regions share the same or similar functions. Additional features of the data table include: (1) A Cutoff function to change the Pearson correlation coefficient cutoff of coexpressed adjacent ORFs displayed in the data table; (2) A Highlight function to highlight the location of any of the coexpressed adjacent ORFs displayed in the data table on the chromosome correlation map; and (3) A direct link from the data table to the Analysis Section enabling any of the coexpressed adjacent ORFs displayed in the table to be investigated in further detail in the other sections of YETI.

7.4: Chromosomal Regions of Coexpression

The correlation map for each of the 16 nuclear chromosomes of *S. cerevisiae* was analysed in the Genome vs Transcriptome Section of YETI to find regions of coexpression. A number of interesting regions were found and subsequently selected and investigated further using the other sections of YETI; a comprehensive account of the findings is presented in the case studies below.

7.4.1: Galactose Metabolism

Using YETI to analyse the correlation map of chromosome 2 reveals a triplet of adjacent ORFs that are all highly coexpressed with one another (Figure 7.3). These

three ORFs are YBR018C/GAL7, YBR019C/GAL10 and YBR020W/GAL1 and are all characterised with the 'galactose metabolism' GO biological process annotation. As these three ORFs are colocated, coexpressed and share the same GO annotation it was decided to investigate them in further detail in YETI.

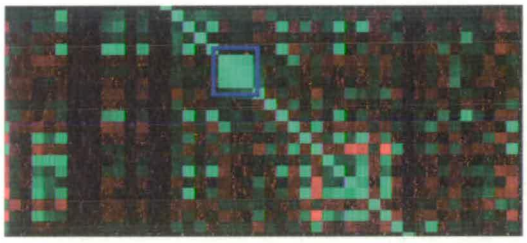


Figure 7.3: Chromosomal correlation map of the galactose genes on chromosome 2
This is an image of the chromosomal region surrounding the three genes involved in galactose metabolism on chromosome 2. The location of the three genes (GAL7, GAL10, GAL1) is highlighted with the blue box. As can be seen the adjacent ORFs are highly coexpressed with one another.

The Transcriptome Section of YETI shows that these three genes (now assigned to the red YETI group) are located right next to each other in the gene expression hierarchical tree (Figure 7.4). As these genes are involved in the same biological process and their expression appears to be tightly coregulated, other genes located in this region of the tree could well be involved in galactose metabolism as well (or a related biological process); this could enable functional roles for any unknown genes in this region to be inferred. To this end, the surrounding genes in the tree were also selected (assigned to the green YETI group) and investigated further. YETI shows that there are indeed three additional genes involved in galactose metabolism located in this region of the tree, namely GAL2, GAL3 and GAL80. There are also three other genes located within this galactose cluster and they are FUR4, MRF1 and MAL12 which are involved in 'uracil transport', 'protein biosynthesis' and 'maltose

catabolism', respectively. However, no genes of unknown function were found in this region so no functional roles could be investigated or inferred in this instance.

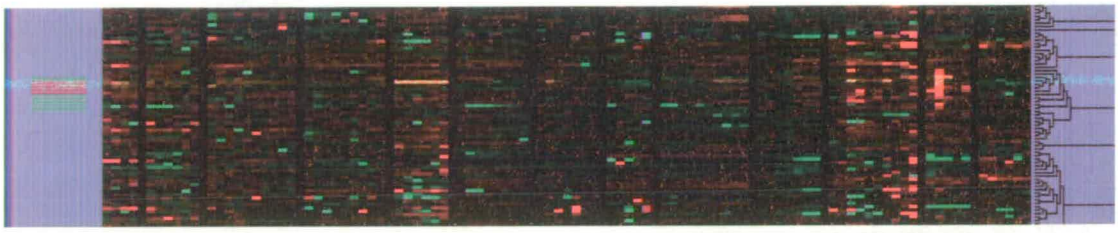


Figure 7.4: The galactose cluster region of the gene expression hierarchical tree
This is a figure of the gene expression hierarchical tree with the location of the three adjacent genes involved in galactose metabolism located on chromosome 2 highlighted in red. The surrounding genes in the tree have subsequently been selected for further investigation and highlighted in green.

The Proteome Section of YETI can be used to investigate whether the identified proteins involved in galactose metabolism are interacting with one another to achieve their biological goals and to also investigate what other proteins they are interacting with; if any proteins of unknown function interact with a number of galactose proteins this could allow a functional role to be inferred. YETI shows that many (but not all) of the identified genes involved in galactose metabolism interact directly with one another forming a large cluster of interactions (Figure 7.5); this cluster also reveals the presence of yet another protein involved in galactose metabolism, namely GAL4 which interacts directly with GAL80. Another interesting observation is that the three original galactose genes colocated on chromosome 2 (GAL7, GAL10 and GAL1) do not appear to interact directly with one another. However, GAL7 and GAL1 both interact with SER3 which is involved in 'serine family amino acid biosynthesis'. These two interactions could well be false positives given that SER3 is involved in such an unrelated biological process; furthermore, these interactions have only been reported once and come from the Ito *et al.* (2001) study which is renowned

for false-positives. The additional genes that were colocated in the gene expression hierarchical tree (FUR4, MRF1 and MAL12) were not found to interact directly with any of the galactose proteins or clusters which suggests that, despite their colocation in the hierarchical tree, they are not directly involved in the process of galactose metabolism. Indeed, investigating these genes individually shows that none of them are directly coexpressed (Pearson >= 0.7) with any of the galactose metabolism genes. On the other hand, although GAL2 (which was also identified from the hierarchical tree) does not interact with the other galactose proteins, it is significantly coexpressed with GAL1 further linking it to the process of galactose metabolism. Overall, the proteins known to be involved in galactose metabolism interact with a number of other proteins involved in a wide variety of biological processes but there does not appear to be any common biological processes among them. Furthermore, there are no proteins of unknown function that interact with any of the galactose proteins so no functional roles could be investigated or inferred in this instance.

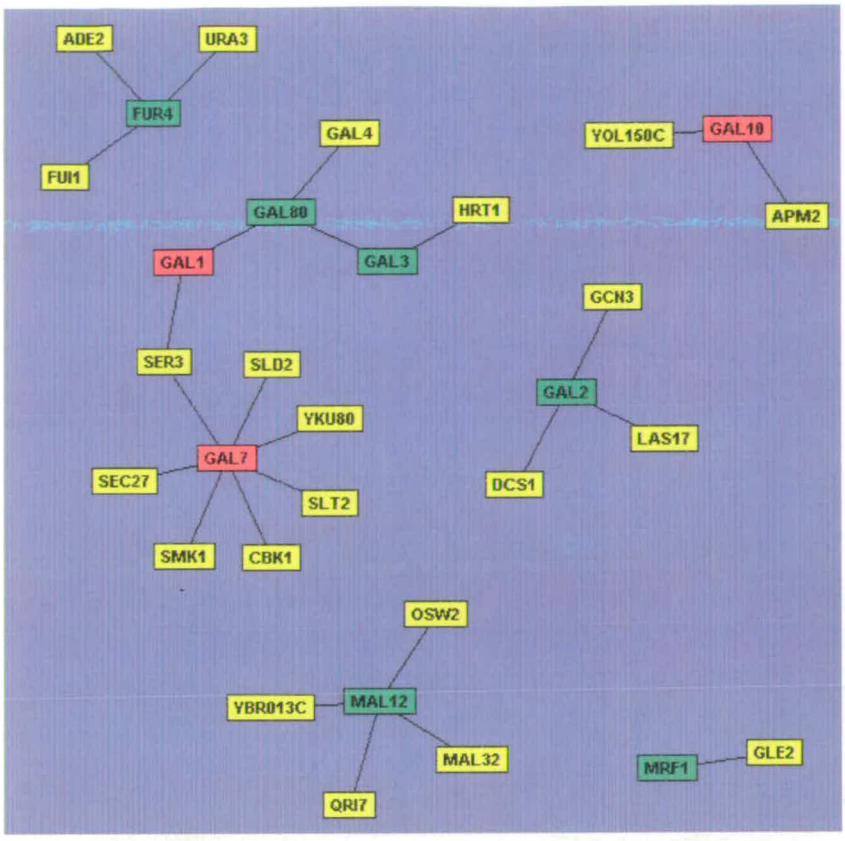


Figure 7.5: Protein interaction map of the identified galactose metabolism proteins
This is a figure of all the protein-protein interactions involving the original three adjacent galactose metabolism genes identified on chromosome 2 (highlighted in red) and the additional genes selected from the gene expression hierarchical tree (highlighted in green). As can be seen many of the galactose (GAL) genes interact directly and indirectly with each other forming a large cluster of interacting proteins.

At this point, it is worth comparing the observation made through using YETI to what is already known about the galactose metabolism pathway. The galactose metabolism pathway has been extensively studied with the majority of components already identified and characterised (for example, see Lohr *et al.*, 1995). It is a classic example of a genetic regulatory switch, in which enzymes required for the transport and catabolism of galactose are expressed only when galactose is present and repressing sugars such as glucose absent (Ideker *et al.*, 2001). An overview of the pathway is presented in Figure 7.6. The first component of the pathway is GAL2 which encodes a galactose permease that transports galactose into the cell. Next are

the enzymatic proteins of the pathway consisting of GAL10 (galactose mutarotase & UDP-glucose 4-epimerase), GAL1 (galactokinase), GAL7 (galactose-1-phosphate uridyl transferase), and PGM2 and PGM1 (both phosphoglucomutases). GAL4, GAL3 and GAL80 are all involved in the regulation of the enzymatic proteins and transporter. GAL4 is a DNA-binding factor that can strongly activate their transcription, but in the absence of galactose GAL80 binds to the activation domain of GAL4 and inhibits its activity. When galactose is present in the cell, it causes the activation of GAL3 which can bind to GAL80 and alter the GAL4/GAL80 complex; this causes the GAL4 activation domain to become available and results in the high expression of the enzymatic and transporter genes (Larschan *et al.*, 2001; Ideker *et al.*, 2001). It is important to note that the transporter gene GAL2 has a higher basal level of expression than the enzymatic genes because there needs to be an initial amount of transporters on the cell membrane to transport the galactose into the cell to begin with.

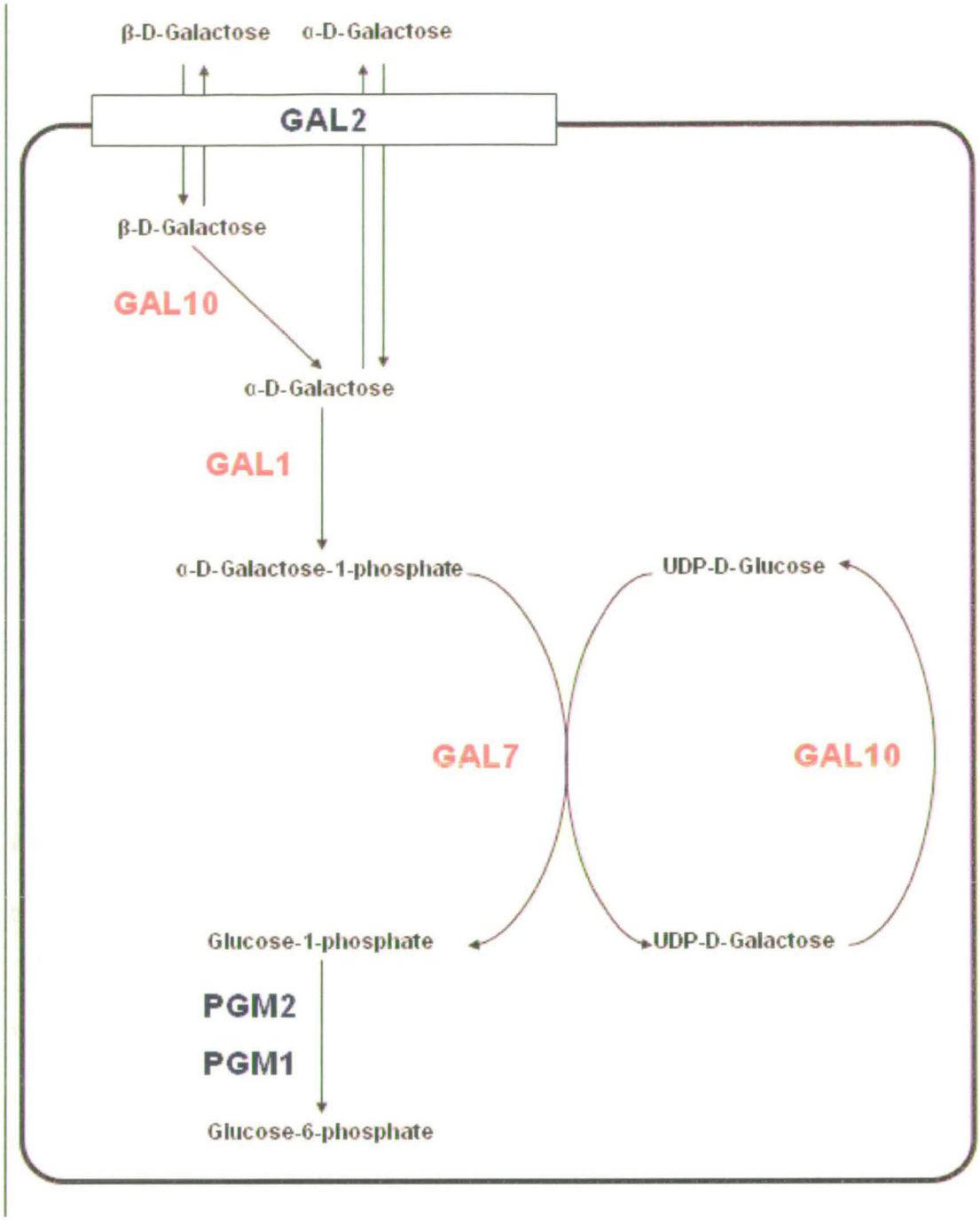


Figure 7.6: Overview of the *S. cerevisiae* galactose metabolism pathway
This figure presents an overview of the galactose metabolism pathway from the transport of galactose into the cell by GAL2 to the production of glucose-6-phosphate by PGM2 and PGM1. Proteins highlighted in red are the galactose proteins colocated on chromosome 2 and proteins highlighted in blue are galactose proteins located on other chromosomes. This figure is based on the galactose metabolism pathway picture from the *Saccharomyces* Genome Database (SGD: Cherry *et al.*, 1998).

Although YETI does not necessarily reveal anything new about the process of galactose metabolism, this case study does demonstrate the potential of YETI as it was able to easily and rapidly identify the majority of this pathway based on the experimental data. Firstly, the Genome vs Transcriptome Section highlighted that three adjacent genes on chromosome 2 (GAL7, GAL1, GAL10) were highly coexpressed. Secondly, the Transcriptome Section showed that these three genes were located in the same region of the hierarchical tree as GAL2; this is now expected as these four genes are the core components of the galactose metabolism pathway and are regulated by the same factors. Thirdly, the Transcriptome Section also showed that GAL80 and GAL3 were located in the same region of the hierarchical tree as the above four genes. Furthermore, the Proteome Section showed that GAL80 interacts directly with both GAL3 and GAL4 (as well as GAL1); this is now expected as the interaction of GAL80 with GAL3 and GAL4 is the main regulatory mechanism of the galactose metabolism pathway. YETI could not assign any new genes of unknown function to this biological process but this is probably to be expected as this pathway is so well studied. However, it is important to note that had the function of any of the GAL genes been unknown then YETI would have successfully highlighted their potential involvement in galactose metabolism based on their chromosomal location, gene expression and/or protein interactions.

However, YETI did not manage to associate PGM1 or PGM2 with the other galactose metabolism genes. This can be explained by the fact that none of the enzymatic proteins of the pathway appear to interact with each other or with the transporter protein; this could be due to poor coverage and false-negatives resulting

in an incomplete protein-protein interaction data set or could be expected as these enzymes may not need to interact with other proteins to fulfil their biological functions. Furthermore, PGM1 and PGM2 do not share similar patterns of expression with the GAL genes because PGM1 and PGM2 are involved in many metabolic pathways (e.g. galactose metabolism, glycogen catabolism, lactose degradation and sucrose biosynthesis) which means the expression of PGM1 and PGM2 differs from the expression of the other GAL genes in the presence of other sugars.

In general, this case study highlights a number of the advantages of YETI. One of the main aims of YETI was to provide clear graphical representations that enable users to easily and rapidly explore the stored data sets and find interesting features. This is exemplified by the chromosome correlation maps which enable users to rapidly explore possible correlations between gene location and expression and easily select any regions of interest to investigate further. Furthermore, the group approach combined with the inter-linked sections of YETI enables users to collectively investigate if and how a group of potentially related genes are working together in order to achieve their biological goal and to also investigate what other genes/proteins they may be working with. This is demonstrated quite well in this case study as starting from a triplet of coexpressed genes involved in galactose metabolism, which YETI automatically highlighted, YETI was able to associate them with the majority of the other galactose genes through the collective investigation of their expression and interaction partners. Although in this instance, nothing new was highlighted about the process of galactose metabolism, it does show

the potential for such an approach in a less well studied biological process or organism.

7.4.2: Allantoin Degradation

Using YETI to analyse the correlation map of chromosome 9 reveals a group of six adjacent ORFs that are all highly coexpressed with one another (Figure 7.7). These six ORFs are YIR027C/DAL1, YIR028W/DAL4, YIR029W/DAL2, YIR030C/DCG1, YIR031C/DAL7, YIR032C/DAL3 which are all characterised with allantoin degradation related GO biological process annotations; allantoin is a nitrogen source that can be degraded to form urea. As all six genes in this cluster were involved in the same biological process and also highly coexpressed together, the genes themselves as well as the overall biological process were investigated in further detail using YETI.

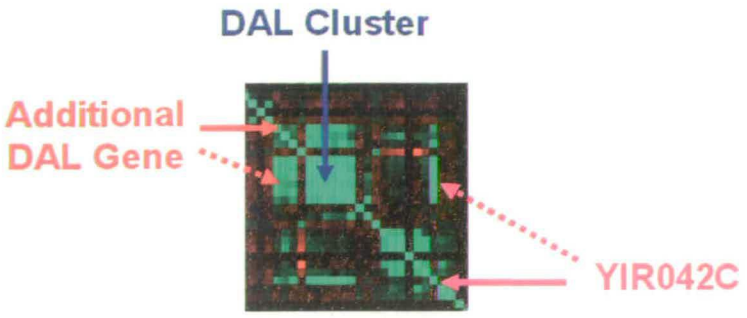


Figure 7.7: Chromosome correlation map of the DAL cluster on chromosome 9
This is an image of the chromosomal region surrounding the six coexpressed adjacent genes involved in allantoin degradation on chromosome 9 (the DAL cluster); the location of the DAL cluster is highlighted with the blue arrow. Upstream is a triplet of genes that are coexpressed with the DAL cluster. This triplet also contains another DAL gene (DAL81) which is highlighted with the red arrow; the dotted red arrow indicates the region displaying the coexpression between the triplet and the DAL cluster. Downstream is a single gene of unknown function called YIR042C (highlighted with the pink arrow) that is also coexpressed with the DAL cluster; the dotted pink line indicates the region displaying the coexpression between YIR042C and the DAL cluster.

As discussed above, YETI highlighted that six adjacent genes (the DAL cluster) on chromosome 9 were highly coexpressed. However, there are also a few other interesting observations that can be made from this region of the chromosome correlation map (Figure 7.7). Firstly, there is a triplet of genes just upstream which are coexpressed with the DAL cluster. Further examination of this triplet reveals the presence of an additional DAL gene, namely DAL81. Secondly, there is a single gene downstream that is also coexpressed with the DAL cluster, namely YIR042C which is currently of unknown function; possible functional roles for YIR042C are discussed later.

The Transcriptome Section of YETI shows that members of the DAL cluster are colocated in the gene expression hierarchical tree (Figure 7.8); this colocation is to be expected as they are so highly coexpressed. As these genes are involved in the same biological process and their expression appears to be tightly coregulated, other

genes located in this region of the tree could well be involved in the same or related biological processes; this could enable possible functions for any unknown genes in this region to be inferred (the concept of guilt by association). Indeed, YETI shows that this region of the tree contains an additional three genes characterised as being involved in the allantoin degradation pathway; specifically: DAL80, DAL5 and DUR3. Furthermore, there are a number of other genes involved in the metabolism of nitrogen compounds; for example: MEP1 and MEP2 (ammonium permeases); ASP3-1, APS3-2, ASP3-3 and ASP3-4 (asparaginases); and GAT1 and GLN3 (transcriptional activators of genes involved in nitrogen catabolite repression). However, there are also a large number of proteins involved in the metabolism of sulphur compounds located in this region of the tree as well; for example: SUL1 and SUL2 (sulphate transport); MET4, MET28 and MET32 (sulphur amino acid metabolism); MET10 (sulphate assimilation); and MET1, MET2, MET3 and MET16 (methionine metabolism). Furthermore, there are also a number of genes of unknown function located in this region of the tree; specifically: YBR147W, YDL183C, YGR125W, YIL165C, YIR042C (which was also identified from the chromosome correlation map), YLR053C and YLR364W. Possible functional roles for these unknown genes are discussed later.

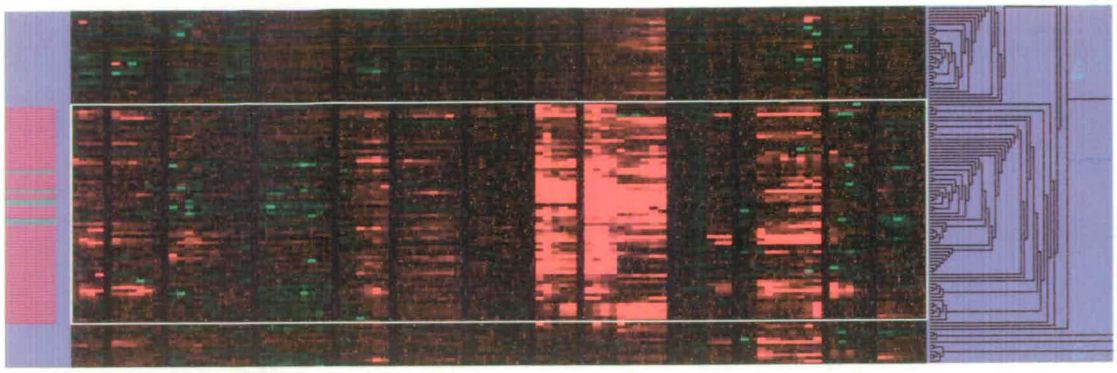


Figure 7.8: The DAL cluster region of the gene expression hierarchical tree
This is a figure of the gene expression hierarchical tree with the location of the six members of the DAL cluster located on chromosome 9 highlighted in green. The other genes in this cluster of the tree have subsequently been selected for further investigation and highlighted in red.

At this point it is again worth comparing the observations made through using YETI to what is already known about the allantoin degradation pathway and the DAL cluster. The DAL cluster is the largest known metabolic gene cluster in yeast (Wong *et al.*, 2005) and consists of six adjacent genes encoding proteins which form the majority of the allantoin degradation pathway that enables *S. cerevisiae* to use allantoin as a sole nitrogen source (Figure 7.9). *S. cerevisiae* is able to import allantoate via the permease DAL5 and both (R) and (S)-allantoin via the permease DAL4; the racemase DCG1 is able to convert (R)-allantoin to (S)-allantoin. The conversion of allantoin to ammonia is carried out by DAL1 (allantoinase), DAL2 (allantoicase) and DAL3 (ureidoglycolate hydrolase) which work sequentially to generate urea. Urea is then degraded to ammonia in a two-step process by the DUR1,2 protein which is a multifunctional enzyme. An additional allontoin related protein is DUR3 which is a plasma membrane urea transporter whose expression is induced by allophanate (the last intermediate of the allantoin degradation pathway). The allantoin degradation pathway genes are regulated by a general signal that responds to the availability of readily utilisable nitrogen sources, and also by

pathway-specific induction by allantoin or the intermediate allophanate. These regulatory effects are mediated by cis-acting DNA elements and the trans-acting factors GLN3, GAT1, DAL80, DAL81, and DAL82 (Cherry *et al.*, 1998; Rai *et al.*, 1999; Scott *et al.*, 2000; Magasanik *et al.*, 2002). A recent study (Wong *et al.*, 2005) showed that the DAL cluster was assembled quite recently in evolutionary terms through a set of genomic rearrangements that happened almost simultaneously. This study showed that six genes involved in allantoin degradation, which were previously scattered around the genome, became relocated to a single subtelomeric site in an ancestor of *S. cerevisiae* (thus forming the DAL cluster). This genomic rearrangement coincided with a biochemical reorganisation of the purine degradation pathway which switched to importing allantoin instead of urate.

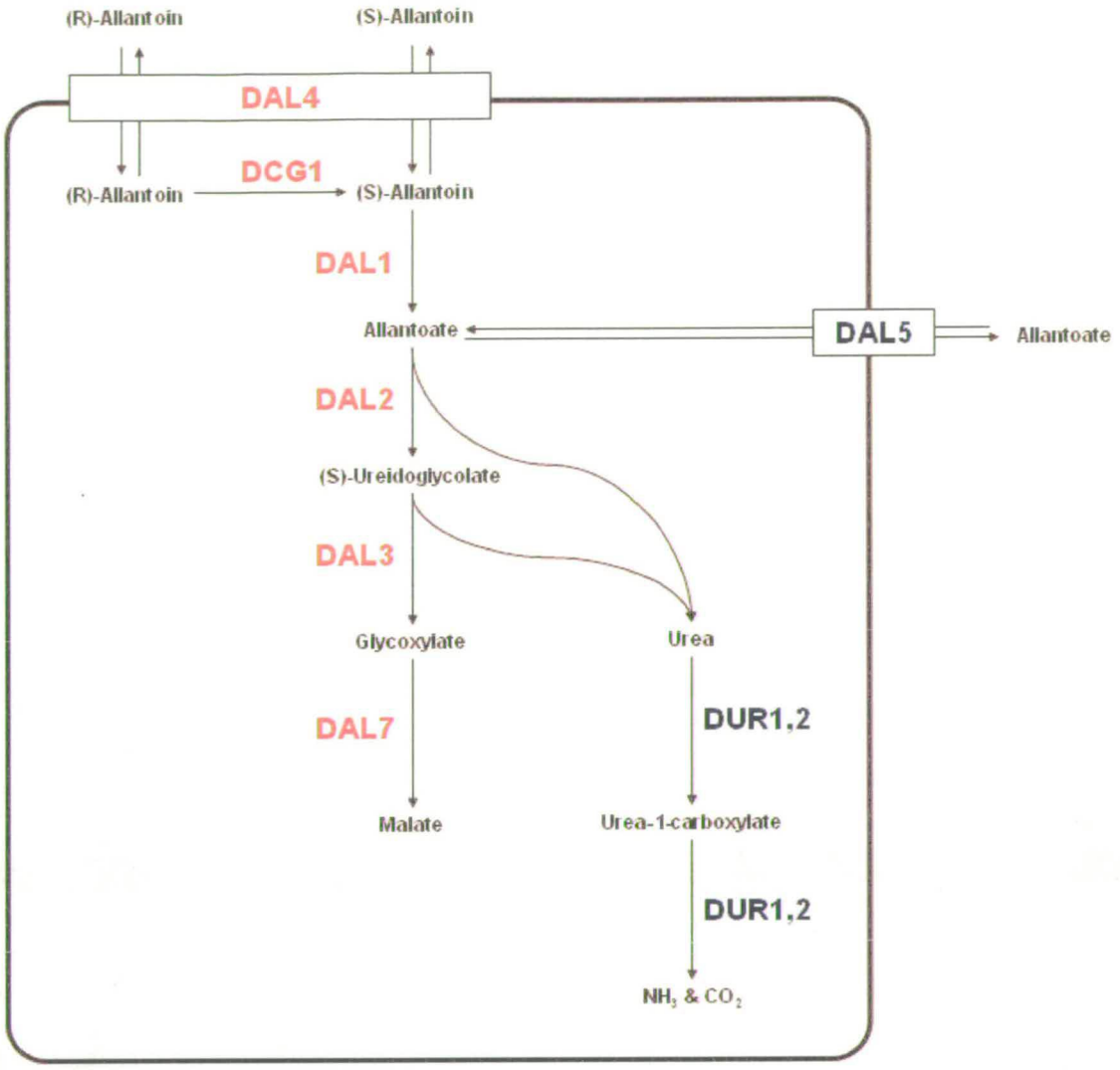


Figure 7.9: Overview of the *S. cerevisiae* allantoin degradation pathway
This figure presents an overview of the allantoin degradation pathway from the transport of allantoin and allantoate into the cell by DAL4 and DAL5, respectively, to the production of urea and malate by DUR1,2 and DAL7, respectively. Proteins highlighted in red are the members of the DAL gene cluster located on chromosome 9 and proteins highlighted in blue are located on other chromosomes. This figure is based on Figure 1 from Wong *et al.* (2005).

Like the galactose metabolism case study, this case study also demonstrates the potential of YETI as it was again able to easily and rapidly identify the majority of this pathway based on the available experimental data. In summary, the Genome vs Transcriptome Section highlighted that six adjacent genes on chromosome 9 (the DAL cluster: DAL1, DAL4, DAL2, DCG1, DAL7, DAL3) were highly coexpressed

with each other; this is now expected as these six proteins are the core components of the allantoin degradation pathway. This section also highlighted a triplet of genes which included DAL81 upstream that were coexpressed with the DAL cluster; this observation is now expected as DAL81 is a positive regulator of genes in multiple nitrogen degradation pathways. The Transcriptome Section showed that the six genes of the DAL cluster were located in the same region of the gene expression hierarchical tree as DAL80, DAL5, DUR3, GLN3 and GAT1 as well as numerous other genes involved in nitrogen compound metabolism; this is now expected as DAL5 is an allantoate permease and DUR3 is a urea transporter induced by allophanate, while DAL80, GLN3 and GAT1 are all involved in the regulation of the allantoin degradation pathway. Although the above findings are now expected, it is again important to note that had the function of any of the above genes been unknown then YETI would have successfully highlighted their potential involvement in allantoin degradation, or the broader nitrogen compound metabolism process, based on their chromosomal location and gene expression patterns.

However, YETI did not manage to associate DAL82 or DUR1,2 with the rest of the allantoin pathway. As DAL82 is a positive regulator of allophanate inducible genes it is quite surprising that it is not located with the DAL cluster in the gene expression hierarchical tree. However, by examining DAL82 individually (via its Datasheet Window) YETI shows that the genes it is most highly coexpressed with are DAL4 (R = 0.782) and DUR3 (R = 0.769) linking it to the allantoin degradation pathway. The lack of association of DAL82 with the other allantoin genes in the gene expression hierarchical tree could be explained by the way the (pairwise average linkage)

hierarchical clustering process proceeds; i.e. the distance between two clusters is calculated as the average distance between all members of the first cluster and all members of the second cluster. The non-association of DUR1,2 could be explained by the fact the degradation of urea is a generic reaction which is involved in many pathways not just allantoin degradation. Therefore, DUR1,2 could have a high basal level of transcription which does not change drastically; this theory seems to be supported by the expression data of DUR1,2 which shows its relative level of expression does not change dramatically in virtually all microarray experiments of the Gasch *et al.* (2000) data set.

Interestingly, the Proteome Section shows that none of the proteins involved in the allantoin degradation pathway interact with one another; in fact, they interact with very few proteins. This is similar to the observation that none of the core proteins involved in galactose metabolism interact directly with one another and again could be explained by an incomplete protein-protein interaction data set or by the fact that the enzymes involved do not need to interact with each other to achieve the biological functions.

As described above, there were a number of genes of unknown function located in the same region of the hierarchical tree as the DAL cluster; specifically: YBR147W, YDL183C, YGR125W, YIL165C, YIR042C, YLR053C and YLR364W. As these unknown genes are located in the same region of the hierarchical tree as many genes involved in sulphur and nitrogen compound metabolism they could well be involved in similar biological processes; although this is quite a broad functional assignment it

is a starting point for further investigation and experimentation. Interestingly, these results are supported to some degree by a recent study which used microarrays to characterise the transcriptional response of *S. cerevisiae* to growth limitation by carbon, nitrogen, phosphorus or sulphur (Boer *et al.*, 2003). In this study, both YIR042C and YLR364W were reported to be part of a group of genes that had 'specifically higher expression under sulphur limitation' along with many other genes involved in the metabolism of sulphur compounds; while YLR053C was reported to be part of a group of genes that had 'specifically higher expression under nitrogen limitation' along with many other genes involved in the metabolism of nitrogen compounds. However, YETI shows that YIR042C is highly coexpressed with mostly nitrogen not sulphur compound metabolism genes; furthermore, the five genes YIR042C is most highly coexpressed with are DUR3 (0.904), DAL5 (0.884), DAL7 (0.866), DAL4 (0.836) and DCG1 (0.813) all of which are involved in the allantoin degradation pathway. In addition, YIR042C was also highlighted on the initial correlation map of chromosome 9 as a gene displaying correlated expression with the DAL cluster. Therefore, the observations presented here suggest that YIR042C is more likely to be involved in nitrogen rather than sulphur compound metabolism and could well be involved in the allantion degradation pathway. However, these observations can only be proven by experiments in the laboratory; possible experiments include gene knockouts combined with growth on various nitrogen or sulphur compound limited mediums to examine any growth defects, and also microarray experiments monitoring gene expression under these mediums.

Overall, this case study is similar to the galactose metabolism case study presented above and highlights the same advantages of YETI; the group approach combined with the inter-linked sections of YETI enables users to collectively investigate if and how a group of potentially related genes are working together in order to achieve their biological goal and to also investigate what other genes/proteins they may be working with. However, in this case study all the associations came from observations of chromosomal location and gene expression while the protein interactions did not yield any useful information. In particular, a large cluster of genes was found in the gene expression hierarchical tree that contained many genes involved in the metabolism of nitrogen and sulphur compounds; this enabled possible (broad) functional roles for a number of unknown genes located in the cluster to be inferred. Therefore, if the protein-protein interaction data set is indeed incomplete, perhaps more information about if and how the allantoin degradation proteins are working together could be yielded from the Proteome Section in the future.

7.4.3: Helicases

Using YETI to analyse the correlation map of chromosome 2 reveals a triplet of adjacent ORFs (YBL113C, YBL112C and YBL111C) located in the left arm telomere that are all highly coexpressed with one another (Figure 7.10). YBL112C and YBL111C are both ORFs of unknown function whereas YBL113C is an ORF for which little is known but has been characterised with the 'helicase activity' GO molecular function annotation. As these three ORFs are all highly coexpressed with one another they could all be involved in similar biological process and perhaps have

similar functions (the concept of guilt by association). Therefore, the two ORFs of unknown function in this region could also have helicase activity and YETI was used to investigate this hypothesis further.

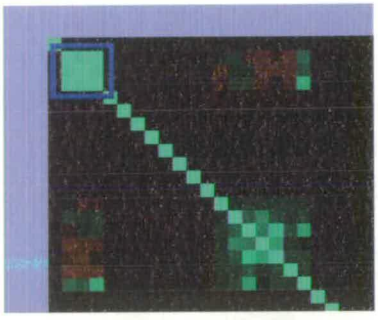


Figure 7.10: Chromosome correlation map of left arm telomere of chromosome 2
This is a screenshot of the chromosome correlation map of the left arm telomere of chromosome 2. There is a triplet of adjacent ORFs (YBL113C, YBL112C, YBL111C) that are all highly coexpressed with one another, located at the end of the chromosome arm (highlighted with the blue box).

The Transcriptome Section of YETI shows that the three adjacent ORFs are all located in the same region of the gene expression hierarchical tree (Figure 7.11). Using YETI to select all the other genes from this region of the tree shows that approximately half of the genes in this region are of unknown function and the other half are characterised with either a 'helicase activity' or 'DNA helicase activity' GO molecular function annotation. Furthermore, the Genome Section shows that all of the ORFs in this region of the hierarchical tree are located in the telomeric regions of the nuclear chromosomes (Figure 7.12). Therefore, given that all the genes in this region of the tree are similarly coexpressed and similarly colocated in the genome, and that half of the genes in this region are already characterised with a helicase activity annotation, naturally leads one to suggest that all the unknown genes in this region of the tree could well have 'helicase activity' as well. Possible laboratory

experiments to validate this observation would be gene knockouts to observe any growth defects. However, given the shear number of genes with potential 'helicase activity' they could well have redundant functions so multiple gene knockouts may well be needed to observe any growth defects.

In particular, this case study highlights the utility of the Genome Section of YETI for investigating whether a group of genes are related through similar genomic locations. Furthermore, it also demonstrates how the expression data set can be analysed in conjunction with other data sets. For example, in this case study a specific cluster of interest was identified in the gene expression hierarchical tree; YETI enabled all the genes within this cluster to be selected and collectively investigated in further detail to examine if they shared similar annotations, if they were located in similar chromosomal regions or if they encoded proteins that interact with one another. However, in this case study (as with the previous case studies) the protein-protein interaction data did not yield any useful information; the corresponding proteins of the genes located in this region of the tree interact with few proteins and none of them interact with each other or with any common proteins.

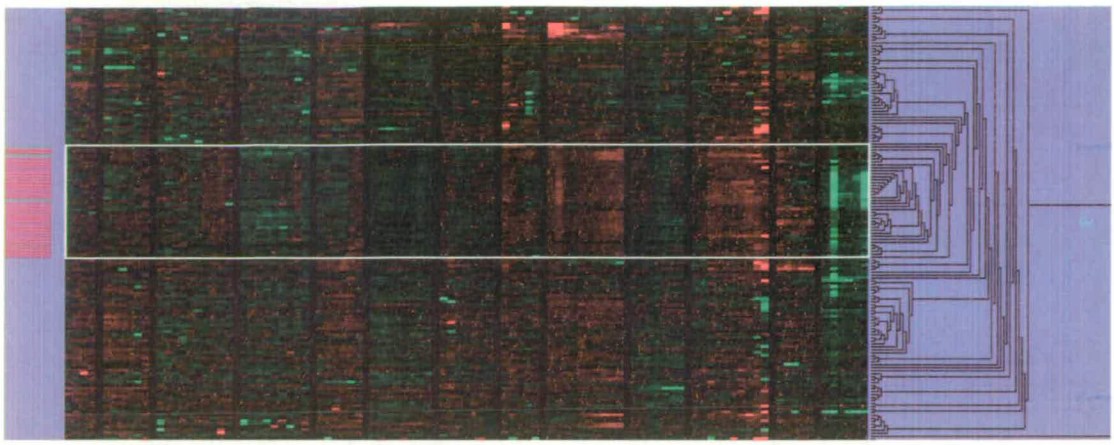


Figure 7.11: The helicase region of the gene expression hierarchical tree
This is a figure of the gene expression hierarchical tree with the location of the three adjacent coexpressed ORFs from the telomeric region of the left arm of chromosome 2 (YBL113C, YBL112C and YBL11C) highlighted in green. The surrounding genes in the tree have subsequently been selected for further investigation and highlighted in red.

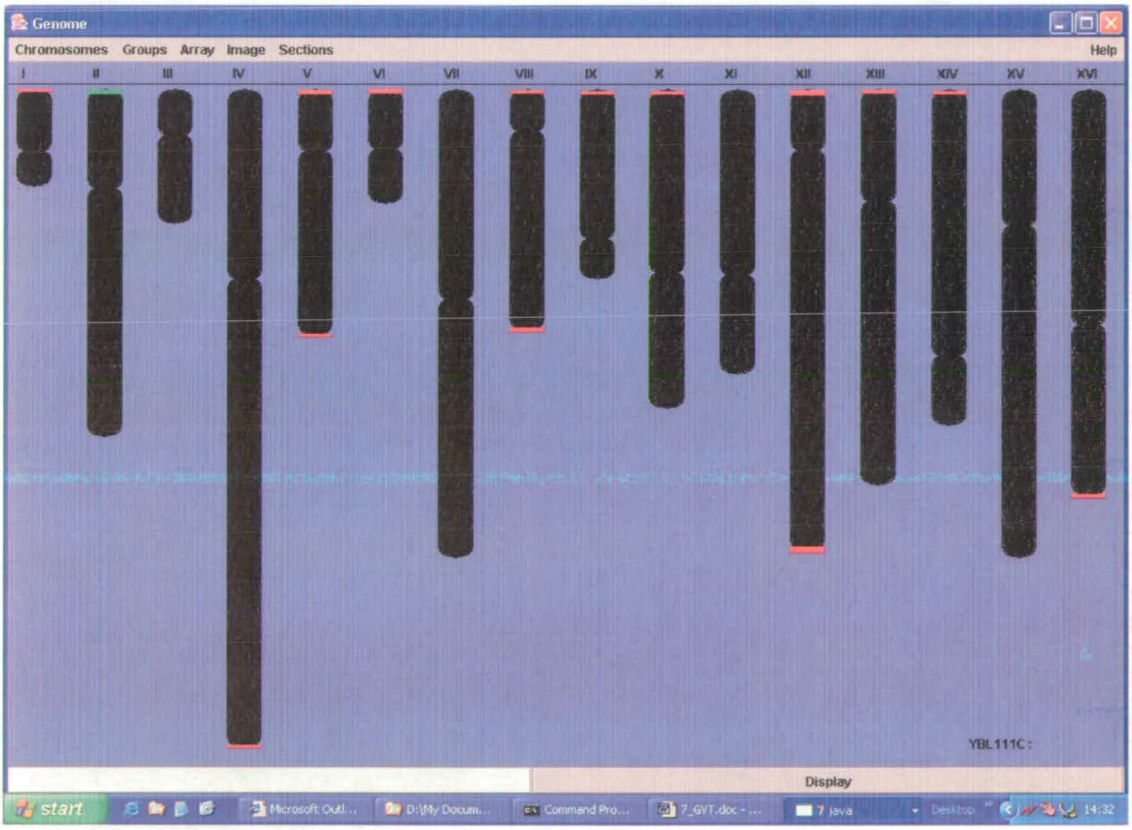
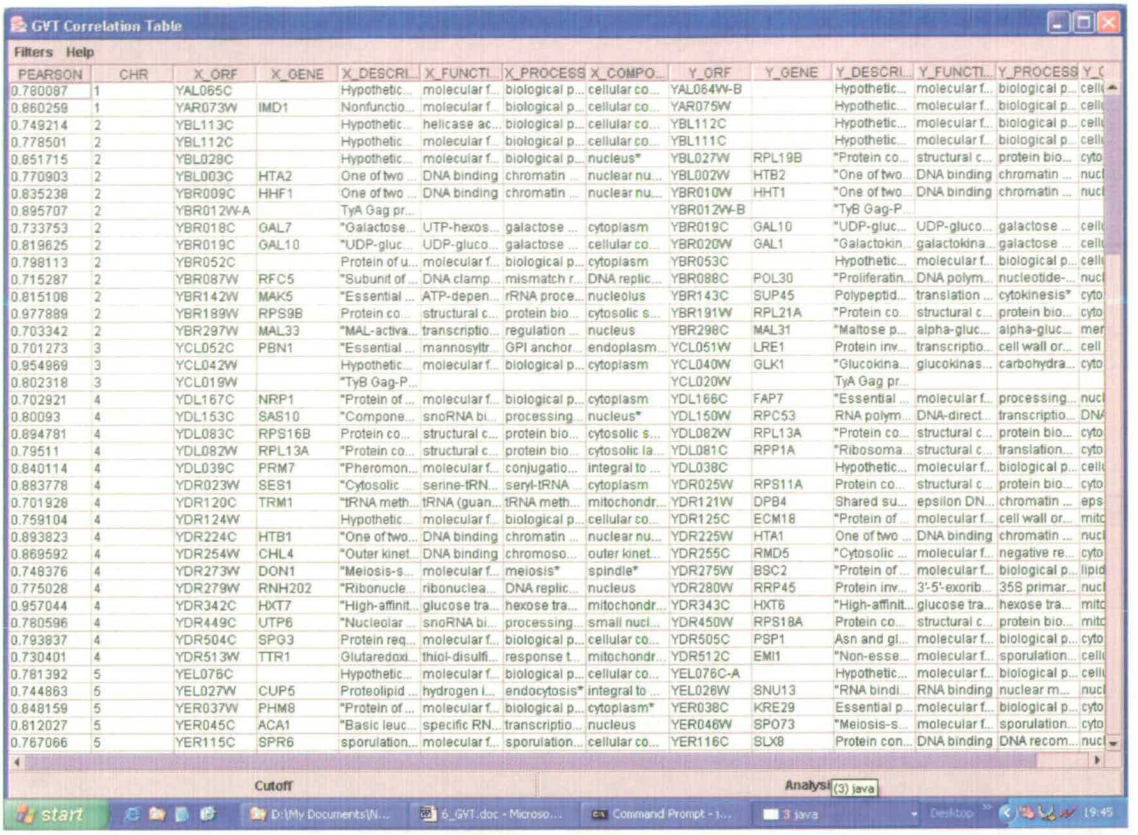


Figure 7.12: The genomic location of the helicase gene expression cluster genes
This is a screenshot of the Genome Section of YETI where the location of all the genes located in the same region of the gene expression hierarchical tree as YBL113C, YBL112C and YBL111C are highlighted on the genome schematic. YBL113C, YBL112C and YBL11C are highlighted in green and all other genes are highlighted in red. As can be seen, all the genes are located in the telomeric regions of the chromosomes.

7.5: All Coexpressed Adjacent ORFs

In addition to displaying the chromosome correlation map for a selected nuclear chromosome of *S. cerevisiae*, this section of YETI can also display a single data table containing information on all the coexpressed adjacent ORFs in the entire *S. cerevisiae* genome (Figure 7.13); the are a total of 158 coexpressed adjacent ORFs in the genome and a statistical analysis of whether this observed number is significant is presented below in section 7.6 of this chapter. This table contains the names and primary GO annotations for each pair of adjacent coexpressed ORFs along with the Pearson correlation coefficient of the pair; the ORF pairs displayed in the table are sorted by chromosome number followed by ORF order. Therefore, this table enables users to rapidly examine all the coexpressed adjacent ORFs in the genome, find larger coexpressed regions on the chromosomes and examine if there are any common GO annotations for coexpressed adjacent ORFs across the genome. The table also has a number of data filters to control which coexpressed adjacent ORFs are displayed which can help users find interesting regions exhibiting coexpression; they can be used, for example, to find all the coexpressed adjacent ORFs whose corresponding proteins also interact with one-another or all the coexpressed adjacent ORFs that are involved in the same biological process. Users are also able to lower the Pearson correlation coefficient cutoff value for coexpressed adjacent ORFs to be displayed in the table; the default cutoff value is 0.7. In addition, the table is linked to the Analysis Section enabling any ORF pairs of interest to be selected and investigated further in the other sections of YETI.



PEARSON	CHR	X_ORF	X_GENE	X_DESCRI...	X_FUNCTI...	X_PROCESS	X_COMPO...	Y_ORF	Y_GENE	Y_DESCRI...	Y_FUNCTI...	Y_PROCESS	Y_C
0.780087	1	YAL065C		Hypothetic...	molecular f...	biological p...	cellular co...	YAL064W-B		Hypothetic...	molecular f...	biological p...	cell
0.860259	1	YAR073W	IMD1	Nonfunctio...	molecular f...	biological p...	cellular co...	YAR075W		Hypothetic...	molecular f...	biological p...	cell
0.749214	2	YBL113C		Hypothetic...	helicase ac...	biological p...	cellular co...	YBL112C		Hypothetic...	molecular f...	biological p...	cell
0.778501	2	YBL112C		Hypothetic...	molecular f...	biological p...	cellular co...	YBL111C		Hypothetic...	molecular f...	biological p...	cell
0.851715	2	YBL028C		Hypothetic...	molecular f...	biological p...	nucleus*	YBL027W	RPL19B	"Protein co...	structural c...	protein bio...	cyto
0.770903	2	YBL003C	HTA2	One of two ...	DNA binding	chromatin ...	nuclear nu...	YBL002W	HTB2	"One of two...	DNA binding	chromatin ...	nucl
0.835238	2	YBR009C	HHF1	One of two ...	DNA binding	chromatin ...	nuclear nu...	YBR010W	HHT1	"One of two...	DNA binding	chromatin ...	nucl
0.895707	2	YBR012W-A		TyA Gag pr...				YBR012W-B		"TyB Gag-P...			
0.733753	2	YBR018C	GAL7	"Galactose...	UTP-hexos...	galactose ...	cytoplasm	YBR019C	GAL10	"UDP-gluc...	UDP-gluco...	galactose ...	cell
0.819625	2	YBR019C	GAL10	"UDP-gluc...	UDP-gluco...	galactose ...	cellular co...	YBR020W	GAL1	"Galactokin...	galactokina...	galactose ...	cell
0.798113	2	YBR052C		Protein of u...	molecular f...	biological p...	cytoplasm	YBR053C		Hypothetic...	molecular f...	biological p...	cell
0.715287	2	YBR087W	RFC5	"Subunit of ...	DNA clamp...	mismatch r...	DNA replic...	YBR088C	POL30	"Proliferatin...	DNA polym...	nucleotide-...	nucl
0.815108	2	YBR142W	MAK5	"Essential ...	ATP-depen...	rRNA proce...	nucleolus	YBR143C	SUP45	Polypeptid...	translation ...	cytokinesis*	cyto
0.977889	2	YBR189W	RPS9B	Protein co...	structural c...	protein bio...	cytosolic s...	YBR191W	RPL21A	"Protein co...	structural c...	protein bio...	cyto
0.703342	2	YBR297W	MAL33	"MAL-activa...	transcriptio...	regulation ...	nucleus	YBR298C	MAL31	"Maltose p...	alpha-gluc...	alpha-gluc...	mer
0.701273	3	YCL052C	PBN1	"Essential ...	mannosyltr...	GPI anchor...	endoplasm...	YCL051W	LRE1	Protein inv...	transcriptio...	cell wall or...	cell
0.954869	3	YCL042W		Hypothetic...	molecular f...	biological p...	cytoplasm	YCL040W	GLK1	"Glucokina...	glucokinas...	carbohydra...	cyto
0.802318	3	YCL019W		"TyB Gag-P...				YCL020W		TyA Gag pr...			
0.702921	4	YDL167C	NRP1	"Protein of ...	molecular f...	biological p...	cytoplasm	YDL166C	FAP7	"Essential ...	molecular f...	processing...	nucl
0.80093	4	YDL153C	SAS10	"Compone...	snoRNA bi...	processing ...	nucleus*	YDL150W	RPC53	RNA polym...	DNA-direct...	transcriptio...	DNA
0.894781	4	YDL083C	RPS16B	Protein co...	structural c...	protein bio...	cytosolic s...	YDL082W	RPL13A	"Protein co...	structural c...	protein bio...	cyto
0.79511	4	YDL082W	RPL13A	"Protein co...	structural c...	protein bio...	cytosolic la...	YDL081C	RPP1A	"Ribosoma...	structural c...	translation...	cyto
0.840114	4	YDL039C	PRM7	"Pheromon...	molecular f...	conjugatio...	integral to ...	YDL038C		Hypothetic...	molecular f...	biological p...	cell
0.883778	4	YDR023W	SES1	"Cytosolic ...	serine-tRN...	seryl-tRNA ...	cytoplasm	YDR025W	RPS11A	Protein co...	structural c...	protein bio...	cyto
0.701928	4	YDR120C	TRM1	"tRNA meth...	tRNA (guan...	tRNA meth...	mitochondr...	YDR121W	DPB4	Shared su...	epsilon DN...	chromatin ...	eps
0.759104	4	YDR124W		Hypothetic...	molecular f...	biological p...	cellular co...	YDR125C	ECM18	"Protein of ...	molecular f...	cell wall or...	mito
0.893823	4	YDR224C	HTB1	"One of two...	DNA binding	chromatin ...	nuclear nu...	YDR225W	HTA1	One of two ...	DNA binding	chromatin ...	nucl
0.869592	4	YDR254W	CHL4	"Outer kinet...	DNA binding	chromoso...	outer kinet...	YDR255C	RMD5	"Cytosolic ...	molecular f...	negative re...	cyto
0.748376	4	YDR273W	DON1	"Meiosis-s...	molecular f...	meiosis*	spindle*	YDR275W	BSC2	"Protein of ...	molecular f...	biological p...	lipid
0.775028	4	YDR279W	RNH202	"Ribonucle...	ribonuclea...	DNA replic...	nucleus	YDR280W	RRP45	Protein inv...	3'-5'-exorib...	35S primar...	nucl
0.957044	4	YDR342C	HXT7	"High-affinit...	glucose tra...	hexose tra...	mitochondr...	YDR343C	HXT6	"High-affinit...	glucose tra...	hexose tra...	mito
0.780596	4	YDR449C	UTP6	"Nucleolar ...	snoRNA bi...	processing ...	small nucl...	YDR450W	RPS18A	Protein co...	structural c...	protein bio...	mito
0.793837	4	YDR504C	SPG3	Protein req...	molecular f...	biological p...	cellular co...	YDR505C	PSP1	Asn and gl...	molecular f...	biological p...	cyto
0.730401	4	YDR513W	TTR1	Glutaredoxi...	thiol-disulfi...	response t...	mitochondr...	YDR512C	EMI1	"Non-esse...	molecular f...	sporulation...	cell
0.781392	5	YEL076C		Hypothetic...	molecular f...	biological p...	cellular co...	YEL076C-A		Hypothetic...	molecular f...	biological p...	cell
0.744863	5	YEL027W	CUP5	Proteolipid ...	hydrogen i...	endocytosis*	integral to ...	YEL026W	SNU13	"RNA bindi...	RNA binding	nuclear m...	nucl
0.848159	5	YER037W	PHM8	"Protein of ...	molecular f...	biological p...	cytoplasm*	YER038C	KRE29	Essential p...	molecular f...	biological p...	cyto
0.812027	5	YER045C	ACA1	"Basic leuc...	specific RN...	transcriptio...	nucleus	YER046W	SPO73	"Meiosis-s...	molecular f...	sporulation...	cyto
0.767066	5	YER115C	SPR6	sporulation...	molecular f...	sporulation...	cellular co...	YER116C	SLX8	Protein con...	DNA binding	DNA recom...	nucl

Figure 7.13: Screenshot of the Genome vs Transcriptome correlation table
This is a screenshot of the Genome vs Transcriptome correlation table which contains information on all the coexpressed adjacent ORFs in the *S. cerevisiae* genome.

Applying the GO biological process filter to the data table reveals that there are a number of common annotations among the adjacent coexpressed ORFs. For example, there are four instances of adjacent coexpressed ORFs involved in the biological process of 'chromatin assembly or disassembly'; further examination of these four pairs shows that all eight genes are histones. However, by far the most common biological process annotation is 'protein biosynthesis' which is typically accompanied by the 'structural constituent of ribosome' molecular function annotation; this molecular function annotation and the proteins associated with it are now discussed in further detail below.

7.5.1: Structural Constituent of Ribosome

The most common GO molecular function annotation represented in the table containing all the coexpressed adjacent ORFs in the *S. cerevisiae* genome was the 'structural constituent of ribosome' annotation; this annotation is typically accompanied with the 'protein biosynthesis' biological process annotation. Out of a total of 158 coexpressed adjacent ORFs found in the genome, there were 24 cases where at least one of the two adjacent ORFs were characterised with the 'structural constituent of ribosome' annotation; furthermore, there were 9 cases where both coexpressed adjacent ORFs were characterised with this annotation and five cases of triplets of coexpressed adjacent ORFs involving this annotation.

There were three cases where an ORF of unknown function was coexpressed with an adjacent 'structural constituent of ribosome' ORF; specifically: YBL028C, YKL137W and YLR063W. As these ORFs are colocated and coexpressed with a 'structural constituent of ribosome' gene, they could well be involved in a related biological process. Indeed, YETI shows that YBL028C is located in the nucleolus and is highly coexpressed with many genes characterised with the 'ribosomal large subunit biogenesis' biological process annotation; furthermore, all of these genes are also characterised as being located in the nucleolus which further suggests that YBL028C could well be involved in this biological process. In addition, YETI shows that YKL137W is coexpressed with just three genes all of which are 'structural

constituent of ribosome' genes while YLR063W is coexpressed with many genes involved in various tRNA and rRNA biological processes.

Due to the frequency of the 'structural constituent of ribosome' annotation appearing, this annotation was investigated in further detail using YETI. The 'structural constituent of ribosome' molecular function annotation alone does not reveal the whole story, as within this group are two sub-groups defined by the accompanying cellular component annotations: (1) Cytosolic Group: consisting of the 'cytosolic small ribosomal subunit (sensu Eukaryota)' [63 ORFs] and the 'cytosolic large ribosomal subunit (sensu Eukaryota)' [93 ORFs]; and (2) Mitochondrial Group: consisting of the 'mitochondrial small ribosomal subunit (sensu Eukaryota)' [35 ORFs] and the 'mitochondrial large ribosomal subunit (sensu Eukaryota)' [44 ORFs]. YETI can effectively be used to collectively investigate and compare the properties of these two groups by assigning the components of the cytosolic subunits to the red group and the components of the mitochondrial subunits to the green group. Interestingly, only 1 out of the 24 cases of coexpressed adjacent ORFs involved a component of a mitochondrial ribosomal subunit; the other 23 cases involved components of both the large and small cytosolic ribosomal subunits. Further examination of this observation in YETI reveals that very few components of the mitochondrial subunits are located next to each other compared to a number of neighbouring cytosolic subunit components.

The Transcriptome Section, displaying the Gasch *et al.* (2000) data set, shows that a large number of the cytosolic subunit components are located in the same region of

the hierarchical tree forming a tight gene expression cluster (Figure 7.14); this cluster contains components of both the large and small cytosolic ribosomal subunits. A similar observation was also made by Gasch *et al.* (2000) who reported a large cluster of genes whose expression was repressed in the majority of environment stress conditions studied; this cluster was found to consist almost entirely of genes encoding ribosomal proteins, however, there was no mention of the fact that they were all components of the cytosolic ribosomal subunits in this study. In this instance, YETI can be used to investigate what other genes are located in this cluster which could allow a biological role for any genes of unknown function to be inferred. YETI shows that the other known genes in the cluster are characterised with molecular function annotations related to the ribosome such as 'translation initiation factor activity', 'RNA binding' and 'uracil phosphoribosyltransferase activity'. There are also three genes of unknown function in this region, namely YKL056C, YMR116C and YNL119W; therefore, these genes could well be components of the cytosolic ribosomal subunits or involved in a related biological process. Indeed, although YMR116C is characterised with a set of unknown GO annotations, it is described as a core component of the ribosome. In contrast to components of the cytosolic subunits, the mitochondrial subunit components form a number of much smaller clusters dispersed fairly evenly throughout the hierarchical tree; however, there is no cluster any where near the size of that observed for the cytosolic subunits.

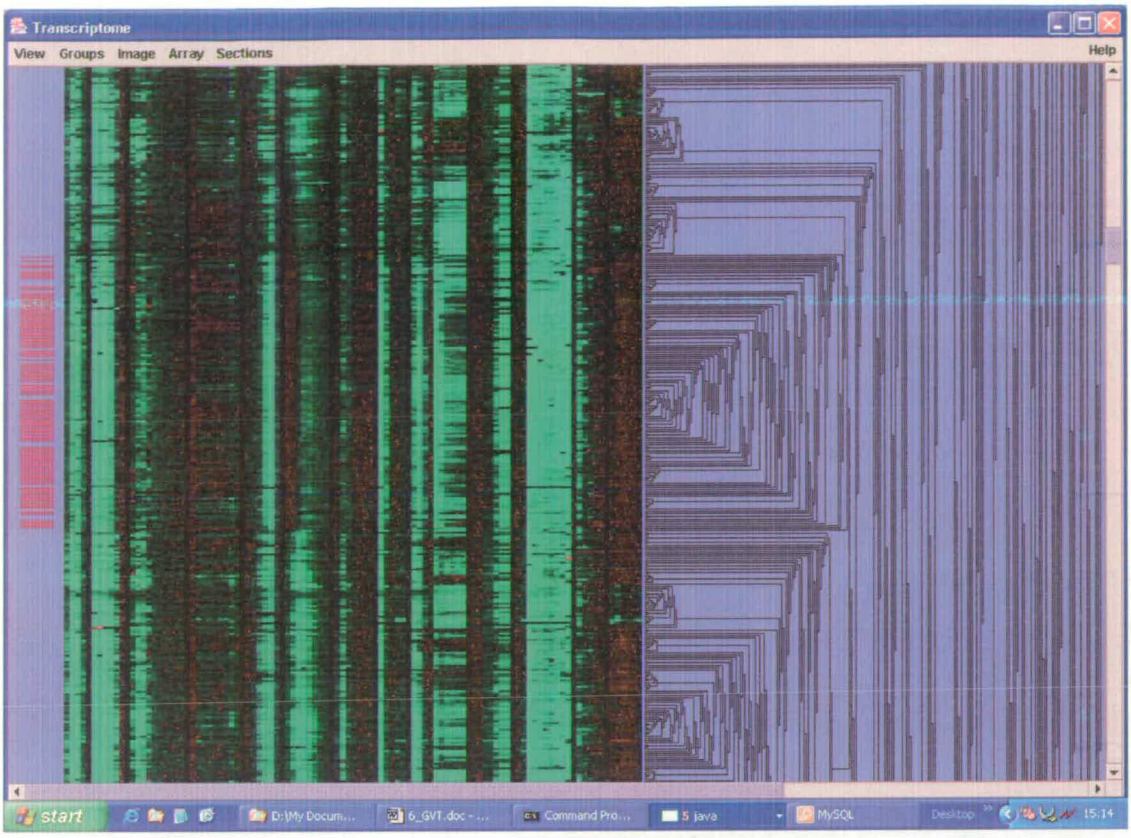


Figure 7.14: The cytosolic ribosomal subunit region of the gene expression hierarchical tree
This is a figure of the region of the gene expression hierarchical tree where a large cluster of cytosolic ribosomal subunit genes (highlighted in red) is observed.

Interestingly, the Proteome Section appears to give contrasting observations to those made in the Transcriptome Section. The Proteome Section shows that the components of the mitochondrial subunits interact highly with one another forming two large connected interaction clusters whereas the components of the cytosolic subunits form a number of much smaller clusters. The interactions of the mitochondrial subunit components were examined further in YETI by assigning genes characterised with the 'mitochondrial large ribosomal subunit' and 'mitochondrial small ribosomal subunit' annotations to the red and green groups, respectively. As can be seen in Figure 7.15, two distinct but connected clusters of proteins are formed. The first cluster is formed by the components of the large

subunit interacting highly with one another and the second cluster is formed by the components of the small subunit also interacting highly with one another. The two clusters are connected by a number of common (non-subunit) proteins such as TIF4631 (translation initiation factor activity) and PRP31 (RNA splicing factor activity) both of which are located in the mitochondrion. There are a number of proteins of unknown function in the two clusters but these tend to interact with just a single ribosomal subunit component; if an unknown protein interacted with many (as opposed to one) ribosomal subunit components this would increase the likelihood that it was involved a related biological process.

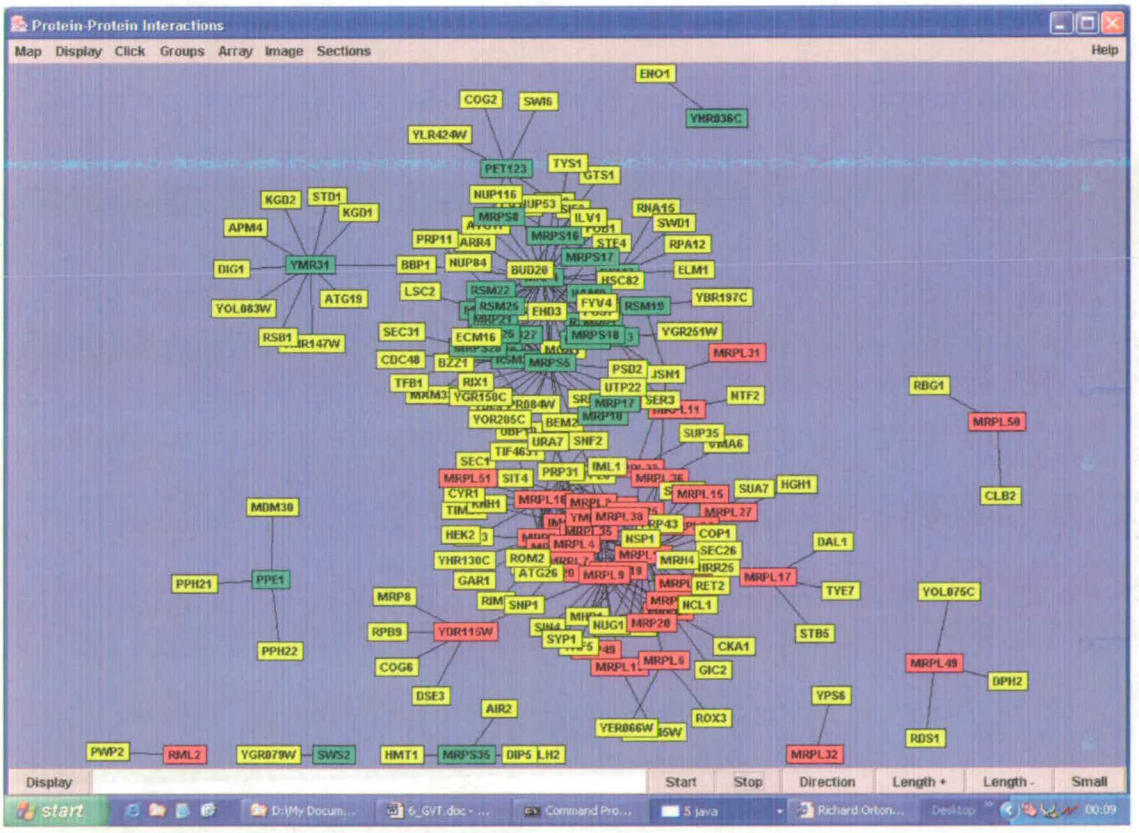


Figure 7.15: Protein-Protein interactions of the small and large mitochondrial ribosomal subunits
This is a figure of all the protein-protein interactions involving all of the small and large mitochondrial ribosomal subunit proteins. Proteins of the large and small ribosomal subunits are highlighted in red and green, respectively. As can be seen the small and large subunits form two distinct but connected interaction cluster.

This case study highlights one of the main advantages of YETI in that it enables the properties of an entire group of functionally related genes to be analysed collectively. This enables user to investigate the dynamics of how they are working together in order to achieve their biological goal and to also examine what other genes or proteins they may working with. As shown above, this can lead to potential biological roles being inferred for genes of unknown function through association with the functional group. Furthermore, this case study demonstrates how YETI can be used to compare the properties of multiple groups; for example, subunits of the same overall complex. In this case study, YETI shows that the cytosolic and mitochondrial ribosomal subunit components appear to have different properties. The majority of cytosolic subunit components are highly coexpressed but do not interact highly with one another. In contrast, the mitochondrial subunit components interact highly with one another but are coexpressed in a number of small clusters. However, the reasons for these differing observations for the cytosolic and mitochondrial subunits is not clear as it would seem likely that subunit components would need to be both coexpressed and be able to interact with one another in order to achieve their biological goals. The observations presented here could be explained by poor interaction data (with false-negatives and poor coverage concerning the cytosolic subunits) or poor expression data (where the conditions studied were not suitable to bring out the potential expression relationships between the mitochondrial subunit components). Alternatively, perhaps the mitochondrial subunit components are only needed in certain combinations for certain conditions.

7.6: Correlation Analysis Results

The Genome vs Transcriptome Section of YETI showed that there are a total of 158 coexpressed adjacent ORFs in the *S. cerevisiae* genome, using the Gasch *et al.* (2000) gene expression data set (Table 7.1). To test whether this observed number was statistically significant it was compared to the number expected derived from a control set of non-adjacent ORFs using the standard cumulative binomial distribution (http://mathworld.wolfram.com/BinomialDistribution.html; Figure 7.16). The P-value obtained for the 158 observed coexpressed adjacent ORFs using the cumulative binomial distribution was 2.95E-46 which suggests that these results are statistically significant. This statistically significant number of observed coexpressed adjacent ORFs is probably to be expected given that Cohen *et al.* (2000) performed their correlation analysis with three different gene expression data sets (cell cycle: Cho *et al.*, 1998; sporulation: Chu *et al.*, 1998; pheromone: Roberts *et al.*, 2000) and reported statistically significant observed numbers in each case. Therefore, the results presented here further suggest that adjacent ORFs in the *S. cerevisiae* genome are more likely to be coexpressed with one another than non-adjacent ORFs. Furthermore, other studies that have combined DNA sequence and expression data have also revealed the existence of chromosomal domains of similarly expressed genes in several other organisms such as *Drosophila melanogaster* (Spellman *et al.*, 2002), *Caenohabditis elegans* (Lercher *et al.*, 2003) and *Arabidopsis thaliana* (Ren *et al.*, 2005).

Category	Number
Total	6,919
Total Pairs	6,903
Total Pairs with Expression Data	4,926
Total Coexpressed Pairs	158
P-Value	2.95E-46

Table 7.1: Genome vs Transcriptome correlation analysis results
This table contains an overview of the Genome vs Transcriptome correlation analysis results. 'Total' corresponds to the total number of genomic features (e.g. ORFs as well as [amongst others] tRNAs, rRNAs and centromeres) currently on the 16 nuclear chromosomes of *S. cerevisiae*; this number excludes dubious ORFs which are highly unlikely to be real genes. 'Total Pairs' corresponds to the total number of pairs of adjacent genomic features. 'Total Pairs with Expression Data' corresponds to the total number of pairs with expression data available in the Gasch *et al.* (2000) study. 'Total Coexpressed Pairs' corresponds to the total number of pairs with expression data that have a Pearson correlation coefficient equal to or above 0.7. 'P-Value' corresponds to the probability of obtaining at least the observed number of coexpressed pairs calculated using the cumulative binomial distribution (Figure 7.16).

$$P(i >= i_0) = \sum_{i=i_0}^{I} p^i (1-p)^{I-i} \left[\frac{I!}{I!(I-i)!} \right]$$

Figure 7.16: Cumulative binomial distribution
This figure shows the cumulative binomial distribution equation used to calculate the probability of obtaining at least the observed number of coexpressed adjacent ORFs by chance. In this case: I = the total number of adjacent ORFs with expression data available analysed; i_0 = the observed number of coexpressed adjacent ORFs; and p = the observed probability of two randomly picked non-adjacent genes having a Pearson correlation coefficient equal to or above 0.7. The observed probability was calculated by generating a control set of 4,926 pairs of non-adjacent ORFs (with expression data) and counting the number of pairs with a Pearson correlation coefficient equal to or above 0.7. A total of ten control sets of 4,926 pairs of non-adjacent ORFs were generated and the number of coexpressed pairs was found to range from 30 to 46 with an average of 39.6 which gives an observed probability of 39.6/4926 = 0.00804.

7.7: Discussion

Chromosome correlation maps (Cohen *et al.*, 2000) enable the visualisation of coexpressed genes along the chromosomes of *S. cerevisiae* and enable users to find chromosomal regions exhibiting coexpression. Although the concept of chromosome correlation maps is by no means new, the main advantage that YETI offers is that

these maps are fully integrated with the rest of the system. This means that if a region of interest is found on a specific chromosomal correlation map it can easily be selected enabling all of the genes within this region to be collectively investigated in further detail in the other sections of YETI; this enables users to examine if and how the selected genes are working together to in order to achieve their biological goals and to also examine what other genes/proteins they may be working with.

The case studies presented in this chapter not only demonstrate the usefulness of chromosome correlation maps in identifying chromosomal regions exhibiting coexpression but also highlight the utility of YETI as a tool to investigate the functions of the genes located within these regions. Although the galactose metabolism case study does not necessarily reveal anything new about this biological process, the fact that YETI was able to easily and rapidly identify the majority of this pathway based on the experimental data could be seen as confirmation that the system strategy works; with the strategy being the ability to select an initial feature of interest and then move through the data sets to see what else can be associated with it. Furthermore, in both the galactose and allantoin case studies YETI was able to identify the majority of the actual pathway components and their associated transcriptional regulators; this could suggest that YETI has a potential use in identifying gene regulatory networks. Furthermore, *S. cerevisiae* is one of the most well studied organisms, therefore if YETI could be applied to a less well studied organism will a fully sequenced genome, expression data and perhaps interaction data it has the potential to yield many interesting observations.

The case studies and analyses presented above were all based on observations made using a specific gene expression data set (Gasch *et al.*, 2000); different gene expression data sets could well highlight different chromosomal regions of coexpression. Indeed, Cohen *et al.* (2000) analysed three different gene expression data sets and reported that the coexpression of an adjacent pair of genes in one data set was not predictive of its coexpression in the other data sets. Therefore, YETI has the potential to highlight many more chromosomal regions of coexpression through the analysis of additional gene expression data sets; any interesting regions that are found can then be investigated in further detail in YETI as demonstrated in the case studies. To the best of our knowledge, Genesis (Sturn *et al.*, 2002) is the only other software tool capable of integrating gene location and gene expression data to generate chromosome correlation maps. However, although Genesis is an effective tool for the visualisation and analysis of gene expression data it does not currently consider protein-protein interaction data. Therefore, as the chromosome correlation maps are effectively integrated into the entire YETI system, any interesting regions that are found can be thoroughly investigated in all the other sections of YETI.

Chapter 8

Discussion

8.1: The Yeast Exploration Tool Integrator

Over the past few years there has been a relative explosion of data in the biological sciences. At the heart of this data explosion is the budding yeast *Saccharomyces cerevisiae* (*S. cerevisiae*) which is one of the most widely studied eukaryotes due to its value as a model organism in biological research; it has a fully sequenced genome that is well annotated and a variety of publicly available functional genomic data sets. Analysis of this vast amount of data is a key challenge and computers in conjunction with effective software tools are an essential part of this process. There has been a rapid increase in the number of software tools available for the visualisation and analysis of individual types of functional genomic data sets. However, there are relatively few tools available that are capable of bringing together a number of different types of data sets for integrated visualisation and analysis. As many new biological insights are likely to emerge from the combined use of data from different functional genomic strategies, there is a need for a new generation of software tools that are capable of effectively utilising the wealth of data available for *S. cerevisiae* enabling users to perform integrative analyses.

The Yeast Exploration Tool Integrator (YETI) is a novel bioinformatics tool for the integrated visualisation and analysis of *S. cerevisiae* functional genomic data sets. The YETI system consists of a database for the storage and management of data and a Java program for the integrated visualisation and analysis of data. YETI utilises publicly available data sets from a number of different functional genomic strategies, such as gene expression microarrays and yeast two-hybrid screens, and provides an

effective means for their integrated visualisation and analysis. YETI consists of a number of individual sections for the visualisation and analysis of functional genomic data sets which are closely inter-linked enabling users to swiftly move between them and investigate all aspects of any genes or proteins of interest as well as providing access to textual information, including Gene Ontology (GO) annotations, at any point. YETI enables users to easily explore the data in an integrated modular fashion, investigate the intricacies of broad biological processes and test specific hypotheses.

The main advantages of YETI are its ease of use and its group approach for analysis combined with its inter-linked sections. YETI was designed with simplicity in mind with simple navigation mechanisms to move through the program, flexible search mechanisms and clear graphical representations of the data in unison with a number of advanced features and functionality. The inter-linked sections effectively integrate a number of functional genomic data sets together enabling users to swiftly move between data sets and investigate all aspects of any features of interest. The group approach enables all the proteins involved in an entire biological process to be collectively examined as a whole to investigate the dynamics of how they are working together to achieve their biological goal and to also examine what other proteins they may be working with.

8.2: Case Studies and Analyses

A number of case studies were presented throughout this thesis which demonstrated the potential and utility of YETI in both single gene and group investigations. Firstly, a number of single gene case studies were presented which demonstrated how YETI could be used to investigate a potential function for a gene of unknown function. In actual fact, an associated computer program originally suggested a potential biological process for all the unknown genes and YETI was subsequently used to test these hypotheses and to try and associate the gene with the suggested biological process. Secondly, a number of much broader case studies were presented which investigated the properties of groups of genes highlighted from the correlation analyses. In addition to demonstrating the utility of YETI, these case studies also resulted in the prediction of potential functions for a number of genes of unknown function; an overview of all the functional predictions of all these case studies is presented in Table 8.1. These functional predictions are fairly tentative and, as always, need thorough validation through experiments in the laboratory.

Case Study	Predictions
5.2	MOH1 'negative regulation of gluconeogenesis'
5.3	YKL056C, YMR116C, YMR321C, YJR124C, YJL193W and YBR025C 'structural constituent of ribosome' 'protein biosynthesis' 'cytosolic small ribosomal subunit (sensu Eukaryota)' or 'cytosolic large ribosomal subunit (sensu Eukaryota)'
5.4	'YMR148W' 'aerobic respiration' or 'mitochondrial electron transport chain'
5.5	'YLR364W' 'sulphur metabolism'
5.6	'IES5' 'chromatin remodelling' 'INO80 Complex'.
6.6	SNZ2 and SNZ3 'nucleus'
7.4.2	YBR147W, YDL183C, YGR125W, YIL165C, YLR053C and YLR364W 'nitrogen compound metabolism' or 'sulphur metabolism' YIR042C 'allantoin degradation'
7.4.3	YBL113C, YBL112C and YBL111C 'helicase activity' or 'DNA helicase activity'
7.5.1	YBL028C 'ribosomal large subunit biogenesis' YKL056C, YMR116C and YNL119W 'structural constituent of ribosome' 'protein biosynthesis' 'cytosolic small ribosomal subunit (sensu Eukaryota)' or 'cytosolic large ribosomal subunit (sensu Eukaryota)'

Table 8.1: Functional predictions of all case studies
This table contains all of the functional predictions made through all of the case studies. The case study number corresponds to the section number of this thesis that the case study is presented in.

The broader case studies all highlight how the group approach combined with the inter-linked sections of YETI enables users to collectively investigate if and how a group of potentially related genes are working together in order to achieve their biological goal and to also investigate what other genes/proteins they may be working with. Perhaps the best illustration of this is case study '7.4.1: Galactose Metabolism'. In this case study, YETI was able to easily and rapidly identify the majority of this pathway, including transcriptional regulators, starting from just a triplet of coexpressed genes and simply based on qualitative exploration of the data. Furthermore, this case study is a good illustration of how all the sections of YETI can be used in conjunction with one another and how an investigation can be expanded out through them. A 'workflow' diagram of how YETI was used in this case study is presented in Figure 8.1; this diagram shows how the investigation progressed through the various sections of YETI and what additional genes were identified at each point and selected for further investigation.

These case studies also highlight the utility of some of the specific features and functions of YETI such as the genome schematic for investigating possible evolutionary relationships between two groups of genes and the chromosome window for investigating the similarity between chromosomal regions; a thorough discussion of the advantages and disadvantages highlighted is presented at the end of each case study. Overall, the case studies clearly show how YETI can easily and rapidly be used to investigate specific genes as well as groups of genes and that the effective integration of functional genomic data sets enabled many interesting

observations to be made. Furthermore, these case studies provide direct examples of the 'typical user questions' that YETI aims to address as detailed in Chapter 2.

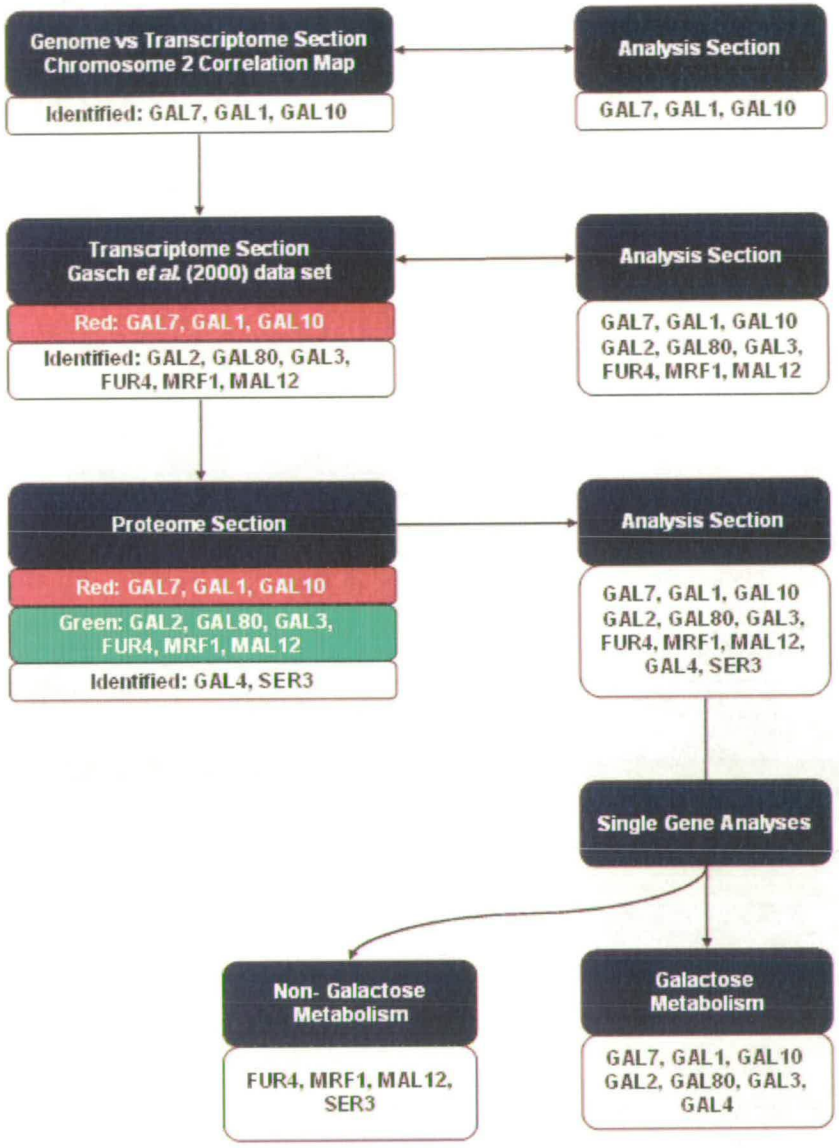


Figure 8.1: Workflow diagram for galactose metabolism case study
This is a workflow diagram which details how YETI was used in the galactose metabolism case study; boxes highlighted in red and green represent what genes were assigned to the red and green groups, respectively, at each stage. Initially, YETI highlighted a triplet of coexpressed genes on the correlation map of chromosome 2; further examination revealed these three genes were all involved in galactose metabolism. The Transcriptome Section showed that these three genes were located in the same region of the hierarchical tree; selecting and investigating the surrounding genes revealed the presence of more genes involved in galactose metabolism. The Proteome Section then showed that a number of the proteins involved in galactose metabolism interacted with one another and also revealed the presence of yet another protein involved in galactose metabolism. Examination of individual Pearson correlation coefficients and interaction confidence scores enabled some associated proteins to be ruled out leaving the core galactose metabolism proteins. A thorough description of this case study can be found in section 7.4.1 of this thesis.

One interesting and important observation is that there appears to be a conflict between the results of the Genome vs Transcriptome correlation analysis and the Genome vs Proteome correlation analysis. The Genome vs Transcriptome correlation analysis indicated that adjacent genes are more likely to be coexpressed with one another than non-adjacent genes. This analysis indicated that there was a statistically significant number of cases in the *S. cerevisiae* genome where adjacent genes are coexpressed. This colocation and coexpression suggests that these adjacent genes are likely to be involved in the same or a related biological process (the concept of guilt by association); indeed, numerous cases of adjacent genes involved in the same overall biological process were highlighted in the case studies above. However, the Genome vs Proteome correlation analysis indicated that there was no tendency for the genes of interacting proteins to be located near each on the genome. In this analysis, only eight neighbouring genes were found to encode protein products that interact with one another. In summary, the Genome vs Transcriptome correlation analysis indicated (through coexpression) that there was a tendency for neighbouring genes to be functionally related while the Genome vs Proteome correlation analysis indicated (through interaction) that there was no tendency for neighbouring to be functionally related.

The Genome vs Transcriptome correlation analysis is unlikely to be incorrect given that it uses good quality data sets and that similar findings are reported elsewhere. The most likely cause of this conflict is an incomplete protein-protein interaction

data used in the Genome vs Proteome correlation analysis and a number of observations support this hypothesis:

1) There are specific interactions missing from the data set that have been reported in scientific studies, for example, Rodriguez-Navarro *et al.* (2002) reported that SNZ2 and SNZ3 could interact directly with THI11, as discussed previously in Chapter 6. This shows that the interaction data set is indeed incomplete from the sense that there are known interactions missing.
2) The 'galactose metabolism' and 'allantoin degradation' case studies showed that none of the core components of these pathways interacted with one another despite their colocation and coexpression. Although this observation could be real, it could also indicate an incomplete interaction data set.
3) The 'structural constituent of ribosome' case study showed that very few of the cytosolic subunit components interacted with one another while the mitochondrial subunit components interacted highly with one another. Although this observation could also be real, it again suggests and incomplete data set.
4) As discussed in Chapter 7, four pairs of coexpressed adjacent histone genes were identified in the genome. However, given that histones almost certainly have to interact with one another in order to form nucleosomes and higher order chromosomal structures, only one of these four pairs are reported to interact with one another. Again, this suggests the possibility of an incomplete data set.

5) Out of the 158 coexpressed adjacent ORFs found in the genome, only one of these pairs encode proteins reported to interact with one another (the histones HTA1-HTB1); intuition suggests that there should be more.

Therefore, the above observations strongly suggest that the protein-protein interaction data set is incomplete. Indeed, various studies have estimated that the ~6,000 *S. cerevisiae* proteins are connected by as many as 40,000 interactions (Wallhout *et al.*, 2000; Tucker *et al.*, 2001; Grigoriev *et al.*, 2003; Uetz *et al.*, 2005). However, the YETI database currently only stores 12,866 unique interactions. Furthermore, this represents the unfiltered data set which is therefore likely to contain many false-positives; the source of this data set is the (GRID; Breitkreutz *et al.*, 2003) database which contains the interactions from many high and low-throughput interaction studies. The incomplete protein-protein interactions is also highlighted by that fact that there is a lack of overlap between the different high-throughput data sets themselves and also with published low-throughput studies which are generally considered to be less prone to false positives and false negatives (Ito *et al.*, 2001; Grunenfelder *et al.*, 2002; Cornell *et al.*, 2004; Uetz *et al.*, 2005). Taken together, this not only suggests that new or improved technologies are needed but also that more interactions could be detected by more exhaustive application of current techniques. Therefore, a more complete protein-protein interaction data set could well give better results for the Genome vs Proteome correlation analysis and eliminate the observed conflict with the Genome vs Transcriptome; this more complete data set will hopefully come with time.

8.3: Improvements to YETI

The obvious improvement that can be made to YETI is higher quality data sets as well as more data sets. The protein-protein interaction data set used in YETI (Breitkreutz *et al.*, 2003) was shown to be incomplete in the case studies and many of the interactions are derived from the yeast two-hybrid technique which is renowned for false-positive errors. Therefore, more protein-protein interaction data sets are needed in conjunction with effective confidence scores to assess their reliability; currently, YETI does apply a number of confidence scores to interactions on importation (such as times reported, cellular location and expression) and more data sets should hopefully come in time. The gene expression data currently utilised in YETI comes from two gene expression microarray studies (Gasch *et al.*, 2000; Gasch *et al.*, 2001) which monitor how *S. cerevisiae* cells respond to a wide variety of environmental conditions and DNA damaging agents. As different data sets are likely to highlight different relationships among the genes of *S. cerevisiae* it would now be useful to incorporate more data sets into the system giving users a choice of which expression data is considered in their investigations. The genome data set currently utilised in YETI comes from the *Saccharomyces* Genome Database (SGD; Cherry *et al.*, 1998) which contains descriptions and annotations on all the genes in *S. cerevisiae*. However, additional information can often be obtained from the other major yeast databases such as MIPS (Mewes *et al.*, 1998) and this additional information could be integrated into the YETI system in the future. Furthermore, the scientific literature contains a wealth of useful information such as reported protein interactions and functional predictions of unknown genes; therefore, the integration

of text mining technologies that can automatically identify and extract this information could be a useful development.

One major addition that could be made to YETI is the introduction of sequence data and sequence analysis techniques into the system. This would be particularly useful when investigating the function of an unknown gene as it would enable users to examine what other genes in *S. cerevisiae* (and potentially other organisms) the unknown gene is related to in sequence and also enable them to examine what functional domains the gene's encoded protein contains. Furthermore, sequence data would enable users to investigate if specific regions of DNA are duplicated and if specific genes share similar promoter regions. In addition, groups of genes related by sequence could be constructed and then be collectively investigated in further detail in YETI.

Another improvement that could be made to YETI is in the way it handles and using GO annotations. Firstly, the FPC Section of YETI simply displays an alphabetical list of all the GO annotations used to characterise the genes of *S. cerevisiae* enabling users to find and subsequently select annotations of interest. However, a graphical representation (for example, see AmiGO; http://www.godatabase.org/) would enable users to browse the GO annotation system and examine the relationship between terms. Furthermore, this would enable users to construct much broader groups of functionally related genes through the selection of high level terms which would also result in the selection of all lower (or child) terms stemming from it. Secondly, when comparing the annotations of genes YETI simply checks if they share the same GO

annotations. For example, the Datasheet Window of YETI provides links to the Analysis Section that enable users to view all the other genes characterised with the same GO annotations. Although this is an essential feature it would also be useful to enable users to examine genes characterised with similar or related annotations and there a number of techniques that can measure the distance in nodes or semantic similarity between annotations (for example: see Lord *et al.*, 2003).

8.4: Extensions to YETI

As discussed in this thesis, we have so far performed 'Genome vs Transcriptome' and 'Genome vs Proteome' correlation analyses and specific correlation sections of YETI were developed to facilitate these investigations. Therefore, the one remaining pair-wise correlation analysis to be performed is between the Proteome and Transcriptome to investigate if there is a tendency for proteins that interact with one another to be encoded by genes that are coexpressed. Currently, there is a relatively simple Proteome vs Transcriptome correlation section in YETI (Figure 8.2). This section simply displays a data table containing information on all the protein-protein interactions whose corresponding genes are coexpressed. The data table contains a number of filters to control what types of protein-protein interactions are displayed and is also linked to the Analysis Section enabling any interactions of interest to be selected and investigated in further detail in the other YETI sections. However, this section is relatively simple and needs further development; for example, an additional visualisation layer on top of the data table could enable correlations between the data sets to be investigated more easily. A number of analyses

investigating correlations between protein interaction and gene expression have already been performed (for example: Ge *et al.*, 2001; Grigoriev *et al.*, 2001; Mrowka *et al.*, 2001; Jansen *et al.*, 2002; Kemmeren *et al.*, 2002). However, the advantage that YETI would offer is that any features highlighted through the analysis could be immediately be investigated in further detail in the other sections.

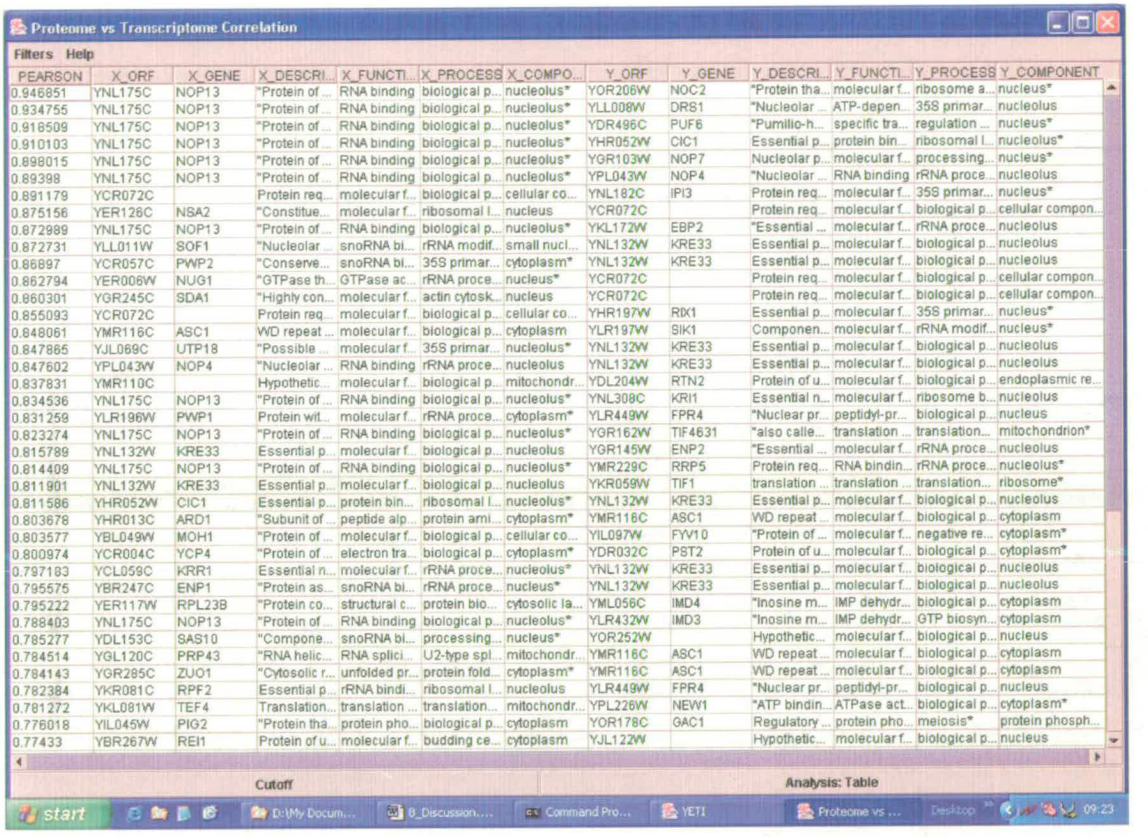


Figure 8.2: Screenshot of the Proteome vs Transcriptome Section
This is a screenshot of the Proteome vs Transcriptome section which currently displays a simple data table containing information on all the protein-protein interactions whose corresponding genes are coexpressed. The table contains a range of information on the interacting proteins such as descriptions and GO annotations as well as the Pearson correlation coefficient between the two corresponding genes.

Although YETI was initially designed for the budding yeast *S. cerevisiae*, it was designed to be a flexible system that could be applied to other organisms with relative ease. The key to this application is the availability of an equivalent genome

data set which is the core data set of the YETI system as it contains the names, locations and descriptions of all the genes present in an organism and it is this data set that links all the other data sets in the system together. If an equivalent genome data set is available, the YETI system could be ported to virtually any other organism with only slight modifications to the program code and the underlying database structure. The Entrez Genome database (Schuler *et al.*, 1996) contains data files for a large number of organisms that can be used as an ideal basic genome data set in YETI. To demonstrate this, a new version of YETI called YETI-O was created which is concerned with the visualisation and analysis of the bacterial genomes available from the Entrez Genome database. To date, YETI-O has been successfully applied to four bacteria (*Bacillus subtilis*, *Escherichia coli*, *Haemophilus influenzae* and *Shewanella oneidensis*) but could be applied to many more with relative ease. Currently, the YETI-O program only has the Analysis and Genome Sections available to the user (Figure 8.3). However, the Transcriptome and Proteome Sections of the original YETI program can also be ported across to YETI-O with relative ease. Only slight modifications to the YETI program code and the underlying database structure would be needed to do this as long as similar gene expression and protein-protein interaction data sets were publicly available. Therefore, YETI has the potential to be a useful tool for many other researchers interested in exploring the functional genomic data sets of other organisms.

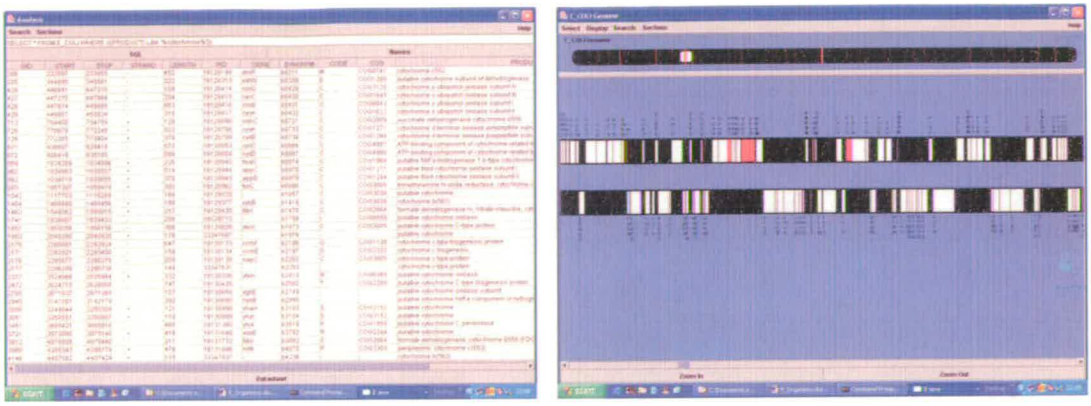


Figure 8.3: Screenshots of YETI-O
These are screenshots of the Analysis (left) and Genome (right) Sections of YETI-O. In this case, the Analysis Section was used to perform a 'cytochrome' keyword search on gene descriptions and the Genome Section was subsequently used to highlight their locations on the chromosomal display.

8.5: Comparison with Other Tools

Essentially, each of the sections of YETI can be viewed as a distinct software tool. The Genome Section can be viewed as a genome and chromosome browser, the Transcriptome Section as a program for the visualisation of gene expression data, and the Proteome Section as a program for the visualisation of protein-protein interactions. There are a number of more advanced tools available to users when the specific sections of YETI are considered individually; for example: Genesis (Sturn *et al.*, 2002) is a more advanced tool for the visualisation and analysis of gene expression data when compared to the Transcriptome Section; Cytoscape (Shannon *et al.*, 2003) is a more advanced tool for the visualisation and analysis of protein-protein interactions when compared to the Proteome Section; and Ensembl (Hubbard *et al.*, 2002) is a more advanced genome browser than the Genome Section. However, the real advantage that YETI offers is that all of these tools (or sections)

are effectively inter-linked together enabling users to seamlessly move between them and investigate all aspects of any features of interest.

Perhaps the most similar tool to YETI is the Genome Information Management System (GIMS; Cornell *et al.*, 2003) which is an object database that integrates genomic data with data on the transcriptome, protein-protein interactions, metabolic pathways and GO annotations. GIMS is a much more powerful analysis tool than YETI as it enables users to perform complex queries over multiple data types; for example, users can retrieve all the mRNAs with a given cellular location that were upregulated by at least a given amount in a given experiment. Although YETI has a number of effective search mechanisms, this type of complex query is currently beyond YETI as it can not utilise the microarray data, for example, in such a quantitative sense. However, YETI is a much more powerful exploration tool than GIMS as it enables users to easily and rapidly explore the data visually, or qualitatively, and select features of interest to investigate further.

Probably the most valuable resource available to *S. cerevisiae* researchers is the *Saccharomyces* Genome Database (SGD; Cherry *et al.*, 1998); indeed, this is where YETI currently gets its core genome data from. The SGD contains a vast amount of data on all the genes in *S. cerevisiae* and contains many useful links to various other scientific websites. However, the SGD still centres around a single gene approach and is primarily concerned with the dissemination of data as opposed to the integrated visualisation and exploration of functional genomic data sets. Essentially, there is nothing that can be done in YETI that can not be done with the existing

computational resources. However, what can be done in YETI almost instantaneously is often cumbersome to do using existing resources. For example, YETI can rapidly show the user if a particular group of genes are located at similar chromosomal locations; however, with existing resources each gene would have to be examined individually and their chromosomal locations investigated textually. Furthermore, YETI eliminates the need for users to visit multiple resources as everything is integrated together in one place. Therefore, we believe that YETI is a useful resource for researchers of *S. cerevisiae* which can take its place alongside the many other resources available; in other words, YETI is not intended as a replacement for any of the existing resources, rather it offers a novel way of exploring the existing data which can yield new interesting observations and hypotheses.

8.6: Conclusion

YETI, like all similar resources, is only as good as the data it uses. Therefore, the future of YETI very much depends on the data it uses. If new protein-protein interaction and gene expression data sets are not continually developed then we will probably fast approach a situation where nothing really new can be gained from the existing data no matter what novel visualisation and analysis techniques are developed. However, in the short term YETI has a solid future ahead of it as new protein interaction data sets will inevitably be produced owing to the fact that the existing interaction data set appears to be so incomplete; new gene expression data sets are also continually produced at present. Furthermore, YETI is a fairly flexible

system which can be expanded with relative ease; therefore, if any new functional genomic strategies are developed in the future YETI could be expanded to integrate this data as well.

Over the past one or two years the focus of research has shifted from bioinformatics to systems biology. Systems biology is concerned with the study of biological systems in terms of their underlying network structure rather than simply their individual molecular components. At first sight, YETI appears to fit in quite nicely as a systems biology tool as its group approach enables the properties of an entire system to be collectively investigated; as opposed to the standard single gene approach. However, the real power of systems biology comes with quantitative modelling techniques that are capable of predicting biological behaviour. Therefore, one could envisage YETI being expanded in the future to become a Systems Biology Markup Language (SBML; www.sbml.org) compatible tool that enables users to import SBML models to collectively investigate the properties of the model components in core sections of YETI and to also enable users to construct SBML models based on observations of biological processes made through using YETI.

In summary, this thesis has detailed the design and development of the Yeast Exploration Tool Integrator and has effectively demonstrated its use in a number of case studies.

Chapter 9

Bibliography

Aloy, P., B. Bottcher, et al. (2004). "Structure-based assembly of protein complexes in yeast." Science **303**(5666): 2026-9.
Aloy, P. and R. B. Russell (2002). "Interrogating protein interaction networks through structural biology." Proc Natl Acad Sci U S A **99**(9): 5896-901.
Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.
Ashburner, M., C. A. Ball, et al. (2001). "Creating the gene ontology resource: design and implementation." Genome Res **11**(8): 1425-33.
Bader, G. D., I. Donaldson, et al. (2001). "BIND--The Biomolecular Interaction Network Database." Nucleic Acids Res **29**(1): 242-5.
Bader, J. S., A. Chaudhuri, et al. (2004). "Gaining confidence in high-throughput protein interaction networks." Nat Biotechnol **22**(1): 78-85.
Barabasi, A. L. and Z. N. Oltvai (2004). "Network biology: understanding the cell's functional organization." Nat Rev Genet **5**(2): 101-13.
Barriot, R., J. Poix, et al. (2004). "New strategy for the representation and the integration of biomolecular knowledge at a cellular scale." Nucleic Acids Res **32**(12): 3581-9.
Bartel, P. L., J. A. Roecklein, et al. (1996). "A protein linkage map of Escherichia coli bacteriophage T7." Nat Genet **12**(1): 72-7.
Battista, G., P. Eades, et al. (1999). "Graph drawing: algorithms for the visualization of graphs."
Begley, T. J., A. S. Rosenbach, et al. (2002). "Damage recovery pathways in Saccharomyces cerevisiae revealed by genomic phenotyping and interactome mapping." Mol Cancer Res **1**(2): 103-12.
Berriz, G. F., J. V. White, et al. (2003). "GoFish finds genes with combinations of Gene Ontology attributes." Bioinformatics **19**(6): 788-9.
Blake, J. A., J. T. Eppig, et al. (2000). "The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse. The Mouse Genome Database Group." Nucleic Acids Res **28**(1): 108-11.
Blumenthal, T. (2004). "Operons in eukaryotes." Brief Funct Genomic Proteomic **3**(3): 199-211.
Boer, V. M., J. H. de Winde, et al. (2003). "The genome-wide transcriptional responses of Saccharomyces cerevisiae grown on glucose in aerobic chemostat cultures limited for carbon, nitrogen, phosphorus, or sulfur." J Biol Chem **278**(5): 3265-74.
Bork, P., L. J. Jensen, et al. (2004). "Protein interaction networks from yeast to human." Curr Opin Struct Biol **14**(3): 292-9.
Braun, E. L., E. K. Fuge, et al. (1996). "A stationary-phase gene in Saccharomyces cerevisiae is a member of a novel, highly conserved gene family." J Bacteriol **178**(23): 6865-72.
Brazma, A., P. Hingamp, et al. (2001). "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data." Nat Genet **29**(4): 365-71.
Brazma, A., H. Parkinson, et al. (2003). "ArrayExpress--a public repository for microarray gene expression data at the EBI." Nucleic Acids Res **31**(1): 68-71.
Breitkreutz, B. J., C. Stark, et al. (2003). "The GRID: the General Repository for Interaction Datasets." Genome Biol **4**(3): R23.

Brown, M. P., W. N. Grundy, et al. (2000). "Knowledge-based analysis of microarray gene expression data by using support vector machines." Proc Natl Acad Sci U S A **97**(1): 262-7.

Burns, N., B. Grimwade, et al. (1994). "Large-scale characterization of gene expression, protein localization and gene disruption in Saccharomyces cerevisiae." Genes & Dev **8**: 1087-1105.

Cherry, J. M., C. Adler, et al. (1998). "SGD: Saccharomyces Genome Database." Nucleic Acids Res **26**(1): 73-9.

Cho, R. J., M. J. Campbell, et al. (1998). "A genome-wide transcriptional analysis of the mitotic cell cycle." Mol Cell **2**(1): 65-73.

Chu, S., J. DeRisi, et al. (1998). "The transcriptional program of sporulation in budding yeast." Science **282**(5389): 699-705.

Cohen, B. A., R. D. Mitra, et al. (2000). "A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression." Nat Genet **26**(2): 183-6.

Cornell, M., N. W. Paton, et al. (2003). "GIMS: an integrated data storage and analysis environment for genomic and functional data." Yeast **20**(15): 1291-306.

Cornell, M., N. W. Paton, et al. (2004). "A critical and integrated view of the yeast interactome." Comp Func Gen **5**: 382-402.

Cornell, M., N. W. Paton, et al. (2001). "GIMS - a data warehouse for storage and analysis of genome sequence and functional data." Proc. 2nd IEEE International Symposium on Bioinformatics and Bioengineering (BIBE): 15-22.

Date, S. V. and E. M. Marcotte (2003). "Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages." Nat Biotechnol **21**(9): 1055-62.

Deane, C. M., L. Salwinski, et al. (2002). "Protein interactions: two methods for assessment of the reliability of high throughput observations." Mol Cell Proteomics **1**(5): 349-56.

Dennis, G., Jr., B. T. Sherman, et al. (2003). "DAVID: Database for Annotation, Visualization, and Integrated Discovery." Genome Biol **4**(5): P3.

Deutschbauer, A. M., R. M. Williams, et al. (2002). "Parallel phenotypic analysis of sporulation and postgermination growth in Saccharomyces cerevisiae." Proc Natl Acad Sci U S A **99**(24): 15530-5.

Drawid, A., R. Jansen, et al. (2000). "Genome-wide analysis relating expression level with protein subcellular localization." Trends Genet **16**(10): 426-30.

Dujon, B. (1996). "The yeast genome project: what did we learn?" Trends Genet **12**(7): 263-70.

Eades, P. (1984). "A Heuristic for Graph Drawing." Congressus Numerantium **42**: 149-160.

Eisen, M. B. and P. O. Brown (1999). "DNA arrays for analysis of gene expression." Methods Enzymol **303**: 179-205.

Eisen, M. B., P. T. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proc Natl Acad Sci U S A **95**(25): 14863-8.

Eisenberg, D., E. M. Marcotte, et al. (2000). "Protein function in the post-genomic era." Nature **405**(6788): 823-6.

Enright, A. J., I. Iliopoulos, et al. (1999). "Protein interaction maps for complete genomes based on gene fusion events." Nature **402**(6757): 86-90.
Fellenberg, M., K. Albermann, et al. (2000). "Integrative analysis of protein interaction data." Proc Int Conf Intell Syst Mol Biol **8**: 152-61.
Ferea, T. L., D. Botstein, et al. (1999). "Systematic changes in gene expression patterns following adaptive evolution in yeast." Proc Natl Acad Sci U S A **96**(17): 9721-6.
Fields, S. and O. Song (1989). "A novel genetic system to detect protein-protein interactions." Nature **340**(6230): 245-6.
Flores, A., J. F. Briand, et al. (1999). "A protein-protein interaction map of yeast RNA polymerase III." Proc Natl Acad Sci U S A **96**(14): 7815-20.
Foury, F., T. Roganti, et al. (1998). "The complete sequence of the mitochondrial genome of Saccharomyces cerevisiae." FEBS Lett **440**(3): 325-31.
Fromont-Racine, M., A. E. Mayes, et al. (2000). "Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins." Yeast **17**(2): 95-110.
Fromont-Racine, M., J. C. Rain, et al. (1997). "Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens." Nat Genet **16**(3): 277-82.
Garrels, J. I. (1996). "YPD-A database for the proteins of Saccharomyces cerevisiae." Nucleic Acids Res **24**(1): 46-9.
Gasch, A. P., M. Huang, et al. (2001). "Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p." Mol Biol Cell **12**(10): 2987-3003.
Gasch, A. P., P. T. Spellman, et al. (2000). "Genomic expression programs in the response of yeast cells to environmental changes." Mol Biol Cell **11**(12): 4241-57.
Gavin, A. C., M. Bosche, et al. (2002). "Functional organization of the yeast proteome by systematic analysis of protein complexes." Nature **415**(6868): 141-7.
Ge, H., Z. Liu, et al. (2001). "Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae." Nat Genet **29**(4): 482-6.
Ge, H., Z. Liu, et al. (2003). "Reply to "Does mapping reveal correlation between gene expression and protein–protein interaction?"" Nat Genet **33**(1): 16-17.
Ge, H., A. J. Walhout, et al. (2003). "Integrating 'omic' information: a bridge between genomics and systems biology." Trends Genet **19**(10): 551-60.
Gelbart, W. M., W. P. Rindone, et al. (1996). "FlyBase: the Drosophila database. The Flybase Consortium." Nucleic Acids Res **24**(1): 53-6.
Ghaemmaghami, S., W. K. Huh, et al. (2003). "Global analysis of protein expression in yeast." Nature **425**(6959): 737-41.
Giaever, G., A. M. Chu, et al. (2002). "Functional profiling of the Saccharomyces cerevisiae genome." Nature **418**(6896): 387-91.
Giot, L., J. S. Bader, et al. (2003). "A protein interaction map of Drosophila melanogaster." Science **302**(5651): 1727-36.
Goffeau, A., B. G. Barrell, et al. (1996). "Life with 6000 genes." Science **274**(5287): 546, 563-7.
Goldberg, D. S. and F. P. Roth (2003). "Assessing experimentally derived interactions in a small world." Proc Natl Acad Sci U S A **100**(8): 4372-6.

Grigoriev, A. (2001). "A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae." Nucleic Acids Res **29**(17): 3513-9.
Grigoriev, A. (2003). "On the number of protein-protein interactions in the yeast proteome." Nucleic Acids Res **31**(14): 4157-61.
Grunenfelder, B. and E. A. Winzeler (2002). "Treasures and traps in genome-wide data sets: case examples from yeast." Nat Rev Genet **3**(9): 653-61.
Han, J. D., N. Bertin, et al. (2004). "Evidence for dynamically organized modularity in the yeast protein-protein interaction network." Nature **430**(6995): 88-93.
Hazbun, T. R. and S. Fields (2001). "Networking proteins in yeast." Proc Natl Acad Sci U S A **98**(8): 4277-8.
Hermjakob, H., L. Montecchi-Palazzi, et al. (2004). "The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data." Nat Biotechnol **22**(2): 177-83.
Hermjakob, H., L. Montecchi-Palazzi, et al. (2004). "IntAct: an open source molecular interaction database." Nucleic Acids Res **32**(Database issue): D452-5.
Ho, Y., A. Gruhler, et al. (2002). "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry." Nature **415**(6868): 180-3.
Hubbard, T., D. Barker, et al. (2002). "The Ensembl genome database project." Nucleic Acids Res **30**(1): 38-41.
Hughes, T. R., M. J. Marton, et al. (2000). "Functional discovery via a compendium of expression profiles." Cell **102**(1): 109-26.
Huh, W. K., J. V. Falvo, et al. (2003). "Global analysis of protein localization in budding yeast." Nature **425**(6959): 686-91.
Ideker, T., V. Thorsson, et al. (2001). "Integrated genomic and proteomic analyses of a systematically perturbed metabolic network." Science **292**(5518): 929-34.
Ito, T., T. Chiba, et al. (2001). "A comprehensive two-hybrid analysis to explore the yeast protein interactome." Proc Natl Acad Sci U S A **98**(8): 4569-74.
Ito, T., K. Tashiro, et al. (2000). "Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins." Proc Natl Acad Sci U S A **97**(3): 1143-7.
Jacobson, I., G. Booch, et al. (1999). "The Unified Software Development Process."
Jansen, R., D. Greenbaum, et al. (2002). "Relating whole-genome expression data with protein-protein interactions." Genome Res **12**(1): 37-46.
Jansen, R., H. Yu, et al. (2003). "A Bayesian networks approach for predicting protein-protein interactions from genomic data." Science **302**(5644): 449-53.
Jelinsky, S. A. and L. D. Samson (1999). "Global response of Saccharomyces cerevisiae to an alkylating agent." Proc Natl Acad Sci U S A **96**(4): 1486-91.
Jeong, H., S. P. Mason, et al. (2001). "Lethality and centrality in protein networks." Nature **411**(6833): 41-2.
Kellis, M., B. W. Birren, et al. (2004). "Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae." Nature **428**(6983): 617-24.
Kemmeren, P., N. L. van Berkum, et al. (2002). "Protein interaction verification and functional annotation by integrated analysis of genome-scale data." Mol Cell **9**(5): 1133-43.

Khatri, P., S. Draghici, et al. (2002). "Profiling gene expression using onto-express." Genomics **79**(2): 266-70.
Kumar, A., S. Agarwal, et al. (2002). "Subcellular localization of the yeast proteome." Genes Dev **16**(6): 707-19.
Kumar, A. and M. Snyder (2002). "Protein complexes take the bait." Nature **415**(6868): 123-4.
Kyrpides, N. C. (1999). "Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide." Bioinformatics **15**(9): 773-4.
Larschan, E. and F. Winston (2001). "The S. cerevisiae SAGA complex functions in vivo as a coactivator for transcriptional activation by Gal4." Genes Dev **15**(15): 1946-56.
Lehner, B. and A. G. Fraser (2004). "A first-draft human protein-interaction map." Genome Biol **5**(9): R63.
Lercher, M. J., T. Blumenthal, et al. (2003). "Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes." Genome Res **13**(2): 238-43.
Li, S., C. M. Armstrong, et al. (2004). "A map of the interactome network of the metazoan C. elegans." Science **303**(5657): 540-3.
Lockhart, D. J., H. Dong, et al. (1996). "Expression monitoring by hybridization to high-density oligonucleotide arrays." Nat Biotechnol **14**(13): 1675-80.
Lord, P. W., R. D. Stevens, et al. (2003). "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation." Bioinformatics **19**(10): 1275-83.
Lu, L., A. K. Arakaki, et al. (2003). "Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the Saccharomyces cerevisiae proteome." Genome Res **13**(6A): 1146-54.
Magasanik, B. and C. A. Kaiser (2002). "Nitrogen regulation in Saccharomyces cerevisiae." Gene **290**(1-2): 1-18.
Marc, P., F. Devaux, et al. (2001). "yMGV: a database for visualization and data mining of published genome-wide yeast expression data." Nucleic Acids Res **29**(13): E63-3.
Marcotte, E. M., M. Pellegrini, et al. (1999). "A combined algorithm for genome-wide prediction of protein function." Nature **402**(6757): 83-6.
Mathe, C., M. F. Sagot, et al. (2002). "Current methods of gene prediction, their strengths and weaknesses." Nucleic Acids Res **30**(19): 4103-17.
Matthews, L. R., P. Vaglio, et al. (2001). "Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs"." Genome Res **11**(12): 2120-6.
Mewes, H. W., K. Albermann, et al. (1997). "Overview of the yeast genome." Nature **387**(6632 Suppl): 7-65.
Mewes, H. W., J. Hani, et al. (1998). "MIPS: a database for protein sequences and complete genomes." Nucleic Acids Res **26**(1): 33-7.
Mortimer, R. K., C. R. Contopoulou, et al. (1992). "Genetic and physical maps of Saccharomyces cerevisiae, Edition 11." Yeast **8**(10): 817-902.
Mortimer, R. K., D. Schild, et al. (1989). "Genetic map of Saccharomyces cerevisiae, edition 10." Yeast **5**(5): 321-403.

Mrowka, R. (2001). "A Java applet for visualizing protein-protein interaction." Bioinformatics **17**(7): 669-71.
Mrowka, R., W. Liebermeister, et al. (2003). "Does mapping reveal correlation between gene expression and protein-protein interaction?" Nat Genet **33**(1): 15-6; author reply 16-7.
Mrowka, R., A. Patzak, et al. (2001). "Is there a bias in proteome research?" Genome Res **11**(12): 1971-3.
Ng, S. K., Z. Zhang, et al. (2003). "Integrative approach for computationally inferring protein domain interactions." Bioinformatics **19**(8): 923-9.
Ogata, H., W. Fujibuchi, et al. (2000). "A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters." Nucleic Acids Res **28**(20): 4021-8.
Oliver, S. (2000). "Guilt-by-association goes global." Nature **403**(6770): 601-3.
Oliver, S. G., Q. J. van der Aart, et al. (1992). "The complete DNA sequence of yeast chromosome III." Nature **357**(6373): 38-46.
Oltvai, Z. N. and A. L. Barabasi (2002). "Systems biology. Life's complexity pyramid." Science **298**(5594): 763-4.
Ooi, S. L., D. D. Shoemaker, et al. (2001). "A DNA microarray-based genetic screen for nonhomologous end-joining mutants in Saccharomyces cerevisiae." Science **294**(5551): 2552-6.
Orton, R. J., W. I. Sellers, et al. (2004). "YETI: Yeast Exploration Tool Integrator." Bioinformatics **20**(2): 284-5.
Padilla, P. A., E. K. Fuge, et al. (1998). "The highly conserved, coregulated SNO and SNZ gene families in Saccharomyces cerevisiae respond to nutrient limitation." J Bacteriol **180**(21): 5718-26.
Peterson, J. D., L. A. Umayam, et al. (2001). "The Comprehensive Microbial Resource." Nucleic Acids Res **29**(1): 123-5.
Planta, R. J., A. J. Brown, et al. (1999). "Transcript analysis of 250 novel yeast genes from chromosome XIV." Yeast **15**(4): 329-50.
Poyatos, J. F. and L. D. Hurst (2004). "How biologically relevant are interaction-based modules in protein networks?" Genome Biol **5**(11): R93.
Quackenbush, J. (2001). "Computational analysis of microarray data." Nat Rev Genet **2**(6): 418-27.
Rai, R., J. R. Daugherty, et al. (1999). "Overlapping positive and negative GATA factor binding sites mediate inducible DAL7 gene expression in Saccharomyces cerevisiae." J Biol Chem **274**(39): 28026-34.
Rain, J. C., L. Selig, et al. (2001). "The protein-protein interaction map of Helicobacter pylori." Nature **409**(6817): 211-5.
Raychaudhuri, S., J. M. Stuart, et al. (2000). "Principal components analysis to summarize microarray experiments: application to sporulation time series." Pac Symp Biocomput: 455-66.
Reiss, D. J. and B. Schwikowski (2004). "Predicting protein-peptide interactions via a network-based motif sampler." Bioinformatics **20 Suppl 1**: I274-I282.
Ren, B., F. Robert, et al. (2000). "Genome-wide location and function of DNA binding proteins." Science **290**(5500): 2306-9.
Ren, X. Y., M. W. Fiers, et al. (2005). "Local coexpression domains of two to four genes in the genome of Arabidopsis." Plant Physiol **138**(2): 923-34.

Roberts, C. J., B. Nelson, et al. (2000). "Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles." Science **287**(5454): 873-80.

Rodriguez-Navarro, S., B. Llorente, et al. (2002). "Functional analysis of yeast gene families involved in metabolism of vitamins B1 and B6." Yeast **19**(14): 1261-76.

Ronne, H. (1995). "Glucose repression in fungi." Trends Genet **11**(1): 12-7.

Said, M. R., T. J. Begley, et al. (2004). "Global network analysis of phenotypic effects: protein networks and toxicity modulation in Saccharomyces cerevisiae." Proc Natl Acad Sci U S A **101**(52): 18006-11.

Saito, R., H. Suzuki, et al. (2003). "Construction of reliable protein-protein interaction networks with a new interaction generality measure." Bioinformatics **19**(6): 756-63.

Schena, M., R. A. Heller, et al. (1998). "Microarrays: biotechnology's discovery platform for functional genomics." Trends Biotechnol **16**(7): 301-6.

Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-70.

Schuler, G. D., J. A. Epstein, et al. (1996). "Entrez: molecular biology database and retrieval system." Methods Enzymol **266**: 141-62.

Scott, S., R. Dorrington, et al. (2000). "Functional domain mapping and subcellular distribution of Dal82p in Saccharomyces cerevisiae." J Biol Chem **275**(10): 7198-204.

Shannon, P., A. Markiel, et al. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome Res **13**(11): 2498-504.

Sharp, P. M. and E. Cowe (1991). "Synonymous codon usage in Saccharomyces cerevisiae." Yeast **7**(7): 657-78.

Shen, X., R. Ranallo, et al. (2003). "Involvement of actin-related proteins in ATP-dependent chromatin remodeling." Mol Cell **12**(1): 147-55.

Sherlock, G., T. Hernandez-Boussard, et al. (2001). "The Stanford Microarray Database." Nucleic Acids Res **29**(1): 152-5.

Spellman, P. T. and G. M. Rubin (2002). "Evidence for large domains of similarly expressed genes in the Drosophila genome." J Biol **1**(1): 5.

Spellman, P. T., G. Sherlock, et al. (1998). "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization." Mol Biol Cell **9**(12): 3273-97.

Spirin, V. and L. A. Mirny (2003). "Protein complexes and functional modules in molecular networks." Proc Natl Acad Sci U S A **100**(21): 12123-8.

Sprinzak, E., S. Sattath, et al. (2003). "How reliable are experimental protein-protein interaction data?" J Mol Biol **327**(5): 919-23.

Stanyon, C. A., G. Liu, et al. (2004). "A Drosophila protein-interaction map centered on cell-cycle regulators." Genome Biol **5**(12): R96.

Sturn, A. (2001). "Cluster Analysis for Large Scale Gene Expression Studies (Masters Thesis)."

Sturn, A., J. Quackenbush, et al. (2002). "Genesis: cluster analysis of microarray data." Bioinformatics **18**(1): 207-8.

Tamayo, P., D. Slonim, et al. (1999). "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." Proc Natl Acad Sci U S A **96**(6): 2907-12.
Tavazoie, S., J. D. Hughes, et al. (1999). "Systematic determination of genetic network architecture." Nat Genet **22**(3): 281-5.
Tong, A. H., G. Lesage, et al. (2004). "Global mapping of the yeast genetic interaction network." Science **303**(5659): 808-13.
Tucker, C. L., J. F. Gera, et al. (2001). "Towards an understanding of complex protein networks." Trends Cell Biol **11**(3): 102-6.
Uetz, P. and R. L. Finley, Jr. (2005). "From protein networks to biological systems." FEBS Lett **579**(8): 1821-7.
Uetz, P., L. Giot, et al. (2000). "A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae." Nature **403**(6770): 623-7.
Uetz, P., T. Ideker, et al. (2002). "Protein-Protein Interactions - A Molecular Cloning Manual (Visualization and integration of protein-protein interactions)." 623-646.
Uetz, P., S. V. Rajagopala, et al. (2004). "From ORFeomes to protein interaction maps in viruses." Genome Res **14**(10B): 2029-33.
Vassarotti, A. and A. Goffeau (1992). "Sequencing the yeast genome: the European effort." Trends Biotechnol **10**(1-2): 15-8.
Velculescu, V. E., L. Zhang, et al. (1997). "Characterization of the yeast transcriptome." Cell **88**(2): 243-51.
Vidal, M. (2001). "A biological atlas of functional maps." Cell **104**(3): 333-9.
von Mering, C., M. Huynen, et al. (2003). "STRING: a database of predicted functional associations between proteins." Nucleic Acids Res **31**(1): 258-61.
von Mering, C., R. Krause, et al. (2002). "Comparative assessment of large-scale data sets of protein-protein interactions." Nature **417**(6887): 399-403.
von Mering, C., E. M. Zdobnov, et al. (2003). "Genome evolution reveals biochemical networks and functional modules." Proc Natl Acad Sci U S A **100**(26): 15428-33.
Walhout, A. J., S. J. Boulton, et al. (2000). "Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm." Yeast **17**(2): 88-94.
Werner-Washburne, M., B. Wylie, et al. (2002). "Comparative analysis of multiple genome-scale data sets." Genome Res **12**(10): 1564-73.
Wightman, R. and P. A. Meacock (2003). "The THI5 gene family of Saccharomyces cerevisiae: distribution of homologues among the hemiascomycetes and functional redundancy in the aerobic biosynthesis of thiamin from pyridoxine." Microbiology **149**(Pt 6): 1447-60.
Winzeler, E. A., D. D. Shoemaker, et al. (1999). "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis." Science **285**(5429): 901-6.
Wodicka, L., H. Dong, et al. (1997). "Genome-wide expression monitoring in Saccharomyces cerevisiae." Nat Biotechnol **15**(13): 1359-67.
Wolfe, K. H. and D. C. Shields (1997). "Molecular evidence for an ancient duplication of the entire yeast genome." Nature **387**(6634): 708-13.
Wong, S. and K. H. Wolfe (2005). "Birth of a metabolic gene cluster in yeast by adaptive gene relocation." Nat Genet **37**(7): 777-82.

Wong, S. L., L. V. Zhang, et al. (2004). "Combining biological networks to predict genetic interactions." Proc Natl Acad Sci U S A **101**(44): 15682-7.
Wood, V., K. M. Rutherford, et al. (2001). "A re-annotation of the Saccharomyces cerevisiae genome." Comparative and Functional Genomics **2**(3): 143-154.
Wuchty, S., Z. N. Oltvai, et al. (2003). "Evolutionary conservation of motif constituents in the yeast protein interaction network." Nat Genet **35**(2): 176-9.
Xenarios, I., D. W. Rice, et al. (2000). "DIP: the database of interacting proteins." Nucleic Acids Res **28**(1): 289-91.
Zanzoni, A., L. Montecchi-Palazzi, et al. (2002). "MINT: a Molecular INTeraction database." FEBS Lett **513**(1): 135-40.
Zeeberg, B. R., W. Feng, et al. (2003). "GoMiner: a resource for biological interpretation of genomic and proteomic data." Genome Biol **4**(4): R28.
Zhang, L. V., S. L. Wong, et al. (2004). "Predicting co-complexed protein pairs using genomic and proteomic data integration." BMC Bioinformatics **5**: 38.
Zhou, M. and Y. Cui (2004). "GeneInfoViz: constructing and visualizing gene relation networks." In Silico Biol **4**(3): 323-33.
Zweiger, G. (1999). "Knowledge discovery in gene-expression-microarray data: mining the information output of the genome." Trends Biotechnol **17**(11): 429-36.