COMPUTATIONAL COMPARATIVE GENOMICS:

GENES, REGULATION, EVOLUTION

by

Manolis (Kellis) Kamvysselis

B.S. Electrical Engineering and Computer Science;
M. Eng. Computer Science and Engineering
Massachusetts Institute of Technology, 1999

Submitted to the Department of Electrical Engineering and Computer Science
In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Computer Science

at the
Massachusetts Institute of Technology
June 2003

Signature of Author .....
Department of Electrical Engineering and Computer Science
May 23, 2003

Certified by .....
Eric S. Lander
Professor of Biology
Thesis Co-Supervisor

Certified by .....
Bonnie A. Berger
Professor of Applied Mathematics
Thesis Co-Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Committee on Graduate Students
Department of Electrical Engineering and Computer Science

Computational Comparative Genomics:

Genes, Regulation, Evolution

by

Manolis (Kellis) Kamvysselis

Submitted to the Department of Electrical Engineering and Computer Science
on May 23, 2003 in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

ABSTRACT

Understanding the biological signals encoded in a genome is a key challenge of computational biology. These signals are encoded in the four-nucleotide alphabet of DNA and are responsible for all molecular processes in the cell. In particular, the genome contains the blueprint of all protein-coding genes and the regulatory motifs used to coordinate the expression of these genes. Comparative genome analysis of related species provides a general approach for identifying these functional elements, by virtue of their stronger conservation across evolutionary time.

In this thesis we address key issues in the comparative analysis of multiple species. We present novel computational methods in four areas (1) the automatic comparative annotation of multiple species and the determination of orthologous genes and intergenic regions (2) the validation of computationally predicted protein-coding genes (3) the systematic de-novo identification of regulatory motifs (4) the determination of combinatorial interactions between regulatory motifs.

We applied these methods to the comparative analysis of four yeast genomes, including the best-studied eukaryote, *Saccharomyces cerevisiae* or baker's yeast. Our results show that nearly a tenth of currently annotated yeast genes are not real, and have refined the structure of hundreds of genes. Additionally, we have automatically discovered a dictionary of regulatory motifs without any previous biological knowledge. These include most previously known regulatory motifs, and a number of novel motifs. We have automatically assigned candidate functions to the majority of motifs discovered, and defined biologically meaningful combinatorial interactions between them. Finally, we defined the regions and mechanisms of rapid evolution, with important biological implications.

Our results demonstrate the central role of computational tools in modern biology. The analyses presented in this thesis have revealed biological findings that could not have been discovered by traditional genetic methods, regardless of the time or effort spent. The methods presented are general and may present a new paradigm for understanding the genome of any single species. They are currently being applied to a kingdom-wide exploration of fungal genomes, and the comparative analysis of the human genome with that of the mouse and other mammals.

Thesis Co-Supervisor: Eric Lander, professor of Biology

Thesis Co-Supervisor: Bonnie Berger, professor of Applied Mathematics

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I am indebted to Eric Lander, Bonnie Berger and Bruce Birren for their constant help, advice, support, and mentorship in all aspects of my thesis and graduate career. Many thanks to my colleague Nick Patterson whose help and advice contributed to chapters 3 and 4, to David Gifford and Gerry Sussman for their advice, and to my friends Serafim Batzoglou, Sarah Calvo, James Galagan, Julia Zeitlinger for invaluable advice and support.

I would like to acknowledge the contribution of Matt Endrizzi and the staff of the Whitehead/MIT Center for Genome Research Sequencing Center, who generated the shotgun sequence from the three yeast species; David Botstein, Michael Cherry, Kara Dolinski, Diana Fisk, Shuai Weng and other members of the Saccharomyces Genome Database staff for assistance and discussions, and for making the data available to the community through SGD; Ed Louis and Ian Roberts who provided the yeast strains; Tony Lee, Nicola Rinaldi, Rick Young and the Young Lab for sharing data about chromatin immunoprecipitation experiments and for discussions; Michael Eisen and Audrey Gasch for sharing information about gene expression clusters and for discussions.

Many thanks to Gerry Fink, Martin Kupiec, Sue Lindquist, Andrew Murray, Heather True-Krobb for discussions and understanding of yeast biology. Many thanks to Jon Butler, Gus Cervini, Ken Dewar, Leslie Gaffney, David Jaffe, Joseph Lehar, Li Jun Ma, Abigail Melia, Chad Nusbaum and members of the WICGR for help and discussions.

I owe my gratitude to my parents John and Anna Kamvysselis, to my siblings Peter and Maria, and to Alexandra Mazalek for their love and constant support.

OVERVIEW

Biological Signals

Understanding the biological signals encoded in a genome is a key challenge of modern biology. These signals are encoded in the four-nucleotide alphabet of DNA and are responsible for all molecular processes in the cell. In particular, the genome contains the blueprint of all protein-coding genes and the control signals used to coordinate the expression of these genes. The well-being of any cell relies on the successful recognition of these signals, and a large number of biological mechanisms have evolved towards this goal. Specific protein complexes are responsible for the copying of a gene segment from DNA to messenger RNA (transcription) and for its eventual translation into protein following the genetic code to assign an amino acid to every tri-nucleotide codon. A specific class of proteins called transcription factors help recruit the transcription machinery to a target gene by binding their specific DNA signals (regulatory motifs) in response to environmental conditions. An abundance of information within the cell guides these processes, involving protein-protein and protein-DNA interactions between a multitude of players, the state of DNA coiling, and other mechanisms that are still not well-understood.

The computational identification of genes however, can only rely on the primary DNA sequence of the organism. Current programs use properties about the protein-coding potential of DNA segments that are unseen by the transcription machinery. In particular, since genes always start with an ATG (start codon) and end in with TAG, TGA, or TAA (one of three stop codons), programs exist that specifically look for these stretches between a start and a stop codon called ORFs (Open Reading Frames). The basic approach is to identify ORFs that are too long to have likely occurred by chance. Since stop codons occur at a frequency of 3 in 64 in random sequence, ORFs of 60 or even 150 amino acids will occur frequently by chance, but longer ORFs of 300 or thousands of amino acids are virtually always the result of biological selective pressure. Hence, simple computational programs can easily recognize long genes, but many small genes will be indistinguishable from spurious ORFs arising by chance. This is evidenced by the considerable debate over the number of genes in yeast[1-5] with proposed counts ranging from 4800 to 6400 genes. The situation is worse for organisms with large,

complex genomes, such as mammals where estimated gene counts have ranged from 30 to 120 thousand genes.

The direct identification of the repertoire of regulatory motifs in a genome is even more challenging. Regulatory motifs are short (typically 6-8 nucleotides), and do not obey the simple rules of protein-coding genes. In any single locus, nothing distinguishes these signals from random nucleotides. Traditionally, their discovery relied on deletion studies of consecutive DNA segments until regulation was disrupted and the control region was identified[6]. With the sequence of multiple genes in the same pathway at hand, it became possible to search for the repetition of these signals in genes controlled by the same transcription factor. Computational methods have been developed to search for enriched sequence motifs in predefined sets of genes (for example, using expectation-maximization[7] or gibbs-sampling[8], reviewed in [9]). As microarray analysis provided genome-wide levels of gene expression under a various experimental conditions, computational methods of gene clustering have resulted in hundreds of such sets of genes. Various computational methods have been used to mine these sets for regulatory motifs, and dozens of candidate motifs have resulted from each search. The vast majority of these candidate motifs are due to noise however, and only a total of about 50 real motifs have currently been discovered.

The current methods of motif identification suffer from a number of limitations. (a) First and foremost is that the weak signal of small motifs is hidden in the noise of relatively large intergenic regions. This inherent signal to noise ratio limits even the best programs from recognizing true motifs in the input data. (b) Additionally, the sets of genes searched, and hence the motifs discovered, are limited by our current biological knowledge of co-regulated sets of genes. The current knowledge is based on the experimental conditions reproduced in the lab, which is likely to be a small fraction of the vast array of environmental responses yeast uses to survive in its natural habitat. (c) Finally, an emerging view of gene regulation has put in question the approaches that search for a single motif responsible for a pathway or environmental response. Pathways are not regulated as isolated components in the cell. Genes and transcription factors have multiple functions and are used in multiple pathways and environmental responses. More importantly, transcription factors do not act in isolation, and protein-protein interactions

between factors are as important as protein-DNA interactions between each individual factor and its target genes. Hence, individual gene sets will be enriched in multiple motifs, and individual motifs will be enriched in multiple gene sets. A comprehensive understanding of regulatory motifs requires a novel, more powerful approach.

Comparative genome analysis of related species should provide such a general approach for identifying functional elements without prior knowledge of function. Evolution relentlessly tinkers with genome sequence and tests the results by natural selection. Mutations in non-functional nucleotides are tolerated and accumulate over evolutionary time. However, mutations in functional nucleotides are deleterious to the organism that carries them, and become sparse or extinct. Hence, functional elements should stand out by virtue of having a greater degree of conservation across the genomes of related species. Recent studies have demonstrated the potential power of comparative genomic comparison. Cross-species conservation has previously been used to identify putative genes or regulatory elements in small genomic regions[10-13]. Light sampling of whole-genome sequence has been used as a way to improve genome annotation[4,14]. Complete bacterial genomes have been compared to identify pathogenic and other genes[15-18]. Genome-wide comparison has been used to estimate the proportion of the mammalian genome under selection[19].

Contributions of this thesis

The goal of this thesis is to develop computational comparative methods to understand genomes. We develop and apply general approaches for the systematic analysis of protein-coding and regulatory elements by means of whole-genome comparisons with multiple related species. We apply these methods to *Saccharomyces cerevisiae*, commonly known as baker's yeast. *S. cerevisiae* is a model organism for which many genetic tools and techniques have been developed, leading to a wealth of experimental information. This knowledge has allowed us to validate our biological predictions and assess the power of the methods developed. We generated high-quality draft genome sequences from three *Saccharomyces* species of yeast related to *S. cerevisiae*. These data provide us with invaluable comparative information currently unmatched by previous sequencing efforts. Starting with the raw nucleotide sequence assemblies of the three newly sequenced species and the current sequence and annotation

of *S. cerevisiae*, we set out to discover functional elements in the yeast genome based on the comparison of the four species.

We first present methods for the automatic comparative annotation of the four species and the determination of orthologous genes and intergenic regions (Chapter 1). The algorithms enabled the automatic identification of orthologs for more than 90% of genes despite the large number of duplicated genes in the yeast genome.

Given the gene correspondence, we construct multiple alignments and present comparative methods for gene identification (Chapter 2). These rely on the different patterns of nucleotide change observed in the alignments of protein coding regions as compared to non-coding regions, specifically the pressure to conserve the reading frame of proteins. The method has high specificity and sensitivity, and enabled us to revisit the current gene catalogue of *S.cerevisiae* with important biological implications.

We then turn to the identification of regulatory motifs (Chapter 3). We present statistical methods for their systematic de-novo identification without use of prior biological information. We automatically identified 72 genome-wide sequence elements, with strongly non-random conservation properties. To validate our findings, we compared the discovered motifs against a list of known motifs, and found that we discovered virtually all previously known regulatory motifs, and an additional 41 motifs. We assign function to these motifs using sets of functionally related genes (Chapter 4), and we discover additional motifs enriched in these sets.

We further present methods for revealing the combinatorial control of gene expression (Chapter 5). We study the genome-wide co-occurrence of regulatory motifs, and discover significant correlations between pairs of motifs that were not apparent in a single genome. We show that these correspond to biologically meaningful relationships between the corresponding factors and that motif combinations can change the specific functional enrichment of target genes, thus increasing the versatility of gene regulation using only a limited number of regulatory motifs.

We finally focus on the differences between the species compared and discover the regions and mechanisms of evolutionary change (Chapter 6). We study rapid gene family expansions and discover that they localize in the telomeres. We show that chromosomal rearrangements and inversions are mediated by specific sequence elements.

We find specific mechanisms of rapid protein change in environment adaptation genes, as well as stretches of unchanged nucleotides suggesting novel functions for uncharacterized genes.

Our results demonstrate the central role of computational tools in modern biology. Our methods are general and applicable to the study of any organism. They are currently being applied to a kingdom-wide exploration of fungal genomes and the comparative analysis of the human genome with that of the mouse and other mammals. Comparison of multiple related species may present a new paradigm for understanding the genome of any single species.

BACKGROUND

0.1. Molecular biology and the study of life.

It is both humbling and bewildering that what separates humans from bacteria is merely the organization and assembly of the same basic bio-molecules. It is the study of these shared foundations of life that gave rise to the discipline of *molecular biology*. In the microscopic level, complex and simple organisms alike are made up of the same unit of life, the *cell*. A cell contains all the information and machinery necessary for its growth, maintenance and replication. It is delimited from its surrounding by a water-impermeable membrane and all communication and transport across the membrane is tightly controlled. Two major types of cells exist, *prokaryotic* cells with simple internal organization, and *eukaryotic* cells, with extensive compartmentalization of functions such as information storage in the nucleus, energy production in mitochondria, metabolism in the cytoplasm, etc. In unicellular organisms, the cell constitutes the complete organism, whereas multi-cellular organisms (typically eukaryotes) can contain up to trillions of cells, and hundreds of specialized cell types. In either case though, a cell can rarely be thought of in isolation, but is constantly interacting with its surrounding, sensing the presence of environmental changes, and exchanging stimuli with other cells that may be part of the same colony or organism.

Within a cell, virtually all functional roles are fulfilled by *proteins*, the most versatile type of macromolecule. Various types of proteins fulfill an immense array of tasks. For example, enzymes catalyze countless chemical reactions; transcription factors control the timing of gene usage; transporters carry molecules inside or outside the cell; trans-membrane channels regulate the concentrations of molecules in the cell; structural proteins provide support and shape to the cell; actins can cause motion; receptors recognize intra- or extra-cellular signals. This incredible versatility of proteins comes from the innumerable combinations of an alphabet of only 20 *amino acid* building blocks, juxtaposed in a single unbranched chain of hundreds or thousands of such amino acids. All amino acids share an identical portion of their structure that forms the protein backbone, to which is attached one of 20 possible side chains of variable size, shape, charge, polarity, hydrophobicity. The precise sequence of amino acids dictates a unique

three-dimensional fold that optimizes electrostatic and other interactions between the side-chains and with the solvent.

DNA in turn carries the genetic information that encodes the precise sequence of all proteins, the signals that control their production, and all other inheritable traits. DNA is also a macromolecule, consisting of the linear juxtaposition of millions of *nucleotides*. It encodes the genetic information digitally, like the bits of a digital computer, in the precise ordering of four types of nucleotides. Like amino-acids, these nucleotides share a fixed portion that forms a (phosphate) backbone to which is connected (via a deoxyribose sugar) a variable portion that is one of four *bases*, abbreviated A, C, G, T. Unlike proteins however, the structure of DNA is fixed. It consists of two strands, like the sidepieces of a ladder, connected by pairs of bases, like the steps of ladder. The two strands are wrapped around each other and form a double-helix. The two phosphate backbones form the outside of the helix, and the base pairs, connected by weak hydrogen bonds, form the interior of the helix. Only two pairings of bases are possible, based on shape and charge complementarity: A always pairs with T and C always pairs with G. This self-complementarity of the DNA structure forms the very basis of heredity: during DNA *replication*, the two strands open locally, and each strand becomes the template for synthesizing the opposite strand, its sequence dictated by base complementarity. The DNA double helix is rarely exposed. It is typically wrapped around histone proteins and packaged in a coiled structure referred to as *chromatin*.

The complete DNA content of an organism is referred to as its *genome*, and is contained in one or more large uninterrupted pieces called *chromosomes*. Prokaryotic cells contain one circular chromosome, and eukaryotic cells contain varying numbers of linear chromosomes (16 in yeast, 23 pairs in human) that are compartmentalized within the cell *nucleus*. Each linear chromosome is marked by a well-defined central region, the *centromere* and the chromosomal endpoints called *telomeres*. In a multi-cellular organism, every cell contains an identical copy of the genome (with extremely few exceptions such as red blood cells that do not have a nucleus). In addition to the chromosomal DNA, cells typically contain additional small pieces of DNA in plasmids (small circular pieces found in bacteria and typically containing antibiotic resistance genes), or mitochondria and chloroplasts (energy production organelles found in

eukaryotes). Genome size varies widely across species, typically 5kb-200kb (kilo-bases) for viruses[20-22], 500kb to 5Mb for bacteria[15], 10-30Mb for unicellular fungi[23,24], 97Mb for the worm[25], 165Mb for the fly[26], 2-3Gb for mammals[19,27], and 100Mb-100Gb for plants[28].

The amino-acid sequence of every protein is encoded within a single continuous stretch of DNA called a *gene*. The transfer of information from the four-letter nucleotide alphabet of DNA to the 20 amino-acid alphabet of proteins is ensured by a process called *translation*. Consecutive nucleotide triplets (*codons*) are translated into consecutive amino-acid residues, according to a precise translation table, referred to as the *genetic code*. There are 64 possible codons and only 20 amino acids, hence the genetic code contains degeneracies, and the same amino acid can be encoded by multiple codons. Additionally, the codon ATG (that codes for Methionine) also serves as a special translation initiation signal, and three codons (TGA, TAG, TAA) are dedicated translation termination signals. These are typically called *start* and *stop* codons. DNA is a *directional molecule*, and so are proteins. DNA is always read and synthesized in the *5' to 3'* direction (named after the 5' and 3' carbons in the carbon-ring of the sugar). Given this directionality of either strand, we can refer to sequences *upstream* (5') or *downstream* (3') of a particular nucleotide on the same strand. The two complementary strands run in opposite direction and are called anti-parallel, hence upstream in one strand is complementary to downstream on the opposite strand. Upstream and downstream are typically used in relation to the coding strand of a gene (containing the sequence ATG). Proteins are synthesized from the N terminus (encoded by the 5' part of the gene) to the C terminus (encoded by the 3' part of the gene).

0.2. Gene regulation and the dynamic cell

DNA is not directly translated into protein, but it is first transferred by complementarity into an intermediary single-stranded information carrier called messenger RNA or *mRNA* in a process called *transcription*. The *Central Dogma* of biology refers to this transfer of the genetic information from DNA to RNA to protein. RNA is similar to DNA, but is single-stranded and contains a different type of sugar connector between the phosphate backbone and the variable base (also the four bases are A,C,G,U instead of A,C,G,T). This difference in structure enables RNA to assume complex three-dimensional folds and perform a variety of cellular functions, only one of

which is information transfer between DNA and protein. In eukaryotic cells, transcription occurs in the nucleus where the DNA resides, and the resulting mRNA molecule is then transferred outside the nucleus where the translation machinery resides. During this transfer, the *transcript* undergoes a maturation step, including the excision (called *splicing*) of untranslated gene portions (called *introns*), and the joining of the remaining portions of the transcribed gene that are typically translated (called *exons*). The splicing of introns is dictated by subtle signals between 6 and 8 bp (base pairs) long that are found mainly at the junctions between exons and introns and within each intron. In prokaryotic cells, transcripts do not undergo splicing and sometimes contain multiple consecutively translated genes of related function.

The process of protein and RNA production, also called *gene expression*, is tightly controlled at multiple stages, but mainly at the stage of *transcription initiation*. This involves the uncoiling of chromatin structure around the gene to be expressed and the recruitment of a number of protein players that include the transcription machinery. These processes are regulated by a specific class of DNA-binding proteins called *transcription factors*. These bind the double-stranded DNA helix in sequence-specific *binding sites*, recognizing electrostatic properties of the nucleotides at each contact point. A *regulatory motif* describes the sequence specificity of a transcription factor, namely, the nucleotide patterns that are in common to the sites bound. Transcription factors are classified according to their effect on the expression of their target genes: an *activator* increases the level of gene expression when bound, and a *repressor* decreases that level. Transcription factor binding is modulated by the protein concentration and localization of the transcription factor, the three-dimensional conformation of the transcription factor that may depend on chemical modifications, protein-protein interactions with other factors that may bind cooperatively or competitively, and chromatin accessibility

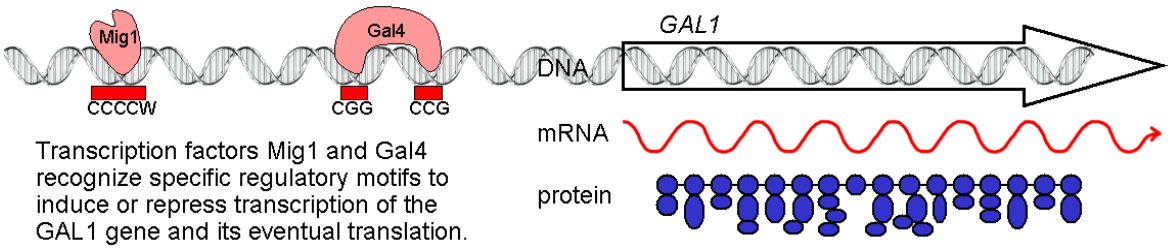


Figure 0.1. The Central Dogma of Biology. DNA makes RNA makes protein

surrounding the binding site. Finally, in addition to transcription initiation, gene expression is regulated at many stages, including mRNA transport and splicing, translation initiation and efficiency, mRNA stability and degradation, post-translational modifications of a protein, and protein stability.

These processes together modulate gene expression in response to environmental changes, and are interlinked in complex *regulatory networks*, responsible for the dynamic nature of the cell. These dynamics create the multitude of specific cell responses to varying environmental stimuli. Gene regulation also creates the incredible variety of cell types found within the same organism. For example heart, liver, lung, nail, skin, eye, neurons, hair, or bone all have the exact same DNA content, but express a different set of genes. Changes in gene expression however, can also be responsible for a number of complex diseases. Understanding the dynamic cell is a major challenge for molecular biology and modern medicine.

0.3. Evolutionary change and comparative genomics

The *evolution* of these complex mechanisms was shaped by the forces of random change and natural selection. Random genomic change can generate new functions or disrupt existing ones, and natural selection favors and keeps the fittest combinations. The *genotypic* differences accumulated at the DNA level lead to observed *phenotypic* differences between individuals of a population. Genomic changes can be as subtle as the mutation, insertion or deletion of individual nucleotides, and as drastic as the duplication or loss of chromosomal segments, entire chromosomes, or complete genomes. Changes in a protein-coding gene can lead to multiple co-existing variants, or *alleles*, of that gene within a population, that differ in specific residues and perform the same function with slight differences. As the result of mating, the progeny will inherit a combination of paternal and maternal alleles for different genes. The random mating of individuals within a populations and the random segregation of chromosomal segments in gamete formation creates new allelic combinations at each generation. The frequency of these allelic combinations will vary through evolutionary time, either by selection for their evolutionary fitness or by random genetic drift. As populations segregate and adapt to their environment, different combinations of alleles dominate in each population. The resulting differences in behavior or chromosomal organization can lead to loss of

reproductive ability across sub-populations and the emergence of new *species*. The emergence of new functions in these changing species allowed adaptation to all niches on land, in the air, underground, or in the deepest oceans, in species as diverse as dinosaurs and amoebae. It is thought that all life in the planet descends from a single ancestral cell that lived around 3.5 billion years ago, and the incredible biodiversity observed today resulted from incremental changes of existing life forms.

The genomes of related species exhibit similarities in functional elements that have undergone little change since the species' common ancestor. Deleterious mutations in these functional regions have certainly occurred, but the individuals carrying them have been at a disadvantage and eventually eliminated by natural selection. Mutations in non-functional regions have no effect to an organism's reproductive fitness, and will accumulate over evolutionary time. Hence, the combined effects of random mutation and natural selection allow comparative approaches to separate conserved functional regions from diverged non-functional regions. Comparative genome analysis of related species should provide a general approach for identifying functional elements without prior knowledge of function, by virtue of having a greater degree of conservation across the genomes of related species. When selecting species for a pairwise comparative analysis, we face a tradeoff between closely related species (with many common functional elements but additional spuriously conserved non-functional regions), and distantly related species (with mostly diverged non-functional regions but fewer common functional elements). The use of multiple closely-related species may present an attractive alternative, exhibiting an accumulation of independent mutations in non-functional regions, while having most biological functions in common.

Recent studies have demonstrated the potential power of comparative genomic comparison. Cross-species conservation has previously been used to identify putative genes or regulatory elements in small genomic regions[10-13]. Light sampling of whole-genome sequence has been studied as a way to improve genome annotation[4,14]. Complete bacterial genomes have been compared to identify pathogenic and other genes[15-18]. Genome-wide comparison has been used to estimate the proportion of the mammalian genome under selection[19].

0.4. Sequence alignment and phylogenetic trees

The comparison of related sequences is typically represented as *sequence alignment* (for an example see figure 3.2). The correspondence of nucleotides across the sequences compared is given by offsetting the nucleotides of each sequence such that matching nucleotides are stacked at the same index across all sequences. To represent insertions or deletions (*indels*), gaps are typically inserted as dashes in the shorter sequence; these could represent a deletion in the sequence containing the gap, or an insertion in the other sequences. Typically, no reordering or repetition of nucleotides is allowed within a sequence, and hence no inversions, duplications, or translocations are represented in a sequence alignment. To construct an alignment of two sequences is equivalent to finding the optimal path in a two-dimensional grid of cells, and dynamic programming algorithms have been developed to align two sequences in time proportional to the product of their lengths, and space proportional to sum of their lengths. The optimal alignment of two sequences minimizes the total cost of insertions, deletions, and nucleotide substitutions (gaps and mismatches), each penalized according to input parameters. These parameters are set to match estimated rates of insertions, deletions and nucleotide substitutions in well-conserved portions of carefully-constructed alignments. For example, substitutions between nucleotides of similar structure are more frequent and hence *transitions* between *purines* (A and G) or between *pyrimidines* (C and T) are penalized less than *transversions* from a purine to a pyrimidine and vice versa. Also, it is typical to penalize gaps using affine functions, namely adding a cost proportional to the size of the gap to a fixed cost for starting a gap. *Global alignments* compare the entire length of the sequences compared, and *local alignments* only align sub-portions of the sequences.

The best *match* of a *query* sequence can be found in a *database* of sequences by scoring the local alignments between the query and each sequence in the database. Constructing the full dynamic programming matrix for each of the sequences in a large database can be costly, and efficient algorithms have been developed to only align a small subset of the database sequences. These algorithms take advantage of the fact that strong matches of a query sequence will typically contain stretches of perfectly conserved residues, and first select all database sequences that contain such stretches. To do so, a

hash table is first constructed for the database, listing all sequences and positions that contain a particular k-mer. After this slow step that need only be performed once, the lookup of all k-mers in a query sequence can be performed rapidly against a large database, constructing a list of *hits*. Local alignments are then constructed around each hit, extending the k-mer matches to longer high-scoring local alignments. These ideas are implemented in the popular program BLAST, and used thousands of times daily to query the genomes of dozens of sequenced species and millions of sequences. One modification of the BLAST algorithm called two-hit Blast only constructs a local alignment when at least two nearby hits are found. This allows the retrieval of more distantly related sequences by searching for shorter k-mers, while still maintaining high specificity by requiring multiple k-mer hits in common.

Multiple sequence alignments can also be constructed for more than two sequences. Constructing the full dynamic programming matrix is exponential in the number of sequences compared and typically impractical for long sequences. Therefore, current algorithms work by extending multiple pairwise alignments between the sequences compared. The similarities between all pairs of sequences can be used to construct a *phylogenetic tree*, summarizing the most likely ancestry of the sequences, linking them hierarchically from the most closely related pair to the most distantly related outgroup. Multiple sequence alignment algorithms typically start by aligning the most closely related sequences, and progressively merge alignments moving up the phylogenetic tree from the leaves to the root. Algorithms to merge two alignments typically use once-a-gap-always-a-gap methods, but more recent algorithms have been developed to locally re-optimize multiple alignment portions by revisiting previously added gaps and improving the overall alignment score.

0.5. Model organisms and yeast genetics.

The shared biology of related species allows one to study a biological process in one organism and apply the knowledge to another organism. Simpler organisms provide excellent models for developing and testing the procedures needed for studying the much more complex human genome. Such *model organisms* include bacteria, yeast, fungi, worms, flies and mice, each teaching us different aspects of human biology. For example, the study of cancer development has flourished by studying mouse models, and

has lead to medical application in humans. *Mutant* strains can be isolated containing specific defects in genes that lead to disease phenotypes. Controlled *crosses* can be used to restore lost functions or inhibit genes at particular stages of development and study their effects on the organism. The shorter the generation time of a model organism, the easier it is to perform multiple crosses.

The yeast *Saccharomyces cerevisiae* in particular provides a powerful genetic system with the availability of a wide array of tools such as gene replacement, plasmids, deletion strains, two-hybrid systems. Yeast is also amenable to biochemical methods, such as the purification and characterization of protein complexes. Because of these experimental advantages, yeast has been the system of choice to study the most basic cellular functions common to eukaryotes such as cell division, cell structure, energy production, cell growth, cell death, cell cycle, gene regulation, transcription initiation, cell signaling, and other basic cell processes. More recently, yeast has become the organism of choice for the development and testing of modern technologies for *genome-wide* experimental studies. The complete parts-list of all genes has radically changed the face of biological research. If a particular phenotype is due to the function of a single protein, it is necessarily encoded by one of these few thousand genes. Additionally, the relatively small number of genes (~6000) allows the simultaneous observation of the complete genome for mRNA expression, transcription factor binding, or protein-protein interactions. The public sharing of yeast strains, materials, and genome-wide experimental data has provided a global view of the dynamic yeast genome unmatched in any other organism.

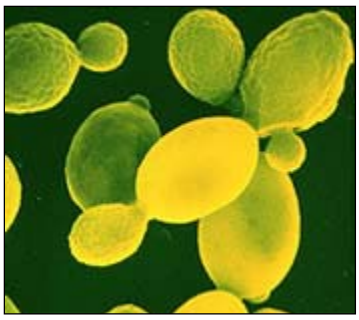


Figure 0.2. The yeast Saccharomyces cerevisiae undergoing cell division.

Yeast also presents an ideal organism for developing computational methods for genome-wide *comparative* analysis. It is the most well-studied eukaryote, and the vast functional knowledge allows the immediate validation of our findings against previous work. Additionally, the strong experimental system allows the experimental follow-up of biological hypotheses raised in the comparative work. The small genome size (250 times

smaller than human) allows the sequencing of multiple yeast species at an affordable cost. Additionally, the small number of repetitive elements allows for easy whole-genome-shotgun assembly (see next section). For all these considerations, we decided to work on yeast.

0.6. Genome sequencing and assembly

We sequenced and assembled the complete genomes of *S. paradoxus*, *S. mikatae* and *S. bayanus*, three yeast species that are close relatives of *S. cerevisiae*, within the *Saccharomyces sensu stricto* group[29]. Their divergence times from the *S. cerevisiae* lineage are approximately 5, 10 and 20 million years (based on sequence divergence of ribosomal DNA sequence). Like *S. cerevisiae*, they all have 16 chromosomes and their genomes contain about 12 million bases. These species were chosen based on their evolutionary relationships (closely enough related that functional elements be conserved, and distant enough that non-functional bases have had enough evolutionary time to diverge).

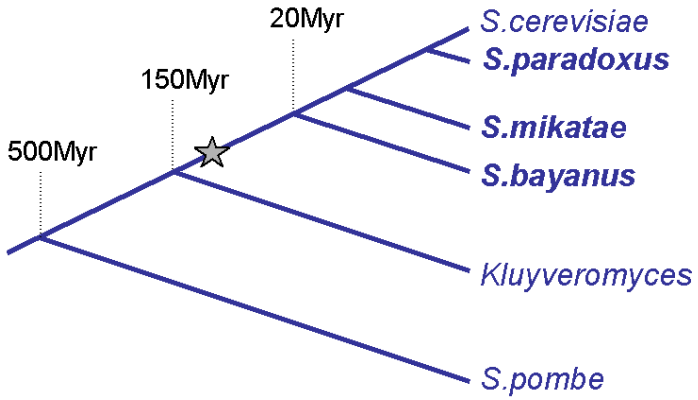


Figure 0.3: Phylogenetic tree of analyzed species. The newly sequenced species are shown in bold. Star denotes inferred genome-wide duplication of the yeast genome. Divergence times are approximate and based on ribosomal DNA sequence divergence

Reading the order of the nucleotides in any one segment of DNA relies on a technology developed by Sanger in 1977 that uses the central agent of DNA replication, *DNA polymerase*. This protein complex recognizes the transition from double-stranded DNA to single-stranded DNA in an incomplete helix, and extends the shorter strand in the 5' to 3' direction. By introducing a small fraction of faulty nucleotides that cause an early termination of the extension reaction, and subsequently comparing the lengths of resulting fragments in each of four reactions, this method infers the sequence of a DNA fragment. The extension reaction can be initiated at any unique segment of DNA by

introducing a complementary segment called a *primer*. This primer binds single-stranded DNA by complementarity, creating the double-strand to single-strand transition recognized by DNA polymerase. Unfortunately, since the Sanger method works by weight separation between fragments of different lengths, it can only determine the sequence of small fragments (currently around 800 nucleotides). The weight difference between fragments of 800 nucleotides and fragments of 801 nucleotides is too small to be detected reliably.

To obtain the sequence of longer stretches of DNA, two methods are possible. One is to synthesize a new primer at the end of 800 nucleotides and use it to sequence the subsequent 800 nucleotides (and so on). Unfortunately, synthesizing new primers is expensive and time-consuming since the primer to be used is not known until the sequence is obtained, and this method is rarely used. An alternative method is to first make many copies of the longer stretch of DNA and randomly break them into small fragments, and then sequence 800 nucleotide *reads* from each of these fragments and re-piece them together computationally (each of the fragments is inserted to a common *vector* whose sequence is known, hence the same primer can be used to sequence the end of each of these fragments). This alternative method is called *shotgun sequencing*, in reference to the random breaking of the longer fragment as if struck by a shotgun. Sequence reads can also be obtained from both ends of a fragment, providing *linking* information between *paired reads*. This method is called *paired-end shotgun sequencing*. The shotgun fragments are typically selected to be of a particular size, providing additional information about the genomic distance between paired sequence reads.

Shotgun sequencing depends heavily on the computational ability to correctly *assemble* the resulting fragments of sequence. *Fragment assembly* searches for sequences common between two sequence fragments (also called *reads*) and unique otherwise, in order to join them into a longer sequence. This is made harder due to sequencing errors that lead to sequence differences between reads that really come from the same part of the genome, as well as repetitive sequences within genomes that lead to identical sequences between reads that come from different parts of the genome. Modern assembly programs produce stretches of continuous sequence called *contigs*, which are

linked into *supercontigs* or *scaffolds*, when their relative order, orientation, and estimated spacing is given by the pairing of reads (Figure 0.4). To assemble complete genomes, two methods are currently in use. *Whole-genome shotgun* (WGS) randomly breaks the complete genome and assembles all fragments computationally. *Clone-based* methods first partition the genome into large fragments (clones) and then use shotgun sequencing for each of the fragments. Clone-based methods are more expensive but more reliable. WGS methods are cheaper but rely more heavily on the ability of subsequent computational assembly programs. Hybrids between WGS and clone-based methods are used nowadays in major sequencing projects. It is also common to use WGS with links of multiple sizes to provide both short-range and long-range connectivity information.

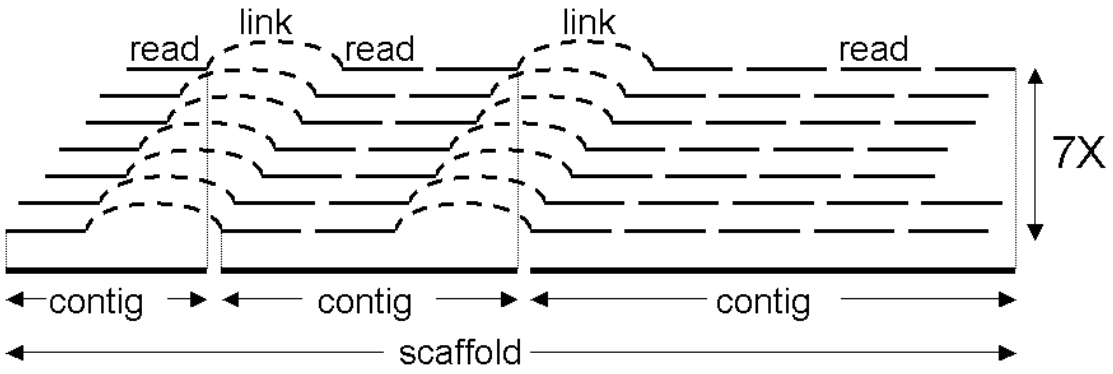


Figure 0.4 Genome Assembly. Overlapping sequence reads are grouped into blocks of continuous sequence (contigs). The pairing of forward and reverse reads provides links across neighboring contigs, grouping them in supercontigs or scaffolds. Each base in the genome is observed on average in 7 overlapping reads.

CHAPTER 1: GENOME CORRESPONDENCE

1.1. Introduction

The first issue in comparative genomics is determining the correct correspondence of chromosomal segments and functional elements across the species compared. This involves the recognition of *orthologous* segments of DNA that descend from the same region in the common ancestor of the species compared. However, it is equally important to recognize which segments have undergone duplication events, and which segments were lost since the divergence of the species. By accounting for duplication and loss events, we ensure that we are comparing orthologous segments.

We decided to use genes as discrete genomic anchors in order to align and compare the species. We constructed a bipartite graph connecting annotated protein-coding genes in *S. cerevisiae* to predicted protein-coding genes in each of the other species based on sequence similarity at the amino-acid level. This bipartite graph should contain the orthologous matches but also contains spurious matches due to shared domains between proteins of similar functions, and gene duplication events that precede the divergence of the species. Determining which matches represent true orthologs and resolving the correspondence of genes across the four species will be the topic of this chapter.

We present an algorithm for comparative annotation that has a number of attractive features. It uses a simple and intuitive graph theoretic framework that makes it easy to incorporate additional heuristics or knowledge about the genes at hand. It represents matches between sets of genes instead of only one-to-one matches, thus dealing with duplication and loss events in a very straightforward way. It uses the chromosomal positions of the compared genes to detect stretches of conserved gene order and uses these to resolve additional orthologous matches. It accounts for all genes compared, resolving *unambiguous* matches instead of simply *best* matches, thus ensuring that all 1-to-1 genes are true orthologs. It works at a wide range of evolutionary distances, and can cope with unfinished genomes containing gaps even within genes.

1.2. Establishing gene correspondence

Previously described algorithms for comparing gene sets have been widely used for various purposes, but they are not applicable to the problem at hand.

Best Bidirectional Hits (BBH)[30,31] looks for gene pairs that are best matches of each other and marks them as orthologs. In the case of a recent gene duplication however, only one of the duplicated genes will be marked as the ortholog without signaling the presence of additional homologs. Thus, no guarantees are given that 1-to-1 matches will represent orthologous relations and incorrect matches may be established.

Clusters of Orthologous Genes (COG)[32,33] goes a step further and matches groups of genes to groups of genes. Unfortunately, the grouping is too coarse, and clusters of orthologous genes typically correspond to gene families that may have expanded before the divergence of the species compared. This inability to distinguish recent duplication events from more ancient duplication events makes it inapplicable in this case, since the genome of *S. cerevisiae* contains hundreds of gene pairs that were anciently duplicated before the divergence of the species at hand[34]. COGs would not distinguish between the two copies of anciently duplicated genes, and many orthologous matches would not be detected (Koonin, personal communication).

We introduce the concept of a Best Unambiguous Subset (BUS), namely a group of genes such that all best matches of any gene within the set are contained within the set, and no best match of a gene outside the set is contained within the set. A BUS builds on both BBHs and COGs to resolve the correspondence of genes across the species. The algorithm, at its core, represents the best match of every gene as a set of genes instead of a single best hit, which makes it more robust to slight differences in sequence similarity. A BUS can be isolated from the remainder of the bipartite gene correspondence graph while preserving all potentially orthologous matches. BUS also allows a recursive application grouping the genes into progressively smaller subsets and retaining ambiguities until later in the pipeline when more information becomes available. Such information includes the conserved gene order (synteny) between consecutive orthologous genes that allows the resolving of additional neighboring genes.

1.3. Overview of the algorithm

We formulated the problem of genome-wide gene correspondence in a graph-theoretic framework. We represented the similarities between the genes as a bipartite graph connecting genes between two species. We weighted every edge connecting two genes by the amino acid sequence similarity between the two genes, and the overall length of the match.

We separated this graph into progressively smaller subgraphs until the only remaining matches connected true orthologs (Figure 1.1). To achieve this separation, we eliminated edges that are sub-optimal in a series of steps. As a pre-processing step, we eliminated all edges that are less than 80% of the maximum-weight edge both in amino acid identity and in length. Based on the unambiguous matches that resulted from this step, we built blocks of conserved gene order (synteny) when neighboring genes in one species had one-to-one matches to neighboring genes in the other species; we used these blocks of conserved synteny to resolve additional ambiguities by preferentially keeping matches within synteny blocks. We finally searched for subsets of genes that are locally optimal, such that all best matches of genes within the group are contained within the group, and no genes outside the group have matches within the group. These best unambiguous subsets (BUS) ensure that the bipartite graph is maximally separable, while maintaining all possibly orthologous relationships.

When no further separation was possible, we returned the connected components of the final graph. These contain the one-to-one orthologous pairs resolved as well as sets of genes whose correspondence remained ambiguous in a small number of homology groups.

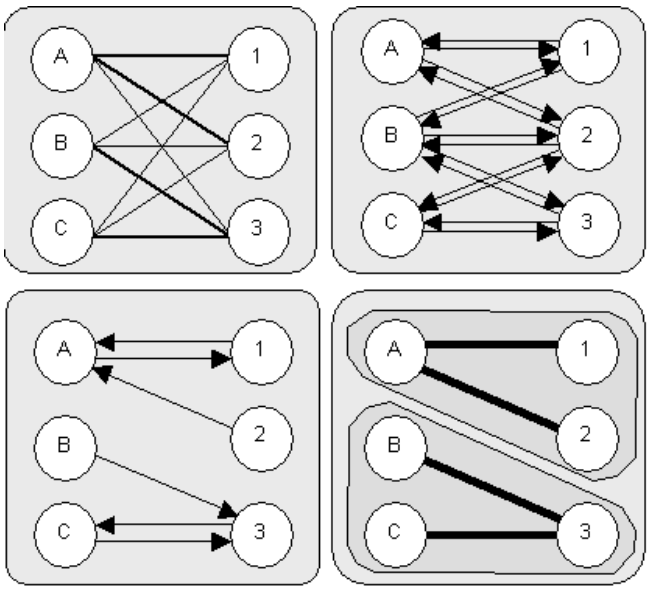


Figure 1.1. Overview of graph separation.
We construct a bipartite graph based on the blast hits. We consider both forward and reverse matches for near-optimality based on synteny and sequence similarity. Sub-optimal matches are progressively eliminated simplifying the graph. We return the connected components of the undirected simplified graph.

1.4. Automatic annotation and graph construction

In this section, we describe the construction of the weighted bipartite graph **G**, representing the gene correspondence across the species compared. We started with the genomic sequence of the species and the annotation of *S. cerevisiae*, namely the start and stop coordinates of genes. We then predicted protein-coding genes for each newly sequenced genome. Finally we connected across each pair of species the genes that shared amino-acid sequence similarity.

The input to the algorithm is based on the complete genome for each species compared. For *S. cerevisiae*, we used the public sequence available from the Saccharomyces Genome Database (SGD) at genome-www.stanford.edu/Saccharomyces. SGD posts sixteen uninterrupted sequences, one for each chromosome. The sequence was obtained by an international sequencing consortium and published in 1996. It was completed by a clone-based sequencing approach and directed sequence finishing to close all gaps. Subsequent to the publication, updates to the original sequence have been incorporated in SGD based on resequencing of regions studied in labs around the world.

The genome sequence of *S. paradoxus*, *S. mikatae* and *S. bayanus* was obtained at the MIT/Whitehead Institute Center for Genome Research. We used a whole-genome shotgun sequencing approach with paired-end sequence reads of 4kb plasmid clones, with lab protocols as described at www-genome.wi.mit.edu. We used ~7-fold

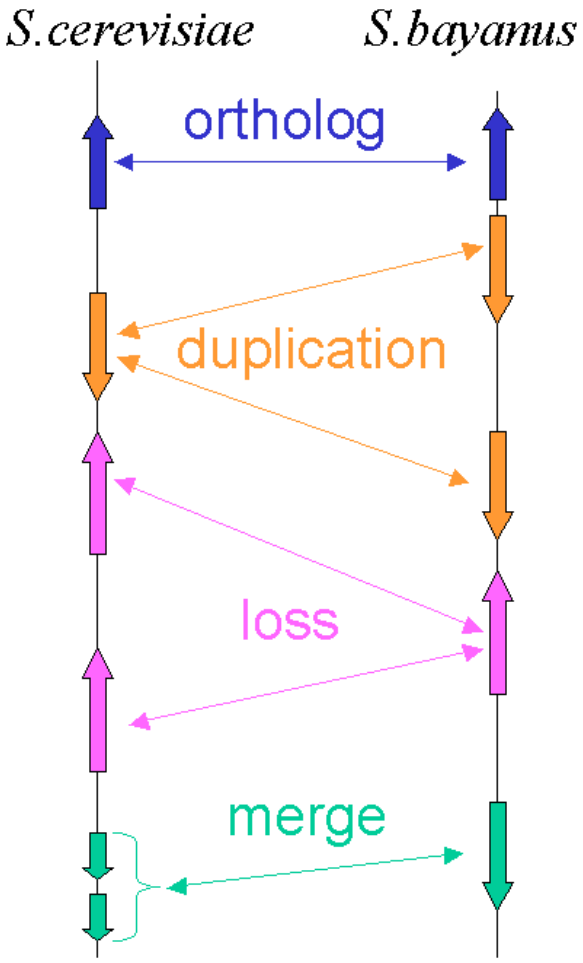


Figure 1.2. Bipartite Graph Construction. Annotated ORFs (vertical block arrows) are connected based on sequence similarity.

redundant coverage, namely every nucleotide in the genome was contained on average in at least 7 different reads. The information was then assembled with the Arachne computer program[35,36] into a draft sequence for each genome. The assembly contains *contigs*, namely continuous blocks of uninterrupted sequence, and *scaffolds* or *supercontigs*, namely uninterrupted blocks of linked contigs for which the relative order and orientation is known. This order and orientation is given by the pairing of reads that originated from the ends of the same 4kb clone. The draft genome sequence of each species has long-range continuity (more than half of the nucleotides are in scaffolds of length 230-500 kb, as compared to 942 kb for the finished sequence of *S. cerevisiae*), relatively short sequence gaps (0.6-0.8 kb, which is small compared to a typical gene), and contains the vast majority of the genome (~95%).

Once the genome sequences are available, we determine the set of protein-coding genes for each species. For *S. cerevisiae*, we used the public gene catalogue at SGD. It was constructed by including all predicted protein coding genes of at least 100 AA that do not overlap longer genes by more than 50% of their length. It was subsequently updated to include additional short genes supported by experimental evidence and to reflect changes in the underlying sequence when resequencing revealed errors. For the three newly sequenced species, we predicted all uninterrupted genes starting with a methionine (start codon ATG) and containing at least 50 amino acids.

We then constructed the bipartite graph connecting all predicted protein coding genes that share amino acid sequence similarities across any two species (Figure 1.2). For this purpose, we first used protein BLAST[37] to find all protein hits between the two protein sets (we used WU-BLAST BlastP with parameters W=4 for the hit size in amino acids, hitdist=60 for the distance between two hits and $E=10^{-9}$ for the significance of the matches reported). Since the similarity between query protein *x* in one genome and subject protein *y* in another genome is sometimes split in multiple blast hits, we grouped all blast hits between *x* and *y* into a single *match*, weighted by the average amino acid percent identity across all hits between *x* and *y* and by the total protein length aligned in blast hits. These matches form the edges of the bipartite graph **G**, described in the following section.

1.5. Initial pruning of sub-optimal matches

Let **G** be a weighted bipartite graph describing the similarities between two sets of genes **X** and **Y** in the two species compared (Figure 1.1, top left panel). Every edge e=(x,y) in E that connects nodes x ∈ X and y ∈ Y was weighted by the total number of amino acid similarities in BLAST hits between genes x and y. When multiple BLAST hits connected x to y, we summed the non-overlapping portions of these hits to obtain the total weight of the corresponding edge. We constructed graph **M** as the directed version of **G** by replacing every undirected edge **e**=(**x**,**y**) by two directed edges (**x**,**y**) and (**y**,**x**) with the same weight as **e** in the undirected graph (Figure 1.1, top right panel). This allowed us to rank edges incident from a node, and construct subsets of **M** that contain only the top matches out of every node.

This step drastically reduced the overall graph connectivity by simply eliminating all out-edges that are not near optimal for the node they are incident from. We defined $\mathbf{M_{80}}$ as the subset of **M** containing for every node only the outgoing edges that are at least 80% of the best outgoing edge (any not in the upper 20% of all scores). This was mainly a preprocessing step that eliminated matches that were clearly non-optimal. Virtually all matches eliminated at this stage were due to protein domain similarity between distantly related proteins of the same super-family or proteins of similar function but whose separation well-precedes the divergence of the species. Selecting a match threshold relative to the best edge ensured that the algorithm performs at a range of evolutionary distances. After each stage, we separated the resulting subgraph into connected components of the undirected graph (Figure 1.1, bottom right panel).

1.6. Blocks of conserved synteny

The initial pruning step created numerous two-cycle subgraphs (unambiguous one-to-one matches) between proteins that do not have closely related paralogs. We used these to construct blocks of conserved synteny based on the physical distance between consecutive matched genes, and preferentially kept edges that connect additional genes within the block of conserved gene order (Figure 1.3). Edges connecting these genes to genes outside the blocks were then ignored, as unlikely to represent orthologous relationships. Without imposing an ordering on the scaffolds or the chromosomes, we

associated every gene x with a fixed position (s, start) within the assembly, and every gene y with a fixed position (chromosome, start) within *S. cerevisiae*. If two one-to-one unambiguous matches (x1, y1) and (x2, y2) were such that x1 was physically near x2, and y1 was physically near y2, we constructed a synteny block B=({x1, x2},{y1,y2}). Thereafter, for a gene x3 that was proximal to {x1, x2}, if an outgoing edge (x3, y3) existed such that y3 was proximal to {y1,y2}, we ignored other outgoing edges (x3, y') if y' was not proximal to {y1,y2}.

Without this step, duplicated genes in the yeast species compared remained in two-by-two homology groups, especially for the large number of ribosomal genes that are nearly identical to one another. We found this step to play a greater role as evolutionary distances between the species compared became larger, and sequence similarity was no longer sufficient to resolve all the ambiguities. We only considered synteny blocks that had a minimum of three genes before using them for resolving ambiguities, to prevent being misled by rearrangements of isolated genes. We set the maximum distance ***d*** for considering two neighboring genes as proximal to 20kb, which corresponds to roughly 10 genes. This parameter should match the estimated density of syntenic anchors. If many genomic rearrangements have occurred since the separation of the species, or if the scaffolds of the assembly are short, the syntenic segments will be shorter and setting *d* to larger values might hurt the performance. On the other hand if the number of unambiguous genes is too small at the beginning of this step, the genes used as anchors will be sparse, and no synteny blocks will be possible for small values of ***d***.

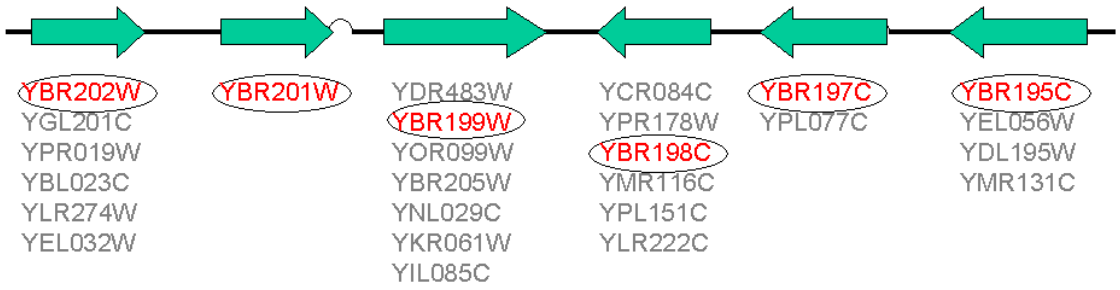


Figure 1.3. The use of synteny. In blocks of conserved gene order (synteny), we preferentially keep those matches that preserve the order of orthologous genes.

1.7. Best Unambiguous Subsets

We finally separated out subgraphs that were connected to the remaining edges in the graph by solely non-maximal edges. These subgraphs are such that the best match of any node within the subset is contained within the subset, and no node outside the subset has its best match within the subset. These two properties ensure that the subsets are both best and unambiguous.

We defined a Best Unambiguous Subset (BUS) of the nodes of $\mathbf{X} \cup \mathbf{S}$, to be a subset **S** of genes, such that $\forall x: x \in S \Leftrightarrow best(x) \subseteq S$, where best(x) are the nodes incident to the maximum weight edges from x. We then constructed $\mathbf{M_{100}}$, following the notation above, namely the subset of **M** that contains only best matches out of a node. Note that multiple best matches were possible based on our definition. To construct a BUS, we started with the subset of nodes in any cycle in M_{100}. We augmented the subset by following forward and reverse best edges, that is including additional nodes if their best match was within the subset, or if they were the best match of a node in the subset. This ensured that separating a subset did not leave any node orphan, and did not remove the strictly best match of any node. When no additional nodes needed to be added, the BUS condition was met.

Figure 1.4 shows a toy example of a similarity matrix. Genes A, B, and C in one genome are connected in a complete bipartite graph to genes 1, 2 and 3 in another genome (ignoring for now synteny information). The sequence similarity between each pair is given in the matrix, and corresponds to the edge weight connecting the two genes in the bipartite graph. The set (A,1,2) forms a BUS, since the best matches of A, 1, and 2 are all within the set,

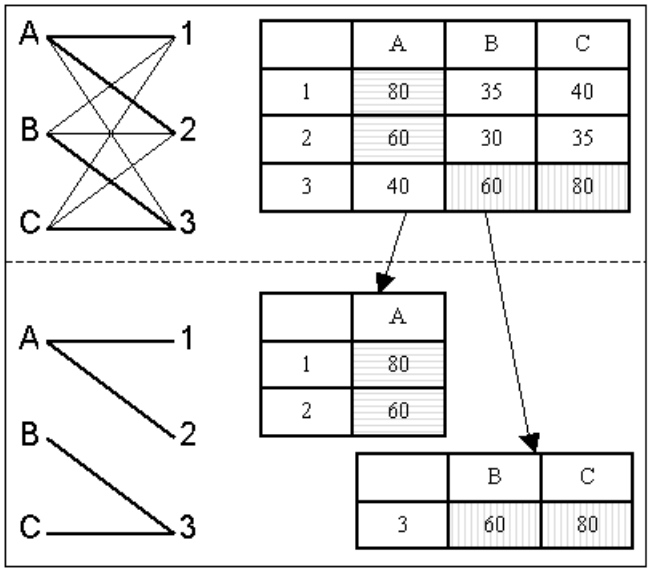


Figure 1.4. Best Unambiguous Subsets (BUS). A BUS is a set of genes that can be isolated from a homology group while preserving all potentially orthologous matches. Given the similarity matrix above and no synteny information, two such sets are (A,1,2) and (B,C,3).

and none of them represents the best match of a gene outside the set. Hence, the edges connecting (A,1,2) can be isolated as a subgraph without removing any orthologous relationships, and edges (B,1), (B,2), (C,1), (C,2), (A,3) can be ignored as non-orthologous. Similarly (B,C,3) forms a BUS. The resulting bipartite graph is shown. A BUS can be alternatively defined as a connected component of the undirected version of M_{100} (Figure 1.1, bottom panels).

This part of the algorithm allowed us to resolve the remaining orthologs, mostly due to subtelomeric gene family expansions, small duplications, and other genes that did not benefit from synteny information. In genomes with many rearrangements, or assemblies with low sequence coverage, which do not allow long-range synteny to be established, this part of the algorithm will play a crucial role.

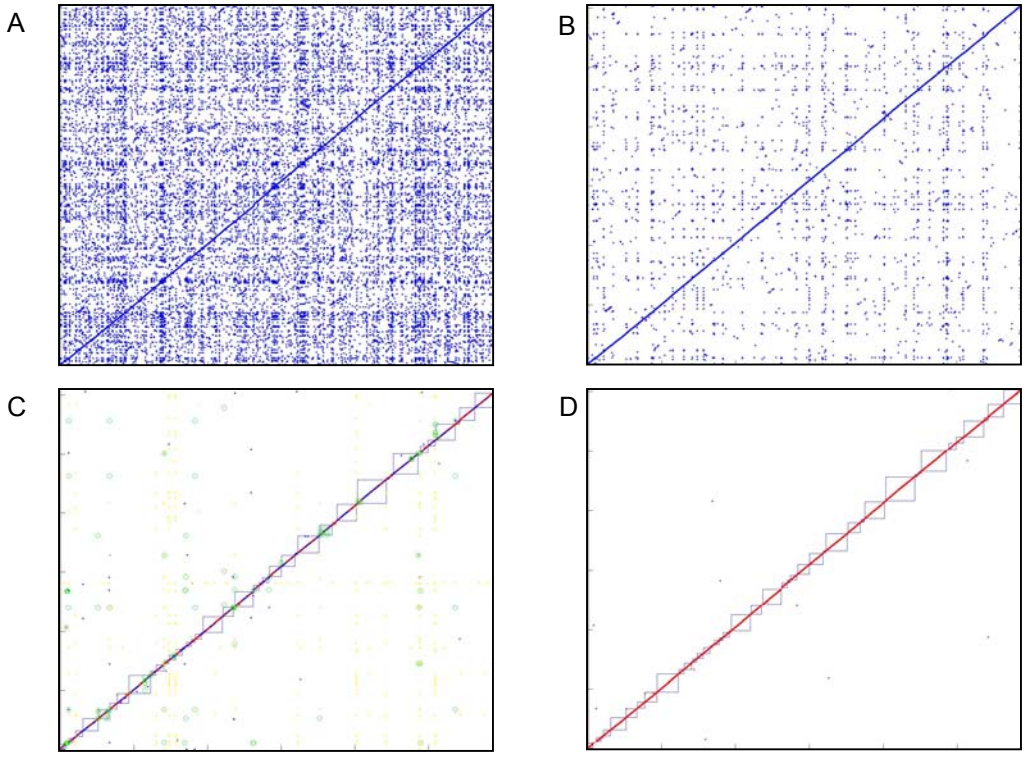


Figure 1.5. Performance of the algorithm. Dotplot representation of the bipartite graph. The 16 chromosomes of *S. cerevisiae* are stacked end-to-end along the y-axis, and the scaffolds of *S. paradoxus* are shown along the x-axis. Every point (x,y) represents an edge between *S. paradoxus* gene y and *S. cerevisiae* gene x. A. Initial bipartite graph. B. Graph resulting from initial disambiguation step. C. Graph resulting from use of BUS and synteny information. D. Unambiguous matches in graph C.

1.8. Performance of the algorithm

We applied this algorithm to automatically annotate the assemblies of the three species of yeast. Our Python implementation terminated within minutes for any of the pairwise comparisons. We successfully resolved the graph of sequence similarities between the four species, and found important biological implications in the resulting graph structure.

Figure 1.5 illustrates the performance of the algorithm for the 6235 annotated ORFs in *S. cerevisiae* and all predicted ORFs in *S. paradoxus*. The graph is initially very dense (panel A), the vast majority of edges representing non-orthologous matches, mostly due to protein domain similarities, ancient duplications that precede the time of the common ancestor of the species compared, and transposable elements. After applying the initial pruning step, many of the spurious matches are eliminated (panel B). The presence of unambiguous matches allows us to build blocks of conserved gene order, and use these to resolve additional matches using the BUS algorithm (panel C). The unambiguous 1-to-1 matches are mostly syntenic for *S. paradoxus*, thus ensuring that we are comparing orthologous regions.

More than 90% of genes have clear one-to-one orthologous matches in each species, providing a dense set of landmarks (average spacing ~2 kb) to define blocks of conserved synteny covering essentially the entire genome. Not surprisingly, transposon proteins formed the largest homology groups. The remaining matches were isolated in small subgraphs. These contain expanding gene families that are often found in rapidly recombining regions near the telomeres, and genes involved in environmental adaptation, such as sugar transport and cell surface adhesion[29]. For additional details see section 6.2.

We have additionally experimented running only BUS without the original pruning and synteny steps. More than 80% of ambiguities were resolved, and the remaining matches corresponded to duplicated ribosomal proteins and other gene pairs that are virtually unchanged since their duplication. The algorithm was slower, due to the large initial connectivity of the graph, but a large overall separation was obtained. Figure 1.6 compares the dotplot of *S. paradoxus* and *S.cerevisiae* with and without the use of synteny. Every point represents a match, the x coordinate denoting the position in the *S.paradoxus* assembly, and the y coordinate denoting the position in the *S.cerevisiae*

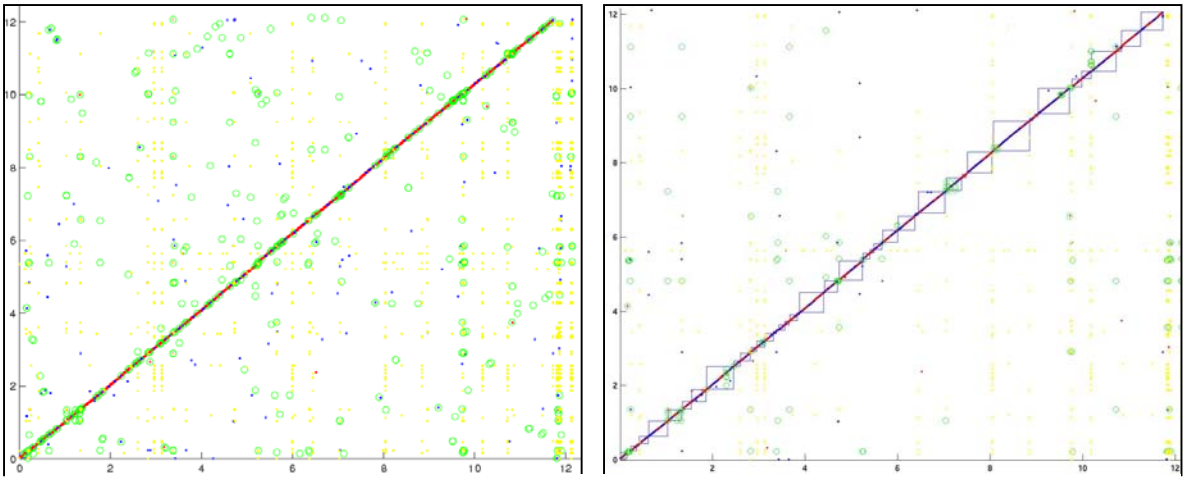


Figure 1.6. The effect of using synteny. Blocks of conserved gene order (blue squares) help resolve additional ambiguities. These are mostly due to pairs of anciently duplicated yeast genes.

genome, with all chromosomes put end-to-end. Lighter dots represent homology containing more than 15 genes (typically transposable elements) and circles represent smaller homology groups (rapidly changing protein families that are often found near the telomeres). The darker dots represent unambiguous 1-to-1 matches, and the boxes represent synteny blocks.

This algorithm has also been applied to species at much larger evolutionary distances, with very successful results (Kellis and Lander, manuscript in preparation). Despite hundreds of rearrangements and duplicated genes separating *S.cerevisiae* and *K.yarowii*, it successfully uncovered the correct gene correspondence between the two species that are more than 100 million years apart.

Additionally, the algorithm works well with unfinished genomes. By working with sets of genes instead of one-to-one matches, this algorithm correctly groups in a single orthologous set all portions of genes that are interrupted by sequence gaps and split in two or multiple contigs. A best bi-directional hit would match only the longest portion and leave part of a gene unmatched. Finally, since synteny blocks are only built on one-to-one unambiguous matches, the algorithm is robust to sequence contamination. A contaminating contig will have no unambiguous matches (since all features will also be present in genuine contigs from the species), and hence will never be used to build a synteny block. This has allowed the true orthologs to be determined and the contaminating sequences to be marked as paralogs.

This algorithm provides a good solution to determining genome correspondence, works well at a range of evolutionary distances, and is robust to sequencing artifacts of unfinished genomes.

1.9. Conclusion.

We have unambiguously resolved the one-to-one correspondence of more than 90% of *S. cerevisiae* genes. This provides us with a unique dataset whereby we can align and compare the evolutionary pressure of nearly every region in the complete yeast genome across four closely related relatives. In presence of gene duplication, some of the evolutionary constraints that a region is under are relieved, and uniform models of evolution would not capture the underlying selection for these sites. By ensuring that the regions compared are orthologous, we can make uniform assumptions about the rate of change of different regions, and apply statistical models for the significance of strong or weak conservation.

In this thesis, we will use the multiple alignments of the four species to discover protein-coding genes based on the pressure to conserve the reading frame of the amino acid translation (Chapter 2). We will also search for unusually strong conservation in non-coding regions to discover recurring patterns that constitute regulatory motifs (Chapter 3). We will assign functions to these motifs (Chapter 4) and discover their combinatorial interactions (Chapter 5) based on their conserved instances. Finally, we will focus on the differences between the species to discover regions and mechanisms of rapid evolutionary change (Chapter 6).

CHAPTER 2: GENE IDENTIFICATION

2.1. Introduction

The genome of a species encodes genes and other functional elements, interspersed with non-functional nucleotides in a single uninterrupted string of DNA. Recognizing protein-coding genes relies on finding stretches of nucleotides free of stop codons (called Open Reading Frames, or ORFs) that are too long to have likely occurred by chance. Since stop codons occur at a frequency of roughly 1 in 20 in random sequence, ORFs of at least 60 amino acids will occur frequently by chance (5% under a simple Poisson model) and even ORFs of 150 amino acids will appear by chance in a large genome (0.05%). This poses a huge challenge for higher eukaryotes in which genes are typically broken into many, small exons (on average 125 nucleotides long for internal exons in mammals[27]).

The basic problem is distinguishing *real genes* – those ORFs encoding a translated protein product – from *spurious ORFs* – the remaining ORFs whose presence is simply due to chance. The current public catalogue of yeast genes lists 6062 predicted ORFs that could theoretically encode proteins of at least 100 amino acids. Only two-thirds of these have been experimentally validated (*known*), and the remaining ~2000 ORFs are currently annotated as *hypothetical*. The total number of real protein-coding genes has been a subject of considerable debate, with estimates ranging from 4,800 to 6,400 genes (in mammalian genomes, estimates have ranged from 28,000 to more than 120,000 genes).

In this chapter, we use the comparative information to recognize real genes based on their conservation across evolutionary time. With the availability of genome-wide alignments across the four species, we first examined the different ways by which sequences change in known genes and in intergenic regions. The alignments of known genes revealed a clear pressure to preserve protein-coding potential. We constructed a computational test for reading frame conservation (RFC) and used it to revisit the annotation of yeast. We showed that more than 500 previously annotated ORFs are not meaningful and discovered 43 novel ORFs that were previously overlooked. We additionally refined the gene structure of hundreds of genes, including translation start,

stop, and exon boundaries. We show that our method has high sensitivity and specificity, and suggest changes that affect nearly 15% of yeast genes.

2.2. Different conservation of genes and intergenic regions

We examined the different types of conservation in genes and intergenic regions. We used the 1-to-1 orthologous anchors (see Chapter 1) to construct a nucleotide-level alignment of the genomes. The strong conservation of local gene order and spacing (Figure 2.1) allowed us to construct genome-wide multiple alignments. We aligned each gene together with its flanking intergenic regions using CLUSTALW[38] for the multiple alignments across the four species. When sequence gaps were present in one or more species, we constructed the alignment in multiple steps. We first aligned the gapless species creating a base alignment. Then we aligned each portion of a partially covered ortholog onto the base alignment, and constructed a consensus for each species based on the individually aligned portions. We marked missing sequence between contigs by a dot and disagreeing overlapping contigs by N. Finally, we constructed a multiple alignment of the four species by merging the piece-wise alignments. With sequence alignments at millions of positions across the four species, it is possible to obtain a precise estimate of the rate of evolutionary change, including substitutions and insertion-deletions (indels), in

Figure 2.1. Strong conservation of local gene order and spacing allows genome-wide multiple alignments. A 50kb segment of *S. cerevisiae* chromosome VII aligned with orthologous contigs from each of the other three species. Predicted ORFs are shown as arrows pointing in the direction of transcription. Orthologous ORFs are connected by dotted lines, and colored by the type of correspondence: red for 1-to-1 matches, blue for 1-to-2 matches and white for unmatched ORFs. Sequence gaps are indicated by vertical lines at ends of contigs, with estimated size of gap shown by the length of the hook.

the tree connecting the species. We counted transitions, transversions, insertions and deletions within these alignments and used these to estimate the rate of evolutionary change between the species. We counted the rate of synonymous and non-synonymous substitutions for every protein coding gene to find evidence of positive selection. The detailed results will be described in chapter 6.

We compared the rate of sequence change at aligned sites across the four species in intergenic and genic (protein-coding) regions (Figure 2.2). We found radically different types of conservation. Intergenic regions typically showed short stretches between 8 and 10 bases of near-perfect conservation, surrounded by non-conserved bases, rich in isolated gaps. Protein-coding genes on the other hand were much more uniform in their conservation, and typically differed in the largely-degenerate third-codon position. The proportion of sites corresponding to a different nucleotide in at least one of the three species is 58% in intergenic regions but only 30% in genic regions – a

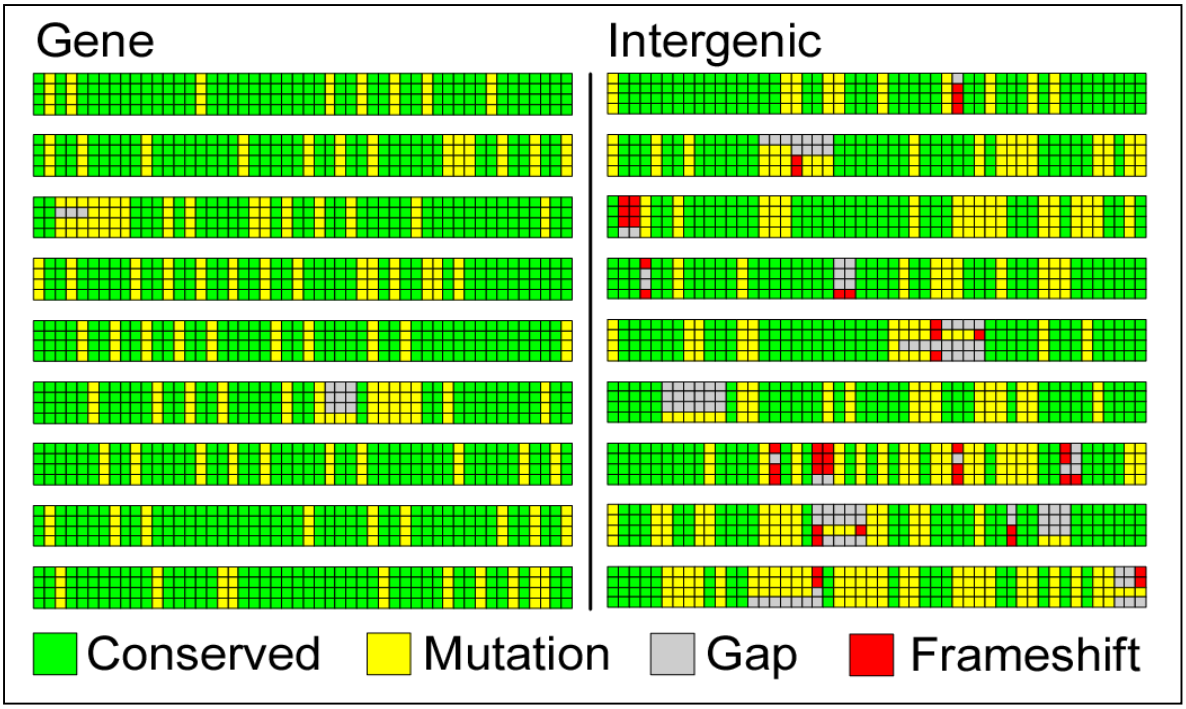


Figure 2.2. Patterns of change in genes and intergenic regions. Schematic representation of two multiple sequence alignments in ORF YMR017W and neighboring intergenic region. Aligned nucleotides across the four species are shown as stacked squares, colored by their conservation: green for conserved positions, yellow otherwise. Alignment gaps are shown in white and frame-shifting insertions (length not a multiple of 3) are shown in red. In addition to the abundance of frame-shift indels shown here, numerous in-frame stop codons are observed in the other three species.

difference of ~2-fold. The difference becomes much greater when one considers the gapped positions in alignments, representing insertion and deletion events (indels). The proportion of indels is 14% in intergenic regions, but only 1.3% in genic regions. The contrast is even sharper for indels whose length is not a multiple of three. These would disrupt the reading frame of a functional protein-coding gene, and are detrimental when they occur in real genes, unless they are compensated by a nearby indel that restores the reading frame. Frame-shifting gaps are found in 10.2% of aligned positions in intergenic regions, but only in 0.14% of positions in genic regions, a 75-fold strong separation. We used these alignment properties to recognize real genes.

2.3. Reading Frame Conservation Test

We developed a Reading Frame Conservation (RFC) test to classify each ORF in *S. cerevisiae* as biologically meaningful or not, based on the proportion of the ORF over which reading frame is locally conserved in each of the other three species. Each species with an orthologous alignment cast a vote for accepting or rejecting the ORF, and the votes were tallied to reach a decision for that ORF.

We evaluated the percent of nucleotides that are in the same frame within overlapping windows of the alignment. For every such window, we labeled each nucleotide of the first sequence by its position within a codon, as 1, 2 or 3 in order, starting at codon offset 1. We similarly labeled the nucleotides of the second sequence, but once for every start offset (1, 2, or 3). We then counted the percentage of gapless positions in the alignment that contained the same label in both aligned species, and selected the maximum percentage found in each of the three offsets of the second sequence (Figure 2.3). The final RFC value for the ORF was calculated by averaging the percentages obtained at overlapping windows of 100 nucleotides starting every 50

```
Scer     CTTCTAGATTTTCATCTT-GTCGATGTTCAAACAACGTGTTA-----TCAGAGAAACAGCTCTATGAGAAATCAGCTGATG
Scer_f1  123123123123123123-123123123123123123123312-----312312312312312312123123123123123123

Spar     TATTCATA-TCTCATCTTCATCAATGTTCAAACAGCGTGTTACAGACACAGAGAAACAGCTTC-TGAGAAGTCAGCCGGTG    RFC
Spar_f1  12312312-312312312312312312312312312312312312312312312312312312312-312312312312312312  → 43%
Spar_f2  23123123-123123123123123123123123123123123123123123123123123123123-123123123123123123  → 34%
Spar_f3  31231231-231231231231231231231231231231231231231231231231231231231-231231231231231231  → 23%
```

Figure 2.3. Reading Frame Conservation Test. Gaps in this alignment between *S. cerevisiae* and *S. paradoxus* change the correspondence of reading frame.

nucleotides. For overlapping ORFs in the *S. cerevisiae* genome ($n = 948$), the RFC was calculated only for the portion unique to each overlapping ORF. For spliced genes ($n = 240$), the RFC was calculated only on the largest exon.

We found that the distribution of frame conservation within each species is bimodal, and we chose a simple cutoff for each species, 80% for *S.paradoxus*, 75% for *S.mikatae* and 70% for *S.bayanus*. If the RFC of the best hit was above the cutoff, a species voted for keeping the ORF tested. If the RFC was below the cutoff and the hit was trusted as orthologous, the species voted for rejecting the tested ORF. Finally, if no orthologous hit could be found due to coverage, a species abstained from voting. We calculated a score between –3 and +3 for every ORF based on the number of species that accepted it (+1) and the number of species that rejected it (-1). We kept all ORFs with a score of 1 or greater, and rejected all ORFs with a score of –1 or smaller. We manually inspected the remaining ORFs.

We also applied this test to 3966 annotated ORFs with associated gene names (Table 2.4). These have been studied and named in at least one peer-reviewed publication, and are likely to be represent real genes. Only 15 of these (0.38%) were rejected (*KRE20, KRE21, KRE23, KRE24, VPS61, VPS65, VPS69, BUD19, FYV1, FYV2, FYV12, API2, AUA1, ICS3, UTR5, YIM2*). We inspected these manually and concluded that all were indeed likely to be spurious. Most lack experimental evidence. For the remainder, reported phenotypes associated with deletion of the ORF seems likely to be explained by fact that the ORF overlaps the promoters of other known genes.

	Accept	Reject
~4000 named genes	99.6%	0.1%
~300 intergenic regions	1%	99%
2000 Hypothetical genes	1500	500

Table 2.4. Testing all annotated protein-coding genes. The RFC test showed strong sensitivity and specificity, accepting virtually all experimentally verified genes (named genes) and rejecting all intergenic regions tested. We further applied this test to all the hypothetical genes and showed that more than 500 currently annotated genes are not real.

To investigate the power of the approach to reject spurious ORFs, we also applied it to a set of controls sequences consisting of 340 intergenic sequences in *S. cerevisiae* with lengths similar to the ORFs tested (Table 2.4). About 96% were rejected as having conservation properties incompatible with a biologically meaningful ORF, showing that the test has high sensitivity. Of the remaining 4% that were not rejected, close inspection shows that three-quarters appear to contain true ORFs. Some define short ORFs with conserved start and stop codons in all four species and others extend *S. cerevisiae* ORFs in the 5'- or 3'-direction in each of the other three species. Thus, at most 1% of true intergenic regions failed to be rejected by the RFC test.

The conservation-based gene identification algorithm we proposed has thus high sensitivity and specificity. In the next section, we apply it systematically for de-novo gene identification in *S. cerevisiae*.

2.4. Results: Hundreds of previously annotated genes are not real

When the yeast genome sequence was completed[23], 6275 ORFs were identified in the nuclear genome that could theoretically encode proteins encoding at least 100 amino acids and that do not overlap a longer ORF by more than half of their length (Figure 2.5). SGD has since updated the catalog based on complete resequencing and re-annotation of chromosome III, re-analysis of other chromosomes and reports in the scientific literature.

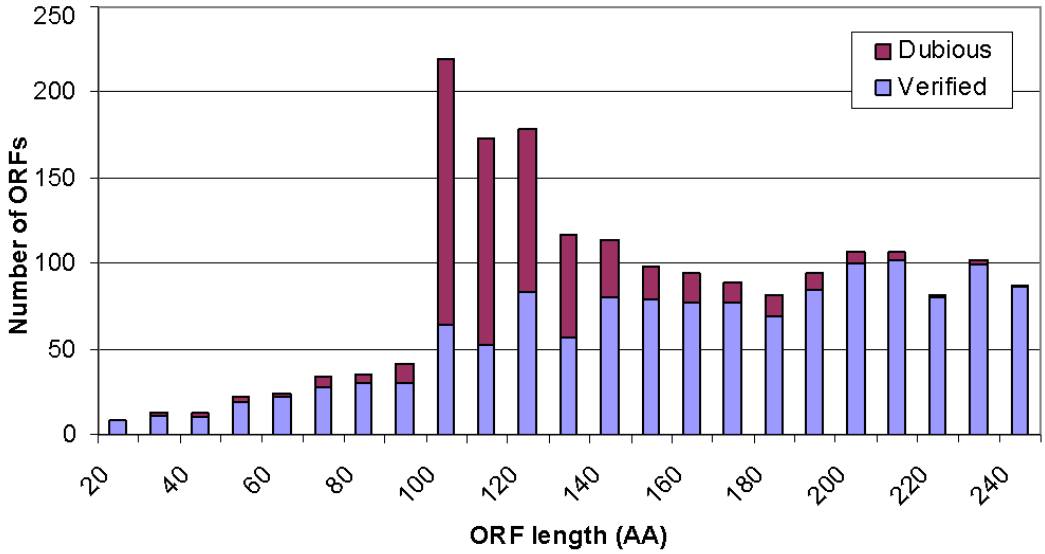


Figure 2.5. Rejected genes are mainly short. These are likely to be occurring by chance alone given the nucleotide composition of the yeast genome. The rejected genes show no evidence of function, such as mRNA expression, protein function, genetically or bio-chemically.

This resulted in a current version (as of May 2002) with 6062 ORFs ≥ 100 amino acids, consisting of 3966 'named' genes (described in at least one publication) and 2096 'uncharacterized' ORFs. SGD also includes a small collection of ORFs < 100 amino acids (see below).

We sought to apply the RFC test to all 6062 ORFs in SGD. A total of 117 could not be analyzed because they were almost completely contained within an overlapping ORF (99 cases, with average non-overlapping portion = 12 bp) or because an orthologous region could not be unambiguously defined in any of the species (18 cases). Of the 5945 ORFs tested, the analysis strongly validated 5550 ORFs. The vote was unanimous in 5458 (~98%) of cases. In the remaining cases, a valid gene appears to have degenerated in one of the four species. A total of 367 ORFs were strongly rejected. These rejections were unanimous in 63% of cases. In most of the remaining cases, *S. paradoxus* was too closely related to *S. cerevisiae* to have accumulated enough frameshifts to allow definitive rejection. The analysis deadlocked (one confirmation, one rejection, one abstention) for 28 ORFs (0.5%). We inspected these, together with the 117 cases that could not be analyzed due to overlaps and found convincing evidence (based on conservation of amino acids, start and stop codons, and presence of indels), that 20 are valid protein coding genes and 105 are spurious. We were unable to reach a judgment in the remaining 20 cases. Overall, a total of 5570 ORFs were accepted, 472 ORFs were rejected, and 20 remain ambiguous.

The vast majority of the rejections (96%) involve uncharacterized ORFs (for an example see Figure 2.6). SGD reports no compelling biological evidence (such as

Figure 2.6. Example of a rejected gene. DNA sequence that was previously though to encode a gene shows an accumulation of frame-shifting insertions and deletions (for color key see Figure 2.2). The sequence in fact does not correspond to a gene, get transcribed, or produce a protein product.

changes in mRNA expression) to suggest that these ORFs encode a true gene. Most of these overlap another well-conserved ORF, but show many insertions and deletions in the non-overlapping portion. The remainder tend to be small (median = 111 aa, with 93% ≤ 150 aa) and show atypical codon usage[23,39,40]. Figure 2.6 illustrates the case of an ORF of 333 bp that is clearly biologically meaningless. The orthologous sequence in all four species is laden with frameshifts (as well as stop codons). Only one rejected ORF, YBR184W, appears to represent a true gene that fails the RFC test because it is evolving very rapidly (see section 6.6).

In summary, the Reading Frame Conservation (RFC) test allowed a major revisiting of the yeast genome annotation. By observing the pattern of indels in the multiple alignment of predicted ORFs, it allowed us to automatically classify them as biologically meaningful or spurious. It reached a decision automatically in 98% of cases, accepting 99% of named ORFs and rejecting 99% of real intergenic regions, showing strong sensitivity and specificity. It resulted in a drastic reduction of the yeast gene count, rejecting nearly 500 ORFs. We next use the comparative information to refine the boundaries of ORFs.

2.5. Refining Gene Structure

Comparative genome analysis not only improves the recognition of true ORFs, it also yields much more accurate definitions of gene structure – including translation start, translation stop and intron boundaries. We used the comparative data to identify sequencing errors and refine the boundaries of true genes. Previous annotation of *S. cerevisiae* has defined the start of translation as the first in-frame ATG codon. However, the actual start of translation could lie 3' to this point, and the earlier in-frame ATG may be due to chance. Alternatively, if sequencing errors or mutations have obscured an earlier in-frame ATG codon, the true translation start could lie 5' to this point. Similarly, the annotated stop codon could be erroneously annotated, due to sequencing errors. Identifying the correct gene boundaries is important for many reasons, both experimental (for example to construct gene probes), as well as computational (for example to search for regulatory motifs).

We examined the multiple alignment of unambiguous ORFs to identify discrepancies in the predicted start and stop codons across the four species. We searched for the first in-frame ATG in each species and compared it to the annotated ATG in *S. cerevisiae*. In the *S. cerevisiae* start was not conserved, we automatically suggested a changed translation start if a subsequent in-frame ATG was conserved in all species and was the first in-frame ATG in at least one species. Otherwise, we searched for a conserved ATG 5' to that point. Similarly, we suggested changes in stop codons when a common stop in all other species disagreed with the *S. cerevisiae* annotation. We manually inspected the alignments to confirm that the suggested start and stop boundary changes agreed with conservation boundaries. We identified merges of consecutive *S. cerevisiae* ORFs, when they unambiguously matched a single ORF in at least one other species, and when their lengths added up to the length of the matching ORF.

We identified 210 cases in which the presumed translational start in *S. cerevisiae* does not correspond to the first in-frame start codon in at least two of the three other species (Figure 2.7 panel 1). In the vast majority of these cases, inspection of the sequence alignments provides strong evidence for an alternative conserved position for the translational start, either 3' or 5' to the previous annotation. We observed a lower overall conservation as well as frame-shifting indels outside the new boundaries. Similarly, we identified 330 cases in which the presumed translational stop codon in *S. cerevisiae* does not correspond to the first in-frame stop codon in at least two of the three

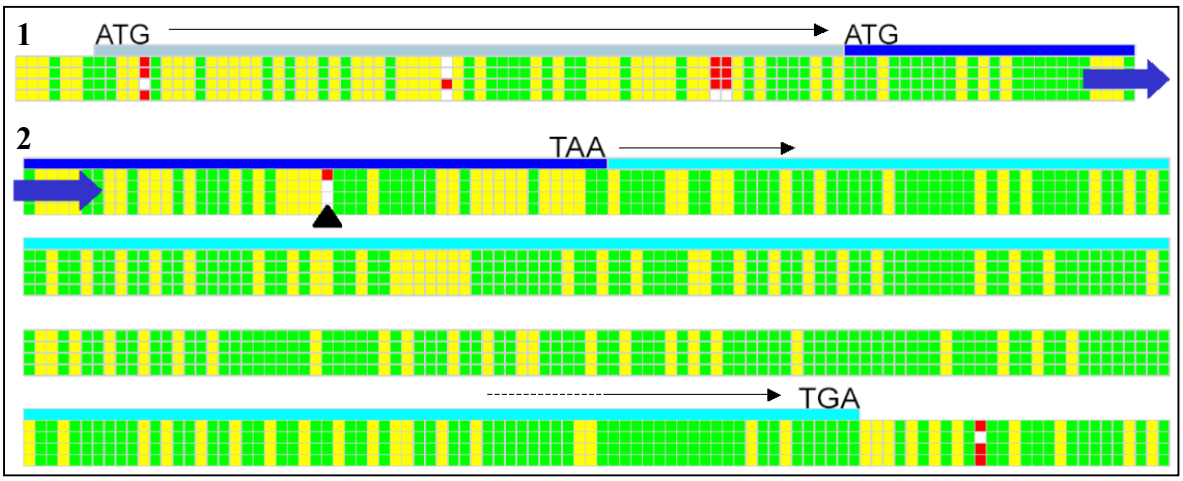


Figure 2.7. Refining gene boundaries. The start and stop codons of more than 300 genes have been refined based on the comparisons. These sometimes reveal sequencing errors in *S. cerevisiae*.

species. In ~25% of these cases, the other three species share a common stop codon and a single base change to the *S. cerevisiae* sequence would result in a stop codon in the corresponding location (Figure 2.7 panel 2). The remaining 75% of cases appear to represent true differences in the location of the translational stop across the species. Thus, stop codons appear to show more evolutionary variability in position than start codons.

We also developed methods for the automatic detection of frame-shifting sequencing errors. When regions of the multiple alignment shifted from one well-conserved reading frame to another well-conserved reading frame, we pinpointed regions of potential sequencing errors in each of the species. A number of these were detected in the reference sequence of *S. cerevisiae*. We confirmed 32 of these computational predictions by resequencing and found that in each case the published sequence was in error, and that the predicted erroneous nucleotide was always within a few base pairs from the experimentally confirmed sequencing error.

We identified 32 cases where two adjacent ORFs in *S. cerevisiae* are joined into a single ORF in all three other species. In every case, a single nucleotide change would suffice to join the ORFs in *S. cerevisiae* (either a substitution altering a stop codon or an indel altering the reading frame). In principle, these cases could represent errors in the genome sequence, mutations private to the sequenced strain S288C, or substitutions fixed in *S. cerevisiae*. We examined 19 cases by resequencing the relevant region in S288C. Our results revealed an error in the published sequence in 11 cases (establishing that there is a single ORF in S288C) and confirmed the published sequence in the remaining 7 cases. Sequencing of additional strains will be required to determine whether these remaining cases represent differences in S288C alone or in S. cerevisiae in general.

We also found two named ORFs (*FYV5* and *CWH36*) that pass the RFC test and cause phenotypes when deleted, but show no significant protein similarity across the four species. In both cases, inspection reveals that the opposite strand encodes a protein that shows strong amino acid conservation. (The latter gene has two introns, increasing the count of doubly spliced genes to 8.) In each case, we postulate that the protein responsible for the reported deletion phenotype is encoded on the opposite strand.

All merges and boundary refinements suggested specific changes to the nucleotide sequence of *S. cerevisiae* (except 3' changes of translation start that required

no change). To validate our predictions, we re-sequenced the sites of predicted sequence discrepancies. We used both forward and reverse reads in two different PCR reactions spanning the site. We examined 4 cases in which the comparative data suggested an earlier start codon and found, by resequencing, that all correspond to errors in the published sequence of S288C. We examined 17 such cases and found that 15 are explained by errors in the published sequence of S288C.

New Introns. We also examined the conservation of introns in the yeast genome. We studied 218 of the 240 ORFs reported in SGD to contain at least one intron (omitting the rest primarily due to lack of an orthologous alignment). In 92% of cases, the donor, branchpoint, and acceptor sites were all strongly conserved with respect to both location and sequence. Moreover, exon boundaries closely demarcated the domains of sequence conservation as measured by both nucleotide identity and absence of indels. Discrepancies were found in 17 cases, of which at least 9 strongly suggest that the previous annotation is incorrect. Five identify a new first exon (Figure 2.8) and four predict that a previously annotated intron is spurious.

We then sought to identify previously unrecognized introns by searching the *S. cerevisiae* genome for conserved splicing signals. We searched for conserved and proximal splice donor and branch signals and manually inspected the resulting alignments. Having constructed multiple alignments of ORFs and flanking intergenic regions, we searched for conserved splicing signals. We used 10 variants of splice donor signals (6-7bp) and 8 variants of branch site signals (7bp) that are found in

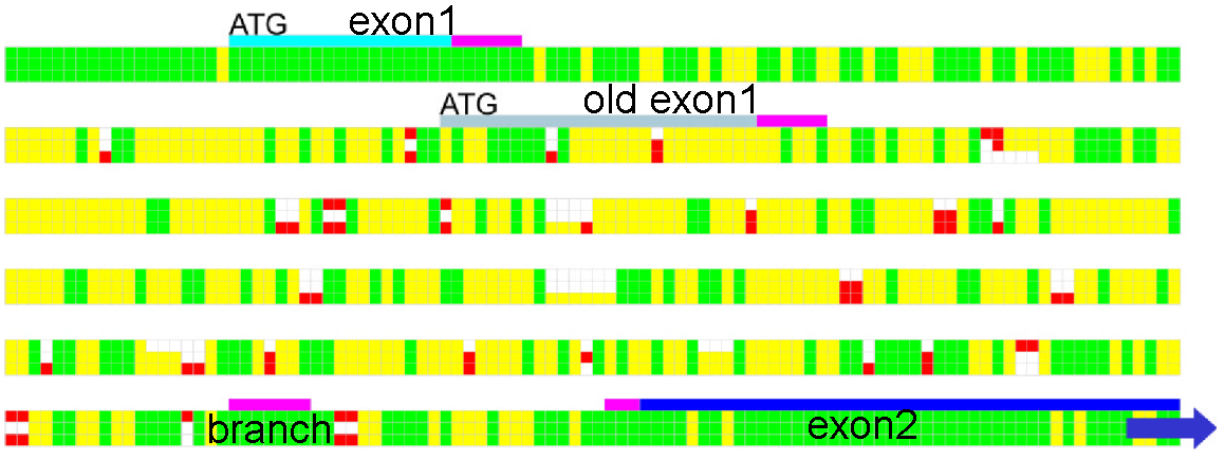


Figure 2.8. Identifying correct splicing. The short first exon was incorrectly annotated in *S. cerevisiae*. A shorter and earlier first exon is conserved across the four species, and corresponds to the correct splicing.

experimentally validated *S. cerevisiae* introns[41]. We searched each species independently but required that orthologous signals appear within 10 bp from each other in the multiple alignment of the region. We also required that branch and donor be no more than 600bp apart, which is the case for 90% of known *S. cerevisiae* introns. We then inspected the multiple alignment surrounding the conserved signals for three properties: (1) a conserved acceptor signal, [CT]AG, 3' of the branch site (2) high RFC 5' of the donor signal and 3' of the acceptor signal. (3) low RFC within the intron. Roughly half of the conserved donor/branch pairs met our additional requirements.

We predict 58 novel introns. Fifty cases affect the structure of known genes (defining new 5'-exons in 42 cases, 3'-exons in 7 cases and an internal splice in one case) and two indicate the presence of new genes. The relationship of the apparent splice signals to existing genes is unclear for the remaining six cases. We visually inspected our predictions and compared our results to experimental studies by Ares and colleagues that identified new introns using techniques such as microarray hybridization[41]. Of our 58 predicted introns, 20 were independently discovered by this group. Of the four annotated introns predicted to be spurious, all four show no experimental evidence of splicing. Our remaining predictions are currently being tested in collaboration with Ares and colleagues.

2.6. Analysis of small ORFs

The power of our method was limited for small ORFs. Smaller regions may indeed show lack of indels due to chance, and hence a high reading frame conservation score may not be meaningful.

We tested 141 ORFs encoding 50-99 amino acids for which some biological evidence has been published and are reported in SGD. Applying the RFC test and inspecting the results, we conclude that 120 appear to be true genes, 18 appear to be spurious ORFs and 3 remain unresolved. SGD also lists 32 ORFs encoding < 50 aa. We did not undertake a systematic search for all such ORFs, because control experiments showed that the RFC test lacked sufficient power to prove the validity of such small ORFs (see below). However, it is able to reject 7 of the 32 ORFs as likely to be spurious. Our yeast gene catalogue thus contains 188 short genes (<100 aa), of which 43 are novel.

To evaluate the predictive power of the RFC test for small ORFs, we additionally tested for presence of in-frame stop codons in the other species. When a small ORF in *S. cerevisiae* showed a strong overall frame conservation, we measured the length of the longest ORF in the same orientation in each orthologous locus. We measured the percent of the *S. cerevisiae* length that was open in each species (no stop codons), and took the minimum of the three percentages (OPEN) across the three additional species. When the reading frame was open in each of the other species, the lengths found were identical to that of *S. cerevisiae*, and OPEN was 100%. When OPEN was below 80%, we concluded that stop codons appeared in the orthologous sequence, and therefore that the RFC test falsely accepted a segment that did not correspond to a true gene. We observed the distribution of OPEN for different values of RFC. For *S. cerevisiae* ORFs between 50 and 100 amino acids (aa), selecting for high RFC automatically selected for high OPEN, and we estimated the test has high specificity. For ORFs between 30 and 50 aa however, only a small portion of the ORFs with high RFC show a high OPEN, and we conclude that the lack of indels within the small interval considered is not due to selective pressure, but instead lack of evolutionary distance between the species aligned.

We further systematically searched the remainder of the *S. cerevisiae* genome and evaluated all ORFs in this size range. Control experiments demonstrated that the RFC test has high power to discriminate reliably between valid and spurious ORFs in this size range. The genome contains 3161 such ORFs, nearly all are readily rejected by the RFC test. However, 43 novel genes were identified. These ORFs not only pass the RFC test, but they also have orthologous start and stop codons. Five of these have been reported in the literature subsequent to the SGD release studied here

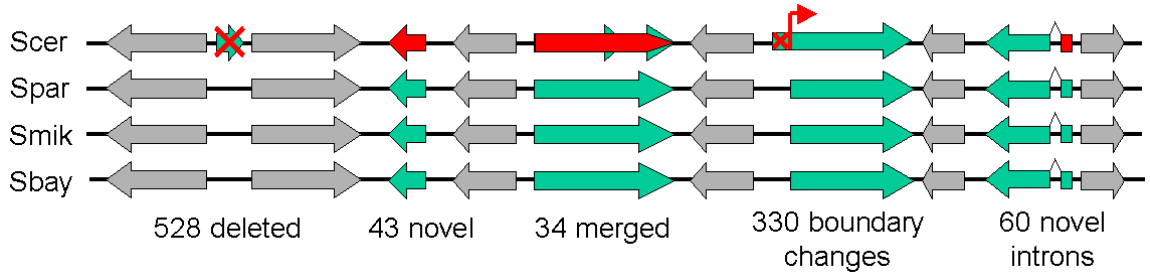


Figure 2.9. Revised yeast catalogue. Our analysis has affected nearly 15% of all genes.

2.7. Conclusion: Revised yeast gene catalog

Based on the analysis above, we propose a revised yeast gene catalog consisting of 5538 ORFs ≥ 100 amino acids. This reflects the proposed elimination of 503 ORFs (366 from the RFC test, 105 by manual inspection and 32 through merger). A total of 20 ORFs in SGD remain unresolved. Complete information about the gene catalog is provided in [29] and will be discussed more fully in a subsequent manuscript in collaboration with SGD and other yeast investigators. The revised gene count is consistent with at least two recent predictions based on light shotgun coverage of related species[4,5]. We believe that this represents a reasonably accurate description of the yeast gene set, because the analysis examines all ORFs ≥ 100 amino acids, the methodology has high sensitivity and specificity and the evidence is unambiguous for the vast majority of ORFs. Nonetheless, some errors are likely to remain. The results could be confirmed and remaining uncertainties resolved by sequencing of additional related yeast species, as well as by other experimental methods.

Despite the intensive study of *S. cerevisiae* to date, comparative genome analysis points to the need for a major revision of the yeast gene catalog affecting more than 15% of all ORFs (Figure 2.9). The results suggest that comparative analysis of a modest collection of species can permit accurate definition of genes and their structure. Comparative analysis can complement the primary sequence of a species and provide general rules for gene discovery that do not rely solely on known splicing signals for gene discovery. Previous studies have shown that such methods are also applicable to the understanding of mammalian genes[42]. The ability to observe the evolutionary pressures that nucleotide sequences are subjected to radically changes our power for signal discovery.

CHAPTER 3: REGULATORY MOTIF DISCOVERY

3.1. Introduction

Regulatory motifs are short nucleotide sequences typically upstream of genes that are used to control the expression of genes, dictating under which conditions a gene will be turned on or off. Direct identification of regulatory elements is more challenging than that of genes. Such elements are typically short (6-15 bp), tolerate some degree of sequence variation and follow few known rules. To date, the majority have been found by experimentation, such as systematic mutation of individual promoter regions; the process is laborious and unsuited for genome-scale analysis.

Computational analysis of single genomes has been successfully used to identify regulatory elements associated with known sets of related genes[7-9]. These methods typically search for frequently-occurring sequence patterns at various distances upstream of coordinately expressed genes, and will be further described in chapter 4. They are however limited by the experimental information available, and hence do not permit a comprehensive direct identification of regulatory elements[43].

Comparative genomics offers various approaches for finding regulatory elements. The simplest approach is to perform cross-species sequence alignment to find *phylogenetic footprints*, regions of unusually high conservation. This approach has long been used to study promoters of specific genes in many organisms[10,12,44-46] and recently was applied across the entire human and mouse genomes[19]. The genome alignments of the four *Saccharomyces* species can similarly be used to study each yeast gene, to help define promoters and other islands of intergenic conservation (Figure 3.2).

Our interest was to go beyond inspection of individual islands of conservation to construct a comprehensive dictionary of regulatory elements used throughout the genome. We investigated the conservation properties of known regulatory motifs and used the insights gained to design an approach for *de novo* discovery of regulatory motifs directly from the genome.

In this chapter, we develop and apply methods for genome-wide motif discovery. We compare our results to a database of experimentally validated regulatory motifs and rediscover virtually all previously known motifs. In chapter 4 we develop methods for

inferring a candidate function for the motifs discovered making use of biological knowledge about genes, and in chapter 5 we explore their combinatorial interactions.

3.2. Regulatory motifs

The current knowledge of gene regulation is based on focused experimental studies of specific examples. The deletion of a transcription factor was shown to disrupt the use of its target genes. Regulatory elements were identified in genetic screens through function-disrupting mutations that reside outside of a protein-coding ORF. Systematic mutagenesis of a particular promoter region (also known as promoter bashing) and testing the resulting effect on gene expression has been used to identify functional blocks in upstream regions of genes. To identify regulatory motifs at a nucleotide level, footprinting methods can be used. These methods expose the bound region to DNA damaging agents that degrade unbound nucleotides, leaving a 'footprint' of the transcription factor on the bound and thus protected nucleotides. Finally, even higher resolution information is obtained through crystal structures of transcription factors bound to DNA. These different methods have produced lists of bound sites for each of a small number of well-studied transcription factors.

The sites bound by these factors exhibit sequence similarities that reveal the binding specificity of each factor, and can be represented in a *regulatory motif*. Representations for these motifs range from consensus sequences listing the nucleotides involved in binding, to weight matrices and graphical models. *Consensus sequences* or *sequence profiles* are the simplest such representation, giving a list of possible bases for each position in the bound site. Some positions are strict and require the presence of a particular nucleotide, others allow for degeneracies. These can be represented compactly using the IUB standard one-letter code (Table 3.1). More complex representations can be used allowing for more detail in the binding specificity.

IUB	Nucleotides	Name	$[p_A,p_C,p_G,p_T]$
A	A	Adenine	[1, 0, 0, 0]
C	C	Cytosine	[0, 1, 0, 0]
G	G	Glutamine	[0, 0, 1, 0]
T	T	Tyrosine	[0, 0, 0, 1]
S	C or G	Strong	[0, ½, ½, 0]
W	A or T	Weak	[½, 0, 0, ½]
R	A or G	PuRine	[½, 0, ½, 0]
Y	C or T	pYrimidine	[0, ½, 0, ½]
M	A or C	aMino group	[½, ½, 0, 0]
K	G or T	Keto group	[0, 0, ½, ½]
B	C or G or T	Not A	[0, ⅓, ⅓, ⅓]
D	A or G or T	Not C	[⅓, 0, ⅓, ⅓]
H	A or C or T	Not G	[⅓, ⅓, 0, ⅓]
V	A or C or G	Not T	[⅓, ⅓, ⅓, 0]
N	A, C, G or T	aNy base	[¼, ¼, ¼, ¼]

Table 3.1. Degenerate nucleotide code.

A *weight matrix* representation of a motif of length L assigns weight vector $w_i = [w_A, w_C, w_G, w_T]$ to every position *i* between 1 and L. The binding strength of a sequence can be scored against a weight matrix by simply adding up the corresponding scores for each position. In a probabilistic framework, the weights can represent the relative frequencies of each nucleotide in real motifs, multiplying across the corresponding weights gives the probability that a sequence ***s*** matches the motif represented by ***m***. Alternatively, if log probabilities are used instead, summing across the matrix gives the corresponding log probability. This probability can be compared to the probability of obtaining ***s*** by chance, to obtain a log-likelihood ratio that the sequence matches the motif. Both consensus sequences and weight matrices model the binding contributions of nucleotide position as independent. More complex Bayesian representations for motifs can be used to capture pairwise and multiple dependencies between positions. As the models become more complex however, the increased power comes at a cost, increasing the number of parameters and possibly overfitting data.

Transcription factors have evolved different ways to contact the DNA double helix, and these are reflected in different types of regulatory motifs. Some factors make one long contact with the DNA helix recognizing between 6 and 8 positions, some of which can be degenerate. One such example is the Mbp1 transcription factor involved in the timing of events such as DNA replication during cell division and recognizes the motif ACGCGT. Other factors contact the DNA at two different points, resulting in motifs with two cores, separated by a stretch of unspecified bases. For example, the binding site recognized by Abf1, a general transcription factor involved in silencing and replication, recognizes the motif RTCRYNNNNNACGR. The DNA-binding domains of other factors are made of two identical parts (and hence called *homodimers*), contacting each other and each contacting the DNA helix. The two parts recognize identical sequences, but on opposite strands, and hence result in motifs that are *reverse palindromes* of themselves. One such example is the Gal4 factor involved in galactose metabolism, recognizing CGGNNNNNNNNNNCCG, namely CGG on one strand spaced by 11 nucleotides (one full turn of the double helix) from its reverse complement, CCG.

3.3. Extracting signal from noise

Computationally, discovering regulatory motifs amounts to extracting signal from noise. When the motifs searched are expected to be more frequent than other patterns of the same length, one can apply discovery algorithms such as Expectation Maximization (EM) or Gibbs sampling (and others reviewed in ref [9]). These were pioneered by Lawrence and coworkers[47], and made popular in software programs like MEME[7,48], AlignACE[8,49,50] or BioProspector[51]. More recent work has extended these methods to incorporate phylogenetic footprinting[45,52-54]. These methods separate the motif discovery problem in two sub-problems. (1) Given a set of starting coordinates $i_1, \ldots, i_n$ in each of the sequences, construct the optimal matrix representation for a motif that starts at each of these positions. (2) Given a matrix representation for a motif m, find the starting positions of the best matches for that motif in each of the sequences. These algorithms start with a random assignment for the start positions and infers the best matrix, then iterates to improve the assignment of start positions to better match the motif. EM algorithms choose the optimal assignment for each of these rounds of iteration. Gibbs sampling algorithms instead sample amidst the best start positions. Both algorithms converge as long as the motif searched is actually frequent in the sequences searched, since probabilistically, the algorithms will be likely to sample these motifs in their iterative steps, and upon sampling them will converge to include them.

These methods have typically been applied to the upstream sequences of small sets of genes, but are not applicable to a genome-wide discovery. Instead, k-mer counting methods have been used to find short sequences that occur more frequently in intergenic regions, as compared to coding regions in a genome-wide fashion[43]. However, these typically find very degenerate sequences (such as poly-A or poly-T) and have shown limited power to separate regulatory motifs from the mostly non-functional intergenic regions. This is largely due to the small number of functional instances of regulatory motifs, as compared to the large number of non-functional nucleotides. The discovery of regulatory motifs still relies heavily on extensive experimentation.

Comparative genomics provides a powerful way to distinguish regulatory motifs from non-functional patterns based on their conservation. In this chapter we first study conservation properties of known regulatory motifs. We use these to construct three tests

to detect the genome-wide signature of motif-like conservation. We use these tests to detect all significant patterns with strong genome-wide conservation, constructing a list of 72 genome-wide motifs. We compare this list against previously identified regulatory motifs and show that our method has high sensitivity and specificity, detecting most previously known regulatory motifs, but also a similar number of novel motifs. In chapter 4, we assign candidate functions to these novel motifs, and in chapter 5, we study their combinatorial interactions.

3.4. Conservation properties of known regulatory motifs

We first studied the binding site for one of the best studied transcription factors, Gal4, whose sequence motif is $CGG(N)_{11}CCG$ (which contains 11 unspecified bases). Gal4 regulates genes involved in galactose utilization, including the *GAL1* and *GAL10* genes that are divergently transcribed from a common intergenic region (Figure 3.2). The Gal4

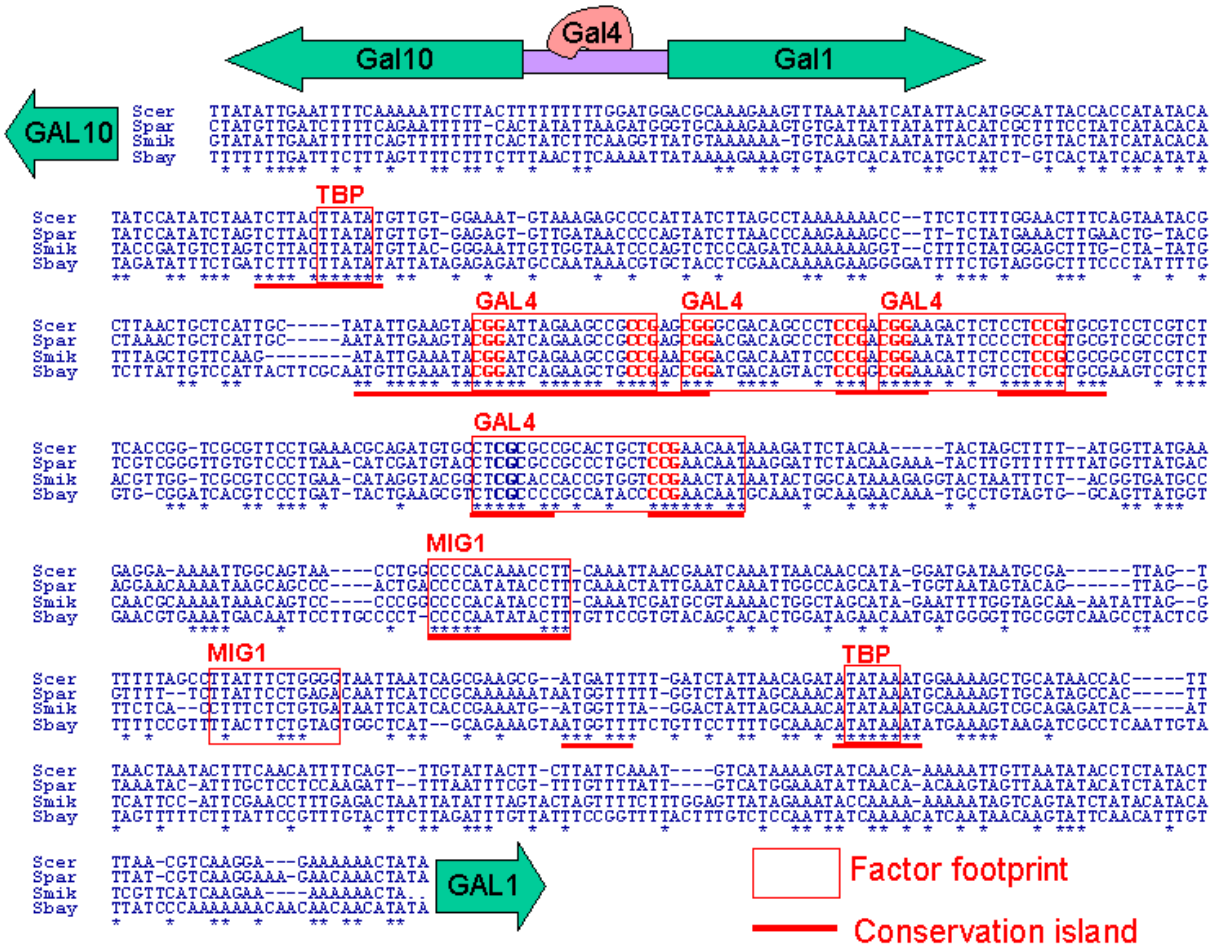
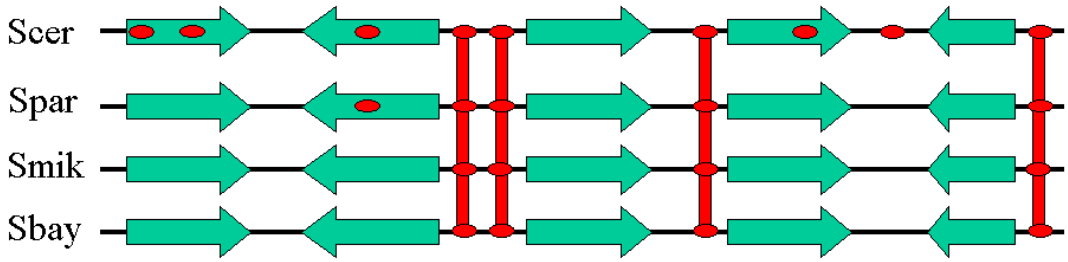


Figure 3.2. Phylogenetic footprinting of the Gal1-Gal10 intergenic region reveals functional nucleotides.

motif occurs three times in this intergenic region, and all three instances show perfect conservation across the four species. In addition, there is a fourth, experimentally validated binding site[55] for Gal4 that differs from the consensus by one nucleotide in *S. cerevisiae*. This variant site is also perfectly preserved across the species.

We then examined the frequency and conservation of Gal4 binding sites across the aligned genomes (Figure 3.3). In *S. cerevisiae*, the Gal4 motif occurs 96 times in intergenic regions and 415 times in genic (protein coding) regions. The motif displays certain striking conservation properties. First, occurrences of the Gal4 motif in intergenic regions have a conservation rate (proportion conserved across all four species) that is ~5-fold higher than for equivalent random motifs (12.5% vs. 2.4%). Second, intergenic occurrences of the Gal4 motif are more frequently conserved than genic occurrences (12.5% vs. 3%). By contrast, random motifs are less frequently conserved in intergenic regions than genic regions (3.1% vs. 7.0%), reflecting the lower overall level of conservation in intergenic regions. Thus, the relative conservation rate in intergenic vs. genic regions is ~11-fold higher for Gal4 than for than random motifs. Third, the Gal4 motif shows a higher conservation rate in divergent vs. convergent intergenic regions (those that lie upstream vs. downstream of both flanking genes); no such preferences is seen for control motifs. These three observations suggest various ways to discover motifs based on their conservation properties (see conservation criteria below).



Evaluate conservation within:	Gal4	Controls
(1) All intergenic regions	12.5%	2.4%
(2) Intergenic : coding	12:3	3:7
(3) Upstream : downstream	12:0	1:1

Figure 3.3. Genome-wide conservation of the Gal4 motif. The six-fold to 11-fold separation between the conservation of Gal4 and that of random control motifs suggests three signatures for motif discovery.

We extended these observations by assembling a catalog of 55 known regulatory sequence motifs (Table 3.4), by starting with two public databases (SCPD[56,57] and YTFD[58]) and curating the entries to select those with the best support in the literature.

	Known motif		Discovered motif			
Factor	Motif	MCS	Motif	Genome-wide	Category- based	MCS
ABF1	**RTCRYnnnnnACG**	50.0	**RTCRY**knnnn**ACG**R	S	S	36.2
UME6	**TCGGCGGCTA**	20.9	**TSGGCGGCTA**WW	S	NC	23.4
CBF1	**RTCACRTG**	19.0	**RTCACGTG**V	S	S	17.6
NDT80	**TCGGCGGCTDW**	18.6	**TSGGCGGCTAW**W	S	NC	23.4
REB1	**TTACCCGG**	17.8	R**TTACCCGR**M	S	S	34.3
MCM1a	T**TWCC**Cn**WWW**R**GGAAA**	16.5	**TTCCn**a**Attn**G**GAAA**	S	S	13.8
SWI6	**ACGCGT**	16.4	**WCGCGT**CGCGt	S	S	10.2
PHO4	**CACGTG**	16.1	RT**CACGTG**V	S	S	17.6
MBP1	**ACGCGTnA**	14.8	**WCGCGT**CGCGt	S	S	10.2
SWI4	**TTTTCGCG**	12.4	**WTTTCGCG**TT	S	S	12.0
DAL81	GATAAG	12.1	–	–	NE	–
RPN4	**TTTTGCCACC**	11.5	**TTTTGCCACC**G	S	NC	11.0
MSN2	**CCCCT**	11.3	hR**CCCYT**WDt	S	NE	7.8
MSN4	**CCCCT**	11.3	hR**CCCYT**WDt	S	NE	7.8
PDR1	**CCGCGG**	9.3	Y**CCGSGG**S	S	NE	6.7
ESR2	**AAAAWTTTT**	8.9	GR**RAAAWTTTT**CACT	S	NC	15.6
MIG1	**CCCCRS**WWWW	8.7	D**CCCCGC**GH	S	NE	8.2
MIG1b	**CCCCGC**	8.4	D**CCCCGC**GH	S	NE	8.2
BAS1	**TGACTC**	8.3	A**TGACTC**WT	S	S	6.1
GCN4	**ATGACTCAT**	8.2	**ATGACTCWT**	S	S	6.1
GAL4	**CGGnnnnnnnnnnnCCG**	8.0	**CGG**CnnMGnnnnnnn**CG**C	S	S	5.0
HSF1b	**TTCTAGAA**	7.8	**TTCTMGAA**GA	S	S	7.0
ESR1	**GATGAG**	7.7	gc**GATGAG**mtgaraw	S	NC	24.7
MET31	AAA**CTGTGGC**	6.8	**SKGTGGS**Gc	S	S	8.1
AFT1	Y**RCACCC**R	6.8	**RVACCC**TD	S	NC	10.3
TEA1	CGGnCGG	6.8	–	–	NC	–
PUT3	**CGG**nnnnnnnnn**CCG**	6.2	**C**C**G**Mnnnnnnnnn**mSG**R	W	NE	5.4
HAP2	**TGATTGG**C	5.7	**TGATTGG**T	–	S	[6.4]
RAP1	**ACACCCA**T**ACATT**T	5.2	**ACACCCA**C**ACATnn**C	S	S	9.9
LEU3	**CCGGnnCCGG**	4.9	**CCSG**TAn**CGG**	S	S	6.5
MCM1b	Y**TTCC**T**AATT**W**G**nnCn	4.8	**TTCCnaAttnG**GAAA	S	S	13.8
INO4	**CATGTGAA**AT	4.1	Gnnn**CATGTGAA**	–	S	[6.8]
INO2	**CATGTG**AAAT	4.1	**CATGTG**	–	S	[4.4]
GLN3	GATAAK	3.8	–	–	NE	–
ADR1	GGAGA	3.7	–	–	NE	–
FKH2	**TTGTTTAC**S**T**	3.6	t**TTGTTTACnT**TT	S	S	10.8
FKH1	**TTGTTTAC**S**T**	3.6	t**TTGTTTACnT**TT	S	S	10.8
RLM1	**CTAWW**W**WT**AG	3.6	**CTAnnTTTAG**	S	S	[4.7]
SWI5	**KGCTGR**	3.4	**TGCTGG**	–	S	[6.1]
HAP1	C**G**Gnnn**TAnCGG**	2.5	**G**CnnT**TAnCGG**	S	NC	4.8
XBP1	M**CTCGARRR**nR	2.5	T**CTCGARRA**	S	NC	12.5
MAC1	TT**TGCTCA**	2.3	**TGCTCA**	–	S	[5.4]
TBF1	**TTAGGG**	2.3	G**KBAGGG**T	S	NC	4.8
MSE	**TTTTGTG**	1.4	**TTTTGTG**TCRC	S	NC	9.9
STE12	R**TGAAACA**	0.7	Y**TGAAACA**	–	S	[12.2]
DIG1	R**TGAAACA**	0.7	Y**TGAAACA**	–	S	[12.2]
MET4	**TGGCAAA**TG	0.7	CGG**TGGCAAA**A	S	NE	–
HAP4	**TnRTTGGT**	0.5	**T**G**ATTGGT**	–	S	[6.4]
SMP1	ACTACTAWWWWTAG	0.4	–	–	NE	–
ACE2	**GCTGGT**	−0.6	T**GCTGGT**	–	S	[7.4]
YAP1	TTACTAA	−1.1	–	–	NE	–
CIN5	TTACTAA	−1.1	–	–	NE	–
RME1	GAACCTCAA	−1.4	–	–	NE	–
HAC1	CAGCGTG	−1.4	–	–	NC	–
GCR1	**GGAAG**	−18.5	**GGAAG**C	–	S	[4.4]

Table 3.4. Genome-wide conservation of known motifs. Matching nucleotides in bold. S=strong match, W=weak match, NE=not enriched, NC=no category available. Category scores in brackets.

We defined a Motif Conservation Score (MCS) based on the conservation rate of the motif in intergenic regions. To evaluate the Motif Conservation Score (MCS) of a motif ***m*** of given length and degeneracy, we compared its conservation ratio to that of random patterns of the same length and degeneracy. We first computed the table ***F*** containing the relative frequencies of two-fold and three-fold degenerate bases, given the *S. cerevisiae* nucleotide frequencies (.32 for A and T, .18 for C and G). For example, W=[AT] (.32*.32) is a more likely two-fold degenerate base than Y=[CT] (.18*.32). We then selected 20 random intergenic loci in *S. cerevisiae*. For each of these loci, we used the order of nucleotides at that locus together with the order of degeneracy levels in ***m*** to construct a random motif. If the first character of ***m*** was two-fold degenerate and the first nucleotide at the selected locus was A, we picked a two-fold degenerate base containing A (W, R or M), their relative frequencies dictated by ***F***, and continued for every character of ***m***. We then counted conserved and non-conserved instances of each of the 20 generated control patterns and computed ***r***, the log-average of their conservation rates. We then counted the number of conserved and non-conserved intergenic instances of ***m***, and computed the binomial probability ***p*** of observing the two counts, given ***r***. We finally reported the MCS of the motif as a z-score corresponding to ***p***, the number of standard deviations away from the mean of a normal distribution that corresponds to tail area ***p***. Nearly all of these sequence motifs are binding sites of known transcription factors. Most of the known motifs show extremely strong conservation, with 60% having MCS ≥ 4 (which is substantially higher than expected by chance). Some of the motifs, however, show relatively modest MCS. These motifs may be incorrect, suboptimal or not well conserved.

3.5. Genome-wide motif discovery

Our methodology for genome-wide motif discovery involves first identifying conserved *mini-motifs* and then using these to construct full motifs (Figure 3.5). Mini-motifs are sequences of the form $XYZn_{(0\text{-}21)}UVW$, consisting of two triplets of specified bases interrupted by a fixed number (from 0 to 21) of unspecified bases. Examples are TAGGAT, ATAnnGGC, or the Gal4 motif itself. The total number of distinct mini-motifs is 45,760, if reverse complements are grouped together.

Conserved mini-motifs are evaluated according to three conservation criteria (CC1-3), based on our observations about the properties of the Gal4 motif. In each case, conservation rates are normalized to appropriate random controls. CC1 (Intergenic conservation) evaluates the conservation rate of a mini-motif in intergenic regions. CC2 (Intergenic-genic conservation) evaluates the stronger conservation in intergenic regions as compared to genic regions. CC3 (Upstream-downstream conservation) evaluates the different conservation of a mini-motif when it occurs upstream vs. downstream of a gene.

CC1: Intergenic conservation. We searched for mini-motifs that show a significant conservation in intergenic regions. For every mini-motif, we counted ***ic*** the number of perfectly conserved intergenic instances in all four species, and ***i*** the total number of intergenic instances in *S.cerevisiae*. We found that the two counts seem linearly related for the large majority of patterns (Figure 3.5 panel A), which can be

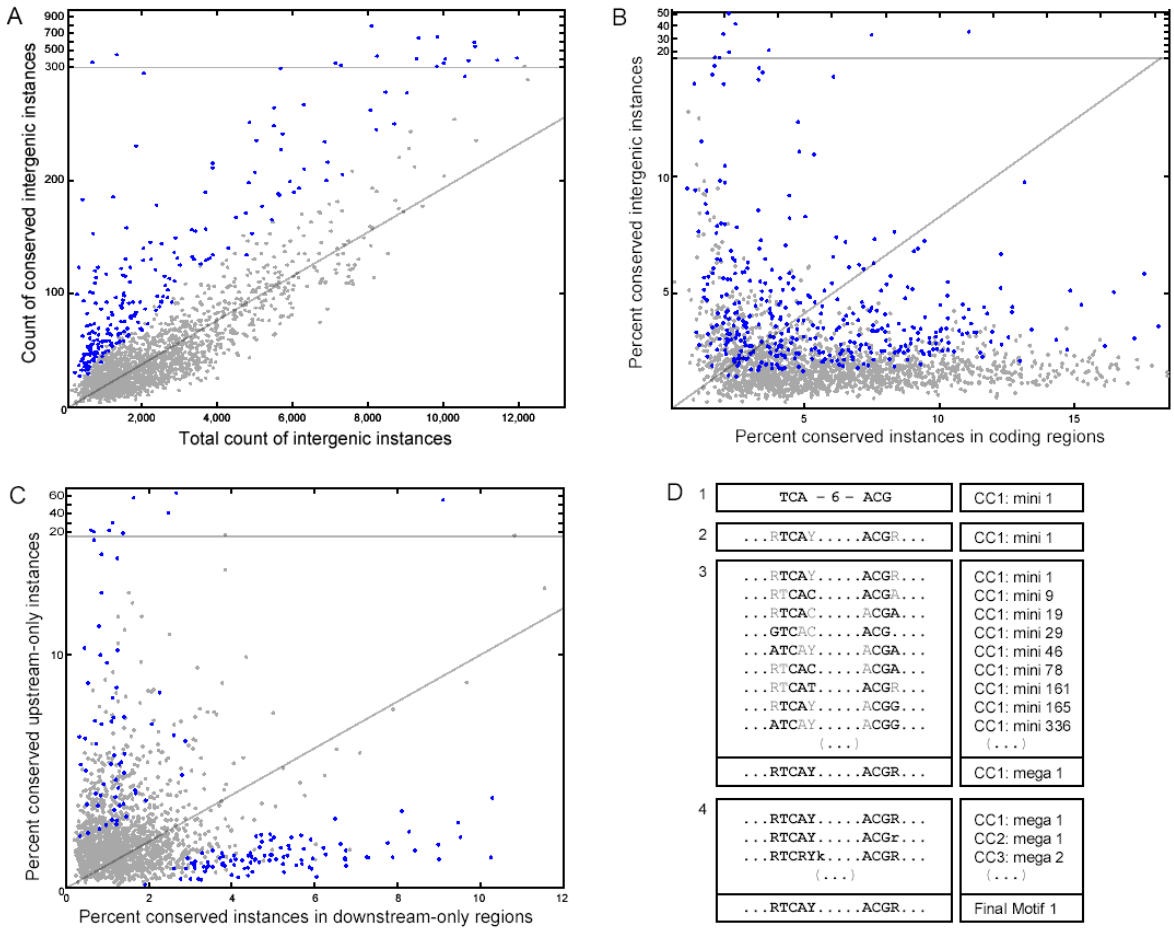


Figure 3.5. Genome-wide motif discovery method. The three conservation tests and motif collapsing.

attributed to a basal level of conservation ***r*** given the total evolutionary distance that separates the four species compared. We estimated the ratio ***r*** as the log-average of non-outlier instances of ***ic/i*** within a control set of all motifs at a given gap size. We then calculated for every motif the binomial probability ***p*** of observing ***ic*** successes out of ***i*** trials, given parameter ***r***. We assigned a z-score ***S*** to every motif corresponding to probability ***p***. This score is positive if the motif is conserved more frequently than random, and negative if the motif is diverged more frequently than random. We found that the distribution of scores is symmetric around zero for the vast majority of motifs. The right tail of the distribution however is much further than the left tail, containing 1190 motifs more than 5 sigma away from the mean, as compared to 25 motifs for the left tail. By comparing the two counts, we estimated that 94% of these 1190 motifs are non-random in their conservation enrichment.

CC2: Intergenic-genic conservation. We searched for motifs that are preferentially conserved in intergenic regions, as compared to coding regions. In addition to *ic* and *i* (see previous section), we counted the number of conserved coding instances *gc*, and the number of total coding instances *g*, for every mini-motif. We observed the ratio of conserved instances that are intergenic $a=ic/(ic+gc)$, and compared it to the total ratio of motif instances that are intergenic $b=i/(i+g)$. Not surprisingly, we found that typically b=25% of all motif instances appeared in intergenic regions, which account for roughly 25% of the yeast genome. Similarly, only a=10% of conserved motif instances appeared in intergenic regions, which reflects the lower conservation of intergenic regions. To correct for this typical depletion in intergenic conservation, we estimated a correction factor $f=a/b$ for mini-motifs of similar GC-content. Then for a given mini-motif, the proportion of all instances found in intergenic regions and the correction for the lower conservation of intergenic regions together gave us $r=f*i/(i+g)$, the expected ratio of conserved intergenic instances for that motif. We evaluated the binomial probability *p* of of observing at least *ic* conserved instances in intergenic regions and *ic+gc* conserved instances overall, given the expected ratio *r*. As in CC1, we computed a z-score *S* for every motif and found a distribution centered around zero for the large majority of motifs, and a heavier right tail. We selected 1110 motifs above 5 sigma and estimated that 97% are non-random as compared to only 39 motifs below -5 sigma.

CC3: Upstream-downstream conservation. We searched for motifs that are differentially conserved in upstream regions and downstream regions. We defined upstream-only intergenic regions in divergent promoters that are upstream of both flanking ORFs, and downstream-only intergenic regions in convergent 3' terminators that are downstream of both flanking ORFs. We then counted ***uc*** and ***u***, the conserved and total counts in upstream-only regions, and similarly ***dc*** and ***d*** in downstream-only regions. We found that upstream-only and downstream-only regions have similar conservation rates, and the ratios ***uc/u*** and ***dc/d*** are both similar to ***ic/i*** for the large majority of motifs. We thus used a simple chi-square contingency test on the four counts (uc,u,dc,d) to find motifs that are differentially conserved. We found 1089 mini-motifs with a chi-square value of 10.83 or greater, which corresponds to a p-value of .001. Given the multiple testing of 45760 mini-motifs, we estimated that roughly 46 will show such a score by chance and that 96% of the selected motifs will be non-random.

The conserved mini-motifs are then used to construct full motifs (Figure 3.5). They are first extended, by searching for nearby sequence positions showing significant correlation with a mini-motif. The extended motifs are then clustered, merging those with substantially overlapping sequences and those that tend to occur in the same intergenic regions. Finally, a full motif is created by deriving a consensus sequence (which may be degenerate). Motifs are typically degenerate, and a single full-motif can be responsible for multiple strong mini-motifs. We now describe methods to recover the full motifs and their degeneracy.

We extended each mini-motif selected by searching for surrounding bases that are preferentially conserved when the motif is conserved. We used an iterative approach adding at every iteration one base that maximally discriminates the neighborhood of conserved motif instances from the neighborhood of non-conserved motif instances. The added base was selected from fourteen degenerate symbols of the IUB code (A, C, G, T, S, W, R, Y, M, K, B, D, H, V). When no such symbol separated the conserved and non-conserved instances with significance above 3 sigma, we terminated the extension. Figure 3.5 panel D shows the top-scoring mini-motif found in CC1 (Row 1), and the corresponding extension (Row 2). We found that many mini-motifs have the same or similar extensions, and we grouped these based on sequence similarity. We measured the

similarity between two motifs as the number of bits in common in the best ungapped alignment of the two motifs, divided by the minimum number of bits contained in either

No.	Discovered motif	Location	MCS	Best category	CCS	Interpretation
1	YCGTnnnnmRYGAY	5′	36.2	ChIP: Abf1	90	Known: Abf1
2	RTTACCCGRM	5′	34.3	ChIP: Reb1	38	Known: Reb1
3	gcGATGAGmtgaraw	5′	24.7	Exp.: cluster 74	62	Known: Esr1 GATGAG
4	TSGGCGGCTAWW	5′	23.4	GO: meiosis	10	Known: Ume6/Ndt80
5	RTCACGTGV	5′	17.6	ChIP: Cbf1	27	Known: Cbf1/Pho4
6	WTATWTACADG	3′	17.4	Exp.: cluster 16 downstream	25	New: mitochondrial downstream
7	GRRAAAWTTTTCACT	5′	15.6	Exp.: cluster 74	37	Known: Esr2
8	TTCCnaAttnGGAAA	5′	13.8	ChIP: Mcm1	29	Known: Mcm1
9	CGTTTCTTTTTCY.	5′	13.5	GO: filamentation	7	New: filamentation
10	TYYTCGAGA.	5′	12.5	Exp.: cluster 86	5	Known: Xbp1 (Hsf1-co-ocuring)
11	TTTTCGCG	5′	12.0	ChIP: Swi4	21	Known: Swi4 fixed gap
11a	TTTT = CGCG	5′	12.0	ChIP: Swi4	–	New: Swi4 variable gap
12	TKACGCGTT	5′	12.0	ChIP: Mbp1	18	Known: Mbp1/Swi6
13	STGCGGnnnttTCTnnG	5′	11.8	GO: filamentation	11	New: filamentation
14	YCTATTGTT	5′	11.5	ChIP: Fkh2	6	New: Rlm1-like
15	TTTTGCCACCG	5′	11.0	GO: proteolysis	25	Known: Rpn4/Met4
16	tTTGTTTACnTTT	5′	10.8	ChIP: Fkh2	28	Known: Fkh1/2
17	RVACCCTD	5′	10.3	–	–	Known: Aft1
18	WCGCGTCGCGt	5′	10.2	ChIP: Mbp1	17	New: double Mbp1
19	GGGTnACCC	5′	10.0	ChIP: Reb1	8	New: Reb1 palindrome
20	GnnATGTGTGGGTGT	5′	9.9	ChIP: Fhl1	5	Known: Rap1
21	TTTTGTGTCRC	5′	9.9	ChIP: Sum1	14	Known: Mse
22	TTTCAnCGCGC	5′	9.8	–	–	New: no category
23	TATTAWTATTATtMtnatta	3′	9.5	–	–	New: no category
24	SCGnHGGS	5′	8.8	GO: filamentation	6	New: filamentation
25	ACAGCCGCRY	5′	8.6	Exp.: cluster 37	6	New: expression cluster 37
26	DCGCGGGGH	5′	8.1	Exp.: cluster 46	8	Known: Mig1b
27	SKGTGGSGc	5′	8.1	ChIP: Met31	5	Known: Met31
28	TTTTn(19)GCKCG	5′	7.8	–	–	Known: no category
29	HRCCCYTWDt	5′	7.8	Exp.: cluster 8	22	Known: Msn2/4
30	TKCCCnnnnGGG	5′	7.3	ChIP: Mcm1	15	Known: Mcm1 (hits tRNA)
31	GTGTCAGTAAt	5′	7.1	ChIP: Sum1	15	New: Sum1
32	RGTTTTTCCG	5′	7.1	ChIP: Rgt1	7	New: Rgt1
33	TTCTMGAAGA	5′	7.0	ChIP: Hsf1	10	Known: Hsf1
34	YCCGSGGS	5′	6.7	GO: filamentation	9	New: filamentation
35	CnCCTTTTATAC	5′	6.5	–	–	New: no category
36	CCSGTAnCGG	5′	6.5	ChIP: Leu3	8	Known: Leu3
37	SKTKCCTT	5′	6.4	GO: filamentation	7	New: filamentation
38	CTCCCCTTAT	5′	6.4	Exp.: cluster 8	11	Known: Msn2/4
39	GCCCGG	5′	6.3	GO: filamentation	10	New: filamentation
40	SGCGCGRB	5′	6.3	–	–	New: no category
41	CTCSGCS	5′	6.2	–	–	New: no category
42	TGnKAGCGCCG	5′	6.2	–	–	–
43	ATGACTCWT	5′	6.1	ChIP: Gcn4	44	Known: Gcn4/Bas1
44	CCGAnnnTCGG	5′	6.1	Exp.: cluster 46	6	New: facilitators palindrome
45	SCGMnnnnnnKCG	5′	6.0	–	–	New: no category
46	CnCCGCGCnnTTTs	5′	6.0	–	–	New: no category
47	TTTTnnnnnnnnnnnngGGGT	5′	5.8	–	–	New: no category
48	TGTRnCAW	3′	5.5	–	–	New: no category
49	YCSknnnnnnnnnKCGG	5′	5.4	Exp: cluster 46	6	Known: Put3
50	CGGnnnnnnnnnnnnnKCGV	5′	5.4	–	–	New: no category
51	WGTGACg	5′	5.3	ChIP: Sum1	14	New: Sum1
52	RTCCCTV	5′	5.3	–	–	New: no category
53	YTCGTTTAGG	5′	5.2	GO: lipid metabolism	5	New: lipid metabolism
54	TYCGKRM	5′	5.2	GO: filamentation	7	New: filamentation
55	CGCnnnnnnnnnnnnnnBCGB	5′	5.1	–	–	New: no category
56	TWCCCCM	5′	5.0	Exp.: cluster 46	7	Known: Mig1 + facilitators
57	CGGCnnMGnnnnnnnCGC	5′	5.0	ChIP: Gal4	7	Known: Gal4
58	CCGSnnnnnGVC	5′	5.0	–	–	New: no category
59	TRTAMATAKWT	3′	4.8	ChIP: Dig1	7	New: Ste12 (hits tRNA)
60	TtTATAnTATATAnA	3′	4.8	Exp: cluster 74 downstream	6	New: downstream cluster 74
61	GKBAGGGT	5′	4.8	GO: glycolysis	6	Known: Tbf1/new: glycolysis
62	GCnnTTAnCGG	5′	4.8	–	–	Known: Hap1
63	GGCSnnnnnGnnnCGCG	5′	4.7	ChIP: Mbp1	6	Known: Mbp1-like
64	TTCTCnnnnnnnCGC	5′	4.7	GO: filamentation	6	New: filamentation
65	SCGKnnnnKCGD	5′	4.5	–	–	New: no category
66	AATATTCTT	3′	4.4	Exp.: cluster 46 downstream	5	New: downstream facilitators
67	CGCGTnnnnnnnnnACG	5′	4.4	ChIP: Swi4	8	New: Swi4-vary gap
68	CCGHVGGM	5′	4.3	–	–	New: no category
69	CGCG = TTTT	5′	4.3	–	–	New: no category
70	CGCGnnnnnGGGS	5′	4.2	Exp.: cluster 46	6	New: expression cluster 46
71	CTGCAGGGR	5′	4.2	GO: filamentation	6	New: filamentation

Table 3.6. Discovered motifs and associated function.

motif. Based on the pairwise motif similarity matrix, we clustered the extended motifs hierarchically, collapsing two groups if the average similarity between their member motifs was at least 70%. We then computed a consensus sequence for every cluster of extended motifs, resulting into a smaller number of mega-motifs for each test (332 for CC1, 269 for CC2 and 285 for CC3). Row 3 shows the first 9 members of the top cluster in CC1, and the resulting mega-motif. Finally, we merged mega-motifs based on their co-occurrence in the same intergenic regions (Row 4). We computed a hypergeometric co-occurrence score between the intergenic regions hit by each mega-motif and again collapsed these hierarchically. We computed a consensus for every cluster, and iterated the co-occurrence-based collapsing step (results not shown). We obtained fewer than 200 distinct genome-wide motifs. Each full motif is assessed for genome-wide conservation by calculating its MCS, and those motifs with MCS ≥ 4 are retained. Each full motif was also tested for enrichment in upstream vs. downstream regions, by comparing its conservation rate in divergent vs. convergent intergenic regions.

3.7. Results and comparison to known motifs

The vast majority of the 45,760 possible mini-motifs show no distinctive conservation pattern. However, ~2400 mini-motifs show high scores by one or more of these criteria (Figure 3.5 panels A, B, C). There is substantial overlap among the mini-motifs produced by the three criteria, with about 50% of those found by one criterion also found by another.

The conserved mini-motifs give rise to a list of 72 full motifs having MCS ≥ 4 (Table 3.6). We omit full motifs with low MCS scores, and those that overlap tRNA genes and may be due to secondary RNA structure. Most of the motifs show preferential enrichment upstream of genes, but six are enriched downstream of genes. These 72 discovered motifs, found with no prior biological knowledge, show strong overlap with 28 of the 33 known motifs having MCS ≥ 4. They include 27 strong matches and 1 weaker match. The 72 discovered motifs also contain matches to 8 of the 22 known motifs with MCS < 4. In these cases, the comparative analysis identified closely related motifs that have higher conservation scores than the known motifs and occur largely at the same genes; these may represent a better description of the true regulatory element. Comparative genomic analysis thus automatically discovered 36 motifs with matches to

most of the known motifs (65% of the full set, 85% of those with high conservation). It also identified 42 additional 'novel' motifs not found in our list of known motifs. In the next chapter, we develop methods to understand these novel motifs and assign a candidate function to each of them.

3.8. Conclusion

Motif discovery amounts to extracting small sequence signals hidden within largely non-functional intergenic sequences. This problem is difficult in a single genome where the signal-to-noise ratio is very small. Previous methods have thus been limited to discovering motifs within small sets of genomic regions. We have conducted a genome-wide exhaustive search for all regulatory motifs. We produced a list of 72 strongly conserved motifs, that includes most previously identified motifs. This ability to directly discover regulatory motifs drastically changes our view of gene regulation. Instead of a case-by-case study, we can now observe complete views of all regulatory building blocks. Our method has re-discovered most previously known regulatory motifs without use of any prior biological function. It should theoretically be applicable to any genome for which no experimental data is available. Additionally, in yeast, we can use the biological information to discover the function of the discovered motifs. We can also use biological function to discover additional motifs. These two goals will be the topic of the next chapter.

CHAPTER 4: REGULATORY MOTIF FUNCTION

4.1. Introduction

In response to environmental changes, a single transcription factor can induce the expression of all genes necessary to fulfill a particular function, such as galactose import and utilization. These genes are typically scattered throughout the genome and targeted by the presence in their upstream regions of a specific regulatory motif recognized by the factor. This regulatory motif will be *enriched* in the upstream regions of these genes, namely it will occur more frequently in these regions than expected by chance as compared to the rest of the genome.

This enrichment of regulatory motifs in functionally related sets of genes can be used in two ways. Given a gene set, an associated motif can be found by searching the upstream intergenic regions for short patterns occurring at an unusual frequency. Alternatively, given a novel motif whose function is unknown, an associated gene set can be found by testing a number of previously defined gene sets (*categories*) for enrichment.

In a single genome, motifs occur frequently by chance, and hence the enrichment observed is sometimes not sufficient to perform either of these two tasks with high sensitivity and specificity. With multiple aligned genomes at hand, most spurious motif instances can be eliminated and the enrichment should become more pronounced. We can use this increased power to assign a candidate function to the motifs discovered in the previous chapter and to discover additional motifs in a category-specific way.

In this chapter, we present methods to distinguish biologically meaningful motif instances under selective pressure from non-functional motif instances. We assign candidate functions to the genome-wide motifs discovered in the previous chapter and find that the majority of discovered motifs show a significant functional enrichment. We also present a new method to discover additional regulatory motifs associated with functional categories. For known factors, we find that our category-based discovery method has great sensitivity and specificity, finding concise binding sites even when previous methods fail. For all 354 categories tested, we find that only a small number of motifs are found and these are shared, reused across categories.

4.2. Constructing functionally-related gene sets.

In yeast, a number of genome-wide experiments have resulted in functional groupings of genes into *gene sets*. These represent possibly co-regulated groups, constructed from gene expression, transcription factor binding and protein function.

Microarray technology enables the simultaneous measurement of gene expression levels for all 6000 annotated yeast genes on a single array. Such arrays contain thousands of spots (one for every gene), each containing multiple single-stranded nucleotide probes complementary to the corresponding predicted yeast gene. When cell extract is washed on the array, the single-stranded mRNA transcripts present in the cell *hybridize* (bind) by complementarity to the appropriate spots in the array. The level of hybridization can be measured by first fluorescently labeling the mRNA transcripts and then measuring the level of fluorescence on each spot using a laser scanner. The higher the hybridization measured at a spot, the higher is the inferred level of mRNA expression for that gene. These genome-wide experiments have been repeated for hundreds of experimental conditions and expression profiles have been constructed for every gene, describing its expression levels in each condition. These profiles can then be clustered computationally[59], typically by their pairwise correlation coefficients, to obtain sets of transcriptionally coordinated sets of genes.

Another technology, *ChIP*, has recently been applied to the genome-wide level to observe the binding locations of a transcription factor across the genome[60,61]. This technology enables the specific targeting of a transcription factor of interest, in order to pull it out of a cell extract. Pulling a transcription factor also selects for the DNA fragments that it is bound to. A researcher can then hybridize these fragments against an array containing probes for promoter regions, and infer which regions are bound by the transcription factor. Current technologies target transcription factors by either constructing an antibody specific to the factor, or by appending to the transcription factor a tag to which an antibody already exists (antibodies are molecules used by our immune system to recognize specific proteins of invading agents like viruses or bacteria; hence the name of Chromatin Immuno-Precipitation abbreviated as ChIP, referring to the use of antibodies to cause the chromatin bound by a factor to precipitate with the factor when

this one is pulled). The DNA is fragmented before precipitation and only a few hundred bases surrounding the bound site are typically pulled.

Genes can also be grouped into *functional categories*, based on the experimentally determined function of the proteins they encode. The function of thousands of yeast genes has been experimentally determined (to various degrees of precision). The scientific papers that describe these functions have been manually curated by the Saccharomyces Genome Database (SGD) group, generating a vast repository of knowledge. This knowledge has been classified hierarchically into Gene Ontology (GO) information or MIPS[62], using a unified language that crosscuts species and organism boundaries. This hierarchy groups at each internal node genes of related function, from the most specific to the most general, in categories such as 'meiotic DNA double-strand break processing', 'cell cycle', or 'metabolism'. Genes of related function will sometimes be part of the same metabolic pathway, required simultaneously for the correct sequence of chemical modifications of a metabolite, and hence likely to be co-regulated. Similarly, proteins that are part of the same protein complex are likely to be co-regulated, since they are required simultaneously for the correct assembly of the protein complex. Experimental methods similar to ChIP can be used to detect protein complexes[63]: an antibody specific to one of the proteins in the complex is used to pull the entire complex out of cell extract; the complex pulled is then fractionated at specific residues and the charge/weight combination of the fragments obtained by Mass Spectroscopy are used to find the precise set of amino acids in the fragment and the corresponding proteins that can result in such amino acid subsets.

4.3. Assigning a function to the genome-wide motifs

We used the biological knowledge captured in these sets of functionally related genes to assign function to the 72 genome-wide motifs discovered in the previous chapter. Since motifs can be degenerate and sometimes conserved in only a subset of the species, we first developed methods to score conserved motif instances. We then evaluated the overlap between the set of intergenic regions with motif scores above a given cutoff, and each functionally-related set of genes. We found a strong overlap with functional sets for most of the genome-wide motifs, and discover novel motif functions.

We used a probabilistic representation to detect conserved motif instances. We interpret every genome-wide motif ***m*** of length ***L*** as a probabilistic model, generator of sequences of length ***L*** over the alphabet {A,C,G,T}. We then evaluated for every genome position, the probability that the sequence was generated by motif ***m***, and compared this to the probability that the sequence was generated at random, given the ratio of A,C,G,T in the genome. We evaluated each species in turn, to obtain a total number of bits in the alignment. Since gaps may exist in the alignment, we did not evaluate the motif match directly on the alignment. Instead, we evaluated the motif in the ungapped sequence of each species in turn, and translated the motif start coordinates based on the alignment. To avoid evaluating each of 12 million start positions in the yeast genome against the motif, we first hashed the four genomes for rapid lookup, and subsequently only search those intergenic regions that contain k-mers in the motif searched. To allow for degenerate matches, we also search for k-mers with one or two degeneracies from the query motif. We then used a simple threshold *t* and obtain the list of all intergenic regions containing conserved instances of the motif with score at least *t*. These instances are either upstream of downstream of each flanking gene, depending on its transcriptional orientation. We could thus generate an 'upstream' list of genes that contain these conserved instances in their upstream regions, and a corresponding 'downstream' list of genes. We compared the overlap between each upstream and downstream gene list against each set of functionally related genes.

We did not expect a perfect overlap where every gene in a category would contain the motif and every gene outside the category would not contain the motif. On one hand, we expected discrepancies due to experimental errors, incomplete annotations and artifacts of the clustering algorithms. But even with perfect data, discrepancies arise from molecular processes that cross-cut functional categories, transcription factor binding that is dependent on additional protein-protein interactions or chromatin structure, expression clusters that are controlled by multiple transcription factors. At the same time, much like spurious motif instances can occur in a single species when motifs are short and degenerate, even conserved motif instances can occur by chance, although less frequent. Similarly, functional motif instances may appear diverged due to alignment errors, or may have genuinely diverged across the species compared.

Thus, we evaluated the overlap between motif presence and functional information probabilistically. Assume that *m* genes contain the motif and *r* genes belong to a particular functional category. At random, if the motif is independent from the category, we expect the same proportion of genes to contain motif instances both inside and outside the category. The probability of observing a deviation from that ratio can be evaluated using the hypergeometric distribution, described in the appendix. If *k* genes are observed in the overlap between the two sets, and *n* genes are present in the yeast genome, we calculate a P-value that the enrichment is observed at random as the hypergeometric sum for all values of *k'* that are greater or equal to *k*. Since we were evaluating the overlap of each motif against a large number of candidate functional categories, we use a Bonferroni correction for multiple hypothesis testing.

We applied these ideas to the motifs we discovered in our genome-wide search. As a control, we used the Gal4 motif (Figure 4.1). Given the biological role of Gal4, we considered the set of genes annotated to be involved in carbohydrate metabolism (126 genes according to the Gene Ontology (GO)[64] classification) with the set of genes that

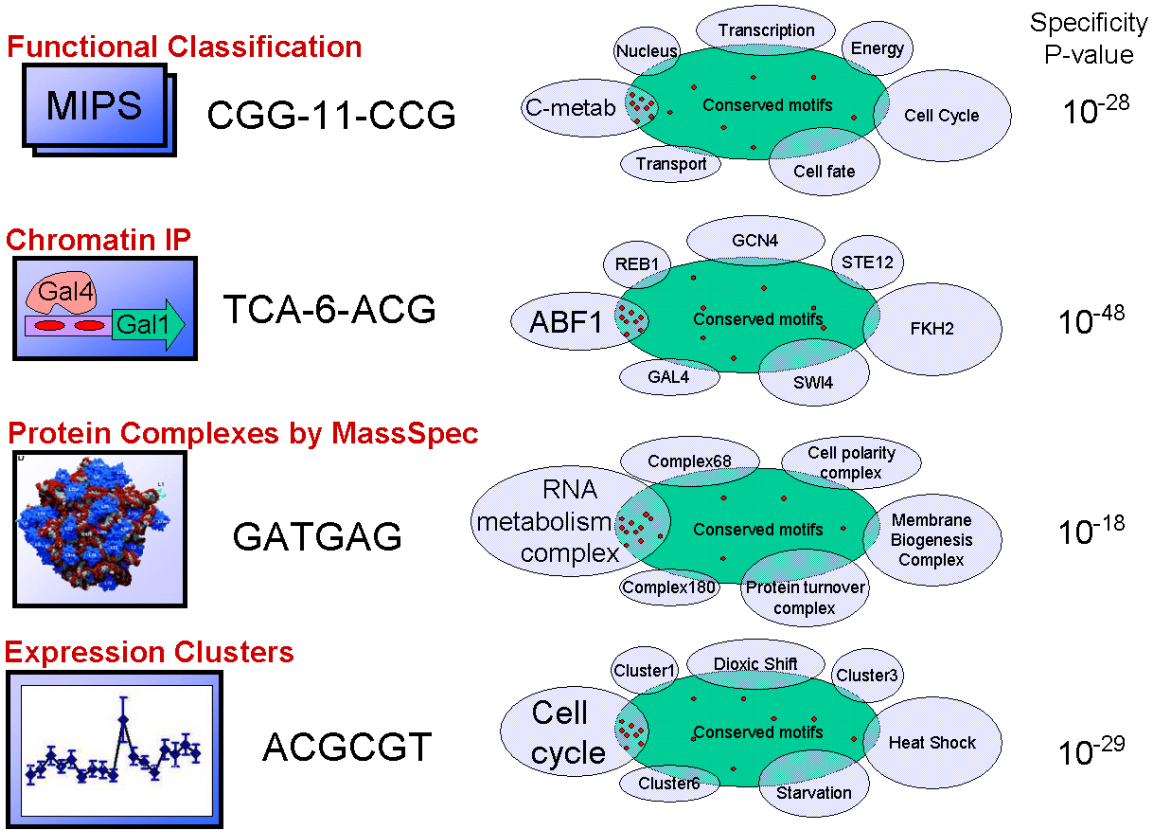


Figure 4.1. Assigning functions to genome-wide motifs based on functionally-related gene sets.

have a Gal4 binding motif upstream. The intergenic regions adjacent to carbohydrate metabolism genes comprise only 2% of all intergenic regions, but 7% of the occurrences of the Gal4 motif in *S. cerevisiae* (3.5-fold enrichment) and 29% of the conserved occurrences across the four species (15-fold enrichment). These results suggest that a function of the Gal4 motif could be inferred from the function of the genes adjacent to its conserved occurrences. Such putative functional assignments can be useful in directing experimentation for understanding the precise function of a motif.

Novel functions for genome-wide motifs

We compared each of the 72 motifs against a collection of 318 yeast gene categories based on functional and experimental data described earlier. These categories consist of 120 sets of genes defined with a common GO classification in SGD[64]; 106 sets of genes whose upstream region was identified as binding a given transcription factor in genome-wide chromatin immunoprecipitation (ChIP) experiments[61]; and 92 sets of genes showing coordinate regulation in RNA expression studies[59]. To measure how strongly the conserved occurrences correlated with the regions upstream (or downstream) of a particular gene category. We require a hypergeometric score of at least 10^{-5} to judge an overlap as significant, after accounting for testing of multiple categories. Most of the 36 discovered motifs that correspond to known motifs showed strong category correlation. Categories with the strongest correlation included those identified by ChIP with the transcription factor known to bind the motif, although many other relevant categories were identified. Of the 42 novel motifs, 25 show strong correlation with at least one category and thus can be assigned a suggestive biological function (Table 3.6).

Some motifs appear to define previously unknown binding sites associated with known transcription factors. Motif 32 is likely to be the binding site for Rgt1, which regulates genes involved in glucose transport[65]; the motif occurs upstream of many such genes, including appearing five times upstream of HXT1, which encodes a high-affinity glucose transporter. Motifs 21, 31 and 51 are all associated with genes whose upstream regions are bound by Sum1, a transcriptional repressor of genes involved in meiosis. The first motif has been previously reported (MSE)[66], but the latter two are novel and occur near genes whose products are involved in chromatin silencing and transcriptional repression.

Other motifs do not match regions bound by known transcription factors, but show strong correlation with functional categories. Motif 9 occurs upstream of genes involved in nitrogen metabolism, including amino acid and urea metabolism, nitrogen transport, glutamine metabolism and carbamoyl phosphate synthesis. Motif 25 is enriched among co-expressed genes (expression cluster 37) whose products function in vesicular traffic and secretion, including GDP/GTP exchange factors essential for the secretory machinery, clathrin assembly factors and many vesicle and plasma membrane proteins. Motifs 9, 13, 26, 34, 37 may play a role in filamentation. They are all enriched in genes co-regulated during environmental changes, involved in signaling and budding and bound by transcription factors involved in filamentation, such as Phd1.

Six motifs show higher conservation downstream of ORFs. Some of these may be in the 3' untranslated region of a transript and play a regulatory role in mRNA localization or stability. The strongest (Motif 6 and [67]) is found at genes whose product localizes to the cytosolic translational machinery, the mtDNA translational machinery or the mitochondrial outer membrane. Downstream motifs are also found enriched in a group of genes repressed during environmental stress (Motif 60 with expression cluster 37) and a group of genes involved in energy production (Motif 66 with expression cluster 46).

Two motifs (Motif 11a and Motif 69) show variable gap spacing, suggesting a new type of degeneracy within the recognition site for a transcription factor complex. Motif 11a corresponds closely to the known motif for Swi4 (Motif 11) but is interrupted by a central gap of 5, 7 or 9 bases; these variant motifs all show strong correlation with genes bound by Swi4 in ChIP experiments.

4.4. Discovering additional motifs based on gene sets

We next explored whether additional motifs could be found by searching specifically for conservation within individual gene categories. We selected mini-motifs based on their enrichment in specific categories and extended them to full motifs. We first evaluated our motif discovery method for ChIP experiments of factors with known motifs, and we found high sensitivity and specificity. We then searched for novel motifs in all 318 functional categories and discovered novel motifs.

The enrichment of regulatory motifs found in co-regulated gene sets has been the primary motivation for motif discovery algorithms such as MEME, AlignAce or BioProspector. These algorithms typically search for frequently occurring motifs within the set and subsequently evaluate the significance of the enrichment observed based on the overall frequency of the motifs throughout the genome. Thus, they search for motifs that are frequent within the set, and filter out those that are also frequent outside the set. We select for both criteria simultaneously by choosing mini-motifs based directly on their category enrichment score. We counted the conserved instances within the category (IN), and the conserved instances outside the category (OUT). We estimated the ratio p=IN/(IN+OUT) that we should expect for the category, based on the entire population of mini-motifs. We then calculated the significance of an observed enrichment as the binomial probability of observing IN successes out of IN+OUT trials given the probability of success p. We assigned a z-score to each mini-motif, as described in the genome-wide search. We extended those mini-motifs of z-score at least 5 sigma by searching for neighboring conserved bases that increase the specificity. We finally collapsed motifs of similar extension based on sequence similarity.

Factor	Known Motif	Hyper	MEME motif (Lee et al)		Category-based motif		Comparison
Abf1	RTCRYnnnnnACG	91.4	T**RTCAY**T-Y--**ACG**RA	good	**RTCACnnnnnACG**A	good	same
Gcn4	ATGACTCAT	47.8	**TGAGTCA**Y	good	**RTGACTCA**	good	same
Reb1	CCGGGTAA	44.7	**SCGGGTAA**Y	good	**CCGGGTAA**C	good	same
Mcm1	TTWCCnwwwrGGAAA	35.9	T**TTCC-AAW-RGGAA**A	good	**TCCnnnnnnGGA**	good	same
Rap1	ACACCCATACATTT	30.0	TTW**ACAYCCRTACAY-Y**	good	**ACCCCA.ACA**	good	same
Cbf1	RTCACRTG	24.2	TR**GTCACGTG**	good	**GTCACGTG**	good	same
Fkh2	TTGTTTACST	20.7	**TTGTTTAC**-TWTT	good	**TGTTTTAC..TT**	good	same
Swi4	CRCGAAAA	19.9	CSMRR**CGCGAAAA**	good	CAA**CRCGAAAA**	good	same
Mbp1	ACGCGT	19.6	G-RR-A-**ACGCGT-R**		A**ACGCGT**CG	good	better (+)
Ste12	RTGAAACA	17.8	GSAASRR-**TGA**TRAWGYA		**YTGAAACA**	good	better (+)
Gal4	CGGnnnnnnnnnnCCG	16.1	**CGG**M---CW-Y--**CCG**		**CGGnnnnnnnnnnnCCGA**	good	better (+)
Swi6	ACGCGT	15.6	**WCGCGT**CGCGTY-C	good	**ACGCGT**	good	same
Pho4	CACGTG	14.2	TTGTA**CACTT**YGTTT		CG**CACGTG**	good	better (+)
Hsf1	TTCTAGAA	14.1	TY**TTCYAGAA--TTC**Y	good	G**TTCTAGAA**nn**TTC**nnG	good	same
Dig1	RTGAAACA	13.6	CCY**TG-AY**TTCW-CTTC		**TGAAACR**	good	better (+)
Ino4	CATGTGAAat	13.4	G..G**CATGTGAAAA**	good	G…**CATGTGAA**	good	same
Fkh1	TTGTTTACST	13.2	C**YTRTTTAY**-WTT	good	**TGTTTAC**	good	same
Leu3	CCGGNNCCGG	13.1	G**CCGG**TMM**CGS**YC–	good	**CCGGnnnCGG**	good	same
Bas1	TGACTC	10.2	CS-CCAATGK--CS		**TGACTCT**A	good	better (+)
Swi5	KGCTGR	9.2	CACACACACACACACACA		T**GCTGG**	good	better (+)
Hap4	TnRTTGGT	8.5	YC**T-ATTSG**-C-GS		**TGATTGGT**	good	better (+)
Rlm1	CTAWWWWTAG	8.4	A-CTSGAAGAAATGCGGT		**CTA..TTTAG**	good	better (+)
Ino2	CATGTGAAat	7.4	G**CATGTGRAAA**	good	**CATGTG**	good	same
Met31	AAACTGTGGC	7.0	GCAC**GTG**ATS		**TGTGGC**	good	same
Ace2	GCTGGT	5.2	GTGTGTGTGTGTGTG		T**GCTGGT**	good	better (+)

Table 4.2. Category-based motif discovery shows increased power to discover concise motifs. Hyper shows the enrichment of the previously published motif in the ChIP experiment corresponding to the factor. For slightly enriched motifs, MEME fails to find the correct motif, but the conservation-based method succeeds. Concise and correct motifs are found in each case.

We first evaluated our ability to detect the 43 known motifs for which ChIP experiments[61] had been performed with the transcription factor that binds the motifs. For each category defined by the ChIP experiment, we undertook category-based motif discovery. Strong category-based motifs were found in 29 cases and these invariably corresponded closely to the known motifs (Table 4.2). These include 11 cases in which the motif had not been found by genome-wide motif discovery, suggesting that a category-based approach can be more sensitive in some cases. No strong category-based motifs were found for the remaining 14 known cases, including 7 cases in which genome-wide analysis yielded the known motif. Analysis of these 14 known motifs showed that none were, in fact, enriched in the ChIP-based category. This may reflect errors in the known motifs in some cases and imperfect ChIP data in others. Genome-wide analysis may simply be more powerful than category-based analysis in some instances. In all, 46 of the 55 known motifs were found by either genome-wide or category-based analysis. The remaining 9 cases may reflect true failures of the comparative genomic analysis or errors in the known motifs.

We compared our results to the motifs discovered by MEME in a single species as reported in Lee et al[61]. Our method showed stronger sensitivity in discovering all motifs for which the ChIP experiment indeed contained the correct motif. Additionally, the method showed strong specificity in the motifs discovered: the motifs were short and concise, and closely matched the published consensus. On the contrary, MEME failed to find the true motif in a number of cases, and when a motif was found it was generally obscured by a number of surrounding spurious bases that are not reported in the known motifs. Thus, we successfully used the additional information that comes from the multiple alignment to improve category-based motif discovery with very satisfactory results. By comparing multiple species, the signal becomes stronger. It allows the search to focus on the conserved bases, eliminating most of the noise. Table 1 summarizes the results. For each factor, we show the published motif, the hypergeometric enrichment score of the motif within the category (Hyper), the motif discovered by MEME and a quality assessment, the motif discovered by our method, as well as the corresponding category-based score and a quality assessment, and finally the comparison of our method to MEME. The performance of MEME degrades for less enriched motifs, but we consistently find the correct motif.

Table 4 **Additional new motifs discovered by category-based analysis**

No.	Category	Category-based motif	Interpretation	Score
1	Exp.: cluster 37	YCCCTTAAA	New: cluster 37 (Msn2/4-like)	[8.5]
2	ChIP: *FHL1* in YPD	ATGTACGGATG	New: Rap1 alternate	[7.6]
3	GO: carbohydrate transport	GTTTTTCCG	New: carbohydrate transport	[7.2]
4	GO: fatty acid beta-oxidation	TTAnnnCCG	New: fatty acid oxidation	[6.3]
5	GO: glycolysis/glyconeogenesis	TAGTGGAAGC	New: glycolysis/glycogenesis	[6.0]
6	Exp.: cluster 37	TCAGCC	New: cluster 37	[5.9]
7	Exp.: cluster 37	CGGnnnnnCGG	New: cluster 37	[5.7]
8	ChIP: *CIN5* in YPD	GnTTAnnTnAGC	New: Cin5 alternate	[5.6]
9	ChIP: *STE12* in butanol	CATTCT	Known: Tec1	[5.4]

Table 4.3. Novel category-based motifs.

We then applied the approach to all 318 gene categories. A total of 181 well-conserved motifs were identified, with many of these being equivalent motifs arising from multiple categories. Merging such motifs resulted in 52 distinct motifs, of which 43 were already found by the analyses described above. The remaining 9 motifs represent new category-based motifs (Table 4.3), including the following.

Three novel motifs are associated with genes that are bound by the transcription factors Rap1, Ste12 and Cin5, respectively. Rap1 is known to bind incomplete or degenerate instances of the published motif and the new motif may confer additional specificity. The motif associated with Ste12 is the known binding site for the partner transcription factor Tec1, suggesting that Ste12 binding is strongly associated with its partner under the conditions examined. Similarly, the novel motif associated with Cin5 may be that of a partner transcription factor. Three novel motifs are associated with the GO category for carbohydrate transport, fatty-acid oxidation and glycolysis-glycogenesis, respectively. Three novel motifs are associated with an expression cluster (cluster 37) that includes many genes involved in energy metabolism and stress response.

4.7. Conclusion

Category-based motif discovery contributes only a modest number of additional motifs beyond those found by genome-wide analysis. This confirms the relatively small number of regulatory motifs in yeast. A limited count is surprising given the large number of coordinately transcribed processes in yeast. The versatility of fine-grain yeast regulation may be rooted in a combinatorial control of gene expression, which will be the topic of the next chapter.

CHAPTER 5: COMBINATORIAL REGULATION

5.1. Introduction

We also used the comparisons to understand combinatorial interactions between regulatory motifs. A simple view of gene regulation where each environmental response is regulated by a dedicated transcription factor would require as many transcription factors and regulatory motifs as there are molecules and environmental changes. This is however not the case. It is estimated that only 160 transcription factors exist in the yeast genome, but yeast cells contain thousand of co-regulated sets of genes. This discrepancy requires a different model of gene regulation that goes beyond a one-to-one correspondence between regulatory motifs and cellular processes.

Our results from the previous chapter indeed point to a model where specific motif combinations are responsible for different cell responses. We saw that a single motif is typically involved in the control of many processes, and that a single process is typically enriched in multiple regulatory motifs. Furthermore, we saw that different processes were enriched in different combinations of regulatory motifs. Protein-protein interactions between the multiple factors bound upstream of every gene may dictate the specific combination of conditions under which the gene will be expressed. Understanding the combinations of regulatory motifs that are biologically meaningful, and the changing target gene sets may explain the versatility of eukaryotic gene regulation using only a small number of regulatory building blocks.

In this chapter, we develop methods to reveal the combinatorial control of gene expression. We construct a global motif interaction map, simply based on proximity of conserved motif pairs without requiring biological knowledge of gene function. We then present evidence for the changing functional specificities of the motif combinations discovered. Finally, we show the genome-wide effect of motif combinations on gene expression change.

5.2. Motifs are shared, reused across functional categories

We saw in the previous chapter that the motifs discovered across different categories largely overlapped. Each motif was discovered on average in three different

categories. This overlap is certainly to be expected between functionally related categories such as the chromatin IP experiment for Gcn4, the expression cluster of genes involved in amino acid biosynthesis, as well as the GO annotations for amino acid biosynthesis, all of which are enriched in the Gcn4 motif, the master regulator of amino acid metabolism.

More surprisingly however, different transcription factors are often enriched in the same motif (which may be due to cooperative binding), and the same motif appears enriched in multiple expression clusters and functional categories. For example, Cbf1, Met4, and Met31 share a motif, and so do Hsf1, Msn2 and Msn4; Fkh1 and Fkh2; Fhl1 and Rap1; Ste12 and Dig1; Swi5 and Ace2; Swi6, Swi4, Ash1 and Mbp1. Also, a single motif involved in environmental stress response is found repeatedly in numerous expression clusters, and in functional categories ranging from secretion, cell organization and biogenesis, transcription, ribosome biogenesis and rRNA processing.

Hence, the set of regulatory motifs that are specific to one functional category seems limited. This can hamper category-based motif discovery methods: no category will be enriched in a single motif, and no motif will be enriched in a single category. Additionally, there are a number of experimental limitations to a category-based approach. For example, the expression clusters we have used, although constructed over an impressive array of experiments, are still limited to the relatively few experimental conditions generated in the lab. Additionally, the functional categories we used are limited to the few well-characterized processes in yeast, and the molecular function of more than 3000 ORFs remains unknown.

A genome-wide approach presents a new and powerful paradigm to understanding the dictionary of regulatory motifs. By discovering in an unbiased way the complete set of conserved sequence elements, we now have the building blocks to subsequent analyses of regulation. To understand the full versatility of gene regulation, we now turn to understanding the combinatorial code of motif interactions. We first show that motif combinations can change the specificity of target genes, not in an additive, but in a combinatorial way. We then present methods to discover interacting motifs from the

genome-wide co-occurrence of their conserved instances, without making use of functional information. We then show that the interactions found are meaningful.

5.3. Changing specificity of motif combinations.

The effect of motif sharing a reuse can be additive or combinatorial. An additive effect simply adds the effect of the co-occurring transcription factors. For example, if each of two factors induces the expression of a gene, and both bind to a particular region, then their effect would be a doubly increased level of transcription for that gene. A combinatorial effect can be more complex. Namely, the combination of two factors may repress expression for a gene, even though either of the factors alone induces its expression.

Similarly, we should find that transcription factor combinations show different functional specificities than either of the transcription factors alone (Figure 5.1). We study here the gene category enrichment of two transcription factors that are known to bind to DNA cooperatively: Ste12 and Tec1. We considered three types of regions: those containing Tec1 motifs but no Ste12 motifs, those containing Ste12 motifs but no Tec1 motifs, and those containing both Ste12 and Tec1 motifs. We then intersected these

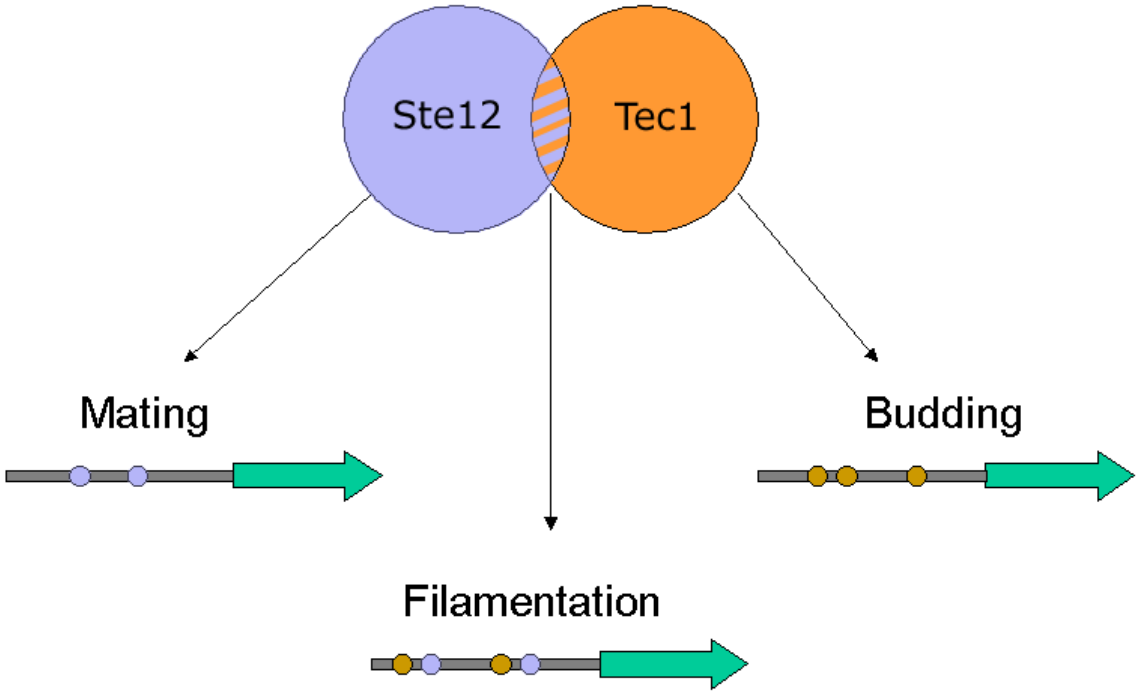


Figure 5.1. Changing specificity of motif combinations increases versatility of gene regulation.

three types of regions against the gene sets described previously.

We found that the regions that contain only the conserved Ste12 motif are enriched for genes involved in mating and pheromone response, while those that contain conserved occurrences of both the Ste12 and Tec1 motifs are enriched for genes involved in filamentous growth. These computational observations are consistent with recent elegant work showing genome-wide evidence that Ste12 and Tec1 indeed cooperate during starvation to induce filamentation-specific genes[68]. We also found that regions that contain only conserved occurrences of the Tec1 motif are enriched for genes involved in budding and cell polarity, suggesting that Tec1 has functions that do not require cooperative binding with Ste12.

5.4. Genome-wide motif co-occurrence map.

We next address the question of discovering these motif interactions in a genome-

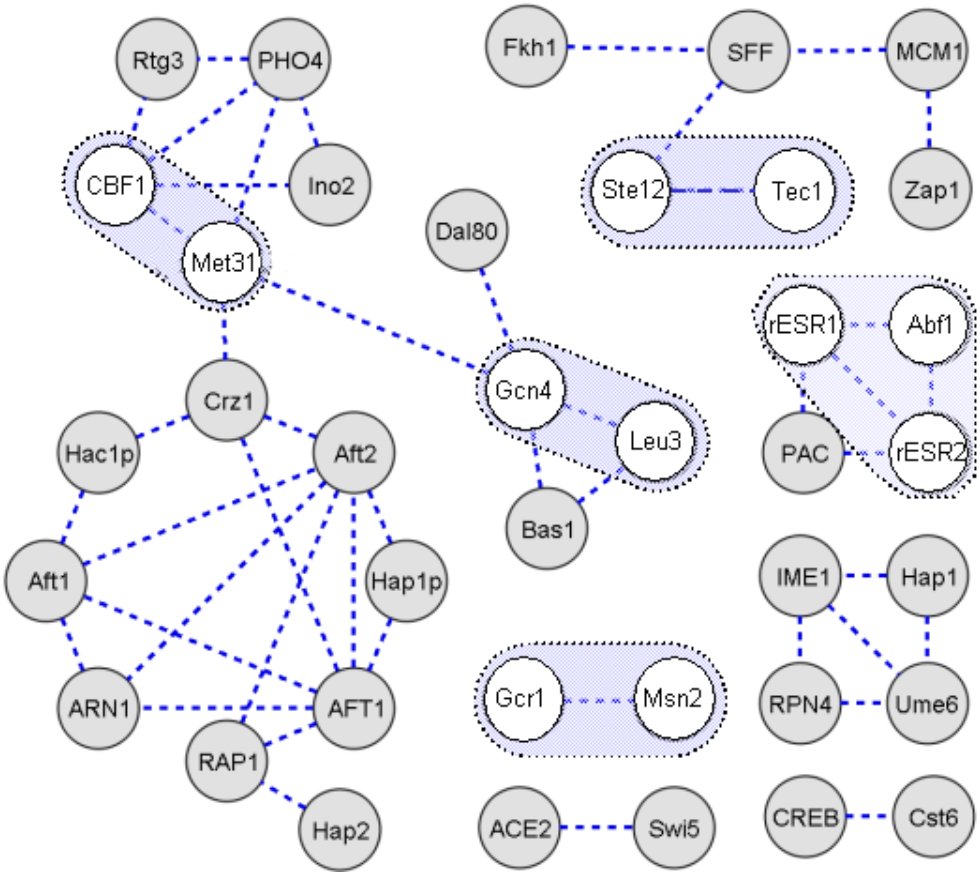


Figure 5.2. Genome-wide motif co-occurrence map reveals biologically meaningful motif relationships and transcription factor interactions

wide fashion. Protein-protein interactions between cooperatively binding transcription factors require that they bind in proximity upstream of their target genes. The regulatory motifs recognized by these factors should therefore co-occur in these intergenic regions of cooperative binding. The spatial orientation and physical distance between these motifs may vary across different genes, the varying distances being compensated by DNA bending that can bring the two sites in proximity. However, motif interactions do not typically cross gene boundaries, that are enforced by chromatin packaging and larger physical distances from one intergenic region to the next. Thus, co-occurrence of regulatory motifs in the same intergenic regions might be a good indicator of interacting transcription factors.

Using the comparison of the four species, we observed the genome-wide co-occurrence patterns of regulatory motifs (Figure 5.2). We searched for motifs that occur in the same intergenic regions more frequently than one would expect by chance. We computed the probability of seeing at least k regions in common when one motif is found in m regions and the other motif is found in r regions, given a total of n intergenic regions using the hypergeometric distribution.

Without using any functional information of gene categories, we found a number of significant motif interactions. These group motifs together into complex motif co-occurrence networks that may form the basis for studying combinatorial regulation of gene expression. These are not apparent in a single genome, where functional instances of the motif are overwhelmed by a much larger number of random occurrences. Cross species conservation greatly decrease this random noise and reveals biologically meaningful correlations.

5.5. Results.

We outlined here a number of biologically significant connections in the motif co-occurrence map. The combinatorial effect between Ste12 and Tec1 was indeed observed at the genome-wide level. The Ste12 and Tec1 motifs show clear correlation, with about 20% of regions having a conserved occurrence of one also having a conserved occurrence of the other. This enrichment is not apparent when considering *S. cerevisiae* alone.

The motif co-occurrence map reveals a number of biologically meaningful interactions. (a) About 60% of regions containing conserved motifs for the transcription factor Leu3 (which regulates branched-chain amino-acid biosynthesis) also contain conserved motifs for Gcn4 (a general factor regulating amino acid biosynthesis, as well as many other processes). (b) About 46% of regions containing conserved motifs for the transcription factor Met31 also contain conserved occurrences of Cbf1. In fact, Cbf1 (which is involves in DNA bending) is known to physically interact and cooperate with the MET regulatory complex. (c) About 34% of regions containing a conserved Gal4 motif also contain a conserved Mig1 motif. In this case, the correlation reflects antagonistic interaction. Gal4 induces galactose metabolism genes in presence of galactose, but Mig1 represses galactose metabolism in presence of glucose. (d) Pairwise co-occurrence connects a group of five motifs: Msn2/4 (general stress response), Rlm1 (response to cell-wall stresses), Pdr1 (pleiotropic drug resistance), Tea1 (Ty element activator) and Tbf1 (Telomere-binding factor). This suggests a possible link between various stress responses and adaptive changes at the genome level[69].

Many additional correlations are seen among known and novel motifs and can be pursued experimentally and computationally to construct comprehensive co-occurrence networks. These can provide information valuable in deciphering biological pathways in yeast.

5.6. Conclusion.

In this chapter, we provide methods to discover meaningful combinatorial interactions between regulatory motifs in a genome-wide way. Motif combinations can change the functional specificity of downstream motifs, and regulate a large number of processes using only a small number of regulatory motifs. This combinatorial nature of yeast regulation allows for a robust and modular regulatory network to adapt to changing environmental conditions. It is possible that additional regulatory motifs are added to the network, modulated by the more stable master regulatory motifs. We can further pursue these ideas to understand the rewiring of regulatory networks across evolutionary time. This may be one of many subtle ways of rapid evolutionary change outlined in the next chapter.

CHAPTER 6: EVOLUTIONARY CHANGE

6.1. Introduction

In previous chapters, we used the stronger conservation of functional elements across related species for the direct identification of genes and regulatory motifs. However, the species compared are not identical. They live in different environments and are subject to different pressures for survival. In the short evolutionary time that separates them, they have undergone a number of evolutionary changes to adapt to their respective environments.

In comparative genomics, both similarities and differences of the species compared can reveal important biological principles. Focusing on the similarities gives us a view of a core cell whose functionality has remained unchanged since the common ancestor of the species. Focusing on the differences gives us a dynamic view of a changing genome, and the mechanisms evolved for rapid adaptation to changing environments.

In this chapter, we focus on the mechanisms of evolutionary change that have become apparent in our comparisons. We show that the ambiguities in gene correspondence found in chapter 1 are localized in rapidly evolving telomeric regions at the chromosome endpoints. We also show that non-telomeric changes in gene order are due to either the inversion of a chromosomal segment (containing fewer than 20 genes) or reciprocal exchanges of chromosomal arms. For both types of events, the sequences at the breakpoints suggest specific mechanisms of chromosomal change. We observed few differences in gene content between the species, suggesting that phenotypic differences may be due to more subtle effects like protein domain changes and changes in gene regulation. Finally, we observed rapidly and slowly evolving genes: at one end of the spectrum, we found evidence of positive selection for rapid change in membrane adhesion proteins, suggesting a small number of mechanisms of rapid change; we also found genes that were surprisingly strongly conserved suggesting new hypotheses for their function.

6.2. Protein family expansions localize at the telomeres.

In the previous chapters, we used unambiguous ORFs and intergenic regions to discover conserved coding and regulatory elements in the yeast genome. In this chapter, we use ORFs with ambiguous correspondence to determine regions of rapid change.

We marked the chromosomal location of all *S. cerevisiae* ORFs that are ambiguous in at least one species. We then constructed ambiguity clusters when two or more ambiguous ORFs within 16kb of each other. We counted the number of ambiguities in each cluster, counting more than one ambiguities for an ORF whose correspondence was ambiguous in more than one species. Only 32 clusters were found containing more than two ambiguities. We ignored two clusters due to regions of low coverage in *S. mikatae* and one cluster corresponding to a previously described inversion.

Most of the ambiguities are strikingly clustered in telomeric regions (Figure 6.1). More than 80% fall into one of 32 clusters of two or more genes (average size ~18 kb, together comprising ~4% of the genome), which correspond nearly perfectly to the 32 telomeric regions of the 16 chromosomes of S. cerevisiae. Only one telomeric region lacks a cluster and only one cluster does not lie in telomeric regions in S. cerevisiae: it is a recent insertion of a segment that is telomeric in the other three species. The rapid structural evolution in the telomeric regions can also be observed at the gene level. The gene families contained within these regions (including the HXT, FLO, PAU, COS, THI, YRF families) show significant changes in number, order, and orientation. The regions also harbor many novel sequences, including protein-coding sequences. Finally, the telomeric regions have undergone 11 reciprocal translocations across the species.

Together, these features define relatively clear boundaries for the telomeric regions on all 32 chromosome arms, with sizes ranging from ~7 kb to ~52 kb on chromosome I-R. The extraordinary genomic churning occurring in these regions - and the telomeric localization of environment adaptation protein families - together probably play a key role in rapidly creating phenotypic diversity over evolutionary time. A high degree of variation in telomeric gene families has also been reported in P. falsiparum[69], the parasite responsible for malaria, and is related to antigenic variation.

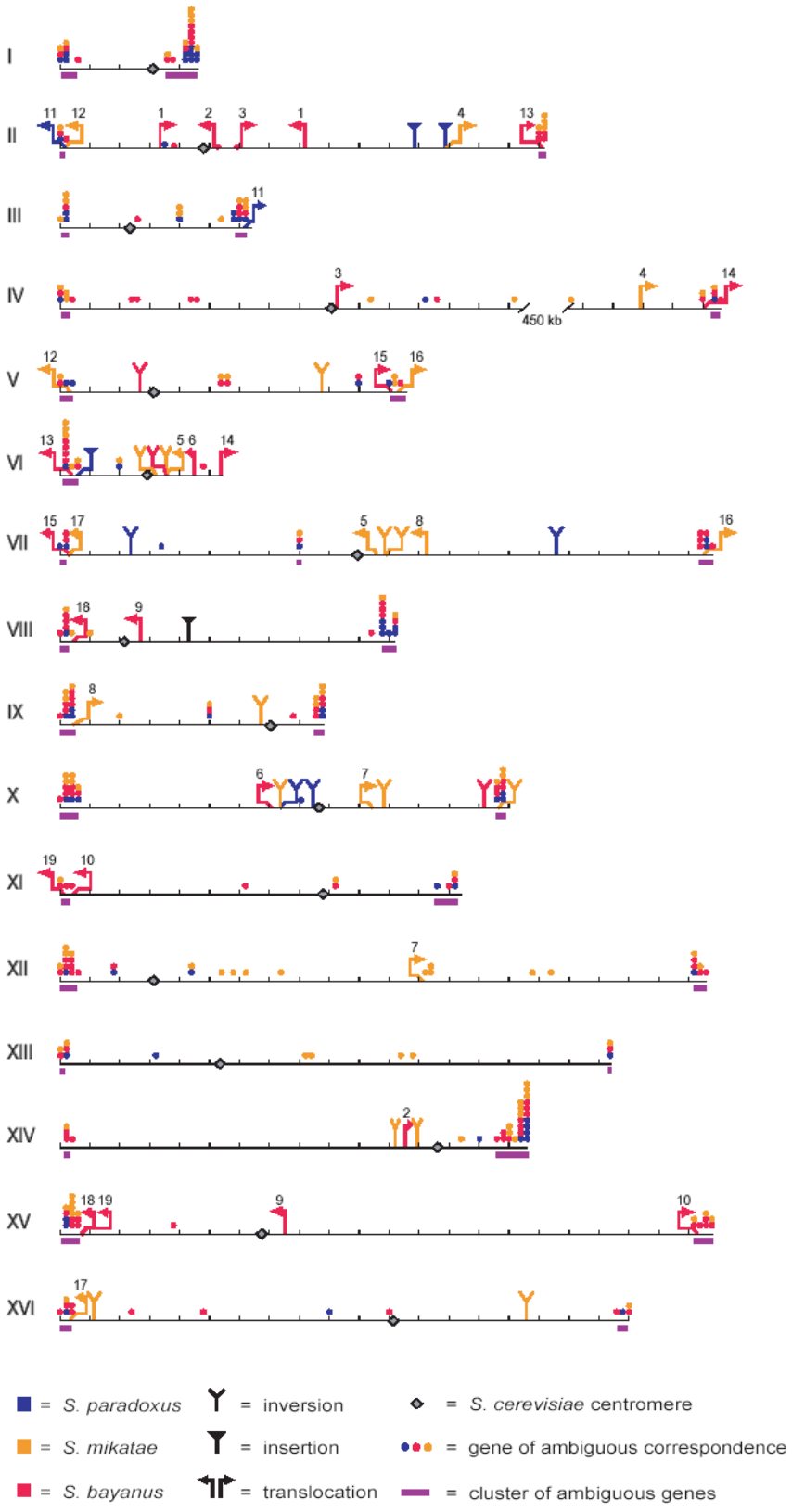


Figure 6.1. Rapid evolution in telomeres. Telomeric protein family expansions can rapidly create phenotypic diversity, potentially an evolutionary advantage.

6.3. Chromosomal rearrangements mediated by specific sequences.

Outside of the telomeric regions, few genomic rearrangements are found relative to *S. cerevisiae* (Figure 6.2). To discover these, we considered consecutive unambiguous matches, marking all changes in gene spacing, gene orientation, and off-synteny matches between scaffolds and orthologous *S. cerevisiae* chromosomes. We found that changes in gene spacing are typically associated with transposon insertions and associated novel genes, as well as tandem duplications. Virtually all changes in gene orientation typically affect between 2 and 10 consecutive ORFs and can be traced to one of 16 multi-gene inversions. The majority of off-synteny matches involve a single ORF and only 20 involve more than 2 consecutive ORFs. Virtually all single-gene off-synteny matches were contained within ancient duplication blocks of *Saccharomyces* as described in [70] and http://acer.gen.tcd.ie/~khwolfe/yeast/nova/. These probably represent previously duplicated genes that were differentially lost in different species, rather than a DNA

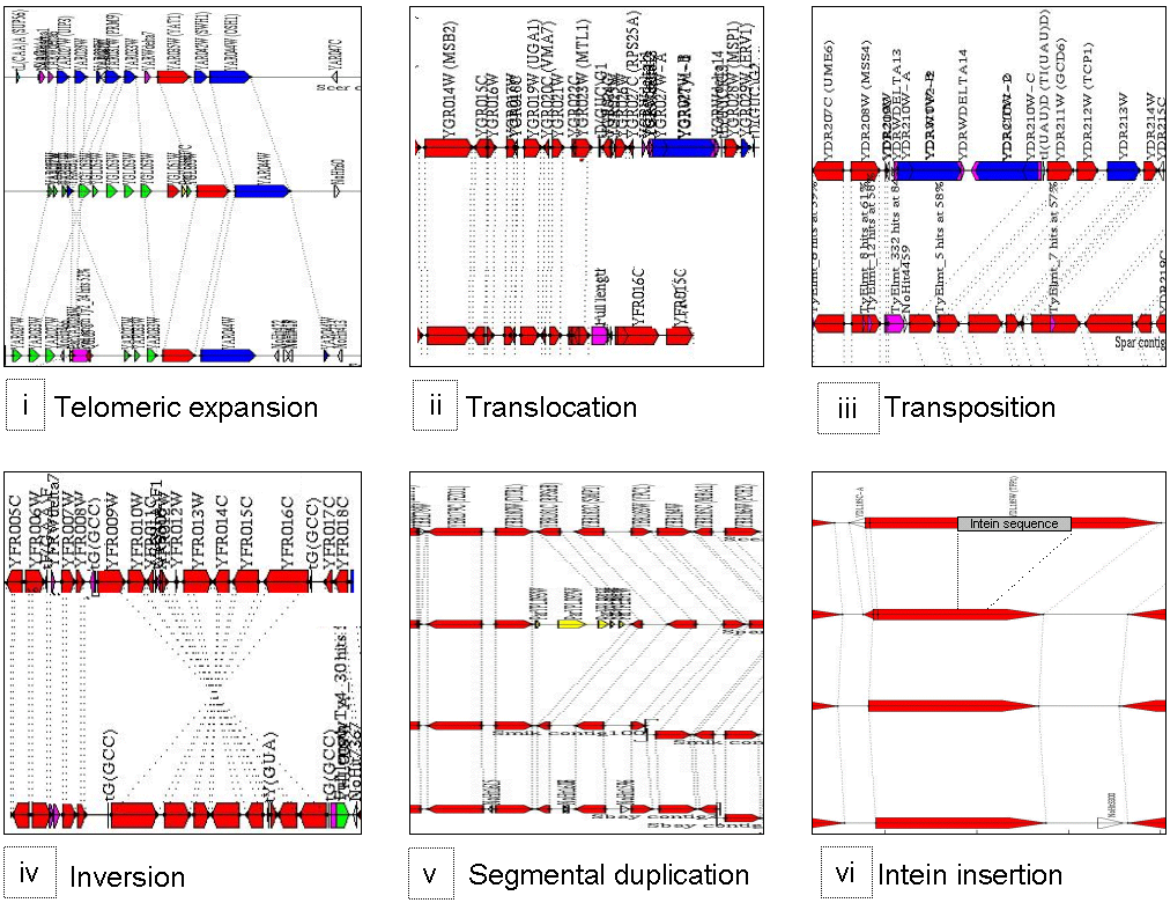


Figure 6.2. The six types of genome rearrangements that separate the species.

break in one of the two lineages, as was previously noted in [71]. Off-synteny matches that involve more than two genes from the same chromosome correspond to one of 20 chromosomal exchanges.

S. paradoxus shows no reciprocal translocations, 4 inversions and 3 segmental duplications. S. mikatae shows 4 reciprocal translocations and 13 inversions. S. bayanus has 5 reciprocal translocations and 3 inversions. The results confirmed four recently reported reciprocal translocations in these species, identified by pulsed-field gel electrophoresis[72], and identified four additional reciprocal translocations that had been missed. The sequence at the chromosomal breakpoints suggested the possible mechanism that underlie the rearrangements. Strikingly, the 20 inversions are all flanked by tRNA genes in opposite transcriptional orientation and usually of the same isoacceptor type; the origins of inversions in recombination between tRNA genes has not previously been noted. The reciprocal translocations occurred between Ty elements in seven cases and between highly similar pairs of ribosomal protein genes in two cases; the implication of Ty elements in reciprocal translocation is consistent with previous reports[44,71-73]. One segmental duplication involves 'donor' and 'recipient' regions that are descendants of an ancient duplication in the yeast genome[70]. Differential gene loss of anciently duplicated genes has been previously reported[74], but this is the first observation of a recent re-duplication event within anciently duplicated regions.

6.4. Small number of novel genes separate the species

We found a very small number of genes unique to one species and absent in the others. We noted above that *S. cerevisiae* contains 18 genes for which we could not identify orthologs in any of the other species, of which 7 encode ≥ 200 aa. These may be species-specific genes in *S. cerevisiae*, but alternatively could simply reflect gaps in the available draft genome sequences.

This uncertainty does not arise, however, in the reverse direction in identifying genes in the related species that lack an ortholog in *S. cerevisiae*. We found a total of 35 such ORFs encoding ≥ 200 aa (with the minimum length chosen to ensure that these are likely to represent valid genes). The list includes 5 genes unique to *S. paradoxus*, 8 genes unique to *S. mikatae* (two of which are 99% identical) and 19 genes unique to *S. bayanus*

(three of which form a gene family with ≥ 90% pairwise identity). There is also one gene represented by orthologous ORFs found in the latter two species only and one represented by orthologous ORFS in all three related species.

These species-specific ORFs are notable with respect to both function and location. The majority (63%) can also be assigned biological function on the basis of strong protein-sequence similarity with genes in other organisms. Most involve sugar metabolism and gene regulation (including one encoding a silencer protein). The majority (69%) are found in telomeric regions and an additional set (17%) are immediately adjacent to Ty elements; these locations are consistent with rapid genome evolution.

A curious coincidence was noted in the region between *YFL014W* and *YFL016W* in *S. cerevisiae*. In the orthologous regions in all four species, we find a species-specific ORF in every case (165, 111, 136 and 228 aa), but these four ORFs show little similarity at the protein level. The amino acid sequence has been disrupted by frame-shifting indels, but a long ORF has been maintained in each case. The explanation for this phenomenon is unclear, but may prove interesting.

6.5. Slow evolution suggests novel gene function.

With sequence alignments at millions of positions across the four species, it is possible to obtain a precise estimate of the rate of evolutionary change in the tree connecting the species.

One notable observation is the difference in substitution rate between *S. cerevisiae* and *S. paradoxus* (Figure 6.3). Using *S. bayanus* as an outgroup, the substitution rate is about 67% lower in the lineage leading to *S. paradoxus*. This observation is consistent regardless of the measure of evolutionary change: mutations,

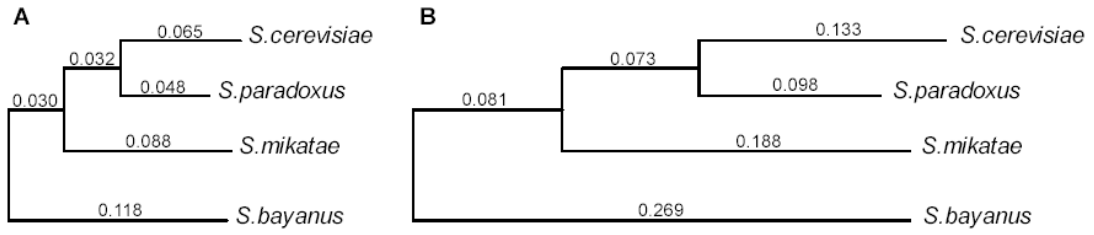


Figure 6.3. Slower mutation rate of *S. paradoxus* observed in genes and in intergenic regionsl

insertions, deletions measured across intergenic regions, genes or degenerate nucleotides in coding sequence all point to the same discrepancy. Hence, we can conclude that *S. paradoxus* is evolving at a slower rate than *S. cerevisiae* or *S. mikatae*. This could be due to generation time, but also life cycle throughout the year. Wild-type species remain dormant most of the year in spores, until the next blooming. This causes fewer cell divisions, hence fewer errors in replicating the DNA.

We can also observe differences in the rate of change of individual genes. One case stands out as an extreme outlier: the mating-type gene MATA2. The gene shows perfect 100% conservation at the amino acid level over its entire length (119 aa) across all four species. More strikingly, the gene shows perfect 100% conservation at the nucleotide level as well (357 bp). This differs sharply for the typical pattern seen for protein-coding genes, which show relaxed constraint in third positions of codons. Notably, the MATA2 gene is the only one of the four mating-type genes (the others being MATα1, MATα2 and MATA1) whose biochemical function remains unknown despite two decades of research[75]. An important clue may be that the sequence of MATA2 is identical in all four species to the 3'-end of the MATα2 gene. Perfect conservation at the nucleotide-level and identity to the terminus of MATα2 suggests that MATA2 may function not only by encoding a protein, but by encoding an anti-sense RNA or a DNA site. Hence, the lack of evolutionary change can suggest additional biological functions responsible for the pressure to conserve nucleotide sequence.

6.6. Evidence and mechanisms of rapid protein change.

Similarly, the unusually high rate of change can be biologically meaningful. The gene analysis described in chapter 2 rejected only a single ORF (*YBR184W*) that is clearly encoding a functional protein. The region containing *YBR184W* corresponds to a large open reading frame in all four species (524, 558, 554 and 556 amino acids, respectively), but the alignment shows unusually low sequence conservation. The sequence has only 32% nucleotide identity and 13% amino acid identity across the four species (Figure 6.4). Pairwise alignments across the species show numerous insertions and deletions, explaining why the gene failed the RFC test. (Interestingly, multiple alignment of all four species simultaneously improves the alignment sufficiently that the gene passes the RFC test; this suggests a way to improve the test.)

The rapid divergence is suggestive of a gene under strong positive selection. We tested this notion by calculating the Ka/Ks ratio (the normalized ratio of amino-acid-altering substitutions to silent substitutions), a traditional test for positive selection[76].

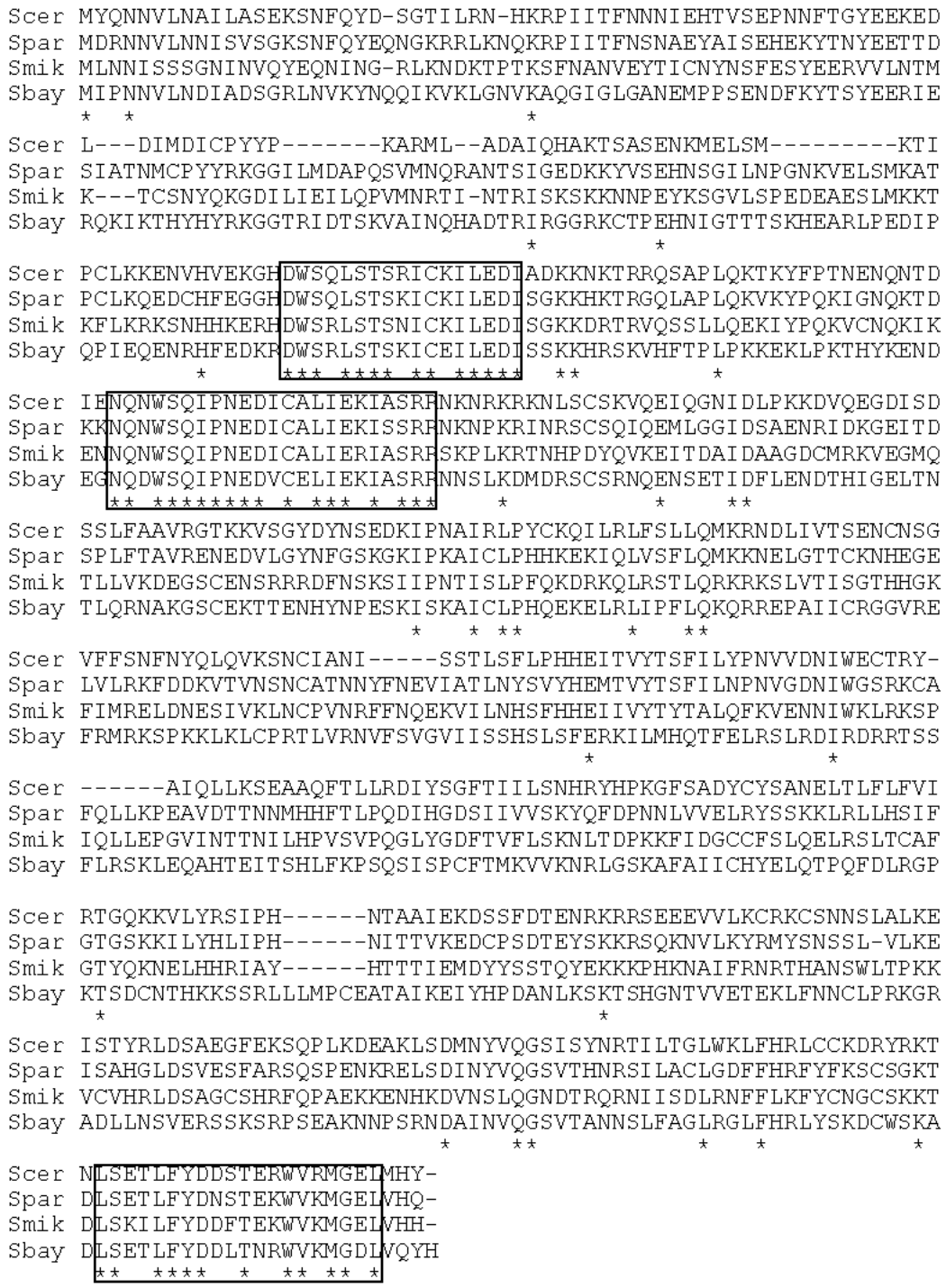


Figure 6.4. Multiple alignment of YBR184W shows only three conserved protein domains.

Whereas typical genes in *S. cerevisiae* show a Ka/Ks ratio of 0.11 ± 0.02, *YBR184W* has a ratio of 0.689. This ratio ranks as the third highest observed among all yeast genes (If three small domains with high conservation are excluded, the ratio rises to 0.774). The two genes with higher Ka/Ks ratio are *YAR068W*, a putative membrane protein, and *YER121W*, whose expression changes under stress.

The protein encoded by *YBR184W* has not been extensively studied, but expression studies show that the gene is induced during sporulation[77] and sequence analysis shows that it is similar to the gene *YSW1* that encodes a spore-specific protein. This is consistent with the observation that many of the best studied examples of positive selection in other organisms are genes related to gamete function. The change might promote speciation by imposing constraints on mating partner selection.

The vast majority of nucleotide changes in protein coding regions are silent or affect individual amino acids. However, a small number of events suggest additional mechanisms of rapid protein change. These events include closely spaced compensatory indels that affect the translation of small contiguous amino acid stretches. They also include the loss and gain of stop codons (by a nucleotide substitution or a frame-shifting indel) that may result in the rapid change of protein segments or the translation of previously non-coding regions[78]. Such events are observed more frequently near telomeric regions and may affect silenced genes or recently inactivated pseudogenes.

Additionally, we found a small number of differences in the length of orthologous proteins. These typically involve changes in the copy number of tri-nucleotide repeats, such as (CAA)n that encodes hydrophobic stretches often involved in protein-protein interactions. The most drastic example is seen for the TFP1 gene, which encodes a vacuolar ATPase. The *S. cerevisiae* gene contains an insertion of 1400 bp that is absent in the three related species. The insertion corresponds to the recent horizontal transfer of a known post-translationally self-splicing intein, VMA1[79].

6.7. Conclusion.

When comparing genomes, similarities and differences alike can reveal biological meaning. In comparing closely related species, the precise ways in which genomes change can reveal important biological insights. From the large-scale chromosomal

changes, to the substitutions of individual nucleotides, we find specific rules and constraints in the ways genomes evolve. Precise signals seem to govern how genomes are read, but also how they change. Evolutionary fitness may come from the combination of a fit genome that outperforms competition in the present, but also a modular genome that enables rapid evolution in times of extreme environmental pressure. The ability to rapidly carry out advantageous changes may be an inherent requirement in creating complexity via modularity. Evolutionary traits may be selected by reversible changes that allowed survival in the past, and will allow survival in the future. Each of the similarities and differences observed merits further experimental study. Understanding how genomes are written, and how they change, will be central to our understanding of the ever-changing book of life.

CONCLUSION

C.1. Summary

In this thesis, we explored the ability to extract a wide range of biological information from genome comparison among related organisms. Our results show that comparative analysis with closely related species can be invaluable in annotating a genome. It reveals the way different regions change and the constraints they face, providing clues as to their use. Even in a genome as compact as that of *S.cerevisiae*, where genes are easily detectable and rarely spliced, much remains to be learned about the gene content. We found that a large number of the annotated ORFs are dubious, adjusted the boundaries of hundreds of genes, and discovered more than 50 novel ORFs and 40 novel introns. Moreover, our comparisons have enabled a glimpse into the dynamic nature of gene regulation and co-regulated genes by discovering most known regulatory motifs as well as a number of novel motifs. The signals for these discoveries are present within the primary sequence of *S.cerevisiae*, but represent only a small fraction of the genome. Under the lens of evolutionary conservation, these signals stand out from the non-conserved noise. Hence, in studying any one genome, comparative analysis of closely related species can provide the basis for a global understanding of a wide range of functional elements.

Our results demonstrate the central role of computational tools in modern biology. The analyses presented in this thesis have revealed biological findings that can not be discovered by traditional genetic methods, regardless of the time or effort spent. Isolated deletion of every single yeast gene has been carried out without resolving the debate on the number of functional genes. Promoter analysis of any single gene could not reveal the subtle regulatory signals that become apparent at the genome-wide level. The approach presented is general, and has the advantage that one can increase its power by increasing the number of species studied. As sequencing costs lower and sequencing capacity increases, obtaining additional genomes becomes only a question of time. The comparison of multiple related species may present a new paradigm for understanding the genome of any single species. In particular, our methods are currently being applied to a kingdom-wide exploration of fungal genomes, and the comparative analysis of the human genome with that of the mouse and other mammals.

C.2. Extracting signal from noise.

For *S. cerevisiae*, our results show that comparative genome analysis of a handful of related species has substantial power to separate signal from noise to identify genes, define gene structure, highlight rapid and slow evolutionary change, recognize regulatory elements and reveal combinatorial control of gene regulation. The power is comparable or superior to experimental analysis, in terms of sensitivity and precision.

In principle, the approach could be applied to any organism by selecting a suitable set of related species. The optimal choice of species depends on multiple considerations, largely related to the evolutionary tree connecting the species. These include the following:

(1) The branch length t between species should be short enough to permit orthologous sequence to be readily aligned. The yeasts studied here differ by t = 0.23-0.55 substitutions per site and are readily aligned. The strong conservation of synteny (covering more than 90% of *S. cerevisiae* chromosomes belong in synteny blocks) allowed the unambiguous correspondence of the vast majority of genes.

(2) The total branch length of the tree should be large enough that non-functional sites will have undergone substantially more drift than functional sites, thereby providing an adequate degree of signal-to-noise enrichment (SNE). For this analysis, the multiple species studied provide a total branch length of 0.83 and a probability of nucleotide identity across all four species in non-coding regions of 49%. The SNE is thus ~2-fold (=1/0.49) for highly constrained nucleotides and correspondingly higher for composite features involving many nucleotides.

(3) The species should represent as narrow a group as possible, subject to the considerations above. Because the comparative analysis above seeks to identify genomic elements common to the species, it can explain only aspects of biology shared across the taxon. In the present case, the analysis identifies elements shared across *Saccharomyces sensu stricto*, a closely related set of species such that the vast majority of genes and regulatory elements are shared.

With these considerations in mind, the question remains as to what is the "right" number of species for comparative analysis. Similarly, one can ask, given a set of

previously sequenced species, what is the optimal choice for the next species to sequence. The answer of course depends on the goal at hand. In discovering genes, the number of species required depends on the length of the genes sought. In discovering motifs, the number of species depends on the motif length, its allowed degeneracy, and the total number of conserved instances. And in each case, the evolutionary distance of the species compared, but also the topology of the phylogenetic tree, will determine our ability to extract signal from noise. We found that genome-wide methods could increase the power of comparative analysis that is based on a handful of species. The answer in the general case merits a much more detailed analysis.

C.3. Analysis of mammalian genomes

What are the implications for the understanding of the human genome?

The present study provides a good model for evolutionary distances (substitutions per site in intergenic regions) relevant to the study of the human. The sequence divergence between *S. cerevisiae* and the most distant relative *S. bayanus* (11% indels and 62% nucleotide identity in aligned positions) is similar to that between human and mouse (12% indels and 66% nucleotide identity in aligned positions).

An important difference between yeast and human is the inherent signal-to-noise ratio (SNR) in the genome. Yeast has a high SNR, with protein-coding regions comprising ~70% of the genome coding for protein or RNA genes and regulatory elements comprising perhaps ~15% of the intergenic regions. The human has a much lower SNR, with the corresponding figures being perhaps ~2% and ~3%[19]. A lower SNR must be offset by a higher SNE. Some enrichment can also be obtained by filtering out the repeat sequences that comprise half of the human genome. Greater enrichment can be accomplished by increasing the number of species studied, taking advantage both of nucleotide level divergence and frequently occurring genomic deletion[19].

Such considerations indicate that it should be possible to use comparative analysis, such as explored here for yeast, to directly identify many functional elements in the human genome common to mammals. More generally, comparative analysis offers a powerful and precise initial tool for interpreting genomes.

C.4. The road ahead

In this thesis, we explored the ability of computational comparative genomics to extract biological signals that govern genes, regulation, and evolution. The nature of these signals however had been previously established experimentally. Knowing that genes were translated into amino acids every three nucleotides was central in our test of reading frame conservation. Knowing that regulatory motifs appear in multiple intergenic regions was crucial to our genome-wide discovery methods. Knowing the kinds of functional sequences to look for allowed us to examine the ways that they change. In each case, our methods relied on well-posed questions based on currently established biological knowledge.

In the future however, it will be important to formulate new hypotheses from genomic data. We cannot begin to imagine the types of information encoded in the human genome. The basis for intelligence, psychology, immunity, development, emotions are all encoded within our cells. New biological paradigms will be needed to explore novel aspects of biology, and their very discovery will reside in genome-wide studies. Development of new technologies, new statistical methods, new computational tools will be needed. An explosion of biological data, but also an explosion in novel experimental techniques has already started. And the only way to proceed is a constant marriage between biology and computer science.

REFERENCES

1. Kowalczuk, M., Mackiewicz, P., Gierlik, A., Dudek, M. R. & Cebrat, S. Total number of coding open reading frames in the yeast genome. *Yeast* **15**, 1031-4 (1999).
2. Harrison, P. M., Kumar, A., Lang, N., Snyder, M. & Gerstein, M. A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res* **30**, 1083-90 (2002).
3. Velculescu, V. E. et al. Characterization of the yeast transcriptome. *Cell* **88**, 243-51 (1997).
4. Blandin, G. et al. Genomic exploration of the hemiascomycetous yeasts: 4. The genome of Saccharomyces cerevisiae revisited. *FEBS Lett* **487**, 31-6 (2000).
5. Wood, V., Rutherford, K. M., Ivens, A., Rajandream, M.-A. & Barrell, B. A Re-annotation of the *Saccaromyces cerevisiae* Genome. *Comparative and Functional Genomics* **2**, 143-154 (2001).
6. Toda, T. et al. Deletion analysis of the enolase gene (enoA) promoter from the filamentous fungus Aspegillus oryzae. *Curr Genet* **40**, 260-7 (2001).
7. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**, 28-36 (1994).
8. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. Systematic determination of genetic network architecture. *Nat Genet* **22**, 281-5 (1999).
9. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16-23 (2000).
10. McGuire, A. M., Hughes, J. D. & Church, G. M. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res* **10**, 744-57 (2000).
11. Loots, G. G. et al. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136-40 (2000).
12. Pennacchio, L. A. & Rubin, E. M. Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* **2**, 100-9 (2001).
13. Oeltjen, J. C. et al. Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res* **7**, 315-29 (1997).
14. Cliften, P. F. et al. Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res* **11**, 1175-86 (2001).
15. Alm, R. A. et al. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. *Nature* **397**, 176-80 (1999).

16. Carlton, J. M. et al. Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. *Nature* **419**, 512-9 (2002).

17. Perrin, A. et al. Comparative genomics identifies the genetic islands that distinguish Neisseria meningitidis, the agent of cerebrospinal meningitis, from other Neisseria species. *Infect Immun* **70**, 7063-72 (2002).

18. McClelland, M. et al. Comparison of the Escherichia coli K-12 genome with sampled genomes of a Klebsiella pneumoniae and three salmonella enterica serovars, Typhimurium, Typhi and Paratyphi. *Nucleic Acids Res* **28**, 4974-86 (2000).

19. Intl_Mouse_Genome_Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).

20. Galabru, J., Rey-Cuille, M. A. & Hovanessian, A. G. Nucleotide sequence of the HIV-2 EHO genome, a divergent HIV-2 isolate. *AIDS Res Hum Retroviruses* **11**, 873-4 (1995).

21. Read, T. D. et al. The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria. *Nature* **423**, 81-6 (2003).

22. Genome_Sciences_Centre. The complete genome of the SARS associated Coronavirus. *Unpublished* (2003).

23. Goffeau, A. et al. Life with 6000 genes. *Science* **274**, 546, 563-7 (1996).

24. Galagan, J. E. et al. The genome sequence of the filamentous fungus Neurospora crassa. *Nature* **422**, 859-68 (2003).

25. The_C._elegans_Sequencing_Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-8 (1998).

26. Adams, M. D. et al. The genome sequence of Drosophila melanogaster. *Science* **287**, 2185-95 (2000).

27. Intl_Human_Genome_Sequencing_Consortium. in *Nature* 860-921 (2001).

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

29. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241-54 (2003).

30. Fitch, W. M. Uses for evolutionary trees. *Philos Trans R Soc Lond B Biol Sci* **349**, 93-102 (1995).

31. Fitch, W. M. Distinguishing homologous from analogous proteins. *Syst Zool* **19**, 99-113 (1970).

32. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631-7 (1997).

33. Tatusov, R. L. et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29**, 22-8 (2001).

34. Keogh, R. S., Seoighe, C. & Wolfe, K. H. Evolution of gene order and chromosome number in Saccharomyces, Kluyveromyces and related fungi. *Yeast* **14**, 443-57 (1998).

35. Batzoglou, S. et al. ARACHNE: a whole-genome shotgun assembler. *Genome Res* **12**, 177-89 (2002).

36. Jaffe, D. B. et al. Whole-genome sequence assembly for Mammalian genomes: arachne 2. *Genome Res* **13**, 91-6 (2003).

37. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).

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

39. Dujon, B. et al. Complete DNA sequence of yeast chromosome XI. *Nature* **369**, 371-8 (1994).

40. Sharp, P. M. & Li, W. H. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**, 1281-95 (1987).

41. Clark, T. A., Sugnet, C. W. & Ares, M., Jr. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* **296**, 907-10 (2002).

42. Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B. & Lander, E. S. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res* **10**, 950-8 (2000).

43. Hampson, S., Kibler, D. & Baldi, P. Distribution patterns of over-represented k-mers in non-coding yeast DNA. *Bioinformatics* **18**, 513-28 (2002).

44. Blanchette, M. & Tompa, M. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res* **12**, 739-48 (2002).

45. McCue, L. et al. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res* **29**, 774-82 (2001).

46. Gelfand, M. S., Koonin, E. V. & Mironov, A. A. Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Res* **28**, 695-705 (2000).

47. Lawrence, C. E. et al. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208-14 (1993).

48. Grundy, W. N., Bailey, T. L., Elkan, C. P. & Baker, M. E. Meta-MEME: motif-based hidden Markov models of protein families. *Comput Appl Biosci* **13**, 397-406 (1997).

49. Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* **296**, 1205-14 (2000).

50. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* **16**, 939-45 (1998).

51. Liu, X., Brutlag, D. L. & Liu, J. S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, 127-38 (2001).

52. Jiao, K. et al. Phylogenetic footprinting reveals multiple regulatory elements involved in control of the meiotic recombination gene, REC102. *Yeast* **19**, 99-114 (2002).

53. Tompa, M. Identifying functional elements by comparative DNA sequence analysis. *Genome Res* **11**, 1143-4 (2001).

54. Blanchette, M., Schwikowski, B. & Tompa, M. Algorithms for phylogenetic footprinting. *J Comput Biol* **9**, 211-23 (2002).

55. Keegan, L., Gill, G. & Ptashne, M. Separation of DNA binding from the transcription-activating function of a eukaryotic regulatory protein. *Science* **231**, 699-704 (1986).

56. Zhu, J. & Zhang, M. Q. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics* **15**, 607-11 (1999).

57. Zhang, M. Q. Promoter analysis of co-regulated genes in the yeast genome. *Comput Chem* **23**, 233-50 (1999).

58. Mewes, H. W. et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **27**, 44-8 (1999).

59. Gasch, A. P. & Eisen, M. B. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol* **3**, RESEARCH0059 (2002).

60. Simon, I. et al. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106**, 697-708 (2001).

61. Lee, T. I. et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804 (2002).

62. Mewes, H. W. et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* **30**, 31-4 (2002).

63. Gavin, A. C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-7 (2002).

64. Dwight, S. S. et al. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res* **30**, 69-72 (2002).

65. Mosley, A. L., Lakshmanan, J., Aryal, B. K. & Ozcan, S. Glucose-mediated phosphorylation converts the transcription factor Rgt1 from a repressor to an activator. *J Biol Chem* (2003).

66. Lindgren, A. et al. The pachytene checkpoint in Saccharomyces cerevisiae requires the Sum1 transcriptional repressor. *Embo J* **19**, 6489-97 (2000).

67. Jacobs Anderson, J. S. & Parker, R. Computational identification of cis-acting elements affecting post-transcriptional control of gene expression in Saccharomyces cerevisiae. *Nucleic Acids Res* **28**, 1604-17 (2000).

68. Zeitlinger, J. et al. Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* **113**, 395-404 (2003).

69. Gardner, M. J. et al. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498-511 (2002).

70. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708-13 (1997).

71. Fischer, G., Neuveglise, C., Durrens, P., Gaillardin, C. & Dujon, B. Evolution of gene order in the genomes of two related yeast species. *Genome Res* **11**, 2009-19 (2001).

72. Fischer, G., James, S. A., Roberts, I. N., Oliver, S. G. & Louis, E. J. Chromosomal evolution in Saccharomyces. *Nature* **405**, 451-4 (2000).

73. Dunham, M. J. et al. Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **99**, 16144-9 (2002).

74. Bon, E. et al. Genomic exploration of the hemiascomycetous yeasts: 5. Saccharomyces bayanus var. uvarum. *FEBS Lett* **487**, 37-41 (2000).

75. Haber, J. E. Mating-type gene switching in Saccharomyces cerevisiae. *Annu Rev Genet* **32**, 561-99 (1998).

76. Hurst, L. D. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet* **18**, 486 (2002).

77. Chu, S. et al. The transcriptional program of sporulation in budding yeast. *Science* **282**, 699-705 (1998).

78. True, H. L. & Lindquist, S. L. A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature* **407**, 477-83 (2000).

79. Koufopanou, V., Goddard, M. R. & Burt, A. Adaptation for horizontal transfer in a homing endonuclease. *Mol Biol Evol* **19**, 239-46 (2002).

APPENDIX

Counting combinations: The number of ways to choose k items without replacement from a total of n is given by (n choose k) :

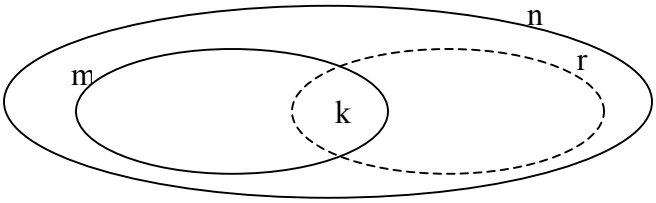
$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!} = \frac{n(n-1)...(n-k+1)}{k!}$$

Binomial distribution: The probability of obtaining k successes out of n trials given a probability p of success for any one trial is given by:

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Hypergeometric distribution: When choosing a random subset of size *r* from *n* items of which *m* belong in a particular category, the probability that *k* of the selected items belong to that category is given by:

$$p(X = k) = \binom{m}{k}\binom{n-m}{r-k} \Big/ \binom{n}{r}$$



Standard normal distribution (or Gaussian distribution)**:** The sum of a large number of independent variables follows a normal distribution of density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

Computing z-scores: Any probability p can be represented as the standard deviations away from the mean of a standard normal distribution corresponding to tail area p.