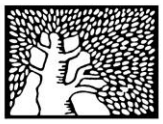


מכון ויצמן למדע

WEIZMANN INSTITUTE OF SCIENCE

Thesis for the degree
Doctor of Philosophy

חבור לשם קבלת התואר
דוקטור לפילוסופיה

By
Michal Lapidot

מאת
מיכל לפידות

אלמנטים רצפיים המשתתפים בבקרת ביטוי גנים: אתרי קישור של פקטורי שיעתוק ותעתיקי אנטיסנס

Sequence Elements Controlling Gene Expression: The Case of Transcription Regulatory Motifs and of Antisense Transcripts

Advisor
Dr. Yitzhak Pilpel

מנחה
ד"ר יצחק פלפל

July 2007

סיוון תשס"ז

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

מוגש למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

Acknowledgments

First and foremost, I wish to express my deepest gratitude to my supervisor Dr. Yitzhak Pilpel for his dedicated guidance, constant encouragement, continuous support and above all friendship. None of this work would have been possible without him. I feel extremely fortunate to have been the first student of the young 'Pilpel lab', watching it grow and thrive into an exceptionally pleasant and collaborative research environment. Much of this unique atmosphere is owed to Tzachi's pleasant personality and to his genuine enthusiasm and broad scientific vision.

I wish to thank all my lab colleagues throughout these years, for their friendship and encouragement, on top of invaluable advice and discussions. I am especially grateful to Ran Kafri and Shai Kaplan who helped form this lab from the very beginning and to Yael Garten, Shai Kaplan, Reut Shalgi, and Shai Shen-Orr whose research contributed directly to this work. I would also like to express my deep admiration to Ilya Venger, Ophir Shalem and Amir Mitchell for their valued friendship and for making this lab a pleasant place to walk into each morning.

I wish to thank my PhD Escorting Committee members Prof. Shmuel Pietrokovski and Prof. Michael Walker, for their time and effort to review and inspect this research, and to acknowledge the Horowitz Center for Complexity Science for supporting my work.

A special recognition is owed to all my collaborators: Prof. David Horn, Prof. Eytan Ruppin, Dr. Zach Solan and Liat Segal from Tel-Aviv University; Prof. Menachem Rubinstein and Roni Golan from the Weizmann Institute; and Prof. Judith Berman from the University of Minnesota, who helped me 'find my way' through the unfinished draft of the *Candida albicans* genome.

I would like to express my profound appreciation to the Weizmann Institute of Science, where I have spent (with a few exceptions) the last 10 years of my life, and which has been much more than a mere study and work place to me. I wish every student would have the privilege of studying in such a stimulating and enriching environment and enjoying the personal approach of the Feinberg Graduate School. I especially wish to thank Dr. Ami Shalit, Yosefa Givoli and the late Roni Golan, of the Graduate School administration, for attending to my every request and making me forget that I am one of many students.

Last but not least, deepest thanks to my family for encouraging me to follow my everlasting curiosity and to constantly seek knowledge, believing in my skills and fully supporting my every choice.

SUMMARY **1**

1 INTRODUCTION **3**

1.1 REGULATION OF GENE EXPRESSION THROUGH THE CONTROL OF TRANSCRIPTION 4
1.1.1 TF binding sites: representation and discovery *4*
1.1.2 An alternative approach for binding site prediction *5*
1.1.2.1 Rational 5
1.1.2.2 The motif analysis workbench 6
1.1.3 Motif dictionaries and their applications *6*
1.1.3.1 The challenge 6
1.1.3.2 Our motif dictionaries 7
1.1.3.3 Applications 8
1.1.4 Functional consequences of changes in transcriptional regulation *9*
1.1.4.1 Variations within individual TF binding sites 9
1.1.4.2 Variations on the promoter level 10
1.1.4.3 Gene regulatory networks and robustness towards mutations 11
1.2 NATURAL ANTISENSE TRANSCRIPTION – A POSSIBLE MECHANISM FOR THE CONTROL OF GENE EXPRESSION AT DIFFERENT LEVELS 12
1.2.1 Definition and extent of the phenomenon *12*
1.2.2 Evidence for a regulated process *14*
1.2.3 Principal mechanisms by which NATs regulate gene expression *16*

2 RESULTS **19**

2.1 REGULATORY MOTIF DICTIONARIES 19
2.1.1 Yeast (S. cerevisiae) motif dictionaries *20*
2.1.1.1 Dictionary construction methodology 20
2.1.1.2 Validation of the dictionary methodology 21
2.1.1.2.1 Assessing published motifs using our scoring method 21
2.1.1.2.2 Coverage of known motifs by the dictionary k-mers 23
2.1.1.2.3 Functional coherence (FC) analysis 23
2.1.1.3 Distinct characteristics of dictionary motifs 24
2.1.1.3.1 Motif length 25
2.1.1.3.2 Motif GC content 25
2.1.1.3.3 Motif GC contrast 26
2.1.1.3.4 Motif Entropy 27
2.1.1.3.5 Motif positional bias 28
2.1.1.3.6 Motif copy number 29
2.1.1.3.7 Motif evolutionary conservation 30
2.1.1.4 Motif Extraction Algorithm (MEX) - a syntax based approach 32
2.1.1.4.1 Evaluating the success of MEX in extracting biological significant motifs 33
2.1.1.4.2 Analysis of biologically significant motifs extracted by MEX 34
2.1.2 Candida albicans stress response dictionaries *37*
2.1.3 Human cell cycle dictionary *40*
2.2 FUNCTIONAL CHARACTERIZATION OF BINDING SITE VARIATIONS 41
2.2.1 Exploiting the yeast motif dictionaries to predict the outcome of a binding site substitution *42*
2.2.1.1 Quantitative measures for the severity of a substitution 42
2.2.1.2 The 'motif landscape analysis' tool 43
2.2.1.3 Differentiating between binding site switching and binding site loss 46
2.2.2 Deducing general properties of expression-altering substitutions *47*
2.2.3 The information content of the substituted position *49*
2.3 THE EVOLUTION OF INTERFERON-A PROMOTERS – AN ADAPTATION TO VARYING VIRAL THREATS? 51
2.3.1 IFN-α promoter scan using a selected set of motifs *53*
2.3.2 Comparisons of IFNA-α promoter motif content *53*
2.3.3 Measurement of IFN-α promoter activity in response to viral stimulus *55*
2.4 ANTISENSE TRANSCRIPTION – A REGULATED MECHANISM FOR THE CONTROL OF GENE EXPRESSION 57
2.4.1 In search for human trans encoded antisense *57*
2.4.1.1 Human *cis*-encoded NAT pairs can target transcripts in *trans* 58
2.4.1.2 Functional categorization of transcripts involved in sense-antisense pairing 61
2.4.2 Coupling NATs mechanisms of action to their regulation *65*

3 METHODS 70

3.1 REGULATORY MOTIF DICTIONARIES 70
3.1.1 Sequence and expression data 70
3.1.1.1 Yeast (*S. cerevisiae*) 70
3.1.1.2 Human 70
3.1.1.3 Candida Albicans 70
3.1.2 Dictionary Construction 71
3.1.2.1 Exhaustive genome scan 71
3.1.2.2 k-mer scoring 71
3.1.2.3 Clustering of dictionary motifs 72
3.1.3 MEX algorithm 73
3.1.4 Expression coherence (EC) score 75
3.1.5 Matching dictionary strings to PWMs. 76
3.1.6 Grouping Harbison's PWM set into distinct clusters 77
3.1.7 Functional Coherence (FC) score 77
3.1.8 Positional Bias p-value 78
3.1.9 Evolutionary conservation 79
3.1.9.1 Data 79
3.1.9.2 Motif conservation calculation 79
3.2 FUNCTIONAL CHARACTERIZATION OF BINDING SITE VARIATIONS 79
3.2.1 Motif positions used to gather statistics on substitution severity 79
3.2.2 The information content of a motif position 80
3.3 THE EVOLUTION OF INTERFERON-A PROMOTERS 81
3.3.1 Promoter Scan 81
3.3.2 Functional enrichment of GO terms 81
3.3.3 Binding site enrichment in IFN-α promoters. 82
3.4 ANTISENSE TRANSCRIPTION 82
3.4.1 Pipeline for whole-genome search of trans antisense hits 82
3.4.1.1 *cis*-NAT dataset and RefSeq mRNA sequences. 82
3.4.1.2 Data validation 82
3.4.1.3 Blast search 83
3.4.1.4 Separation of *cis* hits from *trans* hits 83
3.4.1.5 Annotations of transcripts participating in sense-antisense pairing 84

4 DISCUSSION 85

4.1 REGULATORY MOTIF DICTIONARIES 85
4.1.1 Method strengths 85
4.1.2 The complementary sequence-based approach 86
4.1.3 Method limitations 88
4.1.4 Further applications 89
4.2 FUNCTIONAL CHARACTERIZATION OF BINDING SITE VARIATIONS 89
4.3 REGULATION THROUGH ANTISENSE TRANSCRIPTS 91
4.3.1 Regulation of the regulator 91
4.3.2 A human trans-encoded NAT network 92

PUBLICATIONS RESULTING FROM THIS WORK 103

INDEPENDENT EFFORTS AND COLLABORATIONS 104

APPENDIX I – ABBREVIATIONS 105

APPENDIX II – 40 BIOLOGICAL CONDITIONS 106

APPENDIX III – SUPPLEMENTARY WEB FILES 107

Summary

Gene expression is tightly regulated at multiple stages from transcription to mRNA processing, mRNA transport and translation. This regulation is largely mediated through DNA and RNA sequence elements which are recognized by specific cellular partners, mostly proteins and other RNAs. However, unlike the amino acid translation code, much of the 'regulatory code' is yet to be revealed.

Here, we describe several interrelated projects, all of which integrate genome-wide sequence data with various types of functional information in order to identify regulatory sequence elements and to study the processes they mediate. We focus on two specific processes: the regulation of transcription through binding of transcription factors (TFs) to their designated sites, and regulation through natural antisense transcripts (NATs), which may act at several stages of gene expression. For both processes, we define gene sets which share distinct sequence elements and study their common characteristics in order to deduce the likely regulatory roles of these shared elements.

In the case of transcriptional regulation, we define our gene sets by the presence of a short DNA sequence motif in their promoters. By examining the expression profiles of these gene sets across a range of biological conditions, we identify promoter elements which are linked with coherently expressed gene sets. We hypothesize that such elements actively induce coherent expression, by serving as TF binding sites. We systematically apply this principle to short oligomers residing in the genome-wide promoters of yeast and human and construct catalogues of putative TF biding sites in these organisms. Each binding site is defined by its nucleotide sequence as well as by the expression profiles of the genes it appears to regulate. This motif discovery method overcomes the requirement for significant motif over-representation, posed by common *ab initio* motif finding algorithms. This is obtained by introducing a new statistical model which assesses the probability of obtaining the observed expression coherence for a random set of genes. This model enables the detection of motifs regulating small transcriptional networks, which may be present in the genome in relatively low numbers.

We provide several supports for the ability of our approach to identify functional TF binding sites, including its success at re-discovery of known yeast binding sites,

and the fact that the defined motifs possess many features which are characteristic of known binding sites. We additionally demonstrate the use of our motif collections for the study of evolutionary conservation of stress response among different yeasts.

Our dataset construction method quantifies the effect of a sequence motif on the expression profiles of its regulated genes. This allows us to address a central question in functional genomics: predicting the functional outcome of binding site variations, which exist in the population or between genomes of related species. We systematically compare the expression outcomes of motifs within our dataset, which differ at a single position, and use these comparisons to predict what would be the functional effect of mutating one binding site into another. In cases, in which such predictions can be compared to published experimental evidence, we find good agreement. We further accumulate statistics from multiple substitutions across numerous binding sites in an attempt to deduce general properties that characterize nucleotide substitutions which are likely to alter gene expression. Indeed we find that not all substitutions were 'born equal', and some are more likely to be 'deleterious'. This work serves as a first step towards a larger task - predicting the phenotypic effect of variations in regulatory motifs which exist in the human population.

For the study of antisense-mediated regulation in human, we group together transcripts which contain a sequence match to a complementary transcript, which was previously reported to reside in *cis* to its sense target. We find that many *cis*-acting anti-sense transcripts may have additional targets in *trans*, thus *cis*- and *trans* NAT networks are interlaced and are not distinct phenomena as commonly accepted. We reveal a putative genome-wide NAT network, which displays many to many relations: The same NAT may potentially target multiple targets (both in *cis* and *trans*) and a given mRNA may serve as a potential target of more than one *trans*-encoded NAT.

We find several particular biological functions to be enriched among the genes belonging to our antisense network, suggesting that they may indeed be subject to common regulation. Intriguingly, a similar set of functional categories was recently reported to be enriched in the *trans*-antisense network of *Arabidopsis thaliana*. This is a remarkable correspondence that may represent convergence to similar regulatory regimes of functionally related genes in organisms as distant as human and plant.

We anticipate that similar genome-wide analyses, integrating sequence data with functional information, will prove useful for the study of additional cellular regulatory processes.

1 Introduction

The current work is composed of several related projects, which while may be viewed independently, are also interrelated by their biological motivation as well as the computational frameworks used and the types of analyses applied.

The two major projects (described in results sections 2.1-2.2 and 2.4) provide a global view on the regulation of gene expression in yeast and human, with a focus on two specific processes: the regulation of transcription through binding of transcription factors (TFs) to their designated sites, and regulation through natural antisense transcripts (NATs), which may act at several stages of gene expression (see section 1.2). At the core of both projects is a genome-wide study of gene sets which are defined by a common sequence element; In the case of transcriptional regulation, these sets are defined by the presence of a short DNA sequence motif in their promoters. In the case of antisense-mediated regulation, the gene sets are defined by a sequence match to a complementary transcript, which is transcribed in *trans*, namely from a different genomic location. In both cases, we identify features, common to all genes belonging to the same set, which imply that these genes may be subject to common regulation. Such features include related biological functions, correlated expression profiles or localizations to similar cellular compartments.

Two of the projects which deal with the regulation of transcription (described in results sections 2.2 and 2.3) share an additional aspect; the attempt to characterize promoter sequence variations that affect gene expression and may thus alter gene function. The first project, conducted in yeast on a genome wide scale (results section 2.2), deals with single nucleotide variations within TF binding sites, and introduces computational means to assess their effect on gene expression. The second project, conducted in a specific gene family in human (results section 2.3), deals with variations on a larger scale, i.e. in the composition of binding sites within a promoter. Through characterizing changes in the composition of TF binding sites residing within the promoters of human interferon-α genes, this project studies how different members of the interferon family evolved to respond to different viral stimuli. This project combines a computational search for biologically significant regulatory motifs with accompanying expression experiments, carried out by our collaborators.

1.1 Regulation of gene expression through the control of transcription

1.1.1 TF binding sites: representation and discovery

The regulation of gene expression is mediated mainly through specific interactions of TF proteins with short promoter elements. TF binding sites are short (~6-20 bases) and imprecise; unlike restriction enzymes which recognize unique nucleotide sequences, a single TF protein may interact with a range of related sequences. For most TFs, there appears to be no distinct sequence of nucleotide bases that is shared by all recognized binding sites. However there are typically clear biases in the distribution of bases that occur at each binding site position. These biases are commonly represented by position weight matrices (PWMs), whose components give the probabilities of finding each nucleotide at each binding site position [1, 2].

Binding sites typically comprise a minority of the nucleotides within a promoter region. They are embedded within sequence that is assumed to be non-functional with respect to transcription (although surrounding sequences may contribute to protein-DNA binding through influencing local DNA conformation). Identifying genuine binding sites is a challenging task as the physical extent of a promoter is rarely well defined, and within this ill-defined region we are seeking sparsely distributed, short and degenerate sequence motifs. Several experimental and computational methods have been developed to meet this challenge. Experimental methods include: DNAseI protection (footprinting), i.e. the identification of DNA segments that are protected from nuclease digestion by protein binding [3], electrophoretic mobility shift essays (EMSA) [4, 5], *in vivo* binding assays [6-8] and modification of a putative binding site and assaying transcription *in vitro* or *in vivo*, by transformation with a reporter gene (e.g. luciferase or green fluorescent protein) [9].

Such experimental methods are generally technically challenging and time consuming and thus may be difficult to apply in large scale screening of potential binding sites at a variety of conditions. Advances in genome research, including whole genome sequencing and mRNA expression monitoring have allowed the development of computational methods for *ab initio* binding site prediction. A popular method searches for shared motifs in the promoters of co-expressed genes [10-12]. A complementary approach termed "phylogenetic footprinting" searches for conserved motifs in the promoters of orthologous genes [13-18]. The rational being that binding sites are more likely to be conserved by natural selection than their

putatively non-functional background nucleotides. The choice of species is crucial for obtaining reliable results; Comparing species with a short divergence time may result in many false positives, as conservation is likely to reflect common ancestry rather than purifying selection. Conversely, choosing too distant species, may fail to recover species-specific sites [13, 14]. For instance, about 40% of human functional binding sites are expected to be non functional in rodents [19], and a similar proportion of species specific sites was reported in yeast [20]. Furthermore, the alignment of orthologous intergenic sequences is nontrivial since well conserved sequences of different lengths are interspersed with sequences that show little conservation.

Common to many of the computational methods is the search for "over-represented motifs" i.e. motifs that are observed in the data at a frequency that is significantly higher than that expected by chance given an appropriate background statistical model. Although very successful in producing testable predictions, these methods are prone to both false-positive and false-negative motif predictions. False-positive predictions include motifs that are also present in the promoters of many other genes outside of the gene set from which they were derived (e.g. 'TATA' box). These motifs usually do not determine particular expression patterns. The false negatives are motifs that occur in sets of genes that are smaller than the size threshold required for their detection with sufficient statistical significance.

1.1.2 An alternative approach for binding site prediction

1.1.2.1 Rational

We have proposed an alternative motif discovery and analysis methodology to meet these two opposing challenges [21]. The methodology is based on the previously introduced measure of expression coherence (EC), which quantifies the extent to which a set of genes display similar expression profiles at a given set of biological conditions [22, 23]. While in previous analyses we mainly used a corollary of the EC score definition to detect functional interactions between known motifs [22, 23], in the present work, we utilized this measure for the discovery of novel motifs, as well as for the refinement of previously published ones. These individual motifs are the building blocks of subsequent combinatorial motif reconstructions.

We applied the EC score to sets of genes that contain a given motif in their promoters, and used it to assess the hypothesis that the motif drives the genes' coherent expression at an examined biological condition. This approach ensures that

only motifs that occur in tightly co-expressed genes will score highly. Additionally, we introduced a statistical model that computes the probabilities of obtaining the observed or higher EC score by chance [21]. This model relaxes the requirement that motifs will appear at a particularly high number of genes, and enables the detection of motifs that regulate small transcriptional networks.

1.1.2.2 The motif analysis workbench

We have made our method accessible to the experimental and computational biology communities, through the construction of "The Motif Analysis Workbench", http://longitude.weizmann.ac.il/services.html [21]. This is a WWW interface for the automated analyses of promoter regulatory motifs and the effect they exert on mRNA expression at different biological conditions. The server provides a wide spectrum of analysis tools that allow *de novo* motif discovery as well as motif refinement and in-depth investigation of fully or partially characterized motifs. The discovery and analysis tools are fundamentally different from existing tools in their basic, rational, statistical background and specificity and sensitivity towards true regulatory elements.

1.1.3 Motif dictionaries and their applications

1.1.3.1 The challenge

An important challenge in the study of gene regulation is to produce reliable reference collections ('dictionaries') of TF binding sites in different organisms. Such binding sites can be viewed as the atomic units of highly complex multi-component transcriptional regulatory networks. Their study may thus reveal properties of individual motifs as well as contribute to the understanding of higher levels of transcriptional control. For most organisms, the current lack of such reference data, not only hinders our understanding of transcription regulation, but also results in an over-reliance on the very few experimentally validated binding sites for the evaluation of novel motif finding algorithms. Such over reliance maintains a bias towards previously discovered motifs. There is a need for the introduction of methods, capable of detecting motifs on a genome-wide scale in an unbiased fashion. The most promise for this task lies in an interplay between computational and experimental approaches [24].

Perhaps the most studied TF repertoire is that of the budding yeast *Saccharomyces cerevisiae* (*S. cerevisiae*), in which several works have attempted to identify binding

sites on a genome-wide scale: Phylogenetic footprinting of 7 fully sequenced *Saccharomyces* species was carried out by two separate groups yielding two largely non overlapping sets of conserved regulatory elements [13, 14]. A different work introduced a model in which upstream motifs contribute additively to the log-expression level of a gene. The model was applied to publicly available expression data for *S. cerevisiae* and successfully identified known motifs as well as new putative ones [25]. The most elaborate and extensively cited work is that of Harbison et al. [8], which conducted *in vivo* binding essays for 203 yeast transcriptional regulators, under more than one growth condition. The technique they used is ChIP-chip, chromatin immunoprecipitation followed by hybridization of the precipitated DNA fragments to a microarray with known genomic promoter sequences. ChIP-chip typically detects DNA fragments of size 100–500 base pairs (bp). To identify the considerably shorter TF binding sites, the authors further applied several motif discovery methods (including AlignACE [26] and MEME [27]) to DNA fragments that were bound by the same regulator. Pulling together significant motifs from all programs and filtering them by evolutionary conservation, they were able to define binding sites for 102 TFs. Although not accounting for almost 50% of yeast TFs, this set was adopted by the community as a reliable reference.

1.1.3.2 Our motif dictionaries

The vast experimental based knowledge available in yeast renders this organism ideal for the assessment of new frameworks for *de novo* motif discovery and characterization. We developed a method for the construction of comprehensive motif dictionaries, which are unbiased by prior knowledge (results section 2.1), offering a solution for the over reliance on few validated motifs, described above. Our method forms a quantifiable connection between binding site sequence and the expression profiles of the regulated genes and may be applicable to the genomes of any organism for which both genome-wide promoter sequences and whole genome mRNA profiles are available.

Our dictionary construction method is in fact an expansion of our application of the EC score for measuring a motif's regulatory capacity (described above, section 1.1.2). The method is based on the premise that any nucleotide sequence that resides in the promoter of a gene may potentially contribute to the regulation of that gene's expression. We developed two complementary approaches and demonstrated the

success of both in producing reliable binding sites in the widely studied *S. cerevisiae* genome; The first approach exhaustively enumerates k-mers residing in genes' promoters and uses the EC score to assess their likely affects on gene expression in various biological conditions. The second approach applies the same scoring system to pre-selected candidate motifs (as opposed to all k-mers) (section 2.1.1.4). Pre-selection of motifs may be based on different criteria, we illustrate a well performing syntax based selection [28, 29], but other properties such as evolutionary conservation within related species may also be considered. Following a thorough method validation in budding yeast, we created motif dictionaries for additional organisms including *Candida albicans*, *Caenorhabditis elegans* (*C. elegans*) and human.

1.1.3.3 Applications

Our comprehensive binding site collections are an invaluable source for the study of different aspects of transcription regulation: We investigated the *S. cerevisiae* dictionary in order to define characteristics that likely contribute to the biological function of regulatory motifs (section 2.1.1.3). Such characteristics (e.g. distinct nucleotide composition, positional bias, motif multiplicity, evolutionary conservation) may subsequently be incorporated into more elaborate motif prediction algorithms. Applying comparative genomics to dictionaries of different species may elucidate features of transcriptional regulation that are common to different eukaryotes. We demonstrate the use of such comparative analyses in the study of stress response in evolutionary distinct yeast species (section 2.1.2).

An extremely important application of the motif dictionaries is their utilization for the study of phenotypic effects of binding site variations. In results section 2.2 we describe an elaborate analysis of the *S. cerevisiae* dictionary, conducted in order to characterize and potentially predict the effects of motif variations on gene expression and ultimately on the phenotype. Below (section 1.1.4) we discuss the theoretical background and motivation for this project.

Lastly, the motifs defined in the dictionary project may serve as input for studies aimed at revealing higher levels of gene regulation, some of which are currently conducted in the lab. Regulation of gene expression, in eukaryotes, often involves the coordinated action of multiple TFs [25, 30-32]. It was proposed that the yeast regulatory network utilizes a limited number of TFs and creates alternative diverse combinations between them in a condition-specific modulated fashion [22]. The

regulatory networks of multi-cellular organisms are expected to be significantly more complex than those of unicellular organisms, mainly since they govern differentiation of multiple tissues and cell types and because they control intricate developmental programs. Consequently, these networks are usually subject to the control of numerous regulatory factors, acting combinatorially according to complex interaction rules (logical gates) [33-35].

1.1.4 Functional consequences of changes in transcriptional regulation

Changes in transcriptional regulation comprise a significant component of the genetic basis for phenotypic evolution. Moreover, these changes were suggested to play a key role in speciation [36, 37]. Transcription regulation can be viewed at several levels, from single TF binding sites, through entire promoter organizations to the architecture of complete gene regulatory networks. According to this hierarchical structure, variations ranging from single nucleotide substitutions within individual binding sites, through larger variations on the promoter level, involving multiple nucleotide bases, to network rewiring, may all affect gene expression and alter phenotypes.

1.1.4.1 Variations within individual TF binding sites

A natural consequence of the short and degenerate nature of TF binding sites is that highly similar sites within the same genome are in some cases recognized by the same TF whereas in others serve as targets for distinct TFs. This is also observed in the genomes of related species, where slight changes in binding site sequence, occurring throughout evolution, may either maintain the specificity of the site to the original TF or alternatively lead to its loss or create a site targeted by a different TF [20, 38]. The desire to distinguish between 'neutral' binding site variations, which do not change the recognition range of the site, and 'functional' variations, which may affect gene expression by altering protein-DNA interactions, poses a great challenge. Such a distinction may have important implications; Firstly it should greatly improve the performance of scanning algorithms, which search promoter sequences for matches to predefined PWMs. These algorithms typically regard all mismatches between a promoter sequence and a given PWM's preferences as equal (c.f. ScanACE [26], Match^{TM} [39], MAST [40]). More reliable predictions may be obtained if such mismatches are differentially weighed based on their expected effects on expression.

Identification of genuine sites is also crucial when comparing the promoters of orthologous genes - some across-species variations may change the functionality of a motif in some of the organisms. Another intriguing application is the detection of regulatory site variations, which have the potential to cause phenotypic diversity within a population, and ultimately diseases, which may occur through altering gene expression. Disease-causing binding site variations are known to be wide-spread [41, 42], however so far no attempts have been made for their prediction on a genome wide scale. Most efforts to distinguish disease-causing variations from neutral ones have focused on coding single nucleotide polymorphisms (SNPs) [43-49]. Estimates show that the human population contains thousands of *cis*-regulatory variations [50]. Such high numbers justify a dedicated effort for the development of computational means for predicting deleterious regulatory variations. The project described in section 2.2 of the results lays the foundations for the development of such methods by introducing measures for quantifying the effects of binding site variations on gene expression, and systematically applying them to variations within motifs belonging to the *S. cerevisiae* dictionary. The measures were initially applied to individual binding sites, and subsequently statistics from multiple substitutions across various binding sites were accumulated in an attempt to characterize nucleotide substitutions that alter gene expression. The developed tools and measures demonstrated in yeast can be applied to other organisms and specifically to human.

1.1.4.2 Variations on the promoter level

Variations on the promoter level range from small scale variations such as SNPs, short insertions or deletions (indels) and tandem repeats, to complete promoter rewiring (involving multiple nucleotides) all of which can alter transcription. Variations within promoter sequences surrounding TF binding sites may affect expression by altering local DNA conformation or changing the spacing between binding sites [23, 51].

Promoters are thought to be composed of multiple TF binding sites, arranged in a modular fashion, which facilitates the promoter's rapid evolution towards novel regulatory programs. This is apparent following gene duplication, when functional divergence is observed not only in the encoded protein but also in its *cis*-regulatory sequences. The duplication-degeneration-complementation (DDC) model of promoter evolution [52] proposes that selection can maintain functionally redundant coding

sequences after gene duplication if each copy loses a different promoter module due to random mutation. The different gene copies thus develop distinct temporal and/or spatial expression patterns, which 'justify' the retention of several genes, supposedly encoding similar proteins, in a single genome [53]. In section 2.3 of the results, we describe a study conducted in a human gene family of 13 members, which deals precisely with this aspect of promoter evolution. The 13 genes encode extremely similar proteins, which are known to perform similar functions. The differences and perhaps the key for understanding the need for all 13 genes, lie within their promoters, which have undergone considerable evolution. In this project we attempted to link changes in promoter binding site composition to distinct regulatory programs of the corresponding genes. The hypothesis being that as a result of promoter evolution, different family members are induced in response to distinct stimuli.

Finally, although generally resulting in altered transcription profiles, loss and gain of individual binding sites may also preserve promoter functionality due to compensation. Studies in *Drosophila* have revealed conservation in expression program despite fluidity in the exact composition of regulatory regions [54]. This level of analysis was not touched upon in the present study, but is important to mention as it may account for false predictions; Such compensations when taking place, may result in no phenotypic effects in cases where our analysis would predict an altered expression profile.

1.1.4.3 Gene regulatory networks and robustness towards mutations

Gene regulatory networks are supposedly organized in such a way that they produce consistent transcriptional outputs across a range of TF concentrations and TF-binding site interactions. This may be the result of natural selection to stabilize transcription against environmental variation and genetic background.

It is intriguing to understand how the transcription control network insures its stability with respect to mutations along with an ability to adapt and acquire new functions. Transcription networks are composed of TFs with different binding specificity requirements. Factors with highly sequence-specific binding impose severe constraints on binding sites and increase the sensitivity to mutation. Factors with low specificity of binding confer robustness towards mutations, yet increase the probability of spurious interactions. It was suggested that robustness is maximized by the compromise between these two effects [55]. This notion is supported by the

observation that individual binding sites exhibit different evolutionary constraints [19]. Higher evolutionary rates of some binding sites are expected to be related to a higher flexibility in the binding properties of the corresponding TFs.

Our analyses of the effects of variations at the single motif level are thus directly related to the entire network architecture. By predicting the consequences of variations within different binding sites, we can identify sites bound by highly specific TFs (these sites are highly sensitive to mutation) versus sites targeted by 'promiscuous' TFs (these sites can tolerate mutations with no apparent effect). Both types of sites should be present in an organism's motif repertoire in order to allow for a robust regulatory network.

Compensation effects, such as those described on the promoter level, may exist on the network level as well. Such compensation, termed 'genetic buffering', accounts for cases in which the gene's function is altered or lost, but this loss is not reflected in the phenotypic level. The 'rescue' can stem from a duplicate gene which is 'reprogrammed' to take over the lost function [53] or from a bypassing network path [56]. Network level compensation may be in action when all analyses (both on the single binding site and on the promoter levels) predict that a crucial gene is not expressed at the correct time and place or in the required level, yet no phenotypic effect is apparent.

In the current study we deal with genome-wide (large-scale) discovery of TF binding sites as well as with deciphering promoter organization, within a multi-gene family (mezo-scale). We introduce a *de-novo* motif finding method, as well as employ available scanning algorithms to search promoters for the presence of previously characterized binding sites. We attempt to link variations within individual binding sites and within entire promoter organizations, to changes in gene expression. The higher level of network organization, although complementary to this work, is outside of its scope.

1.2 Natural antisense transcription – a possible mechanism for the control of gene expression at different levels

1.2.1 Definition and extent of the phenomenon

Natural antisense transcripts (NATs) are endogenous RNA molecules containing sequences that are complementary to other transcripts. In numerous individual cases, NATs have been reported to negatively regulate their conjugate sense transcripts at

several levels: transcription, messenger RNA processing, splicing, mRNA stability, mRNA transport and translation. NATs are also linked to monoallelic gene expression through mechanisms that include genomic imprinting, X-inactivation and clonal expression.

NAT pairs are conventionally divided into *cis*-NATs, which are transcribed from opposing DNA strands at the same genomic locus, and *trans*-NATs, which are transcribed from separate loci. *cis*-NAT pairs display, by definition, perfect sequence complementarity, whereas *trans*-NAT pairs often display imperfect complementarity. In addition it is commonly accepted that *trans* NATs may target many sense transcripts to form complex regulatory networks, whereas *cis*-NATs have but a single sense target [57].

cis-NATs can be categorized according to their relative orientation and degree of overlap; head-to-head (5′ to 5′), tail-to-tail (3′ to 3′) or fully overlapping (Figure 1). All genome-wide studies, except one [58], have reported the tail-to-tail orientation to be the most prevalent. Overlapping transcripts might comprise two protein encoding genes, one protein-encoding and one non-encoding gene, or two non-encoding transcripts.

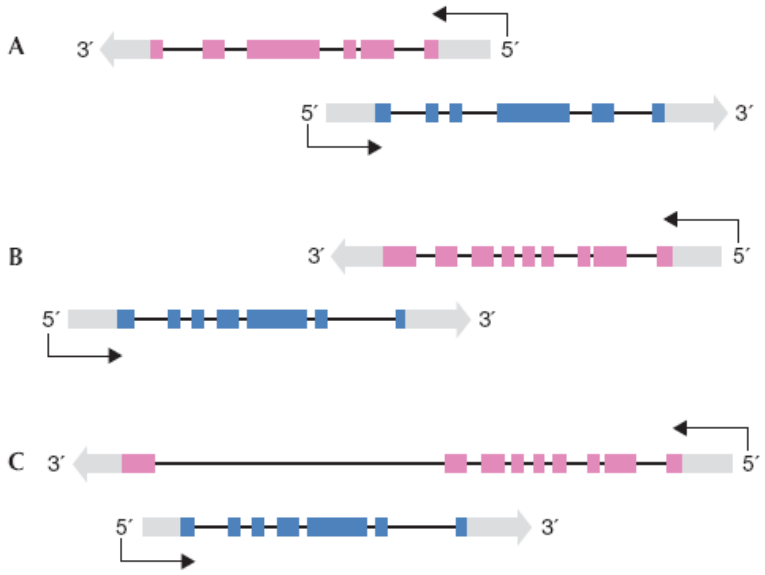


Figure 1: Illustration of *cis*-natural antisense transcript pairs. Three relative orientations are displayed: (A) Head-o-head (5′ to 5′) overlap involving 5′-untranslated regions and coding exons. (B) Tail-to-tail (3′ to 3′) overlap. (C) Fully overlapping (one gene entirely included within the region of the other). Colored boxes represent exons, grey boxes represent untranslated regions

cis-NATs were first detected in viruses [59], then in prokaryotes [60, 61] and finally in eukaryotes [62-64]. In recent years it has become apparent that NATs are

widely prevalent in the genomes of multiple species from viruses to human. Genome-wide computational studies have estimated that the percentage of transcriptional units involved in sense-antisense overlaps, ranges from 5% to 29% in animals [58, 65-70] and from 7% to 9% in plants [71, 72]. Most estimates are based on the alignment of full-length cDNAs and expressed sequence tags (ESTs) to the genome, and the identification of overlapping transcripts on opposite strands. Such a procedure is limited to the detection of *cis*-NATs, implying that the extent of the NAT phenomenon might be much broader. Some predictions were further experimentally validated using methods such as reverse transcription–PCR (RT-PCR) and microarrays containing strand-specific probes [65, 67]. A recent paper suggests that in the human and mouse genomes, the percentage of genes involved in sense-antisense pairing is in fact much higher and approaches 50% [73].

The conservation of this phenomenon across kingdoms implies that NATs may constitute a common mechanism for regulating gene expression, however there remains a possibility that genome-wide antisense transcription is a mere outcome of a 'leaky' RNA transcription machinery. Below (section 1.2.2) we review several lines of evidence supporting the notion that antisense transcription is tightly regulated and moreover that its regulation may be coupled to that of the sense transcript. Additionally we list the different modes by which antisense is currently known to regulate gene expression (section 1.2.3).

1.2.2 Evidence for a regulated process

Most coding *cis*-NAT pairs overlap in their untranslated regions (UTRs) due to alternative polyadenylation (forming transcript variants that differ in their 3′□ termini), or heterogeneous transcription start sites (creating head-to-head overlaps). A key question is whether such alternative (3′□ or 5′) end processing is intentional - forming regulated transcript overlaps - or does it result from 'leakage' of the RNA transcription machinery.

To address this question, Dahary and colleagues examined the evolution of *cis*-NATs [74]. They defined a set of consecutive gene pairs in the human genome and identified their orthologous gene pairs in both mouse and *Fugu*. The human genes were divided into sense-antisense pairs and pairs which are transcribed from the same strand. The authors assumed that if sense-antisense pairs carried a beneficial function, selection would work against their separation in related species. Indeed, 23.3% (55

out of 236) of the human sense-antisense pairs remained consecutive in *Fugu*, compared with only 13.5% (170 out of 1,250) of the same-strand pairs. Moreover, although the *Fugu* genome is much more compact than the human genome, the average distance between sense-antisense gene pairs was only slightly greater in humans than in *Fugu*, whereas same-strand pairs were significantly further apart (up to 11-fold).

If antisense transcription is indeed beneficial, how is it regulated? A recent study [75] mapped the binding sites of three human TFs - SP1, c-Myc and p53 - to chromosomes 21 and 22 using ChIP-chip technology [6]. Surprisingly, 36% of the binding sites mapped within or immediately 3′ to well-characterized protein-encoding genes and were associated with non-encoding RNAs. Moreover, many overlapping sense-antisense transcripts showed correlated expression. Some overlapping transcripts were flanked by binding sites for the same TF, implying that the sense and corresponding antisense transcripts might in fact be co-regulated. Similar results were obtained for the human TF CREB [76].

Additional evidence for genome-wide regulation of antisense transcription was revealed upon the recent completion of the ENCODE (Encyclopedia of DNA Elements) pilot project [77]. This project attempted to functionally annotate the human genome, by mapping a variety of sequence elements (exons, promoters, enhancers, TF binding sites, methylation sites etc.) to 1% (~30 Mb) of its sequence. A computational analyses of "ENCODE-wide" ChIP-chip data (with factors known to mark transcription initiation) revealed a surprisingly high proportion (23%) of promoters located on the anti-sense strand of previously identified coding transcripts [78]. These promoters potentially drive the transcription of anti-sense transcripts.

Sense-antisense gene pairs were reported to be co-expressed or inversely expressed more frequently than would be expected by chance [79]. Moreover, co-expressed and inversely expressed sense-antisense pairs display striking conservation throughout evolution [74, 79]. Both conservation and coupled sense-antisense expression are more prevalent in tail-to-tail NAT pairs, suggesting that such an orientation is not only the most abundant, but also more likely to have a regulatory function [80].

Lastly antisense genes, especially those which are evolutionary conserved, were reported to have particularly short introns in humans, mice and Drosophila [68, 81].

It was suggested that, in the case of the antisense genes, the purpose of short introns is not to allow a high level of expression or spurious expression, as has been shown for other genes [82, 83], but to address the need for a rapid response. This proposition should be constrained by the fact that formal math analysis [84] shows that response time (defined as the time at which an mRNA reaches half of its steady state level) is actually dependent on the rate of degradation and not the rate of production. Naturally short introns would affect mRNA production and not degradation.

Although the above studies suggest that antisense transcription is tightly regulated and evolutionarily conserved, they should be regarded with some caution. The observed co-expression of *cis*-NATs might originate from the known tendency of genes in close proximity (even in the same genomic strand) to be co-expressed, e.g. as a result of local chromatin structure or shared regulatory elements [85, 86]. It is possible that proximal co-expressed genes that were not selected against seem to be 'tailored' by evolution to serve a regulatory purpose. A recent study introduced the intriguing concept of "neutral expression" [87]. The authors of this study argue that mutations that alter gene expression might not always be sufficiently deleterious to be eliminated by purifying selection, and therefore might be fixated in the population by random drift. According to this idea, the possibility that some NATs represent cases of residual transcription cannot be entirely eliminated.

1.2.3 Principal mechanisms by which NATs regulate gene expression

Well-documented examples point to four major mechanisms by which NATs may regulate gene expression [88]: transcriptional interference, RNA masking, double-stranded RNA (dsRNA)-dependent mechanisms and chromatin remodeling:

(i) *Transcriptional interference* - The presence of an overlapping transcriptional unit might stall sense transcription owing to the collision of two bulky RNA polymerase II complexes on opposite strands. This is most apparent in the transcription elongation step as has been shown for the yeast gene pair GAL10 and GAL7 [89]. An additional support for this collision model is the recent finding that for both human and mouse, the expression level of *cis*-NATs decreases as the length of the overlapping region (between sense and antisense transcripts) increases [90]. Furthermore, in *Escherichia coli* , the collision of RNA polymerases was observed by atomic force microscopy

[91]. This observation showed that RNA polymerases do not pass each other or displace one another, but instead stall against each other.

(ii) *RNA masking* - Sense-antisense duplex formation might mask *cis* elements residing in either of the transcripts and hinder processes that require protein-RNA interactions such as splicing, mRNA transport, polyadenylation, translation and degradation. The best characterized example of this mechanism is the antisense transcript for the thyroid hormone receptor gene erbAα, which shifts the balance between two splice variants through the masking of a splice site [92]. A more recent example is that of the human apoptotic gene *FAS* and its antisense *SAF* [93]. Over-expression of *SAF* alters the splicing pattern of *FAS* in a regulated way, suggesting that *SAF* controls the splicing of *FAS*. On a larger scale, a quantitative analysis of genome wide sense-antisense pairs in human and mouse, suggested that the presence of an antisense transcript, complementary to an exon-intron border of the sense gene, increases the rate of retention of the respective intron [73].

(iii) *dsRNA-dependent mechanisms and RNA interference* - There is accumulating evidence that antisense transcripts might function through the activation of dsRNA-dependent mechanisms such as RNA editing and RNA interference (RNAi). Both these mechanisms play a role in the nuclear defense strategy against dsRNA. While RNA editing, which involves the deamination of dsRNA adenosines to inosines [94], was recently shown to be negligible in sense-antisense duplexes [95], RNAi is likely to be a prominent mechanism.

RNAi involves cleavage of dsRNA by the enzyme Dicer into 21-23 nucleotide duplexes. These duplexes are further separated into single strands and become part of the RNA-induced silencing complex (RISC). RISC mediates their degradation or the repression of their translation [96, 97]. Several precedents suggest that sense-antisense transcription can induce gene silencing through an RNAi-dependent mechanism; Salt tolerance in *Arabidopsis* is regulated by two small interfering RNAs (siRNAs) produced from a pair of overlapping protein-encoding genes [98]; The regulation of iron deficiency in cyanobacteria [99] and the maintenance of male fertility in Drosophila [100] were both shown to involve dsRNA. The same mechanism could apply to other eukaryotic *cis*-NAT pairs. In fact, 11 *Arabidopsis* siRNAs have been mapped to complementary regions of overlapping transcripts, suggesting that these

overlapping transcripts might feed into the RNAi machinery [101]. So far, however, there has been no evidence for mammalian antisense transcripts acting through duplex formation.

(iv) *Antisense involvement in methylation and monoallelic expression* - Non-coding antisense transcripts have been reported to induce the methylation and silencing of corresponding genes. For example, thalassemia - a form of anemia - is caused by antisense-induced DNA methylation (and silencing) of the human hemoglobin 2 gene [102]. Antisense transcripts are also involved in X-chromosome inactivation [103], autosomal imprinting [103] and allelic exclusion in B and T lymphocytes [104]. In all these cases, non-coding antisense transcription affects an entire gene cluster, rather than merely the overlapping sense transcript [105, 106]. The silencing effect is probably exerted through the recruitment of histone-modifying enzymes, resulting in chromatin remodeling and transcription silencing.

For additional details and specific examples of each of the described mechanisms, see our recently publish 'Concept Paper' [107].

In section 2.4.2 of the results, we combine the current knowledge regarding the different mechanisms of antisense action with the knowledge regarding its own transcriptional regulation (regulation of the regulator), and present our hypothesis [107], that the regulation of antisense transcription might be tailored to its mode of action. According to this model, an experimentally observed relationship between the expression pattern of a NAT and those of its target genes might indicate the regulatory mechanism that is in action.

In section 2.4.1 of the results we challenge the well accepted notion that *cis* and *trans* encoded NATs represent two separate phenomena; We demonstrate that a single antisense transcript may have both *cis* and *trans* targets and that the same mRNA may potentially be regulated by both *cis* and *trans* encoded NATs. This extends the definition of NATs, suggesting that they form a putative regulatory network, which exhibits many-to-many relations. We further identify properties that are common to targets of the same antisense transcript, and hint to the biological processes that may be regulated by NATs.

2 Results

2.1 Regulatory motif dictionaries

Aim

The aim of the regulatory motif dictionary project was to produce reliable reference collections ('dictionaries') of transcription factor binding sites (TFBS) across a range of eukaryotic species. Such motif collections may serve both for studying the properties that distinguish functional binding sites from their surrounding DNA sequence, and for conducting higher level analyses of the regulation of gene expression.

Major findings and conclusions

We developed a methodology that combines genome-wide promoter sequences with whole-genome mRNA expression data, in order to create regulatory motif datasets that are unbiased by prior knowledge of TFBS. Our method is based on the premise that any nucleotide sequence that resides in a promoter of a gene, may contribute to the regulation of that gene's expression. It employs the previously described EC score [21, 22] to assess the effect of a promoter sequence motif on the expression profile of the corresponding gene. Our dictionary construction method is thus unbiased, applicable to different genomes and moreover forms a quantifiable connection between binding site sequences and the expression profiles of the genes they regulate. Each motif is characterized by its DNA sequence (its 'syntax'), the set of genes that contain it in their promoters, the set of biological conditions in which these genes display coherent expression, and their corresponding expression profiles. The latter comprise the 'semantics' of the motif, namely its likely regulatory function.

Applying this methodology (discussed in detail below), we constructed motif dictionaries for the yeast *S. cerevisiae* across 40 different time series expression experiments, for the yeast *Candida albicans* for three experiments of stress response and for human across the cell cycle. In a complementary effort carried out by Shai Shen-Orr (a former lab member), dictionaries were constructed for the worm *C. elegans*, during embryonic development, using the software package and the statistical tools generated here.

The human dictionary was constructed mainly to assess the feasibility of our method in higher organisms; the *C. albicans* dictionaries were used to study

evolutionary conservation of stress response in yeast, whereas most of the comprehensive analyses were carried out in *S. cerevisiae*; We found that our yeast motifs have a extremely good coverage (91%) of a recently published intensive experiments-based motif dataset [8] , that they are evolutionary conserved, display a high GC content in comparison to their AT rich promoter environment, and tend to appear at specific distances from the transcription start site (TSS). Furthermore we found that many of these motifs regulate gene sets that share common biological functions. These findings both validate the capability of our method to identify true motifs and reveal properties that may distinguish binding sites from biologically meaningless DNA stretches.

2.1.1 Yeast (S. cerevisiae) motif dictionaries

2.1.1.1 Dictionary construction methodology

In order to construct the motif dictionaries we integrated whole genome promoter sequences of *S. cerevisiae* with expression patterns of the corresponding genes in 40 natural and perturbed biological conditions including cell cycle, sporulation, diauxic shift and various stress responses. Each biological condition was represented by a time series experiment, monitoring yeast whole-genome mRNA levels via Affymetrix gene chips or microarrays (see methods 3.1.1.1). All together these experiments measure the mRNA levels of over 6,000 *S. cerevisiae* genes, although an individual experiment may contain only a subset of these genes. We assigned promoter sequences to 5,642 of the genes for which expression data was available. The dictionary methodology attempts to assess the effect of a sequence motif that is present in the promoter of a gene on the gene's expression profile. To quantify this effect we used the EC score [21-23]. In short the EC score measures the extent to which a set of genes display similar expression profiles at a given set of conditions. Formally, the EC score of a set of N genes is defined as the fraction p of gene pairs in the set, for which the Euclidean distance between normalized expression profiles falls bellow a threshold D, EC=p/[0.5*N(N-1)] [22] (see methods section 3.1.4). The EC score can be calculated for any set of genes for which expression data is available. For the purpose of motif assessment, we applied it to genes that contain a given motif in their promoter, and used the score to test the hypothesis that the motif exerts an effect on expression at the examined condition.

We applied two complementary approaches to the task of creating a comprehensive motif dictionary; the exhaustive approach and the syntax based approach (see section 2.1.1.4). In the exhaustive approach, we systematically scanned all k-mers (k ranges from 7-11) that appear in the promoters of *S. cerevisiae* genes. For each such k-mer we hypothesized that the mere existence of this nucleotide sequence in the gene's promoter plays a role in regulating the transcription of the gene. We applied the EC score to all sets of genes containing a given k-mer in their promoter, and asked whether they are similarly expressed in a variety of experimental conditions. We developed a sampling-based approach to assess a p-value to each EC score, given the size of the corresponding set of genes. This p-value estimates the probability of obtaining the observed or higher EC score by chance [21] (see methods section 3.1.4). Further we used false discovery rate (FDR) [108] of 0.1 (allowing 10% false positives) to correct for multiple hypotheses. 8,610 sequence motifs appeared significant in at least one out of the 40 examined experimental conditions. In the syntax based approach, we applied the same procedure (EC score followed by FDR) to a set of pre-selected candidate motifs, instead of to all possible k-mers .Motif pre-selection was based on syntactic rules adapted from the field of linguistics. The following sections refer to the exhaustive approach, whereas the syntax based approach is described in section 2.1.1.4.

2.1.1.2 Validation of the dictionary methodology

We employed two different tests to assess the soundness of our scoring methodology: 1. Given a set of PWMs from the literature, we asked how well do they score using our method (2.1.1.2.1) 2. Running our method without any prior knowledge, we asked how well do the high scoring motifs cover known motifs (2.1.1.2.2). We further examined whether gene sets, which were signified by our method as coherently expressed, also share similar biological functions (2.1.1.2.3). This served as an additional assurance of their biological relevance.

2.1.1.2.1 Assessing published motifs using our scoring method

In order to test whether previously published regulatory motifs score highly using our method, we calculated the EC scores of 102 recently published *S. cerevisiae* binding sites [8] (hereafter referred to as the Harbison set) in the 40 different time series experiments (ExpressDB [109]). 89/102 (87%) of the Harbison binding sites

passed FDR of 0.1 in at least one experimental condition. The Harbison motif set is slightly redundant, because it contains several TF families which recognize highly similar binding sites. Therefore, when clustering all of Harbison's binding sites according to the similarities among their PWM representations, the 102 sites fall into 79 distinct clusters (see methods section 3.1.6). The Harbison motifs that scored significantly in at least one of the 40 examined conditions belong to 68/79 (86%) of these clusters. By calculating the EC scores we were able to assign at least one Harbison binding site to 39 out of the 40 examined conditions, providing a functional description for these sites. As a control we created 102 random gene sets in sizes corresponding to the sets of genes containing each of Harbison's motifs. Only 15/102 (~15%) of these control gene sets appeared significant in at least one condition (Figure2). This is only slightly above what would be expected by chance when applying a false discovery rate of 10%.

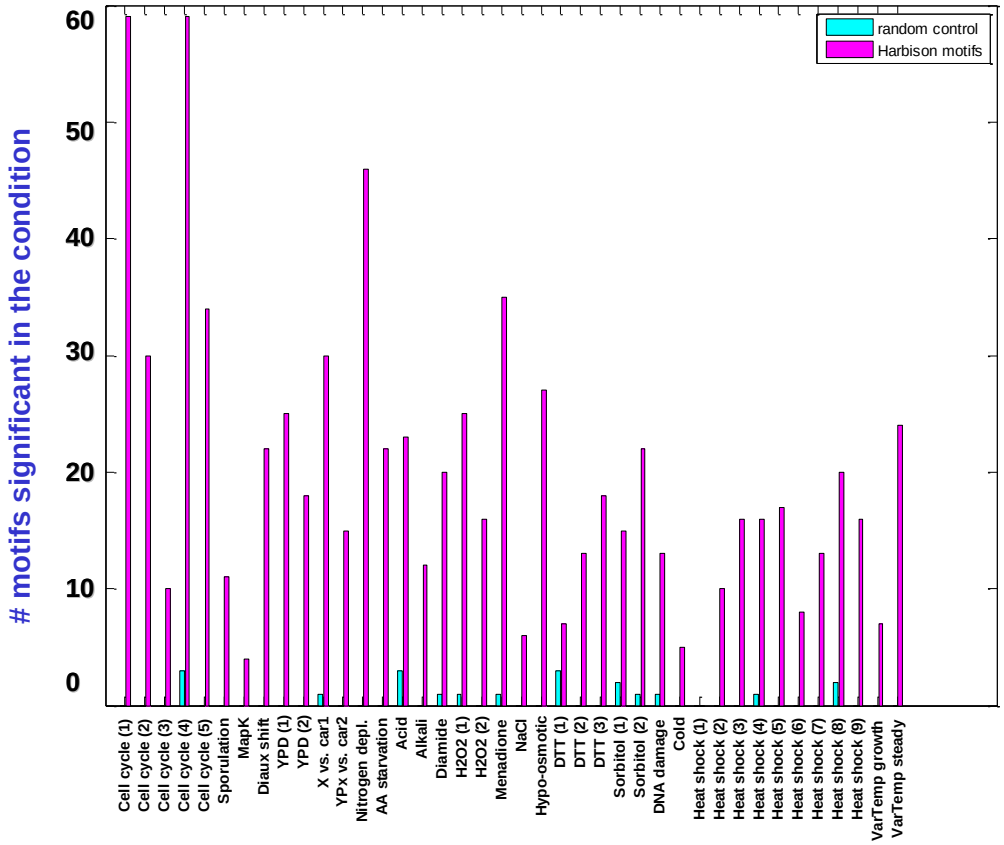


Figure 2: Re-discovery of the Harbison motif set. Number of significantly scoring Harbison motifs (pink) versus a control of random gene sets (cyan). Significance was tested across 40 experimental conditions (x axis). Not all conditions require the same amount of regulators, the largest number of Harbison's TFs appear to regulate (i.e. obtain significant EC scores in the corresponding experiments) cell cycle, nitrogen depletion, oxidative stress (in response to menadione, a superoxide-generating drug), and hypo-osmotic shock.

2.1.1.2.2 Coverage of known motifs by the dictionary k-mers

Applying our method to all k-mers (length 7-11) residing in yeast promoters, with no prior knowledge, we obtained 8,610 putative binding sites. We compared these putative sites to the PWMs of binding sites published in the literature. We used a score between 0-100 that denotes how likely a given string is to be generated from a given PWM (see methods section 3.1.5). Requiring a match score of 99 between dictionary strings and Harbison PWMs, we obtain a coverage of 89/102 of Harbison's motifs (87%), and of 72/79 (91%) of the non redundant set. If we relax the similarity requirement to 95, we obtain a coverage of 99/102 (97%) motifs, falling into 77/79 (97%) clusters (see Table 1). Our coverage of the Harbison set is significantly higher than that of a random motif set of the same size (p-value =10^-5).

Comparison score cutoff	Dictionary coverage	Coverage of known	Unique known clusters
99	1402/8610=16%	89/102=87%	72/79=91%
98	1528/8610=18%	93/102=91%	75/79=95%
97	1719/8610=20%	96/102=94%	75/79=95%
95	2198/8610=25%	99/102=97%	77/79=97%

Table 1: Coverage of the Harbison motif set by our dictionary strings. A scoring method was devised to assess how likely a given string is to be generated from a given PWM. The score is on a scale of 0 to 100 (see methods section 3.1.5). We computed this score for all 8,610 dictionary strings over the 102 Harbison PWMs. The coverage of Harbison's motif set was assessed for several different score cutoffs. Note that one dictionary string may match more than one Harbison PWM, because of redundancy in Harbison's dataset. There are two very long (17 and 18 positions) gapped motif in Harbison's set, for which we have no match, because the dictionary only covers motifs of length 7-11.

2.1.1.2.3 Functional coherence (FC) analysis

For further validation of the regulatory potential of our significant motifs, we examined whether the sets of genes defined by each of the 8,610 dictionary motifs share common biological functions. Such common functionality may indicate a need for common regulation. To assess common functionality, we employed the Functional Coherence (FC) score [110], which uses similarity in Gene Ontology (GO) annotations [111] to quantify the overall functional similarity among a set of genes, in a manner similar to the EC score (see methods section 3.1.7). A set of genes is functionally coherent if its genes are significantly closer to each other in function than expected by chance given the size of the set. 1,440 (17%) out of the 8,610 motifs that were selected based on significant EC scores, also scored significantly in the functional coherence test, using similarity in GO biological process annotations. For

comparison, among 1,000 randomly selected strings from a control set of lowly scoring k-mers (see next section for a full description of the control set), only 3 (0.3%) obtained a significant FC score. This is an additional reassurance that many of our discovered motifs are biologically relevant. FC and EC can be seen as two complementary approaches. A certain overlap is expected between motifs that score highly in EC and those that score highly in FC; Genes that participate in similar biological processes (significant FC) are in many cases (although not always) co-regulated. Genes that are co-regulated (significant EC) are needed in the cell at the same time and thus are likely to belong to the same biological process. However co-expressed genes may also belong to several processes that happen to take place in the cell at the same time.

2.1.1.3 Distinct characteristics of dictionary motifs

Following the validations described above, we refer to our significant motifs as likely *cis*-regulatory elements and use them to investigate characteristics that may be of relevance to their biological function. For this purpose we compiled a control set of 190,211 low scoring k-mers, that were insignificant in all 40 examined biological conditions, and in addition scored especially low (p-value > 0.8, gene set size>8) in at least one of these conditions. We considered various features that may be important for the function of a regulatory motif and for each such feature, defined a quantitative measure, and tested whether it can significantly differentiate between our highly scoring motifs (included in the dictionary) and the control set.

Compared to the control set our significant motifs were found to have high GC content (relative to the yeast AT rich genomic background) (Figure 4), to have high information content (Figure 6), to appear in higher copy numbers (Figure 8) and to display a preference to distinct positions relative to the TSS in different promoters (positional bias) (Figure 7). Additionally our motifs were found to be evolutionary conserved in the promoters of four closely related *Saccharomyces* species (Figure 9). Some of these properties are known to characterize functional binding sites (positional bias [23, 112], multiplicity of sites [21, 113], evolutionary conservation [13, 14]). Below we discuss these analyses in further detail.

2.1.1.3.1 Motif length

The distribution of string lengths differs significantly (ranksum test: $P<10^{-300}$) between dictionary motifs and low scoring motifs. The dictionary motifs peak at length 8, the low scoring at lengths 9 and 10, whereas the entire set of all k-mers of lengths 7-11 peaks at length 11 nucleotides. For comparison we looked at the distribution of PWM width (number of columns) in the recently published Harbison motif set. The mean PWM width is slightly over 9 positions, but this set includes very long PWMs representing gapped motifs which shift the mean. According to the ranksum test the length distribution of the Harbison motifs can not be distinguished from the length distribution of our highly scoring motifs (P= 0.4976), however it is distinguishable from the length distribution of the control motifs (P=0.0297).

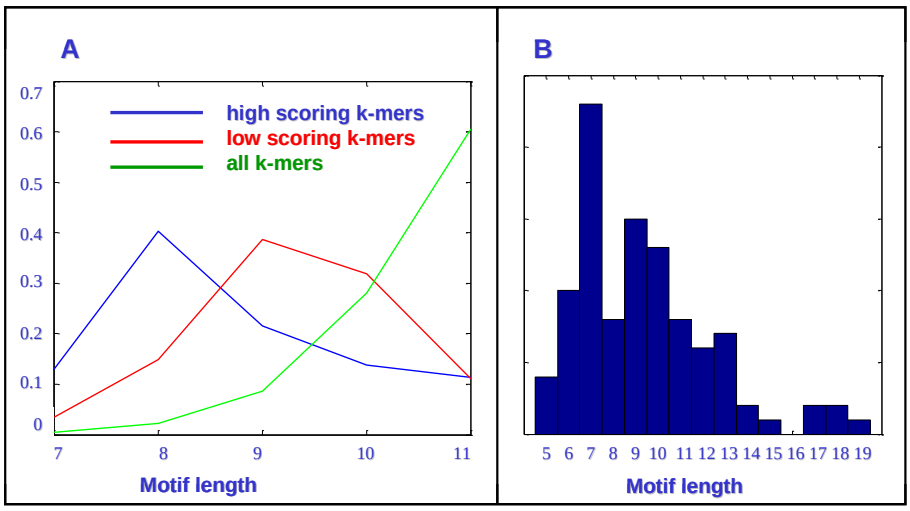


Figure 3: A. Length distribution of high scoring k-mers versus low scoring k-mers. B. Length distribution of the Harbison motif set. This set includes some PWMs that represent gapped motifs and are thus much longer than our un gapped motifs.

2.1.1.3.2 Motif GC content

The nucleotide composition of a motif may be crucial for its function, for instance by allowing it to be readily distinguished from its surrounding promoter sequence. The distribution of normalized GC content (number of GCs/motif length) of the high scoring motifs differs significantly ($P<10^{-300}$) from that of the low scoring motifs, as seen in Figure 4A. The distribution of the low scoring k-mers peaks close to the mean background promoter GC content (36%), whereas the high scoring motifs have a GC content distribution that is comparable to that of the Harbison motif set (ranksum test:

P=0.7445) (Figure 4B). This may allow these motifs to be detectable by the TF on the background of the AT rich (38% GC, 62%AT) yeast genome.

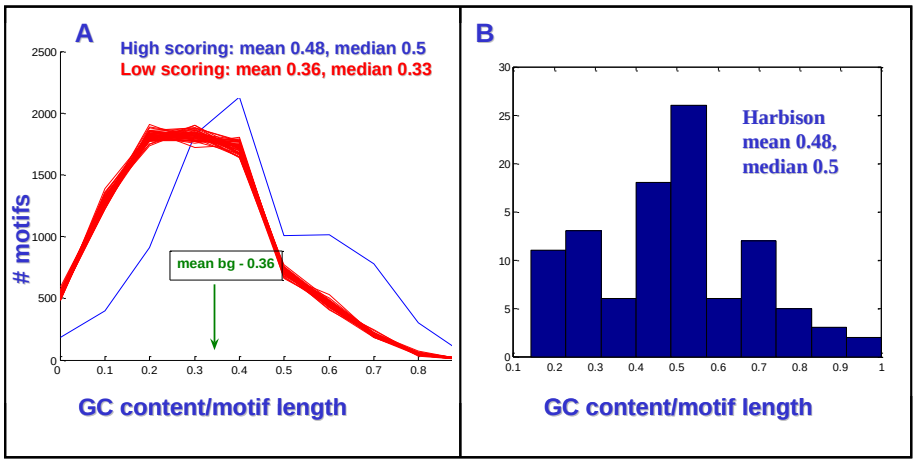


Figure 4: A. Distributions of GC content for high scoring motifs and for 50 random sets of control motifs. The distributions differ significantly ($P<10^{-300}$). The control motifs peak very close to the background promoter GC content (0.36), whereas the high scoring motifs are more GC rich
B. Distribution of GC content for Harbison motifs. The normalized GC content of the high scoring motifs is similar to that of Harbison's set, and significantly higher than that of the control motifs.

2.1.1.3.3 Motif GC contrast

To further assess whether high scoring motifs 'stand out' from their surroundings , we defined a measure termed GC contrast, which gauges the difference in nucleotide composition between the motif and the promoter sequence it is embedded within. We defined GC contrast as the absolute difference between the GC content of a motif and the GC content of its surrounding promoter, averaged over all promoters in which an instance of this motif appears. The absolute value of the difference was used, because we were primarily interested in whether or not the motif is distinguishable from its background. High scoring motifs have a significantly (ranksum $P<10^{-53}$) higher GC contrast than the low scoring motifs as can be seen in Figure 5A (mean-0.166, median-0.135 for high scoring, versus mean 0.127 and median 0.103 for the low scoring). Yet we suspected that the high GC contrast may be a consequence of the GC content signal, since our motifs are GC rich and the promoters are AT rich, the GC contrast may reflect just that. To test this, we plotted GC contrast versus GC content for the dictionary motifs and for the control-set motifs. It is apparent from these plots (Figure 5B) that indeed most of the GC contrast signal stems from the fact that the control set has more motifs with a low GC content (below background) . Given a similar GC content, dictionary motifs and control motifs will have the same average contrast with their surrounding (Figure 4B).

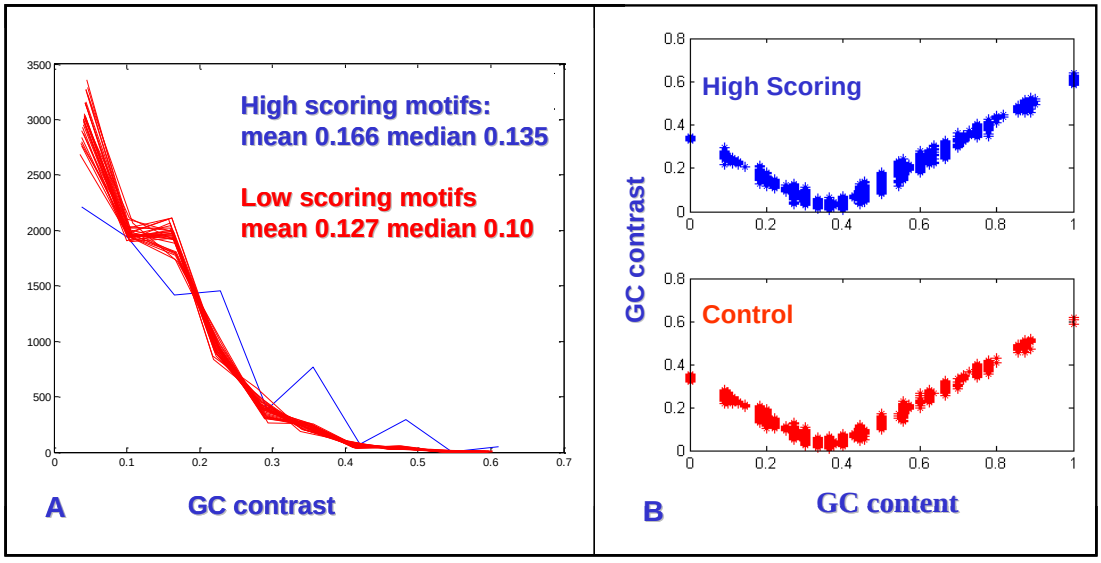


Figure 5: A. Distributions of GC contrast for high scoring motifs and for 30 random sets of control motifs. The distributions differ significantly ($P<10^{-53}$). B. The relation between GC contrast and GC content is similar for both high scoring motifs and control set motifs. This indicates that the differences in the distributions of GC contrast are a consequence of the GC signal.

2.1.1.3.4 Motif Entropy

We defined a measure termed motif Entropy to quantify how evenly are the four nucleotides distributed within a candidate motif. Namely is a functional motif likely to be composed of an equal amount of all 4 nucleotides or mainly of one or two of the nucleotides. The motif Entropy is defined as follows:

$$Entropy = -\sum_{i\in\{A,C,G,T\}} (qi * \log 2(qi))$$

Where i can be any of the four nucleotides and q_i is the frequency of this nucleotide in the motif. For instance, the sequence 'AAAAAAAAAAAA' will have an entropy of 0, 'AAAAAACCCCCC' an entropy of 1, 'AAAACCCCGGG' an entropy of 1.585 and 'AAACCCGGGTTT' an entropy of 2. High scoring motifs that comprise the EC dictionary, have a significantly ($P<10^{-22}$) higher entropy than low scoring motifs (Figure 6A), which is expected because they should have a high information content. However the number of nucleotides a motif is composed of is similar for good motifs and for lowly scoring ones. (Figure 6B). This means that even if a control motif is composed of all 4 nucleotides, the distribution is not even, but instead one or perhaps two of the nucleotides is the most prevalent. A simple count of the number of

different nucleotides within a motif is thus not enough to differentiate between meaningful and nonsense motifs, a more elaborate score such as entropy is needed.

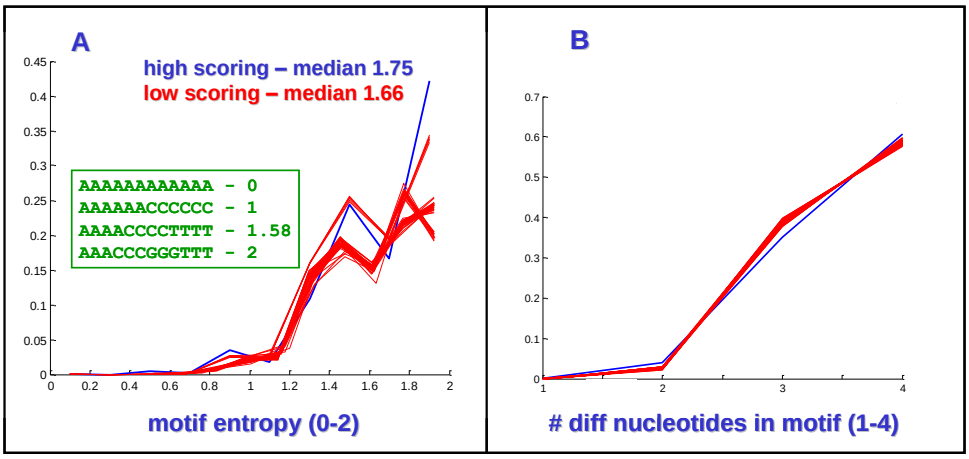


Figure 6: A: Distributions of entropy for high scoring motifs and for 50 random sets of control motifs. The distributions differ significantly ($P<10^{-22}$) B: distribution of the number of different nucleotides comprising high versus low scoring motifs (same distributions)

2.1.1.3.5 Motif positional bias

A majority of the functional motifs are thought to be located at a preferable distance window from the TSS. This positional bias is most likely needed for their function, and specifically for their cooperation with nearby binding sites. To quantify the positional bias of our motifs, we gathered, for each k-mer, the positions (relative to the TSS) of all its genome-wide promoter instances. These positions were sorted into 40 bp wide bins. A positional bias p-value was calculated using a binomial model for the most highly populated interval adapted from Hughes et al. [112] (see methods section 3.1.8). The distributions of positional bias p-values differ significantly (ranksum test - $P<10^{-300}$) between the significant motif set and the control set (Figure 7A). Although there are k-mers from the control set which display a significant positional bias, their preferred positions differ from those preferred by the high scoring motifs (Figure 7B); The most biased motifs among the high scoring set are located mostly at 80-160 nucleotides upstream of the TSS. The first 80 nucleotides are almost devoid of high scoring motifs, probably because of constraints of the basal transcriptional machinery. This is inline with the findings of Harbison et al. [8], which reported very few binding sites in the region 100 bp upstream of the TSS and a sharp peak in binding site number between 100-200 bp. The most biased motifs among the control set are located anywhere between 0 and 240 nucleotides upstream of the TSS, with the majority at a distance of 0-40.

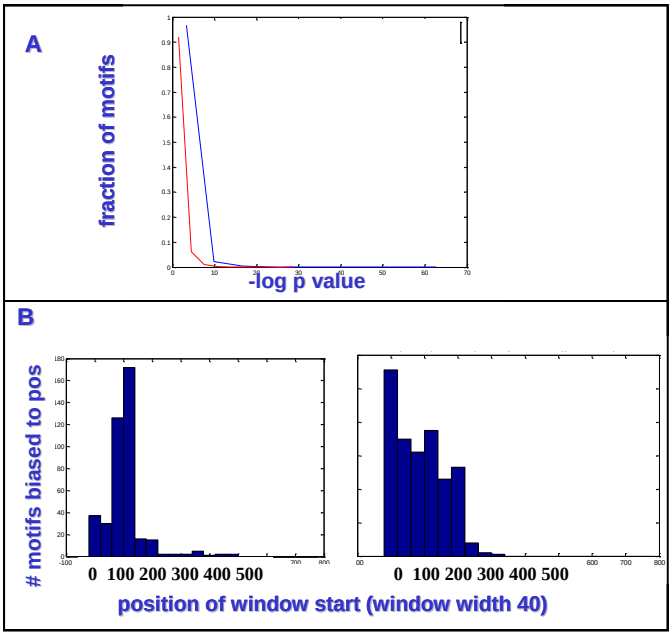


Figure 7: Positional bias A. The distributions of positional bias p-values differ significantly between high (blue) and low (red) scoring motifs. A greater fraction of the higher scoring motifs, appear to have a significant positional bias. The best p-value of the high scoring motifs is 1.08*e-066, while the best p-value for the low scoring motifs is 5.97*e-03. B. The preferred positions are also significantly different between the two sets of motifs (ranksum test – $P=5.25*10^{-4}$). Positionally biased dictionary motifs (left) are primarily located within 80-160 nucleotides from the TSS, whereas the low scoring motifs (right) can be located throughout the first 240 nucleotides.

2.1.1.3.6 Motif copy number

Many functional motifs are present in multiple copy numbers in the promoters of the genes they regulate. We thus observed the distributions of motif copy number per promoter among high scoring motifs and a control set of low scoring ones. The distributions of the maximal number of occurrences of each motif per promoter are significantly different (ranksum test - $P<10^{-135}$). High scoring motifs tend to appear in a larger copy number, this is in line with the common belief that in many cases the same TF binds to multiple binding sites in the same promoter [21, 114]. 29% (2523/8610) of the high scoring motifs appear more than once in a promoter, where the maximal number of occurrences is 27. Only 14% (1171/8610) of the control set motifs appear more than once in a promoter, where the maximal number of occurrences is 11 (see Figure 8).

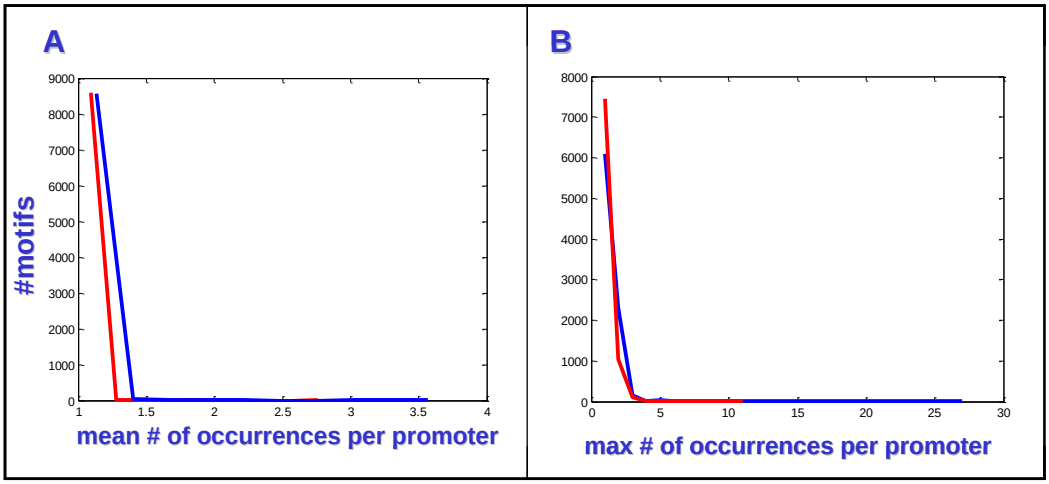


Figure 8: Motif copy number A. The distributions of the mean number of occurrences of each motif per promoter are significantly different (ranksum test - $P<10^{-130}$) for high scoring motifs (blue) versus low scoring motifs (red) . B. Distributions of the maximum number of occurrences of each motif per promoter, high scoring (blue) versus low scoring (red) motifs (ranksum test - $P<10^{-135}$). High scoring motifs tend to appear in a larger copy number.

2.1.1.3.7 Motif evolutionary conservation

Many functional motifs are conserved throughout evolution, thus the evolutionary conservation is another criterion that may differentiate high scoring motifs from biologically meaningless oligomers. Evolutionary conservation across four close *Saccharomyces* species*: S. cerevisiae, S. mikate, S. kudriazevii and S. bayanus* [13]. was assessed as follows: For each motif, we counted the percentage of fully conserved (in all aligned species) positions in each of its instances, and averaged over all motif instances. Our dictionary motifs showed high evolutionary conservation when compared to a control set of randomized k-mers (randomization was used in order to preserve the same GC content). In fact the motifs in our set, which scored significantly in the control of progression through cell cycle, were as evolutionary conserved as a set of motifs that were defined solely based on phylogenetic footprinting [14] (Figure 9). This is striking as conservation was not taken into account in our motif scoring methodology. Moreover it turned out that there is a positive correlation between the normalized EC score (EC score*gene set size) of a motif and its degree of conservation (Figure 10).

When expanding the conservation analysis to the complete set of high scoring motifs, and not only to those regulating cell cycle (see methods section 3.1.9), only 17.6% of our dictionary motifs had a conservation rate higher than the 95th percentile of the control set distribution. This may be partly explained by species specific motifs

that are likely to be present in every genome; About 40% of human functional binding sites are estimated to be non functional in rodents [19]. A similar proportion of species specific sites has been observed in yeast [20]. Furthermore we used quite a strict conservation criteria, requiring a position to be maintained in all 4 species in order to be regarded as conserved. The fact that evolutionary conservation appears to be significant in cell cycle motifs, but not in all high scoring motifs, may suggest especially higher conservation across these species in the context of cell cycle regulation.

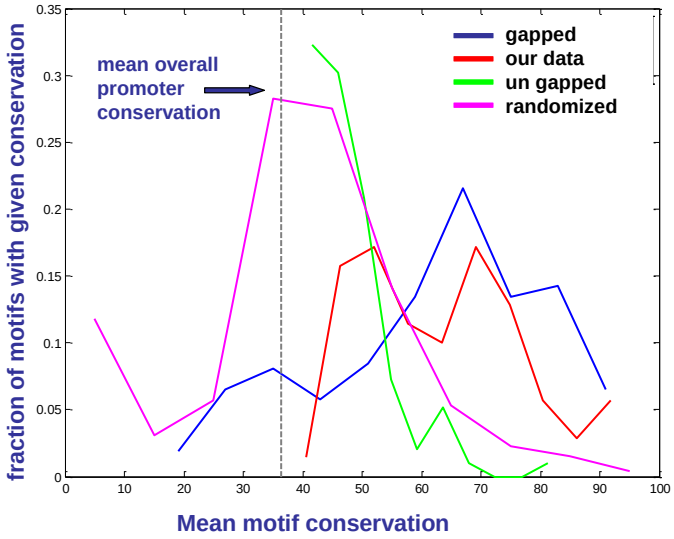


Figure 9: Evolutionary conservation of highly scoring k-mers. Evolutionary conservation in four *Saccharomyces* species [13] was calculated for the highly scoring *S. cerevisiae* cell cycle k-mers (lengths 7-11) and compared with the conservation of two sets of motifs defined by yeast phylogenetic footprinting [13, 14]. A set of the randomized k-mers was used as a control (to preserve the same GC content). The highly scoring k-mers are conserved comparably to one set of published motifs [14] and appear to be more conserved than the second set [13]. The randomized k-mers show an evolutionary conservation that is similar to that of the background promoters (~36%). For each putative motif, evolutionary conservation was calculated by finding the percentage of fully conserved (in all aligned species) positions in each motif instance, and averaging over all motif instances. The background promoter conservation was calculated in a similar manner by counting the number of fully conserved positions in each promoter and averaging over all promoters.

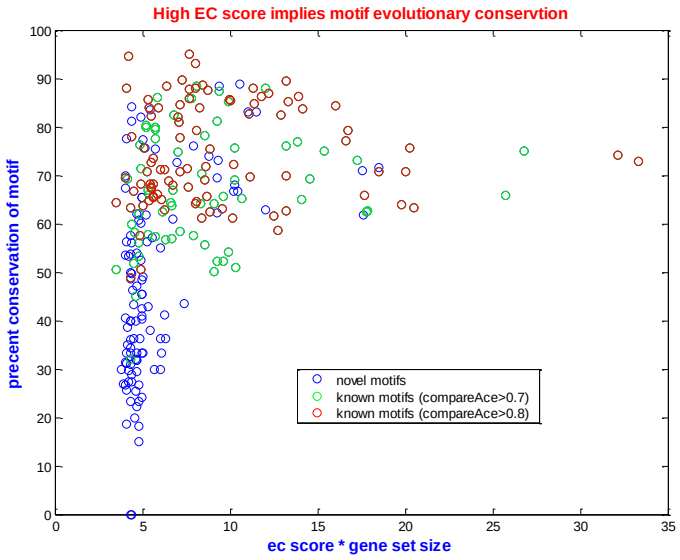


Figure 10: High EC score implies motif evolutionary conservation. Evolutionary conservation in four *Saccharomyces* species [13] was calculated for the highly scoring *S. cerevisiae* cell cycle k-mers (lengths 7-11) and plotted against their normalized EC scores (EC score*gene set size). It is clear from the plot that putative motifs with a higher EC score tend to be more evolutionary conserved. Putative motifs that are similar in sequence to known motifs (CompareAce score>0.8) are marked in red, putative motifs for which there is a known motif with a lower sequence similarity (CompareAce score >0.7) are marked in green. For these motifs our scoring methodology suggests a refined sequence. Potentially novel motifs are marked in blue

2.1.1.4 Motif Extraction Algorithm (MEX) - a syntax based approach

The exhaustive k-mer enumeration has some major limitations: First it is extremely expensive computationally and thus not readily scalable to larger genomes. Second - the binding sites discovered are limited in length, the number of possible k-mers increases exponentially with k, so that scanning sequences longer than 11 nucleotides (4^11 possibilities) is not practical. Thirdly – because no additional information is integrated to limit the search space, multiple false hypotheses are generated and in order to correct for this, a stringent p-value threshold is set by the FDR procedure.

To compensate for these limitations, a complementary approach was applied in collaboration with the research groups of Prof. David Horn and Prof. Eytan Ruppin from Tel Aviv university [29]. They applied a modification of their unsupervised pattern recognition algorithm [28] to the promoter sequences of *S. cerevisiae*. This algorithm was originally designed to extract significant patterns from natural-language corpora, and was adapted to fit the biological task of sequence motif extraction (hence the name Motif Extraction Algorithm – MEX). MEX is based on a statistical model that identifies consecutive chains of interdependencies between

adjacent nucleotide positions. It can thus successfully identify motifs as statistically significant on a genome-wide scale, even without significant over-representation [29] (see methods section 3.1.3). The algorithm readily detects the motif boundaries, as the positions where the series of highly probable transitions begins and terminates. For instance in the English language many words end with ‘ing’, thus observing an ‘in’ at the end of the word, suggests a ‘g’ will follow. After an ‘ing’, any letter of the alphabet can occur (marking the beginning of the next word), and thus the series of high probability transitions is truncated.

Applying MEX to the *S. cerevisiae* promoters produced a set of 8,498 sequence hypotheses. This set despite not being exhaustive has the advantages of (i) no sequence length limitation and (ii) an internal dependency between motif positions that may prove to be biologically significant. Dependencies between regulatory motif positions are known to occur [115, 116], and are not accounted for by the commonly used PWM representation. MEX has the advantage of capturing such inherent dependencies.

2.1.1.4.1 Evaluating the success of MEX in extracting biological significant motifs

To assess the biological functionality of the motifs extracted by MEX, we computed the EC scores of the gene sets containing each motif in the same 40 experimental conditions. 1,873/8,498 (22 %) of MEX’s predictions appeared significant in at least one condition (FDR of 0.1). For comparison – when applying the exhaustive approach – we scanned 1,510,057 hypotheses, 8,610 of which scored significantly (0.6%). In other words we see a striking enhancement in the probability of a k-mer to pass an EC test if it was pre-selected by MEX as a motif that obeys some grammar.

There was an overlap of 849 motifs between the motif sets obtained by the two approaches. 1024 (55%) of MEX’s motifs were not discovered by the exhaustive approach (Figure 11A). These are mostly weaker motifs, that could not be identified within a very noisy background. MEX provides an enrichment in signal which relaxes the p-value thresholds set by FDR, allowing for weaker motifs to be detected as significant. In addition, MEX extracted sequence motifs of length up to 19 nucleotides. 57 of the unique MEX motifs were longer than 11 bases, and thus were not examined by the exhaustive approach (limited to lengths 7-11) (Figure 11B).

Motifs that were detected by the exhaustive approach, but not by MEX most likely

do not obey the inherent position dependencies, selected for by MEX (Figure 11C). It has been reported that some, but not all functional TFBS display such position dependencies [117]. The relative success of MEX in identifying high scoring motifs suggests however that there are some syntactic rules that characterize functional binding sites. This thought is intriguing because it implies that we may be able to identify at least some of the TF binding sites based on their sequence context alone.

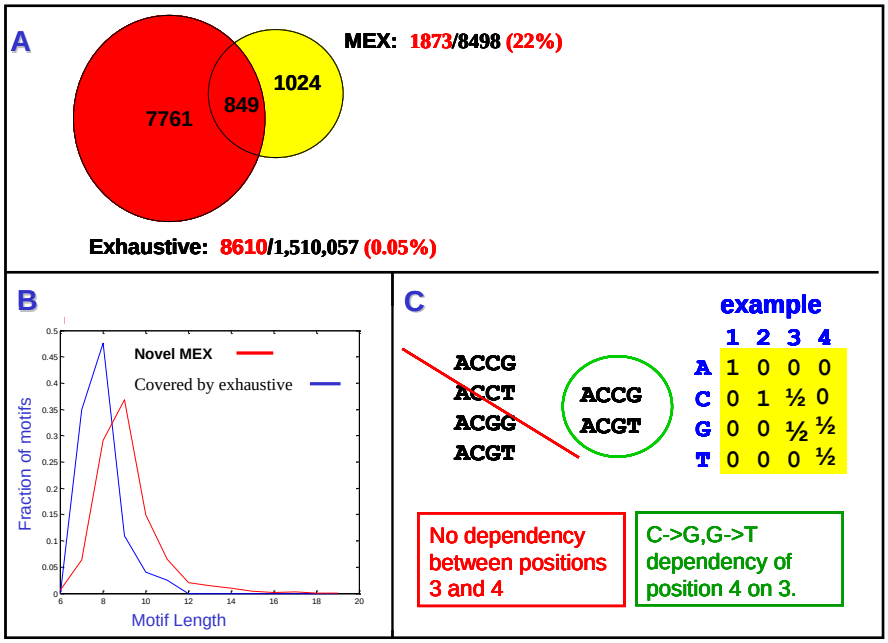


Figure 11: Comparison between motif sets obtained by the exhaustive approach versus the syntax based approach (MEX). A. Overlap between motif sets obtained by the exhaustive dictionary method and by MEX. B. Length distribution of MEX motifs that are covered by the exhaustive search versus novel MEX motifs. MEX has a clear advantage in identifying longer motifs, for which the exhaustive search is computationally too demanding. There are also motifs in the length range of 7-11 which scored significantly in the MEX based dictionary, and not in the exhaustive dictionary. These motifs are weaker and when embedded in a very noisy background (of all possible k-mers), their score is not high enough to pass the threshold set by FDR.C. Motifs in which there is no clear dependency between positions will be missed by MEX. MEX learns simple syntax rules from the promoter sequences and searches for motifs that obey these rules.

2.1.1.4.2 Analysis of biologically significant motifs extracted by MEX

The most significant set of motifs extracted by MEX was further divided into subsets, based on two criteria – the motif's DNA sequence and the biological conditions in which each motif appears to operate (as determined by significant coherence of its target genes in these conditions). This was done by Liat Segal, a joint student of Prof. David Horn and Prof. Eytan Ruppin, our collaborators in Tel Aviv University [29]. Clustering a selected set of 694 motifs (which have both passed FDR of 0.1 and were assigned an EC score with a p-value of 0.001 or lower in at least one

of the examined biological conditions) yielded 20 motif clusters, 14 of which correspond to known PWMs. Interestingly some of these clusters are very similar in sequence (correspond to the same known PWM), yet appear to govern a different set of biological conditions. These motif clusters may represent cases in which two TFs serving distinct biological functions, recognize seemingly identical consensus motifs [118]. Alternatively, such clusters may represent binding sites of the same TF, which yield different regulatory outputs due to slight variations in binding site sequence. This may be a result of different affinities of the TF to the different sites. There are also reported cases in which binding site sequence variations cause the bound TF to adopt different conformations, directing interactions with specific cofactors and resulting in different expression responses. Such TFs have been termed alosteric regulators [119].

One example of such a case is given in four motif clusters that correspond to the recognition sites of two related yeast TF complexes. Both complexes are known to regulate the G1/S transition during cell cycle; The first complex MBF (MCB-binding factor) consists of two protein components Mbp1 and Swi6 and recognizes a site called MCB ('ACGCGT'). The second complex SBF consists of Swi4 and swi6 and binds to a site called SCB ('C(A/G)CGAAA'). The four motif clusters which correspond to MCB and SCB are shown in Figure 12. This figure displays both the motif sequences contained within each cluster, and the biological conditions in which these motifs govern coherent expression; Cluster M1 contains sequence motifs whose common core 'ACGCGA' corresponds to the known SCB consensus site. Cluster M2 contains motifs whose core sequence 'ACGCGT' is the known MCB consensus. Both motif clusters appear to govern coherent expression through cell cycle and in response to various environmental stresses The M1 and M2 clusters provide support for a known difference in binding preferences between MBF and SBF, and proof of concept for our ability to distinguish between two highly similar motifs. The motifs belonging to clusters M3 and M4 contain core sequences, which are slight variations on these known sites. Interestingly these two motif clusters govern a distinct set of biological conditions: M3 governs primarily stress responses and M4 – cell cycle. For these two motif clusters, the change in biological function may be attributed to specific nucleotide changes in the motif core.

Because both MCB and SCB are bound by protein complexes, one may hypothesize that the differences in the biological conditions regulated by the different

clusters may result from different co-factors interacting with the DNA-binding protein in each case. Additional motif clusters are described in our recent paper presented in ISMB 2007 [29].

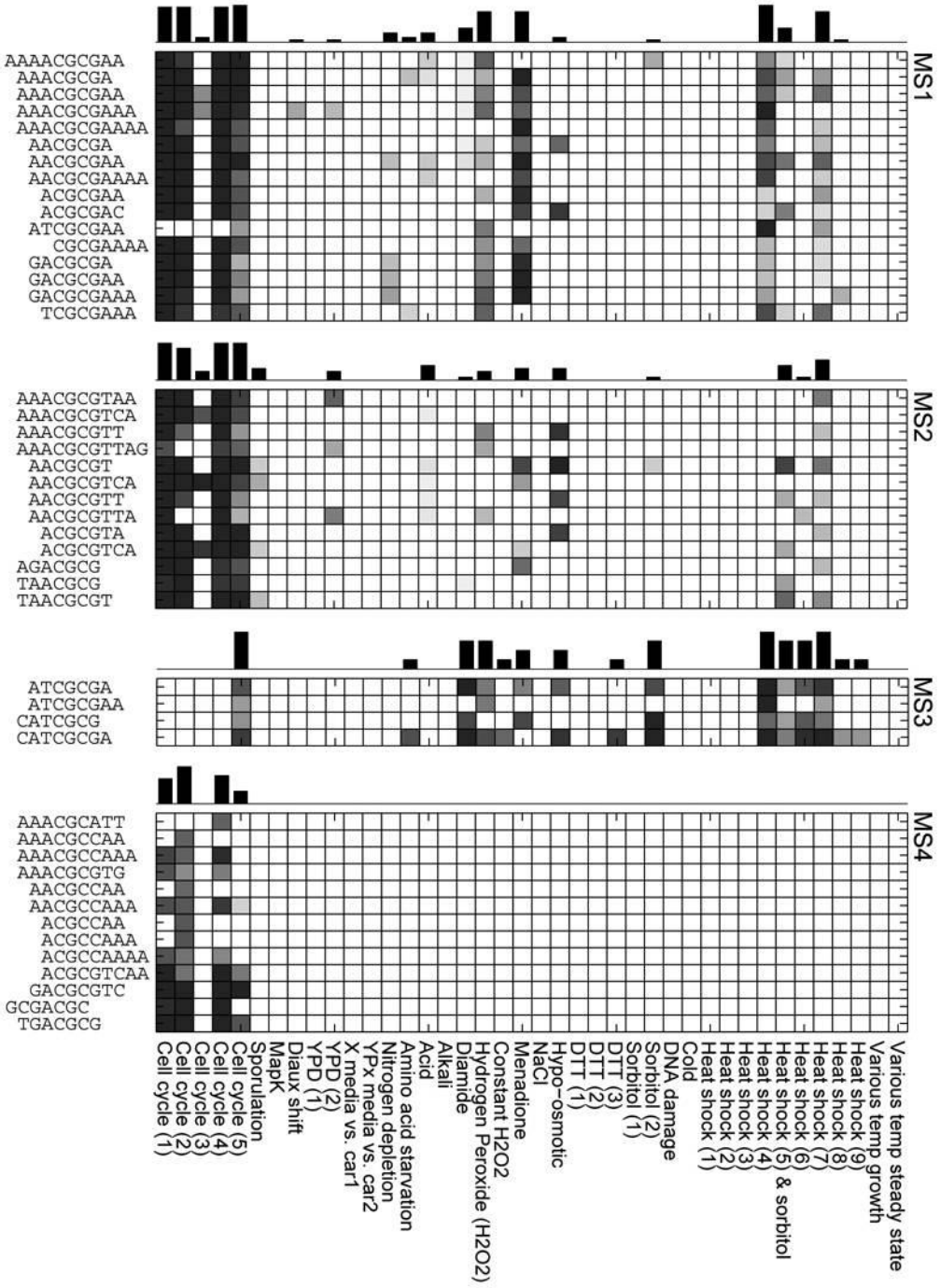


Figure 12: Four clusters of MEX extracted sequence motifs correspond to known MCB and SCB binding sites. For each motif cluster, a matrix of motif by condition, displays the significance (in terms of p-value) of the EC score of each sequence motif at each biological condition. Significant p-values are represented by dark colors, with a grayscale proportional to -log(p) (white implies p>0.001). The bars indicate the percentage of sequence motifs that had a significant EC score in each condition. M1- corresponds to SCB, M2 to MCB, M3 and M4 – related sequences that govern distinct biological conditions.

2.1.2 Candida albicans stress response dictionaries

In order to conduct comparative genomics of binding sites in a related species, we chose to construct motif dictionaries for the yeast *Candida albicans*. The estimated evolutionary distance between *S. cerevisiae* and *C. albicans* is between 140-800 million years (MY). *C. albicans* is a human pathogen, it has the ability to switch between three morphologies; yeast, pseudo-hyphae and true hyphae, which probably contributes to its virulence. *C. albicans* has a diploid genome consisting of ~6,358 short open reading frames (ORFs), the average ORF is of 1,439 bp, and only 217 (3.4%) ORFs contain introns [120].

We constructed exhaustive motif dictionaries (k=7-11) for *C. albicans* in three stress response experiments Heat Shock, Osmotic Shock and Oxidative stress [121] (see methods section 3.1.1.3). The availability of the same stress conditions in both *C. albicans* and *S. cerevisiae* allowed us to gain insight into the common features of regulation in both organisms. Our exhaustive scoring of all possible k-mers revealed the following picture: The majority of k-mers appear non-functional in both species, some k-mers score significantly in *S. cerevisiae*, yet seem to be non-functional in *C. albicans* – these represent *S. cerevisiae* specific motifs, others score highly in *C. albicans* yet seem to be non functional in *S. cerevisiae* – representing *C. albicans* specific motifs. There is also a group of k-mers that score significantly in both species, and thus compose the core of the evolutionary conserved regulation. Figure 13 illustrates this division for all 8-mers in heat shock.

We chose to focus on the motifs that score highly in both organisms. We adapted the common notion of orthology, so that it will cover three levels: (i) Orthology on the motif level – namely the existence of the same motif in the sets of significantly high scoring motifs of both organisms. (ii) Orthology on the expression level, namely cases in which the shared regulatory motif brings about the same expression behavior in the same condition in the two species. (iii) Orthology on the gene set level, namely cases in which the sets of genes regulated by the shared motif in the two species are enriched with orthologous gene pairs.

On the first level, of the motif itself: The C. *albicans* osmotic shock dictionary contains 27 single strings that are roughly clustered into three major motifs (Figure 14). The clustering is based both on the motif sequence and on the expression profiles of the downstream genes. Only two of these motif clusters have a counterpart in *S.*

cerevisiae (PAC and RRPE). The same picture is true for the heat shock (85 motifs) and oxidative stress (13 motifs) dictionaries. In each of these experiments the motifs are clustered into three major regulators, two of which have a corresponding cluster in *S. cerevisiae*. We thus estimate that about 33% of *C. albicans* heat shock regulators are species specific, while the rest are common to different yeast species. PAC and RRPE – two well known regulators of ribosomal RNA (rRNA) processing [23], appear significant in stress response in both *C. albicans* and *S. cerevisiae*. We conducted a detailed study of the three levels of orthology for the PAC motif. The consensus sequence of PAC is: (A/T/C)(G/T/C)CTCATC(G/T/A)(C/A/T). Sequences corresponding to this consensus score highly in both organisms. PAC is known to regulate mostly the transcription of rRNA and of rRNA processing proteins. In response to stress it shuts down the expression of its targets causing a general halt in protein synthesis. As can be seen in Figure 15, both *S. cerevisiae* genes that contain a PAC binding site in their promoters and *C. albicans* genes that contain a PAC site in their promoter are down regulated as an initial response to stress. Down regulation occurs within the first 10 minutes following heat shock. We took all the *C. albicans* genes in the main PAC responding cluster (Figure 15) and compared them to the genes in the corresponding *S. cerevisiae* cluster. 70% of the Candida PAC regulated genes have a reciprocal-best-blast *S. cerevisiae* ortholog, that is also regulated by (the *S. cerevisiae*) PAC, Namely orthology is observed here on all three levels: The same motif governs the same response in the same set of genes.

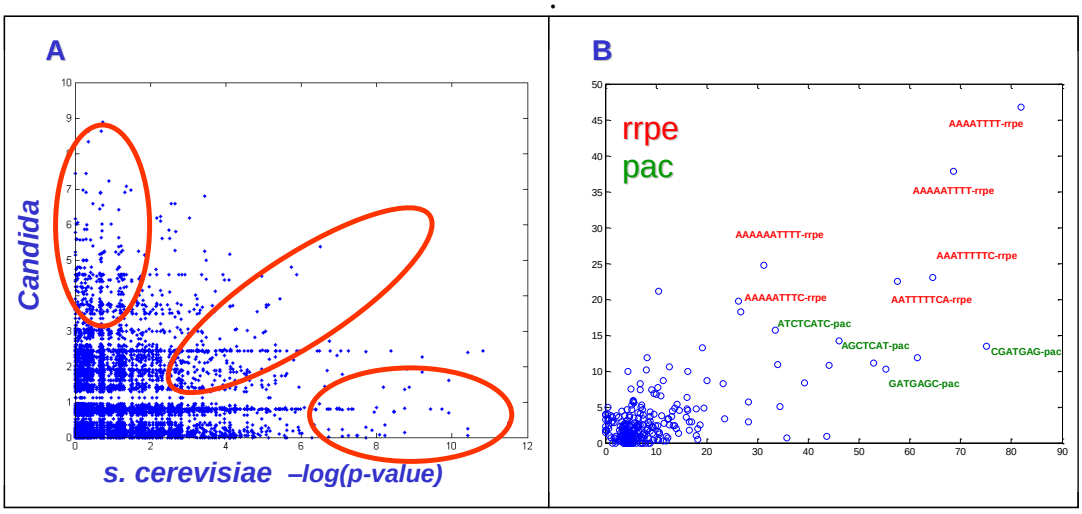


Figure 13: A. Comparison of the contribution of the same 8-mer to the control of expression in response to heat shock in two diverged yeast species. All 8-mers were scored in both *S. cerevisiae* and *C. albicans*. The score of each 8-mer (-log p-value) in *C. albicans* was plotted against its score in *S. cerevisiae*. Three types of 8-mers are marked in red circles: Candida specific, *S. cerevisiae* specific and evolutionary conserved. B. Zooming in on the evolutionary conserved motifs: PAC and RRPE – two

well known regulators of rRNA processing, appear significant in stress response in both yeast species [23].

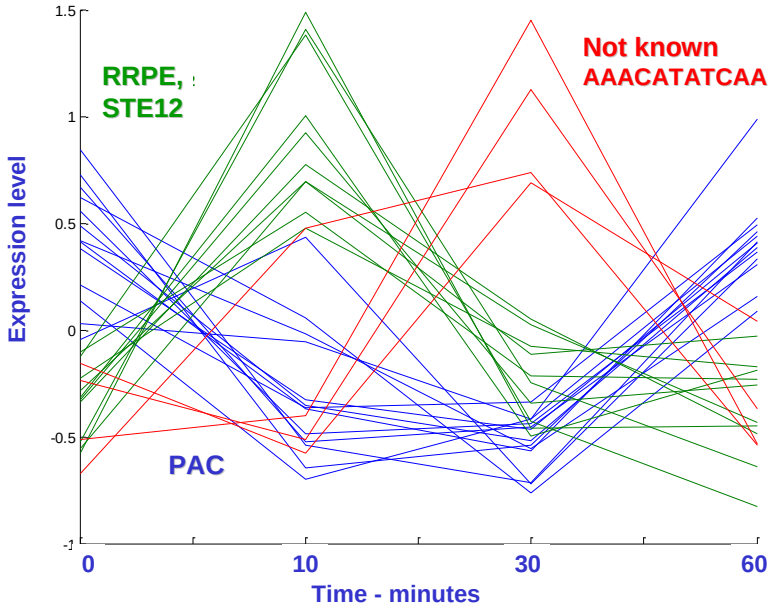


Figure 14: Semantic description of the *C. albicans* osmotic stress dictionary. Each line represents the mean expression profiles of all the genes regulated by one motif. Three main expression profiles are dictated by the motifs – two have a corresponding regulator in *S. cerevisiae*, and one appears to be *C. albicans* specific.

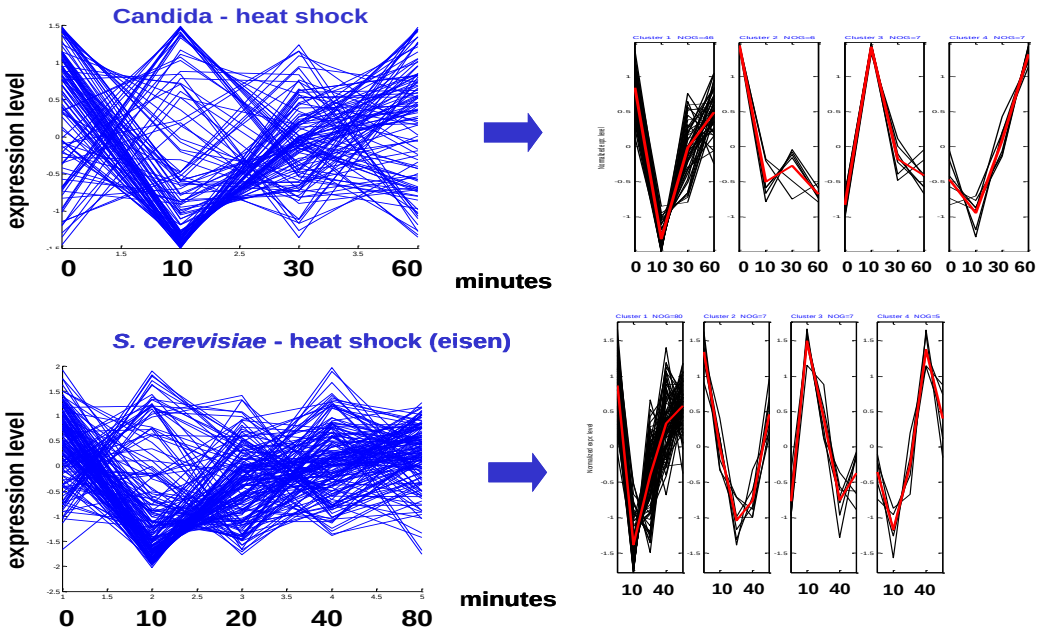


Figure 15: Expression of PAC regulated genes following heat shock. Most genes (main cluster) containing PAC in their promoters are down regulated as an initial response to stress. Down regulation occurs within the first 10 minutes following heat shock in both organisms. Among the genes in the principal response cluster, 70% of the Candida PAC regulated genes have a reciprocal-best-blast S. *cerevisiae* ortholog**.**

2.1.3 *Human cell cycle dictionary*

To assess feasibility of our method in higher organisms, we constructed dictionaries for three cell cycle experiments in human HeLa cells (differing in the cell cycle synchronizations strategy) [122]. Although none of the motifs passed our FDR requirement, a set of 90 motifs scored highly (p-value < 0.05) in all three experiments. Therefore we regarded this set as the human cell cycle dictionary. This set corresponds to 46 known TRANSFAC [123] motifs, some of which are known as cell cycle regulators (for instance E2F,EGR, NFY). The majority of the motifs in this set govern a very tight regulation of the G2->M checkpoint, as illustrated in Figure 16.

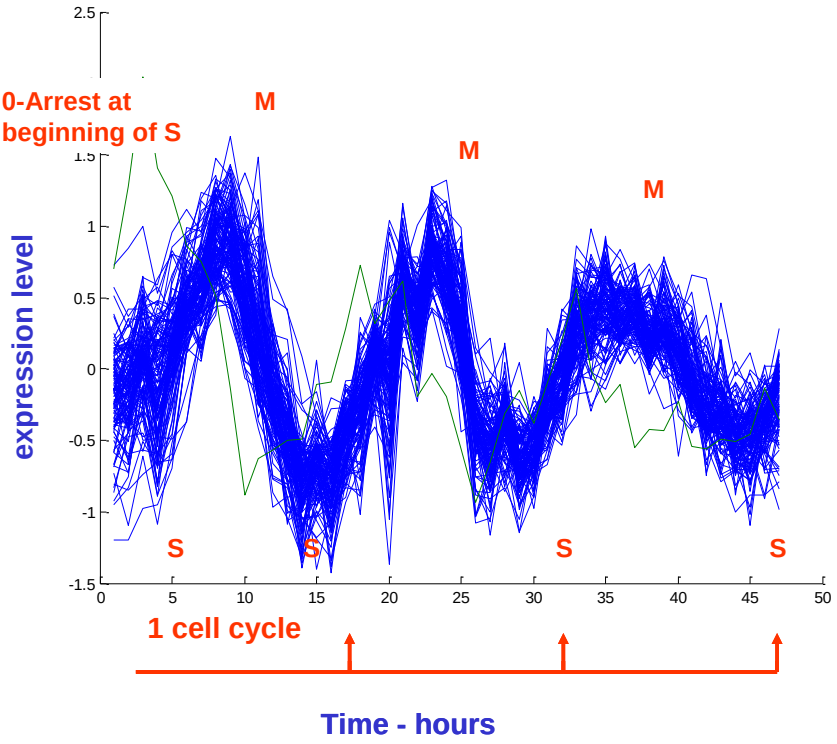


Figure 16: Each line represents the mean expression profiles of all the genes regulated by one motif. A very tight regulation of the G2->M transition is observed. A complete human cell-cycle is 14-16 hours.

2.2 Functional Characterization of Binding Site Variations

Aim

Because TFs typically bind to short degenerate sequences, highly similar sites within the same genome are in some cases recognized by the same TF, whereas in others serve as targets for distinct TFs. This is also observed in the genomes of related species, where slight changes in binding site sequence, occurring throughout evolution, may either maintain the specificity of the site to the original TF or alternatively lead to its loss or create a site targeted by a different TF [20, 38]. The desire to distinguish between 'neutral' binding site variations, which do not change the recognition range of the site, and 'functional' variations, which may affect gene expression by altering protein-DNA interactions, lays at the heart of this work. Such a distinction may allow the prediction of regulatory site variations, which have the potential to cause diseases through altering gene expression.

Major findings and conclusions

An analysis of the yeast binding site dictionaries (described in section 2.1.1) revealed that binding sites with similar syntax may yield different expression patterns of the regulated genes, while binding sites with different syntax may dictate similar expression patterns.

We have developed computational measures to estimate the functional consequence of substituting a single position within a binding site. Applying these measures to binding sites of known TFs we were able to make predictions that were in line with published experimental evidence and with structural data on DNA-protein interactions. This suggests that our methods could complement and in some cases replace time consuming mutation experiments. We further accumulated statistics from multiple substitutions across various binding sites in an attempt to deduce general properties that characterize nucleotide substitutions that are more likely to alter expression. We found that in the yeast genome substitutions that abolish a G or a C tend to have a more severe outcome than substitutions that abolish an A or a T. This may be specific to the yeast genome which has a low GC content, and thus G and C may be important for specificity. We found additional factors that are correlated with the severity of a substitution, such as the Information Content (IC) of the substituted position. These factors can be further integrated to make trustful predictions.

The present work sets the foundations for obtaining a larger goal: predicting the phenotypic effects of regulatory motif variations within human promoters. Such predictions will facilitate the prioritization of human SNPs residing within TFBS, according to their disease-causing potential.

2.2.1 Exploiting the yeast motif dictionaries to predict the outcome of a binding site substitution

2.2.1.1 Quantitative measures for the severity of a substitution

In the process of producing the motif dictionaries, we assigned EC scores, corresponding p-values and likely expression effects to all k-mers residing in yeast promoters, regardless of whether they were ultimately included in the dictionary. This provided a unique source of information for addressing our research question; By comparing the EC scores and the induced expression profiles of k-mers differing in a single position we could predict the outcome of a substitution that transforms one k-mer into the other. Three main scenarios were observed (i) Two k-mers differing at a single position both belong to the dictionary (passed FDR) and regulate genes with a similar expression profile. This implies that the k-mers are recognized by the same TF, and a substitution from one to the other is thus expected to have a very mild effect (Figure 18, green arrows). (ii) The two k-mers belong to the dictionary but regulate genes with a different expression profile. This may imply that they are recognized by different TFs, thus a substitution from one k-mer to the other is predicted to cause binding site switching (Figure 18, blue arrows). (iii) One k-mer belongs to the dictionary whereas the other did not pass the FDR constraint. This implies that substituting the former to the latter may result in binding site loss without acquisition of a new site (Figure 18 red arrows).

We devised three quantitative measures in order to compare the regulatory functions of two k-mers: (1) ΔEC – the difference in EC scores between the set of genes containing k-mer a in their promoters and the set of genes containing k-mer b in their promoters. (2) ΔPV – the difference in the logarithm of p-values assigned to the EC scores of the two gene sets (3) Distance in the mean expression profiles of the two gene sets across a given time series experiment. Each k-mer is represented by the mean expression profile of all genes containing it in their promoters (methods section 3.1.2.2). We measure the distance between the vectors representing the mean

expression profiles of the two gene sets (calculated as 1-corelation coefficient of the two vectors).

2.2.1.2 The ‘motif landscape analysis’ tool

We have developed a computational tool termed ‘motif landscape analysis’ [21] that employs our comprehensive motif dataset in order to systematically predict the outcome of all possible single nucleotide substitutions within a given motif. For a motif of length L this tool examines all 3*L k-mers that are obtained by substituting the motif at each single position. For each such k-mer it computes the three described measures ΔEC, ΔPV and distance in expression profiles between genes containing it in their promoters and genes containing the consensus motif. The results are graphically displayed (Figure 18 right panel).

Applying this tool to the consensus of the yeast sporulation factor Ndt80 (Figure 18 right panel) using the *S. cerevisiae* sporulation expression data, predicted that two out of the three possible substitutions in the second position will not affect expression whereas an A->G substitution at the same position will result in an effect that is not severe (see Figure 18 legend for details). When averaging over all possible single nucleotide substitutions, the second position appears to be the most tolerant towards substitutions (mean ΔEC 0.089, mean ΔPV 0.9677, mean expression distance 0.0358) and the seventh position - the most sensitive (mean ΔEC 0.3381, mean ΔPV 4.1784 , mean expression distance 0.7715) (Figure 19). One possible reason for such a marked difference between the tolerance of different positions within the same motif to substitutions may be that the binding transcription factor forms different contacts with the DNA at each of the positions. Particularly, we may expect the positions that form tight contact to be less permissive to substitutions. Indeed, our results are in good agreement with the structural data of Ndt80 bound to its DNA target [124]; the second ‘permissive’ motif position is the only position which does not form a direct contact with the protein (Figure 17). But do these differences affect TF function? Reassuringly, these results are also supported by recently published *in vivo* reporter expression experiments and *in vitro* binding essays of Ndt80 mutants [125]. This experiment represent the 'wet' analog to our computational experiment – each of the nucleotide positions in Ndt80 was replaced with all possible 3 alternatives. These results too showed that the second position is the most permissive to substitutions, and that, as predicted by us, G is the only nucleotide that when placed at this position

weakens binding affinity and reduces expression level of the reporter gene [125]. This implies that our method can complement and predict the outcome of time consuming mutation experiments.

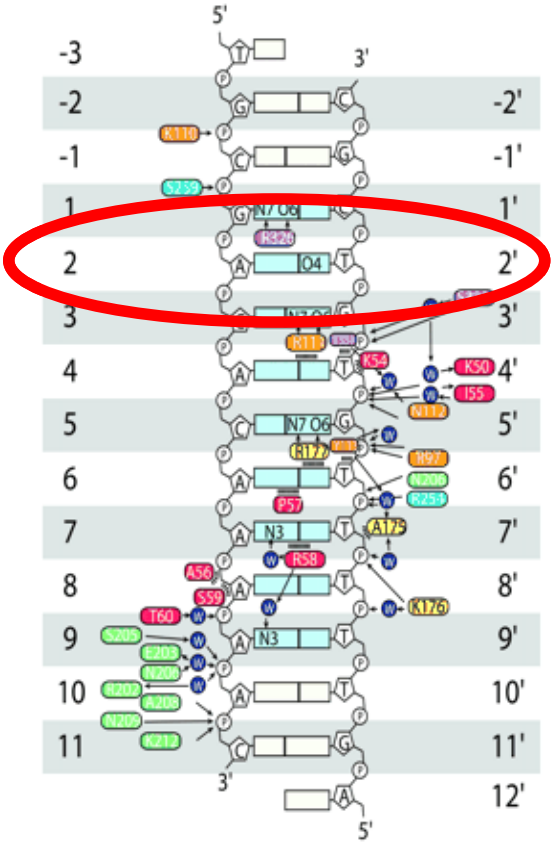


Figure 17: Schematic representation of Ndt80 bound to DNA [124]. The consensus binding site positions are highlighted in light blue. Protein residues appear as colored ellipses. The second position, which was predicted by our method to be the most permissive, does not form a direct contact with the protein.

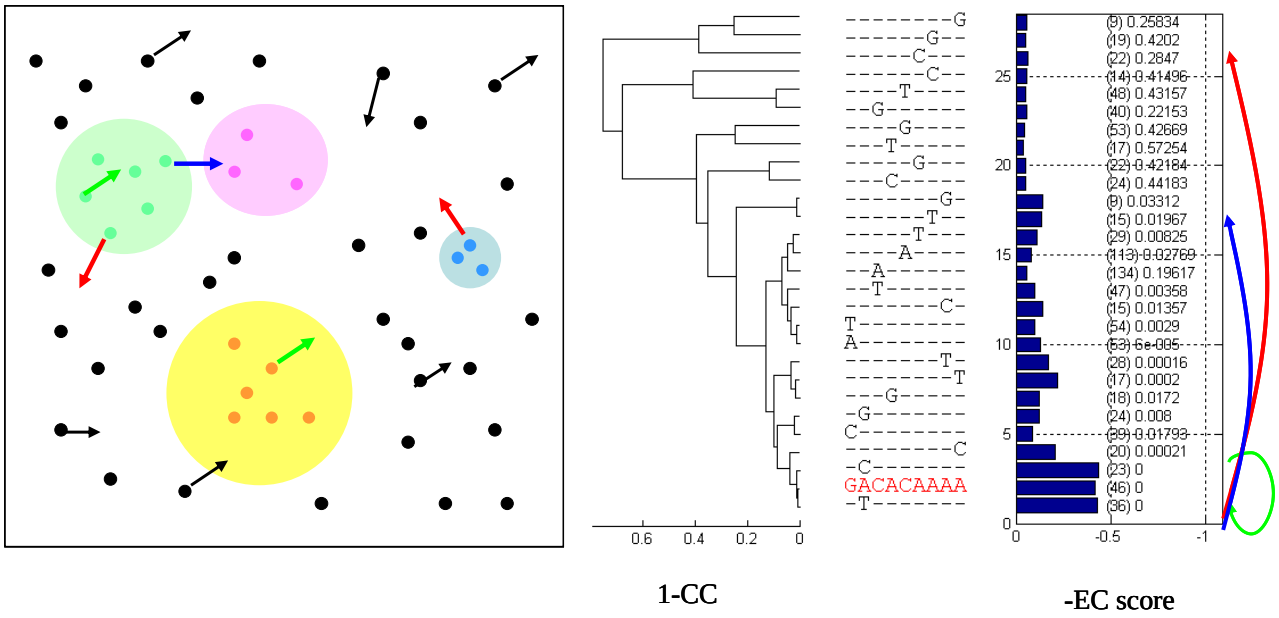


Figure 18: Possible outcomes of binding site substitutions: Left panel a cartoon depicting possible effects of mutations in regulatory motifs. Points represent promoter elements and discs represent transcription factor recognition ranges. Points that are included within the disc of a given TF represent promoter elements that are bound by the TF. Arrows illustrate the result of single nucleotide substitutions within a promoter element. Such a substitution, can cause binding site loss (red arrows), a change in affinity to the same TF (green arrows), or binding site switching - creation of a binding site with higher affinity to a different TF (blue arrows). The right panel illustrates the detection of the same outcomes using our motif landscape analysis tool (as described in detail in [21]). This display captures the effects of single nucleotide substitutions of a given motif on the expression profiles of the downstream genes. The analyzed motif is the yeast Ndt80 sporulation factor (wild type motif marked in red). The dendrogram on the left part of the display shows the similarity in mean expression profiles between gene sets baring variations of the motif in their promoters. The right side of the display shows the similarity within sets of genes that contain the same motif variation in their promoters, as measured by the EC score. The numbers in parentheses correspond to the gene set sizes and the numbers next to them to p-values on the EC scores. The middle section displays the sequence of the motif variation studied in the corresponding row (with a '-' indicating same nucleotide as the wild type motif). A substitution, that is in the recognition range of the same TF, is expected to maintain a high EC score and a similar expression profile (green arrow), A substitution that causes binding site loss, is expected to be recognized by both loss of coherence and a change in the mean expression profile (red arrow). A substitution that creates a new motif, that is in the recognition range of a different TF, is expected to maintain high expression coherence, while altering the mean expression profile (blue arrow). The second motif position appears relatively tolerant to substitutions, 2 out of the 3 possible single nucleotide substitutions of this position do not alter TF recognition (green substitutions). This observation is supported by the recently published structural data of Ndt80 bound to DNA [124]. The second motif position does not form a contact with the protein.

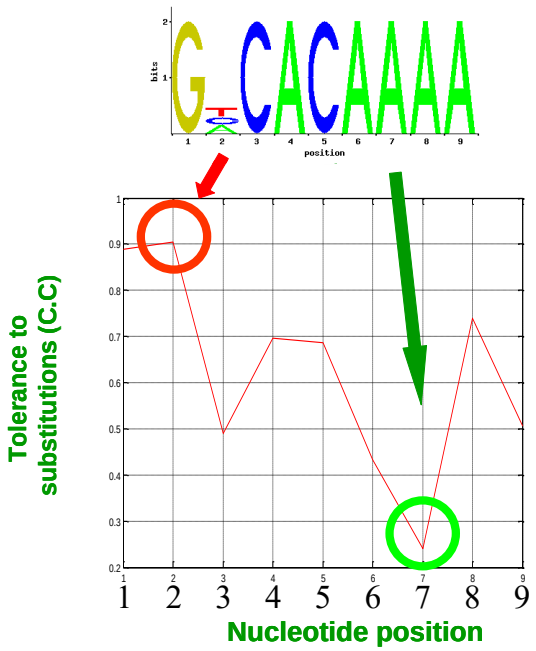


Figure 19: The averaged tolerance to substitution for each nucleotide position within the Ndt80 motif was defined as the averaged correlation coefficient between the averaged expression profiles of the genes that have a perfect match to the consensus motif and the averaged expression profiles of the genes that have each of the three possible substitutions relative to the consensus in that position

2.2.1.3 Differentiating between binding site switching and binding site loss

As described above, our measures can differentiate between cases in which a binding site is lost (observed as loss of expression coherence) and cases in which a site with higher affinity to a different TF is created (observed as preservation of significant expression coherence, along with a change in expression pattern). We illustrate this using the pair of TFs Ndt80 and Sum1, which are known to recognize overlapping binding sites, yet display distinct sequence preferences [125]. Ndt80 is the primary transcriptional activator of middle sporulation genes, whereas Sum1 is a transcriptional repressor of the same genes during mitosis and early sporulation. Both TFs recognize variations of a site termed middle sporulation element (MSE), whose consensus sequence is GNCRCAAAW. Through a combination of *in vivo* reporter expression essays and *in vitro* binding essays of Ndt80 and Sum1 mutants, Pierce et al. defined the specific binding preferences of these two TFs [125]. They found that while positions 3-5 of the MSE are important for binding of both Ndt80 and Sum1, there is a difference in binding preferences at positions 6-7. For these positions, Ndt80 requires strictly an A, whereas Sum1 binds equally to an A and to a T. Indeed our landscape analysis (Figure 18, right panel) shows that mutating position 6 from A to T

results in a change in expression profile, yet coherence remains high, p-value 0.0083. This may be explained by binding site switching from Ndt80 to Sum1. Transitions of the same position into C or G result in binding site loss (p-values 0.4 and 0.3 respectfully). The same applies for position 7, in which transition from A to T maintains a relatively significant EC score (p-value 0.019), whereas substitutions to both C and G lead to complete loss of coherence (p-values 0.4 for both). This position also scored as the most sensitive to mutations – any change will abolish the Ndt80 site, either by switching or complete loss.

2.2.2 Deducing general properties of expression-altering substitutions

Encouraged by our ability to predict the effects of binding site substitutions within a single motif, we attempted to generalize these predictions in order to define universal properties of substitutions that alter gene expression. We used the three measures described above to assess the severity of a substitution from base i to base j in a regulatory motif. Namely: change in EC score, change in the EC p-value and change in mean expression profiles of genes assigned to a motif with nucleotide i versus genes assigned to a variation on the same motif with nucleotide j at the substituted position. This time, instead of analyzing a single motif we accumulated statistics from substitutions of different positions across multiple binding sites. The premise to be tested here is that some universal preferences for particular substitutions exist and that accumulation of statistics from all motifs should reveal them. The alternative to this possibility is that in every motif different nucleotide substitutions are tolerated, and accumulated statistics on all motifs should not reveal a signal. There are twelve possible single nucleotide substitutions from base i to base j (when i can be A,C,G or T, and $j \neq i$). Each severity measure was averaged over all substitutions of the type $n_i \rightarrow n_j$ in any possible motif. The motifs used for this analysis were 339 dictionary motifs that correspond to known TFBS from Harbison's set [8] (see methods 3.2.1). Selecting motifs that match a published set increases the likelihood that the 'wild type' motif is indeed biologically functional. All together we analyzed 2,881 motif positions; typically ~600 (462-745 depending on the identity of i and j in the substitution $n_i \rightarrow n_j$) data points were used to generate each of the twelve substitution type 'penalties'. This elaborate statistics constitute the strength of the method.

Our first question was whether there were substitution types that are more radical than others (in analogy to amino acid substitutions where there are conservative changes that maintain the chemical properties of the residue versus radical changes that form a residue with different characteristics). Interestingly, although there was no single substitution type that appeared more radical than others, there were systematically higher penalties for substitutions that abolished a C or a G in the consensus versus substitutions that abolished an A or a T (Figure 20). Because the yeast promoters are AT rich (64%), this result may suggest that C and G are the nucleotides that confer most of the specificity of a motif to its target, and thus their substitution bears a greater effect on the motif's function, compared to mutations in As and Ts. Indeed we have shown (Figure 4), that our core dataset motifs have a GC content that is significantly higher than that of the background promoter, and is comparable to the GC content of known motifs such as the Harbison set.

This raises a prediction that in other genomes with different promoter GC contents, the penalties might be different, reflecting loss of information content with the elimination of different nucleotides.

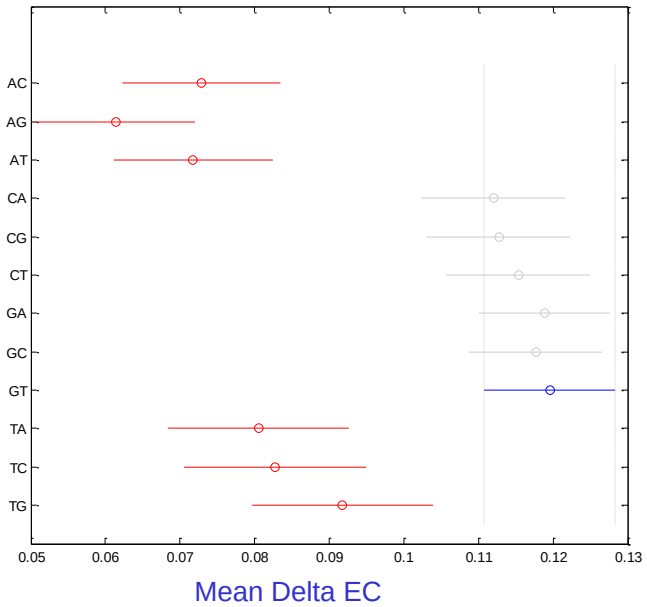


Figure 20: Effects of each of the twelve possible single nucleotide substitutions on expression. The data was accumulated from all (2,881) possible substitutions of each type in a dataset of 339 highly scoring k-mers that correspond to known Harbison PWMs. The 'severity measure' applied is mean delta EC, thus high values correspond to severe substitutions. A clear trend is seen whereby substitutions that abolish an A or a T are less severe than substitutions that abolish a C or a G

2.2.3 The information content of the substituted position

The degree of conservation of the substituted position in the PWM may also affect the severity of the phenotype. This may be analogous to the situation that was shown in protein coding SNPs – substitutions within conserved amino acid positions are more likely to be detrimental [43, 44]. A binding site position is said to be 'conserved' if different instances of this site, which are present in the same genome, tend to have the same nucleotide at the given position. Substitutions of highly conserved positions are expected to have a more dramatic effect on expression compared to substitutions of positions with low conservation. To test this hypothesis we analyzed high scoring k-mers from our dataset which correspond to known Harbison PWMs. For these motifs both the expression measures obtained in the process of creating our dataset (EC, p-value, expression profiles) and the conservation, captured as information content (IC) (methods section 3.2.2) of all PWM positions are available. We could thus assess the correlation between the IC of a position and its sensitivity to substitution based on the previously described severity measures. Indeed a significant correlation exists between the mean expression distance and the IC of a position (table 2). The mean expression distance is also highly correlated to our other two expression based measures mean ΔEC and mean ΔPV. Thus two measures, the identity of the substituted nucleotides, and the conservation of the substituted position serve as good predictors to the effect of the substitution on the expression of the regulated gene.

	Mean ΔEC	Mean ΔPV	Mean Exp Dist	Position IC
Mean ΔEC	1			
Mean ΔPV	0.5402 6.12e-142	1		
Mean Exp Dist	0.3505 4.34e-055	0.3053 1.41e-041	1	
Position IC	0.0827 3.47e-004	0.0531 0.0217	0.1252 5.7785e-008	1

Table 2: Correlations between the three expression measures mean ΔEC, mean ΔPV, mean expression distance and the IC of a PWM position. Data was accumulated for 1,867 positions. In each table cell, the first number is the correlation and the second number is the p-value on this correlation. The different expression measures are highly correlated. There is a correlation between the measure Mean Expression distance and the information content of a position.

It is interesting to point out that the correlation between the two measures mean ΔEC and mean expression distance is significant yet low. The dot plot of mean ΔEC versus mean expression distance (Figure 21A) can be roughly divided into the three cases illustrated in Figure 18: (i) Low ΔEC and low expression distance – these are substitutions that leave the motif in the domain of the same TF (green arrow in Figure 18) (ii) Low ΔEC and high expression distance – these are substitutions resulting in binding site switching (the creation of a site with higher affinity to a different TF, blue arrow). (iii) High ΔEC and high expression distance – these are substitutions that cause binding site loss (red arrow)

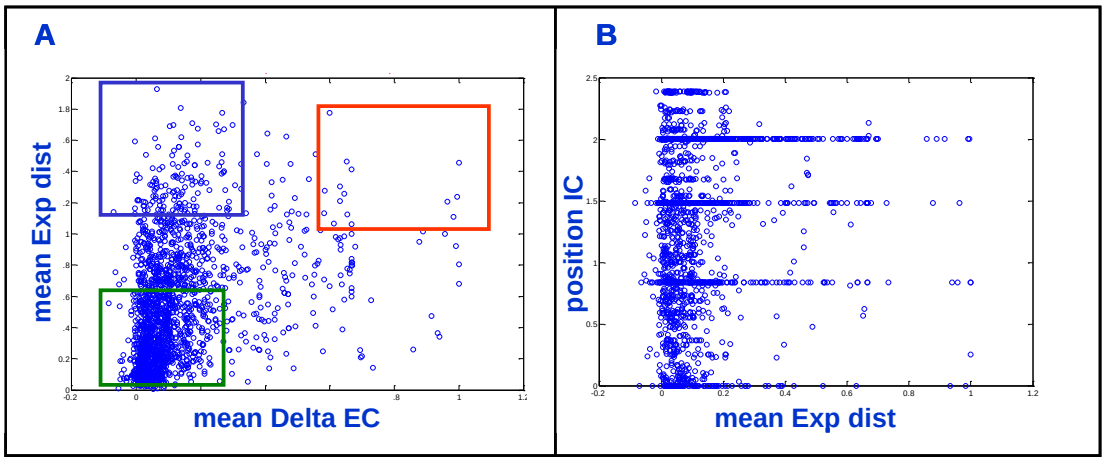


Figure 21: The measure mean expression distance is correlated both to the mean delta EC (left) and to the IC of the position in the motif PSSM (right). The correlations and p-values are listed in table 3.

Additional features including the evolutionary conservation of the substituted position and its vicinity to the protein in the DNA-protein co-crystal structure may also predict the outcome of the substitution. These features can be integrated to form a prioritization scheme that would allow the ranking of existing genome variations by their disease-causing potential. An additional application of the present approach may be in algorithms that assign PWMs to promoters (e.g. PRIMA [126]) as it should provide means to differently weigh mismatches between the PWM preferences and the promoter sequence based on expected effect on expression.

2.3 The Evolution of Interferon-α Promoters – an Adaptation to Varying Viral Threats?

Aim

Like the previously described project, the current project aims at characterizing promoter sequence variations that affect gene expression and gene function. While the previous project dealt with variations on the single nucleotide level, and introduced computational means to assess their effect on gene expression, this project deals with variations on a larger scale, namely changes in the composition of transcription factor binding sites residing within a promoter. Through characterizing the regulatory motif composition of human interferon-α promoters, we wished to study how this gene family evolved to respond to different viral stimuli. For this purpose, we combined a computational search for biologically significant regulatory motifs with accompanying expression experiments, carried out by our collaborators.

Interferons (IFNs) are generally-acting antiviral cytokines, induced at the level of transcription upon viral infection (their basal level is essentially zero). IFNs are divided into three types I, II and III. In human, type I IFNs include 13 members of the IFN-α family and three distant relatives (β, τ, ε).

The 13 human IFN-α genes share very similar coding regions and structures, operate through a common signaling pathway and induce the same set of genes. The coding regions alone can not explain the advantage of maintaining 13 supposedly equivalent genes in the genome. However, the IFN-α promoters are not as similar and may hold the key to the differential roles played by different member of the family. Diverse promoters may have evolved to allow for differential expression in different cell types or to adapt to different viral threats.

Certain viruses are known to prevent IFN induction by coding for proteins which bind IFN activators and prevent them from binding to the IFN promoter. It is thus possible that gene regulatory regions rapidly evolved to circumvent these viral attempts enabling the induction of some IFN-α species while others are neutralized. This theory may be tested through better understanding of the regulation of IFN gene transcription. Previous studies of IFN-α promoters characterized only motifs at the VRE (virus response element -176 to –131 relative to ATG) of each gene, except for IFN-α1 whose promoter was further characterized [127].

The aims of this study were (i) To characterize the promoter motif content of the 13 IFN-α promoter (ii) To study the expression pattern of IFN-α genes in response to different viral infections (iii) To link the two in order to elucidate the mechanisms regulating the differential expression of IFN-α genes upon viral infection. This includes identifying key promoter elements and their corresponding transcription factors (TFs). The promoter analysis was carried out by myself and the experimental procedures by my collaborator Roni Golan (laboratory of Prof. Menachem Rubinstein).

Major findings and conclusions

To characterize IFN-α promoters, we scanned them with a database of known TF binding sites and compared the regulatory motif content of all 78 (13*12/2) promoter pairs. In order to focus on regulatory motifs that are likely to be functional in the IFN-α promoters, we compiled a set of selected motifs which fulfill the following criteria: appear at a preferred location in the promoters of IFN-α genes (positionally biased), enriched in these promoters compared to the entire genome, and annotated as related to the control of immune related genes. We found that some of the IFN-α promoter pairs are highly similar in the content of these selected motifs despite overall sequence divergence. We hypothesized that these promoters respond similarly to viral stimulus, and may thus be used to study the promoter elements that mediate this response. We experimentally tested the response of four IFN-α promoters to induction with Newcastle disease virus (NDV), using a luciferase reporter system. Indeed we identified a pair of promoters α13 and α2 that were predicted by the computational analysis to share significant motifs, despite overall diverged promoter sequences, and are both highly induced by NDV infection. This promoter pair is currently under further investigation to elucidate the specific motifs that participate in its induction. In addition, we found promoter pairs that are completely diverged, with very low similarity in motif content. We predict that these promoters have evolved to respond to different viral stimuli. To test this we intend to experimentally test the expression of IFN-α genes upon exposure to a number of different viruses.

2.3.1 IFN-α promoter scan using a selected set of motifs

The 13 IFN-α promoters (of average length 1,065 bp) were scanned with the TRANSFAC database [123] that consists of 344 human TF binding sites represented by positional weight matrices (PWMs) (see methods section 3.3.1). Many of the PWMs in this dataset are rather degenerate and thus appear spuriously in multiple promoters, introducing false-positive hits. In order to reduce this noise, we used a combination of three different criteria that intend to select motifs that are likely to mediate the regulation of IFN-α genes : (i) Motifs bound by immune-related TFs. TFs were annotated as immune- related both automatically using functional enrichment of GO biological function terms, and manually based on a literature search (see methods section 3.3.2) . 146 binding sites passed this criterion.
(ii) Binding sites located at a preferred position relative to the TSS in the IFN-α promoters (see methods section 3.1.8). Such preference, known as positional bias, is often a hallmark of functional binding sites, because the function of many (though not all) sites depends on their location and on the distance between them and other cooperatively operating sites. 127 motifs appeared to be positionally biased in the promoters of IFN-α genes. (iii) Binding sites enriched in IFN-α promoters relative to all other human promoters (see methods section 3.3.3). 123 motifs passed this criteria.

Applying all three criteria resulted in a set of 45 PWMs corresponding to 35 unique TFs (the TRANSFAC database is redundant and some TFs are represented by more than one PWM). All binding sites known to reside in IFN-α promoters (e.g. IRFs, ISRE) were included in this set supporting the choice of criteria. We focused on this selected motif set in our analysis of IFN-α promoter content.

2.3.2 Comparisons of IFNA-α promoter motif content

To evaluate the similarity between different IFN-α promoters, we compared the motif content of all 78 IFN-α promoter pairs using the following score:
#common motifs/ min(#motifs in promoter1, #motifs in promoter2).
For example if the first promoter had 10 motifs, the second 8 motifs, and 4 motifs were common to both, the score of the pair was 4/8=0.5. We computed all pair-wise promoter similarities using the entire motif dataset (344 motifs) and once again using only the selected set of 45 motifs. The results are displayed as a scatter plot (Figure 22A.); Each pair of IFN-α promoters is represented by a dot whose location indicates

the similarity in selected motif content versus the similarity in all motif content. Three types of promoter pairs are observed (1) Highly similar promoter pairs (marked in red). These promoters share most of their overall sequence and thus also share many motifs. They score highly regardless of the choice of motif set. They are probably induced by the same stimuli (2) Diverged promoter pairs (marked in blue), these pairs have very few motifs in common regardless of the motif set examined. They probably respond to different stimuli. (3) Promoter pairs which are diverged in their overall sequence, but appear closer once focusing on the selected motifs (marked in green). These promoters have a small overlap in overall motif content, but when clearing the 'noise' by using only biologically informative motifs, their overlap increases significantly. We hypothesize that these promoters respond to similar stimuli via the conserved binding sites. Such promoters are thus excellent candidates for experimentally studying the mechanism of induction and the promoter elements necessary for it.

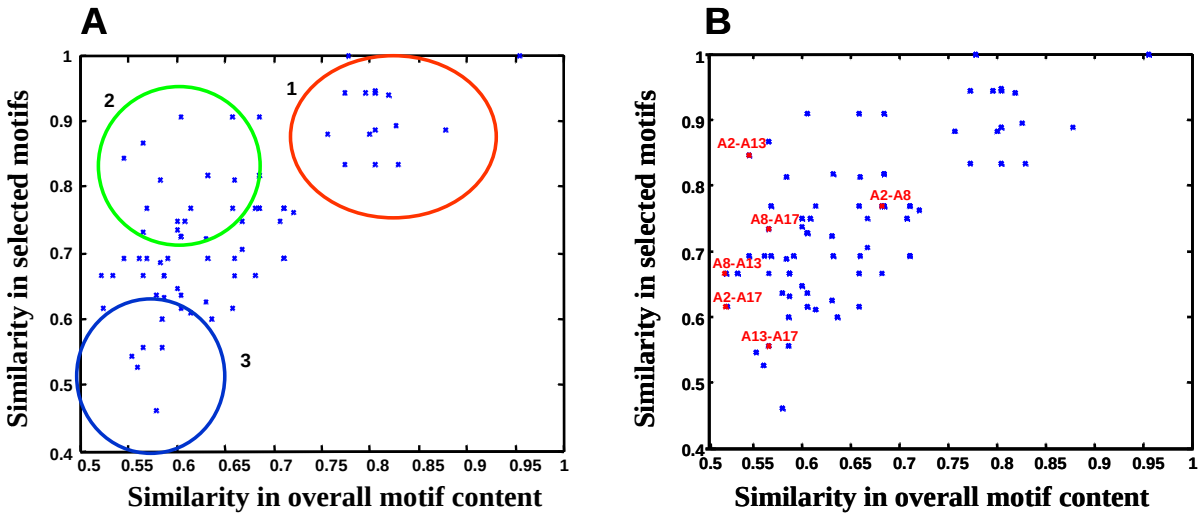
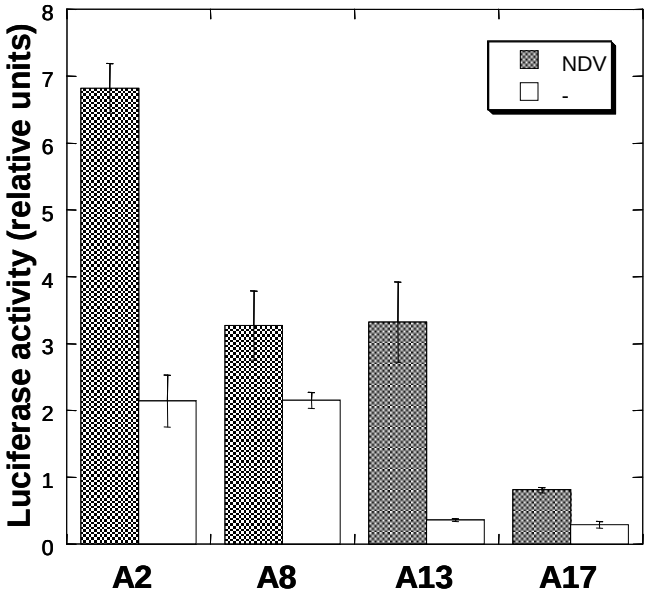


Figure 22: Pair-wise similarities of IFNA-α- promoter motif content. Each x represents a pair of IFNA- α promoters. The location within the dot-plot indicates the degree of similarity between the two promoters when comparing their overall motif content versus content of selected motifs. In both cases the similarity measure used was: #common motifs/min(#motifs in promoter1,#motifs in promoter2). (A) Three types of promoter pairs are observed: Highly similar promoter pairs (red), Promoter pairs which are diverged in their overall sequence, but similar in their selected motifs (green), and Diverged promoter pairs (blue). (B) The promoter pairs formed by the four experimentally essayed IFN-α promoters: A2, A8, A13 and A17 (shown in red). The promoters activity of these genes in response to Newcastle disease virus NDV was tested by a reporter gene assay.

2.3.3 Measurement of IFN-α promoter activity in response to viral stimulus

Our elaborate promoter analysis guided the selection of optimal genes for the subsequent experimental expression essays; Four genes were selected, such that all their corresponding promoter pairs form two types of relations: either diverged promoters, possibly indicating that the genes evolved to respond to distinct stimuli (these promoter pairs are within the blue circle in Figure 22A), or promoters, which maintained common binding sites despite overall divergence (within the green circle, Figure 22A). The latter may indicate a common regulation of gene induction and help in elucidating its components. The selected genes were α2, α8, α13 and α17. The relations (in sense of motif content) between their corresponding promoters are displayed in Figure 22B. The activity of these selected promoters in response to viral stimulus, was tested experimentally using a luciferase reporter system. All four IFN-α promoters were induced by Newcastle disease virus (NDV), but to a different extent (Figure 23). The strongest induction by NDV was achieved for the α13 promoter (9.3 fold relative to a non induced promoter). Second is the promoter of α2 with a 3.2 fold increase and third α17 induced by 2.8 fold. The α8 promoter was induced to the lowest extent 1.5 fold. When looking at the final absolute level of expression, α2 greatly exceeds all the rest. The promoters of α13 and α2 were predicted by the computational analysis to share significant motifs, despite overall diverged promoter sequences (appear in the green circle in Figure 22A). Their high induction levels in response to viral stimulus support this prediction. This pair of promoters is currently under further investigation to elucidate the motifs participating in their induction. This is done by mutating common promoter elements, based on the computational analysis, and repeating the luciferase viral induction essay with the altered promoters. We search for promoter elements which are essential for viral induction, such that their deletion drastically reduces gene expression in response to the virus. In addition, in order to identify IFN-α promoters that evolved to respond to different viral threats, we intend to examine promoter activity in response to a different stimulus such as Sendai virus (SV).

Figure 23 : Induction of human IFN-α promoters by NDV. KG-1 cells were transfected with luciferase vector, pGL3-basic containing promoter (-1 to -1065 bp relative to ATG) of either IFN-α2, IFN-α8, IFN-α13 or IFN-α17 . All transfections were done in the presence of renilla expressing vector (pRL-TK), used as a standard for trnsfection efficiency. Transfected cells were infected with NDV 24 hours post-transfection, and the levels of luciferase activity were determined 19 hours later, and normalized to the levels of Renilla activity (three repeats for each treatment).



2.4 Antisense Transcription – a Regulated Mechanism for the Control of Gene Expression

In this section we cover two subjects, both related to genome-wide antisense transcription (i) A computational work in which we detected and studied genome-wide *trans* targets of human NATs (section 2.4.1). (ii) A summary of the concept presented in our recently published paper [107] (section 2.4.2).

2.4.1 In search for human trans encoded antisense

Aim

NATs are conventionally divided into *cis*-NATs (transcribed from the same genomic locus) and *trans*-NATs (transcribed from separate loci). Most genome-wide attempts to estimate the extent of the NAT phenomenon, were limited to *cis*-NATs, implying that the prevalence of NATs may in fact be broader. These studies aligned full-length cDNAs and ESTs to the corresponding genome and identified overlapping transcripts on opposite strands as *cis*-NAT pairs.

But are *cis* and *trans*-NATs in essence two separate phenomena? The aim of the current research was to assess whether *cis*-encoded antisense, can also target transcripts in *trans*. To test this we used a previously published set of human *cis*-NAT pairs [67] and conducted a BLAST [128] search (confining our search to the opposite strand) against all human mRNAs. Our goal was not only to find potential *trans*-encoded targets, but also to study the common properties of all targets of a single antisense transcript, in order to learn about the biological processes that may be regulated by NATs.

Major findings and conclusions

We have performed a genome-wide search for putative *trans* encoded targets of human NATs, which have been previously reported to act in *cis*. For 39% of our NAT queries, we found at least one *trans* encoded target, in addition to the known *cis*-target. This finding expands the common definition of antisense function, by demonstrating that the same antisense transcript may have both *cis* and *trans* targets. Consequently it appreciably expands the set of human genes which may be regulated

by NATs. We additionally show that the same mRNA may potentially be regulated by both *cis* and *trans* encoded NATs.

Our search revealed a putative regulatory network, which exhibits many-to-many relations: A given NAT may have multiple mRNA targets and a given mRNA may serve as a potential target of more than one *trans*-encoded NAT.

To assess the regulatory potential of this antisense network, we tested whether transcripts participating in sense-antisense pairing have common biological functions or reside in specific cellular compartments. We found particular functions to be enriched among the genes that belong to this network. These functions include GO categories related to: transporter activity, vesicular transport, enzymatic activity, and response to external stimulus through signal transduction. The enriched locations were accordingly mostly membrane fractions, cytoplasm and vesicles. Strikingly, a recent report of a related analysis carried out in the plant *Arabidopsis thaliana* [129] revealed a similar set of functional categories to be enriched in the plant's *trans*-antisense network. This is a remarkable correspondence that may represent convergence to similar regulatory regimes of functionally related genes in extremely remote organisms (see discussion section 4.3.2).

2.4.1.1 Human *cis*-encoded NAT pairs can target transcripts in *trans*

To find out whether *cis*-encoded antisense can target transcripts in *trans*, we used a previously published dataset of 2,667 human *cis*-NAT pairs [67]. Each NAT pair was separated into its sense and antisense components, and each transcript was individually compared to all human RefSeq mRNAs [130], using a BLAST search [128] (see methods section 3.4.1). We searched only for hits on the opposite strand, so that all matches were in fact complementary to the query sequence. All together we searched for putative targets of 2,826 antisense queries (Note that this is much less than 2,667*2; A detailed description of the transcript set that served as query is provided in methods sections 3.4.1.2 and 3.4.1.3). To define a hit, we required a minimal sequence identity of 98%, over at least 30 nucleotides, and a BLAST e-value of $1e^{-9}$ or lower. Furthermore, we removed hits resulting from an alignment within low complexity sequence regions (see methods section 3.4.1.3). In order to separate *cis*-hits from *trans*-hits and to obtain a non-redundant set of targets, we mapped all target sequences onto the genome and merged all targets residing in the same genomic locus into one sequence (methods section 3.4.1.4). Applying this pipe-line, we found

at least one putative *trans* target (in addition to the known *cis* target) for 1107 (39%) of the queries. The number of putative *trans* targets ranged from one (for 300/1107 queries) to 55 (in one case). Figure 24 displays the distribution of the number of putative *trans* hits, for all queries that had such a hit.

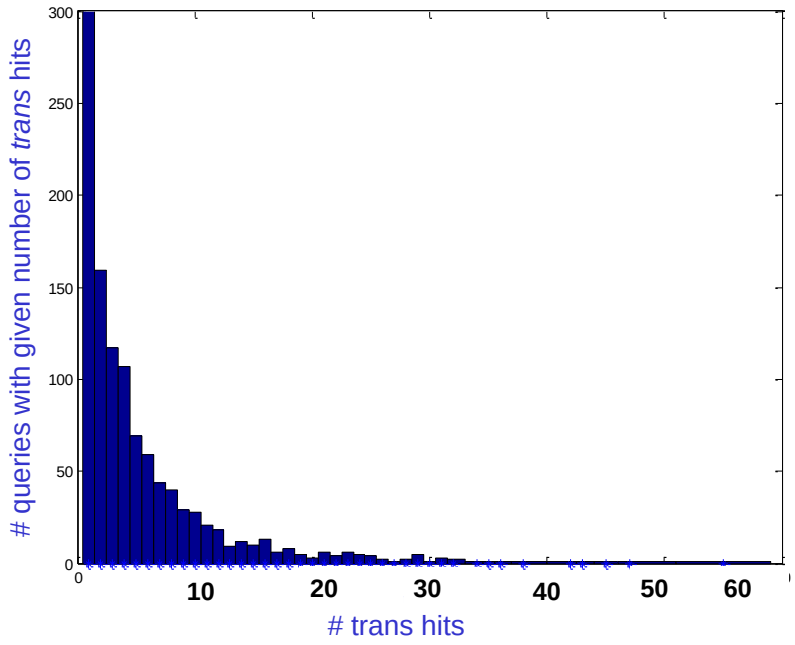


Figure 24: Distribution of the number of *trans* hits per query for the 1,107 antisense queries that had such a hit. The requirement for a hit was a minimal sequence identity of 98%, over at least 30 nucleotides, and a BLAST e-value of $1e^{-9}$ or lower

The 2,826 antisense queries along with their *cis* and *trans* hits, form a putative regulatory network that comprises of 5,252 transcripts. This network exhibits many – to-many relations, whereby a given query may have multiple targets and a given *trans* hit may be the target of more than one query. Analysis of this network revealed that 1,007 of the human RefSeq mRNAs serve as putative targets of at least one non-coding antisense. The number of antisense transcripts that potentially regulate a given *trans* target, ranges from 1 (for 400/1,007 mRNAs) to 119 (in one case). The distribution of the number of putative antisense regulators per *trans* target is displayed in Figure 25. 67 RefSeq mRNAs have the potential to be regulated by both a *cis*-encoded NAT and *trans* encoded NAT(s).

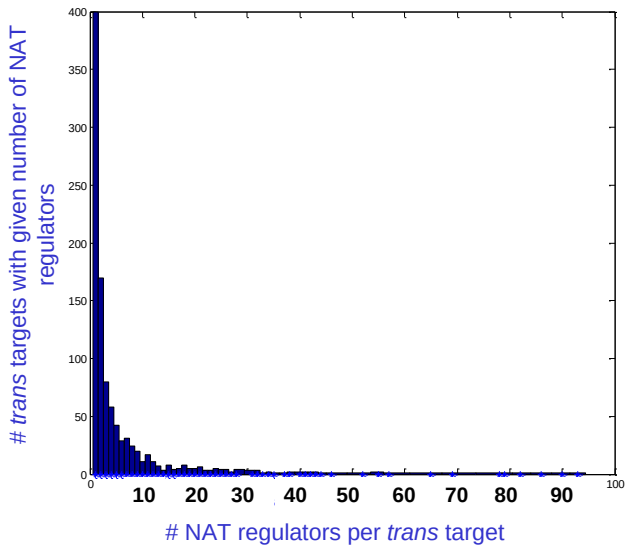


Figure 25: Distribution of number of NATs that have the potential to regulate each RefSeq mRNA. The distribution only covers the 1,007 RefSeq mRNAs that have a complementary *trans* encoded antisense. There are two mRNAs with over 100 potential regulators, which do not appear in the distribution.

We compared the alignment length distribution and the typical percent identity of all *trans* hits to those of all *cis* hits (Figure 26). While *cis* hits were typically longer (minimal alignment length 32 bases, maximum-4,807, median-156), there were also *trans* alignments as long as 2,499 bases (minimum-32, maximum-2,499, median-41). Regarding percent identity, *trans* hits displayed a range of 98% (the minimal search cutoff) to 100%, when 100% identity was only observed for the shorter alignments; 95% of the *trans* hits displaying perfect complementarity were of length 48 bases or shorter, and the longest alignment with 100% identity was of 672 bases (Figure 26, right panel). The *cis* hits should have by definition displayed 100% sequence identity. Surprisingly, 15% of the *cis* targets display a lower identity (Figure 26, left panel). We suspect that this imperfect complementarity between the NAT query sequences and their mRNA targets is a result of sequencing errors in the NAT sequences, which were mostly derived from single pass ESTs.

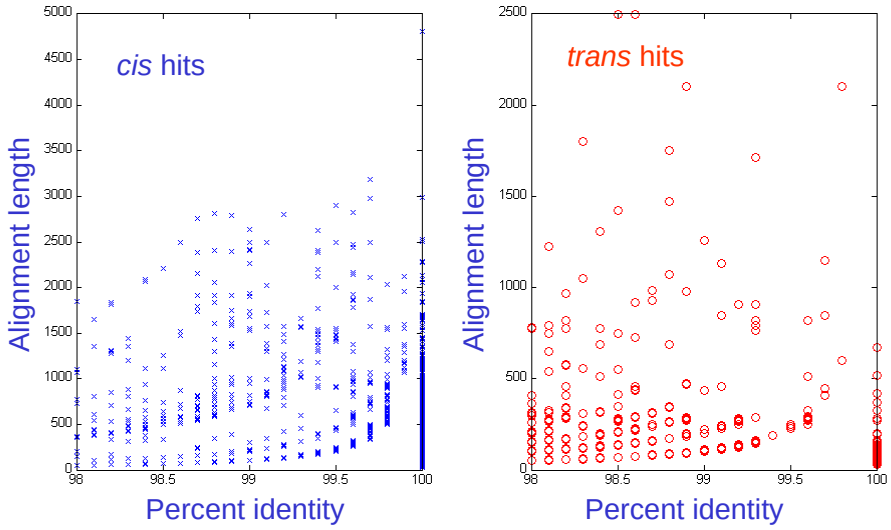


Figure 26: Alignment length versus percent identity for all *cis* hits (blue) and all *trans* hits (red). While *cis* hits display 100% identity across all alignment lengths, *trans* hits display perfect complementarity only for the shorter alignments (up to 672 bases with a median alignment length of 39 bases)

2.4.1.2 Functional categorization of transcripts involved in sense-antisense pairing

In order to examine the regulatory potential of our putative antisense network, we tested whether the set of putative targets of a given antisense, share common molecular functions, are involved in common biological processes or are located at specific sub-cellular locations. For each of the 1,107 NATs, which had at least two putative targets (one in *cis* and at least one in *trans*), we tested whether the set of targets is functionally enriched in any of the GO terms. We used terms in all three categories: molecular function, biological process and cellular localization. Functional enrichment was assessed using the hypergeometric distribution, as described in section 3.3.2. We could only assess functional enrichment for target sets which contained at least two transcripts with an assigned GO term. This further limited our search. Given this limitation, 137 sets appeared enriched with at least one GO term (most sets were enriched with multiple terms). Interestingly, different target sets were enriched with similar or related terms, implying that the entire network may play regulatory functions in a particular set of biological conditions (Table 3).

General category	Reported in plant	GO term	Term type	# Sets enriched with related terms	p-value	# Genes annotated with term	Gene set size	Category size
Transporters and transferases	Yes	transferase activity (GO:0016740)	Mol. Func.	9	9.25E-04	4	4	2,894
		channel or pore class transporter (GO:0015267)	Mol. Func.	1	1.29E-03	2	2	596
		UDP-glycosyltransferase activity (GO:0008194)	Mol. Func.	2	3.46E-04	2	4	127
		intracellular transport (GO:0046907)	Biol Proc.	4	7.85E-04	3	5	727
		secretory pathway (GO:0045045)	Biol Proc.	3	1.81E-03	2	5	227
		protein import (GO:0017038)	Biol Proc.	1	2.16E-04	2	5	78
Ion binding	No	metal ion binding (GO:0046872)	Mol. Func.	7	2.51E-03	5	5	5,011
Enzymatic/Catalytic activity	Yes	oxidoreductase (GO:0016491)	Mol. Func.	2	5.22E-03	3	6	1,117
		ligase (GO:0016874)	Mol. Func.	3	7.31E-04	2	2	449
		enzyme activator (GO:0008047)	Mol. Func.	4	1.62E-03	2	2	297
		protein kinase (GO:0004672)	Mol. Func.	2	8.62E-03	2	3	906
Signal Transduction	Yes	growth factor (GO:0008083)	Mol Func.	2	1.60E-04	2	2	210
		G-protein coupled receptor (GO:0004930)	Mol. Func.	1	4.51E-03	2	2	1,114
		GTP-binding (GO:0005525)	Mol. Func.	1	8.64E-03	2	5	503
		Transcription factor (GO:0003700)	Mol. Func.	3	3.61E-04	3	3	1,182
		G-protein coupled receptor protein signaling pathway (GO:0007186)	Biol. Proc.	6	6.32E-03	2	2	1,319
		regulation of transcription (GO:0045449)	Biol Proc.	7	5.14E-03	3	3	2,864
Nucleic Acid binding	No	DNA binding (GO:0003677)	Mol. Func.	4	4.19E-03	3	3	2,675
		RNA binding (GO:0003723)	Mol. Func.	1	2.48E-03	2	2	827
		adenyl nucleotide binding (GO:0030554)	Mol Func.	4	1.77E-02	2	2	2,206
Motor activity/ cytoskeleton	No	microfilament motor activity (GO:0000146)	Mol Func.	1	2.29E-06	2	3	15
		actin binding (GO:0003779)	Mol Func.	1	1.13E-03	2	3	324
		cytoskeletal protein binding (GO:0008092)	Mol Func.	1	2.21E-03	2	3	455
		cell motility (GO:0006928)	Biol Proc	11	2.16E-03	2	5	248

General category	Reported in plant	GO term	Term type	# Sets enriched with related terms	p-value	# Genes annotated with term	Gene set size	Category size
response to biotic stimulus/ immune	No (there is hormonal response)	cytokine activity (GO:0005125)	Mol. Func.	1	5.03E-03	2	7	264
		response to biotic stimulus (GO:0009607)	Biol. Proc.	10	4.72E-04	2	2	361
metabolism/biosynthesis	Yes	nucleotide metabolism (GO:0009117)	Biol Proc.	20	6.77E-04	2	3	251
		regulation of metabolism (GO:0019222)	Biol Proc.		6.92E-03	3	3	3,161
		protein biosynthesis (GO:0006412)	Biol Proc.		4.54E-04	3	4	814
		biosynthesis (GO:0009058)	Biol Proc.		9.37E-04	3	3	1,624
Apoptosis	No	apoptosis (GO:0006915)	Biol Proc.	17	1.63E-03	2	2	671
Cellular Compartment								
Membrane/membrane fraction	Not tested in plant	intrinsic to plasma membrane (GO:0031226)	Cel. Comp.	13	6.18E-03	2	2	1,304
		Golgi membrane (GO:0000139)	Cel. Comp.		2.00E-03	2	8	143
		endoplasmic reticulum (GO:0005783)	Cel. Comp.		1.88E-03	2	2	720
		organelle membrane (GO:0031090)	Cel. Comp.		7.35E-04	4	12	616
		membrane-bound organelle (GO:0043227)	Cel. Comp.		5.15E-03	6	6	6,895
Vesicle		membrane-bound vesicle (GO:0031988)	Cel. Comp.	4	9.69E-03	2	12	210
cytoplasm/cytoplasmic part		cytoplasm (GO:0005737)	Cel. Comp.	10	3.53E-03	4	4	4,044
cytoskeleton/filaments		cytoskeleton (GO:0005856)	Cel. Comp.	3	4.84E-03	2	2	1,155
extracellular		extracellular region (GO:0005576)	Cel. Comp.	2	7.32E-03	2	2	1,420

Table 3 Functional Enrichment of GO terms: For each of the 1,107 NATs, which had at least two putative targets (one in *cis* and at least one in *trans*), we tested whether the set of targets is functionally enriched in any GO term, using the hypergeometric distribution. We tested out all GO terms in the three categories: Molecular Function (Mol. Func.), Biological Process (Biol. Proc.) and Cellular Component (Cel Comp). We compared our results to the results reported for Arabidopsis, however in plant only molecular function terms were evaluated. The table groups the enriched GO terms into broader functional categories (left column). For each such broad category, we list selected examples of enriched terms, along with their hypergeometric p-value for one set. The table also states for each enriched term, what was the number of antisense target sets in which it was enriched.

The enriched molecular function terms were: Transferase activity, channel and pore transporter activity and intracellular transport; Binding of different cations including calcium ion, iron, zinc, magnesium and transition metal ion; Activities related to signal transduction, including: G-protein coupled receptor activity, GTP binding and transcription factor activity; Enzymatic activity including ligase, hydrolase, kinase and oxidoreductase, and motor related activity, including microfilament motor activity, cytoskeletal protein binding, and actin binding.

The enriched cellular compartments were mostly membranes of different organelles including Golgi apparatus, endoplasmic reticulum and plasma membrane, as well as several types of vesicles, which all relate to the vesicular transport system. There were also different filaments that may mediate transport such as actin cytoskeleton, intermediate filament and contractile fiber.

The enriched biological processes were accordingly: actin filament-based processes, cell surface-receptor linked signal transduction, response to biotic stimulus, transcriptional regulation, intracellular signaling, intracellular transport, protein import, secretory pathway, vesicle mediated transport. Biosynthesis, metabolism and apoptosis were also enriched.

Interestingly a recent paper which studied genome-wide *trans*-encoded antisense in *Arabidopsis thaliana* [129], reported similar enriched GO categories including: various transporter activities, transferase activity, catalytic activity, and signal transduction. Furthermore a specific transferase gene family, UDP-glycosyl transferase, was enriched both in the *Arabidopsis* antisense network and in our network. Some functions, unique to plants, were also reported in *Arabidopsis*, such as cell wall biosynthesis, response to auxin stimulus and chlorophyll binding. Intriguingly all categories enriched in plant, which are not plant specific, were also enriched in human (with the exception of one category ubiquitin-dependent proteolysis, enriched in plants but not in human). In human we observed some additional enriched functions. These findings suggest that the same control mechanism may have developed in both human and plant, but has been optimized to serve slightly different functions in each organism.

We next intend to test whether sense-antisense transcript pairs have a tendency to be co-expressed in the same set of tissues, by examining the expression patterns of transcripts belonging to our antisense network in 22 whole human tissues [131]. In

addition we have mapped known TF binding sites onto the putative promoters of all transcripts participating in our network, in order to see if sense transcripts tend to be regulated by the same TFs as their antisense targets.

2.4.2 Coupling NATs mechanisms of action to their regulation

The fact that the level of some antisense transcripts is correlated with that of the corresponding sense transcript, while in other cases an inverse relation is observed [79], suggests that the mechanisms of NAT action may be diverse. Indeed, as discussed in the introduction section 1.2.3, well-documented NAT examples point to four major mechanisms [88]: transcriptional interference, RNA masking, double-stranded RNA (dsRNA)-dependent mechanisms and chromatin remodeling.

Each mechanism requires different associations between sense and antisense expression patterns; some mechanisms require the concomitant presence of sense and antisense transcripts, whereas others impose their mutual exclusion (Figure 27). We propose that the regulation of sense and antisense transcription is coupled to serve the different regulatory mechanisms employed by the antisense. Below we accompany each of the proposed mechanisms of NAT action, with the resulting relationship between sense and antisense transcript levels:

(i) *Transcriptional interference* - The presence of an overlapping transcriptional unit might stall sense transcription owing to the collision of two bulky RNA polymerase II complexes on opposite strands (see introduction section 1.2.3). Competitive transcriptional interference could be the underlying mechanism when anti-correlated expression levels of sense and antisense are observed. Such interference might alternatively result in the shutdown of both transcripts (Figure 27A).

(ii) *RNA masking* - Sense-antisense duplex formation might mask *cis* elements residing in either of the transcripts and hinder processes that require protein–RNA interactions such as splicing, mRNA transport, polyadenylation, translation and degradation. The best characterized examples of this mechanism are of antisense transcripts which mask splice sites and cause the retention of the corresponding intron [92, 93]). Such a mechanism would result in a correlated expression level of the antisense and the preferred splice variant (Figure 27B).

(iii) *dsRNA-dependent mechanisms and RNA interference* - There is accumulating evidence that antisense transcripts might function through the activation of dsRNA-dependent mechanisms such as RNA editing and RNAi. Such mechanisms require the simultaneous existence of sense and antisense transcripts for duplex formation, and might therefore account for the observed co-expression of numerous sense–antisense pairs (Figure 27C; [79]).

(iv) *Antisense involvement in methylation and monoallelic expression* - non-coding antisense transcripts have been reported to induce the methylation and silencing of corresponding genes [102] and to be involved in X-chromosome inactivation, genomic imprinting [103] and allelic exclusion [104]. Common to all these processes is that antisense transcription affects an entire gene cluster, rather than merely the overlapping sense transcript. The silencing effect is probably exerted through the recruitment of histone-modifying enzymes, resulting in chromatin remodeling and transcription silencing. We therefore predict an inverse expression profile for the antisense and all the genes in the silenced cluster (Figure 27D).

The descriptions above illustrate that the coupling between the transcriptional regulation of the sense and antisense transcripts is characteristic of the regulatory mechanism at hand. We therefore suggest that the relationship between the expression profiles of sense and antisense transcripts can hint at the mechanism at work, as well as at the ultimate biological outcome

To illustrate this point, we predict two biological outcomes that might result from a delay between the initiation of transcription of the sense and antisense transcripts. In the first scenario, the sense transcript is initially transcribed up to a certain level, then antisense transcription begins which subsequently promotes sense degradation. In this case the anticipated outcome of the antisense activation is a delayed shutdown of the sense gene (Figure 28A). In a second scenario antisense transcription precedes sense transcription. This regime is somewhat less intuitive. We speculate that its biological outcome might be the dampening of stochastic fluctuations (noise) in the level of the sense transcripts; the antisense level sets a threshold and only sense transcripts that exceed it are effectively expressed (Figure 28B). Noise dampening was hypothesized

to be obtained by another type of regulatory RNA, microRNAs [132], and we hypothesize that antisense transcripts might fulfill the same function.

Differences in transcription activation times might be encoded by differential affinities of the sense and antisense promoters to a shared transcription factor, assuming that such a regulator is an activator and that it accumulates with time (Figure 28). Indeed a bioinformatics effort conducted in the lab, identified multiple *cis*-NAT pairs, which are flanked by binding sites for a common TF. In many of these cases, either the sense or antisense transcripts had a stronger site for the TF in their promoters. This finding suggests that both scenarios described above (delayed shutdown of the sense transcript and noise dampening) may take place in different circuits. For one such NAT pair, the MDM2 gene and its corresponding antisense, it was experimentally shown that transcription activation of the antisense transcript precedes that of the sense transcript. This NAT pair is regulated by a common TF p53, which may have a stronger site within the promoter of the antisense gene. These computational predictions along with the experimentally validated example support our hypothesis that co-regulation of sense and antisense transcripts via a common TF, may comprise a general mechanism.

In addition, our prediction regarding the capacity of early transcribed antisense to provide 'noise dampening' of the conjugated sense transcript is currently tested by mathematical simulations, through a collaboration with the lab of Tsvi Tlusty (Department of physics of complex systems).

Sense and antisense transcripts might be regulated not only at the transcriptional level, but also at the level of mRNA stability. Therefore, differences in mRNA half-lives of the two transcripts might also be predictive of antisense function. A well-characterized example is the hok–sok system of the R1 plasmid [133] in which differences in mRNA stability result in delayed activation of the sense-encoded protein (the hok-sok system is discussed in further detail in our concept paper [107]).

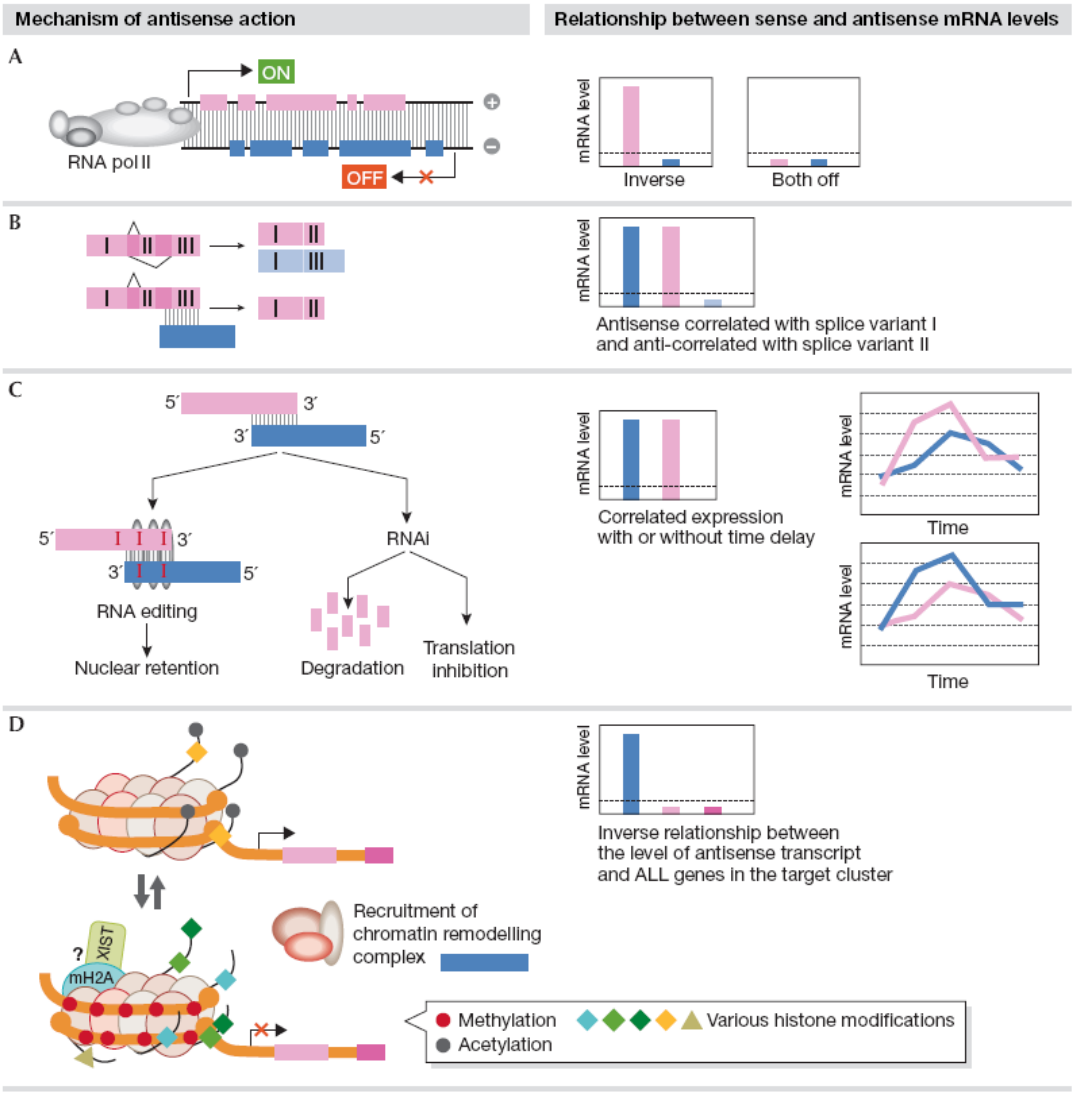


Figure 27: The main mechanisms by which NATs regulate gene expression. Each mechanism is accompanied by what it requires from, or imposes on, the relationship between the levels of sense and antisense transcripts. (**A**) *Transcriptional interference*. Two bulky RNA polymerase II complexes on opposite DNA strands might collide with and stall one another. The interference occurs mostly in the elongation step, resulting in either transcription arrest or transcription in one direction (sense or antisense) only. Such a mechanism might occur in cases in which inverse expression is observed. (**B**) *RNA masking*. A specific case is shown in which the antisense masks a splice site on the sense pre-mRNA sequence. This prevents a given splice variant from being formed and shifts the balance towards splice variants that do not require splicing of the masked region. Such a mechanism could be observed by correlated expression of the antisense and favored splice variant and an inverse relationship with the repressed variant. (**C**) *Double-stranded RNA-dependent mechanisms* such as RNA editing and RNA interference require the simultaneous presence of sense and antisense transcripts for duplex formation, and might therefore account for the observed co-expression of numerous sense–antisense pairs. A delay in expression of sense compared with antisense (or vice versa) is also possible as long as there is a period in which both transcripts are present (see Figure 28). (**D**) *Chromatin remodeling*. Transcription of non-coding antisense transcripts is involved in monoallelic gene expression, including genomic imprinting, X-inactivation and clonal expression of lymphocyte genes. In these processes, antisense transcripts have been suggested to silence the expression of nearby gene clusters by chromatin remodeling, most likely through the recruitment of histone-modifying enzymes. If such mechanisms are in action, an inverse expression profile of the antisense compared with all genes in the silenced cluster would be expected.

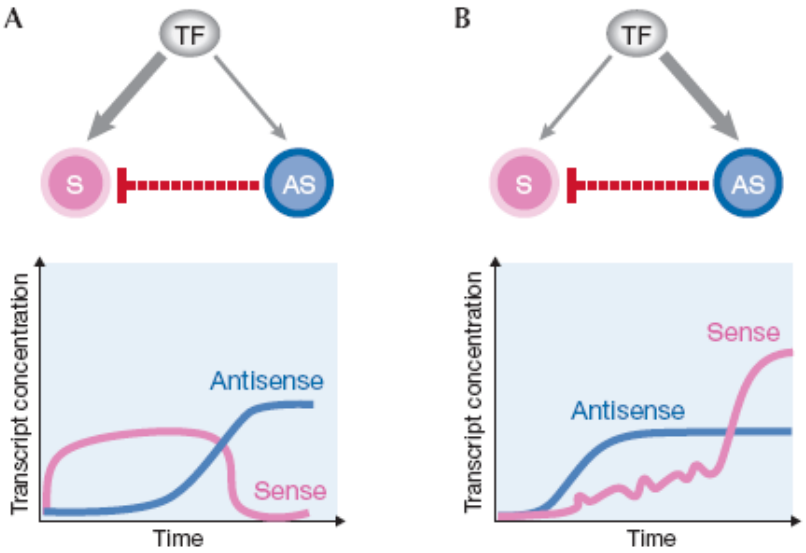


Figure 28: Differences in activation times of the sense compared with the antisense transcript. Such differences might be easily encoded in differential affinities to a shared transcription factor, assuming that this transcription factor is an activator and that it accumulates with time. (A) A higher affinity of the transcription factor to the sense transcript might result in a delayed shutdown, whereby the transcription factor initially activates transcription of the sense messenger RNA up to a certain level and only then is triggered by antisense transcription. The delayed antisense transcription prevents the sense transcript from exceeding the level it has reached when antisense transcription is switched on. (B) A higher affinity of the transcription factor to the antisense transcript. In this case, antisense transcription precedes sense transcription and acts as a buffer for the sense transcript. When the transcription factor accumulates, transcription of sense mRNA begins, but only sense transcripts exceeding the threshold set by the antisense level can be effectively translated. This generates a step-like function in the concentration of the sense transcript. Fluctuations in the amount of sense transcript below the threshold are dampened.

3 Methods

3.1 Regulatory motif dictionaries

3.1.1 Sequence and expression data

3.1.1.1 Yeast (*S. cerevisiae*)

Promoter sequences for 5,651 *Saccharomyces cerevisiae* genes were taken from SGD [134]. Whole-genome mRNA expression data of 40 time series in yeast were obtained from ExpressDB [109]. These time series represent a wide range of natural (e.g. cell cycle) [135-137] and perturbed [138-141] conditions. A detailed description of all analyzed conditions is available in appendix II.

3.1.1.2 Human

Human promoter sequences were downloaded from the UCSC database, UCSC human genome assembly Jul. 2003 [142], and saved locally in a mySQL database. We used a set of 14,252 human promoters, each of 1,200 bp; 1,000 bp from upstream of the transcription start site (TSS), and 200 bp downstream of the TSS. Genes with alternative promoters or with promoters that overlap a genomic region of another gene were not included in this set. The reason is that in most cases we had no knowledge of the alternatively transcribed transcript printed on the DNA chip (the transcript for which we have expression data), and thus preferred to confine to non-ambivalent cases. Expression data for human cell cycle experiments was downloaded from the supplementary web site of Whitfield et al [122] (http://genomewww.stanford.edu/Human-CellCycle/Hela/index.shtml)

3.1.1.3 Candida Albicans

Promoter sequences for 6,282 *C. albicans* genes were downloaded from the *Candida Genome Database* (CGD [143]), assembly 19. In most cases the promoters were of length 1,000 bp upstream of the TSS. Promoters were shorter for 203 genes, located in genomic areas for which the sequence is not yet complete (the *C. albicans* sequencing project is not yet at its final stage). Incomplete promoter sequence was most pronounced for genes located at chromosome ends. Expression data for three stress responses: heat shock (23°C to 37°C), osmotic shock (0.3M NaCl) and oxidative stress (0.4 mM H_2O_2) was taken from [121]. For 4,635 out of the 6,282 *C.*

albicans genes, we were able to identify a mutual *S. cerevisiae* ortholog, defined by reciprocal best Blast hit.

3.1.2 Dictionary Construction

The dictionary construction procedure consisted of four major steps, described schematically in Figure 29 and detailed below.

3.1.2.1 Exhaustive genome scan

Promoter sequences were systematically scanned for all occurrences of every possible k-mer (k varies from 7-11), resulting in an index file listing for each k-mer the set of genes that contain it in their promoters, along with the positions and orientations (strand). Bidirectional promoters (in yeast) were taken twice in different orientations and associated with the corresponding genes. For the purpose of indexing, each k-mer was combined with its reverse complement because TFs are thought to recognize and bind double stranded DNA.

3.1.2.2 k-mer scoring

Following the k-mer indexing step, EC scores in various experimental conditions were calculated for the sets of genes containing each of the k-mers in their promoters. A p-value was assigned to each EC score and false discovery rate (FDR) of 0.1 (allowing 10% false positives) was used to correct for multiple hypotheses [108]. This rate was chosen after trying various ratios ranging from 0.01 to 0.3, and attempting to maximize the number of true positives. In addition to the EC scores and corresponding p-values, each k-mer was characterized by the expression profile it dictates; this was defined, at each time point as the average expression level of all genes assigned to the k-mer. Such averaged profiles were defined for each k-mer across the 40 time series experiments, resulting in 40 vectors per motif. These ‘mean expression vectors’ were subsequently used to cluster k-mers into groups that share sequence as well as functional similarities (see below).

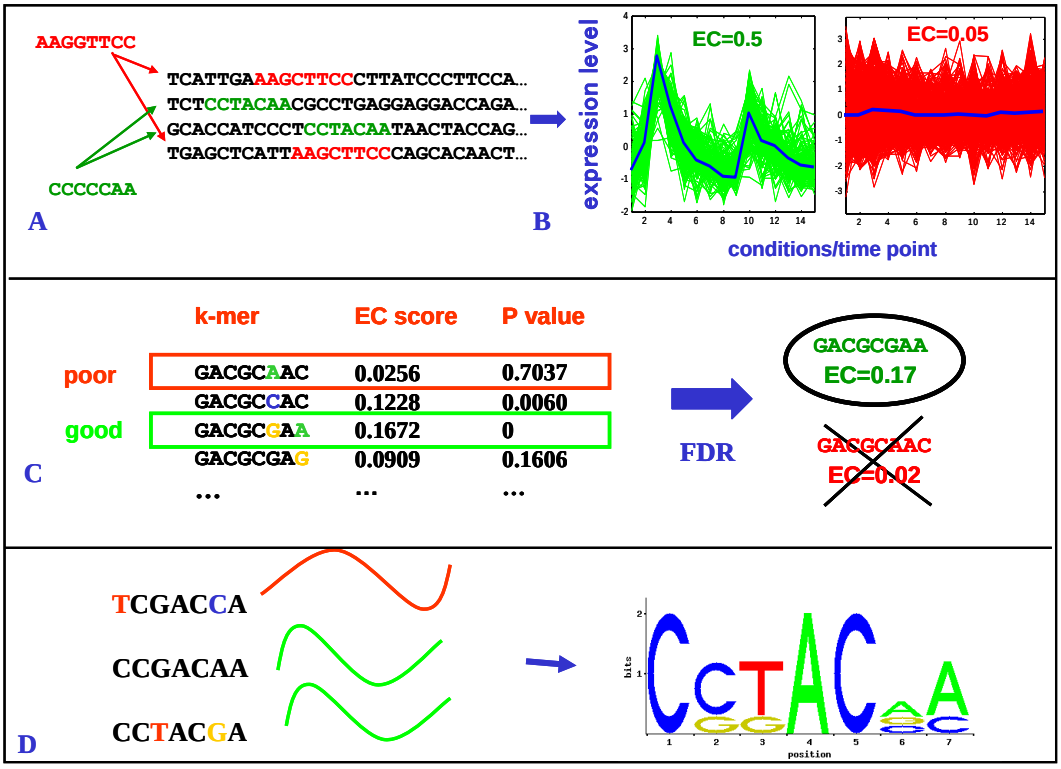


Figure 29: Schematic description of the dictionary construction process:
A. For all possible k-mers (k=7-11), index all genes that contain the k-mer in their promoters.
B. Expression coherence score – calculate EC score and assign a p-value to all genes containing a given k-mer in their promoter, over all expression conditions.
C. FDR – select only k-mers with significant EC scores.
D. Motif Clustering - group together k-mers that are similar in sequence and exert a similar effect on expression.

3.1.2.3 Clustering of dictionary motifs

We employed a two step clustering procedure in order to group together dictionary motifs that share a close sequence and exert a similar effect on expression. In the first step we clustered the motifs according to the mean expression profiles of the genes that contain them in their promoters. We used hierarchical clustering and determined the number of clusters by eye. On average every biological condition displayed 2-4 distinct expression behaviors. An example is seen in Figure 14 of the results, which displays three main expression profiles dictated by the motifs of the *C. albicans* osmotic stress dictionary. Next we clustered motif sequences that govern a similar expression behavior, based on their sequence similarity, using the QT_clust clustering algorithm [144]. Unlike many clustering algorithms, such as k-means, that require the a-priori determination of number of clusters and that give rise to clusters of various extents of tightness, in this algorithm the only input is the minimal cluster tightness, and the output is the number of clusters along with the motif-cluster assignments. The distance between each two motifs was determined by first obtaining

their optimal un-gapped alignment (examining both the motif and its reverse complement), and then counting the number of mismatches within the alignment. We required that the number of mismatches between two sequences within a cluster be lower than 30% of the shorter sequence.

3.1.3 MEX algorithm

MEX is an unsupervised motif extraction algorithm, developed in the framework of a broader algorithm, ADIOS (automatic distillation of structure), which deduces grammar from texts [28]. MEX was designed originally in order to extract patterns from natural-language corpora, and was adapted here to the problem of regulatory motif extraction. MEX is data driven; it identifies motifs that are strings of adjacent nucleotides on promoters within a genome-wide analysis. MEX does not depend on over representation of the motifs in the genome. Instead it uncovers motifs that are significant within the relatively local context of the promoter on which they occur.

Consider a data-set of many sequences of variable length, each such sequence is expressed in terms of an alphabet of finite size N (e.g. N=20 for amino-acids or N=4 for DNA, with each promoter region defining a sequence). The N letters form nodes of a graph on which the sequences in the data will be placed as ordered paths. Each such sequence defines a data-path over the graph. All N letters have an equal number of incoming and outgoing strings, or data-paths. Once all the data-paths (all promoter sequences) are loaded onto the graph, we explore the graph for patterns. An example is demonstrated in Figure 30 where the search path consists of the set of nodes A->B->C->D->E. Other paths may join and leave the search path at various vertices. In this example, the bundle of path sections between B and D display a certain coherence, possibly indicating the presence of a significant pattern.

Two probability functions are defined over the graph for any given search path: The first one, P_R is the right-moving ratio of fan-through (through going flux of strings) to fan-in (incoming flux of strings), which varies along the trial-path. Starting at A we define:

$$\Pr(A;B) = \frac{\text{number of datapaths leading from A to B}}{\text{total number of datapaths leaving A}}$$

$$\Pr(A;C) = \frac{\text{number of datapaths leading from A through B to C}}{\text{number of datapaths leading from A to B}}$$

and so on.

This function increases along the search path, because other paths join to form a coherent bundle, but shows a decrease at E, because many paths leave the search path at D. To quantify this decline of P_R, which is interpreted as an indication of the end of the candidate pattern, we define a 'decrease ratio' D_R, whose value at D is: $D_R(D) = P_R(A;E)/P_R(A;D)$. We require it to be smaller than a cutoff parameter $\eta < 1$. Similarly, we proceed from the right end of the trial-path starting with D and study a left- going ratio of fan-through over fan-in P_L. Thus:

$$Pl(D;C) = \frac{\text{number of datapaths leading from C to D}}{\text{total number of datapaths entering D}}$$

This function will increase, going to the left, and the point (B) at which it shows a considerable decrease, $D_L(D;B) = P_L(D;A)/P_L(D;B) < \eta$, is declared to be the starting point of the putative motif. Finally because the data consists of a finite number of strings that may be quite small, we need to assess whether the decrease in P_R or P_L is statistically significant. P_R and P_L may be regarded as variable order Markov probability functions, as indicated on Fig. 25. We define a threshold α and require the significance p-values of both $D_R(D) < \eta$ and $D_L(B) < \eta$ to be, on the average, smaller than $\alpha < 1$. Significance is defined here in terms of a null-hypothesis stating that $P_R(E) \geq \eta\, P_R(D)$ and $P_L(A) \geq \eta\, P_L(B)$ for the right and left-paths respectively. In other words we wish to reject the possibility that, given the existing number of strings, the probability continues to remain high.

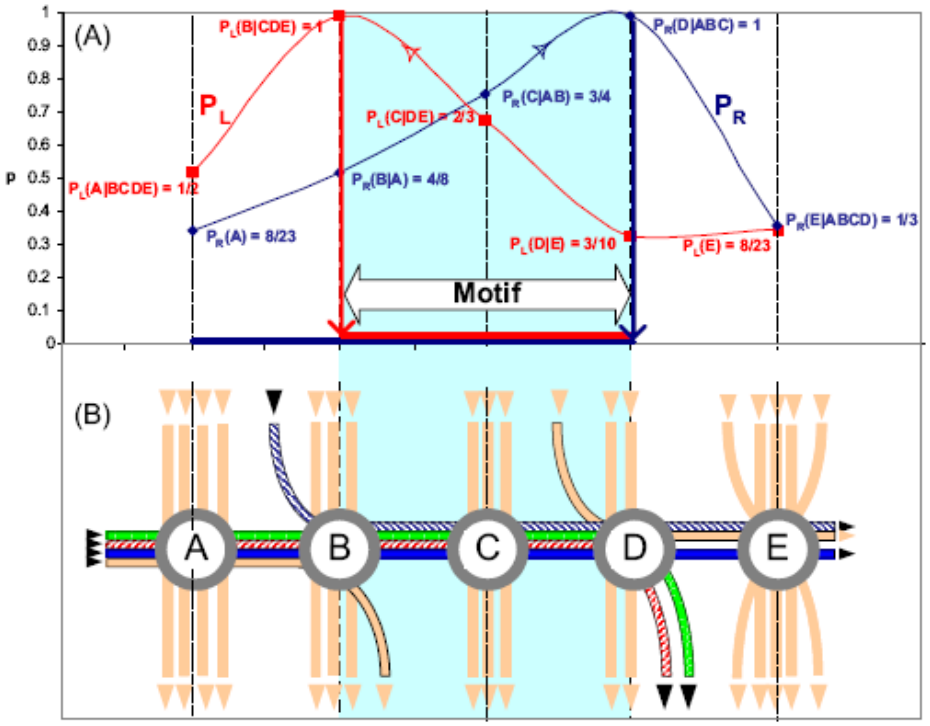


Figure 30: The definition of a motif within the MEX algorithm. Note that the maxima of P_L and P_R define the beginning and the end of the motif. Drops following the maxima signify divergence of paths at these points. The nodes in this figure are labeled by different letters. Note, however, that different letters may also label the same node appearing at different locations on the trial-path. In this example, E and A may represent the same node.

We define sequence motifs as overlapping regions of P_L and P_R whose end-points obey the thresholds conditions. All data-paths (promoter sequences) were used as trial-paths, and both P_L and P_R were calculated from all possible beginning points along the path from right and from left correspondingly. We set η to 1 and α to 0.1 (after testing various values). These two parameters should be chosen so as to best suit the problem at hand.

3.1.4 Expression coherence (EC) score

The EC score is a measure of the extent to which a set of genes is clustered into one or more clusters in expression space. The formal definition of the EC score is the fraction of gene pairs in a given set S, for which the normalized Euclidean distance between expression profiles falls bellow a threshold D.

$$EC(S) = \frac{\left|\left\{g_i, g_{j\neq i} \in S\right\} : ExpDist(g_i, g_j) < D\right|}{|S| * (|S| - 1) \div 2}$$

The threshold D is determined based on the distribution of pair-wise distances between expression profiles of all genes in the genome (or more precisely of all genes for which expression level was measured). The original definition of the EC score [22] used the 5th percentile as the cutoff for defining "close" expression profiles (D). This definition may create a bias towards TFs that exert a very tight regulation and miss regulatory motifs that correspond to factors exerting a more lose regulation. We therefore tested a range of EC definitions, with cutoffs corresponding to the 5th, 10th, 20th, 30th, 40th and 50th percentile of the pair-wise distance distribution. For each definition of EC cutoff we assigned a significance p-value separately. P-values were calculated by random sampling. For each of the 40 expression time series and for each gene set sizes (varying from 3-100 genes), we selected 100,000 random gene sets and computed an EC score for each such set at each cutoff definition. We define the p-value of a given EC score as the fraction of random sets (of the same size and in the same condition) that scored similarly or higher (Note that this sets a lower bound of 10^{-5} on the significance that can be assigned to a given EC score). Since we assume that for a given EC score, the probability to get the same score for random sets of genes drops with the set size, gene sets larger than 100 are assigned an upper bound approximated p-value, using the randomly sampled sets of size 100.

3.1.5 Matching dictionary strings to PWMs.

A scoring method was devised to assess how likely a given string is to be generated from a given PWM. The score is on a scale of 0 to 100. It is computed by summing up the frequencies corresponding to the observed nucleotides over all motif positions, and normalizing this score to a scale of 0-100. The scaling is done by subtracting the minimal possible score and dividing by the range of possible scores. For example for the PWM [A: 0.0191 0.0191 0.9733 0.9733 0.0120, C:0.9500 0.9500 0.0074 0.0074 0.0074, G: 0.0117 0.0117 0.0074 0.0074 0.0074 T:0.0191 0.0191 0.0120 0.0120 0.9733] the lowest possible score 0.0455 is obtained for the string GG(C/G)(C/G)(C/G), the highest possible score 4.8198 is obtained for the string CCAAT. After scaling GGCCC will score 0, CCAAT will score 100 and CCATT will score 79.9 ((3.8585-0.0455)/(4.8198-0.0455)).

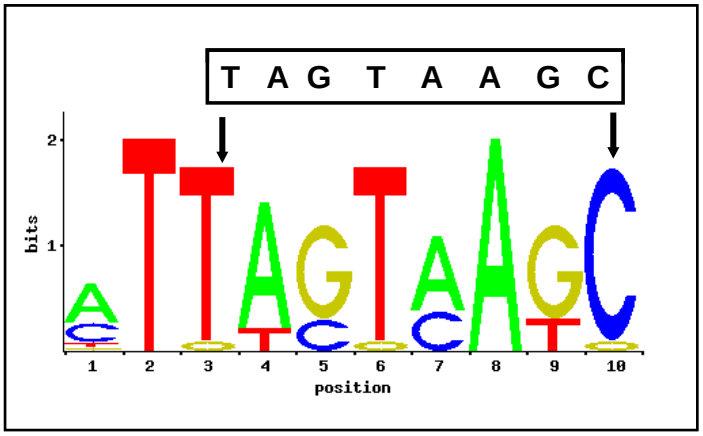


Figure 31: An example for a match between a high scoring k-mer and a known PSSM. In this example, the k-mer TAGTAAGC has a perfect match to the PSSM of the known motif CAD1. The first position of the k-mer corresponds to the third position in the PSSM.

3.1.6 Grouping Harbison's PWM set into distinct clusters

Highly similar PWMs belonging to the Harbison set were grouped together using hierarchical clustering. The similarity measure used for clustering was the compareACE score [112]; this method compares two PWMs by aligning them and calculating the Pearson correlation coefficient between the base frequencies of the aligned matrix portions. To prevent spurious matches, compareACE requires that the aligned portion include at least the six most informative positions in each motif. Using a similarity cutoff of compareACE score> 0.9, the 102 Harbison PWMs were grouped into 77 distinct clusters.

3.1.7 Functional Coherence (FC) score

Functional Coherence (FC) is a term used to describe the extent to which a set of genes is similar in function. Data on biological processes were derived from the GO database [111], which defines a hierarchy of functional annotations. The gene annotations themselves were taken from SGD [134]. The similarity measure between GO functional annotation terms was taken to be the 'semantic similarity', defined by Lord et al. [145]. Given the semantic similarity scores between each pair of GO annotation terms, the similarity score between a pair of genes is defined as follows:

$$Sim(gene_i, gene_j) = \max_{term_i \in gene_i, term_j \in gene_j} \{SemanticSimilarity(term_i, term_j)\}$$

Namely if each gene is annotated by several GO terms, the pair of terms with the maximal similarity is used. The rational is that for our purposes, if two genes participate in several biological processes, we view them as similar even if they have only one process in common. We do not require all their processes to be similar.

The FC score of a set of genes is defined as the fraction of all 'significantly similar' gene pairs out of all pairs of annotated genes in the set, where significantly similar is determined by a threshold θ:

$$FC(S) = \frac{\left|\left\{g_i, g_{j\neq i} \in S : Sim(g_i, g_j) \geq \theta\right\}\right|}{|S| * (|S| - 1) \div 2}$$

The threshold θ was set to be the 95th or 90th percentile scores of the distribution of all pairwise similarity scores of the yeast genome. Genes without annotations, or annotated as ‘biological process unknown' were excluded from the analysis. The FC p-value was calculated using random sampling in a similar manner to the EC p-value described above.

3.1.8 Positional Bias p-value

For each k-mer, we gathered the positions (relative to the TSS) of all its genome-wide promoter instances. These positions were sorted into 40 bp wide bins. Positional bias was assessed using a statistic measure introduced by Hughes et al. [112]. This measure assesses whether the most populated bin (with m motif instances) contains more motif instances than expected by chance given the overall number of motif instances (t), the promoter length (s ~ 600 bp) and the bin width (w).

$$p = \left(\frac{s}{w}\right)\sum_{i=m}^{t}\binom{t}{i}\left(\frac{w}{s}\right)^{i}\left(1-\frac{w}{s}\right)^{t-i}$$

3.1.9 Evolutionary conservation

3.1.9.1 Data

Promoter data for four closely related *Saccharomyces* species *S. cerevisiae, S. mikate, S. kudriazevii and S. bayanus* were taken from Cliften et al. Science 2003 [13] The reference lists of motifs that were defined strictly based on phylogenetic footprinting were taken from both Cliften et al. Science 2003 [13] and Kellis et al. Nature 2003 [14] .

3.1.9.2 Motif conservation calculation

The motif conservation calculation was adapted from Xie et al. Nature 2005 [146]. We defined the motif conservation rate separately for each motif as the ratio of conserved motif instances to total occurrences of the motif in the genome. We regarded a motif instance as conserved if at least 90% of the motif positions were identical across all 4 species. For each motif length (from 7-11), we obtained the distribution of expected conservation rates using a control set of 1,000 random motifs of that length. We took the 95th percentile of the control set distribution as the cutoff defining high conservation and counted the number of motifs with a conservation rate above this cutoff.

3.2 Functional characterization of binding site variations

3.2.1 Motif positions used to gather statistics on substitution severity

To define properties of binding site substitutions which alter gene expression, we accumulated statistics from substitutions of different positions across multiple binding sites. The binding sites used for this analysis were core set motifs that correspond to known TFBS from Harbison's set. As described in the results 1,402 of our core set motifs had at least one corresponding Harbison PWM, with a match score of 99 or above (Table 1). Some of these motifs were similar to more than one Harbison PWM, because some Harbison PWMs belong to the same TF family and are thus very similar to one another (for instance MET31and MET32, HAP2, HAP3 and HAP5, INO2 and INO4, FKH1and FKH2). In such cases we selected the Harbison PWM with the best match score to a given k-mer, as its annotation. Taking only the best match for each k-mer, our data set covers 74 of Harbison motifs. For each of these 74

Harbison motifs, we selected a unique set of representative k-mers according to the following criteria: (i) Biological condition – if there were multiple k-mers that match this TF, but they govern different biological conditions (as determined by the significance of their EC scores in the different conditions), we selected a representative k-mer for each condition. (ii) For k-mers that correspond to the same TF and appear to govern the same biological condition, we performed clustering based on both k-mer sequence similarity and the expression profiles of the genes each k-mer regulates. We then selected a representative k-mer for each such cluster. Applying these criteria, we selected 339 unique k-mers that match 74 of Harbison's TFs. This means that on average a single Harbison TF has about 4 k-mers representing it. Statistics were accumulated for all 2,881 positions within these 339 motifs. For each motif the effect on expression was measured in the condition in which the motif had a significant EC score. If there were several such conditions, the condition with the most significant EC score was chosen.

3.2.2 The information content of a motif position

The information content (IC) of a binding site position is a measure of its conservation among multiple binding site instances. The more conserved the position, the highest its information content. The information content of a position at which the nucleotides A,C,G, and T occur with probabilities p_A, p_C, p_G, and p_T, respectively, is defined as:

$$IC = \sum_{i \in \{A,C,G,T\}} p_i \log 2 \frac{p_i}{q_i}$$

when q_A, q_C, q_G, and q_T are the corresponding background promoter nucleotide frequencies. In the specific case of equal nucleotide background distribution this formula is reduced to:

$$IC = 2 + \sum_{i \in \{A,C,G,T\}} p_i \log 2 p_i$$

and the IC is thus bound from 0 to 2

3.3 The Evolution of Interferon-α Promoters

3.3.1 Promoter Scan

IFN-α promoters were taken to be 1,000 bp upstream and 200 bp downstream of the predicted TSS, but not further than the first ATG. IFN-α genes have short 5' UTRs, hence average promoter length was 1,065 bp. The promoter sequences were downloaded from the UCSC human genome browser (http://genome.ucsc.edu/). The promoters were scanned with a database of 344 human TF binding sites represented by positional weight matrixes (PWMs). The majority of PWMs (337) were downloaded from TRANSFAC [123], while seven PWMs, representing interferon regulatory factor (IRF) elements, were manually added from the MatInspector library [147] and from the literature [148, 149]. The 13 promoters were scanned for the presence of these 344 PWMs via a Perl script utilizing the TFBS module http://forkhead.cgb.ki.se/TFBS/ [150]. This module implements the PWM search algorithm described in [151]. The search cutoff used was 92%. This was the strictest cutoff that still detected all binding sites that are known to reside in IFN-α promoters.

3.3.2 Functional enrichment of GO terms

The GO consortium [111] defines hierarchies of gene annotation terms in three major categories: biological process, molecular function and cellular component. To assess whether a set of genes is functionally enriched with a given GO term, a series of individual tests is conducted. For each GO term we test whether the set of genes annotated by this term (or all terms beneath it in the hierarchy of the GO graph) has a significant overlap with our gene set of interest. The significance of the overlap is evaluated using the hypergeometric p-value. For each group of genes and each GO term, all p-values of overlaps are calculated, and FDR is performed to control for multiple hypotheses. A gene set is said to be functionally enriched with any term with which it had a statistically significant overlap. To automatically annotate TFs as immune related, we conducted functional enrichment tests using molecular function terms, for sets of genes containing each motif in their promoters. We searched for motifs that regulate genes enriched in immune related functions, such as defense response, immune response, response to external stimulus, response to biotic stimulus etc. In addition we searched the literature for indications of the relevance of all TFs to immune response, obtaining a set of 146 immune related TFs.

3.3.3 Binding site enrichment in IFN-α promoters.

To select motifs that are enriched in the promoters of IFN-α genes, relative to all other human genes, we searched for the 344 PWMs in a set of 14,252 promoters from the entire human genome. For each PWM, we compared the fraction of IFN-α promoters in which it appeared to the fraction of non-IFN promoters in which it appeared. We asked that the ratio between these two fractions be greater than 1. 123 motifs passed this criteria.

3.4 Antisense Transcription

3.4.1 Pipeline for whole-genome search of trans antisense hits

3.4.1.1 *cis*-NAT dataset and RefSeq mRNA sequences.

GenBank accession IDs for 2,667 previously published human *cis*-NAT pairs, were downloaded from the supplementary website of Yelin et al. [67]. This data comprised of 744 pairs of protein-coding transcripts, 1,546 pairs in which one transcript is protein-coding and the other non-coding and 377 pairs in which both transcripts are protein-coding. All human RefSeq mRNA sequences were downloaded from NCBI on May 2006. (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/).

3.4.1.2 Data validation

The published *cis*-NAT pairs were based on old assemblies of the human genome. To validate their location and directionality, we positioned all the downloaded accession IDs on the current human genome release, hg18. We used annotation tables downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/).

The non-coding transcripts were mostly ESTs, which are often miss-annotated. Indeed 50% of the ESTs were annotated as having the wrong orientation; Namely both transcripts belonging to a given NAT pair were annotated as residing on the same strand. We corrected the orientation of each transcript using further information such as the orientation of the splice sites within it. Following genome-mapping and validation of correct orientation, we removed *cis*-NAT pairs from the data-set in each of the following cases: (1) There was no genomic location data for at least one member of the pair (2) There were multiple genomic locations for at least one of the transcripts. (3) Both transcripts resided on the same strand, after validation of correct

orientation. (4) Transcripts mapped to different chromosomes, or to different locations within the same chromosome.

This filtering step resulted in a dataset of 2,025 NAT pairs (out of the original 2,667 IDs). Of these, 493 NAT pairs (out of the original 744) were both protein-coding, 1,224 pairs (out of the original 1,546) were composed of one protein-coding transcript and one non-coding transcript, and 308 NAT pairs (out of 377) consisted of two non-coding transcripts.

3.4.1.3 Blast search

Each NAT pair was separated into its sense and antisense components, and each component was individually compared to all human RefSeq mRNAs [130] using nucleotide BLAST (BLASTn) [128]. In the case of the 'mixed' NAT pairs, containing one protein-coding and one non-coding transcripts, we used only the non-coding transcript as search queries. The rational which guided this decision was to use only the potential regulators as queries. For a mixed coding and non-coding NAT pair, a reasonable scenario is that the non-coding transcript regulates the coding transcript, whereas in a pair of all-coding or all-non-coding NATs, it is harder to anticipate, which will be the regulator. There are known examples of two coding antisense transcripts, which regulate each other [98].

All together we obtained 2,826 antisense queries (2*493 + 1,224 + 2* 308). We searched for genome-wide targets for these queries using BLASTn [128] with the following cutoffs: e-value $\leq 1e^{-9}$, percent identity $\geq 98\%$, and alignment length ≥ 30 [57]. The strand flag (S) of the BLASTn algorithm was set to 2, indicating a search only against the complementary strand. We further removed BLAST hits that were within low-complexity sequence regions, using two programs – RepeatMasker (http://www.repeatmasker.org) and Tandem Repeat Finder (http://tandem.bu.edu/trf/trf.intermediate.submit.html). RepeatMaker was run with the flag -noint for masking only low complex/simple repeats.

3.4.1.4 Separation of *cis* hits from *trans* hits

To separate *cis* hits from *trans* hits, we located all hits on the human genome assembly hg18. If there were several *cis* hits, we took the longest. We further merged all *trans* hits that mapped to the same location into one hit. We created a contig of all overlapping *trans* hits and took the sequence that displayed the longest alignment with

the query as the contig representative. After merging hits residing in the same genomic locus, we obtained at least one *trans* hit for 1,107 (39%) of our queries. Altogether 5,252 transcripts appear to participate in sense-antisense pairing.

3.4.1.5 Annotations of transcripts participating in sense-antisense pairing

To study the characteristics of transcripts that participate in sense-antisense pairing, we obtained their annotations in various categories. We used GeneALaCarte (http://www.genecards.org/cgi-bin//BatchQueries/Batch.pl), a batch search utility that automatically extracts annotations for our 5,252 transcripts. GeneALaCarte provides annotations only for transcripts that correspond to an approved HUGO symbol [152]. 2,282/5,252 transcripts had HUGO symbols and hence annotations. For these transcripts we downloaded Gene Ontology (GO) annotations, and expression patterns in 12 normal human tissues [153].

4 Discussion

The integration of genomic sequence data with various types of functional information such as gene expression profiles, Gene Ontology annotations (of biological processes, molecular functions or cellular components), mRNA decay profiles etc. allows us to investigate various regulatory mechanisms which operate in the cell. The same methods may be applied to the study of different mechanisms. For instance, we applied the dictionary construction methodology, introduced here, to promoter sequences in combination with mRNA expression data, and produced comprehensive collections of TF binding sites. The same method applied to 3' UTR sequences in combination with measures of mRNA half lives (indicating transcript stability), was used to construct a catalogue of stabilizing and de-stabilizing sequence motifs [110]. Such motifs are likely bound by proteins which act in mRNA stabilization (or destabilization). Similarly, integration of 3'UTR sequences with GO cellular component annotations, yielded a catalogue of motifs associated with sub-cellular localization [110].

This demonstrates the strength of our method and its wide applicability to a range of cellular processes. Transcriptional regulation (studied here) together with the regulation of mRNA degradation and sub-cellular localization (studied in [110] employing the methodology developed here) are the components which determine the cell's transcriptome.

In the antisense project we applied a slightly different approach in which we initially used sequence information alone in order to group together genes potentially targeted by the same antisense transcript. We then searched for functional annotations enriched within these gene sets. Functional annotations were thus used here for validation purposes as well as for characterizing the regulatory process at hand, and not for dataset construction.

4.1 Regulatory motif dictionaries

4.1.1 Method strengths

We introduced a computational approach, which quantifies the effect of promoter sequence elements on the expression profiles of the corresponding genes, in order to produce unbiased reference collections of eukaryotic TFBS. Applying this method to

the *S. cerevisiae* genome across various natural and perturbed biological conditions, we obtained a dataset of putative regulatory motifs, with a good coverage of known binding sites, as well as novel and refined sites. Although we relied on genomic sequence and gene expression data alone in the derivation of our motif dictionaries, many of the defined motifs exhibit properties known to characterize functional binding sites. These include high evolutionary conservation, high positional bias, distinct nucleotide composition, all of which serve as validations for the biological relevance of our motifs. In addition, we found that sets of genes defined by each of our dictionary motifs are likely to share common biological functions, as might be expected from their common regulation. These findings not only validate our ability to identify functional motifs, but also prove our dataset as useful for the systematic search of additional binding site characteristics.

Our method overcomes the requirement for significant motif over-representation, posed by common *ab initio* motif finding algorithms. This is done here by introducing a new statistical model. This model enables the detection of motifs that may regulate even small transcriptional networks and that may be present in the genome in relatively low numbers. The method is in principle applicable to any organism for which both genome-wide promoter sequences and mRNA monitoring data are available. Indeed we constructed motif dictionaries for *Candida albicans* in response to different stresses, for *C.elegans* during embryonic development (carried out by a lab colleague, Shai Shen-Orr, using the same methods) and for human through the progression of the cell cycle.

4.1.2 The complementary sequence-based approach

Despite this proof of concept, some limitations of the method must be pointed out: (i) The exhaustive k-mer enumeration employed is computationally demanding, as the number of possible sequences increases exponentially with k. This makes it difficult to systematically scan sequences of increasing lengths (ii) Larger genomes likely encompass more elaborate regulatory regions, which would require scanning longer sequences for the presence of each k-mer (iii) The fact that we do not integrate any additional information available to us when defining the motifs has both an advantage and a drawback. The advantage is that we remain unbiased by known binding sites and thus do not limit ourselves to sites which obey conventional criteria, such as evolutionary conservation. The drawback is that we do not restrict our search space to

include only likely hypotheses. Our systematic k-mer scan generates multiple hypotheses, the majority of which are likely false. Such low signal to noise ratio imposes strict statistical criteria on a k-mer in order for it to score significantly. The result is many false negatives, namely weak motifs fail to be recovered.

To overcome these limitations, we demonstrate the performance of a complementary syntax-based approach [29]. In this approach we first generated likely hypotheses, based on syntactic assumptions, and then applied our EC-based functional assessment to these hypotheses alone. We assumed that functional motifs posses inherent position dependencies which may be captured by high-order hidden Markov models [154]. One algorithm which employs such a model is MEX (Motif Extraction algorithm) [28] developed originally for the extraction of patterns from written language corpora. The applicability of a linguistic algorithm to the biological task at hand is intriguing as it implies that rules which govern natural languages may also define 'meaningful' biological sequence. If this holds, such principles may in future allow the inference of function from raw sequence data, much like what is common practice for protein-encoding sequences since the breaking of the genetic code.

Pre-selection of candidate motifs by MEX provided an enormous enrichment in signal; 22% of the motifs extracted by MEX appeared to govern coherent expression as compared to 0.6% of the k-mers scanned via the exhaustive approach. This implies that pre-selection based on inter-position dependencies indeed increases the chance of a k-mer to score significantly. In addition MEX may extract k-mers of increasing lengths without increasing its computational complexity; it readily extracted motifs of length up to 19 nucleotides. Even within the length limitation employed by the exhaustive scan (7-11 bases), some relatively weak motifs passed the imposed statistical criteria only within the signal-enriched background provided by MEX. MEX is expected to fail in cases in which a functional motif does not obey the inherent position dependencies it selects for. It has been reported that while some functional TFBS display such position dependencies, others do not [117].

While we chose to pre-select likely regulatory motifs based on their internal sequence structure, other criteria may also be considered. For instance the k-mer search may be confined to promoter regions which are evolutionary conserved or to small sequence windows which coincide with the boundaries of distinct genomic annotations (e.g. first intron, first exon etc.). An exhaustive k-mer search within restricted sequence windows was employed in the lab for *C. elegans*. This search

detected regulatory motifs in locations other than the traditional upstream region, for instance a gene's first intron or the region immediately downstream to the translation stop site. Moreover it revealed that different motifs tend to appear at different locations and there is a correlation between the location of the motif and the effect it exerts on expression; For example motifs appearing immediately downstream of the gene typically have a down-regulating effect. This illustrates the use of our method for the construction of context dependent motif dictionaries. Such dictionaries may reveal new regulatory schemes.

4.1.3 Method limitations

Despite the proven advantages of our method, one inherent limitation is that we do not take into account the promoter backgrounds in which each motif is embedded. We currently refer to all motif instances as equal, whereas it is well accepted that many aspects, such as a motif's location relative to the TSS and to other binding sites and its surrounding promoter sequence, affect the accessibility of the motif to its binding protein. A recent publication [155] revealed that the locations of nucleosomes are encoded in the genome and may therefore be predicted from DNA sequence. This is a major step towards our ability to incorporate chromatin structure in our model. Other factors which may be readily incorporated include TF locations and multiplicity: These can be accounted for by calculating EC scores only for promoters in which the motif resides at a certain distance from the TSS, or only for promoters which contain multiple motif instances. The latter was successfully implemented by us, and is available on the web through our 'Motif Analysis Workbench' [21] at http://longitude.weizmann.ac.il/services.html . The described refinements may filter non functional motif instances and improve our method's sensitivity.

Another limitation is that we currently assess individual motifs only. Some TFs exert a measurable effect on gene expression only when operating in combinations with additional TFs. For the detection of motifs recognized by such TFs, pairs and triplets of co-occurring motifs should be evaluated. A systematic scan of all k-mer combinations is not practical, however motif combinations of particular interest may be readily assessed. Our ability to re-discover most of the known yeast regulatory motif repertoire confirms that, at least for this simple organism, most motifs give a strong enough signal when operating alone. It is possible that for higher organisms

which likely employ more sophisticated TF combinatorics, a higher proportion of motifs would be missed when searching solely on the individual motif level.

4.1.4 Further applications

The regulatory motif dictionaries have many possible applications, some of which are demonstrated here. By comparing stress response in distant yeast species, we showed that a major component of this regulatory program is evolutionary conserved; In both *S. cerevisiae* and *Candida albicans* similar sequence motifs regulate sets of orthologous genes in a similar manner following stress. In both organisms there are also species specific motifs. The same analysis can be expanded to other biological conditions and to additional organisms in order to reveal the degree of conservation of different eukaryotic regulatory programs.

The individual motifs in our collections comprise the 'building blocks' of sophisticated regulatory networks. They can thus be used for the study of logical gates operating between TFs, and for elucidating higher levels of expression regulation.

An extremely important application illustrated in this work, is the use of the motif dictionaries for studying the phenotypic effects of binding site variations, as discussed bellow.

4.2 Functional characterization of binding site variations

Two important properties of the defined dictionary motifs make them particularly suitable for the study of binding site variations (i) A quantitative link is formed between a motif's nucleotide sequence and the effect it exerts on expression (ii) The motifs are defined as single k-mers, which can subsequently be compared to one another or grouped together according to both sequence and function. Such grouping may reveal new biological insight, which is lost when the starting point of the motif search is a pre-determined PWM.

These advantages were utilized by us in two complimentary projects within this work: Firstly we clustered motifs that are similar both in their nucleotide sequence and in the set of biological conditions they appear to regulate (as judged by coherent expression of their target genes at these conditions). This revealed that motifs with highly similar sequences may operate at distinct sets of biological conditions [29]. Secondly within a given biological condition, we compared the regulatory effects of highly similar sequence motifs, and noticed that they may differ considerably.

These observations may be reasoned if the related sites are bound by different TFs. Alternatively similar sites may be bound by the same TF, yet yield different regulatory outputs, due to different affinities of the TF to the different sites, or to different co-factors interacting with the TF in each case. Whichever is the case, it is clear that sequence considerations alone are not sufficient for predicting a motif's regulatory outcome, as motifs differing at a single position may exert different regulatory affects.

This recognition is highly relevant to an important task; It is often desirable to asses the effects on gene expression of motif variations which are present in the population, or in the genomes of related species. Such assessment would distinguish 'neutral' binding site variations (which do not alter expression) from 'deleterious' ones (which alter expression and may ultimately cause disease). We established a systematic method to predict the phenotypic effects of binding site variations. The idea is simple and appealing: Such effects can be inferred from comparing the regulatory output of related motifs that are present in the same genome. The motif dictionaries are an invaluable source for such comparisons; The regulatory effect of every dictionary motif was compared to that of all k-mers that differ from it by a single nucleotide.

Applying such comparisons to dictionary motifs which correspond to the consensus sequence of known yeast binding sites, we were able to produce reliable predictions. Particularly for Ndt80, a pivotal yeast sporulation factor, our predictions were in good agreement with experimental results in which the site was systematically mutated at each position and the effect on protein binding and expression was monitored. By accumulating statistics for many substitutions across multiple binding sites we observed that not all nucleotide substitutions are similar in severity: In the *S. cerevisiae* genome abolishing a C or a G has a harsher effect on average than abolishing an A or a T. Although this result may be specific to the AT rich *S. cerevisiae* genome, the same method can be easily applied to other genomes, and specifically to human.

The power of our approach stems from its huge statistics – thousands of genes in hundreds of expression time points are utilized, with hundreds of motifs and an even higher number of variations on them. To our disadvantage is the fact that we cannot control for differences in the promoter backgrounds in which our motif variants are embedded (i.e. differences that are outside of the substituted position). The fact that

we obtain statistically significant differences between the effects of different types of substitutions on expression likely indicates that despite uncontrolled sources of variation we extracted genuine signals.

We show that other characteristics, in addition to nucleotide identity, such as the information content of a binding site position are predictive of its sensitivity towards substitution. An intriguing follow-up on this study would be to test the predictive ability of additional features such as the evolutionary conservation of a position within a binding motif and its proximity to the protein in the DNA-protein complex. It is anticipated that high conservation and greater proximity to the binding proteins, two factors which themselves are likely correlated, will be also correlated with the severity of the substitutions [43, 46, 48]. Many such features may be ultimately integrated in order to form a prioritization scheme that would allow the ranking of existing genome variations by their disease-causing potential.

An additional application may be in algorithms that assign PWMs to promoters (c.f. PRIMA [126]) as it should provide means to differentially weigh mismatches between the PWM preferences and the promoter sequence, based on the expected effect on expression. Particularly, at least in the *S. cerevisiae* genome if a PWM contains a C or a G in a given position, and an A or a T in another, potential targets that deviate from the consensus in the first of the two position types are less likely to be assigned to the motif compared to targets that deviate from it in the second of these types.

4.3 Regulation through antisense transcripts

4.3.1 Regulation of the regulator

In recent years, there has been a revolution in our understanding of the regulatory role of non-encoding RNAs. Genome-wide technologies reveal that a significant proportion of all genomes is transcribed, and might thus fulfill regulatory functions [156]. The possibility that transcribed RNAs represent leakage of the transcription machinery exists, but evidence for a selected process is convincing. In our recently published ‘Concept Paper’, we have discussed one type of non-encoding RNA, natural antisense transcripts, and suggested that its transcriptional, and post-transcriptional, regulation is tailored to its various regulatory roles [107].

4.3.2 A human trans-encoded NAT network

Genome-wide computational efforts have shown that NATs are widely prevalent in the genomes of many species. In parallel, evidence accumulated from multiple studies of individual genes, suggests that NATs play an important role in the regulation of gene expression. It is commonly accepted to distinguish between *cis* and *trans* encoded NATS, based on both the genomic locations from which they are encoded and the types of regulatory networks they form. Here we challenge this common distinction by demonstrating that the same NAT may potentially target transcripts in both *cis* and *trans*. We have discovered a putative genome-wide regulatory network, in human, which exhibits many-to-many relations: A given NAT may have multiple mRNA targets and a given mRNA may serve as a potential target of more than one *trans*-encoded NAT.

We found several particular biological functions to be enriched among the genes belonging to this network, suggesting that they may indeed be subject to common regulation. These functions include GO categories related to: transporter activity, vesicular transport, enzymatic activity, and response to external stimulus through signal transduction. Intriguingly, a similar set of functional categories was recently reported to be enriched in the *trans*-antisense network of *Arabidopsis thaliana* [129]. This is a remarkable correspondence that may represent convergence to similar regulatory regimes of functionally related genes in extremely remote organisms. Alternatively, such regimes may have diverged from a common ancestral regulatory mechanism. This scenario is less likely to be the case, as regulatory mechanisms are rarely conserved through such long evolutionary distances. In fact the *Arabidopsis trans*-antisense network was compared to the corresponding networks of two related plants poplar and rice [129]; About half of the transcripts involved in the *Arabidopsis* network, had an ortholog involved in the networks of poplar or rice. However only one *Arabidopsis trans*-NAT pair maintained both transcripts and pairing relationships in the other two plants. For all other NAT pairs even if both transcripts had corresponding orthologs in the related species, the pairing was different. This finding supports the convergence model. It suggests that antisense regulation may be important for only one transcript of a *trans*-NAT pair and that different regulators may have been recruited in different species.

NATs are known to regulate their conjugate sense transcripts at several levels including transcription, mRNA processing, splicing, mRNA stability, mRNA transport and translation (as summarized in our recent paper [107]). In *Arabidopsis* the potential roles of *trans*-NATs in regulating alternative splicing and inducing gene silencing were explored. Here we report an additional observation (which was not tested in plant) that transcripts involved in sense-antisense pairing are predominantly localized to membrane fractions, cytoplasm, cytoskeleton filaments and various vesicles. This finding leads us to hypothesize a potential role of human *trans*-antisense in the control of mRNA intercellular localization.

While in the past mRNAs were thought to be translated exclusively in the cytoplasm, there is today accumulating evidence that mRNAs may be transported to specific cellular locations and translated there upon demand [157, 158]. mRNA localization should be more cost-efficient than protein transport as one mRNA molecule can give rise to several protein molecules. In addition localization of the mRNA limits the presence of the protein at any location other than the target site. It is thus expected that mRNA localization should be tightly regulated.

Localization likely occurs by active transport along the cytoskeleton and is mediated by RNA-binding proteins which couple the mRNA to the localization machinery. These proteins recognize *cis* regulatory elements primarily located at the 3'UTR. Because we observe localization of our transcripts both to cytoskeleton filaments and to distinct membranal organelles, we hypothesize that *trans*-NATs may regulate mRNA localization, for instance by masking the corresponding *cis* elements.

In summary: the interlaced relationships observed between *cis*- and *trans*-NAT pairs suggest that antisense transcripts may form complex regulatory networks, governing distinct cellular processes in organisms as distant as plant and human. There remains a possibility that, notwithstanding the sequence complementarity, the two transcripts of a putative *trans*-NAT pair are not related and rarely form RNA-RNA duplexes within the cell. However the vast potential of *trans*-NAT pairs to form duplexes was recently demonstrated experimentally for RNA extracted from human cells [159]. It thus seems likely that at least some of our putative NAT pairs should form duplexes *in vivo*.

References

1. Berg, O.G. and P.H. von Hippel, *Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters.* J Mol Biol, 1987. **193**(4): p. 723-50.
2. Stormo, G.D., *DNA binding sites: representation and discovery.* Bioinformatics, 2000. **16**(1): p. 16-23.
3. Galas, D.J. and A. Schmitz, *DNAse footprinting: a simple method for the detection of protein-DNA binding specificity.* Nucleic Acids Res, 1978. **5**(9): p. 3157-70.
4. Fried, M. and D.M. Crothers, *Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis.* Nucleic Acids Res, 1981. **9**(23): p. 6505-25.
5. Garner, M.M. and A. Revzin, *A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system.* Nucleic Acids Res, 1981. **9**(13): p. 3047-60.
6. Ren, B., F. Robert, et al., *Genome-wide location and function of DNA binding proteins.* Science, 2000. **290**(5500): p. 2306-9.
7. Lee, T.I., N.J. Rinaldi, et al., *Transcriptional regulatory networks in Saccharomyces cerevisiae.* Science, 2002. **298**(5594): p. 799-804.
8. Harbison, C.T., D.B. Gordon, et al., *Transcriptional regulatory code of a eukaryotic genome.* Nature, 2004. **431**(7004): p. 99-104.
9. Carey, M. and S.T. Smale, *Transcriptional regulation in eukaryotes: concepts, strategies, and techniques*. 1st ed. 2000, Cold Spring Harbor: Cold Spring Harbor Laboratory Press. 640.
10. Tavazoie, S., J.D. Hughes, et al., *Systematic determination of genetic network architecture.* Nat Genet, 1999. **22**(3): p. 281-5.
11. Brazma, A., I. Jonassen, et al., *Predicting gene regulatory elements in silico on a genomic scale.* Genome Res, 1998. **8**(11): p. 1202-15.
12. Wolfsberg, T.G., A.E. Gabrielian, et al., *Candidate regulatory sequence elements for cell cycle-dependent transcription in Saccharomyces cerevisiae.* Genome Res, 1999. **9**(8): p. 775-92.
13. Cliften, P., P. Sudarsanam, et al., *Finding functional features in Saccharomyces genomes by phylogenetic footprinting.* Science, 2003. **301**(5629): p. 71-6.
14. Kellis, M., N. Patterson, et al., *Sequencing and comparison of yeast species to identify genes and regulatory elements.* Nature, 2003. **423**(6937): p. 241-54.
15. McCue, L.A., W. Thompson, et al., *Factors influencing the identification of transcription factor binding sites by cross-species comparison.* Genome Res, 2002. **12**(10): p. 1523-32.
16. McGuire, A.M., J.D. Hughes, and G.M. Church, *Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes.* Genome Res, 2000. **10**(6): p. 744-57.
17. Wasserman, W.W., M. Palumbo, et al., *Human-mouse genome comparisons to locate regulatory sites.* Nat Genet, 2000. **26**(2): p. 225-8.
18. Weitzman, J.B., *Tracking evolution's footprints in the genome.* J Biol, 2003. **2**(2): p. 9.
19. Dermitzakis, E.T. and A.G. Clark, *Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover.* Mol Biol Evol, 2002. **19**(7): p. 1114-21.

20. Doniger, S.W. and J.C. Fay, *Frequent Gain and Loss of Functional Transcription Factor Binding Sites.* PLoS Comput Biol, 2007. **3**(5): p. e99.
21. Lapidot, M. and Y. Pilpel, *Comprehensive quantitative analyses of the effects of promoter sequence elements on mRNA transcription.* Nucleic Acids Res, 2003. **31**(13): p. 3824-8.
22. Pilpel, Y., P. Sudarsanam, and G.M. Church, *Identifying regulatory networks by combinatorial analysis of promoter elements.* Nat Genet, 2001. **29**(2): p. 153-9.
23. Sudarsanam, P., Y. Pilpel, and G.M. Church, *Genome-wide Co-occurrence of Promoter Elements Reveals a cis-Regulatory Cassette of rRNA Transcription Motifs in Saccharomyces cerevisiae.* Genome Res, 2002. **12**(11): p. 1723-31.
24. Elnitski, L., V.X. Jin, et al., *Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques.* Genome Res, 2006. **16**(12): p. 1455-64.
25. Bussemaker, H.J., H.J. Bussemaker, et al., *Regulatory element detection using correlation with expression.* Nat Genet, 2001. **27**(2): p. 167-71.
26. Roth, F.P., J.D. Hughes, et al., *Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.* Nat Biotechnol, 1998. **16**(10): p. 939-45.
27. Bailey, T.L. and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* Proc Int Conf Intell Syst Mol Biol, 1994. **2**: p. 28-36.
28. Solan, Z., D. Horn, et al., *Unsupervised learning of natural languages.* Proc Natl Acad Sci U S A, 2005. **102**(33): p. 11629-34.
29. Segal, L., M. Lapidot, et al., *Nucleotide variation of regulatory motifs may lead to distinct expression patterns.* Bioinformatics, 2007. **in press**.
30. Kel, A., O. Kel-Margoulis, et al., *Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells.* J Mol Biol, 1999. **288**(3): p. 353-76.
31. Halfon, M.S., Y. Grad, et al., *Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model.* Genome Res, 2002. **12**(7): p. 1019-28.
32. Berman, B.P., Y. Nibu, et al., *Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.* Proc Natl Acad Sci U S A, 2002. **99**(2): p. 757-62.
33. Yuh, C.H., H. Bolouri, and E.H. Davidson, *Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene.* Science, 1998. **279**(5358): p. 1896-902.
34. Hurgin, V., D. Novick, and M. Rubinstein, *The promoter of IL-18 binding protein: activation by an IFN-gamma -induced complex of IFN regulatory factor 1 and CCAAT/enhancer binding protein beta.* Proc Natl Acad Sci U S A, 2002. **99**(26): p. 16957-62.
35. Buchler, N.E., U. Gerland, and T. Hwa, *On schemes of combinatorial transcription logic.* Proc Natl Acad Sci U S A, 2003. **100**(9): p. 5136-41.
36. Johnson, N.A. and A.H. Porter, *Rapid speciation via parallel, directional selection on regulatory genetic pathways.* J Theor Biol, 2000. **205**(4): p. 527-42.
37. King, M.C. and A.C. Wilson, *Evolution at two levels in humans and chimpanzees.* Science, 1975. **188**(4184): p. 107-16.

38. Tanay, A., I. Gat-Viks, and R. Shamir, *A global view of the selection forces in the evolution of yeast cis-regulation.* Genome Res, 2004. **14**(5): p. 829-34.
39. MATCH™ [http://www.gene-regulation.com/pub/programs.html#match]
40. Bailey, T.L. and M. Gribskov, *Combining evidence using p-values: application to sequence homology searches.* Bioinformatics, 1998. **14**(1): p. 48-54.
41. Prokunina, L., C. Castillejo-Lopez, et al., *A regulatory polymorphism in PDCD1 is associated with susceptibility to systemic lupus erythematosus in humans.* Nat Genet, 2002. **32**(4): p. 666-9.
42. Zwarts, K.Y., S.M. Clee, et al., *ABCA1 regulatory variants influence coronary artery disease independent of effects on plasma lipid levels.* Clin Genet, 2002. **61**(2): p. 115-25.
43. Ng, P.C. and S. Henikoff, *Predicting deleterious amino acid substitutions.* Genome Res, 2001. **11**(5): p. 863-74.
44. Ng, P.C. and S. Henikoff, *Accounting for human polymorphisms predicted to affect protein function.* Genome Res, 2002. **12**(3): p. 436-46.
45. Ng, P.C. and S. Henikoff, *SIFT: predicting amino acid changes that affect protein function.* Nucleic Acids Res, 2003. **31**(13): p. 3812-4.
46. Wang, Z. and J. Moult, *SNPs, protein structure, and disease.* Hum Mutat, 2001. **17**(4): p. 263-70.
47. Sunyaev, S., V. Ramensky, and P. Bork, *Towards a structural basis of human non-synonymous single nucleotide polymorphisms.* Trends Genet, 2000. **16**(5): p. 198-200.
48. Sunyaev, S., V. Ramensky, et al., *Prediction of deleterious human alleles.* Hum Mol Genet, 2001. **10**(6): p. 591-7.
49. Vitkup, D., C. Sander, and G.M. Church, *The amino-acid mutational spectrum of human genetic disease.* Genome Biol, 2003. **4**(11): p. R72.
50. Rockman, M.V. and G.A. Wray, *Abundant raw material for cis-regulatory evolution in humans.* Mol Biol Evol, 2002. **19**(11): p. 1991-2004.
51. Chiang, D.Y., A.M. Moses, et al., *Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts.* Genome Biol, 2003. **4**(7): p. R43.
52. Force, A., M. Lynch, et al., *Preservation of duplicate genes by complementary, degenerative mutations.* Genetics, 1999. **151**(4): p. 1531-45.
53. Kafri, R., A. Bar-Even, and Y. Pilpel, *Transcription control reprogramming in genetic backup circuits.* Nat Genet, 2005. **37**(3): p. 295-9.
54. Ludwig, M.Z., C. Bergman, et al., *Evidence for stabilizing selection in a eukaryotic enhancer element.* Nature, 2000. **403**(6769): p. 564-7.
55. Sengupta, A.M., M. Djordjevic, and B.I. Shraiman, *Specificity and robustness in transcription control networks.* Proc Natl Acad Sci U S A, 2002. **99**(4): p. 2072-7.
56. Wagner, A., *Robustness against mutations in genetic networks of yeast.* Nat Genet, 2000. **24**(4): p. 355-61.
57. Li, Y.Y., L. Qin, et al., *In silico discovery of human natural antisense transcripts.* BMC Bioinformatics, 2006. **7**: p. 18.
58. Katayama, S., Y. Tomaru, et al., *Antisense transcription in the mammalian transcriptome.* Science, 2005. **309**(5740): p. 1564-6.
59. Barrell, B.G., G.M. Air, and C.A. Hutchison, 3rd, *Overlapping genes in bacteriophage phiX174.* Nature, 1976. **264**(5581): p. 34-41.

60. Tomizawa, J., T. Itoh, et al., *Inhibition of ColE1 RNA primer formation by a plasmid-specified small RNA.* Proc Natl Acad Sci U S A, 1981. **78**(3): p. 1421-5.
61. Wagner, E.G. and R.W. Simons, *Antisense RNA control in bacteria, phages, and plasmids.* Annu Rev Microbiol, 1994. **48**: p. 713-42.
62. Knee, R. and P.R. Murphy, *Regulation of gene expression by natural antisense RNA transcripts.* Neurochem Int, 1997. **31**(3): p. 379-92.
63. Kumar, M. and G.G. Carmichael, *Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes.* Microbiol Mol Biol Rev, 1998. **62**(4): p. 1415-34.
64. Williams, T. and M. Fried, *A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3' ends.* Nature, 1986. **322**(6076): p. 275-9.
65. Chen, J., M. Sun, et al., *Over 20% of human transcripts might form sense-antisense pairs.* Nucleic Acids Res, 2004. **32**(16): p. 4812-20.
66. Misra, S., M.A. Crosby, et al., *Annotation of the Drosophila melanogaster euchromatic genome: a systematic review.* Genome Biol, 2002. **3**(12): p. RESEARCH0083.
67. Yelin, R., D. Dahary, et al., *Widespread occurrence of antisense transcription in the human genome.* Nat Biotechnol, 2003. **21**(4): p. 379-86.
68. Sun, M., L.D. Hurst, et al., *Evidence for variation in abundance of antisense transcripts between multicellular animals but no relationship between antisense transcription and organismic complexity.* Genome Res, 2006.
69. Carninci, P., T. Kasukawa, et al., *The transcriptional landscape of the mammalian genome.* Science, 2005. **309**(5740): p. 1559-63.
70. Kiyosawa, H., I. Yamanaka, et al., *Antisense transcripts with FANTOM2 clone set and their implications for gene regulation.* Genome Res, 2003. **13**(6B): p. 1324-34.
71. Osato, N., H. Yamada, et al., *Antisense transcripts with rice full-length cDNAs.* Genome Biol, 2003. **5**(1): p. R5.
72. Jen, C.H., I. Michalopoulos, et al., *Natural antisense transcripts with coding capacity in Arabidopsis may have a regulatory role that is not linked to double-stranded RNA degradation.* Genome Biol, 2005. **6**(6): p. R51.
73. Galante, P.A., D.O. Vidal, et al., *Sense-antisense pairs in mammals: functional and evolutionary considerations.* Genome Biol, 2007. **8**(3): p. R40.
74. Dahary, D., O. Elroy-Stein, and R. Sorek, *Naturally occurring antisense: transcriptional leakage or real overlap?* Genome Res, 2005. **15**(3): p. 364-8.
75. Cawley, S., S. Bekiranov, et al., *Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs.* Cell, 2004. **116**(4): p. 499-509.
76. Impey, S., S.R. McCorkle, et al., *Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions.* Cell, 2004. **119**(7): p. 1041-54.
77. The ENCODE Project: ENCyclopedia Of DNA Elements [http://www.genome.gov/10005107]
78. Trinklein, N.D., U. Karaoz, et al., *Integrated analysis of experimental data sets reveals many novel promoters in 1% of the human genome.* Genome Res, 2007. **17**(6): p. 720-31.

79. Chen, J., M. Sun, et al., *Genome-wide analysis of coordinate expression and evolution of human cis-encoded sense-antisense transcripts.* Trends Genet, 2005. **21**(6): p. 326-9.
80. Sun, M., L.D. Hurst, et al., *Evidence for a preferential targeting of 3'-UTRs by cis-encoded natural antisense transcripts.* Nucleic Acids Res, 2005. **33**(17): p. 5533-43.
81. Chen, J., M. Sun, et al., *Human antisense genes have unusually short introns: evidence for selection for rapid transcription.* Trends Genet, 2005. **21**(4): p. 203-7.
82. Hurst, L.D., G. McVean, and T. Moore, *Imprinted genes have few and small introns.* Nat Genet, 1996. **12**(3): p. 234-7.
83. Castillo-Davis, C.I., S.L. Mekhedov, et al., *Selection for short introns in highly expressed genes.* Nat Genet, 2002. **31**(4): p. 415-8.
84. Alon, U., *An introduction to systems biology - Design Principles of Biological Circuits*. 1st ed. 2006: Chapman & Hall/CRC.
85. Cohen, B.A., R.D. Mitra, et al., *A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression.* Nat Genet, 2000. **26**(2): p. 183-6.
86. Lercher, M.J., T. Blumenthal, and L.D. Hurst, *Coexpression of neighboring genes in Caenorhabditis elegans is mostly due to operons and duplicate genes.* Genome Res, 2003. **13**(2): p. 238-43.
87. Yanai, I., D. Graur, and R. Ophir, *Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control.* Omics, 2004. **8**(1): p. 15-24.
88. Lavorgna, G., D. Dahary, et al., *In search of antisense.* Trends Biochem Sci, 2004. **29**(2): p. 88-94.
89. Prescott, E.M. and N.J. Proudfoot, *Transcriptional collision between convergent genes in budding yeast.* Proc Natl Acad Sci U S A, 2002. **99**(13): p. 8796-801.
90. Osato, N., Y. Suzuki, et al., *Transcriptional interferences in cis natural antisense transcripts of human and mouse.* Genetics, 2007.
91. Crampton, N., W.A. Bonass, et al., *Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy.* Nucleic Acids Res, 2006. **34**(19): p. 5416-25.
92. Hastings, M.L., C. Milcarek, et al., *Expression of the thyroid hormone receptor gene, erbAalpha, in B lymphocytes: alternative mRNA processing is independent of differentiation but correlates with antisense RNA levels.* Nucleic Acids Res, 1997. **25**(21): p. 4296-300.
93. Yan, M.D., C.C. Hong, et al., *Identification and characterization of a novel gene Saf transcribed from the opposite strand of Fas.* Hum Mol Genet, 2005. **14**(11): p. 1465-74.
94. Bass, B.L., *RNA editing by adenosine deaminases that act on RNA.* Annu Rev Biochem, 2002. **71**: p. 817-46.
95. Neeman, Y., D. Dahary, et al., *Is there any sense in antisense editing?* Trends Genet, 2005. **21**(10): p. 544-7.
96. Meister, G. and T. Tuschl, *Mechanisms of gene silencing by double-stranded RNA.* Nature, 2004. **431**(7006): p. 343-9.
97. Mello, C.C. and D. Conte, Jr., *Revealing the world of RNA interference.* Nature, 2004. **431**(7006): p. 338-42.

98. Borsani, O., J. Zhu, et al., *Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis.* Cell, 2005. **123**(7): p. 1279-91.
99. Duhring, U., I.M. Axmann, et al., *An internal antisense RNA regulates expression of the photosynthesis gene isiA.* Proc Natl Acad Sci U S A, 2006.
100. Aravin, A.A., N.M. Naumova, et al., *Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline.* Curr Biol, 2001. **11**(13): p. 1017-27.
101. Wang, X.J., T. Gaasterland, and N.H. Chua, *Genome-wide prediction and identification of cis-natural antisense transcripts in Arabidopsis thaliana.* Genome Biol, 2005. **6**(4): p. R30.
102. Tufarelli, C., J.A. Stanley, et al., *Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease.* Nat Genet, 2003. **34**(2): p. 157-65.
103. Ogawa, Y. and J.T. Lee, *Antisense regulation in X inactivation and autosomal imprinting.* Cytogenet Genome Res, 2002. **99**(1-4): p. 59-65.
104. Corcoran, A.E., *Immunoglobulin locus silencing and allelic exclusion.* Semin Immunol, 2005. **17**(2): p. 141-54.
105. Sleutels, F., G. Tjon, et al., *Imprinted silencing of Slc22a2 and Slc22a3 does not need transcriptional overlap between Igf2r and Air.* Embo J, 2003. **22**(14): p. 3696-704.
106. Thakur, N., V.K. Tiwari, et al., *An antisense RNA regulates the bidirectional silencing property of the Kcnq1 imprinting control region.* Mol Cell Biol, 2004. **24**(18): p. 7855-62.
107. Lapidot, M. and Y. Pilpel, *Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms.* EMBO Rep, 2006. **7**(12): p. 1216-22.
108. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* J. Roy Stat Soc, 1995. **57**: p. 289-300.
109. ExpressDEb [http://arep.med.harvard.edu/cgi-bin/ExpressDByeast/EXDStart]
110. Shalgi, R., M. Lapidot, et al., *A catalog of stability-associated sequence elements in 3' UTRs of yeast mRNAs.* Genome Biology, 2005. **6**(10): p. R86 1-15.
111. Gene Ontology [http://www.geneontology.org/]
112. Hughes, J.D., P.W. Estep, et al., *Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae.* J Mol Biol, 2000. **296**(5): p. 1205-14.
113. Cunningham, T.S. and T.G. Cooper, *The Saccharomyces cerevisiae DAL80 repressor protein binds to multiple copies of GATAA-containing sequences (URSGATA).* J Bacteriol, 1993. **175**(18): p. 5851-61.
114. Davidson, E., *Genomic regulatory systems*. 2001, San Diego, Calif: Academic Press.
115. Benos, P.V., M.L. Bulyk, and G.D. Stormo, *Additivity in protein-DNA interactions: how good an approximation is it?* Nucleic Acids Res, 2002. **30**(20): p. 4442-4451.
116. Bulyk, M.L., P.L. Johnson, and G.M. Church, *Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors.* Nucleic Acids Res, 2002. **30**(5): p. 1255-61.

117. Tomovic, A. and E.J. Oakeley, *Position dependencies in transcription factor binding sites.* Bioinformatics, 2007.
118. Maerkl, S.J. and S.R. Quake, *A systems approach to measuring the binding energy landscapes of transcription factors.* Science, 2007. **315**(5809): p. 233-7.
119. Lefstin, J.A. and K.R. Yamamoto, *Allosteric effects of DNA on transcriptional regulators.* Nature, 1998. **392**(6679): p. 885-8.
120. Jones, T., N.A. Federspiel, et al., *The diploid genome sequence of Candida albicans.* Proc Natl Acad Sci U S A, 2004. **101**(19): p. 7329-34.
121. Enjalbert, B., A. Nantel, and M. Whiteway, *Stress-induced gene expression in Candida albicans: absence of a general stress response.* Mol Biol Cell, 2003. **14**(4): p. 1460-7.
122. Whitfield, M.L., G. Sherlock, et al., *Identification of genes periodically expressed in the human cell cycle and their expression in tumors.* Mol Biol Cell, 2002. **13**(6): p. 1977-2000.
123. TRANSFAC [http://www.gene-regulation.com/pub/databases.html#transfac]
124. Lamoureux, J.S., D. Stuart, et al., *Structure of the sporulation-specific transcription factor Ndt80 bound to DNA.* Embo J, 2002. **21**(21): p. 5721-32.
125. Pierce, M., K.R. Benjamin, et al., *Sum1 and Ndt80 proteins compete for binding to middle sporulation element sequences that control meiotic gene expression.* Mol Cell Biol, 2003. **23**(14): p. 4814-25.
126. Elkon, R., C. Linhart, et al., *Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells.* Genome Res, 2003. **13**(5): p. 773-80.
127. Ryals, J., P. Dierks, et al., *A 46-nucleotide promoter segment from an IFN-alpha gene renders an unrelated promoter inducible by virus*, in *Cell*. 1985. p. 497-507.
128. Altschul, S.F., W. Gish, et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
129. Wang, H., N.H. Chua, and X.J. Wang, *Prediction of trans-antisense transcripts in Arabidopsis thaliana.* Genome Biol, 2006. **7**(10): p. R92.
130. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic Acids Res, 2007. **35**(Database issue): p. D61-5.
131. Su, A.I., T. Wiltshire, et al., *A gene atlas of the mouse and human protein-encoding transcriptomes.* Proc Natl Acad Sci U S A, 2004. **101**(16): p. 6062-7.
132. Hornstein, E. and N. Shomron, *Canalization of development by microRNAs.* Nat Genet, 2006. **38 Suppl 1**: p. S20-4.
133. Gerdes, K., T. Thisted, and J. Martinussen, *Mechanism of post-segregational killing by the hok/sok system of plasmid R1: sok antisense RNA regulates formation of a hok mRNA species correlated with killing of plasmid-free cells.* Mol Microbiol, 1990. **4**(11): p. 1807-18.
134. Saccharomyces Genome Database [http://www.yeastgenome.org/]
135. Cho, R.J., M.J. Campbell, et al., *A genome-wide transcriptional analysis of the mitotic cell cycle.* Mol Cell, 1998. **2**(1): p. 65-73.
136. Roberts, C.J., B. Nelson, et al., *Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles.* Science, 2000. **287**(5454): p. 873-80.

137. Spellman, P.T., G. Sherlock, et al., *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.* Mol Biol Cell, 1998. **9**(12): p. 3273-97.
138. Jelinsky, S.A., P. Estep, et al., *Regulatory networks revealed by transcriptional profiling of damaged saccharomyces cerevisiae cells: rpn4 links base excision repair with proteasomes.* Mol Cell Biol, 2000. **20**(21): p. 8157-67.
139. Chu, S., J. DeRisi, et al., *The transcriptional program of sporulation in budding yeast.* Science, 1998. **282**(5389): p. 699-705.
140. Gasch, A.P., P.T. Spellman, et al., *Genomic expression programs in the response of yeast cells to environmental changes.* Mol Biol Cell, 2000. **11**(12): p. 4241-57.
141. Causton, H.C., B. Ren, et al., *Remodeling of yeast genome expression in response to environmental changes.* Mol Biol Cell, 2001. **12**(2): p. 323-37.
142. UCSC Genome Browser [http://genome.ucsc.edu/]
143. Candida Genome Database [http://www.candidagenome.org/]
144. Heyer, L.J., S. Kruglyak, and S. Yooseph, *Exploring expression data: identification and analysis of coexpressed genes.* Genome Res, 1999. **9**(11): p. 1106-15.
145. Lord, P.W., R.D. Stevens, et al., *Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.* Bioinformatics, 2003. **19**(10): p. 1275-83.
146. Xie, X., J. Lu, et al., *Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.* Nature, 2005.
147. Quandt, K., K. Frech, et al., *MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data.* Nucleic Acids Res, 1995. **23**(23): p. 4878-84.
148. Braganca, J. and A. Civas, *Type I interferon gene expression: differential expression of IFN-A genes induced by viruses and double-stranded RNA.* Biochimie, 1998. **80**(8-9): p. 673-87.
149. Yeow, W.S., W.C. Au, et al., *Reconstitution of virus-mediated expression of interferon alpha genes in human fibroblast cells by ectopic interferon regulatory factor-7.* J Biol Chem, 2000. **275**(9): p. 6313-20.
150. Lenhard, B. and W.W. Wasserman, *TFBS: Computational framework for transcription factor binding site analysis.* Bioinformatics, 2002. **18**(8): p. 1135-6.
151. Fickett, J.W., *Quantitative discrimination of MEF2 sites.* Mol Cell Biol, 1996. **16**(1): p. 437-41.
152. Wain, H.M., M.J. Lush, et al., *Genew: the Human Gene Nomenclature Database, 2004 updates.* Nucleic Acids Res, 2004. **32**(Database issue): p. D255-7.
153. Yanai, I., H. Benjamin, et al., *Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification.* Bioinformatics, 2005. **21**(5): p. 650-9.
154. Rabiner, L.R. and B.H. Juang, *An introduction to hidden Markov models.* IEEE ASSP Magazine, 1986. **3**: p. 4-16.
155. Segal, E., Y. Fondufe-Mittendorf, et al., *A genomic code for nucleosome positioning.* Nature, 2006. **442**(7104): p. 772-8.
156. Carninci, P., *Tagging mammalian transcription complexity.* Trends Genet, 2006. **22**(9): p. 501-510.

157. St Johnston, D., *Moving messages: the intracellular localization of mRNAs.* Nat Rev Mol Cell Biol, 2005. **6**(5): p. 363-75.
158. Du, T.G., M. Schmid, and R.P. Jansen, *Why cells move messages: the biological functions of mRNA localization.* Semin Cell Dev Biol, 2007. **18**(2): p. 171-7.
159. Rosok, O. and M. Sioud, *Systematic identification of sense-antisense transcripts in mammalian cells.* Nat Biotechnol, 2004. **22**(1): p. 104-8.
160. Eisen, M.B., P.T. Spellman, et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.

Publications Resulting from this Work

Refereed articles

- Segal, L.*, **Lapidot, M.***, Solan Z., Ruppin, E., Pilpel, Y. and Horn, D. *Nucleotide variation of regulatory motifs may lead to distinct expression patterns*. ISMB 2007 in Bioinformatics 2007. (In press)
  * These authors contributed equally to this work.

- **Lapidot, M.** & Pilpel Y. Characterization of the effects of TF binding site variations on Gene Expression. Towards predicting the functional outcomes of regulatory SNPs. RECOMB 2005 Workshop on Regulatory Genomics, LNBI 4023 proceedings, Springer-Verlag, pp. 51-61, 2007.

- **Lapidot, M.** & Pilpel Y. *Genome-wide natural antisense transcription: Coupling its regulation to its different regulatory mechanisms.* EMBO Reports 2006 Dec;7(12):1216-22.

- Shalgi, R., **Lapidot, M**., Shamir, R. and Pilpel Y. *A catalog of stability-associated sequence elements in 3' UTRs of yeast mRNAs*. Genome Biology 2005 6(10): R86

- **Lapidot, M**. & Pilpel Y. *Comprehensive quantitative analyses of the effects of promoter sequence elements on mRNA transcription.* Nucleic Acids Res. 2003 31(13): 3824-3828.

Submitted manuscripts

- **Lapidot, M.** & Pilpel Y. *Functional characterization of variations on regulatory motifs.* Submitted to PLoS Genetics.

Independent efforts and collaborations

I hereby declare that this thesis summarizes my independent efforts.
Two projects were performed in collaboration, and the individual contributions are detailed bellow:

- The syntax-based *S. cerevisiae* dictionary [29]

Zach Solan ran MEX over *S. cerevisiae* promoter sequences, I conducted EC analysis for the MEX-extracted motifs across 40 biological conditions (constructed motif dictionaries), Liat Segal performed clustering and cluster analysis.

- Analysis of interferon-α promoters

I conducted all computational analyses, experiments were performed by Roni Golan.

Appendix I – Abbreviations

BLAST - basic local alignment search tool
bp - base pair
ChIP - chromatin immunoprecipitation
ChIP-chip - ChIP followed by DNA sequence identification using DNA microarrays
dsRNA – double stranded RNA
EC - expression coherence
EMSA- electrophoretic mobility shift essays
ENCODE - encyclopedia of DNA elements
EST - expressed sequence tag
FDR - false discovery rate
GO - gene ontology
IC - information content
IFNs - interferons
IRF - interferon regulatory factor
MEX - motif extraction algorithme
mRNA - messenger RNA
MY - million years
NAT - Natural antisense transcript
ORF - open reading frame
PWM - positional weight matrix
RNAi – RNA interference
rRNA - ribosomal RNA
SNP - sequence nucleotide polymorphism
rSNP - regulatory sequence nucleotide polymorphism
TF - transcription factor
TFBS - transcription factor binding site
TSS - transcriptional start site
UTR – untranslated region

Appendix II – 40 biological conditions

Below are short descriptions of the 40 conditions from which expression data was gathered. These datasets were downloaded from ExpressDB [109].

1. Cho – Mitotic cell cycle [135]
2. Chu – Sporulation [139]
3. Environmental response – Acid [141] (same reference for condition 3-8)
4. Environmental response – Alkali
5. Environmental response –Heat
6. Environmental response -NaCl
7. Environmental response - Peroxide
8. Environmental response - Sorbitol
9. Eisen – Cold shock [160] (same reference for conditions 11-12)
10. Gasch environmental response- Diauxic shift [140] (same for conditions 14-35)
11. Eisen - Dtt
12. Eisen - Heat
13. Jelinsky - DNA Damage [138]
14. Gasch environmental response - 37ºC -25 ºC shock
15. Gasch environmental response - Amino acid starvation
16. Gasch environmental response – Diamide (sulfhydryl-oxidizing agent)
17. Gasch environmental response - Dtt1
18. Gasch environmental response - Dtt2
19. Gasch environmental response - Heat shock 1, 25 ºC-37ºC
20. Gasch environmental response- Heat shock 29 ºC -33 ºC YPD +1M sorbitol
21. Gasch environmental response- Heat shock 29 ºC -33 ºC YPD
22. Gasch environmental response- Heat shock 29 ºC -33 ºC YPD +1M sobitol to YPD
23. Gasch environmental response- Heat shock 2 25 ºC-37ºC
24. Gasch environmental response- constant H_2O_2 (hydrogen peroxide)
25. Gasch environmental response- Menadione (superoxide-generating drug)
26. Gasch environmental response- Hypo-osmotic
27. Gasch environmental response- Nitrogen Depletion
28. Gasch environmental response- Sorbitol (Hyper-Osmotic)
29. Gasch environmental response- Heat shock, from various temperatures (17º, 21 º, 25 º, 29 º, 33 º) to 37ºC.
30. Gasch environmental response- Growth at various temperatures (17º, 21 º, 25 º, 29 º, 33 º, 37 º)
31. Gasch environmental response- Growth at various temperatures
32. Gasch environmental response- X media versus carbon source 1
33. Gasch environmental response- YPD1 25 º C
34. Gasch environmental response- YPD2 30 º C
35. Gasch environmental response- YPx media versus carbon source 2
36. MapK - monitor signal transduction during yeast pheromone response [136]
37. Spellman cell-cycle alpha factor arrest [137] (same for conditions 38-40)
38. Spellman cell-cycle cdc15
39. Spellman cell-cycle cdc28 (reanalysis of Cho's experiment, condition 1)
40. Spellman cell-cycle eluteration

Appendix III – Supplementary web files

The following supplementary files can be found at:

http://longitude.weizmann.ac.il/~lapidotm/PhDThesis

1. Table S1: Significantly scoring *S. cerevisiae* k-mers (our yeast motif dictionary). Lists our 8,610 dictionary motifs, along with their EC scores and p-values in the biological condition in which each motif obtained the most significant score. For motifs that matched at least one of Harbison's PWMs with a match score higher than 99, the highest scoring match is also listed.

2. Table S2: Regulatory motif content of IFN-α promoters.
(i) Lists for each one of the 13 IFN-α promoters, the locations of each of the 45 selected binding sites we searched for (ii) Lists all promoter pairwise similarities.

3. Interface to MySql database containing all sense-antisense pairs forming the human *trans*-antisense network.

4. Links to all published work along with supporting websites (when available).