Sackler Faculty of Exact Sciences, School of Computer Science

Discovering motifs in large genomic datasets

THESIS SUBMITTED FOR THE DEGREE OF
"DOCTOR OF PHILOSOPHY"

by

Chaim Linhart

The work on this thesis has been carried out
under the supervision of
Prof. Ron Shamir
and **Prof. Yosef Shiloh**

Submitted to the Senate of Tel-Aviv University
July 2009

Acknowledgments

This dissertation summarizes most of my research in the last five years. I would like to express my sincere thanks to my advisor, Ron Shamir, for his guidance, advice and support, and for giving me academic freedom to pursue my research interests.

I would like to thank all my friends and collaborators in the Computational Genomics lab. First and foremost, it was a real pleasure to work with Yonit Halperin, a gifted programmer and computer scientist, on the Amadeus and Allegro projects and on the TLRs and *C. elegans* motif pair studies. Lots of thanks to Igor Ulitsky for his contribution to Allegro and for many helpful comments on various issues. Special thanks to Rani Elkon, a good friend with whom I collaborated many times, for sharing with me his knowledge in biology and intelligent insights. I enjoyed lots of constructive and interesting discussions with other members of the lab, especially: Michal Ozery-Flato, Ofer Lavi, Adi Maron-Katz, Seagull Shavit and Michal Ziv-Ukelson.

I am grateful to Yossi Shiloh, my second advisor, for his advice and support in several studies. I would also like to acknowledge additional collaborators on various projects, some of which are not included in this thesis: Shahar Kidron, Limor Broday, Amir Darom (TAU, Faculty of Medicine), Marian Walhout (University of Massachusetts Medical School), Raphaël Clifford (University of Bristol) and Gidi Weber (The Hebrew University of Jerusalem).

Last but not least, I would like to thank my family. Thanks to my parents for their love and support throughout all my academic studies. This work is dedicated to my dear wife, Nami, who helped and encouraged me in so many ways. Finally, I would like to mention my greatest achievements during the past five years, our wonderful children – Yoav, Noga and Yonatan – you're the best!

Preface

This thesis is based on the following collection of six articles that were published throughout the PhD period in scientific journals.

1. **Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets.**
Chaim Linhart*, Yonit Halperin* and Ron Shamir.
Published in *Genome Research* [1].

2. **Allegro: Analyzing expression and sequence in concert to discover regulatory programs.**
Yonit Halperin*, Chaim Linhart*, Igor Ulitsky and Ron Shamir.
Published in *Nucleic Acids Research* [2].

3. **Deciphering transcriptional regulatory elements that encode specific cell-cycle phasing by comparative genomics analysis.**
Chaim Linhart, Ran Elkon, Yosef Shiloh and Ron Shamir.
Published in *Cell Cycle* [3].

4. **Functional genomic delineation of TLR-induced transcriptional networks.**
Ran Elkon*, Chaim Linhart*, Yonit Halperin, Yosef Shiloh and Ron Shamir.
Published in *BMC Genomics* [4].

5. **Faster pattern matching with character classes using prime number encoding.**
Chaim Linhart and Ron Shamir.
Published in *Journal of Computer and System Sciences* [5].

6. **Matching with don't-cares and a small number of mismatches.**
Chaim Linhart and Ron Shamir.
Published in *Information Processing Letters* [6].

(* marks equal contribution)

Abstract

A major challenge in system biology is to delineate the regulatory program of a genome, which describes how the cell controls the amount and exact composition of the proteins it produces from each gene in a given circumstance. A key element in this effort is the computational task of motif discovery, which calls for the identification of recurring sequence patterns, or motifs, in genomic sequences; such motifs represent binding sites of transcription factors and microRNAs, two central components of the cellular transcriptional regulatory program. We studied the practical and theoretical aspects of motif finding. We developed new algorithms, computational models and statistical scores for de-novo motif discovery, and implemented them in an efficient, user-friendly software package. Our approach outperforms existing methods and is applicable to a wide range of motif finding tasks. We dissected the transcriptional programs of two pivotal mammalian cellular processes – the cell cycle and the innate immune response. By using ad-hoc techniques that utilize multiple sources of information, we obtained remarkably high accuracy of binding site prediction, and revealed new regulatory links and modules. On the theoretical side, we developed new efficient algorithms for solving several types of pattern matching problems that are related to motif finding. Our methods employ ideas from number theory, and are conceptually simpler, and in some settings faster, than extant algorithms.

Contents

1. Introduction

1.1 The post-genomic era

In recent years biological data have been accumulating in an overwhelming pace. A key milestone was reached in 2001 with the publication of the first draft of the human genome. Today, public databases contain the full DNA sequences of dozens of species, including model organisms such as yeast, fly and mouse, enabling researchers to apply comparative genomics techniques that analyze the similarities and differences between genomes in order to gain further insight on the structure and evolution of genes. Sequencing the human genome and detecting the genes within it and the proteins they encode are indeed a giant leap for biology, but in our efforts to fully understand the genetic cellular mechanisms, it is just the first step.

The living cell is an amazingly complex machine, constantly performing a myriad of biochemical reactions to sustain itself and carry out a variety of functions in a diverse and ever-changing environment. A single cell operates in many different ways in various conditions and times, for example, during cell-cycle progression or stress response. Different cells in an organism share exactly the same DNA, but nevertheless perform a variety of unique tasks. In order to understand how this machinery works, we need to determine the function of each protein and the interactions among them. Thanks to the maturation of high-throughput experimental techniques, we now have tools with which we can tackle these difficult questions. One of the major technologies in functional genomics is the DNA chip, or microarray, which simultaneously measures the mRNA expression levels of thousands of genes in a given cell-line [7]. Thus, a single chip experiment yields a genome-wide snapshot of the mRNA concentrations in the cell. Comparison of gene expression profiles under various biological conditions and in different time points could indirectly reveal the role some genes play in the studied processes. Large volumes of microarray data are publicly available for many types of cells and biological conditions. Alas, such data do not reveal *how* the cell modifies expression levels and what biochemical interactions each protein participates in, in order to fulfill its task. To do so, we have to infer the entire genetic network of cellular processes that determine how protein concentrations are controlled by other proteins or modified as a result of external stimuli, how proteins interact to form complexes with unique functions, and how this huge number of diverse elements play in concert to generate viable consistent temporal cellular behaviors.

1.2 Transcriptional regulation

The cell is equipped with several tools for regulating the amount and exact composition of the proteins it produces from each gene in a given circumstance - chromatin state, RNA interference (RNAi), RNA editing, and alternative splicing, to name a few. Perhaps the main regulatory mechanism is the transcriptional program, which describes when and to what extent each gene is transcribed to mRNA. Transcription is controlled primarily via regulatory sequence elements, located in the proximity of each gene's coding sequence, that are recognized and bound by specialized proteins, called *transcription factors* (TFs). The set of TFs that bind to the DNA, and the intensity, or *affinity*, of these bindings, may increase or decrease the rate of transcription of the corresponding gene. Thus, different combinations of TFs and binding affinities could produce a huge variety of transcription profiles.

The DNA sequences bound by a TF are called its *binding sites* (BSs), or *cis*-regulatory elements. They are typically very short (6-15 bases) and degenerate - a TF can bind, with varying affinities, to many different sequences that reflect a common pattern, or *motif*, characteristic of the factor. Most BSs are found in the *promoter*, the region upstream of the gene's transcription start site (TSS), though BSs may also exist downstream of the TSS and at large distance from the gene. A special type of TFBSs is the core promoter elements, which are bound by general-purpose TFs and reside very close to the TSS of many genes. TATA-box and INR (Initiator) are well-known examples. Other TFs usually bind further away from the TSS and have more specific function. For example, E2F is a pivotal regulator of cell-cycle progression in human. Some TFs cooperate in the regulation of genes, resulting in more complex and specific transcription profiles. Such recurring sets of TFs are termed *cis*-regulatory modules, or simply *modules*. Reverse-engineering the transcriptional program of an organism requires identifying its TFs, the locations and affinities of their BSs, and the various modules they are organized in. As expected, many aspects of the transcription program, such as the number of TFs, diversity of BSs, and size of modules, are especially complex in higher eukaryotes.

1.2.1 Models for binding site motifs

Identifying the sites bound in-vivo by a specific TF under certain conditions is not an easy task. Methods like DNA footprinting or chromatin immunoprecipitation (ChIP) can be used, but are applicable only to short, hand-chosen genomic loci. The combined strategy of ChIP and promoter microarrays, also termed ChIP-chip, enables genome-wide identification of promoter segments that are bound by specific TFs, in a single experimental assay [8]. Replacing the microarray-based readout with next-generation sequencing technologies, an approach called ChIP-seq, allows the detection of BSs throughout the entire genome [9].

Examination of biologically validated BS sequences reveals the common characteristics shared by the BSs of each TF, as well as their high degree of diversity. This has led to the development of several computational models that attempt to describe BS motifs by simple mathematical constructs. These models can be divided into two main types – pattern-based models, and profile-based models. A simple pattern-based model is the *consensus string*, which consists of a k-mer w (that is, a word w of length k, where k is the length of the motif, typically ranging between 6 and 15), and an integer d that defines the maximum number of allowed mismatches. A k-mer along a *cis*-regulatory sequence is considered a *hit*, i.e., a putative BS, if its Hamming distance from w is d or less. Another way for describing shared patterns is using a *degenerate string*, which specifies the allowed set of nucleotides at each position. Degenerate strings are usually described using the IUPAC code [10]. For example, the motif `AYGAN` represents 8 k-mers, since `Y` stands for two bases (`C`,`T`), and `N` stands for all four bases. The most popular profile-based model is the *position weight matrix* (PWM), also known as position specific scoring matrix (PSSM). This model uses a k×4 frequency matrix $f_{i,b}$ to represent the motif, where $f_{i,b}$ is the probability for observing nucleotide b at position i in the motif. The probability that a given k-mer $w = w_1w_2\ldots w_k$ is a functional BS is simply the product of the corresponding matrix elements, i.e., $\prod_{i=1}^{k} f_{i,w_i}$. The frequency matrix $f_{i,b}$ is often converted into log-likelihood ratios: $P_{i,b} = \log (f_{i,b}/p_b)$, where p_b is the background frequency of base b ([11] gives several justifications for this formula). The similarity score $S(w)$ of the k-mer w to the PWM $P_{i,b}$ is: $S(w) = \sum_{i=1}^{k} P_{i,w_i}$. A k-mer w is considered a hit if $S(w)$ is above some fixed cutoff. Alternatively, the score $S(w)$ can represent the binding affinity of w. Using validated BSs as training sets, parameters for TFBS models have been derived for scores of known TFs in various species, and deposited in databases such as TRANSFAC [12] and JASPAR [13].

1.2.2 MicroRNA binding sites

In addition to TFs, another key regulatory mechanism is controlled by microRNAs (miRNAs), short non-coding RNA molecules. Annealing of a miRNA to its target mRNA, typically in its 3′ untranslated region (3' UTR), triggers the degradation of the mRNA transcript or inhibits protein translation. As with TFs, the BSs of a miRNA share a common motif, typically complementary to 6-8 nucleotides in the miRNA's seed (positions 1-8 in the 5' end of the miRNA). The exact parameters that define when miRNAs bind to their target genes and with what efficacy appear quite complex, and are still not entirely understood [14]. To date, thousands of miRNAs have been identified in dozens of species, and their sequences are accessible via online repositories, such as miRBase [15]. Based on these data, several computational tools attempt to predict the target genes of known miRNAs, e.g., TargetScan [16] and PicTar [17].

For brevity of this introduction, we shall focus on TFs in the following sections; very similar computational challenges and methods apply also for miRNAs, with appropriate modifications.

1.3 Motif finding

In parallel to the advancements in high-throughput experimental techniques, many computational methods have been developed in order to analyze data obtained from these experiments and suggest novel biological hypotheses, which can then be tested by further experiments. As mentioned above, experimentally identifying BSs on a genomic scale is laborious and expensive, considering the large number of TFs and BSs, the vast genomic regions that contain active BSs, and the fact that TFs may bind at different loci under changing conditions. Computational tools offer a cheap and efficient means for locating BSs - by scanning promoter sequences using a given motif, each subsequence is assigned a score that indicates how similar it is to the motif; subsequences whose score is above some threshold are counted as hits, i.e., putative BSs. Unfortunately, since BSs are short and degenerate, the genome contains many subsequences that are identical or very similar to functional BSs, but are merely random occurrences. Thus, extant motif models do not contain enough information to locate functional BSs accurately - at thresholds low enough to recover a large percentage of the true sites, many false positive hits are also reported. Apparently, TFs are guided to their in-vivo binding sites by contextual factors, such as chromatin structure and interactions with other TFs, in addition to their innate DNA binding preferences.

1.3.1 Motif discovery tasks and methods

Several approaches have been proposed in order to increase the specificity of motif search. The most popular practice is to limit the scan to a subset of genes that are known or believed to be co-regulated, e.g., genes that exhibit similar expression profiles in microarray experiments, or genes that have a common biological function. The comparative genomics approach, called phylogenetic footprinting, utilizes information from multiple organisms. Since TFBSs play an important biological role, selective pressure causes them to evolve at a slower rate than non-functional intergenic sequences. In other words, subsequences along orthologous promoters that are significantly conserved among related species are more likely to be active BSs. Other techniques for reducing the number of false-positive hits in a motif scan is to search for modules, i.e., identify BSs of several TFs that tend to co-occur in the same genes, possibly in a fixed order or at conserved distances, or to locate regions with high density of BSs. Examples of tools for BS prediction are MatInspector [18], MotifScanner [19], Cluster-Buster [20], CisOrtho [21], and MONKEY [22].

In addition to locating new putative BSs, computational tools are also often used in order to suggest which TFs regulate a given set of genes. In a common scenario, genes are grouped according to their expression profiles obtained from microarray assays, and the promoters of each group are scanned for *enriched* TFs, i.e., TFs whose hits are statistically over-represented in the promoters in comparison to some background model or to a supplied reference set of genes. This method associates TFs with well-defined biological processes, connecting precious pieces in the transcription network puzzle. PRIMA [23], Clover [24], ROVER [25], OTFBS [26], COMET [27], and CREME [28] fall into this category.

An even more challenging problem than locating new potential BSs of known TFs or determining which TFs are enriched in a group of genes is to identify motifs of yet unknown TFs. As in the previous scenario, we are given a set of co-regulated genes, and we wish to find motifs that are statistically enriched in their promoters, only this time we are not supplied with a list of well-characterized motifs of known TFs. Instead, our goal is to recover novel motifs. Once found, further biological research must be performed in order to discover the proteins, whose BSs are described by these motifs. *De-novo* motif discovery has been tackled using a myriad of algorithmic techniques, such as Expectation Maximization (MEME [29], EMnEM [30], OrthoMEME [31], PhyME [32]), Gibbs sampling (GibbsDNA [33], AlignACE [34], MotifSampler [35]), efficient enumeration (YMF [36], MITRA [37], Multiprofiler [38], WEEDER [39], FootPrinter [40], FIRE [41], Trawler [42]), and neural networks (ANN-Spec [43]), as well as greedy (CONSENSUS [44]), graph-based (WINNOWER and SP-STAR [45]), and randomized (PROJECTION [46]) methods.

It is worthy to note that motif finding is not limited to *cis*-regulatory elements, such as TF and miRNA BSs. The genome is abundant with recurring patterns that correspond to DNA-protein interactions or other biochemical processes. For example, interesting (though not always well-understood) motifs can be found in introns surrounding alternatively-spliced exons, and in recombination hot-spots.

1.3.2 Motif finding and gene expression data

In the above scenarios, we assumed that the set of co-regulated genes is known in advance, or inferred from another source of information, such as gene expression microarrays. In the latter case, the following two-step approach is most commonly used (see examples in [47, 48] and the review in [49]): In the first step, a clustering procedure is executed to partition the genes into groups believed to be co-regulated, based on expression profile similarity [50]. In the second step, a motif discovery tool is applied to search for abundant sequence patterns in the promoters of each group that may represent the BSs of TFs that regulate the corresponding genes. Since both the expression profiles and the promoter sequences of the genes carry information regarding their regulation, a methodology that utilizes both sources of information may give better results than the two-step approach.

Several studies proposed computational schemes for this parallel analysis. Most of these algorithms use a unified probabilistic model over both gene expression and sequence data, and assume a Gaussian distribution of the expression values [51-53]. Additional examples are the algorithms Reduce [54] and Motif Regressor [55], which search for motifs correlated with a *single* condition using linear regression, and assume that the number of BSs and their affinity are linearly correlated with the gene's expression. The algorithm DRIM [56] uses the hypergeometric (HG) score to compute the enrichment of motif occurrences among the top-ranked genes. However, it too is limited to a single condition.

Despite extensive molecular and computational research, it remains exceedingly difficult to accurately predict BSs and discover novel motifs. State-of-the-art software tools often yield unsatisfactory results on large, complex datasets, especially in metazoans [57].

1.4 Pattern matching in computational biology

A basic building-block of BS prediction and motif finding algorithms is a procedure for locating all occurrences of a given pattern along all the *cis*-regulatory sequences of one or more species. In a typical setting, the motif length is 6-15 and the total length of the regulatory regions is roughly 10^8 per genome. Pattern matching also arises in other types of tasks in computational biology, such as degenerate primer design for Polymerase Chain Reaction (PCR) experiments. PCR is a technique for amplifying a specific region of DNA, so that enough copies of it are available for testing or sequencing. The first step in PCR is to synthesize two DNA segments, or *primers*, lying on opposite sides of the target region. As with TFBS motifs, a PCR primer is called degenerate if some of its positions have several possible bases. Thus, a degenerate primer can be described as a pattern with character classes. Degenerate primers can be used to amplify several related genomic sequences in a single PCR experiment. We previously studied the computational problem of designing highly degenerate primers [58], and applied our algorithms in experiments for studying the human and canine olfactory receptor genes [59, 60]. A common problem in the design of degenerate primers is to verify that the primers do not bind to DNA regions others than those they are meant to amplify. Thus, one needs to search for all occurrences of a candidate primer, typically of length 20-30, in the entire genome ($6 \cdot 10^9$ in human). One may also want to allow a small number of mismatches, as the PCR technique usually tolerates a few mismatches.

The pattern matching problem requires finding all occurrences of a pattern p of length m in a text t of length n ($m<n$) over a finite alphabet Σ. Classical string matching, where both p and t are strings, can be solved in linear time using algorithms such as Knuth-Morris-Pratt [61] and Boyer-Moore [62]. Alas, pattern matching becomes much more difficult in many real-life applications, such as those described above - namely, when the pattern is a degenerate string, and when we allow mismatches. Developing efficient algorithms for these types of pattern matching problems is interesting from the theoretical computer science perspective, and may yield important practical improvements as well.

1.4.1 Matching with don’t-cares

A first step in generalizing simple string matching is obtained when we allow the pattern and the text to contain don't-care characters, or wildcards, denoted '*', which match all symbols in Σ. In the context of TFBS motifs, a pattern with don’t-cares is a simple model that allows a single nucleotide per position and an arbitrary number of gaps (i.e., positions, usually marked by ‘N’, in which any nucleotide is allowed). Matching with don’t-cares can be solved using the *match-count* algorithm, which finds the number of matching positions (or, equivalently, the Hamming distance) between the pattern and every length m substring of the text (see, e.g., Chapter 4.3 in [63]). The algorithm, first introduced by Fischer and Paterson [64], computes the contribution of each alphabet symbol to the score independently, as follows. For the symbol $a \in \Sigma$, each occurrence of a in the text and in the pattern is replaced by the number 1, and all other symbols are encoded by 0. The number of matching a's between the pattern and every substring in the text is obtained by computing the convolution between the binary-encoded pattern and text. Using Fast Fourier Transform (FFT), the convolution can be computed in $O(n \log m)$ time under the RAM model of computation, which assumes that arithmetic operations on numbers with w bits take constant time, where $w=O(\log N)$ is the RAM word size and N is the maximal input size. Thus, the total running time of match-count is $O(|\Sigma|\, n \log m)$, as it involves $|\Sigma|$ such convolutions.

The above time complexity was improved over the years using various FFT-based methods. However, removing the dependence on $|\Sigma|$ remained an open problem until recently. Cole and Hariharan were the first to obtain an $O(n \log m)$ time deterministic algorithm [65], which was simplified by Clifford and Clifford [66].

1.4.2 Matching with character classes

Matching with don't-cares can be further generalized by allowing the pattern to contain any non-empty subset, or *class*, of characters at each position. As described earlier, a degenerate PCR primer is a pattern with character classes over the DNA alphabet: $\Sigma=\{A,C,G,T\}$. Another example of a pattern with character classes is the degenerate (IUPAC) string model for a TFBS motif. Most algorithms for matching with character classes have the same worst-case running time as the naïve algorithm – $O(nm)$. Bit-parallelism techniques improve this to $O(nm/w)$, where w is the RAM word size [67]. In general, the best worst-case performance is attained by the match-count algorithm - $O(|\Sigma|\, n \log m)$. It remains an open question whether the dependency on $|\Sigma|$ can be improved (it is unlikely that a speed-up for the $n\log m$ factor can be attained using similar convolution-based techniques, since this will require improving the time complexity of FFT).

1.4.3 Matching with don’t-cares and mismatches

Another important generalization of simple string matching is that of approximate pattern matching, where one would like to find all locations in the text that match the pattern up to a small pre-specified distance. Perhaps the most widely used metric is the Hamming distance, which counts the number of mismatched pattern symbols. As mentioned above, this problem arises in bioinformatics when designing degenerate primers and when searching for occurrences of a TFBS motif – in both cases, we would like to report all locations that match the pattern with at most k mismatches, where k is a small pre-specified number. Currently, the most efficient method for solving pattern matching with k mismatches runs in time $O(n\sqrt{k\log k})$ [68]. As in the case of exact matching, searching for approximate matches becomes much more difficult when we allow don't-cares. This variant, called matching with don't-cares and k mismatches, can be solved using the match-count algorithm in time $O(|\Sigma|\, n \log m)$, or using Abrahamson’s technique, which combines match-count with a divide-and-conquer procedure, in time $O(n\sqrt{m\log m})$ [69]. Very recently, Clifford et al. devised randomized and deterministic FFT-based algorithms that run in time $O(n(k+\log n\ \log\log n)\ \log m)$ and $O(nk^2\log^3 m)$, respectively [70]. The latter was improved by the same authors to $O(nk\ \log^2 m\ (\log^2 k+\log\log m))$ using methods from algebraic coding theory [71]. Even for small values of k, these algorithms are asymptotically slower than matching with don’t-cares and no mismatches. This raises the question whether matching with don’t-cares and k mismatches can be solved in $O(\text{poly}(k)\ n \log m)$ time. Specifically, can we generalize the best known $O(n \log m)$ time complexity of exact matching with don’t-cares so that it still applies when we allow a fixed number of mismatches?

1.5 Summary of articles included in this thesis

1. **Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets.**
   Chaim Linhart, Yonit Halperin and Ron Shamir.
   Published in *Genome Research* [1].

We present a threefold contribution to the computational task of motif discovery, a key component in the effort of delineating the regulatory map of a genome: (1) We constructed a comprehensive large-scale, publicly-available compendium of transcription factor and microRNA target gene sets derived from diverse high-throughput experiments in several metazoans. We used the compendium as a benchmark for motif discovery tools. (2) We developed Amadeus, a highly efficient, user-friendly software platform for genome-scale detection of novel motifs, applicable to a wide range of motif discovery tasks. Amadeus improves upon extant tools in terms of accuracy, running time, output information, and ease of use and is the only program that attained a high success rate on the metazoan compendium. (3) We demonstrate that by searching for motifs based on their genome-wide localization or chromosomal distributions (without using a predefined target set), Amadeus uncovers diverse known phenomena, as well as novel regulatory motifs.

2. **Allegro: Analyzing expression and sequence in concert to discover regulatory programs.**
   Yonit Halperin, Chaim Linhart, Igor Ulitsky and Ron Shamir.
   Published in *Nucleic Acids Research* [2].

A major goal of system biology is the characterization of transcription factors and microRNAs (miRNAs) and the transcriptional programs they regulate. We present Allegro, a method for *de-novo* discovery of *cis*-regulatory transcriptional programs through joint analysis of genome-wide expression data and promoter or 3' UTR sequences. The algorithm uses a novel log-likelihood-based, non-parametric model to describe the expression pattern shared by a group of co-regulated genes. We show that Allegro is more accurate and sensitive than existing techniques, and can simultaneously analyze multiple expression datasets with more than 100 conditions. We apply Allegro on datasets from several species and report on the transcriptional modules it uncovers. Our analysis reveals a novel motif over-represented in the promoters of genes highly expressed in murine oocytes, and several new motifs related to fly development. Finally, using stem-cell expression profiles, we identify three miRNA families with pivotal roles in human embryogenesis.

3. **Deciphering transcriptional regulatory elements that encode specific cell-cycle phasing by comparative genomics analysis.**
   Chaim Linhart, Ran Elkon, Yosef Shiloh and Ron Shamir.
   Published in *Cell Cycle* [3].

Transcriptional regulation is a major tier in the periodic engine that mobilizes cell cycle progression. The availability of complete genome sequences of multiple organisms holds promise for significantly improving the specificity of computational identification of functional elements. Here, we applied a comparative genomics analysis to decipher transcriptional regulatory elements that control cell-cycle phasing. We analyzed genome-wide promoter sequences from 12 organisms, including worm, fly, fish, rodents and human, and identified conserved transcriptional modules that determine the expression of genes in specific cell cycle phases. We demonstrate that a canonical E2F signal encodes for expression highly specific to the G_1/S phase, and that a cis-regulatory module comprising CHR-NF-Y elements dictates expression that is restricted to the G_2 and G_2/M phases. B-Myb binding site signatures occur in many of the CHR-NF-Y target genes, suggesting a specific role for this triplet in the regulation of the cell cycle transcriptional program. Remarkably, E2F signals are conserved in promoters of G_1/S genes in all organisms from worm to human. The CHR-NF-Y module is conserved in promoters of G_2/M regulated genes in all analyzed vertebrates. Our results reveal novel modules that determine specific cell-cycle phasing, and identify their respective putative target genes with remarkably high specificity.

4. **Functional genomic delineation of TLR-induced transcriptional networks.**
   Ran Elkon, Chaim Linhart, Yonit Halperin, Yosef Shiloh and Ron Shamir.
   Published in *BMC Genomics* [4].

The innate immune system is the first line of defense mechanisms protecting the host from invading pathogens such as bacteria and viruses. The innate immunity responses are triggered by recognition of prototypical pathogen components by cellular receptors. Prominent among these pathogen sensors are Toll-like receptors (TLRs). We sought global delineation of transcriptional networks induced by TLRs, analyzing four genome-wide expression datasets in mouse and human macrophages stimulated with pathogen-mimetic agents that engage various TLRs.

Combining computational analysis of expression profiles and cis-regulatory promoter sequences, we dissected the TLR-induced transcriptional program into two major components: the first is universally activated by all examined TLRs, and the second is specific to activated TLR3 and TLR4. Our results point to NF-κB and ISRE-binding transcription factors as the key regulators of the universal and the TLR3/4-specific responses, respectively, and identify novel putative positive and negative feedback loops in these transcriptional programs. Analysis of the kinetics of the induced network showed that while NF-κB regulates mainly an early-induced and sustained response, the ISRE element functions primarily in the induction of a delayed wave. We further demonstrate that co-occurrence of the NF-κB and ISRE elements in the same promoter endows its targets with enhanced responsiveness.

Our results enhance system-level understanding of the networks induced by TLRs and demonstrate the power of genomics approaches to delineate intricate transcriptional webs in mammalian systems. Such systems-level knowledge of the TLR network can be useful for designing ways to pharmacologically manipulate the activity of the innate immunity in pathological conditions in which either enhancement or repression of this branch of the immune system is desired.

5. **Faster Pattern Matching with Character Classes using Prime Number Encoding.**
   Chaim Linhart and Ron Shamir.
   Published in *Journal of Computer and System Sciences* [5].

In *pattern matching with character classes* the goal is to find all occurrences of a pattern of length m in a text of length n, where each pattern position consists of an allowed set of characters from a finite alphabet Σ. We present an FFT-based algorithm that uses a novel prime-numbers encoding scheme, which is $\log n/\log m$ times faster than the fastest extant approaches, which are based on boolean convolutions. In particular, if $m^{|\Sigma|}=n^{O(1)}$, our algorithm runs in time $O(n\log m)$, matching the complexity of the fastest techniques for wildcard matching, a special case of our problem. A major advantage of our algorithm is that it allows a tradeoff between the running time and the RAM word size. Our algorithm also speeds up solutions to approximate matching with character classes problems—

namely, matching with k mismatches and Hamming distance, as well as to the *subset matching* problem.

6. **Matching with don't-cares and a small number of mismatches.**
   Chaim Linhart and Ron Shamir.
   Published in *Information Processing Letters* [6].

In *matching with don't-cares and k mismatches* we are given a pattern of length m and a text of length n, both of which may contain don't-cares (a symbol that matches all symbols), and the goal is to find all locations in the text that match the pattern with at most k mismatches, where k is a parameter. We present new algorithms that solve this problem using a combination of convolutions and a dynamic programming procedure. We give randomized and deterministic solutions that run in time $O(nk^2\log m)$ and $O(nk^3\log m)$, respectively, and are faster than the most efficient extant methods for small values of k. Our deterministic algorithm is the first to obtain an $O(\text{poly}\ (k) \cdot n\log m)$ running time.

2. Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets

Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets

Chaim Linhart,[1] Yonit Halperin,[1] and Ron Shamir[2]

School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

We present a threefold contribution to the computational task of motif discovery, a key component in the effort of delineating the regulatory map of a genome: (1) We constructed a comprehensive large-scale, publicly-available compendium of transcription factor and microRNA target gene sets derived from diverse high-throughput experiments in several metazoans. We used the compendium as a benchmark for motif discovery tools. (2) We developed Amadeus, a highly efficient, user-friendly software platform for genome-scale detection of novel motifs, applicable to a wide range of motif discovery tasks. Amadeus improves upon extant tools in terms of accuracy, running time, output information, and ease of use and is the only program that attained a high success rate on the metazoan compendium. (3) We demonstrate that by searching for motifs based on their genome-wide localization or chromosomal distributions (without using a predefined target set), Amadeus uncovers diverse known phenomena, as well as novel regulatory motifs.

[Supplemental material is available online at www.genome.org. The Amadeus software is available at http://acgt.cs.tau.ac.il/amadeus.]

One of the main cellular regulatory mechanisms is the transcriptional program, which describes when and to what extent each gene is transcribed to mRNA. Transcription is controlled primarily via transcription factors (TFs)—specialized proteins that bind sequence elements, called binding sites (BSs), which are located mainly in each gene's promoter sequence upstream its transcription start site (TSS). Another key regulatory effect is controlled by microRNAs (miRNAs), short noncoding RNA molecules. Annealing of a miRNA to its target mRNA, typically in its 3′ untranslated region (UTR), triggers the degradation of the mRNA transcript or inhibits protein translation. Delineating the regulatory program of a species requires the combination of experimental and computational techniques. To this end, huge volumes of experimental data have been generated in the past decade by means of high-throughput technologies, such as gene expression microarrays (Lockhart and Winzeler 2000) and ChIP-chip location analyses (Wu et al. 2006). In parallel, numerous software tools were developed in order to analyze these data and suggest novel biological hypotheses.

A major computational challenge is identifying recurring sequence patterns, or motifs, in *cis*-regulatory sequences; such motifs represent BSs of TFs/miRNAs. In a typical scenario, given a target set of coregulated genes, one would like to identify TFs whose BSs are statistically overrepresented in the promoters of these genes, compared with some background model or with a supplied reference set of genes. In recent years, a plethora of computational tools have been developed for discovering enriched motifs of known TFs (Elkon et al. 2003; Sharan et al. 2004), as well as for finding novel motifs that represent BSs of yet uncharacterized TFs. The latter task, known as de novo motif discovery, has been tackled using a myriad of algorithmic techniques, such as expectation-maximization (EM) (Bailey and Elkan 1994), Gibbs sampling (Hughes et al. 2000), and efficient enumeration (Pavesi et al. 2001; Sinha and Tompa 2002; Ettwiller et al. 2007). The most common computational models employed by motif finders to describe TF BSs are degenerate (IUPAC) strings (Sinha and Tompa 2002) and position weight matrices (PWMs) (Bailey and Elkan 1994). Commonly used scores for evaluating candidate motifs include likelihood ratio (Bailey and Elkan 1994) and the *Z*-score (Sinha and Tompa 2002; Ettwiller et al. 2007) and hypergeometric (HG) overrepresentation scores (Eskin and Pevzner 2002).

Most studies that describe novel motif discovery algorithms report their success either on synthetically generated data or on a small ad hoc collection of samples constructed by the investigators for their particular analysis. Obviously, such results do not guarantee equally-good performance in many real-life scenarios. Perhaps the most popular large-scale motif finding benchmark is the yeast ChIP-chip data set of Harbison et al. (2004). To the best of our knowledge, the only large-scale metazoan benchmark constructed to date is that of Tompa et al. (2005). In that study, validated TF BSs from the TRANSFAC (Wingender et al. 1996) database were implanted inside real and synthetic promoter sequences. The benchmark contains 52 data sets (eight from yeast, the rest are from metazoans), with an average of seven sequences per data set. Its main drawback is that it does not reflect many real-life scenarios. For example, one would often like to discover motifs in a cluster of coexpressed genes or in a set of sequences bound by a TF in ChIP-chip. In these scenarios, the analyzed set typically consists of dozens or hundreds of genes, of which only an unknown (often modest) fraction contain BSs; moreover, many of the BSs might reside very far from the TSS or in other types of genomic sequences (introns, UTRs, etc.), and the gene set might be regulated by more than one TF. In this work, we constructed the first publicly-accessible, large-scale compendium of metazoan data sets that were obtained by various experimental

[1] These authors contributed equally to this work.
[2] Corresponding author.
E-mail rshamir@tau.ac.il; fax 972-3-640-5384.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.076117.108.

18:1180–1189 ©2008 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/08; www.genome.org

techniques and cover a wide-range of real-life motif discovery scenarios.

Despite extensive research, it remains exceedingly difficult to accurately predict BSs and discover novel motifs, especially in metazoan data sets, due to the short and degenerate nature of BSs, the size and complexity of genetic sequences, and the high levels of noise in results obtained by high-throughput technologies (Tompa et al. 2005). Moreover, most motif discovery tools present only a small amount of information for the discovered motifs, usually in textual format, and cannot analyze large sets of genes due to running time or memory limitations. Perhaps most importantly, a user without advanced computer skills would find it quite difficult to execute some of these software tools and interpret their results.

We developed a new software suite for efficient genome-scale detection of known and novel motifs, called Amadeus (a motif algorithm for detecting enrichment in multiple species). Amadeus evaluates the discovered motifs using one or more of several built-in statistical scores, and is suitable to a broad range of motif finding tasks. It has an intuitive, user-friendly, and highly informative graphical interface. We ran Amadeus on the yeast ChIP-chip benchmark and on our metazoan compendium, and compared the results to those found by five popular motif finding tools. In addition, we used it to perform genome-wide discovery of motifs whose occurrences are localized within human and mouse promoters. This analysis uncovered two novel motifs, both of which are supported by multiple independent studies and are thus likely to represent real BSs of yet uncharacterized TFs. Another type of genome-wide search we performed found motifs whose chromosomal distribution is nonrandom. We believe Amadeus sets a new standard for motif discovery software in terms of accuracy, running time, range of application and ease of use.

Results

Overview of Amadeus

We developed a highly accurate, efficient, and user-friendly motif discovery software, called Amadeus, for finding short sequence patterns that are overrepresented in the promoters or 3′ UTRs of a given set of genes with respect to a large background (BG) set, typically the entire genome. The general architecture of Amadeus is a pipeline of filters, or refinement phases, where each phase receives as input a list of motif candidates and applies an algorithm for refining the list and producing a set of improved candidates, which serve as a starting point for the next phase (Fig. 1). The first phases typically work on a very large number of candidates, such as all possible *k*-mers, and execute simple procedures for choosing the most promising motifs. Successive phases run more complex (and computationally intensive) algorithms in order to converge to better motifs. The default score for evaluating each candidate motif is the HG enrichment score; other scores measure BS localization, strand-bias, and chromosomal preference. Amadeus also searches for pairs of enriched motifs that tend to co-occur in the same sequences and thus represent a putative cooperative *cis*-regulatory module. Finally, a built-in bootstrapping procedure may be applied to correct for multiple testing. See Methods and Supplemental Notes for a detailed description of the algorithm, scores, and additional features.

The output of Amadeus is a nonredundant list of top-scoring motifs. For each motif, a wealth of information is displayed, including the motif's logo, the scores it received, its occurrences localization graph, a list of similar known TF/miRNA motifs from TRANSFAC/miRBase, and the set of genes presumably regulated by the motif (Fig. 2; Supplemental Fig. 2). A graphical TF BS viewer displays the putative BSs of the motifs within the genomic se-

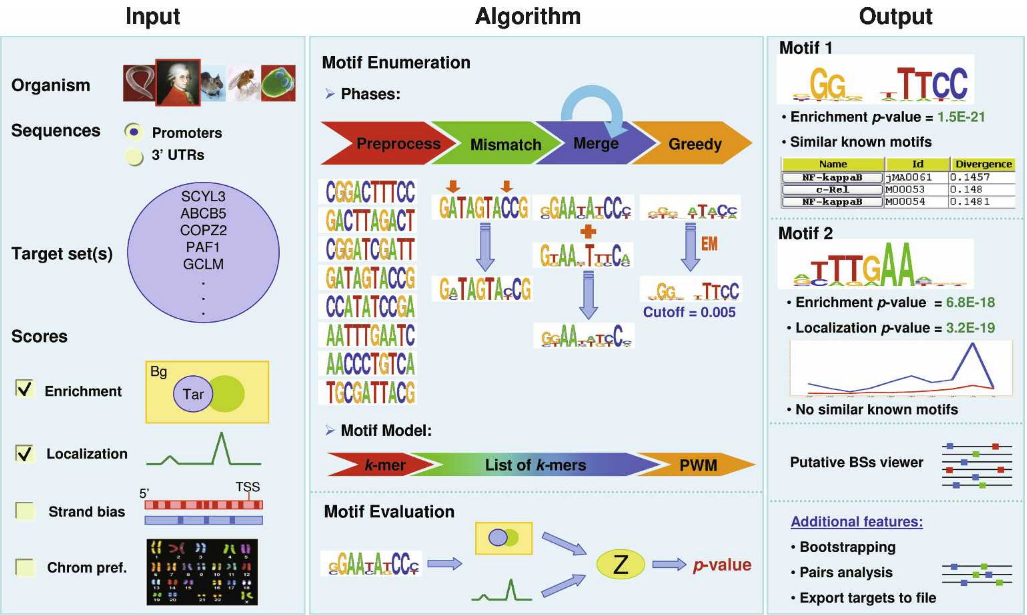
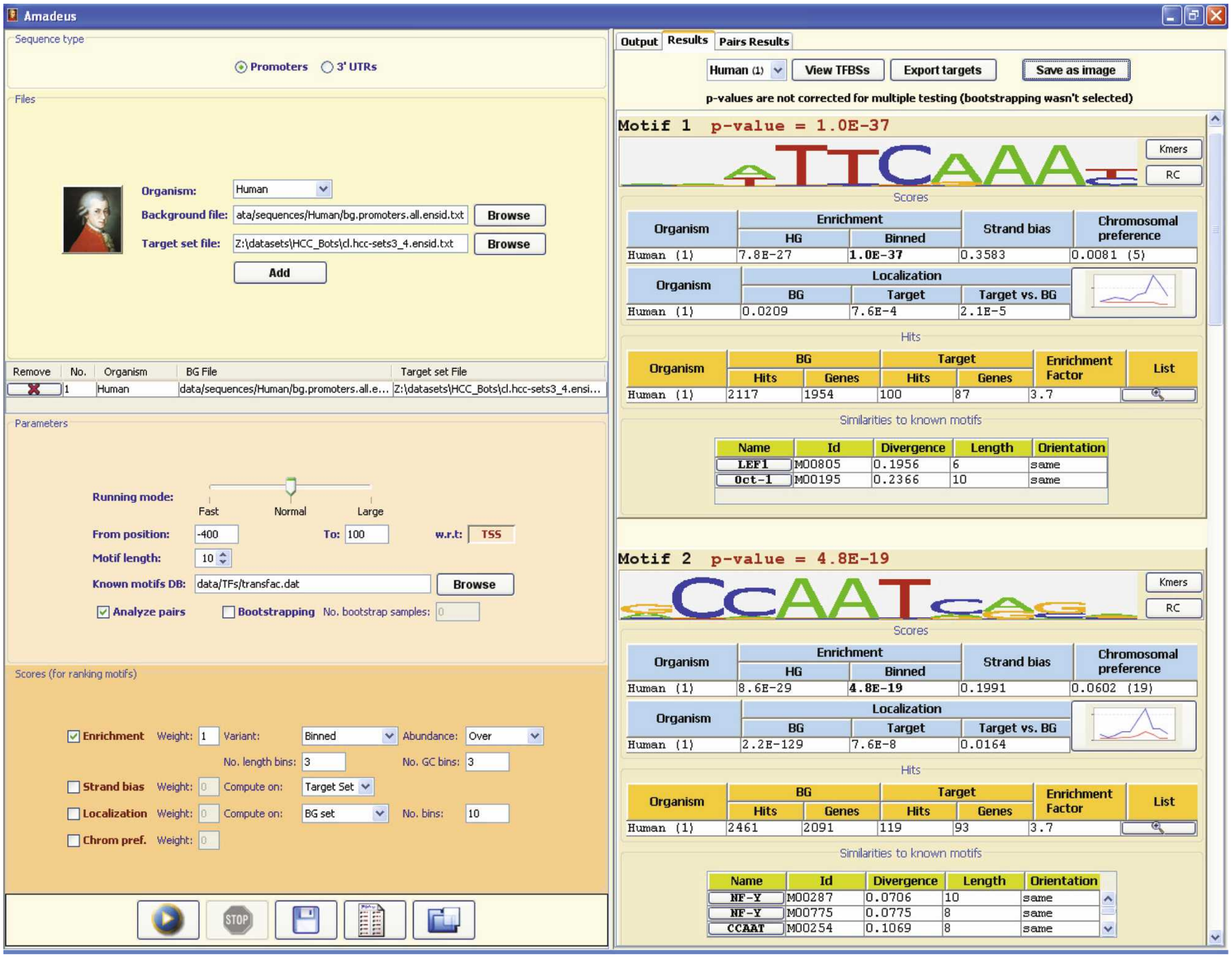


Figure 1. The main components of the Amadeus computational pipeline. The input consists of one or more target gene sets and various parameters such as the score(s) for evaluating the motifs. Starting from all *k*-mers, the algorithm runs a series of refinement phases that eventually converge to a nonredundant list of high-scoring PWMs. These motifs, together with additional information and analyses, are displayed in the graphical output. For more details, see Methods.



Sequence type

● Promoters ○ 3' UTRs

Files

Organism: Human

Background file: ata/sequences/Human/bg.promoters.all.ensid.txt

Target set file: Z:\datasets\HCC_Bots\cl.hcc-sets3_4.ensid.txt

Remove	No.	Organism	BG File	Target set File
☒	1	Human	data/sequences/Human/bg.promoters.all.e...	Z:\datasets\HCC_Bots\cl.hcc-sets3_4.ensi...

Parameters

Running mode: Fast Normal Large

From position: -400 To: 100 w.r.t: TSS

Motif length: 10

Known motifs DB: data/TFs/transfac.dat

☑ Analyze pairs ☐ Bootstrapping No. bootstrap samples: 0

Scores (for ranking motifs)

☑ Enrichment Weight: 1 Variant: Binned Abundance: Over

No. length bins: 3 No. GC bins: 3

☐ Strand bias Weight: 0 Compute on: Target Set

☐ Localization Weight: 0 Compute on: BG set No. bins: 10

☐ Chrom pref. Weight: 0

p-values are not corrected for multiple testing (bootstrapping wasn't selected)

Motif 1 p-value = 1.0E-37

Scores

Organism	Enrichment		Strand bias	Chromosomal preference
	HG	Binned		
Human (1)	7.8E-27	**1.0E-37**	0.3583	0.0081 (5)

Organism	Localization		
	BG	Target	Target vs. BG
Human (1)	0.0209	7.6E-4	2.1E-5

Hits

Organism	BG		Target		Enrichment Factor	List
	Hits	Genes	Hits	Genes		
Human (1)	2117	1954	100	87	3.7	

Similarities to known motifs

Name	Id	Divergence	Length	Orientation
LEF1	M00805	0.1956	6	same
Oct-1	M00195	0.2366	10	same

Motif 2 p-value = 4.8E-19

Scores

Organism	Enrichment		Strand bias	Chromosomal preference
	HG	Binned		
Human (1)	8.6E-29	**4.8E-19**	0.1991	0.0602 (19)

Organism	Localization		
	BG	Target	Target vs. BG
Human (1)	2.2E-129	7.6E-8	0.0164

Hits

Organism	BG		Target		Enrichment Factor	List
	Hits	Genes	Hits	Genes		
Human (1)	2461	2091	119	93	3.7	

Similarities to known motifs

Name	Id	Divergence	Length	Orientation
NF-Y	M00287	0.0706	10	same
NF-Y	M00775	0.0775	8	same
CCAAT	M00254	0.1069	8	same

Figure 2. Screenshot of Amadeus. The *left* panel controls the input parameters (organism, target set, promoter region, scores, etc.). Here, Amadeus was executed on the set of genes expressed in G$_2$ and G$_2$/M phases of the human cell cycle (Whitfield et al. 2002). The top-scoring motif shown in the output panel on the *right* is CHR (cell-cycle genes homology region), a *cis*-regulatory element that was experimentally found in promoters of several G$_2$/M genes (Zhu et al. 2004), and is not represented in TRANSFAC; the second motif is CCAAT-box (NF-Y). For each motif discovered, the output also lists similar patterns from TRANSFAC, information on the localization of its occurrences, and additional statistics. In agreement with recent studies (Linhart et al. 2005; Tabach et al. 2005), the motif-pairs analysis in Amadeus reports the de novo found CHR and NF-Y motifs as a *cis*-regulatory module that is highly specific to the G$_2$ and G$_2$/M cell-cycle phases (Supplemental Fig. 1). A screenshot with additional graphical features is shown in Supplemental Figure 2.

quences. All these data assist the user in dissecting the regulatory network underlying the studied gene set and in focusing on the most promising motifs for further research.

Performance on the yeast transcriptional regulatory map

In their seminal paper, Harbison et al. (2004) constructed a nearly complete map of the transcriptional regulatory code of *Saccharomyces cerevisiae* using ChIP-chip assays. We assessed the performance of Amadeus using the ChIP-chip data of 83 factors that bound more than four promoters (58 on average), and whose binding motifs have been reported in the literature. We executed Amadeus twice on each data set with motif lengths 8 and 10, and compared the two top-scoring motifs obtained from each execution to the corresponding literature motif—as in Harbison et al. (2004), a match was defined if the average Euclidean distance between the columns of the two PWMs, referred henceforth as "PWM divergence," was below 0.18. Amadeus was said to successfully recover a TF BS pattern if at least one of the four motifs it reported (two for each motif length) matched the literature PWM. Under these strict criteria, Amadeus discovered 54 of the known motifs (65% of 83).

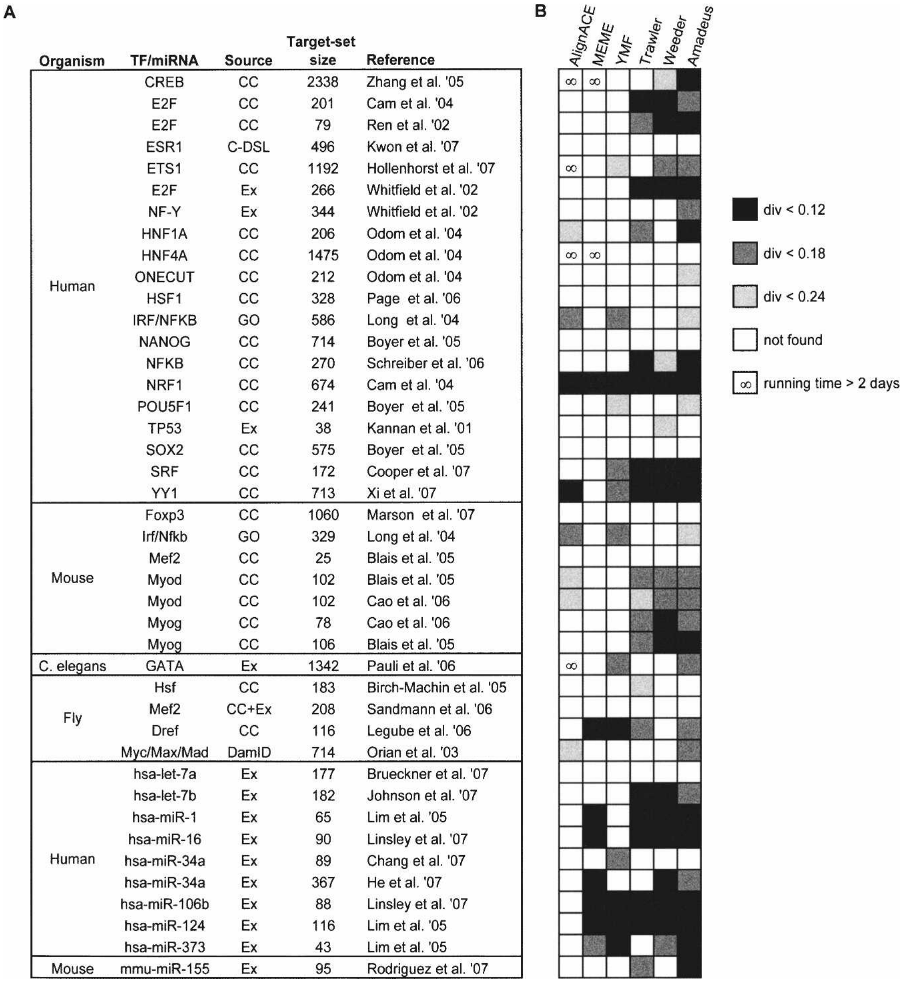
We compared the performance of Amadeus to five popular motif finders that represent an assortment of algorithms and motif evaluation scores—MEME (EM) (Bailey and Elkan 1994), AlignACE (Gibbs sampling) (Hughes et al. 2000), YMF (Sinha and Tompa 2002), Weeder (Pavesi et al. 2001), and Trawler (Ettwiller et al. 2007) (exhaustive search). Of note, Weeder outperformed 13 motif discovery tools by most measures in Tompa's assessment (Tompa et al. 2005), and Trawler was very recently reported to outperform four tools on mammalian data sets (Ettwiller et al. 2007). As in Tompa's study, we did not include in our analysis programs that utilize auxiliary information, such as ChIP binding affinities, known TF BS models, or cross-species sequence conservation. Although Amadeus can incorporate some of this information, we wanted to focus on the core functionality of motif detection that is common to the widest possible range of setups. Each program was run with its default parameters using motif lengths 8 and 10, and the two top-scoring motifs were compared to the correct PWM as described earlier. As shown in Supplemental Figure 3, Amadeus recovered the largest number of motifs (65%); in agreement with Tompa et al. (2005), Weeder outperformed MEME, AlignACE, and YMF, successfully recovering 58% of the PWMs. Interestingly, the performance ranking among all five methods remained unchanged for other PWM divergence cutoffs.

Compendium of target sets of metazoan TFs and miRNAs

Ettwiller et al. (2007) tested the performance of their Trawler program using 10 mammalian ChIP-chip data sets. While this benchmark is larger than most data sets used in the literature, it is still relatively small and represents a single experimental technique. As explained earlier, Tompa's data set (Tompa et al. 2005), the only large-scale metazoan benchmark constructed to date, does not reflect target sets obtained by high-throughput experiments. We therefore set out to construct a comprehensive motif discovery benchmark that is based on a large compendium of experimental studies. We collected diverse types of data sets from several metazoans, as published by independent groups in leading journals. Our compendium, listed in Figure 3, consists of 42 gene sets from human, mouse, fly (*Drosophila melanogaster*), and worm (*Caenorhabditis elegans*) and represents a total of 26 TFs and eight miRNAs. The sets were collected from 29 publications and were obtained by various types of technologies, primarily gene expression microarrays and ChIP-chip location analyses. The number of genes in each target set ranges from 25–2338 with mean 400—57-fold larger than Tompa's sets. For each set, we used the corresponding PWM(s) from TRANSFAC, or the eight-long miRNA seed from miRBase, as the correct motif. A comparison of our compendium to several other motif discovery benchmarks is given in Table 1.

Results on metazoan benchmark

We executed Amadeus and the five other motif finding tools on the metazoan target-set compendium. Here too, each tool was run with motif lengths 8 and 10, and the two top-scoring motifs were compared to the correct PWM(s). The results of each tool are shown in Figure 3; success rates and running times are summarized in Figure 4. Amadeus significantly outperformed all other programs in terms of motif recovery rate—62% success (with PWM divergence cutoff of 0.18); consistent with recent studies (Tompa et al. 2005; Ettwiller et al. 2007), Weeder and Trawler (43% success) performed better than the rest of the tools (10%–27%). We repeated the benchmark comparison using only the top-scoring motif from each execution and with two other PWM similarity measures and obtained very similar results (Supplemental Fig. 4).



A					B					
Organism	TF/miRNA	Source	Target-set size	Reference	AlignACE	MEME	YMF	Trawler	Weeder	Amadeus
Human	CREB	CC	2338	Zhang et al. '05	∞	∞	□	□	░	■
	E2F	CC	201	Cam et al. '04	□	□	□	■	■	▩
	E2F	CC	79	Ren et al. '02	□	□	□	▩	■	■
	ESR1	C-DSL	496	Kwon et al. '07	□	□	□	□	□	□
	ETS1	CC	1192	Hollenhorst et al. '07	∞	□	░	□	▩	▩
	E2F	Ex	266	Whitfield et al. '02	□	□	□	■	■	■
	NF-Y	Ex	344	Whitfield et al. '02	□	□	□	□	□	▩
	HNF1A	CC	206	Odom et al. '04	░	□	□	▩	□	■
	HNF4A	CC	1475	Odom et al. '04	∞	∞	□	□	□	□
	ONECUT	CC	212	Odom et al. '04	□	□	□	□	□	░
	HSF1	CC	328	Page et al. '06	□	□	□	□	□	□
	IRF/NFKB	GO	586	Long et al. '04	▩	□	▩	□	□	░
	NANOG	CC	714	Boyer et al. '05	□	□	□	□	□	□
	NFKB	CC	270	Schreiber et al. '06	□	□	□	■	░	■
	NRF1	CC	674	Cam et al. '04	■	■	■	■	■	■
	POU5F1	CC	241	Boyer et al. '05	□	□	░	□	□	░
	TP53	Ex	38	Kannan et al. '01	□	□	□	□	░	□
	SOX2	CC	575	Boyer et al. '05	□	□	□	□	□	□
	SRF	CC	172	Cooper et al. '07	□	□	▩	■	■	■
	YY1	CC	713	Xi et al. '07	■	□	▩	■	■	■
Mouse	Foxp3	CC	1060	Marson et al. '07	□	□	□	□	□	□
	Irf/Nfkb	GO	329	Long et al. '04	▩	□	▩	□	□	░
	Mef2	CC	25	Blais et al. '05	□	□	□	□	□	□
	Myod	CC	102	Blais et al. '05	░	□	□	▩	▩	▩
	Myod	CC	102	Cao et al. '06	░	□	□	░	▩	▩
	Myog	CC	78	Cao et al. '06	□	□	□	▩	■	▩
	Myog	CC	106	Blais et al. '05	□	□	□	▩	■	■
C. elegans	GATA	Ex	1342	Pauli et al. '06	∞	□	▩	□	□	▩
Fly	Hsf	CC	183	Birch-Machin et al. '05	□	□	□	░	□	□
	Mef2	CC+Ex	208	Sandmann et al. '06	□	□	□	□	□	□
	Dref	CC	116	Legube et al. '06	□	■	■	▩	□	▩
	Myc/Max/Mad	DamID	714	Orian et al. '03	░	□	□	□	□	▩
Human	hsa-let-7a	Ex	177	Brueckner et al. '07	□	□	□	□	□	□
	hsa-let-7b	Ex	182	Johnson et al. '07	□	□	□	■	■	▩
	hsa-miR-1	Ex	65	Lim et al. '05	□	■	□	■	■	■
	hsa-miR-16	Ex	90	Linsley et al. '07	□	■	□	■	■	■
	hsa-miR-34a	Ex	89	Chang et al. '07	□	□	▩	□	□	□
	hsa-miR-34a	Ex	367	He et al. '07	□	■	□	□	■	▩
	hsa-miR-106b	Ex	88	Linsley et al. '07	□	■	■	■	■	■
	hsa-miR-124	Ex	116	Lim et al. '05	□	■	■	■	■	■
	hsa-miR-373	Ex	43	Lim et al. '05	□	▩	■	□	▩	■
Mouse	mmu-miR-155	Ex	95	Rodriguez et al. '07	□	□	□	▩	□	■

■ div < 0.12

▩ div < 0.18

░ div < 0.24

□ not found

∞ running time > 2 days

Figure 3. The metazoan target-set compendium and benchmark results on it. (*A*) The compendium of metazoan TF/miRNA target sets collected from the literature. The "Source" column indicates the experimental procedure or database from which the target set was derived: gene expression microarrays (Ex), ChIP-chip (CC), ChIP-DSL (C-DSL), DamID (van Steensel et al. 2001), or Gene Ontology (GO) database (Ashburner et al. 2000). For additional information and references, see http://acgt.cs.tau.ac.il/amadeus. (*B*) Performance of motif finding tools on each target set—each successful motif recovery is marked by a gray-shaded box, according to the PWM divergence (darker shades of gray indicate higher similarity of the recovered motif to the one in the literature); the ∞ symbol marks long executions (>48 h) that were aborted. Here, Amadeus was run with the HG enrichment score. The success-rate patterns of the six motif finders are almost identical when comparing different target sets of the same TF. For example, in all three E2F data sets, Amadeus, Weeder, and Trawler are the only tools that recovered the correct motif; in the two Myod sets, Amadeus and Weeder succeeded with PWM divergence cutoff 0.18, AlignACE succeeded with cutoff 0.24, and MEME and YMF failed with all cutoffs. This consistency, observed for all six TFs that are represented by more than one set in our compendium, is not a result of large overlaps between the target sets, as such overlaps were avoided in the construction of the compendium. Instead, it is likely to stem from properties inherent to the TFs, such as the extent and type of their BSs degeneracy.

Table 1. Comparison of several medium- and large-scale motif discovery benchmarks

Benchmark	Harbison et al. 2004	Tompa et al. 2005	Ettwiller et al. 2007	Our compendium
Type	Experimental	Synthetic	Experimental	Experimental
Technology	ChIP-chip	Validated BSs	ChIP-chip	ChIP-chip, gene expression, others
Source	In-house	TRANSFAC	Literature (seven publications)	Literature (29 publications)
Species	Yeast	Human, mouse, fly, yeast	Human, mouse	Human, mouse, fly, worm
Regulators	TFs	TFs	TFs	TFs, miRNAs
No. of sets	173	52	10	42
No. of distinct TFs/miRNAs	83 TFs	Unknown	10 TFs	26 TFs, eight miRNAs
Average no. of genes per set	58	7	259	400
Average sequence length per set	35 kbp	8 kbp	210 kbp	383 kbp

The yeast ChIP-chip data sets (Harbison et al. 2004) are a popular benchmark, but they represent a single, relatively simple species and only one technology. Tompa's benchmark (Tompa et al. 2005) is based on validated BSs from the TRANSFAC database—the BSs were chosen by the investigators according to various criteria and implanted inside real and synthetic promoters. Very recently, Ettwiller et al. (2007) developed Trawler, a new motif discovery tool for ChIP experiments, and reported its performance on 10 mammalian ChIP-chip data sets. Our compendium is the first large-scale collection of metazoan gene sets derived from high-throughput experiments; it represents diverse technologies and organisms and consists of both TF and miRNA target sets. Of note, the average set size in our compendium is substantially larger than in all other benchmarks.

We observed a considerable degradation of up to 32% in the success rate of most motif finders on our benchmark relative to their performance on the yeast data sets. Remarkably, the success rate of Amadeus on the metazoan benchmark is comparable to that on the yeast data—62% vs. 65%, respectively. Amadeus is also the fastest tool (10 min per data set, on average); AlignACE and MEME are prohibitively slow on large target sets.

Handling length and GC-content biases

The HG enrichment score might fail to discover the correct motif, or alternatively detect many spurious motifs, when the distribution of the length and/or GC-content of the target set sequences significantly differ from their distribution in the BG set. Biologically meaningful groups of genes with such biases are not uncommon. For instance, genes with GC-rich promoters, such as housekeeping genes, tend to have higher expression rates (Kass et al. 1997; Aerts et al. 2004). Another example is the length bias of 3′ UTRs of tissue-specific genes. For example, genes that are expressed in neuronal tissues have relatively long 3′ UTRs (1300 nucleotides vs. 950 nucleotides in the entire genome) (Sood et al. 2006). To search for enriched motifs in such data sets, we developed a novel score, termed binned enrichment score, which partitions the genes into bins according to the length and GC-content of their *cis*-regulatory sequences and evaluates the overrepresentation of the motif based on its abundance in each bin (see Methods).

Running Amadeus on our metazoan target-sets compendium using the binned enrichment score further improves over the results of the HG score (Fig. 4). One example is the target set of Mef2 (Blais et al. 2005), for which none of the programs we tested recovered the correct motif. The promoters of these genes are longer than average (972 bp vs. 840 bp after masking out repetitive and coding sequences) and have a higher GC-content (53% vs. 49%). Using the binned enrichment score, Amadeus discovers the Mef2 binding pattern as the top-scoring motif. Additional examples that demonstrate the importance of accounting for length and GC biases are given in the Supplemental material. The improved sensitivity of the binned score remained consistent for other PWM similarity measures and cutoffs (5% improvement, on average) (data not shown).

Genome-wide analyses

Another useful application of motif finding is a genome-wide analysis, targeted to uncover regulatory motifs based on the genome alone, without having at hand a set of coregulated genes. We developed three scores for this type of analysis: localization, strand bias, and chromosomal preference (see Methods).

Localization

Many TFs are known to bind more frequently close to their target genes' TSSs than in distant promoter regions (Tabach et al. 2007). Some elements that directly interact with or are part of the basal transcriptional machinery, such as TATA-box and Initiator, are found mainly in core promoters, spanning several dozens of bases around the TSS (Smale and Kadonaga 2003). We implemented a localization score that measures the tendency of a motif to occur at specific locations along the promoters. Applying this score on all human and mouse promoter sequences revealed binding patterns of many known TFs, including core promoter elements (e.g., SP1, NF-Y, TATA) and prominent TFs (e.g., MYC, ATF/CREB), as well as two novel motifs. Some of the discovered motifs exhibit a significant strand bias (i.e., they do not appear at similar rates on both strands) or chromosomal preference (i.e., a nonuniform distribution across chromosomes). The main results are listed in Table 2 (see also Discussion). Using a specially tailored method, FitzGerald et al. (2004) reported on nine motifs that localize in human promoters. Eight of these motifs were found by Amadeus. The ninth motif is the ATG start codon, which was not discovered by Amadeus, since we masked out coding sequences.

A genome-wide analysis of fly promoters uncovered more than 30 motifs with significant localization (for full results, see the Amadeus website). Ohler et al. (2002) searched for motifs that are enriched in the core promoters of the fly genome. They reported on 10 motifs, all of which are among the top 21 motifs we discovered.

Chromosomal preference

Motivated by the observation that coregulated genes may colocalize (Cohen et al. 2000; Boutanaev et al. 2002), we developed a chromosomal-preference score to discover motifs whose occur-

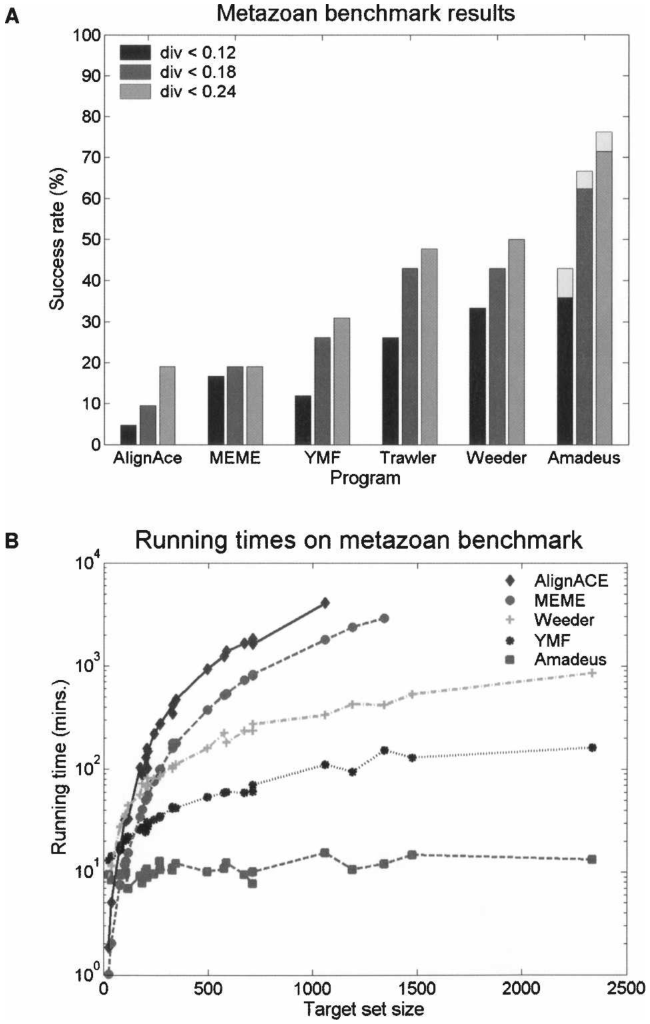


Figure 4. Performance of six motif finding tools on our compendium of metazoan target sets. (*A*) Success rates for three PWM divergence cutoffs, indicated by different shades of gray. The light-gray boxes on *top* of the Amadeus bars show the improved success rates when using the binned enrichment score (instead of the HG score; see Methods). Success rates for other PWM similarity measures and cutoffs are shown in Supplemental Figure 4. (*B*) Running times in logarithmic scale for the TF target-sets (AlignACE and MEME did not finish within 48 h on several sets). Trawler is a web-based tool so we could not measure its running time. For full results, see Supplemental Table 1 and http://acgt.cs.tau.ac.il/amadeus. A detailed comparison of all tested tools is given in Supplemental Table 2.

rences are not distributed evenly across chromosomes (see Methods). Interestingly, when we applied this score to *D. melanogaster* promoters, Amadeus found that the Dref binding motif is overrepresented on the X chromosome (Supplemental Fig. 6). Indeed, Dref was recently associated with the dosage compensation complex (DCC) that equalizes the expression levels of X-linked genes in drosophila males and females.

Recently, Ruby et al. (2006) discovered a new class of small 21-nucleotide RNAs in worms, called 21U-RNAs, that reside mainly in introns and intergenic regions on chromosome IV. They also discovered a conserved motif located ~40 bases upstream to these regions. Running the chromosomal-preference analysis on all *C. elegans* promoters, Amadeus reported this pattern as the top-ranking motif (Fig. 5). Thus, without any prior knowledge on 21U-RNAs in worm and DCC in fly, Amadeus found motifs known to be associated with them, demonstrating another type of biological signal it can uncover.

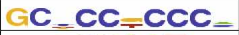
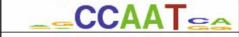
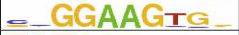
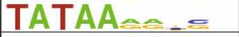
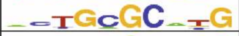
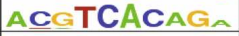
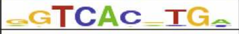
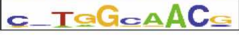
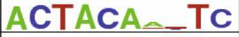
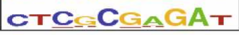
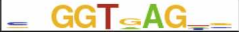
Discussion

In this article, we present a compendium of target sets of metazoan TFs and miRNAs that we used as a benchmark for motif finding tools. To the best of our knowledge, our compendium is the first publicly-available large-scale collection of experimentally derived TF and miRNA target sets and thus constitutes a valuable database for studying gene regulation. We believe it is an improvement over previously published benchmarks (Harbison et al. 2004; Tompa et al. 2005), as it more accurately represents a broad range of gene-regulation motif discovery tasks. The yeast ChIP-chip data sets (Harbison et al. 2004) represent only one, relatively simple, species and only a single type of assay. As explained earlier, Tompa's benchmark (Tompa et al. 2005) is to some extent artificial, or "unrealistically clean"—in the nonsynthetic data sets, each gene has a validated BS in its promoter. As expected in high-throughput techniques, our gene sets are much larger and contain a high rate of false positives, i.e., genes that are not targets of the corresponding TF or that contain a BS further upstream/downstream from the TSS. Moreover, our compendium contains sets of 3′ UTRs targeted by miRNAs; these sets have different statistical properties than promoters bound by TFs (e.g., GC-content and length variance), and thus pose additional computational challenges.

For simplicity of implementation and in order to allow a fair comparison between motif finders and among various types of data sets, our benchmark does not exploit all the available information generated by some of the experimental techniques, such as the binding peaks and affinities derived from ChIP assays. In addition, we did not exploit comparative sequence analysis, a potentially powerful tool that poses additional challenges, on top of the basic motif finding task studied here. For example, recent studies reported a limited cross-species conservation of functional BSs (Borneman et al. 2007; Lin et al. 2007; Odom et al. 2007); thus, in some situations, searching for motifs only within aligned sequences might be unfavorable (for further discussion, see Supplemental material).

We developed Amadeus, a new software platform for de novo motif discovery, and compared its performance to five popular motif finding tools. Amadeus, whose running time depends on the number and length of BG sequences, but not on the size of the target set, was significantly faster than the other programs on most data sets. Unlike the other tools, which performed rather poorly on the metazoan data, Amadeus achieved a high success rate on both the metazoan and yeast benchmarks. We believe this is largely due to the fact that most tools use BG models based on precomputed *k*-mer counts (*k* = 1, 4, 4, 8 in AlignACE, YMF, MEME, and Weeder, respectively), whereas Amadeus utilizes the entire set of promoters (or 3′ UTRs) in the genome as a reference set for testing over-representation. This is especially important in higher eukaryotes that have complex signals in their *cis*-regulatory sequences, which are not likely to be captured by simple BG models. Indeed, on our benchmark, the success rates of extant motif finders correlate with the complexity of their BG models. Trawler is the only extant tool we tested that utilizes a supplied set of BG sequences to assess motif enrichment. However, its BG set is relatively small (it failed to run with more than 2000 BG sequences),

Table 2. Main results of human and mouse genome-wide localization analysis

Name	Motif logo	Localization		Strand bias	Chrom pref.
		peak	*p*-val	*p*-val	*p*-val
A. Known TFs					
SP1		-60	10^{-129}	-	-
NF-Y		-90…-60	10^{-145}	-	10^{-4} (19)
GABP		-30…0	10^{-113}	-	-
TATA		-30	10^{-61}	10^{-15}	10^{-3} (6)
NRF1		-30	10^{-48}	-	-
ATF/CREB		-60	10^{-27}	-	-
MYC		-60…-30	10^{-27}	-	-
RFX1		-60	10^{-9}	-	-
B. Novel					
ACTACAWYTC		-90…-60	10^{-21}	10^{-8}	10^{-4} (19)
CTCGCGAGAT		-60…-30	10^{-7}	-	-
C. Other					
Splice donor site		+30…	10^{-23}	10^{-8}	-

Amadeus was run on all human and mouse promoters and searched de novo for motifs that are significantly localized (i.e., overrepresented at a particular distance from the TSS, measured in bins of size 30) in both species. Approximately 23,000 and 24,000 human and mouse promoters, respectively, were analyzed. Promoters spanned from 500 bp upstream to 100 bp downstream of the TSS. Both known and novel motifs were found. All *P*-values listed in the table are for human. "Peak" refers to the center(s) of the location bin(s) with the largest motif occurrence rate. Amadeus also tests whether the motif occurrences are distributed nonuniformly between the strands ("Strand bias" column, showing the significance in human) or across the chromosomes ("Chrom pref." column, showing also the overrepresented chromosome in human in parentheses).

which in addition to the algorithm and statistical score it employs may explain its moderate performance on metazoan data sets (for more details, see Supplemental material). In conclusion, the success rate and running time of Amadeus scale up better than extant programs in terms of both the size of the data set and the species complexity. Supplemental Table 2 summarizes the main differences between the tools in terms of algorithms, scores, features, and performance.

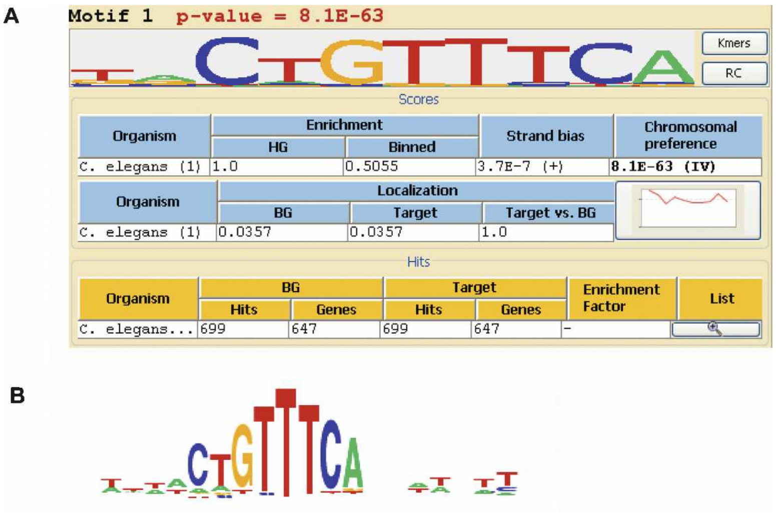
The high accuracy of Amadeus remained consistent under various benchmark settings, e.g., evaluating the performance using other common PWM similarity measures or using the top-scoring motif only (Supplemental Fig. 4). Taken together with the fact that our benchmark contains a large number of diverse data sets, our results indicate that the improved performance of Amadeus is inherent, rather than a product of overfitting or biased choice of parameters.

We developed a novel statistical score for evaluating motif overrepresentation in target sets that are biased with respect to the rest of the genome in terms of sequence length or base composition. Although they are quite common, such biases are often ignored, which might lead to false results. This score improved the performance of Amadeus by 5%.

In order to gain insight into the practical limitations of Amadeus, we examined the target sets in which it failed to discover the correct motif. Evidently, in most cases a large fraction of the reported target genes does not contain a BS within the 1200-bp promoter region we analyzed. For example, Boyer et al. (2005) used promoter arrays against the −8-kb to +2-kb region relative to the TSS. Only 30% of the genes they reported as targets of NANOG contain a binding event within 1 kb upstream of the TSS. Another example is HSF (heat-shock factor), which is represented in our compendium by two target sets—human and fly. In both cases, it seems that the overrepresentation of the BS motif is borderline, which explains why none of the tools we tested accurately recovered the motif. Using a combined analysis of both sets together, a unique feature in Amadeus, we were able to successfully discover the correct binding pattern (Supplemental Fig. 7).

In this study, we also demonstrated application of Amadeus to genome-wide motif analysis, which can be applied to any genome with a sufficient number of *cis*-regulatory sequences without need for target sets from prior experiments. Using various statistical scores, Amadeus discovered an assortment of biological phenomena. Searching for motifs with nonrandom chromosomal distribution in fly and worm revealed the Dref and 21U-RNA–related patterns, respectively, which were found recently using a combination of experimental and computational techniques.

Localization analysis of human and mouse promoters recovered known mammalian TF motifs, the splice donor site, and two novel motifs. The first novel motif (ACTACAWYTC) was also discovered independently by high-throughput location analyses for ESR1 (ER-α) (Kwon et al. 2007), RUNX1, and ETS1 (Hollenhorst et al. 2007). Running Amadeus on these sets reproduced the motif, which apparently has diverse biological functions. Interestingly, the motif has a significant strand bias (the only other localized human TF we found with a strand bias was TATA-box), and like NF-Y, it is over-represented on chromosome 19. Very recently, Sinha et al. (2008) used a decoy corresponding to a variant of this motif, reported in Xie et al. (2005), in order to experimentally validate that it has a regulatory role in cell-cycle progression. The



Motif 1 p-value = 8.1E-63

Organism	Enrichment		Strand bias	Chromosomal preference
	HG	Binned		
C. elegans (1)	1.0	0.5055	3.7E-7 (+)	**8.1E-63 (IV)**

Organism	Localization		
	BG	Target	Target vs. BG
C. elegans (1)	0.0357	0.0357	1.0

Organism	BG		Target		Enrichment Factor	List
	Hits	Genes	Hits	Genes		
C. elegans...	699	647	699	647	-	

Figure 5. Genome-wide chromosomal preference analysis of *C. elegans* promoters. (*A*) Screenshot of Amadeus output, showing the top-scoring motif found in the analysis. The motif is highly overrepresented on chromosome IV ($P = 8 \times 10^{-63}$). (*B*) The motif reported by Ruby et al. (2006), found upstream of many 21U-RNAs, is nearly identical to the one identified de novo by Amadeus.

second motif (CTCGCGAGAT), reported also by FitzGerald et al. (2004), was shown to regulate ARF3 in vivo (Haun et al. 1993).

The localization analysis results obtained by Amadeus on both human/mouse and fly promoters compare favorably with other, specially tailored methods (Ohler et al. 2002; FitzGerald et al. 2004): In a single run, Amadeus discovered all the motifs reported by those methods and supplied additional information on their strand and chromosomal distributions. We therefore believe that Amadeus may be used as a general tool for genome-wide motif discovery tasks, aimed at uncovering sequence patterns with various global features.

In addition to sensitivity, efficiency, and supporting multiple target sets and scores, we focused on developing a friendly and informative graphical user interface in order to make Amadeus easily accessible and beneficial to a wide range of users. Built-in algorithmic features, such as pairs analysis and bootstrapping, as well as various graphical and textual displays of the motifs, the scores they attained, and their putative BSs, make it easier for the user to understand the nature of the discovered motifs and highlight the most biologically interesting ones. The Amadeus software (standalone Java application) and our compendium of TF and miRNA target sets are accessible at http://acgt.cs.tau.ac.il/amadeus. We are continuing to implement novel features in Amadeus and to add newly published data sets to the compendium.

Methods

Genomic sequences and binding patterns

Promoter and 3′ UTR sequences (repeat- and coding-sequence-masked) of human, mouse, fly, and worm were extracted from Ensembl (Birney et al. 2004). Yeast promoters were downloaded from SGD (http://www.yeastgenome.org). Binding patterns of TFs and miRNAs were taken from TRANSFAC (Wingender et al. 1996) and miRBase (http://microrna.sanger.ac.uk/sequences), respectively. For more details, see Supplemental material.

Target sets of metazoan TFs and miRNAs

We collected 42 TF/miRNA target sets from the literature, focusing on sets obtained using high-throughput techniques, such as gene expression microarrays and ChIP-chip assays. We included only TFs and miRNAs whose binding patterns are described in TRANSFAC and miRBase, respectively. Genes were mapped to Ensembl gene IDs using Biomart (http://www.biomart.org). In order to avoid strong dependencies between the target sets, we verified that no two sets of the same TF/miRNA have an overlap greater than 30%.

For the TF data sets, we used promoter sequences spanning from 1000 bp upstream to 200 bp downstream of the TSS, a range that covers most of the promoter array sequences and is often used in computational promoter analysis; for the miRNA data sets, we used full-length 3′ UTRs (coding strand only); repetitive and coding sequences were masked out. The total sequence length of the target sets is 383 kbp on average, much larger than the yeast ChIP-chip data sets (35 kbp) and Tompa's benchmark (8 kbp).

Amadeus software and algorithms

Amadeus executes a series of refinement phases where each phase gets as input a list of motif candidates, applies an algorithm for refining the list, and produces a set of improved candidates, which serves as a starting point for the next phase. Each phase uses a different motif model, which best suits its algorithm and performance requirements. Generally, the first phases use simple motif models and enumerate a very large number of candidates, whereas the final phases evaluate a smaller number of more complex motifs, namely PWMs. Motifs in each phase are evaluated using one or more score functions: enrichment, localization, strand bias, and chromosomal preference, which are combined into a single *P*-value (see below). The phases of Amadeus, in their running order, are *preprocess*, *mismatch*, *merge*, *greedy*, *postprocess*, and *pairs analysis* (see Fig. 1).

In the *preprocess* phase, all *k*-mers are evaluated, where *k*, the motif length, is a user-defined parameter. In the *mismatch* phase, the motif model is changed from a *k*-mer to a list of *k*-mers by introducing degenerate positions into the *k*-mers. Afterward, the *merge* phase combines pairs of similar motifs. This is done recursively until no new high-scoring similar pairs are encountered. The *greedy* phase constructs a PWM from each motif and optimizes it using a greedy EM-like iterative process: In each iteration, it searches for the PWM cutoff that yields the best score, and then the occurrences in the target set that pass this cutoff are used to build a new, refined PWM; this process is repeated as long as the score improves. Finally, in the *postprocess* phase, redundancy is eliminated by removing every motif, for which there exists a higher scoring motif, such that more than 5% of their occurrences overlap. The final list of discovered motifs is then compared with a database of known PWMs (TRANSFAC for TFs, miRBase for miRNAs), and all similarities with PWM divergence below 0.24 are reported. Additional statistics and information are provided for each motif to assist the user in evaluating the results (Supplemental Fig. 2). In the optional *pairs analysis* phase, Amadeus reports pairs of motifs that tend to co-occur within the same *cis*-regulatory sequences.

Other important features we implemented in Amadeus include automatic removal of redundant sequences (to avoid biases in the analysis due to families of paralogous genes with nearly-identical *cis*-regulatory sequences) and bootstrapping (to correct the reported *P*-values for multiple testing, by repeating the entire analysis on randomly selected gene sets). By utilizing highly efficient data structures, designed to minimize running time and memory consumption, Amadeus is able to check a huge number of candidate motifs and analyze quickly whole-genome *cis*-regulatory sequences. For more details, see Supplemental material.

Scores for evaluating motifs

Amadeus evaluates each candidate motif using one or several model-independent score functions, chosen by the user. The *P*-values computed for multiple score functions and/or target sets are combined into a single *P*-value using the Z-transform (Whitlock 2005).

HG enrichment score

Let *B* and *T* ($T \subseteq B$) denote the BG and target sets, respectively, and let *b* and *t* denote the subset of genes from the BG and target set, respectively, that contain at least one occurrence of the motif (hit, in short) in their *cis*-regulatory sequence. The HG enrichment score computes the probability of observing at least $|t|$ target sequences with a motif occurrence, under the null hypothesis that the genes in the target set were drawn randomly, independently, and without replacement from the BG set (Elkon et al. 2003):

$$HG\ score = HG\ tail\ (|B|,|T|,|b|,|t|) = \sum_{i=|t|}^{\min(|T|,|b|)} \frac{\binom{|b|}{i}\binom{|B|-|b|}{|T|-i}}{\binom{|B|}{|T|}}.$$

Binned enrichment score

The genes are divided into n bins according to the GC-content and length of their *cis*-regulatory sequences. Let B_i and T_i be the BG and target set genes, respectively, in the ith bin, and denote by b_i the subset of genes from B_i whose sequence contains a hit. The goal of this score is to account for cases where the fraction of targets is uneven across bins. Suppose that targets within each bin are selected uniformly. Then, in bin i the probability that a selected gene will contain a hit (i.e., belong to b_i) is $|b_i|/|B_i|$. Since the fraction of targets in bin i is $|T_i|/|T|$, it follows that the probability that a selected gene will contain a hit is

$$p_m = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \cdot \frac{|b_i|}{|B_i|}.$$

Assume now that $|T|$ target genes are sampled with replacement from B. Then, the probability for having at least $|t|$ target genes with a motif occurrence is given by the tail of the following binomial distribution:

$$\textit{Binned score} = \textit{Binomial tail}\ (\ |T|, p_m, |t|\ ) = \sum_{i=|t|}^{|T|} \binom{|T|}{i} p_m{}^{i}(1-p_m)^{|T|-i}.$$

Note that this score does not use the number of targets that contain a hit in each bin separately, but rather the total t.

Strand–bias score

The score uses a binomial test to measure the tendency of the motif to occur in one of the strands more often than in the other. A strong strand bias could, for example, indicate that the motif has a post-transcriptional role, as it may be related to the gene's RNA.

Localization score

The score estimates whether the occurrences of the motif tend to cluster at specific distances from the TSS. The hits are partitioned into bins according to their location; for each bin, a binomial test computes the overrepresentation of hits in that bin under the null hypothesis that the hits are distributed randomly among the bins (i.e., according to the total number of k-mers in each bin); finally, the bin with the lowest P-value is chosen and its score is Bonferroni corrected for multiple testing.

We implemented three variants of the localization score: The "BG" and "Target" variants compute the localization of the hits in the BG and target set, respectively (a motif may exhibit localization across the entire genome, or only for target-set genes); in order to account for global location-dependent biases in the nucleotide composition, the "Target vs. BG" variant checks whether the occurrences of the motif in the target-set tend to localize given the distribution of their locations in the rest of the genome. For a detailed explanation, see Supplemental material.

Chromosomal–preference score

In order to test whether the motif is not distributed evenly among the chromosomes, the enrichment of the motif in each chromosome is evaluated using the HG distribution; the smallest P-value is chosen and Bonferroni corrected for multiple testing.

Pairs of co-occurring motifs

In order to find pairs of cooperative TFs (or miRNAs), Amadeus checks the co-occurrence rate of each pair of motifs by computing the following HG tail probability:

$$\textit{Pair score} = \textit{HG tail}\ (\ |T|, |t_1|, |t_2|, |t_{12}|\ ),$$

where T is the target set; t_1 and t_2 are the subsets of target genes that contain at least one occurrence of the first and second motif, respectively; and t_{12} is the subset of target genes that contain hits for both motifs. Applying an EM-like procedure similar to the one used for single motifs, the PWMs comprising the pair of motifs are tuned in order to optimize the co-occurrence score.

Acknowledgments

We thank Gidi Weber and Adi Maron-Katz for their contributions to the source code. Promoter sequences were downloaded by Ran Elkon. We also thank the anonymous referees for their constructive comments. R.S. was supported in part by the Raymond and Beverly Sackler Chair in Bioinformatics and by a grant from the Wolfson Foundation.

References

Aerts, S., Thijs, G., Dabrowski, M., Moreau, Y., and De Moor, B. 2004. Comprehensive analysis of the base composition around the transcription start site in Metazoa. *BMC Genomics* **5:** 34. doi: 10.1186/1471-2164-5-34.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

Bailey, T.L. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2:** 28–36.

Birney, E., Andrews, T.D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., et al. 2004. An overview of Ensembl. *Genome Res.* **14:** 925–928.

Blais, A., Tsikitis, M., Acosta-Alvear, D., Sharan, R., Kluger, Y., and Dynlacht, B.D. 2005. An initial blueprint for myogenic differentiation. *Genes & Dev.* **19:** 553–569.

Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., Rozowsky, J., Seringhaus, M.R., Wang, L.Y., Gerstein, M., and Snyder, M. 2007. Divergence of transcription factor binding sites across related yeast species. *Science* **317:** 815–819.

Boutanaev, A.M., Kalmykova, A.I., Shevelyov, Y.Y., and Nurminsky, D.I. 2002. Large clusters of co-expressed genes in the *Drosophila* genome. *Nature* **420:** 666–669.

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122:** 947–956.

Cohen, B.A., Mitra, R.D., Hughes, J.D., and Church, G.M. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* **26:** 183–186.

Elkon, R., Linhart, C., Sharan, R., Shamir, R., and Shiloh, Y. 2003. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res.* **13:** 773–780.

Eskin, E. and Pevzner, P.A. 2002. Finding composite regulatory patterns in DNA sequences. *Bioinformatics* **18:** S354–S363.

Ettwiller, L., Paten, B., Ramialison, M., Birney, E., and Wittbrodt, J. 2007. Trawler: De novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat. Methods* **4:** 563–565.

FitzGerald, P.C., Shlyakhtenko, A., Mir, A.A., and Vinson, C. 2004. Clustering of DNA sequences in human promoters. *Genome Res.* **14:** 1562–1574.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431:** 99–104.

Haun, R.S., Moss, J., and Vaughan, M. 1993. Characterization of the human ADP-ribosylation factor 3 promoter. Transcriptional regulation of a TATA-less promoter. *J. Biol. Chem.* **268:** 8793–8800.

Hollenhorst, P.C., Shah, A.A., Hopkins, C., and Graves, B.J. 2007. Genome-wide analyses reveal properties of redundant and specific promoter occupancy within the ETS gene family. *Genes & Dev.* **21:** 1882–1894.

Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. 2000. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296:** 1205–1214.

Kass, S.U., Pruss, D., and Wolffe, A.P. 1997. How does DNA methylation repress transcription? *Trends Genet.* **13:** 444–449.

Kwon, Y.S., Garcia-Bassets, I., Hutt, K.R., Cheng, C.S., Jin, M., Liu, D., Benner, C., Wang, D., Ye, Z., Bibikova, M., et al. 2007. Sensitive ChIP-DSL technology reveals an extensive estrogen receptor alpha-binding program on human gene promoters. *Proc. Natl. Acad. Sci.* **104:** 4852–4857.

Lin, C.Y., Vega, V.B., Thomsen, J.S., Zhang, T., Kong, S.L., Xie, M., Chiu, K.P., Lipovich, L., Barnett, D.H., Stossi, F., et al. 2007. Whole-genome cartography of estrogen receptor alpha binding sites. *PLoS Genet.* **3:** e87. doi: 10.1371/journal.pgen.0030087.

Linhart, C., Elkon, R., Shiloh, Y., and Shamir, R. 2005. Deciphering transcriptional regulatory elements that encode specific cell cycle phasing by comparative genomics analysis. *Cell Cycle* **4:** 1788–1797.

Lockhart, D.J. and Winzeler, E.A. 2000. Genomics, gene expression and DNA arrays. *Nature* **405:** 827–836.

Odom, D.T., Dowell, R.D., Jacobsen, E.S., Gordon, W., Danford, T.W., MacIsaac, K.D., Rolfe, P.A., Conboy, C.M., Gifford, D.K., and Fraenkel, E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.* **39:** 730–732.

Ohler, U., Liao, G.C., Niemann, H., and Rubin, G.M. 2002. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.* **3**. doi: 10.1186/gb-2002-3-12-research0087.

Pavesi, G., Mauri, G., and Pesole, G. 2001. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* **17:** S207–S214.

Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D.P. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127:** 1193–1207.

Sharan, R., Ben-Hur, A., Loots, G.G., and Ovcharenko, I. 2004. CREME: *Cis*-Regulatory Module Explorer for the human genome. *Nucleic Acids Res.* **32:** W253–W256.

Sinha, S. and Tompa, M. 2002. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.* **30:** 5549–5560.

Sinha, S., Adler, A.S., Field, Y., Chang, H.Y., and Segal, E. 2008. Systematic functional characterization of *cis*-regulatory motifs in human core promoters. *Genome Res.* **18:** 477–488.

Smale, S.T. and Kadonaga, J.T. 2003. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72:** 449–479.

Sood, P., Krek, A., Zavolan, M., Macino, G., and Rajewsky, N. 2006. Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc. Natl. Acad. Sci.* **103:** 2746–2751.

Tabach, Y., Milyavsky, M., Shats, I., Brosh, R., Zuk, O., Yitzhaky, A., Mantovani, R., Domany, E., Rotter, V., and Pilpel, Y. 2005. The promoters of human cell cycle genes integrate signals from two tumor suppressive pathways during cellular transformation. *Mol. Syst. Biol.* **1:** 2005.0022. doi: 10.1038/msb4100030.

Tabach, Y., Brosh, R., Buganim, Y., Reiner, A., Zuk, O., Yitzhaky, A., Koudritsky, M., Rotter, V., and Domany, E. 2007. Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site. *PLoS One* **2:** e807. doi: 10.1371/journal.pone.0000807.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., et al. 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* **23:** 137–144.

van Steensel, B., Delrow, J., and Henikoff, S. 2001. Chromatin profiling using targeted DNA adenine methyltransferase. *Nat. Genet.* **27:** 304–308.

Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., et al. 2002. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell* **13:** 1977–2000.

Whitlock, M.C. 2005. Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* **18:** 1368–1373.

Wingender, E., Dietze, P., Karas, H., and Knuppel, R. 1996. TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* **24:** 238–241.

Wu, J., Smith, L.T., Plass, C., and Huang, T.H. 2006. ChIP-chip comes of age for genome-wide functional analysis. *Cancer Res.* **66:** 6899–6902.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345.

Zhu, W., Giangrande, P.H., and Nevins, J.R. 2004. E2Fs link the control of G1/S and G2/M transcription. *EMBO J.* **23:** 4615–4626.

Received January 8, 2008; accepted in revised form April 2, 2008.

3. Allegro: Analyzing expression and sequence in concert to discover regulatory programs

1566–1579 *Nucleic Acids Research, 2009, Vol. 37, No. 5* *Published online 16 January 2009*
doi:10.1093/nar/gkn1064

Allegro: Analyzing expression and sequence in concert to discover regulatory programs

Yonit Halperin, Chaim Linhart, Igor Ulitsky and Ron Shamir*

School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

Received November 10, 2008; Revised and Accepted December 16, 2008

ABSTRACT

A major goal of system biology is the characterization of transcription factors and microRNAs (miRNAs) and the transcriptional programs they regulate. We present Allegro, a method for *de-novo* discovery of *cis*-regulatory transcriptional programs through joint analysis of genome-wide expression data and promoter or 3′ UTR sequences. The algorithm uses a novel log-likelihood-based, non-parametric model to describe the expression pattern shared by a group of co-regulated genes. We show that Allegro is more accurate and sensitive than existing techniques, and can simultaneously analyze multiple expression datasets with more than 100 conditions. We apply Allegro on datasets from several species and report on the transcriptional modules it uncovers. Our analysis reveals a novel motif over-represented in the promoters of genes highly expressed in murine oocytes, and several new motifs related to fly development. Finally, using stem-cell expression profiles, we identify three miRNA families with pivotal roles in human embryogenesis.

INTRODUCTION

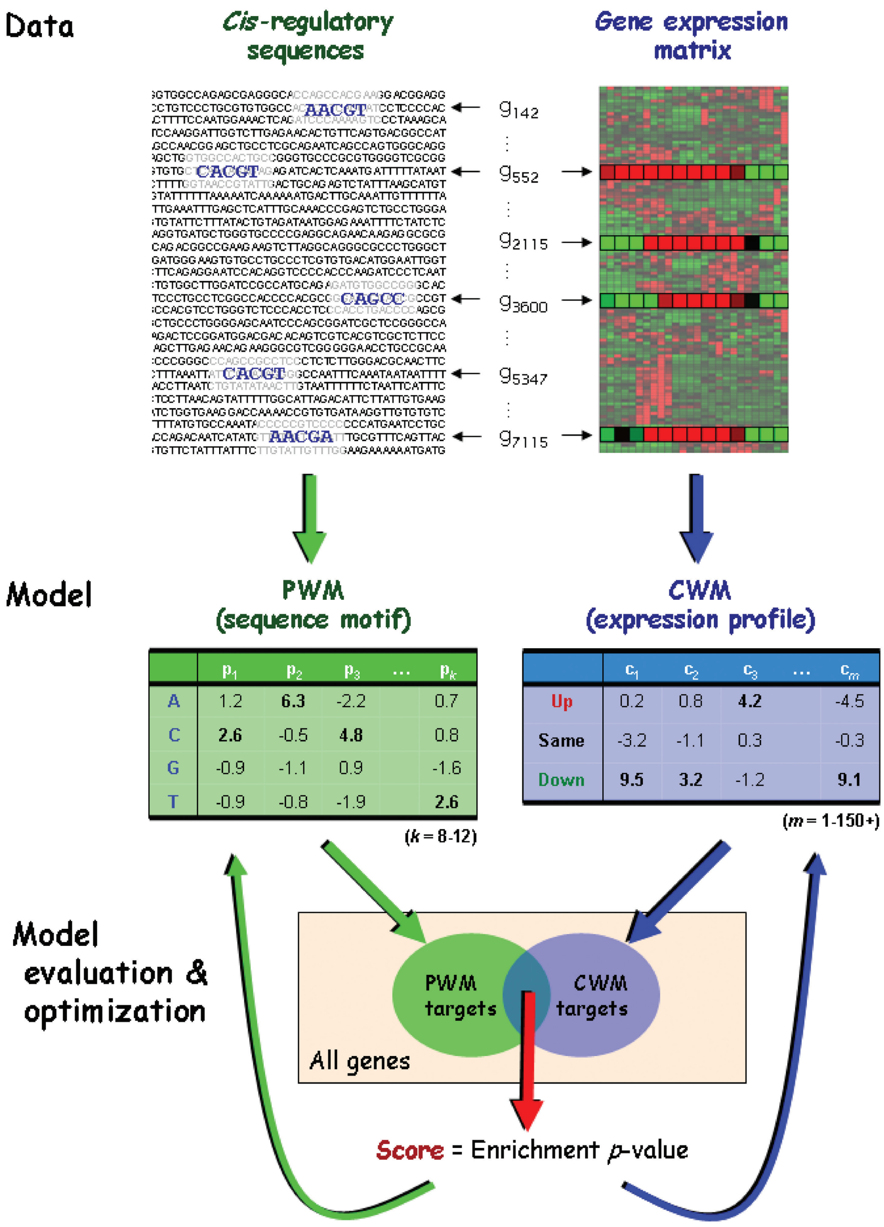
One of the main challenges in molecular biology is to understand the regulatory program that controls mRNA levels. The key components of this program are transcription factors (TFs), proteins that activate or repress transcription of a gene by binding to short DNA sequences, termed transcription factor binding sites (TFBSs), which usually reside in the gene's promoter. The level of translated mRNA of a gene can also be decreased post-transcriptionally, through annealing of microRNAs (miRNAs) to the 3′ UTR of the mRNA. A key step in reverse engineering regulatory networks is computational analysis of genome-wide measurements of mRNA levels obtained from DNA microarray assays in various environmental conditions, biological samples and time-points (henceforth we use the term *condition* to refer to each microarray assay). The purpose of this analysis is to identify groups of genes that are co-regulated, also termed *transcriptional modules* (TMs), and to characterize their regulators. A two-step approach is most commonly used [see examples in (1,2) and the review in (3)]: In the first step, a clustering procedure is executed to partition the genes into groups believed to be co-regulated, based on expression profile similarity (4). In the second step, a motif discovery tool is applied to search for abundant sequence patterns in the promoters (or 3′ UTRs) of each group that may represent the binding sites (BSs) of TFs (or miRNAs) that regulate the corresponding genes.

Despite extensive research, motif discovery has had limited success due to the short and degenerate nature of BSs, and the high levels of complexity of transcriptional networks, especially in metazoans. Since both the expression profiles and the promoter sequences of the genes carry information regarding their regulation, a methodology that utilizes both sources of information may give better results than the two-step approach. Several studies proposed computational schemes for this parallel analysis. Most of these algorithms use a unified probabilistic model over both gene expression and sequence data, and assume a Gaussian distribution of the expression values (5–7). Additional examples are the algorithms Reduce (8) and Motif Regressor (9), which search for motifs correlated with a *single* condition using linear regression, and assume that the number of BSs and their affinity are linearly correlated with the gene's expression. The algorithm DRIM (10) uses the hypergeometric (HG) score to compute the enrichment of motif occurrences among the top-ranked genes. However, it too is limited to a single condition.

Here we present Allegro (A Log-Likelihood based Engine for Gene expression Regulatory motifs Over-representation discovery), a *de-novo* motif discovery platform for simultaneously detecting gene sets with coherent expression profiles and corresponding over-represented sequence patterns. A graphic overview of the Allegro

*To whom correspondence should be addressed. Tel: +972 3 640 5383; Fax: +972 3 640 5384; Email: rshamir@post.tau.ac.il

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.



	p_1	p_2	p_3	…	p_k
A	1.2	**6.3**	-2.2		0.7
C	**2.6**	-0.5	**4.8**		0.8
G	-0.9	-1.1	0.9		-1.6
T	-0.9	-0.8	-1.9		**2.6**

(k = 8-12)

	c_1	c_2	c_3	…	c_m
Up	0.2	0.8	**4.2**		-4.5
Same	-3.2	-1.1	0.3		-0.3
Down	**9.5**	**3.2**	-1.2		**9.1**

(m = 1-150+)

Figure 1. Overview of the Allegro computational approach. Given a genome-wide expression matrix and *cis*-regulatory sequences (promoters or 3′ UTRs), Allegro executes efficient algorithms and statistical analyses to search for transcriptional modules. A transcriptional module is a set of genes sharing a sequence motif, modeled using a PWM, and a common expression profile described using a novel model called CWM. The CWM is analogous to the PWM: it assigns a weight to each discrete expression level in each of the experimental conditions. Allegro uses a multi-phase motif enumeration engine to generate candidate motifs. For each motif, it applies a cross-validation-like procedure to construct a CWM (Supplementary Figure 2), such that there is a significantly large overlap between the targets of the motif (the set of genes whose *cis*-regulatory sequence has an occurrence of the PWM, left arrows at the top) and the targets of the CWM (the genes whose expression levels match the CWM, right arrows). The statistical significance of this overlap is evaluated using one of two enrichment scores: the HG score or the binned enrichment score, which accounts for biases in the length and GC-content of the *cis*-regulatory sequences. The scores obtained by the motifs and their CWMs are iteratively modified to improve the models and eventually converge to high-scoring transcriptional modules.

approach is presented in Figure 1. Unlike existing methods, which rely on statistical assumptions, Allegro uses a novel non-parametric model called *Condition Weight Matrix* (CWM) to describe the expression profile of a group of co-regulated genes. We show that this model represents the expression profiles of sets of co-regulated genes more accurately than do commonly used expression metrics and statistical distributions. Allegro builds upon a motif discovery software platform we recently developed called Amadeus (11). In brief, given a set of co-regulated genes, Amadeus searches for motifs that are over-represented in their *cis*-regulatory sequences with respect to (w.r.t.) the rest of the *cis*-regulatory sequences in the genome or some other background (BG) set (see Supplementary Data for additional information). Allegro utilizes the efficient motif search engine of

Amadeus to enumerate a huge number of candidate motifs and to converge to high-scoring ones. For each candidate motif, Allegro fits a CWM to its putative targets using a cross-validation-like procedure. In order to ascertain whether the motif and the CWM are significantly correlated, Allegro computes one of two enrichment scores: the HG score or the binned enrichment score (11). As we demonstrate, the latter is very useful in cases where the expression profiles are correlated to the length and GC-content of the *cis*-regulatory sequences. Such expression-sequence dependencies are ignored by most existing methods, leading to many false predictions.

To test the performance of our method and highlight its unique features and advantages over existing approaches, we applied Allegro on several large-scale datasets from yeast, fly, mouse and human. In all cases, Allegro successfully recovered binding motifs of TFs and miRNAs that are known to regulate the relevant processes, together with their corresponding expression profiles. In addition, we report on novel transcriptional modules discovered by Allegro in datasets of human and murine tissues, and in *Drosophila* tissues profiled during various stages of development. For example, we discovered a novel motif that is over-represented in the promoters of genes that are highly induced in oocytes and fertilized eggs. Application of Allegro to expression profiles of human stem cell lines highlighted three miRNA families as key players in regulation of cell fate in embryogenesis. The miRNA activities predicted based on these findings are in good agreement with evidence from recent miRNA expression studies. A comparison of our results with those obtained by several current methods for clustering and motif finding indicates that Allegro is more sensitive and accurate. We also demonstrate additional important advantages of our approach, including joint analysis of multiple expression datasets from several organisms, and accounting for correlations between the expression levels of genes and the length and GC-content of their *cis*-regulatory sequences. We believe that Allegro introduces significant novel ideas in computational motif finding and gene expression analysis. On the practical side, our software can serve as an accurate, feature-rich, user-friendly tool for the biological community.

METHODS

Genomic sequences and binding patterns

Promoter sequences (repeat- and coding-sequence-masked) of human, mouse and fly, and 3′ UTR sequences (repeat-masked) of human were extracted from Ensembl (12). Yeast promoters were downloaded from SGD (http://www.yeastgenome.org). Motifs reported by Allegro were compared to known binding patterns of TFs and miRNAs taken from Transfac (13) and miRBase (http://microrna.sanger.ac.uk/sequences), respectively. See Supplementary Data for more details.

Gene expression datasets

The expression dataset for the yeast osmotic-stress response was downloaded from the supplementary material of (14). The values in the data are $\log_2$ of the fold change w.r.t. wild-type (WT) grown on YPD medium at standard osmolarity.

The human cell-cycle dataset was obtained from the supporting web-site of (15) (http://genome-www.stanford.edu/Human-CellCycle/HeLa). Expression values are $\log_2$ of the fold change w.r.t. asynchronously grown HeLa cells.

Human and mouse tissue expression datasets were downloaded from the GNF SymAtlas web-site (http://symatlas.gnf.org/SymAtlas, version 1.2.4, gcRMA-analzyed) (16). We applied quantile normalization (17), as implemented in Expander (18), in order to rescale the expression values in each tissue to a common distribution. We then normalized the values of each gene by computing the $\log_2$ of the fold change w.r.t. the gene's average expression value.

The human stem cells dataset is the stem cell matrix described in (19) (GEO accession number GSE11508). After averaging technical replicates, this dataset contains 124 samples. The full list of cell types used appears in Supplementary Table V. The expression pattern of each gene was normalized to mean 0 and SD 1.

The datasets analyzed in this study are summarized in Supplementary Table I. See Supplementary Data for additional details.

CWM for a motif target set

Denote by B the set of genes in the expression data, and let $e_{g(1)}, \ldots, e_{g(m)}$ denote the *discrete expression levels* (DELs) of gene $g \in B$ ($\forall 1 \le j \le m$, $e_{g(j)} \in \{e_1, \ldots, e_l\}$). The background *condition frequency matrix* (CFM), $R = \{r_{i,j}\}$, holds the frequencies of the DELs in each condition across all genes: $r_{i,j} = |\{g \in B \mid e_{g(j)} = e_i\}|/|B|$. For a candidate motif M, denote by T its target set, i.e. the group of genes whose *cis*-regulatory sequences contain an occurrence of M. As described in the Results section, Allegro samples a training set S from T, and constructs a CFM $F = \{f_{i,j}\}$ based on the DELs of the genes in S: $f_{i,j} = |\{g \in S \mid e_{g(j)} = e_i\}|/|S|$. The training-set sampling procedure is described in the Supplementary Data. Allegro then calculates the CWM, which contains the log-ratios between F and R:

$$\forall 1 \le i \le l, 1 \le j \le m \; CWM(i,j) = \log\left(\frac{f_{i,j}}{r_{i,j}}\right)$$

Allegro uses the CWM to compute the log-likelihood ratio (LLR) score of every gene, as explained below.

LLR score computation

Given the background CFM, $R = \{r_{i,j}\}$, and a CFM, $F = \{f_{i,j}\}$, learnt from the target set of a candidate motif, Allegro computes the LLR score of all the genes, as described in the Results section. The naïve computation takes $O(|B| \cdot |C|)$ time, where B is the set of genes and C is the set of conditions. Different genes may share the same discrete pattern, so the time complexity can be improved to $O(|P| \cdot |C|)$, where P is the set of distinct discrete expression patterns observed in the dataset. For example, in the tissues dataset (16) there are 14 698 human genes but only 2112 distinct expression patterns, so the above observation gives a 7-fold speedup in this case. Another running time improvement is achieved by reducing the average number of operations per discrete pattern in the LLR computation, as follows. In a preprocessing procedure we build a complete weighted graph, G_P, in which the nodes correspond to the patterns in P, and the weight of an edge is the Hamming distance between the two corresponding patterns. We then find a minimum spanning tree (MST) of G_P, denoted T_P. In order to compute the LLR score of all the patterns in P, we scan T_P in preorder, and use the LLR score of each pattern as a basis for computing the scores of its child nodes. Formally, let $v = (e_{v(1)}, \ldots, e_{v(|C|)})$ be a discrete expression pattern. If v is the root of T_P, the LLR is calculated naïvely, as described in the Results. Otherwise, let $u = (e_{u(1)}, \ldots, e_{u(|C|)})$ be the parent of v in T_P, then:

$$\mathrm{LLP}(v) = \mathrm{LLR}(u) + \sum_{j \in D_{uv}} \left(\log\left(\frac{f_{v(j),j}}{r_{v(j),j}}\right) - \log\left(\frac{f_{u(j),j}}{r_{u(j),j}}\right) \right)$$

where D_{uv} is the set of conditions, in which the DELs in u and v differ ($|D_{uv}|$ is the Hamming distance between u and v). Note that since T_P is scanned in preorder, LLR(u) is calculated before LLR(v), as required. In preprocess, we compute a table that contains the value $\log(f_{i,j}/r_{i,j}) - \log(f_{k,j}/r_{k,j})$ for every pair of DELs, e_i and e_k, and every condition c_j. Using this table, LLR(v) can be calculated given LLR(u) in time $c \cdot |D_{uv}|$, where, c is a very small constant. Thus, the total time complexity of computing the LLR score of all patterns is $O(|P| \cdot d + |C|)$, where d is the average Hamming distance in the MST (the second summand, |C|, is the time for the LLR computation of the root). In the human tissues dataset mentioned above, there are 79 tissues, but the average distance in T_P is only 1.31. Thus, using the MST gives a further 59-fold time improvement.

Enrichment scores

For each candidate motif, we use a subset S of its target genes ($S \subset T$) as a training set for learning a CWM, as described in the Results section. The set of all other genes in the expression data, denoted B_s ($B_s = B \backslash S$), is used for evaluating the fit between the CWM and the motif, as follows. Let W ($W \subseteq B_s$) be the set of genes, whose expression pattern obtained an LLR score higher than the current CWM cutoff (Allegro tries several cutoffs), excluding the training-set genes. Denote by b and w the subset of genes from B_s and from W, respectively, that contain at least one occurrence of the motif in their *cis*-regulatory sequence, i.e. $b = B_s \cap T$ and $w = W \cap T$. Allegro computes one of two supported enrichment scores, as specified by the user, to assess whether the motif is over-represented in W, i.e. whether w is significantly larger than expected, given B_s, W and b. The first score, called the HG enrichment score, uses the HG tail distribution to compute the probability of observing at least $|w|$ sequences in W with a motif hit, under the null hypothesis that the genes in W were drawn randomly, independently and without replacement from B_s:

$$\mathrm{HG\,score} = \mathrm{HGtail}(|B_s|, |W|, |b|, |w|) = \sum_{i=|w|}^{\min(|W|,|b|)} \frac{\binom{|b|}{i}\binom{|B_s|-|b|}{|W|-i}}{\binom{|B_s|}{|W|}}$$

The second score, called the binned enrichment score, accounts for cases where the expression values are correlated with the length or GC-content of the *cis*-regulatory sequences. In short, the genes are divided into bins according to the length and GC-content of their *cis*-regulatory sequence. The counts of the number of genes in each bin that passed the LLR cutoff and the number of genes with a hit in their sequence are used in order to estimate the overall enrichment. For exact details, see (11).

Clustering and motif-finding tools

K-means (20) and CLICK (21) were executed using the Expander gene expression analysis software (18). K-means was run twice—with $k = 10$, and with $k = 20$. CLICK was run with the 'homogeneity' parameter set to 0.3. Two motif-finding tools, Weeder and Amadeus, were applied on all clusters found by K-means and CLICK, excluding huge clusters with more than 900 genes. Weeder (v1.3) was executed with the 'medium T100 S' parameters and using the BG model files supplied with the software (22). Amadeus (v1.0) was run with its default settings (11).

GO functional analysis

For each motif discovered by Allegro in the tissues datasets, we ran the TANGO algorithm via the Expander software (18) to test whether the CWM targets of the motif are enriched for Gene Ontology biological process terms. TANGO performs a bootstrapping procedure to correct the enrichment p-values for multiple testing and account for the large overlaps between related GO terms. All results reported here obtained a p-value less than 10^{-9} and a corrected p-value less than 10^{-3}.

RESULTS

We developed a novel method, called Allegro, for simultaneous *de novo* discovery of regulatory sequence motifs and the expression profiles they induce in one or more genome-wide gene expression datasets. Given a candidate motif, Allegro learns an expression model that describes the shared expression profile of the genes, whose *cis*-regulatory sequence contains the motif. It then computes a p-value for the over-representation of the motif within the *cis*-regulatory sequences of the genes that best fit the expression profile. We implemented Allegro and integrated it with our Amadeus motif discovery platform. Amadeus executes a series of refinement phases to converge to high-scoring motifs. Each phase receives as input a list of candidate motifs, applies an algorithm for refining the list, and produces a set of improved candidates that constitute the starting point for the next phase. The output of Amadeus is a non-redundant list of top-scoring motifs, modeled using position weight matrices (PWMs).

Additional scoring functions and features available in Amadeus are described in (11). In the current study, motifs in each phase are evaluated using Allegro. Thus, the motifs reported by the algorithm are those that possess the highest correlation to the expression data in terms of the aforementioned p-value.

In the following sections we introduce the expression model used by Allegro and demonstrate its advantages over commonly used approaches. We then describe the algorithm Allegro applies to ascribe a p-value for a given motif. Finally, we present results of applying Allegro to several large-scale expression datasets representing a diverse set of biological systems and species, and compare them to those obtained by existing tools.

Modeling the expression profile of co-regulated genes

We developed a new method for modeling the expression profile shared by a group of co-regulated genes. Unlike existing approaches, it does not make complex statistical assumptions about the distribution of the expression values in each condition. Furthermore, unlike expression similarity measures employed by clustering techniques, our model is robust against extreme values and can describe profiles that differ across a very small number of conditions. The model is analogous to the PWM model for sequence motifs (23,24), with DNA bases substituted here by discrete expression levels, and the positions along the motif replaced by the experimental conditions.

Given continuous expression values, Allegro first transforms them into *discrete expression levels* (DELs, in short): $e_1, e_2, \ldots, e_l$. The number of expression levels (l) and the range of values that define each one are set by the user. For example, if the expression values are given in $\log_2$ ratios w.r.t. some base condition, then one may use three DELs, as illustrated in Supplementary Figure 1: Expression values above 1.0 are replaced by e_1 (or 'U', for 'Up-regulated'), values between −1.0 and 1.0 are replaced by e_2 (or 'S', for 'Similar to base condition') and values below −1.0 are replaced by e_3 (or 'D', for 'Down-regulated'). The DELs may also be defined using percentiles rather than cutoffs.

Let $c_1, c_2, \ldots, c_m$ be the set of m conditions in the given expression matrix. The expression model assigns to each condition a discrete probability distribution. Define an l x m matrix, called *condition frequency matrix* (CFM), in which column j holds the distribution of the DELs in condition c_j according to the model. Hence, the value in row i and column j is the probability of generating expression level e_i in condition c_j (Supplementary Figure 1). The background CFM, $R = \{r_{i,j}\}$, is computed from the observed DELs of all given genes; i.e. $r_{i,j}$ is the BG frequency of expression level e_i in condition c_j (see Methods section).

Given another CFM $F = \{f_{i,j}\}$, which models the expression levels of a transcriptional module T, we would like to assign to each gene a score that quantifies its similarity to F. To this end, we use the standard likelihood ratio approach, as follows. Let $e_{g(j)}$ $(1 \leq j \leq m)$ denote the DEL of gene g in condition c_j. The LLR score of g is the logarithm of the ratio between the probability of observing these expression levels under the assumption that gene g belongs to T, and the probability of observing them under the null hypothesis:

$$\text{LLR(expression of gene } g) = \sum_{j=1}^{m} \log\left(\frac{f_{g(j),j}}{r_{g(j),j}}\right)$$

The $l \times m$ matrix whose entries are $\log(f_{i,j}/r_{i,j})$ is called the *CWM*. The CWM can be used to classify genes as belonging to the transcriptional module T in the standard way: for a given threshold h, a gene is considered to belong to T if its LLR score is above h. In a sense, the CWM represents an expression motif similarly to the standard sequence motif representation using a PWM. In the next section we explain how the CWM and the threshold h are computed for a putative transcriptional module.

We tested how well the CWM model identifies the expression profile of known transcriptional modules, and compared its performance to that of popular expression metrics: Pearson correlation, Spearman's rank correlation, and Euclidean distance (4). The results show that in most cases (16 out of 18) our model describes the expression profile of TMs more accurately than existing approaches (Supplementary Table II). The experimental procedure and results are detailed in the Supplementary Data.

Learning the expression profile induced by a motif

For each candidate motif, Allegro tries to learn a CWM that describes the expression of (some of) its targets. If the motif represents BSs of a TF that is active in the measured conditions, Allegro will likely find a CWM that is characteristic of the motif's targets; otherwise, the expression values of the target genes are expected to behave like the BG distribution, and no such CWM will be found. Let T denote the set of genes whose *cis*-regulatory sequences contain at least one occurrence, or *hit*, of the motif M. Allegro finds a CWM that models the expression profile of T by executing a cross-validation-like procedure, illustrated in Supplementary Figure 2. First, it samples a training set from T and generates a CFM F based on the frequencies of DELs in that training set. A CWM is computed from F and from the background CFM, as explained earlier (see Methods section). Then, for all genes excluding those in the training set, it computes the LLR score described above. In order to ascertain that the motif M is over-represented in the genes with a high LLR score (i.e. genes whose expression is more similar to the profile represented by F than to the background CFM), Allegro computes one of two enrichment scores developed in Amadeus: the HG score and the binned enrichment score. The latter accounts for biases in the length and nucleotide composition of the regulatory sequences (see Methods section). Note that the training-set genes are ignored when computing the enrichment score in order to avoid over-fitting. The enrichment score is computed for several LLR cutoffs and the best one is chosen and Bonferroni-corrected for multiple testing. Allegro repeats this process for several training sets, which are sampled in a judicious procedure that takes into account both the expression and sequence data (see Supplementary Data). Finally, Allegro chooses the CWM that yielded the best enrichment score, and this score is set as the *p*-value of the motif. We use the term *CWM targets* to refer to the genes that passed the LLR cutoff of the top-scoring CWM. For an arbitrary motif, only a relatively small fraction of the CWM targets are also targets of the motif (i.e. contain a hit for the PWM in their *cis*-regulatory sequence), whereas, for a biologically relevant motif, the overlap between the set of its PWM targets and the set of its CWM targets is significantly large (in the sense of the enrichment score).

As described earlier, Allegro examines a large number of candidate motifs in a series of refinement phases. The motifs in each phase are ranked according to the above enrichment score. We implemented sophisticated data-structures and algorithms in order to speed-up the CWM learning procedure (see Methods section). The output of the Allegro algorithm is a list of transcriptional modules, each one comprised of a sequence motif (PWM) and an expression profile (CWM) that are highly correlated in terms of the genes they match.

Test case: human cell cycle

Whitfield *et al*. studied cell-cycle regulation using cDNA microarrays that measured gene expression profiles of HeLa cells over five time courses (15). In each time course, the cells were synchronized to the same cell-cycle phase by one of three different methods. In order to identify cell-cycle genes and the phases in which they are active, Whitfield *et al*. quantified the periodicity of the expression levels of each gene using Fourier transform, and compared it to that of known cell-cycle genes. Several studies utilized their findings to analyze the transcriptional programs underlying the cell-cycle phases (25–28).

In order to test the ability of our method to uncover transcriptional modules *ab initio* from a large mammalian dataset, we applied it to the cell-cycle data of Whitfield *et al*. The input to Allegro consisted of expression values across 111 time points and of 1200 bps-long promoter sequences of ~15 000 genes. Consistent with biological knowledge and previous studies, the three top-scoring motifs found by Allegro are the BS patterns of E2F and NF-Y (CCAAT-box), and the motif termed CHR (cell-cycle genes homology region), whose binding protein is yet to be discovered (29) (see Supplementary Data for information on how the motifs are matched to known BS patterns). As shown in Figure 2, the expression of the CWM targets of E2F peaks in the G_1/S phase, whereas genes associated with NF-Y and CHR are active in the G_2 and M phases. Importantly, these results were obtained by analyzing the expression and sequence data alone, without using any prior knowledge on periodicity of human cell-cycle or on known phase-specific genes.

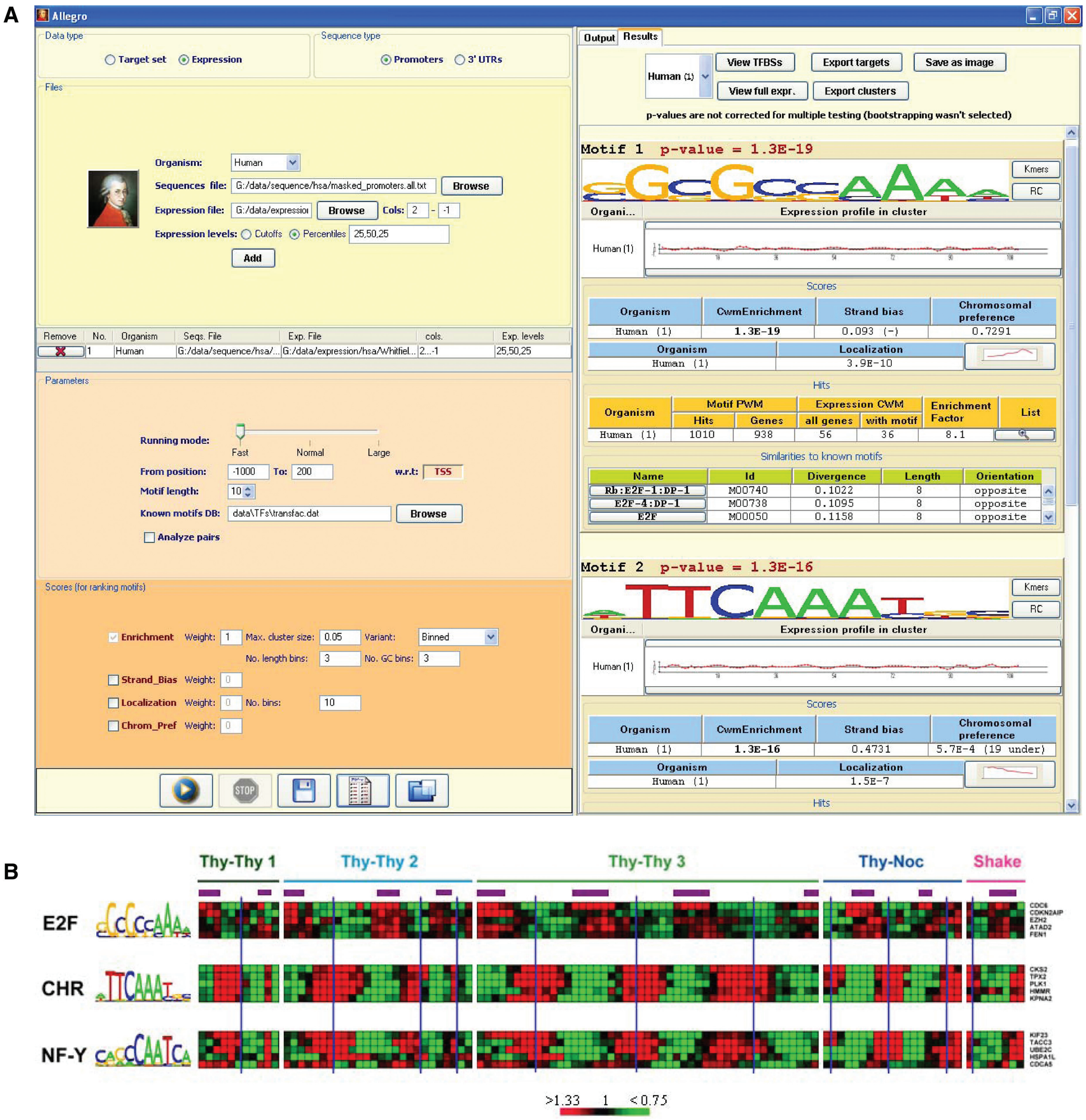
An additional test case on expression data of the innate immune response in mouse is described in the Supplementary Data.

Comparison to the two-step approach: yeast HOG pathway

The *Saccharomyces cerevisiae* high osmolarity glycerol (HOG) pathway is required for osmoadaptation. It contains two branches that activate the protein Hog1 via Pbs2, one containing Ssk1 and the other containing Sho1 and Ste11. O'Rourke *et al*. characterized the roles of Hog1, Pbs2, Ssk1, Sho1 and Ste11 in response to elevated osmolarity using whole-genome expression profiling (14). The expression data contain osmotic shock profiles of the WT strain, and of mutant strains in which components of the HOG pathway were knocked-out. The profiles were monitored at different levels of hyper-osmolarity at several time points. In addition, the transcriptional response of the WT strain to the mating pheromone α-factor was measured at four time points. Overall, the dataset consists of expression values of 5758 genes in 133 conditions.

The seven top-scoring motifs reported by Allegro for this dataset are the RRPE, PAC and STRE (stress response element) motifs, and the BS patterns of Rap1, MBF, Ste12 and Sko1 (Figure 3). Remarkably, all seven motifs are related to osmotic shock (30–33). For example, Msn2 and Msn4 mediate a general stress response through binding to STRE (31,34), and they are also controlled by Hog1 (33). Indeed, the CWM targets of STRE are up-regulated in the time series of exposure to high osmolarity. Another example, which provides further evidence of the sensitivity of our approach, is Sko1, one of the main TFs that control the specific response to hyper-osmolarity (33). Under normal conditions, Sko1 recruits the general repressor complex Tup1–Ssn6 and together they act to repress their target genes. After osmotic shock, Hog1 phosphorylates Sko1, resulting in decreased affinity for Tup1, and Sko1 then activates transcription by an unknown mechanism. Reassuringly, Allegro uncovered the Sko1 binding motif, and its CWM targets are considerably up-regulated in response to high osmolarity only in strains in which Hog1 and Pbs2 were not knocked out. See Supplementary Data for additional analysis of the results.

We applied the standard two-step approach to the HOG dataset to check whether the transcriptional modules discovered by Allegro can also be found using existing techniques. We first performed clustering using three methods—*k*-means with $k = 10$ and $k = 20$ (20), and the CLICK algorithm (21), which resulted in 38 clusters. Four of these clusters were huge (>900 genes, i.e. >20% of the entire gene set) and did not exhibit an interesting expression profile, so we ignored them. We then executed two motif finding tools on each of the 34 remaining clusters: Weeder (22), which out-performed 13 other tools in a large-scale assessment (35), and Amadeus, our recently published software (11). Following (11,35), from each such execution we examined the two top-scoring motifs reported by the motif finder. We thus examined a total of 68 motifs discovered by the clustering and motif-finding pipeline. As listed in Table 1, out of the seven motifs Allegro discovered, only four were found by the two-step approach—RRPE, PAC, MBF and STRE. We also applied the clustering and motif-finding tools developed by Slonim *et al*., Iclust (36) and FIRE (37). Again, only



Allegro

Data type: ○ Target set ● Expression

Sequence type: ● Promoters ○ 3' UTRs

Files

Organism: Human

Sequences file: G:/data/sequence/hsa/masked_promoters.all.txt Browse

Expression file: G:/data/expressio Browse Cols: 2 - -1

Expression levels: ○ Cutoffs ● Percentiles 25,50,25

Add

Remove	No.	Organism	Seqs. File	Exp. File	cols.	Exp. levels
☒	1	Human	G:/data/sequence/hsa/...	G:/data/expression/hsa/Whitfiel...	2...-1	25,50,25

Parameters

Running mode: Fast Normal Large

From position: -1000 To: 200 w.r.t: TSS

Motif length: 10

Known motifs DB: data\TFs\transfac.dat Browse

□ Analyze pairs

Scores (for ranking motifs)

☑ Enrichment Weight: 1 Max. cluster size: 0.05 Variant: Binned

No. length bins: 3 No. GC bins: 3

□ Strand_Bias Weight: 0

□ Localization Weight: 0 No. bins: 10

□ Chrom_Pref Weight: 0

Output Results

Human (1) View TFBSs Export targets Save as image

View full expr. Export clusters

p-values are not corrected for multiple testing (bootstrapping wasn't selected)

Motif 1 p-value = 1.3E-19

Kmers RC

Organi...	Expression profile in cluster
Human (1)	

Scores

Organism	CwmEnrichment	Strand bias	Chromosomal preference
Human (1)	**1.3E-19**	0.093 (-)	0.7291

Organism	Localization
Human (1)	3.9E-10

Hits

Organism	Motif PWM		Expression CWM		Enrichment Factor	List
	Hits	Genes	all genes	with motif		
Human (1)	1010	938	56	36	8.1	

Similarities to known motifs

Name	Id	Divergence	Length	Orientation
Rb:E2F-1:DP-1	M00740	0.1022	8	opposite
E2F-4:DP-1	M00738	0.1095	8	opposite
E2F	M00050	0.1158	8	opposite

Motif 2 p-value = 1.3E-16

Kmers RC

Organi...	Expression profile in cluster
Human (1)	

Scores

Organism	CwmEnrichment	Strand bias	Chromosomal preference
Human (1)	**1.3E-16**	0.4731	5.7E-4 (19 under)

Organism	Localization
Human (1)	1.5E-7

Hits

Figure 2. Results of Allegro on the human cell cycle dataset (15). (**A**) Screenshot of Allegro. The left panel presents the input parameters: organism, expression data file, scores, etc. The top-scoring motifs discovered by Allegro are shown in the output panel on the right. Additional information is displayed for each motif, such as the average expression profile of the CWM targets that contain a hit of the motif, statistics on the number of hits and their locations, similar binding patterns from Transfac or miRBase, and more. Here, the three top-scoring motifs reported by Allegro represent the BS patterns of key regulators of the human cell cycle: E2F, CHR (whose binding TF is unknown), and NF-Y (not shown). (**B**) Expression profiles of the five CWM targets with the highest LLR score of the three motifs found by Allegro. High and low expression values w.r.t. time 0 are colored in red and green, respectively. The purple bars represent S phase and the blue vertical lines indicate mitoses, as reported in (15). In agreement with biological knowledge and previous computational analyses (25–28,50), E2F induces genes mainly in the G_1/S phase, whereas CHR and NF-Y are highly specific to the G_2 and G_2/M phases.

four out of the seven motifs Allegro recovered were reported by FIRE. Specifically, the BS motifs of Sko1 and Ste12 were found by Allegro but not by any other method.

We could not compare Allegro to other published methods that infer motifs by simultaneous analysis of sequence and expression data (5–7), either because they are not publicly available or we could not execute them and obtain reasonable results.

Analysis of multiple datasets: tissue-specific regulators

A unique feature of Allegro is simultaneous analysis of multiple datasets from one or more species. Given several expression matrices and corresponding sequence data, Allegro explores the motif search space as described above. For each candidate motif, it computes its enrichment score in each of the datasets separately; i.e. it finds a CWM whose top-scoring genes have a significantly large

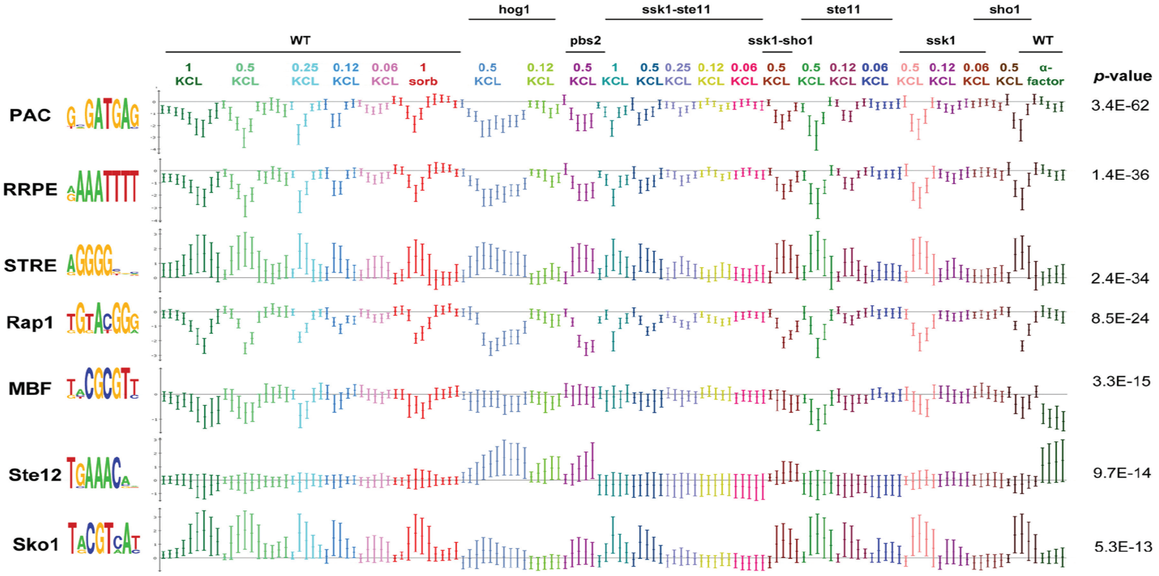


Figure 3. Results of Allegro on the yeast HOG pathway expression dataset (14). Allegro finds the motifs PAC, RRPE, STRE and the binding patterns of Rap1, MBF, Ste12 and Sko1. Each motif is presented together with the average expression profile (±1 SD) of its CWM targets which contain a hit for the motif in their promoter. The titles above the expression series indicate the yeast strain the expression was sampled from: WT, and knockout strains [indicated by the name(s) of the gene(s) that were knocked-out]. The concentrations of KCL and sorbitol are given in molar units.

Table 1. Results of Allegro and existing tools on the yeast HOG MAPK dataset (14)

Biological process	Motif/TF	Reference	K-means/CLICK Amadeus/Weeder	Iclust FIRE	Allegro
General stress response	RRPE	(31,72)	+	+	+
	PAC	(31,72)	+	+	+
	Rap1	(31)	–	+	+
HOG and pheromone response pathways	Sko1	(32,33)	–	–	+
	Ste12	(30,33)	–	–	+
	MBF	(33,73,74)	+	–	+
	Smp1	(32)	–	–	–
	Skn7	(32)	–	–	–
General stress response and HOG pathway	STRE	(31–33)	+	+	+

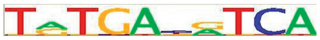
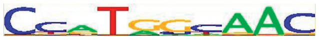
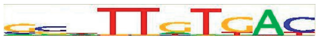
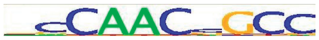
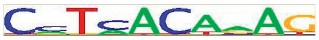
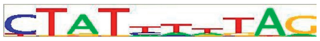
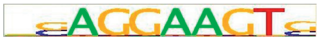
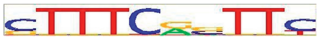
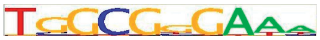
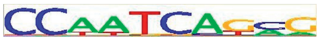
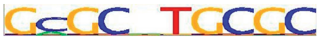
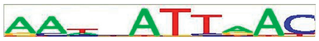
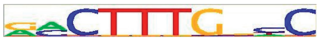
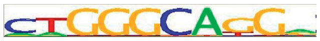
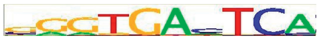
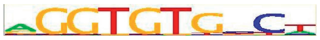
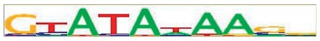
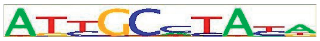
There are nine TFs and motifs known to be involved in the regulation of genes in the studied conditions. In a single execution, Allegro successfully recovered seven of these binding patterns as the seven top-scoring motifs. In contrast, only four motifs were discovered when the two-step approach was applied using various combinations of existing clustering and motif discovery tools.

overlap with the genes that contain the motif in their *cis*-regulatory sequence. Allegro then combines these scores into a single *p*-value using the *Z*-transform (38).

We tested this feature on the human and mouse gene atlas (16) in search of tissue-specific regulators. Given the expression levels of ~15 000 human and mouse genes across 79 human tissues and 61 mouse tissues, Allegro found known and novel motifs. The main results are summarized in Table 2. The motifs reported by Allegro are non-redundant: for every pair of reported motifs—M_1 and M_2—no more than 5% of their hits overlap, i.e. ≥95% of the occurrences of M_1 do not overlap any occurrence of M_2, and vice versa. Thus, each reported motif is likely to represent a biologically distinct binding pattern.

The top-scoring motif is the binding pattern of CREB/ATF, and its target genes are up-regulated in testis tissues (Supplementary Figure 3). Indeed, CREB is known to activate transcription of genes essential for proper germ cell differentiation (39), and its disruption in mice severely impairs spermatogenesis (40,41). Allegro reported four additional testis-specific motifs: RFX, MYB and two novel motifs (motifs 2–5 in Table 2). Members of the RFX and MYB families are expressed at high levels in the testis (42–46). Interestingly, all three known testis-related TF families—CREB, RFX and MYB—have testis-specific gene products (42,46,47). We performed functional analysis on the sets of CWM targets of the motifs found by Allegro in order to identify GO terms over-represented in these sets (see Methods section). Reassuringly, the CWM targets of all five testis-related motifs in both species are highly enriched for spermatogenesis.

Table 2. Main results of Allegro for the combined analysis of the human and mouse tissue gene atlas datasets (16)

	Logo	TF/motif	*p*-value	Tissues	Gene Ontology (BP; CC)
1		CREB/ATF	10^{-32}	**Testis**: Testis, testis germ cell, testis interstitial, testis Leydig cell, testis seminiferous tubule	Spermato-genesis; Flagellum
2		RFX	10^{-24}		
3		–	10^{-23}		
4		MYB	10^{-21}		
5		–	10^{-15}		
6		MEF2	10^{-29}	**Muscle**: Heart, skeletal muscle, tongue	Muscle contraction; Myofibril
7		ETS/ELF	10^{-27}	**Immune system**: Peripheral blood cells, B/T-cells, lymphnode, BM myeloid, thymus	Immune response; Plasma membrane
8		IRF	10^{-15}		
9		E2F	10^{-23}	**Proliferating cells**: Oocyte, embryo, bone marrow, thymus, lymphoblasts, cancers	Cell cycle, DNA replication; Chromosome
10		NF-Y	10^{-13}		
11		NRF1	10^{-14}		
12		HNF1	10^{-22}	**Digestive tract**: Liver, kidney, pancreas, intestine	Metabolism (carboxylic acid, lipid, amine, . . .); Mitochondrion
13		HNF4	10^{-21}		
14		–	10^{-18}	**Keratinocytes**: Epidermis, tongue, digits	Epidermis development, keratinization; Intermediate filament cytoskeleton
15		AP1/FOS	10^{-16}		
16		T-box	10^{-15}		
17		TATA	10^{-14}		
18*		–	10^{-14}	**Oocyte**: Oocyte, fertilized egg	Cell cycle; Nucleus

The table lists all motifs with *p*-value ≤ 10^{-15} (combined score for human and mouse datasets), as well as several motifs with high similarity to known binding patterns (TATA, Nrf-1 and NF-Y). Three of the motifs are apparently novel. In addition, a novel motif that obtained a significant *p*-value (10^{-14}) only in the mouse dataset is listed. Similar known binding patterns from the Transfac database are shown in the 'TF/Motif' column. The 'Tissues' column lists the tissues in which the target genes of each motif are up-regulated. Some tissues were sampled in only one of the two organisms. The 'Gene Ontology' column specifies the most enriched biological process (BP) and cellular component (CC) GO terms in the CWM targets of each motif.

**p*-value, tissues and GO terms for motif #18 are based only on the mouse dataset; oocyte and fertilized egg were not sampled in human.

Additional known TF-tissue associations recovered by Allegro include MEF2, whose target genes are induced in heart, skeletal muscle and tongue (48) (Supplementary Figure 4); HNF1 and HNF4, which induce genes in liver, and, to a lesser extent, in kidney, pancreas and intestine (49) (Supplementary Figure 5); and the cell-cycle regulators E2F, NF-Y and NRF1, whose targets are up-regulated in various types of proliferating cells (25,27,50,51). We also found four motifs whose targets are up-regulated in the epidermis and related tissues, such as tongue and digits: the AP1/FOS-binding pattern, T-box, TATA and a novel motif (motifs 14–17 in Table 2). There is evidence of the involvement of FOS and TBP (TATA binding protein) in the regulation of keratinocyte proliferation (52,53).

Allegro discovered a novel motif whose target genes are highly induced in murine oocytes (motif #18 in Table 2, see also Supplementary Figure 6). Oocytes are not among the tested tissues in human, so we do not know whether this enrichment is conserved. A partial list of the putative targets of the motif is given in Supplementary Table III.

To further test the ability of Allegro to simultaneously analyze multiple expression datasets, we applied it on three datasets that recorded expression levels of fly (*Drosophila melanogaster*) genes during various developmental stages (54–56) (see Supplementary Data). Allegro discovered known and novel motifs associated with various developmental profiles. The 20 top-scoring motifs are listed in Supplementary Table IV. Of note, this list includes the top seven core promoter motifs found by Ohler (57), indicating that core promoter *cis*-regulatory elements play an important role in fly development. Another interesting example is the TAGteam motif, which was recently identified and shown experimentally

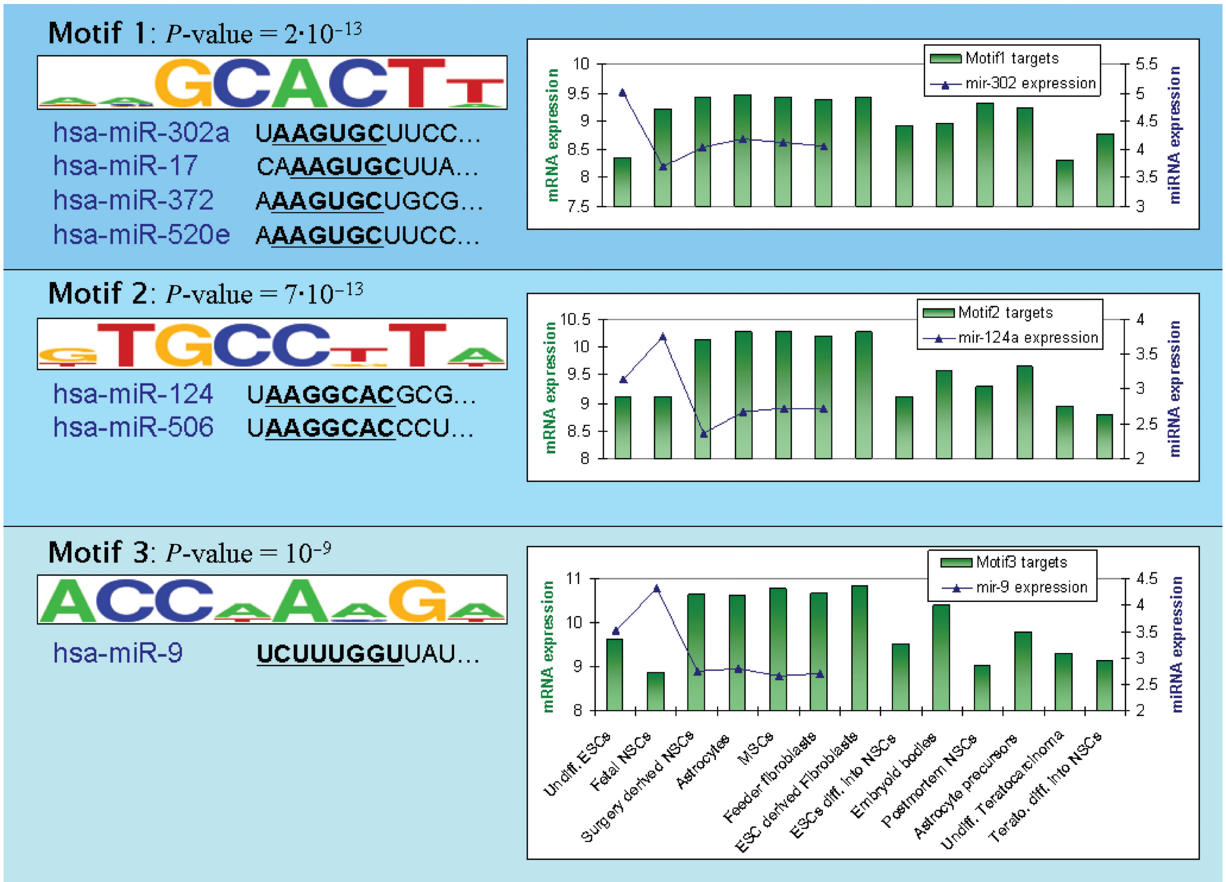


Figure 4. The top three 3′ UTR motifs identified in the stem cells dataset (19). On the left, the motif p-value and logo are presented along with the first 11 bases (starting from the 5′ base of the mature microRNA) of miRNAs with a seed that matches the reverse complement of the motif. For the first motif, only one miRNA from each of the four matching miRNA families is presented. For each motif, the graph on the right shows the average expression values (in $\log_2$ scale) of the corresponding CWM targets that contain a hit for the motif. Each bar represents the average expression level in one of the cell types (ESCs/NSCs/MSCs—embryonic/neural/mesenchymal stem cells; 'Undiff.'—Undifferentiated, 'diff.'—differentiated, 'Terato.'—Teratocarcinoma; see also Supplementary Table V; the full expression profile of targets of motif 1 in all 124 samples is shown in Supplementary Figure 9). The graph also shows the expression levels (in $\log_2$ scale) of the matching miRNA(s): mir-302 for motif 1 (average expression over all mir-302 family members), mir-124 for motif 2 and mir-9 for motif 3. miRNA expression levels are presented only for the cell types profiled in (63). Evidently, the expression profiles of the motif targets and those of the matching miRNAs are anti-correlated, increasing our confidence that the discovered motifs represent miRNAs that are active in the relevant cells.

to induce early zygotic genes (58,59). Allegro recovered this motif and the expression profile it induces (Supplementary Figure 7).

3′ UTR analysis: human stem cells

Stem cells, and in particular embryonic stem cells (ESCs), have a unique ability to differentiate into diverse cell types. This multipotency (or pluripotency in case of ESCs) is maintained by a variety of epigenetic mechanisms, including DNA methylation, chromatin modifications and miRNAs (60). Analysis of sequence motifs in 3′ UTRs of genes up- or down-regulated in various types of stem cells carries the promise of identifying key miRNAs maintaining the stem cell differentiation capabilities. Mueller *et al.* (19) profiled gene expression in 124 cell samples, including a variety of stem cells. The analysis of 3′ UTR motifs in this large dataset is hindered by biases in 3′ UTR length and base composition (Supplementary Table V). For example, proliferating cells, such as ESCs, are known to express genes with 3′ UTRs that are much shorter than those of genes expressed in other cell types (61). In contrast, genes specific to the nervous system are known to have particularly long UTRs (62). This leads to an almost 2-fold difference in 3′ UTR length between genes up-regulated in undifferentiated ESCs and genes up-regulated in fetal neural stem cells (NSCs) (Supplementary Figure 8).

We applied Allegro to search for enriched motifs in the 3′ UTRs of the Mueller *et al.* dataset. Due to the biases mentioned above, we used the binned enrichment score to compute the over-representation of each candidate motif in the set of CWM targets fitted to it. The results are presented in Figure 4. The top-scoring motif (GCACTT) is the reverse complement of the hexamer AAGTGC, which appears in the seed sequences of several miRNA families (mir-17, mir-302, mir-290 and mir-515), all of which are among the most highly expressed miRNAs in human and mouse ESCs (63,64). Indeed, genes reported by Allegro as putative targets of these miRNA families are evidently down-regulated in human ESCs compared to other cell types (Figure 4). Interestingly, as shown in Supplementary Figure 9, these genes are also down-regulated in a subset of NSCs, which were differentiated from ESCs or from teratocarcinoma, indicating that it is possible that the expression of these miRNA families is not down-regulated immediately upon differentiation.

The second most significant motif reported by Allegro is GTGCCTT, which corresponds to the seeds of mir-506 and mir-124a. Inspection of the expression pattern of the CWM targets (Figure 4) shows that genes carrying this motif are generally down-regulated in less differentiated cells (ESCs, NSCs, embryoid bodies and teratocarcinoma) compared to more differentiated ones [mesenchymal stem cells (MSCs), fibroblasts and astrocytes]. Mir-506 did not show any differential expression between ESCs, NSCs and differentiated cells (63), and was not detected in any tissue in a recent comprehensive sequencing effort (62); thus, it is not likely to be the regulator of this gene set. Mir-124a is known to be abundant and functional in the neural cell lineage (65), and is up-regulated in NSCs compared to MSCs and fibroblasts (63). However, it is also up-regulated in NSCs compared to ESCs (63), while the expression levels of the CWM targets do not appear to differ between these two cell types. It is possible, therefore, that the regulation of the CWM targets is carried out by mir-124a alongside other regulatory mechanisms that may or may not involve miRNAs.

The third motif reported by Allegro (ACCAAAG) matches the seed of mir-9. The expression pattern of its targets shows down-regulation in NSCs compared to differentiated cells, with intermediate levels in ESCs and in teratocarcinoma. Mir-9 is expressed specifically in the neural lineage (62,63) and is known to have an active role in neurogenesis (66).

Neither the standard two-step approach (clustering with *k*-means or CLICK, followed by motif finding using Weeder or Amadeus), nor Allegro with the HG enrichment score, recovered the above three motifs. This emphasizes the importance of accounting for sequence biases when conducting *cis*-regulatory motif finding.

DISCUSSION

In this work we present Allegro, a software platform that analyzes genomic sequences and expression datasets to infer transcriptional modules—groups of genes that are co-expressed along all or some of the experimental conditions and share an enriched regulatory motif in their promoters or 3′ UTRs. This single-step methodology, which infers transcriptional modules by simultaneously analyzing the sequence and expression data, utilizes all available information throughout the entire analysis, giving it a clear advantage over the standard two-step approach. Allegro employs a powerful motif enumeration engine and our CWM model to discover sequence motifs and their associated expression profiles without relying on pre-defined types of distribution to model the sequence and expression data. Unlike the vast majority of motif-finding tools, Allegro does not rely on pre-computed *k*-mer counts to construct a sequence model; and, unlike most clustering metrics and existing algorithms for combined sequence-expression analysis, it does not assume a Gaussian distribution of the expression values. Instead, Allegro utilizes the *cis*-regulatory sequences and expression values of all the analyzed genes (typically, the entire genome) as a reference set against which to evaluate the statistical significance of the overlap between each sequence motif and the expression profile fitted to its targets.

Another major contribution of the current study is the CWM, a novel non-parametric model for describing the common expression profile of a group of co-regulated genes. The model gives a likelihood ratio to the group using discrete expression levels. It makes no assumptions about the type of distribution of the expression values, and is robust against extreme values. Unlike similarity metrics, a CWM can describe an expression profile that differs from the background expression levels across a very small number of conditions (even a single condition), and can, in effect, assign a different weight (i.e. contribution) to each condition. Furthermore, a CWM can model more complex transcriptional patterns than existing methods. For example, it can describe the effect of a TF that activates some genes and suppresses others in the same conditions [e.g. Oct4 and Nanog (67)]. As we demonstrated for experimentally derived TF target sets and for functionally related annotated groups of genes, the CWM captures their distinct expression profiles more accurately than commonly used metrics (Supplementary Table II). A detailed discussion on the shortcomings of existing expression similarity measures is given in the Supplementary Data. While in this study we used the CWM in the context of motif finding, it can be applied in other gene expression analysis tasks, such as functional analysis (i.e. identifying GO terms whose genes exhibit a distinct expression profile).

We applied Allegro to several large-scale gene expression datasets in human, mouse, fly and yeast. Our results indicate that in a single run, and without any prior knowledge of known binding patterns or the characteristics of the transcriptional modules (e.g. the number of modules, their size and the overlap between them), Allegro successfully recovers the correct TF/miRNA motifs and reports them as the top-scoring motifs. The transcriptional modules found by Allegro are highly heterogeneous in terms of their expression profiles. For example, the cell-cycle regulators induce very subtle and noisy cyclic patterns in the human cell cycle dataset. The yeast HOG pathway dataset, on the other hand, consists of diverse time-series experiments, and, accordingly, the relevant TFs induce distinct complex expression profiles, some of which differ from the BG distribution in only a small fraction of the conditions.

One of the unique features of Allegro is joint analysis of multiple expression datasets. Unlike some comparative analysis techniques, Allegro does not search for conserved motifs within aligned promoter sequences, since the conservation of TFBSs is, in many cases, very limited across species (68–70). Instead, for each candidate motif it examines, Allegro utilizes the information from all supplied datasets by combining the scores the motif attained on them into a single p-value, thus improving the accuracy of the analysis. We demonstrated this feature on the human and mouse tissues datasets (16), in which Allegro found 18 distinct motifs in various tissue types (Table 2). Notably, some of the tissue-specific motifs obtained borderline p-values in one or both species. Many of these motifs were not reported by Allegro when it was applied on each dataset separately (data not shown), underscoring the importance of combined analysis of multiple datasets for increased sensitivity. For example, E2F received a p-value of 4×10^{-11} in the human data, which is within the range of random scores given the huge number of candidate motifs considered by the algorithm; the combined human–mouse p-value of 10^{-23} is statistically significant. Perhaps this, together with the binned score, is why other methods failed to recover some of the well-known TF-tissue associations. Two cases in point: Elemento *et al.* applied their Iclust and FIRE tools on the human and mouse datasets separately, and did not discover CREB/ATF, RFX, MEF2, IRF, HNF1 and HNF4 (37). When Xie *et al.* searched for conserved promoter elements and tested whether they were tissue-specific (71), they failed to find many of the known TF-tissue associations such as HNF1 and HNF4 in liver, and E2F in proliferating cells. In addition to known TFs, Allegro reported novel motifs that attained statistically significant scores. Experiments are required to verify and study their regulatory roles. Additional novel motifs were discovered by Allegro in fly promoters using three expression datasets of *Drosophila* developmental stages (Supplementary Table IV).

Our analysis of the stem cells dataset demonstrates the ability of Allegro to reverse-engineer transcriptional programs regulated by miRNAs. Using the binned enrichment score, Allegro was able to overcome the two main obstacles in 3′ UTR sequence analysis: length heterogeneity and GC-content bias. The three top-scoring motifs identified by Allegro correspond to three miRNA families, indicating that these families are among the main post-transcriptional regulators in ESCs and NSCs. In particular, the top-scoring motif corresponds to a miRNA seed sequence that was recently shown to be highly dominant in human and mouse ESCs (63,64). The results of Allegro further highlight the importance of the miRNA families carrying this seed sequence in ESC biology. Finally, we show evidence of activity of miRNA carrying this seed sequence in several NSC lines for which miRNA expression profiles are not available. Technologies to accurately measure miRNA expression levels are maturing, but are still inferior in fidelity to mRNA profiling. As we have shown, using sequence analysis and mRNA profiles, we can predict the activity of miRNAs without the direct measurement of miRNA expression.

Due to the flexibility of Allegro's methodology and interface, it is suitable for a broad range of motif discovery tasks. For example, in addition to the HG or binned enrichment score, motifs can be evaluated using other scores we developed previously that measure global features of the distribution of the motif hits: localization along the promoters, strand bias and chromosomal preference (11). Allegro can simultaneously analyze promoter or 3′ UTR sequences and multiple genome-wide expression datasets from several species and combine all available information for optimal results. Running time on a standard PC is between a few minutes and several hours, depending primarily on the size of the expression data. We developed a user-friendly graphical interface, making Allegro accessible to a wide range of users. In order to help the user understand the results of the analysis, Allegro's graphical interface displays additional information and statistics on each reported motif, such as the scores it attained, its putative targets and their expression profile, similar known motifs from Transfac/miRBase, and more. The Allegro software (a standalone Java application) is available at http://acgt.cs.tau.ac.il/allegro.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

This work was supported in part by the Israel Science Foundation (grant 802/08 and Converging Technologies Program grant 1767.07). Igor Ulitsky was supported in part by a fellowship from the Edmond J. Safra Bioinformatics program at Tel Aviv University. Funding for open access charge: ISF 802/08 and 1767.07.

Conflict of interest statement. None declared.

REFERENCES

1. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
2. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.
3. Wyrick,J.J. and Young,R.A. (2002) Deciphering gene expression regulatory networks. *Curr. Opin. Genet. Dev.*, **12**, 130–136.
4. Jiang,D., Tang,C. and Zhang,A. (2004) Cluster analysis for gene expression data: a survey. *IEEE Trans. Knowl. Data Eng.*, **16**, 1370–1386.
5. Holmes,I. and Bruno,W.J. (2000) Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 202–210.
6. Segal,E., Yelensky,R. and Koller,D. (2003) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, **19(Suppl. 1)**, i273–i282.
7. Reiss,D.J., Baliga,N.S. and Bonneau,R. (2006) Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, **7**, 280.
8. Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.
9. Conlon,E.M., Liu,X.S., Lieb,J.D. and Liu,J.S. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA*, **100**, 3339–3344.
10. Eden,E., Lipson,D., Yogev,S. and Yakhini,Z. (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol.*, **3**, 39.
11. Linhart,C., Halperin,Y. and Shamir,R. (2008) Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets. *Genome Res.*, **18**, 1180–1189.

12. Birney,E., Andrews,T.D., Bevan,P., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cuff,J., Curwen,V., Cutts,T. *et al.* (2004) An overview of Ensembl. *Genome Res.*, **14**, 925–928.
13. Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.
14. O'Rourke,S.M. and Herskowitz,I. (2004) Unique and redundant roles for HOG MAPK pathway components as revealed by whole-genome expression analysis. *Mol. Biol. Cell.*, **15**, 532–542.
15. Whitfield,M.L., Sherlock,G., Saldanha,A.J., Murray,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. *et al.* (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.
16. Su,A.I., Wiltshire,T., Batalov,S., Lapp,H., Ching,K.A., Block,D., Zhang,J., Soden,R., Hayakawa,M., Kreiman,G. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
17. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
18. Shamir,R., Maron-Katz,A., Tanay,A., Linhart,C., Steinfeld,I., Sharan,R., Shiloh,Y. and Elkon,R. (2005) EXPANDER–an integrative program suite for microarray data analysis. *BMC Bioinformatics*, **6**, 232.
19. Muller,F.J., Laurent,L.C., Kostka,D., Ulitsky,I., Williams,R., Lu,C., Park,I.H., Rao,M.S., Shamir,R., Schwartz,P.H. *et al.* (2008) Regulatory networks define phenotypic classes of human stem cell lines. *Nature*, **455**, 401–405.
20. MacQueen,J. (1965) Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, CA: University of California Press, pp. 281–297.
21. Sharan,R. and Shamir,R. (2000) CLICK: a clustering algorithm with applications to gene expression analysis. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 307–316.
22. Pavesi,G., Mauri,G. and Pesole,G. (2001) An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, **17(Suppl. 1)**, S207–S214.
23. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
24. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
25. Elkon,R., Linhart,C., Sharan,R., Shamir,R. and Shiloh,Y. (2003) Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res.*, **13**, 773–780.
26. Tabach,Y., Milyavsky,M., Shats,I., Brosh,R., Zuk,O., Yitzhaky,A., Mantovani,R., Domany,E., Rotter,V. and Pilpel,Y. (2005) The promoters of human cell cycle genes integrate signals from two tumor suppressive pathways during cellular transformation. *Mol. Syst. Biol.*, **1**, 0022.
27. Linhart,C., Elkon,R., Shiloh,Y. and Shamir,R. (2005) Deciphering transcriptional regulatory elements that encode specific cell cycle phasing by comparative genomics analysis. *Cell Cycle*, **4**, 1788–1797.
28. Zhu,Z., Shendure,J. and Church,G.M. (2005) Discovering functional transcription-factor combinations in the human cell cycle. *Genome Res.*, **15**, 848–855.
29. Zhu,W., Giangrande,P.H. and Nevins,J.R. (2004) E2Fs link the control of G1/S and G2/M transcription. *Embo J.*, **23**, 4615–4626.
30. Bardwell,L. (2004) A walk-through of the yeast mating pheromone response pathway. *Peptides*, **25**, 1465–1476.
31. Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.
32. Hohmann,S. (2002) Osmotic stress signaling and osmoadaptation in yeasts. *Microbiol. Mol. Biol. Rev.*, **66**, 300–372.
33. O'Rourke,S.M., Herskowitz,I. and O'Shea,E.K. (2002) Yeast go the whole HOG for the hyperosmotic response. *Trends Genet.*, **18**, 405–412.
34. Martinez-Pastor,M.T., Marchler,G., Schuller,C., Marchler-Bauer,A., Ruis,H. and Estruch,F. (1996) The Saccharomyces cerevisiae zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE). *EMBO J.*, **15**, 2227–2235.
35. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
36. Slonim,N., Atwal,G.S., Tkacik,G. and Bialek,W. (2005) Information-based clustering. *Proc. Natl Acad. Sci. USA*, **102**, 18297–18302.
37. Elemento,O., Slonim,N. and Tavazoie,S. (2007) A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell*, **28**, 337–350.
38. Whitlock,M.C. (2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.*, **18**, 1368–1373.
39. Don,J. and Stelzer,G. (2002) The expanding family of CREB/CREM transcription factors that are involved with spermatogenesis. *Mol. Cell Endocrinol.*, **187**, 115–124.
40. Hummler,E., Cole,T.J., Blendy,J.A., Ganss,R., Aguzzi,A., Schmid,W., Beermann,F. and Schutz,G. (1994) Targeted mutation of the CREB gene: compensation within the CREB/ATF family of transcription factors. *Proc. Natl Acad. Sci. USA*, **91**, 5647–5651.
41. Blendy,J.A., Kaestner,K.H., Schmid,W., Gass,P. and Schutz,G. (1996) Targeting of the CREB gene leads to up-regulation of a novel CREB mRNA isoform. *EMBO J.*, **15**, 1098–1106.
42. Morotomi-Yano,K., Yano,K., Saito,H., Sun,Z., Iwama,A. and Miki,Y. (2002) Human regulatory factor X 4 (RFX4) is a testis-specific dimeric DNA-binding protein that cooperates with other human RFX members. *J. Biol. Chem.*, **277**, 836–842.
43. Grimes,S.R. (2004) Testis-specific transcriptional control. *Gene*, **343**, 11–22.
44. Mettus,R.V., Litvin,J., Wali,A., Toscani,A., Latham,K., Hatton,K. and Reddy,E.P. (1994) Murine A-myb: evidence for differential splicing and tissue-specific expression. *Oncogene*, **9**, 3077–3086.
45. Oh,I.H. and Reddy,E.P. (1999) The myb gene family in cell growth, differentiation and apoptosis. *Oncogene*, **18**, 3017–3033.
46. Sitzmann,J., Noben-Trauth,K., Kamano,H. and Klempnauer,K.H. (1996) Expression of B-Myb during mouse embryogenesis. *Oncogene*, **12**, 1889–1894.
47. Huang,X., Zhang,J., Lu,L., Yin,L., Xu,M., Wang,Y., Zhou,Z. and Sha,J. (2004) Cloning and expression of a novel CREB mRNA splice variant in human testis. *Reproduction*, **128**, 775–782.
48. Black,B.L. and Olson,E.N. (1998) Transcriptional control of muscle development by myocyte enhancer factor-2 (MEF2) proteins. *Annu. Rev. Cell Dev. Biol.*, **14**, 167–196.
49. Kuo,C.J., Conley,P.B., Chen,L., Sladek,F.M., Darnell,J.E., , and Crabtree,G.R. (1992) A transcriptional hierarchy involved in mammalian cell-type specification. *Nature*, **355**, 457–461.
50. Dimova,D.K. and Dyson,N.J. (2005) The E2F transcriptional network: old acquaintances with new faces. *Oncogene*, **24**, 2810–2826.
51. Cam,H., Balciunaite,E., Blais,A., Spektor,A., Scarpulla,R.C., Young,R., Kluger,Y. and Dynlacht,B.D. (2004) A common set of gene regulatory networks links metabolism and growth inhibition. *Mol. Cell*, **16**, 399–411.
52. Mehic,D., Bakiri,L., Ghannadan,M., Wagner,E.F. and Tschachler,E. (2005) Fos and jun proteins are specifically expressed during differentiation of human keratinocytes. *J. Invest. Dermatol.*, **124**, 212–220.
53. Fadloun,A., Kobi,D., Pointud,J.C., Indra,A.K., Teletin,M., Bole-Feysot,C., Testoni,B., Mantovani,R., Metzger,D., Mengus,G. *et al.* (2007) The TFIID subunit TAF4 regulates keratinocyte proliferation and has cell-autonomous and non-cell-autonomous tumour suppressor activity in mouse epidermis. *Development*, **134**, 2947–2958.
54. Hooper,S.D., Boue,S., Krause,R., Jensen,L.J., Mason,C.E., Ghanim,M., White,K.P., Furlong,E.E. and Bork,P. (2007) Identification of tightly regulated groups of genes during Drosophila melanogaster embryogenesis. *Mol. Syst. Biol.*, **3**, 72.
55. Arbeitman,M.N., Furlong,E.E., Imam,F., Johnson,E., Null,B.H., Baker,B.S., Krasnow,M.A., Scott,M.P., Davis,R.W. and White,K.P.

(2002) Gene expression during the life cycle of Drosophila melanogaster. *Science*, **297**, 2270–2275.

56. Spellman,P.T. and Rubin,G.M. (2002) Evidence for large domains of similarly expressed genes in the Drosophila genome. *J. Biol.*, **1**, 5.
57. Ohler,U., Liao,G.C., Niemann,H. and Rubin,G.M. (2002) Computational analysis of core promoters in the Drosophila genome. *Genome Biol.*, **3**, 0087.
58. Ten Bosch,J.R., Benavides,J.A. and Cline,T.W. (2006) The TAGt eam DNA motif controls the timing of Drosophila pre-blastoderm transcription. *Development*, **133**, 1967–1977.
59. De Renzis,S., Elemento,O., Tavazoie,S. and Wieschaus,E.F. (2007) Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. *PLoS Biol.*, **5**, 117.
60. Bibikova,M., Laurent,L.C., Ren,B., Loring,J.F. and Fan,J.B. (2008) Unraveling epigenetic regulation in embryonic stem cells. *Cell Stem Cell*, **2**, 123–134.
61. Sandberg,R., Neilson,J.R., Sarma,A., Sharp,P.A. and Burge,C.B. (2008) Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science*, **320**, 1643–1647.
62. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
63. Laurent,L.C., Chen,J., Ulitsky,I., Mueller,F.J., Lu,C., Shamir,R., Fan,J.B. and Loring,J.F. (2008) Comprehensive microRNA profiling reveals a unique human embryonic stem cell signature dominated by a single seed sequence. *Stem Cells*, **26**, 1506–1516.
64. Calabrese,J.M., Seila,A.C., Yeo,G.W. and Sharp,P.A. (2007) RNA sequence analysis defines Dicer's role in mouse embryonic stem cells. *Proc. Natl Acad. Sci. USA*, **104**, 18097–18102.
65. Cao,X., Pfaff,S.L. and Gage,F.H. (2007) A functional study of miR-124 in the developing neural tube. *Genes Dev.*, **21**, 531–536.
66. Krichevsky,A.M., Sonntag,K.C., Isacson,O. and Kosik,K.S. (2006) Specific microRNAs modulate embryonic stem cell-derived neurogenesis. *Stem Cells*, **24**, 857–864.
67. Loh,Y.H., Wu,Q., Chew,J.L., Vega,V.B., Zhang,W., Chen,X., Bourque,G., George,J., Leong,B., Liu,J. *et al.* (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.*, **38**, 431–440.
68. Borneman,A.R., Gianoulis,T.A., Zhang,Z.D., Yu,H., Rozowsky,J., Seringhaus,M.R., Wang,L.Y., Gerstein,M. and Snyder,M. (2007) Divergence of transcription factor binding sites across related yeast species. *Science*, **317**, 815–819.
69. Lin,C.Y., Vega,V.B., Thomsen,J.S., Zhang,T., Kong,S.L., Xie,M., Chiu,K.P., Lipovich,L., Barnett,D.H., Stossi,F. *et al.* (2007) Whole-genome cartography of estrogen receptor alpha binding sites. *PLoS Genet.*, **3**, e87.
70. Odom,D.T., Dowell,R.D., Jacobsen,E.S., Gordon,W., Danford,T.W., MacIsaac,K.D., Rolfe,P.A., Conboy,C.M., Gifford,D.K. and Fraenkel,E. (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.*, **39**, 730–732.
71. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
72. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.*, **296**, 1205–1214.
73. Escote,X., Zapater,M., Clotet,J. and Posas,F. (2004) Hog1 mediates cell-cycle arrest in G1 phase by the dual targeting of Sic1. *Nat. Cell Biol.*, **6**, 997–1002.
74. Gartner,A., Jovanovic,A., Jeoung,D.I., Bourlat,S., Cross,F.R. and Ammerer,G. (1998) Pheromone-dependent G1 cell cycle arrest requires Far1 phosphorylation, but may not involve inhibition of Cdc28-Cln2 kinase, *in vivo*. *Mol. Cell. Biol.*, **18**, 3681–3691.

4. Deciphering transcriptional regulatory elements that encode specific cell-cycle phasing by comparative genomics analysis

[Cell Cycle 4:12, 1788-1797, December 2005]; ©2005 Landes Bioscience

Report

Deciphering Transcriptional Regulatory Elements that Encode Specific Cell Cycle Phasing by Comparative Genomics Analysis

Chaim Linhart[1]
Ran Elkon[2]
Yosef Shiloh[2]
Ron Shamir[1,*]

[1]School of Computer Science; [2]The David and Inez Myers Laboratory for Genetic Research; Department of Molecular Genetics and Biochemistry; Sackler School of Medicine; Tel Aviv University; Tel Aviv, Israel

*Correspondence to: Ron Shamir; School of Computer Science; Tel Aviv University; Tel Aviv, 69978 Isreal; Tel: +972.3.640.5383; Fax: +972.3.640.5384; Email: rshamir@tau.ac.il

Received 08/18/05; Accepted 09/13/05

Previously published online as a Cell Cycle E-publication:
http://www.landesbioscience.com/journals/cc/abstract.php?id=2173

KEY WORDS

E2F, NF-Y, CHR, B-Myb, cell cycle regulation, transcriptional network

ACKNOWLEDGEMENTS

R. E. is a Joseph Sassoon Fellow. This study was supported in part by a research grant from the Ministry of Science and Technology, Israel. R.S was also supported in part by the Israel Science Foundation (grant 309/02).

SUPPLEMENTARY MATERIAL

Supplementary Material can be found at: http://www.landesbioscience.com/cc/supplement/linhartCC4-12-sup.pdf

ABSTRACT

Transcriptional regulation is a major tier in the periodic engine that mobilizes cell cycle progression. The availability of complete genome sequences of multiple organisms holds promise for significantly improving the specificity of computational identification of functional elements. Here, we applied a comparative genomics analysis to decipher transcriptional regulatory elements that control cell cycle phasing. We analyzed genome-wide promoter sequences from 12 organisms, including worm, fly, fish, rodents and human, and identified conserved transcriptional modules that determine the expression of genes in specific cell cycle phases. We demonstrate that a canonical E2F signal encodes for expression highly specific to the G_1/S phase, and that a cis-regulatory module comprising CHR-NF-Y elements dictates expression that is restricted to the G_2 and G_2/M phases. B-Myb binding site signatures occur in many of the CHR-NF-Y target genes, suggesting a specific role for this triplet in the regulation of the cell cycle transcriptional program. Remarkably, E2F signals are conserved in promoters of G_1/S genes in all organisms from worm to human. The CHR-NF-Y module is conserved in promoters of G_2/M regulated genes in all analyzed vertebrates. Our results reveal novel modules that determine specific cell cycle phasing, and identify their respective putative target genes with remarkably high specificity.

INTRODUCTION

The eukaryotic cell cycle is driven by a periodic, tightly controlled network of accumulation and destruction of key regulators and effectors. Precise coordination of cell cycle processes of DNA replication and chromosome segregation is required to ensure that daughter cells receive the requisite complement of genetic material. The fidelity of the cell cycle engine operation is tightly supervised by an intricate checkpoints mechanism acting during different phases of the division cycle. The mobilization of this engine is regulated at three major layers: transcriptional regulation of gene expression, post-translational modulation of protein activity, and modulation of protein stability.[1-3]

In this study, we focus on the transcriptional program associated with cell cycle progression. Prominent among the regulators of this program are the members of the E2F family of transcription factors (TFs). E2F1-3 are positive regulators of cell cycle progression while E2F4-6 play an inhibitory role.[4] Traditionally, the regulatory function of E2F was linked to the G_1 and S phases, but recent studies pointed to the involvement of this family in other cell cycle phases as well.[5-7] Other TFs and regulatory elements were shown to play an important role in driving the cell cycle transcriptional program. The CCAAT binding TF NF-Y was linked to the regulation of G_2/M progression by several studies: NF-Y controls the expression of several key regulators of this phase, including *CDC2*,[8] *CCNB1*[9] and *CCNB2*.[9,10] Furthermore, p53-mediated activation of the G_2/M checkpoint is executed through its inhibition of NF-Y induction of these target genes.[11] CDE (cell cycle dependent element) and CHR (cell cycle homology region) cis-regulatory elements were found in promoters of several cell cycle genes, including *CDC25C*,[12] *CDC2*,[13] *CCNB1*,[14] *CCNB2*,[15] *AURKB* (encoding aurora kinase B)[16] and *PLK*,[17] suggesting that these elements too play a role in controlling G_2/M progression. The B-Myb TF is an E2F-regulated gene induced at G_1/S, whose activity is enhanced during S phase through phosphorylation by cyclin A/Cdk2.[18,19] The transcriptional activity of B-Myb is required for cell cycle progression and it was recently suggested to play a role, together with E2F, in linking the G_1/S and G_2/M transcriptional programs.[14] Another TF, FOXM1, was recently shown to be required for execution of mitosis.[20,21]

While conventional biological studies focus on specific isolated components within a network of interest, the availability of essentially complete genome sequences in many organisms, and the maturation of novel functional genomics technologies, enable systems-level analysis of cellular networks. In a previous study, we applied computational promoter analysis to publicly available cell cycle related functional genomics datasets, delineating on a genomic scale regulatory mechanisms that control the human cell cycle transcriptional program.[22] We identified a significant statistical over-representation of several TF binding site (BS) signatures on promoters of cell cycle regulated genes. Among the most significant observation was the enrichment of E2F and NF-Y signatures in G_1/S and G_2/M promoters, respectively. Here, we employ comparative genomics to further elucidate the cell cycle network regulated by these regulators, and to pinpoint, with high accuracy, the target genes that they control.

A major challenge in computational promoter analysis is the typically very short (8–14 bp) and highly flexible nature of cis-regulatory elements recognized and bound by TFs: Most positions within the binding site motif are not strictly limited to a particular nucleotide, so genome-wide computational scans for putative TF binding sites (TFBSs) inevitably yield many false positive hits[23,24] (we use the term *hit* to refer to computationally-identified putative binding sites). The availability of sequences of many genomes in addition to the human genome greatly boosts the specificity of in silico identification of regulatory elements embedded in the genome.[25,26] Because higher selective pressure imposed on functional elements makes them more conserved than their surrounding non-functional DNA, scanning for evolutionarily conserved elements, an approach called phylogenetic footprinting, markedly reduces false-positive hit rates.[27,28]

We searched for conserved transcriptional regulators of cell cycle progression by integrating several sources of information: promoter sequences from twelve organisms, ranging from worm to human; orthology relationships among genes of these organisms; models of BSs of known TFs; and genome-wide gene expression profiles. We find that E2F signals are conserved in G_1/S regulated genes in all organisms from worm to human, and show that a canonical E2F signal is associated with gene expression that is highly specific to the G_1/S phase. In addition, we define a novel cis-regulatory module comprising CHR-NF-Y cis-elements, demonstrate that it dictates an expression pattern that is tightly restricted to the G_2 and G_2/M phases, and identify with very high specificity the genes that it putatively regulates. We show that the CHR-NF-Y module is conserved in cell cycle regulated promoters in vertebrates. We also observe that B-Myb signature appears in many of the CHR-NF-Y target promoters, suggesting that B-Myb cooperates with CHR-NF-Y, together constituting a triplet with specific functional roles in the regulation of G_2 and M phases.

MATERIALS AND METHODS

Extraction of promoter sequences. Putative promoter sequences were extracted based on gene transcription start site (TSS) annotation from the genome sequences of twelve organisms: two worms (*C. elegans* and *C. briggsae*), two insects (the fruit fly, *Drosophila melanogaster* and a mosquito, *Anopheles gambiae*), three fish (the zebrafish, *Danio rerio*; Fugu, *Fugu rubripes*; and Tetraodon, *Tetraodon nigroviridis*), chicken (Gallus, *gallus*), rat (*Rattus norvegicus*), mouse (*Mus musculus*), dog (*Canis familiaris*) and human. Promoters were extracted using a Perl script based on the application programming interface provided by the Ensembl project.[29] All sequences were extracted from version 27 of Ensembl genome release (Dec 2004), except for the *C. briggsae* promoters that were extracted from the v22 release (this worm was not included in later Ensembl releases). Ortholog genes between these organisms were determined using the EnsMart utility.[30] Only one-to-one mapped genes were taken into account when constructing the orthology maps. Specifically, the human-mouse orthology set contained a total of 16,299 genes.

Models for transcription factor binding sites. TFBSs are commonly modeled by position weight matrices (PWMs). PWMs for known human TFBSs were obtained from the TRANSFAC database (release 8.2, June 2004).[31] Typically, promoter sequences of a set of coregulated genes are scanned using a given PWM, and each subsequence is assigned a score that indicates how similar it is to the PWM; subsequences whose score is above some threshold are counted as hits, i.e., putative BSs. A judicious choice of the threshold value is essential in order to find a good balance between the rates of false positives and false negatives. Hits for the TATA-box cis-element were detected using the PRIMA software that we developed in a previous study, which sets the threshold by scanning randomly generated sequences with similar statistical characteristics to those of the genomic promoters.[22] PRIMA is available for download as part of the EXPANDER gene-expression analysis and visualization software (http://www.cs.tau.ac.il/~rshamir/expander/).[32] The TRANSFAC matrix M00252 was used by PRIMA as the TATA-box model. For ease of implementation and in order to ensure efficient performance, the other TFs in this study were modeled using regular expressions, which were composed and fine-tuned manually, based on TRANSFAC PWMs and a small selected list of known BSs. The following TFBS models were used (the models are written using the IUPAC nucleotide base code, e.g., Y stands for [CT]; "|" denotes "or"):

- E2F:TTT{E2F-core}NNN|(TTN|TNT|NTT){E2F-core}(ANN|NAN|NNA), where E2F-core is (SS|AG)CGSS|SSCG(SS|CT); Canonical E2F hits are those matching TTT{E2F-core}AAN (based on TRANSFAC matrix M00516 and BSs in *E2F1*, *CDC2*, *ORC1*)[33]
- NF-Y: ((RN|NR)CCAATSR)|(RRCCAAT(SN|NR)) (based on M00185 and BSs in *CCNB1*, *CDC2*,[14] *CCNB2*)[10]
- CHR: BNNRTTTRAAH (based on seven CHR BSs summarized in Kimura et al[16] and in *CCNB1*, *CDC2*)[14]
- B-Myb: (MNR|NNY)AACB(NYY|GHB) (based on M00004 and BSs in *CCNB1*, *CDC2*)[14]

Scanning promoters for TF hits. Promoter sequences were scanned by searching for matches of each regular expression in both strands of a predefined interval around the TSS. The intervals used were (positions are relative to the TSS, negative positions are upstream to the TSS): E2F: from -300 to +100; NF-Y: from -400 to 0; CHR: from -400 to +50; B-Myb: from -300 to +200. Each match of a regular expression was considered a hit of the corresponding TF. A hit of a module consisting of a pair of TFs was declared if the scanned promoter interval contained a match for both regular expressions, and the distance between the two matches was at most 200 bp. Hits of the triplet CHR-NF-Y-B-Myb were found by intersecting the promoters containing the pair CHR-NF-Y with those containing CHR-B-Myb. When scanning for evolutionarily conserved hits, additional constraints were applied, as explained below. Hits of the TATA-box element for Supplementary Fig. 2 were located using PRIMA, as described in Elkon et al.[22]

Phylogenetic footprinting constraints. Given the human promoters and their orthologs in one or more other species (typically mouse), each set of orthologous promoters was scanned for conserved hits. In the case of a single TF, a match was considered a hit if the following conservation criteria were fulfilled: (1) Each of the orthologous promoters contained a match for the regular expression in the corresponding interval; (2) All matches were on the same strand; (3) The locations of the matches in each of the non-human promoters differed by at most 100 bp from the location of the match in the human promoter; (4) The Hamming distance (i.e., number of different nucleotides) between each of the non-human matches and the human match was at most H, where H was set on a per-TF basis, as follows: $H = 4$ for E2F, $H = 3$ for NF-Y and CHR and $H = 2$ for B-Myb. For a module of two TFs, two additional constraints were applied: (5) The order of the TFs was

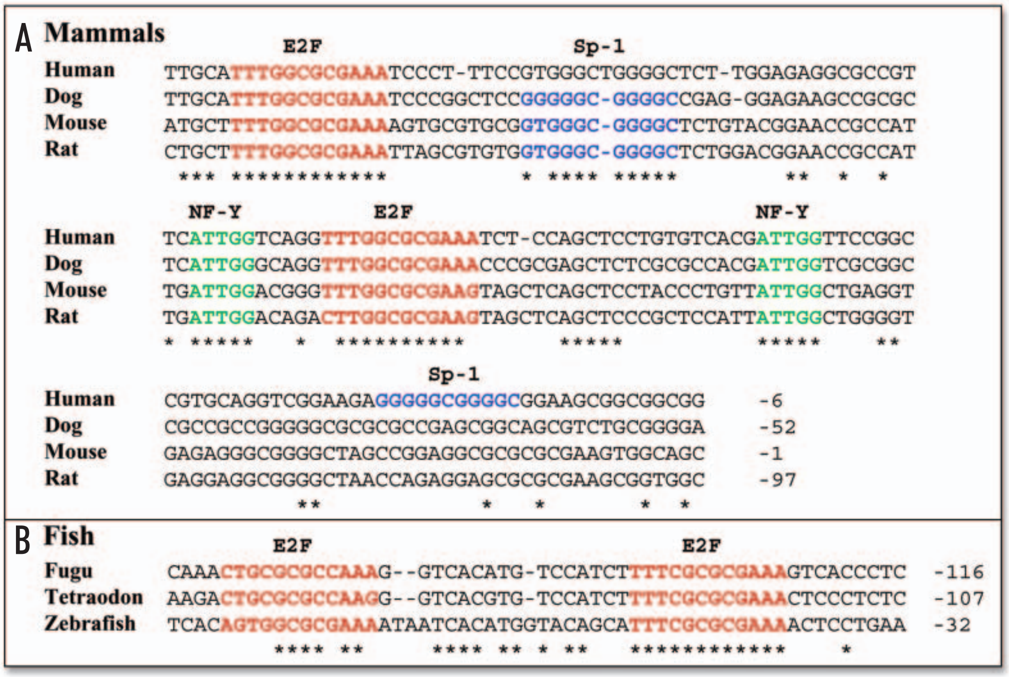

A
Mammals
E2F
Sp-1
Human TTGCATTTGGCGCGAAATCCCT-TTCCGTGGGCTGGGGCTCT-TGGAGAGGCGCCGT
Dog TTGCATTTGGCGCGAAATCCCGGCTCCGGGGGC-GGGGCCGAG-GGAGAAGCCGCGC
Mouse ATGCTTTTGGCGCGAAAAGTGCGTGCGGTGGGC-GGGGCTCTGTACGGAACCGCCAT
Rat CTGCTTTTGGCGCGAAATTAGCGTGTGGTGGGC-GGGGCTCTGGACGGAACCGCCAT
*** ************* * **** ***** ** * *
NF-Y
E2F
NF-Y
Human TCATTGGTCAGGTTTGGCGCGAAATCT-CCAGCTCCTGTGTCACGATTGGTTCCGGC
Dog TCATTGGGCAGGTTTGGCGCGAAACCCGCGAGCTCTCGCGCCACGATTGGTCGCGGC
Mouse TGATTGGACGGGTTTGGCGCGAAGTAGCTCAGCTCCTACCCTGTTATTGGCTGAGGT
Rat TGATTGGACAGACTTGGCGCGAAGTAGCTCAGCTCCCGCTCCATTATTGGCTGGGGT
* ***** * *********** ***** ***** **
Sp-1
Human CGTGCAGGTCGGAAGAGGGGGCGGGGCGGAAGCGGCGGCGG -6
Dog CGCCGCCGGGGGCGCGCGCCGAGCGGCAGCGTCTGCGGGGA -52
Mouse GAGAGGGCGGGGCTAGCCGGAGGCGCGCGCGAAGTGGCAGC -1
Rat GAGGAGGCGGGGCTAACCAGAGGAGCGCGCGAAGCGGTGGC -97
** * * * *
B
Fish
E2F
E2F
Fugu CAAACTGCGCGCCAAAG--GTCACATG-TCCATCTTTTCGCGCGAAAGTCACCCTC -116
Tetraodon AAGACTGCGCGCCAAGG--GTCACGTG-TCCATCTTTTCGCGCGAAACTCCCTCTC -107
Zebrafish TCACAGTGGCGCGAAAATAATCACATGGTACAGCATTTCGCGCGAAAACTCCTGAA -32
**** ** **** ** * ** ************ *

Figure 1. The power of phylogenetic footprinting. Alignment of promoter sequences from multiple species of the G_1/S-induced gene *MCM6* demonstrates the strength of comparative genomics in boosting computational identification of cis-regulatory elements. E2F elements (red), which have been validated experimentally,[36] are perfectly conserved across mammals (A) and fish (B). The alignment also points to other putative functional sites corresponding to NF-Y (green) and Sp1 (blue). The sequences flanking the TFBSs show lower conservation. Interestingly, the Sp1 site in human may have shifted downstream. Numbers next to the sequences indicate their location relative to the TSS (negative means upstream).

identical in all organisms; (6) The distance between the matches of the two TFs in each non-human promoter differed by at most 30 bp from their distance in the human promoter. In promoters that contained several matches for the same TF, all matches were checked.

Enrichment score and factor. The standard hypergeometric score was used to determine whether a certain TF, or module of TFs, is over-represented in a given set of genes. Specifically, let *TS* be a given gene set of interest of size *T* (in this study, a cell cycle phase or the entire cell cycle set), and let *BS* denote a large background set of size *B* (in our case, all the genes that are not included in the cell cycle set). Let *t* and *b* denote the number of promoters in *TS* and *BS*, respectively, in which a hit was identified (either in the single- or multi-species case). Assuming that *TS* is randomly chosen out of *BS*, the probability, or *p*-value, of observing at least *t* hits in *TS* is:

$$p = \sum_{i=t}^{\min\{b+t,T\}} \frac{\binom{T}{i}\binom{B}{b+t-i}}{\binom{B+T}{b+t}}$$

The enrichment factor, denoted by *f*, of a given TF or module is the ratio between its frequency in a specified set of promoters and its frequency in the rest of the genome, i.e., $f = (t/T)/(b/B)$.

Cooccurrence of a pair of TFs. Given a pair of TFs, a cooccurrence score was computed in order to ascertain whether their hits tend to appear together in the same promoters significantly more often than expected by chance. Denote by *m* the number of analyzed genes, let f_a and f_b be the number of promoters that contain a hit for the each TF, and let f_{ab} be the number of promoters with a hit for both TFs. Using the hypergeometric score, the *p*-value for observing f_{ab} or more promoters containing hits for both TFs is:

$$p = \sum_{i=f_{ab}}^{\min\{f_a,f_b\}} \frac{\binom{f_a}{i}\binom{m-f_a}{f_b-i}}{\binom{m}{f_b}}$$

The set of E2F4-bound promoters. Cam et al.[7] used ChIP-on-chip to identify promoters bound by E2F4 in quiescent cells, which were arrested using three methods: mitogen depletion, contact inhibition, and p16^{INK4A} induction. They reported a very high overlap (roughly 80%) between the results obtained by all three methods. Their microarray contained approximately

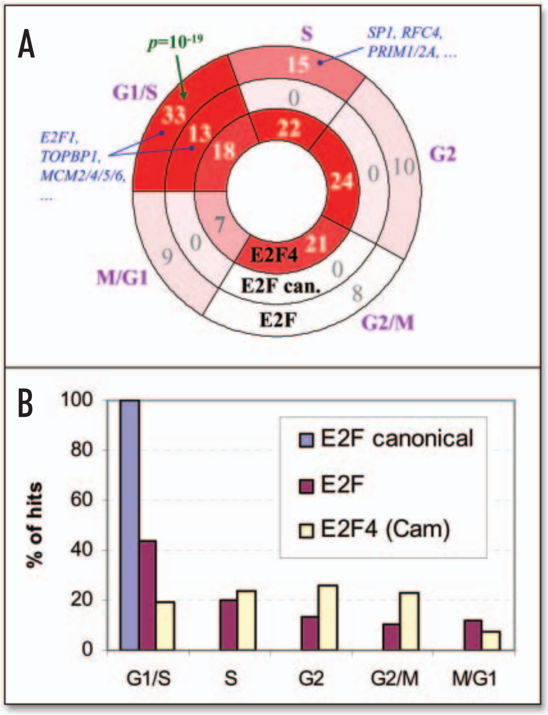


Figure 2. Distribution of E2F signatures and binding among cell cycle phases. (A) The number of promoters that contain a hit for the general E2F signature (outer circle) and canonical E2F signature (middle), and those that were shown to bind E2F4 in quiescent cells (inner) are indicated within each sector of the cycles. Each cycle is partitioned into five sectors corresponding to G_1/S, S, G_2, G_2/M and M/G_1 as defined by Whitfield et al.[41] Color intensities correspond to enrichment *p*-values of the corresponding set relative to all non-cell cycle genes: The general E2F signature is significantly enriched in G_1/S ($p = 10^{-19}$) and, to a lesser extent, in S; the canonical E2F signature is enriched in G_1/S. In contrast, E2F4 binding is highly enriched in each of the first four phases ($p = 10^{-12}$). Interesting representative targets are listed next to selected sectors (blue). (B) Distribution of the cell cycle targets across the five phases, for E2F general and canonical signatures, and for E2F4-bound promoters: The canonical E2F signature appears exclusively in G_1/S promoters, whereas binding of E2F4 is distributed uniformly across the first four phases.

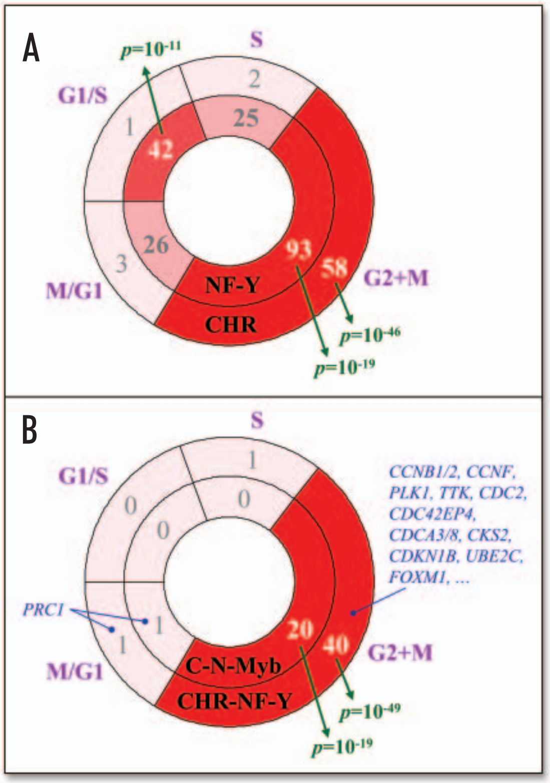


Figure 3. Distribution of NF-Y and CHR signatures among cell cycle phases. As in Figure 2, but here unifying the G_2 and G_2/M sectors, the number of promoters that contain a hit for NF-Y (A, inner circle) and CHR (A, outer) are indicated in each sector; and for the CHR-NF-Y pair (B, outer circle) and the CHR-NF-Y-B-Myb triplet (B, inner). NF-Y is enriched especially in G_1/S and G_2 + M, whereas CHR is highly specific to G_2 + M. The targets of the CHR-NF-Y and CHR-NF-Y-B-Myb modules are almost exclusive to G_2 + M.

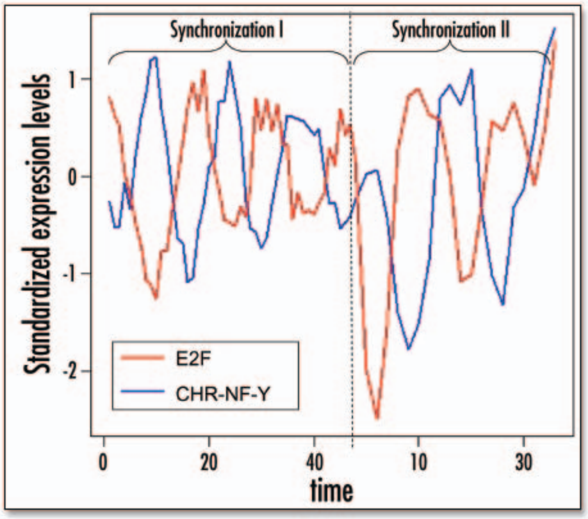


Figure 4. The canonical E2F signature and the CHR-NF-Y module dictate distinct and specific cell cycle phasing. Mean expression patterns over cell cycle progression (Whitfield et al. dataset) of genes containing the canonical E2F hits (13 cell cycle genes) and the CHR-NF-Y module (42 cell cycle genes) sharply peak at G_1/S and G_2/M phases, respectively. Expression levels of each gene were standardized to mean 0 and SD 1 before averaging over gene sets (in order to focus on the pattern rather than on the magnitude of expression). Y-axis represents standardized expression levels. Two synchronization methods were used by Whitfield et al.: Cells were arrested either in S phase using double thymidine block (synchronization I), or in M phase with a thymidine-nocodazole block (synchronization II).

13,000 sequences corresponding to promoter regions from -700 to +200 relative to the TSS. The set of E2F4-bound promoters used in this study consists of 271 promoters that were bound by E2F4 in at least one of the three methods (using a binding threshold of $p < 0.001$), and that are included in our human promoters set.

RESULTS

In this study, applying wide-scale computational promoter analysis, we sought to further elucidate transcriptional mechanisms that drive cell cycle progression. Our main objectives were to identify major cis-regulatory signals that dictate phase-specific expression, and to pinpoint, with high specificity, target genes that are under the control of these promoter elements. Several approaches have been proposed in an effort to increase the specificity of computational search for TFBSs. Phylogenetic footprinting[34,35] builds on the fact that TFBSs play an important biological role and have therefore evolved at a slower rate than non-functional intergenic sequences. Consequently, hits that are conserved across orthologous promoters in related species are more likely to be active BSs. Figure 1 illustrates the power of phylogenetic footprinting. The figure shows aligned promoter sequences of the gene *MCM6*, which encodes a subunit of the replication licensing complex, whose expression peaks at G_1/S,[36] in several mammals and fish. Evidently the promoters of *MCM6* are quite variable—most positions are not perfectly conserved across all species of each group. Remarkably, however, most of the conserved positions reside in contiguous blocks of 5-12bp in length, most of which match signatures of known TFs, namely E2F, NF-Y and Sp1. The role of all three TFs in cell cycle regulation is well established.[4,9,37,38] In the promoters presented in Figure 1, biologically active BSs emerge as islands of conservation, easily distinguishable from the surrounding sequences. Unfortunately, in most cases the identification of TFBSs is more difficult, either because of differences between the orthologous BSs, or because the BSs lie within long stretches of highly conserved promoter regions.

Transcriptional regulation in eukaryotes is to a large extent combinatorial, that is, the spatio-temporal conditions under which a gene is expressed are encoded by the specific combination of cis-regulatory elements embedded in its promoter region (and in the more distant regulatory regions, the enhancers and silencers). Therefore, a second common approach for reducing the rate of false positives in a TFBS scan is to search for a *module* of TFs, that is, a group of TFs whose joint binding activity has a specific transcriptional effect.[22,39,40] Identifying BSs of several TFs that tend to co-occur in the same promoters, possibly in a fixed order or at conserved distances, can eliminate many of the false hits that turn up when searching for each individual TF separately. Again, Figure 1 illustrates this idea: The order of the various BSs and the distances between them are highly conserved within each group of organisms; the only exception is the Sp1 BS, which seems to have drifted downstream in the human promoter.

Based on the aforementioned ideas, we sought to identify evolutionarily conserved transcriptional modules that control cell cycle progression. We first focused on E2F and performed a genome-wide scan for E2F signatures that are conserved between orthologous human-mouse promoters (see Materials and Methods). We found a conserved hit for E2F in 595 promoters out of the 16,299 orthologous promoter pairs included in our analysis. Next, we examined whether these hits are biased for cell cycle regulated promoters, using the cell cycle gene expression dataset published by Whitfield et al.[41] That study employed microarrays to profile gene expression throughout progression of the cell cycle in human Hela cells, and reported 872 cell cycle oscillating genes with periodic expression profiles. Our set of orthologous human-mouse promoters contains promoter sequences for 697 out of these 872 genes. We found that the overlap between the sets of promoters with conserved hits for E2F and the set of cell cycle oscillating genes (hereafter referred to as the 'cell cycle set') contains 75 genes, a statistically highly significant enrichment ($p = 2 \times 10^{-17}$). The list of cell cycle

Table 1 **CHR-NF-Y putative target genes whose expression peaks in the G_2 or G_2/M phases of the cell cycle[#]**

Symbol	Ensembl ID	Description
ARHGAP19	ENSG00000187122	ARHGAP19 (Rho GTPase activating protein 19) is involved in the regulation of members of the Rho GTPase family, that, among other roles, regulate chromosome alignment and cytokinesis.
ATF7IP	ENSG00000171681	Activating transcription factor 7 interacting protein
CCNB1[*+]	ENSG00000134057	Cyclin B1 complexes with CDC2 (Cdk1) to form the maturation-promoting factor (MPF), a master regulator of G_2/M phase.
CCNB2[*+]	ENSG00000157456	Cyclin B2 complexes with CDC2 (Cdk1) to form the M-phase-promoting factor (MPF), a master regulator of G_2/M phase.
CCNF	ENSG00000162063	Cyclin F regulates the nuclear localization of cyclin B1 through a cyclin-cyclin interaction.[42]
CDC2[*+]	ENSG00000170312	CDC2 is a catalytic subunit of the M-phase promoting factor (MPF), which is essential for G_1/S and G_2/M phase transitions of eukaryotic cell cycle.
CDC42EP4 (BORG4)	ENSG00000179604	This protein is a member of the CDC42-binding protein family. Members of this family interact with Rho family GTPases and regulate the organization of the actin cytoskeleton.
CDCA3 (Tome-1)[*+]	ENSG00000111665	Tome-1 (trigger of mitotic entry 1) mediates the destruction of the mitosis-inhibitory kinase, Wee1, via the E3 ligase, SCF.
CDCA8[+]	ENSG00000134690	A component of the mitotic spindle.
CDKN1B (p27)	ENSG00000111276	CDKN1b binds to and prevents the activation of cyclin E-CDK2 or cyclin D-CDK4 complexes, and thus controls the cell cycle progression at G_1.
CENPF	ENSG00000117724	CENPF associates with the centromere-kinetochore complex and may play a role in chromosome segregation during mitosis.
CKS2[+]	ENSG00000123975	CKS2 is required for the first metaphase/anaphase transition of mammalian meiosis.[43]
DEPDC1	ENSG00000024526	DEP domain containing 1.
DEPDC1B	ENSG00000035499	DEP domain containing 1B.
ECT2[+]	ENSG00000114346	ECT2 is related to Rho-specific exchange factors and regulates the activation of CDC42 in mitosis.
FOXM1[+]	ENSG00000111206	FoxM1 is a transcription factor that is required for execution of the mitotic programme and chromosome stability.[21]
GTSE1 (G-2 and S-phase expressed 1)	ENSG00000075218	GTSE1 is only expressed in the S and G_2 phases of the cell cycle, where it colocalizes with cytoplasmic tubulin and micro tubules. In response to DNA damage, the encoded protein accumulates in the nucleus and binds the tumor suppressor protein p53, shuttling it out of the nucleus and repressing its ability to induce apoptosis
H2AFX[+]	ENSG00000188486	H2A histone family, member X
HMGB2[+]	ENSG00000164104	This gene encodes a member of the non-histone chromosomal high mobility group protein family, which are chromatin-associated and ubiquitously distributed in the nucleus of higher eukaryotic cells. HMGB2 was demonstrated to associate with mitotic chromosomes.[52]
HMGB3[+]	ENSG00000029993	This gene encodes a member of the non-histone chromosomal high mobility group protein family, which are chromatin-associated and ubiquitously distributed in the nucleus of higher eukaryotic cells.
HMMR (RHAMM)	ENSG00000072571	The receptor for hyaluronan mediated motility has been reported to mediate migration, transformation, and metastatic spread of murine fibroblasts. Its over-expression results in structural centrosomal abnormalities and and mitotic defects.[53]
KPNA2	ENSG00000182481	Karyopherin-α2 protein interacts with Chk2 and contributes to its nuclear import.[54]
LRRC17	ENSG00000128606	Leucine rich repeat containing 17.
MKI67	ENSG00000148773	The cell proliferation-associated antigen of antibody Ki-67 is widely used in routine pathology as a "proliferation marker" to measure the growth fraction of cells in human tumors.[55]
NUSAP1	ENSG00000137804	NuSAP (nucleolar and spindle associated protein 1) is primarily nucleolar in interphase, and localizes prominently to central spindle microtubules during mitosis. Depletion of NuSAP by RNA interference resulted in aberrant mitotic spindles, defective chromosome segregation, and cytokinesis.[56]
PLK1[*+]	ENSG00000166851	Polo-like kinase 1 (Plk1) is a key regulator of centrosome maturation, mitotic entry, sister chromatid cohesion, the anaphase-promoting complex/cyclosome (APC/C), and cytokinesis.
SFPQ	ENSG00000116560	splicing factor proline/glutamine rich.
SGOL2 (shugoshin-like 2)	ENSG00000163535	Shugoshin prevents dissociation of cohesin from centromeres during mitosis in vertebrate cells.[57]
STK17B	ENSG00000081320	Serine/threonine kinase 17b (apoptosis-inducing).
TACC3	ENSG00000013810	TACC3 is a centrosomal/mitotic spindle-associated protein that is highly expressed in a cell cycle dependent manner in hematopoietic lineage cells.[44]
TMPO (LAP2)	ENSG00000120802	Lamina-associated polypeptide (LAP) 2 is suggested to play a role in targeting mitotic vesicles to chromosomes and reorganizing the nuclear structure at the end of mitosis.[58]

Table 1 **CHR-NF-Y putative target genes whose expression peaks in the G_2 or G_2/M phases of the cell cycle# (continued)**

Symbol	Ensembl ID	Description
TOP2A	ENSG00000131747	This nuclear enzyme is involved in processes such as chromosome condensation, chromatid separation, and the relief of torsional stress that occurs during DNA transcription and replication.
TTK+	ENSG00000112742	TTK was demonstrated to be dynamically distributed from the kinetochore to the centrosome, as cell enters into anaphase, and to phosphorylate the centrosomal protein TACC2 in mitosis.[59]
UACA+	ENSG00000137831	Uveal autoantigen with coiled-coil domains and ankyrin repeats.
UBE2C+	ENSG00000175063	This gene encodes a member of the E2 ubiquitin-conjugating enzyme family that is required for the destruction of mitotic cyclins and for cell cycle progression.[48]

*Genes whose promoter is already reported to be regulated by a CHR element. +Genes whose promoter was found to contain a strong conserved hit for B-Myb (in addition to CHR-NFY hit). #Four additional genes that do not have an official HUGO symbol are not included here (but included in Supplementary Table C).

regulated promoters on which a conserved E2F hit was identified is provided in Supplementary Table A. Whitfield et al. partitioned the cell cycle oscillating genes into five clusters—G_1/S, S, G_2, G_2/M and M/G_1—according to the phase in which their expression peaked. Hereafter we refer to the set of genes assigned either to the G_1/S or to the S clusters as the G_1 + S set, and to the set of the genes assigned either to the G_2 or to the G_2/M clusters as the G_2 + M set. In agreement with current biological knowledge and with results we previously reported,[22] we also observed here, but this time for human-mouse evolutionarily conserved hits, a strong bias of E2F signature for promoters of cell cycle regulated genes that peak at G_1/S phase ($p = 5 \times 10^{-19}$ relative to all the genes that are not in the cell cycle set) and, to a lesser extent, at S phase ($p = 7 \times 10^{-6}$) (Fig. 2A).

Recent functional genomics studies showed that the regulatory role of the E2F family on cell cycle progression extends also to G_2 and M phases.[5,6] To examine this point more closely, we analyzed the dataset published by Cam et al.[7] that used the combination of chromatin immunoprecipitation and promoter microarrays, also known as 'ChIP-on-chip', to identify promoters that are bound by the inhibitory E2F4 in quiescent cells. We checked the overlap between the set of 271 genes, whose promoters are bound by E2F4 (see Materials and Methods), and the cell cycle set. We found a highly significant overlap of 92 common genes ($p = 10^{-67}$). Surprisingly, these genes were not biased to G_1/S phase but were distributed almost uniformly across the first four phases—G_1/S, S, G_2 and G_2/M (Fig. 2A and B). This apparent discrepancy between the near-uniform distribution of E2F4 targets and the strong G_1/S bias of the E2F signature can be explained in several ways. It is possible that the inhibitory E2F4 is recruited to many G_2 + M promoters by physical association with other DNA binding TFs rather than by its direct binding to the DNA. Another option is that E2F binding elements on G_2 + M phase promoters are variants, perhaps with lower binding affinity, of the canonical E2F signature (which was originally defined using mainly G_1 and S phase E2F target promoters). To check this hypothesis, we performed a genome-wide scan for promoters that contain human-mouse conserved E2F signatures, with a strict requirement of adherence to the canonical E2F BS consensus (see Materials and Methods). Only 22 promoters met this stringent genome-wide scan (Supplementary Table B); 13 of them are contained in the cell cycle set. Remarkably, all 13 genes peak at G_1/S phase ($p = 4 \times 10^{-22}$) (Fig. 2).

Our previous computational cell cycle analysis, as well as other experimental studies, indicated that NF-Y plays a major role in regulating cell cycle progression in general, and is especially linked to G_2 and G_2/M phases.[9,11,22] Using our current approach, we identified 1,754 promoters with human-mouse conserved NF-Y hits, of which 186 are in the cell cycle set ($p = 2 \times 10^{-33}$), reflecting that NF-Y regulates a variety of biological processes, and its key function in cell cycle. In agreement with our previous results, NF-Y hits are enriched in all five phases ($p \leq 10^{-4}$ in each phase compared to the non-cell cycle genes), and most prominently in G_1/S (42 genes, $p = 6 \times 10^{-11}$) and in G_2 + M (93 genes, $p = 10^{-19}$) (Fig. 3A).

Zhu et al.[14] recently validated functional NF-Y and CHR elements in the promoters of both *CDC2* and *CCNB1*, the master regulators of G_2 and G_2/M phases. Based on this observation, we tested whether this pair constitutes a recurrent cis-regulatory module. First, scanning for conserved CHR hits, we detected a striking bias for the G_2 + M set (Fig. 3A). We next searched for targets of the pair CHR-NF-Y with some distance constraints (see Materials and Methods). In the entire genome (16K genes), only 71 promoters met our criteria for human-mouse conserved hits of this module (Supplementary Table C), and 42 of them are contained in the cell cycle set ($p = 2 \times 10^{-39}$). Remarkably, 40 of these genes are assigned to G_2 + M ($p = 9 \times 10^{-50}$, Fig. 3B). Moreover, this bias is not explained merely by the hit distributions of each individual TF within the G_2 + M genes—the co-occurrence of the CHR and NF-Y elements is way above the expected rate given the prevalence of each TF separately ($p = 4 \times 10^{-13}$). Thus, CHR-NF-Y emerges as a major regulatory module of the G_2 + M transcriptional program. This module dictates a highly phase-specific expression pattern, which is strongly anti-correlated with the expression imposed by the canonical E2F signature (Fig. 4).

Table 2 **Evolutionary conservation of cell cycle TFs**

		Organism											
		Tetrapods					Fish			Insects		Worms	
TFs	Cell cycle Phases	Human	Mouse	Rat	Dog	Chicken	Zebrafish	Fugu	Tetraodon	Fly	Mosquito	*C. elegans*	*C. briggsae*
E2F	G_1+S	6×10^{-20} (268)	3×10^{-23} (248)	4×10^{-12} (237)	3×10^{-4} (247)	2×10^{-9} (206)	4×10^{-13} (181)	7×10^{-7} (203)	2×10^{-4} (208)	3×10^{-4} (110)	1×10^{-7} (112)	8×10^{-4} (102)	2×10^{-4} (88)
CHR-NF-Y	G_2+M	4×10^{-38} (350)	2×10^{-31} (334)	9×10^{-11} (320)	2×10^{-9} (322)	2×10^{-15} (269)	3×10^{-8} (252)	3×10^{-3} (270)	5×10^{-4} (271)	N.E. (132)	N.E. (120)	N.E. (106)	N.E. (99)

The table shows enrichment *p*-values across 12 organisms of the E2F signature and the CHR-NF-Y module in promoters of genes whose human orthologs have an expression profile that peaks at $G_1 + S$ and $G_2 + M$ phases, respectively. "N.E." denotes "not enriched" ($p > 0.1$). E2F is enriched in $G_1 + S$ across all tested species, whereas CHR-NF-Y is enriched in G_2+M only in vertebrates. The total number of genes in each set is written in parentheses (e.g., our data contains 203 Fugu genes, whose human orthologs are expressed in $G_1 + S$).

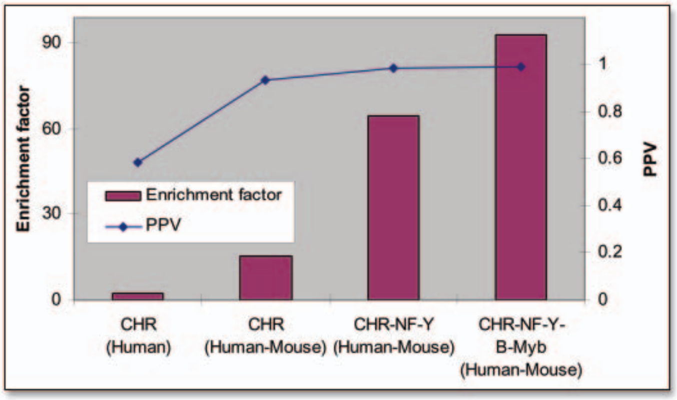


Figure 5. Improving TFBS detection by utilizing comparative genomics and searching for TF modules. The graph shows the dramatic improvement in enrichment factor (bars) and PPV values (curve) when searching for CHR-related hits in G_2 + M genes. Searching for targets of CHR in human yields an enrichment factor f = 2.4 and PPV = 0.58, indicating a low rate of true hits. Utilizing human-mouse conservation criteria improves the performance to f = 15.5, PPV = 0.93. Scanning for conserved modules results in an additional increase in the specificity, reaching a remarkable enrichment factor of 93 and PPV = 0.99 for CHR-NF-Y-B-Myb.

The forty G_2 + M putative CHR-NF-Y targets include several genes that have already been shown to be controlled by CHR and NF-Y elements, but the majority of the targets are reported here for the first time (Table 1). The utilization of phylogenetic footprinting and the fact that these genes were experimentally demonstrated to peak at G_2 + M phases greatly boost the confidence that the hits reported here are biologically significant. Known CHR targets among the G_2 + M hits include *CDC2*, *CCNB1* and *CCNB2*,[13-15] which constitute the Cyclin-CDK complex of the G_2/M phase; and *PLK1*,[17] which plays a major role in controlling centrosome maturation, mitotic entry, sister chromatid cohesion, the anaphase-promoting complex/cyclosome (APC/C), and cytokinesis. The proteins encoded by the novel targets putatively regulated by CHR-NF-Y participate in all major activities that are carried out during G_2 and M phases. Prominent among them are CCNF, which regulates the nuclear localization of cyclin B1 through a cyclin-cyclin interaction;[42] CKS2, which regulates CDKs activity during mitosis and meiosis;[43] CENPF, which associates with the centromere-kinetochore; the mitotic spindle-associated protein TACC3 that functions in chromosome segregation;[44] and the CDCA8, NUSAP1 (nucleolar and spindle associated protein 1) and TTK regulators of the mitotic spindle. Correct alignment of sister chromatide during metaphase and their balanced segregation during anaphase are critical processes executed by intricate complexes of cohesins, the centrosome-kinetochore at centromeres, and the bipolar structure of the mitotic spindle composed of microtubules and associated motor proteins. Recently, the Cdc42 member of the Rho GTPases family and its effector mDia3 were shown to regulate chromosome alignment by stabilizing kinetochore-microtobule attachment.[45,46] The Rho member of this GTPase family is known to regulate cytokinesis by controlling the assembly and the contraction of the myosin-actin network that comprises the contractile ring that is attached to the plasma membrane.[47] Importantly, ECT2, a major regulator of these GTPases-mediated pathways, and two additional proteins (CDC42EP4/Borg4 and the Rho GTPase activating protein ARHGAP19) that are tightly involved in them, are among our putative CHR-NFY targets. We also observed that protein products of many of the known and putative CHR-NF-Y targets are targeted for degradation by the anaphase-promoting complex/cyclosome (APC/C). In this regard, it is noteworthy that *UBE2C*, which encodes for an E2 ubiquitin-conjugating enzyme required for the destruction of mitotic cyclins,[48] is among the CHR-NF-Y putative targets.

A few salient examples of evolutionarily conserved CHR-NF-Y hits are shown in Supplementary Figure 1. As in the case of Figure 1, the BSs appear as islands of conservation along the promoter sequences. We observed that many of the G_2 + M putative CHR-NF-Y targets also contain a conserved cis-element that resembles the signature of B-Myb, as demonstrated in the promoters of *PLK1* (Supplementary Fig. 1A) and *UBE2C* (Supplementary Fig. 1B). Functional B-Myb sites were also identified in promoters of *CDC2* and *CCNB1*.[14] This suggests that B-Myb cooperates with CHR-NF-Y, together constituting a triplet with a specific functional role in regulating G_2 and M phases. We located a total of 31 conserved hits of this triplet; 21 of them are in the cell cycle set ($p = 4 \times 10^{-22}$), and of these 20 are in G_2 + M (see Table 1). Of note, the single cell cycle target gene of the triplet that is not in G_2 + M, PRC1 (protein regulator of cytokinesis 1), is also closely involved in regulation of the mitotic spindle and cytokinesis.[49] Interestingly, in 12 of the 21 promoters, the order of the hits within the triplet is NF-Y, CHR and B-Myb (from 5' to 3' on the coding strand), suggesting a possible structural preference of this module.

Our analysis highlights two major conserved transcriptional regulators of cell cycle progression—E2F and CHR-NF-Y—with key roles in G_1 + S and G_2 + M, respectively. We sought to trace the conservation of these signals along metazoan evolution. To this aim, we first extracted genome-wide promoter sequences of 12 organisms, including worms, insects, fish, chicken, rodents, dog and human (see Materials and Methods). In order to ensure that the quality of the annotated TSS in all organisms suffices for the detection of cis-regulatory elements, we verified that the TATA-box signal peaks at the correct location, in the very proximity of the TSS, in each of the tested species (e.g., the peak value is 7.9 standard deviations in human, and 11.0 in mouse) (Supplementary Fig. 2). Strikingly, we found that the E2F signature is conserved in G_1 + S genes across all organisms, from worm to human: scanning each organism separately for hits of E2F, we found a strong enrichment in the orthologs of human G_1 + S genes across all species ($p \leq 8 \times 10^{-4}$) (Table 2). In contrast, the CHR-NF-Y module, as defined using our BSs models, apparently evolved in vertebrates, as it is enriched in G_2 + M in all vertebrates but in none of the other species.

DISCUSSION

Advances in functional genomics provide broad systems-level views of biological networks for the first time. In this study we conducted computational promoter analysis using genome sequences from multiple organisms and cell cycle gene expression profiles in order to comprehensively delineate the cell cycle transcriptional program. We demonstrate that E2F signals are conserved in G_1 + S regulated genes in all organisms from worm to human, and that a canonical E2F signal encodes for an expression at a very precise timing during cell cycle progression. In addition, we define a novel cis-regulatory module made up of CHR-NF-Y cis-elements, and demonstrate that it determines an expression pattern that is tightly restricted to the G_2 + M phase. Our analysis identifies with high specificity forty G_2 + M genes that are putatively regulated by this module, thereby substantially extending current knowledge on the role of the CHR element in cell cycle regulation. We show that the CHR-NF-Y module is conserved in cell cycle regulated promoters in vertebrates.

TFBS detection has been the subject of numerous studies, but remains a difficult challenge. Existing BSs models do not contain enough information to locate functional BSs accurately. Typically, when promoter sequences are scanned using thresholds that allow recovering a large percentage of the true sites, many false positive hits are also reported.[24] Another difficulty lies in the evaluation of the results: Since there are no large validated gene sets in which the entire list of active BSs of the studied TF have been completely mapped, the specificity and sensitivity values are hard to assess. In this study, we searched for TFs and regulatory modules that are not only over-represented in the set of cell cycle promoters, but are also

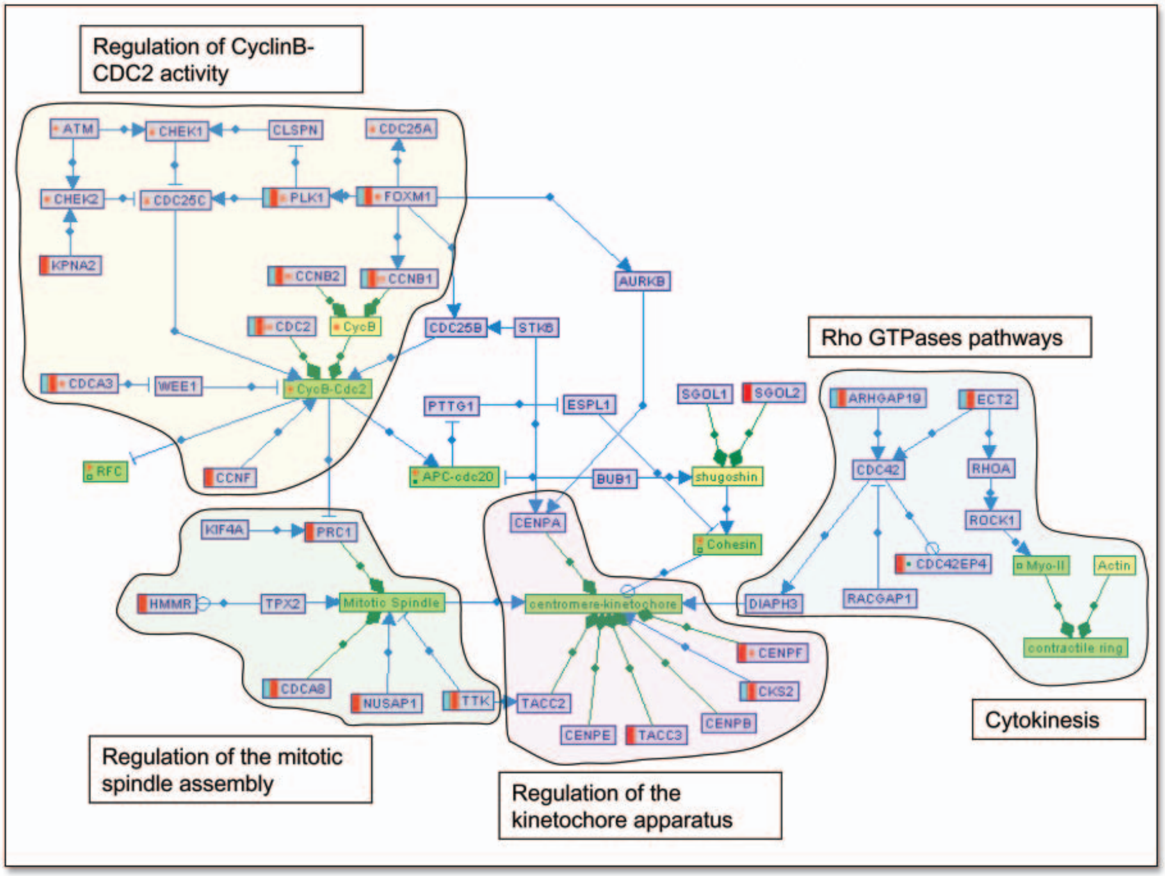


Figure 6. Putative roles for the CHR-NF-Y module in regulation of the G_2 and M phases. The interaction map contains nodes of three types: gray nodes represent single genes (denoted by the official HUGO symbol), yellow nodes represent gene families (e.g., Cyclin B), and green nodes represent protein complexes (e.g., CyclinB-CDC2). Blue edges denote regulation relations (→ for 'activation', ——| for 'inhibition') and green edges denote containment relations among nodes (e.g., CDC2 is contained in the CyclinB-CDC2 complex). Genes whose promoter contains a conserved CHR-NF-Y hit are marked by a red bar to the left of their node; an additional blue bar marks putative targets of the CHR-NF-Y-B-Myb triplet. CHR-NF-Y putative targets participate in all major activities that are carried out during G_2 and M phases, including modulation of CyclinB-CDC2 activity, control of sister chromatide alignment by the centrosome-kinetochore, control of chromosome segregation by the mitotic spindle apparatus, and regulation of the contractile ring assembly for the execution of cytokinesis. The figure was created using our SHARP software and knowledgebase for signaling pathways (http://www.cs.tau.ac.il/~sharp/). A red dot within a node indicates that the node has additional regulations in the SHARP database that are not displayed in the current map. Similarly, a green dot indicates that not all containment relations in which the node is involved are displayed.

highly biased to specific phases. Measuring the phase specificity of the TFs that we identified enables us to approximate their overall specificity, as false hits are expected to be distributed randomly among the genes in all phases. The TFs and modules we report are exceedingly phase-specific: All 13 promoters with a conserved canonical E2F site are G_1/S genes, which constitute a mere 19% of the entire cell cycle set; 95% (40 out of 42) of the promoters with conserved CHR-NF-Y hits are in the G_2 + M phases, which contain 48% of the cell cycle regulated genes.

Another approach to evaluating the accuracy of our results is to compute each TF's or module's enrichment factor, denoted by *f*—the ratio between its frequency in a given set of promoters and its frequency in the rest of the genome (see Methods). Using the latter frequency as an upper-bound estimate of the false-positives rate, the value *1 - 1/f* approximates the positive predictive value, or PPV, which is the fraction of true-positive hits out of all the identified hits in the promoter set (see Tompa et al.).[24] For example, the CHR-NF-Y-B-Myb module has 10 putative targets out of 15,602 genes that are not in the cell cycle set. Thus, we estimate that our scan reports up to one false-positive hit per 1,560 promoters. Searching for the same module within the set of 334 G_2 + M genes yields 20 targets. Using the 1:1560 false-positives rate, we expect the number of false targets in this set to be no more than 0.21 (=334/1560). In other words, at least 19 (or, more accurately, 19.79, which is 99% of 20) of the 20 identified targets should be true hits (i.e., PPV = 0.99).

Remarkably, the enrichment factor *f* increases as more sources of information are added into the scan algorithm (Fig. 5). For instance, searching for hits of CHR in the human genome yields *f* = 2.4 (PPV = 0.58) for G_2 + M phases; utilizing phylogenetic footprinting—requiring each hit to match human-mouse conservation constraints —increases the enrichment factor to *f* = 15.5; and searching for human-mouse conserved modules improves this even further: *f* = 64.4 for CHR-NF-Y, and *f* = 93 for CHR-NF-Y-B-Myb. The latter enrichment factor implies that PPV = 0.99, that is, 99% of the reported G_2 + M hits are expected to be true BSs, as explained above. These enrichment factors and PPV's exemplify the dramatic improvement in TFBS detection accuracy gained by applying comparative genomics techniques and by searching for modules of cooperative TFs.

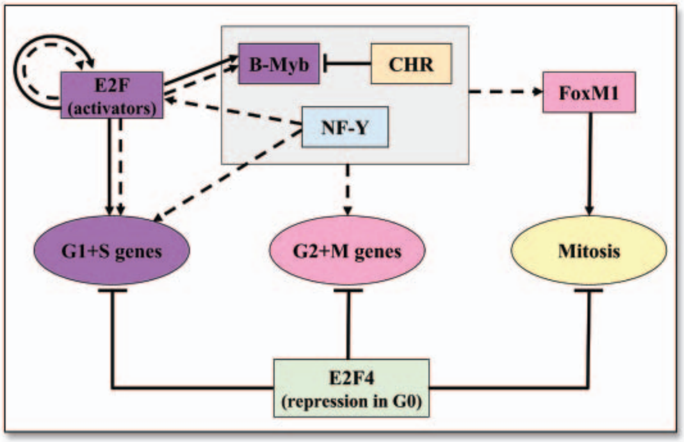


Figure 7. A model of the cell cycle transcriptional program. Integrating current knowledge and our computational results, a model for the propagation of the cell cycle transcriptional program emerges. Interactions supported by experimental data are presented by solid arrows, while those based on our computational analysis are marked by dashed ones. Genes expressed in G_1 or S phases are colored purple (dark gray), and G_2 + M genes are pink (light gray). Missing pieces in this puzzle can be expected to be discovered by combinations of biochemical experiments, functional genomics studies, and computational analyses.

In agreement with current biological knowledge, but without being biased by it, our computational analysis identified E2F as the major transcriptional regulator of the G_1 + S transcriptional network. However, recent studies extended the role of the E2F family in cell cycle regulation beyond the G_1 and S phases. Indeed, we show that the promoters reported to bind the inhibitory E2F4 in quiescent cells are not biased to any specific cell cycle phase. This apparent discrepancy may suggest that E2F regulation on G_2 and M promoters is mediated by a variant of the canonical E2F signature, possibly with lower binding affinity (or even, for some G_2 + M promoters, by non-direct DNA binding). The absolute assignment of the 13 canonical E2F targets to the G_1/S phase supports this hypothesis.

Our analysis points to the CHR-NF-Y cis-module as the major regulator of gene expression in G_2 + M phases. Several cell cycle regulated promoters were reported to be regulated by the CHR element, including *Cyclin A*,[8] CDC25C,[8] CDC2,[8] Cyclin B2,[10] Aurora B,[16] B-Myb[50] and PLK1.[17] However, the importance of the CHR-NF-Y module as a key regulator of G_2 and M phases is not widely appreciated, and is put in the spotlight by our results. We report, with high specificity, 42 mitotic genes that are putatively regulated by this module. Examination of these putative targets suggests that the CHR-NF-Y module regulates all known major activities that are carried out in G_2 and M phases, including modulation of CCNB-CDC2, and the assembly of the kinetochore-centrosome complexes, of the mitotic spindle and its associated motor proteins, and of cytokinesis effectors. A portion of the intricate network putatively modulated by the CHR-NF-Y module is depicted in Figure 6. Given its apparent pivotal role, it is intriguing that the protein that binds the CHR element is yet to be identified.[13] The list of putative CHR hits we provide could guide the empirical identification of this protein. Experimental analysis of CHR elements on several cell cycle-regulated promoters showed that these elements exert a repressive effect on the expression of their target genes. This suggests a model in which CHR and NF-Y play antagonistic roles, with the former acting as a repressor and the latter as an activator of G_2 + M promoters. This model requires experimental examination in which it will also be interesting to study whether the CHR and NF-Y elements are occupied simultaneously by their respective binding TFs or at different times during cell cycle progression.

We observed that many of the promoters that contain a hit for the CHR-NF-Y module also contain a conserved signature of B-Myb, suggesting a combinatorial role for the triplet CHR-NF-Y-B-Myb module. The promoter of B-Myb itself is regulated by E2F and is activated in late G_1/early S phase.[50] B-Myb is known to cooperate with E2F in the activation of *CDC2* and *CCNB1*.[14] In addition, a repressive CHR element was defined in the B-Myb promoter.[50] Furthermore, FOXM1, a TF recently demonstrated to be required for the execution of the mitotic program,[21] is among our twenty putative targets of the CHR-NF-Y-B-Myb triplet. Taken together, a picture of an intricate regulatory network maintained among the transcriptional regulators of cell cycle progression emerges (Fig. 7).

While preparing this manuscript, a computational paper analyzing cell cycle regulation was published by Zhu et al.[51] These authors too pointed out CHR-NF-Y (together with the CDE element) as a major transcriptional regulatory module of G_2 + M genes.

Our methodology and results demonstrate the power of computational analysis applied to functional genomics data in delineating novel aspects of the architecture of the transcriptional network that controls cell cycle progression. High false-positive rates are often a major limiting factor of computational binding site predictions, gravely hampering their experimental examination. Therefore, the significant improvement that we achieved in the specificity of the putative targets can potentially make their empirical validation much more focused and efficient.

References

1. Murray AW. Recycling the cell cycle: Cyclins revisited. Cell 2004; 116:221-34.
2. Castro A, Bernis C, Vigneron S, Labbe JC, Lorca T. The anaphase-promoting complex: A key factor in the regulation of cell cycle. Oncogene 2005; 24:314-25.
3. Fung TK, Poon RY. A roller coaster ride with the mitotic cyclins. Semin Cell Dev Biol 2005; 16:335-42.
4. Dimova DK, Dyson NJ. The E2F transcriptional network: Old acquaintances with new faces. Oncogene 2005; 24:2810-26.
5. Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, Dynlacht BD. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. Genes Dev 2002; 16:245-56.
6. Ishida S, Huang E, Zuzan H, Spang R, Leone G, West M, Nevins JR. Role for E2F in control of both DNA replication and mitotic functions as revealed from DNA microarray analysis. Mol Cell Biol 2001; 21:4684-99.
7. Cam H, Balciunaite E, Blais A, Spektor A, Scarpulla RC, Young R, Kluger Y, Dynlacht BD. A common set of gene regulatory networks links metabolism and growth inhibition. Mol Cell 2004; 16:399-411.
8. Zwicker J, Lucibello FC, Wolfraim LA, Gross C, Truss M, Engeland K, Muller R. Cell cycle regulation of the cyclin A, *cdc25C* and *cdc2* genes is based on a common mechanism of transcriptional repression. EMBO J 1995; 14:4514-22.
9. Manni I, Mazzaro G, Gurtner A, Mantovani R, Haugwitz U, Krause K, Engeland K, Sacchi A, Soddu S, Piaggio G. NF-Y mediates the transcriptional inhibition of the cyclin B1, cyclin B2, and cdc25C promoters upon induced G2 arrest. J Biol Chem 2001; 276:5570-6.
10. Bolognese F, Wasner M, Dohna CL, Gurtner A, Ronchi A, Muller H, Manni I, Mossner J, Piaggio G, Mantovani R, Engeland K. The cyclin B2 promoter depends on NF-Y, a trimer whose CCAAT-binding activity is cell-cycle regulated. Oncogene 1999; 18:1845-53.
11. Imbriano C, Gurtner A, Cocchiarella F, Di Agostino S, Basile V, Gostissa M, Dobbelstein M, Del Sal G, Piaggio G, Mantovani R. Direct p53 transcriptional repression: In vivo analysis of CCAAT-containing G_2/M promoters. Mol Cell Biol 2005; 25:3737-51.
12. Lucibello FC, Liu N, Zwicker J, Gross C, Muller R. The differential binding of E2F and CDF repressor complexes contributes to the timing of cell cycle-regulated transcription. Nucleic Acids Res 1997; 25:4921-5.
13. Liu N, Lucibello FC, Korner K, Wolfraim LA, Zwicker J, Muller R. CDF-1, a novel E2F-unrelated factor, interacts with cell cycle-regulated repressor elements in multiple promoters. Nucleic Acids Res 1997; 25:4915-20.
14. Zhu W, Giangrande PH, Nevins JR. E2Fs link the control of G_1/S and G_2/M transcription. EMBO J 2004; 23:4615-26.
15. Wasner M, Haugwitz U, Reinhard W, Tschop K, Spiesbach K, Lorenz J, Mossner J, Engeland K. Three CCAAT-boxes and a single cell cycle genes homology region (CHR) are the major regulating sites for transcription from the human cyclin B2 promoter. Gene 2003; 312:225-237.

16. Kimura M, Uchida C, Takano Y, Kitagawa M, Okano Y. Cell cycle-dependent regulation of the human aurora B promoter. Biochem Biophys Res Commun 2004; 316:930-6.
17. Uchiumi T, Longo DL, Ferris DK. Cell cycle regulation of the human polo-like kinase (PLK) promoter. J Biol Chem 1997; 272:9166-74.
18. Joaquin M, Watson RJ. Cell cycle regulation by the B-Myb transcription factor. Cell Mol Life Sci 2003; 60:2389-401.
19. Schubert S, Horstmann S, Bartusel T, Klempnauer KH. The cooperation of B-Myb with the coactivator p300 is orchestrated by cyclins A and D1. Oncogene 2004; 23:1392-404.
20. Costa RH. FoxM1 dances with mitosis. Nat Cell Biol 2005; 7:108-10.
21. Laoukili J, Kooistra MR, Bras A, Kauw J, Kerkhoven RM, Morrison A, Clevers H, Medema RH. FoxM1 is required for execution of the mitotic programme and chromosome stability. Nat Cell Biol 2005; 7:126-36.
22. Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. Genome Res 2003; 13:773-80.
23. Stormo GD. DNA binding sites: Representation and discovery. Bioinformatics 2000; 16:16-23.
24. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 2005; 23:137-44.
25. Duret L, Bucher P. Searching for regulatory elements in human noncoding sequences. Curr Opin Struct Biol 1997; 7:399-406.
26. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 2004; 306:636-40.
27. Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW. Identification of conserved regulatory elements by comparative genome analysis. J Biol 2003; 2:13.
28. Sandelin A, Wasserman WW, Lenhard B. ConSite: Web-based prediction of regulatory elements using cross-species comparison. Nucleic Acids Res 2004; 32(Web Server issue):W249-252.
29. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, Down T, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz HR, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark KC, Cameron G, Durbin R, Cox A, Hubbard T, Clamp M. An overview of Ensembl. Genome Res 2004; 14:925-8.
30. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E. EnsMart: A generic system for fast and flexible access to biological data. Genome Res 2004; 14:160-9.
31. Wingender E. TRANSFAC, TRANSPATH and CYTOMER as starting points for an ontology of regulatory networks. In Silico Biol 2004; 4:55-61.
32. Sharan R, Maron-Katz A, Shamir R. CLICK and EXPANDER: A system for clustering and visualizing gene expression data. Bioinformatics 2003; 19:1787-99.
33. Kel AE, Kel-Margoulis OV, Farnham PJ, Bartley SM, Wingender E, Zhang MQ. Computer-assisted identification of cell cycle-related genes: New targets for E2F transcription factors. J Mol Biol 2001; 309:99-120.
34. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J Mol Biol 1988; 203:439-55.
35. Blanchette M, Schwikowski B, Tompa M. Algorithms for phylogenetic footprinting. J Comput Biol 2002; 9:211-23.
36. Ohtani K, Iwanaga R, Nakamura M, Ikeda M, Yabuta N, Tsuruga H, Nojima H. Cell growth-regulated expression of mammalian *MCM5* and *MCM6* genes mediated by the transcription factor E2F. Oncogene 1999; 18:2299-309.
37. Rotheneder H, Geymayer S, Haidweger E. Transcription factors of the Sp1 family: Interaction with E2F and regulation of the murine thymidine kinase promoter. J Mol Biol 1999; 293:1005-15.
38. Chang YC, Illenye S, Heintz NH. Cooperation of E2F-p130 and Sp1-pRb complexes in repression of the Chinese hamster *dhfr* gene. Mol Cell Biol 2001; 21:1121-31.
39. Pilpel Y, Sudarsanam P, Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet 2001; 29:153-9.
40. Sudarsanam P, Pilpel Y, Church GM. Genome-wide cooccurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*. Genome Res 2002; 12:1723-31.
41. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, Botstein D. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell 2002; 13:1977-2000.
42. Kong M, Barnes EA, Ollendorff V, Donoghue DJ. Cyclin F regulates the nuclear localization of cyclin B1 through a cyclin-cyclin interaction. EMBO J 2000; 19:1378-88.
43. Spruck CH, de Miguel MP, Smith AP, Ryan A, Stein P, Schultz RM, Lincoln AJ, Donovan PJ, Reed SI. Requirement of Cks2 for the first metaphase/anaphase transition of mammalian meiosis. Science 2003; 300:647-50.
44. Piekorz RP, Hoffmeyer A, Duntsch CD, McKay C, Nakajima H, Sexl V, Snyder L, Rehg J, Ihle JN. The centrosomal protein TACC3 is essential for hematopoietic stem cell function and genetically interfaces with p53-regulated apoptosis. EMBO J 2002; 21:653-64.
45. Yasuda S, Oceguera-Yanez F, Kato T, Okamoto M, Yonemura S, Terada Y, Ishizaki T, Narumiya S. Cdc42 and mDia3 regulate microtubule attachment to kinetochores. Nature 2004; 428:767-71.
46. Narumiya S, Oceguera-Yanez F, Yasuda S. A new look at Rho GTPases in cell cycle: Role in kinetochoremicrotubule attachment. Cell Cycle 2004; 3:855-7.
47. Glotzer M. The molecular requirements for cytokinesis. Science 2005; 307:1735-9.
48. Townsley FM, Aristarkhov A, Beck S, Hershko A, Ruderman JV. Dominant-negative cyclin-selective ubiquitin carrier protein E2-C/UbcH10 blocks cells in metaphase. Proc Natl Acad Sci USA 1997; 94:2362-7.
49. Jiang W, Jimenez G, Wells NJ, Hope TJ, Wahl GM, Hunter T, Fukunaga R. PRC1: A human mitotic spindle-associated CDK substrate protein required for cytokinesis. Mol Cell 1998; 2:877-85.
50. Liu N, Lucibello FC, Zwicker J, Engeland K, Muller R. Cell cycle-regulated repression of B-myb transcription: Cooperation of an E2F site with a contiguous corepressor element. Nucleic Acids Res 1996; 24:2905-10.
51. Zhu Z, Shendure J, Church GM. Discovering functional transcription-factor combinations in the human cell cycle. Genome Res 2005; 15:848-55.
52. Pallier C, Scaffidi P, Chopineau-Proust S, Agresti A, Nordmann P, Bianchi ME, Marechal V. Association of chromatin proteins high mobility group box (HMGB) 1 and HMGB2 with mitotic chromosomes. Mol Biol Cell 2003; 14:3414-26.
53. Maxwell CA, Keats JJ, Belch AR, Pilarski LM, Reiman T. Receptor for hyaluronan-mediated motility correlates with centrosome abnormalities in multiple myeloma and maintains mitotic integrity. Cancer Res 2005; 65:850-60.
54. Zannini L, Lecis D, Lisanti S, Benetti R, Buscemi G, Schneider C, Delia D. Karyopherin-alpha2 protein interacts with Chk2 and contributes to its nuclear import. J Biol Chem 2003; 278:42346-51.
55. Schluter C, Duchrow M, Wohlenberg C, Becker MH, Key G, Flad HD, Gerdes J. The cell proliferation-associated antigen of antibody Ki-67: A very large, ubiquitous nuclear protein with numerous repeated elements, representing a new kind of cell cycle-maintaining proteins. J Cell Biol 1993; 123:513-22.
56. Raemaekers T, Ribbeck K, Beaudouin J, Annaert W, Van Camp M, Stockmans I, Smets N, Bouillon R, Ellenberg J, Carmeliet G. NuSAP, a novel microtubule-associated protein involved in mitotic spindle organization. J Cell Biol 2003; 162:1017-29.
57. McGuinness BE, Hirota T, Kudo NR, Peters JM, Nasmyth K. Shugoshin prevents dissociation of cohesin from centromeres during mitosis in vertebrate cells. PLoS Biol 2005; 3:e86.
58. Furukawa K. LAP2 binding protein 1 (L2BP1/BAF) is a candidate mediator of LAP2-chromatin interaction. J Cell Sci 1999; 112:2485-92.
59. Dou Z, Ding X, Zereshki A, Zhang Y, Zhang J, Wang F, Sun J, Huang H, Yao X. TTK kinase is essential for the centrosomal localization of TACC2. FEBS Lett 2004; 572:51-6.

5. Functional genomic delineation of TLR-induced transcriptional networks

BMC Genomics

BioMed Central

Research article

Functional genomic delineation of TLR-induced transcriptional networks

Ran Elkon†1, Chaim Linhart†2, Yonit Halperin2, Yosef Shiloh1 and Ron Shamir*2

Address: [1]The David and Inez Myers Laboratory for Genetic Research, Department of Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University, Tel Aviv 69978, Israel and [2]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

Email: Ran Elkon - ranel@post.tau.ac.il; Chaim Linhart - chaiml@post.tau.ac.il; Yonit Halperin - yonithal@post.tau.ac.il; Yosef Shiloh - yossih@post.tau.ac.il; Ron Shamir* - rshamir@post.tau.ac.il

* Corresponding author †Equal contributors

Published: 29 October 2007

BMC Genomics 2007, **8**:394 doi:10.1186/1471-2164-8-394

Received: 4 July 2007
Accepted: 29 October 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/394

Abstract

Background: The innate immune system is the first line of defense mechanisms protecting the host from invading pathogens such as bacteria and viruses. The innate immunity responses are triggered by recognition of prototypical pathogen components by cellular receptors. Prominent among these pathogen sensors are Toll-like receptors (TLRs). We sought global delineation of transcriptional networks induced by TLRs, analyzing four genome-wide expression datasets in mouse and human macrophages stimulated with pathogen-mimetic agents that engage various TLRs.

Results: Combining computational analysis of expression profiles and cis-regulatory promoter sequences, we dissected the TLR-induced transcriptional program into two major components: the first is universally activated by all examined TLRs, and the second is specific to activated TLR3 and TLR4. Our results point to NF-kB and ISRE-binding transcription factors as the key regulators of the universal and the TLR3/4-specific responses, respectively, and identify novel putative positive and negative feedback loops in these transcriptional programs. Analysis of the kinetics of the induced network showed that while NF-kB regulates mainly an early-induced and sustained response, the ISRE element functions primarily in the induction of a delayed wave. We further demonstrate that co-occurrence of the NF-kB and ISRE elements in the same promoter endows its targets with enhanced responsiveness.

Conclusion: Our results enhance system-level understanding of the networks induced by TLRs and demonstrate the power of genomics approaches to delineate intricate transcriptional webs in mammalian systems. Such systems-level knowledge of the TLR network can be useful for designing ways to pharmacologically manipulate the activity of the innate immunity in pathological conditions in which either enhancement or repression of this branch of the immune system is desired.

Background

Immune systems in vertebrates have two basic arms: innate and adaptive immunity. The innate immune system is the first line of defense protecting the host from invading pathogens such as bacteria and viruses. It consists of various types of leukocytes (e.g., blood monocytes, neutrophils, tissue macrophages, dendritic cells) that specialize in phagocytosis (ingesting and digesting pathogens) and in evoking a complex response at the site of infection, collectively known as inflammation. The adaptive immunity arm is capable of specifically recognizing and selectively eliminating foreign microorganisms and molecules. It relies on T and B lymphocytes that express antigen-specific receptors. Upon encountering their specific antigens, these lymphocytes undergo extensive proliferation (clone expansion), maturation and activation. There are multiple cross-talks between the innate and adaptive immunity arms. For example, the phagocytic cells are intimately involved in the activation of the adaptive arm by functioning as antigen presenting cells (APCs) required for the activation of T lymphocytes, and T_H lymphocytes secrete stimulatory cytokines that enhance phagocytosis by the specialized phagocytic cells.

Innate immune responses to pathogens are triggered by recognition of prototypical pathogen components, called pathogen-associated molecular patterns (PAMPs), through cellular pattern recognition receptors (PRRs). Prominent among these pathogen sensors is the family of Toll-like receptors (TLRs). To date, ten and thirteen TLR genes have been cloned in human and mouse, respectively; each of the TLRs appears to recognize a unique set of PAMPs [1,2]. TLR1, 2, 4, 5 and 6 are expressed on the cell surface membrane and recognize bacterial and fungal products, while TLR3, 7, 8 and 9 reside in intracellular endosomes and specialize in detection of pathogens' nucleic acids [3]. For example, lipopolysaccharide (LPS), which is a common structure of the cell wall of Gram-negative bacteria, is recognized by the extracellular TLR4, whereas double-stranded RNA (dsRNA), which is a viral PAMP, triggers the intracellular TLR3 signaling. The function of the other TLRs is less characterized.

After recognition of their ligands, TLRs trigger intricate cellular signaling pathways that endow the cells with antiviral and antibacterial states, which are acquired by the induction of protein effectors that impede viral replication and bacteria growth, and of inflammatory cytokines, chemokines and co-stimulatory molecules that enhance the activation of the adaptive immune response [2,4]. The activation of this broad response is mediated by a signaling cascade that leads to stimulation of several transcription factors (TFs), primarily NF-kB, IRF3/7, and AP-1. Important among the induced cytokines are the interferons (IFNs), whose secretion results in the induction of a set of IFN-stimulated genes (ISGs), which are vital components in the development of antiviral and antimicrobial cellular states [5]. The transactivation of the ISGs is controlled via the JAK/STAT signaling pathway either by an IFNa/b-activated TF complex termed ISGF3 (composed of STAT1, STAT2 and IRF9), which binds to a regulatory element denoted as ISRE (IFN-stimulated response element) [5,6], or by an IFNg-activated STAT1 homodimer complex, which binds primarily to the GAS regulatory element [7].

The transcriptional program spanned by activated TLRs encompasses hundreds of genes. The advent of gene expression microarrays and the availability of complete sequences of the mouse and human genomes enable study of these networks on the system level. Here, we analyzed four publicly available genome-wide datasets that recorded expression profiles in mouse and human macrophages stimulated with various pathogen-mimetic agents, with the goal of obtaining global delineation of the transcriptional network activated by TLRs. Combining computational analyses of gene expression profiles and cis-regulatory promoter sequences, we dissected the TLR-induced transcriptional program into two major components: the first is universally activated by all examined TLRs, and the second is specific to TLR3 and TLR4. Our results identify NF-kB as the key regulator of the universal TLR response and the ISRE element as the key control site of the TLR3/4 specific component, and reveal, on a genomic scale, known and novel target genes regulated by these elements. We also identify novel putative positive and negative feedback loops in these transcriptional programs, further increasing the complexity of the known tightly regulated network induced in response to pathogen invasion. Analysis of the kinetics of the induced network showed that while NF-kB regulates mainly an early-induced and sustained response, the ISRE element functions primarily in the induction of a delayed wave. In addition, we demonstrate that the pair of NF-kB and ISRE elements constitutes a cis-regulatory module that endows its targets with enhanced responsiveness to TLR3/4 activation. By combining expression and promoter analyses, we substantially reduced the high level of noise inherent in genome-wide analysis of such data, and obtained highly reliable results supported by independent datasets from both human and mouse.

Results

We sought to obtain a global view of the transcriptional programs that are induced by activated TLRs, and to identify components common to all TLRs and those specific to some of them. To this end, we used four large-scale gene expression datasets that examined global response in mouse and human macrophages stimulated with various TLR stimulators [8-10] (Table 1). Our analysis flow is schematically sketched in Figure 1 and is described in detail in the sections below. In brief, starting with the mouse datasets, we first partitioned the induced genes into disjoint groups according to the subset of stimulators

Table 1: Summary of datasets analyzed in this study

Dataset	MmBMM	MmRAW	HsM1	HsM2
Reference	Gilchrist et al. (2006) [8]	[11]	Nau et al. (2002) [9]	Jeffrey et al. (2006) [10]
Organism	Mouse	Mouse	Human	Human
Cells	BMM	RAW264.7	Mph	Mph
Stimulators	LPS, CpG, PAM2, PAM3, PIC, R848	LPS, CpG, PAM2, PAM3, PIC, R848	LPS, PIC	LPS
Time-points	0 h, 20 m, 40 m, 1 h, 80 m, 2 h, 8 h*, 24 h*	0 h, 1 h*, 2 h*, 4 h, 8 h*, 24 h*	0 h, 1 h, 2 h, 6 h, 12 h, 24 h	0 h, 4 h
Microarray	Affymetrix MG430 2.0	Two-channel oligonucleotide chip (Operon)	Affymetrix HU6800	Affymetrix HGU133A
# distinct annotated genes	15,277	11,442	5,215	7,981
Replicates	Triplicates	Quadrareplicates	One (two at time 0 h)	Duplicates

* time-points measured only for LPS

to which the genes were responsive. Applying computational analysis of cis-regulatory promoter elements we sought to discover the major TFs that control each of the identified response groups. Next, we analyzed the kinetics of the transcriptional network induced by LPS treatment, and identified the TFs that regulate each kinetic pattern. Finally, we corroborated the results obtained on the mouse datasets by demonstrating their validity in independent human datasets.

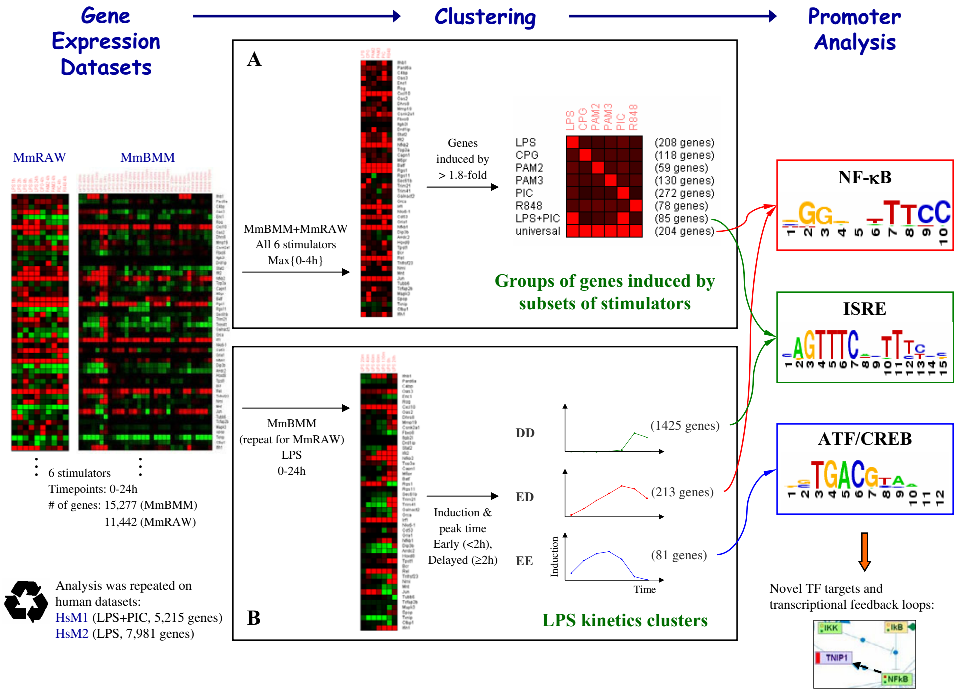


Figure 1
Analysis Flow. A schematic sketch of the major steps in our analysis. Using two comprehensive mouse gene expression datasets, we partitioned the genes into distinct groups according to the subset of TLR stimulators to which they were responsive (A), and identified the TFs that control each response group by computational analysis of cis-regulatory promoter elements. We then characterized three kinetic patterns of the transcriptional network induced by LPS treatment (B), and again discovered the TFs that regulate each pattern. A similar analysis of two independent human datasets confirmed our main findings. Integrating the various sources of information points to novel putative targets of the studied TFs, adding new regulatory links to the transcriptional network of the innate immune system.

Characterization of TLR-induced transcriptional networks

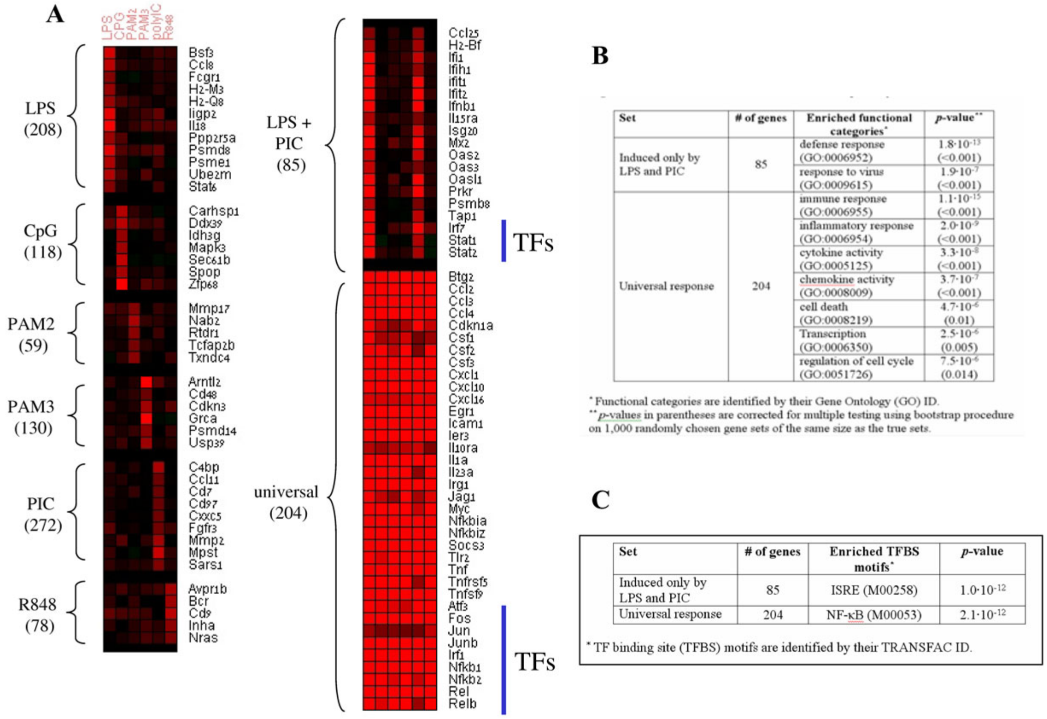
In the first step of the analysis, we analyzed the comprehensive gene expression dataset gathered by the Innate-Immunity System-Biology project [11], in which expression profiles were recorded in two murine macrophage cellular systems (bone marrow-derived macrophage cells (BMM) and the RAW264.7 monocyte macrophage-like cell line) at several time points after exposure to six agents, each in a separate experiment. We began with the mouse datasets because they included more stimulators and denser kinetics than the human datasets. The following are the agents examined in mouse, and the TLRs they activate: LPS – TLR4; PAM2 – TLR2:6; PAM3 – TLR1:2; poly I:C (PIC, in short) – TLR3; R848 – TLR7 and TLR8, and CpG – TLR9 (see Table 2). In order to distinguish agent-specific from common responses, we divided the genes into disjoint groups according to the subset of agents in which they were induced. Each group consisted of genes that were up-regulated by at least 1.8-fold (at any time point) by a particular subset of agents, and did not exceed this factor of induction by all other agents (a list of these genes and their group assignment is provided in Additional File 1). In this analysis we included only the time points common to all probed agents: 20 mins, 40 mins, 1 hr, 80 mins and 2 hrs in the MmBMM dataset, and 4 hrs in the MmRAW dataset. Groups with less than 40 genes were ignored, as they do not contain sufficient information for further statistical analysis. Obviously, in such partition some genes are classified somewhat arbitrarily, e.g., a gene whose induction level is slightly above the 1.8 cutoff in LPS and slightly below 1.8 in all other agents, is assigned to the LPS-specific group. However, the mean expression pattern of each gene group reveals a sharp difference between the average induction level in response to the agent(s) that defines the group and the average induction level in response to all other agents (see Additional File 2), indicating that the borderline genes are a minority within the groups. We identified two induction patterns in addition to the six agent-specific sets (Figure 2A): 1) a large core universal response – 204 genes that were induced by all examined stimulators; and 2) a response only to LPS and PIC (which engage TLR4 and TLR3, respectively) – 85 genes that were induced by LPS and PIC, and did not pass the 1.8-fold threshold in the four other stimulators. Remarkably, both of the above sets are substantially larger than all the other non-agent-specific groups (55 groups in total, all of which contained less than 40 genes, with an average size of only 7 genes), pointing to the major biological role of these two response components in the TLR induced network.

Functional characterization utilizing the standard GO ontology [12] revealed that the universal and TLR3/4-specific responder sets were highly enriched for functions related to the innate immune response, including inflammation, and chemokine and cytokine activities (Figure 2B). Interestingly, no enrichment for any functional category was detected for the agent-specific sets. One explanation could be that these sets contain more false positives, as detection of genes induced only in a single condition is more prone to noise. In addition, it is possible that genes specifically induced by a single stimulator are less functionally characterized.

Our next goal was to identify the regulators that underlie the induction of the TLR-mediated transcriptional programs. We and others have demonstrated that combining computational analysis of cis-regulatory promoter elements with gene expression measurements can identify major transcription factors (TFs) that regulate transcriptional networks, even in complex mammalian systems [13-16]. We applied the promoter analysis algorithm PRIMA [14] implemented in the EXPANDER package [17]. Given a target set and a background set of genes, PRIMA performs statistical tests to identify TFs whose binding site (BS) signatures are significantly more prevalent in the promoters of the target set than in the background set. Here, each of the eight gene sets was considered a target set and the entire set of 10,113 genes present on both arrays used in the MmBMM and MmRAW datasets served as the background set (see Methods). PRIMA identified significant over-representation of the NF-kB binding site signature in the group of genes that were induced by all TLRs ($p = 2 \cdot 10^{-12}$), and of the ISRE element in the set of genes that were induced only by LPS and PIC ($p = 10^{-12}$) (Figure 2C). As in the functional analysis, no over-represented promoter signals were detected for the agent-specific clusters. PRIMA tests are confined to

Table 2: Stimulators used in the mouse MmBMM and MmRAW datasets

Agent	Description	Engaged TLR
LPS	Lipopolysaccharide is a component of the bacterial cell wall (gram-negative bacteria)	TLR4
PAM2	Synthetic diacylated lipopeptide (mimics bacterial lipoproteins)	TLR2:6
PAM3	Synthetic triacylated lipopeptide (mimics bacterial lipoproteins)	TLR1:2
PIC	Polyinosine-polycytidylic acid (Poly I:C) is a synthetic mimic of viral double-stranded RNA	TLR3
R848	Synthetic molecule of the imidazoquinoline family (mimics a viral product)	TLR7/8
CpG	Mimics bacterial and viral CpG DNA motifs	TLR9



B

Set	# of genes	Enriched functional categories*	*p*-value**
Induced only by LPS and PIC	85	defense response (GO:0006952)	$1.8 \cdot 10^{-13}$ (<0.001)
		response to virus (GO:0009615)	$1.9 \cdot 10^{-7}$ (<0.001)
Universal response	204	immune response (GO:0006955)	$1.1 \cdot 10^{-15}$ (<0.001)
		inflammatory response (GO:0006954)	$2.0 \cdot 10^{-9}$ (<0.001)
		cytokine activity (GO:0005125)	$3.3 \cdot 10^{-8}$ (<0.001)
		chemokine activity (GO:0008009)	$3.7 \cdot 10^{-7}$ (<0.001)
		cell death (GO:0008219)	$4.7 \cdot 10^{-6}$ (0.01)
		Transcription (GO:0006350)	$2.5 \cdot 10^{-6}$ (0.005)
		regulation of cell cycle (GO:0051726)	$7.5 \cdot 10^{-6}$ (0.014)

* Functional categories are identified by their Gene Ontology (GO) ID.
** *p*-values in parentheses are corrected for multiple testing using bootstrap procedure on 1,000 randomly chosen gene sets of the same size as the true sets.

C

Set	# of genes	Enriched TFBS motifs*	*p*-value
Induced only by LPS and PIC	85	ISRE (M00258)	$1.0 \cdot 10^{-12}$
Universal response	204	NF-κB (M00053)	$2.1 \cdot 10^{-12}$

* TF binding site (TFBS) motifs are identified by their TRANSFAC ID.

Figure 2

TLR-induced transcriptional programs. (A) Genes that were induced by at least one of the six examined TLR stimulators (induction of at least 1.8-fold at any time point) were partitioned into distinct sets according to their agent-induction pattern. Taking into account sets that contained at least 40 genes, only two complex induction patterns were identified in addition to the six agent-specific patterns: universal and LPS-PIC patterns. Selected genes are shown in the heat-map for each set (a complete list of genes is provided in Additional File 1). The maximum induction of the gene over the examined time points per stimulator is depicted in the heat-map. (B) Enriched GO functional categories were identified in the universal and LPS-PIC sets (*p*-values in parentheses are corrected for multiple testing using a bootstrap procedure on 1,000 randomly chosen gene sets of the same size as the true sets). (C) Highly significant over-represented cis-regulatory elements were identified in the promoters of the universal and LPS-PIC sets, pointing to a pivotal role for NF-kB and ISRE in the induction of these two components of the TLR-induced transcriptional program.

TFs with characterized binding site signatures. Search for novel elements using the MEME motif discovery tool [18] did not find any additional motif, except for the ubiquitous Sp1 signature in several sets. Taken together, the analysis suggests that while NF-kB is universally activated by all TLRs, the TFs that act via the ISRE element (namely, IRF3/7 and the STAT1:STAT2:IRF9 (ISGF3) complex) are activated specifically by the TLR4- and TLR3-mediated signaling pathways. Indeed, many key targets of NF-kB and the ISRE element are in the universal and TLR3/4 sets, respectively, as shown in Figure 2A. Notably, in support of this model, the Nf-kb1, Nf-kb2, Rel and Relb subunits of NF-kB are themselves included in the universal set (that is, they were induced in response to all agents), while Irf7, Stat1 and Stat2, which bind the ISRE, were specifically induced by the LPS and PIC treatments. (Irf9, the third component of the ISGF3 complex, was up-regulated in response to LPS and PIC as well, but only at late time-points – 8 h, 24 h for LPS, 4 h for PIC. As noted above, here we analyzed only time-points 0–4 h, which are common to all the examined TLR-inducing agents.)

Carrying out a similar analysis on the sets of down-regulated genes (using the minimum expression value over time-points 0–4 h in all six agents) did not yield any significant results. However, taking into account the later time-points of 8 h and 24 h (measured only for LPS) identified enrichment of cell-cycle related GO categories and TFs (namely, E2F, NF-Y; data not shown), reflecting proliferation arrest upon pathogen recognition.

Kinetics of the LPS-induced transcriptional response

Expression profiles in response to LPS stimulation were recorded at denser time points (20 mins, 40 mins, 1 hr, 80 mins, 2 hrs, 8 hrs and 24 hrs in the MmBMM dataset, and 1 hr, 2 hrs, 4 hrs, 8 hrs and 24 hrs in the MmRAW dataset), which permitted detailed analysis of the kinetics of the transcriptional program induced by this agent. We partitioned the genes that were induced by LPS (1,719 and 1,239 genes in MmBMM and MmRAW, respectively) into three sets according to the kinetics of their induction, as follows: For each gene we recorded the first time at which it exceeded the 1.8-fold induction threshold, as well as the time at which its expression was highest; we defined three kinetic patterns: 1) Early induction and early peak ('EE' set), containing the genes that peaked (and, obviously, were first induced) before 2 hrs; 2) Early induction and delayed peak ('ED' set) – the genes that were first induced before 2 hrs and peaked at 2 hrs or later; and 3) Delayed induction and delayed peak ('DD' set) – the genes that were first induced (and thus also peaked) at 2 hrs or later (Figure 3A). In both datasets, the 'DD' set was considerably larger than the two other sets, reflecting the fact that the main transcriptional response to LPS exposure was at 2 hrs or later.

Searching for TFs that control these kinetic waves, we applied PRIMA to these six sets (three in each dataset). We identified over-representation of the following BS signatures in both datasets: ATF/CREB in the promoters of genes assigned to the 'EE' set; NF-kB in the 'ED' set; and ISRE in the 'DD' set (Table 3). In addition, enrichment for SRF BS signature was identified in the 'EE' set in MmRAW, and for ETS in the 'DD' set in MmBMM. These results suggest a model in which TFs of the ATF/CREB family modulate an immediate transcriptional response, NF-kB

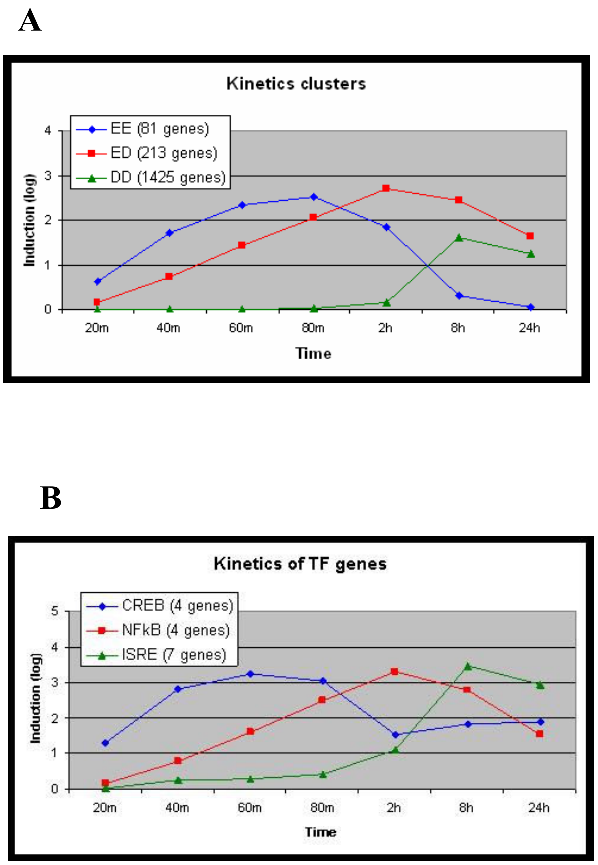


Figure 3

Kinetics of the LPS-induced transcriptional response. (A) Genes that were induced by LPS (by at least 1.8-fold) were divided into three kinetic sets according to the time their expression was first induced and the time it peaked. The 'EE' set contains the early induction, early peak genes; the 'ED' set contains early induction, delayed peak genes; and the 'DD' set contains delayed induction, delayed peak genes. The figure displays the mean expression patterns of the genes assigned to the three kinetic sets in the MmBMM dataset (y-axis is $\log_2$ of induction fold). (B) Mean expression of induced genes that encode for TFs: ATF/CREB (Atf3, Fos, Jun, Junb), NF-kB (Nfkb1, Nfkb2, Rel, Relb), and ISRE (Irf1, Irf2, Irf7, Stat1, Stat2, Stat3, Stat5a). The expression pattern of each TF is highly correlated with that of the kinetic wave, in which the computational promoter analysis found an over-representation of its BSs (compare the kinetic expression of the TF genes (B) and the induced waves (A)).

Table 3: TFBS over-represented in kinetic waves induced by LPS

Kinetics set	Enriched TFBS motifs	Dataset	# of genes	*p*-value
EE	ATF/CREB (M00177)	MmBMM	81	$1.0 \cdot 10^{-8}$
		MmRAW	100	$1.1 \cdot 10^{-5}$
	SRF (M00810)	MmRAW	100	$4.1 \cdot 10^{-6}$
ED	NF-kB (M00053)	MmBMM	213	$2.3 \cdot 10^{-7}$
		MmRAW	133	$2.9 \cdot 10^{-6}$
DD	ISRE (M00258)	MmBMM	1425	$1.7 \cdot 10^{-17}$
		MmRAW	1006	$8.4 \cdot 10^{-11}$
	ETS (M00971)	MmBMM	1425	$1.9 \cdot 10^{-8}$

controls an early response that persists longer, and TFs that act via the ISRE element (members of the IRF and STAT families) regulate mainly the delayed transcriptional response. Importantly, in accordance with this model, we observed that genes that encoded for TFs of the respective families followed a kinetic pattern that was correlated with the one manifested by their putative targets (Figure 3 and Table 3). To further corroborate this kinetic model, we carried out a complementary analysis in which we compared the induction kinetics of putative targets of NF-kB and ISRE based on appearance of strong TF binding site (TFBS) motif hits in their promoters (as identified by PRIMA). Comparing the induction of the putative targets of NF-kB or ISRE, but not both (82 and 112 genes, respectively), indeed showed that targets of NF-kB were induced before targets of ISRE ($p < 0.01$ in both datasets; see Methods). Similar statistical tests showed that genes whose promoter contained an ATF/CREB BS signature peaked at earlier time points than induced genes whose promoter did not contain this cis-regulatory element ($p < 0.0001$ in both datasets).

An additive effect of the pair of NF-kB and ISRE elements

The above results suggest that NF-kB and the IRF-like TFs that act via the ISRE element mainly regulate separate components of the TLRs-induced program and different response waves induced by LPS. Yet, genome-wide scan identified 55 genes whose promoters contained hits for these two regulatory elements. In 27 (49%) of these promoters, the ISRE element is located upstream to the NF-kB putative site, indicating no order bias between the two elements. We next examined whether there is an enhanced effect when NF-kB and ISRE elements co-occur; in other words, do genes whose promoter contains both BSs exhibit a unique expression pattern? We did this by comparing the expression of these genes after exposure to LPS to that of putative targets of each single element separately. Targets of the NF-kB+ISRE pair tended to have higher expression values than genes with only one of these elements (Figure 4). Specifically, when the putative targets of NF-kB were sorted in descending order according to their maximal expression value in MmBMM (over all time points), the top 10% genes were significantly enriched for the NF-kB+ISRE pair ($p < 0.005$; see Methods). The top 10% genes with the ISRE element were also enriched for the pair ($p < 0.05$). This finding points to an additive effect of these two regulatory elements that boosts the induction of the respective target promoters beyond the induction of genes controlled by only one of them. This suggests that the NF-kB and ISRE cis-elements form together a functional regulatory module in promoters of genes that are induced by LPS. An alternative explanation for this observation is that the identification of targets of a single cis-element is more prone to false-positives than that of both elements, and therefore the expression values we obtained for the set of putative targets of NF-kB and ISRE separately were attenuated to a larger extent by false-positives than the expression of putative targets of the module. However, previous studies support

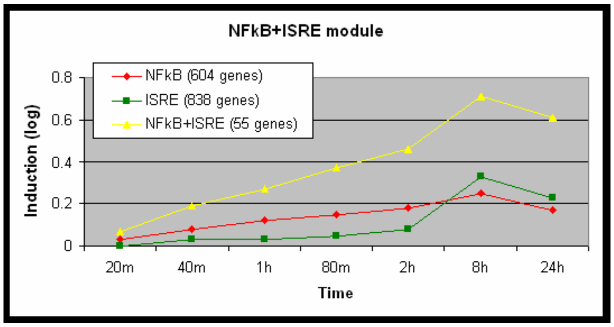


Figure 4

Identification of the NFkB+ISRE cis-regulatory module. Mean expression patterns after exposure to LPS (MmBMM dataset) were computed for three disjoint sets of genes – putative targets of each single element separately (604 NF-kB targets, 838 ISRE targets), and targets of both elements (55 genes), obtained by scanning the promoters of all the genes in the MmBMM dataset. Y-axis is average $\log_2$ of induction fold relative to time 0. Genes whose promoters contain hits for both NF-kB and ISRE elements were more strongly induced by LPS than genes whose promoters contain a hit for only one of these two elements.

the additive effect of the NF-kB+ISRE module, reporting several genes that were co-regulated by NF-kB and ISRE. Doyle et al [1], for example, experimentally demonstrated functional cooperation between NF-kB and IRF3 in the induction of IFNb and IP-10 (CXCL10) in response to LPS.

Corroboration of the findings on independent human macrophage datasets

The results presented hitherto were inferred from analysis of responses of mouse macrophages to various TLR stimuli. Seeking corroboration of our findings in human cells, we analyzed two publicly available datasets that profiled transcriptional responses in immunologically challenged human macrophages. The first study, by Nau et al. [9], examined expression profiles in human monocyte-macrophages at several time points (1 hr, 2 hrs, 6 hrs, 12 hrs and 24 hrs) after stimulation by various agents; among them LPS and PIC are common to the stimuli examined by the mouse datasets we analyzed (this dataset is hereafter called HsM1). The second study, by Jeffery et al. [10] (hereafter called HsM2), profiled transcriptional responses in several human leukocytes challenged with various stimuli, among which monocyte-macrophages treated with LPS for 4 hrs were relevant to our analysis (see Table 1). These two studies provided us with independent data that profiled the transcriptional network induced by activated human macrophages, and allowed us to examine whether our findings on the major roles of NF-kB and ISRE elements in the activation of the transcriptional networks induced by activated TLR4 (LPS) and TLR3 (PIC) are valid also in humans.

Analyzing the HsM1 dataset, we first identified the genes that were induced by LPS alone or by PIC alone, or by both treatments, and subjected these three gene sets to computational promoter analysis. In full accordance with the results obtained on the mouse data, an unbiased search for TFs that underlie the networks induced by LPS and PIC in HsM1 did not identify any signal in the sets of genes that responded specifically to either LPS or to PIC, but did detect a significant over-representation of NF-kB and ISRE elements in the promoters of genes that were induced by both agents (Table 4). This over-representation reflects the superposition of the two components of the TLR-induced transcriptional program: the universal response induced by all TLRs (mediated by NF-kB) and the TLR3/4-specific component (regulated by TFs that act via the ISRE element). These findings were further supported by the second human macrophage dataset that we analyzed: 505 genes were induced by at least 1.8-fold at 4 hrs after LPS treatment in the HsM2 dataset. Unbiased computational promoter analysis again detected only two signals enriched in this gene set: NF-kB ($p = 8.8 \cdot 10^{-8}$) and ISRE ($1.4 \cdot 10^{-12}$).

Table 4: TFBS over-represented in the response induced by LPS and PIC in the HsM1 dataset

Set	# of genes	Enriched TFBS motifs	*p*-value
Induced only by LPS	196	---	---
Induced only by PIC	123	---	---
Induced by both LPS and PIC	75	NF-kB (M00053)	$1.1 \cdot 10^{-7}$
		ISRE (M00258)	$8.3 \cdot 10^{-9}$

Next, we sought to demonstrate that the kinetic model that emerged in the analysis of the mouse datasets remains valid for the human data. Following the analysis applied to the mouse datasets, we partitioned the genes induced by LPS and PIC in the human datasets to the three kinetic sets: 'EE', 'ED' and 'DD' (again, using the 2 hr time point as the boundary between early and delayed time points), according to the kinetics of their activation, and searched for over-represented signals in the promoters of these gene sets. In agreement with the results obtained on the mouse dataset, here too we observed a strong enrichment for NF-kB and ISRE elements in the 'ED' (early induction, delayed peak) and 'DD' (delayed induction and peak) sets, respectively (Table 5). In contrast to the results found on the mouse dataset (Table 3), we did not detect here an over-representation of ATF/CREB in the 'EE' set (representing early induction and peak). This is probably due to the small size of this set and the existence of only a single "early" time-point (1 hr), which might have hindered statistical detection of enriched signals.

Last, we examined whether the additive effect between the NF-kB and ISRE cis-elements could be detected also in the human macrophage datasets. Indeed, the same statistical test we applied to the mouse data revealed that in both HsM1 and HsM2, the 10% most highly induced putative targets of each of the two elements were significantly enriched for genes whose promoter contained a signature for both NF-kB and ISRE (Table 6).

Discussion

In this study we systematically delineated the transcriptional program induced by stimulation of various TLRs in macrophages. We dissected two major components of this program: the first is a core response universally activated by all examined TLRs, and the second is specifically activated by TLR3 and TLR4. Our analysis identified NF-kB and IRF-like TFs binding ISRE as the key regulators of these two components and pointed to their respective target genes on a genomic scale. While the involvement of NF-kB and IRF-like TFs in response to TLR induction has been known before, our study makes novel contributions to several aspects of system-level understanding of the

Table 5: TFBS over-represented in kinetic waves induced by LPS and PIC in the HsM1 dataset

Kinetics set	TFBS motif	Stimulator	# of genes	*p*-value
EE	---	LPS	12	
		PIC	80	---
ED	NF-kB (M00053)	LPS	34	$7.6 \cdot 10^{-7}$
		PIC	7	---
DD	ISRE (M00258)	LPS	225	$1.8 \cdot 10^{-5}$*
		PIC	111	$1.9 \cdot 10^{-7}$

* a similar TFBS motif of the same element (M00972) received *p*-value $8.4 \cdot 10^{-7}$.

transcriptional networks induced by innate immunity: (a) the combined, focused reanalysis of four independent datasets identifying a clean, combinatorial response; (b) revealing the intricate kinetics of the transcriptional response; (c) pinpointing novel specific genes involved in each of the responses; (d) identification of NF-kB and ISRE binding site locations over target genes; and (e) the refinement of the understanding of the regulatory circuitry involved in innate immune response.

Novel targets of NF-kB and ISRE identified in this study (see selected examples in Tables 7 and 8) call for experimental validation. Typically, a genome-wide scan for putative TF targets is prone to a high rate of false positives. However, the candidates we identified are based on diverse evidence that collectively increase the confidence that they are true targets: their induction was triggered by several stimulators in multiple time points and in independent studies on two organisms; and in most cases the respective BS signature was identified in both the human and mouse orthologous promoters.

The repertoire of the TLR universal response includes proinflammatory cytokines and chemokines (e.g., Ccl2-4, Csf1-3 and Cxcl1, which orchestrate innate immunity fight against pathogens), as well as co-stimulatory molecules (e.g., Il23a) that promote the activation of the T-cell branch of the adaptive immunity. The universal response also contains many general stress-responsive genes (e.g., Jun, Fos, Atf3, Egr1-3, Myc) that control cell proliferation and survival. Prominent among the genes specifically induced by TLR3 and TLR4 are the interferon (IFN)-induced genes (Figure 2A). IFN-induced genes comprise potent antiviral molecules (e.g., Mx2, Isg20, Oas2-3, Prkr) and are therefore expected to be induced by TLR3, which is activated by virally derived dsRNA. However, IFNs also have an important role in linking innate and adaptive immunity by regulating the induction of genes that enhance T-cell activation and antigen-presentation capacity in response to pathogen infection (e.g., Il15, Tap1, Psmb8), which explains their induction by bacterial stimuli such as LPS [19,20].

Without any prior knowledge on TLR signaling, our computational promoter analysis revealed NF-kB as the pivotal regulator of the universal-TLR transcriptional response. This finding is in line with current biological knowledge. Several molecular mechanisms through which NF-kB is activated by TLR signaling have been characterized [2,20]. The first depends on Myd88 and is utilized by all TLRs with the exception of TLR3. Activated TLRs recruit Myd88, which then associates with members of the IRAK family, initiating a cascade in which TRAF6 and TAK1 (official symbol: MAP3K7) are sequentially activated. TAK1 in turn promotes downstream activation of the IKK complex, which leads to the activation of NF-kB by directly phosphorylating, and thereby removing the inhibitory effect of, the members of the IkB family on NF-kB (Figure 5). On the other hand, TLR3 activates NF-kB in a Myd88-independent manner: The TRIF adaptor protein (TICAM1) is recruited to activated TLR3, and then directly interacts with TRAF6, which presumably leads to the activation of NF-kB using the same cascade described above for the Myd88-depndent pathway [20] (Figure 5). Substantiating the universal role of NF-kB in the TLR-induced network, we observed that the NFkB1, NFkB2, Rel and Relb subunits of the NF-kB heterodimer were induced by all examined stimuli.

Table 6: Statistical significance of increased expression of NF-kB+ISRE module.

Dataset	MmBMM	MmRAW	HsM1	HsM2
Targets of module vs. targets of NF-kB	0.0045	0.089	0.059	0.0034
Targets of module vs. targets of ISRE	0.041	0.089	0.015	0.0039

The table shows the *p*-values of the enrichment of the module's putative targets within the top 10% targets of NF-kB/ISRE, based on the genes' maximum induction (across all time-points) in response to LPS.

Table 7: Predicted NF-kB target genes in the universal TLR response network.

Symbol	NFkB BS (location)		LPS maximum induction (log_2)				LPS	Validated
	Human	Mouse	MmBMM	MmRAW	HsM1	HsM2	kinetics	BS
CXCL10	GGGAAATTCC (-176)	GGGAAATTCC (-233)	10.42	5.47	7.98	6.9	ED	[41]
RELB	GGGGTTTTCC (-107)	GGGGTTTTCC (-96)	4.3	1.27	0.77	1.46	ED/EE*	[42]
NFKBIA	TGGAAATTCC (-84)	GGGAAACCCC (-81)	4.1	4.25	3.35	2.58	ED	[43]
NFKB2	GGGAATTCCC (-101,-73)	CGGGAATTCC (-102,-74)	3.56	3.82	1.17	1.72	ED	[44]
SDC4	N/F	GGGGAATTCC (-81)	1.53	2.78	1.01	2.02	DD/ED*	[45]
CD69	GGGAAAATCC (-222)	GGGAAAATCC (-220,-155)	8.3	1.82	0.88	3.11	ED/DD*	[46]
BIRC3	GGAAATCCCC (-177)	GGAAATCCCC (-60)	2.97	0.75	3.19	4.03	ED	[47]
MAP3K8	GGAAAACCCC (-724)	CGGAATTTCC (-490)	3.42	0.46	0.65	2.98	ED	---
BATF	N/F	GGGATTTTCC (-233)	4.51	3.07	3.31	1.04	DD	---
IRG1	N/F	TGGAAATTCC (-50)	10.8	7.69	x	x	ED	---
RIPK2	GGGGCTTTCC (-310)	GGGATTTTCC (-521)	2.51	x	x	3.23	ED	---
GCH1	CGGGCTTTCC (-11)	N/F	3.19	0.82	6.19	3.64	ED/DD**	---
TNIP1	GGGGACTTTC (-68)	N/F	3.51	-0.64	3.01	2.46	ED/DD**	---

Promoter sequences matching the PWMs of NF-kB and ISRE were identified by the PRIMA software; mapping between human and mouse orthologous genes was downloaded from the Ensembl web-site; sequences are shown on the coding strand; in cases of multiple matches, the sequence of the first listed match is shown; "N/F" means no putative BS was found. For each gene, the tables indicate (the log_2 of) its maximum fold-induction (over all time-points) in response to LPS in all four datasets, as well as its kinetic pattern (in case of conflicting patterns in different datasets, the default pattern is in MmBMM, * is in MmRAW, and ** is in HsM1). References for validated BSs are given in the last column.

Superimposed on the TLR universal program, we detected a robust TLR3/4-specific response, and demonstrated by promoter analysis that its key regulator is the ISRE element. In addition, our results indicate that this ISRE-mediated response is kinetically delayed compared to the NF-kB-regulated program. These findings too are corrobo-

Table 8: Predicted ISRE target genes in the specific response to LPS and PIC

Symbol	ISRE BS (location)		LPS maximum induction (log_2)				LPS kinetics	Validated BS
	Human	Mouse	MmBMM	MmRAW	HsM1	HsM2		
IFNB1	GGGAGAAGTGAAAGT (-59)	GGGAGAACTGAAAGT (-150)	5.53	0.44	3.28	x	ED/DD**	[2]
TOR3A	GCGGTTTCATTTCCC (161)	ACTGTTTCATTTTCC (-485)	4.08	2.31	x	-0.19	DD	[48]
OAS3	GAAAGAAACGAAACT (-29,108)	GGAGAAAACGAAAGT (-77,0)	5.21	2.85	x	2.42	DD	[49, 50]
OAS2	TCAGTTTCAGTTTCC (49)	TGAGTTTCGATTTCC (-74)	3.24	2.1	2.56	2.53	DD	[50]
OASL	TTGAGAATCGAAACT (-288)	CACAAAAGAGAAACT (-159)	7.93	5.79	2.7	3.98	ED/DD**	[50]
CFB (BF)	CTTGTTTCACTTTCA (-98)	ATAGTTTCTGTTTCC (-148)	8.38	3.32	2.04	x	DD	[51]
TRIM21	GCGGAAACTGAAAGT (9)	GAGGAAACTGAAAGT (-30,4)	2.65	1.52	2.73	0.22	DD	[52]
IFIH1	ATCGAAACAGAAACC (-178)	ATCGAAACAGAAACC (-65)	4.55	2.92	x	3.09	DD/ED*	---
NMI	N/F	ACCGAAAGTGAAAGT (71)	3.31	1.54	1.61	1.36	DD	---
LGP2	TCAGTTTCAGTTTCC (-1)	TCAGTTTCATTTCTA (-1)	1.87	2.07	x	x	DD	---
RTP4 (IFRG28)	ACAGAAACAGAAACT (-39,-15)	TTGGAAACCGAAACT (-84,-58,-35)	2.65	1.59	x	2.33	DD	---
BATF2	GGAGAAACTGAAACT (-2)	GGAGAAACTGAAACT (-95)	5.64	1.9	x	x	DD	---
STAT2	CTAGTTTCGGTTCCG (-353)	CTGGTTTCAGTTTCC (-303)	5.94	2.01	1.5	1.5	DD	---

See legend of Table 7.

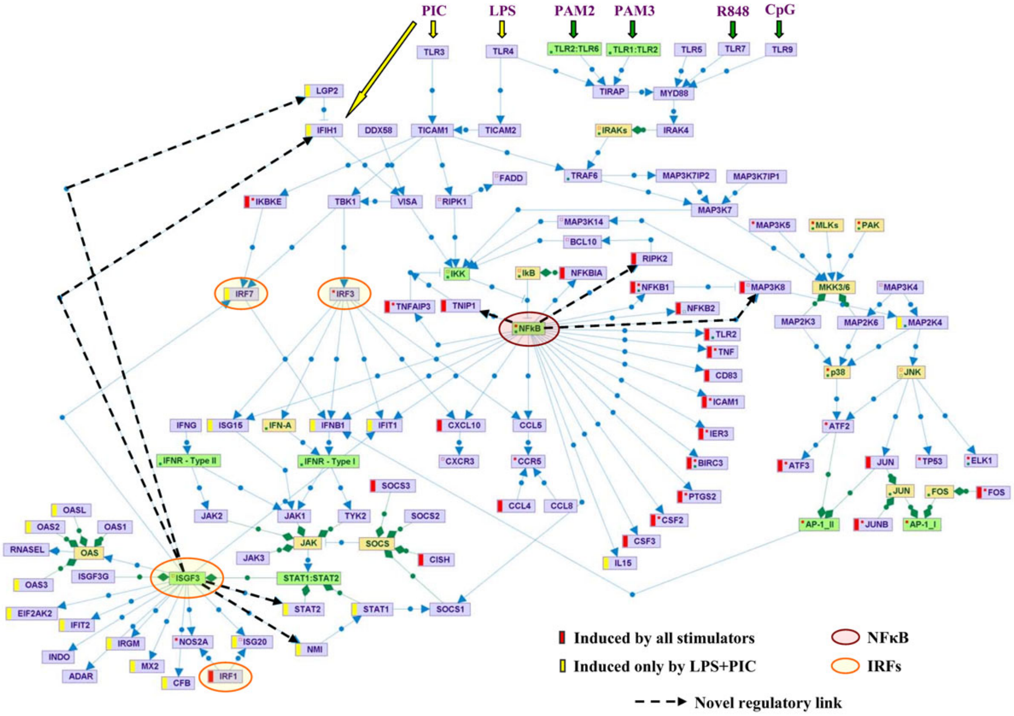


Figure 5

TLR-induced signaling pathways and transcriptional programs. The map, constructed using our SPIKE knowledgebase of signaling pathways [40], presents current knowledge on signaling cascades emanating from activated TLRs and culminating in activation of several key TFs and their respective target genes to achieve robust antiviral and antimicrobial responses. SPIKE maps contain nodes representing three biological entities: gene/proteins (violet nodes); protein complexes (green nodes, e.g., the ISGF3 complex); and gene families (yellow nodes, e.g., the IkB family of NF-kB inhibitors). The map contains two types of edges: Blue edges represent regulations between genes/proteins. Arrowheads correspond to activation, and T-shaped edges (---|) represent inhibition. Green edges represent containment relations between nodes (e.g., the relationships between a complex and its components). Red and green dots within a node indicate that not all the regulation and containment relations stored in SPIKE's DB for that node are displayed on the map. Genes that were universally induced by all examined TLRs are marked by a red bar to the left of the node; genes that were specifically induced by LPS and PIC (which activate TLR4 and TLR3, respectively) are marked by a yellow bar. Novel regulatory links identified in this study that close feedback loops within the TLR-induced network are emphasized in the map by a dashed arrow.

rated by current biological knowledge. The ISRE cis-element is bound by members of the IRF and STAT TF families. Several studies demonstrated the existence of two waves of activation of TFs that act via ISRE by TLR3 and TLR4 [1,20-22]. The emerging model is that IRF3, which is post-translationally activated by TLR3 and TLR4 via a cascade that involves the TRIF (TICAM1) and TRAM (TICAM2) adaptor proteins and their downstream kinases IKKe (IKBKE) and TBK1, promotes an early wave of IFN-b gene induction (Figure 5) [1,23]. Once IFN-b is produced and secreted, it engages the type-I IFN receptor in both paracrine and autocrine fashion, thereby triggering the JAK-STAT signaling cascade that culminates in the activation of the ISGF3 TF complex, which is comprised of STAT1, STAT2 and IRF9 (official symbol: ISGF3G) [24]. ISGF3 induces the expression of IRF7, which in turn further activates the expression of type-I IFNs. In this way, a positive loop is established, which ensures persistent expression of IFN-stimulated genes that enhance the antiviral and antimicrobial cellular state [20]. Strikingly, in full compliance with this model, we observed that IFN-b and the IRF7, STAT1 and STAT2 TFs were specifically induced by LPS and PIC in the datasets we analyzed (Figure 5).

Our analysis points to novel feedback loops in the TLR-induced network, further increasing the known complexity of the regulatory circuits that modulate its induction and repression (see Figure 5 and Tables 7, 8): We identified IFIH1 (also known as MDA5) and LGP2 as novel putative targets regulated by the ISRE element. IFIH1 is a non-TLR cytoplasmic sensor that detects actively replicating viruses [2,25], and triggers the induction of the NF-kB and IRF3 pathways via the activation of the adaptor protein VISA (also known as cardif or IPS-1) [26]. Moreover, it has been recently demonstrated that IFIH1 detects cytoplasmic dsRNA generated during viral replication (while TLR3 detects viral dsRNA phagocytosed in endosomes), and that this sensor also binds to PIC and mediates type I IFN responses to this synthetic analog of viral dsRNA [27]. Therefore, the transcriptional program induced by PIC stimulation probably reflects a combined outcome of the activation of TLR3-mediated and IFIH1-mediated pathways.

Interestingly, the second putative ISRE target we identified, LGP2, is a direct negative regulator of IFIH1 [28]. The simultaneous activation of positive and negative regulators of the same pathway seems to be a recurrent theme in the logic of cellular signaling networks. Another novel putative positive loop in the ISRE-regulated network is mediated by NMI, which enhances the transcriptional activity of STAT-1 [29]. In the NF-kB-regulated transcriptional response, which is universally activated by all examined TLRs, we identified MAP3K8 (also known as TPL-2 and COT) and RIPK2 as novel targets that form positive feedback loops which reinforce the persistent activation of this network [30,31], and TNIP1 as a regulator that forms a negative feedback loop which inhibits the IkK complex, thereby contributing to the turning-off of this response [32].

The kinetic analysis of the response to LPS also suggests a role for the ATF/CREB cis-regulatory element. We identified a significant over-representation of this signature on promoters of genes whose expression peaked at very early time points (before 2 hrs). Two alternative interpretations of the role played by these elements are consistent with this rapid pattern of induction: According to the first, members of the ATF/CREB family activate this early and very short response; the second interpretation ascribes an inhibitory effect to these elements, implying that the TF(s) that act via them repress the expression of their target genes, and therefore the induction of these targets declines shortly after their activation. A recent study by Gilchrist et al. [8] demonstrating that ATF3 negatively regulates a subset of NF-kB target genes induced by TLR4 supports the second interpretation. Notably, the ATF3 gene itself is included in the TLR universal response, pointing to a negative loop that regulates a sub-network of TLR-induced transcriptional program.

The computational promoter analysis ferreted out the major regulators of the two components of the TLR-induced network. This complex transcriptional network is likely regulated by additional TFs, which were not detected by promoter analysis. Indeed, the TLR universal response contains several other TFs in addition to those discussed above (e.g., Egr1-3, c-Myc, Ets2, Fos). This could be explained by the fact that our statistical promoter analysis detects TFs with a relatively high number of direct targets, whose BSs are located within the scanned promoter region and which were responsive beyond a certain threshold in the studied conditions. It is therefore expected to miss TFs that: (a) have a small number of directly induced targets; (b) bind at large distances from the transcription start site; (c) regulate the TLR network by interacting with other TFs rather than directly binding to the DNA; or (d) have a very subtle (though, perhaps, biologically important) influence on the expression of their targets.

Our results suggest mainly distinct programs mediated by the NF-kB and ISRE cis-elements. However, when the two elements co-occured in the same target promoter, we detected an additive effect that boosts the induction of the target genes. This finding further defines the NF-kB+ISRE pair as a functional transcriptional module, and adds several novel candidates to the list of genes reported to be controlled by it [1,33-35] (Tables 7, 8). Importantly, IFN-b is among the genes whose promoters were empirically demonstrated to be under the regulation of the NF-kB+ISRE pair [1].

Conclusion

Our analysis demonstrates the power of functional genomics approaches to delineate intricate transcriptional networks in mammalian systems. Microarray data are often noisy and do not distinguish between direct and secondary responses. Likewise, large-scale promoter scanning for putative TF targets produces many false positives due to the short and degenerate nature of BS signatures. Combining these two sources of information, and augmenting them by utilizing datasets and promoter sequences from both human and mouse, gave us an accurate, system-level delineation of the TLR-induced transcriptional program, and identified highly reliable putative direct targets of its key regulators. The findings reported in this study generalize, on a genomic scale, the current knowledge on the identity, function, kinetics and modular organization of the transcriptional regulators that mobilize the innate immune response, which is often based on studies of specific genes. Such knowledge can be useful for designing ways to pharmacologically manipulate the activity of the innate immunity in pathological conditions in which either enhancement or repression of this branch of the immune system is desired.

Methods

Microarray datasets

The four expression datasets analyzed in this study are summarized in Table 1. We used the original normalized probe expression values, as provided by the authors. In each dataset, we averaged measurements over replicate samples, and then, for each probe, we divided expression values in treated samples by the values in the corresponding control samples (time 0 hr). These fold-change ratios were log (base2)-transformed and averaged over probes that correspond to the same gene. Mapping probes in the MmBMM, HsM1 and HsM2 datasets to Ensembl gene ids was done using annotation files provided by Affymetrix. The MmRAW dataset included the Entrez-Gene id of each probe; we used Biomart [36] to map Entrez-Gene ids to Ensembl gene ids. The HsM1 experiment measured responses of macrophages cultured with LPS derived from *E. coli* (LPS_E) and *Salmonella typhi* (LPS_S). We regarded LPS_E and LPS_S as duplicates and averaged over these two conditions.

Definition of stimulator-induced genes

In all datasets except HsM1, a gene was considered to be induced by a given stimulator if its expression level in one or more of the time points was at least 1.8-fold higher than its expression at time 0. The results we report are not sensitive to the chosen cutoff and remained consistent for a wide range of values (from 1.5- to 2-fold). In HsM1 we used a more stringent threshold of 3.5-fold, since the expression values in this dataset showed a much larger variance, probably because no replicates were performed (except for time 0). This threshold was chosen so that a similar percentage of the genes will be considered induced in HsM1 as in the other datasets.

Groups of genes induced by subsets of stimulators

The two mouse datasets – MmBMM and MmRAW – share 10,113 genes. Using the maximum induction-fold of each of these genes, computed over six time-points (20 mins-2 hrs in MmBMM, and 4 hrs in MmRAW), for each of the six stimulators (LPS, PAM2, PAM3, PIC, R848 and CpG), we partitioned the genes into groups as follows. We enumerated all 63 (= 2^6-1) non-empty subsets of the six stimulators, and for each such subset we collected all the genes that were induced in those stimulators and not induced in the others. Ignoring sets with less than 40 genes, we obtained eight gene sets (Figure 2A): six agent-specific sets (i.e., genes that were induced only in one of the six stimulators), an LPS-PIC specific set, and a universal response set.

In humans, we repeated the above analysis for the LPS and PIC stimulators in the HsM1 dataset. Here, we used all five time-points (1 hr–24 hrs), and an induction threshold of 3.5-fold (see Table 4).

Functional categories analysis

Identification of enriched Gene Ontology (GO) biological processes categories was done using the TANGO algorithm implemented in the EXPANDER package [17]. In brief, TANGO calculates the statistical significance of GO categories' over-representation within a given set of genes by computing the upper tail of the hypergeometric distribution. In order to account for multiple testing, a major challenge in such an analysis due to the strong dependencies among GO categories, TANGO estimates fixed *p*-values using an empirical distribution based on 1,000 randomly chosen gene sets. We report all GO categories with an enrichment *p*-value less than 10^{-5} (before correcting for multiple testing) (see Figure 2B). Association of mouse genes with GO categories was downloaded from the GO web-site [37] (Sep 2006).

Computational promoter analysis

Identification of enriched BS signature of known TFs was done using our PRIMA algorithm [14], which is implemented in the EXPANDER package. PRIMA identifies TFs whose BS signatures are significantly abundant in the promoters of a specified group of genes, given their distribution in the promoters of the entire background set (i.e., all the genes present on the chip). PRIMA uses position weight matrices (PWMs) as models for regulatory sites that are bound by TFs. 498 PWMs that represent human or mouse TFBSs were obtained from the TRANSFAC database (release 10.2, June 2006) [38]. Promoter sequences corresponding to all known human and mouse genes were extracted from the Ensembl project (release 40, Sep 2006) [39]. PRIMA scanned both strands of each promoter sequence in the region from 600 bps upstream to 100 bps downstream of the putative transcription start site (TSS). Repetitive elements were masked out. A detailed description of how PRIMA determines PWM cutoffs, identifies putative TFBSs, and computes enrichment scores is given in [14]. We report TFs with an enrichment *p*-value less than 10^{-5}. We used this stringent threshold due to the large number of PWMs examined. Note, however, that there is a very high level of redundancy in the TRANSFAC database. For example, there are seven different PWMs for NF-kB, which are naturally all very similar. Thus, the actual number of independent multiple tests performed by PRIMA is considerably less than the total number of PWMs. For each of the TFs reported in this study, we chose the PWM that gave the best overall results (in terms of enrichment): M00053 for NF-kB, M00258 for ISRE, and M00177 for ATF/CREB; other PWMs of these TFs often gave very similar *p*-values.

We also subjected each of the eight TLR-induced gene sets (Figure 2) to the MEME program (version 3.0.3) [18]. MEME is a tool for discovering motifs de-novo in a group of related DNA sequences. MEME was run with a 4th-order Markov background model, which we constructed using all the mouse promoter sequences (from 600 bps upstream to 100 bps downstream the TSS). We searched for motifs of length 8 and 10, and used the following options: "-dna -revcomp -mod zoops -evt 0.001 -text -nostatus".

Statistical tests for the kinetics of TF targets

In order to statistically evaluate the difference in the induction time of NF-kB and ISRE targets, we counted the number of putative targets of these elements, denoted s_1 and s_2, respectively, that were induced up to 1 hr after LPS treatment. (genes whose promoter contained both the NF-kB and ISRE signatures were ignored in this test). Given the total number of putative targets (induced at any time-point), denoted t_1 and t_2, respectively, we computed the probability that out of $s_1 + s_2$ early-induced genes, at least s_1 of them are targets of NF-kB. A small probability indicates that statistically significant number of the early-induced genes is regulated by NF-kB. This probability is given by the hypergeometric tail distribution:

$$P = \sum_{i=s_1}^{\min\{t_1, s_1+s_2\}} \frac{\binom{t_1}{i}\binom{t_2}{s_1+s_2-i}}{\binom{t_1+t_2}{s_1+s_2}} \qquad (1)$$

Using a similar statistical test, we showed that the peak time of putative targets of ATF/CREB is significantly earlier than that of all other induced genes. Denoting by t_1 (t_2) the number of LPS-induced genes that are (are not) putative targets of ATF/CREB, out of which s_1 (s_2) reached their maximal expression at or before 1 hr, we computed the hypergeometric probability as above.

Statistical evaluation of increased induction of targets of NF-kB+ISRE

To examine whether there is a significant additive effect between the NF-kB and ISRE elements, we performed the following test: Given the total number of genes whose promoter contains signatures of both NF-kB and ISRE, or only NF-kB, denoted t_1 and t_2, respectively, we checked whether there is an enrichment of NF-kB+ISRE joint targets within the 10% most highly induced NF-kB targets. Here, genes were ranked based on their maximum induction in response to LPS. Let s_1 and s_2 denote the number of NF-kB+ISRE and NF-kB (but not ISRE) targets, respectively, whose induction-fold is above the aforementioned 10% threshold (i.e., $s_1 + s_2 = (t_1 + t_2)/10$). Then, using the standard hypergeometric score (Equation 1), we computed the probability to observe at least s_1 highly-induced NF-kB+ISRE targets, given t_1, t_2 and s_2. For example, in the MmBMM dataset, we found an NF-kB signature in 659 genes, of which 55 also contained an ISRE element; among the 65 NF-kB targets with highest induction by LPS, 12 genes also had an ISRE element. Thus, $t_1 = 55$, $t_2 = 604$, $s_1 = 12$, and $s_2 = 53$, which gives $p = 0.004$.

The above test evaluates the increased expression of putative targets of the pair NF-kB+ISRE with respect to all NF-kB targets. We performed a similar test to check the increased expression of NF-kB+ISRE relative to all ISRE targets.

Authors' contributions

RE and CL conceived the study, carried out the analyses and drafted the manuscript. YH participated in the data analysis and developed the methods for the kinetic analysis. RS and YS participated in the design of the study, led and funded it and reviewed the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

The file lists the genes that were induced by TLR activation and their assignment into the agent specific, LPS+PIC and universal clusters.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-394-S1.xls]

Additional file 2

The mean expression pattern of each gene cluster (representative genes of each cluster are shown in Figure 2A).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-394-S2.xls]

Acknowledgements

We thank Richard Young, Richard Jenner and the Innate-Immunity System-Biology project for making their microarray data publicly available. We thank Ioannis Xenarios, Gideon Schreiber and Diego Jaitin for reading the manuscript and providing helpful advice. R. Elkon was supported by an Eshkol Fellowship from the Ministry of Science, Israel. R. Shamir was supported by the Wolfson Foundation.

References

1. Doyle S, Vaidya S, O'Connell R, Dadgostar H, Dempsey P, Wu T, Rao G, Sun R, Haberland M, Modlin R, Cheng G: **IRF3 mediates a TLR3/TLR4-specific antiviral gene program.** *Immunity* 2002, **17:**251-263.
2. Kawai T, Akira S: **TLR signaling.** *Cell Death Differ* 2006, **13:**816-825.
3. Zhong B, Tien P, Shu HB: **Innate immune responses: crosstalk of signaling and regulation of gene transcription.** *Virology* 2006, **352:**14-21.
4. Oda K, Kitano H: **A comprehensive map of the toll-like receptor signaling network.** *Mol Syst Biol* 2006, **2:**2006 0015.
5. Grandvaux N, Servant MJ, tenOever B, Sen GC, Balachandran S, Barber GN, Lin R, Hiscott J: **Transcriptional profiling of interferon regulatory factor 3 target genes: direct involvement in the regulation of interferon-stimulated genes.** *J Virol* 2002, **76:**5532-5539.
6. Decker T, Muller M, Stockinger S: **The yin and yang of type I interferon activity in bacterial infection.** *Nat Rev Immunol* 2005, **5:**675-687.

7. Schroder K, Hertzog PJ, Ravasi T, Hume DA: **Interferon-gamma: an overview of signals, mechanisms and functions.** *J Leukoc Biol* 2004, **75:**163-189.
8. Gilchrist M, Thorsson V, Li B, Rust AG, Korb M, Kennedy K, Hai T, Bolouri H, Aderem A: **Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4.** *Nature* 2006, **441:**173-178.
9. Nau GJ, Richmond JF, Schlesinger A, Jennings EG, Lander ES, Young RA: **Human macrophage activation programs induced by bacterial pathogens.** *Proc Natl Acad Sci USA* 2002, **99:**1503-1508.
10. Jeffrey KL, Brummer T, Rolph MS, Liu SM, Callejas NA, Grumont RJ, Gillieron C, Mackay F, Grey S, Camps M, Rommel C, Gerondakis SD, Mackay CR: **Positive regulation of immune cell function and inflammatory responses by phosphatase PAC-1.** *Nat Immunol* 2006, **7:**274-283.
11. **The Innate-Immunity System-Biology project** [http://www.systemsbiology-immunity.org]
12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
13. Das D, Nahle Z, Zhang MQ: **Adaptively inferring human transcriptional subnetworks.** *Mol Syst Biol* 2006, **2:**2006 0029.
14. Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y: **Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells.** *Genome Res* 2003, **13:**773-780.
15. Zhu Z, Shendure J, Church GM: **Discovering functional transcription-factor combinations in the human cell cycle.** *Genome Res* 2005, **15:**848-855.
16. Blais A, Tsikitis M, Acosta-Alvear D, Sharan R, Kluger Y, Dynlacht BD: **An initial blueprint for myogenic differentiation.** *Genes Dev* 2005, **19:**553-569.
17. Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, Sharan R, Shiloh Y, Elkon R: **EXPANDER--an integrative program suite for microarray data analysis.** *BMC Bioinformatics* 2005, **6:**232.
18. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proc Int Conf Intell Syst Mol Biol* 1994, **2:**28-36.
19. Hoebe K, Beutler B: **LPS, dsRNA and the interferon bridge to adaptive immune responses: Trif, Tram, and other TIR adaptor proteins.** *J Endotoxin Res* 2004, **10:**130-136.
20. Moynagh PN: **TLR signalling and activation of IRFs: revisiting old friends from the NF-kappaB pathway.** *Trends Immunol* 2005, **26:**469-476.
21. Taniguchi T, Ogasawara K, Takaoka A, Tanaka N: **IRF family of transcription factors as regulators of host defense.** *Annu Rev Immunol* 2001, **19:**623-655.
22. Jenner RG, Young RA: **Insights into host responses against pathogens from transcriptional profiling.** *Nat Rev Microbiol* 2005, **3:**281-294.
23. Fitzgerald KA, McWhirter SM, Faia KL, Rowe DC, Latz E, Golenbock DT, Coyle AJ, Liao SM, Maniatis T: **IKKepsilon and TBK1 are essential components of the IRF3 signaling pathway.** *Nat Immunol* 2003, **4:**491-496.
24. Shuai K, Liu B: **Regulation of JAK-STAT signalling in the immune system.** *Nat Rev Immunol* 2003, **3:**900-911.
25. Kato H, Takeuchi O, Sato S, Yoneyama M, Yamamoto M, Matsui K, Uematsu S, Jung A, Kawai T, Ishii KJ, Yamaguchi O, Otsu K, Tsujimura T, Koh CS, Reis e Sousa C, Matsuura Y, Fujita T, Akira S: **Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses.** *Nature* 2006, **441:**101-105.
26. Meylan E, Curran J, Hofmann K, Moradpour D, Binder M, Bartenschlager R, Tschopp J: **Cardif is an adaptor protein in the RIG-I antiviral pathway and is targeted by hepatitis C virus.** *Nature* 2005, **437:**1167-1172.
27. Gitlin L, Barchet W, Gilfillan S, Cella M, Beutler B, Flavell RA, Diamond MS, Colonna M: **Essential role of mda-5 in type I IFN responses to polyriboinosinic:polyribocytidylic acid and encephalomyocarditis picornavirus.** *Proc Natl Acad Sci USA* 2006, **103:**8459-8464.
28. Yoneyama M, Kikuchi M, Matsumoto K, Imaizumi T, Miyagishi M, Taira K, Foy E, Loo YM, Gale M Jr, Akira S, Yonehara S, Kato A, Fujita T: **Shared and unique functions of the DExD/H-box helicases RIG-I, MDA5, and LGP2 in antiviral innate immunity.** *J Immunol* 2005, **175:**2851-2858.
29. Zhu M, John S, Berg M, Leonard WJ: **Functional association of Nmi with Stat5 and Stat1 in IL-2- and IFNgamma-mediated signaling.** *Cell* 1999, **96:**121-130.
30. Ruefli-Brasse AA, Lee WP, Hurst S, Dixit VM: **Rip2 participates in Bcl10 signaling and T-cell receptor-mediated NF-kappaB activation.** *J Biol Chem* 2004, **279:**1570-1574.
31. Lin X, Cunningham ET Jr, Mu Y, Geleziunas R, Greene WC: **The proto-oncogene Cot kinase participates in CD3/CD28 induction of NF-kappaB acting through the NF-kappaB-inducing kinase and IkappaB kinases.** *Immunity* 1999, **10:**271-280.
32. Mauro C, Pacifico F, Lavorgna A, Mellone S, Iannetti A, Acquaviva R, Formisano S, Vito P, Leonardi A: **ABIN-1 binds to NEMO/IKKgamma and co-operates with A20 in inhibiting NF-kappaB.** *J Biol Chem* 2006, **281:**18482-18488.
33. Jahnke A, Johnson JP: **Synergistic activation of intercellular adhesion molecule 1 (ICAM-1) by TNF-alpha and IFN-gamma is mediated by p65/p50 and p65/c-Rel and interferon-responsive factor Stat1 alpha (p91) that can be activated by both IFN-gamma and IFN-alpha.** *FEBS Lett* 1994, **354:**220-226.
34. Ohmori Y, Hamilton TA: **The interferon-stimulated response element and a kappa B site mediate synergistic induction of murine IP-10 gene transcription by IFN-gamma and TNF-alpha.** *J Immunol* 1995, **154:**5235-5244.
35. Ohmori Y, Schreiber RD, Hamilton TA: **Synergy between interferon-gamma and tumor necrosis factor-alpha in transcriptional activation is mediated by cooperation between signal transducer and activator of transcription 1 and nuclear factor kappaB.** *J Biol Chem* 1997, **272:**14899-14907.
36. **Biomart** [http://www.biomart.org/]
37. **GO** [http://www.geneontology.org]
38. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31:**374-378.
39. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, Down T, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz HR, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, *et al.*: **An overview of Ensembl.** *Genome Res* 2004, **14:**925-928.
40. **SPIKE** [http://www.cs.tau.ac.il/~spike/]
41. Ohmori Y, Hamilton TA: **Cooperative interaction between interferon (IFN) stimulus response element and kappa B sequence motifs controls IFN gamma- and lipopolysaccharide-stimulated transcription from the murine IP-10 promoter.** *J Biol Chem* 1993, **268:**6677-6688.
42. Bren GD, Solan NJ, Miyoshi H, Pennington KN, Pobst LJ, Paya CV: **Transcription of the RelB gene is regulated by NF-kappaB.** *Oncogene* 2001, **20:**7722-7733.
43. Haskill S, Beg AA, Tompkins SM, Morris JS, Yurochko AD, Sampson-Johannes A, Mondal K, Ralph P, Baldwin AS Jr: **Characterization of an immediate-early gene induced in adherent monocytes that encodes I kappa B-like activity.** *Cell* 1991, **65:**1281-1289.
44. Lombardi L, Ciana P, Cappellini C, Trecca D, Guerrini L, Migliazza A, Maiolo AT, Neri A: **Structural and functional characterization of the promoter regions of the NFKB2 gene.** *Nucleic Acids Res* 1995, **23:**2328-2336.
45. Zhang Y, Pasparakis M, Kollias G, Simons M: **Myocyte-dependent regulation of endothelial cell syndecan-4 expression. Role of TNF-alpha.** *J Biol Chem* 1999, **274:**14786-14790.
46. Lopez-Cabrera M, Munoz E, Blazquez MV, Ursa MA, Santis AG, Sanchez-Madrid F: **Transcriptional regulation of the gene encoding the human C-type lectin leukocyte receptor AIM/CD69 and functional characterization of its tumor necrosis factor-alpha-responsive elements.** *J Biol Chem* 1995, **270:**21545-21551.
47. Stehlik C, de Martin R, Binder BR, Lipp J: **Cytokine induced expression of porcine inhibitor of apoptosis protein (iap) family member is regulated by NF-kappa B.** *Biochem Biophys Res Commun* 1998, **243:**827-832.

48. Theofilopoulos AN, Baccala R, Beutler B, Kono DH: **Type I interferons (alpha/beta) in immunity and autoimmunity.** *Annu Rev Immunol* 2005, **23:**307-336.
49. Bottrel RL, Yang YL, Levy DE, Tomai M, Reis LF: **The immune response modifier imiquimod requires STAT-1 for induction of interferon, interferon-stimulated genes, and interleukin-6.** *Antimicrob Agents Chemother* 1999, **43:**856-861.
50. Mashimo T, Glaser P, Lucas M, Simon-Chazottes D, Ceccaldi PE, Montagutelli X, Despres P, Guenet JL: **Structural and functional genomics and evolutionary relationships in the cluster of genes encoding murine 2',5'-oligoadenylate synthetases.** *Genomics* 2003, **82:**537-552.
51. Huang Y, Krein PM, Winston BW: **Characterization of IFN-gamma regulation of the complement factor B gene in macrophages.** *Eur J Immunol* 2001, **31:**3676-3686.
52. Rhodes DA, Ihrke G, Reinicke AT, Malcherek G, Towey M, Isenberg DA, Trowsdale J: **The 52 000 MW Ro/SS-A autoantigen in Sjogren's syndrome/systemic lupus erythematosus (Ro52) is an interferon-gamma inducible tripartite motif protein associated with membrane proximal structures.** *Immunology* 2002, **106:**246-256.

6. Faster pattern matching with character classes using prime number encoding

Journal of Computer and System Sciences 75 (2009) 155–162

Contents lists available at ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Faster pattern matching with character classes using prime number encoding

Chaim Linhart, Ron Shamir *

School of Computer Science, Tel Aviv University, Tel-Aviv 69978, Israel

ARTICLE INFO

Article history:
Received 30 September 2007
Received in revised form 7 August 2008
Available online 22 August 2008

Keywords:
Pattern matching with character classes
Subset matching
Hamming distance
FFT-based pattern matching

ABSTRACT

In *pattern matching with character classes* the goal is to find all occurrences of a pattern of length m in a text of length n, where each pattern position consists of an allowed set of characters from a finite alphabet Σ. We present an FFT-based algorithm that uses a novel prime-numbers encoding scheme, which is $\log n/\log m$ times faster than the fastest extant approaches, which are based on boolean convolutions. In particular, if $m^{|\Sigma|} = n^{O(1)}$, our algorithm runs in time $O(n\log m)$, matching the complexity of the fastest techniques for wildcard matching, a special case of our problem. A major advantage of our algorithm is that it allows a tradeoff between the running time and the RAM word size. Our algorithm also speeds up solutions to approximate matching with character classes problems—namely, matching with k mismatches and Hamming distance, as well as to the *subset matching* problem.

1. Introduction

Generic pattern matching problems require finding all occurrences of a pattern p in a text t. Throughout this paper, we denote by m and n the length of the pattern and the text, respectively ($m < n$). In the classical string matching problem both p and t are strings over a finite alphabet $\Sigma = \{a_1, \ldots, a_\sigma\}$ of size σ. A myriad of efficient algorithms have been developed over the years, the fastest of which solve this problem in linear time, such as the Knuth–Morris–Pratt [1] and Boyer–Moore [2] algorithms.

1.1. Matching with don't-cares

A more general matching problem is obtained when we allow the pattern and the text to contain don't-care characters, or wildcards, denoted '$*$,' which match all symbols in Σ. Formally:

Matching with don't-cares. Given a pattern p and a text t, which may contain don't-cares, find all occurrences of p in t. Here, p is said to occur at location i in t if: $\forall 1 \leqslant j \leqslant m$, $p[j] = t[i+j-1]$ or $p[j] =$ '$*$' or $t[i+j-1] =$ '$*$.'

If the number of don't-cares in the pattern is very small, the problem can be solved in linear time, for example by building a deterministic finite automaton (DFA) that detects all possible words that match the pattern. Another approach is to use the *match-count* algorithm, which finds the number of matching positions (or, equivalently, the Hamming distance) between the pattern and every length m substring of the text (see, e.g., [3, Ch. 4.3]). The algorithm, first introduced by

* Corresponding author.
E-mail addresses: chaiml@post.tau.ac.il (C. Linhart), rshamir@post.tau.ac.il (R. Shamir).

0022-0000/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcss.2008.08.005

Table 1
Summary of the main results described in this paper

Problem	Previous complexity	Prime-code complexity
Matching with character classes	$O(\sigma n \log m)$ [4]	$O(\sigma^{1-\kappa} n \log m)$ if $m^\sigma = n^{\omega(1)}$ $O(n \log(\frac{m}{\kappa}))$ if $m^\sigma = n^{o(1)}$
Hamming distance with char. classes	$O(\sigma n \log m)$ [4]	$O(n\sigma(1 + \log^2 m/\log n))$ if $m^\sigma = n^{\omega(1)}$ $O(n(\log m + \sigma))$ if $m^\sigma = n^{o(1)}$
Subset matching (Monte Carlo)	$O(\sigma n \log(\sigma n))$ [5]	$O(\sigma n \log m)$ if $m^\sigma = n^{\omega(1)}$ $O(n \log n \log(\frac{m}{\kappa})/\log m)$ if $m^\sigma = n^{o(1)}$

Fischer and Paterson [4], computes the contribution of each alphabet symbol to the score independently, as follows. For the symbol $a \in \Sigma$, each occurrence of a in the text and in the pattern is replaced by the number 1, and all other symbols are encoded by 0. The number of matching a's between the pattern and every substring in the text is obtained by computing the convolution between the binary-encoded pattern and text. Using Fast Fourier Transform (FFT), the convolution can be computed in $O(n \log m)$ time under the RAM model of computation, which assumes that arithmetic operations on numbers with w bits take constant time, where $w = O(\log N)$ is the RAM word size and N is the maximal input size. Thus, the total running time of match-count is $O(\sigma n \log m)$, as it involves σ such convolutions. The algorithm can easily be extended to cope with wildcards in the pattern and in the text.

Fischer and Paterson further showed that a similar technique can be applied to solve matching with don't-cares in time $O(\log \sigma \cdot n \log m)$ [4]. Removing the dependence on σ remained an open problem until recently. Indyk introduced a randomized technique for computing boolean products, which yielded an $O(n \log m)$-time Monte Carlo algorithm for wildcard matching and other problems [5]. Kalai gave another elegant Monte Carlo algorithm with the same time complexity, based on integer codes [6]. Cole and Hariharan were the first to obtain an $O(n \log m)$-time deterministic algorithm, by encoding each symbol with a pair of rational numbers [7]. A simpler deterministic algorithm with the same time complexity was presented very recently by Clifford and Clifford [8]. All of the above algorithms compute convolutions using FFT; the main differences between them is in the way they encode the pattern and the text.

1.2. Matching with character classes

Matching with don't-cares can be generalized by allowing the pattern to contain any non-empty subset, or *class*, of characters at each position:

Matching with character classes. Given a pattern p with character classes ($p[j] \subseteq \Sigma$), and a text t, which may contain don't-cares ($t[i] \in \Sigma \cup$ '*'), find all occurrences of p in t. Here, p is said to occur at location i in t if:
$\forall 1 \leqslant j \leqslant m$, $t[i + j - 1] \in p[j]$ or $t[i + j - 1] =$ '*.'

For example, the pattern `a[abcd]r[ab]` matches the text `abracadadrb` at locations 1 (the substring "`abra`") and 8 ("`adrb`"). W.l.o.g., we may assume that the text does not contain don't-cares—otherwise, we can add the don't-care symbol to all the character classes in the pattern, and treat it as a regular symbol in the alphabet. Matching with character classes, as well as with similar types of patterns, has been studied extensively (e.g., [9,10]). Most algorithms, however, have the same worst-case running time as the naïve algorithm—$O(nm)$. Bit-parallelism techniques improve this to $O(nm/w)$, where w is the RAM word size [11]. In general, the best worst-case performance is attained by the match-count algorithm—$O(\sigma n \log m)$.

We present an FFT-based algorithm, whose running time depends on the parameter $\kappa = \log_\sigma(\log n/\log m)$. If $\kappa < 1$, its time complexity is $O(\sigma^{1-\kappa} n \log m)$, which is $\log n/\log m$ times faster than match-count; if $\kappa = 1$, i.e., $m^\sigma = n$, our algorithm computes a single convolution, matching the $O(n \log m)$ running time of the fastest wildcard matching algorithms; and when $\kappa > 1$, our method runs in time $O(n \log(m/\kappa))$, asymptotically converging to the optimal linear-time performance of classical string matching methods as κ approaches infinity. Notably, in the latter case we obtain an improvement for wildcard matching. Our algorithm uses a novel encoding scheme that is based on large prime numbers. The basic idea is to encode the text and the pattern in such a way that at match locations their convolution is congruent to 0 modulo some large number M. Unlike other methods, prime code exploits the entire RAM word, admitting improved performance for a longer word size. Table 1 summarizes the main results presented in this paper.

1.3. Approximate matching with character classes

In many practical scenarios, one would like to find not only perfect matches, but also locations at which the pattern approximately matches the text, that is, matches it up to a specified small distance. A commonly used metric is the number of mismatched pattern positions, or Hamming distance. Returning to the previous example, the pattern `a[abcd]r[ab]` matches the text `abracadadrb` at location 4 (the substring "`acad`") with two mismatches (positions 3 and 4). Finding all pattern occurrences with at most k mismatches is often referred to as the *k-Mismatches problem*. The fastest algorithm for string matching with k mismatches runs in time $O(n\sqrt{k \log k})$ [12]. If the pattern contains character classes, the most efficient algorithm is match-count, running in time $O(\sigma n \log m)$. In fact, match-count provides more information than just the k-mismatches—as explained earlier, it computes the number of mismatches at every text location. We call this the *Hamming distance problem*. In the restricted case of a pattern that contains only single symbols and don't-cares, but not other classes of characters, Abrahamson [10] showed that the Hamming distance can be computed in time $O(n\sqrt{m \log m})$, which is faster than match-count when $\sigma > \sqrt{m/\log m}$. The algorithm we developed for matching with character classes can also solve the k-mismatches and Hamming distance problems with small additional cost. This is an asymptotic improvement over the match-count algorithm.

1.4. Subset matching

In the subset matching problem, both the pattern and the text are composed of character classes. According to the original definition by Cole and Hariharan [13], the pattern matches the text if every character class in the pattern is a subset of the corresponding character class in the text. For consistency with the definition of matching with character classes, we shall switch the roles of the pattern and the text, and obtain the following equivalent problem:

Subset matching. Given a pattern p and a text t, both consisting of character classes ($p[j], t[i] \subseteq \Sigma$), find all occurrences of p in t. Here, p is said to occur at location i in t if: $\forall 1 \leqslant j \leqslant m$, $t[i+j-1] \subseteq p[j]$.

Obviously, matching with character classes is a special case of subset matching. Let s be the total number of characters in the pattern and text, i.e., $s = \sum_{j=1}^{m} |p[j]| + \sum_{i=1}^{n} |t[i]|$. The most efficient algorithms for subset matching are an $O(s \log s)$ Monte Carlo algorithm due to Indyk [5], and an $O(s \log^2 s)$ deterministic algorithm by Cole and Hariharan [7]. Since s might be as large as $\sigma(n+m)$, the above methods have worst-case running times of $O(\sigma n \log(\sigma n))$ and $O(\sigma n \log^2(\sigma n))$, respectively. We develop a randomized variant of our prime-code technique that yields a Monte Carlo algorithm for subset matching. Assuming $\sigma = O(m)$, the algorithm runs in time $O(\sigma n \log m)$ if $\kappa < 1$, $O(n \log n)$ if $\kappa = 1$, and $O(n \log n \log(m/\kappa)/\log m)$ if $\kappa > 1$.

1.5. Motivation

Character classes are commonly used in regular expressions and in many applications of pattern matching in various fields. We briefly describe here two such applications in computational biology. The alphabet in both applications consists of the four DNA bases—$\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.

Transcription factors (TFs) are specialized proteins that bind to regulatory regions in the DNA and control gene expression. A TF usually binds to many different DNA segments that share a common pattern, or *motif*, characteristic of the TF. Such binding-site motifs are often modeled using patterns with character classes. For example, *p53*, the most frequently mutated tumor suppressor in human cancers, binds to two repeats of `[AG][AG][AG]C[AT][AT]G[CT][CT][CT]` [14]. In a typical setting, the TF binding pattern is short ($10 \leqslant m \leqslant 20$), and the goal is to efficiently locate all its occurrences in many long regulatory regions ($n \approx 10^8$).

The second application is in the design of degenerate primers for Polymerase Chain Reaction (PCR) experiments. PCR is a technique for amplifying a specific region of DNA, so that enough copies of it are available for testing or sequencing. The first step in PCR is to synthesize two DNA segments, or *primers*, lying on opposite sides of the target region. A PCR primer is called degenerate if some of its positions have several possible bases. Thus, a degenerate primer can be described as a pattern with character classes. Degenerate primers can be used to amplify several related genomic sequences in a single PCR experiment. We studied the computational problem of designing highly degenerate primers [15], and applied our algorithms in experiments for studying the human and canine olfactory receptor genes [16,17]. A common problem in the design of degenerate primers is to verify that the primers do not bind to DNA regions others than those they are meant to amplify. Thus, one needs to search for all occurrences of a candidate primer, typically of length $20 \leqslant m \leqslant 30$, in the entire genome ($n \approx 6 \cdot 10^9$ in human). One may also want to allow a small number of mismatches (e.g., $k = 3$), as the PCR technique usually tolerates a few mismatches.

Subset matching is applicable in many pattern matching scenarios, such as geometric pattern matching and general pattern matching. Most notably, Cole and Hariharan showed that tree pattern matching, an important problem which has been studied extensively, can be reduced to subset matching in linear time [13]. In computational biology, subset matching can be applied to search for conserved TF binding sites in aligned sequences of multiple species. The straightforward solution is to search for the occurrences of the TF's motif in each species separately, as described earlier, and then check which locations match the motif in all species. An alternative approach is to combine the sequences into a single consensus sequence, in which each position contains the set of bases that appear in that position in one or more species, and apply subset matching to search for the motif in the consensus sequence.

2. Preliminaries

All the algorithms described in this paper assume the RAM model, wherein standard arithmetic on w bit numbers is performed in constant time. Following standard practice, we shall assume that the word size is $w = O(\log n)$ (see, e.g., [5,6]).

Convolution. The convolution, or cross-correlation, of two vectors a, b is the vector $a \oplus b$ such that $(a \oplus b)[i] = \sum_{j=1}^{|a|} a[j]b[i+j-1]$ for $1 \leqslant i \leqslant |b| - |a| + 1$.

Given a pattern p of length m and a text t of length n ($m < n$), both encoded using numbers with w bits, the convolution $p \oplus t$ can be computed in $O(n \log m)$ time, as follows. First, the text is split into n/m pieces of length $2m$, with overlap m between consecutive pieces. The convolution between the pattern and each piece of the text is then computed using FFT in time $O(m \log m)$ per piece (as in [4]).

3. Matching with character classes

In this section we describe our encoding scheme and how it can be applied to solve pattern matching with character classes. Since our algorithm is based on computing convolutions on segments of length $2m$, we may assume w.l.o.g. that $\sigma \leqslant 2m$, as each $2m$-long piece of text contains at most $2m$ distinct symbols.

3.1. Prime code

A prime code assigns to each symbol $a_i \in \Sigma$ a distinct prime number p_i, where $p_1 < p_2 < \cdots < p_\sigma$ (notice that we use p_i to denote the ith prime number, whereas $p[i]$ is the character class at position i in the pattern). Denote $M = p_1 \cdot \ldots \cdot p_\sigma$. We further require that all primes are larger than m (i.e., $p_1 > m$). We first describe how such prime numbers can be found, and then explain how to encode the pattern and the text.

Finding Prime Numbers $p_1, \ldots, p_\sigma > m$: Following are well known bounds on the number $\pi(x)$ of primes less than or equal to x [18]:

$$\forall x \geqslant 17, \quad \frac{x}{\ln x} < \pi(x) < 1.26 \frac{x}{\ln x}.$$

Using these bounds, for $m \geqslant 17$ we get:

$$\pi(5m \ln m) - \pi(m) > \frac{5m \ln m}{\ln(5m \ln m)} - 1.26 \frac{m}{\ln m} > \frac{5m \ln m - 2.6m}{2 \ln m} > 2m \geqslant \sigma.$$

Thus, if we search for prime numbers between $m+1$ and $5m \ln m$, we are guaranteed to find at least σ prime numbers, as required. Since testing for primality takes polynomial time (i.e., testing whether x is prime takes polylog(x) time) [19], the prime numbers we seek can be found in $O(m \cdot \text{polylog}(m))$ time. Alternatively, we could apply Eratosthenes' sieve to obtain all the primes up to $5m \ln m$ in time $O(m \log m \log \log m)$. This can be improved to $o(m \log m)$ time using modern sieves, such as the sieve of Atkin [20]. Notice that each of the primes we obtain is a number with at most $\log_2(5m \ln m) < 2 \log_2 m$ bits. The prime numbers depend only on m—if we are given a list of equal-length patterns, this step of the algorithm needs to be performed only once.

Text Code: The symbol a_i in the text is encoded by the integer M/p_i.

Pattern Code: A character class $[a_{i_1}, \ldots, a_{i_c}]$ in the pattern is encoded by an integer $n_\mathcal{S}$, where $\mathcal{S} = \{i_1, \ldots, i_c\}$, s.t.:

$$n_\mathcal{S} \equiv \begin{cases} 0 \pmod{p_i} & \forall i \in \mathcal{S}, \\ 1 \pmod{p_j} & \forall j \notin \mathcal{S}. \end{cases} \tag{1}$$

The Chinese Remainder Theorem (CRT, in short) guarantees that such integers exist (see, e.g., [21, Ch. 31.5]). In fact, for each subset $\mathcal{S} \subseteq \{1, \ldots, \sigma\}$ there exists a single integer $0 \leqslant n_\mathcal{S} < M$, for which Eq. (1) holds. Moreover, this integer can be found using the CRT, as follows. For each j ($1 \leqslant j \leqslant \sigma$), we obtain a pair of integers r_j, q_j s.t.: $r_j p_j + q_j(M/p_j) = 1$. These integers can be computed using Euclid's gcd algorithm in time $O(\log p_j)$. Denoting $c_j = q_j(M/p_j)$, it follows from the CRT that:

$$n_\mathcal{S} = \sum_{j \notin \mathcal{S}} c_j = 1 - \sum_{j \in \mathcal{S}} c_j \bmod M.$$

Hence, given the coefficients $c_1, \ldots, c_\sigma$, a character class with c symbols is encoded in linear time. Let s_p denote the total number of characters in the pattern, i.e., $s_p = \sum_{j=1}^{m} |p[j]|$. Encoding the entire pattern takes $O(s_p)$ time, plus $O(\sigma \log p_\sigma) = O(m \log m)$ (since $\sigma \leqslant 2m$ and $p_\sigma \leqslant 5m \ln m$) for computing the c_j's, which can be done in pre-processing, as they do not depend on the content of the pattern. The attentive reader may have noticed that we ignored a crucial problem—the numbers we are dealing with might be too large to fit into a single RAM word. We will address this issue in the next section.

3.2. The PMCC algorithm

Our basic algorithm for pattern matching with character classes, called PMCC, is outlined in Fig. 1.

```
INPUT: Pattern $p$ with character classes, text $t$ over alphabet $\Sigma$, $\sigma = |\Sigma|$
OUTPUT: All occurrences of $p$ in $t$
Algorithm PMCC:
1. Pre-processing:
    1a. Find $\sigma$ prime numbers—$p_1, \ldots, p_\sigma > m$
    1b. Set $M \leftarrow p_1 \cdot \ldots \cdot p_\sigma$
    1c. Compute coefficients $c_1, \ldots, c_\sigma$ using the CRT:
        $c_i \equiv 1 \pmod{p_i}$ and $c_i \equiv 0 \pmod{p_j}$ for $j \neq i$
2. Encode the pattern and the text using a prime code:
    2a. Text: Replace the symbol $a_i$ by $M/p_i$
    2b. Pattern: Replace the character class $[a_{i_1}, \ldots, a_{i_c}]$ by $n_{\{i_1,\ldots,i_c\}}$:
        $n_{\{i_1,\ldots,i_c\}} = 1 - (c_{i_1} + \cdots + c_{i_c}) \bmod M$
3. Compute the convolution $p \oplus t$ using FFT
4. Report a match at location $i$ iff $(p \oplus t)[i] \equiv 0 \pmod{M}$
```

Fig. 1. Algorithm PMCC for pattern matching with character classes.

Theorem 1. *If $m^\sigma = n^{O(1)}$, algorithm PMCC solves pattern matching with character classes using a single convolution in time $O(n \log m)$.*

Proof. We first prove that PMCC produces the correct output. Let us compare the pattern to the substring at location i in the text. The value of the convolution at this location is: $(p \oplus t)[i] = \sum_{j=1}^{m} p[j]t[i+j-1]$. Let $[a_{i_1}, \ldots, a_{i_c}]$ be the character class at position j in the pattern, and let a_k be the symbol at position $i+j-1$ in the text. It is easily seen that:

$$p[j]t[i+j-1] = n_{\{i_1,\ldots,i_c\}} \cdot M/p_k \equiv \begin{cases} 0, & \text{if } a_k \in \{a_{i_1}, \ldots, a_{i_c}\}, \\ M/p_k, & \text{otherwise} \end{cases}$$

(all congruences are modulo M). Denote by e_k the number of times the symbol a_k in the text does not match the corresponding character class in the pattern, when the pattern is aligned against text location i. Thus, $(p \oplus t)[i] \equiv R$, where $R = \sum_{k=1}^{\sigma} e_k \cdot (M/p_k)$. Since $p_k > m$ for all k, we get $R < M/m \cdot \sum_{k=1}^{\sigma} e_k$. Obviously, $\sum_{k=1}^{\sigma} e_k \leqslant m$, so $R < M$. Of course, $R \geqslant 0$, and this inequality strictly holds iff $\exists k$, $e_k > 0$. The correctness of the algorithm immediately follows.

We now analyze the running time of PMCC. As explained in Section 3.1, step 1 can be performed in time $O(m \log m)$, and encoding the text and the pattern in step 2 takes $O(n + s_p)$ time. Step 4 takes $O(n)$ time. Henceforth we shall ignore these pre-processing and linear-time phases of the algorithm, and focus on step 3, which determines the overall time complexity. Since we showed that $\log_2 p_i < 2 \log_2 m$ (Section 3.1), it follows that $\log_2 M = \sum_{k=1}^{\sigma} \log_2 p_k < 2\sigma \log_2 m$. Thus, for $m^\sigma = n^{O(1)}$, we get $\log_2 M = O(\log n)$, i.e., M fits into a single machine word, so the convolution in step 3 can be computed in time $O(n \log m)$, as required. □

We now show how to adjust the PMCC algorithm so that it could solve instances with $m^\sigma = n^{\omega(1)}$, and how to improve its performance when $m^\sigma = n^{o(1)}$.

Theorem 2. *Pattern matching with character classes can be solved in time*:

$$\begin{cases} O(\sigma^{1-\kappa} n \log m), & \text{if } 0 \leqslant \kappa \leqslant 1, \\ O(n \log m), & \text{if } \kappa = 1, \\ O(n \log(m/\kappa)), & \text{if } \kappa \geqslant 1 \end{cases}$$

where $\kappa = \log_\sigma(\log n / \log m)$, or, more generally, $\kappa = \log_\sigma(w / \log m)$, where w is the RAM word size.

Proof. *The case $m^\sigma = n^{\omega(1)}$*: In order to ensure that all the numbers the algorithm computes do not exceed $O(\log n)$ bits, we apply a standard trick—we partition Σ into smaller alphabets, $\Sigma = \bigcup \Sigma_j$, each of size at most $\log n / \log m$. For each of these $\lceil (\log m / \log n)\sigma \rceil$ alphabets, we solve the problem using PMCC, ignoring all symbols that are not in the active alphabet—such symbols in the text are replaced by the symbol '$*$,' which is also added to all the character classes in the pattern, as well as to the alphabet. Finally, we report a match at each location for which a match was found over all the alphabets. Denoting $\kappa = \log_\sigma(\log n / \log m)$, the total running time is $O(\sigma \log m / \log n \cdot n \log m) = O(\sigma^{1-\kappa} n \log m)$.

The case $m^\sigma = n^{o(1)}$: In this case, PMCC utilizes only a small part of the RAM word. We can improve its running time by avoiding this waste, as follows. The main idea is to work on κ-tuples. We first rename the pattern using the alphabet Σ^κ, padding the pattern with character classes that consist of the entire alphabet, if required. The new pattern is a pattern with character classes over the new alphabet (notice that this trick does not work for a pattern with don't-cares—renaming it with Σ^κ results in a pattern with character classes, not only single symbols and don't-cares). Next, we rename the text using Σ^κ, each time starting at a different offset $0 \leqslant i < \kappa$. For each offset, we run PMCC and report the matches. Since the text and the pattern are of length n/κ and m/κ, respectively, and all the numbers involved fit into a RAM word ($\log_2 M < 2\sigma^\kappa \log_2(m/\kappa) = O(\log n)$), each offset is handled in time $O(n/\kappa \cdot \log(m/\kappa))$. The total time complexity is therefore $O(n \log(m/\kappa))$. □

3.3. Approximate matching

In this section we describe simple post-processing procedures that can be applied to our PMCC algorithm in order to solve two approximate matching with character classes problems—Hamming distance and matching with k mismatches.

3.3.1. Hamming distance

Interestingly, the prime-code convolution vector $p \oplus t$ contains more information than merely the locations of the matches. As we shall now show, the number of mismatches at every location in the text can easily be derived from it, thus computing the Hamming distance for patterns with character classes more efficiently than match-count.

Theorem 3. *The Hamming distance for patterns with character classes can be computed in time*:

$$\begin{cases} O\big(n\big(\sigma^{1-\kappa}\log m + \sigma\big)\big) = O\big(n\sigma\big(1+\log^2 m/\log n\big)\big), & \textit{if } 0 \leqslant \kappa < 1, \\ O\big(n(\log m + \sigma)\big), & \textit{if } \kappa \geqslant 1. \end{cases}$$

Proof. Recall that for a fixed location i in the text: $(p \oplus t)[i] \equiv R \pmod M$, where $R = \sum_{k=1}^{\sigma} e_k(M/p_k)$, and e_k is the number of mismatches for the symbol a_k in the text. Since $R \equiv e_k(M/p_k) \pmod{p_k}$, we get: $e_k = R \cdot (M/p_k)^{-1} \bmod p_k$. (Notice that the modular inverse $(M/p_k)^{-1}$ is the integer q_k we computed earlier for encoding the pattern.) As described in Section 3.2, if $m^\sigma = n^{O(1)}$ all the numbers computed by our matching algorithm fit into a machine word, so we can compute e_k from $(p \oplus t)[i]$ in constant time. If $m^\sigma = n^{\omega(1)}$, the algorithm performs a separate convolution for each partial alphabet Σ_j; we thus calculate e_k from the convolution vector we computed for the partial alphabet that contains a_k. Therefore, in both cases the Hamming distance $\sum_k e_k$ at every text location can be calculated in total time $O(\sigma n)$, given the convolution vector(s). In fact, we can compute a weighted Hamming distance—$\sum_k w_k e_k$, where each mismatched text symbol is assigned a pre-defined weight w_k. □

3.3.2. k-Mismatches

We would now like to find all text locations at which there are at most k mismatches $(1 \leqslant k < m)$. Obviously, we could compute the Hamming distance and solve the problem in $O(\sigma n)$ time. However, when k is small, as is often the case in practice, there is a more efficient alternative.

Theorem 4. *Matching with k mismatches for patterns with character classes can be solved in time*:

$$\left.\begin{cases} O\big(\sigma^{1-\kappa} n \log m\big), & \textit{if } 0 \leqslant \kappa < 1, \\ O(n \log m), & \textit{if } \kappa \geqslant 1 \end{cases}\right\} + O\big(n \cdot \min\big\{\sigma, k\big(1 + \log \tfrac{\sigma}{k}\big)\big\}\big).$$

Proof. The idea is to identify which symbols have mismatches when the pattern is compared to a substring in the text. To this end, for every text location we perform a binary search using boolean queries on the value of R modulo subsets of the prime numbers, as follows.

Suppose $k < \sigma$, and let $\mathcal{T}$ be a balanced binary tree, whose leaves, ordered from left to right, are the alphabet symbols $a_1, \ldots, a_\sigma$. Each node in $\mathcal{T}$ corresponds to a subset of Σ, comprised of the symbols at the leaves of its subtree. An example is illustrated in Fig. 2. We further assume that $(p \oplus t)[i]$ fits into a single RAM word, so $m^\sigma = n^{O(1)}$. We start at the root of $\mathcal{T}$, and check $R \bmod p_1 p_2 \ldots p_{\lceil \sigma/2 \rceil}$—if it is 0, then there are no mismatches for the symbols $a_1, \ldots, a_{\lceil \sigma/2 \rceil}$, and we prune the left branch; otherwise, at least one of these symbols has mismatches, so we continue the search in the left subtree of the root. Similarly, if $R \bmod p_{\lceil \sigma/2 \rceil + 1} \ldots p_\sigma \neq 0$, we continue to the right subtree. In this manner we traverse $\mathcal{T}$ breadth-first, pruning some of the branches along the way. Each non-pruned branch contains one or more mismatched symbols (that is, at least one of the leaves in its subtree has mismatches). Thus, if the number of non-pruned branches exceeds k, there are more than k mismatched symbols, which clearly implies that there are more than k mismatches at the current text location, so we stop the search. Otherwise, we end up with at most k leaves that correspond to the mismatched symbols. We calculate the exact number of mismatches for each of these symbols, as we have done for the Hamming distance computation, and report a match if their sum does not exceed k. In the example illustrated in Fig. 2 there are three mismatched symbols—a_1, a_2 and a_4, and a total of four mismatches. The total time required for the above search is proportional to the number of branches we traverse, which is $O(k + k \log \frac{\sigma}{k})$.

If $m^\sigma = n^{\omega(1)}$, we cannot perform the above search, since we do not have the value $(p \oplus t)[i]$. Instead, the alphabet is partitioned into partial alphabets—$\Sigma_1, \ldots,$ of size at most $\log n/\log m$, and the matching algorithm computes $\lceil (\log m/\log n)\sigma \rceil$ convolutions, one for each partial alphabet. Thus, we can implement the same breadth-first binary search as for the case $m^\sigma = n^{O(1)}$, except that rather than starting at the root of $\mathcal{T}$, we begin the search from the level in the tree that contains the nodes that correspond to the partial alphabets $\Sigma_1, \ldots$ (or the next level if $\lceil (\log m/\log n)\sigma \rceil$ is not an integer power of 2). For example, if $(\log m/\log n)\sigma = 2$, the matching algorithm computes two convolutions (one for $\Sigma_1 = \{a_1, \ldots, a_{\sigma/2}\}$, and one for $\Sigma_2 = \{a_{\sigma/2+1}, \ldots, a_\sigma\}$), so we start the breadth-first search at the second level of $\mathcal{T}$ (whose nodes correspond to Σ_1 and Σ_2); in order to check whether R mod $p_1 p_2 \ldots p_{\lceil \sigma/2 \rceil}$ is 0, we use the values $(p \oplus t)[i]$ computed in the first

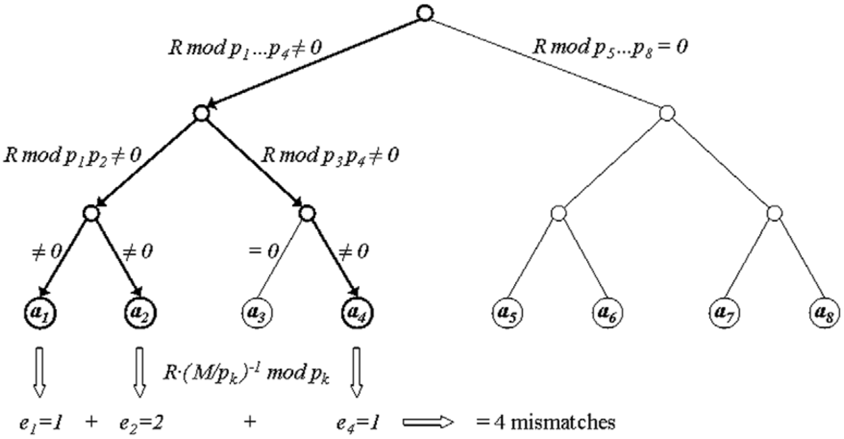


Fig. 2. Illustration of the k-mismatches algorithm for $\sigma = 8$. For every text location, the algorithm traverses the alphabet tree $\mathcal{T}$ breadth-first, continuing the search only in edges with mismatches (thick arrows).

convolution, and for $R \bmod p_{\lceil\sigma/2\rceil+1} \ldots p_\sigma$ we use the output of the second convolution. Searching for the mismatched symbols at a given location takes in this case $O(k + k\log\frac{\sigma}{k} + (\log m/\log n)\sigma) = O(k(1+\log\frac{\sigma}{k}) + \sigma^{1-\kappa})$ time. □

4. Subset matching

In the subset matching problem, both the pattern and the text consist of subsets of Σ (see Section 1.4 for the definition of the problem). Unlike in the previous problem, here a $2m$-long piece of text may contain more than $2m$ different symbols. However, we shall still assume that $\sigma = O(m)$; we shall not analyze the performance of the algorithm for larger alphabets. We now describe a randomized version of the prime code technique for solving subset matching. The difference lies in the way we encode the text; the pattern is encoded as in Section 3.1.

Randomized Text Code: A non-empty character class $[b_{j_1}, \ldots, b_{j_d}]$ in the text is encoded by an integer $r \cdot M/p_{\mathcal{S}_t}$, where $p_{\mathcal{S}_t} = \prod_{k=1}^{d} p_{j_k}$, and r is a random totative[1] of $p_{\mathcal{S}_t}$; in other words:

(i) $1 \leqslant r < p_{\mathcal{S}_t}$,
(ii) r is relatively prime to $p_{\mathcal{S}_t}$,
(iii) r is chosen uniformly among the numbers that fulfill (i) and (ii).

An empty character class in the text is encoded by 0.

Given the primes $p_{j_1}, \ldots, p_{j_d}$, each totative of $p_{\mathcal{S}_t}$ is uniquely characterized by its set of residues $r_{j_1}, \ldots, r_{j_d}$ modulo $p_{j_1}, \ldots, p_{j_d}$, respectively. Therefore, in order to uniformly select a random totative r, we choose random residues, and then compute the corresponding r using the CRT. Similarly to the analysis of the pattern code in Section 3.1, encoding the text takes $O(s_t)$ time, where s_t is the total number of characters in the text, plus $O(m\log m)$ for pre-processing.

4.1. The SSM algorithm

We now give a randomized algorithm, called SSM, for solving subset matching. The algorithm, outlined in Fig. 3, is a variant of PMCC that uses the randomized text code described above.

Theorem 5. *Algorithm SSM is a Monte Carlo algorithm for solving subset matching. If $\sigma = O(m)$ then with probability at least $1 - \frac{1}{n}$ the algorithm reports no false matches in time*:

$$\begin{cases} O(\sigma n \log m), & \text{if } 0 \leqslant \kappa \leqslant 1, \\ O(n \log n), & \text{if } \kappa = 1, \\ O(n \log n \log(m/\kappa)/\log m), & \text{if } \kappa \geqslant 1 \end{cases}$$

where $\kappa = \log_\sigma(w/\log m)$ and w is the RAM word size.

[1] A totative of x is a positive integer smaller than and relatively prime to x.

```
INPUT: Pattern $p$ and text $t$ with character classes, alphabet $\Sigma$, $\sigma = |\Sigma|$
OUTPUT: All occurrences of $p$ in $t$
Algorithm SSM:
1. Pre-processing: Same as in algorithm PMCC (Fig. 1)
2. For $k = 1, \ldots, 2\log_2 n/\log_2 m$:
    3. Encode the pattern and the text using a randomized prime code:
        3a. Text: Replace the character class $[b_{j_1}, \ldots, b_{j_d}]$ by $r \cdot M/p_{\mathcal{S}_t}$:
            $p_{\mathcal{S}_t} = p_{j_1} \cdot \ldots \cdot p_{j_d}$
            $r$ is a random totative of $p_{\mathcal{S}_t}$ (i.e., $1 \leqslant r < p_{\mathcal{S}_t}$, $gcd(r, p_{\mathcal{S}_t}) = 1$)
        3b. Pattern: Replace the character class $[a_{i_1}, \ldots, a_{i_c}]$ by $n_{\{i_1,\ldots,i_c\}}$:
            $n_{\{i_1,\ldots,i_c\}} = 1 - (c_{i_1} + \cdots + c_{i_c}) \bmod M$
    4. Compute the convolution $C_k = (p \oplus t) \bmod M$ using FFT
5. Report a match at location $i$ iff $\forall k\ C_k[i] = 0$
```

Fig. 3. Algorithm SSM for subset matching.

Proof. We first analyze a single iteration k of steps 3 and 4. The convolution C_k at location i in the text is: $(p \oplus t)[i] = \sum_{j=1}^{m} p[j]t[i+j-1]$. Let $\mathcal{S}_p = [a_{i_1}, \ldots, a_{i_c}]$ be the character class at position j in the pattern, and let $\mathcal{S}_t = [b_{j_1}, \ldots, b_{j_d}]$ be the character class at position $i+j-1$ in the text. If $\mathcal{S}_t \subseteq \mathcal{S}_p$, then:

$$p[j]t[i+j-1] = n_{\{i_1,\ldots,i_c\}} \cdot rM/(p_{j_1} \cdot \ldots \cdot p_{j_d}) \equiv 0 \pmod{M}.$$

Thus, if the pattern matches the text at location i, the above holds for all $1 \leqslant j \leqslant m$, and we get $C_k[i] = 0$. Conversely, suppose there is a mismatch at position j in the pattern, and let $a_k \in \mathcal{S}_t - \mathcal{S}_p$. In this case, $n_{\{i_1,\ldots,i_c\}} \equiv 1 \bmod p_k$, and r modulo p_k is a random totative of p_k (since it is a random totative of $p_{\mathcal{S}_t}$), so:

$$p[j]t[i+j-1] \bmod p_k \sim U(1, \ldots, p_k - 1).$$

Fixing the remaining variables, the congruence $(p \oplus t)[i] \equiv 0 \bmod M$ has a unique solution for $p[j]t[i+j-1] \bmod p_k$. Therefore, the probability that the congruence holds is at most $1/(p_k - 1) \leqslant 1/m$. In other words, if the pattern does not match the text at location i, then $C_k[i] = 0$ with probability at most $1/m$.

If we perform $(\log_2 n + c)/\log_2 m$ iterations of steps 2 and 3, each time encoding the text using new random residues, the probability of having any false matches at any position is at most $n/m^{(\log_2 n + c)/\log_2 m} = 1/2^c$. Specifically, for $c = \log_2 n$ the probability for failure is at most $1/n$, and the entire algorithm is $\Theta(\log n/\log m)$ times slower than the deterministic algorithm for pattern matching with character classes (Theorem 2). □

Acknowledgments

We would like to thank Michal Ziv-Ukelson for many helpful discussions, and Ze'ev Rudnick for his help in Number Theory. We also thank an anonymous referee for constructive comments. This study was supported by the Israel Science Foundation (grants 309/02 and 802/08 and Converging Technologies Program Grant 1767.07).

References

[1] D. Knuth, J. Morris, V. Pratt, Fast pattern matching in strings, SIAM J. Comput. 6 (2) (1977) 323–350.
[2] R.S. Boyer, J.S. Moore, A fast string searching algorithm, Commun. ACM 20 (10) (1977) 762–772.
[3] D. Gusfield, Algorithms on Strings, Trees, and Sequences, Cambridge Univ. Press, 1997.
[4] M. Fischer, M. Paterson, String matching and other products, in: R.M. Karp (Ed.), Complexity of Computation, SIAM–AMS Proceedings, 1974, pp. 113–125.
[5] P. Indyk, Faster algorithms for string matching problems: matching the convolution bound, in: Proc. 39th IEEE Annual Symposium on Foundations of Computer Science, 1998, pp. 166–173.
[6] A. Kalai, Efficient pattern-matching with don't cares, in: Proc. 13th Annual ACM–SIAM Symposium on Discrete Algorithms, 2002, pp. 655–656.
[7] R. Cole, R. Hariharan, Verifying candidate matches in sparse and wildcard matching, in: Proc. 34th Symposium on Theory of Computing, 2002, pp. 592–601.
[8] P. Clifford, R. Clifford, Simple deterministic wildcard matching, Inform. Process. Lett. 101 (2) (2007) 53–54.
[9] R. Pinter, Efficient string matching with don't-care patterns, in: A. Apostolico, Z. Galil (Eds.), Combinatorial Algorithms on Words, in: NATO ASI Ser., vol. F12, Springer-Verlag, 1985, pp. 11–29.
[10] K. Abrahamson, Generalized string matching, SIAM J. Comput. 16 (1987) 1039–1051.
[11] R.A. Baeza-Yates, G.H. Gonnet, A new approach to text searching, Commun. ACM 35 (10) (1992) 74–82.
[12] A. Amir, M. Lewenstein, E. Porat, Faster algorithms for string matching with k mismatches, in: Proc. 11th annual ACM–SIAM Symposium on Discrete Algorithms, 2000, pp. 794–803.
[13] R. Cole, R. Hariharan, Tree pattern matching and subset matching in randomized $o(n\log^3 m)$ time, in: Proc. 29th ACM Symposium on Theory of Computing, 1997, pp. 66–75.
[14] J. Hoh, S. Jin, T. Parrado, J. Edington, A. Levine, J. Ott, The p53MH algorithm and its application in detecting p53-responsive genes, Proc. Natl. Acad. Sci. USA 99 (13) (2002) 8467–8472.
[15] C. Linhart, R. Shamir, The degenerate primer design problem: Theory and applications, J. Comput. Biol. 12 (4) (2005) 431–456.
[16] T. Fuchs, B. Malecova, C. Linhart, R. Sharan, M. Khen, R. Herwig, D. Shmulevich, R. Elkon, M. Steinfath, J. O'Brien, U. Radelof, H. Lehrach, D. Lancet, R. Shamir, DEFOG: A practical scheme for deciphering families of genes, Genomics 80 (3) (2002) 295–302.
[17] T. Olender, T. Fuchs, C. Linhart, R. Shamir, M. Adams, F. Kalush, M. Khen, D. Lancet, The canine olfactory subgenome, Genomics 83 (3) (2004) 361–372.
[18] J. Rosser, L. Schoenfield, Approximate formulas for some functions of prime numbers, Illinois J. Math. 6 (1962) 64–94.
[19] M. Agrawal, N. Kayal, N. Saxena, PRIMES is in P, Ann. of Math. 160 (2) (2004) 781–793.
[20] A. Atkin, D. Bernstein, Prime sieves using binary quadratic forms, Math. Comp. 73 (2004) 1023–1030.
[21] T. Cormen, C. Leiserson, R. Rivest, C. Stein, Introduction to Algorithms, 2nd edition, MIT Press, McGraw-Hill, 2001.

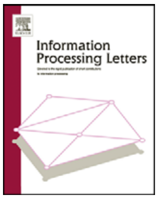
7. Matching with don't-cares and a small number of mismatches

Information Processing Letters 109 (2009) 273–277

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl



Information Processing Letters

Matching with don't-cares and a small number of mismatches

Chaim Linhart *, Ron Shamir

School of Computer Science, Tel Aviv University, Tel-Aviv 69978, Israel

ARTICLE INFO

Article history:
Received 22 May 2008
Received in revised form 11 October 2008
Accepted 30 October 2008
Available online 7 November 2008
Communicated by L.A. Hemaspaandra

Keywords:
Analysis of algorithms
Pattern matching
Approximate wildcard matching

ABSTRACT

In *matching with don't-cares and k mismatches* we are given a pattern of length m and a text of length n, both of which may contain don't-cares (a symbol that matches all symbols), and the goal is to find all locations in the text that match the pattern with at most k mismatches, where k is a parameter. We present new algorithms that solve this problem using a combination of convolutions and a dynamic programming procedure. We give randomized and deterministic solutions that run in time $O(nk^2 \log m)$ and $O(nk^3 \log m)$, respectively, and are faster than the most efficient extant methods for small values of k. Our deterministic algorithm is the first to obtain an $O(\text{poly}(k) \cdot n \log m)$ running time.

1. Introduction

The problem of *pattern matching with don't-cares* requires finding all occurrences of a pattern p of length m in a text t of length n, where the pattern and the text contain don't-cares (or wildcards), often marked as '*', that match all symbols. Fischer and Paterson developed an algorithm for solving this problem that utilizes boolean convolutions, computed using the Fast Fourier Transform (FFT) [1]. Assuming the RAM model, which is the computational model used by most studies on FFT-based pattern matching techniques, its running time is $O(\log|\Sigma| \cdot n \log m)$, where Σ is the alphabet. This time complexity has been improved over the past decade using various FFT-based methods. Cole and Hariharan were the first to obtain an $O(n \log m)$ time deterministic algorithm [2], which was simplified by Clifford and Clifford [3].

In many practical scenarios, one may want to search for approximate matches, that is, locations in the text that match the pattern up to a small pre-specified distance. Perhaps the most widely used metric is the Hamming distance, which counts the number of mismatched pattern symbols. Applications of this variant of approximate matching are very common. For example, in bioinformatics they arise when comparing genes or proteins, and in the context of motif finding and primer design. The Hamming distance between the pattern and the text at every offset can be computed using the *match-count* algorithm, which computes $|\Sigma|$ boolean convolutions in time $O(|\Sigma| n \log m)$ [1]. Abrahamson combined the match-count algorithm with a divide-and-conquer technique to compute the Hamming distance in time $O(n\sqrt{m \log m})$ [4]. Randomized solutions for Hamming distance computation can also be obtained using sketching protocols (e.g., [5,6]).

In this paper we focus on the problem of *matching with k mismatches*. Given a pattern, a text and an integer k, we would like to report all locations in the text that match the pattern with at most k mismatches. This problem has been studied extensively for simple strings (i.e., without don't-cares). Currently, the most efficient method runs in time $O(n\sqrt{k \log k})$ [7]. As in the case of exact matching, searching for approximate matches becomes much more difficult when we allow don't-cares. This variant, which we call *matching with don't-cares and k mismatches*, has received attention only very recently (see details below). Here, we describe new efficient algorithms for matching with don't-cares and k mismatches, which are conceptually simpler, and in some cases faster, than extant techniques.

* Corresponding author.
E-mail addresses: chaiml@post.tau.ac.il (C. Linhart), rshamir@post.tau.ac.il (R. Shamir).

0020-0190/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ipl.2008.10.016

2. Problem definition and preliminaries

Let Σ be a finite alphabet, and denote by '$*$' the don't-care symbol. A text $t = t_1 \ldots t_n$ and a pattern $p = p_1 \ldots p_m$ are strings over $\Sigma \cup$'$*$'. Define $HD(i)$ to be the Hamming distance between p and $t_i \ldots t_{i+m-1}$:

$$HD(i) = \left|\{1 \leqslant j \leqslant m \mid p_j \neq t_{i+j-1} \text{ and } p_j, t_{i+j-1} \neq \text{'}*\text{'}\}\right|.$$

Matching with don't-cares and *k* mismatches. Given a pattern p and a text t with don't-cares, and an integer k, find all occurrences of p in t with at most k mismatches, i.e., report all locations i in t with $HD(i) \leqslant k$.

All the algorithms described in this paper assume the RAM model, wherein standard arithmetic on w-bit numbers are performed in constant time. Following common practice, we shall assume that the word size is $w = O(\log n)$.

Convolution. The convolution of two vectors a, b is the vector $a \oplus b$ such that:

$$(a \oplus b)[i] \overset{\text{def}}{=} \sum_{j=1}^{|a|} a_j b_{i+j-1},$$
$$\text{for } 1 \leqslant i \leqslant |b| - |a| + 1.$$

Given a pattern p of length m and a text t of length n ($m < n$), both encoded using numbers with w bits, the convolution $p \oplus t$ can be computed in $O(n \log m)$ time, as follows. First, the text is split into $\lceil n/m \rceil$ pieces of length $2m$, with overlap m between consecutive pieces. The convolution between the pattern and each piece of the text is then computed using FFT in time $O(m \log m)$ per piece (as in [1]).

3. Related work and previous results

Both match-count and Abrahamson's technique for Hamming distance matching can easily handle don't-cares. Thus, matching with don't-cares and k mismatches can be solved in time $O(|\Sigma| n \log m)$ or $O(n\sqrt{m \log m})$ [1,4]. Intuitively, finding the locations at which the pattern matches the text with at most k mismatches should be easier than computing the exact Hamming distance at all locations. Indeed, Clifford et al. [8] recently developed several faster algorithms for this problem. Their algorithms, as well as the new ones we introduce in this work, extend the elegant technique for wildcard matching reported by Clifford and Clifford [3], which we now describe in brief (note that this technique also appears in [9] in the context of string matching with L_2 distance, and that similar methods based on manipulation of polynomials for solving various pattern matching problems were suggested earlier, e.g., [10,11]).

3.1. Simple matching with don't-cares

The simple algorithm for matching with don't-cares first encodes the pattern and the text, as follows. Each symbol is replaced by a unique positive number, and don't-cares are replaced by 0's. Then, for each location i in the text, the algorithm computes the sum $A_0[i]$:

$$A_0[i] = \sum_{j=1}^{m} x_{i,j}, \tag{1}$$

where:

$$x_{i,j} = p_j t_{i+j-1} (p_j - t_{i+j-1})^2. \tag{2}$$

It is easy to see that $A_0[i] = 0$ if and only if there is an exact match at offset i. The key observation is that this sum can be computed efficiently for all offsets using three FFTs, since:

$$A_0[i] = \sum_{j=1}^{m} p_j^3 t_{i+j-1} - 2\sum_{j=1}^{m} p_j^2 t_{i+j-1}^2 + \sum_{j=1}^{m} p_j t_{i+j-1}^3. \tag{3}$$

For example, the first sum in (3) is a convolution between $p_1^3, \ldots, p_m^3$ and the text $t_1, \ldots, t_n$. Thus, the total running time is $O(n \log m)$ [3].

3.2. Matching with don't-cares and k mismatches

Clifford et al. [8] further developed the above idea and devised an algorithm for solving the 1-mismatch problem—given a pattern and a text that contain don't-cares, the algorithm reports all text locations that match the pattern with at most one mismatch. In short, their algorithm computes (again, with FFTs) an additional array $A_1[i] = \sum_{j=1}^{m} (i + j - 1) x_{i,j}$. If there is a single mismatch at offset i, then the value $B[i] = A_1[i]/A_0[i]$ is the position of the mismatch. Thus, there is one mismatch iff $A_0[i] = x_{i,B[i]-i+1}$, which could easily be verified in constant time per text offset. Clifford et al. used this procedure as a building block for solving the k mismatches with don't-cares problem. They present a randomized algorithm that runs in $O(n(k + \log n \log\log n) \log m)$ time and gives the correct answer with high probability. Their deterministic algorithms, based on tools developed for group testing and for k-selectors, run in time $O(nk^2 \log^3 m)$ and $O(nk \operatorname{polylog} m)$, respectively (the latter with $O(\operatorname{poly} m)$ time preprocessing).

4. Main ideas and results

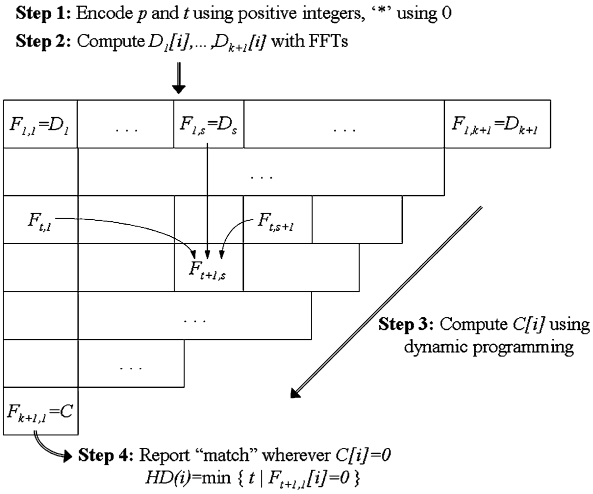
Our approach is based on the fact that at a fixed location i in the text, the number of mismatches between the pattern and the text is the number of non-zero's in the array $x_{i,1}, \ldots, x_{i,m}$. Denote:

$$C[i] = \sum_{1 \leqslant j_1 < j_2 < \cdots < j_{k+1} \leqslant m} x_{i,j_1} \cdot x_{i,j_2} \cdot \ldots \cdot x_{i,j_{k+1}}, \tag{4}$$

where the sum is over all possible $(k + 1)$-tuples of ordered indices from $\{1, \ldots, m\}$. We claim that there is a k-mismatch if and only if $C[i] = 0$. This is because if there are k or less mismatches at location i, then every set of $k + 1$ indices must contain at least one position j' where the pattern matches the text, i.e., $x_{i,j'} = 0$, which implies that $C[i] = 0$. Conversely, if there are more than k mismatches at text location i, and let $j_1, \ldots, j_{k+1}, \ldots$ denote their positions in the pattern, then

```
**Algorithm** $k$-**MISMATCH** (pattern $p$, text $t$, integer $k$):
1. Encode $p$ and $t$ using positive integers, '*' using 0
2. Compute the arrays $D_1[i] = \sum_{j=1}^{m} x_{i,j}, \ldots, D_{k+1}[i] = \sum_{j=1}^{m} x_{i,j}^{k+1}$,
   where $x_{i,j} = p_j t_{i+j-1}(p_j - t_{i+j-1})^2$
3. Compute the array $C[i] = \sum x_{i,j_1} \cdot x_{i,j_2} \cdot \ldots \cdot x_{i,j_{k+1}}$
4. Report a match at location $i$ iff $C[i] = 0$
```

Fig. 1. Algorithm for matching with don't-cares and k mismatches. See also Fig. 2.



Step 1: Encode *p* and *t* using positive integers, '*' using 0
Step 2: Compute $D_1[i], \ldots, D_{k+1}[i]$ with FFTs
$F_{1,1}=D_1$
. . .
$F_{1,s}=D_s$
. . .
$F_{1,k+1}=D_{k+1}$
. . .
$F_{t,1}$
$F_{t,s+1}$
$F_{t+1,s}$
. . .
. . .
$F_{k+1,1}=C$
Step 3: Compute $C[i]$ using dynamic programming
Step 4: Report "match" wherever $C[i]=0$
$HD(i)=\min\{ t \mid F_{t+1,1}[i]=0 \}$

Fig. 2. Our algorithm for matching with don't-cares and k mismatches.

$x_{i,j_1} \cdot x_{i,j_2} \cdot \ldots \cdot x_{i,j_{k+1}} > 0$. Since all $x_{i,j}$'s are non-negative, we get $C[i] \geqslant x_{i,j_1} \cdot x_{i,j_2} \cdot \ldots \cdot x_{i,j_{k+1}} > 0$, as required.

Alas, the value $C[i]$ is a sum of $\binom{m}{k+1}$ products of $k+1$ $x_{i,j}$'s—how can we compute it efficiently? Our main observation is that $C[i]$ can be expressed using a recursion, whose base is made up of $k+1$ arrays of the type $D_s[i] = \sum_{j=1}^{m} x_{i,j}^{s}$. Each of these arrays can be broken up into $O(k)$ convolutions and computed using FFTs. A dynamic programming procedure is then applied to compute $C[i]$ and report the results. An additional obstacle we need to overcome is that the numbers computed by the algorithm are too large to fit inside a single RAM word. We use simple tools from number theory to solve this problem. The total time complexity of our randomized algorithm, which reports the correct locations with high probability, is $O(nk^2 \log m)$. The running time of our deterministic solution is $O(nk^3 \log m)$. It is the first deterministic algorithm that solves matching with don't-cares and k mismatches in $O(\mathrm{poly}(k) \cdot n \log m)$ time. In particular, for constant k, it matches the $O(n \log m)$ time complexity for exact matching with don't-cares [2,3].

5. The algorithm

Our algorithm for matching with don't-cares and k mismatches, called k-MISMATCH, consists of four main steps, outlined in Fig. 1.

In step 1, the pattern and the text are encoded as in [3]—each alphabet symbol is replaced by a unique positive integer, and don't-cares are replaced by 0's. Step 2 computes the arrays $D_s[i]$ for $s = 1, \ldots, k+1$:

$$\begin{aligned} D_s[i] &= \sum_{j=1}^{m} x_{i,j}^{s} = \sum_{j=1}^{m} p_j^s t_{i+j-1}^s (p_j - t_{i+j-1})^{2s} \\ &= \sum_{j=1}^{m} p_j^{3s} t_{i+j-1}^{s} - 2s \sum_{j=1}^{m} p_j^{3s-1} t_{i+j-1}^{s+1} \\ &\quad + \cdots + \sum_{j=1}^{m} p_j^{s} t_{i+j-1}^{3s}. \end{aligned}$$

Thus, each array $D_s[i]$ is a linear combination of $2s+1$ convolutions of the type $p^a \oplus t^b$, so a total of $O(k^2)$ convolutions plus $O(k^2)$ linear-time operations on arrays of length n are required in step 2. In order to perform step 3, we need to define another family of arrays. Let s and t be positive integers, $s + t \leqslant k + 2$. Define the following array:

$$F_{t,s}[i] = \sum_{\substack{1 \leqslant j_1 \leqslant m \\ 1 \leqslant j_2 < \cdots < j_t \leqslant m \\ \forall l > 1 \;\; j_l \neq j_1}} x_{i,j_1}^{s} \cdot x_{i,j_2} \cdot \ldots \cdot x_{i,j_t}.$$

If $s = 1$, we also require $j_1 < j_2$, so that each term occurs only once in the above sum. Informally, $F_{t,s}[i]$ is the sum of all terms of the type $x_{i,j_1}^{s} \cdot x_{i,j_2} \cdot \ldots \cdot x_{i,j_t}$, where the indices $j_1, \ldots, j_t$ are chosen in such a way that each term is taken exactly once. Notice that $F_{1,s}[i] = D_s[i]$, and $F_{k+1,1}[i] = C[i]$.

Lemma 1. *The following recursion holds*:

$$F_{t+1,s}[i] = \frac{1}{c}\big(F_{t,1}[i] \cdot F_{1,s}[i] - F_{t,s+1}[i]\big),$$

$$\textit{where } c = \begin{cases} t+1, & \textit{if } s = 1, \\ 1, & \textit{if } s > 1. \end{cases}$$

Proof. By definition:

$$\begin{aligned} & F_{t,1}[i] \cdot F_{1,s}[i] \\ &\quad = \Bigg(\sum_{1 \leqslant j_1 < \cdots < j_t \leqslant m} x_{i,j_1} \cdot \ldots \cdot x_{i,j_t} \Bigg) \cdot \Bigg(\sum_{1 \leqslant j \leqslant m} x_{i,j}^{s} \Bigg). \end{aligned}$$

Opening the above parentheses, we get two types of terms—one with $t+1$ distinct x's (when the index j in $x_{i,j}^{s}$ is not one of $j_1, \ldots, j_t$), and one with t distinct x's (when $j \in \{j_1, \ldots, j_t\}$). Collecting each type to a separate sum, we get using simple algebra:

$$\begin{aligned} & F_{t,1}[i] \cdot F_{1,s}[i] \\ &\quad = c \cdot \sum_{\substack{1 \leqslant j_1 \leqslant m \\ 1 \leqslant j_2 < \cdots < j_{t+1} \leqslant m \\ \forall l > 1 \;\; j_l \neq j_1}} x_{i,j_1}^{s} \cdot x_{i,j_2} \cdot \ldots \cdot x_{i,j_{t+1}} \\ &\qquad + \sum_{\substack{1 \leqslant j_1 \leqslant m \\ 1 \leqslant j_2 < \cdots < j_t \leqslant m \\ \forall l > 1 \;\; j_l \neq j_1}} x_{i,j_1}^{s+1} \cdot x_{i,j_2} \cdot \ldots \cdot x_{i,j_t} \\ &\quad = c \cdot F_{t+1,s}[i] + F_{t,s+1}[i]. \qquad \square \end{aligned}$$

By Lemma 1, step 3 of the algorithm can be computed using dynamic programming. We first set $F_{1,s}[i] = D_s[i]$ for $s = 1, \ldots, k+1$. We then compute $F_{2,s}[i]$ for

$s = 1, \ldots, k$ using the recursion. We continue in this way, as illustrated in Fig. 2, until we obtain $F_{k+1,1}[i]$, which is the array $C[i]$ we wished to compute. Note that by examining the arrays $F_{1,1}[i], \ldots, F_{k+1,1}[i]$, we can infer the exact number of mismatches at each k-mismatch location:

$$HD(i) = \min\bigl\{t \mid F_{t+1,1}[i] = 0\bigr\}.$$

The number of arrays the algorithm computes in step 3 is $k(k+1)/2$. Since each array is calculated in linear time, the running time of this step is $O(nk^2)$.

The overall running time of the algorithm is dominated by the time taken to perform the $O(k^2)$ FFTs in step 2, which is $O(nk^2 \log m)$. However, as mentioned earlier, there is still one flaw we must address—the algorithm computes numbers as large as $\binom{m}{k+1}|\Sigma|^{4(k+1)} < m^{5(k+1)}$ (see (4)), i.e., numbers with $O(k \log m)$ bits, whereas the RAM model commonly used in the pattern-matching literature permits unit-cost operations only on $O(\log n)$-bit words. To solve this problem, we perform all computations modulo some large prime number q that fits into a single RAM word, as described in the next sections.

5.1. Randomized algorithm

Our randomized algorithm, called k-MISMATCH-RAN, is outlined in Fig. 3. The algorithm randomly chooses two large prime numbers—q_1 and q_2, each with $O(\log n)$ bits, and computes the array $C_q[i] = C[i] \bmod q$, where $q = q_1 q_2$, using the procedure described above (integers in $\mathbb{Z}_q$ fit into a single RAM word, as required). Finally, it reports a match at location i if $C_q[i] = 0$. $C[i]$ is an integer between 0 and some large number N, where $N < m^K$ and $K = 5(k+1)$. Thus, it has at most K prime factors larger than m. We therefore choose q_1 and q_2 randomly and uniformly from the primes within a sufficiently large interval, to guarantee that the probability of reporting a false match is small. The following lemma specifies the required interval.

Lemma 2. *For $n \geqslant 17$ and $K = 5(k+1) \leqslant 5n$, there are more than nK primes in the interval $[n+1, 6n(K+1)\ln n]$. All these primes are $O(\log n)$-bit numbers.*

Proof. Following are well-known bounds on the number $\pi(x)$ of primes less than or equal to x [12]:

$$\forall x \geqslant 17 \quad \frac{x}{\ln x} < \pi(x) < 1.26\frac{x}{\ln x}.$$

Since $\ln(6n(K+1)\ln n) < 4\ln n$ for $n \geqslant 17$, it follows from the above bounds that:

$$\begin{aligned}
&\pi\bigl(6n(K+1)\ln n\bigr) - \pi(n) \\
&\quad > \frac{6n(K+1)\ln n}{4\ln n} - \frac{1.26n}{\ln n} \\
&\quad > \frac{6n(K+1)\ln n - 6n}{4\ln n} \\
&\quad > nK. \qquad \square
\end{aligned}$$

```
Algorithm k-MISMATCH-RAN (p, t, k)
1. Randomly choose two prime numbers—$q_1, q_2 \in [n+1, 6n(K+1)\ln n]$,
   where $K = 5(k+1)$
2. $C_q[i] = k$-MISMATCH$(p, t, k) \bmod q_1 q_2$
3. Report a match at location $i$ iff $C_q[i] = 0$
```

Fig. 3. Randomized algorithm for matching with don't-cares and k mismatches.

```
Algorithm k-MISMATCH-DET (p, t, k)
1. Find prime numbers—$q_1, \ldots, q_K > m$, where $K = 5(k+1)$
2. For each prime $q_r$ do:
     Let $C_r[i] = k$-MISMATCH$(p, t, k) \bmod q_r$
3. Report a match at location $i$ iff $\forall r\ C_r[i] = 0$
```

Fig. 4. Deterministic algorithm for matching with don't-cares and k mismatches.

In order to obtain the primes q_1 and q_2, one can randomly draw numbers from the above interval and check each number for primality. This takes $O(\text{polylog}\, n)$ expected time [13], and can be done in preprocessing, as it depends only on the length of the text. Since there are more than nK primes in the interval, out of which at most K are factors of $C[i]$, it follows that if $C[i] > 0$ the probability that $C_q[i] = 0$ is:

$$\begin{aligned}
&P\bigl(C_q[i] = 0 | C[i] > 0\bigr) < \binom{K}{2}\Big/\binom{nK}{2} \\
&\quad = \frac{K(K-1)}{nK(nK-1)} < 1/n^2.
\end{aligned}$$

In other words, the algorithm reports a false match at a given location i with probability less than $1/n^2$ (a true match is always reported correctly, since $C[i] = 0$ implies $C[i] \equiv 0 \pmod q$). Thus, the probability that the algorithm reports any false match in the entire text is less than $1/n$.

Theorem 1. *Algorithm k-MISMATCH-RAN solves matching with don't-cares and k mismatches in $O(nk^2 \log m)$ time and gives the correct output (i.e., does not report false matches anywhere in the text) with probability at least $1 - \frac{1}{n}$.*

For $k = O(\sqrt{\log n \log\log n})$ our algorithm improves upon the randomized technique of Clifford et al. [8], which runs in $O(n(k + \log n \log\log n)\log m)$ time.

5.2. Deterministic algorithm

The deterministic algorithm, called k-MISMATCH-DET and outlined in Fig. 4, chooses K prime numbers—$q_1, \ldots, q_K > m$, and computes the array $C[i]$ modulo each of these primes separately. Lemma 3 specifies the interval that contains these primes.

Lemma 3. *For $m \geqslant 17$, the interval $[m+1, 19m\ln m]$ contains more than $5m$ primes.*

Proof. Since $\ln(19m\ln m) < 3\ln m$ for $m \geqslant 17$, then using the bounds on $\pi(x)$ (see proof of Lemma 2) we get:

$$\begin{aligned}
&\pi(19m\ln m) - \pi(m) \\
&\quad > \frac{19m\ln m}{3\ln m} - \frac{1.26m}{\ln m}
\end{aligned}$$

$$
\begin{aligned}
&> \frac{19m\ln m - 4m}{3\ln m} \\
&> 5m. \qquad \square
\end{aligned}
$$

It follows from the above lemma that the primes $q_1, \ldots, q_K$ $(K \leqslant 5m)$ can be found in time $o(m \ln m)$ using modern sieve techniques [14], and that they are $O(\log n)$-bit numbers, as required. The algorithm completes by reporting all locations for which $C[i]$ is 0 modulo all K primes. This always yields the correct answer, since:

$$0 \leqslant C[i] < m^K < \prod_{j=1}^{K} q_j.$$

The running time of the deterministic algorithm is $o(m \ln m)$ for finding the prime numbers, plus $O(Knk^2 \log m)$ for computing $C[i]$ modulo each of the K primes. Thus, its total running time is $O(nk^3 \log m)$.

Theorem 2. *Algorithm k-MISMATCH-DET solves matching with don't-cares and k mismatches in $O(nk^3 \log m)$ time.*

Our deterministic algorithm is faster than the $O(nk^2 \cdot \log^3 m)$ time deterministic method of [8] for $k = O(\log^2 m)$.

6. Summary

We presented efficient randomized and deterministic algorithms for matching with don't-cares and k mismatches with running times $O(nk^2 \log m)$ and $O(nk^3 \log m)$, respectively. For small values of k, our algorithms are faster than the recently published methods of Clifford et al. [8]. For small alphabets $(|\Sigma| = O(k^2 \min\{k, \log^2 m\}))$, the match-count algorithm is currently the fastest. Our solution is the first $O(\text{poly}(k) \cdot n \log m)$ time deterministic algorithm. For fixed values of k, this matches the $O(n \log m)$ time complexity of exact matching with don't-cares [2,3]. An interesting open question is whether an $O(f(k)n \log m)$ algorithm can be found with $f(k) = o(k^3)$, or even $f(k) = O(k)$.

Acknowledgements

This study was supported in part by the Israel Science Foundation (grant 802/08). We thank the anonymous referees for their constructive comments.

References

[1] M. Fischer, M. Paterson, String matching and other products, in: R.M. Karp (Ed.), Proc. 7th SIAM–AMS Complexity of Computation, 1974, pp. 113–125.

[2] R. Cole, R. Hariharan, Verifying candidate matches in sparse and wildcard matching, in: Proc. 34th Symposium on Theory of Computing, 2002, pp. 592–601.

[3] P. Clifford, R. Clifford, Simple deterministic wildcard matching, Inform. Process. Lett. 101 (2) (2007) 53–54.

[4] K. Abrahamson, Generalized string matching, SIAM J. Comput. 16 (1987) 1039–1051.

[5] J. Feigenbaum, Y. Ishai, T. Malkin, K. Nissim, M. Strauss, R. Wright, Secure multiparty computation of approximations, ACM Trans. on Algorithms 2 (3) (2006) 435–472.

[6] E. Porat, O. Lipsky, Improved sketching of hamming distance with error correcting, in: A. Apostolico, M. Crochemore, K. Park (Eds.), Proc. 18th Annual Symposium on Combinatorial Pattern Matching, 2007, pp. 173–182.

[7] A. Amir, M. Lewenstein, E. Porat, Faster algorithms for string matching with k mismatches, J. Algorithms 50 (2) (2004) 257–275.

[8] R. Clifford, K. Efremenko, E. Porat, A. Rothschild, k-mismatch with don't cares, in: L. Arge, M. Hoffmann, E. Welzl (Eds.), Proc. 15th Annual European Symposium on Algorithms, 2007, pp. 151–162.

[9] O. Lipsky, E. Porat, Approximated pattern matching with the L_1, L_2 and L_∞ metrics, in: A. Amir, A. Turpin, A. Moffat (Eds.), Proc. 15th Symposium on String Processing and Information, 2008, pp. 212–223 (original manuscript from 2002).

[10] T. Eilam-Tzoreff, U. Vishkin, Matching patterns in strings subject to multi-linear transformations, in: R. Capocelli (Ed.), Sequences: Combinatorics, Compression, Security, and Transmission, Springer-Verlag, 1990, pp. 45–58.

[11] A. Amir, Y. Aumann, G. Benson, A. Levy, O. Lipsky, E. Porat, S. Skiena, U. Vishne, Pattern matching with address errors: rearrangement distances, in: Proc. 17th Annual ACM–SIAM Symposium on Discrete Algorithms, 2006, pp. 1221–1229.

[12] J. Rosser, L. Schoenfield, Approximate formulas for some functions of prime numbers, Illinois J. Math. 6 (1962) 64–94.

[13] M. Agrawal, N. Kayal, N. Saxena, PRIMES is in P, Ann. of Math. 160 (2) (2004) 781–793.

[14] A. Atkin, D. Bernstein, Prime sieves using binary quadratic forms, Math. Comp. 73 (2004) 1023–1030.

8. Discussion

In this thesis we described our study on the theoretical and practical aspects of *cis*-regulatory motif finding. We developed novel statistical scores and algorithms for uncovering new BS patterns of TFs and miRNAs. We implemented our methods in an efficient user-friendly software package, demonstrated its applicability to a wide range of motif finding tasks, and showed that it outperforms existing tools. We applied computational analyses on pivotal mammalian cellular processes and demonstrated the power of our approach to delineate their intricate transcriptional programs. Finally, we developed new efficient algorithms for several pattern matching problems that are related to motif finding. In the future, we hope that the practical tools and techniques we implemented and the theoretical algorithms we developed will contribute to researchers in biology and computer science, respectively.

8.1 The Amadeus/Allegro motif finding platform

The main contribution of this thesis is the Amadeus/Allegro motif discovery software platform. Amadeus, described in Chapter 2, searches for motifs that are over-represented in the *cis*-regulatory sequences of a given target set of genes with respect to a large reference gene set (typically, the entire genome). Over-representation is evaluated either using the standard hypergeometric score or using a novel score, termed *binned enrichment score*, which accounts for biases in the length and nucleotide composition of the sequences of the given target set. The general architecture of Amadeus is a pipeline of filters, or refinement phases, where each phase receives as input a list of candidate motifs and applies an algorithm for refining the list and producing a set of improved candidates, which serve as a starting point for the next phase. The first phases typically work on a very large number of candidates, such as all possible *k*-mers, and execute simple procedures for choosing the most promising motifs. Subsequent phases run more complex (and computationally intensive) algorithms in order to converge to better motifs. Using sophisticated data structures that were tailored for the specific demands of the algorithm, Amadeus achieves high efficiency, and is among the fastest motif finding tools.

Extensive testing showed that Amadeus outperforms other popular motif finding tools in terms of accuracy and running time. Of note, Amadeus attained high motif recovery rates both in yeast and in metazoans, whereas the success rates of extant tools deteriorated in the transition from yeast to more complex organisms. We believe this is largely due to the fact that Amadeus utilizes the entire genome as a reference set for testing over-representation, rather than a probabilistic sequence model inferred from nucleotide counts.

Amadeus supports a second type of genome-wide motif discovery task - identifying motifs based on global spatial features, namely – uneven distribution along the promoters, between the strands, or among the chromosomes. This type of analysis can be applied to any genome with a sufficient number of *cis*-regulatory sequences without need for target sets from prior experiments. We demonstrated how, in a single run, Amadeus recovered many known and novel motifs that are localized along human and mouse promoters, and motifs with non-random chromosomal distribution in fly and in worm.

In Chapter 3 we extended the above functionality of Amadeus using Allegro, a method for simultaneous inference of motifs and their associated expression profiles given genome-wide expression datasets. Unlike existing techniques, which rely on statistical assumptions, Allegro uses a novel non-parametric model called *Condition Weight Matrix* (CWM) to describe the expression profile of a group of co-regulated genes. Allegro builds upon the powerful motif search engine and other features of Amadeus. For each candidate motif, Allegro fits a CWM to its putative targets and computes a statistical score to ascertain whether the sequence and expression patterns (i.e, the PWM and CWM, respectively) are significantly correlated. We applied Allegro on several large-scale datasets from yeast, fly, mouse and human. In all cases, Allegro successfully recovered relevant TF/miRNA motifs, and outperformed the popular two-step approach of first clustering the expression data to identify groups of co-regulated genes and then searching for motifs that are over-represented in each group.

In addition to its applicability to the wide range of use-cases described above, the Amadeus/Allegro platform includes a wealth of features for enhanced usability and performance, such as combined analysis of multiple gene target/expression datasets from one or more organisms, built-in bootstrapping, motif-pairs analysis, and comparison to known TF/miRNA binding patterns from Transfac/miRBase. In order to make Amadeus easily accessible to biologists, we "wrapped" it in an informative, user-friendly graphical interface. Our software is publicly available at http://acgt.cs.tau.ac.il/allegro. To date, it has been downloaded by over 500 researchers around the world. Several groups are incorporating Amadeus as part of their computational pipeline for analyzing gene expression assays.

The Amadeus/Allegro platform can be extended in several ways: (a) improved motif enumeration algorithms for detecting long, gapped motifs (e.g., the yeast GAL4 motif contains a gap of length 11); (b) new statistical scores for computing the conservation of candidate motifs among several species, and the bias in the orientation, order and distance between motifs within a module; (c) analysis of expression values obtained by next-generation sequencing technologies.

8.1.1 Benchmarking motif finding tools

As mentioned in Section 1.3.1, a plethora of motif discovery methods has been described in the literature. Comparing them is an important, and apparently non-trivial, task. Many studies that present new motif finding algorithms demonstrate the improved accuracy of their methods using simulated data or a small ad-hoc collection of datasets. Obviously, such results do not guarantee equally-good performance in many real-life scenarios. As explained in Chapter 2 (see also part B of Supplemental Notes in [1]), a good benchmark for reliably comparing the performance of different tools should be based on a large number of real, heterogeneous, experimentally-derived datasets. We constructed the first such benchmark. To do so, we collected from the literature a compendium of over 40 TF and miRNA target gene sets derived from diverse high-throughput experiments in several metazoans. Our benchmark is publicly available at http://acgt.cs.tau.ac.il/amadeus. We hope other researchers use the benchmark to test and improve their methods, and extend it with additional gene sets from various sources.

8.1.2 Biases in *cis*-regulatory sequences

In the course of our study, we observed that in many gene expression experiments there is a correlation between the expression values of genes and the length or GC-content of their *cis*-regulatory sequences. Obviously, groups of co-regulated genes inferred from such datasets are biased too. Many motif finders ignore these biases, and this often results in failure to discover the correct motifs or in many false predictions. The binned enrichment score we implemented in Amadeus and in Allegro partitions the genes into bins according to the length and GC-content of their *cis*-regulatory sequences and evaluates the over-representation of the motif based on its abundance in each bin. Using this score, Amadeus and Allegro yielded more accurate results.

We believe that this issue should be addressed in other contexts as well, e.g., when assessing the enrichment of known TFBS motifs in a given set of genes. For example, the FAME algorithm developed in our lab computes the enrichment of miRNA targets in a given set of genes [72]. In order to evaluate the statistical significance of the enrichment while accounting for biases in the length and base composition of 3' UTRs and miRNA seeds, FAME constructs a bipartite graph, in which miRNAs are connected with their predicted targets. It then uses degree-preserving permutations to generate random graphs, in which each miRNA is connected with the same number of targets (and vice versa) as in the original graph. These random graphs provide a distribution of the number of targets per miRNA within the given gene set, and this distribution is used to assess the *p*-value of the observed enrichment. The main drawback of such a bootstrapping-based approach is the time complexity of constructing many random samples, which limits the significance of the reported *p*-values (*N* samples are required for *p*-value *1/N*) and prohibits their application in cases where the randomization has to be performed many times (e.g., when searching for *de-novo* motifs).

8.1.3 Comparative sequence analysis

As explained in Section 1.3.1, the phylogenetic footprinting approach attempts to obtain more accurate BS predictions by focusing on promoter (or 3' UTR) regions that are conserved among orthologous genes in related species. However, several recent studies suggest a limited cross-species conservation of functional BSs. For example, Odom et al. studied several evolutionarily-conserved TFs in human and mouse hepatocytes, and found that most of their BSs (41-89%) are species specific; moreover, when a TF binds the promoters of orthologous genes, the BSs reside in aligned regions only in a third of the cases [73]. Lin et al. identified genomic sequences bound by ESR1 (Estrogen Receptor alpha) in breast cancer cells, and found that only 23% of them are conserved among vertebrates [74]. In another work, BSs of two TFs in three yeast species were shown to have diverged considerably faster than ortholog content, perhaps constituting the major cause of phenotypic diversity [75]. In light of such evidence, we suggest to utilize comparative genomics by means of searching for motifs that are concurrently over-represented in target sets of related species, as identified by species-specific experiments (see also Section 8.2). Amadeus supports such joint analysis of multiple target sets from one or more species – the motif search is performed on all target sets in parallel, and the scores attained by each motif on all sets are combined into a single *p*-value using the Z-transform. Allegro can analyze multiple expression datasets in a similar way. In Chapters 2 and 3 we described several examples in which joint analysis of multiple target sets or expression datasets recovered the correct motif, whereas analyzing each dataset separately failed to discover it.

8.1.4 Modeling expression profiles

The most common type of analysis applied to gene expression datasets today is clustering, which partitions the genes into co-expressed groups. Co-expression is usually scored using a similarity measure, such as Pearson or Spearman correlation, or Euclidean distance [50, 76]. One of the shortcomings of such metrics is that all the conditions in the expression matrix have an equal contribution to the overall similarity. Specifically, if a certain transcriptional module is tightly co-expressed across a small fraction of the tested conditions, the similarity among the expression profiles of the genes in the module, measured across all the conditions, might be indistinguishable from their similarity to the expression of the rest of the genome. Likewise, a set of related conditions (e.g., close time-points), in which the expression values are highly correlated, might bias the similarity scores [76]. Another shortcoming of similarity measures is that they are sensitive to extreme values (outliers), that is, an expression pattern that contains a very high or low value in one of the conditions could be too far from its true cluster and might thus be assigned to an unrelated cluster [77]. Furthermore, Pearson correlation might not perform well when the mean expression level of the genes, or their variance, carry important information. For example, a gene that is highly induced in all measured conditions will have high Pearson correlation with a gene that is not expressed in any of these conditions. Euclidean distance is usually computed after standardizing the expression profiles to zero mean and variance one to account for differences in the magnitude of the expression values (e.g., [48, 78, 79]). This again assumes that the mean and the variance of each expression profile are not informative. In an alternative non-parametric approach, when ranking the expression values, Spearman's rank correlation ignores a significant amount of information, and often performs poorly (see Supplemental Table II in [50]).

Another crucial disadvantage of clustering is that it partitions the genes into disjoint groups, whereas transcriptional modules may have large overlaps. For example, two functionally-related TFs could have distinct, but overlapping, target sets. In order to address these issues, biclustering methods have been developed [50, 80]. A biclustering algorithm identifies groups of genes whose expression profiles are correlated over a subset of the conditions, usually allowing for overlaps between the groups. Still, the aforementioned drawbacks of similarity measures apply here as well – namely, equal contribution of all the conditions in a bicluster to the similarity score, biases due to correlated conditions, sensitivity to extreme values, and, in most cases, ignoring information in the mean and variance of expression profiles. Obviously, if the (bi-)

clustering algorithm fails to identify the true gene sets with reasonable accuracy, subsequent analyses of the gene clusters, such as motif discovery, are bound to fail.

In Chapter 3 we introduced a new model, called CWM, for describing the common expression pattern of a set of co-regulated genes. The model gives a likelihood ratio to the group using discrete expression levels. The CWM does not suffer from the aforementioned drawbacks of similarity measures – it is robust against extreme values, it can capture transcriptional modules whose expression profiles differ from the rest of the genome across a small fraction of the conditions, it is not biased by correlations among the conditions, and it does not assume that the expression values follow a pre-defined type of distribution or that they are standardized to zero mean and variance one. The CWM is analogous to the PWM model for sequence motifs, with DNA bases substituted here by discrete expression levels, and the positions along the sequence motif replaced by the experimental conditions. Indeed, since it uses discrete expression levels rather than continuous values, a shortcoming of the CWM approach is that it ignores some of the information present in the data. Our experiments indicate that despite this loss of information, the CWM outperforms similarity metrics in the vast majority of the cases (see Supplemental Table II in [2]), and yields excellent results in the context of Allegro. An open question is how to automatically set and optimize the discretization parameters.

In this study we used the CWM in the context of motif finding. It would be interesting to test its applicability to other gene expression analysis tasks, such as functional analysis (i.e., identifying GO terms whose genes share a common expression profile). We are currently developing an algorithm for identifying groups of genes that exhibit distinct expression profiles across several expression datasets, given a (possibly large) collection of datasets from multiple biological systems and/or species (joint work with Shahar Kidron).

8.1.5 A novel motif pair in *C. elegans*

By applying Allegro on several gene expression datasets in the nematode *C. elegans*, we discovered a pair of novel motifs that seem to be related to the transcriptional regulation of oogenesis (production of oocytes) in adult hermaphrodites. Strikingly, the order between the occurrences of these two motifs, their orientation, and the distance between them are highly conserved along the *C. elegans* genome and in all other available genomes of the *Caenorhabditis* genus, but not in other nematode species. Furthermore, although the length of the gap between the two motifs is almost fixed (up to a couple of bases), its sequence content is not conserved (i.e., the two motifs are not part of a long repeat). The full details are described in [81]. We are currently collaborating with Dr.

Limor Broday (Faculty of Medicine, Tel Aviv University) and Dr. Marian Walhout (University of Massachusetts Medical School) in an attempt to experimentally validate our findings and gain further insights into their biological roles.

8.2 Dissecting regulatory networks

In Chapters 4 and 5 we analyzed two central regulatory networks in mammals – cell cycle and innate immune response. In both cases we utilized multiple sources of information in order to improve the accuracy of BS prediction and to obtain a more reliable and comprehensive picture of the transcriptional network. In Chapter 4 we deciphered the *cis*-regulatory elements that control cell cycle phasing by analyzing genome-wide promoter sequences from 12 species. Our analysis highlighted two major regulators of cell cycle progression. The first is E2F, a family of TFs with prominent roles in cell cycle regulation. We showed that E2F binding signals are conserved in promoters of genes that are transcribed during the G_1/S phase in all the organisms we examined, from worm to human. Interestingly, all 13 cell-cycle genes with a conserved canonical E2F BS (i.e., a site that perfectly matches the E2F consensus motif, possibly indicating high-affinity binding) peak at the G_1/S phase. This may indicate that high-affinity BSs of E2F are specific to the G_1/S phase, whereas E2F regulation in other phases is mediated by sites with lower affinity. We discovered a novel *cis*-regulatory module, made up of the CHR and NF-Y elements and conserved in all analyzed vertebrates, that dictates an expression profile that is almost exclusively restricted to the G_2 and G_2/M phases (40 out of 42 genes). By searching for conserved TF modules that are over-represented in specific phases of the cell cycle, our analysis attained unprecedented accuracy – up to 99% true positives in some cases. Moreover, we demonstrated the dramatic improvement in TFBS detection as more sources of information are incorporated, by applying comparative genomics techniques (i.e., sequence conservation criteria) and searching for modules of cooperative TFs (as opposed to single TFs) with conserved order and distance between their BSs.

The second transcriptional network we delineated is the response induced by Toll-like receptors (TLRs), which are the main pathogen sensors of the innate immune system in vertebrates (Chapter 5). We analyzed four large-scale gene expression datasets in mouse and human macrophages stimulated with various pathogen-mimetic agents that stimulate several TLRs. By combined computational analysis of promoter sequences and multiple expression datasets, and by defining kinetic patterns of transcriptional response, we identified and characterized several distinct regulatory programs. The two main components are: (1) an early-induced universal response, activated by all examined TLRs, and regulated by the NF-κB transcription factor; (2) a delayed wave that is specific to activated TLR3 and TLR4, and induced by TFs that bind to ISRE (Interferon-Stimulated Response Element). We identified new genes that participate in these two transcriptional components and pointed to novel regulatory feedback loops, further increasing the known complexity of the TLR-induced network. Using diverse evidence, primarily the kinetics of expression levels in multiple datasets and conserved BS signatures, we obtained an accurate, system-level delineation of the transcriptional program of the innate immune response.

Our analyses of the cell cycle and immune response transcriptional networks highlight the functions of TFs and their modular organization in these biological processes. The significant improvement we achieved in the specificity of the putative targets can make their empirical validation much more focused and efficient, and assist in easier interpretation of the predicted regulatory networks.

8.3 Pattern matching algorithms

We developed new FFT-based algorithms for solving pattern matching problems that are related to *cis*-regulatory motif finding. Our methods improve on the complexity of existing techniques, and borrow ideas from number theory. In Chapter 6 we described an algorithm for matching patterns with character classes (i.e., degenerate motifs). The algorithm uses a novel encoding scheme that utilizes features of prime numbers. Its running time complexity ranges from near-linear time to the $O(|\Sigma|\, n \log m)$ complexity of the match-count algorithm (see Section 1.4.2), depending on a parameter κ. Interestingly, this parameter incorporates the three main characteristics of the input – the length of the text (n), the length of the pattern (m), and the size of the alphabet ($|\Sigma|$): $\kappa = \log_{|\Sigma|}(\log n / \log m)$. Thus, κ provides a single scale for measuring how difficult a given instance of the problem is, in terms of complexity. In particular, if κ=1 (i.e., $m^{|\Sigma|}=n$), our algorithm runs in time $O(n \log m)$, the same as the best methods for matching with don't-cares, which is a special case of the problem we solve. We also developed variants of our method for solving approximate matching with character classes and the subset matching problem.

In Chapter 7 we presented new efficient algorithms for matching with don't-cares and k mismatches, i.e., locating the occurrences of a consensus motif that contains gaps (see Section 1.4.3). Our method uses a combination of FFT-based convolutions and a dynamic programming procedure. It is conceptually simpler, and for small values of k

faster, than extant techniques - the deterministic algorithm we developed runs in time $O(nk^3\log m)$, compared to $O(nk \log^2 m (\log^2 k+\log\log m))$ time of the fastest existing method, which has been published very recently [71]. In Section 1.4.3 we raised the question whether the best known $O(n \log m)$ time complexity of exact matching with don't-cares still applies when we allow a fixed number of mismatches per pattern occurrence. The answer is yes - our solution is the first $O(\text{poly}(k)\ n \log m)$ time deterministic algorithm for matching with don't-cares and k mismatches. An interesting open question is whether an $O(f(k)\ n \log m)$ algorithm can be devised with $f(k)=o(k^3)$, or even $f(k)=O(k)$; recall that pattern matching with k mismatches (without don't-cares) can be solved in time $O(n\sqrt{k \log k})$, i.e., with sub-linear dependency on k [68].

We are now developing a new efficient algorithm for pattern matching with swaps. In this problem, the pattern is said to match the text at location i if adjacent pattern characters can be swapped, as necessary, so as to make the pattern identical to the substring at location i in the text (note that each character may be involved in at most one swap). The biological motivation for considering the swap operation is gene rearrangement events, which obviously have many additional variants and constraints. Single-character swap is also one of the most typical typing errors. Solving swap matching in time $o(nm)$ was described in 1995 as one of the open problems in non-standard string matching [82]. Since then, a number of efficient algorithms have been described, solving this open problem. Currently, the fastest algorithm for swap matching runs in time $O(\log|\Sigma|\ n \log m)$ [83]. An open question is whether the dependence on $|\Sigma|$ can be removed, as achieved recently for matching with don't-cares (see Section 1.4.1), i.e., can swap matching be solved in time $O(n \log m)$? We believe that the answer is yes.

Bibliography

1. Linhart C, Halperin Y, Shamir R: **Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets**. *Genome Res* 2008, **18**(7):1180-1189.
2. Halperin Y, Linhart C, Ulitsky I, Shamir R: **Allegro: Analyzing expression and sequence in concert to discover regulatory programs**. *Nucleic Acids Res* 2009, **37**(5):1566-1579.
3. Linhart C, Elkon R, Shiloh Y, Shamir R: **Deciphering transcriptional regulatory elements that encode specific cell cycle phasing by comparative genomics analysis**. *Cell Cycle* 2005, **4**(12):1788-1797.
4. Elkon R, Linhart C, Halperin Y, Shiloh Y, Shamir R: **Functional genomic delineation of TLR-induced transcriptional networks**. *BMC Genomics* 2007, **8**:394.
5. Linhart C, Shamir R: **Faster pattern matching with character classes using prime number encoding**. *Journal of Computer and System Sciences* 2009, **75**(3):155-162.
6. Linhart C, Shamir R: **Matching with don't-cares and a small number of mismatches**. *Information Processing Letters* 2009, **109**(5):273-277.
7. Lockhart DJ, Winzeler EA: **Genomics, gene expression and DNA arrays**. *Nature* 2000, **405**(6788):827-836.
8. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E *et al*: **Genome-wide location and function of DNA binding proteins**. *Science* 2000, **290**(5500):2306-2309.
9. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A *et al*: **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing**. *Nat Methods* 2007, **4**(8):651-657.
10. **Nomenclature Committee of the International Union of Biochemistry (NC-IUB): Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984**. *Biochem J* 1985, **229**(2):281-286.
11. Stormo GD: **DNA binding sites: representation and discovery**. *Bioinformatics* 2000, **16**(1):16-23.
12. Wingender E, Hogan J, Schacherer F, Potapov AP, Kel-Margoulis O: **Integrating pathway data for systems pathology**. *In Silico Biol* 2007, **7**(2 Suppl):S17-25.
13. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles**. *Nucleic Acids Res* 2004, **32**(Database issue):D91-94.
14. Bartel DP: **MicroRNAs: target recognition and regulatory functions**. *Cell* 2009, **136**(2):215-233.

15. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics**. *Nucleic Acids Res* 2008, **36**(Database issue):D154-158.

16. Friedman RC, Farh KK, Burge CB, Bartel DP: **Most mammalian mRNAs are conserved targets of microRNAs**. *Genome Res* 2009, **19**(1):92-105.

17. Lall S, Grun D, Krek A, Chen K, Wang YL, Dewey CN, Sood P, Colombo T, Bray N, Macmenamin P *et al*: **A genome-wide map of conserved microRNA targets in C. elegans**. *Curr Biol* 2006, **16**(5):460-471.

18. Quandt K, Frech K, Karas H, Wingender E, Werner T: **MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data**. *Nucleic Acids Res* 1995, **23**(23):4878-4884.

19. Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B: **Toucan: deciphering the cis-regulatory logic of coregulated genes**. *Nucleic Acids Res* 2003, **31**(6):1753-1764.

20. Frith MC, Li MC, Weng Z: **Cluster-Buster: Finding dense clusters of motifs in DNA sequences**. *Nucleic Acids Res* 2003, **31**(13):3666-3668.

21. Bigelow HR, Wenick AS, Wong A, Hobert O: **CisOrtho: a program pipeline for genome-wide identification of transcription factor target genes using phylogenetic footprinting**. *BMC Bioinformatics* 2004, **5**:27.

22. Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB: **MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model**. *Genome biology* 2004, **5**(12):R98.

23. Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y: **Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells**. *Genome Res* 2003, **13**(5):773-780.

24. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z: **Detection of functional DNA motifs via statistical over-representation**. *Nucleic Acids Res* 2004, **32**(4):1372-1381.

25. Haverty PM, Hansen U, Weng Z: **Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification**. *Nucleic Acids Res* 2004, **32**(1):179-188.

26. Zheng J, Wu J, Sun Z: **An approach to identify over-represented cis-elements in related sequences**. *Nucleic Acids Res* 2003, **31**(7):1995-2005.

27. Frith MC, Spouge JL, Hansen U, Weng Z: **Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences**. *Nucleic Acids Res* 2002, **30**(14):3214-3224.

28. Sharan R, Ovcharenko I, Ben-Hur A, Karp RM: **CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments**. *Bioinformatics* 2003, **19 Suppl 1**:i283-291.

29. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers**. *Proc Int Conf Intell Syst Mol Biol* 1994, **2**:28-36.

30. Moses AM, Chiang DY, Eisen MB: **Phylogenetic motif detection by expectation-maximization on evolutionary mixtures**. *Pac Symp Biocomput* 2004:324-335.

31. Prakash A, Blanchette M, Sinha S, Tompa M: **Motif discovery in heterogeneous sequence data**. *Pac Symp Biocomput* 2004:348-359.

32. Sinha S, Blanchette M, Tompa M: **PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences**. *BMC Bioinformatics* 2004, **5**:170.

33. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment**. *Science* 1993, **262**(5131):208-214.

34. Hughes JD, Estep PW, Tavazoie S, Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae**. *J Mol Biol* 2000, **296**(5):1205-1214.

35. Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouze P, Moreau Y: **A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes**. *J Comput Biol* 2002, **9**(2):447-464.

36. Sinha S, Tompa M: **Discovery of novel transcription factor binding sites by statistical overrepresentation**. *Nucleic Acids Res* 2002, **30**(24):5549-5560.

37. Eskin E, Pevzner PA: **Finding composite regulatory patterns in DNA sequences**. *Bioinformatics* 2002, **18 Suppl 1**:S354-363.

38. Keich U, Pevzner PA: **Finding motifs in the twilight zone**. *Bioinformatics* 2002, **18**(10):1374-1381.

39. Pavesi G, Mauri G, Pesole G: **An algorithm for finding signals of unknown length in DNA sequences**. *Bioinformatics* 2001, **17 Suppl 1**:S207-214.

40. Blanchette M, Schwikowski B, Tompa M: **Algorithms for phylogenetic footprinting**. *J Comput Biol* 2002, **9**(2):211-223.

41. Elemento O, Slonim N, Tavazoie S: **A universal framework for regulatory element discovery across all genomes and data types**. *Mol Cell* 2007, **28**(2):337-350.

42. Ettwiller L, Paten B, Ramialison M, Birney E, Wittbrodt J: **Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation**. *Nat Methods* 2007, **4**(7):563-565.

43. Workman CT, Stormo GD: **ANN-Spec: a method for discovering transcription factor binding sites with improved specificity**. *Pac Symp Biocomput* 2000:467-478.

44. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences**. *Bioinformatics* 1999, **15**(7-8):563-577.

45. Pevzner PA, Sze SH: **Combinatorial approaches to finding subtle signals in DNA sequences**. *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:269-278.

46. Buhler J, Tompa M: **Finding motifs using random projections**. *J Comput Biol* 2002, **9**(2):225-242.

47. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization**. *Mol Biol Cell* 1998, **9**(12):3273-3297.
48. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture**. *Nat Genet* 1999, **22**(3):281-285.
49. Wyrick JJ, Young RA: **Deciphering gene expression regulatory networks**. *Curr Opin Genet Dev* 2002, **12**(2):130-136.
50. Jiang D, Tang C, Zhang A: **Cluster analysis for gene expression data: A survey**. *IEEE Transactions on knowledge and data engineering* 2004, **16**(11):1370-1386.
51. Holmes I, Bruno WJ: **Finding regulatory elements using joint likelihoods for sequence and expression profile data**. *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:202-210.
52. Segal E, Yelensky R, Koller D: **Genome-wide discovery of transcriptional modules from DNA sequence and gene expression**. *Bioinformatics* 2003, **19 Suppl 1**:i273-282.
53. Reiss DJ, Baliga NS, Bonneau R: **Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks**. *BMC Bioinformatics* 2006, **7**:280.
54. Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression**. *Nat Genet* 2001, **27**(2):167-171.
55. Conlon EM, Liu XS, Lieb JD, Liu JS: **Integrating regulatory motif discovery and genome-wide expression analysis**. *Proc Natl Acad Sci U S A* 2003, **100**(6):3339-3344.
56. Eden E, Lipson D, Yogev S, Yakhini Z: **Discovering motifs in ranked lists of DNA sequences**. *PLoS computational biology* 2007, **3**(3):e39.
57. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ *et al*: **Assessing computational tools for the discovery of transcription factor binding sites**. *Nat Biotechnol* 2005, **23**(1):137-144.
58. Linhart C, Shamir R: **The degenerate primer design problem: theory and applications**. *J Comput Biol* 2005, **12**(4):431-456.
59. Fuchs T, Malecova B, Linhart C, Sharan R, Khen M, Herwig R, Shmulevich D, Elkon R, Steinfath M, O'Brien JK *et al*: **DEFOG: a practical scheme for deciphering families of genes**. *Genomics* 2002, **80**(3):295-302.
60. Olender T, Fuchs T, Linhart C, Shamir R, Adams M, Kalush F, Khen M, Lancet D: **The canine olfactory subgenome**. *Genomics* 2004, **83**(3):361-372.
61. Knuth D, Morris Jr J, Pratt V: **Fast pattern matching in strings**. *SIAM Journal on Computing* 1977, **6**:323.
62. Boyer R, Moore J: **A fast string searching algorithm**. *Communications of the ACM* 1977, **20**(10):762-772.
63. Gusfield D: **Algorithms on Strings, Trees, and Sequences**: Cambridge University Press; 1997.

64. Fischer M, Paterson M: **String Matching and Other Products**. In: *Complexity of Computation, SIAM-AMS Proceedings.* Edited by Karp RM; 1974: 113-125.

65. Cole R, Hariharan R: **Verifying candidate matches in sparse and wildcard matching**. *Proc 34th Symposium on Theory of Computing* 2002:592-601.

66. Clifford P, Clifford R: **Simple deterministic wildcard matching**. *Information Processing Letters* 2007, **101**(2):53-54.

67. Baeza-Yates R, Gonnet G: **A new approach to text searching**. *Communications of the ACM* 1992, **35**(10):74-82.

68. Amir A, Lewenstein M, Porat E: **Faster Algorithms for String Matching with *k* Mismatches**. *J Algorithms* 2004, **50**(2):257-275.

69. Abrahamson K: **Generalized string matching**. *SIAM Journal on Computing* 1987, **16**:1039-1051.

70. Clifford R, Efremenko K, Porat E, Rothschild A: ***k*-Mismatch with Don't Cares**. *Proc 15th Annual European Symposium on Algorithms* 2007:151-162.

71. Clifford R, Efremenko K, Porat E, Rothschild A: **From coding theory to efficient pattern matching**. *Proc 19th Symposium on Discrete Algorithms* 2009:778-784.

72. Ulitskyi I, Shamir R: **Towards computational prediction of MicroRNA function and activity**. In preparation, 2009.

73. Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E: **Tissue-specific transcriptional regulation has diverged significantly between human and mouse**. *Nat Genet* 2007, **39**(6):730-732.

74. Lin CY, Vega VB, Thomsen JS, Zhang T, Kong SL, Xie M, Chiu KP, Lipovich L, Barnett DH, Stossi F *et al*: **Whole-genome cartography of estrogen receptor alpha binding sites**. *PLoS genetics* 2007, **3**(6):e87.

75. Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M: **Divergence of transcription factor binding sites across related yeast species**. *Science* 2007, **317**(5839):815-819.

76. D'Haeseleer P, Liang S, Somogyi R: **Genetic network inference: from co-expression clustering to reverse engineering**. *Bioinformatics* 2000, **16**(8):707-726.

77. Heyer LJ, Kruglyak S, Yooseph S: **Exploring expression data: identification and analysis of coexpressed genes**. *Genome Res* 1999, **9**(11):1106-1115.

78. Sharan R, Shamir R: **CLICK: a clustering algorithm with applications to gene expression analysis**. *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:307-316.

79. De Smet F, Mathys J, Marchal K, Thijs G, De Moor B, Moreau Y: **Adaptive quality-based clustering of gene expression profiles**. *Bioinformatics* 2002, **18**(5):735-746.

80. Tanay A, Sharan R, Shamir R: **Biclustering algorithms: A survey**. In: *Handbook of Computational Molecular Biology.* Edited by Aluru S: Chapman and Hall/CRC Press; 2006: 1-17 (Ch. 26).

81. Halperin Y: **Discovery of motifs involved in transcriptional regulation**. *M.Sc. thesis.* Tel Aviv University; 2008.

82. Muthukrishnan S: **New results and open problems related to non-standard stringology**. *Proc 6th Combinatorial Pattern Matching Conference* 1995:298-317.

83. Amir A, Cole R, Hariharan R, Lewenstein M, Porat E: **Overlap matching**. *Information and Computation* 2003, **181**(1):57-74.

TEL AVIV UNIVERSITY אוניברסיטת תל-אביב

הפקולטה למדעים מדויקים ע"ש ריימונד ובברלי סאקלר

בית הספר למדעי המחשב

חיפוש מוטיבים ברצפי דנ"א ארוכים

חיבור לשם קבלת תואר "דוקטור לפילוסופיה"

מאת **חיים לינהרט**

בהנחייתם של פרופ' **רון שמיר**

ופרופ' **יוסי שילה**

הוגש לסנאט של אוניברסיטת ת"א

יולי 2009

תמצית

אתגר מרכזי בביולוגיה מערכתית הוא פענוח מפת הבקרה של הגנום, המתארת כיצד התא שולט על הכמות וההרכב המדויק של החלבונים אותם הוא מייצר מכל גן בתנאים נתונים. מרכיב חשוב במאמץ זה הינו פתרון הבעיה החישובית של חיפוש מוטיבים (motif finding), תבניות רצף קטנות שמופיעות בשינויים קלים פעמים רבות לאורך הגנום. מוטיבים מייצגים אתרי קישור של גורמי שעתוק ומיקרו-רנ"א, הרכיבים העיקריים של מנגנון בקרת השעתוק בתא. אנחנו חקרנו את הבעיה של חיפוש מוטיבים מן ההיבט המעשי והתיאורטי. פיתחנו אלגוריתמים, מודלים חישוביים ומבחנים סטטיסטיים חדשים ויישמנו אותם בחבילת תוכנה יעילה וידידותית למשתמש. הגישה שלנו מדויקת ומהירה יותר משיטות קיימות וניתן להשתמש בה למגוון רחב של בעיות חיפוש מוטיבים. ניתחנו את מפות בקרת השעתוק של שני תהליכים מרכזיים בתא האנושי – מחזור התא ומערכת החיסון המולדת. על ידי שילוב מספר מקורות מידע שונים, הצלחנו לנבא אתרי קישור בדיוק רב מאד, וחשפנו רכיבים חדשים במפות הבקרה. בצד התיאורטי, פיתחנו אלגוריתמים יעילים חדשים עבור מספר סוגים של בעיות התאמת תבניות (pattern matching) הקשורות לחיפוש מוטיבים. השיטות שלנו משתמשות ברעיונות מתורת המספרים, והן פשוטות יותר, ובתנאים מסוימים גם מהירות יותר, מפתרונות קיימים.

תקציר

בקרת השעתוק בתא

התא החי הוא מכונה מסובכת להפליא, המסוגלת לבצע מגוון רחב של פעולות בסביבה שמשתנה ללא הרף. סוגים שונים של תאים בגוף פועלים באופן שונה, אף על פי שלכולם אותו דנ"א בדיוק. שיטות ביו-טכנולוגיות חדשות מאפשרות לקבל תמונת מצב של התא בעת פעילותו – למשל, שבבי דנ"א מודדים את רמת הביטוי, כלומר ריכוז הרנ"א השליח (mRNA), של אלפי גנים בו זמנית [7]. על ידי השוואת רמות הביטוי בתאים שונים או בתנאים שונים ניתן ללמוד על תפקידם של הגנים והקשרים ביניהם. על מנת להבין כיצד כל תא קובע את ריכוזי החלבונים ומשנה אותם בעת הצורך, עלינו לפענח את תוכניות הבקרה (regulatory programs) של התא. אחד ממנגנוני הבקרה העיקריים הוא בקרת השעתוק (transcriptional regulation). מנגנון זה מורכב מ**גורמי שעתוק** (transcription factors) - חלבונים מיוחדים שנקשרים לרצפי דנ"א קצרים, לרב בקרבת גנים באזורים הקרויים פרומוטורים. כאשר גורם שעתוק נקשר לפרומוטור של גן מסוים, הוא יכול להעלות או להוריד את רמת הביטוי של הגן. מספר גורמי שעתוק יכולים אף לפעול יחד וליצור תבנית ביטוי מורכבת לגנים אותם הם מבקרים. אתרי הקישור של גורם שעתוק הם קצרים (6-15 בסיסים) ומנוונים – גורם שעתוק יכול לקשור מספר רב של רצפים שונים בעלי תבנית משותפת, או **מוטיב**, שייחודית לאותו גורם שעתוק. מנגנון בקרה נוסף הינו **המיקרו-רנ"א**, מולקולת רנ"א קצרה שנקשרת לרצף ה-3' UTR של גן המטרה, ובכך מעודדת את פירוק הרנ"א השליח של אותו גן או מונעת את תרגומו לחלבון.

חיפוש מוטיבים

פענוח רשתות בקרה מעלה מספר אתגרים חישוביים. ראשית, עבור גורם שעתוק (או מיקרו-רנ"א) עם מוטיב קשירה ידוע, נרצה למצוא את אתרי הקשירה שלו, ומכאן גם את הגנים אותם הוא מבקר, בדיוק רב ככל האפשר [18-22]. שנית, בהינתן קבוצה של גנים בעלי פרופיל ביטוי דומה בניסוי כלשהו, או בעלי פונקציה ביולוגית משותפת, נרצה לגלות את המוטיבים שדרכם מתבצעת הבקרה שלהם. לצורך כך פותחו שיטות שבודקות אילו מבין המוטיבים *הידועים* מועשרים בפרומוטורים (או ב-3' UTRs) של קבוצת גנים נתונה, כלומר אילו תבניות רצף מופיעות במספר רב מן הצפוי של פרומוטורים, באופן שהוא מובהק מבחינה סטטיסטית [23-28]. בעיה קשה עוד יותר היא מציאת מוטיבים *חדשים* שמועשרים ברצפים נתונים. מספר רב של אלגוריתמים פותחו בשנים האחרונות על מנת לנסות לפתור בעיה זו [29-46]. חלקם מנתחים את רמות הביטוי של הגנים ואת רצפי הבקרה שלהם בו-זמנית, על מנת לגלות קבוצות של גנים בעלי פרופיל ביטוי משותף יחד עם המוטיבים שמבקרים כל קבוצה [51-56].

מכיון שמוטיבים הם קצרים ומנוונים, ויש לחפשם בתוך רצפי פרומוטורים ארוכים שמכילים חתימות ביולוגיות שונות ומגוונות, ומכיון שהנתונים מניסויים רחבי היקף (כגון רמות הביטוי של הגנים) מכילים רעש מדידה רב, הרי שבעיית חיפוש המוטיבים מהווה אתגר חישובי קשה שמצריך אלגוריתמים יעילים וציונים סטטיסטיים מתקדמים. הכלים הקיימים כיום לחיפוש מוטיבים לא מניבים תוצאות טובות די הצורך, במיוחד באורגניזמים כמו החולייתנים, שרשתות הבקרה שלהם מסובכות [57].

התאמת תבניות

בבעיית התאמת תבניות (pattern matching) נתונים תבנית באורך m וטקסט באורך n מעל אלפבית Σ, והמטרה היא למצוא את כל המקומות בטקסט שמתאימים לתבנית. כאשר התבנית והטקסט הם מחרוזות פשוטות, ניתן לפתור את הבעיה בזמן ליניארי על ידי אלגוריתמים קלאסיים [61,62]. הבעיה נעשית קשה יותר כאשר התבנית מכילה תווים חופשיים (wildcards / don’t-cares). תו חופשי, שמסומן על פי רב ב-'*', הוא תו שמתאים לכל תו ב-Σ. אלגוריתם ה-match-count פותר את הבעיה בזמן $O(|\Sigma|\, n \log m)$ על ידי חישוב קונבולוציות (convolutions) באמצעות אלגוריתם ה-FFT (ראה, למשל, פרק 4.3 ב-[63]). במהלך השנים האחרונות פותחו שיטות יעילות יותר, כולן מבוססות FFT. זמן הריצה של האלגוריתמים המהירים ביותר כיום הוא $O(n \log m)$ [65,66]. הכללה של הבעיה הנ"ל מתקבלת כאשר כל עמדה בתבנית מכילה קבוצה כלשהי של תווים אפשריים. בעיה זו, שנקראת **התאמת תבניות עם קבוצות תווים** (pattern matching with character classes), קשורה לחיפוש מוטיבים – תבנית עם קבוצות תווים היא מוטיב שמייצג אתרי קישור. האלגוריתם המהיר ביותר כיום לפתרון הבעיה הזו הוא אלגוריתם ה-match-count. הכללה חשובה נוספת של התאמת תבניות עם תווים חופשיים מתקבלת כאשר מרשים מספר קטן של שגיאות בכל מקום בטקסט שמתאים לתבנית. בעיה זו נקראת **התאמת תבניות עם תווים חופשיים ו-k אי-התאמות** (pattern matching with don’t-cares and k mismatches). ניתן לפתור אותה באמצעות אלגוריתם ה-match-count בזמן $O(|\Sigma|\, n \log m)$, או על ידי שיטת Abrahamson, שמשלבת את match-count עם טכניקת "הפרד ומשול", בזמן $O(n\sqrt{m \log m})$. לאחרונה פותחו אלגוריתמים יעילים יותר; המהיר ביותר רץ בזמן $O(nk \log^2 m\ (\log^2 k+\log\log m))$ [70,71]. שאלה פתוחה מעניינת היא האם ניתן לפתור את הבעיה בזמן $O(\mathrm{poly}(k)\, n \log m)$. בפרט, האם ניתן להכליל את סיבוכיות הזמן $O(n \log m)$ של התאמת תבניות עם תווים חופשיים כך שתחול גם כאשר מרשים מספר קטן (קבוע) של שגיאות?

תקציר המאמרים הכלולים בתזה

להלן תקצירי המאמרים עליהם מבוססת עבודה זו:

1. **Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets.**
Chaim Linhart, Yonit Halperin and Ron Shamir.
Published in *Genome Research* [1].

אנו מציגים תרומה משולשת לבעיה החישובית של חיפוש מוטיבים, מרכיב מרכזי במאמץ לפענוח מפת הבקרה של הגנום: (1) יצרנו אוסף גדול ומקיף של קבוצות גנים שמבוקרים על ידי גורמי שעתוק ומיקרו-רנ"א; כל הקבוצות דווחו בספרות ונבנו על סמך סוגים שונים של ניסויים רחבי היקף (high-throughput experiments) במספר מינים של בעלי-חיים (metazoans). השתמשנו באוסף זה על מנת להשוות את הביצועים של שיטות שונות לחיפוש מוטיבים; (2) פיתחנו את אמדאוס, תוכנה יעילה וידידותית למשתמש לגילוי מוטיבים חדשים, שמתאימה למגוון רחב של משימות. אמדאוס מדויקת יותר, מהירה יותר וקלה יותר לשימוש מכלים קיימים, והיא התוכנה היחידה שהשיגה אחוז הצלחה גבוה על הנתונים שבאוסף שלנו; (3) אנו מדגימים שבאמצעות חיפוש מוטיבים על סמך הריכוז שלהם לאורך הפרומוטורים או הפיזור שלהם בין הכרומוזומים (ללא שימוש בקבוצת מטרה נתונה), אמדאוס יכולה לגלות תופעות ידועות מגוונות, וכן לחשוף מוטיבים חדשים.

2. **Allegro: Analyzing expression and sequence in concert to discover regulatory programs.**
Yonit Halperin, Chaim Linhart, Igor Ulitsky and Ron Shamir.
Published in *Nucleic Acids Research* [2].

אחת המשימות המרכזיות בביולוגיה מערכתית היא מיפוי ואפיון גורמי השעתוק והמיקרו-רנ"א שמבקרים את תוכניות השעתוק של התא. אנו מציגים את אלגרו, שיטה לגילוי תוכניות שעתוק חדשות באמצעות ניתוח משותף של נתוני ביטוי גנים (gene expression datasets) ושל רצפי פרומוטורים או 3’ UTRs. האלגוריתם משתמש במודל א-פרמטרי חדש, שמבוסס על ציון נראות, על מנת לתאר את תבנית הביטוי של קבוצת גנים שמבוקרים יחדיו. אנו מראים שאלגרו מדויקת יותר מטכניקות קיימות, ומסוגלת לנתח יחד מספר אוספי נתונים עם למעלה מ-100 תנאים בכל אחד. הרצנו את אלגרו על נתוני ביטוי ממספר אורגניזמים, ואנו מדווחים על תוכניות השעתוק שמצאנו. הניתוח שלנו חושף מוטיב חדש, שמועשר בפרומוטורים של גנים שמבוטאים בביציות של עכברים, ומספר מוטיבים חדשים שקשורים לתהליכי ההתפתחות של הזבוב. לבסוף, על סמך נתוני ביטוי של תאי גזע, אנו מזהים שלוש משפחות של מיקרו-רנ"א עם תפקידי מפתח בהתפתחות העובר באדם.

3. **Deciphering transcriptional regulatory elements that encode specific cell-cycle phasing by comparative genomics analysis.**
   Chaim Linhart, Ran Elkon, Yosef Shiloh and Ron Shamir.
   Published in *Cell Cycle* [3].

בקרת שעתוק הינה מרכיב מרכזי במנוע המחזורי שמניע את מחזור התא. פענוחם של גנומים שלמים של מספר אורגניזמים טומן בחובו את האפשרות לשפר באופן משמעותי את הדיוק של ניבוי חישובי של אלמנטים פונקציונאליים. בעבודה זו השתמשנו בשיטות של גנומיקה השוואתית על מנת לפענח את רכיבי הבקרה ששולטים בשלבים של מחזור התא. ניתחנו רצפי פרומוטורים מ-12 אורגניזמים, כולל תולעת, זבוב, דג, עכבר ואדם, וזיהינו רכיבי שעתוק שמורים, שקובעים את רמות הביטוי של גנים בשלבים השונים של מחזור התא. אנו מראים שתבנית קישור קנונית של E2F קשורה לרמות ביטוי גבוהות בשלב ה-G_1/S, ושרצפי הקישור של NF-Y ו-CHR מופיעים יחד בגנים שמבוטאים ב-G_2 וב-G_2/M. אתרי קישור של b-Myb נמצאים גם כן בגנים המבוקרים על ידי NF-Y ו-CHR, מה שמצביע על תפקיד ייחודי של השלשה הזו בתכנית הבקרה של מחזור התא. תופעה מעניינת נוספת היא שבעוד שהחתימה של E2F שמורה בפרומוטורים של גנים שמבוטאים ב-G_1/S בכל האורגניזמים שבדקנו, מתולעת ועד אדם, הרי שהזוג CHR-NF-Y מופיע בפרומוטורים של גנים של G_2/M רק בקרב החולייתנים. התוצאות שקיבלנו חושפות מרכיבים חדשים ששולטים בשלבי מחזור התא ומזהות ברמת דיוק גבוהה במיוחד את הגנים שמבוקרים על ידם.

4. **Functional genomic delineation of TLR-induced transcriptional networks.**
   Ran Elkon, Chaim Linhart, Yonit Halperin, Yosef Shiloh and Ron Shamir.
   Published in *BMC Genomics* [4].

מערכת החיסון המולדת הינה קו ההגנה הראשון נגד פלישת פתוגנים כמו חיידקים ווירוסים. התגובות של מערכת זו מופעלות על ידי קולטנים בתא שמזהים רכיבים טיפוסיים המשותפים לפתוגנים רבים. החיישנים העיקריים הם ה-TLRs (Toll-like receptors). על מנת לקבל ניתוח גלובלי של רשתות השעתוק המופעלות על ידי TLRs, השתמשנו בארבע קבוצות נתונים של ביטוי גנים במקרופאג'ים בעכבר ובאדם, שעוררו באמצעות חומרים דמויי פתוגנים. גילינו שתי תוכניות עיקריות: האחת אוניברסאלית, כלומר מופעלת על ידי כל ה-TLRs שנבדקו, והשנייה מופעלת רק על ידי TLR3 ו-TLR4. גורמי השעתוק המרכזיים ששולטים בתוכניות האלה הם NF-κB וחלבונים שקושרים ISRE, בהתאמה. ניתחנו גם את הקינטיקה של רשתות הבקרה הללו ומצאנו שבעוד ש-NF-κB מבקר בעיקר תגובה מוקדמת ומתמשכת, הרי שאלמנט ה-ISRE מפעיל גל ביטוי מאוחר יותר. גילינו שהופעה משותפת של שני אתרי הקישור, NF-κB ו-ISRE, באותו פרומוטור גורמת לרמות ביטוי גבוהות יותר של הגן המתאים. התוצאות שקיבלנו מרחיבות את הידע הקיים אודות התוכניות שמופעלות על ידי TLRs, ומדגימות את הכוח של גישות חישוביות בניתוח מדויק של רשתות שעתוק סבוכות ביונקים. הבנה כזו של רשת ה-TLRs יכולה לתרום לתכנון תרופות שישפיעו על פעילות מערכת החיסון המולדת במחלות שונות, שבהן יש לעודד או לדכא ענף זה של מערכת החיסון.

5. **Faster Pattern Matching with Character Classes using Prime Number Encoding.**
Chaim Linhart and Ron Shamir.
Published in *Journal of Computer and System Sciences* [5].

בבעיית התאמת תבניות עם קבוצות תווים (pattern matching with character classes) המטרה היא למצוא את כל ההופעות של תבנית באורך m בתוך טקסט באורך n, כאשר כל עמדה בתבנית מכילה קבוצה של תווים מורשים מתוך אלפבית סופי Σ. אנו מציגים אלגוריתם מבוסס FFT לפתרון הבעיה. האלגוריתם שלנו משתמש בקידוד באמצעות מספרים ראשוניים, והוא מהיר פי $\log n/\log m$ מהשיטות הקיימות המהירות ביותר. בפרט, אם $m^{|\Sigma|}=n^{O(1)}$, האלגוריתם רץ בזמן $O(n \log m)$; זוהי סיבוכיות הזמן של השיטות היעילות ביותר כיום להתאמת תבניות עם תווים חופשיים (wildcard matching), שהיא מקרה פרטי של הבעיה שאנו פותרים. יתרון חשוב נוסף של האלגוריתם שלנו הינו שהוא מאפשר להקטין את זמן הריצה ככל שמילת המחשב (RAM word) ארוכה יותר. האלגוריתם משפר גם את הסיבוכיות של מציאת התאמות מקורבות של תבנית עם קבוצות תווים – חיפוש עם k אי-התאמות וחישוב מרחק Hamming, וכן של בעיית התאמת תתי קבוצות (subset matching).

6. **Matching with don't-cares and a small number of mismatches.**
Chaim Linhart and Ron Shamir.
Published in *Information Processing Letters* [6].

בבעיית התאמת תבניות עם תווים חופשיים ו-k אי-התאמות (pattern matching with don’t-cares and k mismatches) נתונה תבנית באורך m וטקסט באורך n שמכילים תווים חופשיים (תו חופשי הוא תו שמתאים לכל תו אחר) והמטרה היא למצוא את כל המקומות בטקסט שמתאימים לתבנית עם k שגיאות לכל היותר. פיתחנו אלגוריתמים חדשים שפותרים בעיה זו באמצעות שילוב של קונבולוציות ותכנות דינמי – אלגוריתם רנדומי שרץ בזמן $O(nk^2\log m)$, ואלגוריתם דטרמיניסטי שסיבוכיותו $O(nk^3\log m)$. עבור ערכים קטנים של k, האלגוריתמים שלנו מהירים יותר מהשיטות הקיימות היעילות ביותר. האלגוריתם הדטרמיניסטי שלנו הוא הראשון שפותר את הבעיה בזמן $O(\text{poly}\ (k) \cdot n\log m)$.