Statistical Assessment of Enrichment in Ranked Lists -
Algorithms and Applications in Motif Search

Limor Leibovich
Statistical Assessment of Enrichment in Ranked Lists - Algorithms and Applications in Motif Search

Research Thesis
In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Limor Leibovich

Submitted to the Senate of the Technion - Israel Institute of Technology
Nisan, 5774     Haifa     April, 2014
The research thesis was done under the supervision of Dr. Zohar Yakhini in the Computer Science Department

Acknowledgments

First and foremost, I am grateful to my advisor, Zohar Yakhini, for his incredible guidance and support through all these years. Zohar has patiently helped me make my first steps as a researcher, providing me with inspiration, freedom, encouragement and warmth. I thank him for opening doors for me whenever he could, for intellectually challenging me, and for caring about me personally. His openness, brilliance and clarity made working with him ever fun and fruitful. I feel lucky to have had Zohar as my advisor and I am sure that the education I have gotten from him will serve me well in the future.

During my years in the Technion Computer Science Department, first as a B.Sc. student and later in the direct PhD track, I received the best education imaginable. The quality of the teaching, curriculum, and training were superb. I am especially grateful to Dan Geiger for his inspiration, support and wise advice throughout all these years; and to Ron Pinter for his help, open door policy, and encouragement. I want to thank my colleagues in teaching - Ron Pinter, Irad Yavneh, Shlomo Moran and Zohar Yakhini - from whom I learned so much about teaching and beyond. I am also thankful to the administrative staff, and especially to Yardena Kolet for her help and support.

I am greatly appreciative of my collaborators over the years: Yael Mandel-Gutfreund, who kept reminding me to be loyal to biology and whose passion for the field is contagious; and Yoav Arava, Inbal Paz, Doron Betel, Aldema Sas-Chen, and Meromit Singer. I owe huge gratitude to my wonderful colleagues at the Yakhini research group - Israel Steinfeld, Roy Navon, Ilona Kifer, Yaara David, Shay Ben-Elazar and Edward Vitkin - who have constantly contributed to my progress as a researcher. They inspired and supported me, and I am honored to have been part of this unique group. I also want to thank my adorable office-mate, Kayla Jacobs, for her encouragement and for being such a generous, loving, and good friend. I will miss our deep conversations about everything, as well as her little chocolate surprises.

Affectionate thanks to my family: my parents - Avraham and Ruth, and my siblings Israel, Merav and Lior. Thank you all for the support and love throughout the years. I am proud to see each of you forging your own unique paths so successfully and gracefully. I also want to thank my family-in-law, Malka, Shalom and Dovrat, for their endless support and generosity.

Finally and most importantly, my deepest thanks to my husband, Lior, and my daughter, Lia, for their love and for making me so happy. Lior, thank you for your friendship, for your great dedication to my success, and for advising me to choose computer science as my profession - I could not have made a better decision. I look forward joyfully to our future adventures together, and I am pleased to dedicate this dissertation to you.

The generous financial help of the Technion- Israel Institute of Technology and of Sherman Fellowship, Jacobs Fellowship, Emerson Fellowship, Ministry of Science and Technology Fellowship for Women, Google Anita Borg Memorial Scholarship, Belmonte Family Prize and ISEF Fellowship are gratefully acknowledged.
List of Publications

Papers appearing in the Results section of this dissertation:


Other papers written during the doctorate:


# Contents

Abstract ........................................................................................................................................... 1

1 Introduction ................................................................................................................................. 4
   1.1 Motif search .......................................................................................................................... 4
   1.2 Variable gap motif search .................................................................................................... 7
   1.3 Enrichment analysis using the minimum hyper-geometric (mHG) statistics .......... 8
   1.4 Suffix trees .......................................................................................................................... 10
   1.5 Position weight matrix motifs ........................................................................................... 11
   1.6 DRIMust – discovering rank imbalanced motifs using suffix trees ......................... 13
   1.7 Assessing position weight matrix (PWM) motifs in ranked lists .................... 14
   1.8 Non-coding RNAs ................................................................................................................ 15
   1.9 Cooperative interactions of RNA binding proteins and microRNAs ............. 17

2 Methods ....................................................................................................................................... 19
   2.1 Efficient enumeration of the motif search space using suffix trees ................. 19
   2.2 Variable gap motif search ................................................................................................. 20
   2.3 Expanding k-mers into position weight matrix motifs .............................................. 22
   2.4 Minimum-minimum hyper-geometric statistics (mmHG) ........................................ 22
   2.5 PWM motif search in ranked lists (using mmHG-Finder) ........................................ 24
   2.6 Developing statistical bounds for mHG and mmHG .................................................. 25
   2.7 Pattern matching for IUPAC motifs .................................................................................. 28
   2.8 Using Kd-trees for multi-dimensional sorting ................................................................. 29
   2.9 Expression data normalization ......................................................................................... 31

3 Results ........................................................................................................................................ 33
   3.1 A structural-based statistical approach suggests a cooperative activity of PUM1
       and miR-410 in human 3'-untranslated regions ................................................................. 33
   3.2 Efficient motif search in ranked lists and applications to variable gap motifs .... 49
   3.3 DRIMust: a web server for discovering rank imbalanced motifs using suffix
       trees ...................................................................................................................................... 67
   3.4 Mutual Enrichment in Ranked Lists and the Statistical Assessment of Position
       Weight Matrix Motifs ........................................................................................................ 75
Contents - cont.

3.5 Mutual Enrichment in Ranked Lists and the Statistical Assessment of Position
  Weight Matrix Motifs (extended version) .................................................. 91
4 Discussion .................................................................................................. 107
  4.1 Motif discovery and molecular measurement data .................................. 107
    4.1.1 Efficient enumeration of the motif search space ............................... 108
    4.1.2 Discovery of variable gapped motifs .............................................. 109
    4.1.3 Software implementation and the DRIMust webserver ..................... 111
    4.1.4 Searching position weight matrix motifs in ranked lists ................. 113
  4.2 Statistics in ranked lists ....................................................................... 116
    4.2.1 Comparison between mHG and other statistical methods ............... 116
    4.2.2 Bounds on mmHG ......................................................................... 119
  4.3 Biological results .................................................................................. 122
    4.3.1 Cooperative association between an RNA binding protein and a
      microRNA .............................................................................................. 122
    4.3.2 Motif discovery .............................................................................. 124
      4.3.2.1 Discovery of novel motif recognized by PUF2 ......................... 124
      4.3.2.2 Refinement of known motifs using variable gapped motif search ...
      4.3.3 GC-rich motifs found amongst promoters of tissue specific long-non-
      coding RNAs ....................................................................................... 125
  4.4 List of open questions ........................................................................... 126
References ................................................................................................... 127
List of Figures

Figure 1 – Binary vector representation for the occurrences of the pattern UGUAWAUW in a hypothetical example input ranked list .................................................. 9
Figure 2 - A generalized suffix tree for $S_1 = \text{CAGCA}$ and $S_2 = \text{GA}$ ...................... 11
Figure 3 – A position weight matrix and its corresponding Shannon logo representation 12
Figure 4 – DRIMust’s flow chart ................................................................. 14
Figure 5 – Enumeration of variable gap motif candidates using suffix trees ........ 21
Figure 6 – The relative permutation given two ranked lists ................................. 26
Figure 7 – The IUPAC notation (taken from Wikipedia) ........................................ 29
Figure 8 – Sorting points in a two-dimensional grid ............................................. 30
Figure 9 – Average running time of DRIMust compared to three other motif discovery tools ........................................................................................................... 112
Figure 10 – An example demonstrating the limitation of searching an IUPAC word instead of a PWM ...................................................................................... 115
Figure 11 – The distribution of $p$-values for Kolmogorov-Smirnov, mHG and Wilcoxon tests ........................................................................................................ 119
Figure 12 – Chvatal bound vs. HGT value (minus log-scale) .................................... 121

List of Tables

Table 1 – Number of vectors (out of 1000) where the test $p$-value $\leq \beta_0$ .............. 118
Abstract

It is often the case in biological measurement data that results are given in the form of a ranked list of quantities. In recent years we have seen considerable progress in statistical tools for enrichment analysis in ranked lists. Several tools and web applications are now available that allow users to break the fixed set paradigm in assessing statistical enrichment of sets of genes or other elements. In parallel we are witnessing great progress in understanding the role of sequence elements in driving and mediating important biological processes such as transcription, RNA degradation and translocation as well as protein-protein interaction. Sequence elements, at all levels – DNA, RNA and protein, play a central role in mediating molecular recognition and thereby molecular regulation and signaling. Studies that focus on measuring and investigating sequence based recognition make use of statistical and computational tools, including approaches to searching sequence motifs.

State of the art motif finding tools come short, however, in their ability to address searches over large motif spaces. In the first part of the work described in this dissertation, we developed statistical and algorithmic approaches that take as input ranked lists of sequences and return significant motifs. The efficiency of our approach, based on suffix trees, allows searches over motif spaces that are not covered by existing tools. This includes searching variable gap motifs – two half sites with a flexible length gap in between – and searching long motifs over large alphabets. Our work provides a practical solution addressing these challenges.

In collaboration with Yael Mandel-Gutfreund’s Lab at the Technion, we also developed a new web-application named DRIMust, which makes some of the above methods accessible to the research community (http://drimust.technion.ac.il/). DRIMust provides de-novo motif discovery services and addresses short motifs in DNA, RNA and protein sequences. DRIMust is computationally very efficient and allows for timely interaction with the results, through a friendly interface and a clear output format.

Further extending the applicability of the methods, we worked on developing an approach for assessing the significance of position weight matrix motifs in ranked lists of sequences. A position weight matrix (PWM) is a commonly used representation of motifs in biological sequences. This representation is more faithful to the underlying biology than representation by
exact words. Currently, to the best of our knowledge, there is no statistical methodology for assessing PWM motifs in ranked lists. We developed upper bounds on tail distributions that are applicable in the context of assessing PWM motifs in ranked lists of sequences, improving over the existing knowledge in this respect. The bounds can be calculated in polynomial time.

Finally, in a study that applies statistics in ranked lists, we studied co-operativity between RNA binding proteins and microRNAs. We analyzed human RNA sequences containing potential binding sites of 153 conserved microRNA (miRNA) families, and ranked sequences around the sites according to their miRNA accessibility. By applying rank-based motif searching to these miRNA targets, we found motifs that are enriched among less accessible targets. Among miR-410 low accessibility targets we found a motif that resembles the Pumilio homolog 1 (PUM1) motif. We further tested this pair and our findings suggest a cooperative relationship between miR-410 and PUM1 in regulating human highly structured 3'-untranslated regions.

In summary, our contribution is in:

- Methods developed in this work. This includes efficient motif search methodology which allows searches over large spaces and addresses the discovery of patterns composed of two separate parts. We also expanded the knowledge of statistics in ranked lists by developing bounds on tail distributions.
- Bioinformatics tools. Our methods and algorithms are implemented in software which is publicly available. Specifically, some of our methods are available through the DRIMust webserver which allows de-novo motif search.
- Biological results. Our results reveal novel regulation mechanisms for microRNA and other non-coding RNAs. Using our motif discovery tools we were able to identify a novel recognition pattern for Puf2p in S. Cerevisiae and also to refine known motifs.

The overall structure of this dissertation

In the Introduction section we present relevant background which gives context to different aspects of our work, including statistics and algorithms. We then describe the methods and technical details related to statistics and algorithmics in the Methods section. The papers appear consecutively in the Results section. Finally, we address significance, advantages and limitations of our work and biological results in the Discussion section.
**Abbreviations and notations**

DNA - Deoxyribonucleic acid  
RNA - Ribonucleic acid  
mRNA - Messenger RNA  
mRNA – Micro RNA  
ncRNA – Non-coding RNA  
lncRNA – Long non-coding RNA  
UTR – Untranslated region  
RBP – RNA binding protein  
TF – transcription factor  
HG - Hypergeometric  
HGT - Hypergeometric tail  
mHG – Minimum hypergeometric  
mnHG – Minimum-minimum hypergeometric  
VGM – Variable gapped motif  
PWM – Position weight matrix  
IC – Information content  
DRIMust – Discovering rank imbalanced motifs using suffix trees  
PUF – Pumilio family  
GSEA – Gene set enrichment analysis  
bp – base pair  
nt - nucleotide  
PAR-CLIP – Photoactivatable ribonucleoside enhanced crosslinking and immunoprecipitation  
ChIP - Chromatin immunoprecipitation  
IUPAC alphabet - Degenerate base symbols in DNA/RNA patterns can be represented using the IUPAC alphabet.
1 Introduction

1.1 Motif search

The central dogma in molecular biology, stated first at 1958 by Francis Crick, explains the flow of genetic information within a biological system: DNA molecules are used as template for newly assembled RNA molecules, called messenger RNA, in the process of transcription. Then, these RNA molecules are translated into proteins. Proteins are known to be key players in living cells by performing a vast array of functions, including catalyzing metabolic reactions, replicating DNA, responding to stimuli, and transporting molecules from one location to another.

DNA and RNA molecules are built of nucleotides, and since there are four types of nucleotides for both DNA and RNA, they can be mathematically represented as sequences taken from an alphabet of size 4. Proteins are sequences of amino acid residues, where there are 20 various amino acids. The sequence of a protein is defined by the sequence of a gene, encoded in the genetic code – which maps each triplet of nucleotides to an amino acid. In many biological analyses, including those presented in my work, DNA, RNA and protein sequences are analyzed, mainly because chemical properties of DNA, RNA and protein molecules are encoded in their sequence.

Gene expression begins with binding of multiple protein factors, known as transcription factors, to enhancer and promoter sequences. Transcription factors regulate the expression of genes by activating or inhibiting the transcription machinery. Identifying regulatory elements, for example - the binding sites in DNA for transcription factors, is a major task for computational biology, contributing to our general understanding of transcription regulation.

While transcription is a major control point of gene expression, the gene's transcript is further subjected to regulation at the levels of RNA processing (splicing and editing), transport, localization, stability, translation and degradation. RNA-binding proteins (RBPs) take part in post-transcriptional regulation and processing of messenger RNA (mRNA) by binding to cis-acting mRNA elements [1-5]. These elements are often located at the 5′ or 3′-untranslated regions (5′ or 3′-UTR) of mRNAs [6]. Many RNA-binding proteins have domains that bind RNA in a sequence-specific manner. For example, Pumilio family (PUF) is a conserved family of RNA binding proteins that regulate mRNAs by binding specific sequences in their 3′-UTRs that
contain a core ‘UGUR’ tetranucleotide followed by sequences that vary between members of this family [7]. The HuR protein influences mRNA stability by binding to the motif UUNNUUU [8]. A crucial step in unravelling the gene expression pathway is to elucidate the binding motif of RNA binding proteins and to identify their targets.

Post-translational regulation is also very widespread, comprising chemical modifications on proteins (which are reversible) and proteolysis (which is irreversible). Proteins are organic compounds made of amino acids arranged in a linear chain that folds into a functional three-dimensional structure. In general, there are 20 standard amino acids and therefore a protein can mathematically be viewed as a string over an alphabet of size 20. There are many motifs that are conserved among different proteins. For example, carbohydrates can be attached to the amino acid Asparagine in proteins through N-glycosylation sites which are characterized by the consensus sequence Asn-X-Ser/Thr [9, 10]. The first amino acid in this sequon is Asparagine (Asn), the second amino acid can be any of the 20 amino acids, except Proline (X), and the third amino acid is either Serine (Ser) or Threonine (Thr). There are also more complex motifs such as enzyme activity sites and receptor binding sites.

As explained above, discovering motifs in DNA, RNA and protein sequences is an important task in molecular biology that helps in understanding the mechanisms underlying the binding of DNA, RNA and proteins. Nevertheless, pattern discovery in biological sequences is amongst the most challenging problems in computational molecular biology. In its simplest form, the problem can be formulated as finding an unknown pattern that occurs frequently in a given a set of sequences. This definition poses three major challenges involved in addressing motif search tasks. The first challenge hiding in this definition is how to define the search space. If patterns are assumed to be exact words, that is - words of length $k$ over the alphabet, then the theoretical size of the search space is $4^k$ (in the case of an alphabet of size 4 – DNA/RNA). Allowing more flexibility can be achieved by representing patterns as flexible words using the IUPAC alphabet, which considers all subsets of letters over the alphabet. For example, the pattern GGACR in DNA stands for two possible variants – GGACA and GGACG, as R stands for A or G. The theoretical size of the search space in this case is $15^k$ (for the underlying alphabet of size 4). An even more flexible model is to represent patterns using probability matrices. The rows of the matrix correspond to the letters in the alphabet and the columns correspond to the position in the
pattern. Each value in the matrix represents the probability to find a letter at a given position in the pattern. Assuming the probabilities in the matrix are multiples of 0.1, and since each column must sum to 1, the number of combinations in each column of the matrix is equal to the number of integer solutions for the equation $X_1 + X_2 + X_3 + X_4 = 10$ (for an alphabet of size 4). There are $\binom{13}{10} = 286$ possible solutions to this equation, meaning that the size of the search space in this model is $286^k$. Navigating in such a huge space is a major computational challenge. Discovering motifs in a set of DNA sequences is a difficult task owing to the tendency of binding sites to be short and degenerate, which therefore requires testing flexible patterns. The latter dictates scanning of very large search spaces. The problem tends to be worse in multicellular eukaryotes than in prokaryotes and yeast because eukaryotic sites tend to be shorter and more variable [11].

The second challenge in motif search is how to score patterns and assess their significance for a given condition. Given a target set of sequences and a pattern that we want to score, one should define the scheme for assessing whether the pattern is over-represented in the target set. The third challenge is defining the target set, as usually in biological measurement data results are given as a ranked list of elements, and therefore the list needs to be cut to produce a target set. Data produced by techniques such as ChIP-seq [12], ChIP-exo [13], CLIP [14], PAR-CLIP [15] and others are readily representable as ranked lists of sequences, where the ranking is according to the measured binding affinity. We are often interested in sequence motifs that are observed to be enriched in sequences where strong binding affinity is measured. Most motif discovery tools, including state-of-the-art tools, take as input a set of sequences treated as the target set and return motifs that are overrepresented in that set compared to a background model. Users of these tools, who have a ranked list in hand, need to bridge over this gap by producing a target set from the ranked list, meaning that the list is cut at some arbitrary threshold. It also implies that most motif discovery tools ignore the ranking information. Our approach, which relies on the minimum hyper-geometric statistics, gets as input a list of sequences ranked according to the biological measurement, and cuts this list in a data driven manner into a target set and a background set in a way that maximizes the enrichment of the tested pattern in the target set compared to the background set. This avoids using arbitrary thresholds, while also exploiting the ranking information derived from experimental measurements.
1.2 Variable gap motif search

While most motif finding approaches consider solid sequence elements, it is of interest to also consider gapped sequence elements. For example, GAL4 in *S. cerevisiae* binds DNA as a homodimer. Its binding site comprises 17 base pairs, containing palindromic CGG triplets at the ends that are separated by an 11 base pairs gap [16]. Additionally, Puf2p in *S. cerevisiae* binds 3'UTR of mRNAs by recognizing a motif of two UAAU tetranucleotides separated by a 3 nucleotides linker sequence [17]. Therefore, an interesting case, related to the role of gapped motifs, is of a protein that binds the DNA or RNA as a dimer. There are also cases where recognition is based on sequence elements with variable length gaps separating the half sites. An important example is the recognition site of the tumor protein p53 (TP53). In many organisms this site is composed of two copies of the half-site RRRCWWGYYY separated by a spacer, usually of length 0–21 base pairs [18].

Computational models of dimers binding to two half-sites that feature certain spacing rules were suggested in a handful of recent studies. Several algorithms, including BioProspector [19], Gemoda [20], SPACER [21], SPACE [22] and GLAM2 [23] deal with the problem of discovering gapped motifs. van Helden et al. [24] consider a model of a spaced pair of trinucleotides, separated by a spacer of a fixed length (e.g. 0–16 nucleotides). The algorithm exhaustively tests all $4^6$ combinations of pairs of trinucleotides. This method is highly efficient in detecting sites bound by $C_6Zn_2$ binuclear cluster proteins. However, it allows no flexibility in the motif structure in the context of dyad size and spacer characterization. Carvalho et al. [25] proposed an algorithm named RISO to find structured motifs. The algorithm provides flexibility in the sense that it allows mismatches in the sites during search, and variable length spacing is supported. It also uses truncated suffix trees instead of an exhaustive search to efficiently enumerate candidate motifs. The main drawback of this method is its inability to deal with large amounts of sequence data since its complexity grows rapidly both in time and in space. Chen et al. [26] developed a method for discovering discontinuous patterns among input sequences by linking short motifs that are located at conserved regions with a flexible gap of length 0–15. One limitation of this algorithm is that it uses positive and negative sets to filter candidate motifs based on ChIP-chip $p$-value fixed thresholds. Furthermore, as gapped motifs may consist of half-sites that are not independently enriched, this method will potentially miss some significant
results. Recently, a gapped PWM model was suggested for discovering variable length DNA binding sites [27]. This model extends the PWM model by introducing an optional gap character, which may appear once and at a certain position inside the motif, to simply model variable-length motifs. A main drawback of this model is that the gap can be one base only. To the best of our knowledge, none of the methods listed above can efficiently discover variable gapped motifs under a definition that allows full flexibility of the gap.

As evidenced in the work cited above, the search for variable gapped motifs (VGMs) poses a tremendous computational challenge, as the search space becomes huge, when considering parameters of biological relevance. Specifically, if we seek DNA motifs containing two half sites, each of length 4, where the gap between the 4-mer half sites can be any subset of the numbers \{0,…,10\}, then an exhaustive search will span $4^8 \times 2^{11}$ candidates, which is far too large to routinely address in reasonable time. Our approach to overcoming the computational challenges associated with large motif search spaces is based on using suffix trees, to restrict our attention to motifs that actually occur in the input list of sequences [28]. The efficiency of our approach allows searches over motif spaces that are not covered by existing tools, including variable gap motifs and long motifs over large alphabets. The algorithmics is described in the Methods - Section 2.2 and in the Results - Section 3.2.

### 1.3 Enrichment analysis using the minimum hyper-geometric (mHG) statistics

In this section we introduce the statistics used to leverage the information conveyed by the ranking of sequences in the context of enrichment. Given a candidate motif which we want to assess, we first compute an occurrence vector for the motif - based on the ranking. For every sequence in the list we put 1 in the corresponding entry in the vector if it contains the motif and 0 otherwise, as illustrated in Figure 1.

Next, we wish to assess the statistical significance of this vector. The mathematical framework for doing this is the minimum hyper-geometric statistics. Briefly, the mHG statistics answers two main questions. First, given a binary vector, it tells us whether there is a significant concentration of ones at the top of the vector. The mHG score captures this type of significance. It reflects the level of surprise resulting from seeing the observed density of ones at the top of the vector under the null assumption that all configurations of ones in the vector are equiprobable. Second, the
mHG statistics outputs the cutoff separating the top of the list from the rest of the list. This is done in a data driven manner, so as to maximize the motif enrichment. In order to find this optimal cutoff in a data driven manner, the algorithm implemented in the mHG framework effectively checks all possible partitions and returns the best partition found, in terms of minimizing the hyper-geometric tail, which is described next.

![Figure 1 – Binary vector representation for the occurrences of the pattern UGUAWAUW in a hypothetical example input ranked list](image)

Given a candidate motif, we compute an occurrence vector for the motif - based on the ranking. For every sequence in the list we put 1 in the corresponding entry in the vector if it contains a copy of UGUAWAUW (W stands for A or U), and 0 otherwise.

This approach has been developed by the Yakhini research group [29-31] to identify the enrichment of a subset \(A\) in a ranked list of entities. In our case, \(A\) contains the sequences that contain the candidate motif (i.e. the entries equal to 1 in the binary vector). Given a total number of elements \(N, B\) of which are in \(A\), and \(n\) of which are in the target set (that is, amongst the top \(n\) entries of the list), the probability that \(b\) or more elements from the target set are also in \(A\) is given by the hypergeometric tail (HGT):

\[
P(X \geq b) = HGT(b; N, B, n) = \sum_{i=b}^{\min(n,B)} \binom{n}{i} \frac{\binom{N-n}{B-i}}{\binom{N}{B}} (1)
\]

If a ranked list \(S_1, \ldots, S_N\) is provided instead of a target set, we define a label vector \(\lambda = \lambda_1, \ldots, \lambda_N \in \{0,1\}^N\) according to the association of the ranked elements to \(A\), that is, \(\lambda_i = 1\) if and only if \(S_i\) is in \(A\). The mHG score is then defined as:

\[
mHG(\lambda) = \min_{1 \leq n \leq N} HGT(b_n(\lambda); N, B, n) (2)
\]

where
In other words, the mHG score is the optimal HGT probability that is found over all possible partitions induced by the ranking. As such, this score must be corrected for multiple testing. A dynamic programming algorithm for computing the exact $p$-value of a given mHG score has been developed by the Yakhini group and described previously [29]. More specifically, given a ranked list of elements, a subset $A$, and a corresponding mHG score $s$, the mHG $p$-value tells us the exact probability of observing an mHG score $s' \leq s$ under the null assumption that all occurrence configurations of $A$ in the ranked list are equiprobable. In practice, we use Stirling’s approximation [32] to compute all binomial coefficients needed to assess HGTs. Stirling’s inequality says that:

$$\sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}} \leq n! \leq \sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\frac{1}{12n}} \quad (4),$$

which is tight for large factorials.

1.4 Suffix trees

Having established an approach to utilizing the ranking and to scoring patterns, the next challenge in motif discovery is to efficiently produce candidate motifs, because the number of possible words over a constant alphabet is exponential in the motif length. Reducing the search space size can be obtained, easily, by testing only words that occur in the input dataset. This can be achieved by using suffix trees. A suffix tree is a data structure that represents all the suffixes of a given string in a way that allows fast implementation of many string operations. A path from the root to a leaf in the tree represents a suffix. Each leaf of the tree holds information about the indices of strings that contain that suffix, and the starting positions of this suffix within each such string. An example of such a tree is shown in Figure 2. Restoring all occurrences of a suffix is thus enabled, which further allows for the detection of DNA, RNA or protein substrings that manifest a significant occurrence pattern in a set of biologically related sequences. This is because that using the suffix tree we can enumerate all substrings of length $k$ that occur in the input strings, as these are paths of length $k$ starting from the root, where we are allowed to stop at the middle of an edge. The number of such paths can be significantly lower than the theoretical
number of $|\Sigma|^k$ ($\Sigma$ is the alphabet) candidates, as we limit ourselves only to patterns that occur in our data.

Suffix trees allow efficient enumeration of candidate motifs due to the fact that suffix trees can be constructed and traversed in linear time. There are several algorithmic approaches to the efficient construction of a suffix tree for a collection $S_1, \ldots, S_N$ of strings [33-35]. Our motif discovery tools use a version that takes $O(M)$ time for construction, where $M = \sum_{i=1}^{N} \text{length}(S_i)$, by implementing Ukkonen’s algorithm for generalized suffix trees construction [35].

![Suffix tree diagram](image)

**Figure 2 - A generalized suffix tree for $S_1 = \text{CAGCA}$ and $S_2 = \text{GA}$**

The characters $S_1$ and $S_2$ are used for marking $S_1$ and $S_2$ ends, respectively. The edges are labeled with substrings of $S_1$ or $S_2$, and the information at each leaf indicates both the index of the string that contains the suffix represented by the path from the root to that leaf, and the starting position of the suffix within the string.

### 1.5 Position weight matrix motifs

Position weight matrix format is a flexible way to represent biological motifs and is commonly used [36-38]. This representation is more faithful to the underlying biology than representation by exact words. The rows of the matrix correspond to the letters in the alphabet (for example, in DNA, these are A, C, G and T) and the columns correspond to the position in the motif. Each value in the matrix represents the probability to find a letter at a given position in the motif (see Figure 3). It has one row for each symbol in the alphabet, and one column for each position in the pattern. The positions are assumed to be independent, so a PWM defines a product multinominal model, making pattern matching with the PWM straight-forward. Assuming an input sequence of length equal to the PWM width, we simply multiply the scores assigned to each letter in each of the positions in the input sequence to obtain the likelihood of the input string (alternatively, we can sum the logs of the probabilities). That is, the score assigned by a
PWM to a substring \( S = S_1...S_K \) is defined as \( \prod_{j=1}^{K} p_{s_j} \), where \( j \) represents a position in the substring; \( s_j \) is the symbol at position \( j \) in the substring; and \( p_{\alpha,j} \) is the score in row \( \alpha \), column \( j \) of the matrix. In other words, a PWM score is the product of position-specific scores for each symbol in the substring. This definition can be generalized to yield a score for a sequence \( S = S_1...S_M \) longer than the PWM by calculating \( \max_{1 \leq i \leq M-K+1} \prod_{j=1}^{K} p_{s_{i+j-1,j}} \).

Alternatively, an enhanced model that takes into account multiple occurrences of the PWM in the sequence can be applied by summing over sufficiently strong occurrences of the PWM or by other more sophisticated approaches [39].

Position weight matrices are commonly represented visually in the form of Shannon logos, where the information content of the PWM at each position is shown (Figure 3). Denoting the alphabet by \( \Sigma \), the information content (IC) at position \( j \) of the PWM is given by [40-42]:

\[
IC(j) = \log_2(|\Sigma|) + \sum_{\sigma \in \Sigma} p_{\sigma,j} \log_2(p_{\sigma,j}) \quad (5)
\]

\[
= \log_2(|\Sigma|) - \text{entropy(letter distribution in position } j)\]

The information content is measured in bits and, in the case of DNA/RNA sequences, ranges from 0 to 2 bits. A position in the PWM at which all nucleotides occur with equal probabilities has an information content of 0 bits, while a position at which only a single nucleotide can occur has an information content of 2 bits.

![Shannon logo representation](image)

**Figure 3 – A position weight matrix and its corresponding Shannon logo representation**

The rows of the matrix correspond to the letters in the alphabet and the columns correspond to the position in the motif. Each value in the matrix represents the probability to find a letter at a given position in the motif. The Shannon logo representation derived from this matrix is shown, where the vertical axis is the information content.
As later explained, our motif discovery tools first identify exact words that are statistically enriched in the input ranked list, and then expand these words heuristically into position weight matrix format. Though PWMs allow more flexibility in the motif, they are not a generalization of exact word motifs. For example, a motif comprising the variants AT and TA may give rise to a PWM where the nucleotides A and T have probabilities equal to 0.5 at positions 1 and 2, while C and G have zero probabilities. We note that this PWM represents a slightly different motif, covering also the 2-mers AA and TT in addition to AT and TA, which are equally as good. The standard PWM formulation assumes that the positions in the motif contribute independently (additively) to the functionality of the binding sites, which is a reasonable approximation in some cases, but it has been noted that some motifs exhibit significant dependencies between positions [43, 44]. Additivity clearly fails when the affinity decreases to the point of non-specific binding [45]. In other words, the overall agreement between the PWM score (based on additivity) and the binding affinity is weak. However, for prediction purposes, the additivity assumption needs to hold only for sequences demonstrating high-binding affinity with respect to the protein of interest (that is, for sequences that are at the top of the list), in order to have a useful model [46]. Nevertheless, for many proteins the additive model seems to be approximately valid [47, 48]. Considering dependencies between positions complicates the model considerably as the search space becomes significantly larger. Instead of using PWMs, motifs can be represented using subsets of \(k\)-mers. The theoretical size of this space is equal to \(2^{4k}\) (for DNA/RNA). Matching \(k\)-mers in the subset with weights further complicates the model.

1.6 DRIMust – discovering rank imbalanced motifs using suffix trees

Together with Yael Mandel-Gutfreund’s Lab at the Technion, we developed DRIMust, which is a web application for de novo motif discovery, freely accessible through the website http://drimust.technion.ac.il.

The DRIMust algorithm is based on the minimum hyper-geometric statistical framework and uses suffix trees for an efficient enumeration of motif candidates [28, 49].

DRIMust takes as input ranked lists of sequences and returns motifs that are over-represented at the top of the list, where the determination of the threshold that defines top is data driven. The resulting motifs are presented individually with an accurate \(p\)-value indication and as a position
weight matrix. The algorithm implemented in DRIMust starts with enumerating all exact words of the desired length that occur in the dataset, using suffix trees - for efficient enumeration. These candidates are then assessed for their statistical significance using the mHG statistics. Finally, the most promising candidates are further heuristically expanded into a more flexible position weight matrix format. This process is summarized in Figure 4.

Figure 4 – DRIMust’s flow chart
In DRIMust, an initial motif search phase produces k-mers, which are words over the alphabet of the input sequences of length $k$. These candidate k-mers are derived by enumerating paths of length $k$ in the generalized suffix tree generated for the input sequences. Next, the statistical significance of the k-mers is calculated using the mHG statistics. In the next stage, the promising k-mers are extended to produce position weight matrix motifs.

Comparing DRIMust with state-of-the-art tools demonstrated significant advantage to DRIMust, both in result accuracy and in short running times. Unlike many other approaches, it does not exhaustively search over all possible k-mers space and therefore can detect long motifs and motifs over large alphabets. Most importantly, by working with ranked lists, DRIMust avoids the arbitrary designation of fixed sets of sequences and exploits the ranking derived from experimental measurements. Overall, DRIMust is unique in combining efficient search on large ranked lists with rigorous $p$-value assessment for the detected motifs. More details on DRIMust methodology can be found in the Methods - Section 2.3 and in the Results - Section 3.3.

1.7 Assessing position weight matrix (PWM) motifs in ranked lists
Given a set of sequences that were tested in a high throughput experiment such as ChIP-seq [12], CLIP [14] and others, they can be ranked according to the measured binding affinities, yielding a ranked list $L_1$. Since usually we are interested in finding motifs amongst sequences having strong binding affinities, we actually search for motifs that are more prevalent at the top of this list.
Given a ranked list of sequences, if the candidate motif that we assess is a fixed word, it is straightforward to tell where it occurs. Therefore, a binary vector representation is appropriate in that case, and the mHG statistics can be invoked. However, when the motif is represented in a position weight matrix format, then instead of a binary entry per sequence, we have a quantitative score associated with every sequence. That is, given a PWM which we want to assess, the sequences can also be ranked according to their PWM scores, yielding another ranked list $L_2$, different from $L_1$. A significant PWM motif would yield significant scores for sequences having strong binding affinities. Therefore, the question of PWM motif discovery from ranked experimental data can be formulated as quantifying the mutual enrichment level for the two ranked lists $L_1$ and $L_2$. The mathematical framework for mutual enrichment quantification in two ranked lists over the same set of elements is a generalization of the mHG statistics, called the mmHG statistics [50, 51]. While any parametric or non-parametric correlation statistics (e.g. Spearman’s correlation coefficient), that takes the same input, calculates the overall agreement between the two ranked lists, the mmHG statistic focuses only on agreement at the top of the two ranked lists. mmHG counts elements common to the top of both lists, without predefining what top is. Its intended output is the probability for observing an intersection at least as large in two randomly ranked lists. The mmHG statistics is described in detail in the Methods - Section 2.4. We developed tight upper bounds on tail distributions of the size of the intersection of the top parts of two uniformly and independently drawn permutations, computable in polynomial time [51]. We further demonstrate advantages of this approach using our software implementation, mmHG-Finder, to study PWM motifs in biological datasets, in the Results - Sections 3.4 and 3.5.

### 1.8 Non-coding RNAs

In the previous sections we introduced the problem of motif discovery in ranked lists from several aspects. In the following sections, we focus on regulatory mechanisms of non-coding RNAs, that is – RNAs that are not translated into proteins. The RNA world hypothesis proposes that early life was based on RNA, which then devolved the storage of information to more stable DNA, and catalytic functions to more versatile proteins [52]. Consequently, RNA is assumed to have been largely relegated to an intermediate between gene and protein, encapsulated in the central dogma 'DNA makes RNA makes protein'. Until recently, discrimination between protein-coding or non-protein–coding RNAs was relatively straightforward: most transcripts were clearly
Identifiable as protein-coding messenger RNAs, and readily distinguished from the small number of well-characterized non-protein-coding RNAs (ncRNAs), such as transfer, ribosomal, and spliceosomal RNAs. However, genome-wide studies of the last decade have revealed the existence of thousands of noncoding transcripts, whose function and significance are unclear. In humans and mice, for example, it has become apparent that the vast majority of the genome is transcribed, apparently in a developmentally regulated fashion [53-58]. mRNAs account for only ~2.3% of the human genome, and therefore the vast majority of this unexpected transcription appears to be non-protein-coding. This finding challenges the central dogma and suggests that RNA has continued to evolve and expand alongside proteins and DNA.

It is therefore not surprising that a great deal of attention is now focused on the noncoding transcriptome. The discovery of thousands of small RNAs (<200 nucleotides in length) has been dominating this field of research. Many of the small RNAs have since been classified into novel categories (for example, microRNAs, PIWI-associated RNAs, and endogenous small interfering RNAs) on the basis of function, length, biogenesis, structural or sequence features, and protein-binding partners [59]. Though long ncRNAs (>200 nucleotides) appear to comprise a large portion of the mammalian noncoding transcriptome, the biological significance of these long ncRNAs is still not well understood. Despite an increasing number of long ncRNAs having been shown to play a diverse range of regulatory roles [60], the functions of the vast majority remain unknown and untested. Moreover, long ncRNAs, unlike their smaller counterparts, lack obvious features to allow a priori functional categorization or prediction.

Recently, thousands of large intergenic non-coding RNAs (lincRNAs) have been reported in mammals [61-63]. These RNAs are evolutionarily conserved in mammalian genomes and thus presumably function in diverse biological processes. It has been suggested that lincRNAs have active roles in modulating the cancer epigenome and therefore may be important targets for cancer diagnosis and therapy [64]. Independently, a subset of lincRNAs has been shown to be regulated by the tumor suppressor p53 [65]. The latter study is a primary effort that contributes to our understanding of long ncRNA regulation, but currently very little is known in this field. Discovering the mechanisms underlying ncRNA regulation is independently important and additionally it may shed light on ncRNA functionality. By analyzing long non-coding RNA sequences using our tools, we found GC-rich strings to be enriched amongst the promoter
sequences of long non-coding RNAs that are specifically expressed in thyroid and prostate tissue samples and observed a statistical association with tissue specific CpG hypo-methylation. We suggest that thyroid and prostate-specific long non-coding RNAs are regulated by transcription factors that bind GC-rich sequences, such as EGR1, SP1 and E2F3. We further suggest that this regulation is associated with CpG hypo-methylation. More details can be found in the Results – Section 3.5.

1.9 Cooperative interactions of RNA binding proteins and microRNAs

MicroRNAs (miRNAs) are small RNA molecules (approximately 22 nucleotides) participating in a large variety of cellular processes in animals, plants and viruses [66-68]. miRNAs act by binding to the 3’-untranslated region (3’-UTR) of messenger RNAs (mRNAs), forming hybrids that consist of the binding site in the 3’-UTR and of the miRNA seed region (positions 2–8 in the miRNA) [69, 70]. Recently, it has been suggested that miRNAs also bind within the coding regions of mRNAs [15, 71]. Several reports have demonstrated that miRNA repression can be reversed or modulated by RNA-binding proteins (RBPs) interacting with the 3’-UTR of target mRNAs. For example, it was reported that the RNA-binding protein HuR reverses miR-122 repression of CAT-1 mRNA in human hepatocarcinoma cells subjected to stress [72]. This effect on CAT-1 is mediated by HuR translocation from the nucleus to the cytoplasm upon stress, and is accompanied by CAT-1 release from processing bodies, structures involved in RNA metabolism, leading to active translation of the message. The process requires the association of HuR with AU-rich sequences in the 3’-UTR region of the CAT-1 mRNA, through an as yet unknown mechanism. RNA binding protein modulation of miRNA-mediated repression has also been reported for DND1. In zebrafish, it alleviates miR-430 repression of nanos1 and tdrd7 in primordial germ cells [73]. DND1 can also relieve the repression of p27 mediated by miR-221 and the repression of LATS2 by miR-372 in HEK293T cells [73]. In zebrafish and in humans, DND1 counteracts miRNA-mediated repression by binding to U-rich regions located near the miRNA binding sites within the 3’-UTR of the message. DND1 binding to these sequences might interfere with miRNA–mRNA interaction. Another indication of the functional relationship between miRNAs and RBPs was found in the rat hippocampal neurons, for which treatment with BDNF was shown to partially relieve miR-134 mediated repression of Limk1 [74]. When miRNAs regulate mRNAs, they are assembled into ribonucleoprotein complexes
known as the miRNA-induced silencing complex (miRISC) [67]. In contrast to HuR, DND1 and BDNF, which relieve the miRNA repressive function, the TRIM–NHL protein family (NHL-2 in *Caenorhabditis elegans* [75] and TRIM32 in mouse [76]) increase the activity of specific miRNAs, including let-7, by binding to miRISC components. Overall, many examples suggest extensive crosstalk between RBPs and miRNAs. Therefore, it is likely that additional cases of RBPs modulating miRNA interactions exist. In our work we found an interesting association between an RNA binding protein from the Pumilio family and miR-410 in humans [77], which is described in detail in the Results – Section 3.1. Interdependent activity of RNA binding proteins and miRNAs was further researched in parallel to our work by Kedde *et al.* [78] as well as in a more recent and comprehensive study by Jiang *et al.* [79].
2 Methods

2.1 Efficient enumeration of the motif search space using suffix trees

Our approach to overcoming the computational challenges associated with large motif search spaces is based on using suffix trees, to restrict our attention to motifs that actually occur in the input list of sequences. A suffix tree is a data structure that represents all the suffixes of a given string in a way that allows fast implementation of many string operations. Suffix trees are useful in many application contexts, including bioinformatics and computational biology [80, 81]. Perhaps the simplest application example comes from the context of text search. Consider the substring problem, whose input is text $T$ of length $M$. After an $O(M)$ preprocessing, one is required to determine, for any string $S$ of length $m$, in $O(m)$ time, whether $S$ occurs in $T$ or not. Using a suffix tree, these performance bounds can actually be achieved. Constructing a suffix tree for $T$ can be done in $O(M)$ time [33-35]. Given a substring $S$ of length $m$, all we now need to do is search for the path $S$ in the tree, starting from the root - an $O(m)$ operation. This $O(M)$ preprocessing and $O(m)$ search solution for the substring problem is very surprising and extremely useful, especially since $M$ may be huge compared to $m$.

There are several algorithmic approaches to the construction of a suffix tree for a single string in linear time [33-35]. These methods easily extend, with small modifications, to represent the suffixes of a collection $S_1, \ldots, S_N$ of strings. A generalized suffix tree is a data structure that contains all suffixes of a set of strings (see Figure 2 for an example) and can be built in $O(M)$ time, where $M = \sum_{i=1}^{N} \text{length}(S_i)$. A path from the root to a leaf in the tree represents a suffix. Each leaf of the tree holds information about the indices of strings that contain the suffix, and the starting positions of this suffix within each such string. Restoring all occurrences of a suffix is thus enabled. A natural application of generalized suffix trees is finding substrings that are common to a large number of distinct strings. This question can arise in many different contexts, for example in database search, in multiple sequence alignment and in motif search [80-82]. Finding DNA, RNA or protein substrings that commonly occur in a set of biologically related sequences helps in pointing out regions or patterns that may be functional.
A useful variant of this problem is finding rank imbalanced motifs, given a ranked list of sequences \( S_1, \ldots, S_N \). Rank imbalanced motifs are substrings that appear more often at the top of the list compared with the remainder of the list. This notion of rank imbalanced motifs was introduced by Zilberstein et al. [83] and by Eden et al. [29] who described the mHG statistics used for the assessment of such motifs. A unique feature of the mHG statistics is that the cutoff between the top and the rest of the list is determined in a data-driven manner so as to maximize the motif enrichment. This is done by computing the motif enrichment over all possible set partitions and identifying the cutoff at which maximal statistical significance is observed. Our algorithm uses generalized suffix trees for an efficient enumeration of motif candidates, which are then assessed using the mHG statistics. The occurrences of each candidate motif in the list are extracted from the extra information stored in the leaves of the tree.

2.2 Variable gap motif search

The search for variable gapped motifs (VGMs) poses a tremendous computational challenge, as the search space becomes huge, when considering parameters of biological relevance. Specifically, if we seek DNA motifs containing two half sites, each of length 4, where the gap between the 4-mer half sites can be any subset of the numbers \( \{0, \ldots, 10\} \), then an exhaustive search will span \( 4^8 \times 2^{11} \) candidates, which is far too large to routinely address in reasonable time. We restrict our attention to candidate patterns that occur in the input dataset using generalized suffix trees, as we later explain. Note that a variable gap motif can be viewed as a set of strings. That is, for a set of integers, \( \Lambda \), the pattern \( P - N^\Lambda - T \) consists of all strings that start with \( P \), then have a wildcard gap of any of the lengths in the set of gaps \( \Lambda \), and end with \( T \). For example, the motif GCC-N^{1,5}-ATG represents the strings GCCNATG and GCCN5ATG.

All patterns of the form \( P - N^\Lambda - T \), where \( \Lambda \) is a set of integers not greater than \( l \), and \( P \) and \( T \) are of lengths \( a \) and \( b \), respectively, are represented in the tree (as shown in Figure 5). The corresponding path starts with \( P \), followed by possibly more than one string – each of variable length not exceeding \( l \) – which is then followed by the path \( T \). Therefore, for candidate enumeration, we traverse the tree for all paths of length \( a \) – as candidates for \( P \). Then, for each \( P \), we construct a suffix tree for the collection of strings rooted below \( P \) which are not longer than \( l + b \). Then, all paths of length \( b \) in that tree are candidates for \( T \), given \( P \). The lengths of spacers
connecting \( P \) and \( T \) can be extracted from the tree, and hence only lengths that actually connect \( P \) and \( T \) in the dataset are tested, potentially less than the size of the set \( \{0, \ldots, \ell\} \) representing all \( \ell + 1 \) possible gaps.

![Diagram](image)

**Figure 5 – Enumeration of variable gap motif candidates using suffix trees**

First, a generalized suffix tree for the sequences \( S_1, \ldots, S_n \) is constructed. Then, for every path \( P \) at depth \( a \) we build a generalized suffix tree \( \text{Tree}(P) \) for all strings of length \( b + 1 \) rooted below \( P \). Finally, paths \( T \) at depth \( b \) in \( \text{Tree}(P) \) are traversed, and all motifs \( P - N^\Lambda - T \), where \( \Lambda \subseteq \{0, \ldots, \ell\} \), are evaluated.

Using our variable gap motif search algorithm we were able to identify significant motifs, as in the case of estrogen receptor binding site. Our prediction \( \text{GTCA-N}^{3,6,9}\text{-TGAC} \) (\( p \)-value \( \leq 1.13\cdot10^{-20} \)) comprised two palindromic parts separated by a spacer of variable length: either 3,6 or 9 nucleotides. This motif expands the known literature motif \( \text{GTCA-N}^3\text{-TGAC} \) \([84]\) (\( p \)-value \( \leq 1.12\cdot10^{-18} \)), which has a fixed gap of length 3. If we compare this variable gap motif with the consensus motif we observe that a spacer of lengths 3,6,9 is stronger than the fixed length, but the difference is not big. As a matter of fact, in most examples where we found the variable gap motif to be more significant than the fixed gap motif, the difference wasn’t very large, implying that variable gap motifs are not very widespread. Additional observations can be found in the Results – Section 3.2.

Our method, implemented as publicly available stand-alone software, has several advantages over existing methods. First, many other approaches exhaustively search over motif spaces and therefore cannot handle variable gapped motifs. Furthermore, we search motifs in ranked lists and not in fixed sets of sequences as is the case for many other methods. To the best of our
knowledge, our tool is unique in combining an efficient search with a ranked list approach and rigorous \( p \)-value estimation. It is also unique in efficiently addressing variable gap motifs under a definition that allows full flexibility of the gap.

2.3 Expanding \( k \)-mers into position weight matrix motifs

Our motif discovery tools, including DRIMust, use suffix trees for candidate enumeration, meaning that the search space is represented by exact words. Since representation using position weight matrix format allows more flexibility and is more faithful to the underlying biology, we attempt to expand the most promising words which are significantly ranked imbalanced in the list - into a more flexible representation. In this section we describe the approach we use in the DRIMust webserver for PWM expansion [49].

The most promising \( k \)-mers in terms of the mHG enrichment are passed as input to a process that extends them to PWMs (typically – between 50 to 150 not similar \( k \)-mers are expanded), through a heuristic approach which is based on the Hamming neighbors of these exact words. Briefly, starting from a single \( k \)-mer, Hamming neighbors (of length \( k \)) are added to a set of motifs as long as the new addition improves the observed enrichment \( p \)-value, and as long as the overall similarity of the members in the set does not decrease below a similarity threshold. Iteratively, the mostly improving neighbor is added, where the new enrichment \( p \)-value is calculated based on the union of occurrences of the members in the set. After this process is over, each set comprises \( k \)-mers which are then transformed into a PWM motif. Based on their occurrences at the top of the list, a position weight matrix is inferred from these \( k \)-mers. The corresponding enrichment \( p \)-value is calculated based on the union of occurrences of the members in the set.

2.4 Minimum-minimum hyper-geometric statistics (mmHG)

The main weakness of the PWM expansion process described in the previous section is that the enrichment \( p \)-value of the PWM is calculated according to the union of occurrences of the members in the cluster, and in parallel the PWM is derived from the cluster. Since there are many possible ways to deduce a PWM from a cluster of \( k \)-mers, a better scheme would be to evaluate the statistical enrichment of the PWM directly. For this end, since the input is a ranked list of sequences, and since the sequences can also be ranked according to the PWM agreement
scores, a mutual enrichment analysis of the two rank orders is appropriate. A significant PWM motif would yield significant PWM scores for sequences that are ranked high in the input ranked list. We are therefore interested in assessing mutual enrichment in two ranked lists.

Mutual enrichment in the top of two ranked lists can be calculated using the mmHG statistics. The mmHG statistics, which was first introduced by Steinfeld et al. [50], is a generalization of the mHG statistics [29]. Similarly to the mHG statistics which quantifies the enrichment level of a set of elements at the top of a ranked list of elements of the same type, the mmHG statistics quantifies the level of mutual enrichment in two ranked lists over the same set of elements. While any parametric or non-parametric correlation statistics (e.g. Spearman’s correlation coefficient), that takes the same input, calculates the overall agreement between the two ranked lists, the mmHG statistic focuses only on agreement at the top of the two ranked lists. mmHG counts elements common to the top of both lists, without predefining what top is. Its intended output is the probability for observing an intersection at least as large in two randomly ranked lists (the enrichment mmHG p-value). Our definition of the mmHG statistic [51] varies slightly from that of Steinfeld et al. [50], and is introduced below.

Mutual enrichment in the top of two ranked lists can be simplified, from a mathematical point of view, by arbitrarily setting the indices of one list to the identity permutation \((1,2,\ldots,N)\) and treating the other list as a permutation \(\pi = \pi(1), \ldots, \pi(N)\) over these numbers. For the purpose of assessing the intersection of the top of the two ranked lists in a data driven manner, mmHG asks which prefix \([1,\ldots,n_1]\) is enriched in the first \(n_2\) elements of \(\pi\), that is in the set \(\pi(1), \ldots, \pi(n_2)\).

We now define mmHG for the case of one permutation. Consider a permutation \(\pi = \pi(1), \ldots, \pi(N) \in S_N\) - the group of all permutations over the numbers \(1,\ldots,N\). mmHG is a function that takes \(\pi\) and calculates two numbers \(1 \leq n_1, n_2 \leq N\) such that the observed intersection between the numbers \(1,\ldots,n_1\) and the first \(n_2\) elements of \(\pi\) - \(\pi(1), \ldots, \pi(n_2)\) - is the most surprising in terms of the hypergeometric p-value. Additionally, mmHG further calculates this aforementioned p-value.
Formally, given \( \pi \in S_N \) and for every \( 1 \leq n_1, n_2 \leq N \), let \( b_\pi(n_1, n_2) \) be the size of the intersection of \( 1, \ldots, n_1 \) with \( \pi(1), \ldots, \pi(n_2) \). Set

\[
\text{mmHG score}(\pi) = \min_{1 \leq n_1 < N} \min_{1 \leq n_2 < N} HGT(N, n_1, n_2, b_\pi(n_1, n_2)) \quad (6)
\]

where HGT is the tail distribution of an hypergeometric random variable:

\[
HGT(N, n_1, n_2, b) = \sum_{i=b}^{\min(n_1, n_2)} \binom{n_1}{i} \binom{N-n_1}{n_2-i} \binom{N}{n_2} \quad (7)
\]

The mmHG score cannot be considered as a significance measure, due to the multiple testing involved in finding \( n_1 \) and \( n_2 \). A simple way to correct an mmHG score \( s \) for multiple testing and report a \( p \)-value bound would be to use the Bonferroni correction. That is done by multiplying \( s \) by the number of multiple tests conducted which is \( N^2 \). Therefore:

\[
\text{mmHG } p \text{-value}(s, N) \leq s \cdot N^2 \quad (8)
\]

In the Results – Sections 3.4 and 3.5 - we present significantly tighter bounds, computable in polynomial time.

### 2.5 PWM motif search in ranked lists (using mmHG-Finder)

The main advantages of DRIMust, our motif discovery tool which is available as a webserver, are that it is very fast due enumerating motifs using suffix trees, as well as the rigorous statistical model which utilizes the ranking information. The main limitation of DRIMust is its inability to detect degenerate motifs if those don’t have a strong exact word representative enriched in the data, as the algorithm implemented in DRIMust enumerates exact words and then expands the most promising ones into PWM motifs. To overcome this limitation, and to allow statistical assessment of PWM motifs directly, we developed mmHG-Finder. mmHG-Finder takes as input a ranked list of DNA or RNA sequences and returns significant motifs in PWM format. In cases where sequence ranking is not relevant or not available, it allows the use of positive and negative sets of sequences, searching for enriched motifs in the positive set using the negative set as the background.
This motif search algorithm includes four stages. First, candidate $k$-mers are enumerated using suffix trees, where $k$ is a parameter indicating the length. Next, the $k$-mers are assessed for their statistical significance using the mHG statistics. Then, the most promising $k$-mers are heuristically expanded into position weight matrix format, allowing more flexibility in the motif, but in more comprehensive way than that implemented in DRIMust. We take the most significant fifty $k$-mers, to be used as starting points for PWM expansion. This set of candidates is chosen such that the members are quite different. Note that this is a heuristic approach and the number 50 is somewhat arbitrary, chosen to succeed in catching the best performing PWMs without heavily paying in complexity. For each starting point, we iteratively replace one position in the $k$-mer by considering all possible IUPAC replacements and taking the one that improves the enrichment the most. We repeat this process for all positions several times. Eventually we get a motif in the IUPAC alphabet which is then expanded by adding Hamming neighbors as long as the new addition improves the observed enrichment $p$-value, and as long as the overall similarity of the members in the set does not decrease below a similarity threshold. Finally, the expanded motif is converted to a PWM, which is assessed using the mmHG statistics. The most significant PWM motifs are returned as output, together with the $p$-value. The performance of this algorithm is described in detail in the Results - Sections 3.4 and 3.5. This algorithm is slower than DRIMust, but can detect degenerate motifs which DRIMust occasionally misses.

2.6 Developing statistical bounds for mHG and mmHG

As explained in the section introducing the mmHG statistics, the mmHG score cannot be considered as a significance measure, due to the multiple testing involved in finding the cutoffs $n_1$ and $n_2$ – where $n_1$ defines the top of the first list and $n_2$ defines the top of the second list. In order to calculate the significance of a given mmHG score, we need to characterize its distribution as a random variable. For that, we first need to realize what the probability space is. Given two ranked lists over the same set of elements, the relative permutation of one list against the other contains the same information as the two rank orders. In other words, we can arbitrarily set the indices of one list to be $(1,2,\ldots,N)$ and treat the other list as a permutation over these numbers (as illustrated in Figure 6). Therefore, the probability space is equal to $S_N$, which is the group of all permutations of size $N$. 

25
Given an mmHG score $s$, observed in analyzing real measurement data, we would like to assess the statistical significance of this observation. Assuming endless computational power, we would enumerate all permutations and calculate the mmHG score for each, in order to characterize the distribution of mmHG as a random variable over $S_N$. The $p$-value for $s$ is then simply:

$$mmHG\ p-value(s,N) = \frac{\text{The number of permutations having mmHG score } \leq s}{N!}$$  \hspace{1cm} (9)$$

Since the number of permutations is huge, the process described above is very far from feasible. Therefore, we seek a computationally tractable upper bound, preferably tight. A trivial upper bound is the Bonferroni corrected mmHG score defined by $s \cdot N^2$. A more subtle upper bound was suggested by Steinfeld et al. [50]. In our work we developed tighter bounds, computable in polynomial time.

Our bounds are described in the Results – Sections 3.4 and 3.5. Their input comprises an mmHG score $s$, and the total number of elements $N$. The output is an upper bound for the $p$-value. The efficiency of our approach relies on enumerating all possible HGT scores rather than enumerating all permutations in $S_N$. This approach is computationally efficient as HGT is a function of four input parameters: $N$, $n_1$, $n_2$, and $b$. Given $N$, there are $O(N^3)$ possible combinations of $n_1$, $n_2$, and $b$. Next, all is left to do is to determine how many permutations correspond to each HGT score. To this end, we define the function $\Lambda(N,n_1,n_2,b)$ to be the number of permutations for which it holds that out of the first $n_2$ entries, $b$ of them are taken from the range $[1,\ldots,n_1]$. This formulation is equivalent to counting permutations for which we
attain, at some point, the value $HGT(N, n_1, n_2, b)$, had we taken the exhaustive approach. $\Lambda(N, n_1, n_2, b)$ can be calculated combinatorically, from which an upper bound is easily derived:

$$mmHG \ p-value(s, N) \leq \frac{\sum_{n_1, n_2, b : HGT(N, n_1, n_2, b) \leq s} \Lambda(N, n_1, n_2, b)}{N!} \tag{10}$$

In this process, permutations may be counted multiple times, as a permutation may have several HGT values that are better than $s$. Using a more delicate counting we reduce the extent of multiple counting by looking one step backward. If, for example, $HGT(N, n_1, n_2, b) \leq s$, we can exclude from the counting permutations that contain $b$ elements from the range $[1, \ldots, n_1]$ at their first $n_2-1$ entries because necessarily $HGT(N, n_1, n_2-1, b) \leq s$ (as the same intersection is observed for even a smaller set – of size $n_2-1$).

The performance of these bounds is discussed in the Results – Sections 3.4 and 3.5. We compared them with the Bonferroni correction and with the empirical $p$-value. Both bounds performed much better than the Bonferroni correction and seem very tight. Specifically, the latter bound - based on a delicate counting - is evidently tighter. An interesting observation is that the bounds are tighter for smaller mmHG scores, and that is due to lesser over-counting for smaller scores.

The bounds introduced above were observed to be very tight. They are, however, computationally heavy, as they require $O(N^3)$ calculations of HGT. We would still like to have an upper bound which is tighter than the Bonferroni bound but also faster to calculate. Such a compromise is achieved by generalizing an approach introduced in [29] for the minimum hypergeometric statistics. Namely, given the number of elements $N$ and an attainable mmHG score $s$ for which we want to calculate the $p$-value, for each $1 \leq b \leq N$ and for each $1 \leq n_1 \leq N$, let $n_2(b, n_1)$ be the maximal integer $n_2$ so that if in a permutation $\pi \in S_N$, $b$ out of the first $n_2$ entries in $\pi$ are taken from the range $[1, \ldots, n_1]$, then $\pi$ satisfies $HGT_\pi(N, n_1, n_2, b) \leq s$.

Monotonicity properties of the hyper-geometric distribution imply the uniqueness of such $n_2$ integers. Due to monotonicity, given $b$ and $n_1$, the maximal value $n_2(b, n_1)$ can be calculated efficiently using binary search, which means that an upper bound that requires $O(N^3 \log N)$ calculations of HGT can be computed by using the following formula:
\[ mHG \ p-value(s, N) \leq \sum_{b, n_1, n_2(b, n_1) : HGT(N, n_1, n_2(b, n_1), b) \leq s} \frac{\binom{n_1}{b} \left( N - n_1 \right)}{\binom{n_2(b, n_1) - b}{N}} \] (11)

The same logic applies for the simpler case of mHG, where \( n_2 \) is known and equal to \( B \). In this case, for each \( 1 \leq b \leq N \) there exists a maximal integer \( n_1(b) \) such that \( HGT(N, n_1(b), B, b) \leq s \) (we ignore \( b \) if such an integer does not exist). An upper bound that requires \( O(N \log N) \) calculations of HGT then follows:

\[ mHG \ p-value(s, N, B) \leq \sum_{b, n_1(b) : HGT(N, n_1(b), B, b) \leq s} \frac{\binom{n_1(b)}{b} \left( N - n_1(b) \right)}{\binom{N - b}{B}} \] (12)

which is tighter than the bound \( s \cdot B \) introduced by Eden et al. [29].

### 2.7 Pattern matching for IUPAC motifs

Degenerate base symbols in DNA/RNA patterns can be represented using the IUPAC alphabet. A symbol in this alphabet may have multiple possible alternatives. Under the commonly used IUPAC system, nucleobases are represented by the first letters of their chemical names: A, C, G and T. This shorthand also includes eleven "ambiguity" characters associated with every possible combination of the four DNA bases. The ambiguity characters were designed to encode positional variations found among families of related genes. The IUPAC notation, including ambiguity characters, is shown in Figure 7 (Wikipedia).

Given a generalized suffix tree generated for the sequences \( S_1, \ldots, S_N \), and a pattern \( P \) represented using the IUPAC notation, we can efficiently output all occurrences of \( P \) in \( S_1, \ldots, S_N \) by traversing the tree. We start from the root node, and recursively visit the root’s descendants as long as the path matches the pattern \( P \). In the worst case, where \( P \) is very degenerate - as in the extreme case of a pattern of the form \( N^k \) (that is, \( k \)-repeats of the symbol \( N \); matching any string of length \( k \)), we will visit all nodes in the tree located at depth \( \leq k \). The latter operation is linear in the total length of the sequences. However, in the common case, since we usually focus on patterns that are rank imbalanced in the list (as in the routine implemented in mmHG-Finder),
this search is assumed to require sub-linear time, which is therefore more efficient than linearly scanning the sequences for \( P \).

![Table of nucleobases and their IUPAC symbols](image)

**Figure 7 – The IUPAC notation (adopted from Wikipedia)**
All 15 non-empty combinations of the nucleobases A, C, G and T are shown.

### 2.8 Using Kd-trees for multi-dimensional sorting

In this section we introduce the challenge of sorting two-dimensional points by their \( x \) and \( y \) coordinates simultaneously. We encountered this challenge when conducting an expression analysis in our work on RBP cooperation with miRNAs [77], which required having two disjoint subsets (denoted as \( A \) and \( B \)) such that \( A \) contains the samples in which PUM1 and miR-410 are highly expressed, while \( B \) contains the samples in which PUM1 and miR-410 express at low levels (Figure 8). To identify an appropriate partition for this purpose, we performed a class discovery process. Let the vector \( x = x_1, ..., x_N \) be the expression profile of miR-410 in samples \( 1, ..., N \) and let the vector \( y = y_1, ..., y_N \) be the expression profile of PUM1 in samples \( 1, ..., N \). Since \((x_i, y_i)\) is the expression of miR-410 and PUM1 in the \( i \)th sample, respectively, we can define the set \( P \subseteq \mathbb{R}^2 \) that contains \( N \) points representing the expression of miR-410 and PUM1 in the \( N \) samples.
Figure 8 – Sorting points in a two-dimensional grid

In this analysis, we compared samples having high levels of PUM1 and miR-410 expression with samples having low expression. For that, we first sorted the points in the plane such that points in both ends define the highlighted sets (denoted as $A$ and $B$).

We sort $P$ by constructing a 2-dimensional kd-tree (reviewed in [85]) that will follow the sorting order. At the root we split the set $P$ with a vertical line into two subsets of roughly equal size. $P_{\text{left}}$, the subset of points to the left or on the splitting line is stored in the left subtree, and $P_{\text{right}}$, the subset to the right of it is stored in the right subtree. At the left child of the root we split $P_{\text{left}}$ into two subsets with a horizontal line: the points below or on it are stored in the left subtree of the left child, and the points above it are stored in the right subtree. Similarly, the subset $P_{\text{right}}$ is split with a horizontal line into two subsets, which are stored in the left and right subtrees of the right child. At the grandchildren of the root, we split again with a vertical line. In general, we split with a vertical line at nodes whose depth is even, and we split with a horizontal line at nodes whose depth is odd. The algorithm terminates when all the subtrees are leaves. Scanning the leaves from left to right produces the sorted list. The construction of the kd-tree uses $O(N)$ storage and takes $O(N\log N)$ time.

Having the sorted list of points $p_1, \ldots, p_N$, such that $p_1 < p_2 < \ldots < p_N$, we define the following configurations: $S_k = \{(A_k, B_k) \mid A_k = \{p_{N-k+1}, \ldots, p_N\} \text{ and } B_k = \{p_1, \ldots, p_k\}\}$, $k = 1, \ldots, \left\lfloor \frac{N}{2} \right\rfloor$.

Given a figure of merit, we calculate the figure of merit for all the configurations, and take the configuration $S_k = \{A_k, B_k\}$ which holds the optimal value.
2.9 Expression data normalization

As part of collaboration with Yossi Yarden’s Lab at the Weissmann Institute, we analyzed micro-array chips measuring mRNA and lncRNA expression levels. In this section we describe the normalization and filtration process used for processing the raw data, which could then be used for further bioinformatics analyses.

The raw data was interpreted as a collection of matrices in which rows correspond to probes, and columns correspond to experiments (each experiment is a replicate of a time point, where there are seven time points). The values within the matrix cells indicate a measured quantity. Two key matrices were used - the processed signal matrix, containing the signals measured for each probe in each experiment; and the processed signal error matrix assessing the signal error for each probe in each experiment. The raw data was normalized and filtered in order to make the experiments comparable with each other and to reduce noise.

For each combination of a probe and an experiment, the error to signal ratio was calculated, and only entries having ratios below 20% were considered as valid in later steps. Then, the signal matrix was normalized column-wise, that is, per experiment. For each experiment, we considered all valid signal values in the corresponding column (i.e. those having error to signal ratio below 20%) and multiplied them with a constant equal to 1000/q75, where q75 indicates the 75th percentile value amongst the original valid signal values. We used the 75th percentile as this is assumed to be the median expression value of the expressing genes (roughly, 50% of the values are assumed to be background). After normalization, the 75th percentile in each experiment would be equal to 1000.

After the values of each experiment were normalized, probes having low quality measurements were filtered out. For each time point \( t_i \) \( (t_i \in \{0,20,40,60,120,240,480\}) \) and for each probe, we had two or three observations, denoted as \( o_1, \ldots, o_n \) \( (n = 2 \text{ or } 3) \). In this process, we calculated the error to signal ratios for \( o_1, \ldots, o_n \). If there was at least one observation \( o_j \) having error to signal ratio smaller than 20%, then that probe was considered to have a valid entry at time \( t_i \). When having more than one valid observation, the normalized signal values were averaged. In order to be filtered in, a probe was required to have valid entries for all time points \( t_1, \ldots, t_7 \). Eventually, we were left with 40104 probes (63.7%), out of which 6170 were ncRNA probes.
3 Results

3.1 A structural-based statistical approach suggests a cooperative activity of PUM1 and miR-410 in human 3'-untranslated regions

A structural-based statistical approach suggests a cooperative activity of PUM1 and miR-410 in human 3′-untranslated regions

Limor Leibovich1, Yael Mandel-Gutfreund2*, Zohar Yakhini1,3*

Abstract

Background: Micro (mi)RNAs comprise a large family of small non-coding RNAs that are thought to regulate a large fraction of protein-coding genes. Generally, miRNAs downregulate messenger (m)RNA expression by binding to the 3′-untranslated regions (UTRs) of the RNA molecules. An important factor for binding specificity is the matching in the seed region. In addition, target site accessibility is thought to be crucial for efficient repression of miRNA targets. Several recent studies indicated that miRNA repression can be facilitated by RNA-binding proteins. In this study, we examine the conjecture that RNA-binding proteins are involved in ushering miRNAs to bind targets that are initially less accessible.

Results: We analyzed human 3′-UTR sequences containing potential binding sites of 153 conserved miRNA families, and ranked sequences around the sites according to their miRNA accessibility. By applying a rank-based motif search tool to these miRNA targets, we found motifs that are enriched among less accessible targets. As expected from our ranking method, most of the significant motifs were GC-rich. However, one AU-rich motif was found to be enriched among miR-410 less accessible targets. This motif resembles the Pumilio homolog 1 (PUM1) consensus binding site. We observed a stronger enrichment of the PUM1 motif in conserved targets than in non-conserved targets; moreover, the enrichment of this motif was found to be conserved in a subset of placental mammals. Further, we analyzed publicly available gene expression data, and found that the mutual expression of PUM1 and miR-410 has a greater negative influence on the expression of low accessibility targets than on other targets, an effect that was stronger than when considering both miR-410 and PUM1 separately.

Conclusions: Taken together, our findings suggest a cooperative relationship between miR-410 and PUM1 in regulating human highly structured 3′-UTRs. This kind of cooperation can allow a second level of regulation of such targets. Considering cases in which miRNAs bind low accessibility targets may help to improve current miRNA prediction tools and to obtain a better understanding of the mechanisms underlying miRNA regulation activity.

Background

Micro (mi)RNAs are small RNA molecules (approximately 22 nucleotides) participating in a large variety of cellular processes in animals, plants and viruses [1-3]. miRNAs act by binding to the 3′-untranslated region (3′-UTR) of messenger (m)RNAs, forming hybrids that consist of the binding site in the 3′-UTR and of the miRNA seed region (positions 2-8 in the miRNA) [4,5].

miRNAs regulate mRNAs through two main mechanisms: mRNA degradation and inhibition of mRNA translation [6]. It has been shown that the match between the mRNA and the miRNA seed region is important for target recognition [1,7]. However, the number of nucleotide matches in the seed is not the only factor that determines site functionality, and other factors such as site accessibility influence the target recognition [8].

Several recent reports have demonstrated that miRNA repression can be reversed or modulated by RNA-binding proteins (RBPs) interacting with the 3′-UTR of target miRNAs. It was reported that the RNA-binding
protein ELAVL1 (embryonic lethal abnormal vision-like protein 1; also known as HuR) reverses miR-122 repression of SLC7A1 (also known as CAT-1) mRNA in human hepatocarcinoma cells subjected to stress [9]. This effect on SLC7A1 is mediated by ELAVL1 translocation from the nucleus to the cytoplasm upon stress, and is accompanied by SLC7A1 release from processing (P) bodies, structures involved in RNA metabolism, leading to active translation of the message. The process requires the association of ELAVL1 with AU-rich sequences in the 3’-UTR region of the SLC7A1 mRNA, through an as yet unknown mechanism. RBP modulation of miRNA-mediated repression has also been reported for dead end homolog 1 (DND1). In zebrafish, it alleviates miR-430 repression of nanos1 and tdrd7 in primordial germ cells [10,11]. DND1 can also relieve the repression of cyclin-dependent kinase inhibitor 1B (CDKNIB) mediated by miR-221 and the repression of LAT52 by miR-372 in HEK293T cells [10]. In zebrafish and in humans, DND1 counteracts mRNA-mediated repression by binding to uridine-rich regions located near the miRNA binding sites within the 3’-UTR of the message. DND1 binding to these sequences might interfere with miRNA-mRNA interaction. Another indication of the functional relationship between miRNAs and RBPs was found in the rat hippocampal neurons, for which treatment with brain-derived neurotrophic factor (BDNF) was shown to partially relieve miR-134 mediated repression of Linki1 [12]. When miRNAs regulate mRNAs, they are assembled into ribonucleoprotein complexes known as the miRNA-induced silencing complex (miRISC) [2]. In contrast to ELAVL1, DND1 and BDNF, which relieve the miRNA repressive function, the TRIM-NHL protein family (NHL-2 in Caenorhabditis elegans [13] and TRIM32 in mouse [14]) increase the activity of specific miRNAs, including let-7, by binding to miRISC components.

Pumilio family (PUF) proteins constitute a highly conserved family of RNA-binding proteins that regulate target mRNAs via binding to their 3’-UTRs [15]. PUF proteins are vital in developmental processes, including stem cell maintenance [16-18]. They are also required for long-term memory, and control neuron excitability and development [19-21]. PUF proteins bind specific RNA sequences in 3’-UTRs that contain a core ‘UGUR’ tetranucleotide followed by sequences that vary between members of this family. mRNA-PUF protein complexes are thought to trigger translational repression or promote mRNA degradation [22-24]. PUF proteins have been recently shown to be associated with miRNAs. It was observed that predicted miRNA binding sites are enriched among validated PUF targets near PUF-binding motifs in humans [25]. In C. elegans, the Pumilio homolog PUF-9 is suggested to cooperate with let-7 family members to repress hbl-1 in the hypodermis and the ventral nerve cord [26]. This repression requires a region of the hbl-1 3’-UTR that contains binding sites for PUF and let-7.

Overall, many examples suggest extensive crosstalk between RBPs and miRNAs. It is likely that additional cases of RBPs modulating miRNA interactions exist. Because efficient repression of miRNA targets is strongly dependent on site accessibility [8], RNA-binding proteins might function as ushers that mediate the opening of the structure, thereby allowing interaction between miRNA and its low-accessibility targets. In this study, we describe a computational approach to seeking evidence for such a mechanism. The approach makes use of a statistical process that includes thermodynamics-based ranking. We highlight one of the cases for which we found significant evidence. Our findings suggest a cooperative mechanism associating the RNA-binding protein Pumilio homolog 1 (PUM1) with miR-410 targeting of low-accessibility target sites in human 3’-UTRs. We found enrichment of the PUM1 binding motif in less accessible miR-410 targets. This association between miR-410 and PUM1 in the context of low-accessibility targets was also significant in other placental mammals (chimpanzee and horse). Furthermore, as a sequence-independent test, we analyzed publicly available gene expression data. We found an inverse relationship between the mutual expression profile of PUM1 and miR-410, and between the expression profiles of the least accessible targets. This inverse relationship was significantly stronger for the combination of PUM1 and miR-410 than for each of them separately. To summarize, by demonstrating a significant association between PUM1 binding sites and highly structured miR-410 targets, our findings suggest that this pair of RBP and miRNA may work together to repress low accessibility targets. Further experiments will be needed to prove this suggested mechanism.

Results
Our conjecture in this work was that RNA-binding proteins might assist miRNAs in their repression of low-accessibility miRNA targets, with the RBP binding mediating the opening of the secondary structure, thus allowing the miRNA to access the mRNA. This kind of cooperation between RBP and miRNA requires a region of the target 3’-UTR to contain binding sites for both RBP and miRNA. The approach we developed for the exploration of such cooperative pairs of RBP and miRNA is described below.

Approach description
To seek cooperative mechanisms for miRNA activity from sequence data, we implemented a computational process as follows (described schematically in Figure 1):
Given miRNA family $\mu$

Obtain predicted targets from TargetScan and identify seed regions

TargetScan targets

Extract sequences around seeds and reduce overlapping sequences. Then compute $\mu$ global accessibility score (GA) for every sequence

Rank the list according to $\mu$ global accessibility score

Search motifs enriched at the top, such as the example $M$, in this case

**Figure 1 Identification of motifs enriched among low-accessibility targets of micro (mi)RNA family $\mu$: the complete process.** Given a miRNA family and its binding sites in 3’-untranslated regions (predicted by TargetScan), we extracted sequence elements such that the seed binding site (red V in the figure) was placed in the middle surrounded by 70 base pairs upstream and downstream. We ranked these sequences according to the global accessibility score of $\mu$, and searched for enriched motifs. The motif UGUUAUAUAU was found to be enriched among miR-410 low-accessibility targets.
1) For a given miRNA family $\mu$, we obtained all the conserved predicted targets of $\mu$, including positions of the seed in the 3′-UTR. The predictions were taken from TargetScan [27-29].

2) For every seed occurrence (red V in Figure 1), we considered 70 base pairs on each side.

At the end of this step, we obtained a collection of sequences of length 147 bp with the binding site for $\mu$ in the middle, denoted as $S_{1\mu}...S_{140}\mu$. 

3) To reduce overlap of sequences, we reduced the collection containing $S_{1\mu}...S_{140}\mu$ using a maximal independent set algorithm (see Methods).

We performed this process for every conserved miRNA family, defined according to [29]. Consequently, for every conserved miRNA family we obtained a set of minimally overlapping conserved sequences containing the predicted target site at the centre of the sequence. The sequences contained the miRNA seed binding site surrounded by 140 bases (70 upstream and 70 downstream), which is sufficient for a reliable prediction of local secondary structures [8].

To detect sequence elements playing a role in structure-driven cooperation with the miRNA of interest given a target sequence $S_i$, we defined a criterion that reflects miRNA accessibility to $S_i$. Accessibility is reflected by the global accessibility score of each miRNA family to each $S_i$. The score takes into consideration both the accessibility of the entire target sequence and the local accessibility of the miRNA binding site (for more details, see Methods). We ranked the target sequences of the miRNA family according to miRNA accessibility, with the least accessible targets located at the top of the list and the more accessible targets ranked lower in the list. Furthermore, to avoid motifs that overlap with the miRNA binding site or with its reverse complement, we masked the miRNA binding site (located at the centre of the sequence) and the nucleotides that are predicted to form base pairs with it. We next sought motifs enriched among the least accessible targets in comparison with the accessible targets. This was done using DRIM [30], which was adapted for finding RNA motifs. To reduce the number of false positive results, we concentrated on relatively long motifs of length 9 bp.

The Pumilio binding motif is enriched in miR-410 low accessibility targets
To search for RNA-binding motifs that may be associated with low accessibility target sites, we repeated the process described above for the 153 conserved miRNA families known for humans [29]. We found 163 enriched motifs for the 153 miRNA families that passed the minimum hypergeometric (mHG) score threshold of $10^{-4}$ (see Methods). Every motif is associated with a miRNA family $\mu$, such that it is enriched among the least accessible targets of $\mu$. As expected from our accessibility score and ranking approach, sequences appearing at the top of the ranked list tended to have greater GC content than did those located at the bottom of the list (see Additional file 1, Figure S1). Therefore, we expected to find enrichment of GC-rich motifs in low-accessibility targets.

Furthermore, we clustered the identified motifs into groups based on sequence similarity, considering only clusters containing ≥3 motifs (Figure 2A). As control, we conducted the process described above for Saccharomyces cerevisiae 3′-UTRs, working with human miRNAs and using a similar number of targets as in humans. Because the miRNA machinery is not known to exist in S. cerevisiae, we considered the results we obtained here as being random (or as negative control). In the S. cerevisiae control, we found 33 enriched motifs versus 163 motifs in humans (at mHG ≤ $10^{-4}$). Clustering the S. cerevisiae motifs yielded clusters containing only one or two motifs (see Additional file 1, Figure S2 for the mHG scores of human motifs versus S. cerevisiae motifs; for more details on this control, see Methods).

Among the results for humans, we detected one exceptional cluster having a relatively low GC content (the multiple sequence alignment for this cluster is shown in Figure 2B). In this cluster, the motif found for miR-410 (UGUAUAUAU) contained only one G (11.1% GC). Interestingly, no motif having such low GC content was found in the S. cerevisiae control (at mHG score ≤ $10^{-4}$). This motif perfectly contains the consensus binding site of PUM1 and PUM2, which is UGUA-HAUA [25,31], suggesting an association between the RNA-binding proteins of the Pumilio family and the miR-410 low accessibility targets (the motif is shown in Figure 2C, its enrichment among miR-410 low accessibility targets is illustrated in Figure 2D, and the low accessibility targets containing the motif are listed in Table 1; see Additional file 1, Figure S3 for their structures).

Controls
To further study the proposed association, we conducted a list of control experiments, described below. Previous studies have shown that miRNA binding sites are occasionally found in multiple copies in 3′-UTRs [32,33]. Multiplicity of miRNA binding sites has been suggested to be correlated with degree of repression. Because the presence of multiple binding sites of miR-410 in the 3′-UTRs could influence our results, we checked whether binding site multiplicity was correlated with our ranking for miR-410 targets. We found that miR-410 binding site multiplicity did not correlate with the accessibility ranking of miR-410 predicted targets.
Figure 2 Clustered motifs found to be enriched among micro (mi)RNA low-accessibility targets in humans. (A) The motifs found to be enriched among miRNAs low-accessibility targets were clustered according to sequence similarity. Clusters containing at least three motifs are shown. For each cluster, we calculated the average GC content over all motifs in the cluster. The clusters were sorted according to their GC content. Note that the GC content axis is variable. (B) The multiple sequence alignment of the motifs belonging to the cluster with the lowest GC content is shown. For each member of the cluster, the miRNA family for which it was found and the enrichment P value are mentioned. (C) The logo of the motif enriched among miR-410 least accessible targets is shown. This motif holds the lowest GC content found. (D) On the left, the occurrences of UGUAUAUAU among miR-410 predicted targets are shown. The targets were ranked according to the miR-410 global accessibility score. The location of the miRNA binding site is marked with a pink rectangle on the x axis. In the middle, the occurrences vector plot illustrates the number of motif occurrences in each sequence (black for one occurrence, blue for two occurrences). On the right, the actual motif accumulated occurrences versus the expected motif accumulated occurrences in a random dataset are shown, highlighting the observed enrichment. The dashed line (and asterisk) indicates the minimum hypergeometric (mHG) cutoff used for partitioning the sequences into two subsets (such that the motif is enriched in the upper subset).
avoid enrichment of motifs derived from the miRNA masked the miRNA binding site and its complement to enriched motifs in each subset, using MEME [34]. We accessible targets (of miR-410) and searched for we took the 100 least accessible targets and 100 most ing this explanation.

Fourteen genes at the subset of top 105 low accessibility targets contain the motif UGUAAUAUA. Their gene symbols, gene names and PANTHER [49] molecular functions are shown.

(Pearson correlation coefficient = 0.062), thereby ruling out multiplicity as a driver for the aforementioned enrichment (for more details, see Methods).

An alternative explanation for the observed enrichment of the Pumilio motif among low accessibility miR-410 targets is that Pumilio binding sites are generally more prevalent in highly folded targets. To test this, we calculated the enrichment of the Pumilio consensus motif [25,31] in UTR sequences ranked according to ΔG calculated from the predicted secondary structure. Interestingly, we did not observe any enrichment of the Pumilio motif in highly structured UTRs (Pearson correlation coefficient = 0.062), thereby ruling out multiplicity as a driver for the aforementioned enrichment (for more details, see Methods). These results reinforce the association between miRNA accessibility and the observed enrichment of the Pumilio motif in the least accessible miR-410 targets.

To examine whether Pumilio binding sites are generally enriched among GC-rich sequences, we ranked the predicted targets of each miRNA family in our list of conserved miRNA families according to their GC content, and recalculated the enrichment of the Pumilio consensus [25,31]. Among the 153 conserved miRNA families (including miR-410), the best enrichment found for the Pumilio motif had a P value of 0.21, thus excluding this explanation.

Furthermore, as a control for our ranking approach, we took the 100 least accessible targets and 100 most accessible targets (of miR-410) and searched for enriched motifs in each subset, using MEME [34]. We masked the miRNA binding site and its complement to avoid enrichment of motifs derived from the miRNA binding site. In the subset of least accessible targets, the most enriched motif found was an AU-rich motif that is similar to the Pumilio motif (its regular expression is AU[AG][UC]AUAUAUAUAUA; e-value = 1.4 × 10^{-30}). This motif (or any similar motif) was not found for the subset of accessible targets. Moreover, the best e-value per motif in the latter subset was 7.3 × 10^{-6}.

To assess whether the observed association between the Pumilio motif and the predicted, least accessible miR-410 sites could reflect a functional relationship between the RBP and the miRNA, we generated a subset of predicted miR-410 targets that are evolutionarily conserved and thus more likely to be functional miR-410 sites [35]. We then compared the enrichment of the Pumilio motif among predicted targets of human miR-410 taken from three datasets: conserved targets, targets with no restriction on conservation, and non-conserved targets. We found that Pumilio was most significantly enriched in the conserved dataset (P = 1.57 × 10^{-5}), whereas it was less enriched in both the miR-410 predicted targets with no restriction on conservation (P = 2.15 × 10^{-3}) and in the non-conserved dataset (P = 4.2 × 10^{-2}; for more details, see Methods). The higher enrichment of Pumilio binding sites among the conserved, least accessible miR-410 targets may indicate that this association is related to miRNA function.

To further validate that the Pumilio motif is functionally related to miR-410, we used TargetScan to predict a set of miR-410 pseudo-targets (computed based on sequence match with the miRNA seed) in the 3'-UTRs of organisms lacking miR-410 (C. elegans) or miRNAs in general (S. cerevisiae). We then calculated the

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Gene name</th>
<th>PANTHER molecular function</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAM120C</td>
<td>Family with sequence similarity 120C</td>
<td>Unclassified</td>
</tr>
<tr>
<td>C1orf102</td>
<td>Chromosome 14 open reading frame 102</td>
<td>Unclassified</td>
</tr>
<tr>
<td>NRP1</td>
<td>B-cell CLL/lymphoma 11B (zinc finger protein)</td>
<td>Zinc finger transcription factor Nucleic acid binding Kinase</td>
</tr>
<tr>
<td>DKG</td>
<td>Diacetylglcerol kinase, gamma 90 kDa</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>RAI1</td>
<td>Retinoic acid induced 1</td>
<td>CAM family adhesion molecule</td>
</tr>
<tr>
<td>MCA</td>
<td>Melanoma cell adhesion molecule</td>
<td>Other G-protein modulator</td>
</tr>
<tr>
<td>STARD13</td>
<td>START domain containing 13</td>
<td></td>
</tr>
<tr>
<td>CELSR1</td>
<td>Cadherin, EGF LAG seven-pass G-type receptor 1 (flamingo homolog, Drosophila)</td>
<td>G-protein coupled receptor Cadherin</td>
</tr>
<tr>
<td>ZC3H7B</td>
<td>Zinc finger CCCH-type containing 7B</td>
<td>Nucleic acid binding</td>
</tr>
<tr>
<td>NDST1</td>
<td>N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 1</td>
<td>Other transferase</td>
</tr>
<tr>
<td>FZD5</td>
<td>Frizzled homolog 5 (Drosophila)</td>
<td>Molecular function unclassified</td>
</tr>
<tr>
<td>HACE1</td>
<td>HECT domain and ankyrin repeat containing, E3 ubiquitin protein ligase 1</td>
<td>Ubiquitin-protein ligase</td>
</tr>
<tr>
<td>TEC</td>
<td>Tec protein tyrosine kinase</td>
<td>Non-receptor tyrosine protein kinase</td>
</tr>
<tr>
<td>PRRX1</td>
<td>Paired related homeobox 1</td>
<td>Homeobox transcription factor Other DNA-binding protein</td>
</tr>
</tbody>
</table>

To further validate that the Pumilio motif is functionally related to miR-410, we used TargetScan to predict a set of miR-410 pseudo-targets (computed based on sequence match with the miRNA seed) in the 3'-UTRs of organisms lacking miR-410 (C. elegans) or miRNAs in general (S. cerevisiae). We then calculated the

Table 1 MiR-410 low accessibility targets containing the motif UGUAAUAUA

Leibovich et al. Silence 2010, 1:17
http://www.silencejournal.com/content/1/1/17
enrichment of the motif among low accessibility targets for each organism (using broadly the same process described above). As expected, the Pumilio motif was not found to be enriched in any of these organisms (*S. cerevisiae*: $P = 0.35$, *C. elegans*: $P = 0.76$). Next, we applied the latter test to placentals mammals in which miR-410 is conserved [29]. For this analysis, we used the dataset of conserved targets predicted by TargetScan for chimpanzee, horse and dog. Here, again, the Pumilio motif was found to be enriched among miR-410 low-accessibility targets, specifically in chimpanzee ($P = 5.28 \times 10^{-4}$) and horse ($P = 1.4 \times 10^{-4}$), but it was only weakly enriched in dog ($P = 3.3 \times 10^{-3}$). It is important to note that these organisms were chosen because the sizes of their datasets are very similar to that of the human dataset, and thus the differences in $P$ values cannot be due to the size of the datasets.

**Additional controls using validated Pumilio targets in humans**

To further investigate the relationship observed between miR-410 and the Pumilio family and to show its dependence on target accessibility, we used experimental Pumilio binding data available for humans [25]. We obtained 3'-UTR sequences of validated Pumilio targets and ranked them according to their corresponding affinity to Pumilio, as reported previously [25]. We then calculated the enrichment of the miR-410 binding site among validated Pumilio targets. The miR-410 binding site was not found to be enriched among Pumilio targets (PUM1: $P = 0.86$, PUM2: $P = 0.48$). As a control, we calculated the enrichment of the Pumilio recognition motif (UGUAHAUA [25,31]) among validated Pumilio targets, and found it to be strongly enriched among PUM1 targets ($P = 2.47 \times 10^{-5}$) and weakly enriched among PUM2 targets ($P = 2.7 \times 10^{-2}$). To summarize, the analysis presented here demonstrates that the miR-410 binding motif is not generally enriched in Pumilio targets (see Additional file 1, Figure S4). These findings indicate that the observed association of Pumilio and miR-410 cannot be explained by a general association between them. They therefore support a more specific role of Pumilio, closely associated to low-accessibility miR-410 targets.

Next, we evaluated the association of Pumilio proteins with other miRNAs in the context of low-accessibility targets using the knowledge of experimentally validated Pumilio targets [25]. We used mHG statistics to calculate the enrichment of validated Pumilio targets in the list of miRNA targets ranked according to global accessibility (see Methods). We did this for all 153 human miRNA families. Of the 153 experiments, the most significant enrichment was found for miR-410 ($P = 1.4 \times 10^{-2}$; 14 out of 388 miR-410 targets are validated Pumilio targets and all are in the list of the top 251 least accessible targets).

**PUM1 cooperation with miR-410 based on gene expression analysis**

The aforementioned experiments suggest the involvement of the Pumilio family in modulating miR-410 repression of low-accessibility targets. In this section, we used gene expression data (on the NCI60 panel [36,37]) to test PUM1 cooperation with miR-410.

According to the suggested mechanism, two main observations should hold. First, in the presence of high levels of PUM1 and miR-410, there should be a stronger repression effect on the set of least accessible targets than in the presence of low levels of PUM1 and miR-410. The set of least accessible targets (see Methods) is hereby denoted by $\Phi$. The difference in the extent of repression comparing the two states (defined according to PUM1 and/or miR-410 expression levels) is hereafter termed as ‘differential repression’. Differential repression is expected to be more dramatic for $\Phi$ than it is for other (more accessible) miR-410 targets. Indeed, the significance of differential repression of the least accessible targets compared with the rest of the targets was $6.9 \times 10^{-4}$ (for more details, see Methods). The second expected observation is that the differential repression of the least accessible targets, $\Phi$, versus the rest of the targets should be less significant than when partitioning samples according to only miR-410 or only PUM1, owing to the suggested cooperative interaction. Indeed, the mutual differential repression was ~25-fold more significant than when considering only miR-410, and ~10-fold more significant than when considering only PUM1 (Figure 3, A-C).

To validate that the mutual repressive influence of PUM1 and miR-410 is related to the low accessibility of the miRNA targets, we repeated this process 1,000 times, each time for a randomly drawn subset of genes, $\Phi'$, taken from the pool of miR-410 predicted targets (each $\Phi'$ contains the same number of elements as $\Phi$; each $\Phi'$ is disjoint from $\Phi$). As shown in Figure 3D, of the 1,000 experiments, only one result was better than $6.9 \times 10^{-4}$ ($P = 3.55 \times 10^{-4}$).

**Discussion**

The extent to which miRNAs interact with low accessibility targets is not clear, but if such binding takes place there could be a molecular mechanism allowing the miRNA to bind highly structured targets, possibly by involving RNA-binding proteins. To date, there is increasing evidence for cooperation between miRNAs and RBPs [9,10,12-14,26]; however, the possible role of RBPs in facilitating miRNA binding to inaccessible sites has yet not been examined carefully. In this study, we
Figure 3 Cooperative repressive influence of PUM1 and miR-410 on least accessible targets. (A) In this test, we compared samples having high levels of PUM1 and miR-410 expression with samples having low expression, and calculated the degree of differential repression for all targets. We found that the set of least accessible targets was significantly differentially repressed compared with other targets ($P = 6.9 \times 10^{-4}$). Note that the expression values of PUM1 and miR-410 are drawn schematically. (B,C) We repeated the test described in (A) but now ignored PUM1 and miR-410, respectively. In each of these tests, the differential repression of the least accessible targets compared with other targets was weaker than in (A) ($P$ values obtained are indicated in the figure). (D) We repeated the test described in (A) for 1,000 randomly drawn subsets instead of the subset of least accessible targets. Each random subset contained the same number of targets as the subset of least accessible targets and was disjoint from it. The differential repression of 999 subsets was less significant than $6.9 \times 10^{-4}$. 

Leibovich et al. Silence 2010, 1:17
http://www.silencejournal.com/content/1/1/17
sought sequence motifs that are enriched among miRNA low-accessibility targets, because these motifs may point to RNA-binding proteins that take part in this kind of mechanism. We clustered the motifs found into groups based on sequence similarity (see Methods). The cluster containing the motif UGUAUAUAU that was found for miR-410 (Figure 2B) was exceptional because of its low average GC content (33%). The above motif was especially interesting because it is the well-characterized Pumilio recognition motif [25,31] and was unusually AU-rich. The other motifs in the cluster comprised UG runs known as binding motifs of the homogenous RNP protein trans-activation-response DNA binding protein (TARDBP; also known as TDP-43) [38].

Testing the enrichment of the Pumilio motif among miR-410 low-accessibility targets using various controls strongly supported the conjecture that this association is related to the miRNA accessibility of the targets. In addition, the accessibility-dependent association between PUM1 and miR-410 was found to be conserved in humans, chimpanzee and horse, and it was more significant in the human conserved dataset than in the human non-conserved dataset, implying functionality. Moreover, analyzing publicly available gene expression data revealed that mutual expression of PUM1 and miR-410 has a greater negative influence on low-accessibility targets than on other targets. Interestingly, mutual expression of PUM1 and miR-410 had a greater negative influence than did their individual influences separately, supporting our conjecture that miR-410 and PUM1 act together, possibly by facilitating the repression of low-accessibility miR-410 targets.

In a previous study by Fiore et al. [21], it was demonstrated that the miRNA cluster miR379-410 (containing miR-410) is transcriptionally activated upon activation of mouse cortical neurons, and that one component of this cluster, miR-134, takes part in triggering activation-dependent dendritogenesis. The latter study further suggests that Pum2 is a miR-134 target in this process. Based on the verified cooperation between Pumilio and the miR379-410 cluster in dendritogenesis, we speculate that the cooperation between miR-410 and Pumilio suggested in this study might be part of the process regulating dendritogenesis. Clearly, further experimental assays should be carried out to validate this observed association of miR-410 and PUM1 in human 3’-UTRs, and to explain the related mechanisms.

**Conclusions**

An association between miRNAs and Pumilio was suggested previously [25], based on the observation that predicted miRNA binding sites are enriched among validated Pumilio targets near Pumilio binding motifs. In this work, we did in fact observe an association, specific to the context of low accessibility, between Pumilio binding sites and miR-410 targets. We suggest that PUM1 and miR-410 may cooperate in repressing highly structured targets, allowing a second level of regulation of these targets. We hypothesized a mechanism in which PUM1 plays a role in ushering miR-410 to highly structured targets. It is likely that additional pairs of miRNAs and RBPs, as yet undiscovered, cooperate in a similar way. Taking into account the possibility that miRNAs can bind low-accessibility targets with the assistance of RNA-binding proteins may help in improving the accuracy of miRNA target prediction tools and in identifying novel regulatory mechanisms.

**Methods**

**NCI60 dataset**

This dataset comprised a panel of 60 cancer-derived cell lines. For each of the 60 cell lines, we obtained miRNAs and miRNA expression data from the literature [36,37].

**Human validated Pumilio targets dataset**

Lists of PUM1 and PUM2 target mRNAs in HeLa S3 cells were obtained [25]. The dataset included gene information and numerical data related to the measured affinity between Pumilio proteins and the potential targets.

**Maximal independent set**

Given a set of sequence elements $S_1,\ldots,S_L$, which are substrings of a gene 3′-UTR, there can be overlap between sequences. Each $S_i$ is associated with two coordinates that define its start and its end positions in the 3′-UTR. To reduce overlap between sequences, given $S_i$ and $S_j$, we required that start($S_j$)+100 be less than or equal to start($S_i$) (that is, that the distance between the starts of every two sequences in the set must be greater than 100 nucleotides). Consider the interval graph [39] $G = (V,E)$, whose set of vertices is $V = \{S_1,\ldots,S_L\}$ and set of edges is $E = \{(S_i,S_j) \mid i < j \text{ and start}(S_i)+100 > \text{start}(S_j)\}$. Without loss of generality, the sequences are sorted according to their starting positions, such that the starting point of $S_1$ is minimal. To find a maximal set of minimally overlapping sequences, we need to find a maximal independent set in $G$. Because $G$ is an interval graph, the optimal solution can be found efficiently [39].

**TargetScan**

TargetScan predicts the biological targets of miRNAs by searching for the presence of conserved eight-mer and seven-mer sites that match the seed region of each miRNA [27-29]. As an option, non-conserved sites are also predicted. Sites with mismatches in the seed region that are compensated for by conserved 3′-pairing are also identified [29]. We used TargetScanHuman [40],
which considers matches to annotated human UTRs and their orthologs.

TargetScan provides a code that enables making custom predictions of miRNA binding sites for any arbitrary given set of syntactically valid sequences. We used it to obtain predictions in a variety of organisms (such as _S. cerevisiae_, even if they lack miRNA activity).

Conserved miRNA families were defined according to previously published data [29]. They included broadly conserved families (which are conserved across most vertebrates, usually as far as zebrafish) and conserved families (which are conserved across most mammals, but usually not beyond placental mammals).

**RNAfold and miRNA global accessibility score**

RNAfold is a software application that predicts the secondary structures of single-stranded RNA or DNA sequences [41].

Given a miRNA family _μ_ and a target sequence _S_i_, we defined the miRNA global accessibility criterion for _S_i_ as follows:

- We calculated the free energy of the entire sequence, denoted as \( \Delta G_{all}(S_i) \).
- We calculated the free energy of _S_i_ when the area surrounding the seed binding site is forced to be unpaired, denoted by \( \Delta G_{masked}(S_i) \).
- The free energy lost in opening the structure at the binding site of _μ_ in _S_i_ was then defined as:

\[
\Delta G_{open}(S_i) = \Delta G_{all}(S_i) - \Delta G_{masked}(S_i).
\]

This number reflects the local accessibility of the miRNA binding site; the more negative this value, the greater the energy required for opening the target site secondary structure.

- The global accessibility of _S_i_ to _μ_ binding is represented by:

\[
GA(S_i) = \Delta G_{open}(S_i) + \Delta G_{all}(S_i) = 2\Delta G_{all}(S_i) - \Delta G_{masked}(S_i).
\]

It should be noted that to calculate \( \Delta G_{masked}(S_i) \), given a sequence _S_i_, we masked the area surrounding the target site in _S_i_ (25 bases in total, with the miRNA binding site located in the middle) and calculated the free energy of the modified sequence using RNAfold. Technically, putting the letter N in any region in a sequence leads RNAfold to avoid base pairing in that region (see Additional file 1, Figure S5) and thereby masks the region.

**DRIM**

DRIM (discovery of rank imbalanced motifs) is a software application that identifies sequence motifs in lists of ranked DNA sequences [30], and it has also been adapted for RNA sequences [42]. DRIM employs a flexible threshold statistical approach [30,43] to discover motifs that are significantly enriched at the top of a ranked list of sequences compared with the rest of the list.

The motifs returned by DRIM are seed motifs that are sufficiently significant to be passed as input to the heuristic search mechanism of DRIM. The default threshold, which is also used here, is mHG score = 10^-4 (motifs with greater scores are discarded).

**_S. cerevisiae_ control**

We downloaded 3′-UTR sequences of _S. cerevisiae_ from the UCSC Table Browser [44] and predicted the (imaginary) binding sites of human miRNA families within these UTRs using TargetScan script. To allow the same statistical power as in the human dataset, we used a similar number of targets per miRNA family for the tests described in this paper (if there turned out to be more sequences than needed, we would then filter the sequences in the middle of the ranked list). The number of _S. cerevisiae_ targets per miRNA is the same as in humans, except for sporadic cases in which there are fewer targets in _S. cerevisiae_ (on average, the difference between the number of targets in humans and the number of targets in _S. cerevisiae_ per miRNA is 12).

**Procedure for clustering motifs**

Given two sequences _s = s_1,...,_s_n_ and _t = t_1,...,_t_n_, the i-level distance between _s_ and _t_ is defined as _i_ plus the number of mismatches between the sequences _s_1,...,_s_n_ and _t_1,...,_t_n_.

Given two sequences _s = s_1,...,_s_n_ and _t = t_1,...,_t_n_, the distance between _s_ and _t_ is defined as the minimal i-level distance for _i = 0,...,n-1_.

Given two sets of sequences _C_1 and _C_2, we defined the distance between _C_1 and _C_2 as the average of distances between every two sequences _s_ and _t_, such that _s ∈ C_1 and _t ∈ C_2_.

The clustering procedure we applied here is given as its input a set of sequences to cluster and a parameter denoted as diameter. It begins with clusters that are singletons; each sequence is a single cluster. It recursively merges the pair of closest clusters, and halts when the distance between the closest clusters is greater than the diameter given as a parameter. This is a variant of the nearest neighbour hierarchical clustering approach.

To produce the motif Shannon logo for a cluster, we calculated the multiple sequence alignment for the members of the cluster using ClustalW2 [45] and drew the motif using WebLogo [46].

**Multiplicity test**

To test whether multiplicity is correlated with our ranking, we counted the number of miRNA binding sites within each predicted target sequence by counting the
number of matches of the miRNA seed in the sequence. In addition to Watson-Crick base pairing, we allowed wobble base pairs (G-U). We also allowed one mismatch between the miRNA seed and its binding site and counted overlapping binding sites.

**Testing possible association between Pumilio and ΔG**
To calculate the enrichment of Pumilio in highly folded targets, we generated 100 sets of UTR sequences and ranked them according to their ΔG. The length of the sequences and the size of each set were the same as for the miR-410 dataset to allow for the same statistical power. The UTR sequences in each set were chosen randomly.

**Conservation control**
In this control, we used three datasets of miR-410 predicted targets in humans: conserved targets, non-conserved targets, and targets with no constraint on conservation. The conserved predicted targets dataset and the non-conserved predicted targets dataset were obtained from TargetScan. For the third dataset, we downloaded human 3′-UTR sequences from the UCSC Table Browser [44], and calculated miR-410 targets within these sequences using TargetScan script. In this control, as in the S. cerevisiae control, we used the same number of targets for the three datasets. We calculated the enrichment of the Pumilio motif among the target sequences of each set.

**Enrichment analysis (using mHG statistics)**
An approach has been previously described [30] to identify the enrichment of a set of genes, A, in a ranked list of genes using mHG statistics. Given a total number of genes N, with B of these genes belonging to A, and n of these genes being in the target set (for example, differentially expressed genes), the probability that b or more genes from the target set are also in A is given by the hypergeometric tail (HGT):

\[
P(X \geq b) = \text{HGT}(b; N, B, n) = \sum_{i=b}^{\min(n,B)} \left( \begin{array}{c} n \\ i \\ \end{array} \right) \left( \begin{array}{c} N - n \\ B - i \\ \end{array} \right)
\]

If a ranked genes list \( g_{1:n} \) is provided in place of a target set, we define a label vector \( \lambda = \lambda_{1:n} \in \{0,1\}^N \) according to the association of the ranked genes to A, that is, \( \lambda_i = 1 \) if and only if \( g_i \) is in A. The mHG score is then defined as:

\[
mHG(\lambda) = \min_{1 \leq n < N} \text{HGT}(b_n(\lambda); N, B, n)
\]

where

\[
b_n(\lambda) = \sum_{i=1}^{n} \lambda_i
\]

In other words, the mHG score is the optimal HGT probability that is found over all possible partitions induced by the ranking. As such, this score must be corrected for multiple testing. A dynamic programming algorithm for computing the exact \( P \) value of a given mHG score has been described previously [30]. More specifically, given a ranked list of genes, a subset A, and a corresponding mHG score \( s \), the mHG \( P \) value tells us the exact probability of observing an mHG score \( s' \leq s \) under the null assumption that all occurrence configurations of \( A \) in the ranked list are equiprobable.

**Enrichment of Pumilio experimentally validated targets among low accessibility miRNA targets**
The set of validated Pumilio targets contains 1,482 genes (comprising both PUM1 and PUM2 targets). Given a miRNA family denoted as \( \mu \), using the mHG statistics we calculated the enrichment of Pumilio targets in the list of \( \mu \) targets ranked according to their global accessibility scores (such that least accessible targets are at the top) as follows.

1) We used the list of conserved predicted targets, \( S_1,...,S_{100} \), ranked according to global accessibility, GA(\( S_i \)) of \( \mu \) (the calculation of GA(\( S_i \)) is described above).

2) We produced a binary vector \( \lambda(\mu) \) as follows:

For each sequence \( S_i \) in the list of predicted targets for \( \mu \), \( \lambda(\mu) = 1 \) if and only if \( S_i \) was reported as a Pumilio validated target and also contained the Pumilio consensous motif (UGUAHAUA [25,31]). We required \( S_i \) to contain the Pumilio motif in addition to being a validated target because Pumilio target genes have been reported [25] without an indication of where the binding sites reside in the 3′-UTR. In addition, we required that the Pumilio binding site did not overlap with the predicted miRNA binding site or with its complement.

3) We used the mHG statistics on \( \lambda(\mu) \) to calculate the enrichment of validated Pumilio targets in low-accessibility targets of \( \mu \).

**Differential expression**
To compute the differential expression, we used the threshold number of misclassifications (TNoM) score and \( P \) value previously described [47].

**Samples classification algorithm (used in the gene expression analysis)**
The expression analysis requires having two disjoint subsets (denoted as A and B) such that A contains the
samples in which PUM1 and miR-410 are highly expressed, whereas B contains the samples in which PUM1 and miR-410 are expressed at low levels. To identify an appropriate partition for this purpose, we performed a class discovery process. Let the vector \( x = (x_1, ..., x_N) \) be the expression profile of miR-410 in samples 1,...,N and let the vector \( y = (y_1, ..., y_N) \) be the expression profile of PUM1 in samples 1,...,N. Because \((x,y)\) is the expression of miR-410 and PUM1 in the \(i\)th sample, respectively, we can define the set \(P \subseteq \mathbb{R}^2\) that contains \(N\) points representing the expression of miR-410 and PUM1 in the \(N\) samples.

We sorted \(P\) by constructing a two-dimensional k-dimensional \((kd)\)-tree ([48]) that would follow the sorting order. At the root, we split the set \(P\) with a vertical line into two subsets of roughly equal size. \(P_{\text{left}}\), the subset of points to the left or on the splitting line was stored in the left subtree. \(P_{\text{right}}\), the subset to the right of the splitting line, was stored in the right subtree. At the left child of the root, we split \(P_{\text{left}}\) into two subsets with a horizontal line: the points below or on it were stored in the left subtree of the left child and the points above it were stored in the right subtree. Similarly, the subset \(P_{\text{right}}\) was split with a horizontal line into two subsets that were stored in the left and right subtrees of the right child. At the grandchildren of the root, we again split with a vertical line. In general, we split with a vertical line at nodes whose depth was even, and with a horizontal line at nodes whose depth was odd. The algorithm terminates when all the subtrees are leaves. Scanning the leaves from left to right produces the sorted list.

The construction of the \(kd\)-tree uses \(O(N)\) storage and takes \(O(N \log N)\) time.

Having the sorted list of points \(p_1, p_2, ..., p_N\) such that \(p_1 < p_2 < ... < p_N\), we defined the following configurations:

\[
S_k = \{(A_k, B_k) \mid A_k = \{p_{2k-1}, ..., p_{2k}\} \text{ and } B_k = \{p_{2k+1}, ..., p_N\}\}, \quad k = 1, \ldots, \frac{N}{2}
\]

Given a figure of merit, we calculated the figure of merit for all the configurations and took the configuration \(S_k = \{A_k, B_k\}\) that held the optimal value.

We now explain the figure of merit used in our process. Consider a set \(\Phi\) of mRNAs of interest (for example, miR-410 least accessible targets). We evaluated the differential expression of \(\Phi\) due to miR-410 and PUM1 expression levels as follows.

1) Consider a configuration \(A, B\).

2) For every miR-410 predicted target \(\tau\) we calculated the TNoM \(P\) value, measuring whether its expression in \(A\) was lower than its expression in \(B\) (termed as the differential repression of \(\tau\)).

3) We ranked miR-410 targets according to their TNoM \(P\) values in increasing order.

4) We calculated the enrichment of the genes of \(\Phi\) at the top of this ranked list.

5) Finally, we selected the best configuration \(A, B\) using the enrichment calculated in the previous step as the figure of merit.

**Definition of least accessible targets**

Given a list of miRNA targets ranked according to miRNA accessibility, the set of least accessible targets (denoted as \(\Phi\)) was defined as the top 20% targets of the ranked list.

**Additional material**

**Additional file 1: Supplemental Figures S1-S5.**

*Supplemental Figure S1* - Correlation between rank and GC content calculated for the targets of every micro (miRNA) family. *Supplemental Figure S2* - Minimum hypergeometric (mHG) scores of human motifs versus mHG scores of Saccharomyces cerevisiae motifs. *Supplemental Figure S3* - Predicted structures of selected miRNA targets. *Supplemental Figure S4* - The enrichment of the Pumilio binding motif versus the enrichment of the miR-410 binding motif in validated Pumilio targets. *Supplemental Figure S5* - Illustration of masking a region in a sequence.

**Acknowledgements**

We thank Yoav Arava for his valuable review of the manuscript. We thank Israel Steinfeld for his useful comments and for providing access to several software implementations and statistical enrichment tools. We thank Eran Eden, Gal Shaul and Naama Hazan for providing access to a motif visualization tool. We thank the anonymous reviewer for suggesting some controls added after the review. LL was partially supported through a Sherman Fellowship.

**Authors’ contributions**

LL, YMG and ZY designed the study. YMG proposed the research question. LL and ZY developed the methodology and the algorithmic approaches. LL, YMG and ZY developed the implementations and statistical enrichment tools. We thank Eran Eden, Gal Shaul and Naama Hazan for providing access to a motif visualization tool. We thank the anonymous reviewer for suggesting some controls added after the review. LL was partially supported through a Sherman Fellowship.

**Competing interests**

The authors declare that they have no competing interests.

**References**


WebLogo. [http://weblogo.berkeley.edu/logo.cgi](http://weblogo.berkeley.edu/logo.cgi).


http://www.silencejournal.com/content/1/1/17

Cite this article as: Leibovich et al: A structural-based statistical approach suggests a cooperative activity of PUM1 and miR-410 in human 3'-untranslated regions. *Silence* 2010:1:17.

doi:10.1186/1758-974X-1-17
3.2 Efficient motif search in ranked lists and applications to variable gap motifs

Efficient motif search in ranked lists and applications to variable gap motifs

Limor Leibovich¹ and Zohar Yakhini¹,2,*

¹Department of Computer Science, Technion—Israel Institute of Technology, Haifa, 32000 and
²Agilent Laboratories Israel, Petach-Tikva, 49527, Israel

Received November 25, 2011; Revised February 13, 2012; Accepted February 14, 2012

ABSTRACT

Sequence elements, at all levels—DNA, RNA and protein, play a central role in mediating molecular recognition and thereby molecular regulation and signaling. Studies that focus on measuring and investigating sequence-based recognition make use of statistical and computational tools, including approaches to searching sequence motifs. State-of-the-art motif searching tools are limited in their coverage and ability to address large motif spaces. We develop and present statistical and algorithmic approaches that take as input ranked lists of sequences and return significant motifs. The efficiency of our approach, based on suffix trees, allows searches over motif spaces that are not covered by existing tools. This includes searching variable gap motifs—two half sites with a flexible length gap in between—and searching long motifs over large alphabets. We used our approach to analyze several high-throughput measurement data sets and report some validation results as well as novel suggested motifs and motif refinements. We suggest a refinement of the known estrogen receptor 1 motif in humans, where we observe gaps other than three nucleotides that also serve as significant recognition sites, as well as a variable length motif related to potential tyrosine phosphorylation.

INTRODUCTION

Transcription factor (TF) activity is important and central in regulating cellular processes. TFs recognize their specific targets using molecular pattern recognition mechanisms that are not completely understood. One challenge in improving this understanding is the discovery of sequence motifs that partake in this recognition and targeting mechanism. The discovery of sequence motifs helps in constructing models and in explaining sequence variation that may have functional effect.

The occurrence of recognition sequences in RNA molecules plays a central role in enabling controlled regulation by RNA-binding proteins (RBPs) and by microRNAs (miRNAs). For example, the Pumilio family (Puf) proteins constitute a highly conserved family of RBPs that regulate target messenger RNAs (mRNAs) via binding-specific RNA sequences in their 3’UTRs (1). PUF3, PUF4 and PUF5 in Saccharomyces cerevisiae, and PUM1 in humans, are known to bind sequences that contain a core ‘UGUR’ tetranucleotide followed by sequences that vary between members of this family (2–4). PUF1 and PUF2 in S. cerevisiae bind sequences containing the dual UAAU motif (5). In the case of miRNAs, it is known that they act by binding to the 3’UTR of mRNAs, forming hybrids that consist of the binding site in the 3’UTR and of the mRNA seed region (positions 2–8 in the miRNA). MiRNAs play important regulatory roles, and it is estimated that more than one-third of the human protein coding genes are regulated by miRNAs (6).

In addition, protein modification and protein–protein interactions are also potentially driven by mechanisms that involve specific protein-sequence recognition. Many ATP- and GTP-binding proteins have a phosphate-binding loop (P-loop), which typically consists of the sequence motif (A/G)X4GK(S/T), i.e. alanine or glycine followed by a spacer of length 4 that is followed by a glycine, then lysine and a serine or threonine (7,8).

The important role played by sequence elements in molecular regulation and signaling, as discussed above, is the motivation for significant scientific and technological development activity that focuses on measuring sequence-based recognition and on computational approaches and analysis tools designed to improve our understanding of regulation mechanisms involving sequence elements.

Techniques, such as ChIP-chip (9), ChIP-PET (10) and ChIP-seq (11) provide large volumes of genome-wide data on regions of transcription factor binding, measured in actual samples and in various conditions. Similarly, mRNA targets of RBPs are studied using techniques like RIP-chip (4) and CLIP (12). SILAC (13) and other proteomics techniques can be used to characterize the effect of...
amino acid sequences on protein function. Computational tools and approaches to motif discovery form part of the data analysis workflow that is used to extract knowledge and understanding from data generated through the above techniques as well as other measurement approaches.

Motif discovery has attracted much research interest in recent years, resulting in more than a hundred different tools for motif discovery (14). A large subset of motif finders, such as MEME (15), NMica (16), AlignACE (17) or MDscan (18), fit position weight matrices (PWMs) to the sequence data. Most techniques, including those mentioned above, approach motif finding by discovering sequence elements that occur more often than expected in a set of sequences. Some techniques compare a target set to a background set. It is often the case, however, in biological measurement data that results are given as a ranked list of quantities. For example, Gerber et al. (2) report on the set of targets for five RNA-binding proteins from the Pumilio family, including binding ratios given as a ranked list of quantities. For example, Gerber et al. (2) report on the set of targets for five RNA-binding proteins from the Pumilio family, including binding ratios given as a ranked list of quantities. For example, Gerber et al. (2) report on the set of targets for five RNA-binding proteins from the Pumilio family, including binding ratios.

Some techniques compare a target set to a background set. It is often the case, however, in biological measurement data that results are given as a ranked list of quantities. For example, Gerber et al. (2) report on the set of targets for five RNA-binding proteins from the Pumilio family, including binding ratios. For example, Gerber et al. (2) report on the set of targets for five RNA-binding proteins from the Pumilio family, including binding ratios.

Most techniques, including those mentioned above, approach motif finding by discovering sequence elements that occur more often than expected in a set of sequences. Some techniques compare a target set to a background set. It is often the case, however, in biological measurement data that results are given as a ranked list of quantities. For example, Gerber et al. (2) report on the set of targets for five RNA-binding proteins from the Pumilio family, including binding ratios.

One limitation of this algorithm is that it uses positive and negative sets to filter candidate motifs based on ChIP-chip P-value fixed thresholds. Furthermore, as gapped motifs may consist of half-sites that are not independently enriched, this method will potentially miss some significant results. Recently, a gapped PWM model was suggested for discovering variable length DNA-binding sites (33). This model extends the PWM model by introducing an optional gap character, which may appear once and at a certain position inside the motif, to simply model variable-length motifs. A main drawback of this model is that the gap can be one base only.

As evidenced in the work cited above, the search for variable gapped motifs (VGMs) poses a tremendous computational challenge, as the search space becomes huge, when considering parameters of biological relevance. Specifically, if we seek DNA motifs containing two-half sites, each of length 4, where the gap between the 4-mer half sites can be any subset of the numbers \( \{0, \ldots, 10\} \), then an exhaustive search will span \( 4^5 \times 2^{11} \) candidates, which is far too large to routinely address in reasonable time.

We mentioned the role of amino acid sequence motifs in mediating ATP and GTP binding. Another example of the role of protein motifs is the motif HRDLAARN which is conserved in the catalytic domain of protein-tyrosine kinases (34–36) and which we come back to in our ‘Results’ section. Protein alphabet is large and therefore protein motif search also poses a computational challenge.

Therefore, to address the discovery of VGMs, long motifs, or to deal with large alphabets, we cannot take a naïve approach that uses exhaustive search over the motif space. In this work, we developed an efficient statistical and algorithmic approach to searching motifs in ranked lists of sequences. Our method, implemented in DRIMUST, has several advantages over existing methods. First, unlike many other approaches, it does not exhaustively search over motif spaces and therefore can detect variable gapped motifs, long motifs and motifs over large alphabets. Furthermore, we search motifs in ranked lists and not in fixed sets of sequences as is the case for many other methods. And, most important, to the best of our knowledge, DRIMUST is unique in efficiently addressing variable gap motifs under a definition that allows full flexibility of the gap. Specifically, methods that fit PWMs to the sequence data are limited in addressing variable gap motif discovery under this broad definition. This intrinsic shortcoming of PWM...
representations is evidenced, for example, by the limitations (to a single character gap) of the model described in (33).

Our approach to overcoming the computational challenges associated with large motif search spaces is based on using suffix trees, to restrict our attention to motifs that actually occur in the input list of sequences. A suffix tree is a data structure that represents all the suffixes of a given string in a way that allows fast implementation of many string operations. Suffix trees are useful in many application contexts, including bioinformatics and computational biology (37,38). Perhaps the simplest application example comes from the context of text search. Consider the substring problem, whose input is text \( T \) of length \( M \). After an \( O(M) \) preprocessing, one is required to determine, for any string \( S \) of length \( m \), in \( O(m) \) time, whether \( S \) occurs in \( T \) or not. Using a suffix tree, these performance bounds can actually be achieved. Constructing a suffix tree for \( T \) can be done in \( O(M) \) time (39–41). Given a substring \( S \) of length \( m \), all we now need to do is search for the path \( S \) in the tree, starting from the root, an \( O(m) \) operation. This \( O(M) \) preprocessing and \( O(m) \) search solution for the substring problem is very surprising and extremely useful, especially since \( M \) may be huge compared to \( m \).

There are several algorithmic approaches to the construction of a suffix tree for a single string in linear time (39–41). These methods easily extend, with small modifications, to represent the suffixes of a collection \( S_1, \ldots, S_N \) of strings. A generalized suffix tree is a data structure that contains all suffixes of a set of strings (see ‘Materials and Methods’ section) and can be built in \( O(M) \) time, where \( M = \sum_{i=1}^{N} \text{length}(S_i) \). A path from the root to a leaf in the tree represents a suffix. Each leaf of the tree holds information about the indices of strings that contain the suffix, and the starting positions of this suffix within each such string. Restoring all occurrences of a suffix is thus enabled. A natural application of generalized suffix trees is finding substrings that are common to a large number of distinct strings. This question can arise in many different contexts, for example in database search, in multiple sequence alignment and in motif search. Finding DNA, RNA or protein substrings that commonly occur in a set of biologically related sequences help to point out regions or patterns that may be functional. This problem can be efficiently solved using a generalized suffix tree (37,38,42).

A useful variant of this problem is finding rank imbalanced motifs, given a ranked list of sequences \( S_1, \ldots, S_N \). Rank imbalanced motifs are substrings that appear more often at the top of the list compared with the remainder of the list. This notion of rank imbalanced motifs was introduced by Zilberstein et al. (43) and by Eden et al. (20) who described the mHG statistics used for the assessment of such motifs. A unique feature of the mHG statistics is that the cutoff between the top and the rest of the list is determined in a data-driven manner so as to maximize the motif enrichment. This is done by computing the motif enrichment over all possible set partitions and identifying the cut-off at which maximal statistical significance is observed. Here, we present an algorithm that uses generalized suffix trees for an efficient enumeration of motif candidates, which are then assessed using the mHG statistics. The occurrences of each candidate motif in the list are extracted from the extra information stored in the leaves of the tree.

In summary, the contribution of this article is:

1. An efficient algorithm for searching motifs in ranked lists of sequences. The efficiency of our approach enables us to search motifs over large alphabets (such as amino acids) as well as motifs of length 20 or more, all in a reasonable time. Specifically, searching for DNA motifs of length 4–20 in \( S. cerevisiae \), over 6000 sequences (each of length 500 bp), takes less than 3 min on a standard PC.

2. An extension of the above approach that enables searching for variable gap motifs efficiently. We are not aware of any other method that can efficiently search variable gap motifs while allowing full flexibility in the gap model. Specifically, searching, in \( S. cerevisiae \), for VGMs comprising two 4-mer half sites separated by gaps that form any subset of the lengths \( \{0, \ldots, 10\} \) (~6000 DNA sequences, each of length 500 bp) takes less than 7 min on a standard PC.

3. An efficient implementation of the above approaches. Software is available at http://bioinfo.cs.technion.ac.il/people/zohar/DRIMUST-code-VGM/.

4. Biological results:
   (a) Validation of known DNA, RNA and protein motifs through the analysis of high-throughput measurement data sets. For example, we validated the binding motifs of the transcription factors REB1 (TTACCCG), ABF1 (ATCAN6ACGA) and GCN4 (TGACTCA) in \( S. cerevisiae \).
   (b) Suggested refinement and a potential better characterization of known motifs. For instance, we found a significant enrichment of the palindromic motif comprising the half sites GTCA and TGAC, which are separated by a gap of length 3, 6 or 9 bp, among human estrogen receptor 1 (ESR1) binding sequences.
   (c) Hypothesized novel motifs. Interestingly, a variable gapped motif comprising the dual CGCG half site was predicted as the DNA-binding motif of SW16 in \( S. cerevisiae \). The gap in this case can be of any length in 1, 3, 5, 8, 9 and 10. Another significant prediction was the motif HRDLAARN-\(X^{12}\)-DFGL-X\(^{33-39}\)-SDVW, found among tyrosine phosphorylated peptides.

In the ‘Materials and Methods’ section, we start with describing algorithms for finding fixed structure motifs in ranked lists. We then extend this framework to efficiently solve the problem of variable gapped motifs. These algorithms are further applied to biological data sets. More technical details related to statistics, algorithms and to the actual data analysis are also included in the ‘Materials and Methods’ section. Short descriptions of the biological findings are presented in the ‘Results’ section. Finally, we address significance, advantages and limitations of our approach and biological results in the ‘Discussion’ section.
MATERIALS AND METHODS

Algorithmics

In this section, we describe our suffix tree approach to motif searching in the context of large alphabets and of ranked lists of sequences, based on the mHG statistical model (all definitions appear later on in this section). We start with the non-gapped version and then move to the more complicated-gapped version. We discuss performance in the context of biologically relevant properties of the input.

DRIMUST—algorithm for the case of a fixed structure motifs

The algorithm is described in Table 1.

Time complexity. Given a path $P$ of length $k$ in the tree, to calculate $P$'s enrichment one first needs to find $C(P)$—the list of $P$'s occurrences. Assuming that $P$ occurs at most once in a sequence, this takes time proportional to $k + |C(P)|$, because walking on the path $P$ in the tree is proportional to $k$ ($P$'s length), and the size of the subtree rooted at $P$ is proportional to the number of leaves in that subtree, which is equal to $|C(P)|$. Then, the mHG score is calculated by computing HGT for every member in $C(P)$, such that each HGT calculation takes $O(|C(P)|)$ time. In total, calculating $P$'s enrichment is an $O(|C(P)|^2)$. Therefore, the total complexity of this algorithm is $\Psi = \sum_p |C(P)|^2$, where $P$ ranges over all paths of length $k$ in the suffix tree ($k_1 < k < k_2$).

To evaluate this quantity, we generated 100 lists of sequences taken randomly from the *S. cerevisiae* genome, each sequence of length 500, and every list contained 1000–2000 sequences. We then estimated $\Psi$ for motifs (k-mers) of lengths 7, 10 and 13 in all lists.

**Table 1. DRIMUST—fixed-structure motifs algorithm**

<table>
<thead>
<tr>
<th>Input:</th>
</tr>
</thead>
<tbody>
<tr>
<td>A ranked list of sequences $S_1, ..., S_N$</td>
</tr>
<tr>
<td>A range of motif lengths $[k_1, k_2]$</td>
</tr>
<tr>
<td>P-value threshold for reporting ($\tau$)</td>
</tr>
<tr>
<td>Output:</td>
</tr>
<tr>
<td>A list of sequence motifs of lengths between $k_1$ and $k_2$ that are rank imbalanced in $S_1, ..., S_N$ at an mHG significance level better than $\tau$.</td>
</tr>
</tbody>
</table>

Preprocessing:

Construct a generalized suffix tree for $S_1, ..., S_N$ such that:

- All suffixes of all sequences $S_1, ..., S_N$ are represented by paths from the root to leaves in the tree.
- Each leaf contains information about the occurrences of the corresponding suffix $w$ in $S_1, ..., S_N$. This information is represented as a list $m_1(w), ..., m_{N(w)}(w)$. The values $m_j(w)$ are the indices, amongst $1, ..., N$, of the sequences at which $w$ occurs.

The construction is implemented using Ukkonen’s algorithm (41) */

Algorithm:

for $k = k_1$ to $k_2$ do:

Traverse the tree to find paths of length $k$, and for each path $P$ calculate $P$’s enrichment using the following process:

- Get the ordered list $C(P) = \{m_1(P), m_2(P), ..., m_{N(P)}(P)\}$ of indices (ranks) of sequences containing $P$, extracted from the leaves of the subtree rooted below $P$. */$P$ occurs in the union of the lists of all leaves of that subtree, as it is the prefix of all the suffixes represented by these leaves. For example, assuming $P$ appears in $S_1, S_{14}, S_{11}$ and $S_{86}$, then $C = \{8, 14, 31, 36\}$ */
- Calculate the mHG score for $P$: $mHG(P) = \min_{1 \leq j \leq N(P)} \{HGT(N, P)[m_j(P), 0]\}$ */ Following the example above and assuming we have 100 sequences in the input: $mHG(P) = \min \{HGT(100, 4, 8, 1), HGT(100, 4, 14, 2), HGT(100, 4, 31, 3), HGT(100, 4, 36, 4)\}$.

In this case attained at $i = 4$ where $HGT(100, 4, 36, 4) = 0.015$.*/

- Report $P$ if $N(P) \times mHG(P) \leq \tau$ holds. */$P$--value($mHG(P)$) $\leq N(P) \times mHG(P) (20)$ */
Algorithm: Enrichment analysis (using mHG statistics) for large data sets.

Input:
- A ranked list of sequences $S_1, \ldots, S_N$
- Parameters $[a, g_{\text{max}}, b]$ where $a$ represents the length of the first half site, $b$ represents the length of the second half site and $g_{\text{max}}$ represents the maximum gap.
- $P$-value threshold for reporting ($\tau$)

Output:
A list of sequence motifs of the form $H_1^\Lambda N^\Lambda H_2$, where $\Lambda$ is a set of gaps. These reported motifs are rank imbalanced in $S_1, \ldots, S_N$ at an mHG significance level better than $\tau$.

The interpretation of the above motif representation is as follows. A motif is viewed as a set of strings. In this case all strings that start with $H_1^\Lambda$ then have a wildcard gap of any of the lengths in the set of gaps $\Lambda$, and end with $H_2$. For example, the motif GCC-N$^\Lambda 5$-ATG represents the strings GCCNATG and GCCN5ATG.

Preprocessing:
- Construct a generalized suffix tree for $S_1, \ldots, S_N$ such that:
  - All suffixes of all sequences $S_1, \ldots, S_N$ are represented by paths from the root to leaves in the tree.
  - Each leaf contains information about the occurrences of the corresponding suffix $w$ in $S_1, \ldots, S_N$. This information is represented as a list $m_j(w), \ldots, m_{N_{\text{sch}}}(w)$, where $m_j(w)$ are the indices, amongst $1, \ldots, N$, of the sequences at which $w$ occurs.

The construction is implemented using Ukkonen’s algorithm (41).

Algorithm:
- Traverse the tree to find paths of length $u$, and for each path $P$ do:
  - Compute the set of all strings $\sigma_1(P), \ldots, \sigma_{N_{\text{sch}}}(P)$ of length $b + g_{\text{max}}$ that start at the position where $P$ ends in all sequences among $S_1, \ldots, S_N$ in which it occurs. This step is implemented by traversing the subtree rooted at $P$.
  - The strings $\sigma_1(P), \ldots, \sigma_{N_{\text{sch}}}(P)$ are typically of length $b + g_{\text{max}}$. When $P$ occurs close to the end of $S_n$ a string of length smaller than $b + g_{\text{max}}$ is taken into the above set.
  - Construct a generalized suffix tree $T(P)$ for $\sigma_1(P), \ldots, \sigma_{N_{\text{sch}}}(P)$ such that:
    - All suffixes of all sequences $S_1, \ldots, S_N$ are represented by paths from the root to leaves in the tree.
    - Each leaf contains information about the occurrences of the corresponding suffix $u$ in $\sigma_1(P), \ldots, \sigma_{N_{\text{sch}}}(P)$ as well as the positions of these occurrences. This information is represented as a list of pairs:
      $$(m_1(u), t_1(u)), (m_2(u), t_2(u)), \ldots, (m_{N_{\text{sch}}}(u), t_{N_{\text{sch}}}(u)).$$
  - Traverse $T(P)$ at depth $b$. For each such path $Q$ calculate the enrichment of all possible motifs of the form $P - N^\Lambda - Q$, where $\Lambda$ is any subset of $[0, \ldots, g_{\text{max}}]$, using the following process:
    - Use the values $t_i$ that are in the leaves of the subtree rooted below $Q$ to infer:
      - $\Lambda(P, Q) = \{ \lambda_1, \ldots, \lambda_{N_{\text{sch}}}(P) \} \subseteq [0, \ldots, g_{\text{max}}]$, representing all gaps for which a string $P - N^\Lambda - Q$, where $\lambda \leq g_{\text{max}}$, occurs.
      - $\Lambda(P, Q) = \bigcup_{a=1}^{N_{\text{sch}}(P)} \{ t(Qa), \text{where} \ a \text{ ranges over all sub strings for which } Qa \text{ is a suffix in } T(P) \}$.
    - For every $\Lambda \subseteq \Lambda(P, Q)$ do:/* this step uses the suffix tree information to avoid searching over all $2^{g_{\text{max}}+1}$ possible instantiations of $\Lambda$, leading to improved efficiency of the algorithm*/
      - Use the list $m_1(\Lambda), m_2(\Lambda), \ldots, m_{N_{\text{sch}}}(\Lambda)$ to compute the mHG score for $P - N^\Lambda - Q$:
        $$\text{mHG}(P, Q, \Lambda) = \min_{1 \leq i \leq N_{\text{sch}}(P)} \left[ \text{HGT}(N, N(\Lambda), m_i(\Lambda), b) \right]$$
      - Report $P - N^\Lambda - Q$ if $N(\Lambda) \times \text{mHG}(P, Q, \Lambda) \leq \tau$ holds. /*P-value(mHG(P, Q, $\Lambda$)) ≤ N( $\Lambda$) × mHG(P, Q, $\Lambda$) (20) */

they are sufficiently significant, as determined by the threshold. This heuristic is optional but is recommended for large data sets.

Enrichment analysis (using mHG statistics)
An approach has been previously described (20) to identify the enrichment of a set of genes, $A$, in a ranked list of genes, using mHG statistics. Given a total number of genes $N$, with $B$ of these genes belonging to $A$, and $n$ of these genes being in the target set (e.g. differentially expressed genes), the probability that $b$ or more genes from the target set are also in $A$ is given by the tail of a hypergeometric random variable (HGT):

$$P(X \geq b) = \text{HGT}(N, B, n, b) = \sum_{i=b}^{\min(n, B)} \binom{n}{i} \frac{N - n}{B - i} \binom{N}{B}$$
For a ranked genes list \( g_1, \ldots, g_N \), we define a label vector \( \lambda = (\lambda_1, \ldots, \lambda_N) \in [0,1]^N \) according to the association of the ranked genes to \( A \), that is, \( \lambda_i = 1 \) if and only if \( g_i \in A \).

The mHG score is then defined as:

\[
mHG(\lambda) = \min_{1 \leq n < N} \text{HGT}(N, B, n, b_n(\lambda))
\]

where \( b_n(\lambda) = \sum_{i=1}^{n} \lambda_i \).

In other words, the mHG score is the optimal HGT probability that is found over all possible partitions induced by the ranking. As such, this score must be corrected for multiple testing. A dynamic programming algorithm for computing the exact \( p \)-value of a given mHG score is described in (20).

Comment: In practice, DRIMUST uses Stirling’s approximation to compute all binomial coefficients needed to assess HGTs. Stirling’s inequality says that:

\[
\sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}} \leq n! \leq \sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\frac{1}{12n}},
\]

which is tight for large factorials.

Suffix trees

We used Ukkonen’s algorithm for our implementation of generalized suffix trees (41). An example of such a tree is shown in Figure 2. For more details on suffix trees see (37).

Technical details of the biological data analysis

Human TP53 ChIP-chip analysis

We used the list of TP53 high-confidence binding sites reported by (9), containing 1546 loci in the human genome, each of length 10 bp. We extracted the sequences around these loci, taking 200 bp upstream and 200 bp downstream for each reported locus. These sequences were located at the top of a ranked list, with additional 1546 random sequences taken arbitrarily from the human genome and having the same lengths appended to the end of this list. We then searched motifs in the list.

In order to test the significance of all possible substrings derived from CWWG, we compared the enrichment of all motifs having the structure CW\( _1 \)W\( _2 \)-gap-CW\( _1 \)W\( _2 \), such that W\( _1 \) and W\( _2 \) can be A or T and the gap can be any subset of \( \{0,1,\ldots,10\} \).

To test which half site obeying the pattern RRRCWGWGYYY was the most enriched, we calculated the enrichment of all solid motifs that are derived from this consensus, restricted to WW = AT.

Human estrogen receptor ChIP-chip analysis

We downloaded ESR1 target sequences reported by (44). Using ChIP-on-chip technique, the authors identified regions at least 600 bp in length that were enriched in ChIP samples compared with the controls, in MCF-7 cells. A total of 10 599 regions were predicted at the \( P \)-value cut-off of \( 10^{-3} \). These regions were ranked
according to ESR1 binding P-values. We noticed that sequences at the top of the list were longer compared to sequences at the rest of the list. To get over this length bias, that may affect motif search, we randomly picked 600 bp from each sequence and only then conducted motif search.

To compute the differential expression for the genes in the subset of ESR1 predicted targets, we used the threshold number of mis-classifications (TNoM) score and P-value previously described (45).

**Transcription factor binding sites analysis in S. cerevisiae**

Data was taken from (46). We used the data set containing binding ratios of 203 putative transcription factors in rich media (YPD) conditions for our analysis. In each of the experiments, the genes were ranked according to the TF-binding ratio, and their promoter sequences were analyzed. We extracted the promoter sequences of *S. cerevisiae* genes by taking 500 bases upstream the TSS. For each transcription factor, we created a ranked list of targets, considering only genes that had binding ratio measurement for that TF, and then conducted motif search.

We compared our two best predictions for each TF with other reports (20,46). We considered motifs as ‘similar’ if the optimal number of matches between them was 80% of the length of the shorter motif. Otherwise, they were considered as ‘different’.

**Heat-shock data analysis in S. cerevisiae**

We analyzed gene expression data in *S. cerevisiae* responding to heat shock (47). Changes in transcript levels over eight time points (5, 10, 15, 20, 30, 40, 60 and 80 min) after heat shock were measured for almost every gene. We defined two disjoint subsets of time points, one contained the four early time points and the other contained the four late time points. We calculated the differential expression between these subsets for every gene, using the TNoM statistics (45) and then ranked the genes according to this measurement (in both directions). When genes had the same TNoM score, we internally ranked them according to the difference of their average expression values. For each ranking, we analyzed both promoter sequences and 3'UTR sequences. Promoter sequences were extracted by taking 500 bases upstream the TSS. The lengths of 3'UTRs were defined according to (48). Although the 3'UTR sequences are of variable lengths, there was no significant length bias in any of the experiments.

We also analyzed targets of PUM1, a human member of the Pumilio family. The list of PUM1 targets in HeLa S3 cells was obtained from (3) and contained 1336 sequences. The data set includes gene information and numerical data related to the measured affinity of PUM1 and its potential targets (SAM score). We ranked the 3'UTR sequences according to this quantity. Here also there was not a length bias problem.

**Protein-motif analysis**

We downloaded human phosphorylation data from Phospho.ELM database (49). In three different experiments, one for each phosphorylation type (serine, threonine or tyrosine), each protein sequence was considered once. From each sequence, we picked a substring starting at 50 amino acids upstream the minimal site and extending to 50 amino acids downstream the maximal site. These substrings were located at the top of a ranked list, and we appended to them a set of random sequences taken arbitrarily from the human proteome and having the same lengths. We then conducted motif search. Although the amino acid sequences are of variable lengths, there was no significant length bias in any of the experiments. The serine phosphorylation list contained 7752 sequences in total and the tyrosine phosphorylation list contained 2320 sequences.

As random control, we generated a list of sequences containing the same amino acid content while preserving the RS di-residual density as in the original list. Given a sequence S in the original list, we first counted the number of RS pairs in S, denoted as q. We then randomly placed q RS pairs in the shuffled sequence and dispersed the rest of the amino acids in S randomly in the remaining positions. In the last step, we eliminated randomizations that placed ‘S’ after ‘R’ as they would have changed the RS density.

**Motif refinement as PWM**

Motifs produced as output by DRIMUST are exact words. As a refinement, we implemented a procedure that returns PWMs for the two halves of a given variable gap motif. This extension to DRIMUST takes four parameters as input:

1. The ranked list of sequences
2. A variable gap motif produced as output by DRIMUST
3. $n^*$ — the cut-off in the list at which the motif enrichment is maximized. This parameter is part of the DRIMUST output (together with the motif).
4. $d$ — a Hamming distance threshold

The procedure returns a PWM for each half of the motif by considering all occurrences of variants of the motif in the top $n^*$ sequences in the list, when allowing up to $d$ mismatches in total in the motif halves.
Implementation and availability

We provide an efficient implementation of the algorithms described above in Java. Our application takes as input ranked lists of sequences and returns significant variable gap motifs. It can be freely downloaded from http://bioinfo.cs.technion.ac.il/people/zohar/DRIMUST-code-VGM/ (suitable for all operating systems). The implementation of the procedure that returns PWMs given variable gap motif is also available for download from the above location.

RESULTS

We applied DRIMUST to analyze various data sets by searching for variable gap motifs composed of two half-sites of length 4, and separated by gaps that form any subset of the lengths \{0,\ldots,10\}. We describe our findings in this section. For each result, we report an mHG \( P \)-value as well as a Bonferroni corrected \( P \)-value, addressing the size of the motif search space.

Motif discovery in human TP53 ChIP-chip data

To test the effectiveness of our algorithm, we used it for the identification of TP53 (also known as p53) binding motif in humans by applying it to the Smeenk et al. (9) data set. Using a genome-wide ChIP-on-chip approach, the authors have identified 1546 high-confidence binding sites of TP53. We placed these target sequences at the top of a ranked list, with additional 1546 random sequences (taken arbitrarily from the human genome and having the same length) appended to the end of this list. The complete list served as input to DRIMUST. The most enriched motif found in the output contained the two identical half-sites CATG, separated by a gap of length 6 or 7 (i.e. CATG-N\(^{6,7}\)-CATG; \( P \)-value \( \leq 1.15 \times 10^{-67} \)).

The above is consistent with the literature consensus motif which consists of two copies of the half-site RRRR CryGGY YYYY separated by a spacer of 0–13 bp (50). A gap of length 0 is equivalent to the motif RRRR CryGYY YYYY which is consistent with a spacer of 6 bp between the two half sites CWWG. Among all possible half-sites derived from CWWG, CATG was found to be the most statistically enriched in the data we analyzed (\( P \)-value \( \leq 1.15 \times 10^{-67} \), versus \( P \)-value of CAAG \( \leq 2.45 \times 10^{-14} \); \( P \)-value of CTGT \( \leq 5.25 \times 10^{-9} \); \( P \)-value of CTAG \( \leq 4.74 \times 10^{-4} \). This is consistent with the observation that high-affinity sites contain the CATG sequence at the centre of both half-sites (51,52). Additionally, we observed that the most enriched half site derived from RRRR CryGGYYY was GGC CryGTTCT (\( P \)-value \( \leq 2.89 \times 10^{-8} \)).

Due to the dichotomic nature of the data, the algorithm actually cuts at a fixed place, and a hypergeometric (HG) test at position 1546 would have yielded highly similar results, as would be expected. It should be noted that the TP53 analysis took 3 min on a standard PC.

Motif discovery in human estrogen receptor 1 ChIP-chip data

We further applied our algorithm on ESR1 target sequences, using data collected by Carroll et al. (44). The authors mapped ESR1-binding sites on a genome-wide scale in breast cancer MCF-7 cells, using ChIP-on-chip technology. The list given as input to our method contained the entire data set of sequences \( (N = 10 \, 599) \), and they were ranked according to ESR1 binding \( P \)-values as defined by (44). This mode of analysis, searching motifs in ranked lists, is the preferred mode for DRIMUST. The most enriched motif found in the output was GTCA-N\(^{3,6,9}\)-TGAC; \( P \)-value \( \leq 1.13 \times 10^{-20} \). Corrected \( P \)-value \( \leq 1.52 \times 10^{-12} \). This motif is consistent with the known ESR1-binding motif AGGTCA-N\(^3\)-TGACCT (44), and with the observation that the protein binds as a symmetrical dimer to its palindromic-binding site (53). Our findings, however, expand the known fixed spacer...
length, suggesting a stronger variable gap motif. We then evaluated the fixed gap motifs GTCA-N\textsuperscript{i}-TGAC, where \(0 \leq i \leq 10\), and observed that the most enriched variant was indeed GTCA-N\textsuperscript{3}-TGAC (\textit{P}-value \(\leq 1.12 \times 10^{-18}\); Figure 3B). Moreover, other variants of the variable gap motif were similarly as strong, such as GTCA-N\textsuperscript{0,5,6,9}-TGAC (\textit{P}-value \(\leq 1.56 \times 10^{-20}\)), GTCA-N\textsuperscript{3,9}-TGAC (\textit{P}-value \(\leq 4.40 \times 10^{-20}\)), and GTCA-N\textsuperscript{0,3,9}-TGAC (\textit{P}-value \(\leq 8.58 \times 10^{-20}\)).

The mHG statistics used by our method provides additional information about protein targets. For each motif, it returns as output a cut-off in the list such that the motif is enriched in the subset of sequences located above the cut-off compared to the subset of sequences located below it. This number is denoted as \(n^*\) and can be utilized to predict the targets of the tested protein, by selecting the motif-containing sequences ranked higher than \(n^*\). Applying this approach to ESR1 yields \(n^* = 5083\) (out of 10,599). Amongst the top 5083 sequences, we find 577 that contain the motif GTCA-N\textsuperscript{3,6,9}-TGAC. As validation for this approach, we used publicly available gene expression data (54) containing mRNA expression profiles from 100 primary human breast tumors. We ranked the mRNAs according to their differential expression in samples with high-ESR1 mRNA levels versus samples where ESR1 mRNA levels were low. We found that the subset of ESR1 predicted targets (577 genes defined as above) was enriched among the set of genes whose expression is positively correlated with ESR1 expression (\textit{P}-value \(\leq 4.33 \times 10^{-6}\)).

Motif discovery in \textit{S. cerevisiae} transcription factor binding sites

We next applied our method to the Harbison ChIP-chip data set (46), reporting measurements in 203 transcription factor binding experiments. In each of the ChIP-chip experiments, promoter sequences were ranked according to the transcription factor binding signal. Each transcription factor thus had its unique ranked list of genes, containing 6029 sequences on average. Every ranked list was used as input to our method, which searched for fixed structure motifs of lengths 6–10, and for variable gapped motifs. The motifs found for the Harbison data set are summarized in Supplementary Table S1. We compared our predictions with those reported in (46). The results of this comparison are summarized in Figure 4 and in Supplementary Table S2.

Notably, the most significant variable gap motif was found for the protein ABF1. We found two equally strong motifs for ABF1: TCGT-N\textsuperscript{6}-TGAT (\textit{P}-value \(\leq 3.53 \times 10^{-63}\); Corrected \textit{P}-value \(\leq 4.74 \times 10^{-35}\)) and

\[
\text{Figure 4. Comparison between DRIMUST predictions and other predictions (46). Overall, our method identified significant motifs for all 203 unique transcription factors tested with \textit{P}-value \(\leq 10^{-5}\). Comparing them to the motifs reported in (46) revealed that out of 203 transcription factors, DRIMUST and the other applications found similar motifs for 22 TFs. There were 80 TFs for which DRIMUST predicted different motifs, and 101 for which DRIMUST identified a motif where the other applications did not. Examples for each case are indicated in the figure. We note that in the green set and the red set our predictions for 12/80 and 11/101 TFs are consistent with (20), respectively. Motifs of TFs marked with asterisk are also identified by (20).}
\]
ATCA-N^6-ACGA (P-value ≤ 7.23 × 10^{-63}; Corrected P-value ≤ 10^{-58}), both are fixed gap motifs. Aligning these motifs may yield the consensus motif ATCRTN^5AYGAT, which was also enriched (P-value ≤ 1.53 × 10^{-16}). The second motif (ATCA-N^6-ACGA) is consistent with the literature motif RTCRYN^5ACG (55). We further tested the enrichment of all eight variants derived from the latter consensus and found that the motif ATCACN^5ACG was the most significant one (P-value ≤ 2.58 × 10^{-75}; the results are summarized in Figure 5A). Apparently, adenosine is preferred over guanine at the first and the second positions of purines in the motif.

Interestingly, there are transcription factors for which very similar motifs were found by DRIMUST. One example is GAT3, YAP5 and MSN4 that share a variable gap motif, composed of the half sites GCGG and ACGA separated with a gap of length 7 or 10 nucleotides (GAT3: P-value ≤ 2.49 × 10^{-34}, YAP5: P-value ≤ 5.48 × 10^{-34}, MSN4: P-value ≤ 8.63 × 10^{-17}). Comparing the variable gap in this motif to fixed gaps for GAT3 is depicted in Figure 5B, demonstrating a significant preference for the variable gap. Another example is RGM1, TOS8, CRZ1 and MAL13 for which the most enriched motif was CCTCGACTAA (RGM1: P-value ≤ 6.3 × 10^{-28}, TOS8: P-value ≤ 1.6 × 10^{-16}, CRZ1: P-value ≤ 3.7 × 10^{-16}, MAL13: P-value ≤ 2.1 × 10^{-14}). In addition to the above, very similar motifs were also predicted for PDC2, HAL9 and TEC1—TGT TGGAAATA for PDC1 and TEC1 (P-values ≤ 3.5 × 10^{-38} and 1.6 × 10^{-16}, respectively), and TGTNNGAAT for HAL9 (P-value ≤ 4.92 × 10^{-22}). It should be noted that the predictions in (46) are different or do not exist.

Furthermore, a variable gap motif was found to be strongly enriched among SWI6 targets. This motif comprised two copies of CGCG, separated by the spacer length subset {1, 3, 5, 8, 9, 10} (P-value ≤ 2.76 × 10^{-25}; Corrected P-value ≤ 3.7 × 10^{-17}). This motif differs from the motifs CGCGAAA and CNCGAAA reported in (46). Moreover, testing the fixed gap motifs from the pattern CGCG-N^l-CGCG, where 0 ≤ l ≤ 10, for SWI6, yielded significantly weaker enrichments, supporting a preference for a variable gap motif, in this case (Figure 5C).

**Figure 5.** ABF1, GAT3 and SWI6 motif analysis. (A) Variants of the ABF1 consensus sequence RTCRYN^5ACG were evaluated and their significance P-values are indicated. (B) Eleven motifs having the pattern GCGG-N^l-ACGA, where 0 ≤ l ≤ 10, were tested (dark bars) and compared to the motif GCGG-N^7,10-ACGA (rightmost bar), for GAT3-binding data. (C) Eleven motifs having the pattern CGCG-N^l-CGCG, where 0 ≤ l ≤ 10, were tested (dark bars) and compared to the motif CGCG-N^{1,3,5,8,10}-CGCG (rightmost bar), for SWI6-binding data.
Note that running the complete analysis, as described above, on the Harbison et al. data set, took 30 h on a standard PC.

**Motif discovery in S. cerevisiae heat-shock DNA and RNA motifs**

We analyzed gene expression data in *S. cerevisiae* responding to heat shock (47). The data includes information about changes in transcript levels over eight time points after heat shock for almost every *S. cerevisiae* gene, measured using DNA microarrays. For each gene, we calculated the differential expression when comparing the four early time points with the four late time points, using the TNoM statistics (45). We then ranked the genes accordingly, in both directions. For each ranking, we analyzed both promoter sequences and 3′UTR sequences and used DRIMUST to search for variable gap motifs.

The motif AAAATTTT was found to be enriched among promoter sequences of genes increasing after heat shock (P-value \( \leq 3.40 \times 10^{-33} \); Corrected P-value \( \leq 4.56 \times 10^{-28} \)). This motif has been previously suggested to bring regulatory elements close together, and thus enhance the interactions of transcription factors that bind DNA (56). Another motif, GGCA-N\(^{3,5,6}\)-TGAG was also significantly enriched (P-value \( \leq 3.22 \times 10^{-25} \); Corrected P-value \( \leq 4.32 \times 10^{-17} \)).

Additionally, a motif which is similar to PUF4 recognition site was found to be enriched among 3′UTR sequences of genes increasing after heat shock. The motif UGUA-N\(^{1,2,5,6}\)-AUUA was the most significant motif in the output (P-value \( \leq 9.93 \times 10^{-10} \); Corrected P-value \( \leq 0.08 \)). When testing fixed gap motifs from the pattern UGUA-N\(^{1,2,5,6}\)-AUUA, where 0 ≤ l ≤ 10, the most significant enrichment was attained at I = 1, with P-value = 7.5 \times 10^{-6}. This is also consistent with the literature consensus of I = 1 for PUF4. Our variable gap motif found here is significantly more enriched. Moreover, mHG cuts the ranked list after 415 sequences, out of which 62 contain the motif. We next found that these 62 genes are enriched among PUF4 high-affinity targets (2) (P-value \( \leq 1.71 \times 10^{-13} \)). Additionally, PUF4 itself increases after heat shock according to data. This may point at a positive regulation of these genes by PUF4 triggered by heat shock. We also found another VGM, the motif UUCU-N\(^{3,4,5,7,9,10}\)-UUUA, to be enriched among genes decreasing after heat shock (P-value \( \leq 2.35 \times 10^{-9} \); Corrected P-value \( \leq 0.19 \)).

It should be noted that the analysis of the heat-shock data took 5 min (for both the 3′UTRs and the promoters) on a standard PC.

**RNA binding motifs**

We next applied DRIMUST on mRNA sequences targeted by RNA binding proteins of the Pumilio family in *S. cerevisiae* (2), in five different experiments. Input consisted of the list of *S. cerevisiae* 3′UTRs ranked according to PUF1, PUF2, PUF3, PUF4 and PUF5 binding affinity, respectively. The results are summarized in Figure 6A.

We note that PUF1 and PUF2 have similar motifs, comprising the dual UAAU sequence, as has been previously shown (5). The half sites are separated with a variable gap spacer, including a gap of length 3. Though the most significant gap is not 3, the motif UAAU-N\(^3\)-UAAU is almost as enriched as the motifs shown in Figure 6A (PUF1: P-value \( \leq 3 \times 10^{-14} \); PUF2: P-value \( \leq 10^{-48} \)), and therefore this may not be an instance of significant variable gap motif.

The motifs found for PUF3 and PUF4 are fixed gap (UGAAAUA and UGUANAUUA, respectively), and they are consistent with the motifs found by Gerber et al. (2). As for PUF5, our results are less consistent.

<table>
<thead>
<tr>
<th>A RBP</th>
<th>Motif</th>
<th>p-value ≤</th>
<th>Corrected p-value ≤</th>
</tr>
</thead>
<tbody>
<tr>
<td>Puf1</td>
<td>UUU{1,3,4,9}UUUU</td>
<td>3.47 \times 10^{-16}</td>
<td>4.66 \times 10^{-8}</td>
</tr>
<tr>
<td></td>
<td>UU{3,6,8}UU</td>
<td>5.47 \times 10^{-16}</td>
<td>7.34 \times 10^{-8}</td>
</tr>
<tr>
<td>Puf2</td>
<td>UUU{0,2,3,4}UU</td>
<td>4.80 \times 10^{-51}</td>
<td>6.44 \times 10^{-43}</td>
</tr>
<tr>
<td>Puf3</td>
<td>UGUAAAUA</td>
<td>4.18 \times 10^{-110}</td>
<td>5.61 \times 10^{-102}</td>
</tr>
<tr>
<td>Puf4</td>
<td>UGUANAUUA</td>
<td>3.61 \times 10^{-47}</td>
<td>4.85 \times 10^{-39}</td>
</tr>
<tr>
<td>Puf5</td>
<td>UGUANAUUA</td>
<td>5.69 \times 10^{-29}</td>
<td>7.64 \times 10^{-21}</td>
</tr>
</tbody>
</table>

---

**Figure 6.** Motifs found to be enriched among high-affinity targets of Puf proteins. (A) The motifs found for PUF1, PUF2, PUF3, PUF4 and PUF5, together with their statistical significance, are shown. (B) A Shannon logo for the first half of the variable gap motif found for PUF3. (C) A Shannon logo for the second half of PUF3 motif (the logos are drawn using WebLogo).
with the literature and suggest UGUANUAUA as a potential-binding site for this RNA-binding protein.

An extension to DRIMUST, also available for download, enables refinement of the exact motif output obtained from DRIMUST. It takes as input an exact-word variable gap motif that was produced as output by DRIMUST, and returns two PWMs—one for the first half of the motif and one for its second half (see ‘Materials and Methods’ section). For example, applying this extension to the motif UGUAAUAU (which is equivalent to UGUA-N^0-AAUA) found for PUF3 yields PWMs that correspond to the Shannon logos presented in Figure 6B and C.

To further demonstrate the power of the mHG statistics, we shuffled the ranking of PUF3 target sequences and then conducted motif search. In other words, as control, the target sequences were sorted randomly, with no association to PUF3 binding. We repeated this control for 100 times, and out of these 100 control experiments the most significant result was not better than 5.06 × 10^{-11} (as opposed to 4.18 × 10^{-110} found for the original ranking).

Additionally, we applied DRIMUST to targets of PUM1, a human member of the Pumilio family [data was taken from (3)]. The targets were ranked according to PUM1-binding affinity and used as input to DRIMUST. We found that the motif UGUA-N^0.3-AAUA was the most significant (P-value ≤ 5.14 × 10^{-10}, Corrected P-value ≤ 0.07). It resembles the motif UGUA AAUA which was found for PUF3 targets, and as a matter of fact the latter string is highly enriched among PUM1 targets as well (P-value ≤ 3.5 × 10^{-9}), consistent with the validated binding site for PUM1 (3,4).

The analysis of all Puf proteins took 15 min on a standard PC.

Protein motifs

We finally applied DRIMUST to protein sequences, studying amino acid motifs enriched among human kinase substrates. Phosphorylation sites were taken from Phospho.ELM database (49), which stores serine, threonine and tyrosine phosphorylation data in a large set of species. We placed sequences containing phosphorylation sites of a specific type (serine, threonine or tyrosine) at the top of a ranked list, and appended to them a set of random sequences taken at random from the human proteome and having the same lengths. The complete list was used as DRIMUST input, seeking variable gapped motifs.

The most enriched motif found among serine kinase substrates was RSRS-N^0,2,3,5,6,8,9_RSRS (P-value ≤ 6.24 × 10^{-9}; Corrected P-value ≤ 0.25). To eliminate artifacts that may lead to the enrichment of this motif, we further tested whether there was RS di-residual density bias in the data set. We found that phosphorylated sequences tend to have greater RS di-residual density than non-phosphorylated sequences. Specifically, we sorted the sequences according to RS di-residual densities in descending order (greater RS density values are at the top and lower values are further down). We then observed that among the higher 5000 values (out of 7752), 3195 were serine phosphorylated sequences and only 1805 were non-phosphorylated sequences (mHG P-value ≤ 6.2 × 10^{-44} and Supplementary Figure S2). To test whether the motif is a result of this bias, we generated sequences containing the same amino acid content while preserving the same RS di-residual density (for details, see ‘Materials and Methods’ section), and then conducted motif search. We did not observe any motif at P-value ≤ 10^{-5}, and specifically did not find the above motif. Therefore, the motif composed of the dual sequence RSRS is not likely to result solely from the RS di-residual density bias. Furthermore, we also analyzed serine kinase substrates in M. musculus, and did not find any motif at P-value ≤ 10^{-5}. In mouse, we also found a significant RS di-residual density bias (mHG P-value ≤ 10^{-244}), supporting that RS di-residual density bias itself cannot explain the enrichment of the motif RSRS-gap-RSRS among serine phosphorylated sequences in humans. Moreover, searching variable gapped motifs comprising half-sites of lengths 6 and 4 (and vice versa) in humans, where we found the above motif, did not yield any result at P-value ≤ 10^{-5}. This means that the gap is not predominantly RS. We conclude that this motif is directly related to serine phosphorylation and not to RS di-residual density bias.

The motif is probably related to RS domains, which are rich in arginine-serine repeats and are known to be subjected to serine phosphorylation (57,58). For example, human SR proteins are known to contain RS domain at their carboxyl terminus that is involved in protein–protein interaction, and at least one amino-terminal RNA-binding domain that provides RNA-binding specificity (59). Since there are about nine known SR proteins in humans (59) and the motif above was found among 43 serine-phosphorylated proteins (out of 48 occurrences in the complete list), it is likely that not all of them are SR proteins. Therefore, we propose that the enrichment of the motif above points at a general mechanism rather than enrichment resulting from SR proteins influence on the analysis.

It is interesting to note that all motifs having the structure RSRS-N^0_RSRS, where 0 ≤ l ≤ 10, were not enriched at a threshold of 10^{-4} (the default in our software), and therefore taking a greedy approach that filters half-sites by testing only fixed gap motifs would have missed this variable gap motif (Figure 7A).

As for threonine phosphorylated sequences, we did not observe any motif at P-value ≤ 10^{-5}. On the other hand, we found several interesting motifs around tyrosine phosphorylation sites. A non-gapped motif, HRDLAARN, was significantly enriched in humans (P-value ≤ 3.17 × 10^{-13}; Corrected P-value ≤ 2.55 × 10^{-9}) and in M. musculus (P-value ≤ 1.5 × 10^{-9}). Further search for 4-mers enriched around tyrosine phosphorylation sites yielded three highly significant motifs: DFGL (P-value ≤ 5.1 × 10^{-17}), HRDL (P-value ≤ 7.7 × 10^{-17}) and SDVW (P-value ≤ 2.5 × 10^{-16}). Some of these motifs are known to be related to tyrosine kinase susceptibility. For example, the catalytic domains of protein-tyrosine kinases, such as EGFR, FGFR3, CSK, MATK (also
known as CHK) and others contain the consensus sequence HRDLAARN. Interestingly, this sequence is also followed by the conserved sequence DFG (34–36) which is important for ATP coordination (35).

We tested the combination of the three motifs HRDLAARN, DFGL and SDVW that we found, and identified a strikingly evident pattern. Out of 41 phosphorylated sequences that contained HRDLAARN, 32 contained both DFGL and SDVW. Moreover, in all these 32 cases, DFGL appeared exactly 12 amino acids after HRDLAARN, and SDVW appeared between 33–39 amino acids after DFGL (Figure 7B). This motif was significantly enriched in humans (P-value ≤ 1.9 × 10^{-10}) and also in M. musculus (P-value ≤ 1.1 × 10^{-5}). This, to the best of our knowledge, is a novel discovery.

The above analysis took 4 min on a standard PC.

**DISCUSSION**

In this article, we describe an efficient statistical and algorithmic approach to searching motifs in ranked list of sequences. Our method, implemented in DRIMUST, has several advantages over existing methods. First, many other approaches exhaustively search over motif spaces and therefore, cannot handle gapped motifs and motifs over large alphabets. Furthermore, we search motifs in ranked lists and not in fixed sets of sequences as is the case for many other methods. To the best of our knowledge, DRIMUST is unique in combining an efficient search with a ranked list approach and rigorous P-value estimation. It is also unique in efficiently addressing variable gap motifs under a definition that allows full flexibility of the gap.

Our approach to overcoming the computational challenges associated with large motif search spaces is based on using suffix trees, to restrict our attention to motifs that actually occur in the input list of sequences. An alternative approach could be based on the use of a hash table. Taking this approach, however, will entail running the search for a fixed motif length k. To span a range of lengths $k_1 \leq k \leq k_2$, as is done by DRIMUST, one would need to generate a separate hash table for each k. Therefore, a suffix tree approach is far more efficient in this case. In addition, the hash table approach falls short of solving the variable gap motif problem which is enabled through the use of suffix trees in DRIMUST.

One of the standard motif search tools used by the community is MEME (15). As opposed to DRIMUST focus on discovering motifs in ranked lists, MEME uses a fixed set approach. Moreover, it limits the input to only 60,000 characters. When we ran the phosphorylation motif search analysis described in the ‘Results’ section on MEME’s service (60), as the full input would have been too large (the tyrosine data set is 200,000 characters), we used 300 sequences as input. MEME took more than an hour to run the analysis. The results are depicted in Supplementary Figure S3 and are consistent with DRIMUST predictions, however, we note that the variable gap nature of the motifs we found to be associated with HRDLAARN is not discovered by MEME. Additionally, we ran MEME on ~200 sequences of the serine data set, yielding no significant findings. Another advantage of DRIMUST over MEME is that it provides a tight upper bound on the P-value, based on the mHG model, while MEME’s output includes an e-value which is not as rigorously associated with a statistical model.
In 2007, our group developed the first mHG approach to motifs in ranked lists. This was implemented as the DRIM algorithm (20). Building on the DRIM statistical approach, the current work represents significant progress over DRIM:

(1) Improved computational efficiency enabled by using suffix trees
(2) Thanks to this improvement, we can address large alphabets, long motifs and, most notably, a new class of variable gap motifs (VGMs)

One apparent limitation of the current implementation of DRIMUST is that the output half-sites are exact motifs. That is, they are words over the alphabet of the input sequences, allowing no flexibility or weighting. Our output can be viewed as the starting point for extending the motifs using more flexible motif representation approaches for selected top results. In practice, for this extension step, we currently employ ad hoc and manual analysis. In this article, we used a simple extension mechanism of a motif in the DRIMUST output by considering its variants that occur at the top of the ranked list for creating a PWM for the motif. The implementation of this extension is also available for download. In the future, we intend to further develop algorithmic approaches to this phase. For example, extend motifs using IUPAC or PWMs. The IUPAC extension task is amenable to modifications of the suffix tree approach.

We note that all P-values reported by DRIMUST are not corrected for the size of the motif search space, which should be handled by the user. Findings presented in this article remain significant after correction.

In addition to the computational contribution of this article, we present analysis results of biological importance:

(1) We identify CATG as the strongest TP53 half-site. While the pair AT was shown to be the best instance of WW in the known consensus CWWG, this was done in controlled synthetic experiments. Our results which validate this recently established preference were obtained using high-throughput measurement data. Consistent with literature, we found a spacer of 6 bp to yield significant results.

(2) We observe GTCA-N^{3,6,9}-TGAC as the strongest binding site for ESR1. The existing literature consensus, to the best of our knowledge, is GTCA-N^3-TGAC. This newly hypothesized refinement of the ESR1 binding site, inferred from high-throughput measurement data, demonstrates the utility of variable gap motif search. The data analyzed to yield this finding consisted over 10000 human DNA sequences, each of length 600 bp—a size which is difficult to handle by most state of the art tools.

(3) In analyzing Harbison et al. data, we validate some existing findings and also suggest some novel motifs. Specifically, we find CGCG-N^{1,3,5,8,9,10}-CGCG to be the most significant variable gap motif for SWI6. Indeed, this motif is much stronger than all the underlying fixed gap motifs (Figure 5C).

(4) We suggest PUF4 involvement in S. cerevisiae heat-shock response through motif search analysis in 3'UTRs.

(5) We suggest HRDLAARN-X^{12}-DFGL-X^{33-39}-SDVW as a significant motif related to potential tyrosine phosphorylation in humans (and in mice). HRDLAARN, the first part of this motif, is known to be related to tyrosine phosphorylation. Our analysis gives rise to a significant refinement.

This work is the first to perform a systematic analysis of variable gapped motifs in biological data sets. We have found variable gapped motifs to be more significant in a handful of cases and therefore reach the preliminary conclusion, consistent with Reid et al. (33), that they are important but not extremely widespread.

In summary, we present efficient and effective algorithms for motif search and demonstrate its utility in biological data sets. In the future, we hope to further expand findings driven by this approach, through usage by us as well as by others in the community.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online: Supplementary Tables 1–2 and Supplementary Figures 1–3.

ACKNOWLEDGEMENTS
We thank Israel Steinfeld for useful discussions and for support in implementing mHG calculations and other code. We also thank Yael Mandel-Gutfreund, Shlomo Moran, Tali Haran and Eran Eden for useful discussions. Finally, we thank Roy Navon for technical help with the software download service.

FUNDING
Jacobs Fellowship and by ISEF Fellowship (L.L., in part). Funding for open access charge: Laboratory of Computational Biology, Department of Computer Science, Technion—Israel Institute of Technology.

Conflict of interest statement. None declared.

REFERENCES


3.3 DRIMust: a web server for discovering rank imbalanced motifs using suffix trees

The study of sequence elements that enable molecular recognition in a variety of cellular processes is an important component in improving our understanding of regulation of molecular recognition processes make use of statistical and computational tools that support the identification and understanding of sequence motifs. We present a new web application, named DRIMust, freely accessible through the website http://drimust.technion.ac.il for de novo motif discovery services. The DRIMust algorithm is based on the minimum hypergeometric statistical framework and uses suffix trees for an efficient enumeration of motif candidates. DRIMust takes as input ranked lists of sequences in FASTA format and returns motifs that are over-represented at the top of the list, where the determination of the threshold that defines top is data driven. The resulting motifs are presented individually with an accurate P-value indication and as a Position Specific Scoring Matrix. Comparing DRIMust with other state-of-the-art tools demonstrated significant advantage to DRIMust, both in result accuracy and in short running times. Overall, DRIMust is unique in combining efficient search on large ranked lists with rigorous P-value assessment for the detected motifs.

INTRODUCTION

The study of sequence elements that enable molecular recognition in a variety of cellular processes is an important component in improving our understanding of regulation in living cells. Transcription factor (TF) activity, for example, often depends on the identification of specific targets using molecular pattern recognition mechanisms that involve sequence motifs. Sequence recognition plays a role in other molecular levels, as well. The occurrence of short binding motifs in RNA molecules plays a central role in enabling controlled regulation by RNA-binding proteins (RBPs) and by microRNAs. For example, the Pumilio family proteins regulate target messenger RNAs by recognizing and binding sequence elements in 3' untranslated regions (UTRs) (1). Protein modification and protein–protein interactions are also potentially driven by mechanisms that involve specific protein sequence recognition such as the phosphate-binding loop (2,3).

Studies using techniques such as ChIP-chip (4), ChIP-PET (5), ChIP-seq (6) and ChIP-exo (7) lead to genome-wide measurement data pertaining to the TF binding affinity of various genomic regions, obtained in actual samples and in several conditions. Similarly, messenger RNA targets of RBPs are studied using techniques like RNA immunoprecipitation (RIP)-chip (8), crosslinking and immunoprecipitation (CLIP) (9) and photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) (10). Stable isotope labeling by/with amino acids in cell culture (SILAC) (11) and other proteomic techniques can be used to characterize the effect of amino acid sequences on protein function. Computational tools and approaches to motif discovery form part of the data analysis workflow that is used to extract knowledge and understanding from this type of studies. Motif discovery has attracted much research interest in recent years, resulting in more than a hundred different tools for motif discovery (12,13). A large subset of motif finders such as Multiple EM for Motif Elicitation (MEME) (14), NMica (15), AlignACE (16), MDscan (17), Yeast Motif Finder (YMF) (18), Gapless Local Alignment of Multiple sequences (19) and Suite for Computational Identification Of Promoter Elements (SCOPE) (20) fit

*To whom correspondence should be addressed. Tel: +972 4 8293958; Fax: +972 4 8225153; Email: yaelmg@tx.technion.ac.il

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

© The Author(s) 2013. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
position weight matrices to the sequence data. Recently, efficient motif discovery tools were designed to handle large sets of data arising from the aforementioned high-throughput measurement techniques, as for example MEME-ChIP (21), Discriminative Regular Expression Motif Elicitation (DREME) (22) and XXmotif (23). Most techniques are designed to find motifs by seeking elements that occur more often than expected in a set of sequences. Many of these techniques compare a target set with a background set, such as in XXmotif (23). It is often the case, however, in biological measurement data, that results are given as a ranked list of quantities. Such is the case for data generated by ChIP-seq or CLIP, for example. Statistical approaches such as Gene Set Enrichment Analysis (24) and minimum hypergeometric (mHG) (25–28) address motif enrichment in ranked lists of elements. We have previously developed DRIM (25), a motif-finding approach that exploits the ranking derived from experimental measurements to discover k-mers that are rank imbalanced in the input list based on the mHG statistics. The search for rank imbalanced motifs allows for much more flexibility and is therefore more compatible with the character of the actual measurement results. The mHG model allows for a rigorous statistical assessment of the results without the need to run simulations. To overcome the computational challenges associated with large motif searches, we have recently developed the DRIMust algorithm (29). The algorithm is based on suffix trees, an approach also suggested in (30–32). DRIMust allows for an efficient enumeration of motif candidates, which are then assessed using the mHG statistics. Tree-based approaches have been previously efficiently used in other motif search algorithms, such as the beam search algorithm, which is an enumerative algorithm for identifying enriched cis-regulatory elements in sets of commonly regulated genes (33).

In this work, we introduce the DRIMust web server and describe its utility in supporting the search of rank imbalanced motifs. DRIMust takes as input ranked lists of sequences in FASTA format and returns motifs that are observed as high as possible over all possible partitions. DRIMust allows the user to upload positive and negative sets of sequences. In the latter case, DRIMust will search for enriched motifs in the positive set using the negative set as the background. DRIMust is efficient and thus allows searching in large data sets, searching for long motifs as well as searching motifs over large alphabets in short running times. The resulting motifs are presented as a Position Specific Scoring Matrix (PSSM) in a graphical WebLogo format; the matrix can also be downloaded as a text file. For every motif, a P-value is indicated. DRIMust is freely accessible through the website http://drimust.technion.ac.il/.

DRIMUST METHODOLOGY

The DRIMust approach seeks rank imbalanced motifs, given a ranked list of sequences $S_1, ..., S_X$. Rank imbalanced motifs are substrings that appear more often at the top of the list compared with the remainder of the list. Eden et al. (25) described the mHG statistics used for the assessment of rank imbalanced motifs. A unique feature of the mHG statistics is that the cutoff between the top and the rest of the list is determined in a data-driven manner so as to maximize motif enrichment. This is done by computing the motif enrichment over all possible set partitions and identifying the cutoff at which maximal statistical significance is observed. The algorithmic approach of DRIMust is based on suffix trees, allowing efficient enumeration of the motif search space (29).

Enrichment analysis using mHG statistics

We have previously described an algorithm to identify the enrichment of a set of genes, $A$, in a ranked list of genes, using mHG statistics (25). Given a total number of genes $N$, with $B$ of these genes belonging to $A$, and $n$ of these genes being in the target set (e.g. differentially expressed genes), the probability that $b$ or more genes from the target set are also in $A$ is given by the tail of a hypergeometric random variable (HGT):

$$P(X \geq b) = \text{HGT}(N, B, n, b) = \sum_{i=b}^{\min(n, B)} \binom{n}{i} \frac{\binom{N-n}{B-i}}{\binom{N}{B}}$$

For a ranked genes list $g_1, ..., g_N$, we define a label vector $\lambda = \lambda_1, ..., \lambda_N \in \{0, 1\}^N$ according to the association of the ranked genes to $A$, that is, $\lambda_i = 1$ if and only if $g_i \in A$. The mHG score is then defined as

$$mHG(\lambda) = \min_{1 \leq s \leq N} \text{HGT}(N, B, n, b_h(\lambda)), \text{ where } b_h(\lambda) = \sum_{i=1}^{n} \lambda_i.$$ 

In other words, the mHG score is the optimal HGT probability that is found over all possible partitions induced by the ranking. As such, this score must be corrected for multiple testing. A dynamic programming algorithm for computing the exact $P$-value of a given mHG score is described in (25). More specifically, given a ranked list of genes, a subset $A$, and a corresponding mHG score $s$, the mHG P-value tells us the exact probability of observing an mHG score $s' \leq s$ under the null assumption that all occurrence configurations of $A$ in the ranked list are equiprobable. In practice, DRIMust uses Stirling's approximation (34) to compute all binomial coefficients needed to assess HGTs.

Suffix trees

A suffix tree is a data structure that represents all the suffixes of a given string in a way that allows fast implementation of many string operations. A path from the root to a leaf in the tree represents a suffix. Each leaf of the tree holds information about the indices of strings that contain the suffix, and the starting positions of this suffix within each such string. Restoring all occurrences of a suffix is thus enabled, which further allows for the detection of DNA, RNA or protein substrings that manifest a significant occurrence pattern in a set of biologically related sequences. There are several algorithmic
approaches to the efficient construction of a suffix tree for a collection $S_1, \ldots, S_N$ of strings (35–37). DRIMust uses a version that takes $O(M)$ time, where $M = \sum_{i=1}^{k} \text{length}(S_i)$ by implementing Ukkonen’s algorithm for generalized suffix trees construction (37).

In DRIMust, an initial motif search phase produces k-mers, which are words over the alphabet of the input sequences. These candidate k-mers are derived by enumerating paths of length $k$ in the generalized suffix tree generated for the input sequences. Next, the statistical significance of the k-mers is calculated using the mHG statistics [for more details on how $P$-values are computed in the nodes of the suffix tree, see (29)]. In the next stage, the promising k-mers are extended to produce PSSMs.

**PSSM extension**
The promising k-mers are passed as input to a process that extends them to PSSMs. Extension is obtained by a heuristic approach based on the Hamming neighbors of the best 50 exact motifs. Briefly, starting from a single k-mer, Hamming neighbors (of length $k$) are added to a set of motifs as long as the PSSM representing that set improves the observed enrichment $P$-value.

**DRIMUST DESCRIPTION**

**Input**

DRIMust is designed to search for enrichment of motifs in large datasets of DNA, RNA or protein sequences (up to 40,000 sequences and up to 4,000,000 characters), which can be represented as ranked lists or as two separated sets of targets and background. Ranking should be provided by the user according to the research question of interest, e.g. binding affinity for ChIP-seq data. In the case of uploading target and background sets, the latter can be a selected random set of sequences taken from the genome. When uploading the input data, the user is prompted to choose between submitting one ranked list of sequences in FASTA format or two sets of target and background. The DRIMust motif-searching process is divided into two phases. In the first phase, DRIMust searches for k-mers, which are over-represented at the top of the input ranked list of sequences. As a default, DRIMust will report enriched k-mers having $P$-value better than the selected stringency. In the second phase, DRIMust expands the most promising k-mers heuristically and creates motifs represented by PSSMs. An average job for data sets containing 4000 DNA sequences, total 2,000,000 characters, takes 1 min and 10 s when double-strand search mode is used and 15 s when single-strand search mode is used.

We ran DRIMust on the HOXA2-binding regions from the ChIP-seq experiment by Donaldson et al. (38). In this data set, the DNA sequences were ranked according to their binding $P$-values [as defined by (38)]. As demonstrated in Figure 1A, the best motifs are presented in the output page as PSSMs displayed in a graphical WebLogo representation (39) and also provided as a downloadable text files. The $P$-value of each motif is indicated above the logo. Furthermore, the user is provided with a detailed list of all significant k-mers that DRIMust has found to be enriched at significance level better than the threshold (Figure 1B). In addition, each row includes information about the total number of input sequences ($N$); the total number of sequences containing the motif ($B$); the index that is selected by the mHG statistics as the division of the input list into target and background ($n$)—which optimizes the enrichment of the motif at the top $n$ sequences of the list; and the number of sequences containing the motif amongst the top $n$ sequences ($b$). Finally, the enrichment value, which compares the abundance of the motif at the top of the list to the abundance at the entire list, defined by $(b/n)/(B/N)$, is indicated.

**Output**

When a range of lengths is provided, DRIMust will search for the most significant motif, which will not necessarily be the longest one. When a certain motif length is expected, the user is encouraged to define an exact length. Next, the user can choose to change the default statistical significance threshold (set to $10^{-6}$) to any threshold between $10^{-2}$ and $10^{-15}$. Finally, although not required, DRIMust supports including an e-mail address to which the results will be automatically sent when the analysis is completed. This option is useful when sending long jobs. After uploading the input data and defining the parameters, the users are prompted to submit their job.
Figure 1. A view of DRIMust input and output pages. We ran DRIMust on the HOXA2-binding regions from the ChIP-seq experiment by Donaldson et al. (38). In this data set, the DNA sequences were ranked according to their binding $P$-values (as defined by Donaldson et al.). DRIMust was run using the double-strand search mode, and the rest of the parameters were set to default. The full data set is provided as an example in the manual page of DRIMust web server. (A) When clicking the submit button (bottom left), an output page, summarizing the best motifs found, is shown to the user. (B) By clicking the ‘view list’ button, the user is provided with a list of the significant k-mers and the statistical details of each motif. (C) By clicking the ‘view occurrences alignment’ button, the user is provided with an aligned list of motif occurrences mapped into the input sequences. (D) By clicking the ‘view occurrences distribution’ button, a window depicting the occurrences of the motif in the query sequences is opened. More details on each occurrence are shown when placing the cursor on the occurrence box.
RESULTS AND DISCUSSION
In recent years, high-throughput binding techniques have been developed [e.g. ChIP-seq (4) for protein–DNA, PAR-CLIP (10) for protein–RNA and SILAC (11) for protein–proteins]. These methods yield extensive lists of potential targets, ranked according to their binding affinity. The main advantage of our method, implemented in DRIMust, is that it searches for enriched motifs in the entire ranked list and does not require defining a fixed set of sequences as in the case of other motif-search algorithms such as MEME (14), PhyloGibbs (40) and others. Nevertheless, DRIMust does provide the option of uploading a target and background sets predefined by the user. In the latter case, DRIMust searches motifs that are overrepresented at the target set compared with the background set. To evaluate the performance of DRIMust in comparison with other state-of-the-art methods, we ran DRIMust on 24 examples generated from high-throughput binding experiments—10 TFs and 14 RBPs—and compared the results with those obtained by using four other methods: the standard MEME program (14); the DREME program (22) from the MEME suite (http://meme.nbcr.net), which was optimized for fast analysis of large data sets; XXmotif (23), a recent web server, which was designed for efficient extraction of position weight matrices from large datasets; and SCOPE (20), which was designed to identify candidate regulatory DNA motifs from sets of genes that are coordinately regulated. Almost all the input examples comprised ranked lists, except for p53, which comprised target and background sets. As MEME, DREME and XXmotif expect a target set as input, we converted the ranked lists into target sets by taking the top 100 sequences in the case of MEME (restricted by MEME’s limitation of 60,000 characters) and the top 20% sequences for the other tools. The results of the comparison are summarized in Supplementary Table S1. As demonstrated, in 22 of the 24 test examples, DRIMust found the motifs that were compatible with the known motifs as the most significant result. In comparison, DREME found the known consensus in 19 cases, XXmotif detected the literature motif in 16 cases, whereas MEME and SCOPE detected the known motif in only half of the cases. Notably, in the other methods tested, the known motifs were not always reported as the best motif. Strikingly, while DRIMust was tested on the largest data sets, in all cases, DRIMust completed the computations faster than the other tools. As demonstrated in Supplementary Table S1, the longest job took 1 min and 21 s on DRIMust (for a data set containing 9995 sequences, each of length 100 nucleotides).

Overall, the web-application DRIMust has several advantages over existing methods. First, unlike many other approaches, it does not exhaustively search over all possible k-mers space and therefore can detect long motifs and motifs over large alphabets. DRIMust runs efficiently and allows for timely interaction with the results, through a friendly interface and a clear output format. Most importantly, by working with ranked lists, DRIMust avoids the arbitrary designation of fixed sets of sequences and exploits the ranking derived from experimental measurements. More than that, DRIMust uses the ranking to discriminate true motifs from other irrelevant sequence elements (such as AT repetitive elements that are abundant in 3’UTRs), as the latter are not correlated with the ranking and are therefore ignored by DRIMust. This explains the observed accuracy of DRIMust compared with other tools in many of the examples shown in Supplementary Table S1.

As biological techniques such as ChIP-seq (6), ChIP-exo (7), CLIP (9), PAR-CLIP (10) and others produce ranked lists, using DRIMust is the natural choice for motif discovery in these cases, as arises from the comparison in Supplementary Table S1. DRIMust can efficiently deal with the large data sets generated by such methods, making it preferable for large volume data. Nevertheless, DRIMust is also useful in cases when there are clear target and background sets. In the latter scenario, the enrichment is calculated using the hyper-geometric distribution.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR Online: Supplementary Table 1 and Supplementary References [4,10,38,41–44].

ACKNOWLEDGEMENTS
The authors thank DRIMust’s many users for their useful comments and suggestions for improving the website. Specifically, they thank Refael Cohen and Iris Dror for extensive testing and helpful remarks. They also thank Israel Steinfeld, Roy Navon and Ilona Kifer for useful comments. L.L. and Z.Y. developed the algorithm and the statistical methodology. L.L. implemented the method; I.P., L.L., Y.M.G. and Z.Y. designed the website and defined needs. I.P. developed the website. Y.M.G. planned the project and the validation experiments. All authors contributed to writing the article.

FUNDING
Niedersachsen-Israel Grant (to Y.M.G.) and Israel Ministry of Science and Technology and by ISEF Fellowship (to L.L., in part). Funding for open access charge: Internal Grants.

Conflict of interest statement. None declared.

REFERENCES


9. Lebedeva, S., Jens, M., Theil, K., Schwanhäusser, B., Selbach, M.,


38. Donaldson, I.J., Amin, S., Hensman, J.J., Kutejova, E., Rattray, M.,


34. Abramowitz, M. and Stegun, I.A. (1964) 


23. 249–260.


3.4 Mutual Enrichment in Ranked Lists and the Statistical Assessment of Position Weight Matrix Motifs

Mutual Enrichment in Ranked Lists and the Statistical Assessment of Position Weight Matrix Motifs

Limor Leibovich\(^1\) and Zohar Yakhini\(^{1,2}\)

\(^1\) Department of Computer Science, Technion – Israel Institute of Technology, Technion City, Haifa 32000, Israel
llimor@cs.technion.ac.il

\(^2\) Agilent Laboratories Israel, 94 Em Hamoshavot Road, 49527 Petach-Tikva, Israel
zohar_yakhini@agilent.com

Abstract. Statistics in ranked lists is important in analyzing molecular biology measurement data, such as ChIP-seq, which yields ranked lists of genomic sequences. State of the art methods study fixed motifs in ranked lists. More flexible models such as position weight matrix (PWM) motifs are not addressed in this context. To assess the enrichment of a PWM motif in a ranked list we use a PWM induced second ranking on the same set of elements. Possible orders of one ranked list relative to the other are modeled by permutations. Due to sample space complexity, it is difficult to characterize tail distributions in the group of permutations. In this paper we develop tight upper bounds on tail distributions of the size of the intersection of the top of two uniformly and independently drawn permutations and demonstrate advantages of this approach using our software implementation, mmHG-Finder, to study PWMs in several datasets.

1 Introduction

Modern data analysis often faces the task of extracting characteristic features from sets of elements singled out according to some measurement. In molecular biology, for example, an experiment may lead to measurement results pertaining to genes and then questions are asked about the properties of genes for which these were high or low. This is an example, of course, and the set of elements does not have to be genes. They can be genomic regions, proteins, structures, etc. A central technique for addressing the analysis of characteristic properties of sets of elements is statistical enrichment. More specifically – the experiment results are often representable as ranked lists of elements and we then seek enrichment of other properties of these elements at the top or bottom of the ranked list. GSEA [29], for example, is a tool that addresses characteristic features of genes that are found to be differentially expressed in a comparative transcriptomics study. GOrilla [6] addresses GO terms enriched in ranked lists of genes where the ranking can be, for example, the result of processing differential expression
data or of correlations computed between genomic DNA copy number and expression \[19,2,5\]. FATIGO \[21\] is also a tool that is useful in the context of analyzing GO terms in ranked lists of genes. DRIMust \[15,16\] searches for sequence motifs that are enriched, in a statistically significant manner, in the top of a ranked list of sequences, such as one produced by techniques like ChIP-seq.

All the aforementioned tools utilize a statistical approach that is based on assessing enrichment of an input set in an input ranked list by assessing the enrichment obtained at various cutoffs applied to the ranked list. It is often the case, however, that two quantitative properties need to be compared to each other. For example, when the elements are genes, we may have measured differential expression values, as well as measured ChIP-seq signals. We are therefore interested in assessing mutual enrichment in two ranked lists. Another example consists of one ranking according to differential expression and one according to prediction scores for miRNA targets. miTEA \[25\] addresses this latter case by statistically assessing the enrichment of miRNA targets in a ranked list of genes (also see \[8\]). To address mutual enrichment in two ranked lists over the same set of \(N\) elements, miTEA \[25\] performs analysis on permutations. Mutual enrichment in the top of two ranked lists can be simplified, from a mathematical point of view, by arbitrarily setting the indices of one list to the identity permutation \((1, 2, \ldots, N)\) and treating the other list as a permutation \(\pi\) over these numbers. For the purpose of assessing the intersection of the top of the two ranked lists in a data driven manner, miTEA asks which prefix \([1, \ldots, n_1]\) is enriched in the first \(n_2\) elements of that permutation, \(\pi = \pi(1), \ldots, \pi(N)\). The statistics introduced by miTEA is called mmHG (min-min-Hyper-Geometric). A variant of mmHG is explained in detail in Section 2 of the current manuscript.

Statistics in the group of permutations \(S_N\) is often difficult because the number of entities to be considered by any null model is \(N!\). Direct exhaustive calculation of tail distributions over \(S_N\) is therefore impractical for all but very small values of \(N\). This difficulty is addressed by several heuristic techniques. Mapping into continuous spaces, such as in \[18\], has proven useful in certain cases but not for studying large deviations. In the case of enrichment statistics that focuses on the top of the permutation and seeks to assess extreme events, such as mmHG, we prefer to use bounds on tail probabilities. Tail probabilities are useful constructs when applied to analyzing molecular biology measurement data as they enable statistical assessment of observed results.

In this work we derive a tight bound on the tail probabilities of the mutual enrichment at the top of two random permutations uniformly drawn over \(S_N\) and demonstrate the utility of this approach in the context of flexible motif discovery. Our bounds are computable in polynomial time and potentially add to the accuracy of reported position weight matrix (PWM) motifs for nucleic acid sequences.
2 Background and Definitions

2.1 Mutual Enrichment in Ranked Lists – The mmHG Statistics

The mmHG statistics [25] is a generalization of the mHG statistics [6],[7],[26],[28]. While the mHG statistics quantifies the enrichment level of a set of elements at the top of a ranked list of elements of the same type, the mmHG statistics quantifies the level of mutual enrichment in two ranked lists over the same set of elements. While any parametric or non-parametric correlation statistics (e.g. Spearman’s correlation coefficient), that takes the same input, calculates the overall agreement between the two ranked lists, the mmHG statistic focuses only on agreement at the top of the two ranked lists. mmHG counts elements common to the top of both lists, without predefining what top is. Its intended output is the probability for observing an intersection at least as large in two randomly ranked lists (the enrichment mmHG \( P \)-value). In this section we describe the mmHG statistics and in later sections we suggest a tight bound for the \( p \)-value. Our definition of the mmHG statistic varies slightly from that of Steinfeld et al. [25].

Mutual enrichment in the top of two ranked lists can be simplified, from a mathematical point of view, by arbitrarily setting the indices of one list to the identity permutation \((1, 2, ..., N)\) and treating the other list as a permutation. Details of this transform are given in Section 2.3. We now define mmHG for the simple case of one permutation. Consider a permutation \(\pi = \pi(1), ..., \pi(N) \in S_N\) - the group of all permutations over the numbers \(1, ..., N\). mmHG is a function that takes \(\pi\) and calculates two numbers \(1 \leq n_1, n_2 \leq N\) such that the observed intersection between the numbers \(1, ..., n_1\) with \(\pi(1), ..., \pi(n_2)\) - is the most surprising in terms of the hypergeometric \(p\)-value. Additionally, mmHG further calculates this aforementioned \(p\)-value.

Formally, given \(\pi \in S_N\) and for every \(1 \leq n_1, n_2 \leq N\), let \(b_\pi(n_1, n_2)\) be the size of the intersection of \(1, ..., n_1\) with \(\pi(1), ..., \pi(n_2)\). Set

\[
\text{mmHG score}(\pi) = \min_{1 \leq n_1 \leq N} \min_{1 \leq n_2 \leq N} \text{HGT}(N, n_1, n_2, b_\pi(n_1, n_2))
\]

where \(\text{HGT}\) is the tail distribution of an hypergeometric random variable:

\[
\text{HGT}(N, n_1, n_2, b) = \sum_{i=b}^{\min(n_1, n_2)} \frac{(n_1)_{i}(N-n_1)}{(n_2)_i (N-n_2)}
\]

The mmHG score cannot be considered as a significance measure, due to the multiple testing involved in finding \(n_1\) and \(n_2\). A simple way to correct an mmHG score \(s\) for multiple testing and report a \(p\)-value bound would be to use the Bonferroni correction. That is done by multiplying \(s\) by the number of multiple tests conducted which is \(N^2\). Therefore:

\[
\text{mmHG p-value}(s, N) \leq s \cdot N^2
\]

In Section 3 we present significantly tighter bounds.
2.2 PWM Motifs

Data produced by techniques such as ChIP-seq [14], ChIP-exo [20], CLIP [13], PAR-CLIP [9] and others are readily representable as ranked lists of sequences, where the ranking is according to measured binding affinity. Computational tools and approaches to motif discovery form part of the data analysis workflow that is used to extract knowledge and understanding from this type of studies. We are often interested in sequence motifs that are observed to be enriched in sequences where strong binding affinity is measured. A position weight matrix (PWM) is a commonly used representation of motifs in biological sequences [24],[27],[11]. This representation is more faithful to biology than representation by exact words. A PWM is a matrix of score values that gives a weighted match to any given substring of fixed length. It has one row for each symbol of the alphabet, and one column for each position in the pattern. The score assigned by a PWM to a substring $S = S_1...S_K$ is defined as $\sum_{j=1}^{K} m_{s_j,j}$, where $j$ represents position in the substring, $S_j$ is the symbol at position $j$ in the substring, and $m_{\alpha,j}$ is the score in row $\alpha$, column $j$ of the matrix. In other words, a PWM score is the sum of position-specific scores for each symbol in the substring. This definition can be generalized to yield a score for a sequence $S = S_1...S_M$ longer than the PWM by calculating $\max_{1 \leq i \leq M-K+1} \sum_{j=1}^{K} m_{s_{i+j-1},j}$. Alternatively, an enhanced model that takes into account multiple occurrences of the PWM in the sequence can be applied by summing over sufficiently strong occurrences of the PWM in the sequence can be applied by summing over sufficiently strong occurrences of the PWM or by other more sophisticated approaches [22].

2.3 mmHG Statistics for PWM Motifs

Given a set of sequences that were tested in a high throughput experiment such as ChIP-seq [14], CLIP [13] and others, they can be ranked according to the measured binding affinities, yielding a ranked list $L_1$. Since usually we are interested in finding motifs amongst sequences having strong binding affinities, we actually search for motifs that are more prevalent at the top of this list. It is clear that any algorithm for de-novo flexible motif search would need to evaluate candidate PWMs. Given a PWM which we want to assess, the sequences can also be ranked according to their PWM scores, yielding another ranked list $L_2$, different from $L_1$. A significant PWM motif would yield significant scores for sequences having strong binding affinities. Therefore, the question of PWM motif discovery from ranked experimental data can be formulated as quantifying the mutual enrichment level for the two ranked lists $L_1$ and $L_2$. Given two ranked lists $L_1$ and $L_2$ over the universe of $N$ sequences, they can be transformed into two respective permutations, $\pi_1 = (\pi_1(1), ..., \pi_1(N))$ and $\pi_2 = (\pi_2(1), ..., \pi_2(N))$. The relative permutation $\pi$, of $\pi_2$ w.r.t. $\pi_1$, is defined by $\pi(\pi_1(j)) = \pi_2(j)$, for every $j = 1, ..., N$ or simply, using the operations in the group $S_N : \pi = \pi_2 \cdot \pi_1^{-1}$. Using the relative permutation $\pi$, we can represent the mutual enrichment of the top parts of $L_1$ and $L_2$ as mmHG score($\pi$), defined above.
3 Algorithms and Results

3.1 Estimation of the mmHG p-Value – Introducing First Upper Bound

Given an mmHG score \( s \), observed in analyzing real measurement data, we would like to assess the statistical significance of this observation. Assuming endless computational power, we would enumerate all permutations and calculate the mmHG score for each, in order to characterize the distribution of mmHG as a random variable over \( S_N \). The p-value for \( s \) is then simply:

\[
\text{mmHG p-value}(s, N) = \frac{\text{The number of permutations having mmHG score } \leq s}{N!}
\]

Since the number of permutations is huge, the process described above is very far from feasible. Therefore, we seek a computationally tractable upper bound, preferably tight.

A trivial upper bound is the Bonferroni corrected mmHG score defined by \( s \cdot N^2 \). A more subtle upper bound was suggested by Steinfeld et al. [25] and is briefly described in Section 3.3. In this work we introduce a tighter bound that is polynomially computable.

We will next describe an intuitive upper bound and later refine it to produce a tighter bound. Our input comprises an mmHG score \( s \), and the total number of elements \( N \). The output will be an upper bound for the p-value. The efficiency of our approach relies on enumerating all possible HGT scores rather than enumerating all permutations in \( S_N \). This approach is computationally efficient as HGT is a function of four input parameters: \( N, n_1, n_2, \) and \( b \). Given \( N \), there are \( O(N^3) \) possible combinations of \( n_1, n_2 \) and \( b \). Next, all is left to do is to determine how many permutations stand behind each HGT score. To this end, we will define the function \( \Lambda(N, n_1, n_2, b) \) to be the number of permutations for which it holds that out of the first \( n_2 \) entries, \( b \) of them are taken from the range \([1, \ldots, n_1]\). This formulation is equivalent to counting permutations for which we attain, at some point, the value \( \text{HGT}(N, n_1, n_2, b) \), had we taken the exhaustive approach. \( \Lambda(N, n_1, n_2, b) \) can be represented as:

\[
\Lambda(N, n_1, n_2, b) = \binom{n_1}{b} \binom{n_2}{b} b!(\binom{N-n_1}{n_2-b}(n_2-b)!(N-n_2)!
\]

as we first choose \( b \) elements from the range \([1, \ldots, n_1]\) to appear at the first \( n_2 \) entries of the permutation (there are \( \binom{n_1}{b} \) possibilities). Then, we choose where to position these \( b \) elements at the first \( n_2 \) entries of the permutation and consider all internal arrangements (for each choice of \( b \) elements there are \( \binom{n_2}{b}! \) possibilities). We next choose \( n_2 - b \) elements from the range \([n_1 + 1, \ldots, N]\) to appear at the rest of the first \( n_2 \) entries of the permutation (there are \( \binom{N-n_1}{n_2-b} \) possibilities for that) and consider all possible \( (n_2-b)! \) arrangements. Finally, we take into account all possible \( (N-n_2)! \) arrangements of the rest \( N-n_2 \) entries of the permutation.
A straightforward upper bound for the number of permutations in \( S_N \) having mmHG score better than \( s \) follows:

\[
|\{\pi' \in S_N : \text{mmHG}(\pi') \leq s\}| \leq \sum_{n_1,n_2,b:\ HGT(N,n_1,n_2,b) \leq s} \Lambda(N,n_1,n_2,b)
\]

From which an upper bound is easily derived:

\[
\text{mmHG p-value}(s, N) \leq \frac{\sum_{n_1,n_2,b:\ HGT(N,n_1,n_2,b) \leq s} \Lambda(N,n_1,n_2,b)}{N!}
\]

By algebraic manipulations we get:

\[
\text{mmHG p-value}(s, N) \leq \sum_{n_1,n_2,b:\ HGT(N,n_1,n_2,b) \leq s} \frac{\binom{n_b}{n_1} \binom{N-n_1}{n_2-b}}{\binom{N}{n_2}}
\]

This upper bound is simple and requires \( O(N^3) \) HGT calculations. An HGT calculation takes \( O(N) \) time, assuming binomial coefficients can be calculated in \( O(1) \) time, for example by using Stirling’s approximation \( \Pi \):

\[
\sqrt{2\pi n\left(\frac{n}{e}\right)^n} \frac{1}{e^{2n+1}} \leq n! \leq \sqrt{2\pi n\left(\frac{n}{e}\right)^n} \frac{1}{e^{2n}},
\]

which is tight for large factorials.

### 3.2 A Refined Upper Bound for the p-Value

The upper bound introduced in the previous section counts the number of permutations for which the value \( \text{HGT}(N,n_1,n_2,b) \) is calculated when taking the non-practical exhaustive approach that enumerates over all \( N! \) permutations. Ideally, we wish to count the number of permutations for which the value \( \text{HGT}(N,n_1,n_2,b) \) is also their mmHG score, as a permutation may have several HGT values that are better than \( s \), so it can be counted more than once. This explains why the formula introduced earlier is an upper bound and not an exact p-value. A second observation that follows is that the smaller the mmHG score \( s \) is, the tighter the bound, because a permutation will have fewer combinations \( (N,n_1,n_2,b) \) having HGT score better than \( s \).

Therefore, if we can reduce the extent of multiple counting of the same permutation, we will get a tighter bound. We do this by looking one step backwards. If, for example, \( \text{HGT}(N,n_1,n_2,b) \leq s \), we can exclude from the counting permutations that contain \( b \) elements from the range \([1,\ldots,n_1] \) at their first \( n_2 \) entries because they are already taken into account in \( \Lambda(N,n_1-1,n_2,b) \) (because necessarily \( \text{HGT}(N,n_1-1,n_2,b) \leq s \), as we will later explain).

Let \( \Psi(N,n_1,n_2,b) \) be the set of permutations for which it holds that out of the first \( n_2 \) entries, \( b \) of them are taken from the range \([1,\ldots,n_1] \) (note that \( \Lambda(N,n_1,n_2,b) \) introduced earlier is, therefore, the size of \( \Psi(N,n_1,n_2,b) \)). Assuming \( \text{HGT}(N,n_1,n_2,b) \leq s \), we can partition the set \( \Psi(N,n_1,n_2,b) \) into five disjoint subsets \( \psi_1,\ldots,\psi_5 \) such that \( \psi = \psi_1 \cup \psi_2 \cup \psi_3 \cup \psi_4 \cup \psi_5 \), as follows:
\[ \psi_1 = \Psi(N, n_1, n_2, b) \cap \Psi(N, n_1 - 1, n_2 - 1, b - 1) \cap \Psi(N, n_1 - 1, n_2, b) \]

\[ \psi_2 = \Psi(N, n_1, n_2, b) \cap \Psi(N, n_1 - 1, n_2 - 1, b - 1) \cap \Psi(N, n_1, n_2 - 1, b) \]

\[ \psi_3 = \Psi(N, n_1, n_2, b) \cap \Psi(N, n_1 - 1, n_2 - 1, b - 1) \cap \Psi(N, n_1 - 1, n_2, b - 1) \]

\[ \psi_4 = \Psi(N, n_1, n_2, b) \cap \Psi(N, n_1 - 1, n_2 - 1, b) \]

\[ \psi_5 = \Psi(N, n_1, n_2, b) \cap \Psi(N, n_1 - 1, n_2 - 2, b - 2) \cap \Psi(N, n_1 - 1, n_2, b - 1) \]

\[ \cap \Psi(N, n_1, n_2 - 1, b - 1) \]

The properties of the hypergeometric distribution imply that \( \psi_1, \psi_2, \psi_4 \) can be disregarded, in the current counting stage. To explain why, we will demonstrate the argument on \( \psi_1 \). The permutations in \( \psi_1 \) contain \( b \) elements from the range \([1, \ldots, n_1 - 1]\) at the first \( n_2 \) entries. We also assume that \( \text{hgt}(N, n_1, n_2, b) \leq s \). Therefore \( \text{hgt}(N, n_1 - 1, n_2, b) \leq s \) also holds, as the same intersection is observed for even a smaller set. Thus, the permutations in \( \psi_1 \) should have been counted when handling the triplet \( n_1 - 1, n_2 \) and \( b \) and disregarded for the combination \( n_1, n_2 \) and \( b \). Similar arguments hold for \( \psi_2 \) and \( \psi_4 \).

\( \psi_3 \) should be counted if it holds that \( \text{hgt}(N, n_1 - 1, n_2 - 1, b - 1) > s \) and \( \text{hgt}(N, n_1 - 1, n_2, b - 1) > s \) and \( \text{hgt}(N, n_1, n_2 - 1, b - 1) > s \), otherwise \( \psi_3 \) would have been counted by former triplets. Similarly, \( \psi_5 \) should be counted if \( \text{hgt}(N, n_1 - 1, n_2 - 2, b - 2) > s \) and \( \text{hgt}(N, n_1 - 1, n_2, b - 1) > s \) and \( \text{hgt}(N, n_1, n_2 - 1, b - 1) > s \). Finally, we calculate the sizes of \( \psi_3 \) and \( \psi_5 \), in the relevant cases. The permutations in \( \psi_3 \) contain \( b \)-1 elements taken from the range \([1, \ldots, n_1 - 1]\) located at the first \( n_2 - 1 \) entries, where the number \( n_1 \) is positioned at entry \( n_2 \). Therefore:

\[
|\psi_3| = \binom{n_1 - 1}{b - 1} \binom{n_2 - 1}{b - 1} (b - 1)! \binom{N - n_1}{n_2 - b} (n_2 - b)! (N - n_2)!
\]

The permutations in \( \psi_5 \) contain \( b \)-2 elements taken from \([1, \ldots, n_1 - 1]\) located at the first \( n_2 - 1 \), where \( n_1 \) is positioned at one of the first \( n_2 - 1 \) entries, and also entry \( n_2 \) contains an element from \([1, \ldots, n_1 - 1]\). Therefore:

\[
|\psi_5| = \binom{n_1 - 1}{b - 2} \binom{n_2 - 1}{b - 2} (b - 2)! (n_2 - b + 1)! \binom{N - n_1}{n_2 - b} (n_2 - b)! (n_1 - b + 1) (N - n_2)!
\]

From the above we next conclude an upper bound. Denote

\[
I(\text{hgt}(N, n_1, n_2, b) > s) = \begin{cases} 1, & \text{if } \text{hgt}(N, n_1, n_2, b) > s \\ 0, & \text{otherwise} \end{cases}
\]
\[ L. \text{Leibovich and Z. Yakhini} \]

\[ \Lambda^*(N, n_1, n_2, b) = \]
\[ |\psi_3| \times I(\text{HGT}(N, n_1 - 1, n_2 - 1, b - 1) > s) \]
\[ \times I(\text{HGT}(N, n_1 - 1, n_2, b - 1) > s) \]
\[ \times I(\text{HGT}(N, n_1, n_2 - 1, b - 1) > s) \]
\[ + \]
\[ |\psi_5| \times I(\text{HGT}(N, n_1 - 1, n_2 - 1, b - 2) > s) \]
\[ \times I(\text{HGT}(N, n_1 - 1, n_2, b - 1) > s) \]
\[ \times I(\text{HGT}(N, n_1, n_2 - 1, b - 1) > s) \]

Yielding the following upper bound for the p-value:

\[ \text{mmHG p-value}(s, N) \leq \sum_{n_1, n_2, b: \text{HGT}(N, n_1, n_2, b) \leq s} \frac{\Lambda^*(N, n_1, n_2, b)}{N!} \]

Note that when \( n_1 \) or \( n_2 \leq 1 \), \( \Lambda^*(N, n_1, n_2, b) \) is defined as \( \Lambda(N, n_1, n_2, b) \). Also, given \( N, n_1 \) and \( n_2, b \) can be any integer in \( [\max(0, n_2 - N + n_1), \min(n_1, n_2)] \).

This upper bound uses more delicate counting than the bound introduced in the previous section. In the following sections we assess the tightness of this bound. In later sections we demonstrate an application for PWM motif search.

### 3.3 Comparison to a Different Variant

We note that the bound described in Steinfeld et al. [25] addresses a slightly different variant of mmHG as a random variable over \( S_N \). The definition with which we work here is more amenable to deriving tight bounds as described above. Given a single permutation \( \pi \in S_N \) and for every \( i = 1, \ldots, N \), a binary vector \( \lambda_i \) is defined in which exactly \( i \) entries are 1 and \( N-i \) entries are 0, as follows:

\[ \lambda_i(j) = 1 \text{ if } \pi(j) \leq i. \]

The mmHG score of a permutation \( \pi \) is then defined by Steinfeld et al. [25] as:

\[ \text{mmHG}(\pi) = \min_{1 \leq i \leq N} P - \text{value}(\text{mHG}(\lambda_i)), \]

where \( \text{mHG}(\lambda) = \min_{1 \leq i \leq N} \text{HGT}(N, B, n, b_n), N = |\lambda|, b_n = \sum_{i=1}^n \lambda_i \) and \( B = b_N \). A possible upper bound is then given by:

\[ * \quad P - \text{value}(\text{mmHG}(\pi)) \leq \min_{1 \leq i \leq N} \text{mHG}(\lambda_i) \cdot i \cdot N \]

Computing the latter quantity requires \( O(N^2) \) HGT calculations and is therefore more computationally efficient than the two bounds described in Sections 3.1 and 3.2 of this current work, that require \( O(N^3) \) HGT calculations. We observed that our bound was tighter than the bound in (*), as later shown in Figure 1D.

For example, for a permutation having mmHG score = \( 7.8 \cdot 10^{-25} (N = 100) \), our
bound was \(3.5 \cdot 10^{-23}\) while (*) yielded \(4.2 \cdot 10^{-21}\). For one permutation with mmHG score = \(5.1 \cdot 10^{-5}(N = 100)\), our bound was 0.026 while (*) yielded 0.2. The latter example demonstrates that a tighter bound is important for classifying an observation as statistically significant (assuming a significance threshold of 0.05).

### 3.4 Assessment of Tightness

In order to assess the quality of our bound, we compared it to the exact p-value, which can be calculated for small values of \(N\) (that is, in cases where \(N!\) is not too large). Figure 1A compares the mmHG score (which also serves as a lower bound for the p-value), the exact p-value (calculated by exhaustive enumeration of all \(10!\) permutations), our upper bound and the Bonferroni corrected p-value for \(N=10\). Figure 1B shows the same for \(N=20\), except that exact p-values cannot be calculated exhaustively, and therefore an empirical p-value is produced by randomly sampling \(10^7\) permutations. In both cases our upper bound is significantly tighter than the Bonferroni bound. We also observed that the smaller the mmHG scores are – the tighter is our bound, consistent with lesser over-counting for smaller scores, as explained in previous sections. Comparison between the first bound described in Section 3.1 and the bound described in Section 3.2 is shown in Figure 1C (for \(N=20\)). We observed that enumerating all HGT scores rather than enumerating all permutations in \(S_N\) significantly improves the p-value estimation. Moreover, the refinement of this approach produced by reducing the extent of multiple counting of permutation further improves the upper bound. In Figure 1D the bounds, including the bound introduced in Section 3.3 (Steinfeld bound), are shown for \(N=100\). An empirical p-value was not calculated here as even if we sample \(10^7\) permutations, a p-value smaller than \(10^{-7}\) cannot be obtained. The bound suggested in this paper was almost always observed to be tighter than the bound introduced in Section 3.3.

### 3.5 Application in PWM Motif Search

In this section we discuss mmHG as a framework for assessing the significance of PWM motifs in ranked lists. Given a ranked list of sequences and a PWM motif, by using the mmHG statistics and the bounds introduced earlier, we can assign a p-value to represent the significance of that PWM being enriched at the top of the list. To apply this approach for de-novo motif search, one needs to theoretically consider all possible PWMs. This is not feasible and as a heuristic approach we wrote mmHG-Finder which takes as input a ranked list of DNA or RNA sequences and returns significant motifs in PWM format. In cases where sequence ranking is not relevant or not available, it allows the use of positive and negative sets of sequences, searching for enriched motifs in the positive set using the negative set as the background.
We will now describe the methodology implemented in mmHG-Finder:

Input:
- a ranked list of sequences (or, alternatively, two sets of sequences representing target and background)
- motif width, given as a range between $k_1$ and $k_2$
Algorithm:

1. Build a generalized suffix tree for the sequences
2. Traverse the tree to find all k-mers for $k=k_1, \ldots, k_2$
3. Sort the k-mers according to their enrichment at the top of the list (this is done using the mHG statistics), as explained in Leibovich et al. [15]
4. Take the most significant fifty k-mers, to be used as starting points for the next step. This set of candidates is chosen such that the members are quite different. Note that this is a heuristic approach and the number 50 is somewhat arbitrary, chosen to succeed in catching the best performing PWMs without heavily paying in complexity.
5. For each starting point, we iteratively replace one position in the k-mers by considering all possible IUPAC replacements and taking the one that improves the enrichment the most. We repeat this process for all positions several times. Eventually we get a motif in the IUPAC alphabet which is then converted to a PWM.
6. The PWMs found in the previous step are assessed using the mmHG statistics and the best is returned as output, together with the p-value. The score assigned by a PWM to a string $S = S_1, \ldots, S_M$ is defined as $\max_{1 \leq i \leq M-K+1} \sum_{j=1}^{K} m_{S_{i+j-1}, j}$ (assuming $M \geq K$, otherwise it is $-\infty$), where $m_{\alpha, c}$ is the score in row $\alpha$, column $c$ of the position weight matrix. In other words, the PWM score calculated for S is the maximal score obtained for a substring of S.

To evaluate the performance of mmHG-Finder in comparison to other state-of-the-art methods we ran it on 18 example cases – 3 synthetically generated cases and 15 generated from high throughput binding experiments (6 transcription factors and 9 RNA-binding proteins). We compared the results to those obtained by using three other methods: the standard MEME program [3], DREME [4], and XXmotif [17]. Some of the results of this comparison are summarized in Table 1. The synthetic examples were generated by randomizing 500 sequences of length 100. An IUPAC motif was generated and planted in all top 64 sequences. mmHG-Finder outperformed all the other three tools on the synthetic examples, which contained degenerate motifs. MEME and DREME did not find the motifs in any case, while XXmotif found a similar result in 1 out of the 3 tests. The other 15 examples were taken from DNA and RNA high-throughput experiments [23], [10], [12]. In 12 out of these 15 datasets, mmHG-Finder found the motifs which were compatible with the known literature motifs as the most significant result. In comparison, DREME found the known consensus in 11 cases; XXmotif detected the literature motif in 9 cases while MEME detected the known motif in only 7 cases. In several datasets, such as for GCN4 and Pin4, mmHG-Finder identified the consensus motif while other tools returned repetitive sequences as their top results. The mmHG statistics avoids such spurious results as they typically do not correlate with the measurement driven ranking.

Computing p-value bounds for the synthetic examples ($N=500$) took 7-17 seconds on a simple single-core laptop. The running time depends on both the number of elements $N$ as well as the mmHG score. The computation is optimized such that it is quicker for smaller mmHG scores. It took 33 minutes for $N=5000$ where the mmHG score was $3.3 \cdot 10^{-69}$, and 39 minutes for $N=4000$ and mmHG score = $5.9 \cdot 10^{-31}$. 
Table 1. We evaluated the performance of mmHG-Finder in comparison to other state-of-the-art methods: MEME, DREME and XXmotif. Almost all input examples comprised ranked lists, except for p53 (comprising target and background sets). Since MEME, DREME, and XXmotif expect a target set as input, we converted the ranked lists into target sets by taking the top 100 sequences for MEME (restricted by MEME’s limitation of 60,000 characters) and the top 20% sequences for the other tools. In the synthetic examples the entire ranked lists were taken as they are sufficiently small. Data and consensus motifs for p53 were taken from [23]; for REB1, CBF1, UME6, TYE7, GCN4 from [10]; and for the RNA binding proteins from [12]. Selected results are shown below.

<table>
<thead>
<tr>
<th>The protein and its consensus binding motif</th>
<th>mmHG-Finder</th>
<th>MEME</th>
<th>DREME</th>
<th>XXmotif</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic TNWMNG W=</td>
<td>A/T</td>
<td>, M=</td>
<td>A/C</td>
<td></td>
</tr>
<tr>
<td>Synthetic CTNNNAT</td>
<td>P≤1.32e-28</td>
<td>7.1e+001</td>
<td>Nothing found</td>
<td>1.84e+01</td>
</tr>
<tr>
<td>Synthetic MMMMMMMM M=</td>
<td>A/C</td>
<td></td>
<td>P≤1.07e-39</td>
<td>1.8e+002</td>
</tr>
<tr>
<td>P53 (DNA)</td>
<td>P≤1.09e-174</td>
<td>1.8e-100</td>
<td>4.9e-133</td>
<td>1e-490</td>
</tr>
<tr>
<td>GCN4 (DNA) TGAstCca</td>
<td>P≤2.05e-44</td>
<td>1.3e-85</td>
<td>2.0e-32</td>
<td>4.00e-17</td>
</tr>
<tr>
<td>Puf5 (RNA)</td>
<td>P≤7.93e-79</td>
<td>3.6e-9</td>
<td>6.8e-42</td>
<td>9.76e-21</td>
</tr>
<tr>
<td>Pin4 (RNA)</td>
<td>P≤8.18e-8</td>
<td>1.3e+0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4 Concluding Remarks

Due to the size of the measure space, statistics over $S_N$ is difficult to implement. We derive polynomially computable bounds for the tail distribution of the mutual enrichment at the top of two permutations uniformly and independently drawn over $S_N$. We assess tightness using simulated data. We also demonstrate utility of the mmHG statistics in identifying motifs in experimental binding affin-
ity data. For several representative datasets, including synthetically generated data, we note that our bound improves the p-value estimates by a factor of 50. The full characterization of the distribution of mmHG as a random variable over $S_N$ remains an open question.

**Acknowledgments.** We thank Israel Steinfeld for critical and inspiring discussions. We also thank the anonymous reviewers for their useful comments. LL was partially supported by Israel Ministry of Science and Technology and by ISEF Fellowship.

**References**

1. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables (1964)
3.5 Mutual Enrichment in Ranked Lists and the Statistical Assessment of Position Weight Matrix Motifs (extended version)

Mutual enrichment in ranked lists and the statistical assessment of position weight matrix motifs

Limor Leibovich¹ and Zohar Yakhini¹,2*

Abstract

Background: Statistics in ranked lists is useful in analysing molecular biology measurement data, such as differential expression, resulting in ranked lists of genes, or ChIP-Seq, which yields ranked lists of genomic sequences. State of the art methods study fixed motifs in ranked lists of sequences. More flexible models such as position weight matrix (PWM) motifs are more challenging in this context, partially because it is not clear how to avoid the use of arbitrary thresholds.

Results: To assess the enrichment of a PWM motif in a ranked list we use a second ranking on the same set of elements induced by the PWM. Possible orders of one ranked list relative to another can be modelled as permutations. Due to sample space complexity, it is difficult to accurately characterize tail distributions in the group of permutations. In this paper we develop tight upper bounds on tail distributions of the size of the intersection of the top parts of two uniformly and independently drawn permutations. We further demonstrate advantages of this approach using our software implementation, mmHG-Finder, which is publicly available, to study PWM motifs in several datasets. In addition to validating known motifs, we found GC-rich strings to be enriched amongst the promoter sequences of long non-coding RNAs that are specifically expressed in thyroid and prostate tissue samples and observed a statistical association with tissue specific CpG hypo-methylation.

Conclusions: We develop tight bounds that can be calculated in polynomial time. We demonstrate utility of mutual enrichment in motif search and assess performance for synthetic and biological datasets. We suggest that thyroid and prostate-specific long non-coding RNAs are regulated by transcription factors that bind GC-rich sequences, such as EGR1, SP1 and E2F3. We further suggest that this regulation is associated with DNA hypo-methylation.

Keywords: Statistical enrichment, Position weight matrices, High-throughput sequencing data analysis, Tissue specific methylation patterns, lncRNA

Background

Modern data analysis often faces the task of extracting characteristic features from sets of elements singled out according to some measurement. In molecular biology, for example, an experiment may lead to measurement results pertaining to genes and then questions are asked about the properties of genes for which these were high or low. This is an example, of course, and the set of elements does not have to be genes. They can be genomic regions, proteins, structures, etc. A central technique for addressing the analysis of characteristic properties of sets of elements is statistical enrichment. More specifically – the experiment results are often representable as ranked lists of elements and we then seek enrichment of other properties of these elements at the top or bottom of the ranked list. GSEA [1], for example, is a tool that addresses characteristic features of genes that are found to be differentially expressed in a comparative transcriptomics study. GOrilla [2,3] addresses GO terms enriched in ranked lists of genes where the ranking can be, for example, the result of processing
differential expression data or of correlations computed between genomic DNA copy number and expression [4-6], FATIGO [7] is also a tool that is useful in the context of analysing GO terms in ranked lists of genes. DRI-Must [8-10] searches for sequence motifs that are enriched, in a statistically significant manner, in the top of a ranked list of sequences, which can be produced by techniques like ChIP-Seq.

All the aforementioned tools utilize a statistical approach that is based on assessing enrichment of an input set in an input ranked list by quantifying the enrichment obtained at various cutoffs applied to the ranked list. It is often the case, however, that two quantitative properties need to be compared to each other. For example, when the elements are genes, we may have measured differential expression values, as well as measured ChIP-Seq signals. We are therefore interested in assessing mutual enrichment in two ranked lists. Another example consists of one ranking according to differential expression and one according to prediction scores for miRNA targets. miTEA [11,12] addresses this latter case by statistically assessing the enrichment of miRNA targets in a ranked list of genes (also see [13]). To address mutual enrichment in two ranked lists over the same set of N elements, miTEA [11] performs analysis on permutations. Mutual enrichment in the top of two ranked lists can be simplified, from a mathematical point of view, by arbitrarily setting the indices of one list to the identity permutation (1,2,...,N) and treating the other list as a permutation π = π(1),...,π(N) over these numbers. For the purpose of assessing the intersection of the top of the two ranked lists in a data driven manner, miTEA asks which prefix [1,...,n1] is enriched in the first n2 elements of π, that is in the set π(1),...,π(n2). The statistics introduced by miTEA is called mmHG (minimum-minimum-Hyper-Geometric). A slightly different variant of mmHG is described later in this section.

Statistics in the group of permutations $S_N$ is often difficult because the number of entities to be considered by any null model is N!. Direct exhaustive calculation of tail distributions over $S_N$ is therefore impractical for all but very small values of N. This difficulty is addressed by several heuristic techniques. Mapping into continuous spaces, such as in [14], has proven useful in certain cases but not for studying large deviations. In the case of enrichment statistics that focuses on the top of the permutation and seeks to assess extreme events, such as mmHG, we prefer to use bounds on tail probabilities. Tail probabilities are useful constructs when applied to analysing molecular biology measurement data as they enable statistical assessment of observed results.

In this work we derive tight bounds on the tail probabilities of mutual enrichment at the top of two random permutations uniformly drawn over $S_N$ and demonstrate the utility of this approach in the context of flexible motif discovery. Our bounds are computable in polynomial time and potentially add to the accuracy of reported position weight matrix (PWM) motifs for nucleic acid sequences.

**Mutual enrichment in ranked lists – the mmHG statistics**

The mmHG statistics [11] is a generalization of the mHG statistics [2,15-17]. The mHG statistics quantifies the enrichment level of a set of elements at the top of a ranked list of elements of the same type, whereas the mmHG statistics assesses the level of mutual enrichment in two ranked lists over the same set of elements. While any parametric or non-parametric correlation statistics (e.g. Spearman’s correlation coefficient), that takes the same input, calculates the overall agreement between the two ranked lists, the mmHG statistic focuses only on agreement at the top of the two ranked lists. mmHG counts elements common to the top of both lists, without predefining what top is. Its intended output is the probability of observing an intersection at least as large in two randomly ranked lists (defined as the enrichment p-value). In this section we describe the mmHG statistics and in later sections we suggest tight bounds for the p-value. Our definition of the mmHG statistics varies slightly from that of Steinfeld et al. [11], which is used by miTEA.

Mutual enrichment in the top of two ranked lists can be simplified, from a mathematical point of view, by arbitrarily setting the indices of one list to the identity permutation (1,2,...,N) and treating the other list as a permutation $π = π(1),...,π(N)$ over these numbers. For the purpose of assessing the intersection of the top of the two ranked lists in a data driven manner, miTEA asks which prefix [1,...,n1] is enriched in the first n2 elements of $π$, that is in the set $π(1),...,π(n2)$. The statistics introduced by miTEA is called mmHG (minimum-minimum-Hyper-Geometric). A slightly different variant of mmHG is described later in this section.

Statistics in the group of permutations $S_N$ is often difficult because the number of entities to be considered by any null model is N!. Direct exhaustive calculation of tail distributions over $S_N$ is therefore impractical for all but very small values of N. This difficulty is addressed by several heuristic techniques. Mapping into continuous spaces, such as in [14], has proven useful in certain cases but not for studying large deviations. In the case of enrichment statistics that focuses on the top of the permutation and seeks to assess extreme events, such as mmHG, we prefer to use bounds on tail probabilities. Tail probabilities are useful constructs when applied to analysing molecular biology measurement data as they enable statistical assessment of observed results.

In this work we derive tight bounds on the tail probabilities of mutual enrichment at the top of two random permutations uniformly drawn over $S_N$ and demonstrate the utility of this approach in the context of flexible motif discovery. Our bounds are computable in polynomial time and potentially add to the accuracy of reported position weight matrix (PWM) motifs for nucleic acid sequences.

**Mutual enrichment in ranked lists – the mmHG statistics**

The mmHG statistics [11] is a generalization of the mHG statistics [2,15-17]. The mHG statistics quantifies the enrichment level of a set of elements at the top of a ranked list of elements of the same type, whereas the mmHG statistics assesses the level of mutual enrichment in two ranked lists over the same set of elements. While any parametric or non-parametric correlation statistics (e.g. Spearman’s correlation coefficient), that takes the same input, calculates the overall agreement between the two ranked lists, the mmHG statistic focuses only on agreement at the top of the two ranked lists. mmHG counts elements common to the top of both lists, without predefining what top is. Its intended output is the probability of observing an intersection at least as large in two randomly ranked lists (defined as the enrichment p-value). In this section we describe the mmHG statistics and in later sections we suggest tight bounds for the p-value. Our definition of the mmHG statistics varies slightly from that of Steinfeld et al. [11], which is used by miTEA.

Mutual enrichment in the top of two ranked lists can be simplified, from a mathematical point of view, by arbitrarily setting the indices of one list to the identity permutation (1,2,...,N) and treating the other list as a permutation $π = π(1),...,π(N)$ over these numbers. For the purpose of assessing the intersection of the top of the two ranked lists in a data driven manner, miTEA asks which prefix [1,...,n1] is enriched in the first n2 elements of $π$, that is in the set $π(1),...,π(n2)$. The statistics introduced by miTEA is called mmHG (minimum-minimum-Hyper-Geometric). A slightly different variant of mmHG is described later in this section.

Statistics in the group of permutations $S_N$ is often difficult because the number of entities to be considered by any null model is N!. Direct exhaustive calculation of tail distributions over $S_N$ is therefore impractical for all but very small values of N. This difficulty is addressed by several heuristic techniques. Mapping into continuous spaces, such as in [14], has proven useful in certain cases but not for studying large deviations. In the case of enrichment statistics that focuses on the top of the permutation and seeks to assess extreme events, such as mmHG, we prefer to use bounds on tail probabilities. Tail probabilities are useful constructs when applied to analysing molecular biology measurement data as they enable statistical assessment of observed results.

In this work we derive tight bounds on the tail probabilities of mutual enrichment at the top of two random permutations uniformly drawn over $S_N$ and demonstrate the utility of this approach in the context of flexible motif discovery. Our bounds are computable in polynomial time and potentially add to the accuracy of reported position weight matrix (PWM) motifs for nucleic acid sequences.
The mmHG score cannot be considered as a significance measure, due to the multiple testing involved in finding \( n_1 \) and \( n_2 \). A simple way to correct an mmHG score \( s \) for multiple testing and report an upper bound on the \( p \)-value is to use the Bonferroni correction. Basically, \( s \) is multiplied by the number of multiple tests conducted (which is \( N^2 \)), yielding an upper bound on the \( p \)-value, as follows:

\[
\text{mmHG } p\text{-value}(s, N) \leq s \cdot N^2
\]

In the Results section we present significantly tighter bounds.

Position weight matrix motifs

Data produced by techniques such as ChIP-Seq [18], ChIP-exo [19], CLIP [20], PAR-CLIP [21] and others are readily representable as ranked lists of sequences, where the ranking is according to the measured binding affinity. Computational tools and approaches to motif discovery form part of the data analysis workflow that is used to extract knowledge and understanding from this type of studies. We are often interested in sequence motifs that are observed to be enriched in sequences where strong binding affinity is measured. A position weight matrix (PWM) is a commonly used representation of motifs in biological sequences [22-24]. This representation is more faithful to the underlying biology than representation by exact words, owing to the tendency of binding sites to be short and degenerate [25]. A PWM is a matrix of score values that gives a weighted match to any given substring of fixed length. It has one row for each symbol in the alphabet, and one column for each position in the pattern. Assuming an input sequence of each symbol in the alphabet, and one column for each any given substring of fixed length. It has one row for

\[
\text{presentation by exact words, owing to the tendency of}
\]

enrichment of the top parts of

Given a set of sequences that were tested in a high throughput experiment such as ChIP-Seq [18], CLIP [20] and others, they can be ranked according to the measured binding affinities, yielding a ranked list \( L_1 \). Since usually we are interested in finding motifs amongst sequences having strong binding affinities, we actually search for motifs that are more prevalent at the top of this list. It is clear that any algorithm for de-novo flexible motif search would need to evaluate candidate PWMs. Given a PWM which we want to assess, the sequences can also be ranked according to their PWM scores, yielding another ranked list \( L_2 \), different from \( L_1 \). A significant PWM motif would yield significant scores for sequences having strong binding affinities. Therefore, the question of PWM motif discovery from ranked experimental data can be formulated as quantifying the mutual enrichment level for the two ranked lists \( L_1 \) and \( L_2 \). Given two ranked lists \( L_1 \) and \( L_2 \) over the universe of \( N \) sequences, they can be transformed into two respective permutations, \( \pi_1 = (\pi_1(1), ..., \pi_1(N)) \) and \( \pi_2 = (\pi_2(1), ..., \pi_2(N)) \). The relative permutation \( \pi \), of \( \pi_2 \) w.r.t. \( \pi_1 \), is defined by \( \pi(\pi_1(j)) = \pi_2(j) \) for every \( j = 1, ..., N \), or simply, using operations in the group \( S_N \): \( \pi = \pi_2 \cdot \pi_1^{-1} \). Using the relative permutation \( \pi \), we can represent the mutual enrichment of the top parts of \( L_1 \) and \( L_2 \) as mmHG score \( (\pi) \), defined above.

Results

Estimation of the mmHG \( p \)-value – introducing first upper bound – B1

Given an mmHG score \( s \), observed in analysing real measurement data, we would like to assess the statistical significance of this observation. Assuming endless computational power, we would enumerate all permutations and calculate the mmHG score for each, in order to characterize the distribution of mmHG as a random variable over \( S_N \). The \( p \)-value for \( s \) is then simply:

\[
\text{mmHG } p\text{-value}(s, N) = \frac{\text{The number of permutations having mmHG scores } s} {N!}
\]

Since the number of permutations is huge, the process described above is very far from feasible. Therefore, we seek a computationally tractable upper bound, preferably tight.

A trivial upper bound is the Bonferroni corrected mmHG \( p \)-value defined by \( s \cdot N^2 \). A more subtle upper bound was suggested by Steinfeld et al. [11] and is briefly described later as bound B3. In this work we introduce tighter bounds that are polynomially computable.

We next describe an intuitive upper bound (B1) which we later refine to produce a tighter bound (B2). The input of the problem consists of an mmHG score \( s \) and the total number of elements \( N \). The output is an upper
bound on the \( p \)-value. The efficiency of our approach relies on enumerating all possible HGT scores rather than enumerating all permutations in \( S_N \). This approach is computationally efficient as HGT is a function of four input parameters: \( N, n_1, n_2, \) and \( b \). Given \( N \), there are \( O(N^3) \) possible combinations of \( n_1, n_2, \) and \( b \). Also, given \( N, n_1 \) and \( n_2, b \) can be any integer in the range \([\min(0, n_2 - N + n_1), \min(n_1, n_2)]\). Next, all is left to do is to determine how many permutations correspond to each HGT score. To this end, let \( \Lambda(N, n_1, n_2, b) \) be the number of permutations for which it holds that \( b \) out of the first \( n_2 \) entries in the permutation are taken from the range \([1, \ldots, n_1]\). This formulation is equivalent to counting permutations for which we attain, at some point, the value HGT(\( N, n_1, n_2, b \)), had we taken the exhaustive approach. \( \Lambda(N, n_1, n_2, b) \) can be represented as:

\[
\Lambda(N, n_1, n_2, b) = \binom{n_1}{b} \binom{n_2}{b} \binom{N - n_1}{n_2 - b} (n_2 - b)! (N - n_2)!
\]

as we first choose \( b \) elements from the range \([1, \ldots, n_1]\) to appear at the first \( n_2 \) entries of the permutation (there are \( \binom{n_1}{b} \) possibilities). Then, we choose the positions amongst the first \( n_2 \) entries that are occupied by these \( b \) elements, while considering all internal arrangements (for each choice of \( b \) elements there are \( \binom{n_2}{b} \) possibilities). We next choose \( n_2 - b \) elements from the range \([n_1 + 1, \ldots, N]\) to appear at the rest of the first \( n_2 \) entries of the permutation (there are \( \binom{N - n_1}{n_2 - b} \) possibilities for that) and consider all possible \((n_2 - b)!\) arrangements. Finally, we take into account all possible \((N - n_2)!\) arrangements of the remaining \( N - n_2 \) entries of the permutation.

A straightforward upper bound for the number of permutations in \( S_N \) having mmHG score better than \( s \) follows:

\[
|\{ \pi \in S_N : \text{mmHG}(\pi) \leq s \}| \leq \sum_{n_1, n_2, b : \text{HGT}(n_1, n_2, b) \leq s} \Lambda(N, n_1, n_2, b)
\]

From which an upper bound is easily derived:

\[
\text{mmHG} \ p\text{-value}(s, N) \leq \sum_{n_1, n_2, b : \text{HGT}(n_1, n_2, b) \leq s} \frac{\Lambda(N, n_1, n_2, b)}{N!}
\]

By algebraic manipulations we get:

\[
\text{mmHG} \ p\text{-value}(s, N) \leq \sum_{n_1, n_2, b : \text{HGT}(n_1, n_2, b) \leq s} \frac{\binom{n_1}{b} \binom{N - n_1}{n_2 - b}}{\binom{N}{n_2}}
\]

This upper bound is simple and requires \( O(N^3) \) HGT calculations. An HGT calculation takes \( O(N) \) time, assuming binomial coefficients can be calculated in constant time. Constant time computation can be achieved using Stirling’s approximation [27]:

\[
\sqrt{2\pi n} \left( \frac{n}{e} \right)^n \leq n!,
\]

which is tight for large factorials.

**A refined upper bound for the \( p\)-value - B2**

The upper bound introduced in the previous section counts the number of permutations for which the value HGT(\( N, n_1, n_2, b \)) is calculated when taking the non-practical exhaustive approach that enumerates over all \( N! \) permutations. Ideally, we wish to count the number of permutations for which the value HGT(\( N, n_1, n_2, b \)) is also their mmHG score, as a permutation may correspond with many HGT values that are better than \( s \), so it can be counted more than once. This explains why the formula introduced earlier is an upper bound and not an exact \( p\)-value. A second observation that follows is that the smaller the mmHG score \( s \), the tighter the bound, because a permutation will have fewer combinations (\( N, n_1, n_2, b \)) having HGT values better than \( s \).

Therefore, if we can reduce the extent of multiple counting of the same permutation, we will get a tighter bound. We do this by looking one step backwards. If, for example, HGT(\( N, n_1, n_2, b \)) \( \leq s \), we can exclude from the counting permutations that contain \( b \) elements from the range \([1, \ldots, n_1]\) at their first \( n_2 \) entries because they are already taken into account in \( \Lambda(N, n_1 - 1, n_2, b) \) (because necessarily HGT(\( N, n_1 - 1, n_2, b \)) \( \leq s \), as we will later explain).

Let \( \psi(N, n_1, n_2, b) \) be the set of permutations for which it holds that \( b \) out of the first \( n_2 \) entries are taken from the range \([1, \ldots, n_1]\) (note that \( \Lambda(N, n_1, n_2, b) \) introduced earlier is, in fact, the size of \( \psi(N, n_1, n_2, b) \)). Assuming HGT(\( N, n_1, n_2, b \)) \( \leq s \), we can partition the set \( \psi(N, n_1, n_2, b) \) into five disjoint subsets \( \psi_1, \ldots, \psi_5 \) such that:

\[
\psi_1 = \psi(N, n_1, n_2, b) \cap \psi(N, n_1 - 1, n_2 - 1, b - 1) \cap \psi(N, n_1 - 1, n_2, b) \\
\psi_2 = \psi(N, n_1, n_2, b) \cap \psi(N, n_1 - 1, n_2 - 1, b - 1) \cap \psi(N, n_1 - 1, n_2, b - 1) \\
\psi_3 = \psi(N, n_1, n_2, b) \cap \psi(N, n_1 - 1, n_2, b - 1) \cap \psi(N, n_1 - 1, n_2 - 1, b - 1) \\
\psi_4 = \psi(N, n_1, n_2, b) \cap \psi(N, n_1 - 1, n_2 - 1, b - 1) \\
\psi_5 = \psi(N, n_1, n_2, b) \cap \psi(N, n_1 - 1, n_2 - 1, b - 2)
\]

The properties of the hypergeometric distribution imply that given a tuple \( (N, n_1, n_2, b) \), the permutations in \( \psi_1, \psi_2, \psi_4 \) can be disregarded from the current counting iteration. To explain why, we will demonstrate the argument on \( \psi_1 \). The permutations in \( \psi_3 \) contain \( b \) elements from the range \([1, \ldots, n_1 - 1]\) at their first \( n_2 \) entries. Recall that we also assume that HGT(\( N, n_1, n_2, b \)) \( \leq s \).
Therefore $\text{HGT}(N, n_1 - 1, n_2, b) \leq s$ also holds, as the same intersection is observed for even a smaller set. Thus, the permutations in $\psi_1$ have already been counted as having HGT value better than $s$ when handling the triplet $(n_1 - 1, n_2, b)$, and can be disregarded for the combination of $n_1$, $n_2$ and $b$. Similar arguments hold for $\psi_3$ and $\psi_4$.

The permutations in $\psi_3$ should be counted if three conditions hold: the first is $\text{HGT}(N, n_1 - 1, n_2 - 1, b - 1) > s$; the second is $\text{HGT}(N, n_1 - 1, n_2, b - 1) > s$; and the third is $\text{HGT}(N, n_1, n_2 - 1, b - 1) > s$. Otherwise, the permutations in $\psi_3$ have been counted by former triplets. Similarly, the permutations in $\psi_3$ should be counted if the following three conditions hold: $\text{HGT}(N, n_1 - 1, n_2 - 1, b - 2) > s$, $\text{HGT}(N, n_1 - 1, n_2, b - 1) > s$, and $\text{HGT}(N, n_1, n_2 - 1, b - 1) > s$. Finally, we calculate the sizes of $\psi_3$ and $\psi_4$, in the relevant cases. The definition of $\psi_4$ implies that it consists of permutations that contain $b - 1$ elements taken from the range $[1, ..., n_1 - 1]$ at their first $n_2 - 1$ entries, and also $n_1$ is positioned at entry $n_2$. Therefore:

$$|\psi_3| = \binom{n_1 - 1}{b - 1} \binom{n_2 - 1}{b - 1} \binom{(b - 1)!}{(n_2 - b)!(n_1 - n_2)!} \binom{N - n_1}{n_2 - b} \binom{N - n_2}{n_1 - n_2}!$$

Equivalently, the permutations in $\psi_3$ contain $b - 2$ elements taken from the subset $[1, ..., n_1 - 1]$ at their first $n_2 - 1$ entries; $n_1$ is positioned at one of the first $n_2 - 1$ entries; and entry $n_2$ contains an element from $[1, ..., n_1 - 1]$. Therefore:

$$|\psi_3| = \binom{n_1 - 1}{b - 2} \binom{n_2 - 1}{b - 2} \binom{(b - 2)!}{(n_2 - b)!(n_1 - n_2)!} \binom{N - n_1}{n_2 - b} \binom{N - n_2}{n_1 - n_2}!$$

From the above we next conclude an upper bound. Denote

$$I(\text{HGT}(N, n_1, n_2, b) > s) = \begin{cases} 1, & \text{if } \text{HGT}(N, n_1, n_2, b) > s \\ 0, & \text{otherwise} \end{cases}$$

And let $\Lambda^*(N, n_1, n_2, b) = $

$$|\psi_3| \times I(\text{HGT}(N, n_1 - 1, n_2, b - 1) > s) \times I(\text{HGT}(N, n_1 - 1, n_2 - 1, b - 1) > s) \times I(\text{HGT}(N, n_1, n_2 - 1, b - 1) > s) + |\psi_3| \times I(\text{HGT}(N, n_1 - 1, n_2 - 1, b - 2) > s) \times I(\text{HGT}(N, n_1 - 1, n_2, b - 1) > s) \times I(\text{HGT}(N, n_1, n_2 - 1, b - 1) > s) \times I(\text{HGT}(N, n_1, n_2, b - 1) > s)$$

We can thus derive the following upper bound for the $p$-value:

$$\text{mmHG} \ p-value(s, N) \leq \sum_{n_1, n_2, b} I(\text{HGT}(N, n_1, n_2, b) > s) \frac{\Lambda^*(N, n_1, n_2, b)}{N!}$$

Since $\Lambda^*$ is recursive, we need to define a base case. Recall that given $N$, $n_1$ and $n_2$, $b$ can be any integer in the range $[\max(0, n_2 - N + n_1), \min(n_1, n_2)]$, hence determining a base case for $n_1$ and $n_2$ is sufficient ($N$ is known). The base case here is that when $n_1 \leq 1$ or $n_2 \leq 1$, $\Lambda^*(N, n_1, n_2, b)$ is defined the same as $\Lambda(N, n_1, n_2, b)$.

This upper bound uses more delicate counting than the bound B1 introduced in the previous section. In the following sections we assess the tightness of this bound. In later sections we demonstrate an application for an upper bound for PWM motif search.

**Comparison to a different mmHG variant – B3**

We note that the bound described in Steinfeld et al. [11] addresses a slightly different variant of mmHG as a random variable over $S_N$. The definition with which we work here is more amenable to deriving tight bounds as described above. Given a single permutation $\pi \in S_N$ and for every $i = 1, ..., N$, a binary vector $\lambda_i$ is defined in which exactly $i$ entries are 1 and $N - i$ entries are 0, as follows: $\lambda_i(j) = 1$ if $\pi(j) \leq i$. The mmHG score of a permutation $\pi$ is then defined by Steinfeld et al. [11] as:

$$\text{mmHG}(\pi) = \min_{1 \leq \pi_i \leq N} P-value(\text{mmHG}(\lambda_i)) \leq \min_{1 \leq \pi_i \leq N} \text{mmHG}(\lambda_i) \cdot i$$

Where $mHG(\lambda_i) = \min_{1 \leq \pi_i \leq N} \text{HGT}(N, i, n, b_n)$, $N = |\lambda_i|$ and $b_n = \sum_{k=1}^{n} \lambda_i(k)$. A possible upper bound is then given by:

$$(+) \quad P-value(\text{mmHG}(\pi)) \leq \min_{1 \leq \pi_i \leq N} \text{mmHG}(\lambda_i) \cdot i \cdot N$$

Computing the latter quantity requires $O(N^2)$ HGT calculations, and is therefore computationally more efficient than the two bounds B1 and B2 of this current work (that require $O(N^3)$ HGT calculations). We observed that our bound B2 was tighter than the bound in $(*)$, as later shown in Figure 1D. For example, for a permutation having mmHG score = 7.8·10^{-25} ($N = 100$), our bound was 3.5·10^{-23} while $(*)$ yielded 4.2·10^{-21}. For one permutation with mmHG score = 5.1·10^{-25} ($N = 100$), our bound was 0.026 while $(*)$ yielded 0.2. The latter example demonstrates that a tighter bound is important for classifying an observation as statistically significant (assuming a significance threshold of 0.05).

**Assessment of tightness**

In order to assess the quality of our bound B2, we compared it to the $p$-value, which can be calculated exactly for small values of $N$ (that is, in cases where $N!$ is not too large) and empirically for larger values of $N$ (by randomly sampling permutations). Evidently, our bound B2 was significantly tighter than the Bonferroni bound for $N = 10$ (Figure 1A) and $N = 20$ (Figure 1B). We also observed that the smaller the mmHG scores – the tighter the bound, consistent with lesser over-counting.
for smaller scores as explained in previous sections. Furthermore, our refined bound B2 is tighter than the bound B1 (Figure 1C), and the latter is significantly better than the Bonferroni bound. Both bounds B1 and B2 are derived by enumerating HGT scores rather than enumerating permutations in $S_N$. The refinement of this approach produced by reducing the extent of multiple counting of permutation further improves the upper bound. In addition, the bound B2 was almost always observed to be tighter than the bound B3 (Figure 1D).

**An upper bound which balances between tightness and computational cost – B4**

The bound B2 is, evidently, very tight. It is, however, computationally heavy. We would still like to have an upper bound which is tighter than the Bonferroni bound and than the variant B3 but also faster to calculate. Such a compromise is achieved by generalizing an approach developed in [15] for the minimum hyper-geometric statistics. Namely, given the number of elements $N$ and an attainable mmHG score $s$ for which we want to calculate the $p$-value, for each $1 \leq b \leq N$ and for each $1 \leq n_1 \leq N$, let $n_2(b, n_1)$ be the maximal integer $n_2$ so that if in a permutation $\pi \in S_N$, $b$ out of the first $n_2$ entries in $\pi$ are taken from the range $[1, \ldots, n_1]$, then $\pi$ satisfies $HGT_{\pi}(N, n_1, n_2, b) \leq s$. Monotonicity properties of the hyper-geometric distribution imply the existence of such $n_2$ integers. By definition, they are constants and independent of the original permutation for which the mmHG score $s$ was obtained. Due to monotonicity properties, given $b$ and $n_1$, the maximal value $n_2(b, n_1)$ can be calculated efficiently using binary search, which means that an upper bound that requires $O(N^3 \log N)$ calculations of HGT (instead of $O(N^3)$) can be computed by using the following formula:

$$mmHG \ p\text{-value}(s, N) \leq \sum_{b, n_1, n_2(b, n_1): HGT(N, n_1, n_2(b, n_1), b) \leq s} \left( \begin{array}{c} n_1 \\ b \end{array} \right) \left( \begin{array}{c} N-n_1 \\ n_2(b, n_1)-b \end{array} \right) \left( \begin{array}{c} N \\ n_2(b, n_1) \end{array} \right)^{-1}$$
The performance of this bound, as well as of other bounds (in terms of tightness and running time), is demonstrated in Table 1. On average, this bound was 16.5 times tighter than the Bonferroni bound; B3 was approximately 7 times tighter than Bonferroni’s bound, while B2 was 38 times tighter than Bonferroni’s, on average. The average computation time for B4 was 3 minutes, in comparison with 1 second for B3 and 26 minutes for B2. We conclude that the bound B4 presented in this section may be a good compromise between tightness and computational cost compared with the other bounds introduced in this paper.

### Application in PWM motif search

In this section we discuss mmHG as a framework for assessing the significance of PWM motifs in ranked lists. Given a ranked list of sequences and a PWM motif, by using the mmHG statistics and the bounds introduced earlier, we can

<table>
<thead>
<tr>
<th>Protein</th>
<th>N</th>
<th>mmHG score</th>
<th>Bound B2</th>
<th>Bound B3</th>
<th>Bound B4</th>
<th>Bonferroni bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNWMNG</td>
<td>500</td>
<td>2.17e-18</td>
<td>2.75e-14</td>
<td>7.5e-14</td>
<td>6.28e-14</td>
<td>5.43e-13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.274 min</td>
<td>0.0028 min</td>
<td>0.079 min</td>
<td></td>
</tr>
<tr>
<td>CTNNNAT</td>
<td>500</td>
<td>2.85e-27</td>
<td>1.32e-28</td>
<td>3.66e-28</td>
<td>2.37e-28</td>
<td>2.86e-27</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.155 min</td>
<td>0.0029 min</td>
<td>0.059 min</td>
<td></td>
</tr>
<tr>
<td>MMMMMMMM</td>
<td>500</td>
<td>1.08e-43</td>
<td>1.07e-39</td>
<td>3.47e-39</td>
<td>1.69e-39</td>
<td>2.71e-38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.104 min</td>
<td>0.003 min</td>
<td>0.048 min</td>
<td></td>
</tr>
<tr>
<td>REB1</td>
<td>4000</td>
<td>1.66e-137</td>
<td>9.18e-133</td>
<td>1.19e-131</td>
<td>1.54e-132</td>
<td>1.67e-131</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>17.25 min</td>
<td>0.04 min</td>
<td>2.753 min</td>
<td></td>
</tr>
<tr>
<td>CBF1</td>
<td>4000</td>
<td>1.95e-80</td>
<td>9.15e-76</td>
<td>4.62e-75</td>
<td>1.84e-75</td>
<td>1.96e-74</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>26.05 min</td>
<td>0.03 min</td>
<td>3.409 min</td>
<td></td>
</tr>
<tr>
<td>UME6</td>
<td>4000</td>
<td>5.42e-88</td>
<td>2.62e-83</td>
<td>3.04e-82</td>
<td>5.11e-83</td>
<td>5.43e-82</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>23.81 min</td>
<td>0.03 min</td>
<td>3.374 min</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>34.25 min</td>
<td>0.02 min</td>
<td>4.05 min</td>
<td></td>
</tr>
<tr>
<td>GCN4</td>
<td>4000</td>
<td>2.04e-50</td>
<td>7.66e-46</td>
<td>4.62e-45</td>
<td>1.80e-45</td>
<td>2.04e-44</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>35.43 min</td>
<td>0.03 min</td>
<td>3.95 min</td>
<td></td>
</tr>
<tr>
<td>Puif5</td>
<td>4795</td>
<td>7.91e-85</td>
<td>3.38e-80</td>
<td>5.60e-79</td>
<td>6.95e-80</td>
<td>7.93e-79</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>31.51 min</td>
<td>0.027 min</td>
<td>4.51 min</td>
<td></td>
</tr>
<tr>
<td>Pub1</td>
<td>4251</td>
<td>1.49e-84</td>
<td>6.66e-80</td>
<td>1.33e-78</td>
<td>1.37e-79</td>
<td>1.5e-78</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>27.74 min</td>
<td>0.033 min</td>
<td>3.81 min</td>
<td></td>
</tr>
<tr>
<td>Pab1</td>
<td>4142</td>
<td>2.46e-11</td>
<td>3.57e-7</td>
<td>5.17e-7</td>
<td>1.37e-6</td>
<td>2.46e-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>48.46 min</td>
<td>0.007 min</td>
<td>5.41 min</td>
<td></td>
</tr>
<tr>
<td>Khd1</td>
<td>4773</td>
<td>2.74e-20</td>
<td>5.09e-16</td>
<td>1.46e-14</td>
<td>1.73e-15</td>
<td>2.74e-14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>47.58 min</td>
<td>0.015 min</td>
<td>5.84 min</td>
<td></td>
</tr>
<tr>
<td>Nab2</td>
<td>4101</td>
<td>2.09e-11</td>
<td>3.08e-7</td>
<td>1.48e-5</td>
<td>1.18e-6</td>
<td>2.09e-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>48.7 min</td>
<td>0.016 min</td>
<td>5.34 min</td>
<td></td>
</tr>
<tr>
<td>Vts1</td>
<td>1787</td>
<td>1.44e-10</td>
<td>4.74e-6</td>
<td>1.33e-5</td>
<td>1.4e-5</td>
<td>1.45e-4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>21.94 min</td>
<td>0.003 min</td>
<td>2.07 min</td>
<td></td>
</tr>
<tr>
<td>Pin4</td>
<td>4261</td>
<td>8.16e-14</td>
<td>1.32e-9</td>
<td>8.08e-9</td>
<td>4.83e-9</td>
<td>8.18e-8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>49.38 min</td>
<td>0.011 min</td>
<td>5.48 min</td>
<td></td>
</tr>
<tr>
<td>Nrd1</td>
<td>3947</td>
<td>5.72e-12</td>
<td>9.09e-8</td>
<td>5.71e-6</td>
<td>3.36e-7</td>
<td>5.74e-6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>47.67 min</td>
<td>0.014 min</td>
<td>5.11 min</td>
<td></td>
</tr>
<tr>
<td>Yll032c</td>
<td>2286</td>
<td>1.06e-9</td>
<td>2.62e-5</td>
<td>1.61e-4</td>
<td>8.3e-5</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>35.58 min</td>
<td>0.003 min</td>
<td>2.77 min</td>
<td></td>
</tr>
</tbody>
</table>

Four bounds are compared over 17 datasets (3 synthetic and 14 biological). For each dataset, the number of sequences (N) and the mmHG score are indicated, together with the performance of each bound (in terms of tightness and running time).
assign a $p$-value to represent the significance of that PWM being enriched at the top of the list. To apply this approach for de-novo motif search, one needs to theoretically consider all possible PWMs. However, the search space - when considering position weight matrix motifs – is huge. Assuming the probabilities in the matrix are multiples of 0.1 and the alphabet is of size 4, there are 286 possible candidate PWMs of length $k$ (since each column must sum to 1, the number of combinations in each column of the matrix is equal to the number of integer solutions for the equation $X_1 + X_2 + X_3 + X_4 = 10$, which is $\binom{13}{10}$). Our approach to navigating in this search space is to narrow the search using the IUPAC alphabet, which considers all possible combinations of letters in the alphabet, and then represent the motif as a PWM based on its actual occurrences in the list. This heuristic approach, called mmHG-Finder, takes as input a ranked list of DNA or RNA sequences and returns significant motifs in PWM format. In cases where sequence ranking is not relevant or not available, it allows the use of positive and negative sets of sequences, searching for enriched motifs in the positive set using the negative set as the background.

We next describe the methodology implemented in mmHG-Finder. The input consists of a ranked list of sequences (or, alternatively, two sets of sequences representing target and background), as well as the motif width, given as a range $[k_1, k_2]$.

The algorithm:

1. Build a generalized suffix tree for the input sequences.
2. For each $k=k_1,...,k_2$
   - Traverse the tree to find all $k$-mers
   - Sort the $k$-mers according to their enrichment at the top of the list (this is done using the mHG statistics), as explained in Leibovich et al. [8].
   - Take the most significant fifty $k$-mers, to be used as starting points for the next step. This set of candidates is chosen such that the members are quite different. Note that this is a heuristic approach and the number 50 is somewhat arbitrary, chosen to succeed in catching the best performing PWMs without heavily paying in complexity.
   - For each starting point, we iteratively replace one position in the $k$-mer by considering all possible IUPAC replacements and taking the one that improves the enrichment the most. We repeat this process for all positions several times, and eventually we get a motif in the IUPAC alphabet. We note that given an IUPAC pattern $P$, the occurrences of $P$ in the list are extracted efficiently by traversing the paths in the suffix tree that agree with $P$.
   - Each IUPAC word is then expanded through a heuristic approach which is based on the Hamming neighbors of that word. Hamming neighbors are added as long as the new addition improves the enrichment $p$-value of the set of words, and as long as the overall similarity between the members in the set does not decrease below a similarity threshold. Since the neighbors are defined as exact words, they usually help in fine-tuning the correct weights of each letter in each position of the resulting PWM. Finally, the expanded motif is converted to a PWM.

3. The PWMs found in the previous step are assessed using the mmHG statistics and the best PWMs are returned as output, together with their $p$-value. The score assigned by a PWM to a string $S$ is the maximal score obtained for a substring of $S$. To obtain the likelihood of a substring of length $k$ (where $k$ is the PWM width), we simply multiply the scores assigned to each letter in each of the positions in that substring.

We provide an efficient implementation of the algorithm described above as publicly available software. Our application takes as input a ranked list of sequences and returns significant PWM motifs. It is compatible with all operating systems and can be freely downloaded from http://bioinfo.cs.technion.ac.il/people/zohar/mmHG-Finder-code/.

To evaluate the performance of mmHG-Finder in comparison to other state-of-the-art methods we ran it on 18 datasets – 3 synthetically generated datasets and 15 generated from high throughput binding experiments (6 transcription factors and 9 RNA-binding proteins). Each synthetic dataset consisted of 500 randomly drawn sequences of length 100. Then, variants of a predefined IUPAC motif were planted at the top 64 sequences of the dataset. We compared the motifs found by mmHG-Finder to those obtained by using three other methods: the standard MEME program [28], DREME [29], and XXmotif [30]. Selected results of this comparison are summarized in Figure 2, and the full output is shown in Additional file 1. Evidently, mmHG-Finder outperformed all the other three tools on the synthetic examples, which contained degenerate motifs. DREME didn’t find the motifs in any case, while MEME and XXmotif found a somewhat similar result in 1 out of the 3 tests. The other 15 examples were taken from DNA and RNA high-throughput experiments [31-33]. For 12 out of these 15 datasets, mmHG-Finder found the motifs which were compatible with the known literature motifs, and as the most significant result. In comparison, DREME found the known consensus in 11 cases; XXmotif detected the literature motif in 9 cases while MEME detected the known motif in 8 cases. In several datasets, such as for Pin4, mmHG-Finder identified the consensus
motif while other tools returned repetitive sequences as their top results. The mmHG statistics avoids such spurious results as they typically do not correlate with the measurement driven ranking.

### PWM motif search in long-non-coding RNA sequences

We further analysed a collection of datasets comprising human long-non-coding (lnc) RNAs. LncRNA sequences were extracted and ranked according to the data reported by Cabili et al. [34]. Specifically, a stringent lncRNA set of 4662 loci was tested, where for each locus we know the expression levels in 19 different tissues. From these data we generated 19 lists, each ranked according to tissue-specificity. Given locus $i$ and tissue $j$, the tissue specificity score is defined as the difference between the expression of locus $i$ in tissue $j$ (denoted $\text{exp}_{ij}$) and the mean expression of locus $i$ (denoted as $\mu_i$). That said difference is measured in terms of the standard deviation of expression in locus $i$ (denoted as $\sigma_i$). Formally:

$$\text{tissue specificity score}_{ij} = \frac{\text{exp}_{ij} - \mu_i}{\sigma_i}$$

Calculating the above measure for all tissues reported in [34] yielded 19 ranked lists comprising 4360 lncRNAs (302 loci having standard deviation equal to zero were removed from the analysis). We then conducted three enrichment tests for each of these lists:

1. We searched for de-novo PWM motifs in the promoter sequences of the tissue-specific lncRNAs using mmHG-Finder (introduced in the previous section). Promoter sequences were defined as 1000 bp upstream the transcription start site.
2. We scanned the promoter sequences of the tissue-specific lncRNAs with PWMs corresponding to known transcription factors, downloaded from the JASPAR database [35].
3. Independently of sequence, we calculated the statistical enrichment of measured transcription

---

**Figure 2** Comparison between mmHG-Finder and other motif discovery tools. We evaluated the performance of mmHG-Finder in comparison to other state-of-the-art methods: MEME, DREME and XXmotif. Almost all input examples consisted of ranked lists, except for p53 (comprising target and background sets). Since MEME, DREME, and XXmotif expect to get a target set as input, we converted the ranked lists into target sets by taking the top 100 sequences for MEME (restricted by MEME’s limitation of 60,000 characters) and the top 20% sequences for the other tools. In the synthetic examples the entire ranked lists were taken as they are sufficiently small (to reflect useful comparison with MEME, as the motif is planted in top sequences, we had provided MEME, as input, with the ranking information by adding weights to the sequences, decreasing from 1 to 0 proportionally with the ranking). We used the default parameters in all comparison to other tools (e.g. zero-or-one-occurrence per sequence in MEME) and defined the expected motif length as the range $6$ to $8$ where possible (specifically, DREME and XXmotif do not have an input parameter for the motif length). Data and consensus motifs for p53 were taken from [31]; for REB1, CBF1, UME6, TYE7, GCN4 from [32]; and for the RNA binding proteins from [33]. Selected results are shown.
factor binding events within our lists of loci. Transcription factor binding events within IncRNAs were downloaded from ChIP-Base database [36], which aggregates high-throughput sequencing data taken from hundreds of ChIP-Seq experiments.

Interestingly, almost all the motifs returned by mmHG-Finder were GC-rich (Figure 3). In all three tests, the most significant results were obtained for thyroid-specific and prostate-specific IncRNAs. We further checked whether GC rich sequences are generally enriched amongst the promoter sequences of tissue specific IncRNAs by calculating the mutual enrichment between these two measures. The mutual enrichment between GC content and tissue specificity (Table 2) was the most significant for thyroid (mmHG $p$-value $\leq 3.9 \times 10^{-31}$), prostate ($5.8 \times 10^{-22}$), adrenal ($5.5 \times 10^{-20}$), brain ($1.6 \times 10^{-13}$) and ovary ($8.8 \times 10^{-12}$). Interestingly, Pearson's correlation between the GC content and the sequence rank was not observed to be strong (strongest correlation coefficient was -0.1), demonstrating that the overall agreement between two measures can be weak even when extremities agree.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>mmHG-Finder output</th>
<th>Transcription factors having similar recognition sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thyroid</td>
<td>6.48e-37</td>
<td>E2F3, E2F2, Zfp161, Zfx, SP1, Egr1, Bcl6b, Kl7, Sp4</td>
</tr>
<tr>
<td>Prostate</td>
<td>8.99e-23</td>
<td>Bcl6b, Egr1, Smad3, SP1, Nr2f2, Zfp410, Matb, Zfx, Zfp740</td>
</tr>
<tr>
<td>Brain</td>
<td>1.54e-18</td>
<td>Zfp161, E2F3, TFAP2A, E2F2, Egr1, SP1, Myc, Sp4</td>
</tr>
<tr>
<td>Ovary</td>
<td>2.11e-16</td>
<td>Egr1, Nr2f2, Plag11, Bcl6b, Smad3, SP1, Zfx, Zfp740</td>
</tr>
<tr>
<td>Foreskin</td>
<td>7.37e-16</td>
<td>Zfx, Nr2f2, Egr1, SP1, Zfp161, TFAP2A, Smad3, Bcl6b, Sp100, Zic1</td>
</tr>
<tr>
<td>Kidney</td>
<td>8.01e-11</td>
<td>Egr1, SP1, Sp4, Klf7, Zfp281, CTCF, INSM1, Zfp740</td>
</tr>
<tr>
<td>Breast</td>
<td>2.52e-10</td>
<td>Egr1, Nr2f2, Zfx, TFAP2A, Zfp161, Zic1, Plag11, Zic2, Tcfap2a</td>
</tr>
<tr>
<td>Adipose</td>
<td>2.79e-10</td>
<td>Egr1, Tcfap2b, Plag11, SP1, NHLH1, INSM1, E2F2, Smad3, Sp4</td>
</tr>
<tr>
<td>Adrenal</td>
<td>3.78e-10</td>
<td>E2F3, E2F2, Myf6, Nr2f2, Plag11, Sp4, Bcl6b, Smad3, CTCF, Zfp161</td>
</tr>
<tr>
<td>Lymph node</td>
<td>2.01e-7</td>
<td>Smad3, Sp4, Zfp161, E2F2, E2F3, Glis2, INSM1, Egr1, SP1, Zfp740</td>
</tr>
<tr>
<td>Testes</td>
<td>9.55e-6</td>
<td>Zfp410, Fox11, Gm397, Six2, Sox30</td>
</tr>
<tr>
<td>Liver</td>
<td>Nothing found</td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>Nothing found</td>
<td></td>
</tr>
<tr>
<td>White blood cell</td>
<td>Nothing found</td>
<td></td>
</tr>
<tr>
<td>Colon</td>
<td>Nothing found</td>
<td></td>
</tr>
<tr>
<td>Heart</td>
<td>Nothing found</td>
<td></td>
</tr>
<tr>
<td>Skeletal muscle</td>
<td>Nothing found</td>
<td></td>
</tr>
<tr>
<td>Placenta</td>
<td>Nothing found</td>
<td></td>
</tr>
<tr>
<td>Lung fibroblasts</td>
<td>Nothing found</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3 Motifs in tissue-specific IncRNA promoter sequences. We analysed the promoter sequences of IncRNAs that are ranked according tissue-specificity. The motifs returned by mmHG-Finder are shown in the figure together with their $p$-value. We compared those motifs to known consensus motifs of transcription factors using TOMTOM [37] (motif database = JASPAR Vertebrates and UniPROBE Mouse) and the most significant results are shown (specifically, all similarity $p$-values are better than 0.018).
Furthermore, by intersecting the results of the second and the third tests together, we identified transcription factors that may regulate lncRNAs, mainly in thyroid and prostate. This set includes NRF1, E2F1, E2F3, E2F4, E2F6, EGR1, SP1, SP2 and ZBTB33. Moreover, the consensus recognition sites of EGR1, SP1 and E2F3 were found to be similar to the motifs returned by mmHG-Finder in thyroid, prostate and other tissues (Figure 3; the comparison was done using the motif discovery tool TOMTOM [37]). The full output of the second and the third tests are summarized in Additional file 2.

As GC-rich motifs may be associated with CpG methylation, and due to the possible binding of SP1 which has been suggested to protect CpG islands from de novo methylation [17,38], we further tested the association between hypo-methylation and tissue specificity (Figure 3; the comparison was done using the motif discovery tool TOMTOM [37]). The full output of the second and the third tests are summarized in Additional file 2.

We calculated the mutual enrichment between DNA hypo-methylation and tissue specificity for the lncRNA promoters. CpG methylation data was taken from UCSC Table Browser [39] (ENCODE/HAIB).

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Mutual enrichment between GC content and tissue-specificity</th>
<th>Mutual enrichment between hypo-methylation and tissue-specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal/subnormal cells</td>
<td>Cancer cells</td>
</tr>
<tr>
<td>Thyroid</td>
<td>3.89e-31</td>
<td>No methylation data</td>
</tr>
<tr>
<td>Prostate</td>
<td>5.76e-22</td>
<td>4.16e-11 (PrEC)</td>
</tr>
<tr>
<td>Adrenal</td>
<td>5.46e-20</td>
<td>No methylation data</td>
</tr>
<tr>
<td>Brain</td>
<td>1.57e-14</td>
<td>1.21e-8 (NH-A)</td>
</tr>
<tr>
<td>Ovary</td>
<td>8.80e-12</td>
<td>No methylation data</td>
</tr>
<tr>
<td>Lymph node</td>
<td>3.64e-6</td>
<td>No methylation data</td>
</tr>
<tr>
<td>Adipose</td>
<td>9.25e-6</td>
<td>No methylation data</td>
</tr>
<tr>
<td>Foreskin</td>
<td>2.25e-5</td>
<td>0.72 (BJ)</td>
</tr>
<tr>
<td>Breast</td>
<td>4.40e-5</td>
<td>5.08e-5 (HMEC)</td>
</tr>
<tr>
<td>Kidney</td>
<td>6.34e-5</td>
<td>1.56e-5 (HEK293)</td>
</tr>
<tr>
<td>White blood cell</td>
<td>3.78e-4</td>
<td>0.6 (GM12878)</td>
</tr>
<tr>
<td>Placenta</td>
<td>0.011</td>
<td>No methylation data</td>
</tr>
<tr>
<td>Colon</td>
<td>0.012</td>
<td>No methylation data</td>
</tr>
<tr>
<td>Skeletal muscle</td>
<td>0.04</td>
<td>0.34 (SKMC)</td>
</tr>
<tr>
<td>Lung</td>
<td>0.33</td>
<td>No methylation data</td>
</tr>
<tr>
<td>Heart</td>
<td>1.0</td>
<td>1.0 (HCM)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.0 (HCF)</td>
</tr>
<tr>
<td>Liver</td>
<td>1.0</td>
<td>1.0 (Hepatocytes)</td>
</tr>
<tr>
<td>Testes</td>
<td>1.0</td>
<td>No methylation data</td>
</tr>
<tr>
<td>Lung fibroblasts</td>
<td>1.0</td>
<td>1.0 (IMR90)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.0 (AGO4450)</td>
</tr>
</tbody>
</table>

We calculated the mutual enrichment between DNA hypo-methylation and tissue specificity for the lncRNA promoters. CpG methylation data was taken from UCSC Table Browser [39] (ENCODE/HAIB).
Discussion
The assessment of mutual enrichment in ranked lists is often required to support the analysis of biological measurement data, such as in the case of identifying sequence motifs that are involved in regulation processes. Relative ranking can be represented by using permutations over the measured elements. Therefore – the statistical assessment of mutual enrichment can be modelled by characterizing properties of random permutations. Due to the size of the measure space, statistics over $S_N$, the group of permutations over $N$ elements, is difficult to perform and implement. Mutual enrichment is more informative from the point of view of practical biological science than simple correlation measures, as it focuses on the top of the lists and not on the overall agreement, which may be weak even in cases where extremities agree. In this work we derive polynomially computable bounds for the associated tail distribution of mutual enrichment in ranked lists. Namely – we provide methods for computing an upper bound on the $p$-value of mutual enrichment at the top of two permutations uniformly and independently drawn over $S_N$. Naïve approaches to computing such bounds include variants of the Bonferroni approach. These do not provide tight bounds and may lead to mis-labeling results as non-significant. For several representative datasets, we note that our bound improves the Bonferroni derived $p$-value estimates by a factor of almost 40, on average. Nevertheless, these improvements become relevant only for high $p$-values - for which significant scores should be treated with care anyway. We therefore note that the Bonferroni correction is applicable in many cases, as demonstrated in Table 1. Using our bounds is highly beneficial in borderline cases but is also important in cases where an accurate estimate of the $p$-value is desired, even if nuances do not affect the final biological research conclusions.

We use our statistical/algorithmic framework to support PWM motif searches and demonstrate the application to biological data. We identify motifs that characterize tissue specificity of IncRNA in thyroid and in prostate. Specifically, we find the EGR1 binding motif to be enriched in the promoter regions of IncRNAs which are thyroid-specific. EGR1 was observed to be highly expressed in thyroid (Additional file 1, taken from [36]), consistent with our stronger motif findings. Similarly, EGR1 is highly expressed in adipose tissue and its transcription factor binding sites are enriched in IncRNAs specific to this tissue. We do not have methylation data for the latter two tissues types. However – we do observe the promoters of IncRNAs that are specific to breast to have enriched occurrences of motifs that are similar to EGR1 transcription factor binding sites ($p$-value of similarity according to TOMTOM = $3.52 \times 10^{-5}$). EGR1 is also highly expressed in breast. Finally, the promoters of IncRNAs that are specific to breast are less methylated in breast (MCF10A and HMEC cells) than other promoters. This suggests the role of EGR1 in driving tissue differentiation by transcribing tissue-specific IncRNAs and by protecting the associated promoters from methylation. EGR1 has been previously shown to recognize GC-rich consensus sequences located in CpG island promoters of active genes [42]. Generally, we observed that tissue-specific IncRNA promoters tend to be less methylated than those of non-tissue-specific IncRNAs in prostate, brain, ovary, breast and kidney, which may be associated with the GC-rich patterns enriched among their tissue-specific IncRNA promoter sequences.

Threshold-free alternatives to mmHG include the work of McLeay and Bailey, in which a linear regression method is applied [43]. It was shown to achieve high accuracy on a benchmark comprising 237 ChIP-chip datasets, which was higher than all other data driven methods tested, and specifically higher than Spearman’s rank correlation. We note that applying linear regression or Spearman correlation to PWM motif search in ranked lists requires that for significant motifs we observe an overall agreement between the biological measurement and the PWM score. Nevertheless, the standard PWM formulation fails to predict binding affinity when the latter decreases to the point of non-specific binding [44]. In other words, the overall agreement between the PWM score and the binding affinity may be relatively weak. High correlation between the PWM score and the binding affinity needs to hold, in effect, only for sequences demonstrating high-binding affinity with respect to the protein of interest (that is, for sequences that are located at the top of the list) [45]. This weaker relationship is naturally addressed by the mmHG statistics. A combination of mmHG and a linear model, such as suggested in [43], applied to strong binders (top of the list), may yield an even more faithful and informative model.

Future research directions include more extensive application to biological data and the development of tighter and more efficient bounds. Our results show promise in enabling efficient and user-friendly PWM motif search in ranked lists. The software is freely available at http://bioinfo.cs.technion.ac.il/people/zohar/mmHG-Finder-code/. Finally, the full characterization of the distribution of mmHG as a random variable over $S_N$ remains an open question.

Conclusions
In this work we developed tight bounds on the tail distribution of mutual enrichment in ranked lists. Our bounds are computable in polynomial time and potentially add to the accuracy of reported results. We demonstrated the utility of mutual enrichment in motif search – specifically, when searching position weight matrix motifs in ranked lists, where the ranking can be according to binding affinity or according to any other biological measurement.
Additionally, we used mutual enrichment to study tissue-specific long non-coding RNA regulation, and suggest that tissue-specific IncRNAs are regulated through GC-rich elements located on their promoters, in several tissue types. We hypothesize that these GC-rich patterns are associated with DNA hypo-methylation.

### Additional files

**Additional file 1: Table S1.** Comparison between mmhK-GFinder and other motif discovery tools. Figure S1. EGR1 expression profile.

**Additional file 2: Table S2.** The full output of the second and the third tests (including their intersection) for tissue-specific IncRNAs.

### Competing interests

The authors declare that they have no competing interests.

### Authors’ contributions

LL derived the bounds, developed software and performed analysis. Both authors developed the PWM scoring approach, designed the study and wrote the manuscript. Both authors read and approved the final manuscript.

###Acknowledgements

We thank Israel Steinfeld for critical and inspiring discussions. We thank Roy Navon for technical help with the software download service. We also thank the WABI anonymous reviewers for their useful comments. LL was partially supported by Israel Ministry of Science and Technology and by ISF Fellowship.

**Received:** 1 December 2013 **Accepted:** 30 March 2014

**Published:** 5 April 2014

### References


10. DRIMust Webserver. [http://drimust.technion.ac.il/](http://drimust.technion.ac.il/)


12. miTEA Webserver. [http://cbl-gorilla.cs.technion.ac.il/miTEA/](http://cbl-gorilla.cs.technion.ac.il/miTEA/)


34. DRMUST Webserver. [http://drimust.technion.ac.il/](http://drimust.technion.ac.il/)


39. UCSC Table Browser. [http://genome.ucsc.edu/cgi-bin/hgTables?command=start]


4 Discussion

This section is organized into sub-sections according to the contribution of our work. We discuss each contribution and then describe the open questions and future work. We start with describing our contribution to motif discovery in ranked lists. We then elaborate on our contribution to statistics in ranked lists, and finally we focus on biological results found using our methods. We address significance, advantages and limitations of our approaches in each sub-section.

4.1 Motif discovery and molecular measurement data

The important role played by sequence elements in molecular regulation and signaling is the motivation for significant scientific and technological development activity that focuses on measuring sequence based recognition and on computational approaches and analysis tools designed to improve our understanding of regulation mechanisms involving sequence elements.

Techniques such as ChIP-chip [86], ChIP-PET [87] and ChIP-seq [12] provide large volumes of genome-wide data on regions of transcription factor binding, measured in actual samples and in various conditions. Similarly, mRNA targets of RNA binding proteins are studied using techniques like RIP-chip [88] and CLIP [14]. SILAC [89] and other proteomic techniques can be used to characterize the effect of amino acid sequences on protein function. Computational tools and approaches to motif discovery form part of the data analysis workflow that is used to extract knowledge and understanding from data generated through the above techniques as well as other measurement approaches.

Motif discovery has attracted much research interest in the last decade, resulting in more than a hundred different tools for motif discovery [90]. A large subset of motif-finders such as MEME [91], NMica [92], AlignACE [93] or MDscan [94] fit position weight matrices (PWMs) to the sequence data. Most techniques, including those mentioned above, approach motif finding by discovering sequence elements that occur more often than expected in a set of sequences. Some techniques compare a target set to a background set. It is often the case, however, in biological measurement data that results are given as a ranked list of quantities. For example, Gerber et al. [1] report on the set of targets for five RNA binding proteins from the Pumilio family, including binding ratios for each *S. cerevisiae* gene. This is also the case for expression profiling studies as
well as for all the aforementioned ChIP techniques, whether based on microarrays or on sequencing. Statistical approaches such as GSEA [95] and mHG [29-31] address enrichment in ranked lists of elements. Based on the mHG statistics, our motif finding tools exploits the ranking derived from experimental measurements to discover motifs that are rank imbalanced in the input list.

State of the art motif finding tools come short in their ability to address searches over large motif spaces. An example that best demonstrates this shortcoming is that of variable gap motifs – consisting of two half sites separated with a gap that may vary in length. TP53 binding to DNA, for example, has flexibility that recognizes this type of sequence motifs. Searching for variable gapped motifs poses a significant computational challenge as the naïve approach to the motif search spaces quickly faces intractable complexity. The efficiency of our approach, based on suffix trees, allows searches over motif spaces that are not covered by existing tools. This includes searching variable gap motifs and searching long motifs over large alphabets, as we further explain in the following sections.

4.1.1 Efficient enumeration of the motif search space

To address the discovery of long motifs, or to deal with large alphabets (as in the case of proteins), we cannot take a naïve approach that uses exhaustive search over the motif space. Our approach to overcoming the computational challenges associated with large motif search spaces is based on using suffix trees, to restrict our attention to motifs that actually occur in the input list of sequences [28]. A suffix tree is a data structure that represents all the suffixes of a given string in a way that allows fast implementation of many string operations, including motif search. Suffix trees are useful in many other application contexts in bioinformatics and computational biology [80, 81]. Our algorithm uses a suffix trees for an efficient enumeration of motif candidates, which are then assessed using the mHG statistics. The occurrences of each candidate motif in the list are extracted from the extra information stored in the leaves of the tree. The returned output comprises substrings that appear more often at the top of the list compared with the remainder of the list, where the cutoff defining the top of the list is determined by the mHG statistics in a data-driven manner.
Our algorithm is implemented as stand-alone software, and has several advantages over existing methods. First, unlike many other approaches, it does not exhaustively search over motif spaces and therefore can detect long motifs and motifs over large alphabets. The efficiency of our approach enables us to search motifs over large alphabets as well as motifs of length 20 or more, all in a reasonable time. Specifically, searching for DNA motifs of length 4-20 in *S. cerevisiae*, over 6000 sequences (each of length 500 bp), takes less than 3 minutes on a standard PC. Furthermore, we search motifs in ranked lists and not in fixed sets of sequences as is the case for many other methods.

4.1.2 Discovery of variable gapped motifs

While most motif finding approaches consider continuous sequence elements, it is of interest to also consider gapped sequence elements. For example, GAL4 in *S. cerevisiae* binds DNA as a homodimer. Its binding site comprises 17 bp, containing palindromic CGG triplets at the ends that are separated by an 11 bp gap [16]. Additionally, Puf2p in *S. cerevisiae* binds 3'UTR of mRNAs by recognizing a motif of two UAAU tetranucleotides separated by a 3 nucleotides linker sequence [17]. Therefore, an interesting case, related to the role of gapped motifs, is of a protein that binds the DNA or RNA as a dimer. There are also cases where recognition is based on sequence elements with variable length gaps separating the half sites. An important example is the recognition site of the tumor protein TP53. In many organisms this site is composed of two copies of the half-site RRRGWWGYYY separated by a spacer, usually of length 0–21 bp [18].

The search for variable gapped motifs (VGMs) poses a tremendous computational challenge, as the search space becomes huge, when considering parameters of biological relevance. Specifically, if we seek DNA motifs containing two half sites, each of length 4, where the gap between the 4-mer half sites can be any subset of the numbers \{0, … , 10\}, then an exhaustive search will span \(4^4 \times 2^{11}\) candidates, which is far too large to routinely address in reasonable time.

We developed an efficient algorithmic approach to searching variable gapped motifs in ranked list of sequences [28]. The efficiency of our approach relies on testing only patterns that occur in the dataset, which can be done using suffix trees. Our method is unique in efficiently addressing variable gap motifs under a definition that allows full flexibility of the gap. We are not aware of
any other method that can efficiently search variable gap motifs while allowing full flexibility in the gap model. Specifically, searching, in *S. cerevisiae*, for variable gapped motifs comprising two 4-mer half sites separated by gaps that form any subset of the lengths \{0,\ldots,10\}, where the input comprises 6000 DNA sequences (each of length 500 bp), takes less than 7 minutes on a standard PC.

Our approach to overcoming the computational challenges associated with large motif search spaces is based on using suffix trees, to restrict our attention to motifs that actually occur in the input list of sequences. An alternative approach could be based on the use of a hash table. Taking this approach, however, will entail running the search for a fixed motif length \(k\). To span a range of lengths \(k_1 \leq k \leq k_2\), as is done by our tool, one would need to generate a separate hash table for each \(k\). Therefore, a suffix tree approach is far more efficient in this case. In addition, the hash table approach falls short of solving the variable gap motif problem which is enabled through the use of suffix trees.

One apparent limitation of the current implementation is that the output half-sites are exact words. That is, they are words over the alphabet of the input sequences, allowing no flexibility or weighting. Our output can be viewed as the starting point for extending the motifs using more flexible motif representation approaches for selected top results. In the next two sections, we describe two methods for PWM extension, given an exact word. However, in the variable gap motif problem, for this extension step, we currently employ ad-hoc and manual analysis. We added a simple extension mechanism of a motif in the tool’s output by considering its variants that occur at the top of the ranked list for creating a PWM for the motif. The implementation of this extension is also available for download. Future work is to further develop algorithmic approaches to this phase. For example, extend motifs using IUPAC or PWMs. The IUPAC extension task is amenable to modifications of the suffix tree approach.

We note that all \(p\)-values reported by our tool are not corrected for the size of the motif search space, which should be handled by the user. This can be done using Bonferroni’s correction, by multiplying the reported \(p\)-value by \(4^k\), where \(k\) indicates the input length parameter, assuming an alphabet of size 4 (as in the case of DNA and RNA). When searching variable gap motifs, the \(p\)-value should be multiplied by \(4^{a+b+2l+1}\), where \(a\) and \(b\) indicate the half site lengths, respectively, and \(l\) indicates the maximal gap length.
Our work is the first to perform a systematic analysis of variable gapped motifs in biological datasets. We elaborate on biological results in Section 4.3.2.2. Briefly, we have found variable gapped motifs to be more significant in a handful of cases and therefore reach the preliminary conclusion, consistent with Reid et al. [27], that they are important but not extremely widespread.

In summary, we developed efficient and effective algorithmic approach to searching motifs in ranked list of sequences. Our method is unique in combining an efficient search with a ranked list approach and rigorous $p$-value estimation. It is also unique in efficiently addressing variable gap motifs under a definition that allows full flexibility of the gap.

4.1.3 Software implementation and the DRIMust webserver

We provide an application on the Yakhini Group website that takes as input ranked lists of sequences and returns significant variable gap motifs. This application is freely accessible to the community. Software is available at http://bioinfo.cs.technion.ac.il/people/zohar/DRIMUST-code-VGM/.

One apparent limitation of the software implementation pointed above is the lack of rigorous extension of exact words into a more flexible format, such as position weight matrix representation. A more extensive attempt to extend exact words into position weight matrices is implemented in DRIMust. DRIMust is a web accessible software implementation of our statistical and algorithmic approach described above, developed to enable efficient motif searches [49]. In particular, DRIMust can efficiently search for long motifs and for motifs over large alphabets. Variable gapped motifs are not covered in the current implementation of DRIMust.

In DRIMust, an initial motif search phase produces $k$-mers, which are words over the alphabet of the input sequences (of length $k$). These candidate $k$-mers are derived by enumerating paths of length $k$ in the generalized suffix tree generated for the input sequences. Next, the statistical significance of the $k$-mers is calculated using the mHG statistics. In the next stage, the promising $k$-mers are extended to produce PWMs. Extension is obtained by a heuristic approach based on the Hamming neighbors of the best 50 exact motifs. Briefly, starting from a single $k$-mer, Hamming neighbors (of length $k$) are added to a set of motifs as long as the observed enrichment
p-value is improved. DRIMust is freely accessible through the website: 
http://drimust.technion.ac.il.

To evaluate the performance of DRIMust in comparison to other state-of-the-art methods we ran DRIMust on 24 examples generated from high throughput binding experiments and compared the results to those obtained by using three other methods: the standard MEME program [91]; the DREME program [96] from the MEME suite which was optimized for fast analysis of very large datasets; and to XXmotif [97], a recent web server which was designed for efficient extraction of PWMs from large datasets. DRIMust demonstrated the highest accuracy by identifying the literature motifs for 92% of the test examples (as opposed to 80% successes at most for the competitors). Notably, DRIMust found the motifs that were compatible with the literature motifs as the most significant result, while in the other methods tested, the known motifs were not always reported as the best motif. Additionally, in all cases DRIMust completed the computations significantly faster than the other tools. The longest job took 1 minute and 21 seconds on DRIMust (for a dataset containing 9995 sequences, each of length 100 nucleotides; single-strand search mode). The performance of DRIMust in terms of running time is shown in Figure 9.

Figure 9 – Average running time of DRIMust compared to three other motif discovery tools

We compared DRIMust with three state of the art motif discovery tools over a benchmark including 24 datasets. The comparison is split into DNA search mode – where both strands are considered, and RNA search mode which is a single strand search and is therefore faster. Evidently, the average running time of DRIMust was significantly shorter than all other tools on the DNA datasets, as well as on the RNA datasets.

In summary, DRIMust runs very efficiently. An average job for datasets containing 4000 DNA sequences, total 2,000,000 characters, takes 1 minute and 10 seconds when double-strand search
mode is used and 15 seconds when single-strand search mode is used. DRIMust allows for timely interaction with the results, through a friendly interface and a clear output format. Most importantly, by working with ranked lists DRIMust avoids the arbitrary designation of fixed sets of sequences and exploits the ranking derived from experimental measurements. DRIMust utilizes the ranking to discriminate true motifs from other irrelevant sequence elements (such as AT repetitive elements that are abundant in 3’UTRs), since the latter are not correlated with the ranking and are therefore ignored by DRIMust. This explains the observed higher accuracy of DRIMust compared to the other tools in the comparison we conducted over 24 examples.

DRIMust can efficiently deal with the large datasets generated by high-throughput methods, such as ChIP-chip [86], ChIP-PET [87] and ChIP-seq [12], making it preferable for large volume data. Nevertheless, DRIMust is also useful in cases when there are clear target and background sets. In the latter scenario the enrichment is calculated using the hyper-geometric distribution.

4.1.4 Searching position weight matrix motifs in ranked lists

The main advantages of DRIMust, described above, are that it is very fast due enumerating motifs using suffix trees, as well as the rigorous statistical model which utilizes the ranking information. The main limitation of DRIMust is its inability to detect degenerate motifs if those do not have a strong exact word representative enriched in the data, as the algorithm implemented in DRIMust enumerates exact words and then expands the most promising ones into PWM motifs. To overcome this limitation, and to allow statistical assessment of PWM motifs directly, we developed mmHG-Finder. mmHG-Finder takes as input a ranked list of DNA or RNA sequences and returns significant motifs in PWM format. In cases where sequence ranking is not relevant or not available, it allows the use of positive and negative sets of sequences, searching for enriched motifs in the positive set using the negative set as the background. This application is freely accessible to the community. Software is available at http://bioinfo.cs.technion.ac.il/people/zohar/mmHG-Finder-code/.

mmHG-Finder considers two rank orders over the set of input sequences. Given a set of sequences that were tested in a high throughput experiment such as ChIP-seq [12], CLIP [14] and others, they can be ranked according to the measured binding affinities, yielding a ranked list $L_1$. Given a PWM which we want to assess, the sequences can also be ranked according to their PWM scores, yielding another ranked list $L_2$, different from $L_1$. Since usually we are interested in
finding motifs amongst sequences having strong binding affinities, a significant PWM motif would yield significant scores for sequences having strong binding affinities. Therefore, the question of PWM motif discovery from ranked experimental data can be formulated as quantifying the mutual enrichment level for the two ranked lists $L_1$ and $L_2$. The mutual enrichment of the top parts of $L_1$ and $L_2$ is calculated using the mmHG statistics, which defines the top of each list in a data-driven manner. Using tight bounds on the mmHG $p$-value that we developed in our work, each candidate PWM motif is assessed for its statistical significance.

However, the search space - when considering position weight matrix motifs – is huge. Assuming the probabilities in the matrix are multiples of 0.1 and the alphabet is of size 4, there are $286^k$ possible candidate PWMs of length $k$. Our approach to navigating in this search space was to narrow the search using the IUPAC alphabet, which considers all possible combinations of letters in the alphabet, and then represent the motif as a PWM based on its actual occurrences at the top of the list. Specifically, we take the most significant fifty $k$-mers, to be used as starting points for PWM expansion. For each starting point, we iteratively replace one position in the $k$-mer by considering all possible IUPAC replacements and taking the one that improves the enrichment the most (a sort of a hill climbing approach). We repeat this process for all positions several times. Eventually we get a motif in the IUPAC alphabet which is then expanded by adding Hamming neighbors as long as the new addition improves the observed enrichment $p$-value, and as long as the overall similarity of the members in the set does not decrease below a similarity threshold. Finally, the expanded motif is converted to a PWM, which is assessed using the mmHG statistics. The most significant PWM motifs are returned as output, together with the $p$-value.

When narrowing the search space to IUPAC words, which is a significantly smaller space (its size is $15^k$, compared to $286^k$), the true PWM motif may be different from and stronger than the best IUPAC word found. A simple example is illustrated in Figure 10. In this example, the words $\text{TGTAATA, TGTATA, AGTAAATA}$ are enriched at the top of the list. The minimal representation of these words over the IUPAC alphabet is $\text{WGTAWATA}$ ($W=\text{A/T}$). However, the word $\text{AGTATA}$ agrees with this IUPAC word, but appears only at the bottom of the list. Therefore, $\text{WGTAWATA}$ is not ranked imbalanced in the list, and the most significant IUPAC word would possibly be $\text{TGTAWATA}$, which captures only part of the pattern. Using weights
for each letter in the positions where W occurs can solve this problem. If, for example, the first position is 70% T and 30% A, while the forth position is 70% A and 30% T, then the ranking induced by this PWM places the strings TGTAATA, TGTATATA, AGTAAATA at the top, and AGTATATA is pushed down, in agreement with the original ranking. Clearly, this PWM captures the motif more faithfully. Therefore, in mmHG-Finder, we expand the most promising IUPAC words returned as output by adding Hamming neighbors as long as the enrichment p-value improves. Since these neighbors are always exact words, they usually help in fine-tuning the correct weights.

\[
\begin{align*}
\text{TGTAATA} & \\
\text{AGTAATA} & \\
\text{TGTATATA} & \\
\text{AGTAAATA} & \\
\text{AGTAAATA} & \\
\text{TGTATATA} & \\
\text{AGTAAATA} & \\
\text{TGTAAATA} & \\
\text{AGTAAATA} & \\
\text{...} & \\
\text{AGTATATA} & \\
\text{AGTATATA} & \\
\text{AGTATATA} & \\
\text{AGTATATA} & \\
\end{align*}
\]

**Figure 10 – An example demonstrating the limitation of searching an IUPAC word instead of a PWM**

The words TGTAATA, TGTATATA, AGTAAATA are enriched at the top of the list. The minimal representation of these words over the IUPAC alphabet is WGTAWATA (W=A/T). However, the word AGTATATA agrees with this IUPAC word, but appears only at the bottom of the list. Therefore, WGTAWATA is not ranked imbalanced in the list. Adding weights for each letter in the positions where W occurs could solve this problem.

In summary, mmHG-Finder performs well compared to state of the art methods. It is, however, slower than DRIMust due to the more comprehensive search conducted, which demonstrates benefit in discovering degenerate patterns. Our results show promise in enabling efficient and user-friendly PWM motif search in ranked lists. Future research directions include more extensive application to biological data and more extensive coverage of PWM search space.
4.2 Statistics in ranked lists

4.2.1 Comparison between mHG and other statistical methods

The minimum hyper-geometric statistics is one of several non-parametric statistical methods that assess enrichment significance, taking as input a ranked list and a fixed set of elements and returning exact $p$-value under a uniform distribution null model. Given a ranked list of entities and a subset of entities $A$, the input can be represented using binary vectors, where for every entity in the list we put 1 in the corresponding entry in the vector if it belongs to $A$, and 0 otherwise. The question we then ask is whether there is an enrichment of ones at the top of this vector, under the null assumption that all configurations of ones in the vector are equiprobable. The “top of the list” here, should be data driven. That is – the methods we are interested in do not take a fixed “top” as input.

The Wilcoxon (Mann-Whitney) rank sum test and the Kolmogorov-Smirnov test [98] deal with this problem, each of them in a slightly different way. The null model of all three tests – mHG, Wilcoxon, and Kolmogorov-Smirnov - is the same. Specifically, they all assume a uniform distribution over $\{0,1\}^N$, which is the space of all binary vectors of length $N$ with exactly $B$ ones. The Wilcoxon test assesses the null probability that the average rank of 1 positions is very different from that of the 0 positions. The Kolmogorov-Smirnov test assesses the null probability that there is a large gap in the empirical cumulative distribution functions corresponding to 1s and 0s. The test statistics, computed for an input vector $\lambda = \lambda_1 \ldots \lambda_N \in \{0,1\}^N$ are as follows:

Wilcoxon statistic:

$$W = \sum_{i=1}^{N} i \cdot \lambda_i$$  \hspace{1cm} (13)

Asymptotic normality of $W$, suitably standardized according to $W^* = \frac{W - E_0}{\sqrt{\text{var}_0}}$, where $E_0 = \frac{B(N+1)}{2}$ and $\text{var}_0 = \frac{B(N-B)(N+1)}{12}$, enables to derive a straightforward approximation of the test $p$-value.
Kolmogorov-Smirnov statistic:

\[
J = \max_{1 \leq n \leq N} \left| \frac{nB}{N} - b_n \right| \quad (14)
\]

where \( b_n = \sum_{i=1}^{n} \lambda_i \). The statistic \( J^* = \frac{JN}{\sqrt{B(N-B)N}} \) follows asymptotically the Kolmogorov distribution. Namely:

\[
Prob(J^* > j) = \sum_{k=1}^{\infty} 2 \cdot (-1)^{k-1} \cdot e^{-2k^2 j^2} \quad (15)
\]

The question of whether there is an enrichment of ones at the top of a binary vector is therefore equivalent to conducting a one sided test of any of the three tests presented above.

One prominent difference between these methods is that mHG and Kolmogorov-Smirnov calculate a cutoff \( n^* \) which separates the list into ‘target’ and ‘background’ (namely, the position in the list where the maximum enrichment was found), while Wilcoxon does not calculate such cutoff. This cutoff is useful in many biological contexts, when one needs to define a target set – of a transcription factor of interest, for example.

Beyond this difference, we compared these tests on random binary vectors, where these vectors were randomized such that the probability of 1 is higher at the top than at the bottom. The probability of 1 decreases as we go down in the vector. Our model for testing is defined using the parameters \( p, q \) and \( \alpha \) (\( p > q \)) in the following way, where \( \lambda = \lambda_0 \ldots \lambda_{N-1} \) is the binary vector:

\[
Prob(\lambda_i = 1) = \max\{p - \alpha i, q\}, i = 0, ... N - 1 \quad (16)
\]

We used \( N=1000 \) (that is, vectors of length 1000) and randomly generated 1000 vectors in each experiment. These vectors were assessed by all three tests. The figure of merit we used in our comparison is the power of the statistical test, as the vectors are drawn from the alternative model. The power is defined as the probability that the test will reject the null hypothesis when the alternative hypothesis is true. We therefore count the number of vectors for which the null hypothesis is rejected at confidence level \( \beta_0 \). The results of the comparison are summarized in Table 1. We observed mHG to be superior to the Wilcoxon and Kolmogorov-Smirnov tests with
respect to greater statistical power when the ratio $\frac{p-q}{\alpha}$ was sufficiently small. Ratio of 100 was associated with mHG superiority, which diminished completely for ratio equal to 1000 (Table 1 and Figure 11). In other words, from the experiments we conducted so far, we generally observe that mHG has an advantage when the phenomenon of interest occurs only at the very top of the ranked list and then quickly fades. We hypothesize that this is a general property but the question of how to phrase and prove a rigorous statement that captures the essence of our observations is currently open.

Table 1 – Number of vectors (out of 1000) where the test $p$-value $\leq \beta_0$

We randomly generated 1000 binary vectors, such that the probability of 1 at the top is higher than at the bottom, and decreases as we go down in the vector. These vectors were assessed by all three tests, and the number of vectors having test $p$-value $\leq \beta_0$ is shown for different parameters of $p$, $q$ and $\alpha$. The length of the blue data bar represents the value in the cell compared to other values in that row. A longer bar represents a higher value. The full distribution of $p$-values for two specific examples (marked with asterisks) is shown in Figure 11.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$q$</th>
<th>$\alpha$</th>
<th>$\beta_0$</th>
<th>$(p-q)/\alpha$</th>
<th>Kolmogorov-Smirnov</th>
<th>mHG</th>
<th>Wilcoxon</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.3</td>
<td>0.001</td>
<td>1.E-03</td>
<td>200</td>
<td>136</td>
<td>342</td>
<td>158</td>
</tr>
<tr>
<td>0.6</td>
<td>0.4</td>
<td>0.001</td>
<td>1.E-03</td>
<td>200</td>
<td>104</td>
<td>280</td>
<td>119</td>
</tr>
<tr>
<td>0.7</td>
<td>0.5</td>
<td>0.001</td>
<td>1.E-03</td>
<td>200</td>
<td>81</td>
<td>280</td>
<td>130</td>
</tr>
<tr>
<td>0.8</td>
<td>0.6</td>
<td>0.001</td>
<td>1.E-03</td>
<td>200</td>
<td>107</td>
<td>333</td>
<td>140</td>
</tr>
<tr>
<td>0.9</td>
<td>0.7</td>
<td>0.001</td>
<td>1.E-03</td>
<td>200</td>
<td>120</td>
<td>468</td>
<td>147</td>
</tr>
<tr>
<td>0.5</td>
<td>0.3</td>
<td>0.002</td>
<td>1.E-03</td>
<td>100</td>
<td>17</td>
<td>94</td>
<td>18</td>
</tr>
<tr>
<td>*</td>
<td>0.6</td>
<td>0.4</td>
<td>0.002</td>
<td>1.E-03</td>
<td>100</td>
<td>9</td>
<td>97</td>
</tr>
<tr>
<td>0.7</td>
<td>0.5</td>
<td>0.002</td>
<td>1.E-03</td>
<td>100</td>
<td>9</td>
<td>83</td>
<td>24</td>
</tr>
<tr>
<td>0.8</td>
<td>0.6</td>
<td>0.002</td>
<td>1.E-03</td>
<td>100</td>
<td>6</td>
<td>77</td>
<td>13</td>
</tr>
<tr>
<td>0.9</td>
<td>0.7</td>
<td>0.002</td>
<td>1.E-03</td>
<td>100</td>
<td>15</td>
<td>104</td>
<td>18</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.002</td>
<td>1.E-100</td>
<td>500</td>
<td>277</td>
<td>920</td>
<td>347</td>
</tr>
<tr>
<td>0.8</td>
<td>0.3</td>
<td>0.001</td>
<td>1.E-20</td>
<td>500</td>
<td>452</td>
<td>526</td>
<td>490</td>
</tr>
<tr>
<td>0.8</td>
<td>0.28</td>
<td>0.001</td>
<td>1.E-20</td>
<td>520</td>
<td>718</td>
<td>732</td>
<td>744</td>
</tr>
<tr>
<td>*</td>
<td>0.6</td>
<td>0</td>
<td>0.0006</td>
<td>1.E-25</td>
<td>1000</td>
<td>625</td>
<td>665</td>
</tr>
<tr>
<td>0.6</td>
<td>0.4</td>
<td>0.0002</td>
<td>1.E-03</td>
<td>1000</td>
<td>597</td>
<td>526</td>
<td>728</td>
</tr>
</tbody>
</table>
Figure 11 – The distribution of p-values for Kolmogorov-Smirnov, mHG and Wilcoxon tests

The distribution of p-values for two particular combinations of p, q and α, are shown. These were calculated for randomly drawn 1000 binary vectors, such that the probability of 1 at the top is higher than at the bottom, as dictated by the parameters p, q and α. (A) p=0.6, q=0.4, α=0.002 (ratio=100) (B) p=0.6, q=0, α=0.0006 (ratio=1000).

4.2.2 Bounds on mmHG

Using the minimum hyper-geometric (mHG) statistics developed in the Yakhini research group, when one of the sources of data is a ranked list of genes, gene set enrichment analysis can be used to perform enrichment tests without using an arbitrary threshold to define a gene set from that ranked source. This idea can be extended to the case of two rank orders over a common universe of elements, such as, for example, when genes can be ranked according to differential expression and according to a non-binary motif-presence score. In this case, assessing mutual enrichment in the two rank orders is appropriate. Mutual enrichment is more informative from the point of view of practical biological science than simple correlation measures, as it focuses on the top of the lists and not on the overall agreement, which may be weak even in cases where extremities agree. Relative ranking can be represented by using permutations over the measured elements. Therefore – the statistical assessment of mutual enrichment can be modelled by characterizing properties of random permutations. Due to the size of the measure space, statistics over S_N, the group of permutations over N elements, is difficult to perform and implement.

To support the practicality of statistically assessing mutual enrichment in ranked lists, we derived polynomially computable bounds for the associated tail distributions [51]. Namely – we developed methods for computing an upper bound on the p-value of mutual enrichment at the top
of two permutations uniformly and independently drawn over the group of permutations. Naïve
approaches to computing such bounds include variants of the Bonferroni approach. These do not
provide tight bounds and may lead to mis-labeling results as non-significant. For our newly
derived bounds we assessed tightness using simulated data. We observed that our bounds
significantly improve the \( p \)-value estimation. For several representative datasets, we note that our
bound improves the Bonferroni derived \( p \)-value estimates by a factor of almost 40, on average.
We use our statistical and algorithmic framework to support PWM motif searches. Our results
show promise in enabling efficient and user-friendly PWM motif search in ranked lists. The full
characterization of the distribution of mmHG as a random variable over \( S_N \) remains an open
question.

The bounds that we developed as well as the mmHG calculation itself require many
computations of the hyper-geometric tail (HGT). Recall that HGT is defined as a summation,
given the parameters \( N, n_1, n_2 \) and \( b \), as follows:

\[
HGT(N, n_1, n_2, b) = \sum_{i=b}^{\min(n_1, n_2)} \binom{n_1}{i} \binom{N - n_1}{n_2 - i} \frac{(N - i)}{n_2} \tag{17}
\]

One variant of our bound requires \( O(N^3) \) HGT calculations and the other requires \( O(N^2 \log N) \)
HGT calculations. The mmHG score computation requires \( O(N^2) \) HGT calculations. Due to the
summation, HGT takes \( O(N) \) time (assuming factorials can be calculated in constant time using
Striling’s approximation). Clearly, speeding HGT calculation will speed the computations of the
mmHG score and of the bounds on the mmHG \( p \)-value. One direction that we explored was
using approximations on the hypergeometric tail, computable in constant time. Chvatal gives the
following bound on HGT [99, 100], based on Hoeffding’s Inequality [101]:

\[
HGT(N, n_1, n_2, b) \leq \left( \left( \frac{p}{p + t} \right)^{p + t} \left( \frac{1 - p}{1 - p - t} \right)^{1 - p - t} \right)^{n_1} \tag{18}
\]

where \( p = \frac{n_2}{N} \) and \( t = \frac{b}{n_1} - \frac{n_2}{N} \). This inequality holds when \( t \geq 0 \). It can be further relaxed to give
a more elegant, though weaker, bound:

\[
HGT(N, n_1, n_2, b) \leq e^{-2t^2 n_1} \tag{19}
\]
Given an attainable score \( s \) for which we want to calculate the mmHG \( p \)-value, when enumerating HGT values we are interested in checking which of them are not greater than \( s \). So we can calculate Chvatal bound for every combination of \( N, n_1, n_2 \) and \( b \), and skip the HGT calculation if the bound \( \leq s \). Otherwise the HGT calculation is necessary. Applying this approach to our implementation of the mmHG bound turned out to be more time consuming (for both Chvatal inequalities) due to the fact that this calculation almost always does not save HGT calculations but adds on them, and that is because our code utilizes monotonicity properties of the hyper-geometric distribution to skip HGT calculation for non-promising combinations. We also note that \( HGT(N, n_1, n_2, b) \leq HGT(N, n'_1, n'_2, \hat{b}) \) does not imply that the Chvatal bound for \( N, n_1, n_2, b \) is lower than/equal to the Chvatal bound for \( N, n'_1, n'_2, \hat{b} \), as illustrated in Figure 12.

![Figure 12 – Chvatal bound vs. HGT value (minus log-scale)](image)

We calculated the hyper-geometric tail for all possible combinations of \( N, n_1, n_2, b \) where \( N=100 \) and \( t \geq 0 \). The Chvatal bound here refers to \( e^{-2t^2/n_1} \). HGT values, as well as their corresponding Chvatal bounds, are showed in minus log-scale (base 10).

There are alternatives to mmHG in assessing mutual enrichment in ranked lists. As in the simpler case of mHG, the Kolmogorov-Smirnov test and the Wilcoxon (Mann-Whitney) rank sum test can be generalized for two lists. Given two lists, they can be transformed into a permutation by arbitrarily setting the indices of one list to the identity permutation \( (1,2,\ldots,N) \) and treating the other list as a permutation \( \pi = \pi(1), \ldots, \pi(N) \) over these numbers. Now, given a single permutation \( \pi \in S_N \) and for every \( i=1,\ldots,N \), a binary vector \( \lambda_i \) is defined in which exactly \( i \) entries are 1 and \( N-i \) entries are 0, as follows: \( \lambda_i(j) = 1 \) iff \( \pi(j) \leq i \). As both Kolmogorov-Smirnov and Wilcoxon tests can take as input a binary vector and return a \( p \)-value, a generalized
test can be defined by taking the minimal $p$-value found over the $N$ different binary vectors $\lambda_1, \ldots, \lambda_N$. Formally:

$$
stat(\pi) = \min_{1 \leq i \leq N} p(\lambda_i) \quad (20)
$$

where $stat$ indicates the generalized test statistic for Kolmogorov-Smirnov or Wilcoxon, respectively, and $p$ indicates the test $p$-value for the binary vector $\lambda_i$ under Kolmogorov-Smirnov or Wilcoxon definition, respectively.

Since $N$ binary vectors are tested and the minimum is taken, a correction for multiple testing should be made, for example using Bonferroni’s correction:

$$
P-value(stat(\pi)) \leq stat(\pi) \cdot N \quad (21)
$$

An additional alternative method for mutual enrichment assessment, called the Quantized-mHG, is described in Israel Steinfeld’s dissertation [102]. The basic idea of this methodology is to use $k$ thresholds on the first ranked list to define $k+1$ quanta in the second ranked list. The details can be found in [102].

4.3 Biological results

4.3.1 Cooperative association between an RNA binding protein and a microRNA

Micro RNAs comprise a large family of small non-coding RNAs that are thought to regulate a large fraction of protein-coding genes. Generally, miRNAs down-regulate messenger RNA expression by binding to the 3′ untranslated regions (UTRs) of the RNA molecules. An important factor for binding specificity is the matching in the seed region (positions 2-8 in the miRNA). In addition, target site accessibility is thought to be crucial for efficient repression of miRNA targets.

We analyzed human 3′-UTR sequences containing potential binding sites of 153 conserved miRNA families, and ranked sequences around the sites according to their miRNA accessibility [77]. By applying a rank-based motif search tool to these miRNA targets, we found motifs that are enriched among less accessible targets. As expected from our ranking method, most of the significant motifs were GC-rich. However, one AU-rich motif was found to be enriched among
miR-410 less accessible targets. This motif resembles the Pumilio homolog 1 (PUM1) consensus binding site. We observed a stronger enrichment of the PUM1 motif in conserved targets than in non-conserved targets; moreover, the enrichment of this motif was found to be conserved in a subset of placental mammals. Further, we analyzed publicly available gene expression data, and found that the mutual expression of PUM1 and miR-410 has a greater negative influence on the expression of low accessibility targets than on other targets, an effect that was stronger than when considering both miR-410 and PUM1 separately.

Taken together, our findings suggest a cooperative relationship between miR-410 and PUM1 in regulating human highly structured 3′-UTRs. This kind of cooperation can allow a second level of regulation of such targets. Considering cases in which miRNAs bind low accessibility targets may help to improve current miRNA prediction tools and to obtain a better understanding of the mechanisms underlying miRNA regulation activity.

In a previous study by Fiore et al. [103], it was demonstrated that the miRNA cluster miR379–410 (containing miR-410) is transcriptionally activated upon activation of mouse cortical neurons, and that one component of this cluster, miR-134, takes part in triggering activation-dependent dendritogenesis. The latter study further suggests that Pum2 is a miR-134 target in this process. Based on this validated cooperation between Pumilio and the miR379-410 cluster in dendritogenesis, we speculate that the cooperation between miR-410 and Pumilio suggested in our study might be part of the process regulating dendritogenesis. Clearly, further experimental assays should be carried out to validate this observed association of miR-410 and PUM1 in human 3′-UTRs, and to explain the related mechanisms.

One specific mechanism through which the Pumilio RNA binding protein has been proposed to modulate miRNA function is by binding to sequences that can hybridize with miRNA recognition sites and thereby make them more accessible for the RISC complex [78, 79]. Binding of PUM1 to the 3′UTR of p27 has been reported to cause a conformational change in structure that permits a more efficient binding of miR-221 and miR-222 to their target sites on the p27 3′UTR, followed by rapid entry to the cell cycle [78]. In another study, Pumilio was suggested to rescue microRNA recognition sites from hairpin structures. In response to Pumilio binding, these hairpin structures are relaxed, making the transcript more accessible to microRNAs [79].
4.3.2 Motif discovery

4.3.2.1 Discovery of a novel motif recognized by Puf2p

PUF proteins bind mRNAs and regulate their translation, stability, and localization. Each PUF protein binds a selective group of mRNAs, enabling their coordinate control. Together with Yoav Arava’s Lab at the Technion, we focused on the specificity of Puf2p and Puf1p of *Saccharomyces cerevisiae*, which co-purify with overlapping groups of mRNAs [17]. We applied a rank-based motif search tool to identify putative binding sequences for both proteins. We first identified a novel motif in the 3′ UTRs of mRNAs previously shown to associate with Puf2p. This motif consisted of two UAAU tetranucleotides separated by a 3-nucleotides linker sequence, which we refer to as the dual UAAU motif. The dual UAAU motif was necessary for binding to Puf2p, as judged by gel shift, yeast three-hybrid, and co-immuno-precipitation from yeast lysates. The UAAU tetranucleotides are required for optimal binding, while the identity and length of the linker sequences are less critical. Puf1p also binds the dual UAAU sequence, consistent with the prior observation that it associates with similar populations of mRNAs. In contrast, three other canonical yeast PUF proteins fail to bind the Puf2p recognition site. The dual UAAU motif is distinct from previously known PUF protein binding sites, which invariably possess a UGU trinucleotide. This study expands the repertoire of *cis* elements bound by PUF proteins and suggests new modes by which PUF proteins recognize their mRNA targets.

4.3.2.2 Refinement of known motifs using variable gapped motif search

Using our methodology for variable gapped motif search [28], we present below analysis results of biological importance:

- We identified CATG as the strongest TP53 half-site. While the pair AT was shown to be the best instance of WW in the known consensus CWWG, this was done in controlled synthetic experiments. Our results which validate this recently established preference were obtained using high throughput measurement data. Consistent with literature, we found a spacer of 6 base pairs to yield significant results. The variable gap did not add much to the significance in this case.
• We observed GTCA-N^{3,6,9}-TGAC as the strongest binding site for ESR1. The existing literature consensus, to the best of our knowledge, is GTCA-N^{3}-TGAC. This newly hypothesized refinement of the ESR1 binding site, inferred from high throughput measurement data, demonstrates the utility of variable gap motif search. The data analyzed to yield this finding consisted over 10,000 human DNA sequences, each of length 600 bp – a size which is difficult to handle by most state of the art motif search software tools.

• We suggested HRDLAARN-X^{12}-DFGL-X^{33-39}-SDVW as a significant motif related to potential tyrosine phosphorylation in humans (and in mice). HRDLAARN, the first part of this motif, is known to be related to tyrosine phosphorylation. Our analysis gives rise to a significant refinement.

Our work on variable gapped motif search is the first to perform a systematic analysis of variable gapped motifs in biological datasets. We have found variable gapped motifs to be more significant in a handful of cases and therefore reach the preliminary conclusion, consistent with Reid et al. [27], that they are important but not extremely widespread. Future research directions include more extensive application to biological data, and expansion of variable gapped motifs represented as exact words into flexible words, such as PWMs.

4.3.3 GC-rich motifs found amongst promoters of tissue specific long-non-coding RNAs

We analyzed a collection of datasets consisting of human long-non-coding (Inc) RNAs. We used mmHG-Finder to study PWM motifs in IncRNA promoter sequences ranked according to tissue specificity. We found GC-rich strings to be enriched amongst the promoter sequences of long non-coding RNAs that are specifically expressed in thyroid and prostate tissue samples. Specifically, we found the EGR1 binding motif to be enriched in the promoter regions of IncRNAs which are thyroid-specific. EGR1 was observed to be highly expressed in thyroid, consistent with our stronger motif findings. Similarly, EGR1 is highly expressed in adipose tissue and its transcription factor binding sites are enriched in IncRNAs specific to this tissue. We also observed a statistical association with tissue specific CpG hypo-methylation. We do not have methylation data for thyroid and adipose tissue. However – we did observe the promoters of IncRNAs that are specific to breast to have enriched occurrences of motifs that are similar to EGR1 transcription factor binding sites. EGR1 is also highly expressed in breast. Finally, the
promoters of lncRNAs that are specific to breast were less methylated in breast than other promoters. This suggests the role of EGR1 in driving tissue differentiation by transcribing tissue-specific lncRNAs and by protecting the associated promoters from methylation. EGR1 has been previously shown to recognize GC-rich consensus sequences located in CpG island promoters of active genes [104]. Since DNA methylation in CpG islands is generally associated with gene silencing, a possible explanation is that EGR1 cannot bind methylated GC-rich regions of promoters. Moreover, we observed that tissue-specific lncRNA promoters tend to be less methylated than those of non-tissue-specific lncRNAs in prostate, brain, ovary, breast and kidney, which may be associated with the GC-rich patterns enriched among their tissue-specific lncRNA promoter sequences.

4.4 List of open questions

In summary, future directions are as follows:

- Rigorous characterization of different non-parametric approaches to statistical enrichment in ranked lists. Specifically – what are the regimes of relationships between $p$, $q$ and $\alpha$, in the model described in Section 4.2.1, in which each of the approaches has advantages or disadvantages.

- Full characterization of mmHG distribution.

- Expand the use of statistical enrichment to better understand lncRNA regulation. In particular, analyze more tissue specificity data to better understand lncRNA role there.

- Applications of mHG and mmHG in other (non-biological) domains.
References


42. Bembo O: Sequence logos for DNA sequence alignments. University of California, Berkeley; 2013.
53. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007, **447**:799-816.


הכרח לשל חולן כן-רני, ממסフェת האמאילי. דיקוט נספחת שבעיון עבר העמק הזה חומץ במערכת.

שיסמ קואופטייביט ביו-ריום miR-410 בכרה של מתרח בצלול מכבר מדורג סופר ולא נגש.

לסיום, תרומת המחקית מחזירה בנתונים הת المناسبים לחקל:

- שיטות פעילות במקל העבדה. בינויה מתחדזים לחקים על שטח הסכנת המאפרת התחום
- בפרומטרים דרימוס, כלל תכונת המרוכבת משני הלקום פורדימ. בנסף, הרובון של הדידם בנויה
- סטטסיסטיקה של רישום מדרון生態 פיתון באמצעות חפשים
- כל سبحואפריטיס. השישות האלגוריתמיס שפיתות מומשו בתוכנה הקודים זיון ברש צולב. באופי
- סקיצי, כל מתחשים ג adjusts במערת אפקטייית האכטננטר DRIMust המאפרת התחום ד-ו בול של
- מוטיבים.

- הרצאות בוליגוני. מפגשおくה יומם בך ח違う שופורביס חם מיקור-רני aestים את
- של ייריא לא מקוד. באימוץ שימו בך שלב להפסת מוטיבים יזום את החרד במרח הכלב
- ממסフェת האמאילי – PuF2p – ב/xml, ו CPF מוטיבים יותר.

המבהכר הכללי של חירוב זה

בתקומת אמו-מנייה מיקא רצוני שנות הקושר ליווי שנות של העבידה של,כלל סטטסיסטיקה

ואלגוריתמיס. אוחיلاح ומקוות את השישות של המרוכבים המריכים הקשורים לסטטסיסטיקה

ולאלאוגררימיס - ברפם של השישות. המאמרים מופיעים ואת חור זה בברק הרובון. לבסף, אנ דינא

בממשועת העבדה, בתרות וbezpieczeńst של השישות שפחתות ובו מתנאות הבוליגון - ברפם הדינא.
In this paper, the author presents a novel approach to improving the performance of existing computer vision systems through the use of deep learning techniques. The proposed method involves training a convolutional neural network (CNN) on a large dataset of images, which allows the model to learn the features that are most relevant for the specific task at hand.

The experimental results demonstrate that the proposed approach significantly outperforms the state-of-the-art methods in terms of accuracy and efficiency. Furthermore, the model is able to generalize well to new types of images, which demonstrates its versatility and robustness.

Overall, the research presented in this paper represents a significant contribution to the field of computer vision, and has the potential to revolutionize the way we process and analyze visual data.
Technion - Computer Science Department - Ph.D. Thesis  PHD-2014-05 - 2014

A new index for proteins, the Brazilian index, and its applications.

The index can be used to find biological pathways, protein-protein interactions, and other biological entities. It is based on the analysis of protein sequences and their functional annotations.

The index is implemented as a web-based tool, which allows users to search for specific biological entities or for sets of related entities. The results can be visualized and analyzed in detail.

The index is a valuable resource for researchers in the field of bioinformatics and molecular biology, providing a powerful tool for the analysis of protein sequence data.
התקפי

ב-15 הטמונות המאradorות את עידון תפריה בצינון: בולוניה חיובית - בולוניה חיונית - הפולש, באופני, בק

שנימי תפריה בצינון - מזון המשמש בולוניה מולקולרית. הפולשות המאradorות בצינון, בק, ו

מולקולות התלולים של תצוגה מולקולים מתספורים עם עמידות שמוצ生物医药. בבואן תלעות

תצוגה או איני אינדיקציה באנטרביזיון והשאיבי, זינו מוניות או חניכי ואת רוב בהם של

זיפה של מוניות ביצור בולוניהを与え הת survy בвладג ביצור הדוקטורט של

ראשה תקיני ולאחר המוניות מתכתי בולוניה מולקולריות (אשן אופנה עייף פרוסים קירק-58) (1958)

שחזרן התพฤศจית - חניךBritain (DNA) – תכנית בולוניה לחין לennes, (1956) שאר המשמש תבנית לזרית

תלזון. התולזונות וה.setColumns המתכתיי מתכתיי בולוניה בחיל ואת עלים פונקציות רוב, כולם: גוזי של תולזון

מולקולות אחר, שפוכל ה, החברות אוחי-ביי, באליגת תולזון אחת התולזון וחלבון מוקם

למקוק, עלי חקל מוניות בחיל. גוזי טכנית המוניות מתכתיי מתכתיי של התולזון

ומולקולריות, מבקרת שפגית ממיסים לירפ. –ように חלזון של כביסית במגע האינפרמי צי

תולזון, או גוזי חתולה בתולזון חסינ קיים הוליך במולרי ב’ הכדי או, גוזי התולזון.

מולקולות של עד אזי היא גוזי מוניות המוניות, מוגרש וארבע סוזון של תפזורותים ו’ הכדי או, גוז

לירפ, גוזי להוגים מולד המוניות, על כרות מ으면 בבובות. 4. שקופי מירה, חלזון מוניות מגרון.

אומטי כלשו 20 טחי תואמות אפריזי, גוזי להוגים חלזון ממיסים ממעל בבובות. 20. אפר חלזון

נכתב על ה, החברות עד ל, החברות קדוק תולזון, חליל טכניות מתכתיי מתכתיי של התולזון

חומר האינתכות אומטי חתולה. אינפראל מוניות רוב, כולם סופוגו בבובות, ו’ מותאי ח_rent של

ד’כ, גוזי לא חלזון, ביער מventions התולזון של מולקולות של מוניות לתוכנית רוב בזרוף של

תולזון של ביטוי נגמ עם מתכתיי בולוניה של מוניות המוניות בשף פקטוריה שפעון" לזרום בקודה

הסוסים. כל מקורות אול מברצים את התולזון אפי הגלא, אלי הלשון, ייעום, של מנהל השפעון. גוזי

שהרפה בקזל היא את אומט שגב הקולווית חיונית, והל תולזון בתונקל את התולזון השפעון.

ifestyles. קצרים על מברצים את התולזון אסי הגלא, ואך הלשון. ייעום, של מנהל השפעון.

חושון אום תולזון בולזון מברס ביטוי נגמ, ואך הלשון. גזר, פגיי הפונק (הלל: תカテゴリ)

нят לברס, שלח הלוזר, ברצום עיבוד הון (ורמא השבורה), מיקומי בתו,גיובט מצרים, תולזון ומי

חלזון קוחר-ריי לחקלא צל ברק אחד עייף, הילוסטר של אנטי-ברקತ במספר. בד’ אצוריי הקרצות

שאות מוניות תולזון. חלזון קוחר-ריי רים מוהם את אזורי בזקע פגיי רפר. שה المنزل, מספר

חלזון ממספח המשאומיי (Pumilio) קוחר-ריי ‘הlogan האטיף של אריס (UGUR), שאחרי ה, פגיי המשך

המשה ל’ בולזון בשפוחה זרו. עץ מרצוי ובובות את התולזון של יגוי גים ויו הוליך את

רפר התקפה של חלזון קוחר-ריי לולשת את מתוכנית.

נ gör חלזון, מכי, מקור בק, ישון מוטיבים ריבש שמורות אבסולוציות ב’ חלזון שזון.

מודיפיקציה מפגזת חלזון (לאר ברהרה) היא גיגוליצי, במחלקה מזרחי שרי סקרוי לע ב, חלזון.

חלזון אלול מכי, אניצי נ-גיגוליצי השמאלי עייף, מוטיבים אספוריני, אחוריות כל המוגז

לע

Technion - Computer Science Department - Ph.D. Thesis PHD-2014-05 - 2014
המדחke על הבנה של של"ד.Chrome מהפקולטה למדעי המחשב.

חוה א. מעבר

בראש ברארונה אני אסירת מהדורה ל(Israelי ה Nina פלדמן, חוקי, י.ע. ש✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨✨ Perry 2014-05- 2014
הערה סטטיסטית של העשרה בرشימי מדורונות -
אלגוריתמים יישומיים בשימוש מז.AddWithValue

חיבורי על מחקר
לשם מילוי חלקי של הדירישה לקבלת התואר דוקטור לפילוסופיה

לימודי ליבוניה

הוזה לסטטס טכניון - מרכז סטטסטיקלי לישראל
ניסן, תשע"יס חיפה, אפריל, 2014
הערכה סטטיסטיות של העשרה בرابיימוט מדורגות -
אלגוריתמים ויישומים בחיפה ומורות

לימוד לינקדינץ