TEL AVIV UNIVERSITY אוניברסיטת תל-אביב

Sackler Faculty of Exact Sciences, Blavatnik School of Computer Science

Discovering motifs using high-throughput *in vitro* data

THESIS SUBMITTED FOR THE DEGREE OF
“DOCTOR OF PHILOSOPHY”

by

Yaron Orenstein

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

Submitted to the Senate of Tel-Aviv University
September 2014

Acknowledgments

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

I would like to thank all my friends and collaborators in the Computational Genomics lab. I would also like to acknowledge additional collaborators on various projects, some of which are not included in this thesis. Last but not least, I would like to thank my family. Thanks to my parents for their love and support throughout all my academic studies. This work is dedicated to my dear wife, Liat, who helped and encouraged me in so many ways. Finally, I would like to mention my greatest achievements during the last year, our wonderful child – Eyal – you're the best!

Preface

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

1. **Assessment of algorithms for inferring positional weight matrix motifs of transcription factor binding sites using protein binding microarray data**
Yaron Orenstein, Chaim Linhart and Ron Shamir.
Published in *PLoS ONE* [1].

2. **RAP: Accurate and fast motif finding based on protein-binding microarray data**
Yaron Orenstein, Eran Mick and Ron Shamir.
Published in *Journal of Computational Biology* [2].

3. **Design of shortest double-stranded DNA sequences covering all *k*-mers with applications to protein-binding microarrays and synthetic enhancers**
Yaron Orenstein and Ron Shamir.
Published in *Bioinformatics* [3].

4. **A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data**
Yaron Orenstein and Ron Shamir.
Published in *Nucleic Acid Research* [4].

Abstract

A major challenge in system biology is to delineate the regulatory program of a genome, which describes how the cell controls the amount and exact composition of the proteins it produces from each gene in a given circumstance. A major factor in gene regulation is the binding of transcription-regulating proteins to the specific DNA sequences. Technological advancements in recent years have made it possible to take a deep look into cell activity and specifically protein-DNA binding. These new technologies can measure the intensities of thousands and sometimes millions of interactions in a single experiment. The experimental data accumulated by new technologies require efficient and accurate computational analysis to infer the binding preferences of the tested proteins. In this thesis, we studied the practical and theoretical aspects of binding site inference from high-throughput data. We developed new algorithms for inferring compact and accurate binding models from high-throughput data produced by *in vitro* technologies, and implemented them efficiently. Our approach outperforms existing methods and is applicable to data generated by the state-of-the-art technologies. On the theoretical side, we developed new efficient algorithms for solving several combinatorial problems in the field of sequence design. Our methods employ ideas from graph theory, and are faster and conceptually simpler than extant algorithms.

Contents

1. Introduction

1.1 The "Big Data" era

In recent years technologies that measure biological processes have been advancing in an overwhelming pace. Technologies today can measure thousands – and sometimes millions - of values in a single experiment. These can provide an unprecedented view into the living cell. One type of such experiments accurately measures interactions between molecules in a high-throughput manner. Consequently, in each experiment the amount of data produced is enormous. While in the past, biological insights could be achieved by manual interpretations, it is impossible to do so based on such data. Efficient and accurate algorithms are required to process the vast data and derive significant conclusions. As in many other fields, we are in the "Big Data" era.

The living cell is an amazingly complex machine, constantly performing a myriad of biochemical reactions to sustain itself and carry out a variety of functions in a diverse and ever-changing environment. In order to understand how this machinery works, we need to determine the function of each element in that machine and how the functional elements are regulated in the cell. One of the main mechanisms in regulation is through protein-DNA binding. Observing this process *in vitro* can provide important insights regarding its function in the cell.

Thanks to the maturation of high-throughput experimental techniques, we now have tools with which we can study these questions. Two high-throughput technologies measure thousands and even millions of interactions in a single experiment. The first is the "DNA chip", or microarray, which simultaneously measures thousands of interactions using hybridization of mRNAs to an array of pre-designed sequences [5]. The second technology is deep sequencing, which reads millions of DNA sequences simultaneously [6]. In both technologies, a single experiment yields a snapshot of concentrations and strength of interactions in a given tissue or cell-line (*in vivo*) or outside the cell (*in vitro*). While measurements in the cell provide a detailed view of the cell's state, in many cases *in vivo* measurements may be too complex or affected by other confounding factors. In some scenarios, measuring interactions *in vitro* may provide a cleaner view of the studied process. The models inferred from *in vitro* data can later be applied and validated by *in vivo* experiments. Overall, both *in vivo* and *in vitro* experiments are important to advance the research in any biological field.

1.2 Transcriptional regulation

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

The DNA sequences bound by a TF are called its *binding sites* (BSs), or *cis*-regulatory elements. They are typically very short (6-15 bases) and degenerate - a TF can bind, with varying affinities, to many different sequences that reflect a common pattern, or *motif*, characteristic of the factor. Most BSs are found in the *promoter*, the region upstream of the gene's transcription start site (TSS), though BSs may also exist downstream of the TSS and at large distance from the gene, in locations termed *enhancers*. Some TFs cooperate in the regulation of genes, resulting in more complex and specific transcription profiles. Reverse-engineering the transcriptional program of an organism requires identifying its TFs, the locations and affinities of their BSs, and the various modules they are organized in.

Deciphering the transcriptional regulation *in vivo* is a difficult task. While TF binding is sequence-specific, it is affected by many factors. First, the DNA has to be accessible for binding by the TF. Second, other TFs may compete for the same binding sites, making it harder for the TF to bind to its potential binding sites. Third, in some instances, the TFs may only bind cooperatively, but current technologies cannot distinguish between cooperative and direct binding. On top of that, the set of binding site sequences present in the genome may be limited and not reflect all possible binding sequences. In such cases, one cannot derive the full range of TF binding affinities from these data. Learning the DNA binding preferences of a TF from *in vivo* data is hence hampered by assay complexity.

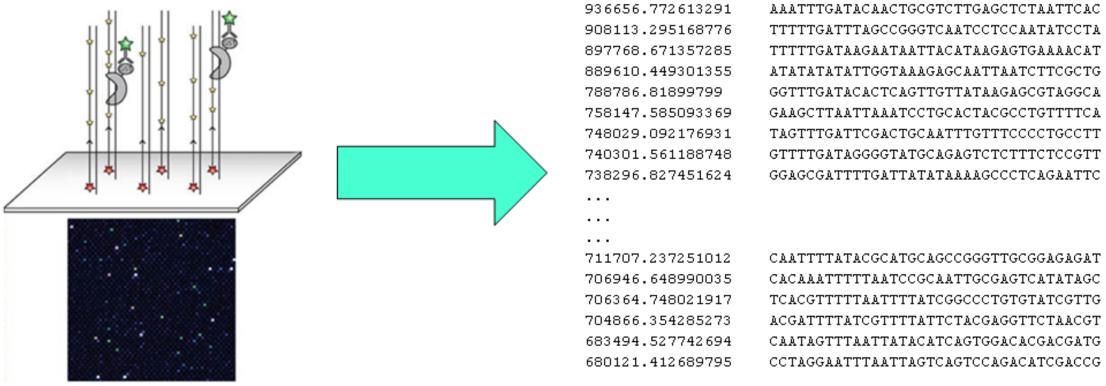
In contrast, *in vitro* data may enable a cleaner high-resolution measurement of TF-DNA binding preferences, as there are fewer confounding factors. First, one can guarantee that no binding sites are inaccessible due to compressed chromatin. Second, there are no competing TFs, as the experiment is performed using a single TF in a synthetic environment. Third, barring technological artifacts, the binding is due to direct TF-DNA binding. Last, in some cases, the sequences can be combinatorially designed to cover all k-mers of the desired length k. In other cases, they are randomly generated such that together they are guaranteed to cover nearly all k-mers. So, if technological biases can be handled, the TF-DNA binding signal is expected to be much clearer.

1.2.1 Technologies for measuring TF-DNA binding

Identifying the sites bound *in vivo* by a specific TF and their affinity is not an easy task. Methods like DNA footprinting or chromatin immunoprecipitation (ChIP) can be used, but are applicable only to short, hand-chosen genomic loci. The combined strategy of ChIP and promoter microarrays, also termed *ChIP-chip*, enables genome-wide identification of promoter segments that are bound by a specific TF, in a single experimental assay [7]. Replacing the microarray-based readout with next-generation sequencing technologies, an approach called *ChIP-seq*, allows the detection of BSs throughout the entire genome [8].

Measuring protein-DNA binding *in vitro* gives a cleaner view of the TF binding preferences, but lacks the genomic context. Since *in vitro* experiments are cheaper and easier, it is highly appealing to use *in vitro* models with complementary genomic information to predict *in vivo* binding. *In vitro* technologies can measure thousands of binding events simultaneously, and report the binding intensity to each possible DNA *k-mer* (a word of length *k*). Using this information, the effects of mutations in the binding sites can be predicted and ultimately help understand individual differences and cross-species divergence. Techniques that measure TF-DNA binding *in vitro* include *protein binding microarrays* (PBMs), based on microarrays, and *high-throughput SELEX* (HT-SELEX), based on deep sequencing.

Universal protein binding microarrays are designed to measure the binding intensity in high-throughput and unbiased manner [9]. Each array contains around 41,000 DNA sequences of length 36bp each. These are designed to cover together all DNA 10-mers [10]. The tested protein binds the sequences, and its binding intensity is measured using a florescence tag (see Figure 1). The same array can be used to test other proteins, as its design is universal. Hundreds of experiments were deposited in the public database UniPROBE [11].



```
936656.772613291    AAATTTGATACAACTGCGTCTTGAGCTCTAATTCAC
908113.295168776    TTTTTGATTTAGCCGGGTCAATCCTCCAATATCCTA
897768.671357285    TTTTTGATAAGAATAATTACATAAGAGTGAAAACAT
889610.449301355    ATATATATATTGGTAAAGAGCAATTAATCTTCGCTG
788786.81899799     GGTTTGATACACTCAGTTGTTATAAGAGCGTAGGCA
758147.585093369    GAAGCTTAATTAAATCCTGCACTACGCCTGTTTTCA
748029.092176931    TAGTTTGATTCGACTGCAATTTGTTTCCCCTGCCTT
740301.561188748    GTTTTGATAGGGGTATGCAGAGTCTCTTTCTCCGTT
738296.827451624    GGAGCGATTTTGATTATATAAAAGCCCTCAGAATTC
...
...
...
711707.237251012    CAATTTTATACGCATGCAGCCGGGTTGCGGAGAGAT
706946.648990035    CACAAATTTTTAATCCGCAATTGCGAGTCATATAGC
706364.748021917    TCACGTTTTTAATTTTATCGGCCCTGTGTATCGTTG
704866.354285273    ACGATTTTATCGTTTTATTCTACGAGGTTCTAACGT
683494.527742694    CAATAGTTTAATTATACATCAGTGGACACGACGATG
680121.412689795    CCTAGGAATTTAATTAGTCAGTCCAGACATCGACCG
```

Figure 1. PBM experiment. A protein binds a pre-designed set of DNA sequences. Its binding is measured using a florescence tag and this image is scanned to produce the binding intensities of each sequence. (Source: [9])

High-throughput SELEX measures the binding of a single protein to millions of random oligos [12-14]. The initial pool, before any binding, is a set of pseudo-random oligos with no specific design. In each cycle of the process, the set of bound oligos is retrieved, amplified and sequenced. The set of filtered oligos is then used as the initial sequence set for the next cycle. Hence, the proportion of the bound oligos increases from one cycle to the next. The output is a set of sequence files, each of a different cycle, starting from the initial pool. Figure 2 shows a schematic of the process.

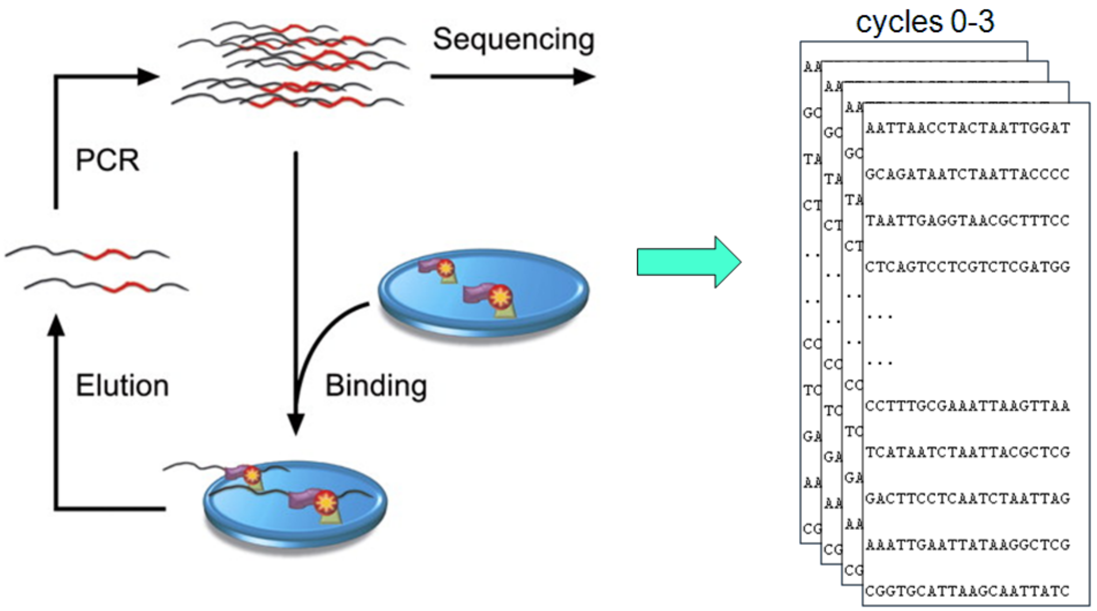


Figure 2. HT-SELEX experiment. A protein binds a pool of random DNA sequences. The bound sequences are filtered and amplified by PCR. A fraction of the resulting set is sequenced and another fraction is used as the initial pool for the next cycle. (Source: [14])

1.2.2 Models for binding site motifs

Several computational models have been developed for describing BS motifs. The most popular model is the *position weight matrix* (PWM), also known as position specific scoring matrix (PSSM) [15]. This model (see Figure 3) uses a 4×k frequency matrix $f_{b,i}$ to represent the motif, where $f_{b,i}$ is the probability for observing nucleotide b at position i in the motif. An inherent property of this model is position-independence: probabilities at different positions are assumed to be independent. The probability that a given k-mer $w = w_1w_2\ldots w_k$ is a functional BS is simply the product of the corresponding matrix elements, i.e., $\prod_{i=1}^{k} f_{w_i,i}$. The matrix can also be viewed as an energy-based model, where instead of frequencies it holds the free energy contributions of the four nucleotides in each position [16]. Among the advantages of the PWM model are its simplicity, small number of parameters and an intuitive visualization [17]. The *logo* format (Figure 3) visualizes the matrix by drawing the different nucleotides in each position in size according to their weights and ordered by their weights. The total height of each position is inversely proportional to its entropy, which corresponds to the strictness of each position.

	1	2	3	4	5	6
A	0.1	0.8	0	0.7	0.2	0
C	0	0.1	0.5	0.1	0.4	0.6
G	0	0	0.5	0.1	0.4	0.1
T	0.9	0.1	0	0.1	0	0.3

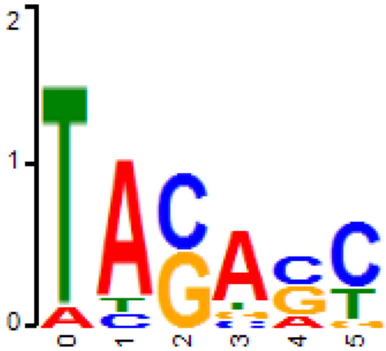


Figure 3. An example of a PWM and its logo illustration. The matrix represents the binding preference of a TF to the different nucleotides in each position. The logo provides a visualization of the matrix.

While the PWM model is very popular and useful, it might be too simplistic for some TFs. An inherent assumption of the model is position-independence, which means that each position adds to the total binding score independently of the other positions. This assumption has been shown to be untrue for some TFs [18]. Other models extend the position weight matrix by additional features. The most prominent features are di-nucleotide dependencies. To avoid the complexity of having too many features, usually only adjacent positions are considered as dependence between neighboring positions were observed more often than between non-neighboring ones [19]. The most comprehensive model, which makes no assumptions, is the complete k-mer model [13]. In this model, every possible DNA k-mer has a binding score representing the affinity of the TF to it. One disadvantage of both models is the huge number of parameters and the risk of over-fitting.

Using validated BSs as training sets and high-throughput experimental techniques such as PBM and HT-SELEX, parameters for TFBS models have been derived for scores of known TFs in various species, and deposited in databases such as TRANSFAC [20], ScerTF [21], UniPROBE [11] and JASPAR [22].

1.3 Motif finding

Over the past several years, a variety of computational methods were developed to analyze PBM and HT-SELEX experimental data and suggest novel biological hypotheses, which can then be tested by further experiments. Unfortunately, since BSs are short and degenerate, and DNA probes contain many putative sites, it is difficult to distinguish between specific binding and non-specific (background) binding. Moreover, each technology suffers from biases, which produce artifacts that eventually distort the measured intensities. Algorithms aim to extract the signal, i.e. the binding preferences of the TF, and distinguish it from the noise (background binding and technological biases).

1.3.1 Motif discovery in genomic sequences

In *de novo* motif discovery, given a set of co-regulated genes, the goal is to find motifs that are statistically enriched in their promoters. Once found, further biological research must be performed in some cases in order to discover the proteins whose BSs are described by these motifs. *De-novo* motif discovery has been tackled using a myriad of algorithmic techniques, such as Expectation Maximization (MEME [23], EMnEM [24], OrthoMEME [25], PhyME [26]), Gibbs sampling (GibbsDNA [27], AlignACE [28], MotifSampler [29]), efficient enumeration (YMF [30], MITRA [31], Multiprofiler [32], WEEDER [33], FootPrinter [34], FIRE [35], Trawler [36], Amadeus [37]), and neural networks (ANN-Spec [38]), as well as greedy (CONSENSUS [39]), graph-based (WINNOWER and SP-STAR [40]), and randomized (PROJECTION [41]) methods.

An extension of this problem is to find motifs *de novo* in a set of ranked or weighted sequences. The weight of a sequence corresponds to the probability or intensity of the binding of the TF to it. Weights may be assigned to different genomic loci based on microarray florescent intensity (in ChIP-chip) or the number of bound sequence reads covering each locus (in ChIP-seq). In other applications, each gene may be given a score based on the change in its expression, and this score is assigned to its promoter sequence. Methods that use weights or a ranked list of genes include DRIM [42], PREGO [43] and MatrixREDUCE [44]. Other methods were specifically designed to infer models from ChIP data (MEME-Chip [45], MDScan [46] ChIPMunk [47] and TherMos [48]). A survey of motif finding tools can be found in [49, 50]. The evolution of motif finding algorithms is described in [51].

1.3.2 Motif finding in PBM data

The problem of inferring a motif from high-throughput *in vitro* data requires algorithms that are tailored to these specific data. A naïve solution is to use methods developed for motif finding in genomic sequence. The set of DNA probes or sequences can be divided into positive and negative sets, according to their binding intensity [1]. A more informative way is to use the measured binding intensities as sequence weights and provide them to one of the tools that work on weighted sequences [52]. Unfortunately, applying these methods has costly running time and produces models that are less accurate compared to models produced by technology-specific methods.

Several approaches have been proposed for inferring accurate binding models from PBM data. The most popular practice is to first derive scores for all possible *k-mers*. These scores depend on the binding intensities of the probes the k-mer appears in. Some methods use average or median binding intensity, while others use enrichment scores, such as Wilcoxon-Mann-Whitney test [53]. The top scoring k-mer is identified as the *consensus* or *seed*. A binding model is inferred by optimizing a function of the data. It may be a model that has the best fit to the ranking of the probes, or to their binding intensities. In either case, a time-consuming optimization procedure learns the model parameters (e.g., maximum likelihood using gradient descent and Levenberg-Marquardt algorithm). Methods for inferring binding site models from PBM data include Seed-and-Wobble [9], RankMotif++ [54] and BEEML-PBM [55]. An international competition on predicting PBM binding intensities was conducted in 2010 [56]. Description of the best performing methods can be found in [52].

1.3.3 Motif finding in HT-SELEX data

Binding model inference from HT-SELEX data is slightly different than from PBM data. As opposed to PBM technology, each DNA oligo represents a binding site, but the intensity is not reported. Instead, it can be computationally derived for k-mers of length smaller than the oligo, since these appear in thousands of oligos. K-mer scores are derived based on their frequency in the different cycles of the experiment. The ratio statistic for a k-mer in cycle i is the ratio of the k-mer's frequency in cycle i and its frequency in cycle i-1. It represents the enrichment of each k-mer between the cycles and thus is an estimate of the binding preference of the TF to this DNA word. The first reported method for inferring binding models from HT-SELEX data was BEEML [12]. It uses the frequencies from two cycles of enrichment to learn the binding preferences based on a free energy model. A method due to Toivonen *et al.* uses k-mer frequencies as scores and constructs a model based on k-mers at Hamming distance ≤1 from the consensus [14]. Another method developed for SELEX-seq data uses k-mer ratios (after correction for biases and artifacts) to derive a complete k-mer list as the binding model [13].

Recently, Jolma *et al.* published hundreds of HT-SELEX experiments [57]. For the first time, a large-scale comparison between HT-SELEX and PBM experiments on the same TFs was possible. Such a comparison may highlight the advantages and disadvantages of each technology, as well as reveal biases and artifacts of each technology. Such insights may later help in developing improved algorithms using these data.

1.4 Combinatorial sequence design in computational biology

Microarray technologies and other techniques that use sets of DNA sequences necessitate design of sequences with specific properties. The set of DNA probes in an experiment, also called *oligonucleotides* (*oligos* in short), determine the space and spectrum of measurements. In general, the wish is to measure a wide spectrum of oligos in order to enable a complete view of the biological process. Typically, the set of oligos is limited by several factors, such as capacity, cost, potential interactions between probes and other experimental considerations.

DNA sequence design is a well-studied area. Microarray probes that measure mRNA quantities were designed to capture transcription profiles of specific organisms [58, 59]. Other designs aim to measure structural variations of genomes, such as genes copy number and SNP detection [60, 61]. In many applications, there is a risk of self- and cross-hybridization of the oligos, which makes them inaccessible. Some designs aim to avoid this risk while preserving high coverage [62].

PBMs measure protein-DNA binding. The microarray is designed to cover all possible k-mers. This enables an unbiased measurement of TF-DNA binding preferences, since all possible k-long binding sites are represented on the array. Ideally, the array would contain 4^k probe sequences, each covering a different k-mer uniquely. However, since the space on the device is limited, this strategy is already unfeasible today for k=8. Instead, a smaller number of longer probes are used, so that each probe contains multiple overlapping k-mers, and together the probes cover all possible k-mers. In the implementation by Bulyk's lab, each microarray contains approximately 41,000 36bp-long probe sequences that together cover all 10-mers [9].

1.4.1 Designing a minimum-length sequence to cover all k-mers

The most compact sequence that covers all k-mers is a *de Bruijn sequence* [9, 63]. A de Bruijn sequence of order k over alphabet ∑ is a cyclic sequence of length $|\sum|^k$, such that each word of length k over $\sum$ appears exactly once. To design a set of oligos from the de Bruijn sequence, it is cut into overlapping subsequences which serve as the oligos. The overlap length is k-1, so each k-mer is present in an oligo. For example, in the PBM array design of [9] all 10-mers are covered in 36bp-long probes, each covering 27 unique 10-mers. Thus, $\lceil 4^{10}/27 \rceil$ probes are required to cover all 10-mers.

There are several methods to generate a de Bruijn sequence of order k over alphabet $|\sum|$. One way to generate de Bruijn sequences is by de Bruijn graphs. A complete *de Bruijn graph* of order k is a directed graph containing $|\sum|^k$ vertices; each vertex represents a unique k-mer. An edge (u, v) exists between two vertices if and only if the (k-1)-suffix of u equals the (k-1)-prefix of v. Thus, each edge represents a unique (k+1)-mer. An Euler tour in a graph traverses each edge exactly once. Thus, such a tour in a complete de Bruijn graph represents a de Bruijn sequence of order k+1 [64]. Another method to generate de Bruijn sequences is based on the theory of Galois fields. Linear shift feedback registers generate a stream of characters where each character is a function of the preceding l characters in the stream [65]. A small subset of these functions can be used to generate a de Bruijn sequence. Universal PBM arrays were designed using linear shift feedback registers with unique properties. The sequences have improved coverage of gapped k-mers and uniform coverage of words of length longer than 10, the order of the de Bruijn sequence used in the PBM design [10]. In general, the number of different de Bruijn sequences over alphabet of size n and order k is $(n!)^{n^{k-1}}/n^k$, making it infeasible to enumerate all of them for realistic k values.

1.4.2 Utilizing the DNA reverse-complement property

In many technologies that utilize sets of DNA probes, the probes are double-stranded. In double-stranded DNA each strand is matched with its *reverse complement*. A *complementarity relation* is a symmetric non-reflexive relation. For DNA, A=complement(T) and C=complement(G). In the reverse complement sequence of sequence S, denoted RC(S), each letter is replaced with its complement and letters are placed in reverse order. For example, ACGG=RC(CCGT). One example for a technology using double-stranded DNA probes is PBMs, which measure the binding of a protein to double-stranded DNA probes [9]. Another example arose in the context of synthetic enhancers: double-stranded DNA sequences were inserted into the zebra fish genome, and their effect on the limb formation during its development was measured [66].

Instead of using the de Bruijn sequence to generate probes containing all k-mers, a major saving in the number of probes may be achieved by utilizing the reverse complementary nature of the probes. The set of probes is designed to cover all k-mers. However, whenever a k-mer is covered by a probe, so is its reverse complement. Theoretically, for each k-mer it is enough for the set of probes to cover either the k-mer or its reverse complement. We call a sequence with this property a *reverse complementary de Bruijn sequence*.

The problem that this reasoning raises is how to generate a minimum-length reverse complementary de Bruijn sequence over a finite alphabet Σ. A solution for odd k was presented (without proof) in [67]. A full solution is given later in this thesis. In parallel to us, a method that generates the smallest set of probes of a specific length to cover all k-mers utilizing the reverse complement property was developed, but its running time is prohibitive even for moderate k values [66]. A polynomial time solution to the related problem of finding a maximum-length sequence such that each k-mer appears at most once, in either orientation, was given for odd k in [68].

1.5 Summary of articles included in this thesis

1. **Assessment of algorithms for inferring positional weight matrix motifs of transcription factor binding sites using protein binding microarray data.**
   Yaron Orenstein, Chaim Linhart and Ron Shamir.
   Published in *PLoS ONE* [1].

The new technology of protein binding microarrays (PBMs) allows simultaneous measurement of the binding intensities of a transcription factor to tens of thousands of synthetic double-stranded DNA probes, covering all possible 10-mers. A key computational challenge is inferring the binding motif from these data. We present a systematic comparison of four methods developed specifically for reconstructing a binding site motif represented as a positional weight matrix from PBM data. The reconstructed motifs were evaluated in terms of three criteria: concordance with reference motifs from the literature and ability to predict *in vivo* and *in vitro* bindings. The evaluation encompassed over 200 transcription factors and some 300 assays. The results show a tradeoff between how the methods perform according to the different criteria, and a dichotomy of method types. Algorithms that construct motifs with low information content predict PBM probe ranking more faithfully, while methods that produce highly informative motifs match reference motifs better. Interestingly, in predicting high-affinity binding, all methods give far poorer results for *in vivo* assays compared to *in vitro* assays.

2. **RAP: Accurate and fast motif finding based on protein-binding microarray data.**
   Yaron Orenstein, Eran Mick and Ron Shamir.
   Published in *Journal of Computational Biology* [2].

The novel high-throughput technology of protein-binding microarrays (PBMs) measures binding intensity of a transcription factor to thousands of DNA probe sequences. Several algorithms have been developed to extract binding-site motifs from these data. Such motifs are commonly represented by positional weight matrices. Previous studies have shown that the motifs produced by these algorithms are either accurate in predicting in vitro binding or similar to previously published motifs, but not both. In this work, we present a new simple algorithm to infer binding-site motifs from PBM data. It outperforms prior art both in predicting in vitro binding and in producing motifs similar to literature motifs. Our results challenge previous claims that motifs with lower information content are better models for transcription-factor binding specificity. Moreover, we tested the effect of motif length and side positions flanking the “core” motif in the binding site. We show that side positions have a significant effect and should not be removed, as commonly done. A large drop in the results quality of all methods is observed between in vitro and in vivo binding prediction. The software is available on acgt.cs.tau.ac.il/rap.

3. **Design of shortest double-stranded DNA sequences covering all *k*-mers with applications to protein-binding microarrays and synthetic enhancers.**
   Yaron Orenstein and Ron Shamir.
   Published in *Bioinformatics* [3].

Novel technologies can generate large sets of short double-stranded DNA sequences that can be used to measure their regulatory effects. Microarrays can measure *in vitro* the binding intensity of a protein to thousands of probes. Synthetic enhancer sequences inserted into an organism’s genome allow us to measure *in vivo* the effect of such sequences on the phenotype. In both applications, by using sequence probes that cover all *k*-mers, a comprehensive picture of the effect of all possible short sequences on gene regulation is obtained. The value of *k* that can be used in practice is, however, severely limited by cost and space considerations. A key challenge is, therefore, to cover all *k*-mers with a minimal number of probes. The standard way to do this uses the de Bruijn sequence of length 4^k. However, as probes are double stranded, when a *k*-mer is included in a probe, its reverse complement *k*-mer is accounted for as well. Here, we show how to efficiently create a shortest possible sequence with the property that it contains each *k*-mer or its reverse complement, but not necessarily both. The length of the resulting sequence approaches half that of the de Bruijn sequence as *k* increases resulting in a more efficient array, which allows covering more longer sequences; alternatively, additional sequences with redundant *k*-mers of interest can be added. The software is freely available from our website http://acgt.cs.tau.ac.il/shortcake/.

4. **A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data.**
   Yaron Orenstein and Ron Shamir.
   Published in *Nucleic Acid Research* [4].

Understanding gene regulation is a key challenge in today's biology. The new technologies of protein-binding microarrays (PBMs) and high-throughput SELEX (HT-SELEX) allow measurement of the binding intensities of one transcription factor (TF) to numerous synthetic double-stranded DNA sequences in a single experiment. Recently, Jolma *et al.* reported the results of 547 HT-SELEX experiments covering human and mouse TFs. Because 162 of these TFs were also covered by PBM technology, for the first time, a large-scale comparison between implementations of these two *in vitro* technologies is possible. Here we assessed the similarities and differences between binding models, represented as position weight matrices, inferred from PBM and HT-SELEX, and also measured how well these models predict *in vivo* binding. Our results show that HT-SELEX- and PBM-derived models agree for most TFs. For some TFs, the HT-SELEX-derived models are longer versions of the PBM-derived models, whereas for other TFs, the HT-SELEX models match the secondary PBM-derived models. Remarkably, PBM-based 8-mer ranking is more accurate than that of HT-SELEX, but models derived from HT-SELEX predict *in vivo* binding better. In addition, we reveal several biases in HT-SELEX data including nucleotide frequency bias, enrichment of C-rich k-mers and oligos and underrepresentation of palindromes.

2. Assessment of algorithms for inferring positional weight matrix motifs of transcription factor binding sites using protein binding microarray data

Assessment of Algorithms for Inferring Positional Weight Matrix Motifs of Transcription Factor Binding Sites Using Protein Binding Microarray Data

Yaron Orenstein, Chaim Linhart, Ron Shamir*

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

Abstract

The new technology of protein binding microarrays (PBMs) allows simultaneous measurement of the binding intensities of a transcription factor to tens of thousands of synthetic double-stranded DNA probes, covering all possible 10-mers. A key computational challenge is inferring the binding motif from these data. We present a systematic comparison of four methods developed specifically for reconstructing a binding site motif represented as a positional weight matrix from PBM data. The reconstructed motifs were evaluated in terms of three criteria: concordance with reference motifs from the literature and ability to predict *in vivo* and *in vitro* bindings. The evaluation encompassed over 200 transcription factors and some 300 assays. The results show a tradeoff between how the methods perform according to the different criteria, and a dichotomy of method types. Algorithms that construct motifs with low information content predict PBM probe ranking more faithfully, while methods that produce highly informative motifs match reference motifs better. Interestingly, in predicting high-affinity binding, all methods give far poorer results for *in vivo* assays compared to *in vitro* assays.

Citation: Orenstein Y, Linhart C, Shamir R (2012) Assessment of Algorithms for Inferring Positional Weight Matrix Motifs of Transcription Factor Binding Sites Using Protein Binding Microarray Data. PLoS ONE 7(9): e46145. doi:10.1371/journal.pone.0046145

Editor: Ying Xu, University of Georgia, United States of America

Received July 23, 2012; **Accepted** August 27, 2012; **Published** September 28, 2012

Funding: This study was supported in part by the European Community's Seventh Framework grant [HEALTH-F4-2009-223575] for the TRIREME project; and by the Israel Science Foundation (grant no. 802/08). YO was supported by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rshamir@tau.ac.il

Introduction

Understanding gene regulation is a fundamental problem in biological research. A principal way to regulate gene expression in the cell is via transcription, which is governed primarily by transcription factors (TFs). A TF is a protein that binds to the promoter region of a gene at specific sequences, called TF binding sites (TFBSs). The binding of one or several TFs enables or impedes the transcription of the gene. A TF binds to similar short nucleotide sequences at different affinities. Finding these cis-regulatory elements and modeling the affinity of TF binding to them is a central challenge in understanding gene regulation.

The most common computational model for describing a TFBS motif is a position weight matrix (PWM) [1]. The TFBS is represented by a $4 \times k$ matrix, where k is the motif length. Each column contains four probabilities, representing the nucleotide frequencies at that position. This relatively simple model is highly popular since it is compact, effective and easy to interpret.

New technologies have enabled comprehensive mapping of protein-DNA binding affinities. The main technology to measure *in vivo* protein occupancy is chromatin immunoprecipitation (ChIP). In the ChIP-chip method, the protein-bound DNA segments are hybridized to a pre-designed microarray [2], whereas the ChIP-seq method uses deep sequencing to read the bound DNA segments [3]. A recent promising technology in this field is the protein binding microarray (PBM) [4]. This microarray contains ~41,000 synthesized, 60 bp-long double-stranded DNA probes, each containing 36 bp of unique sequence, designed so that every possible 10-mer is contained in exactly one probe sequence. A single *in vitro* experiment measures the binding intensity profile of a specific TF to each probe, thereby providing complete coverage of the binding affinity of the TF to all possible 10-mers. Often, two experiments with different array designs are performed with the same TF, providing *paired* profiles.

Numerous computational methods for finding a motif in a target set of promoters have been developed over the last two decades [5–7]. Predicting binding sites based on PBM data is different: the experimental data are much more comprehensive, covering all possible 10-mers, but are generated *in vitro* and in a high-throughput (and hence noisy) fashion. Therefore, several methods were recently developed specifically for identifying TFBS motifs from PBM profiles. Here we compare methods that represent the motifs as PWMs. We do not include methods that use more complex models [8], since we choose to focus on simpler, more compact models.

In this paper we present a systematic comparison of four algorithms for identifying TFBS motifs from PBM profiles: Seed-and-Wobble (SW) [4], RankMotif++ (RM) [9], BEEML-PBM (BE) [10] and the algorithm Amadeus-PBM (AM) introduced here (see **Table 1**). In 2005, a systematic comparison of computational methods for motif discovery in promoters clarified some of the issues and the difficulties in that domain, and led to progress in that research area [11]. We hope that our study will have a similar effect regarding methods for analyzing PBM data.

Table 1. Properties of the tested methods.

Program	Operating principle	Reference
Seed-and-Wobble	Ranks all 8-mers according to Wilcoxon-Mann-Whitney rank-sum score. The top scoring 8-mer is used as a seed, its positions are "wobbled" and its length is extended in order to improve match to the data. http://the_brain.bwh.harvard.edu/PBMAnalysisSuite/index.html	[4]
RankMotif++	Aims to predict the ranking of the probes according to their binding intensity. Maximizes the likelihood of the ranking function, using the three top 7-mers as seeds. http://morrislab.med.utoronto.ca/software.html	[9]
BEEML-PBM	Estimates the position and background biases from the data, then optimizes the parameters of a binding energy model using BEEML algorithm, explicitly taking the biases into account. http://stormo.wustl.edu/beeml/	[10]
Amadeus-PBM	Seeks enriched PWMs in 1000 top ranking 9-mers compared to the background set of all 9-mers, using Amadeus motif finding algorithm. http://acgt.cs.tau.ac.il/amadeus//	Described here

doi:10.1371/journal.pone.0046145.t001

Results

Concordance with SELEX-based reference motifs from the literature

We used each method to find motifs using PBM data, and compared the results to previously reported motifs for the same TFs, obtained using independent experiments. Each motif was learned using the data from two paired experiments performed with the same TF. For each TF, we measured the distance between the PBM-based PWM to the PWM of the same TF as published in JASPAR [12]. For this test we used all mouse PBM datasets from the SCI09 study [13,14] that had a corresponding PWM in JASPAR, excluding those for which the JASPAR PWMs were constructed using PBM data. This set contained 58 PWMs. Most were constructed based on *in vitro* SELEX experiments, which are still the main source of TF motifs.

The AM PWMs were the most similar to JASPAR, with average Euclidean distance (± estimated standard deviation) 0.178 ± 0.11. The average for SW was 0.193 ± 0.1, for RM was 0.21 ± 0.09, and for BE was 0.227 ± 0.1 (**Table 2**). The difference between AM and SW was not significant ($p=0.17$, Wilcoxon rank-sum test) and both were significantly better than RM and BE ($p=0.001$ and $p=0.0005$ compared to AM, respectively).

We then focused on high-quality predictions of the four methods. We say that a motif is successfully recovered by a method if the Euclidean distance of the predicted PWM from the reference PWM is below a predetermined cutoff. As in [15], we used three cutoffs for the distance. AM attained a higher success rate using all cutoffs (**Figure 1**). A similar comparison of mouse motifs in TRANSFAC [16] and yeast motifs in ScerTF [17], and a parallel comparison, using p-value for the significance of the similarity [18], showed a similar advantage to AM (**Figure S1**).

Visual inspection suggested that the PWMs produced by AM and SW are easier to interpret and look distinct in logo format (**Figure 2**). To quantify this observation, we calculated the average information content for each PWM (see **Methods S1**). Averaged over the PWMs computed from all 115 available paired mouse PBM sets, the information scores for the raw PWMs were 1.03, 0.61, 0.42 and 0.53 bits for AM, SW, RM and BE, respectively, with AM scoring significantly higher ($p<10^{-15}$, Wilcoxon rank-sum test). After trimming the PWMs to discard flanking positions with low information, the information averages were 1.03, 1.09, 0.54 and 0.61 bits, respectively ($p=1.2\cdot10^{-7}$ when comparing SW to AM and $<10^{-15}$ when comparing AM and SW to RM and BE). The full comparison results are available in **Table S1**.

Predicting *in vitro* binding intensities

Next, we tested the prediction of binding intensities by the four methods on 115 pairs of mouse PBM profiles [13,14] following the procedure in [9]. Each method learned a PWM according to one PBM experiment; this PWM was used to rank the probes of its paired array. The goal was to correctly rank the positive probes, i.e. those with highest affinity measurements. The set of positive probes (denoted 4σ, see **Methods S1**) contained an average of 912 probes per array. We also evaluated larger sets of positive probes using more permissive cutoffs (denoted 3σ, 2σ and 1σ; an average of 1580, 3215 and 8224 probes per array, respectively).

When testing on 4σ top probes set (**Table 2** and **Figure 3**), BE had significantly best Spearman and AUC scores ($p<0.0025$, Wilcoxon rank-sum test), while AM and RM were essentially equal ($p=0.41$ and $p=0.44$, respectively), and significantly better than SW ($p<10^{-4}$). Using the sensitivity measure, BE was again best

Table 2. Summary of the comparison. Boldface indicates significantly better performance than the other methods (including equal top performance).

	Similarity to reference motifs	*In vitro* binding prediction			*In vivo* binding prediction			Running time
	Average Euclidean distance	Spearman rank coefficient	Sensitivity at 1% FP	AUC	Spearman rank coefficient	Sensitivity at 1% FP	AUC	Seconds
AM	**0.178**	0.27	0.342	0.876	**0.152**	0.089	**0.653**	**30**
SW	**0.193**	0.244	0.305	0.866	**0.145**	**0.118**	**0.659**	7200
RM	0.21	0.264	0.295	0.881	**0.158**	0.092	**0.655**	3600
BE	0.227	**0.308**	**0.411**	**0.891**	**0.146**	0.084	**0.665**	900

doi:10.1371/journal.pone.0046145.t002

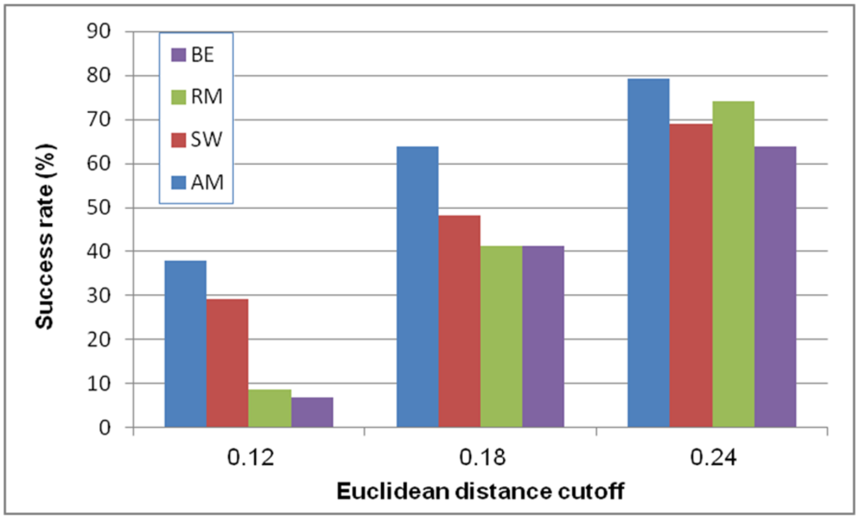


Figure 1. Similarity to experimentally established PWMs. For 58 TFs, we compared the motifs produced from their PBM profiles by each method, to the known motif from JASPAR database. Distance was measured using Euclidean distance. Three distance cutoffs were used, and the fraction of recovered motifs with distance below the cutoff is the success rate. BE: BEEML-PBM, RM: RankMotif++, SW: Seed-and-Wobble, AM: Amadeus-PBM, JR: JASPAR.
doi:10.1371/journal.pone.0046145.g001

($p<10^{-15}$), AM second best ($p=3.8\cdot10^{-6}$ compared to SW), and RM and SW were roughly the same ($p=0.18$). Hence, BE showed consistently best performance in all three measures, followed by AM. Interestingly, BE gave the poorest AUC and Spearman scores on a few samples. On larger probe sets (**Figure 4**), BE performed best, followed by RM. The AUC and sensitivity criteria deteriorated for all methods, as expected due to the increasing difficulty in ranking lower-affinity probes. The Spearman score improvement results from its bias to larger sets, so it is more meaningful for comparison of sets of similar sizes. Full results are available in **Table S2**.

Predicting *in vivo* binding intensities

Since PBM and SELEX are *in vitro* assays, which may introduce biases, we also tested the methods' abilities to predict binding intensities for *in vivo* experiments. Our evaluation included ChIP-chip datasets of 32 yeast TFs (69 experiments) that had also PBM profiles [19,20]. A PWM learned according to the profiles of both PBMs (when available) is tested against the data from a ChIP-chip experiment. To evaluate the prediction on the high intensity promoters, where binding is expected to be strongest, we used the positive promoter set as those with reported p-values below 0.001.

All methods performed quite similarly on the AUC and Spearman rank coefficient criteria (**Figure 5**). Using the sensitivity measure, SW was better than the other three ($p<0.02$), AM and BE were roughly the same ($p=0.39$) and significantly better than RM ($p<0.04$). Hence, SW showed consistently best performance in all three measures, while AM and BE were second best (**Table 2** and **Figure 4**). Full results are available in **Table S3**.

Running times

We ran each method on the same 10 examples using a single core of an Intel® Xeon® CPU E5410 @ 2.33 GHz, with 6 MB of cache and 16 GB of memory. On average, AM runs for 30 seconds (including pre-processing), while BE, RM and SW run for about 15 minutes, one hour and more than two hours,

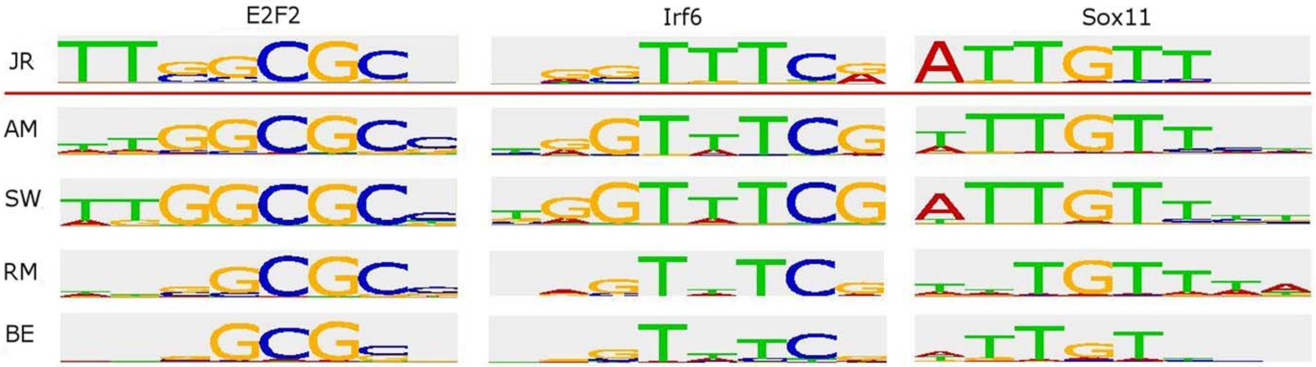


Figure 2. Examples of generated motifs. The figure shows examples of the motifs produced by each method and the corresponding JASPAR motif. For three proteins, the PWM logos produced by each method and the experimentally and independently established motif in the JASPAR database are shown. AM was trained on motif length 8, while for BE, RM and SW only the most informative contiguous positions were kept. We chose TFs whose motifs had information content most similar to the averages of the different methods.
doi:10.1371/journal.pone.0046145.g002

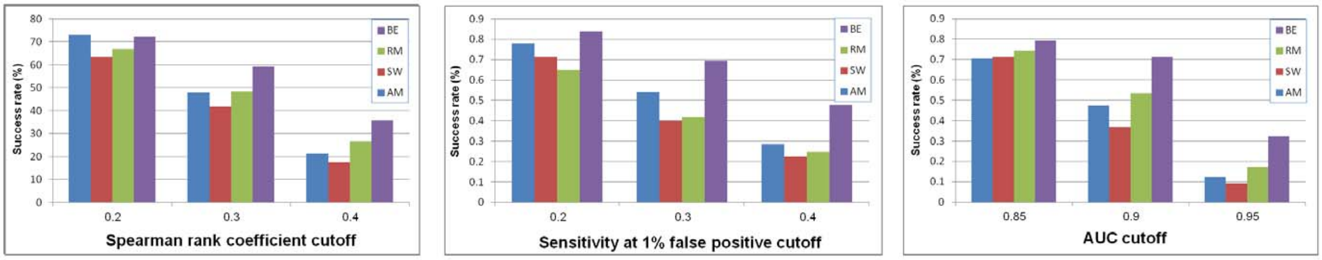


Figure 3. Success rates in probe ranking of a paired PBM. For each TF and method, the PWM was learned using one array and used to infer probe intensity ranking in its paired array. Ranking was gauged on a set of top positive probes (4σ set) according to three measures: Spearman rank coefficient, sensitivity at 1% false positive and AUC (see **Methods S1** for all mathematical terms). For each quality measure, three distance cutoffs were used, and the fraction of TFs with score equal or better to the cutoff is the success rate. The results show the success rate over 230 samples (115 paired arrays).

doi:10.1371/journal.pone.0046145.g003

respectively (**Table 2**). BE currently uses SW results as seeds, thus SW's running time should be added to the total running time of BE. Hence, AM provides a speedup by a factor of 30–200.

Similarity between the algorithms

We evaluated the similarity between the PWMs produced by the four algorithms (**Figure 6A**). In terms of PWM distance, the pairs AM/SW and RM/BE were more similar than others. Note that the comparison is not symmetrical, since it uses the eight most informative contiguous positions in the first PWM (corresponding to a column in the table). Large asymmetries (e.g., SW-RM and RM-SW) reflect the fact that these positions are not clearly detectable in RM and BE PWMs (see also **Figure 2**). On average, the distance between PWMs from different methods is similar to the distance between these and the reference PWMs (**Table 2**).

We also compared the probe ranking that the PWMs of the different algorithms induce (**Figure 6B–D**). We used a PWM inferred by one algorithm on a PBM to rank the probe set of the paired PBM, and measured sensitivity and AUC for these probes ranking produced by another algorithm. Results tended to show more symmetry, with pairs involving BE obtaining best scores, in agreement with the good performance of BE in ranking (**Figure 3** and **Table 2**). Additionally, we focused on rankings of the 4σ probe set and compared them using Spearman rank coefficient. PWMs inferred by two algorithms on a PBM to rank the 4σ probe set of the paired PBM, and compared the two rankings using Spearman score. Again pairs with BE got the highest scores, and remarkably, all pair scores were much higher than their similarity scores to original binding intensities (Spearman rank coefficient, 0.5–0.6 compared to 0.24–0.31, respectively).

Discussion

We have described an assessment of four tools for extracting binding site motifs from PBM data. All four methods report their results in the form of a positional weight matrix (PWM). **Table 2** summarizes the comparison. All tools were run with their recommended default parameters; tuning the parameters could improve the results of some methods and affect the relative ranking in our test criteria.

The reference motifs stored in databases are strongly dependent on experimental sources. Most TRANSFAC and JASPAR motifs that we used were created based on SELEX, an *in vitro* assay of limited accuracy and throughput. Still, the relative performance of the methods was essentially the same when tested on three different databases of two species, which indicates robustness of our conclusions.

The best results in similarity of reference mouse motifs to predicted motifs from PBMs (**Figure 1**) were comparable to the similarity of reference metazoan motifs to predicted motifs obtained using a state-of-the-art motif finder that uses promoter sequences [15]. On one hand, PBM profiles cover the spectrum of possible sequences more comprehensively. On the other hand, they include only relatively short motifs. To conclude, no clear winner has yet emerged between PBM technology and traditional motif finding methods in finding PWMs that are closest to reference motifs.

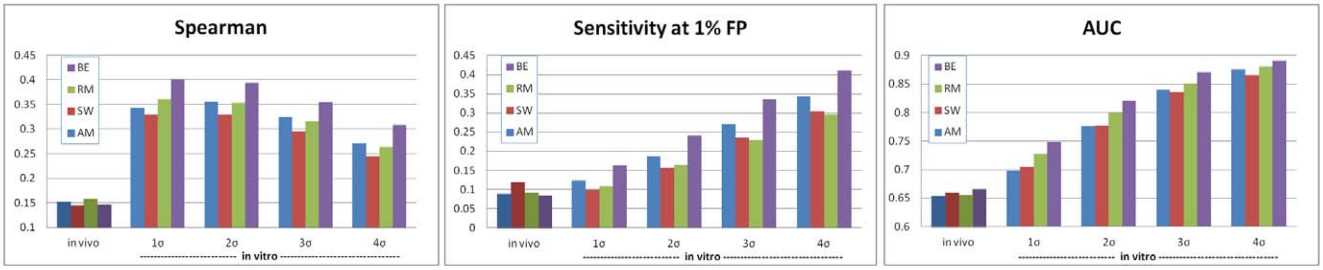
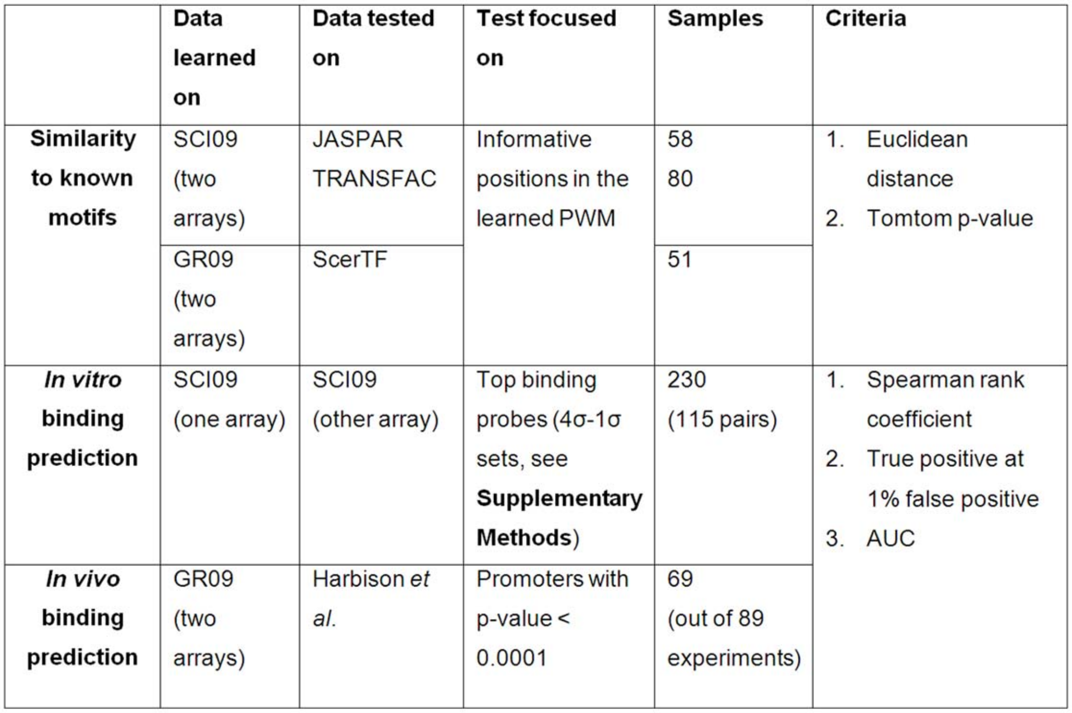


Figure 4. Quality of binding prediction for *in vivo* and *in vitro* data of different sizes. For each of the four algorithms, the quality of the motifs inferred from PBMs in ranking the top binding probes as measured *in vivo* (by ChIP-chip experiments) and *in vitro* (by PBMs) was evaluated. The *in vivo* test included 69 yeast ChIP-chip experiments data (with an average of 61 promoters per experiment). The *in vitro* test included 230 mouse PBMs covering 115 TFs, and used several definitions for the sets of top binding promoter sequences (4σ to 1σ, with averages of 912, 1580, 3215 and 8224 top probes, respectively, see text). Ranking quality was measured by the Spearman rank coefficient, the sensitivity at 1% false positive (FP) and the area under the ROC curve (AUC) (see **Methods S1**). The average ranking quality is reported in each case.

doi:10.1371/journal.pone.0046145.g004

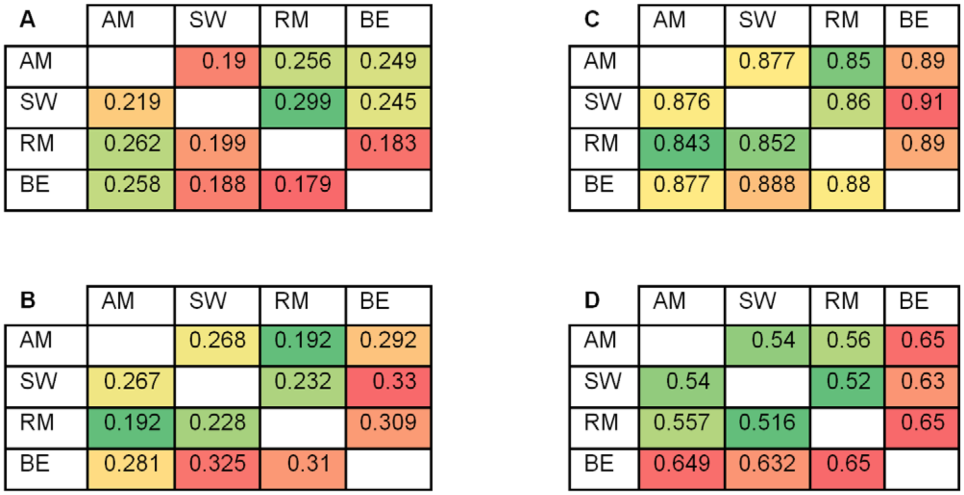


	Data learned on	Data tested on	Test focused on	Samples	Criteria
Similarity to known motifs	SCI09 (two arrays)	JASPAR TRANSFAC	Informative positions in the learned PWM	58 80	1. Euclidean distance 2. Tomtom p-value
	GR09 (two arrays)	ScerTF		51	
***In vitro* binding prediction**	SCI09 (one array)	SCI09 (other array)	Top binding probes (4σ-1σ sets, see **Supplementary Methods**)	230 (115 pairs)	1. Spearman rank coefficient 2. True positive at 1% false positive 3. AUC
***In vivo* binding prediction**	GR09 (two arrays)	Harbison *et al.*	Promoters with p-value < 0.0001	69 (out of 89 experiments)	

Figure 5. Test data and evaluation criteria. The table lists the data and evaluation criteria used in each benchmark.
doi:10.1371/journal.pone.0046145.g005

When using binding intensities of one PBM as input and predicting the ranking of probe intensities of another array for the same TF, BE showed best performance. When using PBM binding intensities to predict ranking of promoter intensities in a ChIP-chip experiment for the same TF, SW performed best. We note that there is still only a modest number of TFs with data from both ChIP-chip and PBM; a larger benchmark for *in vivo* prediction, containing also TF binding in metazoans, is needed.

The performance results can be explained by the different goals of the algorithms. RM was designed to optimally rank all probes, so it tries to capture both high-affinity and low-affinity binding information. This explains why it performs less accurately when analyzing the top-binding probes but performs better on very large



A	AM	SW	RM	BE
AM		0.19	0.256	0.249
SW	0.219		0.299	0.245
RM	0.262	0.199		0.183
BE	0.258	0.188	0.179	

B	AM	SW	RM	BE
AM		0.268	0.192	0.292
SW	0.267		0.232	0.33
RM	0.192	0.228		0.309
BE	0.281	0.325	0.31	

C	AM	SW	RM	BE
AM		0.877	0.85	0.89
SW	0.876		0.86	0.91
RM	0.843	0.852		0.89
BE	0.877	0.888	0.88	

D	AM	SW	RM	BE
AM		0.54	0.56	0.65
SW	0.54		0.52	0.63
RM	0.557	0.516		0.65
BE	0.649	0.632	0.65	

Figure 6. Similarity between methods. (A) For each pair of methods, the Euclidean distance between the PWMs of the two methods is reported. Before the comparison, the column method's PWM is trimmed to eight most informative contiguous positions. (B–D) ranking based comparisons. For each pair of methods, the probe ranking defined according to the column's method is used as reference, and the ranking of the row's method is evaluated using AUC (B) and sensitivity at 1% false positive (C). In (D), for each pair of methods, the 4σ positive sets of the paired PBM are first ranked by each method, and the Spearman rank coefficient of those rankings is computed. In all tables, the average over 230 PBM experiments is reported. Red colour corresponds to greater similarity.
doi:10.1371/journal.pone.0046145.g006

positive sets (**Figure 4**). The same applies to BE. The inclusion of information from low-intensity binding yields better ranking of low-affinity binding probes, but creates PWMs with lower information content (**Figure 2**). In contrast, AM was designed to identify specific binding motifs; it trains only on the 1000 top-binding 9-mers, and so it only uses information on the specific binding of the protein. Interestingly, SW is best for *in vivo* binding, hinting that longer motifs with a stringent core might be better for this data.

The comparison of the prediction results for *in vitro* and *in vivo* data (**Figure 4**) is striking: The quality of the results is much poorer on *in vivo* data, according to all evaluation criteria (similar results were reported in [21]). This is in spite of the fact that the *in vivo* data consisted of yeast motifs, which are easier to find than mice motifs [5,15]. There can be several explanations of this finding:

1. The length of the probes on the PBM (36 bp) is much shorter than the whole yeast promoters targeted by ChIP-chip (an average of 474 bp). As a result, scoring and ranking yeast promoters is harder.
2. Biases caused by the PBM technology lead to systematic distortion in the reconstructed motifs, compared to *in vivo* motifs. If this is the case, revealing and correcting these biases is essential for using the motifs for *in vivo* analysis.
3. The methods tailored specifically for PBMs may overfit this type of data.
4. The complexity of *in vivo* assays distorts the raw binding signals, which look more like the PBM-based motifs in a cleaner *in vitro* environment.

One interesting phenomenon we encountered was secondary motifs: For some PBMs, SW and AM identified a second, completely different motif in addition to the primary one (**Figure S2**). This phenomenon was first reported in [14]. Agius *et al.* suggested that the secondary binding motifs arise as an artefact of the PBM experiment [8]. Zhao and Stormo suggested that secondary motifs are a result of a biased analysis of the PBM data [10], but Morris *et al.* challenge this conclusion [22]. We tested the benefit of using primary and secondary motifs discovered by SW for *in vitro* binding prediction. While there was a significant improvement in performance, it was still worse than BE (data not shown). Jauch et al. recently obtained a crystal structure of the TF Sox4 domain bound to DNA and concluded that two positions in the binding motif are dependent [23]. Such dependency can be manifested by two PWM motifs. Indeed, SW and to some extent AM recover two motifs that reflect this dependence (**Figure S3**). We agree with the conclusion in [21] that more matching PBM and *in vivo* datasets are needed in order to shed more light on this phenomenon.

An interesting insight arises from the comparison of the methods (**Figure 6D**). In terms of the Spearman score of probe ranking, all methods are much more similar to each other than to the true binding intensities. This suggests that all methods capture similar information, while missing other pertinent effects (e.g., background or technological biases). On the other hand, predicting the top probes of another method was harder than finding true positive probes (**Figure 6D**). Overall, BE had highest pairwise ranking-based scores, concordant with our conclusion that it predicts true binding best (**Table 2**). In terms of distance between PWMs, higher similarities between AM and SW, and between BE and RM, reflect the observation that the former pair produce clear, stringent motifs, while the latter generate more variable, ranking-oriented motifs.

Protein-DNA interactions can occur in a broad range of intensities, and involve both specific and low-affinity (less specific) binding. PBM data enable analysis of the full spectrum of DNA binding affinities of a TF. The binding specificity of a protein can be represented using various models, which differ in expressiveness, compactness, redundancy and interpretability. Our analysis suggests that a PWM models the specific *in vitro* binding quite accurately, obtaining an average AUC of 0.9 on the top probes. The fact that results of all methods tend to deteriorate as the positive sets grow (**Figure 4**), and the success of more complex models in ranking [8] suggest that less specific binding may be better captured by other models. The lower success of all methods in predicting *in vivo* binding questions the transformability of PBM-based results to the *in vivo* domain. Deeper analyses using more data are required on this point.

Our study gauged performance using three criteria: similarity to reference literature motifs, and ability to rank *in vitro* and *in vivo* bindings. The tested methods show a tradeoff between ranking quality and motif similarity. Degenerate motifs are better at *in vitro* binding prediction at the cost of lower information content and similarity to literature motifs. Potential improvement may be achieved by novel methods that strive to optimize both criteria simultaneously.

Materials and Methods

Algorithms

We compared four algorithms: Seed-and-Wobble (SW) [4], RankMotif++ (RM) [9], BEEML-PBM (BE) [10] and Amadeus-PBM (AM), a new algorithm presented here (see **Methods S1**). The computational approaches of the algorithms are summarized in **Table 1**. Software for BE, RM and SW was downloaded from the authors' websites and run using the default parameters. The full details are in **Methods S1**.

PBM data

We downloaded PBM data from UniPROBE [13]. This database contains, for each TF, paired probe intensity profiles measured on two different arrays. We used the SCI09 dataset, which contains paired profiles of 115 mouse proteins [13,14], and the GR09 dataset, which contains profiles of 89 yeast TFs [20] (**Figure 5**).

Reference PWM data

To compare predicted PWMs to experimentally obtained PWMs, we used three databases of reference PWMs: JASPAR [12] and TRANSFAC [16] for mouse motifs and the new yeast motif database ScerTF [17] (**Figure 5**). We included in the comparison only reference PWMs that were produced without using PBM data.

ChIP-chip data

We downloaded the ChIP-chip data for yeast TFs from Harbison *et al.* [19]. These data provide large-scale *in vivo* binding for many TFs. Our test used 69 experiments (32 TFs) that had PBM profiles in UniPROBE as well as ChIP-chip measurements.

Comparison and evaluation

We tested the quality of PWMs produced by each method in three ways: by comparison to reference PWMs from the literature (mostly SELEX-based), by their accuracy in predicting *in vitro* binding in PBMs, and by their accuracy in predicting *in vivo* binding as measured by ChIP-chip. In addition, we evaluated how similar the methods are in a pairwise comparison using the same criteria.

To compare a predicted PWM to a reference one, the Euclidean distance between the two PWMs was calculated, as in [15] (for a description of all evaluation criteria see **Methods S1**). The information content of each matrix was also measured in order to evaluate its degeneracy. Each algorithm was trained using the data from both arrays for the same TF. PWMs were also compared using the Tomtom algorithm [24].

For testing the quality of *in vitro* binding prediction, we followed the method of [9]. Since two (paired) binding profiles were available for each TF, a PWM was trained on one profile (the "training array") and used to rank the probes in the other profile (the "test array"). Given a PWM, the probes of the test array were ranked using the sum occupancy score (see **Methods S1**). This ranking was compared to the measured ranking of the probes in the test array according to three criteria: Spearman rank coefficient, sensitivity at 1% false positive rate and area under the ROC curve (AUC) (see **Methods S1** for all definitions). The comparison was done on the probes that showed high binding intensity in the test array (the positive probe set [9]).

To test the quality of *in vivo* binding predictions, we used similar criteria. For each TF, we trained each method using both paired binding profiles (when available) and tested how well the method predicts the ranking of the strongest bound yeast promoters (see **Methods S1**). Predicted and experimental rankings were compared using the same three criteria.

In computing similarity between different methods, we used four criteria. First, we measured the distance between the PWMs inferred by each method. Second, for each method, using the PWM learned on one array, we ranked the set of positive probes in the paired array, and then measured the Spearman rank coefficient between the rankings of each two methods. Third and fourth, we used one method to rank the probes of the paired array, and tested the prediction of the other method using sensitivity at 1% false positive and AUC (see **Methods S1** for computational details).

Statistical significance of the comparison

For each comparison we evaluated its significance using the Wilcoxon rank-sum test [25]. Since the gauged measurements do not distribute normally, we used a non-parametric statistical test.

Supporting Information

Figure S1 Similarity to experimentally established PWMs. (A) TRANSFAC motifs. For 80 proteins available in TRANSFAC we compared the motifs produced from their PBM data by each of the tested methods to the motif available in TRANSFAC. Distance was measured using Euclidean distance. Three distance cutoffs were used, 0.12, 0.18 and 0.24, and the fraction of recovered motifs with distance below the cutoff is the success rate. (B): ScerTF motifs. The same tests on 51 motifs from the ScerTF database. AM: Amadeus-PBM; SW: Seed&Wobble; RM: Rankmotif++; BE: BEEML-PBM.
(TIF)

Figure S2 Shadow motifs. Examples of the primary and secondary motifs found by Amadeus for Pou2f3 (A) and Sox1 (B). p-values for the motif enrichment (hypergeometric score) are indicated above each motif. Note that even the second ranked motifs obtain extremely high significance.
(TIF)

Figure S3 Sox4 primary and secondary motifs as found by Seed-and-Wobble (SW) and Amadeus-PBM (AM). Jauch et al. reported two motifs: CTTTGTT and AATTGTT (23). (A) The two top motifs recovered by AM. The first motif of Jauch et al. was recovered correctly; the second was partially recovered. (B) The two top motifs recovered by SW. Both motifs from Jauch et al. were inferred correctly. Logos taken from UniPROBE database (13).
(TIF)

Table S1 Results of each of the four methods on different reference motifs from the literature. Each line gives the Euclidean distance between a PWM learned on PBM data and a PWM from another source. On the right-hand side, TOMTOM results are reported, giving the statistical significance of PWM similarity.
(XLS)

Table S2 Results of each of the four methods on SCI09 PBM dataset for different positive probe set sizes (4sigma to 1sigma). Each 2 consecutive lines refer to the paired PBM version of the same TF. The one listed under "PBM training data" is used for training, and the scores reported are for testing on the other one.
(XLS)

Table S3 Results of each of the four methods on Harbison *et al.* dataset. Each line gives the result of *in vivo* binding prediction on data taken from on experiment.
(XLS)

Methods S1 Supplementary methods and results.
(DOC)

Acknowledgments

We are grateful to Quaid Morris for providing us with results of SW and RM runs on the SCI09 data, including trimming uninformative positions. We thank Gary Stormo and Yue Zhao for help with using the BEEML-PBM software and Hershel Safer for comments on the manuscript.

Author Contributions

Conceived and designed the experiments: YO CL RS. Performed the experiments: YO. Analyzed the data: YO CL RS. Contributed reagents/materials/analysis tools: YO CL. Wrote the paper: YO RS.

References

1. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28–36.
2. Aparicio O, Geisberg JV, Struhl K (2004) Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. Curr Protoc Cell Biol Chapter 17: Unit 17 17.
3. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497–1502.
4. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, 3rd, et al. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol 24: 1429–1435.
5. Li N, Tompa M (2006) Analysis of computational approaches for motif discovery. Algorithms Mol Biol 1: 8.
6. Das MK, Dai HK (2007) A survey of DNA motif finding algorithms. BMC Bioinformatics 8 Suppl 7: S21.
7. Sandve GK, Drablos F (2006) A survey of motif discovery methods in an integrated framework. Biol Direct 1: 11.
8. Agius P, Arvey A, Chang W, Noble WS, Leslie C (2010) High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions. PLoS Comput Biol 6.

9. Chen X, Hughes TR, Morris Q (2007) RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. Bioinformatics 23: i72–79.
10. Zhao Y, Stormo GD (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nature Biotechnology 29: 480–483.
11. Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23: 137–144.
12. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao XB, et al. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Research 38: D105–D110.
13. Newburger DE, Bulyk ML (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res 37: D77–82.
14. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, et al. (2009) Diversity and complexity in DNA recognition by transcription factors. Science 324: 1720–1723.
15. Linhart C, Halperin Y, Shamir R (2008) Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets. Genome Res 18: 1180–1189.
16. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, et al. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374–378.
17. Spivak AT, Stormo GD (2012) ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. Nucleic Acids Res 40: D162–168.
18. Tanaka E, Bailey T, Grant CE, Noble WS, Keich U (2011) Improved similarity scores for comparing motifs. Bioinformatics 27: 1603–1609.
19. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99–104.
20. Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, et al. (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. Genome Res 19: 556–566.
21. Gordan R, Murphy KF, McCord RP, Zhu C, Vedenko A, et al. (2011) Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. Genome Biol 12: R125.
22. Morris Q, Bulyk ML, Hughes TR (2011) Jury remains out on simple models of transcription factor specificity. Nat Biotechnol 29: 483–484.
23. Jauch R, Ng CK, Narasimhan K, Kolatkar PR (2012) The crystal structure of the Sox4 HMG domain-DNA complex suggests a mechanism for positional interdependence in DNA recognition. Biochem J 443: 39–47.
24. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS (2007) Quantifying similarity between motifs. Genome Biol 8: R24.
25. Fay MP, Proschan MA (2010) Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Stat Surv 4: 1–39.

3. RAP: Accurate and fast motif finding based on protein-binding microarray data

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 20, Number 5, 2013
© Mary Ann Liebert, Inc.
Pp. 375–382
DOI: 10.1089/cmb.2012.0253

RAP: Accurate and Fast Motif Finding Based on Protein-Binding Microarray Data

YARON ORENSTEIN, ERAN MICK, and RON SHAMIR

ABSTRACT

The novel high-throughput technology of protein-binding microarrays (PBMs) measures binding intensity of a transcription factor to thousands of DNA probe sequences. Several algorithms have been developed to extract binding-site motifs from these data. Such motifs are commonly represented by positional weight matrices. Previous studies have shown that the motifs produced by these algorithms are either accurate in predicting *in vitro* binding or similar to previously published motifs, but not both. In this work, we present a new simple algorithm to infer binding-site motifs from PBM data. It outperforms prior art both in predicting *in vitro* binding and in producing motifs similar to literature motifs. Our results challenge previous claims that motifs with lower information content are better models for transcription-factor binding specificity. Moreover, we tested the effect of motif length and side positions flanking the "core" motif in the binding site. We show that side positions have a significant effect and should not be removed, as commonly done. A large drop in the results quality of all methods is observed between *in vitro* and *in vivo* binding prediction. The software is available on acgt.cs.tau.ac.il/rap.

Key words: motif finding, protein-binding microarray, protein-binding site.

1. INTRODUCTION

Gene expression is regulated mainly by proteins that bind to short DNA segments. These proteins, termed transcription factors (TFs), bind to short DNA sequences with variable affinity. These sequences, called binding sites (BSs), are usually found upstream to the gene transcription start site. This TF-BS binding regulates gene expression, either by encouraging or impeding gene transcription.

Many technologies have been developed to measure the binding of TFs to DNA sequences. Chromatin immunoprecipitation (ChIP) extracts bound DNA segments, which are then either hybridized to a predesigned DNA microarray (Aparicio et al., 2004) or directly sequenced (Johnson et al., 2007; Rhee and Pugh, 2011). These technologies can produce reasonably accurate *in vivo* binding profiles. However, they present some difficulties. The binding is tested against genomic sequences only, which have sequence biases (e.g., they do not cover all k-mers uniformly and thus can affect constructed models). In addition, many binding events are due to cooperative binding by more than one TF. Moreover, accurate modeling of these binding events must account for other significant factors that affect binding, such as nucleosome occupancy and chromatin state.

Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel.

In vitro technologies, such as protein-binding microarray (PBM) (Berger et al., 2006) and MITOMI (Fordyce et al., 2010), measure binding of a TF to thousands of synthesized probe sequences. The sequences are designed to cover all DNA k-mers and so give an unbiased measurement of TF binding to a wide spectrum of sequences. The binding is due to TF affinity without additional binding effects found *in vivo* (albeit, with some technological biases). Current implementations of PBMs cover all DNA 10-mers and are available in two different array designs. Another technology, based on high-throughput sequencing, measures binding to random k-mers, with complete coverage of all 12-mers (Nutiu et al., 2011). The latter study showed that some TFs bind to motifs of length greater than 10 and emphasized the importance of greater k-mer coverage.

Several algorithms were developed for the specific task of learning binding-site motifs from protein-binding microarray data. These include Seed-and-Wobble (SW) (Berger et al., 2006), RankMotif++ (RM) (Chen et al., 2007), and BEEML-PBM (BE) (Zhao and Stormo, 2011). All produce the binding-site motif as a position weight matrix (PWM). For each position, the binding preference is given by a probability distribution over four nucleotides. Agius et al. (2010) developed a much more complex model based on a collection of 13-mers, but we will focus here on the PWM model, which is far more common and transparent. A previous study by our group compared these different methods using several evaluation criteria (Orenstein et al., 2012). Weirauch et al. (2012) compared methods for TF-binding prediction using PBMs. A key observation that emerged from both studies is a dichotomy of current motif construction methods: Some produce motifs that accurately predict *in vitro* binding; other methods produce motifs with higher information content that are more similar to literature motifs. No method performed well in both tasks.

The current state of affairs of PBM-based motif prediction raises several questions: Can one develop a method that produces motifs that are both similar to literature motifs and accurately predict *in vitro* binding? What is the best model for TF-binding preference? Is it a PWM with low or high information content motifs? What is the best length of the binding site that can be learned from PBM data?

In this study, we address all these questions. We developed a new simple method to extract binding-site motifs, represented in PWM format, from PBM data. In spite of its simplicity, the method produces motifs that achieve top performance both in predicting *in vitro* binding and in similarity to known motifs. By comparing the performance of motifs of different lengths we conclude that longer motifs are better and that inclusion of flanking positions—even with relatively low information—has a positive effect on predicting binding affinity. We also give evidence to a large gap between the quality of *in vitro* and *in vivo* binding prediction.

2. RESULTS

2.1. *The RAP algorithm*

We developed a new method for finding binding-site motifs using PBM data. The method works in four phases. (1) *Ranking phase*: rank all 8-mers by the average binding intensity of the probes in which they appear. (2) *Alignment phase*: align the top 500 8-mers to the top-scoring 8-mer using star alignment (Altschul and Lipman, 1989). 8-mers must align with an overlap of at least five positions, at least four matches, and at least three consecutive matches, otherwise, they are discarded. (3) *PWM phase*: use the aligned 8-mers to build a PWM. The core matrix is of length 8. In each column of the PWM, the nucleotide probabilities are calculated according to a weighted count in the corresponding column of the alignment. (4) *Extension phase:* the matrix is extended to both sides according to the original probes that contain each of the aligned 8-mers. In each peripheral position, the probe sequences and their scores are used to calculate nucleotide probabilities in a similar fashion as for the core positions. The method is called RAP (for rank, align, PWM). Its running time is less than 2 seconds for one PBM data file, where most of the time is needed to read the file.

2.2. *Performance comparison: predicting* in vitro *binding*

We tested RAP, SW (Berger et al., 2006), RM (Chen et al., 2007), and BE (Zhao and Stormo, 2011) in predicting high-affinity binding. Most TFs studied by PBMs to date were measured in a pair of experiments using two different array designs. This allows an elegant way to test performance, as suggested in Chen

Table 1. Predicting *In Vitro* Binding

Method/criterion	*AUC*	*TP1FP*	*Spearman*
RAP	0.880	0.435	0.293
BEEML-PBM	0.873	0.418	0.283
Seed-and-Wobble	0.858	0.332	0.239
RankMotif++	0.869	0.292	0.245

The table shows average results in three different criteria for each method over 316 PBM pairs. In each experiment, a PWM was learned using one array and then used to rank probes of its paired array. This ranking was compared to the original probe ranking using AUC, sensitivity at 1% false positive (TP1FP), and Spearman rank coefficient. AUC, area under the curve; PWM, position weight matrix; RAP, rank, align, PWM; PBM, protein-binding microarray.

et al. (2007): The binding site is learned according to one array and tested on the other. For this test, we used all TFs in the UniPROBE database that had such paired experiments. PBM experiments that had a positive set of less than 20 probes (see Methods) were excluded from the testing results. In total, the results reported below cover 316 PBM experiments.

A PWM learned using one array was used to rank the probes of its paired array. This ranking was compared to the ranking according to true binding intensity using three criteria: area under the ROC curve (AUC), sensitivity at 1% false positive (TP1FP), and Spearman rank coefficient (see Methods). RAP achieves best average performance in all criteria, followed by BE, RM, and SW in this order (Table 1). The advantage of RAP over BE is not significant in all criteria, while the advantage of both RAP and BE over RM and SW is significant ($p < 0.05$, Wilcoxon rank-sum test). In terms of median performance, BE is slightly better than RAP. Figure 1 shows a dot plot comparison of RAP and BE.

2.3. Performance comparison: similarity to literature motifs

We compared motifs learned by the different methods to motifs learned from non-PBM technologies. We used 58 mouse and 51 yeast PWMs taken from the JASPAR (Bryne et al., 2008) and ScerTF (Spivak and Stormo, 2012) databases, respectively, and calculated the similarity to PWMs learned by the different methods (on two paired PBM profiles together, when available). We measured dissimilarity using Euclidean distance. In addition, we calculated the average information content (IC) of the PWMs of each method (see Methods). The results are summarized in Table 2.

RAP achieves best similarity, followed closely by SW (p-value = 0.14, Wilcoxon rank-sum test), while RM and BE are far less similar to literature motifs (p-value < 0.0003). SW had the highest average IC (1.33), significantly higher than RAP, RM, and BE in that order. Figure 2a shows a boxplot of similarity to known motifs, with colors depicting the IC of each PWM. On average higher IC correlates with lower Euclidean distance, e.g., about -0.4 correlation for BE and RM. Figure 2b shows examples of PWMs in logo format.

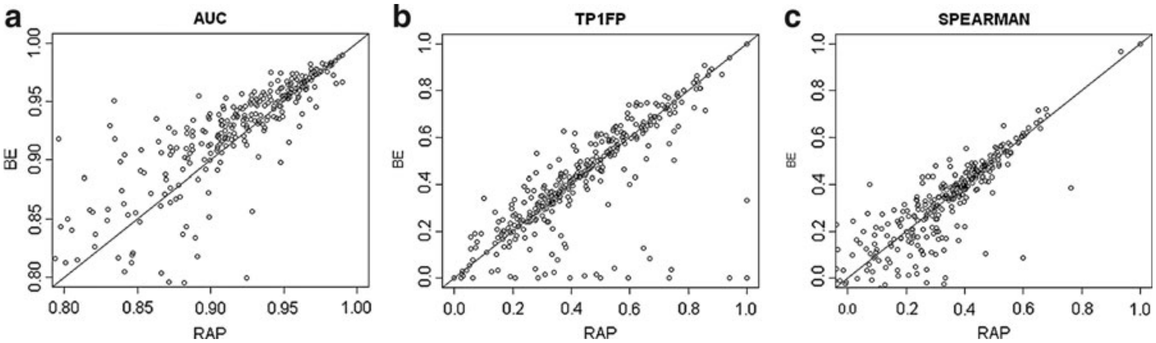


FIG. 1. Comparison of RAP and BEEML-PBM (BE) in predicting *in vitro* binding. Data and performance criteria are as in Table 1. Each dot corresponds to a PBM experiment, and the x- and y-axis are RAP and BE performance results for that experiment, respectively. Note that in the AUC plot experiments with low score are not shown. RAP, rank, align, PWM; PBM, protein-binding microarrays.

Table 2. Dissimilarity to Literature Motifs and Information Content

Method/criterion	*Dissimilarity*	*Information content*
RAP	0.197	0.992
Seed-and-Wobble	0.201	1.330
RankMotif++	0.222	0.884
BEEML-PBM	0.232	0.689

We calculated Euclidean distances between PWMs learned by each method and the corresponding matrices in JASPAR and ScerTF (51 mouse and 58 yeast PWMs, respectively). The table shows average distance for each method. Information content averages are calculated for the PWMs learned by each method.

2.4. *The effect of motif length and flanking sequences*

We tested the effect of motif length and flanking sequences on the ability to predict *in vitro* binding. We took the PWMs produced by the different methods, and for different values of *k*, we kept the *k* contiguous positions with the highest IC. In another test, since different TFs may have different lengths, we also trimmed side positions by using an IC threshold. Figure 3 summarizes the results of both tests.

For most methods, longer motifs are better. The performance of RAP, BE, and RM declined as motif length decreased. On the other hand, SW's performance peaked at length 11 and decreased for longer motifs (Fig. 3a). All four methods did not benefit from trimming flanking positions with low IC (Fig. 3b). Both BE and RM deteriorated sharply as the cutoff increased, since they produce PWMs with low IC (compare Table 2). RAP and SW were barely affected.

2.5. *Predicting* in vivo *binding*

We also tested the performance of the methods in predicting *in vivo* binding. We used the Harbison et al. (2004) data set and its definition of a positive promoter set, focusing on 69 yeast ChIP-chip experiments with corresponding TFs in the UniPROBE database. We used the PWM learned using PBM data to predict ranking of yeast promoter sequences and compared it to the true ranking reported by Harbison et al., using the same three criteria. The results are summarized in Table 3.

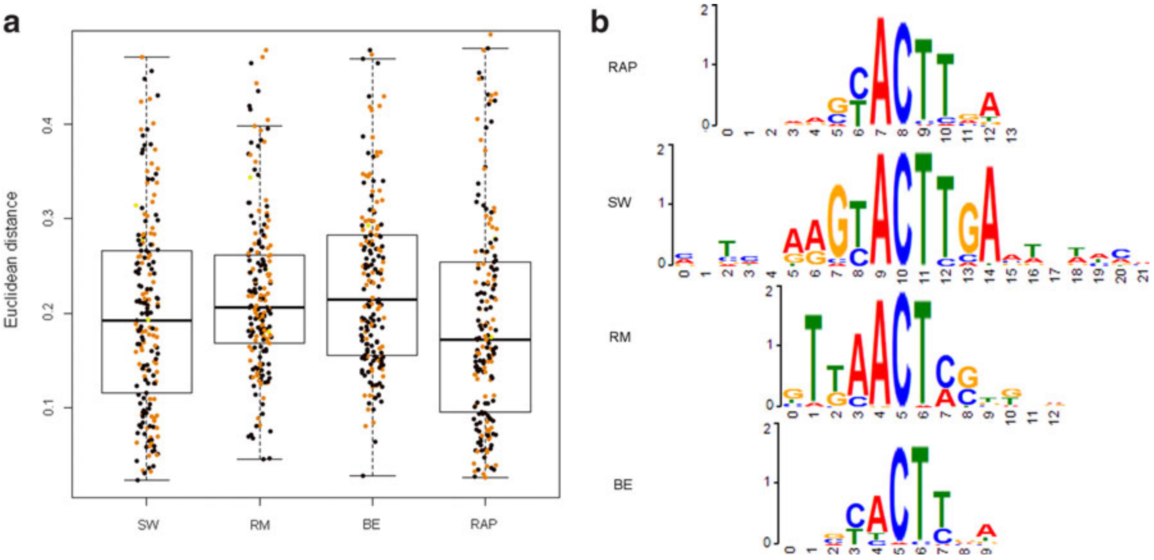


FIG. 2. Dissimilarity to literature motifs and logo comparison. **(a)** Dissimilarity values boxplots. Binding sites were learned by each method, and the Euclidean distance was measured against 109 PWMs from JASPAR and ScerTF. Dots correspond to transcription factors (TFs) where height is the distance and color reflects the information content (IC) of the PWM. Black: IC > 1.3; orange: 1.3 ≥ IC > 0.9; yellow: IC ≤ 0.9. **(b)** PWM logos for Ceh-22 protein.

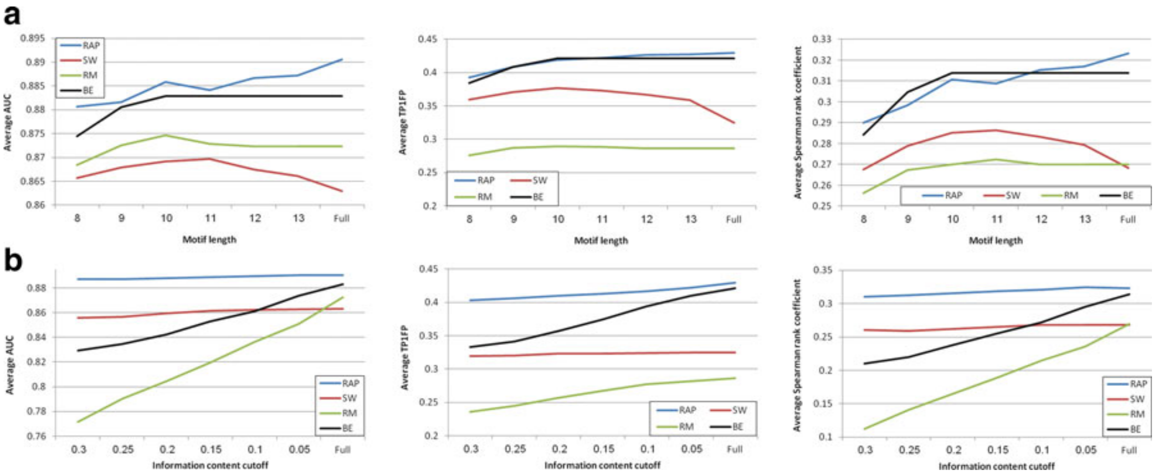


FIG. 3. Effect of motif length and information content on predicting *in vitro* binding. **(a)** Performance as a function of motif length. For each PWM, we kept the *k* most informative contiguous positions and tested the ability of the resulting motif to predict *in vitro* binding. When the motif length was smaller than *k*, we used all positions. Average results of three criteria are shown in the graphs. **(b)** Performance as a function of IC cutoff. For each PWM, we removed all contiguous side positions with IC below the cutoff until reaching the first position with higher IC. The graphs show average results using the same three criteria.

In terms of AUC and Spearman score, all methods performed roughly equally. SW and RM performed slightly (but not significantly) better in the sensitivity and Spearman criterion, respectively. Notably, all methods performed much worse in predicting *in vivo* binding than in predicting *in vitro* binding (compare Table 1).

3. DISCUSSION

We have developed RAP, a new algorithm to extract binding-site motifs in PWM format from protein-binding microarray data. Previous studies observed that algorithms for this task fall into two categories (Orenstein et al., 2012; Weirauch et al., 2012). Some algorithms predict *in vitro* binding well but produce motifs that show low resemblance to motifs reported in the literature. Others match literature motifs (extracted using other technologies) well, but are less successful in *in vitro* binding prediction. This raised the question whether the dichotomy is inevitable. Here we show this is not the case. The RAP algorithm achieved top performance in both criteria. In terms of *in vitro* binding, it is on a par with BE; its motifs are as similar to literature motifs as those of SW. Notably, its running time is a couple of seconds, 2–3 orders of magnitude faster than the other algorithms.

We note that while RAP is slightly better on average than BE, the latter was slightly better in median. For more TFs, BE results are better than RAP's (Fig. 1). But for some, it fails to capture the binding preference correctly. For example, BE achieves AUC < 0.5 for 10 TFs, while only one such case exists for RAP.

TABLE 3. PREDICTING *IN VIVO* BINDING

Method/criterion	*AUC*	*TP1FP*	*Spearman*
RAP	0.662	0.108	0.149
Seed-and-Wobble	0.659	0.118	0.145
RankMotif++	0.655	0.092	0.158
BEEML-PBM	0.665	0.084	0.146

The table shows average results for each method over 69 ChIP-chip experiments. In each experiment, a PWM learned using PBM data was used to rank yeast promoter sequences. This ranking was compared to the original promoter ranking using AUC, sensitivity at 1% false positive (TP1FP), and Spearman rank coefficient.

Hence, BE performs slightly better in more samples, but has a few failures, whereas RAP is robust and produces accurate motifs in almost all cases.

In spite of its very simple algorithm, RAP was shown to be powerful and quite accurate. What explains RAP's performance? Like other methods, it combines information about binding intensities of 8-mers using their occurrences in multiple probes in order to evaluate robustly the 8-mers binding intensity. Unlike other methods, it then focuses solely on the top-binding 8-mers. Star alignment of the top 500 8-mers to the top-ranked one is a simple yet effective way to extract an initial core motif, which is then extended using the original probes. Our tests showed that the use of 500 top 8-mers is optimal, with performance dropping when more k-mers are used. It is possible that a part of the advantage of RAP is gained by focusing on the top 8-mers: They are informative enough to reveal a PWM with good binding-prediction quality, and this approach avoids noise and reduced IC that would be caused by incorporating information from lower intensity probes.

Previous studies suggested that TF-binding preference is best modeled by low IC motifs (cf. Weirauch et al., 2012). This is a natural conjecture derived from the dichotomy of previous methods, since literature motifs tend to have high IC. The RAP algorithm goes against this suggestion: It produces motifs with relatively high IC, which are on par with the best in predicting *in vitro* binding. (SW motifs have substantially higher IC, but they do not perform highly in both criteria).

Our tests of the effect of motif length on performance showed that peripheral positions do affect TF binding. For RM, BE, and RAP, the performance deteriorated as the motif was shortened. Only SW (whose performance was generally lower) did worse for motifs of length ≥ 11. Hence, while the core motifs are easier to comprehend, keeping flanking positions in the model is beneficial. As current PBM techniques are limited to covering all 10-mers (or 12-mers) (Nutiu et al., 2011), producing larger arrays would allow more accurate inference of longer motifs. Our analysis also shows that using IC cutoffs to remove flanking positions is too crude and is particularly damaging to low IC motifs. Hence, both tests suggest that side positions should be kept in the model, in agreement with conclusions reported in Nutiu et al. (2011). To our knowledge, this is the largest-scale rigorous test of the effect of motif length.

Our results show that all algorithms give much poorer prediction on *in vivo* compared to *in vitro* data (Table 3): AUC drops from 0.88 to 0.665, and sensitivity deteriorates from 0.435 to 0.118. While the complexity of the *in vivo* environment may explain this in part, the severity of the gap in the quality of the results questions our ability to carry over the powerful results achievable using PBMs to the natural environment. More complex *in vivo* models that could combine ''naked'' *in vitro* motifs with epigenetic marks and other parameters may help to narrow this gap.

In summary, we developed a new algorithm and showed that it is highly accurate in both predicting *in vitro* binding and producing interpretable motifs. Our results question the claim that TF binding preference is best modeled with low IC motifs and highlight the importance of using long motif models and of learning peripheral positions correctly. Carrying these results over to *in vivo* predictions remains an important challenge.

4. METHODS

4.1. Data

We downloaded all paired PBM profiles from the UniPROBE database (Robasky and Bulyk, 2011), obtaining 364 PBM profiles (182 pairs). From these we removed all PBM profiles, where the size of the positive set (see definition below) on the test array was smaller than 20. This resulted in 316 PBM profiles to test the methods performance.

We compared similarity between motifs learned from PBM data and motifs learned by independent technologies. For this aim, we used 58 mouse TFs that had a PBM profile in the SCI09 study (Badis et al., 2009) and had a model not based on PBM in the JASPAR database (Bryne et al., 2008). Similarly, we collected 51 yeast TFs that had a PBM profile in the GR09 study (Zhu et al., 2009) and were present in ScerTF database (Spivak and Stormo, 2012). The motif was learned using the PBM profile or two paired profiles, when available, and compared against the PWM from the database.

For the *in vivo* binding prediction we used Harbison et al. ChIP-chip dataset (Harbison et al., 2004). The positive promoter set included all promoters with p-value < 0.001 according to Harbison et al. (2004). We used all experiments with a TF in the GR09 dataset (Zhu et al., 2009). This resulted in 69 different experiments.

4.2. Comparison criteria

We tested the ability of each method to predict *in vitro* binding of another PBM array. A binding site in PWM format was learned using one PBM profile. The PWM was used to rank all probe sequences of the paired array. For each probe an occupancy score was calculated, which is the sum of the probability of the TF to bind over all positions (Tanay, 2006). This score is used to rank all probes. For probe sequence *s* and PWM Θ of length *k*, the sum occupancy score is

$$f(s, \Theta) = \sum\nolimits_{t=0}^{|s|-k} \prod\nolimits_{i=1}^{k} \Theta_i[s_{t+i}]$$

where $\Theta_i(x)$ is the probability of base *x* in position *i* of the PWM. The ranking due to the occupancy score is compared to the original probe ranking according to the binding intensity. A positive set of probes is defined as the probes with binding intensity greater than the median by at least 4 * (MAD/0.6745), where MAD is the median absolute deviation (MAD = 0.6745 for the normal distribution *N(0,1)*) (Chen et al., 2007). Three criteria are used to gauge the ranking: AUC of ROC curve, sensitivity (true positive rate) at 1% false positive (TP1FP), and Spearman rank coefficient among the positive set (Orenstein et al., 2012).

For interpretability and motif similarity we used average IC and average Euclidean distance. The IC for vector (v_1, v_2, v_3, v_4) (where $\sum_i v_i = 1$) is defined as $2 + \sum_i v_i \log(v_i)$ (Schneider et al., 1986). The IC for a PWM is the average IC of the most informative eight contiguous positions, since peripheral positions tend to be of low IC and bias the results. To measure the similarity between two motifs we used Euclidean distance (Harbison et al., 2004). For two PWMs, we tried all possible offsets, with an overlap of at least five positions, and chose the one with minimal average Euclidean distance between columns. Motif logos were plotted using http://demo.tinyray.com/weblogo

In vivo binding prediction is tested in the same fashion as for probe ranking. Yeast promoters are ranked according to occupancy score using a PWM learned by one of the methods. The ranking is compared to the original ranking by p-value. AUC, TP1FP, and Spearman rank coefficient are used to gauge the ranking (Orenstein et al., 2012).

4.3. Implementation details

The method is implemented efficiently in Java. Each nucleotide is coded by 2 bits. The 8-mers are kept in a hash table, together with pointers to the original probes in which they appear. The software is available on acgt.cs.tau.ac.il/rap.

ACKNOWLEDGMENTS

This study was supported in part by the European Community's Seventh Framework grant [HEALTH-F4-2009-223575] for the TRIREME project, and by the Israel Science Foundation (grant no. 802/08). Y.O. was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University and by a Dan David PhD Fellowship. We thank David Glauber for his help during the early phase of this study.

DISCLOSURE STATEMENT

All authors state that no competing financial interests exist.

REFERENCES

Agius, P., Arvey, A., Chang, W., et al. 2010. High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions. *PLoS Comput Biol* 6, e1000916.

Altschul, S.F., and Lipman, D.J. 1989. Trees, stars, and multiple biological sequence alignment. *SIAM Journal on Applied Mathematics* 49, 197–209.

Aparicio, O., Geisberg, J.V., and Struhl, K. 2004. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Curr. Protoc. Cell Biol.* Chapter 17, Unit 17 7.

Badis, G., Berger, M.F., Philippakis, A.A., et al. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science* 324, 1720–3.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., et al. 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.* 24, 1429–35.

Bryne, J.C., Valen, E., Tang, M.H., et al. 2008. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res*. 36, D102–6.

Chen, X., Hughes, T.R., and Morris, Q. 2007. RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. Bioinformatics 23, i72–9.

Fordyce, P.M., Gerber, D., Tran, D., et al. 2010. De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nat. Biotechnol.* 28, 970–5.

Harbison, C.T., Gordon, D.B., Lee, T.I., et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99–104.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. Science 316, 1497–502.

Nutiu, R., Friedman, R.C., Luo, S., et al. 2011. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat. Biotechnol.* 29, 659–64.

Orenstein, Y., Linhart, C., and Shamir, R. 2012. Assessment of algorithms for inferring positional weight matrix motifs of transcription factor binding sites using protein binding microarray data. *PLoS ONE* 7, e46145.

Rhee, H.S., and Pugh, B.F. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell 147, 1408–19.

Robasky, K., and Bulyk, M.L. 2011. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res*. 39, D124–8.

Schneider, T.D., Stormo, G.D., Gold, L., and Ehrenfeucht, A. 1986. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* 188, 415–31.

Spivak, A.T., and Stormo, G.D. 2012. ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. Nucleic Acids Res. 40, D162–8.

Tanay, A. 2006. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.* 16, 962–72.

Weirauch, M.T., Cote, A., Norel, R., et al. 2012. Evaluation of methods for modeling transcription factor sequence specificity. To appear in *Nat. Biotechnol*.

Zhao, Y., and Stormo, G.D. 2011. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.* 29, 480–3.

Zhu, C., Byers, K.J., McCord, R.P., et al. 2009. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.* 19, 556–66.

Address correspondence to:
Ron Shamir
Blavatnik School of Computer Science
Tel-Aviv University
P.O.B. 39040
Tel-Aviv
Israel

E-mail: rshamir@tau.ac.il

4. Design of shortest double-stranded DNA sequences covering all k-mers with applications to protein-binding microarrays and synthetic enhancers

BIOINFORMATICS

Vol. 29 ISMB/ECCB 2013, pages i71–i79
doi:10.1093/bioinformatics/btt230

Design of shortest double-stranded DNA sequences covering all *k*-mers with applications to protein-binding microarrays and synthetic enhancers

Yaron Orenstein and Ron Shamir*
Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel

ABSTRACT

Motivation: Novel technologies can generate large sets of short double-stranded DNA sequences that can be used to measure their regulatory effects. Microarrays can measure *in vitro* the binding intensity of a protein to thousands of probes. Synthetic enhancer sequences inserted into an organism's genome allow us to measure *in vivo* the effect of such sequences on the phenotype. In both applications, by using sequence probes that cover all *k*-mers, a comprehensive picture of the effect of all possible short sequences on gene regulation is obtained. The value of *k* that can be used in practice is, however, severely limited by cost and space considerations. A key challenge is, therefore, to cover all *k*-mers with a minimal number of probes. The standard way to do this uses the de Bruijn sequence of length 4^k. However, as probes are double stranded, when a *k*-mer is included in a probe, its reverse complement *k*-mer is accounted for as well.

Results: Here, we show how to efficiently create a shortest possible sequence with the property that it contains each *k*-mer or its reverse complement, but not necessarily both. The length of the resulting sequence approaches half that of the de Bruijn sequence as *k* increases resulting in a more efficient array, which allows covering more longer sequences; alternatively, additional sequences with redundant *k*-mers of interest can be added.

Availability: The software is freely available from our website http://acgt.cs.tau.ac.il/shortcake/.

Contact: rshamir@tau.ac.il

1 INTRODUCTION

Gene regulation is a central focus of biological research. The main factors that regulate gene expression are transcription factors (TFs). These proteins bind to short DNA sequences, either in promoters or enhancers, and by that encourage or impede gene transcription. TFs bind to different DNA sequences with different affinity and specificity. Understanding TF-binding specificity and its effect on gene expression and the final phenotype is a fundamental goal in the study of gene regulation.

Recent technologies measure the binding intensity of a TF to many DNA sequences [e.g. protein-binding microarray (PBM) (Berger *et al*., 2006) and MITOMI [Fordyce *et al*., 2010)]. These technologies synthesize a large set of DNA sequences and measure the binding intensity of the TF to each of those sequences. Some technologies use random DNA sequences (Nutiu *et al*., 2011). Others use sequences that cover all possible DNA *k*-mers, as they provide a complete picture of the binding spectrum (Berger *et al*., 2006; Fordyce *et al*., 2010). A similar approach was also used to test binding *in vivo*. A recent study used synthesized enhancer oligomers designed to cover all 6mers to test their effect on limb formation in zebrafish (Smith and Ahituv, 2012).

De Bruijn sequences are the most compact sequences that cover all *k*-mers (Berger *et al*., 2006; Fordyce *et al*., 2010). The length of a de Bruijn sequence of order k over alphabet $|\Sigma|$ is $|\Sigma|^k$, where the DNA alphabet is $\Sigma = \{A, C, G, T\}$. Because of the exponential dependency on k and small space on the experimental device, these technologies are limited to a small value of k. The most popular technology, PBM, was used in hundreds of experiments to date using arrays with $k = 10$. To create p-long probe sequences, the sequence is split into intervals of length p with $k - 1$ overlap ($p = 36$ is used in PBMs).

Despite the universal and high-throughput nature of these technologies, the data produced are still limited. For many TFs, binding depends on >10 DNA positions, usually with six to eight core positions and additional side positions that have a significant contribution (Nutiu *et al*., 2011; Orenstein *et al*., 2013). A recent study from the Taipale Laboratory using HT-Selex showed that many TFs have longer motifs that are not covered well by an all 10mer array (Jolma *et al*., 2013). The RankMotif++ algorithm for PBM data also generates motifs of length >10 in most cases (Chen *et al*., 2007). Covering all k-mers for a greater value of k will lead to improved understanding of TF binding.

As the probes are double-stranded DNA segments, one can save by using the reverse complementarity of DNA: whenever a k-mer is included, its reverse complement is included as well, and there is no need to cover it again. This brings up the following question: a sequence S is called a *reverse complementary complete sequence* of order k (RC complete sequence for short) if for each k-mer either the k-mer or its reverse complement are included in S. Can we construct an optimal (minimum length) RC complete sequence? Theoretically, if for each k-mer T the sequence S includes either T or its reverse complement but not both, one could save a factor of nearly 2 compared with the length of a de Bruijn sequence.

Ministeris and Eisen (2006) and Philippakis *et al*. (2008) proposed the use of (regular) de Bruijn sequences for designing probes for PBMs. Philippakis *et al*. used linear feedback shift registers to generate a de Bruijn sequence with good coverage of gapped k-mers. This approach was used for constructing two microarrays that are in use today with $k = 10$ (Berger *et al*., 2006). The idea of exploiting reverse complementarity was raised by Ministeris and Eisen (2006), who sketched an algorithm for it without proof. In fact, as we shall show, the algorithm of

*To whom correspondence should be addressed.

Mintseris and Eisen (2006) does not provide an optimal solution for even values of k. In the context of sequence assembly, Medvedev *et al.* (Medvedev and Brudno, 2009; Medvedev *et al.*, 2007) solved the problem of constructing a minimum length sequence that covers a given set of k-mers, using reverse complementarity. Although their algorithm can be applied to solve the problem raised in this study, they do not address it directly. When applied to our problem, their algorithm requires $O(k^2 \log^2(|\Sigma|)|\Sigma|^{2k})$ time. As we shall see, our algorithm is much faster.

In this study, we address the problem of constructing an optimal RC complete sequence. We first give a lower bound for the length of such a sequence. We prove that for odd k, there exists a sequence that achieves the lower bound and show how to construct it in time complexity that is linear in the output sequence length. For odd k, the algorithm constructs two tours that are reverse complementary to each other and together cover all edges of the de Bruijn graph and is identical to Mintseris and Eisen (2006). Then, we show how to adjust the algorithm to handle the case of even k, achieving a saving factor approaching 2 as k increases. We give two solutions: a simple near-optimal one requiring linear time and a more complex ($O(k|\Sigma|^{5k/4}\log(|\Sigma|))$ time) solution that guarantees optimality of the resulting sequence. In particular, this implies that the lower bound is not tight for even k. We implemented the algorithm and we demonstrate the saving it achieves. The produced sequences are nearly half the length compared with a regular de Bruijn sequence.

The article is organized as follows. We first provide formal definitions and preliminaries. We then present a lower bound for the length of an optimal sequence based on k-mer counts. Then, we present an algorithm that works in linear time on the de Bruijn graph and prove that it solves the problem for odd k. We conclude by describing the two possible solutions for even k and report on experimental results with all the algorithms.

2 PRELIMINARIES

We start with some basic definitions of graphs and sequences. For more details see, e.g. West *et al.* (2001).

A *directed graph* (digraph or simply a graph) $G = (V, E)$ is a set of vertices $V = \{v_1, v_2, \ldots, v_n\}$ and a set of edges $E = \{e_1, e_2, \ldots, e_m\}$. Each edge is an ordered pair of vertices (v_i, v_j), and we say the edge is directed from v_i to v_j. The *indegree* of vertex v is the number of edges entering v. Similarly, the *outdegree* is the number of edges outgoing from v. A vertex is *balanced* if its indegree equals its outdegree. A *path* in a digraph is a sequence of vertices, $v_{i_1}, \ldots, v_{i_k}$, such that for each $1 \le j < k$ there is an edge $(v_{i_j}, v_{i_{j+1}})$. A *cycle* is a path where $i_1 = i_k$. A digraph is *strongly connected* if for every pair of vertices u, v there exists a path from u to v and a path from v to u. *A strongly connected component* in a digraph is a maximal set of vertices that induces a strongly connected subgraph.

An *Eulerian tour* through a digraph G is a cycle that traverses all edges in G, such that each edge is traversed exactly once. If a digraph contains an Eulerian tour, we call it *Eulerian*. A digraph is Eulerian if and only if it is strongly connected and all vertices are balanced (West *et al.*, 2001).

A *de Bruijn sequence* of order k over alphabet Σ is a minimum length sequence that covers each k-mer over Σ exactly once. For convenience, we define the *length* of the sequence as the number of k-mers in it. Hence, a sequence of length t contains $t + k - 1$ characters. A de Bruijn sequence has length $|\Sigma|^k$, which is the minimum possible for covering all k-mers.

Given sequences a, b over alphabet Σ, the *overlap* between a and b, denoted $ov(a, b)$, is the largest suffix of a that is also a prefix of b.

A *de Bruijn graph* of order k is a digraph in which for every possible k-mer $x_1, \ldots, x_k$ there is a vertex denoted by $[x_1, \ldots, x_k]$. There is an edge from u to v if and only if $u = [x_1, \ldots, x_k]$ and $v = [x_2, \ldots, x_{k+1}]$, that is, $|ov(u, v)| = k - 1$. Each edge represents a unique $(k + 1)$-mer. For example, the edge (u, v) above represents $(x_1, \ldots, x_{k+1})$. To distinguish vertices from edges, we will use square brackets for vertices. Hence, $(x_1, \ldots, x_{k+1})$ is the edge between $[x_1, \ldots, x_k]$ and $[x_2, \ldots, x_{k+1}]$. Obviously, for each vertex v the indegree and outdegree are $|\Sigma|$, and the graph is strongly connected. Thus, a de Bruijn graph is Eulerian. Any Eulerian tour represents a de Bruijn sequence of order $k+1$. Each edge and vertex in the graph is represented by $O(k \log(|\Sigma|))$ bits. Throughout the article, we assume this number of bits is contained in one computer word; hence, we deduce that it takes $O(1)$ time to find an edge or a vertex.

A *complementarity* relation between characters is a symmetric non-reflexive one-to-one relation. The alphabet of DNA is $\Sigma = \{A, C, G, T\}$ with the complementarity relation $\bar{A} = T$ and $\bar{C} = G$. By symmetry also $\bar{T} = A$ and $\bar{G} = C$. The *reverse complement* of sequence $(x_1, \ldots, x_k)$, denoted $RC(x_1, \ldots, x_k)$, is defined as the sequence obtained by reversing the original sequence and replacing each character by its complement, i.e. $RC(x_1, \ldots, x_k) = (\bar{x}_k, \ldots, \bar{x}_1)$. For example, $RC(CGAA) = TTCG$. A sequence s is called a *palindromic reverse complementary* sequence or in short a *palindrome*, if $s = RC(s)$. For example, $ACGT$ is a palindrome. We define a *reverse complementary complete sequence* of order k over alphabet Σ (RC complete sequence for short) as a sequence such that for each k-mer s, at least one of s and $RC(s)$ are in the sequence. Note that unlike a regular de Bruijn sequence, the definition of an RC complete sequence does not require minimality. An RC complete sequence is *optimal* if it is of minimum length.

3 RESULTS

3.1 A lower bound for the length of an RC complete sequence

First, we derive a lower bound for the length of an RC complete sequence from k-mer counts.

Proposition 1. Denote by $n^*(k)$ the length of an optimal RC complete sequence of order k.

$$n^*(k) \ge \begin{cases} \frac{|\Sigma|^k}{2}, & \text{if } k \text{ is odd} \\ \frac{|\Sigma|^k + |\Sigma|^{k/2}}{2}, & \text{if } k \text{ is even} \end{cases} \quad (1)$$

Proof. We consider separately the cases of odd and even k. For odd k, there are no palindromes, as the middle position in a k-mer differs from its reverse complement. Each k-mer must be represented in the sequence by itself or its reverse complement.

Thus, a lower bound for the minimum length is half the number of unique k-mers, which is $|\Sigma|^k/2$. For even k, some k-mers are palindromes. For palindromes, the first $k/2$ characters define the last $k/2$ characters. Hence, there are exactly $|\Sigma|^{k/2}$ different palindromes. All palindromes must appear at least once in any RC complete sequence, whereas for the non-palindromic k-mers, either they or their reverse complement must appear in the sequence. Thus, for even k, $n^*(k) \geq \frac{|\Sigma|^k - |\Sigma|^{k/2}}{2} + |\Sigma|^{k/2}$. ■

We shall show later that $n^*(k)$ is tight for odd k, but not for even k.

3.2 Constructing an optimal RC complete sequence for odd k

In this section, we prove constructively that for odd k there exists an RC complete sequence that achieves the lower bound of Proposition 1 and is thus optimal. The proof modifies the Euler tour algorithm (West *et al.*, 2001). The modified algorithm was presented without proof in Mintseris and Eisen (2006). The algorithm for generating the sequence will work on the de Bruijn graph of order $k-1$. Every k-mer is represented in the graph as an edge, the graph is strongly connected and all vertices are balanced. As there are no palindromes of odd length, every edge has a unique reverse complement counterpart that is different from it. This defines a perfect matching M on the edges of the graph.

Given a directed path F in the graph, its *reverse complement path* is defined as the path R in which each edge (u, v) in F is replaced by the edge $(\bar{v}, \bar{u})$. For example, for the path $(ACG) \rightarrow (CGG) \rightarrow (GGT)$, its reverse complement is $(ACC) \rightarrow (CCG) \rightarrow (CGT)$ (Fig. 1). We will refer to F and R as forward and reverse paths, respectively.

The following theorem provides a necessary and sufficient condition for the existence of an RC complete sequence that achieves the lower bound.

Theorem 1. *For odd k, an RC complete sequence s achieves the lower bound (Proposition 1) if there exist two edge-disjoint paths with no repeating edges, corresponding to s and RC(s), that together cover all edges of the de Bruijn graph of order $k-1$.*

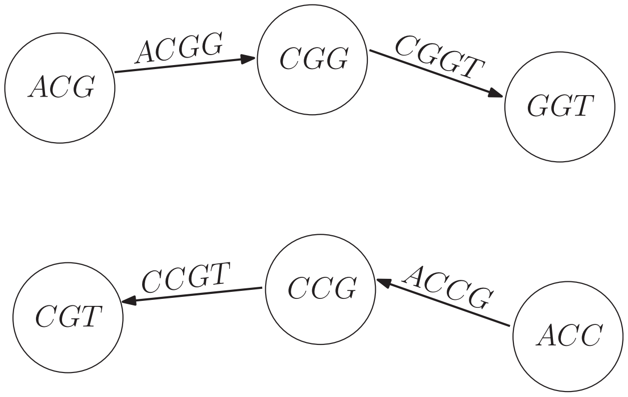


Fig. 1. An illustration of forward and reverse paths (top and bottom, respectively). The forward path traverses the edges in their direction. The corresponding reverse path traverses the reverse complementary edges in reverse direction

Proof. ⇒ Observe that the lower bound assumes one occurrence of either w or $RC(w)$ but not both in the sequence for each k-mer w. Assume an RC complete sequence s^* achieves the lower bound. Then, because of its minimality, it contains no repeating k-mers; therefore, it must correspond to a path F in the de Bruijn graph with no repeating edges. The ordered set of k-mers in s^* corresponds to consecutive edges in F. Note that the reverse complement sequence $t^* = RC(s^*)$ is also a path R in the graph: the k-mers in R are the reverse complement of those in F; therefore, consecutive edges form a path in the graph traversed in reverse order. As each k-mer or its reverse complement is covered in s^*, it is also true that each k-mer or its reverse complement is covered by t^*, and the two paths F and R, corresponding to the two sequences, together cover all edges.

⇐ Suppose there are two edge-disjoint paths F and R with no repeated edges that together cover all edges. As they are reverse complement of each other, and together cover all edges, for each k-mer w, the sequence s (corresponding to path F) must contain either w or $RC(w)$ (otherwise, some edges would have been uncovered). Hence, s is an RC complete sequence. The same argument holds for $RC(s)$ (corresponding to path R). As each contains exactly half the edges, the length of each of them equals the lower bound ■.

Before presenting the algorithm for finding an optimal RC complete sequence, we remind the reader of the algorithm for finding an Eulerian cycle in a digraph (Fleischner, 1990). The algorithm starts from an arbitrary source vertex. Initially all edges are unmarked. It traverses a path of unmarked edges in arbitrary order. Each traversed edge is marked; therefore, no edge is traversed more than once. The algorithm also maintains a set A of the visited vertices that are still active, i.e. they have outgoing unmarked edges. When the last unmarked edge outgoing from a vertex is traversed, the vertex is removed from A. If the algorithm reaches a dead end, it starts another traversal from another vertex in A. A dead end can only be achieved when closing a cycle (i.e. returning to the source vertex), as in any other vertex there is always a free incoming edge and a free outgoing edge (as for every vertex except the source the unmarked outdegree and the unmarked indegree are equal). If not all edges have been traversed, A is not empty, and the process can start from a new source. In the end, as the graph is strongly connected and all cycles start from visited vertices (except for the initial vertex), the cycles can be joined to form one Eulerian cycle. The running time of the algorithm is linear in the number of vertices and edges.

Algorithm 1 finds an optimal RC complete sequence in a de Bruijn graph of order $k-1$ when k is odd. The algorithm imitates the Euler path algorithm but maintains both a forward sequence and a reverse complement sequence simultaneously. The collection of cycles traversed so far is kept in $\mathcal{F}$ and the corresponding reverse complement cycles set is $\mathcal{R}$.

Algorithm 1. Find forward and reverse paths that cover all edges in a de Bruijn graph $G = (V, E)$ of even order $k-1$.

```
(1) Initially all edges are unmarked, $\mathcal{F} = \mathcal{R} = \emptyset$, and $A = \{u\}$,
    an arbitrary vertex.
```

```
(2)  Although $A \neq \emptyset$ do
(3)      $F = R = \emptyset$.
(4)      Pick any starting vertex $v = [x_1, \ldots, x_{k-1}]$ from $A$.
(5)      Although there exists an unmarked edge $e = (x_1, \ldots, x_k)$ outgoing from $v$ do
(6)          Append $e$ to $F$. Prepend $RC(e)$ to $R$.
(7)          Mark $e$ and $RC(e)$.
(8)          Set $v = [x_2, \ldots, x_k]$; $A = A \cup \{v\}$.
(9)      Remove $v$ from $A$.
(10) If $F \neq \emptyset$, add $F$ to $\mathcal{F}$; add $R$ to $\mathcal{R}$;
(11) Merge the cycles in $\mathcal{F}$ to obtain a single forward path.
     Do the same for $\mathcal{R}$.
```

Theorem 2. For odd k, Algorithm 1 returns forward and reverse paths that cover together all edges of the graph and represent two optimal RC complete sequences. The algorithm runs in $O(|V|)$ time.

Proof. We prove the theorem using several lemmas. We first show that if the forward path F reaches a dead end, then so does the reverse path R, and in that case, a cycle is closed (Lemma 1). Note that each pair F, R constructed in Steps 4–7 are reverse complementary paths by the way they are constructed. Then, we show that the cycles in $\mathcal{F}$ can be merged into one cycle (Lemma 2). Third, we deduce that a strongly connected component is covered by $\mathcal{F}$ and $\mathcal{R}$ (Lemma 3). Finally, we conclude that $\mathcal{F}$ and $\mathcal{R}$ cover all edges, as there is only one strongly connected component in any de Bruijn graph (Corollary 1). As each edge is traversed once, the paths are of length $\frac{|\Sigma|^k}{2}$ and, hence, optimal.

Lemma 1. If the forward traversal reaches a dead end, then so does the reverse. Both paths close a cycle in this case.

Proof. Distinguish two cases in which the forward path reaches a dead end:

Case 1. F reaches a vertex v and R reaches a vertex $u \neq v$, and all outgoing edges from v were already traversed. We prove that in that case, F must close a cycle. Assume to the contrary that F contains no edge outgoing from v. In that case, all outgoing edges were traversed by R. Then, all incoming edges must have been traversed by R as well, as each time R reached v, it must have exited it as well. The only exception is if v is also the first (last added) vertex u in R, contradicting our assumption that $u \neq v$. Therefore, all incoming and outgoing edges were covered by R, contradicting the fact that F just entered v. We conclude that F has an edge outgoing from v and thus it closed a cycle.

Denote by $(x_1, \ldots, x_k)$ the last edge traversed by F. All edges of the form $(x_2, \ldots, x_k, a)$, where $a \in \Sigma$, were traversed. Hence, the reverse edges of the form $(\bar{a}, \overline{x_k}, \ldots, \overline{x_2})$ were traversed as well. The last edge traversed by R was $(\overline{x_k}, \ldots, \overline{x_1})$, outgoing from the vertex $[\overline{x_k}, \ldots, \overline{x_2}]$. All incoming edges to this vertex have already been traversed, as they are the reverse complements of the edges outgoing from v, which were traversed by F. Thus, R reaches a dead end as well. R closes a cycle because of a symmetrical argument to that made for F.

Case 2. F and R reach the same vertex v simultaneously. Denote the incoming edge used by F $(x_1, x_2, \ldots, x_k)$. Then, the reverse outgoing edge, which is traversed by R, is $(\overline{x_k}, \ldots, \overline{x_2}, \overline{x_1})$. From the fact that both reach the vertex simultaneously, we get that $[x_2, \ldots, x_k] = [\overline{x_k}, \ldots, \overline{x_2}]$. Hence, in all previous traversals of this vertex F and R also reached the vertex simultaneously. Moreover, the forward and reverse paths reach a dead end together at v. Hence, all incoming and outgoing edges were already traversed, and they are all of the form $(a, x_2, \ldots, x_n)$ and $(\overline{x_n}, \ldots, \overline{x_2}, \bar{a})$, for all $a \in \Sigma$. Thus, both paths close a cycle ■.

Lemma 2. The cycles in $\mathcal{F}$ can be merged into one cycle.

Proof. According to Lemma 1, when F is added to $\mathcal{F}$, it is a cycle in the graph. Thus, $\mathcal{F}$ is a set of cycles. The first cycle starts from an arbitrary vertex, but all other cycles start from a vertex of another cycle in $\mathcal{F}$ (denote *encompassing* cycle). Thus, each inner cycle can be merged into its encompassing cycle, forming one merged cycle. This is true to all cycles, except for the initial cycle ■.

Lemma 3. The merged cycle of $\mathcal{F}$ and $\mathcal{R}$ either cover two strongly connected components separately or one strongly connected component together.

Proof. Cycles are added to $\mathcal{F}$ and $\mathcal{R}$ as long as there are unmarked edges. If there are no shared vertices between $\mathcal{F}$ and $\mathcal{R}$, then both sets cover edges of different components. As each set is added edges until all are traversed, they cover two strongly connected components separately. Else, there is at least one shared vertex; thus, they cover the same component. The component is strongly connected, as no edges are left to traverse ■.

Corollary 1. $\mathcal{F}$ and $\mathcal{R}$ cover all edges of a de Bruijn graph.

Proof. Following Lemma 3, as there is only one strongly connected component in a de Bruijn graph, $\mathcal{F}$ and $\mathcal{R}$ cover it together ■.

This completes the proof of Theorem 2 ■.

3.3 Two solutions for the case of even *k*

Algorithm 1 cannot be applied when k is even. A palindrome is represented by one edge in the de Bruijn graph (like any other k-mer). The algorithm must traverse both an edge and its reverse complement edge on the forward and reverse paths; therefore, for a palindromic edge, both paths should use the same edge, which is impossible.

One possible way to rectify the problem is by adding one more copy of each palindromic edge to the de Bruijn graph. Note that in the resulting (multi-) graph, the number of edges is exactly twice the lower bound. Adding the parallel edges would solve the problem discussed earlier in the text, but it will make some vertices unbalanced; therefore, the resulting graph is not Eulerian. Such a graph cannot be represented as a union of two reverse complementary edge-disjoint paths.

A more aggressive augmentation that overcomes this difficulty is adding a *cycle* for every palindromic edge. This would preserve the balance of all vertices and the strong connectivity as well. If, in addition, the added non-palindromic edges have a perfect matching between reverse complementary edges, the algorithm can be applied.

We present two possible augmentations. One is simple, based on the ideas aforementioned, and near-optimal; the other is optimal but requires a more complex augmentation.

3.3.1 A simple near-optimal augmentation In this approach, for each palindromic edge, we add to the de Bruijn graph all possible cyclic shifts of it. More formally, let $k = 2l$. For the palindrome $e = (x_1, \ldots, x_l, \bar{x}_l, \ldots, \bar{x}_1)$, we add k edges corresponding to all possible cyclic shifts of e. Obviously, as these edges form a cycle, all vertices remain balanced. In fact, this cycle contains two edges that are palindromes, $(x_1, \ldots, x_l, \bar{x}_l, \ldots, \bar{x}_1)$ and $(\bar{x}_l, \ldots, \bar{x}_1, x_1, \ldots, x_l)$; therefore, only one cycle is added for both, and the cycle doubles both palindromic edges. It is easy to see that the remaining $2l - 2$ edges are in fact $l - 1$ matching pairs of reverse complementary edges. For each edge that represents the cyclic shift starting at position i, for $1 < i < k/2$, the matching edge starts at $k + 2 - i$. Hence, a perfect matching exists after adding the new cycles. In total, during the edge augmentation process, for each pair of palindromic k-mers, we add k edges. For example, for the palindromes ACGT and GTAC, we add ACGT, CGTA, GTAC and TACG (Fig. 2). The added edges CGTA and TACG match each other. The added palindromes match the original edges in the graph. The resulting augmented graph contains $|\Sigma|^k + k \cdot \frac{|\Sigma|^{k/2}}{2}$ edges, where the first term is the number of edges in the original de Bruijn graph, and the second is k for each pair of palindromes.

In some cases, the number of added edges can be reduced. If the palindrome $(x_1, \ldots, x_k)$ is periodic, then the number of cyclic shifts needed to return to the original k-mer is the length of the period. For example, the period of (ATAT…T) and (TATA…A) is 2. Only two edges suffice in this case, the edges (ATAT…T) and (TATA…A). This also applies to (CGCG…G) and (GCGC…C). Therefore, each two periodic palindromes that are a cyclic shift of each other require an addition of a number of edges equal to the length of their period. Hence, a smaller augmented graph and a shorter RC complete sequence can be obtained by considering the different possible periods, which can only be of even length, as each period is a palindrome.

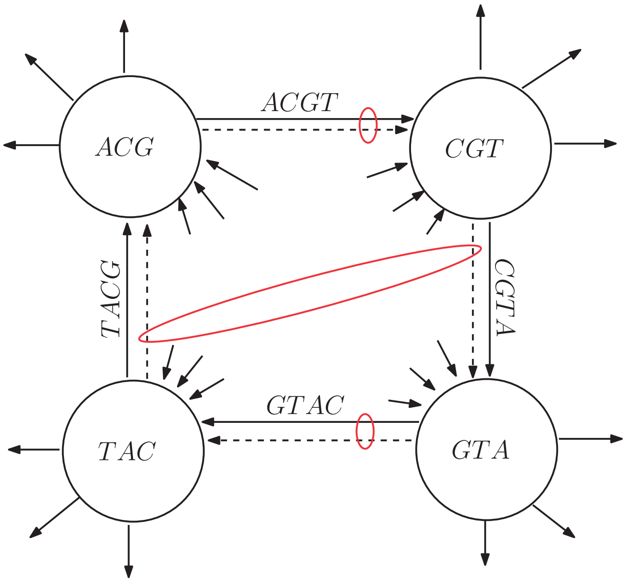


Fig. 2. A cycle and edge matching. For the pair of palindromes ACGT and GTAC, all cyclic shifts of these palindromes are added once (dashed edges). In the matching, palindromic edges in the original cycle are paired with their added copies (encircled by small red ovals). Other non-palindromic added edges are paired (encircled by a large red oval)

Denote by $\varphi(k)$ the set of even integers that divide k, and by $\delta(k)$ the exact number of additional edges.

Theorem 3.

$$\delta(k) = \sum_{i \in \varphi(k)} \frac{i}{2} \cdot \left(|\Sigma|^{i/2} - \max_{j \in \varphi(i/2)} |\Sigma|^{j/2} \right) \qquad (2)$$

Proof. All k-mer palindromes are divided to pairs, which are cyclic shifts of each other. For each pair, all distinct cyclic shifts are added. The number of shifts is equal to the length of the period of the k-mer. The periods can only be even, as the periodic sequences are palindromes by themselves. The number of i-periodic palindromes is $|\Sigma|^{i/2}$. These contain shorter periods, for which edges have already been counted. Thus, $|\Sigma|^{j/2}$ is subtracted, where j is the maximum even integer that divides $i/2$. The number of edges added for each pair of i-periodic palindromes is i ■.

Theorem 4. Running Algorithm 1 on the augmented graph produces forward and reverse paths that together cover all edges of the graph and represent two RC complete sequences.

Proof. Algorithm 1 can be run on graphs that satisfy the following properties:

(i) The graph is strongly connected.
(ii) All vertices are balanced.
(iii) There exists a perfect matching of the edges, such that each pair of edges represent a k-mer and its reverse complement.

The original de Bruijn graph of order k satisfies (1) and (2), and there exists a perfect matching for all non-palindromic k-mers in it. Added edges cannot disturb the connectivity. The addition of cycles preserves the balance. Each added palindromic k-mer matched the edge representing the same k-mer in the original graph. As discussed earlier in the text, the added non-palindromic edges form a perfect matching. Thus, Algorithm 1 can be run on the augmented graph. According to Theorem 2, it produces a forward and reverse path that together covers all edges of the augmented graph.

Each k-mer is represented in the augmented graph as an edge. All edges are covered together by the forward and reverse paths. For each path and for each k-mer, either it or its reverse complement is covered by the path. Thus, the paths represent RC complete sequences ■.

Algorithm 1 produces two sequences, forward and reverse, each of which is an RC complete sequence (Fig. 3). The length of the produced sequences is the number of edges divided by two. For each pair of palindromic edges, at most k edges were added, and by Theorem 3 exactly $\delta(k)$ edges were added in total. Hence, the length of the sequence is $(|\Sigma|^k + \delta(k))/2$, which is bounded by $(|\Sigma|^k + \frac{|\Sigma|^{k/2}}{2} \cdot k)/2$. This is an addition of $\Theta(\sqrt{L}\log(L))$ characters, where L denotes the lower bound in Proposition 1 for an RC complete sequence of even order k.

3.3.2 An optimal augmentation We now present another augmentation that has higher time complexity but leads to an optimal RC complete sequence. As before, starting from the de Bruijn graph $G = (V, E)$, all palindromic edges are doubled, resulting in a graph $G' = (V, E \cup E')$. We temporarily disregard the reverse complementarity matching constraints. As a result of the edge doubling, there are unbalanced vertices in G'. We rectify this by adding short paths between unbalanced vertices. By adding paths of minimum total length, we will obtain a third graph $G^2 = (V, E \cup E' \cup E'')$ in which all degrees are balanced and it has minimum number of edges. Finding an optimal set of edges E'' can be done by solving a maximum weight-matching problem on a related graph. In fact the problem is equivalent to the Chinese postman problem (Edmonds and Johnson, 1973) [the Chinese postman problem is used in Medvedev and Brudno (2009) and Medvedev *et al.* (2007) and is also mentioned in Mintseris and Eisen (2006) as a solution on the original de Bruijn graph]. We shall later show that G^2 can be modified to satisfy the reverse complementarity matching requirement without losing optimality. Hence, applying Algorithm 1 on it will produce an optimal RC complete sequence.

Finding an optimal set of edges E'' is done by solving a maximum weight-matching problem in a bipartite graph, where vertices with greater indegree than outdegree constitute one part, and the vertices with greater outdegree than indegree are the other. The edge weights are k minus the number of characters on the path from one vertex to the other. More formally, let V^- (V^+) be the set of vertices with indegree greater (smaller) than outdegree in G'. For $k = 2l$, there are $|\Sigma|^{k/2} - |\Sigma|$ vertices in V^- of the form $u = [x_2, \ldots, x_l, \bar{x}_l, \ldots, \bar{x}_1]$ and the same number of vertices in V^+ of the form $v = [x_1, \ldots, x_l, \bar{x}_l, \ldots, \bar{x}_2]$ [note that $|\Sigma|$ palindromes of period 2 are already balanced (e.g. ATA…T)]. We define a complete bipartite graph $H = (V^-, V^+, F)$, where the weight of edge (u, v) is the maximum overlap between the suffix of u and the prefix of v (i.e. $|ov(u, v)|$). The length of the shortest path $p(u, v)$ between u and v is $k - |ov(u, v)|$ (Fig. 4). We are looking for a maximum weight matching in H. The procedure is summarized in Algorithm 2, Steps 1–5.

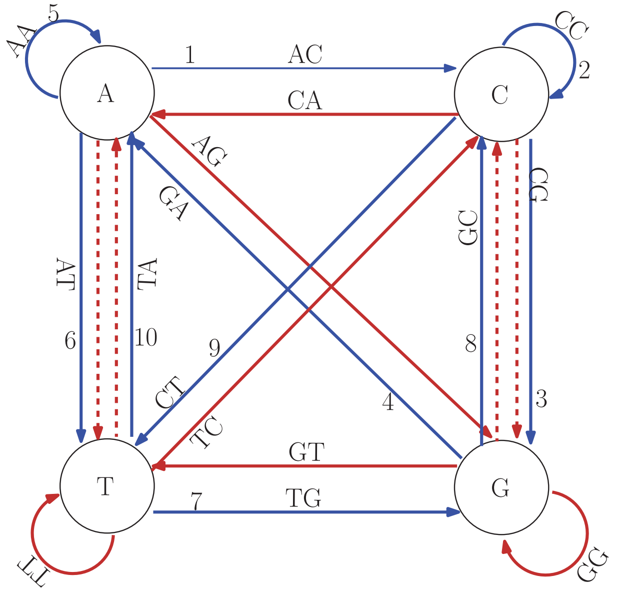


Fig. 3. An augmented de Bruijn graph of order 1 and an example of forward and reverse paths in it. The dashed edges are added edges. The blue and brown paths represent the forward and reverse paths, respectively. Numbers on edges are the order of the edges in the forward path. The sequences are ACCGAATGCT and AGCATTCGGT for forward and reverse paths, respectively

Algorithm 2. Find an optimal augmentation for a de Bruijn graph $G = (V, E)$ of odd order.

```
1. Add to $G$ the set $E'$ of palindromic edges.
   The resulting (multi-)graph is $G' = (V, E \cup E')$.
2. Define $V^+ = \{v \in V | (v, u) \in E' \wedge (u, v) \notin E'$ for some $u\}$
       $V^- = \{u \in V | (v, u) \in E' \wedge (u, v) \notin E'$ for some $v\}$.
3. Define a complete bipartite graph $H = (V^-, V^+, F)$
   with edge weights $w(x, y) = |ov(x, y)|$.
4. Find a maximum weight-matching $M$ in $H$.
5. Define $G^2 = (V, E \cup E' \cup E'')$
   where $E'' = \{(u, v) \in p(x, y) | (x, y) \in M\}$.
6. Modify $M$, so that each cycle in the graph $(V, E' \cup E'')$
   contains exactly two palindromic edges (Lemma 6).
```

The graph G^2 produced in Step 5 of Algorithm 2 is strongly connected with all vertices balanced, but it is not guaranteed to satisfy the third property of Theorem 4, i.e. having a perfect matching among reverse complementary edges, which is needed to apply Algorithm 1. We now prove that it can be modified to satisfy this property without losing optimality. In fact, as $E \cup E'$ has a perfect matching, we only need to prove this property on the added edges E''. Once this is done, Algorithm 1 can be applied to produce two reverse complementary paths that cover all edges.

To establish Algorithm 2, we prove several lemmas:

Lemma 4. The shortest path from palindrome A to the palindrome B is the reverse complementary of the shortest path from B to A.

Proof. Denote $A = (x_1, \ldots, x_k)$ and $B = (y_1, \ldots, y_k)$ two palindromes. Let $(x_i, \ldots x_k, y_1, \ldots, y_{i-1})$ for any $2 \le i \le k$ be an edge in the shortest path from A to B. Its reverse complement is $(\overline{y_{i-1}}, \ldots, \overline{y_1}, \overline{x_k}, \ldots, \overline{x_i})$, which, as A, B are palindromes, which is the same as $(y_{k-i+2}, \ldots, y_k, x_1, \ldots, x_{k-i+1})$, an edge in the shortest path from B to A ■.

Lemma 5. No cycle in $(V, E' \cup E'')$ contains a single palindrome.

Proof. Suppose there exists a cycle containing only one palindrome. The shortest path to return to the palindrome is t cyclic shifts of the palindrome where t is the length of its period. Let $(x_1, \ldots, x_l, \overline{x_l}, \ldots, \overline{x_1})$ be the palindrome. Its cyclic shift $(\overline{x_l}, \ldots, \overline{x_1}, x_1, \ldots, x_l)$ is another palindrome. Thus, the cycle includes more than one palindrome ■.

LEMMA 6. Every cycle in $(V, E' \cup E'')$ can be decomposed into cycles containing exactly two palindromes each, without decreasing the total weight of the matching.

PROOF. The proof is by induction on n, the number of palindromes in the cycle. For the induction base, $n = 1$ is impossible by Lemma 5, and $n = 2$ is trivially true. Induction step, for $n \geq 3$, denote by X, Y, Z and W palindromes in the cycle, where W, X, Y and Z appear in this order in the cycle. Let $x = |ov(W, X)|$, $y = |ov(X, Y)|$, $z = |ov(Y, Z)|$ and let w be the sum of overlaps of all palindromes between Z and W (inclusive). In case $n = 3$, $Z = W$ and $w = 0$. Without loss of generality, let y be a maximum overlap. The total sum of overlaps is $x + y + z + w$ (Fig. 5).

Remove X and Y and form a cycle of these two palindromes. As X, Y are palindromes, $ov(X, Y) = ov(Y, X)$; therefore, the contribution of this cycle to the matching is $2y$. The total overlap of the remaining cycle is w plus the overlap between W and Z, which is at least $min(x, z)$. To see this, denote by $Pref(X, i)$ the i-long prefix of string X, and denote by $Suf(X, i)$ the i-long suffix of X. If $x \leq z$, $Suf(W, x) = Pref(X, x) = \overline{Pref(Y, x)} = Suf(Y, x) = Pref(Z, x)$, where the first, second and fourth equalities follow from the overlap assumptions and the second, third and fourth use the palindrome property. If $z \leq x$, similarly $Suf(W, z) = Pref(Z, z)$. Hence, $|ov(W, Z)| \geq min(z, x)$. The total weight of the two cycles in the new matching is at least $2y + w + min(x, z)$. Hence, the difference between the new matching and the previous one is at least $2y + w + min(x, z) - x - y - z - w = y + min(x, z) - x - z = y - max(x, z) \geq 0$, where the last inequality follows by the choice of y as a maximum overlap.

The remaining cycle has $n - 2$ palindromes, and by the induction step, it is breakable to cycles of size two ■.

PROPOSITION 2. There exists a maximum weight matching in which all the added edges form reverse complementary pairs. Any maximum weight matching can be modified to such matching.

PROOF. Consider the graph G^2 produced in Step 5 of Algorithm 2. If $E' \cup E''$ contains cycles of more than two palindromes, by Lemma 6, they can be decomposed into cycles of two palindromes. The new matching is of the same size, and for each cycle with exactly two palindromic edges, the remaining edges match in reverse complementary pairs (Lemma 4) ■.

The maximum weight-matching problem, also known as the assignment problem (West *et al.*, 2001), can be solved by the Hungarian method in $O(|V|^2 log|V| + |V||E|)$ time (Kuhn, 2006). As $|V| = \Theta(|\Sigma|^{k/2})$ and $|E| = \Theta(|V|^2) = \Theta(|\Sigma|^k)$, the running time is $O(|\Sigma|^{3k/2})$. An improvement to this algorithm (Kao *et al.*, 1997), when the edge weights are integers, runs in $O(\sqrt{|V|}|E|log(|V|N))$ time, where N is the largest edge weight. In our case $N = k$, which gives $O(k|\Sigma|^{5k/4}\log(|\Sigma|))$ running time. The post-processing of the matching (Lemma 6) requires finding two palindromes with maximum overlap. This can be done in total time linear in the number of palindromes, as overlap lengths are integers in the range of 0 to k, and thus can be sorted using count sort. Hence, we conclude

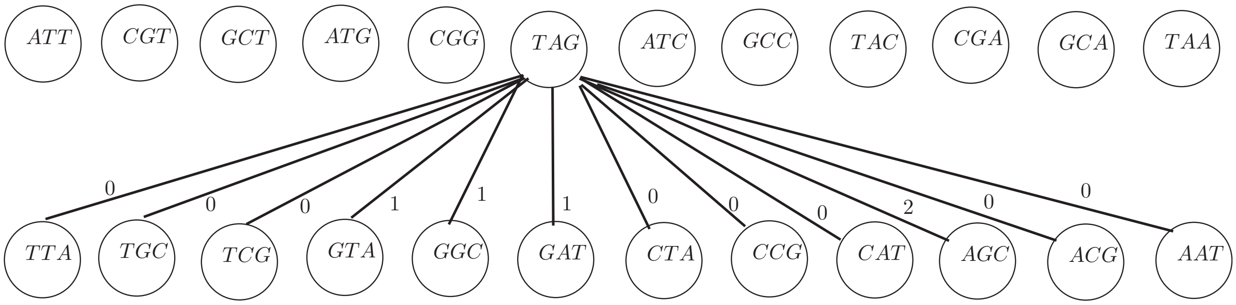


Fig. 4. The bipartite graph for matching unbalanced vertices (Algorithm 2). On the top are the vertices with greater indegree, and on the bottom are the vertices with greater outdegree. Weights on the edges are the maximum overlap between the vertices' sequences. Only the edges out of one vertex are drawn (the graph is a complete bipartite graph). Note that only unbalanced vertices corresponding to $(k - 1)$-long prefixes and suffixes of palindromes are included

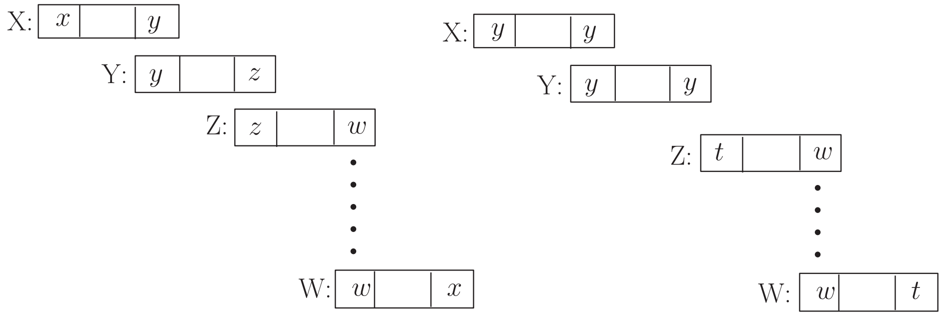


Fig. 5. Breaking down cycles with more than two palindromes. Left: Palindrome overlaps in a cycle found by the maximum matching. The rectangles at the ends indicate overlap between contiguous palindromes. Right: Partition into two cycles, one containing only the palindromes X and Y with a maximum overlap y. As $t \geq min(x, z)$, the partition does not decrease the total contribution of the cycles to the weighted matching (Lemma 6)

THEOREM 5. An optimal RC complete sequence for even k can be produced in time $O(k|\Sigma|^{5k/4}\log(|\Sigma|))$.

Summarizing Theorems 2 and 5 we obtain

THEOREM 6. For every value of k, an optimal RC complete sequence can be obtained in time polynomial in the size of a de Bruijn graph of order $k-1$.

4 EXPERIMENTAL RESULTS

Table 1 shows the results of the two algorithms for even k. As we can see, the sequence obtained by the algorithm is of length nearly half that of the original de Bruijn sequence. For example, for $k=12$, the minimum length is within 0.15 per cent of $4^{12}/2$ and within 10, 116 characters from the theoretical lower bound.

Table 2 lists the number of probes of length p needed to cover all k-mers, by cutting an optimal RC complete sequence to p-long probes with overlaps of $k-1$. As we can see, the saved factor in using the RC complete sequence is roughly 2. A comparison to the Table 1 of (Mintseris and Eisen, 2006) shows that the sequence produced in (Mintseris and Eisen, 2006) is sub-optimal.

Running times: The simple near-optimal algorithm runs in time roughly linear in $|\Sigma|^k$. For example, for $k=8$, 10 and 12 the running times are 1.5, 26 and 445 s, respectively. The optimal algorithm requires 5, 126 and 2937 s, respectively.

5 SUMMARY AND DISCUSSION

In this article, we studied the problem of constructing a minimum length sequence that covers each k-mer or its reverse complement at least once. The problem has applications in construction of dense double-stranded probe arrays for *in vitro* measuring of protein–DNA binding (Berger *et al*., 2006; Fordyce *et al*., 2010), and for design of synthetic enhancers for *in vivo* developmental studies (Smith and Ahituv, 2012). For the case of odd k, we provided a proof that a simple modification of the Eulerian tour algorithm applied to the de Bruijn graph of order $k-1$ gives an optimal solution. The algorithm requires linear time in the output sequence length, and it cuts the sequence length in half compared with using a regular de Bruijn sequence.

The problem is a bit more involved for even k, and here we provided two algorithms, a linear time near-optimal algorithm and a more complex polynomial algorithm that produces an optimal sequence. The length of the sequence produced by the optimal algorithm is slightly shorter, and both algorithms nearly halve the total length of the sequence.

The following related problem was studied by Medvedev *et al.* (Medvedev and Brudno, 2009; Medvedev *et al*., 2007): what is the minimum length sequence that contains a given set of k-mers? Their solution is based on bidirected graphs, which are similar to de Bruijn graphs, with the difference that a k-mer and its reverse complement are represented by the same vertex, and the edges represent the possible ways that double-stranded

Table 1. Length of reverse complementary de Bruijn sequences produced by the two algorithms for even k

k	2	4	6	8	10	12	14
Original	16	256	4096	65 536	1 048 576	16 777 216	268 435 456
Lower bound	10	136	2080	32 896	524 800	8 390 656	134 225 920
Algorithm 1	10	142	2140	33 262	526 840	8 400 808	134 275 060
Optimal	10	142	2140	33 262	526 816	8 400 772	134 274 844
Saving factor	1.6	1.8	1.91	1.97	1.990	1.997	1.999

Note: The top row is the length of a regular de Bruijn sequence that does not exploit complementarity. The next row contains the theoretical lower bound on RC complete sequence length (Proposition 1). The next two rows are the lengths of the sequence computed by the two algorithms of Section 3.3.1 and 3.3.2. The saving factor is the ratio between the original sequence length and length of the optimal RC complete sequence. Note that the lower bound is not tight.

Table 2. Number of probes needed to cover all k-mers as a function of probe length and k

k	6	7	8	9	10	11	12	13	14	14-DB
25	107	432	1848	7711	32 926	139 811	600 056	2 581 111	11 189 571	22 369 622
30	86	342	1447	5958	25 087	104 858	442 146	1 864 136	7 898 521	15 790 321
35	72	283	1188	4855	20 263	83 887	350 033	1 458 889	6 103 402	12 201 612
40	62	241	1008	4096	16 995	69 906	289 682	1 198 373	4 973 143	9 942 054
45	54	211	876	3543	14 634	59 919	247 082	1 016 801	4 196 089	8 388 608
50	48	187	774	3121	12 850	52 429	215 405	883 012	3 629 050	7 255 013
55	43	168	693	2789	11 453	46 604	190 927	780 336	3 197 021	6 391 321
60	39	152	628	2521	10 330	41 944	171 445	699 051	2 856 912	5 711 393
65	36	139	574	2300	9408	38 131	155 570	633 103	2 582 209	5 162 221
70	33	128	528	2115	8637	34 953	142 386	578 525	2 355 700	4 709 394

Note: The table contains the number of probes obtained by cutting an optimal RC complete sequence to short segments with overlaps. Left column: probe length; top row: k. Right column: number of probes needed when using a regular de Bruijn sequence for $k=14$.

strings can overlap. These graphs were originally conceived by Kececioglu and Myers (1995) and actually discovered earlier by Edmonds (1967). Medvedev *et al.* stated, without proof, that an Eulerian path can be found in a bidirected graph in the same way as in a regular de Bruijn graph (Lemma 1), but they did not consider explicitly the problem of covering all k-mers and did not make the distinction between even and odd k. In fact, some vertices in a bidirected graph of odd order (when edges represent k-mers of even length) are unbalanced, and thus an Eulerian tour does not exist. Although their method can be applied to our problem, it is slower than ours: they require $O(k^2 \log^2(|\Sigma|)|\Sigma|^{2k})$, whereas our algorithms requires $O(|\Sigma|^k)$ for odd k and $O(k|\Sigma|^{5k/4} log(|\Sigma|))$ for even k.

Beyond the theoretical interest, our results are applicable to current (Berger *et al.*, 2006; Fordyce *et al.*, 2010; Smith and Ahituv, 2012) and future technologies that require complete coverage of double-stranded DNA k-mers. In PBM, although it is desirable to have redundancy in covering k-mers, space on the arrays is limited. By essentially halving the needed sequence length, space is freed on the array to select additional redundant probes with desired properties. Similarly, in designing synthetic enhancer sequences, by using shorter sequences, experiments can be simplified.

In current technologies, the de Bruijn (or RC complete) sequence is cut into probes of length p with overlap $k-1$ (Table 2). There is no constraint that forces these probes to come from a single sequence. A variant of the problem we studied is as follows: what is the minimum number of double-stranded DNA probe sequences of length p that together cover all k-mers? As our solution for an RC complete sequence of even k covers, a few k-mers more than once and direct design of probe sequences of length p might reduce the number of probes needed to cover all k-mers.

A heuristic solution to that problem was recently proposed by Riesenfeld and Pollard (Riesenfeld and Pollard, 2012). They studied the following problem: given k and m, design a set of m double-stranded DNA probes (of equal or almost equal length, denoted as ℓ) that together cover all k-mers. Their algorithm repeatedly searches for disjoint ℓ-long paths between unbalanced vertices. After removal of all such paths, it finds two reverse-complementary cycles. One cycle is cut into probes (with overlaps of $k-1$) of length ℓ or $\ell+1$. If the program terminates, an optimal set of oligomers is found; however, there is no theoretical guarantee that it will terminate. In our tests, for $k=6$, their program terminates in a few seconds, whereas for $k=8$, it takes >1 h and for $k=10>2$ weeks. For some values of m, the produced probes are not of equal length. A modest reduction in the number of oligomers is obtainable compared with our design: for example, for $k=6$ and probe length 15, the algorithm of Riesenfeld and Pollard produced 208 oligomers compared with 210 in our design. For greater values of k, the running time was already prohibitive (for $k=12$, it kept running for >1 month), and thus we could not test the performance for these values. Our algorithm, on the other hand, produces an output for values of $k \leq 10$ in just a few seconds, whereas for $k=12$, the linear algorithm takes <10 min and the optimal <1 h. The time is polynomial (or even linear) in the output sequence size, independent of probe length or the number of oligomers.

Our study raises several additional open questions. First, following (Philippakis *et al.*, 2008), can one design an optimal RC complete sequence with improved coverage of gapped k-mers? Second, it is known that the number of distinct de Bruijn sequences is $(k!)^{k^{n-1}}/k^n$. What is the number of different optimal RC complete sequences? Third, can one construct an optimal RC complete sequence for even k in linear time? Fourth, is there a closed formula for the length of an optimal RC complete of even order?

Funding: This study was supported in part by the Israel Science Foundation (grant no. 802/08), and by the I-CORE Program of the Planning and Budgeting Committee and the Israel Science Foundation (grant no. 41/11). Y.O. was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University and by a Dan David PhD Fellowship.

Conflict of Interest: none declared.

REFERENCES

Berger,M. *et al.* (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.

Chen,X. *et al.* (2007) Rankmotif++: a motif-search algorithm that accounts for relative ranks of *k*-mers in binding transcription factors. *Bioinformatics*, **23**, i72–i79.

Edmonds,J. (1967) An introduction to matching. In: *Notes of Engineering Summer Conference*. University of Michigan, Ann Arbor.

Edmonds,J. and Johnson,E. (1973) Matching, Euler tours and the Chinese postman. *Math. Program.*, **5**, 88–124.

Fleischner,H. (1990) *Eulerian Graphs and Related Topics*. Vol. 1. North-Holland, Amsterdam and New York.

Fordyce,P. *et al.* (2010) De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nat. Biotechnol.*, **28**, 970–975.

Jolma,A. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327.

Kao,M. *et al.* (1997) All-cavity maximum matchings. *Algorithms Comput.*, **1350**, 364–373.

Kececioglu,J. and Myers,E. (1995) Combinatorial algorithms for DNA sequence assembly. *Algorithmica*, **13**, 7–51.

Kuhn,H. (2006) The Hungarian method for the assignment problem. *Naval Res. Logist. Q.*, **2**, 83–97.

Medvedev,P. and Brudno,M. (2009) Maximum likelihood genome assembly. *J. Comput. Biol.*, **16**, 1101–1116.

Medvedev,P. *et al.* (2007) Computability of models for sequence assembly. *Algorithms Bioinform.*, 289–301.

Mintseris,J. and Eisen,M. (2006) Design of a combinatorial DNA microarray for protein-DNA interaction studies. *BMC Bioinformatics*, **7**, 429.

Nutiu,R. *et al.* (2011) Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat. Biotechnol.*, **29**, 659–664.

Orenstein,Y. *et al.* (2013) Rap: Accurate and fast motif finding based on protein-binding microarray data. *J. Comput. Biol.*, [Epub ahead of print, March 6 2013].

Philippakis,A. *et al.* (2008) Design of compact, universal DNA microarrays for protein binding microarray experiments. *J. Comput. Biol.*, **15**, 655–665.

Riesenfeld,S. and Pollard,K. (2012) Computing MRCC libraries and related types of DNA oligomer libraries. https://github.com/sriesenfeld/MRCC-Libraries (1 April 2013, date last accessed).

Smith,R. and Ahituv,N. (2012) Deciphering the vertebrate regulatory code using short synthetic enhancers in vivo. http://zendev.ucsf.edu/projectview.php?project=6mer (1 April 2013, date last accessed).

West,D. *et al.* (2001) *Introduction to Graph Theory*. Vol. 2. Prentice Hall, Upper Saddle River, NJ.

5. A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP

Nucleic Acids Research Advance Access published February 5, 2014

Nucleic Acids Research, 2014, ***1–10***
doi:10.1093/nar/gku117

A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data

Yaron Orenstein and Ron Shamir*

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

Received November 11, 2013; Revised December 22, 2013; Accepted January 16, 2014

ABSTRACT

Understanding gene regulation is a key challenge in today's biology. The new technologies of protein-binding microarrays (PBMs) and high-throughput SELEX (HT-SELEX) allow measurement of the binding intensities of one transcription factor (TF) to numerous synthetic double-stranded DNA sequences in a single experiment. Recently, Jolma *et al*. reported the results of 547 HT-SELEX experiments covering human and mouse TFs. Because 162 of these TFs were also covered by PBM technology, for the first time, a large-scale comparison between implementations of these two *in vitro* technologies is possible. Here we assessed the similarities and differences between binding models, represented as position weight matrices, inferred from PBM and HT-SELEX, and also measured how well these models predict *in vivo* binding. Our results show that HT-SELEX- and PBM-derived models agree for most TFs. For some TFs, the HT-SELEX-derived models are longer versions of the PBM-derived models, whereas for other TFs, the HT-SELEX models match the secondary PBM-derived models. Remarkably, PBM-based 8-mer ranking is more accurate than that of HT-SELEX, but models derived from HT-SELEX predict *in vivo* binding better. In addition, we reveal several biases in HT-SELEX data including nucleotide frequency bias, enrichment of C-rich k-mers and oligos and underrepresentation of palindromes.

INTRODUCTION

The questions of how, when and where genes are expressed have been fundamental in the field of cell research in the past decades. Transcription factors (TFs) are known to be the main regulators of gene transcription and thus have been a subject for extensive study. These proteins bind to specific short DNA sequence, mainly in the promoter and enhancer regions, and by that impede or encourage transcription. They bind with variable affinity, depending on the sequence and on other factors, and this affinity affects transcription. Learning and modeling the binding preferences of TFs is a central goal in gene regulation research.

Many high-throughput technologies have been developed to study TF binding. Technologies that measure *in vivo* binding include ChIP-chip (1), ChIP-seq (2) and the recently developed ChIP-exo (3). However, measuring *in vivo* binding may not reveal the full picture. First, the accessible sites may not cover the full spectrum of possible DNA k-mers. Second, *in vivo* binding is affected by additional factors, such as chromatin structure, nucleosome positioning and co-factors. As opposed to *in vivo* binding, *in vitro* binding is purely because of direct TF–DNA interaction (or cooperative binding of specific factors) and allows sampling of the full spectrum of DNA k-mers. Technologies that measure *in vitro* binding include protein-binding microarray (PBM) (4) and mechanically induced trapping of molecular interactions (5), both of which measure the binding of a specific protein to a set of oligo sequences designed to cover all k-mers. A newer technology is high-throughput SELEX (HT-SELEX), which consists of several cycles of incubating the DNA-binding protein with a mixture of DNA sequences, enrichment of the bound DNA sequences, sequencing a sample of them and feeding them to the next cycle (6–8).

PBMs have gained great popularity, thanks to their high-throughput and unbiased nature. The public database UniPROBE contains experiments of >400 TFs (9). Although the models derived from this technology have been used extensively, it is unclear how accurate these models are in predicting *in vivo* binding. Several studies have shown that using these positional weight matrix (PWM) models to predict *in vivo* binding leads to

*To whom correspondence should be addressed. Tel: +972 3 6405383; Fax: +972 3 640 5384; Email: rshamir@tau.ac.il
Present address:
Ron Shamir, School of Computer Science, Tel-Aviv University, P.O.B. 39040, Tel-Aviv 69978, Israel.

poorer results compared with *in vitro* binding prediction (10,11). This performance gap can be explained by several reasons related to *in vivo* binding, such as indirect binding and inaccessibility of genomic DNA. Another possible explanation is that these models include PBM-specific biases. Thus, an independent *in vitro* measurement is required to evaluate the validity of these models.

Recently, a study covering >500 TFs in >800 HT-SELEX experiments was conducted by the Taipale laboratory (12). For the first time, a high number of TFs have available experimental data in two independent *in vitro* technologies: 162 TFs were tested both in HT-SELEX and PBM experiments by the Taipale and Bluyk laboratories, respectively. Jolma *et al*. (12) compared SELEX models with PBM models by length and presented several examples where the SELEX models are more accurate than PBM models based on ChIP-seq data. However, a much broader systematic comparison of the binding models produced by each technology is required.

In this study we aim to analyze and measure the similarities and differences between the two technologies. First, we ask how well HT-SELEX-derived PWM models predict PBM binding. Second, to compare the methods without depending on inferred binding models, we study how well the top k-mers of the two technologies correlate, and which technology is better in k-mer ranking. Third, we test which technology produces better models in predicting *in vivo* binding. Fourth, we uncover biases in HT-SELEX technology. We aim to highlight the advantages of each technology compared with the other. Our observations may help in developing a new method to learn binding models based on HT-SELEX data.

MATERIALS AND METHODS

Data

PBM data and PBM-derived PWM models were downloaded from UniPROBE database (9). We used normalized PBM probe data, as available in the database (i.e. the median signal intensity values and corresponding nucleotide probe sequences). Only the 36 bp of unique sequence were used. HT-SELEX experimental data and HT-SELEX-PWM models were downloaded from (12). Human ChIP-seq data were downloaded from ENCODE (13).

Binding prediction

PWMs were used to represent TF binding preferences (14). For each TF, the set of PWMs reported was used for the binding prediction. In many cases, multiple models were available. In general, we did not distinguish between mouse and human and between the full protein and the binding domain only. For each sequence (either PBM probe or a ChIP-seq peak), the maximum sum occupancy score over the set of PWMs was its predicted binding intensity. For probe sequence s and PWM Θ of length k, the sum occupancy score is

$$f(s,\Theta) = \sum_{t=0}^{|s|-k} \prod_{i=1}^{k} \Theta_i[s_{t+i}]$$

where $\Theta_i(x)$ is the probability of base x in position i of the PWM. A PBM probe is defined as a positive hit for Θ if its binding intensity is greater than the median by at least 4 * (MAD/0.6745), where MAD is the median absolute deviation from the binding intensity median (MAD = 0.6745 for the normal distribution N(0,1)) (15). The positive ChIP-seq peaks are defined as the 500 peaks with the smallest reported *P*-value. We used the 250 bp around the center of the peak as the positive sequence and the 250-bp-long genomic sequence 300 bp downstream of the peak center as the negative sequence. Spearman rank coefficient, sensitivity at 1% false-positive and area under the receiver operating characteristic curve were used to gauge the binding prediction (see (15) for details). For ChIP-seq data, when several experiments were available for the same TF, the average area under curve (AUC) over these experiments is reported.

Model independent comparison

For each experiment, the scores of the top 100 8-mers according to one technique were compared with their scores in the other technique. PBM 8-mers were scored by average (or median) binding intensity. For a probe p_i, let $s(p_i)$ be its intensity. The score of 8-mer w is the average binding intensity: $avg(w) = (\sum_{w \in p_i} s(p_i))/(\sum_{w \in p_i} 1)$.

HT-SELEX 8-mers were scored by either their frequency or ratio of frequencies (frequency in cycle i divided by frequency in cycle i-1). The top 100 8-mers according to their PBM scores were selected, and Pearson correlation was calculated between the PBM scores and the HT-SELEX scores on these 8-mers. Similarly, the top 100 HT-SELEX 8-mers were chosen and their HT-SELEX scores were compared with their PBM scores using Pearson correlation.

Logo drawing

Motif logos were plotted using http://demo.tinyray.com/weblogo.

RESULTS

HT-SELEX-derived models predict PBM binding accurately for most TFs

We first used the HT-SELEX-derived PWM models published in (12) to predict bound probes in PBM experiments and compared their performance with PBM-derived PWM models. We used the SCI09 data set of (16), which includes 115 paired PBM experiments of 104 mouse TFs [in paired experiments, two array designs are used to study the same TF, and so a model learned on one array can be evaluated on the other, see (15)]. For 128 PBM experiments (out of 230), an HT-SELEX-derived model was available for the same TF; this set covers 56 different TFs. For some TFs, Jolma *et al.* reported several PWMs, either because of multiple experiments or because of construction of several PWMs by their algorithm. Occasionally, for a TF analyzed by PBM, both a primary motif and a secondary motif are reported. When multiple PWMs were reported for the same TF by one technology, we assigned to each sequence the highest score obtained by such a model. We used five algorithms to generate PWMs from PBM experiments: Amadeus-PBM (10), Seed-and-Wobble (4), RankMotif++ (15), BEEML-PBM (17) and RAP (18). The performance of the models generated by each algorithm was reported in (18). For each paired experiment, these models were learned on one array and tested on the other to avoid overfitting. Testing of a model was by predicting the binding intensity for each probe in the other array and comparing it with the measured binding intensity. Scores for the comparison were the Spearman rank coefficient on the positive probes, the sensitivity (true positive ratio) at 1% false-positive and AUC of the receiver operating characteristic curve (see Methods). We report the average results in Table 1 (for complete results see Supplementary Table S1).

The results show good agreement between the two technologies (Table 1 and Figure 1A). The average accuracy of HT-SELEX models is significantly lower than that obtained by PBM-derived models (e.g. AUC of 0.825 compared with 0.899 for the best PBM-derived models, *P*-value = $7.68 \cdot 10^{-14}$ Wilcoxon signed-rank test). This is expected because the evaluation is using PBM measurements. In an additional test on two other PBM data sets covering 115 human and mouse E26 transformation-specific (ETS) and homeodomain TFs tested on a single array (19,20), HT-SELEX-derived models achieved an average AUC of 0.928 (see Supplementary Information). These results may reflect properties of specific TF families.

We found no significant difference between binding models based on mouse and human proteins and between models based on full proteins and binding domains; in both cases the two models performed essentially equally in predicting PBM binding that used mouse binding domains (see Supplementary Information). Note that sample sizes were small and broader tests are still needed.

For some TFs, the HT-SELEX prediction results were poorer than those achieved by PBM models. We define a set of HT-SELEX-derived models for the same TF as a failure if it achieved an AUC lower by at least 0.1 than the average of the five PBM models. HT-SELEX models failed in 20 TFs (covered by 42 experiments), including all Sox, E2F and Rfx proteins, as well as the individual TFs Hnf4a, Rara, Rxra, Smad3, Sry and Zscan4 (Figure 1A and B). These failures occur in particular TF families, including the E2F, Sox, NR, Rfx, MAD and znfC2H2 families [experiments on HMG and znfC2H2 proteins had a low success rate (12)]. The high-mobility group (HMG) super-family includes the Sox, Lef and Tcf protein families. It was suggested that for this family of proteins the DNA structure plays a larger role for binding site recognition than sequence specificity (21), which may explain the failure for this protein family. The recent observation that E2F1 and Smad1 ChIP-seq peaks do not contain the *in vitro* binding site (22) may explain the failures for E2F and Smad3. Figure 1C presents the differences in the models for some of these cases.

A model-independent comparison

To avoid dependency on model learning, we performed a model-independent comparison. For each HT-SELEX experiment, we selected one arbitrary PBM experiment of the same TF from Cell08, SCI09 or EMBO10 studies. This resulted in 238 PBM-SELEX data sets. We chose to summarize the measurements of each method using 8-mer statistics, and focus on the top ranking 8-mers, which are expected to contain most of the information relevant for TF binding. For PBM 8-mer scores, we used average binding intensity, which is an accurate estimate of binding affinities (18). For HT-SELEX 8-mer scores, we tested two options: 8-mer frequency and 8-mer ratio (frequency in cycle i divided by frequency in cycle i-1) for all cycles (see Methods). With these scores at hand, for each data set we used the set of top 100 8-mers, according to one technology, and calculated the Pearson correlation of its scores with the scores of the same set in the other. Figure 2 shows the results for the different cycles, different scores and different selection of top 8-mers. Complete results are available in Supplementary Table S2. Using the Spearman rank correlation provided similar results (data not shown).

The results show that frequency scores give consistently better correlation with PBM scores than ratios. Hence, for the data analyzed in this study, frequency is superior to ratio, and we used it henceforth. The highest average correlation (just over 0.74) is achieved at cycle 3, when the top 8-mers are selected by PBM data, and HT-SELEX 8-mers are ranked by frequency (Figure 2A). The k-mer ranking becomes more specific as the cycles progress [as was noted in (12)]. At some point it becomes too specific, overrepresenting a small number of top k-mers and thus less accurate for medium- and low-affinity k-mers; we refer to this phenomenon as overspecification. Figure 2B

Table 1. Accuracy of HT-SELEX- and PBM-based PWM models in predicting TF binding to PBMs

Model based on	HT-SELEX	PBM				
Algorithm	Jolma *et al.*	Amadeus-PBM	Seed-and-Wobble	RankMotif++	BEEML-PBM	RAP
Spearman rank coefficient	0.282	0.230	0.272	0.301	0.335	0.339
Sensitivity at 1% false-positive	0.288	0.327	0.293	0.277	0.403	0.400
AUC	0.825	0.877	0.872	0.882	0.899	0.898

Note. Results show average Spearman rank coefficient, sensitivity at 1% false-positive and AUC for predicting positive binding in 128 paired PBM experiments (covering 56 different TFs). PBM data were taken from (16) and HT-SELEX models were taken from (12). Prediction results for the different PBM-based algorithms were taken from (18). For each experiment the PWM models learned by HT-SELEX or by the other PBM array were used to predict the bound probes (see Methods).

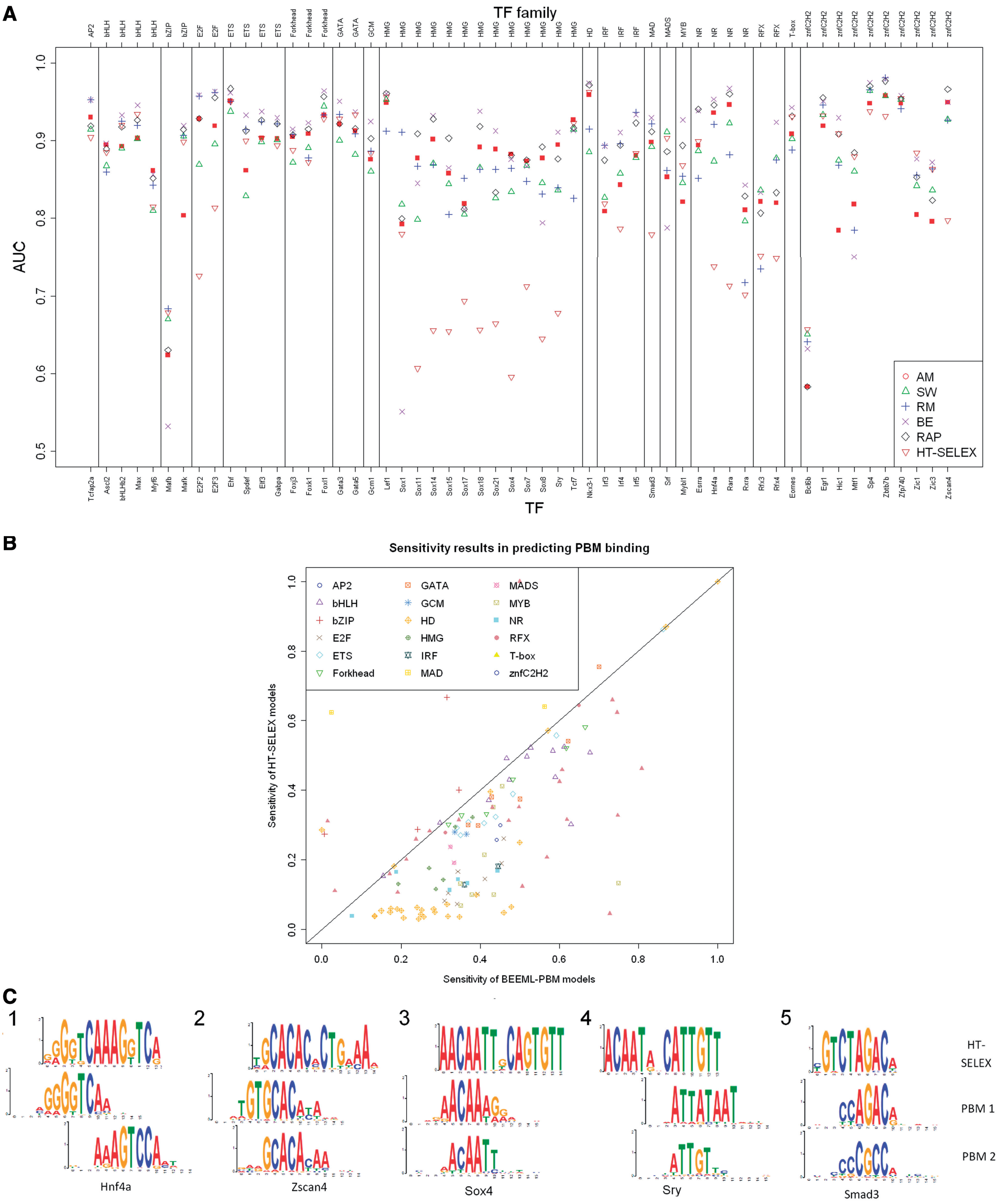


Figure 1. Quality of binding prediction based on PBM data. (**A**) Accuracy in predicting PBM binding. For each PBM experiment, PBM probes are ranked according to motifs inferred by five PBM algorithms (AM = Amadeus-PBM, SW = Seed-and-Wobble, RM = RankMotif++, BE = BEEML-PBM and RAP) and by the HT-SELEX-derived models. This ranking is compared with the true ranking by calculating the AUC for predicting the bound PBM probes. Each dot is the average result of one algorithm in two or four experiments (TF names are listed at the bottom, TF family names are at the top, as given in Jolma *et al.*). (**B**) Sensitivity results in predicting PBM binding. For each PBM experiment, the bound probes were predicted using BEEML-PBM and HT-SELEX PWM models. The plot shows the sensitivity (true positive rate) at 1% false-positive rate of these predictions. Colors correspond to protein families. (**C**) Disagreement between HT-SELEX- and PBM-derived models. The logos are of the PWMs learned from HT-SELEX (top), and PBM (middle and bottom) taken from Jolma *et al.* and UniPROBE, respectively. The middle and

(continued)

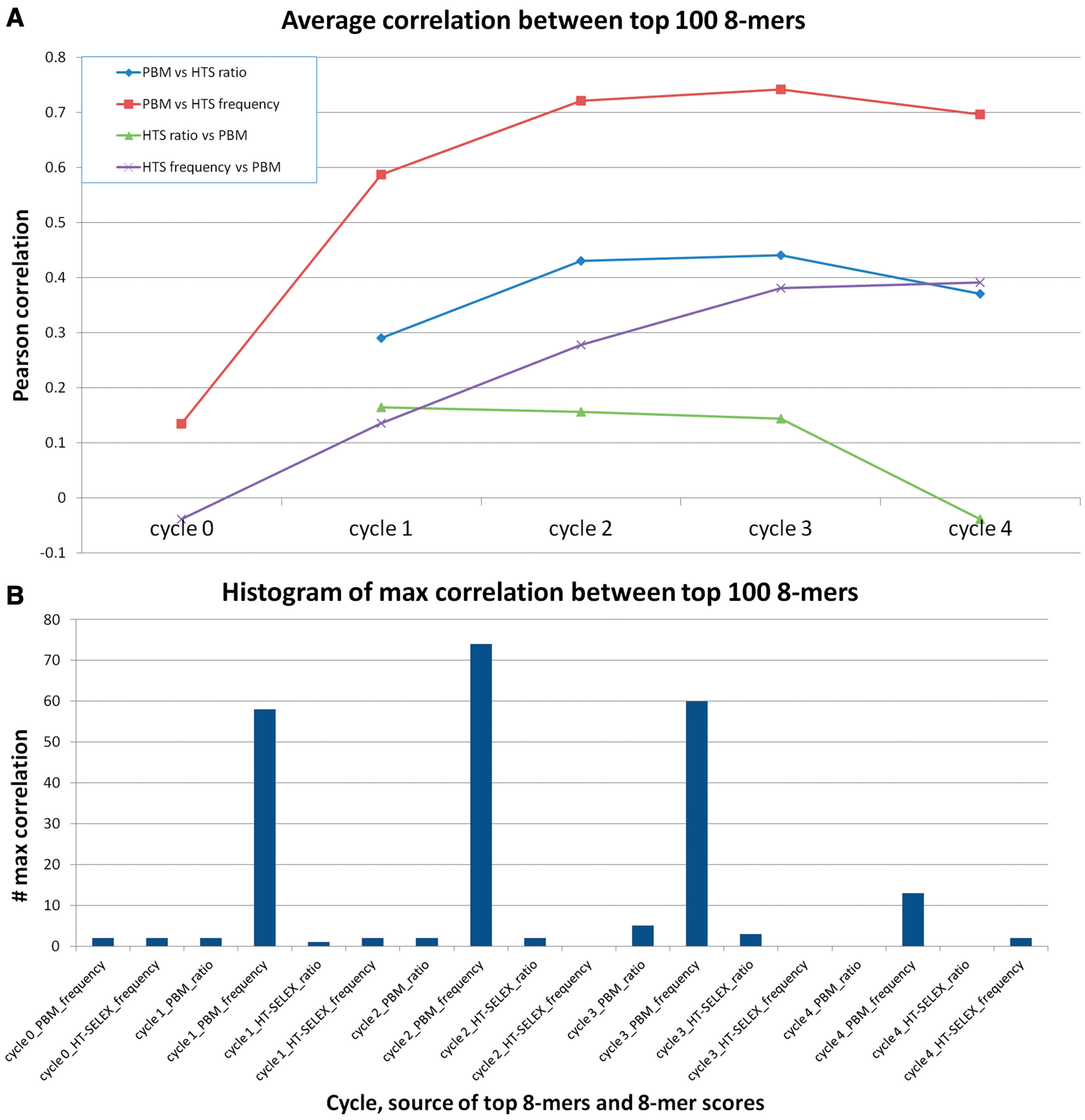


Figure 2. Correlation between the top 8-mers as ranked by PBM and HT-SELEX data. For each HT-SELEX experiment 8-mers were scored by frequency or by the ratio of the frequency to the frequency in the previous cycle. The 8-mers of a PBM experiment on the same TF were scored by average binding intensity. For the 100 top scoring 8-mers according to PBM, the correlation between their PBM scores and their HT-SELEX frequency and ratio scores was computed. Similarly, for the 100 top scoring 8-mers according to HT-SELEX frequency (ratio), their correlation with the PBM scores was computed. (**A**) Average correlation in each cycle. Bar names indicate the technology used to determine the top 100 8-mers. The plot is based on average correlation over 238 TFs. (**B**) Distribution of the maximum correlation for different parameter combinations. The plot shows the number of times the maximum correlation is achieved by each combination of cycle, source of top 8-mers and HT-SELEX 8-mers score. (Because only 39 HT-SELEX experiments included data for a fifth cycle, we excluded it from the comparison; none of these experiments had maximum score at the fifth cycle).

shows, for each combination of cycle, source of top 8-mers and HT-SELEX 8-mer score and the number of times the maximum correlation is achieved by that combination. Cycles 1, 2 and 3 have the highest numbers, supporting the idea of a trade-off between specificity and variability.

The results also suggest that 8-mers ranking using PBM is more reliable than using HT-SELEX. The top 100 PBM 8-mers have greater correlation than the top 100 HT-SELEX 8-mers. Identification of these 8-mers is important for learning the binding preference of the protein. At the current read coverage of HT-SELEX experiments, PBM

Figure 1. Continued
bottom models learned from PBM for each TF are the primary and secondary models, respectively. 1, 2: examples where HT-SELEX produces motifs that are similar to the primary PBM model, but too long for PBM technology; 3, 4: cases where HT-SELEX models agree with PBM secondary model; 5: an example where the HT-SELEX model disagrees with both PBM models. (TCF3 was excluded from the analysis because each technology tested a different TF with that name: a bHLH Tcf3 was tested by HT-SELEX, whereas the HMG Tcf3 was tested by PBM).

data are more robust in identifying the top 8-mers. Sequencing a larger sample of the bound oligos may improve 8-mer scores and thus affect the binding models derived from them.

No significant differences were observed when comparing mouse versus human models as well as full protein versus binding domains (see Supplementary Information). Using median binding intensity to score PBM 8-mers instead of the average showed similar results (data not shown).

HT-SELEX models predict *in vivo* binding more accurately than PBM models

We compared the performance of PBM PWM models with HT-SELEX PWM models in predicting *in vivo* binding. We used human ChIP-seq data from the ENCODE project (13) for TFs that had both PBM and HT-SELEX data. In total, 15 human TFs covered by 111 ChIP-seq experiments were included in this comparison. The top 500 peaks in each experiment were used as a positive set, taking for each peak 250 bp around its center. The negative set consisted of 250-bp-long sequences taken from flanking sequences 300 bp downstream of each positive sequence. This choice is aimed to select negative sequences with statistical features, such as GC-content and k-mer counts, similar to those of the positive ones (23). PBM and HT-SELEX PWM models were taken from UniPROBE database (9) and Jolma *et al.* (12), respectively. When multiple models were reported by one technology, we assigned to each genomic sequence the highest score obtained by such a model. We did not distinguish between human and mouse TFs because Jolma *et al.* (12) reported conservation of binding specificities between these species. Average AUC over the set of ChIP-seq experiments for each TF is reported. Complete results are shown in Supplementary Table S3.

Our results show that HT-SELEX models are more accurate in predicting *in vivo* binding (average AUC of 0.756 compared with 0.715, P-value = $9 \cdot 10^{-5}$ Wilcoxon signed-rank test) (Figure 3A). Trimming the PWM to the eight most informative positions results in average AUC of 0.732 and 0.719 (P-value = 0.18 Wilcoxon signed-rank test), respectively, hinting that the advantage of HT-SELEX models may be due to the addition of flanking positions. We note that the test set is too small to draw definitive conclusions, but we believe it points to an advantage of HT-SELEX models in predicting *in vivo* binding. For Tcf7, Srf, Mafk, Gata3 and Hnf4a HT-SELEX models, AUC is greater than that of PBM models by > 0.05 (Figure 1C and 3B). When excluding secondary PBM models, for Tcf7 and Mafk the average AUC increased from 0.61 to 0.81 and from 0.87 to 0.92, respectively, suggesting that some secondary models are wrong. At the same time, for Hnf4a the AUC dropped from 0.86 to 0.65. Similar results were observed on mouse ChIP-seq experiments downloaded from the ENCODE project (data not shown). Using the upstream sequences as control gave similar results (data not shown). When using a larger set of 1000 peaks, the advantage of HT-SELEX was smaller but still significant (data not shown).

We checked the effect of the source organism on predicting *in vivo* binding in human. Similarly, we compared the prediction quality based on experiments with full proteins compared to experiments using only the TF binding domains. None of the comparisons showed a significant difference (see Supplementary Information).

HT-SELEX experiments show systematic biases

Binding models for HT-SELEX use the most frequent k-mer in some cycle as a seed (6). To study the performance of these models on PBM data, we selected the most frequent 8-mer from each cycle and compared it with the top PBM 8-mer (determined by average binding intensity), when PBM data for the same TF were available (see Methods). We define a positive identification if the top 8-mer is identical with up to two mismatches to the top PBM 8-mer allowing an offset of up to two positions between the aligned sequences. The results are summarized in Figure 4A. Notably, in a substantial number of experiments, the most frequent HT-SELEX 8-mer in the last cycles did not match the top PBM 8-mer. Only 184 of 225 (81%) of the top HT-SELEX 8-mers in cycle 4 matched the top PBM 8-mer. Complete results are available in Supplementary Table S5.

Among the most frequent 8-mers in the different cycles, we observed many A-rich and C-rich 8-mers. To quantify this phenomenon, we focused on poly(A) and poly(C) 8-mers, defined as 8-mers containing at least 7 As or 7 Cs, respectively. Figure 4A shows an overabundance of such 8-mers as the most frequent 8-mers, especially in cycles 0–2. When comparing the distributions of poly(A), poly(C) and of other 8-mers in each cycle over all experiments, poly(A) and poly(C) 8-mers are much more abundant in the initial pool than the other 8-mers (median frequency $1.0 \cdot 10^{-3}$ and $5.66 \cdot 10^{-4}$ in cycle 0 and $9.4 \cdot 10^{-4}$ and $9.43 \cdot 10^{-4}$ in cycle 1, respectively, P-value < $3 \cdot 10^{-5}$ assuming a uniform null 8-mer distribution).

Moreover, certain 8-mers behaved differently in terms of their frequency changes between cycles. The poly(C) 8-mers were magnified from cycle to cycle much more than other 8-mers (Figure 4B). We also tested palindromic 8-mers (i.e. 8-mers that are identical to their reverse complement). We observed that palindromic 8-mers are less frequent initially (P-value = 0.002 in cycle 0 assuming a uniform null 8-mer distribution) and are less magnified than the rest of the 8-mers (Figure 4B, P-value = $2.2 \cdot 10^{-6}$ using a K–S test for comparing the rate of change between cycle 3 and cycle 4 of the palindromes with the other non-poly(A) and non-poly(C) 8-mers). Complete results are available in Supplementary Table S6. Ratio-based statistics showed the same phenomenon (data not shown).

Several reasons can explain the uneven abundance and magnification of k-mers. First, it can arise from technological artifacts. PCR biases have been observed and studied (24), and sequence bias is known to exist in high-throughput sequencing technologies, including the technologies used in Jolma *et al.* study (Illumina Genome Analyzer IIX and Hiseq2000) (25). We observed that nucleotide frequencies in the data are far

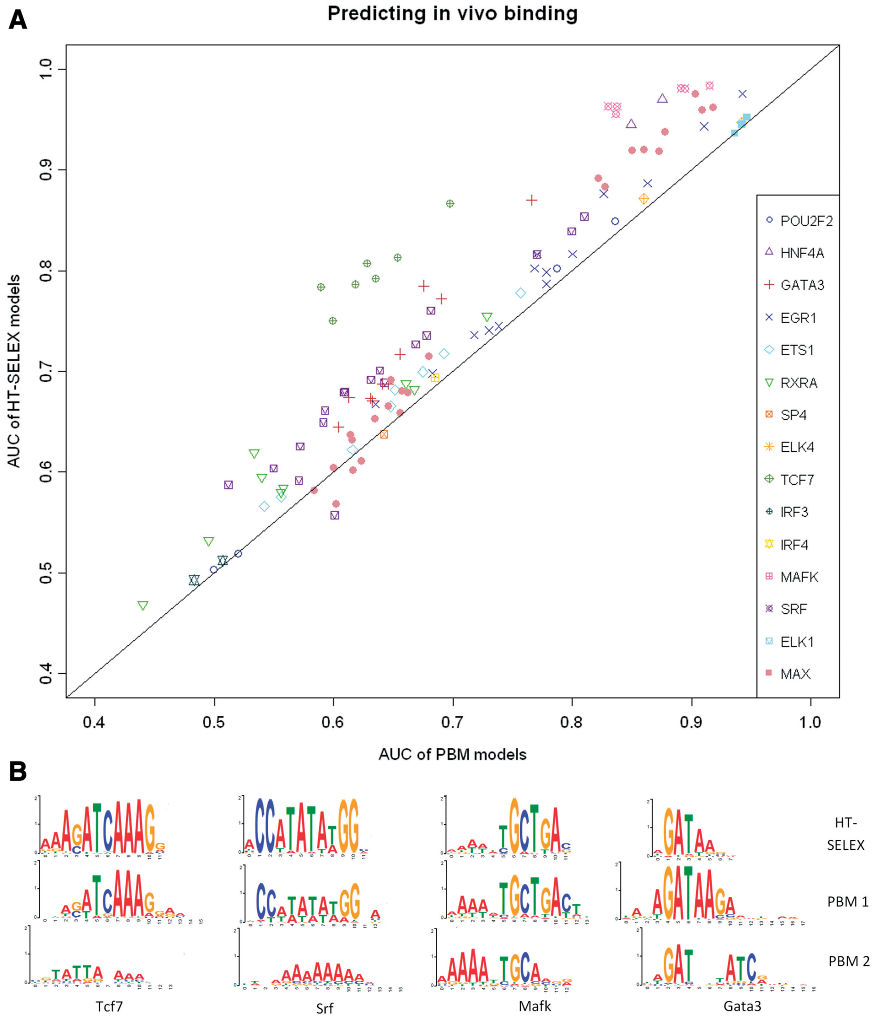


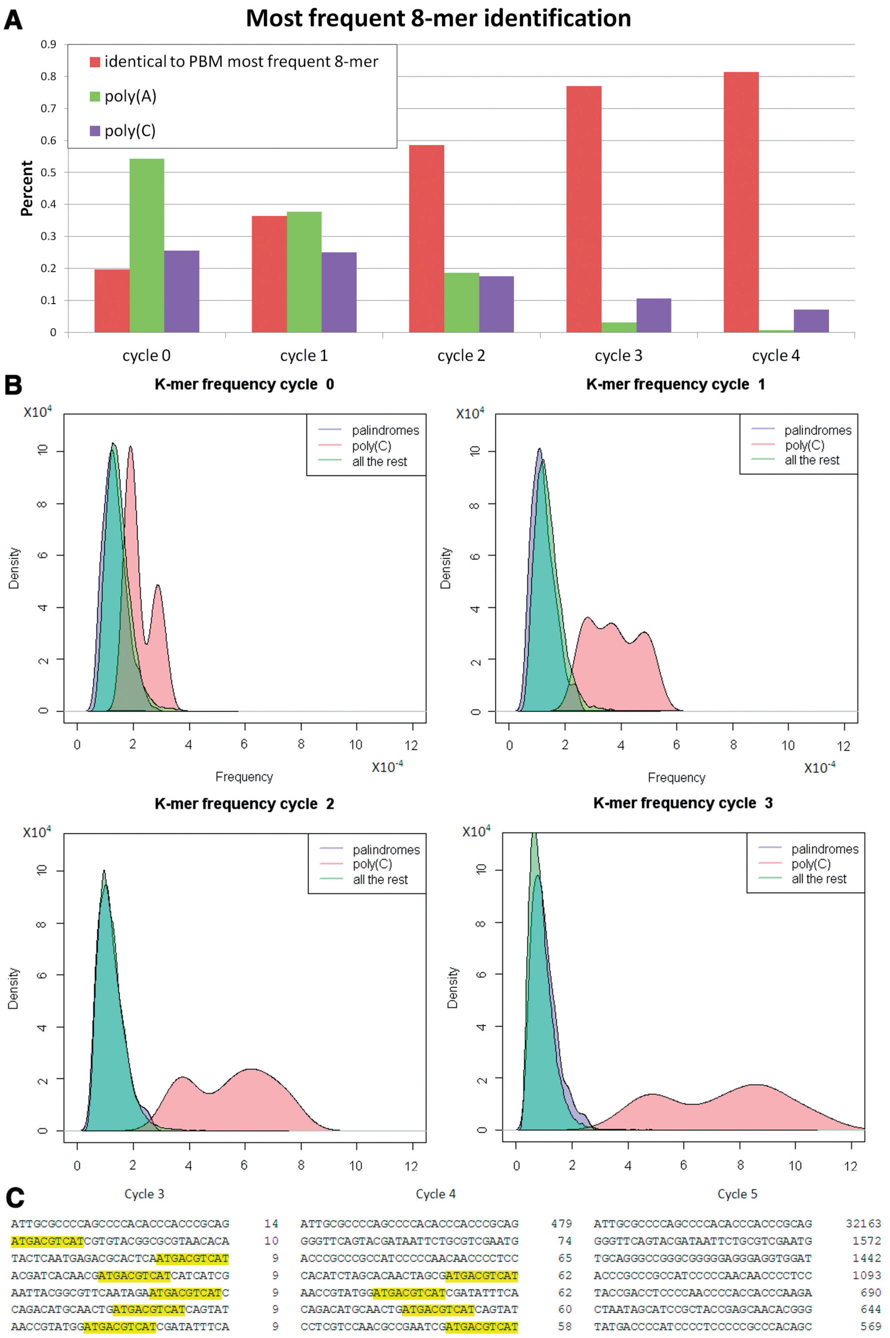
Figure 3. Predicting *in vivo* binding using HT-SELEX- and PBM-derived PWM models. The PWMs learned from HT-SELEX and PBM were taken from Jolma *et al*. and UniPROBE, respectively. *In vivo* binding was measured by the ENCODE project using ChIP-seq. (**A**) AUC results for each ChIP-seq experiment for which HT-SELEX and PBM experiments on the same TF are available. (**B**) Examples where HT-SELEX predicts *in vivo* binding better. For all these examples, the average AUC achieved by the HT-SELEX models exceeds that of the PBM models by >0.05.

from uniform, which can be explained by biased oligo generation (see Supplementary Information). Note that both the oligo generation and sequencing processes are strand-specific, so the frequencies of A and T (and of G and C) need not be equal. The systematic overrepresentation of specific k-mers has been observed both *in vivo* [in ChIP-seq data (26)] and *in vitro* [in PBMs (27), where it was termed 'sticky k-mers']. According to Jiang *et al*., in PBM the set of sticky k-mers are all A-rich except CCCCGCCC, in partial agreement with our observations on HT-SELEX data. An alternative explanation suggested by a recent theoretical study was that TFs bind non-specifically to homogenous sequences (28). The underrepresentation of palindromes may be due to the formation of secondary structures that hinder PCR of such sequences (See Supplementary Information).

False oligos are common in HT-SELEX

Because whole reads (oligos) are sequenced and selected by the HT-SELEX technology, we also conducted an analysis of the abundance and magnification properties of oligos. For each TF, we identified the most frequent oligos in the last cycles. For the 100 most frequent oligos, we defined as false oligos those that do not contain any of the seeds reported in (12) allowing one mismatch. We also measured the oligo enrichment ratio, defined as the oligo's frequency in the last cycle divided by its frequency in the previous cycle.

The false oligos were on average 25% of the 100 most frequent oligos in the last cycle. In 113 experiments (of 547), at least 50 of the 100 most frequent oligos in the last cycle were false. We observed two characteristics common to them. First, they tended to have more skewed nucleotide distribution than true oligos, with high frequency of one nucleotide (C in 75% of the cases). In all, 35% of the false oligos had one nucleotide composing at least 50% of the sequence, compared with 14% in the true oligos. Second, they tended to be extremely magnified, rising from a low count (or zero) in one cycle to a high count in the next. For example, 41% of



Cycle 3		Cycle 4		Cycle 5	
ATTGCGCCCCAGCCCCACACCCACCCGCAG	14	ATTGCGCCCCAGCCCCACACCCACCCGCAG	479	ATTGCGCCCCAGCCCCACACCCACCCGCAG	32163
ATGACGTCATCGTGTACGGCGCGTAACACA	10	GGGTTCAGTACGATAATTCTGCGTCGAATG	74	GGGTTCAGTACGATAATTCTGCGTCGAATG	1572
TACTCAATGAGACGCACTCAATGACGTCAT	9	ACCCGCCCGCCATCCCCCAACAACCCCTCC	65	TGCAGGGCCGGGCGGGGGAGGGAGGTGGAT	1442
ACGATCACAACGATGACGTCATCATCATCG	9	CACATCTAGCACAACTAGCGATGACGTCAT	62	ACCCGCCCGCCATCCCCCAACAACCCCTCC	1093
AATTACGGCGTTCAATAGAATGACGTCATC	9	AACCGTATGGATGACGTCATCGATATTTCA	62	TACCGACCTCCCCAACCCCACCACCCAAGA	690
CAGACATGCAACTGATGACGTCATCAGTAT	9	CAGACATGCAACTGATGACGTCATCAGTAT	60	CTAATAGCATCCGCTACCGAGCAACACGGG	644
AACCGTATGGATGACGTCATCGATATTTCA	9	CCTCGTCCAACGCCGAATCGATGACGTCAT	58	TATGACCCCATCCCCTCCCCCGCCCACAGC	569

Figure 4. Systematic biases in HT-SELEX technology. (**A**) Properties of the most frequent 8-mer in different cycles. For each cycle, the fraction of times the most frequent 8-mer in the HT-SELEX experiment was poly(A), poly(C) or matched the most frequent 8-mer computed from PBM data is presented (see text). (**B**) The 8-mer frequency density plots for each cycle. The 8-mers were partitioned into three categories: palindromes, poly(C) and all the rest. For each category, a smoothed density plot of its 8-mer frequencies is shown. (**C**) Abundant false oligos in Atf7 HT-SELEX experiment. For cycles 3, 4 and 5, the seven most frequent oligos are shown along with their counts. The consensus sequence is highlighted in yellow (none of the top seven oligos in cycle 5 contain the consensus).

the false oligos were not observed in the one-before-last cycle, compared with 19% of the true oligos (note that an oligo present in a particular cycle may have not been observed because of limited sampling). Figure 4C shows an example of Atf7 HT-SELEX experiment. Complete results are available in Supplementary Table S8. Of the previous studies, we observed similar biases in (6) and (8), but not in (7) (see Supplementary Information).

DISCUSSION

Protein–DNA binding has been in the focus of gene regulation studies for years. In the past, binding sites were defined based on few examples and thus had low resolution and limited accuracy. With technological developments, the ability to measure and predict binding sites has improved. A large leap came in the form of PBMs, which measure *in vitro* the binding intensity of a specific TF to thousands of probes, designed to cover all 10-mers (4). Binding models derived from these data performed well on other PBM data but less so on *in vivo* data (10). One possible explanation was that they reflect PBM artifacts together with the specific binding. How well PBM models represent *in vivo* TF–DNA binding remained an open question.

The emergence of new high-throughput *in vitro* technologies allowed us to deepen our understanding on this question. The HT-SELEX technology measures TF–DNA binding using high-throughput sequencing (6–8). Recently, Jolma *et al.* (12) reported HT-SELEX experiments covering hundreds of TFs, where many of them had been tested on PBM as well. This gave the first opportunity to compare on a large-scale models derived from two independent high-throughput *in vitro* technologies. Through this comparison, we could identify some of the advantages and disadvantages of each technology and determine how relevant *in vitro* models are to *in vivo* binding. A small-scale comparison by Jolma *et al.* (12) covering 14 models reported a few differences.

Our comparison shows that for most TFs the PBM and HT-SELEX technologies produce PWM models that are in good agreement. On average over 246 PBM experiments, the AUC when using the HT-SELEX-derived model for predicting PBM probe binding was 0.875. Moreover, in a model-independent comparison, the average correlation between HT-SELEX 8-mer counts in cycle 3 and PBM average binding intensities over the set of top 100 PBM-ranked 8-mers was 0.74. We observed that PBM-based 8-mer ranking is more accurate and robust than HT-SELEX-based ranking, and that the ranking 8-mers by their occurrence frequency in the Jolma *et al.* HT-SELEX data is better than ranking by between-cycle ratio score. We speculate that this is due to the relatively low read coverage in these experiments [compared with SELEX-seq data, where ratio-based scores were used (7)]. Although each HT-SELEX experiment reported hundreds of thousands of oligos, the SELEX-seq experiments had millions. We conclude that high coverage is necessary to derive accurate ratio scores. For some families of TFs, the two technologies give discordant results, perhaps because of differences in DNA structure [e.g. the HMG proteins, for which structure plays a larger role in binding (21)]. In comparison with *in vivo* data from ChIP-seq experiments, HT-SELEX models had better binding prediction, partly because of the ability to model the side positions more accurately. However, the set of TFs for which HT-SELEX, PBM and ChIP-seq data were available was rather modest, and larger tests are needed.

In analyzing the similarity between the top 8-mers determined by PBM and by HT-SELEX in each cycle, we observed the previously reported phenomenon of overspecification. Although 8-mer frequencies in the initial HT-SELEX cycles are too non-specific and similar to the initial pool (i.e. closer to random), the last cycles can, in some cases, be too specific. There is a trade-off between better coverage of top k-mers in later cycles, which can improve the binding model accuracy, and overrepresentation of few top k-mers, which can make the model too narrow, disregarding weaker binding motifs. This was noted in (6) and in previous studies using the SELEX technology (29).

In the course of our analysis, we observed and characterized several strong biases in many experiments in the HT-SELEX technology. First, we found a systematic bias toward certain types of k-mers [similar but not identical to the ‘sticky k-mers’, reported for PBM data (27)]. For many TFs, in the last cycle C-rich 8-mers are among the most frequent (Figure 4). For example, in 7% of the experiments the most frequent 8-mer in the last cycle contained at least 7 Cs. These phenomena can be explained by PCR and sequencing biases (25) or perhaps by non-specific TF binding (28). Moreover, when measuring oligo (whole read) frequencies, we found that in some experiments the oligos with the highest frequency and those whose frequencies increased fastest between cycles did not contain the binding site; we call them ‘false oligos’. We observed these phenomena in the previous studies (6) and (8), but not in (7). Slattery *et al.* were the only ones to isolate bound oligos through a mobility shift assay, which suggests that this phase removes false oligos and thus improves the quality of the data.

Our analysis suggests that each of the HT-SELEX and PBM technologies has its advantages. PBM data are more accurate and robust in 8-mer ranking; HT-SELEX seems to be superior in *in vivo* binding prediction and allows better learning of longer motifs. We recommend using higher read coverage in HT-SELEX experiments, as was done in (7), to produce more sensitive models. We note that our comparisons and conclusions are limited to the specific technological implementations of HT-SELEX and PBM tested, for which the large-scale overlap exists. Unfortunately, we could not compare SELEX-seq and context-genomic PBMs because of fewer data sets.

Our study aimed to provide deeper and broader analysis of the properties of HT-SELEX experiments and to put them in the context of other high-throughput technologies for evaluating TF–DNA binding *in vivo* and *in vitro*. In the future, we plan to extend this work in several directions. First, we intend to use the new insights to design better motif finding algorithms based on HT-SELEX data. Second, we can learn a binding model based on the biomechanical mechanism of TF–DNA binding using regression methods that use k-mer counts [as in (8)]. Third, we plan to learn more complex binding models. More specifically, we plan to incorporate in the models 2-mer features as well as DNA shape features, as was done recently using custom PBM (30), and demonstrated using existing motif databases (31). The rich and broadly available HT-SELEX data provide a great opportunity to improve our understanding of TF–DNA binding.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Jussi Taipale for fruitful discussions and for providing them with the HT-SELEX oligo batch numbers and Kobi Perl for his comments on the manuscript.

FUNDING

Israel Science Foundation (ISF) [802/08, 317/13]; Edmond J. Safra Center for Bioinformatics at Tel Aviv University, the Dan David Foundation, and the Israeli Center for Research Excellence (I-CORE), Gene Regulation in Complex Human Disease, center 41/11 (to Y.O). Funding for Open Access charge: ISF grant [317/13] and I-CORE.

Conflict of interest statement. None declared.

REFERENCES

1. Aparicio,O., Geisberg,J.V. and Struhl,K. (2004) Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences *in vivo*. *Curr. Protoc. Cell Biol.*, **Chapter 17**, Unit 17.7.
2. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, **316**, 1497–1502.
3. Rhee,H.S. and Pugh,B.F. (2012) ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr. Protoc. Mol. Biol.*, **Chapter 21**, Unit 21 24.
4. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
5. Fordyce,P.M., Gerber,D., Tran,D., Zheng,J., Li,H., DeRisi,J.L. and Quake,S.R. (2010) De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nat. Biotechnol.*, **28**, 970–975.
6. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J., Sillanpaa,M.J. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861–873.
7. Slattery,M., Riley,T., Liu,P., Abe,N., Gomez-Alcala,P., Dror,I., Zhou,T., Rohs,R., Honig,B., Bussemaker,H.J. *et al.* (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, **147**, 1270–1282.
8. Zhao,Y., Granas,D. and Stormo,G.D. (2009) Inferring binding energies from selected binding sites. *PLoS Comput. Biol.*, **5**, e1000590.
9. Robasky,K. and Bulyk,M.L. (2011) UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **39**, D124–D128.
10. Orenstein,Y., Linhart,C. and Shamir,R. (2012) Assessment of algorithms for inferring positional weight matrix motifs of transcription factor binding sites using protein binding microarray data. *PLoS One*, **7**, e46145.
11. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.
12. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
13. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
14. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
15. Chen,X., Hughes,T.R. and Morris,Q. (2007) RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. *Bioinformatics*, **23**, i72–i79.
16. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
17. Zhao,Y. and Stormo,G.D. (2011) Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat. Biotechnol.*, **29**, 480–483.
18. Orenstein,Y., Mick,E. and Shamir,R. (2013) RAP: accurate and fast motif finding based on protein-binding microarray data. *J. Comput. Biol.*, **20**, 375–382.
19. Berger,M.F., Badis,G., Gehrke,A.R., Talukder,S., Philippakis,A.A., Pena-Castillo,L., Alleyne,T.M., Mnaimneh,S., Botvinnik,O.B., Chan,E.T. *et al.* (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, **133**, 1266–1276.
20. Wei,G.H., Badis,G., Berger,M.F., Kivioja,T., Palin,K., Enge,M., Bonke,M., Jolma,A., Varjosalo,M., Gehrke,A.R. *et al.* (2010) Genome-wide analysis of ETS-family DNA-binding *in vitro* and *in vivo*. *EMBO J.*, **29**, 2147–2160.
21. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.*, **79**, 233–269.
22. Bailey,T.L. and Machanick,P. (2012) Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.*, **40**, e128.
23. Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
24. Aird,D., Ross,M.G., Chen,W.S., Danielsson,M., Fennell,T., Russ,C., Jaffe,D.B., Nusbaum,C. and Gnirke,A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.
25. Hansen,K.D., Brenner,S.E. and Dudoit,S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.*, **38**, e131.
26. Barski,A. and Zhao,K. (2009) Genomic location analysis by ChIP-Seq. *J. Cell Biochem.*, **107**, 11–18.
27. Jiang,B., Liu,J.S. and Bulyk,M.L. (2013) Bayesian hierarchical model of protein-binding microarray k-mer data reduces noise and identifies transcription factor subclasses and preferred k-mers. *Bioinformatics*, **29**, 1390–1398.
28. Afek,A. and Lukatsky,D.B. (2013) Genome-wide organization of eukaryotic preinitiation complex is influenced by nonconsensus protein-DNA binding. *Biophys. J.*, **104**, 1107–1115.
29. Klug,S.J. and Famulok,M. (1994) All you wanted to know about SELEX. *Mol. Biol. Rep.*, **20**, 97–107.
30. Gordan,R., Shen,N., Dror,I., Zhou,T., Horton,J., Rohs,R. and Bulyk,M.L. (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.*, **3**, 1093–1104.
31. Yang,L., Zhou,T., Dror,I., Mathelier,A., Wasserman,W.W., Gordan,R. and Rohs,R. (2013) TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Res.*, **42**, D148–D155.

6. Discussion

In this thesis we described our study on the theoretical and practical aspects of motif finding in high-throughput *in vitro* data. We developed novel algorithms for inferring TF-DNA binding preferences from these data. We implemented our methods efficiently, demonstrated their applicability to data generated by a range of technologies, and showed that they outperform existing tools. We applied computational analyses to compare between implementations of two different technologies that measure protein-DNA binding. We highlighted the advantages and disadvantages of each technology and observed several biases in the new HT-SELEX data. Finally, we developed new efficient algorithms for a sequence design problem that are related to PBMs. In the future, we hope that the practical tools and techniques we implemented and the theoretical algorithms we developed will be of use to researchers in biology and computer science, respectively.

6.1 Inferring binding site motifs from PBM data

One contribution of this thesis is the algorithms for motif discovery in PBM data. Protein binding microarrays are a leading technology to measure in a high-throughput and unbiased manner the DNA-binding preferences of a TF *in vitro.* Hundreds of TFs were examined using PBMs and the datasets were deposited in public databases. Several methods have been developed for the task of inferring a binding site motif from PBM data, including the method developed by us, Amadeus-PBM, described in Chapter 2. We assessed the performance of Amadeus-PBM and extant methods and found that they fall into two disjoint categories, where each category is better at a different task. Following these insights, we developed the RAP algorithm. This method performs best in all benchmarks, as described in Chapter 3. Through the development of RAP, we learned more about the characteristics of models representing protein-DNA binding preferences.

6.1.1 Amadeus-PBM algorithm

Amadeus-PBM, described in Chapter 2, searches for motifs that are over-represented in the top 1000 9-mers of a given PBM experiment. The k-mers are ranked by the average binding intensity. From our experience this score produces an accurate and robust ranking. Over-representation of the motif in the top of the ranked list is evaluated using the standard hypergeometric score. The general architecture of Amadeus is a pipeline of filters, or refinement phases [37].

Amadeus-PBM produces interpretable motifs in a very short time (less than 30 seconds). On extensive and large-scale tests, we found that the models produced by the algorithm resemble motifs from public databases. These were learned from data of independent technologies, implying that Amadeus-PBM models do not suffer from over-fitting the biases in PBM data or artifacts of the technology. On the other hand, the models are less accurate in predicting the binding of another PBM experiment on the same TF with a different array design. The success of other methods may result from learning specific technological biases and incorporating this information in the model. In our tests for predicting *in vivo* binding, there is no clear winner and the results are much worse than predicting *in vitro* binding. In an international competition carried out to discover the TF based on PBM experiment data, Amadeus-PBM performed best (tied with another algorithm) [56]. The algorithm is implemented as part of the Amadeus software package and benefits from its user-friendly interface. The software is publicly available at http://acgt.cs.tau.ac.il/amadeus.

In addition to Amadeus-PBM's applicability in inferring a binding site model from PBM, the platform includes a wealth of features, such as combined analysis of multiple datasets from one or more organisms and one or more technologies (e.g., PBM and ChIP), built-in bootstrapping, motif-pairs analysis, and comparison to known TF binding sites from public databases (e.g., TRANSFAC and JASPAR). Amadeus-PBM is easily accessible to biologists, since it is “wrapped” in an informative, user-friendly graphical interface. Amadeus-PBM is in fact a generic scheme that can use any score to rank the k-mers and any motif finding algorithm to infer a model based on the top ranking k-mers. This scheme is applicable to data produced from any technology measuring protein-DNA binding. For example, the algorithm has been applied to MITOMI data [63]. It is successfully employed in the lab of Dr. Doron Gerber from Bar-Ilan University to produce models based on their MITOMI experiments.

6.1.2 RAP algorithm

In Chapter 3 we described the RAP algorithm for inferring binding site motifs from PBM data. The algorithm performs the same k-mer ranking as Amadeus-PBM. Then, it aligns the top k-mers and produces a weighted matrix based on this alignment and the k-mer scores. As opposed to other algorithms that learn the model's parameters based on the complete dataset, RAP relies on a set of high-affinity k-mers and their weights. It improves over Amadeus-PBM since it uses k-mer scores and extends the PWM to more than 10 positions. Amadeus-PBM performs better than RAP when the derived k-mer scores are not as accurate and robust as in PBM (as we observed for some MITOMI data). We tested RAP and competing methods on three large-scale benchmarks. In terms of similarity to known motifs and predicting the binding of a PBM experiment on the same TF with a different array design, RAP performed best (tied with a different method in each task). The task of predicting *in vivo* binding based on the *in vitro* models remains difficult: all methods performed much worse in this task. Notably, in a recent study combining chromatin accessibility information together with PBM-derived binding models, RAP models performed best in predicting *in vivo* biding [69]. The RAP algorithm was implemented efficiently and runs in a couple of seconds. It is freely available at http://acgt.cs.tau.ac.il/RAP.

6.1.3 Predicting *in vivo* binding from PBM data

The area of learning DNA-binding preferences of proteins from PBM data has been extensively studied. State-of-the-art methods produce models that predict *in vitro* binding quite accurately. In contrast, constructing accurate models for *in vivo* binding is a harder challenge. In all studies, while predicting *in vitro* binding was very accurate (average AUC reaching almost 0.9), predicting *in vivo* binding was much worse (average AUC 0.7). When testing PBM-derived models on ChIP-seq experiments available on ENCODE (considered more accurate than the ChIP-chip experiments used in Chapter 2), the results were effectively the same. The average AUC was 0.69 over 24 TFs that had both a PBM model and a ChIP-seq experiment. A comparison of the average AUC of different methods revealed that no method was doing better than the others (see Figure 4). The low accuracy may be due to the complexity of the cellular environment and also due to the simplicity of the produced models. Longer models, as produced by HT-SELEX, are more accurate in *in vivo* binding prediction, achieving an average AUC of 0.74 over 51 TFs, hinting that some of the signal directing protein-DNA binding lies in the flanking regions of the core motif. Other probable factors that have to be taken into account in *in vivo* binding in addition to sequence-specific features include nucleosome positioning, competing TFs and cooperating TFs.

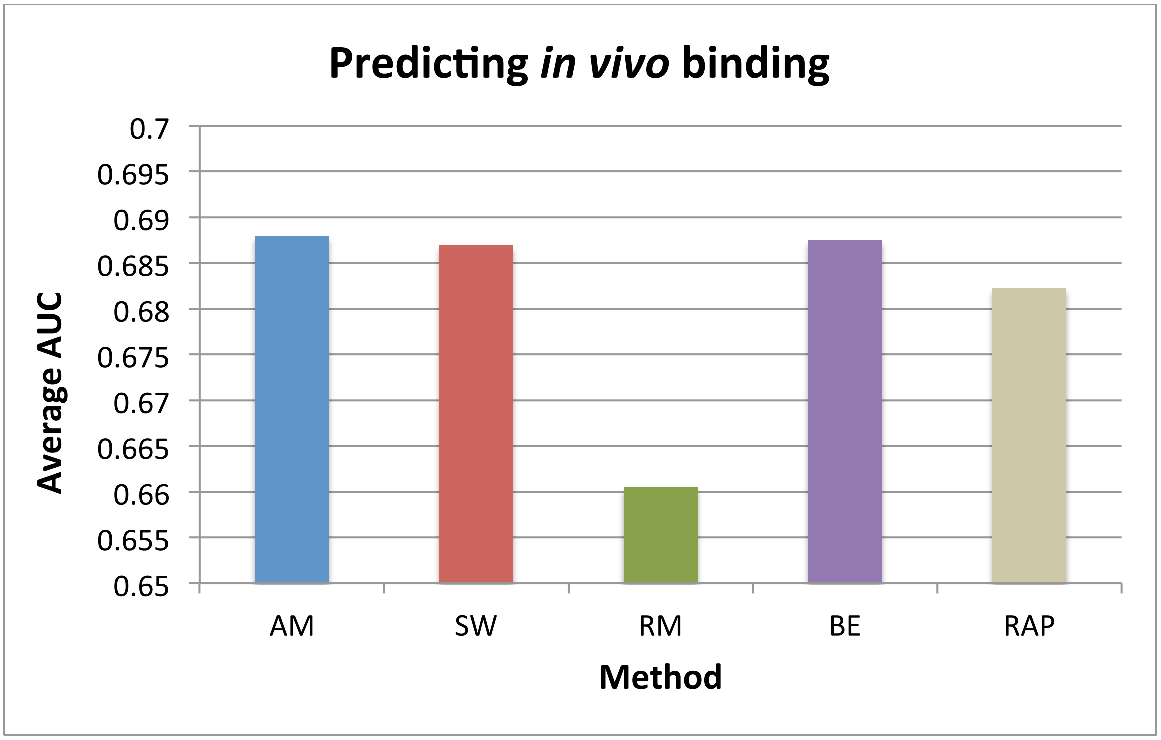


Figure 4. Average AUC of PBM-based models in predicting *in vivo* binding. Models were generated using different methods based on PBM data and used to predict ENCODE ChIP-seq binding for 24 TFs. Note that the average AUC of 0.69 is higher than that achieved on yeast ChIP-chip data (compare Figure 4 in Chapter 2), probably due to the improved accuracy of ChIP-seq experiments. AM: Amadeus-PBM, SW: Seed-and-Wobble, RM: RankMotif++, BE: BEEML-PBM.

6.1.4 Benchmarking tools for motif discovery in PBM data

Several tools for motif discovery in PBM data have been described in the literature (see Section 1.3.2). Comparing them on a large-scale is important to understand the advantages and disadvantages of each method, and highlight the limitations in the state-of-the-art. A good benchmark for reliably comparing the performance of different tools should be based on a large number of real, heterogeneous, experimentally-derived datasets. We conducted the first large-scale comparison and used benchmarks on three different axes. The first compares the models to previously indentified motifs. The second evaluates the performance of the inferred model in predicting *in vitro* binding by predicting the binding of a PBM experiment on the same TF, but with a different array design. Last, we evaluated *in vivo* binding prediction by predicting binding in high-throughput ChIP experiments. We hope other researchers use these benchmarks to test and improve their methods, and extend it with additional datasets from various sources.

6.1.5 The PWM model

An ongoing debate exists in the field of protein-DNA binding modeling regarding the validity of the PWM model. The PWM is the most popular model for representing DNA binding preferences. Our studies have shown that the model is quite accurate for predicting *in vitro* binding. When using the models derived from PBM data to predict the binding of another PBM experiment, the average AUC was 0.9. Moreover, as described in Chapter 5, when using HT-SELEX-derived models to predict the binding of a PBM experiment, prediction accuracy was also very high - reaching average AUC of 0.875. As noted before, this model may be too simplistic for predicting *in vivo* binding.

What are the characteristics of an accurate PWM? In Chapter 2, we observed that methods that generate binding models from PBM data fall into two categories: some produce interpretable models, similar to known motifs, and others produce models that are more accurate for predicting *in vitro* binding. The models produced by our RAP algorithm bridge between these two: the models are both interpretable and accurate. In terms of information content, which is a measure of degeneracy, interpretable models are stricter, while accurate models are more degenerate. RAP is somewhere in the middle, which may explain, in part, its high performance in both categories.

Moreover, we found that the flanking positions in the motif affect the binding, with smaller effect than the core positions. Longer motifs (up to 14bp-long) produced by RAP algorithm perform better in predicting *in vitro* binding. Still, the task of learning these flanks is not easy. For example, the models inferred by Seed-and-Wobble algorithm perform better without the flanks, hinting that these flanks are not learned well. On the other hand, RAP successfully learns the side positions despite the limited coverage of PBM arrays. In our study of HT-SELEX-derived models in Chapter 5, the flanks significantly improved the performance of predicting *in vivo* binding. Other studies hypothesized that these positions determine the DNA-shape locally, and by that affect the binding [70]. In conclusion, the addition of flanking positions to the PWM model improves its performance, if learned correctly subject to the data constraints.

6.2 Comparing HT-SELEX and PBM

In Chapter 5 we conducted a large-scale comparison between two implementations of high-throughput *in vitro* technologies for measuring protein-DNA binding. As noted previously, PBM-derived models are very accurate in predicting binding intensities of another PBM experiment, but much worse in predicating *in vivo* binding (as measured in ChIP experiments). Data from an independent high-throughput *in vitro* technology was needed to validate these models. HT-SELEX technology 'came to the rescue'. A recent study tested hundreds of human TFs in HT-SELEX experiments, 162 of which had a PBM experiment. For the first time, this comparison was possible.

We performed a large-scale comparison using three different benchmarks. The first used the models derived from HT-SELEX to predict PBM binding. This was compared to models derived from PBM data of another array or using cross-validation, if such a model was not available. Second, since model inference highly depends on the algorithm, we performed a model-independent comparison. A list of top k-mers was generated from one technology, and their binding scores were compared to the scores derived from the other technology. Last, we used a third independent technology, ChIP-seq, to decide which models are more accurate. We believe that the ideas implemented in these benchmarks may be useful to other comparisons and method assessments. As technologies are evolving quickly and producing data in the hundreds and thousands, such comparisons are often being called for and can be conducted on a large scale.

In our comparison, we found that, on the whole, models derived from these technologies mostly agree. The average AUC of HT-SELEX-derived models in predicting PBM binding intensities was 0.825, compared to 0.9 using PBM-derived models on paired experiments. The disagreements are limited to several TF families, such as Sox proteins and zinc fingers. When using a dataset of unpaired PBM experiments, the average AUC of HT-SELEX-derived models was 0.925, implying that for the protein families in this dataset, homeodomain and ETS, the technologies are in good agreement. We derived three conclusions from the model-independent comparison. First, compared to the HT-SELEX data produced by the Taipale lab, PBM data are more robust in ranking a set of k-mers according to their binding intensities. Second, the read coverage in the experiment greatly affects the k-mer ratio statistic. For accurate and robust estimation of the binding affinities, a read coverage of at least a million for each k-mer is required. With such coverage, accurate ratios may be estimated, and the ranking may be as good as or even better than that achieved by PBM technology. Third, over-specification may occur at later cycles. High-affinity k-mers may be over-enriched at the expense of low-affinity k-mers. Notably, in our last test we found that HT-SELEX-derived models are more accurate in predicting *in vivo* binding. We believe that this is mostly due to their ability to generate longer motifs, since when we removed these positions the advantage was no longer significant.

When comparing the accuracy of *in vivo* peak binding prediction based on PBM and HT-SELEX models for the same TFs, using ENCODE ChIP-seq peaks and BEEML-PBM-derived models, HT-SELEX remains superior (see Figure 5). Interestingly, the results for individual TFs are quite different from those obtained by Seed-and-Wobble-derived models, demonstrating the effect the algorithm of choice has on prediction accuracy (compare Figure 3 in Chapter 5).

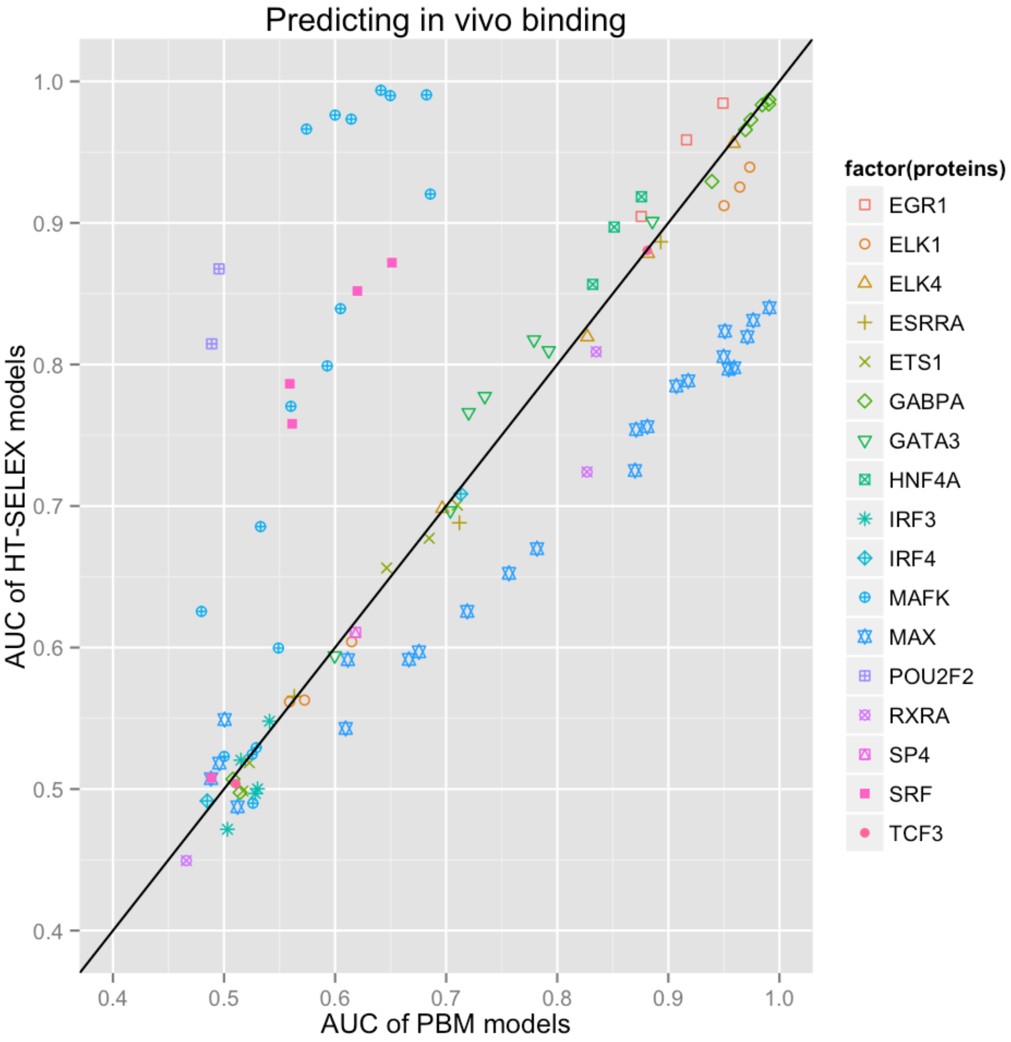


Figure 5. Comparison of PBM and HT-SELEX-derived PWM models in predicting *in vivo* binding. For 17 TFs that had a PBM and HT-SELEX-derived models and ChIP-seq experiments from ENCODE, a PWM model was used to predict the top 500 peaks. This figure is analogous to Figure 3 in Chapter 5 but is using the most updated ENCODE ChIP-seq data and BEEML-PBM-derived models rather than Seed-and-Wobble.

6.2.1 Systematic biases in HT-SELEX technology

Through our comparison, we revealed several biases in HT-SELEX technology. First, some k-mers are enriched in all experiments (in the PBM technology, such enriched k-mers were called 'sticky k-mers' [71]). These are usually C-rich k-mers. Their abundance makes it difficult to identify the TF's consensus sequence. Removing them is not an easy task, since some TFs bind C-rich motifs. We give two possible explanations to the this phenomenon: one is due to biases in the technology, such as sequencing biases and PCR biases; the other is due to non-specific binding of TFs to homogenous oligos. Second, we observed 'false oligos' in many experiments. These are oligos that are the most frequent, but do not contain the binding site. They show extremely high amplification rate from cycle to cycle and are homogenous in their nucleotide composition. Taking it all into account, biases in HT-SELEX must be overcome in order to derive accurate binding models and benefit from the richness of the data.

6.2.2 Future plans

The new data provide several opportunities to further study the mechanisms behind TF-DNA binding. Since HT-SELEX measures the binding to longer motifs, additional features may be derived from sequences flanking the core. These may be added to the PWM as side positions or as local DNA shape features, as was proved useful in a recent study [70]. Biomechanical models based on free energy contributions may be learned from high-quality data (using algorithms such as BEEML and FeatureREDUCE) to improve accuracy. Such models require high-quality data, so in order to employ them successfully data quality has to improve, either computationally or experimentally. Hopefully, using the published data and other data, more can be learned on the binding mechanisms of different protein families, and reveal the mechanisms that differentiate between proteins in the same family.

6.3 Sequence design algorithms

In Chapter 4 we developed a new algorithm for solving a sequence design problem with applications to protein binding microarrays and synthetic enhancers. Both of these technologies require a set of probe sequences that cover together all possible DNA k-mers for some k. In the first, the space on the array is limited, while in the second the number of experiments that can be performed is bounded. Since the set of probes are double-stranded DNA sequences, when a k-mer appears in one strand, its reverse complement appears in the other. Thus, a saving of up to 50% in the number of sequences is possible. A reverse complement de Bruijn sequence is a sequence that for each k-mer either the k-mer or its reverse complement is covered. The theoretical problem is how to design a shortest reverse complement de Bruijn sequence.

We solved this problem optimally. First, we gave a lower bound for the length of such a sequence based on k-mer counts. Then, we described an algorithm for finding two reverse complement Euler tours in a de Bruijn graph. The algorithm works on graphs with certain properties. A de Bruijn graph of even order (k-1) has these properties. Thus, for generating a reverse complement de Bruijn sequence of order k, when k is odd, the algorithm can be run on a de Bruijn graph of order (k-1). Two sequences are produced, represented as the Euler cycles found by the algorithm. The running time is linear in the size of the graph, which is $\Theta(|\sum|^k)$. It is optimal since this is the length of the output sequence.

For even k, the problem is more complex due to palindromes. Palindromes are reverse complements of themselves and can only be of even length. A de Bruijn graph of odd order (k-1) contains palindromes. The algorithm that constructs two Euler tours cannot be run on such graphs. We provided two solutions to this problem, both by augmenting a de Bruijn graph with additional edges. The first is sub-optimal and runs in linear time and the second is optimal, but runs in higher polynomial time. The first augments the graph systematically by adding all cyclic shifts of palindromes. Then the algorithm can find two reverse complement Euler tours that together cover all edges. An optimal augmentation can be achieved by solving a maximum weight matching. The matching finds the smallest set of edges to add, so that in the augmented graph two reverse complementary Euler tours exist. Its running time is $\Theta(k\ |\sum|^{5k/4} \log(|\sum|))$. The length of the output sequence is equal to that obtained by the sub-optimal algorithm for $k \leq 8$ and is only slightly less for greater k's. The algorithm due to Riesenfeld *et al.* [66] aims to produce the smallest set of sequences that cover all DNA k-mers, while utilizing the reverse complementarity property of double-stranded DNA. Unfortunately, it has prohibitive running time for realistic k values. In comparison, our optimal algorithm terminates in less than one hour for $k \leq 12$, while Riesenfeld's algorithm on k=12 did not terminate after more than a month.

Our algorithm for finding two reverse complement Euler tours can be applied in other sequence design problems. For example, we are now developing a new efficient algorithm for a similar problem, in which the sequence is allowed to include each k-mer at most once, in either orientation. The biological motivation comes from microarray sequence design that avoids self- and cross-hybridizations. We have developed an algorithm that produces an optimal solution in polynomial time, based on minimum-cost maximum-flow algorithm in de Bruijn graphs. We believe that the ideas applied in our algorithm are useful to other sequence design problems based on de Bruijn graphs.

6.4 The road ahead

Biological technologies progressed tremendously in recent years. They can measure today in a high-throughput manner millions of interactions in a single experiment. As part of these developments, protein-DNA binding can be measured accurately over a wide spectrum of sequences. The vast data produced by each experiment cannot be analyzed manually. Computational methods were critical in the processing these data and producing an accurate and compact model to represent TF-specific DNA binding preferences.

We believe that techniques for measuring protein-DNA binding will continue to improve thanks to reduced costs of microarrays and deep sequencing platforms. The HT-SELEX technique demonstrates the benefit of high-throughput sequencing in measuring protein-DNA binding [12-14]. One of its main advantages over previous techniques is the ability to measure motifs of length longer than 20bp [4]. Its accuracy will continue to improve with greater read coverage as the cost of deep sequencing continues to decrease. Universal PBMs were recently extended by context-genomic PBMs, which measure the binding of a TF to a pre-defined set of genomic sequences. For example, to test the effect of different flanking sequences on the binding, all sequences containing a specific core motif were placed on one array [70, 72]. As production of microarrays and oligo printing becomes cheaper, it is now possible to design arrays to test the binding preference of TFs to specific genomic sequences *in vitro*.

Technological advancements have been made in measuring *in vivo* binding as well. While ChIP-chip measures *in vivo* binding to a pre-defined set of promoters, and ChIP-seq can detect *in vivo* binding to regions of around 100bp, ChIP-exo can measure TF-DNA *in vivo* binding in nearly single base-pair resolution [73]. Moreover, techniques have been developed to measure other confounding factors that affect *in vivo* binding, such as nucleosome occupancy and other epigenetic marks [74, 75]. On top of that, it is even possible to manipulate an organism's genome to test the effect of different genomic or synthetic regulatory elements (in promoter or enhancer regions) on its phenotype [66, 76]. All of these will surely help in improving our understanding of *in vivo* binding and developing more accurate predictive models.

On the computational side, we see more complex binding models emerging to replace the 'good old PWM'. A long-standing debate has been going on the accuracy of the PWM model [56, 77]. More and more studies are starting to criticize the position-independence assumption and suggest more complex model, mostly adding position-dependent features, such as di-nucleotide and 3-mers [77]. The benefit of these additional features may be explained by their effect on local DNA shape features [78]. While these models have been shown to be more accurate, and it is possible to infer them from the new high-throughput data, they are rarely used. There are two main challenges. The first is the interpretability. Models gain popularity when accompanied by a user-friendly and intuitive visualization, which is still missing for the more complex models. Second, it is difficult for a new model to reach broad impact, when most bioinformatics tools and pipelines accept as input a PWM. Still, this seems to be the direction in which the community is going.

Attempts to improve *in vivo* binding prediction include new epigenetic data. The most useful kind of data is nucleosome occupancy, which demarcates in the genome accessible regions where the TF can bind. Studies have used the new DNAse I hypersensitivity data together with a sequence-specific binding model to improve *in vivo* binding prediction [69, 79]. In addition, a recent study has been looking at cooperative TF binding to improve the predictions based on regions flanking the binding site [80]. Other kinds of information may be used in the future to improve *in viv*o binding prediction.

Last, sequence design problem have been and still are relevant in various applications. Microarrays are being used extensively and it is becoming easier to implement and modify genomic DNA sequences *in vivo*. Application in these platforms raise interesting sequence design problems, which can be solved using combinatorial methods and models, such as de Bruijn graphs and linear shift feedback registers. On the downside, some of these problems may become less relevant in the near future, as microarray applications are taken over by deep sequencing (e.g. HT-SELEX replacing PBM) due to the decreasing cost and higher resolution of the latter.

To conclude, there are many more problems to be solved in order to advance our understanding of protein-DNA binding. The technological advancements require new computational tools, and pose new challenges in modeling and predicting protein-DNA binding both *in vivo* and *in vitro*.

Bibliography

1. Orenstein Y, Linhart C, Shamir R: **Assessment of algorithms for inferring positional weight matrix motifs of transcription factor binding sites using protein binding microarray data.** *PLoS ONE* 2012, **7:**e46145.
2. Orenstein Y, Mick E, Shamir R: **RAP: accurate and fast motif finding based on protein-binding microarray data.** *J Comput Biol* 2013, **20:**375-382.
3. Orenstein Y, Shamir R: **Design of shortest double-stranded DNA sequences covering all k-mers with applications to protein-binding microarrays and synthetic enhancers.** *Bioinformatics* 2013, **29:**i71-79.
4. Orenstein Y, Shamir R: **A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data.** *Nucleic Acids Res* 2014, **42:**e63.
5. Lockhart DJ, Winzeler EA: **Genomics, gene expression and DNA arrays.** *Nature* 2000, **405:**827-836.
6. Soon WW, Hariharan M, Snyder MP: **High-throughput sequencing for biology and medicine.** *Mol Syst Biol* 2013, **9:**640.
7. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, et al: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **290:**2306-2309.
8. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al: **Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.** *Nat Methods* 2007, **4:**651-657.
9. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, 3rd, Bulyk ML: **Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities.** *Nat Biotechnol* 2006, **24:**1429-1435.
10. Philippakis AA, Qureshi AM, Berger MF, Bulyk ML: **Design of compact, universal DNA microarrays for protein binding microarray experiments.** *J Comput Biol* 2008, **15:**655-665.
11. Robasky K, Bulyk ML: **UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions.** *Nucleic Acids Res* 2011, **39:**D124-128.
12. Zhao Y, Granas D, Stormo GD: **Inferring binding energies from selected binding sites.** *PLoS Comput Biol* 2009, **5:**e1000590.
13. Slattery M, Riley T, Liu P, Abe N, Gomez-Alcala P, Dror I, Zhou T, Rohs R, Honig B, Bussemaker HJ, Mann RS: **Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins.** *Cell* 2011, **147:**1270-1282.
14. Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, Taipale M, Vaquerizas JM, Yan J, Sillanpaa MJ, et al: **Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities.** *Genome Res* 2010, **20:**861-873.

15. Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16:**16-23.

16. Zhou Q: **On weight matrix and free energy models for sequence motif detection.** *J Comput Biol* 2010, **17:**1621-1638.

17. D'Haeseleer P: **What are DNA sequence motifs?** *Nat Biotechnol* 2006, **24:**423-425.

18. Benos PV, Bulyk ML, Stormo GD: **Additivity in protein-DNA interactions: how good an approximation is it?** *Nucleic Acids Res* 2002, **30:**4442-4451.

19. Zhao Y, Ruan S, Pandey M, Stormo GD: **Improved models for transcription factor binding site identification using nonindependent interactions.** *Genetics* 2012, **191:**781-790.

20. Wingender E, Hogan J, Schacherer F, Potapov AP, Kel-Margoulis O: **Integrating pathway data for systems pathology.** *In Silico Biol* 2007, **7:**S17-25.

21. Spivak AT, Stormo GD: **ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species.** *Nucleic Acids Res* 2012, **40:**D162-168.

22. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles.** *Nucleic Acids Res* 2004, **32:**D91-94.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

37. Linhart C, Halperin Y, Shamir R: **Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets.** *Genome Res* 2008, **18:**1180-1189.

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

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

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

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

42. Eden E, Lipson D, Yogev S, Yakhini Z: **Discovering motifs in ranked lists of DNA sequences.** *PLoS Comput Biol* 2007, **3:**e39.

43. Tanay A: **Extensive low-affinity transcriptional interactions in the yeast genome.** *Genome Res* 2006, **16:**962-972.

44. Foat BC, Morozov AV, Bussemaker HJ: **Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE.** *Bioinformatics* 2006, **22:**e141-149.

45. Ma W, Noble WS, Bailey TL: **Motif-based analysis of large nucleotide data sets using MEME-ChIP.** *Nat Protoc* 2014, **9:**1428-1450.

46. Liu XS, Brutlag DL, Liu JS: **An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments.** *Nat Biotechnol* 2002, **20:**835-839.

47. Kulakovskiy IV, Boeva VA, Favorov AV, Makeev VJ: **Deep and wide digging for binding motifs in ChIP-Seq data.** *Bioinformatics* 2010, **26:**2622-2623.

48. Sun W, Hu X, Lim MH, Ng CK, Choo SH, Castro DS, Drechsel D, Guillemot F, Kolatkar PR, Jauch R, Prabhakar S: **TherMos: Estimating protein-DNA binding energies from in vivo binding profiles.** *Nucleic Acids Res* 2013, **41:**5555-5568.

49. Das MK, Dai HK: **A survey of DNA motif finding algorithms.** *BMC Bioinformatics* 2007, **8 Suppl 7:**S21.

50. Sandve GK, Drablos F: **A survey of motif discovery methods in an integrated framework.** *Biol Direct* 2006, **1:**11.

51. Zambelli F, Pesole G, Pavesi G: **Motif discovery and transcription factor binding sites before and after the next-generation sequencing era.** *Brief Bioinform* 2013, **14:**225-237.

52. Annala M, Laurila K, Lahdesmaki H, Nykter M: **A linear model for transcription factor binding affinity prediction in protein binding microarrays.** *PLoS One* 2011, **6:**e20059.

53. Moses LE, Emerson JD, Hosseini H: **Analyzing data from ordered categories.** *N Engl J Med* 1984, **311:**442-448.

54. Chen X, Hughes TR, Morris Q: **RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors.** *Bioinformatics* 2007, **23:**i72-79.

55. Zhao Y, Stormo GD: **Quantitative analysis demonstrates most transcription factors require only simple models of specificity.** *Nat Biotechnol* 2011, **29:**480-483.

56. Weirauch MT, Cote A, Norel R, Annala M, Zhao Y, Riley TR, Saez-Rodriguez J, Cokelaer T, Vedenko A, Talukder S, et al: **Evaluation of methods for modeling transcription factor sequence specificity.** *Nat Biotechnol* 2013, **31:**126-134.

57. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al: **DNA-binding specificities of human transcription factors.** *Cell* 2013, **152:**327-339.

58. Lipson D, Webb P, Yakhini Z: **Designing specific oligonucleotide probes for the entire S. cerevisiae transcriptome.** *Algorithms in Bioinformatics, Proceedings* 2002, **2452:**491-505.

59. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci U S A* 2004, **101:**6062-6067.

60. Barrett MT, Scheffer A, Ben-Dor A, Sampas N, Lipson D, Kincaid R, Tsang P, Curry B, Baird K, Meltzer PS, et al: **Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA.** *Proc Natl Acad Sci U S A* 2004, **101:**17765-17770.

61. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: **High density synthetic oligonucleotide arrays.** *Nat Genet* 1999, **21:**20-24.

62. Ben-Dor A, Karp R, Schwikowski B, Yakhini Z: **Universal DNA tag systems: a combinatorial design scheme.** *J Comput Biol* 2000, **7:**503-519.

63. Fordyce PM, Gerber D, Tran D, Zheng J, Li H, DeRisi JL, Quake SR: **De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis.** *Nat Biotechnol* 2010, **28:**970-975.

64. Klein A: *Stream ciphers.* 1st edn. New York: Springer; 2013.

65. Lewis TG, Payne WH: **Generalized Feedback Shift Register Pseudorandom Number Algorithm.** *Journal of the Acm* 1973, **20:**456-468.

66. Smith RP, Riesenfeld SJ, Holloway AK, Li Q, Murphy KK, Feliciano NM, Orecchia L, Oksenberg N, Pollard KS, Ahituv N: **A compact, in vivo screen of all 6-mers reveals drivers of tissue-specific expression and guides synthetic regulatory element design.** *Genome Biol* 2013, **14:**R72.
67. Mintseris J, Eisen MB: **Design of a combinatorial DNA microarray for protein-DNA interaction studies.** *BMC Bioinformatics* 2006, **7:**429.
68. M. D'Addario NKaSR: **Designing q-Unique DNA Sequences with Integer Linear Programs and Euler Tours in De Bruijn Graphs.** *German Conference on Bioinformatics* 2012**:**82-92.
69. Zhong S, He X, Bar-Joseph Z: **Predicting tissue specific transcription factor binding sites.** *BMC Genomics* 2013, **14:**796.
70. Gordan R, Shen N, Dror I, Zhou T, Horton J, Rohs R, Bulyk ML: **Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape.** *Cell Rep* 2013, **3:**1093-1104.
71. Jiang B, Liu JS, Bulyk ML: **Bayesian hierarchical model of protein-binding microarray k-mer data reduces noise and identifies transcription factor subclasses and preferred k-mers.** *Bioinformatics* 2013, **29:**1390-1398.
72. Mordelet F, Horton J, Hartemink AJ, Engelhardt BE, Gordan R: **Stability selection for regression-based models of transcription factor-DNA binding specificity.** *Bioinformatics* 2013, **29:**i117-125.
73. Rhee HS, Pugh BF: **Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution.** *Cell* 2011, **147:**1408-1419.
74. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, Rando OJ: **Genome-scale identification of nucleosome positions in S. cerevisiae.** *Science* 2005, **309:**626-630.
75. Eads CA, Danenberg KD, Kawakami K, Saltz LB, Blake C, Shibata D, Danenberg PV, Laird PW: **MethyLight: a high-throughput assay to measure DNA methylation.** *Nucleic Acids Res* 2000, **28:**E32.
76. Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E: **Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters.** *Nat Biotechnol* 2012, **30:**521-530.
77. Siddharthan R: **Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix.** *PLoS One* 2010, **5:**e9722.
78. Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B: **The role of DNA shape in protein-DNA recognition.** *Nature* 2009, **461:**1248-1253.
79. Gusmao EG, Dieterich C, Zenke M, Costa IG: **Detection of Active Transcription Factor Binding Sites with the Combination of DNase Hypersensitivity and Histone Modifications.** *Bioinformatics* 2014.
80. Munteanu A, Ohler U, Gordan R: **COUGER--co-factors associated with uniquely-bound genomic regions.** *Nucleic Acids Res* 2014, **42:**W461-467.