BIOINFORMATICS DISCOVERY NOTE

Vol. 1 no. 1 2002
Pages 1–11

Identifying Transcription Factor Binding Sites through Markov Chain Optimization

Kyle Ellrott[1, 4], Chuhu Yang[2], Frances M. Sladek[3] and Tao Jiang[1]

[1]Department of Computer Science, University of California, Riverside, CA, 92521, USA, [2]Genetics/Bioinformatics Program, University of California, Riverside, CA, 92521, USA, [3]Department of Cell Biology and Neuroscience, University of California, Riverside, CA, 92521, USA and [4] Protein Informatics Group, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37831-6480, USA

ABSTRACT

Even though every cell in an organism contains the same genetic material, each cell does not express the same cohort of genes. Therefore, one of the major problems facing genomic research today is to determine not only which genes are differentially expressed and under what conditions, but also how the expression of those genes is regulated. The first step in determining differential gene expression is the binding of sequence-specific DNA binding proteins (*i.e. transcription factors*) to regulatory regions of the genes (*i.e. promoters* and *enhancers*). An important aspect to understanding how a given transcription factor functions is to know the entire gamut of binding sites and subsequently potential target genes that the factor may bind/regulate. In this study, we have developed a computer algorithm to scan genomic databases for transcription factor binding sites, based on a novel Markov chain optimization method, and used it to scan the human genome for sites that bind to *hepatocyte nuclear factor 4* α (HNF4α). A list of 71 known HNF4α binding sites from the literature were used to train our Markov chain model. By looking at the window of 600 nucleotides around the transcription start site of each confirmed gene on the human genome, we identified 849 sites with varying binding potential and experimentally tested 109 of those sites for binding to HNF4α. Our results show that the program was very successful in identifying 77 new HNF4α binding sites with varying binding affinities (*i.e.* a 71% success rate). Therefore, this computational method for searching genomic databases for potential transcription factor binding sites is a powerful tool for investigating mechanisms of differential gene regulation.

Contact: jiang@cs.ucr.edu

Analysis of the human genome as well as the genomes of a variety of other organisms recently showed that an unexpectedly large portion of the genetic content of cells is devoted to the regulation of gene expression (Lander *et al.*, 2001; Venter *et al.*, 2001). This is consistent with the long established tenet of cell and developmental biology that even though every cell in the body contains the same complement of genes, only a subset of those genes are expressed in a given tissue at a given time during development and under given physiological and pathological conditions. One of the first steps in determining whether or not a gene is expressed is the binding of specialized transcriptional activators or repressors (*i.e. transcription factors*) to specific DNA sequences (*i.e. promoters* and *enhancers*), near the start site of transcription, typically within a couple hundred nucleotides. Subsequently, these factors recruit general transcription factors and RNA polymerase that transcribes the DNA into RNA, which is then translated into protein. Therefore, cell- and tissue-specific gene expression occurs due to the presence, or absence, of different cohorts of transcription factors. Whereas these transcription factors exhibit sequence-specific DNA binding properties there is almost always variability in the sequence that they recognize, *i.e.* they bind more than one sequence although a consensus sequence can often be established. This variability is thought to be critical to the fine tuning of the regulation of gene expression but it also makes it very difficult to definitively identify all potential binding sites without the aid of computational techniques.

A prime example of promiscuity in DNA binding is *hepatocyte nuclear factor 4* α (HNF4α), a highly conserved factor originally found in liver but also present in kidney, intestine, pancreas, and stomach (Sladek and Seidel, 2001). HNF4α is a member of the nuclear receptor superfamily of ligand-dependent transcription factors that includes steroid and thyroid hormone and vitamins A and D receptors as well as a large number of receptors for which ligands have not yet been identified (*i.e. orphan receptors*). Whereas HNF4α is considered an orphan receptor in terms of ligand binding it is not an orphan in terms of target genes and binding sites. Over 55 target genes for HNF4α have been identified experimentally and can be grouped according to the function of the genes that they encode. The categories include genes involved in glucose, lipid, steroid, xenobiotic and amino acid metabolism and transport as well as genes involved in blood maintenance, liver differentiation and hepatitis B viral infections. Through these target genes HNF4α is associated with several different human diseases. For example, the HNF4α gene is mutated in an inherited form of diabetes, maturity onset diabetes of the young 1 (MODY1) (Sladek and Seidel, 2001), and mutations in HNF4α binding sites are known to be the cause not only of another form of MODY (MODY3) but also of certain types of hemophilia (Sladek and Seidel, 2001). Therefore, knowing the entire complement of genes that HNF4α regulates and the DNA sequences in those genes that HNF4α binds is critical to determining the role of HNF4α in human diseases and potential susceptibilities to those diseases.

The experimental techniques to determine whether a given transcription factor such as HNF4α binds any given DNA sequence in vitro are well established. Complementary oligonucleotides (oligo's) containing the binding site in question are synthesized, annealed, radiolabeled, and analyzed for binding in a gel shift assay in which oligo's bound by the transcription factor are separated from unbound oligo's by differential migration through a native polyacrylamide gel under an electrical field. However, using this method to determine whether a given factor binds hundreds or thousands of potential sites is not only very time consuming and labor intensive but also costly. Therefore, the use of computational methods to identify potential transcription factor binding sites is highly desired, particularly now that the entire genomes of a variety of organisms have been sequenced. [†]

Many algorithmic methods have been proposed in the literature for the identification of transcription factor binding sites, *e.g.* (Brazma *et al.*, 1998; Sinha and Tompa, 2000; Stormo, 2000; Tan *et al.*, 2001; Thakurta and Stormo, 2001; Tompa, 1999; Vanet, Marsan, and Sagot, 1999; Zhu and Zhang, 1999). While some of the methods attempt to find sites in a genome that have interesting characteristics and may bind to *any* transcription factor, others try to identify only binding sites for specific transcription factors. In this study, we are concerned with the latter paradigm.

Most transcription factors have a number of known (*i.e.* experimentally verified) binding sites that can be found by searching the literature. For example, at least 71 distinct binding sites for HNF4α have been previously identified, each consisting of 13 nucleotides as shown in Table 1. From such a list of known binding sites, one can build a model of the binding sequences that characterizes the sequences in some (often probabilistic) way. Then combinatorial matching and/or statistical tesing techniques can be applied to search a genome (or some selected regions of the genome, such as regions that are upstream of verified or predicted genes and are near transcription start sites) for putative binding sites that "match" the model with a high "confidence".

The simplest way of modelling a binding sequence is to use *consensus* sequences or *regular expressions* (Quandt *et al.*, 1995). For example, the well-known Kozak's sequence that marks the start site of translation can be described as a regular expression, $GCC[A|G]CCATGG$, where A or G are acceptable for the fourth position. These methods, of course, represent an extremely simplistic view of the binding sequence and only work well for short, highly conserved sequences. The most popular representations of transcription factor binding sequences are perhaps *position specific score matrices* (PSSM's) (Stormo, 2000). A PSSM for a binding sequence essentially describes the frequency of each nucleotide at each particular position of the binding sequence in the form of *scores* (which are usully based on log ratios of frequencies) (Stormo, 2000). Once a PSSM has been established from the known binding sequences (*i.e.* the training set), one can determine how well a potential binding sequence matches the PSSM by scoring the nucleotide at each position of the sequence against the PSSM (at corresponding positions) and summing up the scores. PSSM is the primary method for expressing transcription factor binding sequences in the Transfac database (Wingender *et al.*, 2000). Since a PSSM treats each position of the binding sequence independently from the other positions, it does not capture any potential dependence that may exist between positions and thus may not work well when the positions are strongly dependent on each other. One possible way to improve the performance of PSSM is to use *maximal dependence decomposition* (MDD) that attempts to capture dependence between positions (Burge and Karlin, 1997). Instead of creating a single PSSM based on the entire training set, MDD creates a tree of PSSM's, each estimated from a subset of training sequences. The correct PSSM used to score a sequence is then determined by "tokens" that occur in the positions that are determined to be dependent on other positions. This method allows for the use of information about dependent and independent positions, but requires a training set that is large enough to be partitioned into smaller, yet still adequate, training subsets.

For most transcription factors, the number of known binding sites is relatively small. *E.g.*, although a relatively large number of HNF4α target genes have been characterized, only 71 distinct binding sites were known at

[†] Although a promising experimental technique, called genome-wide location analysis (Iyer *et al.*, 2001; Lieb *et al.*, 2001; Ren *et al.*, 2000), based on DNA microarray technology has been developed recently, it is presently still too complex and expensive to be used on large genomes such as the human genome. Moreover, the technique is designed to find target genes of transcription factors rather than their specific binding sites.

the beginning of our project. However, a simple statisical dependence analysis (*i.e.* χ^2 test) on the 71 binding sequences (of 13 nucleotides) for HNF4α revealed significant dependence between several pairs of positions, *e.g.* positions 4 and 8, positions 4 and 11, *etc*.

A natural extension of the PSSM method is to use *Markov chains* or *hidden Markov models* (HMM's). Here, a binding sequence is represented as a Markov chain (or HMM) that gives the probability of each nucleotide occurring at a particular position depending on the nucleotides at preceding positions. Markov chains and HMM's have been used extensively in biomolecular sequence analysis and, in particular, transcription factor binding sequence identification (Durbin *et al.*, 1998). Although these models allow dependences among positions to be encoded in the state transition probabilities, not all dependence are treated equally. Intuitively, dependence between two positions is *directly* represented in the Markov chain (or HMM) if the positions are adjacent in the Markov chain (or within close proximity in the case of high order Markov chains/HMM's); otherwise it is only *indirectly* represented. Correlation among non-adjacent positions could especially be important for transcription factor binding sites since the binding between a DNA molecule and a protein molecule is essentially a 3-dimensional geometrical matching process that may involve cooperation between nucleotides (or amino acid residues) at non-adjacent positions of the primary DNA (or protein) sequence. For example, HNF4α is a dimer consisting of two cooperative "arms" that bind to different regions of the target sequence. However, the existing work on using Markov chains and HMM's to identify transcription factor binding sites typically arranges the states in the same order as the positions in the binding sequence, and hence may not capture the most significant inter-position dependence.

In this paper, we propose an enhancement to the above Markov chain based algorithms for finding transcription factor binding sites. Given a set of training sequences (*i.e.* known binding sequences for some particular transcription factor), we first estimate the pairwise dependence between positions in the target binding sequences through a simple statistical (*e.g.* χ^2) analysis. The Markov chain is then ordered so that most pairs (or groups, in the case of high-order Markov chain) of significantly dependent positions are adjacent (or within close proximity). The Markov model is then trained using the training sequences, and the completed model is used to scan genomic sequences of interest to identify potential binding sites. We note in passing that correlations and dependences between positions in regulatory sequences have also been previously studied (Agarwal and Bafna, 1998).

To demonstrate the utility of the above method, we have followed the method to create an (optimized) Markov chain model for the HNF4α binding sequences. As a first application, the model was used to scan an area from $-500bp$ to $+100bp$, relative to the transcription start site, for each of the approximately $9,500$ verified genes obtained from the UCSC Goldenpath human genome annotation (see `http://genome.ucsc.edu/`). The scan yielded a total of 849 sites with varying binding potential. We then selected a subset of 109 sites, and tested their binding affinities *in vitro* using a gel shift assay. This resulted in the identification of 77 new sites in the human genome (including 69 new sequences) that bind HNF4α with a certain affinity (*i.e.* a 71% success rate). This finding significantly impacts the study of HNF4α because only 71 binding sites were known to exist in all genomes at the start of the project. We have also compared the optimized Markov model with the "unoptimized" one where the positions are sequentially ordered according to that in the binding sequence, in terms of (i) information content (or relative entropy) of the model and (ii) accuracy in predicting binding sites, and found that the optimized model is superior in both categories. Encouraged by the success of the test on the HNF4α data, we think that this improved Markov chain approach will be very useful in identifying binding sites for many transcription factors.

The rest of the paper is organized as follows. Section 2 describes the improved Markov chain algorithm for identifying transcription factor binding sites, and the algorithm for ordering the Markov chain to capture the most significant inter-position dependence. In Section 3, we present the experimental results on the HNF4α data and some comparisons of different Markov models. Section 4 concludes the paper with some possible further improvements.

AN IMPROVED MARKOV CHAIN ALGORITHM

In this section, we outline our algorithm for identifying transcription factor binding sites (TFBS's) through Markov chain optimization. The key is an algorithm for ordering the Markov chain to capture the most significant dependence among positions in the target binding sequence. For convenience, we will illustrate the steps in the algorithms mostly in terms of the HNF4α example, although the approach should work for any transcription factor with an adequate set of known binding sites.

A (nonstationary) Markov chain of length n is a probabilistic model that describes the probability distribution of sequences of n states $s_1, s_2, \ldots, s_n$ by means of *transition probabilities*, where the transition probability $P(s_i = q|s_{i-1} = p)$ defines the probability of state $s_i = q$ given state $s_{i-1} = p$. This definitions can be easily extended to *high-order* Markov chains to allow the state at a particular position to depend on states at several preceding positions. For example, a 3rd-order Markov chain, which was the

1. Input: a set of known binding sequences for the target transcription factor and a genomic sequence.
2. Extract regions of the genome that are likely to contain binding sites for the target transcription factor.
3. Perform a dependence analysis (*e.g.* χ^2 test) on the training sequences to find an ordering of the positions in the Markov chain to be created so that most of the significantly dependent positions are adjacent (or within close proximity).
4. Train the (high-order) Markov model using the known binding sequences.
5. Determine a threshold based on the mean and standard deviation of the scores of the training sequences under the model.
6. Scan the regions extracted in Step 2 with the Markov model to create a list of ranked candidate binding sites.

Fig. 1. An improved Markov chain algorithm for finding TFBS's.

model used in the HNF4α project, has transition probabilities of the form $P(s_i = q|s_{i-1} = p, s_{i-2} = v, s_{i-3} = u)$. Third-order Markov models are especially useful in scanning genomes for motifs because they are capable of capturing 4-letter words that may be of (*e.g.* functional) significance (Sinha and Tompa, 2000). The framework of our Markov chain algorithm for finding TFBS's is similar to existing Markov chain algorithms, except that we explicitly order the positions in the Markov chain to maximize the inter-position dependence captured in the model. The algorithm is outlined in Figure 1.

The set of 71 training sequences for HNF4α are shown in Table 1 (Antes *et al.*, 2000; Hauch *et al.*, 1994; Lahuna *et al.*, 2000; Nicolas-Frances *et al.*, 2000; Pinaire *et al.*, 1999; Ozeki *et al.*, 2001; Sladek and Seidel, 2001; Swenson *et al.*, 1999; Yanai *et al.*, 1999). Each sequence consists of two similar segments (*direct repeats*) of 6 nucleotides each separated by a "spacer" (of one nucleotide). Actually, there are two more binding sequences known for HNF4α that contain two "spacers"; but these sequences were not included in the training of our Markov model. Since most known TFBS's occur near the transcription start sites of genes, we focused our attention on regions surrounding known transcription start sites in the given genome. This not only reduced the search space and thus the running time, but also reduced the number of false positives. A similar idea was also considered (Tan *et al.*, 2001). In the first search for HNF4α binding sites, we used all regions containing $-500bp$ through $+100bp$ of the transcription start site of each of the approximately $9,500$ verified genes in the UCSC Goldenpath human genome annotation, for a total of about 6 million bps.

(Pearson) χ^2 test is a standard method for studying independence between two distributions (Hays and Winkler, 1971). In our case, we are not concerned with independence as much as we are with dependence. We will use a liberal interpretation of χ^2 test to determine which distributions are "less independent" than other distributions, thus sorting out pairs of distributions that are not independent. [‡] To define the χ^2 values for a given set of training sequences, let $f_i(x)$ denote the (observed) frequency of nucleotide x at position i, and $O_{i_1,i_2}(x_1, x_2)$ the (observed) frequency of nucleotide x_1 occurring at position i_1 and nucleotide x_2 occurring at position i_2. We can calculate the expected frequency for x_1 to occur at position i_1 and x_2 to occur at position i_2 as $E_{i_1,i_2}(x_1, x_2) = f_{i_1}(x_1)f_{i_2}(x_2)/N$ (assuming the positions are independent), where N is the sample size (*i.e.* the total number of training sequences). Let $X = \{A, C, G, T\}$ denote the set of nucleotides. The χ^2 value for positions i_1 and i_2 is defined as

$$\chi^2(i_1, i_2) = \sum_{x_1 \in X} \sum_{x_2 \in X} \frac{(O_{i_1,i_2}(x_1, x_2) - E_{i_1,i_2}(x_1, x_2))^2}{E_{i_1,i_2}(x_1, x_2)}.$$

For example, the χ^2 values for the 13 positions in the known HNF4α binding sequences are shown in Table 2. Noticing that the χ^2 test has $(4 - 1) \cdot (4 - 1) = 9$ degrees of freedom, we can compute the p values of the χ^2 values (Hays and Winkler, 1971), as shown in Table 3. Here, each p value represents the probability that a pair of positions are independent. In the tables, the rows and columns are numbered from 1 to 13 and smaller p values (*i.e.* larger χ^2 values) indicate less probability of independence (or more probability of dependence). Usually, a p value less than or equal to 0.05 is accepted

[‡] Strictly speaking, non-independence does not always imply dependence in statistics theory. Although the χ^2 test worked well for the HNF4α project, one may also consider other statistical tests for dependence.

HNF4α Binding Site	HNF4α Target Gene	HNF4α Binding Site	HNF4α Target Gene
AGTTCAaGGATCA	apolipoprotein AI	GGGGTCaAGGGTT	apolipoprotein AI
AGGGTAaAGGTTG	apolipoprotein AII	GTCACAaAAGTCC	apolipoprotein AIV
GGTCCAaAGGGCG	apolipoprotein B	AGGCCAaAGTCCT	apolipoprotein CII
TGGGCAaAGGTCA	apolipoprotein CIII	GGTCCAgAGGGCA	apolipoprotein CIII
AGTCCAgAGGTCA	apolipoprotein CIII	GAGTCAaAGGTCA	cellular retinol binding protein II
AGTTCAaAGTTCA	intestinal fatty acid binding protein	AGGTCAaAGATTG	transferrin
GGCAAGgTTCATA	transthyretin	GGGGCTaAGTCCA	α-1-anti-trypsin
GGGTTAaAGGTTG	sex hormone-binding globulin	GGGTCAaGGGTCA	sex hormone-binding globulin
CGGGTAaAGGTGA	medium-chain acyl CoA	AGGACAaAGGTCA	acyl-CoA oxidase
GGGCCAaAGGTCT	3-hydroxy-3-methylglutaryl-CoA	AGACCAaAGTCCG	cytochrome 2A4
GGACCAaAGTCCA	cytochrome 2C1	GGTCCAaAGTCCA	cytochrome 2C2
AGACCAaAGTGCA	cytochrome 2C3	TCCTGAaACTGGG	cytochrome 2C9
AGGGCAaAGGCAA	cytochrome 2D6	GTACCAaAGTCCA	cytochrome 3A1
TGGACTtAGTTCA	cytochrome 7	AGGGCAaTGACGT	cytochrome 7
CGGCCAaAGGTCA	phospho-enol-pyruvate carboxykinase	GGGCCAgAGTCCA	liver-type pyruvate kinase
GGAGTAaAGTTCA	aldolase B	AGATCAaAGAGCA	tyrosine amino transferase
GGTTTAaAGTTCA	ornithine transcarbamylase	AGTTCAgAGGTTA	ornithine transcarbamylase
GGATCAaAGGTCC	ornithine transcarbamylase	GGCTTAaAGTTCA	ornithine transcarbamylase
GGGTCAaAGGCAC	aldenhyde dehrogenase 2	AGGGCAaAGGTCA	Factor VII
CGGGCAaAGTTCT	Factor VII	GGGGCAtAAGTCT	Factor VIII
CTAGCAaAGGTTA	Factor IX	AGTGGTaAGGTCG	Factor IX
GTACCAaAGTACA	Factor IX	GGAGCAaAGTCCA	Factor X
AGGTCGaGAGGTC	erythropoietin	AGTGTAgAGCCCA	antithrombin III
AGGTCAaAGGCTG	antithrombin III	AGTCCAaAGTTCA	hepatocyte nuclear factor 1 α
GGTCCAaAGTTCA	hepatocyte nuclear factor 1 α	GGGTCAcAGTGCA	macrophage stimulating protein
AGGTCTcAGGTCA	macrophage stimulating protein	CTGCCAaGGGCCA	α-1-microglobulin & bikunin
AGTCAAaAGTCCA	α-1-microglobulin & bikunin	GTCTAAgAGTCCA	α-1-microglobulin & bikunin
GGGGTAaAGGTTC	hepatitis B virus enhancer I	AGTCCAaGAGTCC	hepatitis B virus enhancer II
AGGTTAaAGGTCT	hepatitis B virus nucleocapsid	AGTCCAaAGGTCC	woodchuck hepatitis virus enhancer II
GGGCCAaGGGTCA	human immunodef. virus long terminal repeat	AGGTCAgGGTCCA	hepatocyte growth factor-like protein
GGGGCAaAGTCAA	prolactin receptor	GGGCTGaAGTCCA	hepatocyte nuclear factor 1a
CGGGCAaAGGCCA	hepatocyte nuclear factor 6	AGAACAaAGAGCA	apolipoprotein B
GGTTCAaAGGTCT	3-keto acyl-CoA thiolase B	ACGGGAgACGGGA	angiotensinogen
CTTGGAaCCGGGG	angiotensinogen, weaker site	AGGTCAgGGTCCC	aldehyde dehydrogenase 2
TGTCCAaAGTCCA	dihydrodiol dehydrogenase 4	TGATCAgACAAAG	biliary glycoprotein
AAACCAaAGTTCA	guanylyl cyclase C		

Table 1. The 71 known HNF4α binding sequences.

as a convincing evidence of dependence. Hence, Table 3 illustrates many pairs of dependent positions.

Given a matrix of χ^2 values and their p values for all pairs of positions, we wish to construct an ordering of the positions that will maximize the overall dependence among all "neighboring" positions. Here, the neighborhood size depends on the order of the Markov model employed. For example, for a basic (1st-order) Markov model, a neighborhood contains two positions, but for a 3rd-order Markov model, a neighborhood should contain four positions. For a transcription factor binding sequence of length n, there are $n!$ possible orderings of positions in the Markov chain, which could be too many to search exhausitively. So we propose the following simple greedy algorithm instead. Suppose that the order of the Markov model considered is k. For a pair of positions i_1, i_2, where $i_1 \neq i_2$, define the *dependence score*, denoted $g(i_1, i_2)$, as $-\log p(i_1, i_2)$ (assume that $p(i_1, i_2) > 0$). The algorithm starts by picking the two positions with the greatest probable dependence, or in other words, the highest dependence score. Then, we pick a position such that its total dependence score with the two chosen positions is maximized. This is continued until $k + 1$ positions are chosen. Then we pick a position such that its total dependence score with a subset of k chosen positions is maximized. This defines a partial order with two end positions sandwiching k unordered positions. We next add a position at either end of the partial order to maximize its

213.00	19.07	8.06	9.97	8.17	5.08	9.09	12.13	16.00	12.80	5.62	11.92	6.38
19.07	213.00	18.78	3.91	40.01	1.65	2.96	10.28	37.82	1.71	15.94	29.39	4.69
8.06	18.78	213.00	21.89	25.09	7.44	8.00	12.62	19.56	20.28	12.55	7.97	6.43
9.97	3.91	21.89	213.00	15.84	10.85	12.93	14.75	18.44	17.98	8.29	18.38	7.38
8.17	40.01	25.09	15.84	213.00	16.78	8.13	30.45	62.49	17.97	23.14	38.17	15.93
5.08	1.65	7.44	10.85	16.78	213.00	16.84	13.46	28.73	12.37	15.20	15.02	11.26
9.09	2.96	8.00	12.93	8.13	16.84	213.00	3.36	17.61	11.85	13.38	2.04	6.93
12.13	10.28	12.62	14.75	30.45	13.46	3.36	213.00	58.43	24.53	17.60	23.61	16.76
16.00	37.82	19.56	18.44	62.49	28.73	17.61	58.43	213.00	40.91	43.46	41.03	32.39
12.80	1.71	20.28	17.98	17.97	12.37	11.85	24.53	40.91	213.00	29.63	13.43	12.23
5.62	15.94	12.55	8.29	23.14	15.20	13.38	17.60	43.46	29.63	213.00	20.86	4.73
11.92	29.39	7.97	18.38	38.17	15.02	2.04	23.61	41.03	13.43	20.86	213.00	15.96
6.38	4.69	6.43	7.38	15.93	11.26	6.93	16.76	32.39	12.23	4.73	15.96	213.00

Table 2. The χ^2 test on the positions in the HNF4α binding sequences.

0.0000	0.0245	0.5274	0.3528	0.5167	0.8265	0.4283	0.2057	0.0668	0.1716	0.7771	0.2178	0.7011
0.0245	0.0000	0.0270	0.9171	0.0000	0.9958	0.9656	0.3281	0.0000	0.9951	0.0679	0.0005	0.8597
0.5274	0.0270	0.0000	0.0092	0.0028	0.5905	0.5339	0.1801	0.0207	0.0162	0.1837	0.5364	0.6956
0.3528	0.9171	0.0092	0.0000	0.0700	0.2860	0.1655	0.0980	0.0303	0.0353	0.5049	0.0309	0.5973
0.5167	0.0000	0.0028	0.0700	0.0000	0.0522	0.5209	0.0003	0.0000	0.0354	0.0058	0.0000	0.0683
0.8265	0.9958	0.5905	0.2860	0.0522	0.0000	0.0512	0.1425	0.0007	0.1931	0.0854	0.0902	0.2578
0.4283	0.9656	0.5339	0.1655	0.5209	0.0512	0.0000	0.9480	0.0398	0.2218	0.1461	0.9908	0.6438
0.2057	0.3281	0.1801	0.0980	0.0003	0.1425	0.9480	0.0000	0.0000	0.0035	0.0400	0.0049	0.0524
0.0668	0.0000	0.0207	0.0303	0.0000	0.0007	0.0398	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
0.1716	0.9951	0.0162	0.0353	0.0354	0.1931	0.2218	0.0035	0.0000	0.0000	0.0005	0.1437	0.2002
0.7771	0.0679	0.1837	0.5049	0.0058	0.0854	0.1461	0.0400	0.0000	0.0005	0.0000	0.0132	0.8564
0.2178	0.0005	0.5364	0.0309	0.0000	0.0902	0.9908	0.0049	0.0000	0.1437	0.0132	0.0000	0.0677
0.7011	0.8597	0.6956	0.5973	0.0683	0.2578	0.6438	0.0524	0.0001	0.2002	0.8564	0.0677	0.0000

Table 3. The *p* values corresponding to the χ^2 values in Table 2. Note that, some off-diagonal values are shown as 0 due to limited precision in the presentation.

total dependence score with the neighboring end position and $k - 1$ of the unordered positions. This results in a partial order with two ordered positions at each end and $k - 1$ unordered positions in the middle. The process is continued until $2k + 1$ positions are chosen and a total (linear) order is formed. We then repeatedly add new positions at either end in a straghtforward way until all positions are included. A pseudo-code of the algorithm is given in Figure 2.

In the algorithm, P denotes the set of all positions to be ordered, C denotes the set of positions that have been selected, and R denotes the remaining positions. Note that, when $|C| < 2k+1$, the positions in C form a partial order consisting of s linearly ordered positions, followed by a subset of t (unordered) positions, which is then followed by another set of s linearly ordered positions. In particular, when $|C| \leq k + 1$, the positions in C simply form an (unordered) subset (*i.e.* $t > 1$ and $s = 0$). When $k + 1 < |C| < 2k + 1$, the sizes of a linear order and the middle subset always add up to $k + 1$ (*i.e.* $t > 1$ and $s + t = k + 1$), because of the way the algorithm works. When $|C| \geq 2k+1$, the positions in C form a linear order (*i.e.* $t = 0$).

Figure 3 (the first row) illustrates the ordering of positions based on the p values in Table 3 for HNF4α and a 3rd-order Markov model. One (generic) way of measuring the effectiveness of a Markov model is to consider its *relative entropy* (also called *Kullback-Leibler distance*) with respect to the background distribution. Here, the relative entropy of two distributions M (the Markov model) and B (the background, collected from the regions extracted in Step 2 in Figure 1) is defined as $\sum_x M(x) \ln \frac{M(x)}{B(x)}$, where x is an oligo consisting of k nucleotides. As illustrated in Table 4 for the HNF4α data, an "optimized" ordering of positions may in fact increase

```
1. Find two positions with the maximum dependence score and put them in $C$.
2. Set $R = P - C$.
3. while $|R| > 0$ do
    (a) Suppose that $C$ is a partial order of the form $l_1, \ldots, l_s, \{m_1, \ldots, m_t\}, r_1, \ldots, r_s$,
        where $2s + t = |C|$, $t \neq 1$, and if $t > 1$ then either $s = 0$ or $s + t = k + 1$.
    (b) If $|C| \leq k$ then for each $c_1 \in R$, define
            $W(c_1) = \sum_{c_2 \in \{m_1, \ldots, m_t\}} g(c_1, c_2)$
    (c) Elseif $|C| < 2k + 1$ then for each $c_1 \in R$ and $c_2 \in \{m_1, \ldots, m_t\}$, define
            $L(c_1, c_2) = \sum_{i=1}^{s} g(c_1, l_i) + \sum_{c_3 \in \{m_1, \ldots, m_t\} - \{c_2\}} g(c_1, c_3)$
            $R(c_1, c_2) = \sum_{i=1}^{s} g(c_1, r_i) + \sum_{c_3 \in \{m_1, \ldots, m_t\} - \{c_2\}} g(c_1, c_3)$
            $W(c_1, c_2) = \max\{L(c_1, c_2), R(c_1, c_2)\}$
            $W(c_1) = \max_{c_2 \in \{m_1, \ldots, m_t\}} W(c_1, c_2)$
    (d) Else (i.e. $|C| \geq 2k + 1$) for each $c_1 \in R$ define
            $L(c_1) = \sum_{i=s-k+1}^{s} g(c_1, l_i)$
            $R(c_1) = \sum_{i=1}^{k} g(c_1, r_i)$
            $W(c_1) = \max\{L(c_1), R(c_1)\}$
    (e) Find a position $c_1 \in R$ such that $W(c_1)$ is maximized.
    (f) Move $c_1$ from R to $C$ and modify the partial order in $C$ appropriately.
4. Output $C$ (as a linear order).
```

Fig. 2. A greedy algorithm for ordering positions in a Markov chain.

Ordering obtained by the greedy algorithm												
5	6	8	4	12	11	3	9	10	2	0	1	7
Ordering after the addition of new binding sequences												
7	5	8	4	11	12	10	3	9	2	0	1	6

Fig. 3. Ordering of positions based on the χ^2 analysis for the HNF4α data. The second ordering was obtained by using the initial binding sequences plus the new sequences identified in this project.

the relative entropy of the Markov model. Note that, a 3rd-order Markov chain of length 13 has 10 *effective* transitions.

After establishing the ordering of positions in the Markov model as described above (*i.e.* the first row of Figure 3), the model is trained by a standard method (Durbin *et al.*, 1998). The score of a sequence is simply the probability that the sequence is generated by the model. The final step of the procedure is to determine a score threshold so that only sites that score above the threshold will be output. In the HNF4α project, we considered thresholds using the formula, $\mu - \sigma \cdot J$, where μ and σ are the mean and standard deviation of the scores of the training sequences, and J is a control parameter. (But other methods are also possible.) One could choose an appropriate value for J by looking at the percentage of known binding sites that are recovered by the program. In the case of HNF4α, when J was set to 0.67, our program produced a list of 849 potential binding sites including almost all of the expected known binding sites.

Transition	Before Ordering	After Ordering	After New Data
0	1.477	2.753	2.752
1	4.676	7.791	6.897
2	2.824	5.662	6.993
3	6.696	1.733	1.828
4	3.984	1.937	2.565
5	3.547	8.462	4.447
6	9.192	5.063	12.413
7	8.728	4.634	4.083
8	8.427	7.025	6.212
9	3.202	4.345	5.580
Total	52.752	49.405	53.771

Table 4. The relative entropy of the 3rd-order Markov model for HNF4α. "Before Ordering" means the model where the positions are ordered sequentially as in the binding sequences. "After Ordering" and "After New Data" mean models based on the orders given in Figure 3. The results in the table show that the second order in Figure 3 is (slightly) better than the sequential order in terms of relative entropy, while the first order is (slightly) worse.

RESULTS ON HNF4α BINDING SITES

In this section, we demonstrate the effectiveness of the TFBS identification algorithm in Figure 1 by experimentally validating the HNF4α binding sites found by the algorithm in the human genome.

As described in the previous section, we have created an (optimized) 3rd-order Markov model using the 71 known HNF4α binding sequences listed in Table 1 and scanned selected regions of the human genome with a carefully (and empirically) chosen threshold, resulting in a list of 849 potential binding sites. These sites were divided into groups of approximately one hundred and given a binding potential index from 9 (highest potential) down to 1 (lowest potential), based on the scores computed by the algorithms. We then selected 3 to 28 sites from each group, for a total of 109 sites, to test experimentally. The selection was based on a variety of factors including similarity of the potential target gene to known HNF4α target genes, expression in appropriate tissues, repeated occurrences of a gene family, a known association with a human disease, and complete randomness. Whereas we realize that such a selection introduced a bias, it was a "knowledge based" bias and was meant to strike a balance between demonstrating general effectiveness of the algorithm and identifying new HNF4α target genes that might be of interest for future study.

For each of the 109 sites selected, oligo's containing 22 or 25 nucleotides including the HNF4α binding (motif) sequence and some short flanking sequences were synthesized by Genelink (Hawthorne, NY). The complementary oligo's were annealed according to standard protocols (Ausubel *et al.*, 1990) and tested for binding to HNF4α using competitions in a gel shift assay, essentially as previously described in (Jiang *et al.*, 1995).

Briefly, crude nuclear extracts from mammalian cells containing over expressed rat HNF4α1 were incubated with 0.5 ng of a double-stranded (ds) oligo containing a well characterized HNF4α binding site from the human apolipoprotein B promoter (ApoB.85.47) (Maeda *et al.*, 2002) in the absence or presence of 200-fold molar excess of the ds oligo's containing the sites to be tested. The ApoB.85.47 oligo was radiolabeled while the sites being tested were not labeled. After 30 minutes at room temperature, the reaction was loaded onto a native low ionic strength polyacrylamide gel and subjected to an electrical field to separate the oligo bound to HNF4α (shifted band) from unbound oligo followed by autoradiography. Non radiolabeled oligo's containing high affinity HNF4α sites competed for binding with the radiolabeled oligo and resulted in the absence of a shifted band; they were termed *strong binders*. Oligo's with sites that bind HNF4α less well yielded a reduced amount of the shifted band (*weak binders*) and oligo's with sites that do not bind HNF4α did not change the amount of the shifted band (*non-binders*). Reactions were run in parallel with oligo's containing or lacking known HNF4α binding sites (positive and negative controls, respectively) and antisera specific to HNF4α was used to verify that all the binding observed in the crude extract was due to HNF4α. All shift reactions were loaded in duplicate onto 2 gels and all oligo's giving a negative result (*i.e.* no competition) were re-tested to ensure that those oligo's were indeed added to the reaction. Some non-binders were also radiolabeled and analyzed directly for binding to HNF4α. The actual experiments will be described in greater detail (Yang *et al.*, 2002).

The results of the gel shift analysis on the 109 selected sites as listed in Table 5 indicate that the algorithm was very successful in predicting binding sites for HNF4α. Overall, 45 sites were found to bind HNF4α strongly (41%), 32 were found to bind HNF4α weakly (29%), and 32 were found not to bind HNF4α (29%). More importantly, although the number of sites tested in each group was rather small, the general trend of strong to weak binders as predicted by the algorithm was verified experimentally. Namely, the largest percentage of strong binders were in the set of sites predicted by the algorithm to bind the best (62% in group 9) and the largest percentage of non-binders were in the group of sites predicted to bind the least well (over 50% in the last three groups). Of equal importance was the fact that at least one of the sites predicted by the algorithm was subsequently identified independently as a bona fide HNF4α binding site by another research group. Interestingly, that site was in the promoter region of the HNF4α gene itself (Hatzis and Talianidis, 2001). A more thorough discussion of the biological importance of the potential HNF4α target genes identified by the algorithm will be reported in (Yang *et al.*, 2002).

		Binding Results		
Group	Total Tested	Strong	Weak	Not bind
9	21	13 (62%)	4 (19%)	4 (19%)
8	28	16 (57%)	7 (25%)	5 (18%)
7	18	5 (28%)	9 (50%)	4 (22%)
6	7	3 (43%)	2 (29%)	2 (29%)
5	8	3 (38%)	1 (13%)	4 (50%)
4	8	3 (38%)	3 (28%)	2 (25%)
3	8	1 (13%)	2 (25%)	5 (63%)
2	8	1 (13%)	3 (38%)	4 (50%)
1	3	0 (0%)	1 (33%)	2 (67%)
Total	109	45 (41%)	32 (29%)	32 (29%)

Table 5. The results of the *in vitro* DNA binding experiments.

We have also compared the performance of our "optimized" (3rd-order) Markov model with the performance of the "unoptimized" model where positions are sequentially ordered as in the HNF4α binding sequences, in terms of ranking/scoring the confirmed (previously and in this paper) binding sites among all predicted sites. The results are shown in Table 6. As a reference, the table also includes the average rank numbers achieved by respective 2nd,- 1st-, and 0th-order Markov models. Note that a 0th order Markov model is exactly a PSSM. Although this comparison may be biased in favor of the "optimized" models (because many of the confirmed sites were chosen based on the output of this model), it still shows that the ordering of positions has greatly improved the performance because the "unoptimized" model ranked these confirmed binding sites poorly (in other words, many of these sites would not have been picked up if the "unoptimized" model were used to make predictions instead). [§] The table also demonstrates that higher order Markov models generally perform much better than lower order Markov models, perhaps due to variability in the HNF4α binding sequences.

Whereas the results of the gel shift assay indicated that the algorithm was very good at predicting HNF4α binding sites, there is room for improvement since a significant number of sites predicted to bind the best did not bind at all (19% and 18% in groups 9 and 8, respectively). Some of the possible improvements will be discussed in the next section.

	Before Ordering	After Ordering	After New Data
3rd-Order	405	278	237
2nd-Order	1681	1432	720
1st-Order	14955	16832	13347
PSSM	86694	86694	86694

Table 6. The average ranks of the confirmed HNF4α binding sites in the predicted lists by 3rd-, 2nd-, 1st- and 0th-order Markov models with three different orderings of positions.

DISCUSSION AND CONCLUSION

For a majority of transcription factor binding site searches, position specific score matrixes have become the norm. Our experiment shows that the information in position dependence is important to consider, and can help in the search for more new binding sites. While the main contribution of our work is the novel idea of ordering positions in Markov model to capture the most significant inter-position dependence, our work has greatly advanced the number of known HNF4α binding sites. Although the *in vitro* analysis does not tell if HNF4α actually binds any of the 77 positively tested sites *in vivo*, nor if the transcription of the adjacent genes are actually activated, it does serve as a powerful complementary tool to *in vivo* studies for identifying potential target genes of a given transcription factor. We would predict that a combination of efficient computer search and *in vitro* validation will become an effective approach for the identification of TFBS's.

There are a number of ways to improve the algorithm. For example, the Markov model for HNF4α binding sequences was created without taking into account the background distribution of 13-nucleotide oligo's in the human genome (or in the selected transcription regions). Incorporating such background information into the model and score function would likely improve its predication accuracy. Our experience has shown that the ordering of positions in the Markov model can greatly affect the prediction. We also intend to study alternative formulations of the dependence score, such as $\log(1 - p(i_1, i_2))$, and see if they could be more effective in ordering positions.

A useful aspect of combining *in vitro* experiments with computer search is that more training data is accumulated

[§] The poor performance of the "unoptimized" model could perhaps be attributed to the lack of training data too; but then this is a reality that we face in the search for TFBS's.

in the process. This new data can be potentially very useful in training the Markov model and making it more accurate, especially when the initial training set is not very large. For example, using the 77 new HNF4α binding sites identified in the first round of experiments, we have re-ordered and re-trained our Markov model. The relative entropy and prediction accuracy (in terms of ranking the confirmed binding sequences) are given in Tables 4 and 6. A comparison with the (ordered) model without the new data shows that both the relative entropy and the prediction accuracy have improved. Moreover, the *in vitro* experiments also provide negative examples (*i.e.* the non-binders). This negative information can be incorporated into the Markov model although the training will be slightly more complicated and time consuming.

Acknowledgment. This work was supported by NIH grant DK53892 (to FMS) and NSF grants CCR-9988353 and ITR-0085910 (to TJ). Carlotta Domeniconi (UCR) helped with debugging the Markov chain code, and Jim Kent (UCSC) provided tips on writing software to read the UCSC database.

REFERENCES

Agarwal, P. and Bafna, V. (1998) Detecting non-adjoining correlations within signals in DNA. *Proc. 2nd Annual International Conf. on ComputationalMolecular Biology (RECOMB), New York, NY*, 2–8.

Antes, T.J. *et al.* (2000) Identification and characterization of a 315-base pairenhancer, located more than 55 kilobases 5' of the apolipoprotein Bgene, that confers expression in the intestine. *J. Biol. Chem.*, **275(34)**, 26637–26648.

Ausubel, F.M. *et al.* (1990) *Current Protocols in Molecular Biology. John Wiley & Sons, NY*.

Bailey, T.L. and Elkan, C.P. (1995) The value of prior knowledge in discovering motifs. *Proc. 3rd ISMB*, 21–29.

Brazma, A. *et al.* (1998) Predicting gene regulatory elements *in silico* in a genomic scale. *Genome Research*, **15**, 1202–1215.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, **268**, 78–94.

Durbin, M. *et al.* (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins andNucleic Acids. Cambridge University Press*.

Hatzis, P. and Talianidis, I. (2001) Regulatory mechanisms controlling humanhepatocyte nuclear factor 4α gene expression. *Molecular and Cellular Biology*, **21**, 7320–7330.

Hauch, W. *et al.* (1994) Transcriptional control of the human biliaryglycoprotein gene, a CEA gene family member down-regulated incolorectal carcinomas. *Eur J Biochem.*, **223(2)**, 529–541.

Hays, W. and Winkler, R. (1971) *Statistics: Probability, Inference, and Decision.Holt, Rinehart and Winston, Inc., New York.*

Iyer, V. *et al.* (2001) Genomic binding sites of the yeast cell-cycletranscription factors SBF and MBF. *Nature*, **409**, 533–538.

Jiang, G. *et al.* (1995) Exclusive homodimerization of orphan receptor hepatocyte nuclear factor 4defines a new subclass of nuclear receptors. *Molecular and Cellular Biology*, **15**, 5131–5143.

Lahuna, O. *et al.* (2000) Involvement of STAT5 (signal transducer and activator oftranscription 5) and HNF-4 (hepatocyte nuclear factor 4) in thetranscriptional control of the hnf6 gene by growth hormone. *Mol Endocrinol.*, **14(2)**, 285–294.

Lander, E. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lieb, J. *et al.* (2001) Promoter-specific binding of Rap1revealed by genome-wide map of protein-DNA association. *Nature Genetics*, 327–334.

Maeda, Y. *et al.* (2002) Repression of hepatocyte nuclear factor 4α bytumor suppressor p53: involvement of the ligand binding do-mainand hisitone deacetylase activity. *Molecular Endocrinology*, **16**, 402–410.

Nicolas-Frances, V. *et al.* (2000) The peroxisome proliferator response element (PPRE) present at positions-681/-669 in the rat liver 3-ketoacyl-CoA thiolase B genefunctionally interacts differently with PPARα and HNF-4. *Biochem Biophys Res Commun.*, **269(2)**, 347–351.

Ozeki, T. *et al.* (2001) Co-operative regulation of the transcription of human dihydrodioldehydrogenase (DD)4/aldo-keto reductase (AKR)1C4 gene by hepatocytenuclear factor (HNF)-4alpha/gamma and HNF-1α. *Biochem J.*, **355(Pt 2)**, 537–544.

Pinaire, J. *et al.* (1999) Activity of the human aldehyde dehydrogenase 2 promoter is influenced by thebalance between activation by hepatocyte nuclear factor 4 andrepression by perosixome proliferator activated receptor δ,chicken ovalbumin upstream promoter-transcription factor, andapolipoprotein regulatory protein-1. *Adv Exp Med Biol.*, **463**, 115–121.

Ren, B. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Quandt, K. *et al.* (1995) MatInd and MatInspector - New fast and versatile tools for detection ofconsensus matches in nucleotide sequence data. *Nucleic Acids Research*, **23**, 4878–4884.

Sinha, S. and Tompa, M. (2000) A statistical method for finding transcription factor binding sites. *Proc. 8th ISMB.*

Sladek, F.M. and Seidel, S.D. (2001) Hepatocyte nuclear factor 4α. *Nuclear Receptors and Disease. (eds. T. Burris, E. R.B. McCabe), Academic Press, London*, 309–361.

Swenson, E.S. *et al.* (1999) Hepatocyte nuclear factor-4 regulates intestinal expression of theguanylin/heat-stable toxin receptor. *Am J Physiol.*, **276(3 pt 1)**, G728–G736.

Stormo, G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16(1)**, 16–23.

Tan, K. *et al.* (2001) A comparative genomic approach to prediction of new members of regulons. *Genome Research*, **11**, 566–584.

Thakurta, D.G. and Stormo, G.D. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17(7)**, 608–621.

Tompa, M. (1999) An exact method for finding short motifs in sequences,with application to the Ribosome binding site problem. *Proc. 7th ISMB*, 262–270.

Vanet, A., Marsan, L. and Sagot, M.F. (1999) Promotor sequences and algorithmic methods for identifying them. *Res. Microbiology*, **150**, 779–799.

Venter, C. *et al.* (2001) The Human Genome. *Science*, **291**, 1145–1434.

Wingender, E. *et al.* (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.

Yanai, K. *et al*. (1999) Regulated expression of human angiotensinogen gene by hepatocytenuclear factor 4 and chicken ovalbumin upstreampromoter-transcription factor. *J Biol Chem.*, **274(49)**, 34605–34612.

Yang, C. *et al*. (2002) *Manuscript in preparation*.

Zhu, J. and Zhang, M.Q. (1999) SCPD: a promotor database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.