High-throughput SELEX–SAGE method for quantitative modeling of transcription-factor binding sites

Emmanuelle Roulet[1], Stéphane Busso[1], Anamaria A. Camargo[2], Andrew J.G. Simpson[2], Nicolas Mermod[1]*, and Philipp Bucher[3]*

Published online: 8 July 2002, doi:10.1038/nbt718

The ability to determine the location and relative strength of all transcription-factor binding sites in a genome is important both for a comprehensive understanding of gene regulation and for effective promoter engineering in biotechnological applications. Here we present a bioinformatically driven experimental method to accurately define the DNA-binding sequence specificity of transcription factors. A generalized profile[1] was used as a predictive quantitative model for binding sites, and its parameters were estimated from *in vitro*–selected ligands using standard hidden Markov model training algorithms[2,3]. Computer simulations showed that several thousand low- to medium-affinity sequences are required to generate a profile of desired accuracy. To produce data on this scale, we applied high-throughput genomics methods to the biochemical problem addressed here. A method combining systematic evolution of ligands by exponential enrichment (SELEX)[4] and serial analysis of gene expression (SAGE)[5] protocols was coupled to an automated quality-controlled sequence extraction procedure based on Phred quality scores[6]. This allowed the sequencing of a database of more than 10,000 potential DNA ligands for the CTF/NFI transcription factor. The resulting binding-site model defines the sequence specificity of this protein with a high degree of accuracy not achieved earlier and thereby makes it possible to identify previously unknown regulatory sequences in genomic DNA. A covariance analysis of the selected sites revealed non-independent base preferences at different nucleotide positions, providing insight into the binding mechanism.

The reliability and accuracy of existing computer tools for identifying transcription-factor DNA-binding sites are largely unknown but commonly believed to be rather low. Experimental biologists have often studied the binding specificity of transcription factors qualitatively, without attempting to estimate the parameters of a quantitative method for predicting target sites. Consequently, computational biologists have often had to rely on data inappropriate for developing such software tools[7]. In the work presented here, we combined experimental techniques with bioinformatics and proceeded in the opposite direction. First we chose a computational prediction method and determined the type and amount of experimental data needed using computer simulations. We then devised and applied experimental protocols for producing the required data in a cost-effective manner. Finally, we generated a binding-specificity model from the data in a highly automated fashion.

The experimental system chosen for developing and testing this approach was the transcription factor CTF/NFI, a protein that binds DNA as a homodimer and recognizes palindromic sequence motifs resembling TTGGC(N_5)GCCAA. An approximate quantitative model of its binding specificity was available from a previous study[8] in which we represented its binding specificity by a sequence profile (Fig. 1A), which is an extension of a weight matrix, the most commonly used descriptor for transcription-factor binding sites. It consists of two reverse-complementary weight matrices that characterize the half-site motifs recognized by each of the two subunits. Five additional parameters relate to different binding modes suggested to be relevant by previous experiments. Those include various spacer lengths between the two half-sites and alternative modes of interaction in which only one of the subunits is in contact with the DNA. A binding score can be computed for any potential binding site simply by adding up the corresponding weights.

According to the statistical mechanical theory of sequence-specific DNA–protein interactions[9], the parameters of a profile represent negative free-energy contributions to the total binding energy (Fig. 1B). Each column in the half-site matrices characterizes a hypothetical base-pair acceptor site on the protein surface. The weights for different binding modes represent relative free energies of different conformations. Therefore, the binding score defined by the profile should be inversely proportional to the total free energy of the protein–DNA complex. Note however that the absolute scale of the parameters in the profile is arbitrary (see legend to Fig. 1). As a consequence, only the differences between alternative weights are proportional to differences in free energy.

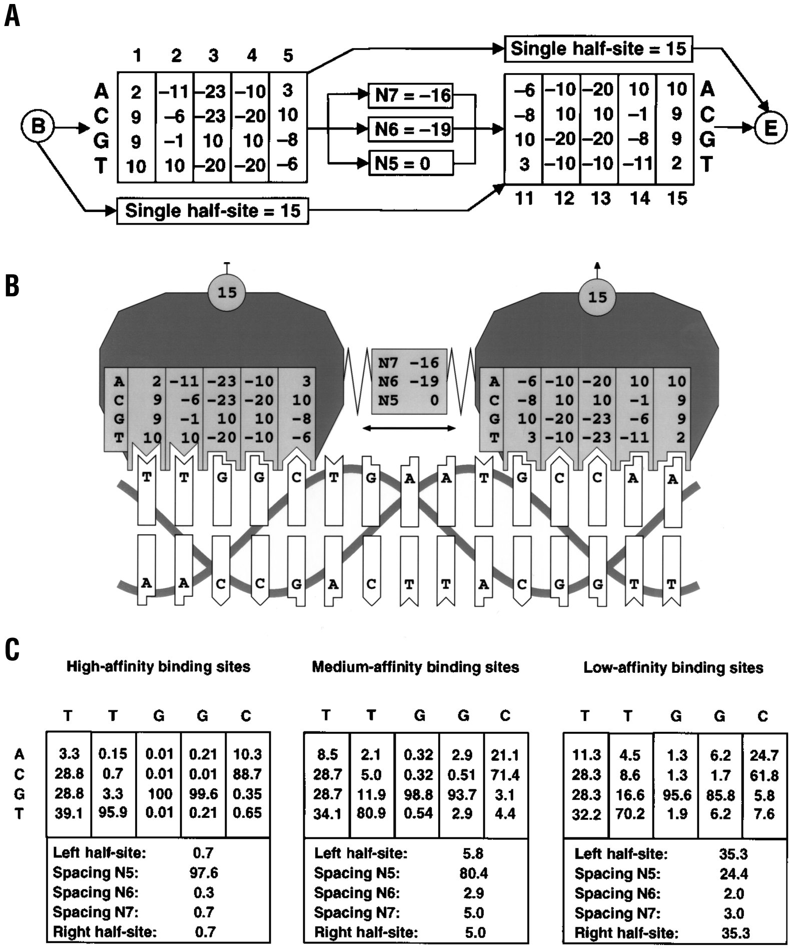
Several methods have been used to estimate profile parameters from experimental binding data. In our earlier study, these were estimated by measuring the effects of a series of mutations in a high-affinity binding site using band-shift assays[8]. Alternatively, binding sites can be selected from libraries of random sequences using an *in vitro* selection–amplification approach known as SELEX[4]. The latter approach allows the generation of larger data sets that can include up to 200 distinct sequences but that, more typically, contain 20–70 sites[10,11]. However, it has the disadvantage that the model parameters cannot be directly read out from the data, since the model-building process necessitates sequence alignments as an intermediate step. The relationship between the type and amount of such data and the accuracy of the model has never been investigated either theoretically or experimentally. Consequently, the number and affinity of the binding sequences required to generate a model of a given precision are not known. We addressed this question first by a computer simulation experiment.

Some underlying assumptions of the model building process need to be introduced to explain the simulation experiment. A profile (Fig. 1A) can be calculated from a base-frequency model (Fig. 1C). In computational sequence analysis, the latter is sometimes called a hidden Markov model (HMM)[2]. Profiles and HMMs of the same architecture are interconvertible[1], and the position-specific base probabilities $p(i,b)$ can be transformed into corresponding profile weights $w(i,b)$ by a logarithmic function:

$$w(i,b) = c_i + \log_a p(i,b)$$

where c_i is a column-specific free constant that can be exploited for scaling purposes (see Fig. 1A legend). To simulate SELEX data, we

[1]Laboratory of Molecular Biotechnology, Center for Biotechnology UNIL-EPFL, and Institute of Animal Biology, University of Lausanne, 1015 Lausanne, Switzerland. [2]Laboratory of Cancer Genetics, Ludwig Institute for Cancer Research, Sao Paulo, 01509-010, Brazil. [3]Swiss Institute for Experimental Cancer Research, Swiss Institute of Bioinformatics, 1066 Epalinges, Switzerland. *Corresponding authors (Philipp.Bucher@isrec.unil.ch and Nicolas.Mermod@iba.unil.ch).



High-affinity binding sites

	T	T	G	G	C
A	3.3	0.15	0.01	0.21	10.3
C	28.8	0.7	0.01	0.01	88.7
G	28.8	3.3	100	99.6	0.35
T	39.1	95.9	0.01	0.21	0.65

Left half-site:	0.7
Spacing N5:	97.6
Spacing N6:	0.3
Spacing N7:	0.7
Right half-site:	0.7

Medium-affinity binding sites

	T	T	G	G	C
A	8.5	2.1	0.32	2.9	21.1
C	28.7	5.0	0.32	0.51	71.4
G	28.7	11.9	98.8	93.7	3.1
T	34.1	80.9	0.54	2.9	4.4

Left half-site:	5.8
Spacing N5:	80.4
Spacing N6:	2.9
Spacing N7:	5.0
Right half-site:	5.0

Low-affinity binding sites

	T	T	G	G	C
A	11.3	4.5	1.3	6.2	24.7
C	28.3	8.6	1.3	1.7	61.8
G	28.3	16.6	95.6	85.8	5.8
T	32.2	70.2	1.9	6.2	7.6

Left half-site:	35.3
Spacing N5:	24.4
Spacing N6:	2.0
Spacing N7:	3.0
Right half-site:	35.3

Figure 1. CTF/NFI sequence-specific DNA–protein interaction profiles. (A) Previously published profile[8]. The position-specific weights associated with the four possible nucleotides are indicated within boxes. Arrows indicate alternative paths corresponding to different binding modes. The score of a potential binding site is computed as the sum of the corresponding weights. The scale is such that a decrease of ten units corresponds to a tenfold decrease in apparent DNA binding affinity[8]. Maximal weight at each matrix position, 10; highest possible score, 100 (arbitrary scaling conventions). (B) Physical interpretation of a transcription-factor binding site profile. Following Berg and von Hippel[9], the parameters represent free-energy contributions attributed to interactions between base pairs and base pair–acceptor sites, or to different conformations corresponding to different binding modes. (C) HMMs for high-, medium-, and low-affinity binding sites computed from the CTF/NFI profile shown in (A) and used to generate the training sets for the computer simulation experiment.

proceeded in the opposite direction, converting the CTF/NFI profile parameters first into several corresponding HMMs using the inverse of the above formula:

$$p(i,b) = c_i a^{w(i,b)}$$

Note that c_i serves here as a normalization factor ensuring that the base probabilities add up to 1. Frequencies and free-energy values of different binding modes are related by a similar formula, implied by the general conversion recipe for profiles and HMMs[1]. The logarithmic base a sets the average score of the sequences described by the resulting frequency model. The Berg–von Hippel theory[9] of sequence-specific protein–DNA interactions proposes that DNA ligand collections selected under different conditions can be described by frequency matrices related to the same energy–weight matrix via different logarithmic bases. Both temperature and substrate concentration influence the parameter a. The above formulae are approximations, as the theory makes some simplifying assumptions, most notably that the observed position-specific base frequencies depend only on the mean and not on the exact shape of the affinity distribution of the selected sites.

In a SELEX experiment, DNA ligand collections selected for ever higher affinities after each cycle correspond to frequency models (HMMs) obtained with increasing exponential bases a (see ref. 12 for a mathematical model of the SELEX process). We exploited this feature to generate frequency models of binding sites corresponding to low, medium, and high average affinities (Fig. 1C). To simulate corresponding SELEX data sets, we generated DNA sequences from these frequency models and added random bases on either side to reach a total length of 25 bp. New frequency models were then trained from such simulated data sets of various sizes and average affinities, using the Baum–Welch expectation-maximization algorithm, as described previously for the STAT DNA-binding proteins[3]. The precision of a new frequency model was then measured by computing the average difference between the re-estimated and original log frequencies. The results (see Supplementary Table 1 online) indicate that more precise models can be obtained from collections of lower-affinity binding sites, presumably because these exhibit unfavorable bases more often, providing more precise frequency estimates.

We next proceeded to the selection of CTF/NFI binding sites by SELEX, which consists of the selection of protein-bound oligonucleotides by native gel electrophoresis and PCR amplification of gel-extracted sequences (Fig. 2A). Typically, multiple selection cycles are required to separate binding sites from random library sequences, a process that usually enriches for a few high-affinity sites. However, the results of the computer simulations suggest that at least 2,000 low- to medium-affinity binding sites are needed to achieve an average log ratio error of 0.1, or 10% in terms of base frequencies. To be able to generate a data set with such properties, we had to modify the SELEX method in two important ways (Fig. 2A). To prevent the selection of only high-affinity binding sites, we monitored the stringency of the binding conditions at each selection cycle by including a radiolabeled 25-bp oligonucleotide probe of moderate affinity that does not hybridize with the PCR primers. The concentration of the DNA library, added to the reaction mixture as a competitor, was adjusted such that 50% of the radiolabeled probe complex was competed away, ensuring the selection of medium-affinity sites from the library. The second modification—concatenating the *in vitro*–selected binding sites in an adaption of the SAGE[5] protocol—served to increase the sequencing throughput.

We analyzed the enriched CTF/NFI sites in the oligonucleotide populations recovered after each of the four SELEX cycles (referred to as the Selex1 to Selex4 libraries) by analytical band-shift assays (see Supplementary Fig. 1 online). This indicated that the Selex3 library was of the desired average affinity, and we therefore subjected this library and representative samples from the other libraries to high-throughput sequencing. Because accurate estimation of base frequencies as low as 0.5% requires sequencing error rates of at least one order of magnitude lower, we developed a perl script that exploits the base quality scores computed by the Phred program[6] for sequence quality control. This allowed the extraction of 1,088 high-quality sequences from Selex1, 1,475 from Selex2, 6,912 from Selex3, and 361 from Selex4. A total of 1,201 sequences were also determined from the unselected Selex0 library. We estimated the overall error rate in the extracted data sets to be lower than 0.02% per base pair.

We confirmed that the sequences from the Selex0 library were not biased in a way that would affect the modeling procedure. Individual base frequencies in the supposedly random part of the oligonucleotides ranged from 24.4% to 26.4%, and the dinucleotide frequency distribution was also close to uniform. Next, we tried to confirm that the sequences from the Selex3 library were indeed in the targeted affinity range by scoring the sequences from all five libraries with the CTF/NFI profile. We observed a steady enrichment in CTF/NFI binding sites but virtually no change in the affinity distribution, as intended by the SELEX protocol modifications (Fig. 2B). The major peak of the Selex3 population is centered at a score around 80, only slightly above 77, the score of the radiolabeled oligonucleotide used for controlling the selection conditions. The minor peak coincides with the maximum of the distribution of the unselected Selex0 library and presumably represents contamination by unbound DNA. This was not a matter of concern as the algorithm chosen for building binding-site models can tolerate a certain amount of noise in the data. Overall, these analyses confirmed that the modified SELEX–SAGE protocol produced the required type and quantity of sequence data. To our knowledge, this constitutes the largest database of DNA binding sites ever produced, being almost two orders of magnitude larger than any previously reported SELEX database. The database is publicly available at http://www.isrec.isb-sib.ch/selex_nf1/.

We then built a new model from the Selex3 data using the same algorithm as in the simulations except that we added four match positions to the model, one on either side of each half-site block, to test whether bases adjacent to the TTGGC consensus motif could also have an influence on binding affinity. The 3′-flanking position showed a clear preference for adenine bases and was thus incorporated into the new model. The 5′-flanking position did not show any bias in base composition and was thus omitted from the model. The new profile computed from the frequency model (see Supplementary Fig. 2 online) is shown in Figure 2C.

Table 1. Strongly correlated dinucleotide pairs in sequences of the major spacing class of CTF/NFI binding sites

Dinucleotide pairs	f_I	f_j	f_j^{exp}	f_{ij}^{obs}	n_{ij}^{exp}	n_{ij}^{obs}	*P* value	M_{ij}	$n_{ij}^{obs}/n_{ij}^{exp}$
Within half-sites									
C_1 A_2	0.252	0.053	0.013	0.006	94.9	41	$<10^{-10}$	0.005	0.4
G_2 A_4	0.096	0.036	0.004	0.000	24.9	0	9.4×10^{-8}	0.005	0.0
A_4 C_5	0.036	0.661	0.024	0.010	171.1	71	$<10^{-10}$	0.017	0.4
A_4 T_5	0.036	0.036	0.001	0.006	9.3	44	$<10^{-10}$	0.008	4.7
A_5 C_6	0.263	0.165	0.043	0.066	308.7	469	$<10^{-10}$	0.013	1.5
C_5 C_6	0.661	0.165	0.109	0.082	775.6	580	$<10^{-10}$	0.017	0.7
Across half-sites									
A_1 T_{14}	0.130	0.053	0.007	0.001	49.0	10	8.3×10^{-10}	0.005	0.2
A_2 T_{11}	0.053	0.263	0.014	0.006	99.1	45	$<10^{-10}$	0.005	0.5
G_2 C_{14}	0.096	0.096	0.009	0.003	65.5	20	5.2×10^{-10}	0.005	0.3
A_4 G_{11}	0.036	0.661	0.024	0.034	171.1	244	$<10^{-10}$	0.013	1.4
A_4 T_{11}	0.036	0.263	0.010	0.002	68.1	14	$<10^{-10}$	0.008	0.2
A_5 T_{11}	0.263	0.263	0.069	0.047	491.2	332	$<10^{-10}$	0.010	0.7

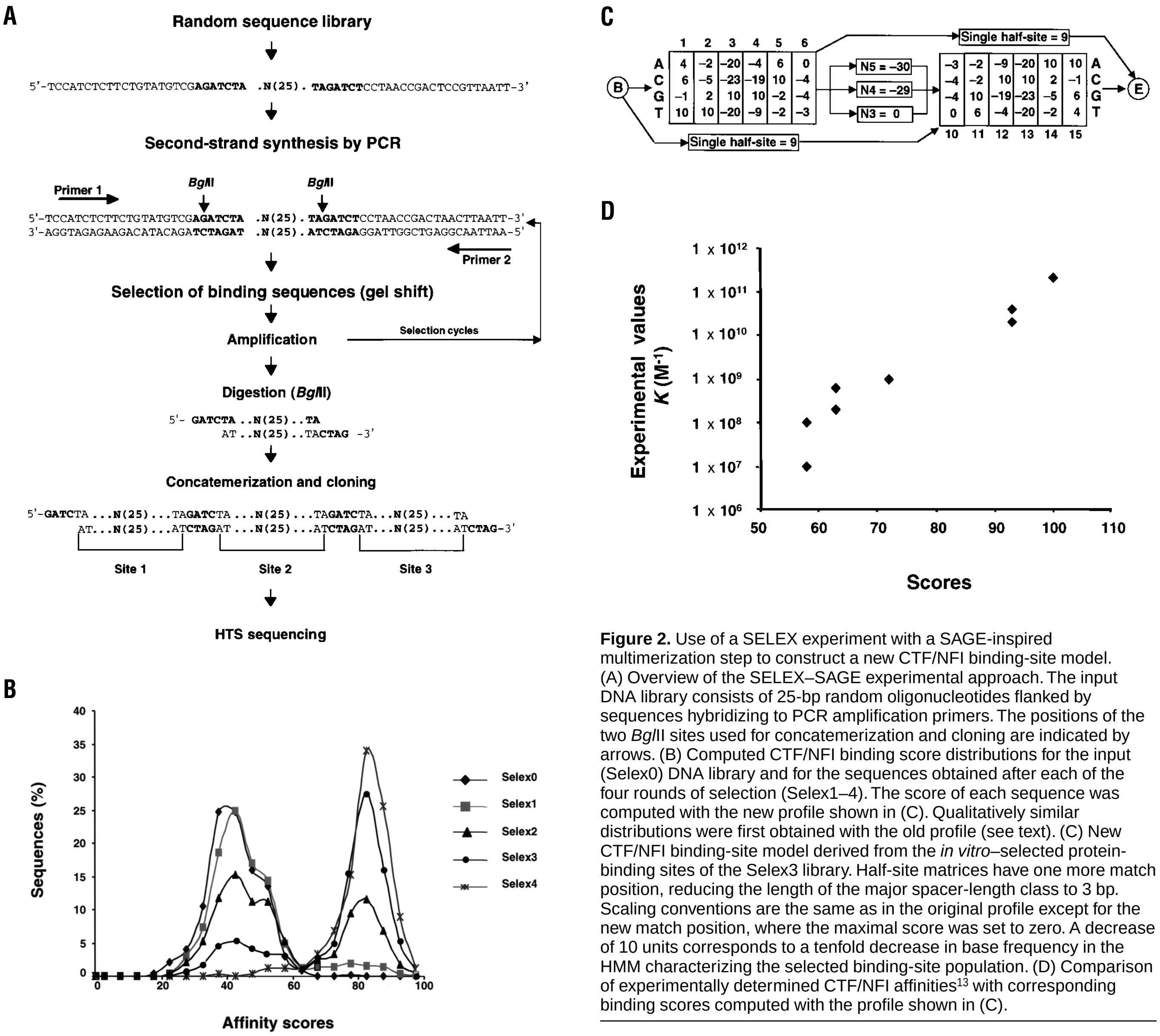
Three alternative measures were taken to assess the strength of a correlation: a *P* value based on a χ^2 test, mutual information (M_{ij}) in bits, and the ratio between the observed and expected number of occurrences ($n_{ij}^{obs}/n_{ij}^{exp}$). f_i, f_j, Frequencies of the first and the second base of the dinucleotide pair at their respective binding site positions; f_{ij}^{exp}, f_{ij}^{obs}, expected and observed frequencies of the dinucleotide *ij* as computed from f_i, f_j. Mutual information measures the global increase of nonrandomness in a binding-site population induced by a dinucleotide correlation. High values can thus be achieved only by pairs of bases that frequently occur at the corresponding positions. In contrast, the ratio values may directly relate to free energies of cooperative interactions and thus are more interesting from a physicochemical viewpoint. This list includes all correlated dinucleotide pairs satisfying one of the following conditions: (i) $M_{ij} > 0.01$ and $n_{ij}^{obs}/n_{ij}^{exp} > 1.2$; (ii) $M_{ij} > 0.005$ and $n_{ij}^{obs}/n_{ij}^{exp} > 2$; (iii) $M_{ij} > 0.01$ and $n_{ij}^{obs}/n_{ij}^{exp} < 0.8$; (iv) $M_{ij} > 0.005$ and $n_{ij}^{obs}/n_{ij}^{exp} < 0.5$.

The accuracy of the new model was assessed by cross-validation, yielding an average error rate of 7.1% between independent profiles, as expected from the simulation data (see Supplementary Fig. 3 online). We next tested the new model by computing binding scores for a set of CTF/NF1 binding sequences for which absolute affinities had previously been determined experimentally by another laboratory using Scatchard plots[13]. The predicted and experimentally determined binding affinities were in excellent correlation (Fig. 2D).

Several significant differences were noted between the new and old profiles. For instance, the new model is much less permissive for half-site spacing variants. Such variants have been consistently found to be inactive in transfected cells[8]. There were also specific changes in base frequencies. For example, the introduction of adenine bases at positions 2 and 4 is much less deleterious with the new profile, which again correlates well with the previously unexplained mild effects of these substitutions in transfection assays[8]. Thus, these differences clearly represent an improvement over the old model.

Finally, we scanned the Eukaryotic Promoter Database (EPD)[14] using the new profile for human promoters with high-scoring CTF/NFI binding sites. The promoters of the vimentin, interleukin-3, tissue plasminogen activator, and cytochrome P450-17 genes were identified as likely physiological targets and therefore selected for transfection in a cell line naturally devoid of CTF/NFI protein. The former two promoters were indeed induced upon co-expression of CTF/NFI protein (data not shown). Remarkably, we observed activation by CTF/NFI protein from single binding sites on transfected promoters. Other sequences correspond to previously identified functional CTF/NFI sites (data not shown). These three independent tests demonstrate that our new binding-site model derived from a high-throughput SELEX–SAGE experiment reliably and accurately predicts *in vitro* and *in vivo* binding sites.

The Selex3 data set is perhaps the first binding-site collection large enough to challenge some of the inherent assumptions of profile-based models. A question debated at length is whether individual base pairs of a binding site interact independently with the protein surface. A clear case of non-independent interactions was recently demonstrated for two adjacent positions in phage P22 Mnt repressor binding sites[15]. To investigate whether such effects also occur in CTF/NFI target sequences, we subjected 3,602 sites of the major spacer-length class to covariance analysis (Table 1). Not surprisingly, the strongest dinucleotide correlations were found at adjacent positions within the same half-site. For instance, the dinucleotide AT at positions 4 and 5 is strongly over-represented relative to AC, with highly significant *P* values. A strong covariance was also noted between positions 5 and 11. This explains well the previously observed non-additive effects of substitutions at these positions[8]. Even though some of the weaker correlations may reflect PCR artifacts (such as selection against perfect palindromes), these results make clear that the assumption of independent base-pair interactions represents an oversimplification of the CTF/NFI DNA-binding mechanism. More sophisticated binding-site models such as weight-array matrices[16] or maximal-dependence decomposition models[17] would be more appropriate from a physical perspective. The overall gain in prediction accuracy expected from such models should be relatively minor, but



	1	2	3	4	5	6
A	4	–2	–20	–4	6	0
C	6	–5	–23	–19	10	–4
G	–1	2	10	10	–2	–4
T	10	10	–20	–9	–2	–3

10	11	12	13	14	15	
–3	–2	–9	–20	10	10	A
–4	–2	10	10	2	–1	C
–4	10	–19	–23	–5	6	G
0	6	–4	–20	–2	4	T

Figure 2. Use of a SELEX experiment with a SAGE-inspired multimerization step to construct a new CTF/NFI binding-site model. (A) Overview of the SELEX–SAGE experimental approach. The input DNA library consists of 25-bp random oligonucleotides flanked by sequences hybridizing to PCR amplification primers. The positions of the two *Bgl*II sites used for concatemerization and cloning are indicated by arrows. (B) Computed CTF/NFI binding score distributions for the input (Selex0) DNA library and for the sequences obtained after each of the four rounds of selection (Selex1–4). The score of each sequence was computed with the new profile shown in (C). Qualitatively similar distributions were first obtained with the old profile (see text). (C) New CTF/NFI binding-site model derived from the *in vitro*–selected protein-binding sites of the Selex3 library. Half-site matrices have one more match position, reducing the length of the major spacer-length class to 3 bp. Scaling conventions are the same as in the original profile except for the new match position, where the maximal score was set to zero. A decrease of 10 units corresponds to a tenfold decrease in base frequency in the HMM characterizing the selected binding-site population. (D) Comparison of experimentally determined CTF/NFI affinities[13] with corresponding binding scores computed with the profile shown in (C).

remains to be determined. Significant pair correlations taken together are estimated to contribute <1 bit of information to the binding specificity defined by the new model, which represents ~12 bits.

In summary, we have obtained an accurate, quantitative binding-site model for a mammalian transcription factor using high-throughput SELEX–SAGE experiments and appropriate bioinformatics tools. Large-scale application of this approach could lead to reliable binding site–prediction tools for all transcription factors of a given organism. Such tools would facilitate the comprehensive understanding of gene regulation and the rational design of control regions for biotechnological applications.

Experimental protocol

***In vitro* selection and amplification of protein-binding sites.** The DNA serving as input in the selection procedure consisted of 25-bp double-stranded oligonucleotides flanked by primer sequences (Fig. 2A). Library DNA was added as competitor to CTF/NFI protein incubated with 1 ng of the medium-affinity 25-bp ^{32}P-labeled oligonucleotide probe 5′-GTCCCTGGGCGTGCAGCCCATGCAC-3 as described previously[8]. The amount of library DNA was titrated such that 50–80% of the radiolabeled complex was competed away. After electrophoresis, the DNA–protein complexes were eluted from the gel and PCR amplified using standard procedures.

Construction of the CTF/NFI binding site clone libraries. The input oligonucleotide library (Selex0) and the output of the four selection cycles (Selex1–Selex4) were digested with *Bgl*II after large-scale PCR amplification. The resulting 36-bp fragments were gel purified and multimerized using the SAGE protocol[5]. Concatemers of 400–500 bp were cloned into *Bam*HI-cleaved pZero vector (Invitrogen, Groningen, The Netherlands) for sequencing.

Extraction of binding sites from sequence trace files. The Phred program[6] was used for automatic sequence extraction and base quality score assignments. Individual sites were excised with the program pfsearch from the pftools package (see below) and a circular profile reflecting the tandemly repeated structure of the inserts. The minimal base quality was recorded for each extracted site. High-quality sites were then extracted from the primary sites database with a perl script selecting those confirmed by double-strand sequencing, or those with a minimal base quality score ≥20.

Sequence analysis software and procedures. The following public programs and software packages were used in this work: pftools version 2.2 (P. Bucher, unpublished; available at ftp://ftp.isrec.isb-sib.ch/sib-isrec/pftools/); SAM version 1.3.3 (ref. 18;http://www.cse.ucsc.edu/research/compbio/sam.html); HMMER version 1.8.4 (S.R. Eddy, unpublished; available at http://hmmer.wustl.edu). Binding scores for oligonucleotide sequences were computed with the program pfsearch (pftools). Simulated SELEX data were generated by first converting the CTF/NFI profile (Fig. 1A) into an equivalent HMM with the program ptoh (pftools), and then by generating random instances from the HMM with the program hmme (HMMER). Exponential bases of 1.14, 1.19, and 1.36 were used to convert the same profile into different HMMs representing low-, medium-, and high-affinity binding sites, respectively. New HMMs were derived from the simulated SELEX data and from the Selex3 database with the program buildmodel (SAM). The details of the computational recipe can be found on our website (http://www.isrec.isb-sib.ch/selex_nf1/). The new profile (Fig. 2C) was computed from the new HMM with the program htop (pftools) using a logarithmic base of 1.26 for conversion (corresponding to the $10 \times \log_{10}$ scale used in the old profile). The profile weights were subsequently rescaled manually to conform to the conventions applied in the old profile (see Fig. 1A legend).

The covariance analysis was done on a set of 3,602 15-mer sites extracted from the Selex3 library with binding score ≥65 according to the new profile. Each sequence was presented in both orientations. The calculation of the χ^2 test variable and the mutual information value[2] was based on a 2×2 contingency table representation of the corresponding base frequencies: we considered only the presence or absence of the specific bases under consideration at each position.

Note: Supplementary information is available on the Nature Biotechnology website.

Acknowledgments
We thank Victor Jongeneel for support and suggestions, Roman Chrast and Stylianos Antonarakis for help with the SAGE procedure, Khalil Kadaoui for assistance, and Alan McNair for helpful comments on the manuscript. The financial support of the Ludwig Institute for Cancer Research, the Etat de Vaud, and the Swiss National Science Foundation (grants 31-63933.00 and 31-59370.99) are gratefully acknowledged.

Received 28 September 2001; accepted 10 May 2002

1. Bucher, P., Karplus, K., Moeri, N. & Hofmann, K. A flexible motif search technique based on generalized profiles. *Comput. Chem.* **20**, 3–29 (1996).
2. Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids* (Cambridge Univ. Press, Cambridge, United Kingdom, 1998).
3. Ehret, G.B. *et al.* DNA binding specificity of different STAT proteins. Comparison of *in vitro* specificity with natural target sites. *J. Biol. Chem.* **276**, 6675–6688 (2001).
4. Klug, S.J. & Famulok, M. All you wanted to know about SELEX. *Mol. Biol. Rep.* **20**, 97–107 (1994).
5. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
6. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **3**, 175–185 (1998).
7. Roulet, E., Fisch, I., Junier, T., Bucher, P. & Mermod, N. Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA. *In Silico Biol.* **1**, 21–28 (1998).
8. Roulet, E. *et al.* Experimental analysis and computer prediction of CTF/NF-I transcription factor DNA binding sites. *J. Mol. Biol.* **297**, 833–848 (2000).
9. Berg, O.G. & von Hippel, P.H. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193**, 723–750 (1987).
10. Goodman, S.D., Velten, N.J., Gao, Q., Robinson, S. & Segall, A.M. *In vitro* selection of integration host factor binding sites. *J. Bacteriol.* **181**, 3246–3255 (1999).
11. Fields, D.S., He, Y.Y., Al-Uzri, A.Y. & Stormo, G.D. Quantitative specificity of the Mnt repressor. *J. Mol. Biol.* **271**, 178–194 (1997).
12. Vant-Hull, B., Payano-Baez, A., Davis, R.H. & Gold, L. The mathematics of SELEX against complex targets. *J. Mol. Biol.* **278**, 579–597 (1998).
13. Meisterernst, M., Gander, I., Rogge, L. & Winnacker, E.L. A quantitative analysis of nuclear factor I/DNA interactions. *Nucleic Acids Res.* **16**, 4419–4435 (1988).
14. Perier, R.C., Praz, V., Junier, T. & Bucher, P. The eukaryotic promoter database EPD. *Nucleic Acids Res.* **28**, 302–303 (2000).
15. Man, T.K. & Stormo, G.D. Non-independence of Mnt repressor-operator interactions determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.* **29**, 2471–2478 (2001).
16. Zhang, M.Q. & Marr, T.G. A weight array method for splicing signal analysis. *Comput. Appl. Biosci.* **9**, 499–509 (1993).
17. Burge, C.B. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
18. Hughey, R. & Krogh, A. Hidden Markov models for sequence analysis. Extension and analysis of the basic method. *Comput. Appl. Biosci.* **12**, 95–107 (1996).

An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments

X. Shirley Liu[1], Douglas L. Brutlag[2], and Jun S. Liu[3]*

Published online: 8 July 2002, doi:10.1038/nbt717

Chromatin immunoprecipitation followed by cDNA microarray hybridization (ChIP–array) has become a popular procedure for studying genome-wide protein–DNA interactions and transcription regulation. However, it can only map the probable protein–DNA interaction loci within 1–2 kilobases resolution. To pinpoint interaction sites down to the base-pair level, we introduce a computational method, Motif Discovery scan (MDscan), that examines the ChIP–array-selected sequences and searches for DNA sequence motifs representing the protein–DNA interaction sites. MDscan combines the advantages of two widely adopted motif search strategies, word enumeration[1–4] and position-specific weight matrix updating[5–9], and incorporates the ChIP–array ranking information to accelerate searches and enhance their success rates. MDscan correctly identified all the experimentally verified motifs from published ChIP–array experiments in yeast[10–13] (STE12, GAL4, RAP1, SCB, MCB, MCM1, SFF, and SWI5), and predicted two motif patterns for the differential binding of Rap1 protein in telomere regions. In our studies, the method was faster and more accurate than several established motif-finding algorithms[5,8,9]. MDscan can be used to find DNA motifs not only in ChIP–array experiments but also in other experiments in which a subgroup of the sequences can be inferred to contain relatively abundant motif sites. The MDscan web server can be accessed at http://BioProspector.stanford.edu/MDscan/.

Although the 10 to 1,000 binding loci selected by ChIP–array experiments may contain false positives, those with high ChIP–array enrichment are more likely to represent true positives with multiple protein–DNA binding sites. MDscan takes advantage of this knowledge by first searching the highly ChIP–array-enriched fragments thoroughly, generating multiple candidate motif patterns, and then updating and refining the candidate motifs using other less likely sequences, guided by statistical scoring functions derived from Bayesian statistical formulation[7]. We applied MDscan to both simulated and biological data sets and compared its performance with BioProspector[9], CONSENSUS[5], and AlignACE[8].

In simulation studies, nine motif models were manually created (Table 1A), representing three different motif widths and three

[1]Stanford Medical Informatics, [2]Department of Biochemistry, Stanford University, Stanford CA 94305. [3]Department of Statistics, Harvard University, 1 Oxford Street, Cambridge MA 02138.
**Corresponding author (jliu@stat.harvard.edu).*