

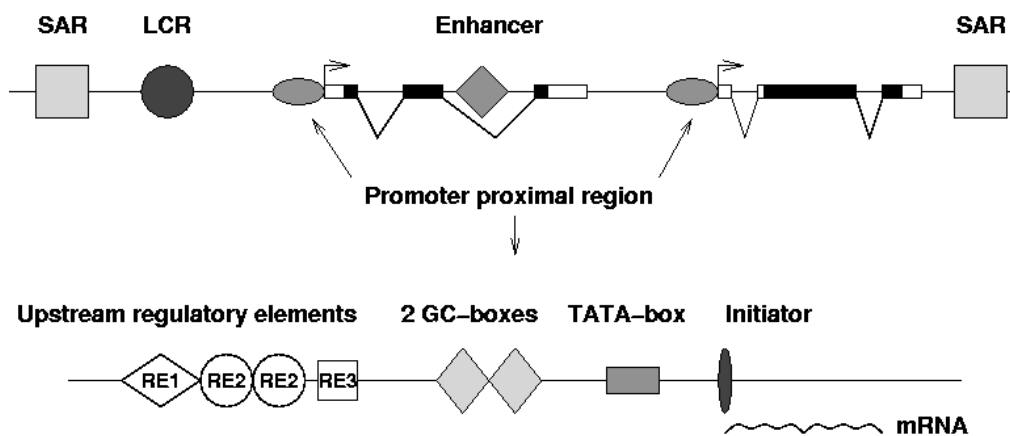
Current Motif Discovery Tools and their Limitations

Philipp Bucher

SIB / CIG Workshop

3 October 2006

Architecture of Eukaryotic Transcriptional Control Regions



Trendy Concepts and Hypotheses

- Transcription regulatory elements act in a context-dependent manner. The autonomous functional unit is a regulatory region consisting of several transcription factor binding sites.
- Transcription regulatory regions are recognized by large protein complexes consisting of several DNA binding and non-binding transcription factors
- The regulation of transcription initiation takes place at different levels: DNA methylation, chromatin remodeling, pre-initiation complex assembly, re-initiation at assembled complexes. Some of the epigenetic regulatory states are inherited through mitosis.
- Transcription regulatory regions occur everywhere in the genome: in promoters, introns, downstream regions, probably also in coding regions. They may act over very long distances in linear DNA sequence space.

Experimental Techniques For Studying Gene Regulatory Elements

- Promoter mapping experiments: nuclease protection and primer extension analysis. **New technologies: RACE, 5'SAGE, CAGE:**
- Reverse genetics: Introduction of mutations into native regulatory regions and study of their effects in experimental gene expression systems (e.g. transfected cells, transgenic organisms)
- *In vitro* protein binding experiments: DNA footprinting, electrophoretic mobility shift assays, SELEX experiments
- In vivo DNA and chromatin structure analysis: DNA methylation, DNase I sensitivity assay, in vivo footprinting, chromatin IP etc. **New technologies: ChIP-chip.**
- Global gene expression profiling of cells, tissues or organisms with trans-regulatory defects

Limitations of Experimental Approaches

Transcription regulatory events are not directly observable. Most data are open to different interpretations, for instance:

- Expression conditions in a reverse genetics experiment (e.g. DNA and protein concentrations) may not resemble physiological conditions.
- A *cis*-regulatory element binding a particular protein *in vitro* may not bind the same protein *in vivo*.
- Point mutations leading to lower transcription rates may be explained by: (i) the destruction of an activator element, (ii) the accidental creation of a repressor element.

Some data require advanced computational tools in order to be converted into a testable hypothesis. Known hard problems are:

- The derivation of a transcription factor binding site predictor from a set of example sequences
- The extraction of a regulatory network from a set of gene expression profiles

Computational Approaches to Gene Regulatory elements

Commonly addressed problems and corresponding bottlenecks:

- Finding a common sequence motif in a set of sequences known to contain a binding site to the same transcription factor or to confer the same regulatory property to an adjacent gene. **Bottleneck:** appropriate data sets, computer algorithms. **New perspective:** ChIP-chip data.
- Identification of sequence motifs that are over-represented at a particular distance from transcription initiation sites. **Bottleneck:** large sets of experimentally mapped promoters. **Hope comes from mass genome annotation (MGA) data (CAGE, etc.).**
- Identification of transcription factor binding sites and other sequence elements in DNA regulatory sequences. **Bottleneck:** accurate and reliable motif descriptions.
- Development of promoter prediction algorithms. **Current bottleneck:** large sets of experimentally mapped promoters. **Perspective:** MGA data.
- Identification and interpretation of conserved non-coding sequence regions between orthologous genes of related organisms. **Current bottleneck:** concepts and models of gene regulatory regions.

Transcription Factor Binding Sites: Features and Facts

Degenerate sequence motifs

Typical length: 6-20 bp

Low information content: 8-12 bits (1 site per 250-4000 bp)

Quantitative recognition mechanism: measurable affinity of different sites may vary over three orders of magnitude

Regulatory function often depends on cooperative interactions with neighboring sites

Representation of sequence families, domains, and motifs

Exact word:

A T G

PROSITE pattern:
(regular expression)

<x(0,1)-[STA]-x(0,1)-W-[DENQH]-x-[YI]-x-[DEQ]

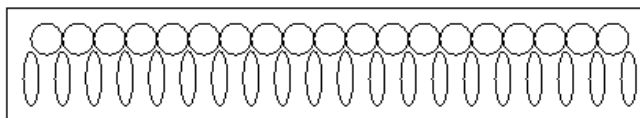
Consensus sequence:

T A T A W A W R

Weight matrix:

	1	2	3	4	5	6	7	8
A	0	9	0	9	8	9	6	4
C	2	0	3	0	1	0	0	2
G	1	0	0	2	0	4	3	4
T	5	4	9	5	7	1	5	0

Generalised profile:
(hidden Markov model)



Weight matrices – Position specific scoring matrices

A weight matrix or position-specific scoring matrix (PSSM) is a table of numbers containing scores for each residue at each position of a fixed-length (gap-free) motif.

There are two types of numerical representations:

- frequency matrix: reflects position-dependent frequencies of residues
- Scoring matrix: contains additive weights for computing a match score

Weight matrices or PSSMs are quantitative, fixed-length motif descriptors. Unlike regular expressions, they can distinguish between mild and severe mismatches.

For qualitative definition of a match, a cut-off score needs to be defined.

Generation of a Weight Matrix from Unaligned Sequences

Sequences:

```

A G G C G T G G G T A A G G C T T A G
G T G G G A C G G T A A G G C C A G C C G
T C C G G C G T A A C T A A G C C C C G
G G G G C G T C T A A A G C C C C G C G
T A T A G C C C A C T T A C T A G A G C T
A G A C T T A A A T A A A G G C G T A
C C A C C T A T C G T A A T C A G G T A
C C C G G T T G T G C A A A G G T T G C
    
```

Alignment:

```

      A G G C G T G G G G T A A G T T A G C C C A G
      G T G C G G A C C C G G A G
      C C G G C G C C C C C C C C G C G
T A T G C A C T G T A G C C C T A A A G G T G G G C G T
      C A C T T C C C C T A A A A G G T G A T T G T T G
      C G G G T T G G C A C A A A A A G A C C
    
```

Frequency matrix:

A:	2	1	9	0	8	7	6	6	1	1
C:	4	3	0	1	0	0	0	0	1	4
G:	4	1	0	0	0	0	4	3	7	3
T:	0	5	1	9	2	3	0	1	1	2

Score matrix:

A:	-1	-2	5	-5	4	4	3	3	-2	-2
C:	2	1	-5	-2	-5	-5	-5	-5	-2	2
G:	2	-2	-5	-5	-5	-5	2	1	4	1
T:	-5	2	-2	5	-1	1	-5	-2	-2	-1

It is not always easy to guess where the binding sites to a transcription factor are located in DNA sequences which are somewhat longer than the actual binding sites

HSV-1 Delayed Early Promoters

		-40	-30	-20	-10	0
		/	/	/	/	/
HSV-1 82K AlkExo	CAGCACCAAGGAGAGGGCTTAACTCTGGGAGGCCAGCCACCGACGACAGTATCGC					
HSV-1 42K	ATGGGTGCGGTATATGCACTTCCATAAGACTCTCCCCACCGCCACAGAG					
HSV-1 39K dUTPase	CGTGTGCGATATAACACAGCCCATCGAGGCCATGCCTACATAAAAGGGCACCA					
HSV-1 33K	GGCCGGGGACCCAGATGTTTCTTAAAGGGCGTGCCTCCGCGCCATGCACC					
HSV-1 21K	CGACGTACCGATGAGATCAATAAAAGGGGGCGTGAGGACCGGGAGGCGGCCAG					
HSV-1 5 kb	CCCCACCCCTGCGCGATGTGCATAAAAGGCCAGCGGGGTGGTTTAGGGTACCA					
HSV-1 RNR2	GGTCCGCCCTTCTGGTCCACGCAATAAGCGCGGACTAAAAACAGGGATGTACTA					
HSV-1 tk	CGCGGTCCCAAGTCCACTTCGCATATTAAGGTGACGGTGTGGCCTCGAATACC					
HSV-1 dbp	CGGCACGCCCCAGGTAAAGTGTACATATACCAACCGCATACCAGACGCACCCG					
HSV-1 gB 3.3 kb	CCACTCAGCGCGCGCCTGGCGATATATTCGGAGCTGATTATCGCCACCACAC					
HSV-1 gD	AGGGGTATAACAAGTCTGTCTTTAAAAAGCAGGGGTTAGGGAGTTGTTCGGTC					
HSV-1 gE	GGAGAGGGCCCGCGCCATTTAAGCGTGTGTGTGACTTTGCCCTCTCTG					
HSV-1 ICP 18.5	AATTATTGCTACGACATCCCGTCTTGTGTGTTCGGTGTCTATATCTCTGGGC					

Algorithms for defining weight matrices

Word counting algorithms (for *ab initio* discovery of consensus sequences)

- Chose word length, define number of allowed mismatches
- Count number of occurrences of each word in input sequences, most frequent words are motif candidates

Iterative refinement algorithms:

1. Starting with a consensus sequence-like matrix (same score for all matches, same score for all mismatches) find best match(es) in each input sequences.
2. Compile new set of putative matches, derive frequency matrix, compute score, and repeat step 1 with new matrix
3. Iterate step 1 and 2 until the matrix does not change anymore.

Motif discovery algorithms – details.

Computation of log-odds scores from base counts:

$$w_{bi} = \log \frac{(n_{bi} + 1)/(N_i + 4)}{e_b}$$

n_{bi} denotes the number of bases b at position i , N_i the total number of bases at position i , and e_b the expected fraction of bases b , usually estimated from the base composition of the input sequences.

Variants of the iterative alignment procedure:

- Expectation-Maximization: Instead of picking the best motif of within a sequence, compute motif probability scores for each subsequence and take weighted average over subsequences (useful in case of two equally good motif occurrences)
- Gibbs-sampling: Instead of taking the best motif per sequence, choose randomly one of the most probable, according to probability distribution (may overcome local optimum problem)

Motif discovery – a multiple local alignment problem

The motif discovery problem may be viewed as a multiple local alignment problem:

For a given set of input sequences:

Find one or several un-gapped multiple alignments (sets of fixed-length sequences) minimizing the entropy function.

Constraints about motif occurrences:

Each motif must occur exactly once,

at most once,

or one or several times

per sequence.

Limitations of Motif Discovery Algorithms

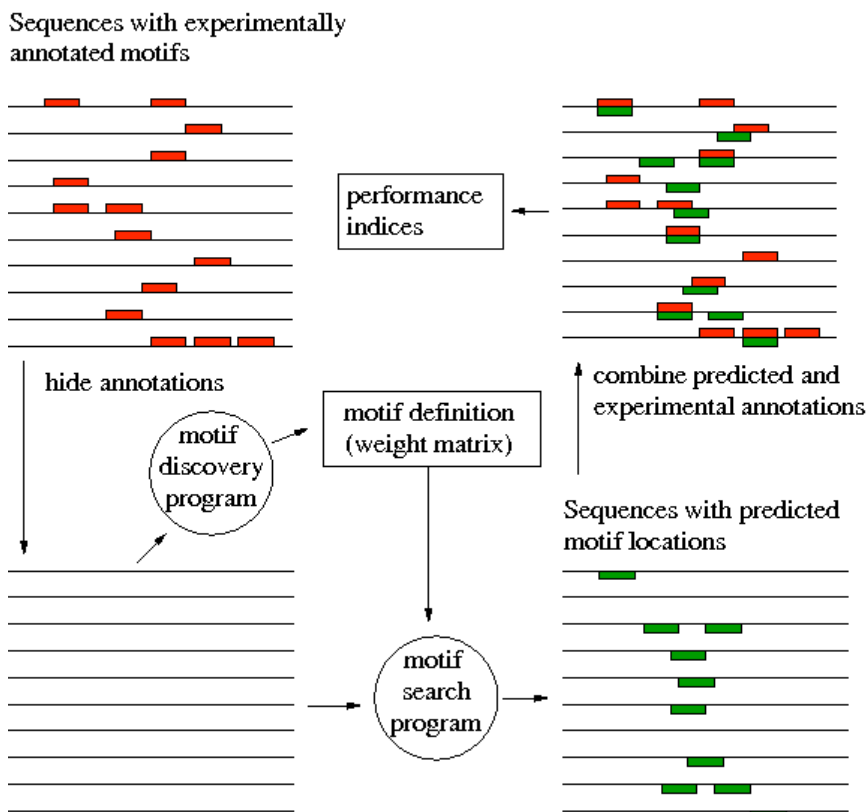
A recent papers show that computational motif discovery is disappointingly ineffective:

Assessing computational tools for the discovery of transcription factor binding sites

Martin Tompa^{1,2}, Nan Li¹, Timothy L Bailey³, George M Church⁴, Bart De Moor⁵, Eleazar Eskin⁶, Alexander V Favorov^{7,8}, Martin C Frith⁹, Yutao Fu⁹, W James Kent¹⁰, Vsevolod J Makeev^{7,8}, Andrei A Mironov^{7,11}, William Stafford Noble^{1,2}, Giulio Pavesi¹², Graziano Pesole¹³, Mireille Régnier¹⁴, Nicolas Simonis¹⁵, Saurabh Sinha¹⁶, Gert Thijs⁵, Jacques van Helden¹⁵, Mathias Vandenbogaert¹⁴, Zhiping Weng⁹, Christopher Workman¹⁷, Chun Ye¹⁸ & Zhou Zhu⁴

The prediction of regulatory elements is a problem where computational methods offer great hope. Over the past few years, numerous tools have become available for this task. The purpose of the current assessment is twofold: to provide some guidance to users regarding the accuracy of currently available tools in various settings, and to provide a benchmark of data sets for assessing future tools.

Benchmarking Protocol for Motif Discovery Algorithm



Bad Performance indices used by Tompa et. al. 2005

- nTP is the number of nucleotide positions in both known sites and predicted sites,
- nFN is the number of nucleotide positions in known sites but not in predicted sites,
- nFP is the number of nucleotide positions not in known sites but in predicted sites, and
- nTN is the number of nucleotide positions in neither known sites nor predicted sites.

Finally, it is enlightening to consider various single statistics that in some sense average (some of) these quantities. Following Pevzner & Sze¹, define the (nucleotide level) performance coefficient as:

- $nPC = nTP / (nTP + nFN + nFP)$.

Following Burset & Guigó³, define the (nucleotide level) correlation coefficient as:

$$nCC = \frac{nTP \cdot nTN - nFN \cdot nFP}{\sqrt{(nTP + nFN)(nTN + nFP)(nTP + nFP)(nTN + nFN)}}$$

and the (site level) average site performance as:

- $sASP = (sSn + sPPV) / 2$.

- sTP be the number of known sites overlapped by predicted sites,
- sFN be the number of known sites not overlapped by predicted sites, and
- sFP be the number of predicted sites not overlapped by known sites.

At either the nucleotide ($x = n$) or site ($x = s$) level, one can then define:

- *Sensitivity*: $xSn = xTP / (xTP + xFN)$, and
- *Positive Predictive Value*: $xPPV = xTP / (xTP + xFP)$.

The sensitivity gives the fraction of known sites (or site nucleotides) that are predicted, and the positive predictive value gives the fraction of predicted sites (or site nucleotides) that are known.

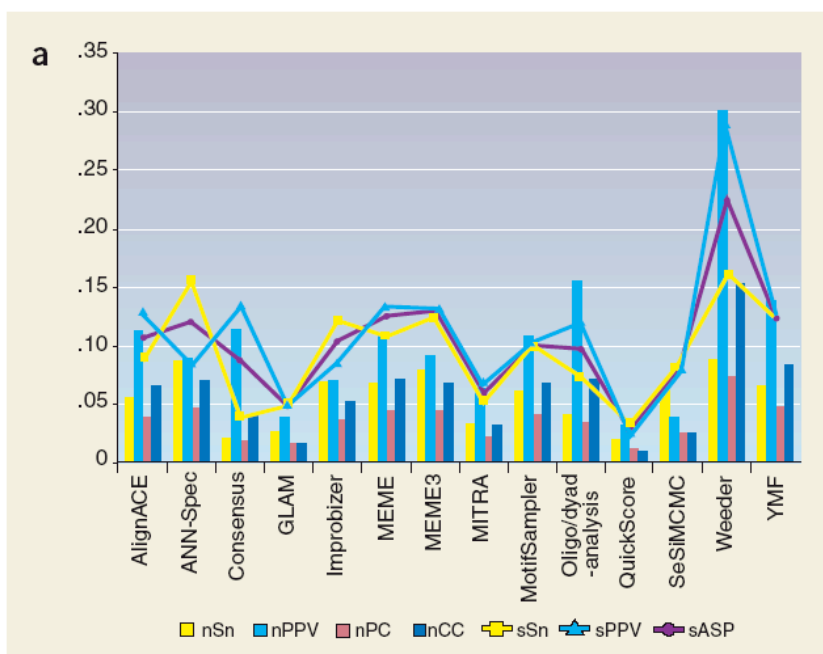
At the nucleotide level one can also define:

$$Specificity: nSP = nTN / (nTN + nFP).$$

Finally, it is enlightening to consider various single statistics that in some sense average (some of) these quantities. Following Pevzner & Sze¹, define the (nucleotide level) performance coefficient as:

- $nPC = nTP / (nTP + nFN + nFP)$.

Bad Performance of Motif Discovery algorithms on Eukaryotic Benchmark Data Sets (Results from Tompa et. al. 2005)



Similar results are obtained with prokaryotic benchmark datasets

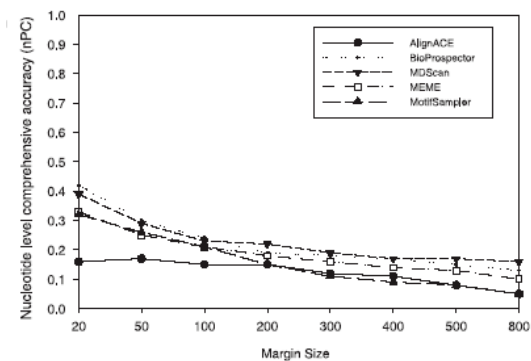
Limitations and potentials of current motif discovery algorithms

Jianjun Hu^{1,2}, Bin Li² and Daisuke Kihara^{1,2,3,4,*}

¹Department of Biological Sciences, ²Department of Computer Science, ³Markey Center for Structural Biology and ⁴The Bindley Bioscience Center, College of Science, Purdue University, West Lafayette, IN 47907, USA

Algorithms	Nucleotide level			
	nPC	nSn	nSp	nF
AlignACE	0.128	0.198	0.152	0.172
BioProspector	0.174	0.205	0.270	0.233
MDScan	0.149	0.177	0.230	0.200
MEME	0.158	0.259	0.199	0.225
MotifSampler	0.153	0.179	0.237	0.204
Random	0.050	0.061	0.083	0.070

BindingSite level			
sPC	sSn	sSp	nF
0.234	0.355	0.335	0.345
0.294	0.424	0.374	0.397
0.240	0.328	0.355	0.341
0.295	0.461	0.436	0.448
0.302	0.331	0.476	0.390
0.100	0.161	0.146	0.153



Modeling of a Transcription Factor Binding Site from High Throughput SELEX Data Using a Hidden Markov Modeling Approach

Emmanuelle Roulet, Nicolas Mermod (Center for biotechnology UNIL-EPFL, Lausanne, Switzerland)

Anamaria A Camargo, Andrew JG Simpson (Ludwig Institute of Cancer Research, Sao Paulo, Brazil)

Philipp Bucher (Swiss Institute for Experimental Cancer Research and Swiss Institute of Bioinformatics, Epalinges s/Lausanne, Switzerland)

Nat. Biotechnol. 20, 31-835 (2002)

Motivation and Goals of the Project

Motivation: Accurate and reliable computational tools to predict transcription factor binding sites are still not available.

Potential reasons:

1. Lack of adequate experimental data
2. Lack of adequate computational models
3. Lack of an adequate method to estimate the parameters of a computational model from the experimental data

Goal: To develop a combined computational-experimental protocol to derive an accurate predictive model of the sequence specificity of a DNA-binding protein

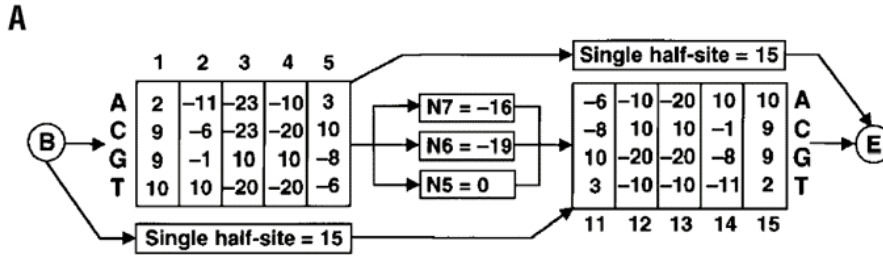
Potential benefits:

1. Being able to predict transcription factor binding in genome sequences.
2. Insights into molecular mechanisms of sequence-specific protein-DNA interactions
3. Ability to rationally design gene control regions of desired properties for biotechnological applications

Our Approach to the Problem of Characterizing the Sequence-Specificity of a DNA Binding Transcription Factor

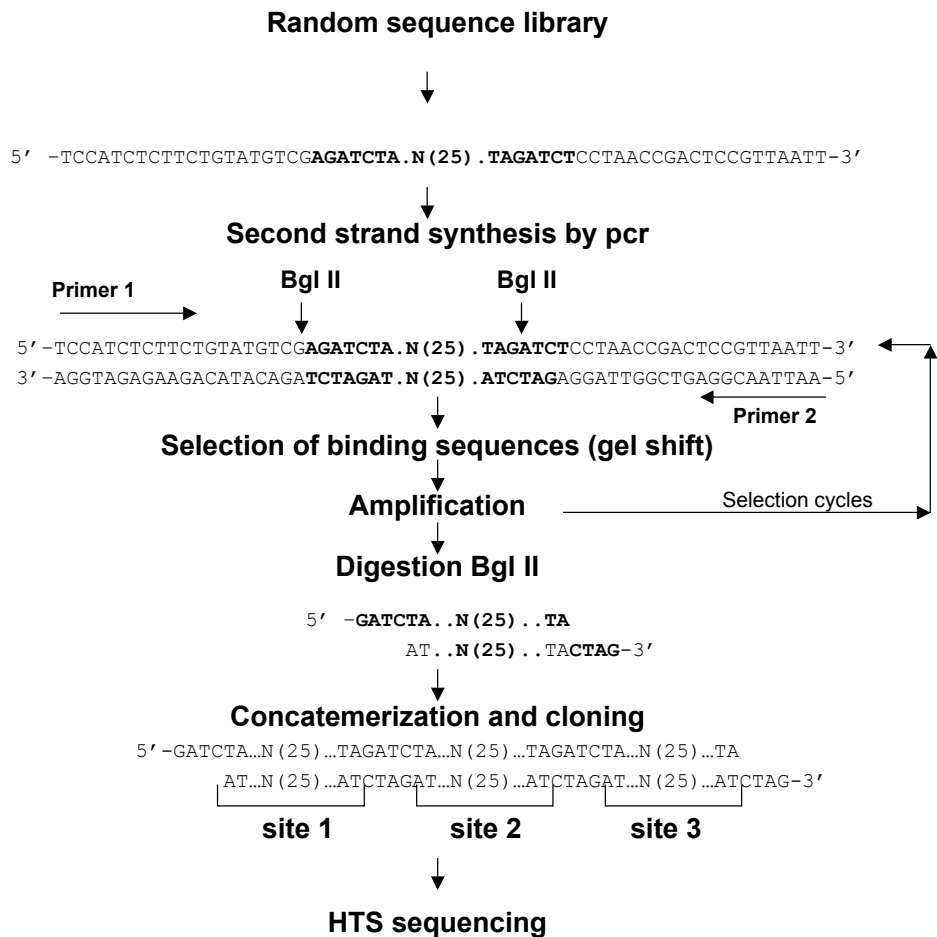
1. Choice of a quantitative predictive model for representing the binding specificity. Our choice: **a profile-HMM**
2. Choice of an experimental method to generate data for estimating the model parameters. Our choice: **a SELEX experiment**
3. Choice of a machine learning algorithm to estimate the model parameters from the data. Our choice: **the Baum-Welch HMM training algorithm**
4. Validation of the approach and optimization of the experimental parameters by a computer simulation of step 2 and 3
5. Adjustment of experimental protocol to produce the necessary data as suggested by the computer simulation
6. Generation of the experimental data
7. Building a binding site model from the data
8. A posteriori validation of the model by cross-validation and comparison with independent experimental results

Old CTF/NFI Binding Site Profile

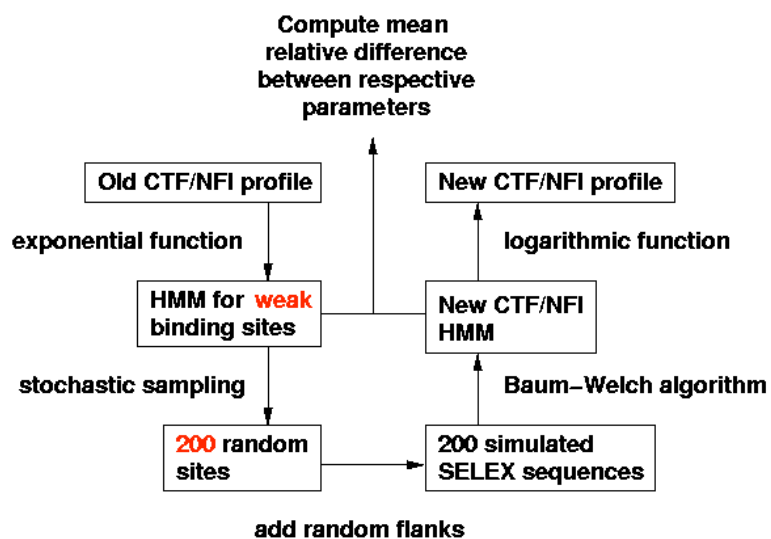


Example: TGGGCATATAGCCAC

Score: $10 - 1 + 10 + 10 + 10 + 0 + 10 + 10 + 10 + 10 + 9 = 88$



Computer Simulation Experiment:



Repeat same experiment for medium and strong binding sites and for smaller and bigger training set sizes.

Supplementary Table 1. Average log frequency error in profiles trained from various simulated training sets

Training set size:	20	50	100	200	500	1,000	2,000	5,000	10,000
Affinity (average binding score)									
Low binding affinity (70.4)	0.80	0.46	0.41	0.32	0.20	0.10	0.08	0.07	0.06
Medium binding affinity (81.6)	1.08	0.76	0.59	0.33	0.26	0.26	0.14	0.08	0.06
High binding affinity (95.5)	2.01	1.58	1.15	0.82	0.61	0.41	0.27	0.16	0.10

The effects of the training set parameters (number and affinity range of the sequences) on the accuracy of the reconstructed binding site models were determined as follows: Samples of 20 to 10,000 DNA sequences were generated from the frequency models for low, medium or high affinity binding sites shown in Figure 1C. New models were trained from these computer-simulated data sets using the Baum-Welch HMM training algorithm (see Experimental Protocol). The error rate of the re-estimated models is expressed as the average difference between the decimal logarithms of the original and corresponding re-estimated base frequencies. Each number reflects the mean of three independent simulation experiments.

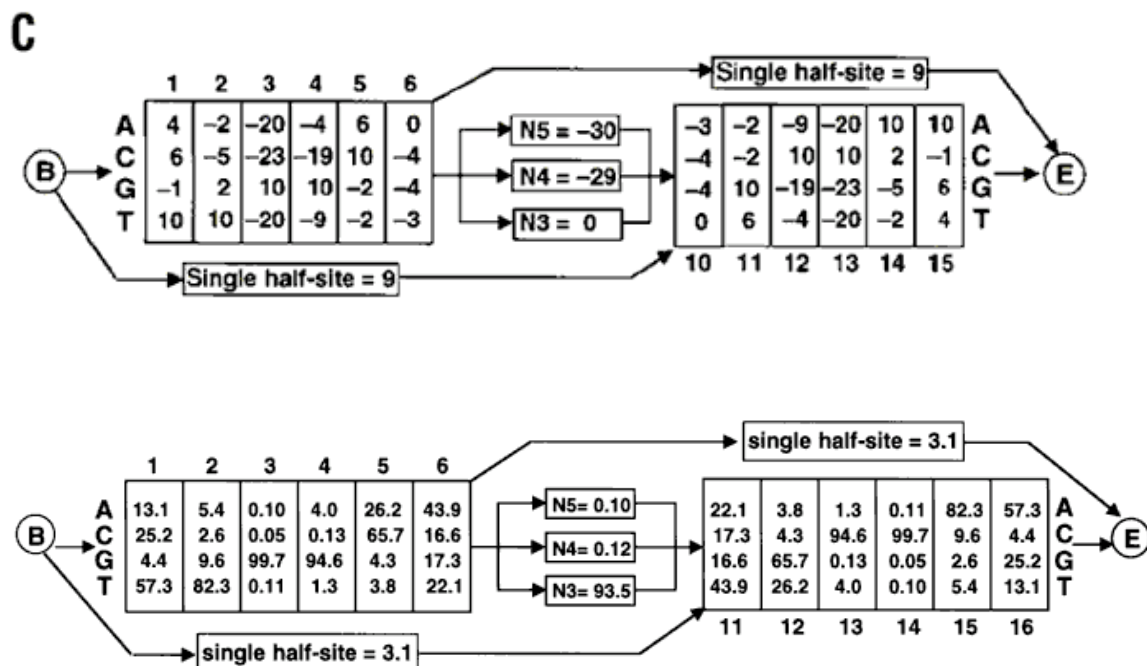
Adjustments of the SELEX Protocol in Response to the Results of the Simulations

Data required for reaching an average relative error of less than 10% in a binding site model according to the computer simulations:

2000 – 5000 low or medium affinity binding sites.

Necessary adjustments of the SELEX protocol:

1. Control of the stringency of the binding conditions via a radio-labeled non-amplifiable oligonucleotide of known affinity present in the binding reaction mixture.
2. Reduction of sequencing costs by multimerization of the binding sites before cloning using an enzymatic protocol borrowed from SAGE analysis.

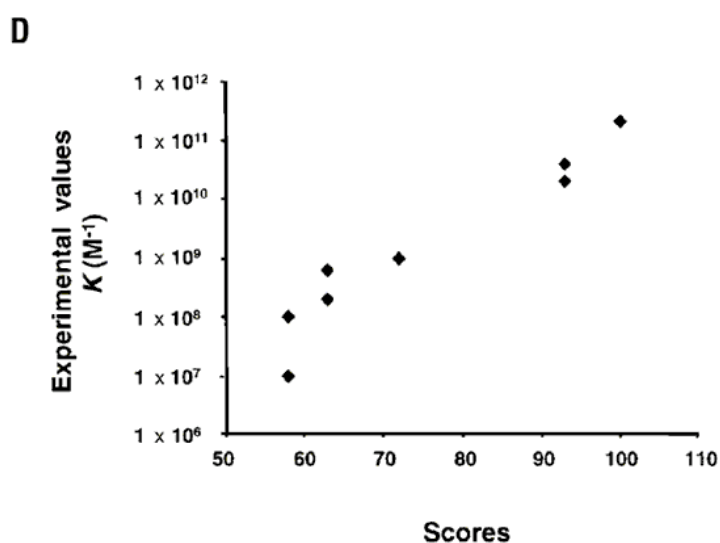


Supplementary Figure 2. Frequency model (HMM) for the CTF/NFI binding sites present in the Selex3 population.

		Subset 1						Subset 2							
		1	2	3	4	5	6	1	2	3	4	5	6		
A		12.9	5.5	0.11	4.3	26.6	43.9	13.3	5.4	0.2	3.7	25.8	43.9	A	
C		25.5	2.3	0.09	0.18	65.1	16.8	24.8	3.0	0.1	0.11	66.2	16.5	C	
G		4.6	9.4	99.6	93.9	4.4	17.3	4.3	10	99.5	95	4.2	17.4	G	
T		57	82.8	0.16	1.6	3.8	22.1	57.5	81.6	0.17	1.1	3.8	22.2	T	
		Left half-site				3.23		Left half-site				2.65			
		Spacing N3				93.32		Spacing N3				94.54			
		Spacing N4				0.12		Spacing N4				0.11			
		Spacing N5				0.16		Spacing N5				0.05			
		Right half-site				3.23		Right half-site				2.65			

Supplementary Figure 3. Cross-validation of the accuracy of the new profile. To assess the reproducibility of the HMM training step, we split the Selex3 sequence collection into two subsets of equal size and trained two independent models. From the previous computer simulations we expected an average relative difference between corresponding base frequency estimates of about 10%. In good agreement with this projection, we observed an average difference of 7.1% between the two models.

Quality Assessment of the New Model: Comparison of Predicted Binding Scores with *in vitro* measured Binding Constants



Data from Meisterernst et al. (1988). Nucl. Acids Res. 16, 4419-4435