IN SILICO PREDICTION OF *CIS*-REGULATORY ELEMENTS

ALBIN SANDELIN

KAROLINSKA INSTITUTET

STOCKHOLM 2004

From the Center for Genomics and Bioinformatics
Karolinska Institutet, Stockholm, Sweden

IN SILICO PREDICTION OF *CIS*-REGULATORY ELEMENTS

Albin Sandelin

KONGL CAROLINSKA MEDICO CHIRURGISKA INSTITUTET

Stockholm 2004

All previously published papers were reproduced with permission from the publisher
"*In silico* prediction of *cis*-regulatory elements"
Printed by Akademitryck
Holms Gård 59098 Edsbruk, Sweden

ISBN: 91-7349-879-3

“Sometimes, MacLeod, not even the sharpest blade is enough”

Ramirez, Highlander

ABSTRACT

As one of the most fundamental processes for all life forms, transcriptional regulation remains an intriguing and challenging subject for biomedical research. Experimental efforts towards understanding the regulation of genes is laborious and expensive, but can be substantially accelerated with the use of computational predictions. The growing number of fully sequenced metazoan genomes in combination with the increasing use of high-throughput methods such as microarrays has increased the necessity of combining computational methods with laboratorial. Computational '*in-silico*' methods for the prediction of transcription factor binding sites are mature, yet critical problems remain unsolved. In particular, the rate of falsely predicted sites is unacceptably high with current methods, due to the small and degenerate binding sites targeted by transcription factors. In addition to the false prediction rate, this restriction limits the ability of pattern discovery algorithms to find mediating binding sites in promoters of co-expressed genes. The latter problem constitutes a bottleneck when analyzing regulatory sequences in complex eukaryotes, as regulatory sequences generally are spread over extended genomic regions.

This thesis describes the development of algorithms and resources for transcription factor binding site analysis in addressing:
site prediction, where a model describing the binding properties of a transcription factor is applied to a sequence to find functional binding sites
pattern discovery, where over-represented patterns are sought in sets of promoters

Initially, an open-access database (JASPAR) was created, holding high quality models for transcription factor sites. The database formed part of the foundation for the subsequent project (ConSite), where a set of methods were developed for utilizing cross-species comparison in binding site prediction (*'phylogenetic footprinting'*) to enhance predictive selectivity. In this study, we could show that ~85% of false predictions were removed when only analyzing promoter regions conserved between human and mouse.
The current statistical framework for modeling binding properties of transcription factors is inadequate for some regulatory proteins, most notably the medically important nuclear hormone receptors. A Hidden Markov Model framework capable of both predicting and classifying nuclear hormone receptor response elements was developed. In a case study, we showed that nuclear receptor genes have a high potential for cross-or auto regulation using the pufferfish genome as a predictive platform.
Pattern discovery in promoters of multi-cellular eukaryotes is limited by the low strength of patterns buried in extended genomic sequence. Methods for improving both sensitivity and evaluation of resulting patterns were developed. We showed that comparison of newly found patterns to databases of experimentally verified profiles is a meaningful complement to other means to evaluate patters. Furthermore, we showed that structural constraints that are shared by families of transcription factors can be integrated as prior expectations in pattern finder algorithms for a significant increase in sensitivity.

ORIGINAL PUBLICATIONS

I

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. & Lenhard, B.
JASPAR: an open-access database for eukaryotic transcription factor binding profiles.
Nucleic Acids Res **32**, D91-4 (2004)

II

Lenhard, B., **Sandelin, A.**, Mendoza, L., Engstrom, P., Jareborg, N. and Wasserman, W. W.
Identification of conserved regulatory elements by comparative genome analysis.
J Biol, **2**, 13 (2003)

III

Sandelin, A., Wasserman, W. W. and Lenhard, B.
ConSite: web-based prediction of regulatory elements using cross-species comparison.
Nucleic Acids Res, (accepted) (2004)

IV

Sandelin, A. and Wasserman, W. W.
Prediction of Nuclear Hormone Receptor Response Elements.
submitted to *Mol. Endocrin* (2004)

V

Sandelin, A., **Höglund, A**., Lenhard, B. and Wasserman, W. W.
Integrated analysis of yeast regulatory sequences for biologically linked clusters of genes.
Funct. Integr. Genomics, **3**, 125-34 (2003)

VI

Sandelin, A. and Wasserman, W. W.
Constrained Binding Site Diversity within Families of Transcription Factors Enhances Pattern Discovery Bioinformatics.
J. Mol.Biol, **(accepted)**, (2004)

RELATED PUBLICATIONS

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., Yamanaka, I., Kiyosawa, H., Yagi, K., Tomaru, Y., Hasegawa, Y., Nogami, A., Schonbach, C., Gojobori, T., Baldarelli, R., Hill, D. P., Bult, C., Hume, D. A., Quackenbush, J., Schriml, L. M., Kanapin, A., Matsuda, H., Batalov, S., Beisel, K. W., Blake, J. A., Bradt, D., Brusic, V., Chothia, C., Corbani, L. E., Cousins, S., Dalla, E., Dragani, T. A., Fletcher, C. F., Forrest, A., Frazer, K. S., Gaasterland, T., Gariboldi, M., Gissi, C., Godzik, A., Gough, J., Grimmond, S., Gustincich, S., Hirokawa, N., Jackson, I. J., Jarvis, E. D., Kanai, A., Kawaji, H., Kawasawa, Y., Kedzierski, R. M., King, B. L., Konagaya, A., Kurochkin, I. V., Lee, Y., Lenhard, B., Lyons, P. A., Maglott, D. R., Maltais, L., Marchionni, L., McKenzie, L., Miki, H., Nagashima, T., Numata, K., Okido, T., Pavan, W. J., Pertea, G., Pesole, G., Petrovsky, N., Pillai, R., Pontius, J. U., Qi, D., Ramachandran, S., Ravasi, T., Reed, J. C., Reed, D. J., Reid, J., Ring, B. Z., Ringwald, M., Sandelin, A., Schneider, C., Semple, C. A., Setou, M., Shimada, K., Sultana, R., Takenaka, Y., Taylor, M. S., Teasdale, R. D., Tomita, M., Verardo, R., Wagner, L., Wahlestedt, C., Wang, Y., Watanabe, Y., Wells, C., Wilming, L. G., Wynshaw-Boris, A., Yanagisawa, M., Yang, I., Yang, L., Yuan, Z., Zavolan, M., Zhu, Y., Zimmer, A., Carninci, P., Hayatsu, N., Hirozane-Kishikawa, T., Konno, H., Nakamura, M., Sakazume, N., Sato, K., Shiraki, T., Waki, K., Kawai, J., Aizawa, K., Arakawa, T., Fukuda, S., Hara, A., Hashizume, W., Imotani, K., Ishii, Y., Itoh, M., Kagawa, I., Miyazaki, A., Sakai, K., Sasaki, D., Shibata, K., Shinagawa, A., Yasunishi, A., Yoshino, M., Waterston, R., Lander, E. S., Rogers, J., Birney, E. and Hayashizaki, Y.
Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs.
Nature, **420**, 563-73 (2002)

Wasserman, W. W. and **Sandelin, A**.
Applied Bioinformatics for the Identification of Regulatory Elements.
Nat Rev Genet, **5**, 276-287 (2004)

CONTENTS

THE BIOLOGY OF GENE REGULATION 13
Laboratorial approaches to study gene regulation 15

THE COMPUTATIONAL BIOLOGY OF GENE REGULATION 16
Modeling of TF binding properties 16
Prediction of TFBS in genomic sequences using models describing TF binding specificity 19
Discovering motifs in promoters from co-regulated genes 19
Discovery of TFBS patterns in genomic sequences using pattern finding 20

PRESENT INVESTIGATION 22
Paper I: JASPAR: an open-access database for eukaryotic transcription factor binding profiles 23
Paper II: Identification of conserved regulatory elements by comparative genome analysis 24
Paper III: ConSite: web-based prediction of regulatory elements using cross-species comparison 24
Paper IV: Prediction of Nuclear Hormone Receptor Response Elements 26
Paper V: Integrated analysis of yeast regulatory sequences for biologically linked clusters of genes 28
Paper VI: Constrained Binding Site Diversity within Families of Transcription Factors Enhances Pattern Discovery Bioinformatics 30

PERSPECTIVES 32

ACKNOWLEDGEMENTS 35

REFERENCES 38

LIST OF ABBREVIATIONS

bp	Base Pairs
ChIP	Chromatin Immunoprecipitation
CRM	Cis-Regulatory Module
DNA	Deoxyribonucleic Acid
EM	Expectation Maximization
HMM	Hidden Markov Model
IC	Information Content
IUPAC	International Union of Pure and Applied Chemsitry
NHR	Nuclear Hormone Receptor
NR	Nuclear Receptor
PCR	Polymerase Chain Reaction
PFM	Position Frequency Matrix
PWM	Position Weight Matrix
RNA	Ribonucleic Acid
SELEX	Systematic Evolution of Ligands by Exponential Enrichment
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TSS	Transcription Start Site

FOREWORD

This thesis describes research in computational biology. Computational biology is an interface between biochemistry, biology, computer science, statistics and related fields. As in all cross-scientific disciplines, communication between scientists can be difficult due to the complexity and inconsistency of the vocabulary in each field. Depending on the background of the reader, certain passages will be harder to read. Most of the text requires a basic to advanced understanding of life science and/or computer science and an interest in both fields.

In this thesis, the major emphasis lie on biology, as the focus is a biological set of problems. Fundamental computer science concepts such as dynamic programming and graph theory, although used extensively, will not be explained in depth. Excellent textbooks in both molecular/cellular biology[1,2] and computer science[3] can be recommended for a more comprehensive coverage of both fields.

THE BIOLOGY OF GENE REGULATION

"I will tell you what knowledge is. To know when you know something and to know when you do not, that is knowledge"

— Konfucius

One of the most central properties shared by all life forms is the ability to store and propagate information: a necessity for evolution[4]. The series of discoveries of how diverse life forms can be coded into a string of chemical entities named nucleotides[5-8] gave birth to the scientific discipline of molecular biology, which again is changing with the sequencing of the genetic material of vertebrates[9-11] nematodes[12]plants[13], insects[14], fungi[15] prokaryotes[16,17]and archea[18].

Nucleic acids in four variations (coded A,C,G,T) can be combined linearly to form a string of deoxyribonucleotides, DNA. Cellular DNA is typically organized in two intergripping chains, creating the iconographic double helix. In multi-cellular organisms, each DNA molecule, wrapped up with associated proteins in the cell nucleus, forms the chromosomes.

Triplets of nucleotides code for amino acids, the building blocks of proteins. Regions of such triplets form instructions on how to construct a certain protein: a protein-coding gene. The information flow from gene to protein is divided into two processes (**Figure 1**):
Transcription, where a protein complex, RNA polymerase II (PolII) reads the gene nucleotide sequence and polymerizes a single-stranded RNA sequence (similar to single-stranded DNA).
Translation, where the RNA-strand is translated to a protein in the ribosome machinery.

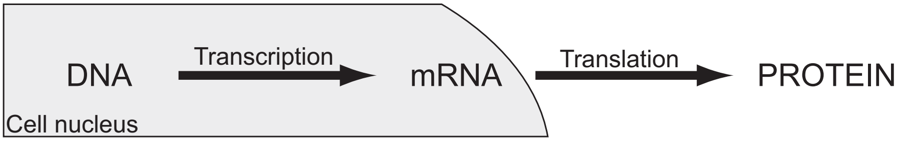


Figure 1
Information flow in eukaryotic cells: from DNA via mRNA to proteins.

Most proteins are only used at specific time points (for instance in a certain phase in development or when the cell reacts to environmental stimuli) or in specialized cells. It follows that most genes are inactive most of the time, and that cells require a mechanism to determine activation. Gene regulation is thus one of the most fundamental mechanisms for any living organism. Regulation can be achieved in many stages: transcriptional, translational and post-translational processes can be modified[1]. This work will solely address the first form, transcriptional regulation. In prokaryotes, transcriptional regulation constitutes the dominating type of gene control[20]. In eukaryotes, the regulation of transcription is the basis for both cellular response to external stimuli (for instance hormones or neurotransmittors[2]) and development[2,21,22]. It follows that we cannot understand cellular biology without a fuller understanding of transcriptional regulation. Therefore, a central goal in cellular biology is to produce a comprehensive map describing the regulatory networks of cells[23].
In an over-simplified model of the mechanism of transcriptional regulation, the PolII complex will bind to the transcriptional start site (TSS) with the help of the DNA-binding TBP-protein and start transcription. The complex must be stabilized by additional

DNA-binding proteins, called general transcription factors, and other interacting proteins (cofactors)[24]. In addition, for precise transcriptional control, regulatory DNA-binding proteins termed transcription factors (TFs) are required to bind to specific sequences in the DNA (transcription factor binding sites, TFBS)[1,2]. TFs can either occur in the near proximity of the TSS, within introns or in distal locations (up to hundreds of kilobasepairs away from the protein coding sequence)[25]. TFs often operate in 'modules' – a set of TFs binding relatively close to each other, presumably interacting directly or indirectly at protein level[25,26]. At a higher level, the global structure of DNA – the chromatin superstructure – has a fundamental role in the regulation of genes[25,27,28]. DNA is wrapped around histone proteins, forming nucleosomes, which in turn are packed in more complex structures[29] (**Figure 2**). There is no widely accepted model describing the positioning of nucleosomes in the genomic DNA, nor the dynamic interplay between TFs and chromatin[29-32].

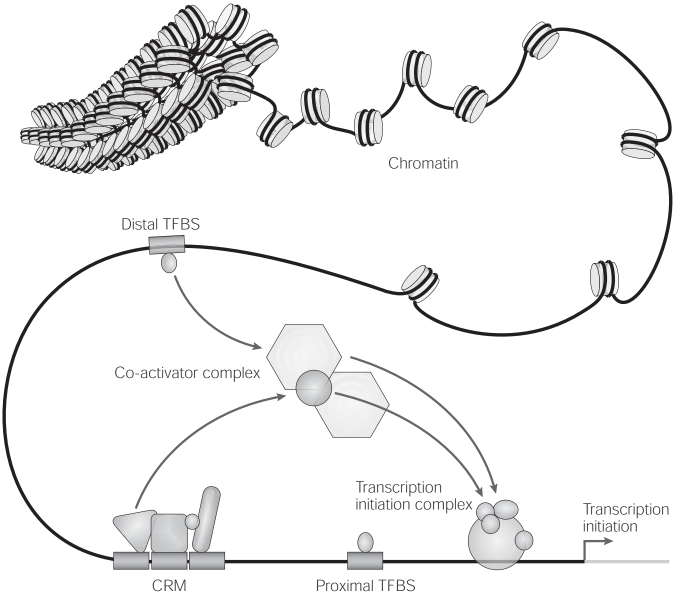


Figure 2
Components of transcriptional regulation. Transcription factors (TFs) bind to specific sites (transcription factor binding sites; TFBS) that are either proximal or distal to a transcription start site. Sets of TFs can operate in functional *cis*-regulatory modules (CRMs) to achieve specific regulatory properties. Interactions between bound TFs and cofactors stabilize the transcription-initiation machinery to enable gene expression. The regulation that is conferred by sequence-specific binding TFs is highly dependent on the three-dimensional structure of chromatin. Figure from Wasserman, W. W. & Sandelin, A. Applied Bioinformatics for the Identification of Regulatory Elements. *Nat Rev Genet* **5**, 276-287 (2004)

Over evolution, a small number of different protein templates have evolved for mediating sequence-specific binding to DNA. Examples include the Zn-finger, the Helix-loop-helix and the Forkhead structures[33,34]. While the number of structural classes is small, the number of TFs in each class is considerable. For instance, Zn-finger genes are one of the most frequently occurring gene types in the human genome[9].

Sequence specific binding is generally achieved by the insertion of one or more protein α-helices of the TF into the DNA major groove, where hydrogen bonds are formed between specific amino acid residues and nucleotides. This process often involves homo- or heterodimerization of TFs, presumably to achieve higher stability and selectivity in binding[33,34]. Sites bound by TFs are short, usually ranging from 5-12 bp. In addition, most TFs tolerate considerable variation in their targeted sites – a principal difference to many other well-studied DNA-binding proteins (e.g. restriction enzymes)[35].

Laboratorial approaches to study gene regulation

Elucidation of the transcription factors responsible for the activity of each gene is a primary goal in cell biology. A variety of laboratorial techniques has been developed to this end. Methods range from large-scale measurements of thousands of genes to the study of individual basepair mutations in a TFBS.

Although no laboratorial investigations were undertaken in this thesis project, the data sources underlying both databases and algorithmic developments in this work originate from laboratorial investigations. For this reason, a brief introduction to some commonly used laboratorial approaches for studying gene regulation is included:

In-situ labeling: Probes consisting of oligonucleotides can be constructed to locate both genes in chromosomes and expressed mRNA in cells. In short, a labeled probe is introduced in cells and will hybridize with exposed complementary sequences (for instance an expressed mRNA). This method has been used to study the distribution of specific mRNAs in cells in tissues[36].

Micro-arrays: Instead of studying the expression of a single gene, micro-array methods measure the expression of thousands of genes simultaneously. DNA oligonucleotides or cDNA probes are fixed in spots at a glass slide surface. Simplified, samples of mRNA from cells are labeled with fluorescent markers and exposed to the array. Expressed mRNAs will hybridize with probes on the array and produce a signal[37-39]. In this way, genes that have similar expression profile over many samples can be identified.

Reporter construct studies: In-situ labeling and micro-array studies can indicate where and when a gene is expressed, but not how the regulation takes place. The locations of binding sites of transcription factors are commonly identified with reporter constructs, where the target gene is fused with a signal molecule, such as GFP. As the expression level of the gene can be measured, systematic deletion of promoter regions can identify regions harboring functional binding sites. Subsequent in-depth mutations and deletions of single nucleotides in putative binding sites can confirm a functional binding site.

***In vitro* site selection:** The compilation of a significant number of binding sites for a given transcription factor using reporter constructs is possible, but expensive and time-consuming. If we are interested only in the type of sites preferentially bound by the factor, *in vitro* site selection assays (often called SELEX) can be used[40]. In such assays, a TF is initially exposed to a pool of random DNA oligomers. The subset of oligomers bound by the TF are isolated and amplified by PCR, to form a new pool of oligonucleotides. The process is iterated to identify critical properties of TF binding sites. However, sites identified in SELEX studies might not be fully representative of functional sites[41].

Chromatin immunoprecipitation (ChIP): A TF bound to DNA *in vivo* can be covalently cross-linked to its cognate binding site, using formaldehyde. Isolated DNA is then mechanically broken into smaller fragments, and exposed to antibodies capable of recognizing the bound TF. Using this technique, *in vivo* binding sites and proximal regions can be purified and sequenced[42,43].

THE COMPUTATIONAL BIOLOGY OF GENE REGULATION

"Faithless is he that says farewell when the road darkens, "
— J. R. R. Tolkien

The importance of unearthing the regulatory mechanisms of genes has already been discussed. As indicated above, experimental techniques towards this end exist, but are laborious and expensive, in particular in the many cases where no prior information is available. Experimental elucidation of functional TFBS can be substantially accelerated with the use of computational predictions[44].

Modeling of TF binding properties

The prediction of TFBS and the connected modeling of TF binding properties is one of the most well-studied problems in computational biology. A brief introduction to the established methods is necessary at this point. For in-depth reviews, see[44-46].

As TFs generally tolerate some variability in their target binding sites, a model describing TF binding properties must be trained on multiple functional sites. Collections of sites for a TF are typically retrieved from functional investigations or *in vitro* site selection assays[40]. In most cases, an alignment of such sites forms the input for training the model.

Consensus models

Consensus sequences are commonly used in molecular biology for describing the static binding properties of restriction enzymes[35] and general transcription factors such as the TBP[47]. A set of known binding sites are aligned, and a consensus nucleotide symbol is assigned to describe the nucleotide composition in each column of the alignment, usually following IUPAC conventions (**Box 1**). The disadvantage with this approach is that a single symbol cannot quantitatively describe the nucleotide distribution within a column. On the other hand, consensus sequences are suitable for fast visual representation.

Profile models

A quantitative matrix (or profile) model can be constructed by simply counting occurrences of nucleotides in each alignment column. A matrix built out of nucleotide counts in this way is referred to as a position frequency matrix (PFM). A normalized PFM (that is, each column summing to 1) can be viewed as a table of probabilities for observing certain nucleotides in a given position. The chance of observing a particular site is then the product of the relevant cell probabilities, taken from each column.

PFMs are often visualized graphically as sequence logos[48]. In this representation, the conservation in each column is calculated in terms of information content (bits)[49] (**Box1, Box2**). Each nucleotide occurrence is then scaled with the total information content in that position. Sequence logos enable fast visual assessments of pattern characteristics, and constitute a significantly richer description than consensus sequences.

For scoring purposes, a PFM is converted to a position weight matrix (PWM), which essentially is a log-odds representation of the PFM (**Box 2**)[46,50]. In this process, a pseudo-count is added to each cell of the PFM, to correct for small samples of binding sites. The choice of pseudo-count function varies between different research groups[51]; in this work, it is simply the square root of the number of contributing sites.

Box 1 | Building models for predicting transcription factor binding sites

The first step towards building models for predicting TF binding sites involves collecting data. To illustrate the process we use the transcription factor MEF2 as an example.

Data collection

A set of experimentally validated binding sites for MEF2 were collected from the literature and aligned (a). The quality of the collection of binding sites has a strong impact on the downstream models for predicting additional sites. Note the diversity between the sites; for instance, only 50% of the nucleotides are identical between sites 1 and 8.

Model building

Consensus sequence model: A consensus sequence is defined by selecting a degeneracy nucleotide symbol for each position (column) in the alignment (b). Unusual binding sites can have an extreme effect on the consensus (e.g. site 8).

Position Frequency Matrix (PFM): To more accurately reflect the characteristics at each position, a matrix containing the number of observed nucleotides at each position is created (c). For instance, the first column in the alignment (a) consists of 0 A:s, 3 C:s, 2G:s and 3 T:s, making the corresponding first matrix column{0,3,2,3}.

Position Weight Matrix (PWM): The frequency matrix is usually converted to a PWM using a formula (Box2) that converts normalized frequency values to a log-scale (d). Using a matrix model, one can generate a quantitative score for any DNA sequence by summing the values corresponding to the observed nucleotide at each position (e). For large and representative collections of binding sites, the scores are proportional to binding energies51.

Sequence logo: The specificity in each column of the alignment can be measured in terms of information content. A sequence logo scales each nucleotide by the total bits of information times the relative occurrence of the nucleotide at the position (f). Sequence logos enable fast and intuitive

a

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

b

B R M C W A W H R W G G B M

Consensus sequence

c Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

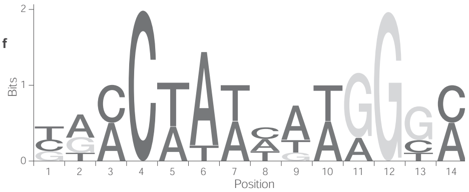
d Position weight matrix (PWM)

A	-1.93	0.79	0.79	-1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	-1.93	-1.93	0.79
C	0.45	-1.93	0.79	1.68	-1.93	-1.93	-1.93	0.45	-1.93	-1.93	-1.93	-1.93	0.00	0.79
G	0.00	0.45	-1.93	-1.93	-1.93	-1.93	-1.93	-1.93	0.66	-1.93	1.30	1.68	1.07	-1.93
T	0.15	0.66	-1.93	-1.93	1.07	0.66	0.79	0.00	0.00	0.79	-1.93	-1.93	-0.66	-1.93

e Site scoring

0.45	-0.66	0.79	1.68	0.45	-0.66	0.79	0.45	-0.66	0.79	0.00	1.68	-0.66	0.79
T	T	A	C	A	T	A	A	G	T	A	G	T	C

= 5.23, 78% of maximum



Box 2 | Formulae linked to methods for the analysis of regulatory sequences

Corrected probabilities of observing a given nucleotide can be calculated using equation 1.

Corrected probability calculation:
$$p(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{b' \in \{A,C,G,T\}} s(b')} \qquad (1)$$

$f_{b,i}$ = counts of base b in position i; N = number of sites; $p(b,i)$ = corrected probability of base b in position i; $s(b)$ = pseudocount function

A position weight matrix (PWM) is constructed by dividing the nucleotide probabilities in (a) by expected background probabilities and converting the values to a log-scale (see equation 2).

PWM conversion:
$$W_{b,i} = \log_2 \frac{p(b,i)}{p(b)} \qquad (2)$$

$p(b)$ = background probability of base b; $p(b,i)$ = corrected probability of base b in position i; $W_{b,i}$ = PWM vaue of base b in position i

The quantitative PWM score for a putative site is the sum of the PWM values for each nucleotide in the site (see equation 3).

Evaluation of sequences:
$$S = \sum_{i=1}^{w} W_{l_i,i} \qquad (3)$$

L_i = the nucleotide in position i in an input sequence; S = PWM score of a sequence; w = width of the PWM

Probability values (1) can be used to determine the total information content (in bits) in each position (see equation 4).

Information content calculation:
$$D_i = 2 + \sum_b p_{b,i} \log_2 p_{b,i} \qquad (4)$$

D_i = information content in position i; $p(b,i)$ = corrected probability of base b in position i

Figures from Wasserman, W. W. & Sandelin, A. Applied Bioinformatics for the Identification of Regulatory Elements. *Nat Rev Genet* **5**, 276-287 (2004)

A TFBS is evaluated by summing the relevant PWM cell values, analogous to the calculation of the probability of observing the site (as described above). For longer sequences, a PWM is slid over the sequence to evaluate all possible TFBS start locations. It has been shown that a PWM score is directly proportional to the binding energy of the TF-DNA interaction[46,52]. Thus, the PWM model can both be viewed as a statistical and energy-based model.

As all possible sites start locations will generate a score, some cutoff is needed to distinguish likely sites from the background. There is some controversy in the field if such a cutoff should be a static score, relative (fraction of score range) or probability-based (i.e. how likely is this score)[53]. While the static score has relevance if we are interested in the strength of the interaction, such scores cannot readily be compared between different factors, as the score ranges are different. Therefore, relative or probability-based cutoffs are commonly used. The advantages of probability-based cutoffs are immediately recognizable: we get an assessment how likely it is to observe a certain score. However, it is not given that the site with the lowest p-value is the most likely candidate for being a functional site, nor that the nucleotide distribution in the promoter flanking the site has a direct impact on the binding thermodynamics. On the other hand, this measure might indicate if there are other equally good sites competing for the limited number of TF molecules[53].

In this model framework, there is an implicit assumption that individual positions in the binding sites are independent (i.e. the nucleotide distribution in one position does not affect the distribution in another)[46,54]. In a few cases, sufficient data has been available to assess the validity of that assumption by building models incorporating high-order interactions[55,56]. While the predictive specificity increases with such models, the improvement is not dramatic. Thus, given the sparse binding site collections available, the profile model is in this respect an adequate framework[54].

Hidden Markov Models

A potential limitation with profile models is the inability to model insertions and deletions[50], thus, the variable spacing between half-sites that is observed for some TFs cannot be incorporated directly into the model. In those cases, a more flexible model framework is required. This problem is analogous to describing protein domains based on gapped multiple alignments. The Hidden Markov Model framework has been used extensively in computational biology to address this and other problems[50,57-59]. Several reviews and textbooks describing the theory and utility of HMMs have been published[50,57]. In the field of gene regulation bioinformatics, the HMM framework can be viewed as an extension of the profile model, enabling a richer description of sequence characteristics, including variable spacing and higher-order interactions.

Briefly, an HMM model consists of a set of "states", where each state can emit symbols (for instance nucleotides) based on some defined probability distribution. The *emission probability* for a certain symbol is specific for each state. States are connected to one or several other states in a chain-like structure (the chain configuration is usually manually chosen to fit with the problem at hand). The probability of moving from one state to another is termed a *transition probability*. Generally, a specific start and end state is defined, which do not emit symbols. Any specific path through the states producing a given sequence of symbols will have a defined probability (effectively the product of all emission probabilities for respective symbol and all transitions probabilities for each move between states). The Viterbi algorithm calculates the optimal (that is, the most probable) path through the states starting at the start state and ending at the end state that produces a given sequence[50]. This algorithm is usually employed in classification problems, where states are labeled to represent certain biological properties. The total probability of the model emitting a certain sequence can be calculated using the Forward algorithm, which sums the probability of all possible routes producing the same sequence[50].

Prediction of TFBS in genomic sequences using models describing TF binding specificity

A profile model, as described above, can be used to predict putative binding sites in genomic sequences for a certain TF. Both the sensitivity and selectivity are affected by the choice of cutoffs. Two key observations have emerged from previous research:

a) A typical binding profile produces, on average, one prediction per 500-1500 bp, depending on settings and model characteristics. This high rate of predictions is biologically unrealistic[60].

b) A significant portion (95%) of sites predicted as above are potential *in vitro* binding sites (although not necessarily needed for regulation of the target gene)[61]. This implies that the models employed are adequate descriptions of *in vitro* binding to DNA.

These two observations demonstrate that, while the models can describe DNA binding properties of TFs adequately, all information required to distinguish a functional site in genomic DNA is not contained within the binding site in itself or the interface between TF and DNA. Thus, in practical terms, almost all predictions made using solely this approach will be non-functional. This statement will be referred to as the *futility theorem* [62].

If we hold the futility theorem for true, the inevitable question is: where is the rest of the required information? Currently, our understanding of the transcriptional process is far from complete[24]. However, it is clear from the body of research on transcriptional regulation that many aspects of the nuclear environment are not incorporated in the profile model, for instance:

Complexity of nuclear DNA: When we scan a single sequence with a profile model, we implicitly assume all regions of analyzed DNA to be equally accessible. However, we know DNA in the nucleus to be involved in an immensely complex dynamic chromatin superstructure[19,63]. We expect that a given region of DNA at many time-points simply is not accessible for transcription factors. It is likely that the regulation of the DNA superstructure is as significant in the regulation of a gene as the actual transcription factors[64].

Modularity of TFs: A single TF is rarely solely responsible for the regulation of a gene. For instance, we know that many tissue-specific genes are regulated by modules of TFs[23,65]. Incorporation of several profile models in a prediction increases the signal strength considerably[66,67]. Modeling of cis-regulatory TF modules is an active subfield of TFBS prediction research[44], but is not the focus in this work.

Discovering motifs in promoters from co-regulated genes

One of the practical limitations with the approaches described above is that a model must be constructed before scanning a sequence for putative sites. Using that approach, we cannot discover sites for TFs for which we have no models. *Pattern discovery* aims to find statistically over-represented patterns in a set of sequences[46]. Applied to promoter analysis, the input sequences are promoters of genes suspected to be co-regulated (regulated by at least one common TF), while over-represented sub-sequences are hypothetical TFBS. Pattern discovery algorithms are divided into two categories, based on underlying methodology:

Word-based: where the occurrence of each 'word' of nucleotides of a certain length is counted and compared to a background distribution[68-70]. An advantage with these methods is that they are comparatively fast and the statistical background well understood. On the other hand, a word-based description of TF binding properties is often inadequate, as TFs are known to tolerate variations within binding sites. Word- based methods have not been used extensively in this work but are presented for reference.

Probabilistic: where the most over-represented pattern (a matrix description) is sought, using random selection at some point in the algorithm. The problem of finding an optimal pattern (and thus evaluate evaluating all possible solutions) is equivalent to finding an optimal local multiple alignment, which is proven NP-complete[71 *]. Therefore, algorithms that can identify over-represented patterns more efficiently are required. Gibbs Sampling[72] and the related Expectation Maximization (EM)[73,74] are the most popular of such probabilistic algorithms in the field. A brief overview of a basic Gibbs sampling algorithm (as described in[72]) is necessary for understanding details in papers IV and V. The central concept behind both Gibbs Sampler and EM methods is to iteratively evolve an initial random pattern into a more specific one.

As input, we have a set of nucleotide sequences *S*, and a proposed width of the sought pattern. Initially, one starting point for a 'site' is randomly selected on each sequence in *S*. One sequence, *Z,* is then removed from the set. A profile model, similar to a PWM, is built from the sites found at the starting points in the remaining sequences (using both a nucleotide background distribution and pseudo-counts). The PWM is slid over the removed sequence *Z*, evaluating each possible 'site' location. One of the sites in *Z* is chosen randomly from a distribution that is proportional to the scores of the sites. In other words, high-scoring sites are more likely to be chosen than low-scoring sites. The sequence *Z* is then incorporated in *S*, with the annotated starting point (as chosen above). The procedure is iterated by choosing another *Z*, until either a) the pattern or the pattern strength does not change between iterations, or b) a set maximum of iterations have been reached. The related EM methods are based on a similar algorithm, but always take the highest scoring site in *Z* instead of choosing it from some distribution.

Discovery of TFBS patterns in genomic sequences using pattern finding

Pattern finding algorithms have been applied to various biological sequence data, for instance in the identification of protein domains in amino acid sequences[72]. Discovery of regulatory patterns in non-coding DNA presents specific challenges, related to the size of promoter sequences and the limited information contained in patterns[75]. In other words, the concepts underlying the futility theorem are equally true for pattern finding. In order to apply pattern finders to longer promoter lengths, two key problems must be addressed:

* NP-complete is a computer science term that refers to problems that are computationally intractable

Pattern drowning: Pattern finding algorithms have been successfully utilized in bacteria and yeast to identify key regulators in biological systems, often in combination with micro-array data[70,76,77]. The major constraint for pattern finders consists of the limited information contained in a TF protein-DNA interface. When pattern finders are applied to longer promoter sequences (~>500bp), the lack of information results in an inability to find the sites forming the pattern. Since promoters of multi-cellular eukaryotes often span 1000 bp or more, this limitation is severe.
Biological relevance: Probabilistic algorithms in particular are prone to output patterns with limited biological relevance, albeit a high over-representation (for instance repeat regions). Certain progress has been made in this area; including advanced background models and the development of maximum *a posteriori* scores (MAP scores: the posterior probability of the alignment given the data[72]). MAP scores are however not perfect estimators of biological significance, as they are dependent on input sequence length, pattern width and number of promoters[78].

The two problems have been addressed in a variety of ways:
Improved background models: The background model used in early pattern finders assumed the genome to be composed of nucleotides drawn randomly from some distribution. It is clear that this description is an over-simplification. Genomes contain many distinctly non-random features that require higher-order models (for instance repeat regions[79] and CpG islands[80]). Various applications of pattern finders have proved that the incorporation of richer background models increases the chance of finding relevant patterns. Different solutions have been proposed: the popular MEME pattern finder uses background models based on scaled Dirichlet distributions[81,82], while the ANN-Spec program during execution finds patterns in a positive and a background set of sequences at the same time[83].
Cross-species comparison: It is often worthwhile to include evolutionary information in the pattern finding process. This can either be achieved by simply including the promoters from orthologous genes (two or more genes separated by speciation, sharing a common ancestor[84,85]) in the analysis, or restrict the search only to conserved regions. As in other forms of cross-species comparison, the choice of species and the related evolutionary divergence between the sequences is important. In particular, human-mouse comparisons have proven valuable[62].
Incorporation of pattern constraints: One of the reasons why pattern finders are challenged by TFBS discovery is the built-in assumption that all equally over-represented patterns are as likely to be functional. In past efforts, some constraints been incorporated into pattern finder algorithms, including restricting patterns to a subset of positions (based on the expectation that only a few positions within a site interacts with the TF)[62] and site palindromicity[86].

PRESENT INVESTIGATION

"What is this thing, anyway?" said the Dean, inspecting the implement in his hands. "It's called a shovel," said the Senior Wrangler. "I've seen the gardeners use them. You stick the sharp end in the ground. Then it gets a bit technical."

— *Terry Pratchett*

The long-term goal of gene regulation bioinformatics is to enhance promoter analysis, ideally to be comparable to experimental techniques in both sensitivity and selectivity. For clarity, this work addresses metazoan transcriptional regulation, with particular emphasis on multicellular eukaryotes.
As stated in the introduction, many new developments are needed, including;

- **compilation of curated model collections**
- **enhanced site detection methods (addressing the futility theorem)**
- **development of new model frameworks**
- **enhanced pattern discovery (addressing the pattern drowning problem)**

The publications presented address each of these aspects:

- A high-quality model collection was created (the JASPAR database , paper I)
- TFBS analysis using cross-species comparison was proven to remove ~85% of false predictions (papers II, III)
- A HMM model for nuclear hormone receptor response elements was developed, and revealed high cross-regulatory potential for nuclear receptor genes in the pufferfish (*Fugu rubripes*) genome (paper IV)
- Pattern comparison algorithms enable enhanced assessment of pattern finding results (paper V). The introduction of structural constraints in pattern finding algorithms increases sensitivity significantly (paper VI)

Paper I: JASPAR: an open-access database for eukaryotic transcription factor binding profiles

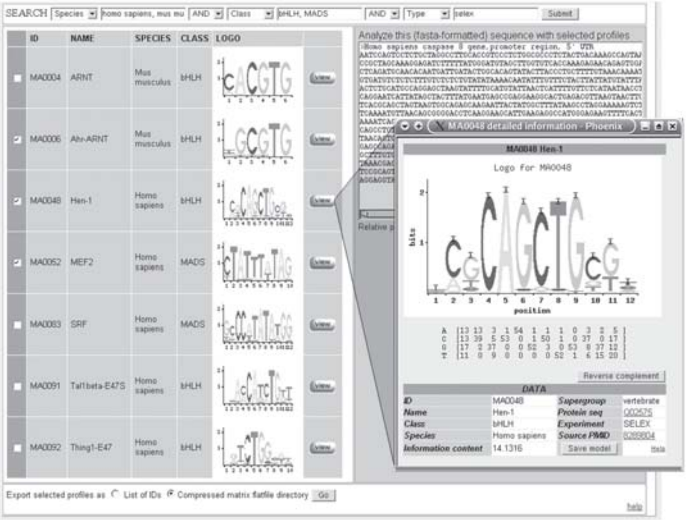
In the initial phase of the thesis project, it quickly became apparent that a high-quality collection of matrix models was needed by the group and in the field. The models in TRANSFAC[87] - the leading TF database, were at that time point not sufficiently curated for our needs and had a large amount of redundancy (many models describing the same factor).

Initially the dataset was collected with the primary purpose to serve as the basis in the construction of generalized profiles for TF classes (Paper V). For this reason, the aim was to obtain as good TF class coverage as possible.

Briefly, scientific publications describing binding preferences of TFs were identified, subjected to critical review and, if judged adequate, incorporated into the database. TF building models were constructed by applying pattern-finding algorithms on the sets of binding sequences retrieved from each publication.

The database, named JASPAR, was later used as the primary profile collection for the ConSite phylogenetic footprinting server and tied into the TFBS scientific programming modules[88].

As the demand for convenient access to this collection grew, a web-based database interface was implemented (*http://jaspar.cgb.ki.se*). In this interface, researchers are able to retrieve profiles according to various criteria, including profile similarity (described in paper IV), and graphically assess results (**Figure 3**). Currently, JASPAR is the only open-access TF profile collection, and is used in many biological servers and programs, including the SockEye[89] visualization tool. In addition, JASPAR is used as a part of the characterization of promoter sequences in the large-scale assessment of full-length mouse cDNAs (the FANTOM consortium 2004, unpublished). Gradually, the collection of profiles will grow, based on both external and group contributions.



SEARCH Species: homo sapiens, mus mu AND Class: bHLH, MADS AND Type: selex

	ID	NAME	SPECIES	CLASS	LOGO	
□	MA0004	ARNT	Mus musculus	bHLH		view
☑	MA0006	Ahr-ARNT	Mus musculus	bHLH		view
☑	MA0048	Hen-1	Homo sapiens	bHLH		view
☑	MA0052	MEF2	Homo sapiens	MADS		view
□	MA0083	SRF	Homo sapiens	MADS		view
□	MA0091	Tal1beta-E47S	Homo sapiens	bHLH		view
□	MA0092	Thing1-E47	Homo sapiens	bHLH		view

Export selected profiles as ○ List of IDs ● Compressed matrix flatfile directory

Analyze this (fasta-formatted) sequence with selected profiles

```
>Homo sapiens caspase 8 gene.promoter region, 5' UTR
[illegible]
```

Relative

MA0048 detailed information - Phoenix

MA0048 Hen-1

Logo for MA0048

```
A  [13 13  3  1 54  1  1  1  0  3  2  5 ]
C  [13 39  5 53  0  1 50  1  0 37  0 17 ]
G  [17  2 37  0  0 52  3  0 53  8 37 12 ]
T  [11  0  9  0  0  0  0 52  1  6 15 20 ]
```

DATA			
ID	MA0048	*Supergroup*	vertebrate
Name	Hen-1	*Protein seq*	Q02575
Class	bHLH	*Experiment*	SELEX
Species	Homo sapiens	*Source PMID*	8289804
Information content	14.1316	Save model	Help

Figure 3
Screenshot of the JASPAR database web interface

Paper II: Identification of conserved regulatory elements by comparative genome analysis
Paper III: ConSite: web-based prediction of regulatory elements using cross-species comparison

As outlined in the introduction, the 'futility theorem' states that while profile models can identify functional binding sites with high sensitivity, the number of false positives is too high for meaningful analysis. As in most prediction problems in computational biology, this is a signal-to-noise problem, where the noise overshadows the signal. There are two principal ways to improve the efficiency: increase the signal or decrease the noise.

Cross-species comparison is one of the principal concepts in computational biology[90,91]. Incorporation of evolutionary information has proven valuable in many sub-fields, including gene finding[59], structure prediction[92] and pattern finding in amino acid sequence[93]. Cross-species comparison is based on the hypothesis that nucleotides or amino acid residues conserved over evolution in related sequences are of particular functional importance. While this hypothesis can be shown to hold empirically, a more pleasing explanation lies in the Darwinian algorithm of selection[4]: critically important nucleotides are subject to higher selective pressure and are thus less likely to mutate to other nucleotides. In order to accurately use phylogenetic footprinting, we need pairs of orthologous sequences, not merely homologues sequences (which might not be subject to the same selective constraints during evolution). As an extension to this rule, analyzing coding sequences for TFBS using phylogenetic footprinting is close to meaningless (even though TFBS can be located in such regions), as the selective pressure in coding sequences is dominated by the properties of the coded protein.

In the field of TFBS prediction, cross-species comparison has been successfully used for improved predictions on a small number of genes[94,95]. For such a strategy to be worthwhile on a global scale, we require that the selectivity gain is substantial – thereby filtering out a significant fraction of falsely predicted sites, while the rate of true predictions must be comparable to standard, single-sequence methods (in other words, sensitivity should be more or less unchanged and selectivity drastically improved).

The ConSite project merged several resources and algorithmic developments to achieve a standardized set of linked methods for phylogenetic footprinting. The method performs three discrete steps, concluding with a prediction of binding sites based both on TF binding models and sequence conservation. Input promoter sequences are either entered by the user or retrieved semi-automatically using a novel expert system based on the Genelynx databases[96]. Alignment of non-coding sequences presents particular challenges – short stretches of similarity are buried in larger, non-conserved regions. Shaped by the work by Mendoza and Wasserman (unpublished work), we used the global ORCA aligner, which combines the BLAST[97] and Needleman-Wunsch[98] algorithms. The degree of conservation in the alignment was assessed by letting a fixed frame incrementally slide over the alignment, observing the number of identical nucleotides within the frame. Given a user-set cutoff, only those windows of sufficiently high sequence identity are used for further analysis. Once conserved regions are defined, a set of chosen TF models from the JASPAR database is employed to scan the regions. Unique to ConSite, both input sequences are scanned, and sites are only retained when predicted in corresponding positions in the alignment.

Two test sets were collected to assess the performance of the methods; one small set of hand-curated sites from literature studies, and one large data-mined set, based on mappings of TRANSFAC[87] sites onto the human and mouse assemblies. The latter set is the largest reference collection for phylogenetic footprinting studies to date. In summary, while sensitivity is slightly affected by phylogenetic footprinting, the number of false positives is reduced by ~85% in both sets. The two test sets differ slightly in the sensitivity tests, presumably due to both differences in the set of models used (for example, the model set in the latter test-case has a significantly lower information content) and the quality of annotations (**Figure 4**).

The ConSite methods are accessible to any researcher in an intuitive, graphical web interface (*http://phylofoot.org)*, integrating different output formats and parameter choices.

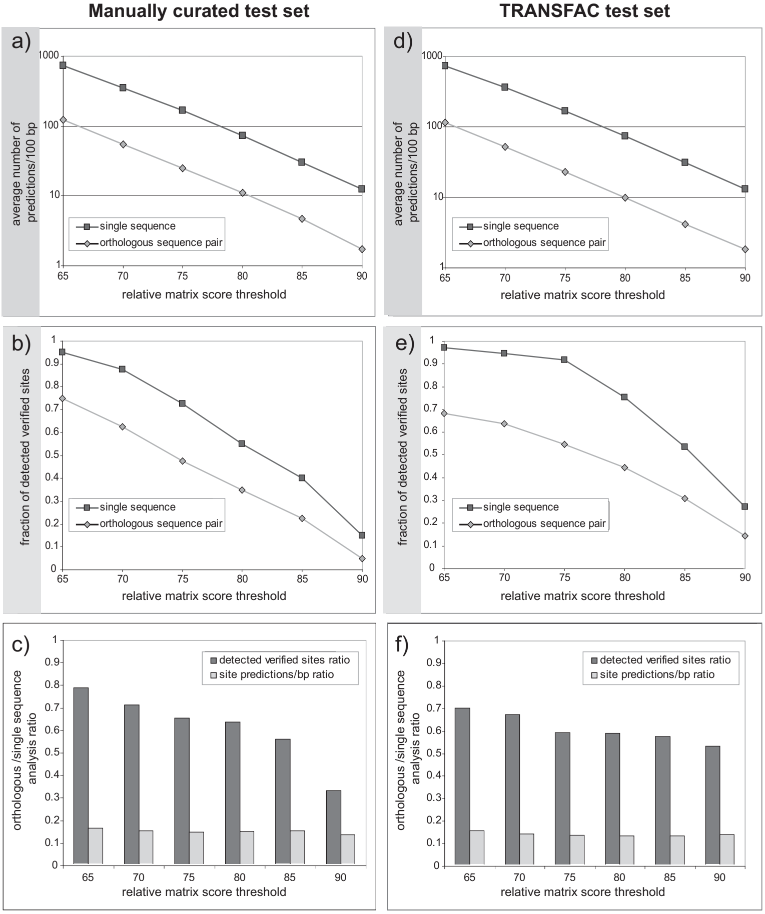


Figure 4
The impact of phylogenetic footprinting analysis. Both **(a-c)** a high-quality set (14 genes and 40 verified sites), and **(d-f)** a larger collection of promoters (57 genes and 110 sites, from the TRANSFAC database) were analyzed. (**a,d)** Comparison of the selectivity (defined as the average number of predictions per 100 bp, using all models) between orthologous and single-sequence analysis modes. **(b,e)** Comparison of the sensitivity (the portion of 40 or 110 verified sites, respectively, that are detected with the given setting) between orthologous and single-sequence analysis modes. **(c,f)** Ratios of the number of sites detected in single-sequence mode to the number detected in orthologous-sequence mode; the pair: single-sequence ratios are displayed for both sensitivity (detected verified sites) and selectivity (all predicted sites).

Paper IV: Prediction of Nuclear Hormone Receptor Response Elements

Protein-DNA interfaces are subject to significant selective pressure – protein and DNA counterparts are co-evolving[99]. Thus, while two binding sites of the same TF may vary in nucleotide sequence, spatial deviations (such as insertions) occur rarely[34]. The inability of profile model framework to describe variable spacing within a single binding site is therefore usually not a concern. However, TFs often bind as dimers to DNA, in some cases with a variable spacing between the two sites recognized by each monomer (often known as 'half-sites')[28,29,33,34]. The nuclear hormone receptor (NHR) class of TFs is perhaps the most well studied group of TFs that has this property[100,101]. Dimers of this class recognize a two consecutive consensus sites ('AGGTCA'), which can be differently spaced and/or have different strand orientations[100]. It is clear that a normal profile model cannot describe these characteristics adequately. The nuclear receptor field is currently lacking mature computational methods for the prediction of NHR response elements.

The Hidden Markov Model (HMM) framework is a suitable candidate for modeling sequences displaying insertions and deletions[50]. In this work, a HMM framework was constructed to model the generalized DNA-binding properties of known nuclear hormone receptors (**Figure 5**). The model should both be able to find nuclear receptor binding sites in genomic sequence and classify found sites correctly (site configuration and number of spacer nucleotides).

A collection of validated NHR TFBS were collected from the biomedical literature. The set was used to train the model, using a simple maximum-likelihood procedure. Cross-validation tests showed the model to be highly sensitive and reasonably selective (given the futility theorem).

As a case example, we applied the model to the compact pufferfish (*Fugu rubripes*) genome[11]. We found that there is a high potential for nuclear receptor genes to be cross-regulated by other nuclear hormone receptors. The different distributions of over-represented site configurations when comparing different types of nuclear receptors suggest that the type of NRs involved in cross-regulation varies depending on the type of target gene.

The model can be used in an intuitive web-interface, located at *http://mordor.cgb.ki.se/NHR-scan.*

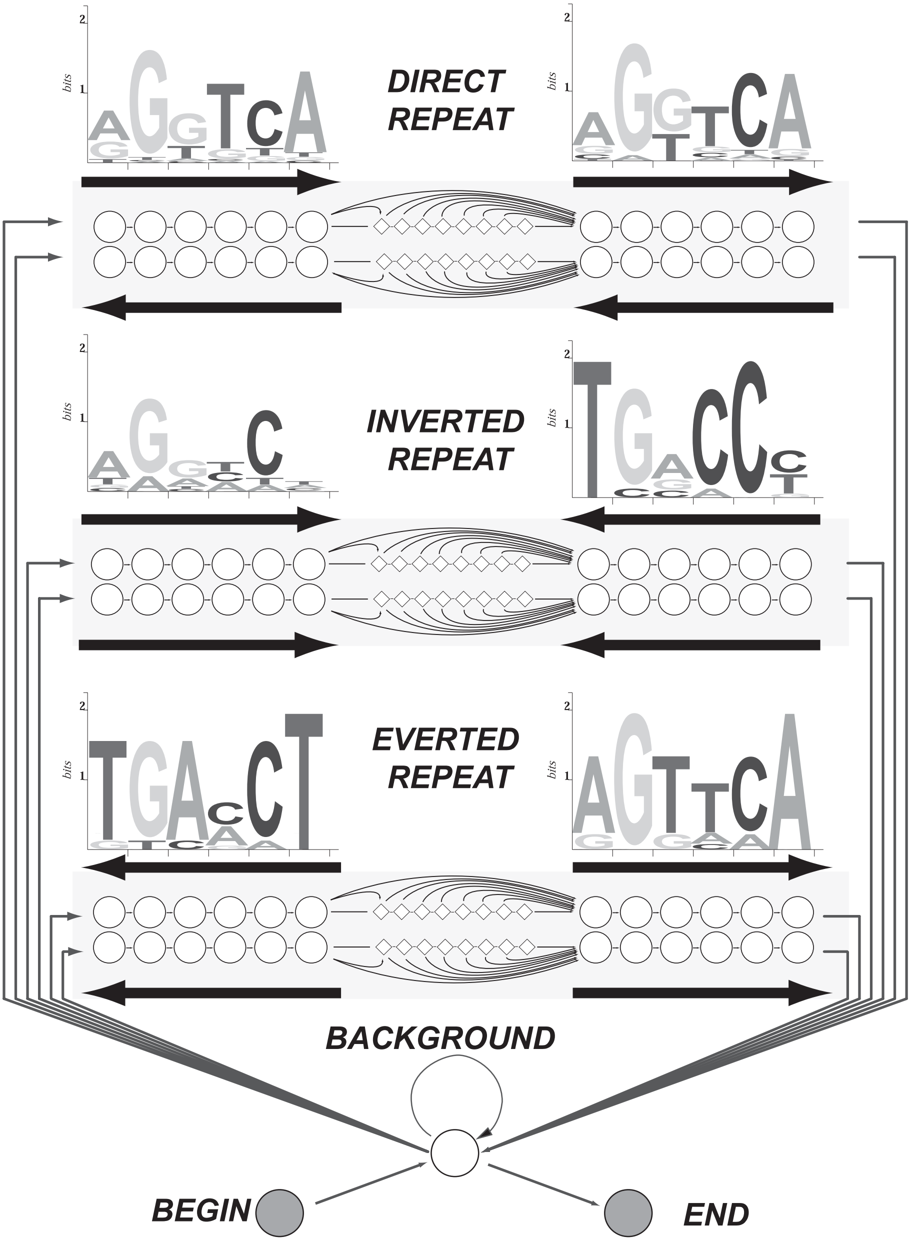


Figure 5
Graph representation of HMM framework: For all three match-states, the halfsite models (excluding pseudo-counts) are shown using sequence logos. Each match state consists of a pair of chains, corresponding to forward/reverse strand

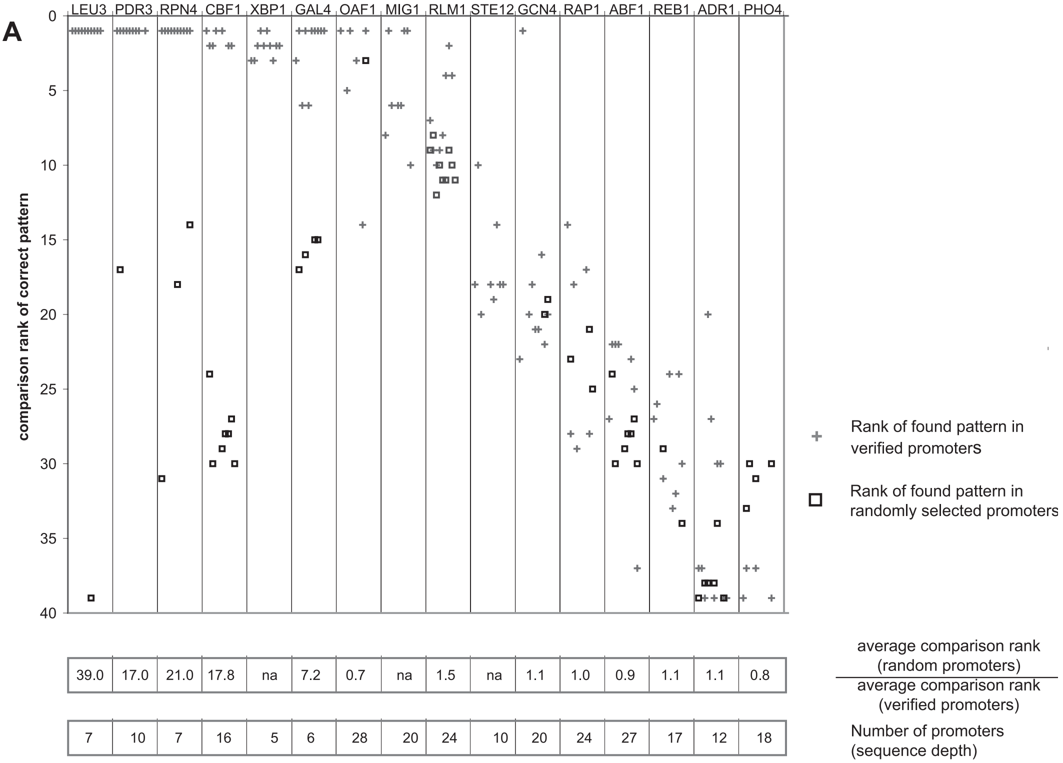
Paper V: Integrated analysis of yeast regulatory sequences for biologically linked clusters of genes

As outlined above, one of the limitations with many pattern finders is the inability to differentiate between biologically relevant patterns and non-functional (albeit over-represented). Maximum a posteriori probability (MAP) scores have some utility in this regard[72,78,102], but is more correlated with statistical over-representation than biological function[78]. In many cases researchers are interested in both over-representation and what TFs that the proposed pattern might originate from.

The YRSA project aimed for a merging of a state-of-the-art pattern finder (Gibbs Motif Sampler[72,102]) and a novel pattern comparison algorithm that compares newly found patterns with experimentally verified TF models. Since pattern finding algorithms often are applied on upstream sequences of genes identified with micro-array technology in bakers yeast, *Saccharomyces cerevisiae*[70,103,104], the application is yeast-centric (even if the approach holds for other organisms).

For the comparison of patterns, a modified Needleman-Wunsch[98] algorithm (Matrixaligner) was implemented. In contrast from the original algorithm, Matrixaligner allows for the opening of at maximum one continuous gap in the profile alignment. This constraint addresses situations where TFs bind as hetero-dimers with variable spacing. As in the original algorithm, Matrixaligner evaluates the optimal alignment given a scoring function of pair of sequence positions (originally nucleotides or amino acids, but in this case profile columns).

To evaluate the system, a set of yeast 'regulons' (here defined as a set of genes known to be regulated by the same TF) was assembled, based on literature data. The YRSA system could find the relevant sites and classify the mediating TF in the majority of cases (**Figure 6**). In a set of case examples, we could show that the YRSA system can confirm old results, expand previous findings and help deliver new biological insights. The case examples culminated in the finding that the MCB (MluI cell cycle box) element is a likely regulator of DNA-damage response genes, which is consistent with MCBs known role in the regulation of the cell-cycle[104].



	LEU3	PDR3	RPN4	CBF1	XBP1	GAL4	OAF1	MIG1	RLM1	STE12	GCN4	RAP1	ABF1	REB1	ADR1	PHO4
average comparison rank (random promoters) / average comparison rank (verified promoters)	39.0	17.0	21.0	17.8	na	7.2	0.7	na	1.5	na	1.1	1.0	0.9	1.1	1.1	0.8
Number of promoters (sequence depth)	7	10	7	16	5	6	28	20	24	10	20	24	27	17	12	18

Figure 6
Systematic estimation of pattern detection specificity using a curated collection of co-regulated genes targeted by characterized TFs. Sets of genes known to be targeted by a TF with available binding profile were analyzed in the YRSA system. **A)** Ranks of detected patterns with MAP scores exceeding the random score threshold. The ratio of the average scores for target and random promoter sets are indicated beneath the figure. In some cases, no significant patterns were found in the random promoter sets

Paper VI: Constrained Binding Site Diversity within Families of Transcription Factors Enhances Pattern Discovery Bioinformatics

The ‘pattern drowning’ experienced in pattern finding when analyzing extended promoter sequences quickly becomes a insurmountable problem when moving from prokaryotes and yeast to multi-cellular eukaryotes, where regulatory regions frequently are scattered over thousands of basepairs of upstream sequence. Many researchers strive to improve the sensitivity of pattern finders by improving background models[74,83,105], as mentioned in the introduction. While many of these improvements are significant, additional developments are needed.

From a biological viewpoint, probabilistic pattern finders are naïve, as (correcting for nucleotide background distributions) all equally over-represented patterns are considered equally good solutions. It is likely that the ‘pattern space’ in biological systems is more constrained, as pattern characteristics are directly dependent on a few distinct DNA-protein interface structures[34]. It is generally recognized that most structurally related TFs bind similar target sequences. If binding models for representative members of a structural family could be merged into a single generalized description, a set of such models could be viewed as focal points in the solution space of pattern finders.

Using the Matrixaligner algorithm described in paper V and the JASPAR database of profiles (Paper I), an algorithm was constructed for the construction of ‘familial binding profiles’ (FBPs) – meta-models describing shared binding characteristics of a class of structurally related TFs. In brief, all profiles belonging to a class were compared to each other, producing an empirical p-value for each pair associated with the similarity of the profiles. The contribution of each profile to the FBP was weighted by a factor inversely proportional to the average p-value score to all other profiles in the set. The profile with the highest average p-value score was used as a positional template to align all profiles within the class. Given the available data, 11 FBPs corresponding to the major TF structural classes were constructed

Comparisons of profiles to FBPs can be utilized for prediction of the structural class of the mediating TF, similar to the comparisons with database profiles in papers I and IV. This application has utility when assessing patterns originating from pattern finding algorithms applied to micro-array data, where no or little information about mediating factors is available. In tests using external and internal data (i.e. cross-validation), close to 90% of the profiles in the test set could be correctly classified.

Probabilistic pattern finders can be intentionally influenced by prior expectations. The exact mechanism differs in different programs (for instance, in the Gibbs Motif Sampler[72,102] the priors are used as pseudocounts when evolving patterns are constructed, while ANN-Spec[83] modifies the initial perceptron in the integrated neural network). FBPs can be used as such prior knowledge, thereby focusing pattern detection towards sites associated with a TFs of a certain structural class.

This approach was tested quantitatively by seeking known binding sites embedded in iteratively extended promoter sequence using two pattern finder programs, with and without incorporated prior knowledge (**Figure 7**). In the ten cases tested, pattern finding using prior knowledge in the form of FBPs had a dramatically improved sensitivity, measured as the promoter length at which the pattern finder results were indistinguishable from the control.

In a case example, this approach could identify functional anti-oxidant response elements in extended promoter sequences, which was not possible without the usage of an FBP as prior knowledge. Furthermore, the pattern was successfully used to classify the type of mediating transcription factor.

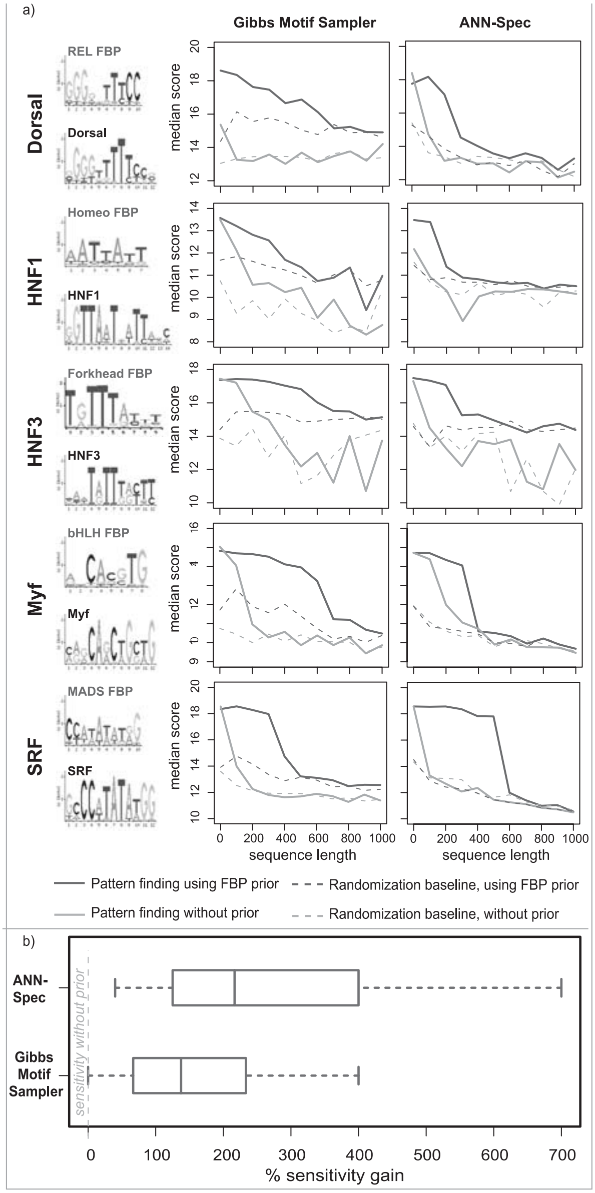


Figure 7

Incorporation of FBPs improve pattern discovery sensitivity. (a) Pattern detection of known sites for Dorsal, HNF1, HNF3, Myf and SRF factors was performed in successively extended sequences, using two different pattern finding algorithms (Gibbs Motif Sampler and ANN-Spec). Resulting patterns from each extension were compared to the respective in vitro model in the JASPAR database (y-axis). Unbiased pattern finding analysis (grey line) and pattern finding intentionally biased towards the respective factor's corresponding FBP (REL, Homeo, Forkhead, bHLH and MADS-box) (black line) was evaluated. As baseline (broken lines), the same procedure was applied with 10 bp wide, randomly selected promoter sequences as starting points instead of true binding sites. (b) Percentage sensitivity gains using prior knowledge, measured as the extension value where pattern finder results (continuouslines) are indistinguishable (i.e. intersects) from background (broken lines).

PERSPECTIVES

Alice "We don't know how to make an invisible robot."
Dogbert "Do you know how to make an empty box?"
—Scott Adams

In this work, different methods for improving computational prediction of *cis*-regulatory elements have been presented – ranging from improved models to algorithm developments. While the methods presented provide certain improvements, they are not conclusive solutions to the problem. Detailed discussions of merits and disadvantages of the presented methods can be found in respective papers. Here, some key issues for future in-silico prediction of *cis*-regulatory elements in light of these findings will be discussed.

Model collections

The profile model is, as stated in the introduction, the most commonly used framework in the field. There is a clear discrepancy between the number of known TFs and the number of high-quality models, owing to the considerable number of verified sites needed to build an adequate model. The possibility to evaluate *in vitro* binding using large-scale site selection assays is promising[106], but has not yet produced significant numbers of profiles. Chip-based chromatin immunoprecipitation[43,107,108] is also a viable option.

The number of profiles in the JASPAR database (Paper I) will grow over time, as new profiles are added when new experimental data becomes available. A regulatory region annotation tool is under construction, in which laboratorial scientists will be able to commit experimentally validated sites and models. This resource will be coupled to JASPAR database.

TFBS prediction

Cross-species comparison is an effective strategy for increasing the selectivity of predictions. In paper II, we show that phylogenetic footprinting can reduce the amount of false predictions by ~85% at uniform settings, using the largest test set to date. This type of analysis is dependent on multiple factors: alignment algorithms, model collections, choice of settings, and methods to assess conservation. These, as well as other aspects, are targets for future improvements. The need for comprehensive model collections has been discussed above – but there is also a need for reference testing sets for evaluating new methods. In the case of phylogenetic footprinting, verified sites are only the initial requirement. Sites must subsequently be mapped on the genome, and relevant orthologous promoters have to be identified. The large test set introduced in paper II is the largest such set to date, but amounts to a minute sample of all functional binding sites in human/mouse.

The definition of orthologous sequences is a related problem, which has not been discussed in this work. In many cases, the distinction between paralogous and orthologous[84,85] sequences is non-trivial. Several important resources addressing this problem has emerged in recent years[109-112].

As computational biologists are in need of experimental data, bridges between the theoretical and laboratory communities of biologists are desirable. Because of this, the implementation of algorithms as user-friendly tools is important. In paper III, enhancements to the ConSite interface are described. In particular, the selection of input sequences was enhanced. As the selection of orthologous promoters often is regarded one of the hardest parts of the analysis, an improved method for semi-automatic retrieval of orthologous human-mouse promoters, was implemented. In the future, more precise TSS locations can be located with the help of forthcoming CAGE data from the RIKEN consortiums[113,114].

Current applications have not significantly moved beyond pair-wise comparisons. Incorporation of multiple sequences is possible, but for meaningful comparison, the contribution of each sequence should be weighted by their evolutionary distance to other sequences. Some new approaches have emerged in this direction[115-117].

Ultimately, cross-species comparison should be regarded a convenient shortcut for helping us describe a cellular reality beyond our current comprehension. TFs in cells have no explicit information about conservation of sequences, but can still find their functional sites.

Model frameworks

The profile model has been shown to be an adequate descriptor of *in vitro* binding specificity[61]. However, it is in some cases inadequate even for *in vitro* situations. In paper IV, we showed that a more advanced HMM framework is more suitable in the case of nuclear hormone receptor TFs.

Even though the profile model and related frameworks are adequate for describing an *in vitro* situation, it is clear that many aspects important for transcriptional regulation are not incorporated in current model frameworks, most notably the influence of chromatin structure. Some pioneering bioinformatics efforts on this topic is emerging[118-120]. Profile models can be integrated in more advanced statistical frameworks (for instance Support Vector Machines[121] or Neural Networks[122] that take some of these aspects into account. However, the lack of a more thorough understanding of the biology of the nucleus might raise obstacles for major advances in this area.

Pattern finding

Pattern finding in multi-cellular eukaryotes is limited by the small amount of information contained within TFBS patterns and the length of the surrounding sequences. Pattern finding can be stated as a purely algorithmic problem[75,123], but when applied to biological questions, the sought answers are often hard to define mathematically. For instance, it is not given that the most over-represented site in a set of sequences is the most biologically relevant. In paper V, we introduce an algorithm for comparing a pattern retrieved from pattern finding to a set of already known (biologically functional) patterns. This enables researchers both to identify putative mediating TFs and the novelty of the pattern found, much like the customary BLAST analysis when characterizing new genes or proteins. In the current implementation, the pattern database only covers 10% of the known TFs in yeast. As stated above, the discrepancy between the number of models and the number of known TFs is a recognized problem. However, large-scale chromatin IP evaluations of the binding characteristics of all yeast TFs are under way[43], which will consolidate the utility of YRSA and similar approaches.

While both improved background models and methods to evaluate results are necessary, we have shown that incorporation of structural constraints as prior expectations in the pattern finder process can increase sensitivity dramatically (paper VI). As discussed in knowledge-based priors into pattern finders are currently immature, and will benefit from both statistical and biological perspectives. For instance, the assessment of the significance of profile-to-profile scores and hierarchical classifications schemes may prove fertile grounds for statisticians, while the incorporation of additional data sources (for instance the definition of invariant nucleotides on a structural basis) is an interesting biological problem.

ACKNOWLEDGEMENTS

I would like to present my sincere thanks to everyone that in any way has supported and helped me during my thesis project. In particular:

Supervisors

Wyeth Wasserman, main supervisor

Kindest man in the world. Great with science and with people. Thank you for endless enthusiasm, empathy and big ideas. Moreover, for initially spotting some deeply hidden potential.

Sven Petterson, co-supervisor

For good discussions and a critical eye

Boris Lenhard, co-supervisor

For heaps of kind help, both practical and theoretical. For being a demanding yet understanding scientist.

Past and current group members, in no particular order

Danielle Kemmer

For being a great roommate. You know, somehow I managed to do something in all those hours after all, other than juggling and listening to music.

William Krivan

A terrific roommate from my first days at the CGB. Thank you for your all your attitude and humor.

Annette Höglund

Thanks for pleasant cooperation, great enthusiasm and to actually get the whole group to take 'fika' at once at repeated times. This takes skill.

Luis Mendoza

For kind help and good cooperation

Pär Engström

Expert surfer and room-mate. For putting up with my many quirks as a long-time roommate. For many valuable scientific, and non-scientific, discussions and sarcastic views of life. Surf's up!

Johan Geijer

For good cooperation and a relaxed, open attitude.

Wynand Alkema

For good cooperation, constructive criticisms and a particularly un-sportsmanlike attitude when playing kubb.

Jing Sheng

For enjoyable discussions and your patience with my data management errors

Christian Storm

For pleasant company and good comments

Elena Herzog

For being a very special person, and good company. And for inviting me to a spectacular wedding

Bill Wilson

For enthusiasm and excellent input

Sara Bruce

For putting up with Pär and me during your project work, your questioning attitude and sharp mind.

KI/UBC/SU/KTH people

I would like to thank everyone I have interacted with at the **Center for Genomics and Bioinformatics** (Karolinska Institutet) **Center for Molecular Medicine and Therapeutics** (University of British Columbia), **Stockholm Bioinformatics Center** and other departments. In particular:

Lars Arvestad
For much appreciated help around alignment algorithms

Timothy Bailey
For the many hours spent on my algorithmic problems

Lukas Käll and Marcus Wistrand
For many helpful discussions around shared HMM issues, science and life in general

Hui Gao
For a cheerful attitude, and expert help on microarrays and nuclear receptors

David Fredman
For helping out, and for always having a weird comment handy.

Mark Reimers
For having a huge interest in biology and willingness to share your wealth of statistical knowledge.

Niclas Jahreborg
For good cooperation and a good attitude

Jens Lagergren
For many words of encouragement and help around computational biology issues.

CMMT people, including **David Arenillas, Jochen Brumm, Shannan J. Ho Sui, Jonathan Lim, Stefanie Butland, Graeme Campbell, Joanne Fox, Sohrab Shah, Jessica Sawkins, Scott McMillan, Miroslav Hatas** and others
For lots of help during my stays in Canada and a friendly atmosphere.

Johan Nilsson
For co-teaching the Biomedical Graduate School bioinformatics course. Responsible for introducing Clint Eastwood as a molecular biologist icon

Yoshahide Hayashizaki,
For fantastic, contagious energy, and scientific courage.

Claes Wahlestedt
For laying the foundation for my interest in working at the CGB, and for help and support during the years.

Bent Nagstrup Terp
Unix guru. Thanks for all computer help during the years, and thanks for reliving me of taking care of the fungal farm. Will never forget the demonstration of 'The Last Alliance vs the Orcs on the Slope of Mount Doom' using cutlery at lunch.

Others

Stockholm Graduate School of Biomedical Research class of 1999-2000
For support and laughs during the years

Eva Severinsson
For all the helpful comments and encouragements over the years, and for coordinating an excellent graduate program.

Uncle Teofil
For many pleasant stories and unbiased perspectives

Kelvar and Tulkas
Cuddling therapists. For an uncomplicated view of life.

Andreas Sandahl, Emma Gunderblad, Niklas Ahlgren, Alexis Voisin, Magnus Bergström, Ann Karlsson and other wild-card muchkins
For not-so-serious yet strangely meaningful gaming.

Finally

Ann, light of my life
For everything.

REFERENCES

1. Alberts, B. et al. *Molecular Biology of the Cell* (Garland Pub, New York, 2002).
2. Lodish, H. et al. *Molecular Cell Biology* (W H Freeman & Co., 1999).
3. Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. *Introduction to Algorithms* (MIT press, Camebridge, Massachusetts, 2001).
4. Darwin, C. *The Origin of Species* (1880).
5. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-8 (1953).
6. Nirenberg, M. & Leder, P. Rna Codewords and Protein Synthesis. The Effect of Trinucleotides Upon the Binding of Srna to Ribosomes. *Science* **145**, 1399-407 (1964).
7. Brenner, S. RNA, ribosomes, and protein synthesis. *Cold Spring Harb Symp Quant Biol* **26**, 101-10 (1961).
8. Hoagland, M. B., Stephenson, M. L., Scott, J. F., Hecht, L. I. & Zamecnik, P. C. A soluble ribonucleic acid intermediate in protein synthesis. *J Biol Chem* **231**, 241-57 (1958).
9. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
10. Waterston, R. H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).
11. Aparicio, S. et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**, 1301-10 (2002).
12. Consortium, T. C. e. S. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-8 (1998).
13. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815 (2000).
14. Adams, M. D. et al. The genome sequence of Drosophila melanogaster. *Science* **287**, 2185-95 (2000).
15. Goffeau, A. et al. Life with 6000 genes. *Science* **274**, 546, 563-7 (1996).
16. Tamas, I. et al. 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**, 2376-9 (2002).
17. Blattner, F. R. et al. The complete genome sequence of Escherichia coli K-12. *Science* **277**, 1453-74 (1997).
18. Kawarabayasi, Y. et al. Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, Aeropyrum pernix K1. *DNA Res* **6**, 83-101, 145-52 (1999).
19. Felsenfeld, G. Quantitative approaches to problems of eukaryotic gene expression. *Biophys Chem* **100**, 607-13 (2003).
20. Maloy, R. S., Cronan, J. E. & Freifelder, D. *Microbial genetics* (Jones and Bartlett Publishers, London, 1994).
21. Weintraub, H. et al. The myoD gene family: nodal point during specification of the muscle cell lineage. *Science* **251**, 761-6 (1991).
22. Thummel, C. S. Mechanisms of transcriptional timing in Drosophila. *Science* **255**, 39-40 (1992).
23. Davidson, E. *Genomic Regulatory Systems. Development and Evolution* (Academic Press, San Diego, 2001).

24. Lemon, B. & Tjian, R. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev* **14**, 2551-69 (2000).

25. Blackwood, E. M. & Kadonaga, J. T. Going the distance: a current view of enhancer action. *Science* **281**, 61-3 (1998).

26. Wolberger, C. Multiprotein-DNA complexes in transcriptional regulation. *Annu Rev Biophys Biomol Struct* **28**, 29-56 (1999).

27. Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**, 285-94 (1999).

28. Wu, J. & Grunstein, M. 25 years after the nucleosome model: chromatin modifications. *Trends Biochem Sci* **25**, 619-23 (2000).

29. Kadonaga, J. T. Eukaryotic transcription: an interlaced network of transcription factors and chromatin-modifying machines. *Cell* **92**, 307-313 (1998).

30. Anderson, J. D. & Widom, J. Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites. *J Mol Biol* **296**, 979-87 (2000).

31. Polach, K. J. & Widom, J. Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation. *J Mol Biol* **254**, 130-49 (1995).

32. Polach, K. J. & Widom, J. A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *J Mol Biol* **258**, 800-12 (1996).

33. Branden, C. & Tooze, J. *Introduction to protein structure* (Garland Publishing, Inc., New York, 1999).

34. Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. An overview of the structures of protein-DNA complexes. *Genome Biol* **1**, REVIEWS001 (2000).

35. Roberts, R. J. Restriction enzymes and their isoschizomers. *Nucleic Acids Res* **16 Suppl**, r271-313 (1988).

36. Ingham, P. W. & Martinez Arias, A. Boundaries and fields in early embryos. *Cell* **68**, 221-35 (1992).

37. Hegde, P. et al. A concise guide to cDNA microarray analysis. *Biotechniques* **29**, 548-50, 552-4, 556 passim (2000).

38. Lockhart, D. J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol* **14**, 1675-80 (1996).

39. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-70 (1995).

40. Pollock, R. & Treisman, R. A sensitive method for the determination of protein-DNA binding specificities. *Nucleic Acids Res* **18**, 6197-204 (1990).

41. Shultzaberger, R. K. & Schneider, T. D. Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepanciesbetween natural selection and SELEX. *Nucleic Acids Res* **27**, 882-7 (1999).

42. Orlando, V. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem Sci* **25**, 99-104 (2000).

43. Shannon, M. F. & Rao, S. Transcription. Of chips and ChIPs. *Science* **296**, 666-9 (2002).

44. Wasserman, W. W. & Krivan, W. In silico identification of metazoan transcriptional regulatory regions. *Naturwissenschaften* **90**, 156-66 (2003).

45. Wasserman, W. W. & Sandelin, A. Applied Bioinformatics for the Identification of Regulatory Elements. *Nat Rev Genet* **5,**276-287 (2004).

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

47. McClure, W. R. Mechanism and control of transcription initiation in prokaryotes. *Annu Rev Biochem* **54**, 171-204 (1985).

48. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **18**, 6097-100 (1990).

49. Shannon, C. E. A mathematical theory of communication. *Bell Syst Tech J* **27**, 379-423 (1948).

50. Durbin, R., Eddy, S., Krogh, H. & Mitchison, G. *Biological sequence analysis* (Cambridge University Press, Cambridge, 1999).

51. King, O. D. & Roth, F. P. A non-parametric model for transcription factor binding sites. *Nucleic Acids Res* **31**, e116 (2003).

52. 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**, 723-50 (1987).

53. Claverie, J. M. & Audic, S. The statistical significance of nucleotide position-weight matrix matches. *Comput Appl Biosci* **12**, 431-9 (1996).

54. Benos, P. V., Bulyk, M. L. & Stormo, G. D. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res* **30**, 4442-51 (2002).

55. Udalova, I. A., Mott, R., Field, D. & Kwiatkowski, D. Quantitative prediction of NF-kappa B DNA-protein interactions. *Proc Natl Acad Sci U S A* **99**, 8167-72 (2002).

56. Barash, Y., Elidan, G., Friedman, N. & Kaplan, T. Modeling Dependencies in Protein-DNA Binding Sites. *RECOMB '03* (2003).

57. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755-63 (1998).

58. Pedersen, A. G., Baldi, P., Brunak, S. & Chauvin, Y. Characterization of prokaryotic and eukaryotic promoters using hidden Markov models. *Proc Int Conf Intell Syst Mol Biol* **4**, 182-91 (1996).

59. Reese, M. G., Kulp, D., Tammana, H. & Haussler, D. Genie—gene finding in Drosophila melanogaster. *Genome Res* **10**, 529-38 (2000).

60. Fickett, J. W. Quantitative discrimination of MEF2 sites. *Mol Cell Biol* **16**, 437-41 (1996).

61. Tronche, F., Ringeisen, F., Blumenfeld, M., Yaniv, M. & Pontoglio, M. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J Mol Biol* **266**, 231-45 (1997).

62. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. Human-mouse genome comparisons to locate regulatory sites. *Nat Genet* **26**, 225-8 (2000).

63. O'Brien, T. P. et al. Genome function and nuclear architecture: from gene expression to nanoscience. *Genome Res* **13**, 1029-41 (2003).

64. Wolffe, A. P. & Guschin, D. Review: chromatin structural features and targets that regulate transcription. *J Struct Biol* **129**, 102-22 (2000).

65. Davidson, E. H. *Genomic regulatory systems: development and evolution* (Academic Press, San Diego, 2001).
66. Wasserman, W. W. & Fickett, J. W. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* **278**, 167-81 (1998).
67. Krivan, W. & Wasserman, W. W. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res* **11**, 1559-66 (2001).
68. Brazma, A., Jonassen, I., Eidhammer, I. & Gilbert, D. Approaches to the automatic discovery of patterns in biosequences. *J Comput Biol* **5**, 279-305 (1998).
69. Brazma, A., Jonassen, I., Vilo, J. & Ukkonen, E. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res* **8**, 1202-15 (1998).
70. Bussemaker, H. J., Li, H. & Siggia, E. D. Regulatory element detection using correlation with expression. *Nat Genet* **27**, 167-71 (2001).
71. Akutsu, T., Arimura, H. & Shimozono, S. On approximation algorithms for local multiple alignment. *Proceedings of the fourth annual international conference on Computational molecular biology*, 1-7 (2000).
72. Lawrence, C. E. et al. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208-14 (1993).
73. Lawrence, C. E. & Reilly, A. A. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* **7**, 41-51 (1990).
74. Bailey, T. L. & Elkan, C. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* **3**, 21-9 (1995).
75. Keich, U. & Pevzner, P. A. Finding motifs in the twilight zone. *Bioinformatics* **18**, 1374-81 (2002).
76. Liu, X., Brutlag, D. L. & Liu, J. S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, 127-38 (2001).
77. Cho, R. J. et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **2**, 65-73 (1998).
78. Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* **296**, 1205-14 (2000).
79. Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9**, 657-63 (1999).
80. Antequera, F. Structure, function and evolution of CpG island promoters. *Cell Mol Life Sci* **60**, 1647-58 (2003).
81. Bailey, T. L. & Gribskov, M. The megaprior heuristic for discovering protein sequence patterns. *Proc Int Conf Intell Syst Mol Biol* **4**, 15-24 (1996).
82. Brown, M. et al. Using Dirichlet mixture priors to derive hidden Markov models for protein families. *Proc Int Conf Intell Syst Mol Biol* **1**, 47-55 (1993).
83. Workman, C. T. & Stormo, G. D. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput*, 467-78 (2000).
84. Fitch, W. M. Distinguishing homologous from analogous proteins. *Syst Zool* **19**, 99-113 (1970).

85. Fitch, W. M. Homology a personal view on some of the problems. *Trends Genet* **16**, 227-31 (2000).

86. van Helden, J., Rios, A. F. & Collado-Vides, J. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res* **28**, 1808-18 (2000).

87. Matys, V. et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**, 374-8 (2003).

88. Lenhard, B. & Wasserman, W. W. TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics* **18**, 1135-6 (2002).

89. Montgomery, S. B. et al. Sockeye: A 3D Environment for Comparative Genomics. *Submitted* (2003).

90. Ureta-Vidal, A., Ettwiller, L. & Birney, E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet* **4**, 251-62 (2003).

91. Frazer, K. A., Elnitski, L., Church, D. M., Dubchak, I. & Hardison, R. C. Cross-species sequence comparisons: a review of methods and available resources. *Genome Res* **13**, 1-12 (2003).

92. Korf, I., Flicek, P., Duan, D. & Brent, M. R. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 Suppl 1**, S140-8 (2001).

93. Sonnhammer, E. L., Eddy, S. R. & Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405-20 (1997).

94. Gumucio, D. L. et al. Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. *Mol Cell Biol* **12**, 4919-29 (1992).

95. Loots, G. G. et al. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136-40 (2000).

96. Lenhard, B., Hayes, W. S. & Wasserman, W. W. GeneLynx: a gene-centric portal to the human genome. *Genome Res* **11**, 2151-7 (2001).

97. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-402 (1997).

98. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**, 443-53 (1970).

99. Szafranski, P. On the evolution of the bacterial major sigma factors. *J Mol Evol* **34**, 465-7 (1992).

100. Owen, G. I. & Zelent, A. Origins and evolutionary diversification of the nuclear receptor superfamily. *Cell Mol Life Sci* **57**, 809-27 (2000).

101. Mangelsdorf, D. J. et al. The nuclear receptor superfamily: the second decade. *Cell* **83**, 835-9 (1995).

102. Thompson, W., Rouchka, E. C. & Lawrence, C. E. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res* **31**, 3580-5 (2003).

103. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* **16**, 939-45 (1998).

104. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863-8 (1998).

105. Grundy, W. N., Bailey, T. L., Elkan, C. P. & Baker, M. E. Meta-MEME: motif-based hidden Markov models of protein families. *Comput Appl Biosci* **13**, 397-406 (1997).
106. Roulet, E. et al. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol* **20**, 831-5 (2002).
107. Iyer, V. R. et al. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533-8 (2001).
108. Lee, T. I. et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804 (2002).
109. Arvestad, L., Berglund, A., Lagergren, J. & Sennblad, B. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* **19**, I7-I15 (2003).
110. Tatusov, R. L. et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
111. Wheeler, D. L. et al. Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res* **32**, D35-40 (2004).
112. Storm, C. E. & Sonnhammer, E. L. Comprehensive analysis of orthologous protein domains using the HOPS database. *Genome Res* **13**, 2353-62 (2003).
113. Okazaki, Y. et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563-73 (2002).
114. Shiraki, T. et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A* (2003).
115. Blanchette, M., Schwikowski, B. & Tompa, M. Algorithms for phylogenetic footprinting. *J Comput Biol* **9**, 211-23 (2002).
116. Wang, T. & Stormo, G. D. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **19**, 2369-80 (2003).
117. Siepel, A. & Haussler, D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol* **21**, 468-88 (2004).
118. Levitsky, V. G., Podkolodnaya, O. A., Kolchanov, N. A. & Podkolodny, N. L. Nucleosome formation potential of eukaryotic DNA: calculation and promoters analysis. *Bioinformatics* **17**, 998-1010 (2001).
119. Cremer, M. et al. Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells. *Chromosome Res* **9**, 541-67 (2001).
120. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**, 292-301 (2001).
121. Vapnik, V. N. *Statistical Learning Theory* (Wiley-Interscience, New York, 1998).
122. Baldi, P. & Brunak, S. *Bioinformatics: The Machine Learning Approach* (MIT Press, 2001).
123. Keich, U. & Pevzner, P. A. Subtle motifs: defining the limits of motif finding algorithms. *Bioinformatics* **18**, 1382-90 (2002).