Identification of Transcription Factor Binding Site Motifs on DNA by Means of Statistical Physics

Aymeric Fouquier d'Hérouël

TECHNISCHE
UNIVERSITÄT
DARMSTADT

KTH
VETENSKAP
OCH KONST

Master of Science Thesis

Stockholm, Sweden
October 2005

Typeset in $\LaTeX\,2_{\varepsilon}$

TRITA-FYS 2005:67
ISSN 0280-316X
ISRN KTH/FYS/--05:67--SE

Examensarbete inom ämnet fysik för avläggande av civilingenjörsexamen i Teknisk Fysik (F) vid Kungliga Tekniska Högskolan.

Tryck: Universitetsservice US AB

Abstract

The problem of identifying binding sites for transcription factors as such in experimentally yet unaccessed or unaccessible sequence domains is an interesting task for both biologically and statistically inclined research and several methods have been proposed to solve this riddle, some of them purely information-theoretical, some others assuming a statistical mechanism within.

We develop a novel sampling algorithm for the detection of transcription factor binding sites, based on a multiple local alignment tool known as “Gibbs sampler” and on a description of regulatory sequences by a matrix of binding-free-energies, describing the interaction between a factor protein and DNA.

Finally, we test the algorithm on artificial as well as biological data, finding the predictions made by our sampler to be in good agreement with experiments for transcription factors with short palindromic binding sites, as the FruR protein in E.coli.

Keywords: Transcription Factor Binding Sites, Energy Matrices, Multiple Local Sequence Alignment, QPMEME, Gibbs Sampling, Monte Carlo, Simulated Annealing

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Nucleic acid polymers, discovered in the late 19$^{\text{th}}$ century, were soon interpreted as the carriers for the blueprint or "hereditary code-script" of life, as stated by Erwin Schrödinger [Schrödinger, 1944]. Giving a possible mechanistic background to the Darwinistic theory just recently accepted these days, life was also to be determined by those polymer sequences and hope emerged, one would understand the machinery by reading the blueprint. Effort was thus concentrated on the extraction of genetic material and with the first complete sequencing of a DNA genome in 1978 - bacteriophage ϕX174's [Sanger et al., 1978] - began an era of biochemical decryption, which finally supplied nearly complete genomes of Saccharomyces cerevisiae in 1996 and of Homo sapiens in 2003 and steadily pursues its task of mapping more and more species' sequences. The availability of a sequence, although, doesn't imply its fluent readability and comprehensibility, i.e. the present task is to understand a text written in an unknown language or even a possibly encrypted document in that language.

One knows for certain today, how a sequence of DNA codes for proteins as well as it is possible to estimate which parts really code for genetic information and which parts have regulatory functions, but still leaving enough regions with yet unknown role or sorely to determinate some structural conformation; whereas the latter certainly also influences or even co-regulates genetic mechanisms. The whole "playground", those mechanisms act on, is thus rather complex and participating elements - in terms of abstract information or concrete objects - can in general be expected to show high degrees of cooperativity and codependency. Indeed, research of the last decades showed that complex network of chemical signals and interactions underlay the mechanisms of gene regulation. Moreover there turned out to be a high grade of dynamism, as the products of these mechanisms can, on their part, act as input signals and thus add some feedback to the mechanism itself.

In this present work, we concentrate on a small but nonetheless crucial part of the genetic machinery: the transcription initiating mechanism. More specifically, we are interested in predicting binding sites on DNA for so called transcription factor proteins (TFs), molecules acting as flags on DNA and regulating the transcription of genes being coded by the corresponding sequence region they bind to. Bioinformatics proved in recent years to be able to develop successful tools for the analysis of the genetic machinery ([Bussemaker et al., 2001], [Bailey and Elkan, 1994], [Lawrence et al., 1993]). One of the keywords here is certainly "sequence alignment", where one attempts to identify functional parts by comparison, assuming some level of similarity. Some of the jargon and some basic methods of bioinformatics will be presented, but the intention of this thesis is to describe and determine a method inspired by statistical physics to identify TF binding sites.

Our aim naturally decomposes into two aspects. First, it is of interest to be able to find new binding sites for already known TFs, both within the genome of a specific species, and on its relatives'. Second, the ability to predict binding sites *ex nihilio*, and hence to postulate yet unknown TFs would be an important step in the understanding of gene regulation networks, as one could then search well directed for the potentially involved TFs. A successful realisation of latter is probably unachievable yet, so we will develop a method which hopefully achieves the first task well while not failing totally on the second one.

We will illustrate the method by running on artificial genomes, i.e. random sequences and constructed test-cases, and thereafter confirm its functionality on the experimentally rather well explored genome of Escherichia coli. Attempts to predict new binding sites are finally made on the more complex genomic sequence of Saccharomyces cerevisiae.

1.1 Remark

Goal of this work being to develop a novel alignment algorithm and to apply it to biological data, it is an attempt to leap into the realms of computational biology and consists thus. Both descriptions of the biological, information theoretical and physical contexts are certainly not exhaustive – many details merely being touched on – but were kept to a basic level, hopefully allowing an easy access and understanding of this combined work.

Regarding the bioinformatics part of this theses, all source codes and data sets used for the evaluation are available from the author upon request.

1.2 Outline

Following this introduction, the biological background from the naive physicist's point of view is briefly described in chapter two. Thereafter, in chapter three, we present some theoretical requisites, necessary to construct our method and to evaluate its results. Chapter four is dedicated to the classical information theoretical as well as to a more physically inspired description of sequence patterns and related similarity scores. We proceed with the identification of regulatory motifs in chapter five, enumerating some general aspects and established methods, before presenting some special approaches. The chapter culminates in the description of our contribution to the world of alignment algorithms. Chapters six and seven deal with the acquisition of genomic data and the illustration of some representatively yielded results, followed by a short discussion in chapter eight.

Chapter 2

Molecular biology of nucleic acid polymers

The very compact paper by [Watson and Crick, 1953] on “A structure for Deoxyribose Nucleic Acid”, proposing a then new geometrical interpretation of DNA’s conformation as double helix consisting of complementary strands, surely counts as one of the most momentous publications in molecular biology and chemistry. This chapter summarises some basic knowledge about the structure and function of DNA and enumerate the different types of nucleic acid polymers, as well as their role in biology.

2.1 Constituents and structure

The two strands of deoxyribo nucleic acid polymers are chains of of phosphate (${PO_4}^{3-}$) alternating with a pentose sugar ring of deoxyribose (C_5OH_7). On the latter can be attached one of the bases adenine (A) or guanine (G) - purines, consisting of a six-membered and a five-membered nitrogen-containing ring - and thymine (T) or cytosine(C) - pyrimidines, having only a six-membered nitrogen-containing ring - respectively. Figure 2.1 gives an overview of the primary structure with the most probable pairing between opposing bases shown. Principally, purines bind to pyrimidines via hydrogen bonds and it is therefore possible to attach T to G and C to A, but those lead to sterical problems in a double strand as can be deduced from the figure. We will hence only consider A-T and G-C bindings, so that we can well define the complement of a single stranded sequence.

Looking at the ring of deoxyribose, we note that it consists of an oxygen and cycles over four carbons. The last carbon is again attached to the oxygen as well as to a fifth carbon. Again referring to figure 2.1, we can determine an orientation of the strands, given by the structure of the sugar ring. As the adjacent phosphates bind to the third and fifth carbon of the ring, we denote both directions by $5'$ and $3'$ respectively. Speaking of single

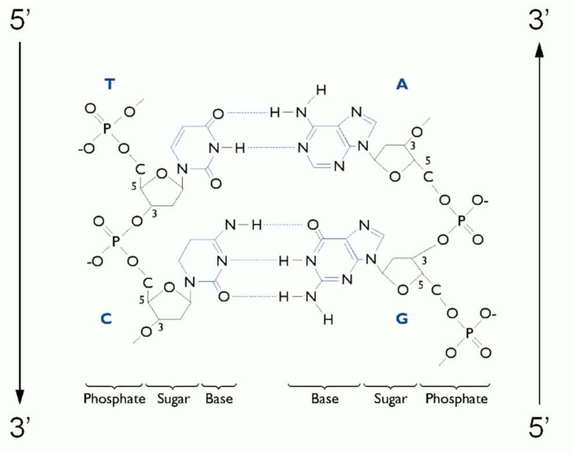


Figure 2.1: Schematic of A-T and G-C base pairing in DNA double-strand

stranded sequences, those are conventionally given in $5' \rightarrow 3'$ direction. The corresponding complement strand then of course reads in the $3' \rightarrow 5'$ direction.

2.2 Types of nucleic acid polymers

Not being a central issue in this work, we shall just shortly recall the most common types of nucleic acid polymers involved in gene production. The main actors here are DNA and RNA, ribose nucleic acid, differing from the former by an OH instead of an H bound to the second carbon of the pentose. This steric hindrance makes the RNA backbone to favour less compact conformations as the DNA’s. Moreover, RNA incorporates the demethylated version of thymine, namely uracil, which is “cheaper” to produce in terms of energy expense. The incorporation of precious thymine in DNA is linked to a reparation process which shall not be discussed here. For details see [Stryer, 2000].

Two types of RNA deserve mention in the actual context, as they play a special role in protein production. Messenger (m)RNA is transcribed from DNA and carries the genetic code of a protein which is to be synthesised by the ribosomes. The secondary structure of mRNA is usually not well defined. Transfer (t)RNA, on the other hand, has a rather well defined secondary structure - the famous clover leaf. Its function is to carry specific amino acids, which are to be processed in the ribosomes. However, since we shall not concentrate on the protein translation mechanism, but on transcription, we refer again to [Stryer, 2000] or equivalent textbooks for further details.

2.3 Hardwired information

Theories about the genetic regulation in organism development have been rather dogmatic throughout history. Religion was eventually replaced by science, which on its turn only introduced the "Central Dogma of Life", as stated by its critics. Figure 2.2 represents a slightly modified version of the assumed flow, which originally includes no feedback.

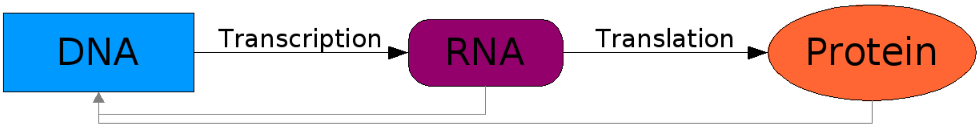


Figure 2.2: Watson and Crick's Central Dogma of Life including feedback

In a simplified picture, DNA is to be transcribed by RNA polymerase to messenger RNA, which on its turn is translated into proteins by the ribosomes. However, both RNA and proteins are chemically active and can thus influence the transcription process itself. The active role of RNA as a reactant or catalyst, i.e. as an enzyme, and thus its possibility to engage in active feedback on DNA - beyond serving as static template during reproduction - is not yet clarified, but what is known is that the whole mechanism of transcription depends highly on the cooperation of TF proteins, as illustrated below. The possible existence of some sort of self-induced mutation mechanism by directed production of enzymes which on their turn might produce mutagens should not be disallowed either, but such threads are out of the scope of the present work and shall not be of any further concern here.

We can thus summarise the function of DNA to consist of a most obvious part: storing static information of genes - and a more subtle part: storing processing directives, how this information has to be interpreted in the actual context of molecule concentrations, temperature, pressure, radiation, etc.

2.4 Transcription regulation mechanism

We assume a simple model of transcription regulation, illustrated in figure 2.3. For clarity, only a single strand of DNA is considered and two special domains are identified: the regulatory region and the gene sequence. The latter is transcribed to mRNA by RNAP, while the former is to bind a set of TFs which may either enhance - by facilitating RNAP-DNA binding - or reduce - by inhibiting the binding site of RNAP - transcription. Transcribed mRNA is then processed by ribosomes, producing signals which may give some feedback on the regulating TFs, e.g. by altering their functionality or by being TFs themselves. From this elementary loop, we can construct networks of arbitrary complexity to model the control of specific transcription based functions or even a whole cell cycle.

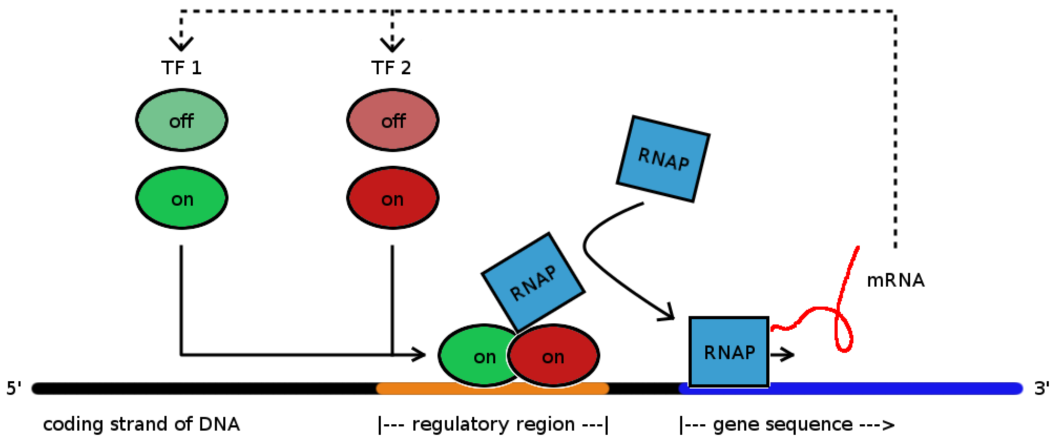


Figure 2.3: Snapshot of the TF specific region in a simplified regulatory network

To get an idea of corresponding binding sites, it is instructive to illustrate some structural examples for TFs. Searching the protein data base (PDB) [Berman et al., 2000], one can easily find representatives for both common and exotic structural classes. Figure 2.4 shows a homo-dimer of *Phosphate System Positive Regulatory Protein PHO4* (PDB ID 1A0A, [Shimizu et al., 1997]) with a common "helix-turn-helix" (HTH) structure motif, bound to DNA. The figure suggests a "site-gap-site" binding-site structure, dictated by the long α-helices of the homo-dimer "clasping" the double-helix. A dimer of the *Nuclear Factor-κB p52* (PDB ID 1A3Q, [Cramer et al., 1997]) is shown beneath, which evidently binds in a more complex and thus specific way. Looking at both structures, one can also guess their order of occurrence in evolution. The smaller TF was extracted from budding yeast cells and its HTH structure motifs can be found in many other organism, both "archaic" and "modern" ones. The larger TF was identified in human cells and its obviously more complex structure suggests a more recent development of the protein.

Budding yeast's protein PHO4 contributes to the organism's phosphatase system and is hence an important part of the nutrient processing apparatus. Human's nuclear factor κB-p52 is a member of the large NF-κ/Rel family of TF proteins, involved in multiple regulatory systems intrinsic to the organism, but also in some dictated by viruses as Herpes and HIV. More details on this versatile TF can be found in biomedical literature.

2.5 Post-transcriptional regulation of gene expression

Until 1993 the mechanism of gene expression was believed to be determined only by the transcriptional part of the central dogma of life. Any valid mRNA being transcribed was supposed to lead to the production of a corresponding protein; no pre-translational feedback originating from the transcription products had been observed until [Lee et al., 1993]

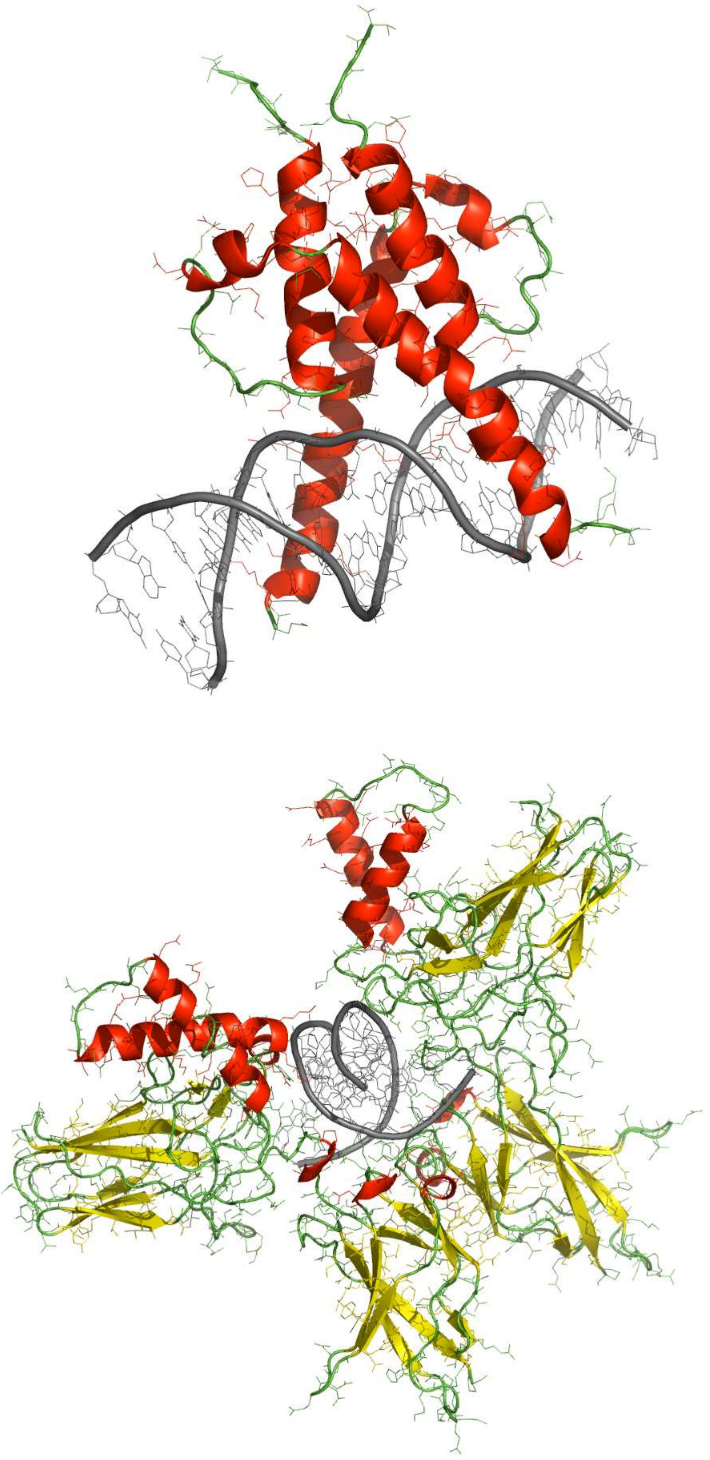


Figure 2.4: Cartoon representations of the dimeric TF proteins 1A0A and 1A3Q (PDB)

identified small fragments of RNA with surprising properties in the worm C. elegans. Those short fragments – named micro (mi)RNA – were excluded from mRNA of the lin-4 gene being prepared for translation and showed reverse complementarity in their genomic sequence to the mRNA of the lin-14 gene. Without going into the details and functions of the named genes, the lin-4 miRNA was shown by [Wightman et al., 1993] to be able to bind lin-14 mRNA and thus reduced the translation of lin-14 by acting as a steric obstacle during the processing of the lin-14 mRNA by ribosomes. Although more hints and clues suggesting another kind of regulatory apparatus were found all along the way, it took almost a decade for the miRNA research to concentrate on the capacities of those short nucleotide sequences. Among others, [Lagos-Quintana et al., 2001] describe the beginning of a broader understanding of the regulatory machinery.

We can summarise today's basic knowledge on miRNA in a simplified manner: miRNA is found in the so called UTR (untranslated region) of a corresponding mRNA sequence; the name of this emplacement already suggests that miRNA is not related to protein synthesis; the short miRNA is separated from the carrying mRNA and remains active, i.e. keeps the ability to bind a complementary sequence; it may inhibit sites on cDNA or other mRNA, thus down-regulating or altering the expression level of a corresponding gene. Three types of down-regulation have been observed so far:

- *mRNA degradation:* Bound mRNA is degraded and can thus not be translated anymore. This ability of miRNA has been mainly observed in plants, as for instance by [Rhoades et al., 2002].
- *translational inhibition:* Bound mRNA remains intact but the processing by ribosomes is repressed. This behaviour has been observed by [Wightman et al., 1993] to take place in animal cells.
- *transcriptional inhibition:* Methylation of DNA, i.e. the addition of a methyl group (-CH_3) to each nucleotide, is known to be reducing the transcription rate of any gene situated on this so called *silent* DNA. miRNA is supposed to play a role in inducing the methylation of the DNA it may bind to. This type of inhibition is assumed to take place in animals and lower eukaryotes, while it has been directly observed in plants [Bao et al., 2004].

This short excursion into the world of post-transcriptional regulation is to prevent the possible thought that one is just in grasping distance of understanding the whole mechanism underlying the Central Dogma of Life. TF proteins probably play the major role in this game but yet unknown participants may be undisclosed any time, enlarging the necessary set of rules – on the other hand making the application of rules possible.

Since we remain in the still vast domain of regulation by TFs, we continue with some theoretical formalism which will be useful when explaining our method for motif identification.

Chapter 3

Theoretical requisites

Before discussing various representations of binding site motifs, we present some general theoretical concepts which will be of use in the following. From an information theoretical point of view, we encounter the concept of entropy and argue how to improve the statistics of a small sample using Bayesian inference. Passing into statistical physics, we draft shortly the ideas behind Monte Carlo methods and present the technique of simulated annealing as a tool for optimising also ”non-physical” systems.

3.1 Information theory

3.1.1 Entropy and Kullback-Leibler divergence

Shannon’s well-known definition of entropy [Shannon and Weaver, 1963] in the context of information theory states for any discrete probability distribution $p(x)$, describing the probability for a set of messages $\{x\}$ to be provided by some device, that

$$S[p] = -\sum_x p(x) \log p(x)$$

which keeps validity also for continuous distributions if one goes over to an integral formulation. It is analogous to the definition of entropy in statistical mechanics of canonical ensembles, where the probability of a configuration $\{x\}$ is related to the free-energy $F(\{x\})$, according to Boltzmann’s law

$$p(x) \propto \exp\left(-\frac{F(x)}{k_B T}\right).$$

The probability distribution $p(x)$ may also be related to another distribution $q(x)$, leading to the concept of relative entropy, defined as

$$H[p \mid q] = \sum_x p(x) \log \left(\frac{p(x)}{q(x)} \right) .$$

This functional was chosen to figure under the section of information theory, for it known since 1951 by the name of Kullback-Leibler divergence [Kullback and Leibler, 1951] and as such mostly found application in classical coding theory. Its physical interpretation in a statistical mechanical context was exemplified by [Qian, 2001] as being associated to the free-energy difference of a non-equilibrium system – with states distributed according to $p(x)$ – to its equilibrium – described by the distribution $q(x)$.

In the context of motif representation and identification, we will encounter the expression for the relative entropy as a probabilistic measure for the information content of an achieved result, the so-called information score.

3.1.2 Pseudocount regularisers for small-sample-statistics

Fitting a predictive model to experimental data has always been an important statistical task and the theory of Bayesian inference offers a useful framework for doing so. The very basics of the theory can be stated as relation between the *prior* distribution $P(\rho)$ of some model parameter ρ and the *posterior* distribution $P(\rho|S)$, where some observed data S is conditioning the probability.

The definition of the conditional probability $P(\rho|S)$ for the random variables ρ and S yields

$$\begin{aligned} & P(\rho|S)p(S) = P(\rho; S) = P(S; \rho) = P(S|\rho)p(\rho) \\ \Rightarrow \quad & P(\rho|S) = \frac{P(\rho)P(S|\rho)}{P(S)} = \frac{P(\rho)P(S|\rho)}{\int d\rho' P(\rho')p(S|\rho')} . \end{aligned} \tag{3.1}$$

In the present context, we are going to face the task of estimating actual occurrence frequencies

$$p_\alpha = \lim_{L\to\infty} \frac{n_\alpha}{L} \quad \text{with} \quad \alpha = \mathrm{A, G, C, T} ,$$

from finite and sometimes small genomic samples of length $|S| \ll L$ by counting out the occurrences n_α of the nucleotide α. The results are intrinsically flawed, the error increasing with decreasing sample sizes. Pictorially spoken, we cannot deduce reliable occurrences n_α for the four nucleotides in the extreme case of a very short sequence, as we cannot safely test a coin for being Laplacian by tossing it only a few times. In the above terminology, we want to find the probability of having a certain set of occurrence frequencies

$$P(\rho|S) \quad \text{with} \quad \rho = (p_\mathrm{A}, p_\mathrm{C}, p_\mathrm{G}, p_\mathrm{T}) ,$$

given the observation of a genomic sequence S. From this probability we can then compute a better estimate for the frequencies than from merely counting occurrences in the finite sample. Thus we evaluate

$$P_S(\alpha) \equiv P(\alpha|S) = \int d\rho\, \rho\, P(\rho|S)$$

by integration over the four dimensional probability space of ρ. The form of the *posterior* $P(\rho|S)$ is not clear *a priori*, but we can use Bayes' rule from (3.1) to find a description in more accessible terms.

$$P_S(\alpha) = \int d\rho \, \frac{p_\alpha \, P(\rho)P(S|\rho)}{\int d\rho' P(\rho')P(S|\rho')}\,, \tag{3.2}$$

where the likelihood $P(S|\rho)$ of generating the sequence S is clearly multinomial, i.e.

$$P(S|\rho) = \frac{[\sum_\alpha n_\alpha]!}{\prod_\alpha n_\alpha!} \cdot \prod_\alpha p_\alpha^{n_\alpha}\,, \tag{3.3}$$

if one assumes the nucleotide occurrence frequencies to be independent. It remains to find a suitable *prior* $P(\rho)$. For simplicity in the calculus, we assume it to be a Dirichlet distribution, although this might be criticised, as in [Stephens and Donnelly, 2003], to be an oversimplification of the problem. Nevertheless, the results obtained by assuming a Dirichlet prior appears rather plausible from a naive point of view.

Leaving the details of this computation to appendix A, we find as final result

$$P_S(\alpha) = \frac{n_\alpha + \beta_\alpha}{\sum_{\alpha'} (n_{\alpha'} + \beta_{\alpha'})}$$

with the free set of parameters β_α, mnemonically called *pseudocounts*, which can be chosen to give certain desired properties to $P_S(\alpha)$. We will not further discuss the philosophy of adjusting pseudocounts to sample sizes but will use the widely accepted assumption that

$$\beta_\alpha = 1 \,\forall\, \alpha$$

is an acceptable pseudocount for arbitrary sample sizes. Clearly, as $|S| \to \infty$, the impact of any finite β_α vanishes and for the case in which no observation was made, a pseudocount of one leads to the reasonable prediction of $p_\alpha = 0.25$ for all four nucleotides, being the best “zero knowledge” guess.

Whenever it comes to counting out nucleotide occurrences, we will make use of the additive pseudocount regulariser. For small samples it just slightly smooths the expected distribution of nucleotides, while it has no major effect when counting out large samples. Since many bioinformatics methods make use of such additive regularisers, it appeared useful to summarise its Bayesian origin here. From a biological point of view, we do not expect the occurrence frequency for any nucleotide to be exact zero, thus motivating the use of pseudocounts when evaluating any genomic data.

3.2 Statistical physics

3.2.1 Monte Carlo sampling on random fields

Representing multidimensional stochastic systems by a lattice of random variables leads to the notion of random fields, where the interdependence between different lattice elements plays an important role for its macroscopic behaviour. Random fields have become popular in physics when describing spin-systems and many-particle-systems or texture like structures, as polymer conglomerates or convecting media. In information theory, they find application when it comes to image treatment, satisfiability problems and in many other fields.

The aim of Monte Carlo sampling is to simulate the realisation of a random field, which in our context will be just a one dimensional lattice of N binding site positions a_k, $k = 1, \ldots, N$, in a set of nucleotide sequences. Like in the application of Monte Carlo methods – for the underlying theory we refer to [van Kampen, 1985] – on physical systems, one starts by guessing some initial configuration for the lattice elements and associate a value $H(\{a_k\})$ to the system, representing its energy. Then, one sequentially updates each element, keeping the new configuration $\{a_k\}'$ if yields a lower value for H. Yielding a higher value, it is accepted only according to a certain probability distribution, which may be adapted to the very problem. We will make use of the Ansatz by [N.Metropolis et al., 1953] and accept a new state raising H with probability

$$\exp\left(-\frac{H(\{a_k\}') - H(\{a_k\})}{k_B T}\right) .$$

3.2.2 Simulated annealing

Lowering the temperature T, while performing a Monte Carlo sampling, is referred to as simulated annealing – the name being motivated by the slow annealing processes in metallurgy. This strategy has been successfully applied to optimisation problems of various kinds, some of the more famous having been described by [Kirkpatrick et al., 1983]. The technique reduces the likelihood of converging towards a configuration with only locally minimal energy.

Chapter 4

Representation of regulatory motifs

Speaking of a binding-site motif is talking about diffuse patterns sharing a certain degree of similarity. Different models of representation have hence been elaborated and we describe the most common ones as well as some quantities which are useful to classify the descriptions. A short overview of common information theoretical terms is followed by the description of a physically motivated one.

4.1 Information theoretical models

On the information theoretical side, we present the description via the so called consensus sequence and via frequency matrices. From the latter, we deduce the information score of such a matrix, thus assessing the quality of an alignment.

4.1.1 Consensus sequences

The binding site for a TF is naturally described via its nucleotide sequence. Considering a set of binding site sequences, one will eventually observe that nucleotides at some positions may vary, while others remain the same in all sequences. Figure 4.1 shows experimentally reported binding sites of the FruR TF, involved in the carbon metabolism of Escherichia coli bacteria. Capital letters in the table denote the identifies binding sites. The most

aagccaaag	CTGAATCGATTTT	atgatttgg
cgttgcgag	CTGAATCGCTTAA	cctggtgat
gttagcgtg	GTGAATCGATACT	ttaccggtt
tagtcgatc	GTTAAGCGATTCA	gcaccttac

Table 4.1: Binding sites of the FruR TF in E.coli

straight forward representation would be by naming the sets of occurring nucleotides at each position. Using the IUPAC[1] symbols for nucleotides, summarised in table 4.2, we find as consensus sequence the motif shown in table 4.3.

Symbol	A	C	G	T	U	R	Y	S	W	K	M	B	D	H	V	N	-
Base	Adenine	Cytosine	Guanine	Thymine	Uracil	A or G	C or T	G or C	A or T	G or T	A or C	C or G or T	A or G or T	A or C or T	A or C or G	any base	gap

Table 4.2: IUPAC symbols for nucleic acids

aagccaaag	CTGAATCGATTTT	atgatttgg
cgttgcgag	CTGAATCGCTTAA	cctggtgat
gttagcgtg	GTGAATCGATACT	ttaccggtt
tagtcgatc	GTTAAGCGATTCA	gcaccttac
nnnnnnnnn	STKAAKCGMTWHW	nnnnnnnnn

Table 4.3: Consensus sequence of the FruR binding sites

Obviously, this description lacks some crucial information. First, nothing is stated about the relative occurrence frequencies of different bases at the same position. Second, no accurate statement is made on the importance of an individual position. Although one can assume the positional importance of a symbols to decrease with the amount of nucleotides it stands for, this remains a rather coarse distinction. To keep information on the occurrence frequencies, it might be more useful to refer to the motif's sequence logo[2]. This kind of motif representation was introduced by [Schneider and Stephens, 1990]. Figure 4.1 shows the corresponding logo for the collection of FruR binding sites. The height of an entry corresponds to its relative occurrence frequency in the collection of sequences and we can readily convert the logo to a consensus sequence.

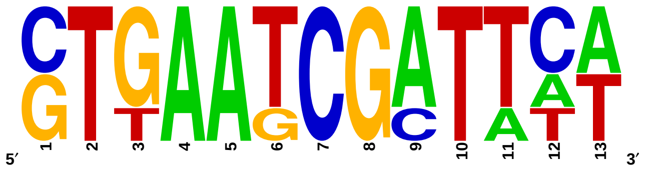


Figure 4.1: Sequence logo of FruR binding sites with relative nucleotide frequencies

[1] International Union of Pure and Applied Chemistry

[2] Created using the on-line service *http://weblogo.berkeley.edu/* by Gavin E. Crooks, Gary Hon, John-Marc Chandonia and Steven E. Brenner, Computational Genomics Research Group, Department of Plant and Microbial Biology, University of California, Berkeley.

4.1.2 Weight matrices and information score

The next step would then be to count the relative occurrences f^i_α of each nucleotide at each position of the motif in our collection of $N = 4$ sequences. Doing so, we calculate the entries of a frequency matrix

$$f^i_\alpha = \frac{n^i_\alpha + 1}{N + 4}$$

considering a pseudocount of one for each nucleotide. With n^i_α we denote the occurrence count of α at position i of the alignment, e.g. $n^1_{\mathrm{G}} = 2$ in the FruR alignment. Relating the frequency matrix to some nucleotide occurrence probability p_α, we reach the concept of weight matrices

$$w^i_\alpha = \log \frac{f^i_\alpha}{p_\alpha} \tag{4.1}$$

which has been successfully applied by [Stormo and Hartzell, 1989] for the description of protein binding cites. Assuming typical nucleotide occurrence probabilities for the upstream regions of the E.coli genome, i.e. $p_{\mathrm{A}} \approx p_{\mathrm{T}} \approx 0.3$ and $p_{\mathrm{G}} \approx p_{\mathrm{C}} \approx 0.2$, we find the weight matrix represented in table 4.4 for the FruR alignment we used as example so far. It is worth to note the usefulness of pseudocounts at this point. Due to the alignment being a rather small statistical sample, we might not observe a certain base at a certain position. Taking the logarithm in (4.1) leads thus to entries of $-\infty$ in the weight matrix, making the description useless.

A	-0.88	-0.88	-0.88	0.73	0.73	-0.88	-0.88	-0.88	0.51	-0.88	-0.18	-0.18	0.22
G	0.63	-0.47	0.92	-0.47	-0.47	0.22	-0.47	1.14	-0.47	-0.47	-0.47	-0.47	-0.47
C	0.63	-0.47	-0.47	-0.47	-0.47	-0.47	1.14	-0.47	0.22	-0.47	-0.47	0.63	-0.47
T	-0.88	0.73	-0.18	-0.88	-0.88	0.51	-0.88	-0.88	-0.88	0.73	0.51	-0.18	0.22

Table 4.4: Positional weight matrix for the FruR alignment

Related to the weight matrix is the so called information score – a measure for some alignment *not* to occur by chance. We begin with the probability of observing the set of positional occurrences $\{n^i_\alpha\}$ of the alignment, given the probabilistic model defined by the set $\{p_\alpha\}$. Considering the alignment of N sequences of length l, this is given by the product over multinomials

$$P(\{n^i_\alpha\} \mid \{p_\alpha\}) = \prod_{k=1}^{l} N! \prod_{\alpha} \frac{[p_\alpha]^{n^k_\alpha}}{n^k_\alpha} .$$

The multinomial after the first product sign is just the generation probability for a column in the alignment. To get the probability of generating the whole alignment, we have to build the product over all such columns. Writing the occurrences as $n^i_\alpha = N \cdot f^i_\alpha$ and making use of Stirling's approximation $\log(n!) \approx n \log n - n$, we find

$$\begin{aligned} P(\{n_\alpha^i\} \mid \{p_\alpha\}) &= \prod_{k=1}^{l} \exp(\log N!) \prod_\alpha \exp\left(N f_\alpha^k \log \frac{p_\alpha}{(N f_\alpha^k)!} \right) \\ &\approx \exp\left(-N \sum_{k,\alpha} f_\alpha^k \log \frac{f_\alpha^k}{p_\alpha} \right) \equiv \exp(-nI)\,, \end{aligned}$$

where we have implicitly defined the information score – or information content, as it is called in the derivation by [Schneider et al., 1986] – of an alignment described by f_α^i as

$$I = \sum_{k,\alpha} f_\alpha^k w_\alpha^k\,.$$

This quantity increases at the same time as the probability for the alignment described by $\{n_\alpha^i\}$ to occur by chance from the probabilistic model decreases. The information score gives hence the possibility to assess the quality of an alignment described by some weight matrix. Summing only over different nucleotides, we get a positional information score

$$I(k) = \sum_\alpha f_\alpha^k w_\alpha^k\,.$$

Sequence logos in chapter 7 are scaled with this positional information score, allowing the visual assessment of the importance of a specific position in the alignment.

4.2 Physical models

4.2.1 Energy matrices

Instead of considering the relative nucleotide frequencies in an alignment, one could use a description of the binding behaviour of a site to its corresponding TF. Assuming the nucleotides of a TF binding site to contribute independently of each other to a total free energy of binding of the whole protein, we can describe the interaction by an energy matrix with numerical entries ε_α^i. The free energy of binding of a specific TF-DNA interaction – describes by the binding sequence S of length l– is then just the sum over corresponding matrix entries times the proportionality factor $k_B T$, i.e.

$$E(S) = \sum_{i=1}^{l} \varepsilon_{\alpha_i^S}^i\,, \tag{4.2}$$

where we denote the nucleotide at position i in the sequence with α_i^S. Speaking about a threshold value which is to be provided by a specific binding, we will also refer to the free energy of binding as *discrimination energy.*

4.2.2 Berg & von Hippel theory

Presenting a statistical-mechanical theory for the functioning of TF driven regulatory systems, [Berg and von Hippel, 1987] make the plausible assumption that the set $M^\varphi = \{S_i^\varphi, S_2^\varphi, \dots\}$ of possible binding-sites corresponding to a specific factor φ is defined by a limited range around the discrimination energy E^φ, hence

$$M_\varphi = \{S : |E(S) - E^\varphi| \leq \Delta E^\varphi\} \ .$$

Finding the occurrence frequencies for a nucleotide at a specific position of the motif defined by this set – not knowing the corresponding sequences *a priori* – is equivalent to the problem of finding the occupation probabilities of energy-levels in a microcanonical ensemble of non-interacting particles. We can thus write

$$f_\alpha^i(E^\varphi) = \frac{\exp(-\lambda \varepsilon_\alpha^i)}{\sum_{\alpha'} \exp(-\lambda \varepsilon_{\alpha'}^i)}$$

with the energy matrix entries ε_α^i and the selection parameter λ which plays the role of an inverse temperature although obviously not being related to the one of the biological environment. It can rather be interpreted as coupling factor between the properties of a TF being represented by the energy matrix and properties of a binding-site motif motif, represented by the frequency matrix. Assuming the contribution of different sequence positions to be additive, one can also find the discrimination energy by evaluating the average

$$E^\varphi = \sum_{i=1}^{l} \sum_{\alpha} \varepsilon_\alpha^i f_\alpha^i \, .$$

We will recognise the fundamentals of this theory in chapter 5 when describing a method of energy matrix estimation. Especially the ensemble interpretation will be of great use, allowing the computation of statistical quantities from the free energy of binding.

Chapter 5

Identification of regulatory motifs

In this chapter, we discuss general schemes of motif identification, whereafter representatives are described explicitly. Finally, a new alignment algorithm is introduced, which will be evaluated in the following chapters.

5.1 General procedures

5.1.1 Pattern deduction

Given the abundance of genomic material, there has been made attempts to identifying regulatory motifs by deducing some set of sequences directly from the genome or by fitting a certain model to some known gene expression data. Such methods, as e.g. described in [Bussemaker et al., 2000] where the authors try to build a dictionary of words – hence deduce a genomic language – by comparing the occurrence probabilities of nucleotide strings in the genome in question, or the one developed by [Bussemaker et al., 2001] to fit a set of regulatory motifs to the experimental data of gene expression levels, try thus to deduce a descriptive pattern by fitting a motif model to the global model defined by the genome or the experimental expression data.

5.1.2 Sequence alignment

Another approach is to compare sub-sequences of a genome with each other. Assuming regulatory elements with the same function to share a similar sequence allows the attempt of identifying those similarity motifs through a local alignment. An important requisite to local alignment procedures is the choice of sequences where on which one wants to identify the regulatory elements. Each sequence is supposed to share the regulatory site *a priori*. Sequences not containing the motif are in general hard to identify and having too many of those makes the alignment difficult or even impossible. Known similarity search methods via alignment were developed e.g. by [Smith and Waterman, 1981], [Altschul et al., 1990] or [Lawrence et al., 1993]. Sequence alignment can be seen as an attempt to find the best local or global model describing a configuration of maximal similarity between different subsequences in the data.

5.1.3 Genome wide model matching

After building a motif model with one of the above procedures, the search for binding sites on the whole genome can begin. Knowing the weight or energy matrix $m^i_\alpha \in \{w^i_\alpha, \varepsilon^i_\alpha\}$ of a motif of length l, we can evaluate it on the genomic sequence G of length L by calculating the score

$$E(a) = \sum_{k=0}^{l-1} m^k_{G_{a+k}} \, ,$$

where G_i is the nucleotide at position i in the genome and $a \in [1; L - l + 1]$ is the first position of an assumed regulatory motif. In the case of energy matrices, the role of E is clearly the free energy of binding of a corresponding TF, thus we can easily decide if a site might act as TF binding site by finding a negative score. Having a weight matrix, no intrinsic threshold is given and one has to decide which score to take as limit for predictions.

5.2 Representative algorithms

5.2.1 Motif detection by fitting to expression data

Regulatory element detection using correlation with expression by [Bussemaker et al., 2001] is based on an assumption of additivity. The expression level of a gene $g \in G$ is to be constituted of

$$A_g \approx C + \sum_{\mu \in M} F_\mu N_{\mu g} = \log_2 \left(\frac{[\text{mRNA}]_{\text{obs}}}{[\text{mRNA}]_{\text{ref}}} \right) \tag{5.1}$$

where A_g is to be interpreted as the $\log_2$ of the fraction of mRNA abundance for the gene g in the observed organism, compared to a reference, hence the experimental input. $N_{\mu g}$ is the integer matrix of occurrences of motif $\mu \in M$ in the relevant upstream region of gene g. F_μ is the parameter to be fitted in this model and C represents a general baseline of expression for all g.

To simplify the situation, one rescales $A_g \rightsquigarrow a_g$ and $N_{\mu g} \rightsquigarrow n_{\mu g}$ using the transformation

$$\bullet_{[\mu]g} \rightsquigarrow \frac{\delta \bullet_{[\mu]g}}{\sqrt{|G| \langle \delta \bullet^2_{[\mu]} \rangle}}$$

with $\delta\bullet_{[\mu]g} = \bullet_{[\mu]g} - \langle\bullet_{[\mu]}\rangle$ and $\langle\delta\bullet^2_{[\mu]}\rangle = \mathrm{var}(\bullet_{[\mu]g})$, leading to a new set of parameters $F_\mu \rightsquigarrow f_\mu$ in the simplified model with normalised vectors of dimension $|G|$ with elements a_g and $n_{\mu g}$:

$$\mathbf{a} \approx \sum_{\mu\in M} f_\mu \mathbf{n}_\mu$$

One now aims to fit the model as best as possible to experimental data, by solving

$$\min_{f_\mu}\left\{\chi^2 = \left|\mathbf{a} - \sum_{\mu\in M} f_\mu \mathbf{n}_\mu\right|^2 \;:\; f_\mu \in \mathbb{R}\right\}$$

$$\sum_{\mu'\in M} f_{\mu'}(\mathbf{n}_\mu \cdot \mathbf{n}_{\mu'}) = \mathbf{n}_\mu \cdot \mathbf{a}$$

The initial parameters C and F_μ can be extracted from the fit via

$$C = \langle A\rangle - \sum_{\mu\in M} F_\mu \langle N_\mu\rangle,$$

$$F_\mu = f_\mu\sqrt{\frac{\langle\delta A^2\rangle}{\langle\delta N_\mu^2\rangle}},$$

which becomes obvious by inserting those relations into equation (5.1). Assuming $|M| = 1$ and thus having $f = \mathbf{n}_\mu \cdot \mathbf{a}$ from (5.1), the error simplifies to

$$\chi^2_\mu = \mathbf{a}^2 - 2(\mathbf{n}_\mu \cdot \mathbf{a}) \cdot (\mathbf{a} \cdot \mathbf{n}_\mu) + (\mathbf{n}_\mu \cdot \mathbf{a})^2 = 1 - (\mathbf{n}_\mu \cdot \mathbf{a})^2 \equiv 1 - \Delta\chi^2_\mu$$

Starting with a single motif in the model M and given some experimental data, e.g. from micro-array experiments in $\mathbf{n}_\mu$, one can compute the error reduction $\Delta\chi^2_\mu$ for all motifs one wants to consider and rank those by the largest $\Delta\chi^2_\mu$. After adding the motif which achieves the maximal $\Delta\chi^2_\mu$ to the set M, one computes the residual

$$\mathbf{a}' = \mathbf{a} - \sum_{\mu\in M} f_\mu \mathbf{n}_\mu$$

and its corresponding error reductions $\Delta{\chi'}^2_\mu = (\mathbf{n}_\mu \cdot \mathbf{a}')^2$ from all remaining motifs. Again adding the best “reducer”, the method is iterated and thus populating the model.

The significance of a motif can be measured statistically by evaluating the probability of it's $\Delta\chi^2_\mu$ to be maximal. One defines therefore

$$Z_\mu \equiv |G|^{1/2}(\mathbf{a} \cdot \mathbf{n}_\mu),$$

which describes the correlation of $\mathbf{a}$ and $\mathbf{n}_\mu$ in units of the variance of their product. Accepting the variance to be $|G|^{1/2}$, the mean zero and Z_μ to follow the normal distribution, the probability for it to be within the range of its maximum $Z_{max} = (|G|\Delta\chi^2_{\max})^{1/2}$ is

$$P(Z_\mu \in (-Z_{max}, +Z_{max})) = \frac{1}{\sqrt{2\pi}} \int\limits_{-Z_{max}}^{+Z_{max}} dZ \ \exp\left(-\frac{Z^2}{2}\right) \tag{5.2}$$

Since the computed set of motifs comprises $|M|$ elements and since all Z_μ are confined to the above interval, the composed probability for this event considering the whole set is just the product of all elemental probabilities, thus the $|M|^{\text{th}}$ power of equation (5.2).

The negated event hence describes the probability for yielding exactly the maximum Z_{max}:

$$P_c(|Z_\mu| = Z_{max}) = 1 - \left[\frac{2}{\sqrt{2\pi}} \int\limits_{0}^{Z_{max}} dZ \ \exp\left(-\frac{Z^2}{2}\right)\right]^{|M|} ,$$

where the integral was slightly simplified using that $\exp(-Z^2)$ is even. The significance of a motif thus scales with the absolute decrease of P_c.

5.2.2 Gibbs sampling

Originating form a statistical mechanical model applied to the description of image data, the algorithm known as “Gibbs sampler” was designed to allow a restoration of “defective” image regions by Bayesian means [Geman and Geman, 1984]. The idea behind the Gibbs sampler is to find the marginal probability density $f(x)$ of a random variable x which occurs in the joint density $f(x, y_1, \dots, y_n)$, i.e.

$$f(x) = \int \cdots \int f(x, y_1, \dots, y_n) \ dy_1 \cdots dy_n \tag{5.3}$$

Of course the problem is uninteresting if the joint density is given such as the integral is (easily) solvable. Whereas if it is hardly possible to solve (5.3) in a closed form or if the joint density is unknown, one has to use other means. We will see in the following that, for the latter case, Gibbs Sampling even allows to approximate the joint density with arbitrary precision. In [Casella and George, 1992], from where we take the elementary description in this section, a structured introduction comes along with illustrative examples.

The tools for successful sampling are the conditional probabilities relating the random variables $x, y_1, \dots, y_n$. From those we can both sample the marginal distributions and reconstruct the joint density, or rather approximate it by simulation draws. Here distribution and density is used in the same context, as the method holds for both. In practise, either the conditional densities are rather easy to construct or they are given within the formulation of the problem, as is the case in motif sampling on DNA. From our toolbox of distributions, we can now make draws for the variables, setting the initial conditions $Y_1^0, \dots, Y_n^0$ arbitrarily. Thus one draws

$$X^i \sim f(x \mid y_1 = Y_1^i, \dots, y_n = Y_n^i),$$

$$Y_j^{i+1} \sim f(y_j \mid x = X^i, y_1 = Y_1^{i+1}, \dots, y_{j-1} = Y_n^{i+1}, y_j = Y_1^i, \dots, y_n = Y_n^i),$$

and by iterating this, a Markov chain is created, as each set of variables $(X^i, Y_1^i, \dots, Y_n^i)$ only depends explicitly on the previous one. With a collection $\mathbb{S}$ of m such sets, we can now approximate the expectation value of of the conditional probabilities and thus approximate the marginal distributions:

$$E[f(x \mid y_1, \dots, y_n)] = \frac{1}{m} \sum_{i=1}^{m} f(x \mid y_1, \dots, y_n) = \frac{1}{m} \sum_{i=1}^{m} X^i,$$

and in the limit $m \to \infty$ this leads to the integral

$$\int dy_1 \cdots \int dy_n f(x \mid y_1, \dots, y_n) f(y_1) \cdots f(y_n) = f(x)$$

which is just the marginal we wanted. For long enough Markov chains, we hence reconstruct the desired distribution. From Monte Carlo theory we know that the reliability of the draws requires a certain *burn-in period*, hence we might choose to drop the first sets of our collection $\mathbb{S}$.

So far, the solution seemed rather straight forward, but one has to remember that the convergence of the Markov chain, especially the answer to the question if it does so at all, highly depends on the given conditional densities and has to be analysed for every problem individually. Moreover, the *burn-in-period* is likely to depend on the choice of initial conditions, which thus have to be set with care. Further details of these and related problems shall not be discussed here, and we refer to the literature, e.g. by [Walsh, 2004].

Given a set $\mathbb{S} = \{S_1, S_2, \dots, S_N\}$ of N DNA sequences, we want to align them to extract motif patterns of a given length l. For simplicity, we assume all sequences of $\mathbb{S}$ to have the same length L. The method described in [Lawrence et al., 1993] is tailored for amino acid coding residual sequences and can thus be easily applied to a sequence of nucleotides. Beside $\mathbb{S}$, two structures are being used in the algorithm: the first one consists of the occurrence probabilities $q_{i,\alpha}$ for the bases at each position of the pattern – with $i = 1, \dots, l$, $\alpha = \mathrm{A, G, C, T}$ – and the background probabilities p_α describing occurrences in the non-pattern regions of the sequences; the second is the alignment itself in form of the starting position a_k of the pattern in each sequence, i.e. $k = 1, \dots, N$.

In terms of the above formalism, we can identify the second set (alignments) with the variables, we want to reconstruct the distributions for, while the first set (probabilities) represents the conditions. The steps are then as follows:

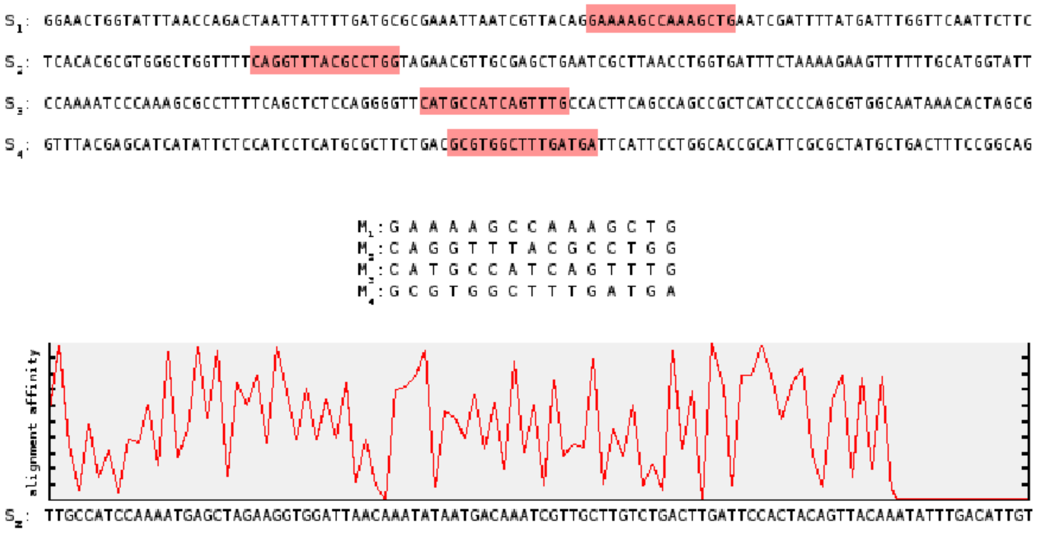
Initialisation: Random positions a_k are chosen for each sequence.

1. *Predictive update step:* The sequence S_z is removed from $\mathbb{S}$, where z can be chosen arbitrarily or in a specific order covering all elements of $\mathbb{S}$. For the remaining set we calculate the p_α and the $q_{i,\alpha}$ according to

   $$p_\alpha = \frac{n^b_\alpha + 1}{L(N-1)+4} \quad \text{and} \quad q_{i,\alpha} = \frac{n^m_{i,\alpha} + 1}{N-1+4}$$

   where occurrences in background and model are denoted with n^b and n^m, respectively. As discussed before, we settle for a simple correction assuming a pseudocount of one for each of the for nucleotides.

2. *Sampling step:* Every segment x of length l in S_z is being considered as instance of the pattern and we calculate the probabilities P_x and Q_x of constructing x from the p_α, respectively from the $q_{i,\alpha}$. Those segments are then assigned the weight $A_x = Q_x/P_x$ describing the likelihood of x to be generated by the actual description and before putting z back into $\mathbb{S}$, we pick a new a_z randomly, but according to the weights.



```
S1: GGAACTGGTATTTAACCAGACTAATTATTTTGATGCGCGAAATTAATCGTTACAGGAAAAGCCAAAGCTGAATCGATTTTATGATTTGGTTCAATTCTTC
S2: TCACACGCGTGGGCTGGTTTTCAGGTTTACGCCTGGTAGAACGTTGCGAGCTGAATCGCTTAACCTGGTGATTTCTAAAAGAAGTTTTTTGCATGGTATT
S3: CCAAAATCCCAAAGCGCCTTTTCAGCTCTCCAGGGGTTCATGCCATCAGTTTGCCACTTCAGCCAGCCGCTCATCCCCAGCGTGGCAATAAACACTAGCG
S4: GTTTACGAGCATCATATTCTCCATCCTCATGCGCTTCTGACGCGTGGCTTTGATGATTCATTCCTGGCACCGCATTCGCGCTATGCTGACTTTCCGGCAG

M1:G A A A A G C C A A A G C T G
M2:C A G G T T T A C G C C T G G
M3:C A T G C C A T C A G T T T G
M4:G C G T G G C T T T G A T G A

Sz: TTGCCATCCAAAATGAGCTAGAAGGTGGATTAACAAATATAATGACAAATCGTTGCTTGTCTGACTTGATTCCACTACAGTTACAAATATTTGACATTGT
```

Figure 5.1: Gibbs sampling on S_z with the model defined by M_1 to M_4

Figure 5.1 visualises those steps. The current alignment with motif length $l = 15$ is represented by the coloured boxes in S_1 to S_4, each of length $L = 100$, a_1 to a_4 thus being the starting positions of those boxes on the sequences. From the assumed model, the affinity A_x is computed for each segment of length l starting at $a_z = 1 \ldots 86$. The $q_{i\alpha}$

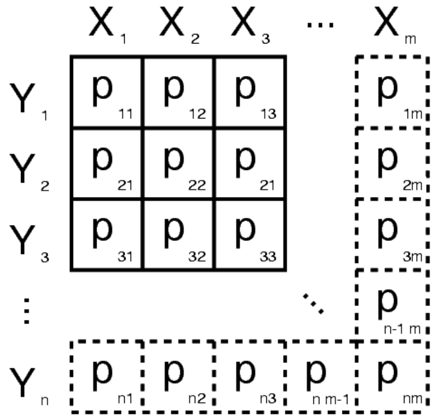


Figure 5.2: Array of conditional probabilities for the events X_i and Y_j with $p_{ij} \leq 0$ and $\sum_{ij} p_{ij} = 1$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The sampling on DNA can be mapped to a similar description.

are calculated from the nucleotides comprising the model and the p_α are evaluated from all the others.

If, at some iteration, a set of "correct" alignments is chosen, i.e. some of the a_k correspond to a highly non-background-pattern, the algorithm will tend to lock the remaining alignments to satisfy the "correct" pattern. That way, the method converges to a (although possibly local) optimum of alignment. A slightly more rigorous proof can be found in [Casella and George, 1992], where convergence is shown for finite dimensional conditional arrays, as illustrated in 5.2. To avoid getting stuck on a local optimum, we can add a step, say every m^{th} iteration, where all alignments are temporarily shifted some bases to both directions, to check if this yields to better A_x.

In the former description, one has to use a given length for the pattens. Making a set of runs with different l each is the simplest way to search for "arbitrary" pattern lengths. Although, one has then to find arguments for one of the results.

Albeit not knowing the motif patterns, we need to feed the algorithm with coarse sequences, associated with binding, making the method differ from zero-knowledge-approaches. The strength of the procedure is, as stated in the literature, the usual speed of convergence with which one gets a solution to the highly non-trivial alignment-problem.

5.2.3 Energy matrix estimation

The recently developed Quadratic Programming Method for Energy Matrix Estimation by [Djordjevic et al., 2003] makes a different approach to the problem of uncovering binding-site motifs. One rather looks at the binding free energy of a transcription factor bound to the DNA molecule, which can be expanded in interaction terms of different order, all depending on the sequence S of length L:

$$E(S) = \sum_{i}^{L} \sum_{\alpha}^{4} \varepsilon_{\alpha}^{i} S_{\alpha}^{i} + \sum_{i,j}^{L} \sum_{\alpha,\beta}^{4} J_{ij}^{\alpha\beta} S_{\alpha}^{i} S_{\beta}^{j} + \sum_{i,j,k}^{L} \sum_{\alpha,\beta,\gamma}^{4} Q_{ijk}^{\alpha\beta\gamma} S_{\alpha}^{i} S_{\beta}^{j} S_{\gamma}^{k} + \ldots \tag{5.4}$$

where all subscripts are integers starting at one and the greek letters are control variables for the four base types at arbitrary convention. Since the involved tensors gain in rank in every sum, making computation more and more laborious, and as one assumes the corrections by the higher order terms to be small, only the linear approximation is used for modelling. s_{α}^{i} switches the interaction, being one if the i^{th} base is α and zero otherwise. It is convenient to think in terms of matrices, leading to

$$E(S) \approx \operatorname{tr}(\boldsymbol{\varepsilon}^T \cdot \mathbf{s}) = \operatorname{tr}\left[\begin{pmatrix} \varepsilon_A^1 & \varepsilon_G^1 & \varepsilon_T^1 & \varepsilon_C^1 \\ \vdots & \vdots & \vdots & \vdots \\ \varepsilon_A^L & \varepsilon_G^L & \varepsilon_T^L & \varepsilon_C^L \end{pmatrix} \cdot \begin{pmatrix} S_A^1 & \ldots & S_A^L \\ S_G^1 & \ldots & S_G^L \\ S_T^1 & \ldots & S_T^L \\ S_C^1 & \ldots & S_C^L \end{pmatrix}\right]$$

This can equally be described by a computationally less exhaustive inner product of the vectors

$$\boldsymbol{\varepsilon} = (\varepsilon_A^1, \varepsilon_G^1, \varepsilon_T^1, \varepsilon_C^1; \ldots; \varepsilon_A^L, \varepsilon_G^L, \varepsilon_T^L, \varepsilon_C^L)^T$$

and

$$\mathbf{s} = (S_A^1, S_G^1, S_T^1, S_C^1; \ldots; S_A^L, S_G^L, S_T^L, S_C^L)^T$$

with the approximated sequence energy

$$E(S) \approx \boldsymbol{\varepsilon}^T \cdot \mathbf{s}$$

since all off-diagonal elements in the above matrix computation are irrelevant due to the tracing. In the following, only the vectorial representation will be used for $\mathbf{s}$ and $\boldsymbol{\varepsilon}$, although we will keep using the common term of energy matrix while speaking of a vectorial representation.

The problem of finding the free binding energy can now be tackled using the method suggested in [Djordjevic et al., 2003]. A corresponding *thought experiment* is the following:

One mixes a large number of randomly generated DNA sequences of length L to a solution with known TF concentration, hence with fixed chemical potential μ. The probability to generate the sequence S be P_S. The likelihood of observing exactly a set $\mathcal{O}$ of sequences and no other is

$$\mathcal{P} = \prod_{S \in O} [\gamma P_S f(E(S) - \mu)] \prod_{S' \notin O} [1 - \gamma P_{S'} f(E(S') - \mu)] \tag{5.5}$$

where $f(E-\mu)$ is the Fermi distribution. The factor γ describes the extraction probability of a bound sequence. The aim of the algorithm is to maximise the likelihood $\mathcal{P}$. Making use of the expansion of $\log(1-x) \approx -x$, one can simplify equation (5.5) to

$$\begin{aligned} \mathcal{P} &= \prod_{S \in O} [\gamma P_S f(E(S) - \mu)] \exp\left(\sum_{S' \notin O} \log[1 - \gamma P_{S'} f(E(S') - \mu)] \right) \\ &\approx \prod_{S \in O} [\gamma P_S f(E(S) - \mu)] \exp\left(\sum_{S' \notin O} [-\gamma P_{S'} f(E(S') - \mu)] \right) \end{aligned} \tag{5.6}$$

To get rid of the product and the exponential function, it is rather convenient to consider the logarithm of this probability, referred to as the logarithmic likelihood $\mathcal{L} = \log \mathcal{P}$. From equation (5.6) we hence get

$$\mathcal{L} = N_S \log \gamma + \sum_{S \in O} \log[P_S f(E(S) - \mu)] - \gamma \sum_{S' \notin O} [P_{S'} f(E(S') - \mu)]$$

Maximising $\mathcal{L}$ corresponds to maximising the probability of extracting bound sequences, hence to identify those relevant during transcription, the general aim of all here presented algorithms.

A short excursion shows why one assumes the binding probability to be Fermi-Dirac distributed. One considers the simple reaction equation describing the binding of a TF:

$$\text{TF} \; + \; \text{DNA} \; \underset{K_{diss}}{\overset{K_{bind}}{\rightleftharpoons}} \; \text{TF} \circ \text{DNA}$$

This represents a pair of coupled ordinary first-order differential equations and the system can be interpreted as one of two states, bound and unbound, separated by the free energy of binding E. One can now look at the steady-state of the bound complex' concentration

$$\partial_t[\text{TF} \circ \text{DNA}] = K_{bind}[\text{TF}][\text{DNA}] - K_{diss}[\text{TF} \circ \text{DNA}] \equiv 0$$

leading to

$$\frac{K_{bind}}{K_{diss}} = \frac{[\text{TF} \circ \text{DNA}]}{[\text{TF}][\text{DNA}]} = K \cdot \exp(-\beta E) \tag{5.7}$$

where equality to the right comes from the two-state model. $\beta = k_B T$ is the Boltzmann factor and E stands for the free energy of binding, while K is a proportionality constant with unit of an inverse concentration.

Now the probability for the DNA sequence S to be bound to a TF is given by

$$P_{bound}(S) = \frac{[\mathrm{TF} \circ \mathrm{DNA}]}{[\mathrm{TF} \circ \mathrm{DNA}] + [\mathrm{DNA}]}$$

into which one can insert the statement of equation (5.7), thus

$$\begin{aligned} P_{bound}(S) &= \left(1 + \frac{[\mathrm{DNA}]}{[\mathrm{TF} \circ \mathrm{DNA}]}\right)^{-1} = \left(1 + \frac{\exp(\beta E(S))}{K \cdot [\mathrm{TF}]}\right)^{-1} \\ &= \frac{1}{1 + \exp[\beta(E(S) - \mu)]} \end{aligned} \tag{5.8}$$

with the chemical potential $\mu = k_B T \log(K \cdot [\mathrm{TF}])$, just leading to the Fermi distribution.

To further simplify equation (5.6), one considers the border case of all TFs being bound. The most radical requirement to satisfy this claim is $T \to 0$ or equivalently $\beta \to \infty$. In this limit, the Fermi distribution goes over into the Heaviside distribution $\Theta(E - \mu)$. Further on, one can assume the energies $E(S')$ of unobserved sequences to be distributed according to $\rho_\varepsilon(E)$, allowing the notation

$$\sum_{S' \notin O} P_{S'} f(E(S') - \mu) = \int_{-\infty}^{\infty} dE \rho_\varepsilon(E) f(E - \mu) \xrightarrow{T \to 0} \int_{-\infty}^{\mu} dE \rho_\varepsilon(E). \tag{5.9}$$

The continuous density function ρ_ε can, on its part, be approximated by a Gaussian distribution as long as one is close to the mean energy. This is assumed to hold for the set of unobserved sequences. Keeping the temperature finite, this leads to a simplified likelihood function

$$\mathcal{L} = N_S \log \gamma + \sum_{S \in O} \log[P_S f(E(S) - \mu)] - \gamma \int dE \rho_\varepsilon(E) f(E - \mu), \tag{5.10}$$

which is yet to be maximised in the $(\varepsilon, \mu, \gamma)$ space.

$$\partial_{\varepsilon^i_\alpha} \mathcal{L} = -N_S \log \gamma \sum_{S \in O} [1 - f(E(S) - \mu)] \cdot \beta S^i_\alpha + \gamma \int dE f(E - \mu) \partial_{\varepsilon^i_\alpha} \rho_\varepsilon(E) \equiv 0$$

$$\partial_\mu \mathcal{L} = N_S \log \gamma \sum_{S \in O} [-f(E(S) - \mu)] \cdot \beta - \gamma\beta \int dE \rho_\varepsilon(E) f(E - \mu)[1 - f(E - \mu)] \equiv 0$$

$$\partial_\gamma \mathcal{L} = \frac{N_S}{\gamma} - \int dE \rho_\varepsilon(E) f(E - \mu) \equiv 0\,.$$

The extraction factor γ turns out to be independent of the other variables

$$\gamma = \frac{N_s}{\int dE \rho_\varepsilon(E) f(E - \mu)}$$

and inserting this result into (5.10) yields a simplified maximisation problem

$$\max_{\boldsymbol{\varepsilon},\mu}\left\{N_S\left(\log N_S-\int\limits_{-\infty}^{\mu}dE\rho_\varepsilon(E)\right)\right\},$$

which – since N_S is constant – is equivalent to the evaluation of

$$\min_{\boldsymbol{\varepsilon},\mu}\left\{\int\limits_{-\infty}^{\mu}dE\rho_\varepsilon(E)=\operatorname{erf}\left(\frac{\mu-\bar{E}}{\sigma}\right):\boldsymbol{\varepsilon}\in\mathbb{R}^4\times\mathbb{R}^L,\ \mu\in\mathbb{R}\ :\ E(S)\leq\mu\ \forall\ S\in O\right\}$$

Here we finally assumed ρ_ε to be Gaussian with mean $\bar{E}$, which can be arbitrarily chosen by shifting the energy scale, since this doesn't affect the minimisation problem. Setting further

$$\mu=\max_{S\in O}E(S)\,,$$

to ensure the observed states states to be bound, the problem can be reduced to minimising the variance of the Gauss distribution. It is hence to find

$$\left\{\boldsymbol{\varepsilon}\in\mathbb{R}^4\times\mathbb{R}^L\ :\ \sigma^2(\boldsymbol{\varepsilon})=\min_{\boldsymbol{\varepsilon}}\sigma^2(\boldsymbol{\varepsilon})\right\}$$

Rescaling all energies to units of the shifted chemical potential $\mu-\bar{E}\equiv-1$ leads to the task

$$\begin{cases}\underset{\boldsymbol{\varepsilon}}{\text{minimise}} & \sigma^2(\boldsymbol{\varepsilon})=\sum\limits_{i,\alpha}\varepsilon^i_\alpha P(\varepsilon^i_\alpha)\\ \text{subject to} & E(S)=\boldsymbol{\varepsilon}^T\cdot\mathbf{s}\leq-1\ \forall\ S\in O\end{cases}\tag{5.11}$$

The probabilities $P(\varepsilon^i_\alpha)$ are yielded from a statistical model in the following, where we will see that they correspond to the probabilities $P(\alpha_i)$ of observing the nucleotide α at position i. The problem of estimating the energy vector $\boldsymbol{\varepsilon}$ – with entries ε^i_α – was thus reduced to a quadratic optimisation problem, which can be handled by the method of *quadratic programming*. Let us first summarise the system of equations (5.11) as

$$\begin{cases}\underset{\boldsymbol{\varepsilon}}{\text{minimise}} & \frac{1}{2}\boldsymbol{\varepsilon}^T\mathbf{P}\boldsymbol{\varepsilon}\\ \text{subject to} & \boldsymbol{\varepsilon}^T\cdot\mathbf{S}+\mathbf{1}\leq\mathbf{0}\end{cases}$$

with the matrix $\mathbf{S}$ comprising the row vectors $\mathbf{s}_{(j)}$ as entries, corresponding to the j^{th} of the N_S extractions each, thus

$$\mathbf{S}=(\mathbf{s}^T_{(1)},\mathbf{s}^T_{(2)},\ldots,\mathbf{s}^T_{(N_\mathbf{s})})\,.$$

$\mathbf{1}$ and $\mathbf{0}$ are the N_S dimensional vectors of ones and zeros, respectively. $\mathbf{P}$ can be interpreted as the Hessian of the variance . Since $p_\alpha\leq0$ and $\sum p_\alpha=1$, $\mathbf{P}$ is at least positive semi-definite, leading to a convex optimisation problem. Any identified optimum is hence a global one. Lagrange's optimisation with constraints requires the system – introducing the vector of multipliers $\boldsymbol{\lambda}$ – to be

$$\begin{cases} \underset{\boldsymbol{\varepsilon},\boldsymbol{\lambda}}{\text{minimise}} & \frac{1}{2}\boldsymbol{\varepsilon}^T \mathbf{P} \boldsymbol{\varepsilon} + \boldsymbol{\lambda}\left(\boldsymbol{\varepsilon}^T \cdot \mathbf{S} + \mathbf{1}\right)^T \\ \text{subject to} & \mathbf{P} \cdot \boldsymbol{\varepsilon} + \mathbf{S} \cdot \boldsymbol{\lambda} = \mathbf{0} \wedge \boldsymbol{\lambda} \geq 0 \end{cases} \tag{5.12}$$

Inserting $\boldsymbol{\varepsilon} = -\mathbf{P}^{-1} \cdot \mathbf{S}\boldsymbol{\lambda}$ from the new constraint back into the optimisation leads to the dual form of the problem

$$\begin{cases} \underset{\boldsymbol{\lambda}}{\text{maximise}} & -\frac{1}{2}\boldsymbol{\lambda}^T \cdot \mathbf{S}^T \cdot \mathbf{P}^{-1} \cdot \mathbf{S} \cdot \boldsymbol{\lambda} + \boldsymbol{\lambda}^T \cdot \mathbf{1} \\ \text{subject to} & \boldsymbol{\lambda} \geq \mathbf{0} \end{cases} \tag{5.13}$$

This dual problem is equivalent to the primal one [Nash and Sofer, 1996] and has only $\boldsymbol{\lambda}$ left as free parameter. Efficient numerical methods are available for solving this problem. In the current application, we use the solver package kindly contributed by Klaus Schittkowski [Schittkowski, 2003].

Zero entries in the sequence matrix $\mathbf{S}$ lead inevitably to the loss of some information, "flattening" the manifold to find the optimum on, so we allow some shifting operation to act on the sequences, defining for the entries of each sequence vector $\mathbf{s}_{(j)}$

$$\left(\hat{\mathbf{s}}_{(j)}\right)_\alpha^i \equiv \left(\mathbf{s}_{(j)}\right)_\alpha^i - p_\alpha \, . \tag{5.14}$$

Such a transformation from $\mathbf{S}$ to $\hat{\mathbf{S}}$ does not affect the optimisation problem but just rescales the chemical potential, which was arbitrarily set to unity, as in equation (5.12), and we solve the problem with this shifted version of the sequence vector.

To reconstruct the estimated energy matrix we have to evaluate

$$\boldsymbol{\varepsilon} = -\mathbf{P}^{-1} \cdot \hat{\mathbf{s}} \cdot \boldsymbol{\lambda}$$

and we obtain the energy matrix $\boldsymbol{\varepsilon}$ in terms of $\mu \equiv -1$.

Solving the dual problem yields a great computational benefit when optimising via quadratic programming. The matrix of the dual quadratic form is $\mathbf{S}^T \cdot \mathbf{P}^{-1} \cdot \mathbf{S}$ and therewith of rank N_S instead of $4L$ for P in the primal problem. In general, we have $N_S \ll 4L$ and although one has to perform a matrix inversion and two multiplications, this has only to be done once, while the optimisation procedure evaluates the quadratic form numerous times.

Figure 5.3 visualises the idea behind the likelihood maximisation at some finite temperature. The domain of overlapping Fermi-Dirac and Gauss distributions represents the probability of having unobserved binders. By minimising the variance of the Gaussian, this probability diminishes, maximising the likelihood of only finding the observed set of sequences in the thought experiment.

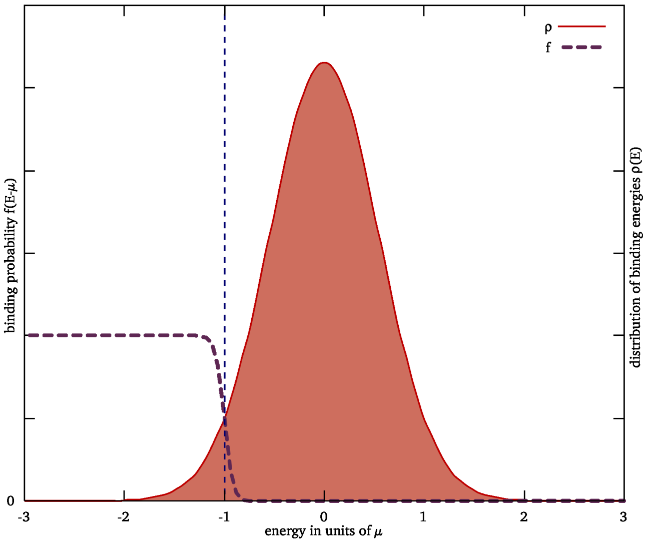


Figure 5.3: Distribution of binding energies and sequence binding probability

One has now to calibrate respective to a known set of binding sequences, in order to find an estimate for the binding-free-energies $\boldsymbol{\varepsilon}$. With this estimate, it is possible to find other sequences subject to the constraint set by the chemical potential by evaluating the energy matrix on the whole genome.

Statistical model

We consider a sequences of the kind $S = (\alpha_1, \alpha_2, \ldots, \alpha_L)$ of length L with elements $\alpha_i \in \mathcal{A}$ from the genomic alphabet $\mathcal{A} = \{\mathrm{A}, \mathrm{G}, \mathrm{C}, \mathrm{T}\}$. From here we build a statistical model, allowing the computation of the variance we want to minimise in order to perform the maximisation of (5.5).

Assuming the occurrence probabilities of the bases in the genome to be independent of position and neighbourhood (independent base statistics), we get a straight forward model genomic background carrying a generation probability for a sequence S of

$$P(S) = \prod_{i=1}^{L} \prod_{\alpha} P(\alpha)^{\delta_{\alpha_i \alpha}},$$

where $P(\alpha)$ denotes the probability of finding base α at an arbitrary position in the genome. Interpreting this probability as the Boltzmann factor of the corresponding chemical potential in a grand-canonical ensemble, i.e.

$$P(\alpha) = \exp(\beta \mu_\alpha),$$

we can identify a partition function in the space of possible sequence "states"

$$\mathfrak{Z}(\beta) = \sum_S \exp\left(-\beta[E(S) - \mu_S]\right) \equiv \sum_{\alpha_i} \cdots \sum_{\alpha_L} \prod_{i=1}^{L} P(\alpha) \cdot \exp(-\beta \varepsilon^i_\alpha).$$

It was implicitly assumed that $\mu_S = \mu_{\alpha_1} + \cdots + \mu_{\alpha_L}$. Further on, the partition function factorises, when using independent base statistics, to

$$\mathfrak{Z}(\beta) = \prod_{i=1}^{L} \sum_{\alpha} P(\alpha) \cdot \exp(-\beta \varepsilon^i_\alpha).$$

We now want to use a two-point statistical model for the genomic background, since an independent nucleotide description turns out to be too inaccurate [Djordjevic et al., 2003]. The partition function of a sequence $S = (\alpha_1, \alpha_2, \ldots, \alpha_L)$ depends hence on the conditional probabilities

$$P(\alpha_i \mid \alpha_{i-1}) = P(\text{find } \alpha_i \text{ at position } i \text{ given a preceding } \alpha_{i-1})$$

where subscripted α represent the base at a specific position. Note however that the conditional probabilities are independent of the absolute position of α_i in the sequence, but only depend on the actual and preceding base. A sequence is thus regarded as the realisation of a Markov chain, leading to

$$\mathfrak{Z}(\beta) = \sum_{\alpha_i} \cdots \sum_{\alpha_L} \prod_{i=1}^{L} P(\alpha_i \mid \alpha_{i-1}) \cdot \exp(-\beta \varepsilon^i_{\alpha_i}) \quad \text{with} \quad P(\alpha_1 \mid \alpha_0) \equiv P(\alpha_1).$$

The L sums over each possible sequence element α_i can be summarised by the sum over all possible sequences S and therewith differentiating $\log \mathfrak{Z}$ with respect to β leads to

$$\begin{aligned}
\partial_\beta \log \mathfrak{Z}(\beta) &= \frac{\partial_\beta \mathfrak{Z}(\beta)}{\mathfrak{Z}(\beta)} \\
&= \frac{\sum_S \sum_{j=1}^{L} \varepsilon^j_{\alpha_j} \prod_{i=1}^{L} P(\alpha_i \mid \alpha_{i-1}) \exp(-\beta \varepsilon^i_{\alpha_i})}{\sum_S \prod_{i=1}^{L} P(\alpha_i \mid \alpha_{i-1}) \exp(-\beta \varepsilon^i_{\alpha_i})} \\
[\ldots]_{\beta=0} &= \sum_S \sum_{j=1}^{L} \varepsilon^j_{\alpha_j} \prod_{i=1}^{L} P(\alpha_i \mid \alpha_{i-1}) \\
&= \sum_{\alpha \in \mathcal{A}} \sum_{j=1}^{L} \varepsilon^j_{\alpha_j} P(\alpha_j) \equiv \langle E \rangle . \qquad (5.15)
\end{aligned}$$

We make use of the short-hand notation $P(\alpha_j)$, meaning the *position dependent* probability to find α at position j in the sequence. This result can be identified with the mean free energy of binding, whose positional dependence we thus introduce as

$$\bar{\varepsilon}_j \equiv \sum_{\alpha\in\mathcal{A}} \varepsilon^j_{\alpha_j} P(\alpha_j)\,. \tag{5.16}$$

Calculating the second derivative of $\mathfrak{Z}$, we obtain

$$\begin{aligned}
\partial_\beta{}^2 \log \mathfrak{Z}(\beta) &= \frac{\partial_\beta{}^2\mathfrak{Z}(\beta)}{\mathfrak{Z}(\beta)} - \left(\frac{\partial_\beta \mathfrak{Z}(\beta)}{\mathfrak{Z}(\beta)}\right)^2 \\
&= \frac{\sum_S \sum_{i,j} \varepsilon^i_{\alpha_i}\varepsilon^j_{\alpha_j} \prod_k P(\alpha_k \mid \alpha_{k-1}) \exp(-\beta\varepsilon^k_{\alpha_k})}{\sum_S \prod_k P(\alpha_k \mid \alpha_{k-1}) \exp(-\beta\varepsilon^k_{\alpha_k})} \cdots \\
&\quad - \left(\frac{\sum_S \sum_i \varepsilon^i_{\alpha_i} \prod_k P(\alpha_k \mid \alpha_{k-1}) \exp(-\beta\varepsilon^k_{\alpha_k})}{\sum_S \prod_k P(\alpha_k \mid \alpha_{k-1}) \exp(-\beta\varepsilon^k_{\alpha_k})}\right)^2
\end{aligned} \tag{5.17}$$

$$\begin{aligned}
[\ldots]_{\beta=0} &= \sum_S \sum_{i,j} \varepsilon^i_{\alpha_i}\varepsilon^j_{\alpha_j} \prod_k P(\alpha_k \mid \alpha_{k-1}) \ldots \\
&\quad - \left(\sum_S \sum_i \varepsilon^i_{\alpha_i} \prod_k P(\alpha_k \mid \alpha_{k-1})\right)^2 \\
&= \sum_{\alpha,\beta} \sum_{i,j} (\varepsilon^i_{\alpha_i} - \bar{\varepsilon}_i)(\varepsilon^j_{\beta_j} - \bar{\varepsilon}_j) P(\alpha_i;\beta_j) \equiv \sigma^2
\end{aligned} \tag{5.18}$$

with the joint probability $P(\alpha_i;\beta_j)$ of having a sequence with α_i at position i and β_j at position j. The formula in (5.18) can be recognised as the variance of the binding energies with respect to the probabilistic model defined by P, which we developed explicitly for the Markovian case with nearest neighbour dependency.

Narrow binding energy window

Here, we argue how to approximate the discrete energy distribution by a Gaussian, as mentioned after equation (5.9). Discrete binding energies can be represented by the Dirac distribution when averaging over all (unobserved) sequences s.

$$\rho(E) = \langle \delta(E - \varepsilon \cdot s) \rangle_S$$

This can be rewritten in integral form, using the orthogonality property of the Fourier kernel

$$\rho(E) = \frac{1}{2\pi} \sum_S P(S) \int_{-\infty}^{+\infty} d\omega \exp\left(i\omega[E - \boldsymbol{\varepsilon}\cdot\mathbf{s}]\right) \equiv \frac{1}{2\pi} \int_{-\infty}^{+\infty} d\omega \exp\left(i\omega E + \log Y(\omega)\right)$$

where the function $Y(\omega)$ was implicitly defined as

$$Y(\omega) \equiv \sum_S P(S)\, \exp(-i\omega\boldsymbol{\varepsilon}\cdot\mathbf{s})\,.$$

Complex analysis allows an assession of the integral representation. Therefore one allows the integration variable to take complex values, i.e. $\omega = \alpha + i\gamma \mid \alpha, \gamma \in \mathbb{R}$, and summarises the exponent in $F(\omega) = i\omega E + \log Y(\omega)$. Expanding this function to second order around the saddle point ω^* leads to

$$F(\omega) \approx F(\omega^*) + \frac{1}{2}\,\frac{\partial^2 F}{\partial\omega^2}\bigg|_{\omega^*} (\omega-\omega^*)^2 = i\omega^* E + \log Y(\omega^*) + \frac{1}{2}\,\frac{\partial^2 \log Y(\omega)}{\partial\omega^2}\bigg|_{\omega^*} (\omega-\omega^*)^2$$

since the first derivative of $F(\omega)$ is to vanish in ω^*. Moreover, we can note that ω^* has to be purely imaginary, as we claim the energy to be purely real. Steepest descent approximation dictates an integration path going through the extremal point ω^*, so one can write

$$\rho(E) \approx \frac{1}{2\pi}\exp(-\gamma^* E + \log Y(i\gamma^*)) \int_{-\infty}^{+\infty} d\alpha\, \exp\left(\alpha^2 \frac{1}{2}\,\frac{\partial^2 \log Y(\omega)}{\partial\omega^2}\bigg|_{i\gamma^*}\right)$$

and identifying

$$\beta \equiv -\gamma \;\wedge\; \mathcal{Z}(\beta) \equiv Y(-i\beta)$$

we get

$$\begin{aligned} \rho(E) &\approx \frac{1}{2\pi}\exp(\beta^* E + \log \mathcal{Z}(\beta^*)) \int_{-\infty}^{+\infty} d\alpha\, \exp\left(-\alpha^2 \frac{1}{2}\,\frac{\partial^2 \log \mathcal{Z}(\beta)}{\partial\beta^2}\bigg|_{\beta^*}\right) \\ &= \exp(\beta^* E + \log \mathcal{Z}(\beta^*)) \Bigg/ \sqrt{2\pi\, \frac{\partial^2 \log \mathcal{Z}(\beta)}{\partial\beta^2}\bigg|_{\beta^*}}\,. \end{aligned}$$

Importance of a second order correction

In the actual computation of the variance, we settle for a first order approximation of the binding energy distribution, neglecting any higher order correction

$$\rho(E) \approx \exp(\beta E + \log \mathcal{Z}(\beta))$$

leading to the variance

$$\sigma^2 = \frac{\partial^2 \log \mathcal{Z}(\beta)}{\partial\beta^2}\bigg|_0\,.$$

This is equivalent to neglecting the normalisation constant in the above density. Keeping the factor, $\rho(E)$ can be written logarithmically as

$$\log \rho(E) \approx \beta E + \log \mathcal{Z}(\beta) - \frac{1}{2} \log \left(2\pi \frac{\partial^2 \log \mathcal{Z}(\beta)}{\partial \beta^2} \right).$$

It is instructive to analyse the error made by the above approximation. Assuming the density to be approximately Gaussian, we can compute the mean energy by differentiating its logarithm, thus finding the extremal value.

$$\begin{aligned} \partial_E \log \rho(E) &= \beta - \frac{1}{2} \partial_E \log(2\pi {\partial_\beta}^2 \log \mathcal{Z}(\beta)) \\ &= \beta - \frac{1}{2} \frac{{\partial_\beta}^3 \log \mathcal{Z}(\beta)}{[{\partial_\beta}^2 \log \mathcal{Z}(\beta)]^2} \end{aligned} \tag{5.19}$$

Having accepted

$$\partial_\beta \log \mathcal{Z}(\beta) = -\bar{E},$$

we immediately get an extremal

$$\beta^* = \frac{1}{2} \cdot \frac{{\partial_\beta}^3 \log \mathcal{Z}(\beta)}{[{\partial_\beta}^2 \log \mathcal{Z}(\beta)]^2}. \tag{5.20}$$

Computing the inverse of the variance as the negative second derivative of the density's logarithm with respect to E leads to

$$\begin{aligned} {\partial_E}^2 \log \rho(E) &= \partial_E \beta - \frac{1}{2} \partial_E \frac{{\partial_\beta}^3 \log \mathcal{Z}(\beta)}{[{\partial_\beta}^2 \log \mathcal{Z}(\beta)]^2} \\ &= \partial_E \beta \left(1 + \frac{1}{2} \partial_\beta \frac{{\partial_\beta}^3 \log \mathcal{Z}(\beta)}{[{\partial_\beta}^2 \log \mathcal{Z}(\beta)]^2} \right). \end{aligned}$$

Hence, we can write the variance as

$$\begin{aligned} \sigma^2 &= -\partial_\beta E \left(1 + \frac{1}{2} \partial_\beta \frac{{\partial_\beta}^3 \log \mathcal{Z}(\beta)}{[{\partial_\beta}^2 \log \mathcal{Z}(\beta)]^2} \right)^{-1} \qquad (5.21) \\ &= -\partial_\beta E \left(1 - \frac{1}{2} \partial_\beta \frac{{\partial_\beta}^3 \log \mathcal{Z}(\beta)}{[{\partial_\beta}^2 \log \mathcal{Z}(\beta)]^2} \right) \\ &= {\partial_\beta}^2 \log \mathcal{Z} \left(1 - \frac{1}{2} \left\{ \frac{{\partial_\beta}^4 \log \mathcal{Z}}{[{\partial_\beta}^2 \log \mathcal{Z}]^2} - 2 \frac{[{\partial_\beta}^3 \log \mathcal{Z}]^2}{[{\partial_\beta}^2 \log \mathcal{Z}]^3} \right\} \right) \\ &= {\partial_\beta}^2 \log \mathcal{Z} - \frac{1}{2} \frac{{\partial_\beta}^4 \log \mathcal{Z}}{{\partial_\beta}^2 \log \mathcal{Z}} + \left[\frac{{\partial_\beta}^3 \log \mathcal{Z}}{{\partial_\beta}^2 \log \mathcal{Z}} \right]^2 \qquad (5.22) \end{aligned}$$

where one assumed the second term of the parenthesis in (5.21) to be small compared to unity. The differentials are all evaluated at β^* which optimises (5.19) and assuming further that this value is close to zero, we can linearise as follows

$$\partial_\beta{}^j \log \mathcal{Z}\big|_{\beta^*} \approx \partial_\beta{}^j \log \mathcal{Z}\big|_0 + \beta^* \cdot \partial_\beta{}^{j+1} \log \mathcal{Z}\big|_0 \; .$$

Inserting this approximation into the expression for the standard deviation (5.22) leads to

$$\begin{aligned}\sigma^2 &= \partial_\beta{}^2 \log \mathcal{Z}\big|_0 + \beta^* \cdot \partial_\beta{}^3 \log \mathcal{Z}\big|_0 + \ldots \\ &\quad - \frac{1}{2}\frac{\partial_\beta{}^4 \log \mathcal{Z}\big|_0 + \beta^* \cdot \partial_\beta{}^5 \log \mathcal{Z}\big|_0}{\partial_\beta{}^2 \log \mathcal{Z}\big|_0 + \beta^* \cdot \partial_\beta{}^3 \log \mathcal{Z}\big|_0} + \left[\frac{\partial_\beta{}^3 \log \mathcal{Z}\big|_0 + \beta^* \cdot \partial_\beta{}^4 \log \mathcal{Z}\big|_0}{\partial_\beta{}^2 \log \mathcal{Z}\big|_0 + \beta^* \cdot \partial_\beta{}^3 \log \mathcal{Z}\big|_0}\right]^2\end{aligned}$$

The fractions can be expected to have only negligible influence for a nearly Gaussian $\rho(E)$, while one might want to consider the correction in first order. Computing the extremal β^* numerically from (5.20) makes this consideration possible and lead to a slight improvement of QPMEME.

5.3 Quadratic Programming Sampler – QPS

As anchor serves the above discussed iterative Gibbs sampling algorithm. Conveniently, the method by which the quality of the alignment in a specific sequence is judged in the sampling steps can be easily replaced by another. While the implementation by [Lawrence et al., 1993] makes implicit use of a weight matrix description, as previously presented in [Berg and von Hippel, 1987], we implement a sampler with a decision making step based on the energy matrix description of [Djordjevic et al., 2003]. The general structure of the technique is, starting from some initial configuration – the alignment – of a system – the input sequences – of components (whose states might be chosen arbitrarily), to sequentially change the state of each component. The new state is chosen according to the probability of transition from the former state. Under certain circumstances[1] such a method leads to a stationary or cyclic configuration, compare e.g. [Honerkamp, 2000].

Let us translate this general scheme to a method for finding TF binding sites. Therefore, some assumptions have to be made to formulate a model on which a physical formalism can be applied. We consider several sequences, i.e. whole gene upstreams or sub-streams of such, sharing a common motif of given length – the binding site – which we want to align, revealing the pattern corresponding to that motif. Not knowing any better, we can start by merely guessing alignment positions in each sequence. Now this configuration is to be updated, so we exclude a sequence for which we try to find a better alignment. The remaining ones give us a model of assumed motifs, defining a corresponding “imaginary”

[1] A stationary solution is always found if the transition rates of the underlying Markov process satisfy the detailed balance condition. In other words, if it corresponds to a reversible system.

TF protein. Screening all possible binding sites of this factor on the excluded sequence, we get an "affinity landscape" and drawing a new alignment from the probability distribution defined by the latter, we can build up an algorithm in analogy to the mentioned Gibbs sampler.

The following describes the method in some more detail before we illuminate its functionality on a set of simple constructed examples and apply it on real genomic data.

5.3.1 Structure

First, we need to count out the nucleotide pair occurrences in the genome on which we will perform the alignment, to be able to build our probabilistic model. If we know the positions of all gene coding sequences, called open reading frames (ORF), we can exclude those from the counting, since we expect TF binding sites to be mainly situated in non-coding regions.

Then, a set of upstream sequences has to be prepared which are supposed to share a regulatory element. This is conveniently done in the FASTA format described by [Pearson and Lipman, 1988]. Input files thus consist of a comment line, identified by a ">" as first character and followed by the nucleotide sequence in the following lines. Table 5.1 gives an example input file. There is no further requirement to the comment line than the first character, so we chose to keep information about the TF whose binding site is embedded in the sequence, its length and absolute position on the genome, as well as the strand it is found on and the function (activator, repressor, dual, unknown) of the factor abbreviated by its initial letter. The upstreams are read in and each is provided with a randomly chosen position for the assumed binding site motif. If the base occurrences in the given upstreams are not expected to differ much from those in all non-ORFs, one can of course build the statistics from those upstreams. When available, however, we rely rather on the whole non-ORF region of a genome.

```
>BetI 21 (328605:328625) + R
GTGGCGTCGATCAGTTGTCTGCGCCGGATCGACTGCATCCCCAATTTGGGCATTTTCGCCACTCCATTCATCAGCG
GTGTTTATCTATTAAAGCGGTTATTGATTGGACGTTCAATATAAAATGTGTCTTAATTGTTACGAATTTGATTTTA
AATAGTAACAATAACAGTGGGGATACTGGATGACAGACCTTTCACACAGCAGGGAAAAGGACAAAATCA

>BetI 21 (328605:328625) - R
TGATTTTGTCCTTTTCCCTGCTGTGTGAAAGGTCTGTCATCCAGTATCCCCACTGTTATTGTTACTATTTAAAATC
AAATTCGTAACAATTAAGACACATTTTATATTGAACGTCCAATCAATAACCGCTTTAATAGATAAACACCGCTGAT
GAATGGAGTGGCGAAAATGCCCAAATTGGGGATGCAGTCGATCCGGCGCAGACAACTGATCGACGCCAC
```

Table 5.1: FASTA input of two sequences containing a BetI binding site

An iteration begins with the definition of the order in which the sequences are updated to avoid caveats or the possible – though improbable – case of never updating some data sets. One of the sequences is then excluded and from the remaining alignments, one can

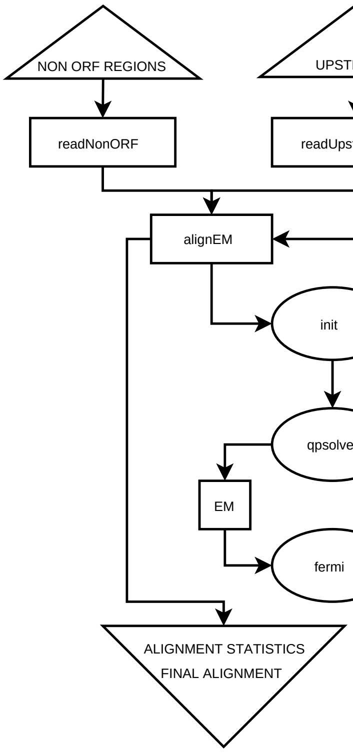


Figure 5.4: Simplified implementation flowchart of the Quadratic Programming Sampler

calculate an energy matrix, describing the imaginary transcription factor. The new alignment in the excluded sequence is then drawn from the distribution of binding probabilities. Upon convergence to some stationary alignment configuration or after a given number of iterations, a final alignment, respective a table of alignment occurrences is produced. The latter makes then sense, when no stationary configuration was reached or where one may assume some cycle to occur. The most common alignments can then be processed further, e.g. tested for actual similarity with experimental data or used to predict new binding sites on the genome. A program was written in *C++* to carry out the computations. Both independent nucleotide statistic and a nearest neighbour model of sequence generation were implemented, and we makes use of the *QL* package by [Schittkowski, 2003] to solve the quadratic optimisation problem. Figure 5.4 shows a flowchart of the implementation. After invoking *readUpstreams* and*readNonORF*, the alignment function *alignEM* is called. Here the quadratic program is prepared and solved, whereafter the computed energy matrix is used to draw a new alignment.

5.3.2 Drawing new alignments

Why the new alignment in the excluded sequence can be drawn from Fermi-Dirac distributed values can easily be seen, considering our simple model of binding between a TF and its binding site on DNA, controlled by the binding and dissociation constants K_{bind} and K_{diss}. This has already been explained in (5.8), i.e for a motif sequence M and a TF described by $\boldsymbol{\varepsilon}$ we get

$$P(M)_{bound} = f(\beta[E(M)+1])\,,$$

where the chemical potential has been set to unity. Recalling that "zero temperature" is a crucial assumption made by [Djordjevic et al., 2003], leading to a Heaviside step function instead of a Fermi-Dirac distribution for the binding probabilities

$$P(M)^{T=0}_{bound} = \begin{cases} 1 & E < \mu \\ 0 & E > \mu \end{cases} .$$

Evaluating all possible binding sites on the excluded sequence, we would get either forbidden or equally probable positions. Building a distribution from which we draw a new alignment which is proportional to $P(M)_{bound}$ instead of $E(M)$ seems plausible from a physical point of view and turns out to lead to both better results and faster convergence in numerical experiments. However, we do not make the "zero temperature" assumption when evaluating $P(M)_{bound}$, since it has a major drawback. When it comes to sequences sharing the actually expected motif – described by the energy matrix computed in the iteration – only partially and thus possibly not show a free energy of binding which is less than the chemical potential at the relevant position, we will assign a zero probability to that position. A slightly different energy matrix might show the position as a valid binding site, whereas immediately rejecting it might lead away from the "correct" alignment. Therefore, we smooth out the Heaviside distribution by reintroducing a finite temperature and thus evaluate via

$$f_t(E(M)+1) = \frac{1}{1+\exp(t^{-1}[E(M)+1])}$$

with the numerical parameter t which turns out to lead to useful results when set to $t \approx 0.05$.

This of course leads to the question how far the evaluated energy matrix is valid for finite temperatures, as the method was developed in the $T \to 0$ limit. We expect the error due to the reintroduction of a temperature to be negligible when ensuring $t \ll 1$.

5.3.3 Accepting new alignments

Having drawn a new possible alignment on the excluded sequence, we have to decide upon keeping or rejecting it. There again, the variance of the free energy of binding plays a major role. Minimising it meant to make the energy matrix more specific, allowing less sequences different to those of the present alignment to be recognised as binder. Thus, sampling for the alignment with minimal variance leads to an alignment of maximal similarity, with respect to our probabilistic model. We perform hence the simulated annealing procedure on the level of the variance $\sigma^2(a_k)$ corresponding to an alignment a_k, accepting a draw on sequence S_k – following [N.Metropolis et al., 1953] – with probability

$$P(a_k \to a'_k) = \begin{cases} \exp(-\gamma\Delta\sigma^2) & \text{if} \quad \sigma^2(a'_k) > \sigma^2(a_k) \\ 1 & \quad\quad \text{otherwise} \end{cases}$$

with $\Delta\sigma^2 = \sigma^2(a_k) - \sigma^2(a'_k)$ in the role of a configuration energy and γ in the role of an inverse temperature which is reduced as we iterate the sampling. The schedule for γ is kept inverse-logarithmically to ensure a slow cooling after a short exposure of the system to “high temperature”. Specifically, we apply in the i^{th} iteration the temperature

$$\gamma(i) = \frac{\gamma_0}{\log(i)}\,,$$

where γ_0 is initially given as parameter to the program. The right choice of γ_0 is crucial for the convergence behaviour of the algorithm. A too high initial value will lead to very long execution times, since we did not implement an adaptive temperature schedule in the present version. A too low choice might result in the system getting stuck in a configuration being far from optimal. A good choice would be to set γ_0 just above the standard deviation of the variance's development. Since this information is not available when starting the sampler, we have to estimate it from several initial observations. This, of course, could be readily added to the implementation.

5.3.4 Convergence detection

We need to decide when to stop iterating and assume the current alignment to have converged to an acceptable solution. Again applying a simple heuristic, we keep a history of the alignments yielded in the last complete[2] iterations. If no change occurred for a certain number – which is also given as parameter to the program – of complete iterations, the current alignment is assumed to be final.

Details on calling the program as well as some sample in- and output is left to appendix B.

5.3.5 Identification of non-contributing sequences

The current implementation does not identify sequences non-contributing sequences in the data set yet. This is problematic when applying the program to “real world” problems where one would like to find TF binding sites in a set containing possibly totally uncorrelated sequences concerning a common motif. We report this discussion to chapter 8, nevertheless one should keep in mind this missing feature when examining the results of a sampling run. We show a realistic case in chapter 7, where such an identification procedure would be needed.

[2] By a complete iteration we mean a round of updating every sequence in the data set.

Chapter 6

Material acquisition

More and more genome sequences are being made available due to the persistent sequencing efforts of the last decades. At the same time – although at a much slower rate – the information is being understood, annotations are made on the sequences and interactions are being discovered. We can thus enjoy a great abundance of data, both to make predictions on unknown grounds and to test methods under development by reproducing known results.

6.1 Sequenced genomes

The world's largest database of genomic sequences containing more than 100 giganucleotides of raw data and representatives of most yet sequences genomes is publicly available from NCBI[1] via

http://www.ncbi.nlm.nih.gov/Genbank

Applying our algorithm on genomic data of the common K-12 MG1655 strain of the Escherichia coli bacterium known from every day life and of the not less familiar budding yeast Saccharomyces cerevisiae S228C, we also made use of specialised databases.

6.1.1 Escherichia coli K-12 MG1655

The genome has been published by [Blattner et al., 1997] and is available from NCBI. For comprehensible annotations on known TF binding site emplacements, we use the public database RegulonDB by [Salgado et al., 2004]. To illustrate the difference in nucleotide

[1] National Center for Biotechnology Information, U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA. Part of the U.S. National Institute of Health.

A	G	C	T	
0.2462	0.2537	0.2542	0.2459	
0.0728	0.0513	0.0553	0.0668	A
0.0576	0.0582	0.0827	0.0551	G
0.0701	0.0747	0.0586	0.0509	C
0.0457	0.0694	0.0576	0.0731	T

Table 6.1: Independent base and pair occurrences in the whole E. coli genome

statistics between the whole genome and non-ORF regions only, the tables 6.1 and 6.2 show the explicitly counted frequencies for single nucleotides (first row) and nearest neighbour pairs (first position in column, second position in row). The difference between the statistics of both samples is due to the overlaying triplet statistics of the amino acid coding codons in ORF regions which is not present in the non-ORF domains.

A	G	C	T	
0.2852	0.2071	0.2091	0.2985	
0.0970	0.0541	0.0636	0.0705	A
0.0475	0.0415	0.0498	0.0682	G
0.0542	0.0566	0.0430	0.0553	C
0.0864	0.0545	0.0527	0.1050	T

Table 6.2: Independent base and pair occurrences in the non-ORF regions of E. coli

6.1.2 Saccharomyces cerevisiae S288C

Both genome and comprehensible annotations are available from the Saccharomyces Genome Database (SDB) by [Balakrishnan et al., 2005], where the information on yeast is usually more up to date than on other public databases. The statistics evaluated from the non-ORF domain are shown in table 6.3.

A	G	C	T	
0.3293	0.1709	0.1709	0.3289	
0.1238	0.0543	0.0565	0.0946	A
0.0554	0.0313	0.0279	0.0564	G
0.0505	0.0349	0.0315	0.0543	C
0.0993	0.0508	0.0550	0.1235	T

Table 6.3: Independent base and pair occurrences in the non-ORF regions of S. cerevisiae

6.2 Known TF binding sites

A more generic information source on gene expression regulation is TRANSFAC – a database by [Wingender et al., 2001], containing annotations on many higher organisms' genomes. When applying our method to other data than bacteria and fungi, considering the use of this database is highly recommended. For our purposes, the information from RegulonDB and SDB is largely sufficient.

To test the algorithm, we prepared – for different TFs – collections of sequences containing an experimentally reported binding site each. The sites were then flanked on both sides with their next 100 neighbour nucleotides, resulting in sequences of length $200 + l$ for a binding site of length l. For the bacterial genome, this might easily lead to overlaps with gene coding regions, which we did not take into account, as the typical intergenic non-ORF is of some hundred nucleotides and binding sites are typically found within tens of nucleotides before the transcription start. On the other hand, we expect other problems to be dominating the causes for alignment failures, as we discuss in chapter 8. In yeast, the typical distance is of several hundred up to a few thousand and TF binding sites are typically situated within a few hundred nucleotides before the transcription start. One might argue the sampling being too unrealistic as a test for "real" yeast sequence alignment tasks. However, looking for a confirmation of general functionality, realism remains secondary in the beginning.

Chapter 7

Results

A short description of alignment successes with constructed samples is followed by a more complex benchmark, where we reconstruct “known” alignments by dissimulating small binding sites in larger sequences. Finally we succeed in “predicting” binding sites of an “unknown” TF in a more realistic yeast sample. The reason for all quotation marks will become clear shortly.

7.1 Artificial data

We tested the principal functionality of our sampling algorithm on simple constructed sequence sets. As a first test bench, one can take the input presented in table 7.1. Each line of characters represents one of the input sequences, denoted by S_i with $i = 1, 2, 3$. The first line just visualises the starting positions for the assumed or guessed motif - here of length three. Bold characters define the actual alignment - here AAT, TCG, GAA - corresponding to a list of numbers - here $(2, 5, 3)_3$, to implicitly introduce the alignment notation $(a_1, a_2, \ldots, a_N)_l$ for the alignment of a motif of length l in N sequences. The “non-ORF” input for the statistics was taken to be the same as the sequence data.

a_i:	1	2	3	4	5	-	-
S_1	A	**A**	**A**	**T**	C	G	A
S_2	A	A	A	A	**T**	**C**	**G**
S_3	T	C	**G**	**A**	**A**	A	A

Table 7.1: Simple test bench for assumed motif of length 3

Simulating with annealing and an initial numerical “temperature” of $\gamma_0 = 1.0$, we find the expected alignment $(4, 5, 1)_3$ in more than 95% of the runs when accepting convergence if we find no changes in the last 5 complete iterations. Lowering γ_0 leads to alignments like $(1, 1, 5)_3$ or "frustrated" results as $(5, 5, 2)_3$ and we observe in fact a minimal variance for $(4, 5, 1)_3$.

Another example input is presented in table 7.2, which is just the former one with the adenines replaced by less alignable data. The non-ORF is again taken to be identical to the sequence input. Again finding the expected alignment $(4, 5, 1)_3$, the algorithm proves to work principally at least on simple collections.

a_i:	1	2	3	4	5	-	-
S_1	G	A	T	T	C	G	A
S_2	A	C	C	A	T	C	G
S_3	T	C	G	G	G	C	A

Table 7.2: Alternative test bench for assumed motif of length 3

7.2 Co-regulated genes

The following results come from aligned upstream regions of co-regulated genes, i.e. sets of genes being activated and/or repressed by the same TF. Known binding sites are identified and placed as described before. The expected alignment from experiments is thus $(100, 100, \ldots, 100)_l$ if nothing else is mentioned.

7.2.1 Escherichia coli

FruR - fructose repressor protein

FruR TFs are involved in the carbon metabolism of E. coli and is known to act as regulator for a dozen of genes. The structure of the protein's DNA binding domain is illustrated in figure 7.1. QPS was applied on the set of the twelve upstream regions of co-regulated genes with a convergence parameter of five and an initial annealing temperature of $\gamma_0 = 0.05$. Running the sampler 100 times on the data set leads to the results summarised in table 7.3, where the first column denotes the amount by which the corresponding alignment was found. The first two alignment sets in the list show the best correspondence to the experimentally reported alignment in the first seven sequences. One could think of playing with the initial temperature to try to reconstruct the experimental result, but comparing the actual sequences of both sampled and experimental alignment from tables 7.4 and 7.7 suggests that the achieved alignment might not be worse than the reported one, in terms of the variance of a corresponding energy matrix. Indeed we find the variance for the

Figure 7.1: DNA binding region of the FruR TF protein

#	variance	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}
2	0.0490531	99	101	98	100	100	100	98	156	189	127	173	116
4	0.0493923	100	102	99	101	101	101	99	157	190	128	174	117
1	0.0510720	96	182	146	97	183	183	95	159	138	124	20	185
1	0.0513208	99	101	98	100	74	100	98	156	189	127	173	116
1	0.0518154	100	102	99	101	87	87	99	157	190	128	174	117
1	0.0520743	47	177	148	68	27	27	1	151	182	122	158	124
1	0.0522605	65	107	104	136	80	80	104	124	111	133	179	23
1	0.0524654	187	95	73	122	171	171	0	150	181	49	155	121
2	0.0526585	41	129	100	132	15	115	100	33	107	129	175	118
1	0.0528457	183	2	78	131	75	75	99	157	190	56	174	117

Table 7.3: Top ten occurrences of different variances and alignments on the FruR data set

experimental alignment to be 0.0714964, when applying QPS to just compute the energy matrix of the experimental motif. On the other hand, excluding the last sequences from the FruR set, keeping only S_1 to S_7, leads to better alignability. Table 7.5 shows the top ten variances for the reduced data set and one observes immediately the occurrence of 38 alignments compatible[1] with $(100, 102, 99, 101, 101, 101, 99)_{15}$. Figures 7.2, 7.3 and 7.4 show the sequence logos corresponding to all three FruR alignments discussed here. Here one can directly see the similarity between the results. The "misalignment" of S_8 till S_{12} in the original data set does not seem to distort the result and all three logos appear to

[1] only differing by small shifts or gaps

```
tgcga GCTGAATCGCTTAAC ctggt
gcgat GCTGAAAGGTGTCAG ctttg
tgact CTTGAATGGTTTCAG cactt
cactg ACTGAAACGTTTTTG cccta
ccaaa GCTGAATCGATTTTA tgatt
ccaaa GCTGAATCGATTTTA tgatt
tccta GCTGAAGCGTTTCAG tcgat
gcggt CCGCAGGCGGCACTG cttac
catcc CCAAAGGCGCTTCTG tttaa
ggcag CCAGAAGGGAGTCAG gctga
tggga TATGAGGCGGTACAG tcatt
tccat CCTCATGCGCTTCTG acgcg
```

Table 7.4: Alignment with minimal variance on the FruR data set

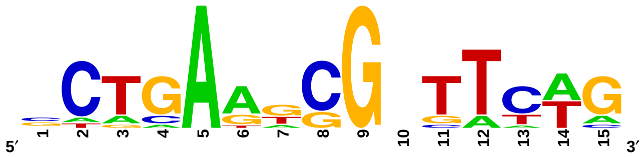


Figure 7.2: FruR sequence logo of the sampled alignment

#	variance	a_1	a_2	a_3	a_4	a_5	a_6	a_7
16	0.0375483	100	102	99	101	101	101	99
13	0.0380083	99	101	98	100	100	100	98
9	0.0395157	98	100	97	99	99	99	97
3	0.0399036	99	101	98	100	74	74	98
1	0.0408882	96	98	95	97	97	97	95
4	0.0435589	96	184	88	97	71	71	95
4	0.044195	162	185	89	67	125	125	129
1	0.0451398	37	99	96	98	111	111	96
1	0.046401	195	82	8	93	150	150	189
4	0.0471001	100	102	99	101	54	54	99

Table 7.5: Top ten occurrences of different variances and alignments on the FruR 1-7 data set

be quite compatible. FruR is thus a first example where the identification of a useful alignment with QPS on real data ist principally possible.

gcgag	CTGAATCGCTTAACC	tggtg
cgatg	CTGAAAGGTGTCAGC	tttgc
gactc	TTGAATGGTTTCAGC	acttt
actga	CTGAAACGTTTTTGC	cctat
caaag	CTGAATCGATTTTAT	gattt
caaag	CTGAATCGATTTTAT	gattt
cctag	CTGAAGCGTTTCAGT	cgatt

Table 7.6: Alignment with minimal variance on the FruR 1-7 data set

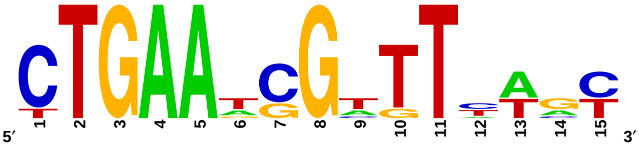


Figure 7.3: FruR sequence logo from the reduced data set

cga	GCTGAATCGCTTAAC	ctg
gat	GCTGAAAGGTGTCAG	ctt
ctc	TTGAATGGTTTCAGC	act
tga	CTGAAACGTTTTTGC	cct
aag	CTGAATCGATTTTAT	gat
aag	CTGAATCGATTTTAT	gat
tag	CTGAAGCGTTTCAGT	cga
gtt	GCTGAATCGTTAAGG	tag
gtg	GTGAATCGATACTTT	acc
aca	GTTAACCGATTCAGT	gcc
gac	CTGAATCAATTCAGC	agg
atc	GTTAAGCGATTCAGC	acc

Table 7.7: Experimentally reported binding sites for the FruR TF

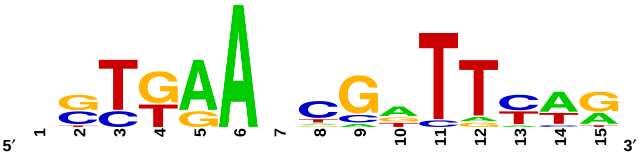


Figure 7.4: FruR sequence logo of the experimentally reported motif

LexA - part of the SOS response system

An important mechanism in bacterial cells is the SOS response system, recognising unpaired DNA segments – which is often due to some damage on the complementary strand

– and repairing possible faults. Among other TFs, LexA is involved in the initiation of SOS responses. It binds as homo-dimer to DNA sites, which should give some degree of symmetry to the corresponding sequence. The DNA binding region of a LexA monomer is illustrated in figure 7.5. Again running QPS 100 times on the collection containing experimentally reported binding sites, we find the alignment corresponding to minimal variance to be $(7, 18, 151, 29, 158, 29, 26, 7, 63, 139)_20$, which is far from being compatible to the experimental result. In fact, no alignment of our results showed compatibility to the alignment from the database. Inspecting the tables 7.8 and 7.9, one might tend to criticise the experimental results which possibly are not accurate. The sampled alignment is justi-

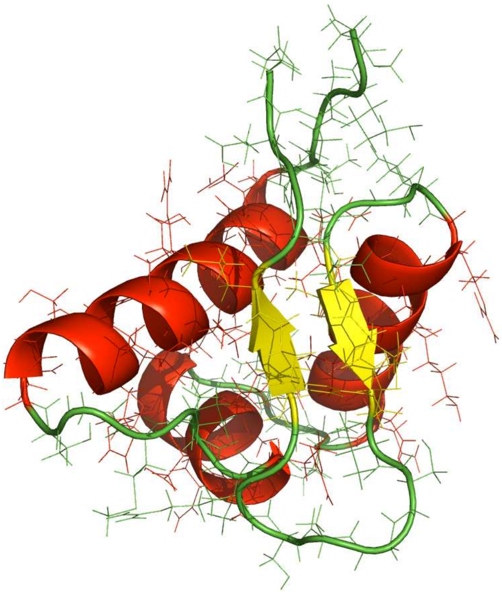


Figure 7.5: DNA binding region of the LexA TF protein

agcag	CTGGCTGCGCTTATCGACAG	ttatc
taagg	CCGGAGTTTTATCTCGCCAC	agagt
cactt	CAGGCTATGCACATCGTTCT	tcgtc
cgacc	GTGATGCGGTGCGTCGTCAG	gctac
gaggc	CAGTTCAGGCACGACGCCGC	cgtag
ttgaa	CAAGCGATGCTCGACGCCGG	gctga
ccgct	CATGTTTCGCGCGGCGCTAC	gcaaa
ccgct	CATGTTTCGCGCGGCGCTAC	gcaaa
tctgg	CTGACGGTTTGCGCCGCCAG	cggga
tcccg	CTGGCGGCGCAAACCGTCAG	ccaga

Table 7.8: Alignment with minimal variance on the LexA data set

fyable by the amassment of – in E. coli non-ORFs rather less probable – C/G occurrences.

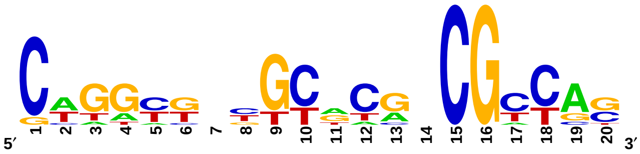


Figure 7.6: LexA sequence logo of the sampled alignment

We can compare both sequence logos in the figures 7.6 and 7.7, where a low positional information might point to a case of cooperative binding, non contribution or misalignment.

gac	GCCTGGCTTTCAGGGCAGCG	tta
gat	GAACTGTTTTTTTATCCAGT	ata
tac	TGTACATCCATACAGTAACT	cac
tga	TACTGTATGAGCATACAGTA	taa
aat	AAGCTGGCGTTGATGCCAGC	ggc
caa	ATCTGTATATATACCCAGCT	ttt
ctt	TTGCTGTATATACTCACAGC	ata
cag	CATAACTGTATATACACCCA	ggg
ttg	ACCTGAATGAATATACAGTA	ttg
caa	TACTGTATATTCATTCAGGT	caa

Table 7.9: Experimentally reported binding sites for the LexA TF

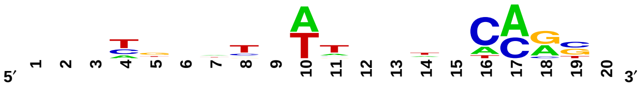


Figure 7.7: FruR sequence logo of the experimentally reported sites

LexA is therewith a factor where we cannot be sure of the results from QPS.

Crp - a bacterium's favourite

To round up the results on E. coli, we attempt to align binding sites for the Crp TF. RegulonDB lists 149 interactions between the factor and DNA with a binding site length of 22 nucleotides and another 47 interactions with a binding site length of 19 nucleotides. With 157 documented distinct[2] binding sites it is the most varied TF of the bacterium. It's beautiful DNA binding domain structure is illustrated in figure 7.8. The sequence logo from experimentally reported binding site of length 22 is shown in figure 7.9. Trying to sample for the binding sites turned out to be rather hopeless. Even on smaller collections, comprehending not more than five sequences were we able to identify the observed binding sites, pointing to a binding behaviour which is not sufficiently described by our models, i.e. cooperativity or some conformal adaptivity[3] of the TF protein.

Figure 7.8: DNA binding region of the Crp TF protein

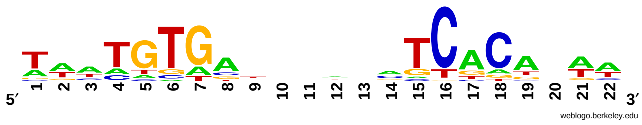


Figure 7.9: Crp sequence logo of the experimentally reported sites

[2] Sites situated on the complementary strand of another one are not counted.

[3] This assumption is motivated by the appearance of of the DNA binding domain and is highly speculative.

7.2.2 Saccharomyces cerevisiae

Gal4 - compatible for alignment

Moving on to a higher organism, we find again good evidence for the applicability of our sampler. Yeast's Gal4 data set disclosed a a fairly compatible alignment of a palindromic binding site motif with $(141, 54, 141, 122, 123, 59, 101, 103, 102, 100, 102, 100, 100)_{21}$. The palindrome is easily recognised in table 7.10. The binding sites from the database, shown in table 7.11 show each a comparable palindrome motif, although most sequences are slightly shifted against each other.

```
gatca CGGTCAACAGTTGTCCG agcgc
gatca CGGTCAACAGTTGTCCG agcgc
aagta CGGATTAGAAGCCGCCG agcgg
aagta CGGATTAGAAGCCGCCG agcgg
ccgag CGGGCGACAGCCCTCCG acgga
ccgag CGGGCGACAGCCCTCCG acgga
acgtt CGGTCCACTGTGTGCCG aacat
tcgca CGGACTCCATTTCCCCG gacct
aagct CGGAGTATATTGCACCG atccg
tttac CGGCGCACTCTCGCCCG aacga
cgccg CGGAGTGCTCTTCGCCG agata
acaat CGGGGCAGACTATTCCG gggaa
tcgcc CGGACATCACCCGCCCG gcaca
```

Table 7.10: Most common alignment on the Gal4 data set

```
cagct TGGCTATTTTGTGAACA ctgta
atttt TGGGTTAAGGAAAATGA cagaa
ggaac TTTCAGTAATACGCTTA actgc
ttaac TGCTCATTGCTATATTG aagta
attga AGTACGGATTAGAAGCC gccga
gcgtc CTCGTCTTCACCGGTCG cgttc
cacgt TCGGTCCACTGTGTGCC gaaca
aactc GCACGGACTCCATTTCC ccgga
ggaag CTCGGAGTATATTGCAC cgatc
tttac CGGCGCACTCTCGCCCG aacga
ggcgc CGCGGAGTGCTCTTCGC cgaga
acaat CGGGGCAGACTATTCCG gggaa
tcgcc CGGACATCACCCGCCCG gcaca
```

Table 7.11: Experimentally reported binding sites for the Gal4 TF

Rtg3 - retrograde regulation protein

The next TF binding site sampling showed again – with the alignment $(18, 150, 41, 64, 76)_6$ – no compatibility to the experimental data. Considering the length of six and the amount of known binding sequences of five, this result is certainly not very surprising. The binding site is “lost” in the data set and too small for our sampler to be of statistical significance. Taking advantage of the small size of the expected binding site, we illustrate the evaluation of the final energy matrix computed by QPS and corresponding to the alignment on genomic data. Our alignment of Rtg3 yields

ataaa	GGTGTC	ttaca
gtacc	GGTTTC	ctttt
aaaga	GCTTTC	acaaa
agact	GCTGTC	gcgat
aagaa	GGTTTC	tgcaa

Table 7.12: Most common alignment on the Rtg3 data set

tac	GGGTCA	cgc
ctt	GTGACC	tga
aca	CAGATA	caa
ctt	GGTCAC	cta
tga	TGAGTG	acc

Table 7.13: Experimentally reported binding sites for the Rtg3 TF

$$(\varepsilon^i_\alpha) = \begin{pmatrix} +0.0964 & -0.2497 & +0.1008 & +0.0992 \\ +0.0577 & -0.1099 & -0.1099 & +0.0679 \\ +0.0174 & +0.0190 & +0.0275 & -0.1634 \\ -0.0166 & -0.1178 & -0.0081 & -0.1178 \\ +0.0508 & +0.0754 & +0.0689 & -0.1228 \\ +0.1156 & +0.1231 & -0.2365 & +0.1321 \end{pmatrix},$$

which we can evaluate, e.g., evaluate on the collection of input sequences, finding the position dependent free energies of the binding of the TF being represented by the energy matrix. Computing (4.2), we find the results shown in figure 7.10.

Yst02r - a realistic sample

Closing this result section, we show the identification of binding sites on a more realistic data set. A large scale assessment of bioinformatics tools for regulatory element detection has been initiated by [Tompa et al., 2005]. Providing data sets of human, mouse, fruit-fly

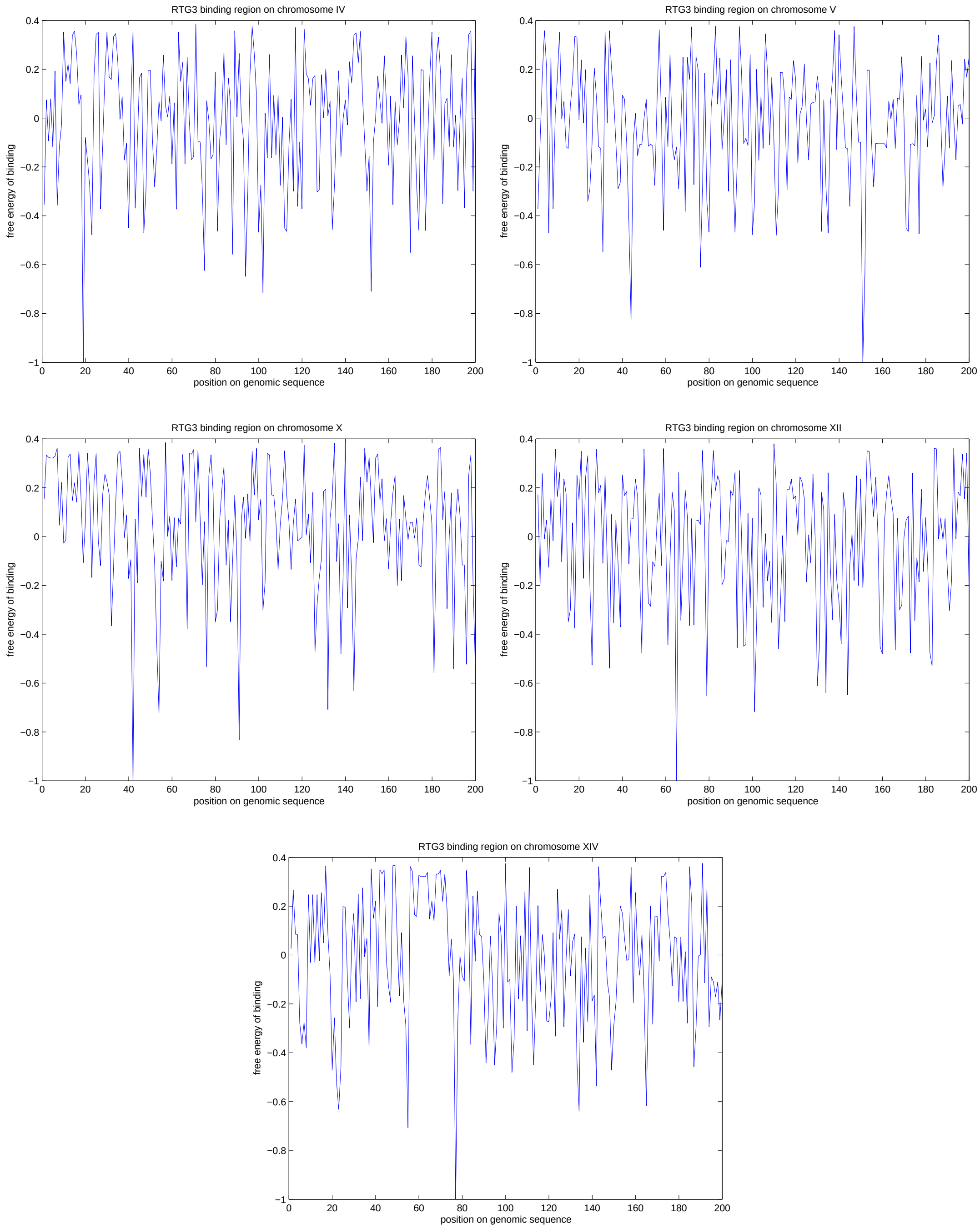
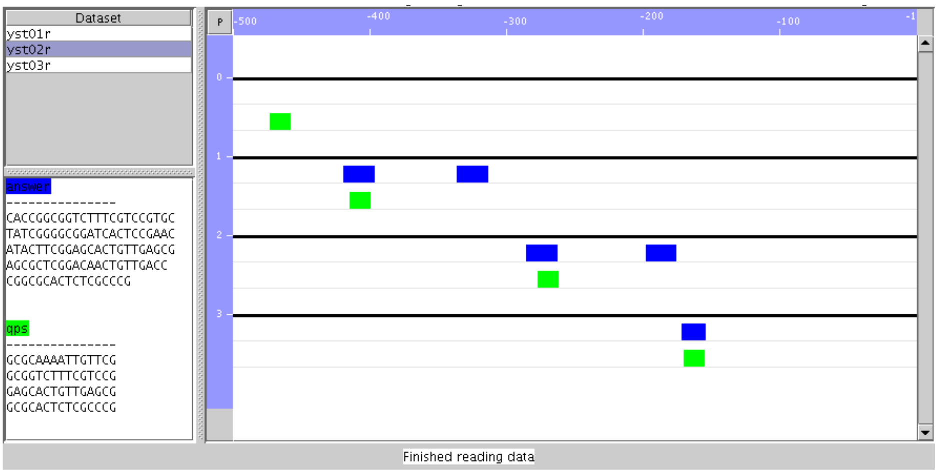


Figure 7.10: Evaluation of the assumed RTG3 TF's energy matrix with clearly identifiable binding sites from the sampling on experimentally reported binding regions



```
answer
---------------
CACCGGCGGTCTTTCGTCCGTGC
TATCGGGGCGGATCACTCCGAAC
ATACTTCGGAGCACTGTTGAGCG
AGCGCTCGGACAACTGTTGACC
CGGCGCACTCTCGCCCG


qps
---------------
GCGCAAAATTGTTCG
GCGGTCTTTCGTCCG
GAGCACTGTTGAGCG
GCGCACTCTCGCCCG
```

Finished reading data

Figure 7.11: Assessment result for the Yst02r data set

and yeast containing TF binding sites, participants are required to make predictions on lengths and locations.

In a short run of QPS, we were successfully predicting binding site motifs in a small collection of four sequences, each 500 nucleotides long, by assuming a length of 15 for the sites, which is rather typical for yeast factors. Great euphoria was nevertheless shorty damped, when making totally wrong predictions for other collections of yeast data. The results of three prediction attempts can be looked at on the assessment homepage[4]. Figure 7.11 shows the results of our successful “prediction”.

[4] http://bio.cs.washington.edu/assessment/

Confirmation ID: qps.25CB4EB40D330A8D6D1BD13CD20CDFA0

Chapter 8

Discussion

8.1 Possible improvements

As already observed by [Berg and von Hippel, 1987] and [Schneider et al., 1986], sequence positions in an experimentally reported set of binding sites leading to a lower information score often coincide with cooperative binding of two neighbouring sites with the corresponding TF. A high positional information score – on the other hand – is mostly associated to independent binding. High occurrences of cooperativity in a binding-site motif lead hence necessarily to difficulties when trying to identify such a binding site with our alignment method with its assumption of independent energy contributions. However, the formalism of QPMEME could be extended to consider nearest neighbour dependencies, leading to a second order expansion of (5.4) to

$$E(S) = \sum_{i}^{L} \sum_{\alpha}^{4} \varepsilon_{\alpha}^{i} S_{\alpha}^{i} + \sum_{i,j}^{L} \sum_{\alpha,\beta}^{4} J_{ij}^{\alpha\beta} S_{\alpha}^{i} S_{\beta}^{j}$$

with the coupling constants satisfying

$$J_{ij} \equiv J_{ij} \delta_{i,j+1} \,.$$

Another crucial improvement which remains to be implemented is the ability to detect sequences which probably do not contribute to the expected motif. We have seen in the results section that the hard constraint of assuming exactly one binding site per input sequence may make the final alignment less specific. Also we need to consider the possibility of multiple binding sites per sequence.

The possible existence of conformationally flexible TFs, i.e. the consideration of variably gapped binding sites in the theory might also be a useful improvement, however this may be solved by abandoning the one-site-per-sequence constraint and allowing, allowing gapped palindromes as a next consequence.

8.2 Conclusion

We have developed a method of sequence alignment which proved being applicable for the identification of regulatory motifs on DNA. Extensive benchmarking has to be performed on the algorithm and improvements have to be implemented, before it can be used to solve realistic problems. As a further step, the classes of motifs regulatory motifs being identifyable by aligning with QPS should be characterised.

The present implementation may be a first step towards the development of an independent classifier of regulatory elements, decoding a genome for its regulatory networks.

Appendix A

Bayesian inference of pseudocount regularisers

Here we show how to compute the probability from (3.2) of observing a nucleotide α, having observed a sample sequence S

$$P_S(\alpha) = \int d\rho \, \frac{p_\alpha \, P(\rho)P(S|\rho)}{\int d\rho' P(\rho')P(S|\rho')} \, .$$

We start by writing the probability from (3.3) of observing a sequence S given the nucleotide occurrence probabilities $\rho = (p_\mathrm{A}, p_\mathrm{C}, p_\mathrm{G}, p_\mathrm{T})$, replacing the factorial by the more general Γ-function.

$$P(S|\rho) = \Gamma\left(1 + \sum_\alpha n_\alpha\right) \prod_\alpha \frac{{p_\alpha}^{n_\alpha}}{\Gamma(1+n_\alpha)}$$

with

$$\Gamma(x) = \int_0^\infty dt \, t^{z-1} e^{-t} \, .$$

Assuming a Dirichlet prior for the distribution of occurrence probabilities

$$P(\rho) = \Gamma\left(\sum_\alpha \beta_\alpha\right) \prod_\alpha \frac{{p_\alpha}^{\beta_\alpha - 1}}{\Gamma(\beta_\alpha)}$$

with the set of parameters β_α, we are able to evaluate the denominator of (3.2) as

$$\begin{aligned}
\int d\rho \, P(\rho)P(S|\rho) &= \frac{\Gamma\left(\sum_\alpha \beta_\alpha\right)\Gamma\left(1 + \sum_\alpha n_\alpha\right)}{\prod_\alpha \Gamma(\beta_\alpha)\Gamma(1+n_\alpha)} \int d\rho \prod_\alpha {p_\alpha}^{\beta_\alpha + n_\alpha - 1} \\
&= \frac{\Gamma\left(\sum_\alpha \beta_\alpha\right)\Gamma\left(1 + \sum_\alpha n_\alpha\right)}{\Gamma\left(\sum_\alpha \beta_\alpha + n_\alpha\right)} \prod_\alpha \frac{\Gamma\left(\beta_\alpha + n_\alpha\right)}{\Gamma(\beta_\alpha)\Gamma(1+n_\alpha)} \, .
\end{aligned}$$

Knowing this, we can readily evaluate (3.2) as

$$\begin{aligned} P_S(\alpha) &= \frac{\Gamma\left(\sum_{\alpha'} \beta_{\alpha'} + n_{\alpha'}\right)}{\prod_{\alpha'} \Gamma\left(\beta_{\alpha'} + n_{\alpha'}\right)} \int d\rho \, p_\alpha \prod_{\alpha'} {p_{\alpha'}}^{\beta_{\alpha'} + n_{\alpha'} - 1} \\ &= \frac{\Gamma\left(\sum_{\alpha'} \beta_{\alpha'} + n_{\alpha'}\right)}{\prod_{\alpha'} \Gamma\left(\beta_{\alpha'} + n_{\alpha'}\right)} \frac{\prod_{\alpha'} \Gamma\left(\delta_{\alpha,\alpha'} + \beta_{\alpha'} + n_{\alpha'}\right)}{\Gamma\left(1 + \sum_{\alpha'} \beta_{\alpha'} + n_{\alpha'}\right)} \end{aligned}$$

which is found using the definition of the Γ-function. Further, by the fact that $\Gamma(x+1) = x\Gamma(x)$, we can simplify the result to the form

$$P_S(\alpha) = \frac{\beta_\alpha + n_\alpha}{\sum_{\alpha'} \beta_{\alpha'} + n_{\alpha'}} \, ,$$

which we already stated in chapter 3.1.2. The choice of a suitable set of pseudocount parameters β_α may depend on the genome or genomic region in question. It might thus be more appropriate to choose higher pseudocounts for nucleotides where one expects a higher occurrence probability from large sample countings, while reducing the pseudocount for other nucleotides. A similar calculation as the one presented here was performed by [Karplus, 1995] in the context of protein sequence statistics.

Appendix B

Implementation details

The implementation is separated in two programs: `statistics`, which is a small helper counting out the nucleotide and -pair occurrences and `qps`, which is the sampler itself. We show here the syntax of calling the programs. This information is also given when invoking the programs with the wrong number of parameters.

```
Syntax: statistics <input> <statistics>

       input     (file) input regions in FASTA format
  statistics     (file) output for occurrence statistics
```

```
Syntax: qps <input> <statistics> <width> <iterations> <seed> <temperature> <anneal> <verbose>

       input     (file) input regions in FASTA format
  statistics     (file) base and base pair occurrence statistics
       width  (integer) assumed motif width
  iterations  (integer) parameter for the convergence test
        seed  (integer) random seed
 temperature     (real) numerical starting temperature
      anneal  (boolean) simulated annealing on/off
     verbose  (boolean) verbose output on/off
```

The output of `qps` should look as follows, here invoked with the BetI sequence file from table 5.1 as input:

```
QPSampler
*********
Random seed: 1129159545

final EM evaluated as:
-0.027786       +0.044369       +0.043419       +0.050666
+0.026009       +0.014694       +0.016337       -0.046650
-0.005485       +0.003757       +0.003175       -0.063951
```

```
+0.052269       -0.046608       +0.055651       +0.059185
-0.103433       -0.128215       -0.062832       -0.076861
-0.039632       +0.041423       +0.039001       +0.033312
-0.003069       -0.008002       -0.104217       -0.000222
+0.055430       -0.049521       +0.049524       +0.053865
+0.014458       +0.017890       +0.021711       -0.057002
-0.013354       +0.000961       -0.058313       -0.044789
+0.032146       +0.030079       -0.073166       +0.026383
-0.074303       -0.002648       -0.003386       -0.011222
-0.053172       +0.011437       +0.012500       +0.022103
+0.036263       +0.028660       +0.029529       -0.044255
-0.001634       +0.040298       -0.011254       +0.044469
+0.011827       +0.037633       +0.040277       -0.003455
-0.039990       +0.030562       +0.030074       +0.035285
-0.011248       +0.014858       +0.020153       -0.016337
-0.045510       +0.022681       +0.024497       +0.023752
-0.047632       +0.022315       +0.020489       +0.025796
+0.044633       +0.036216       -0.009067       +0.002763

best binder  : -1.04004
worst binder :  0.584765
variance     :  0.0212085

convergence after 106 steps

final alignment: 0.17 s
sequence 0 - position 102       tattg   ATTGGACGTTCAATATAAAAT   gtgtc
sequence 1 - position 105       tttat   ATTGAACGTCCAATCAATAAC   cgctt

final temperature: 1.07e-02
```

It is to consider that the base ordering of the energy matrix is given by A, G, C & T and not in the widely used alphabetical order. A sequence positions from left to right corresponds to a matrix entry from top to bottom.

Acknowledgements

I am very grateful to some people who helped me a lot during the accomplishment of this work. First of all, I want to thank my supervisor and examiner Erik Aurell at KTH Stockholm for his patient and very motivating way of guiding me through my thesis and giving me the opportunity to throw a first glance on the auspicious field of statistical biophysics and bioinformatics. Massimo Vergassola from the Institut Pasteur contributed with discussions very useful to me and my thanks go to his kind hospitality.

Barbara Drossel, my examiner at TU Darmstadt, made it possible for me to carry out this work abroad and I am thankful for her feedback and the occasions she gave me to present it in the condensed matter group's internal seminar. Klaus Schittkowski granted the use of his quadratic programming code, sparing me some tidious and less efficient coding.

To Claes Malmnäs goes my special thank for his sparks of inspiration and very enriching conversations. Thanks to him, Maria Werner and John Ardelius, I enjoyed a very pleasant working atmosphere. My good friend Christian Kempf gave me useful last minute advices on the diction in my manuscript, which a appreciate much.

Anna Stockbring deserves my innermost gratitude for being there for me, kindly motivating me and for all the patience she had waiting for me, when I decided to work for some more "twenty minutes" on the spot.

Bibliography

[Altschul et al., 1990] Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410.

[Bailey and Elkan, 1994] Bailey, T. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 2:28–36.

[Balakrishnan et al., 2005] Balakrishnan, R., Christie, K., Costanzo, M., Dolinski, K., Dwight, S., Engel, S., Fisk, D., Hirschman, J., Hong, E., Nash, R., Oughtred, R., Skrzypek, M., Theesfeld, C., Binkley, G., Lane, C., Schroeder, M., Sethuraman, A., Dong, S., Weng, S., Miyasato, S., Andrada, R., Botstein, D., and Cherry, J. (2005). Saccharomyces genome database. WWW resource, http://www.yeastgenome.org/.

[Bao et al., 2004] Bao, N., Lye, K., and Barton, M. (2004). Microrna binding sites in arabidopsis class iii hd-zip mrnas are required for methylation of the template chromosome. *Dev. Cell*, 7:653–662.

[Berg and von Hippel, 1987] Berg, O. and von Hippel, P. (1987). Selection of dna binding sites by regulatory proteins. i. statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, 193:723–750.

[Berman et al., 2000] Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. (2000). The protein data bank. *Nucleic Acids Research*, 28:238–242.

[Blattner et al., 1997] Blattner, F., III, G. P., Bloch, C., Perna, N., Burland, V., Riley, M., Collado-Vides, J., Glasner, J., Rode, C., Mayhew, G., Gregor, J., Davis, N., Kirkpatrick, H., Goeden, M., Rose, D., Mau, B., and Shao, Y. (1997). The complete genome sequence of escherichia coli k-12. *Science*, 277:1453–1474.

[Bussemaker et al., 2000] Bussemaker, H., Li, H., and Siggia, E. (2000). Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical anaysis. *Proc. Natl. Acad. Sci.*, 97:10098–10100.

[Bussemaker et al., 2001] Bussemaker, H., Li, H., and Siggia, E. (2001). Regulatory element detection using correlation with expression. *Nature Genetics*, 27:167–171.

[Casella and George, 1992] Casella, G. and George, E. (1992). Explaining the gibbs sampler. *Am. Stat. Assoc.*, 46:167–174.

[Cramer et al., 1997] Cramer, P., Larson, C., Verdine, G., and Muller, C. (1997). Structure of the human nf-kappab p52 homodimer-dna complex at 2.1 a resolution. *EMBO J.*, 16:7078–7090.

[Djordjevic et al., 2003] Djordjevic, M., Sengupta, A., and Shraiman, B. (2003). A biophysical approach to transcription factor binding site discovery. *Genome Research*, 13:2381–2390.

[Geman and Geman, 1984] Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE-PAMI*, 6:721–741.

[Honerkamp, 2000] Honerkamp, J. (2000). *Statistical Physics*. Springer, second edition.

[Karplus, 1995] Karplus, K. (1995). Regularizers for estimating distributions of amino acids from small samples. In *Proc. Int. Conf. Intell. Sys. Mol. Biol.*, volume 3, pages 188–196.

[Kirkpatrick et al., 1983] Kirkpatrick, S., Gelatt, C., and Vecci, M. (1983). Optimization by simulated annealing. *Science*, 220:671–680.

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Math. Stat.*, 22:79–86.

[Lagos-Quintana et al., 2001] Lagos-Quintana, M., R, R. R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed rnas. *Science*, 294:853–858.

[Lawrence et al., 1993] Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A., and Wootton, J. (1993). Detecting subtle sequence signals: A gibbs sampling strategy for multiple alignment. *Science*, 262:208–216.

[Lee et al., 1993] Lee, R., Feinbaum, R., and Ambros, V. (1993). The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *Cell*, 75:843–854.

[Nash and Sofer, 1996] Nash, S. and Sofer, A. (1996). *Linear and Nonlinear Programming*, pages 464–475. McGraw-Hill Int.

[N.Metropolis et al., 1953] N.Metropolis, Rosenbluth, A., M.N.Rosenbluth, and Teller, A. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092.

[Pearson and Lipman, 1988] Pearson, W. and Lipman, D. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444–2448.

[Qian, 2001] Qian, H. (2001). Relative entropy: Free energy associated with equilibrium fluctuations and nonequilibrium deviations. *Phys. Rev. E*, 63:042103.

[Rhoades et al., 2002] Rhoades, M., Reinhart, B., Lim, L., Burge, C., Bartel, B., and Bartel, D. (2002). Prediction of plant microrna targets. *Cell*, 110:513–520.

[Salgado et al., 2004] Salgado, H., Gama-Castro, S., Martinez-Antonio, A., Diaz-Peredo, E., Sanchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jimenez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martinez, C., and Collado-Vides, J. (2004). Regulondb version 4.0: Transcriptional regulation, operon organization and growth conditions in escherichia coli k-12. *Nucleic Acids Res.*, 32:303–306.

[Sanger et al., 1978] Sanger, F., Coulson, A., Friedmann, T., Air, G., Barrell, B., Brown, N., Fiddes, J., III, C. H., Slocombe, P., and Smith, M. (1978). The nucleotide sequence of bacteriophage ϕx174. *J. Mol. Biol.*, 125:107–248.

[Schittkowski, 2003] Schittkowski, K. (2003). Ql: A fortran code for convex quadratic programming - user's guide. Technical report, Department of Mathematics, University of Bayreuth.

[Schneider and Stephens, 1990] Schneider, T. and Stephens, R. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, 18:6097–6100.

[Schneider et al., 1986] Schneider, T., Stormo, G., Gold, I., and Ehrenfeucht., A. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188:415–431.

[Schrödinger, 1944] Schrödinger, E. (1944). *What is Life? The Physical aspects of the Living Cell.* Cambridge University Press.

[Shannon and Weaver, 1963] Shannon, C. and Weaver, W. (1963). *The mathematical theory of communication.* Univ. of Illinois Press.

[Shimizu et al., 1997] Shimizu, T., Toumoto, A., Ihara, K., Shimizu, M., Kyogoku, Y., Ogawa, N., Oshima, Y., and Hakoshima, T. (1997). Crystal structure of pho4 bhlh domain-dna complex: flanking base recognition. *EMBO J.*, 16:4689–4697.

[Smith and Waterman, 1981] Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197.

[Stephens and Donnelly, 2003] Stephens, M. and Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, 73:1162–1169.

[Stormo and Hartzell, 1989] Stormo, G. and Hartzell, G. (1989). Identifying protein-binding sites from unaligned dna fragments. *Proc. Natl. Acad. Sci.*, 86:1183–1187.

[Stryer, 2000] Stryer, L. (2000). *Biochemistry.* W.H. Freeman and Company, fourth edition.

[Tompa et al., 2005] Tompa, M., Li, N., Bailey, T., Church, G., Moor, B. D., Eskin, E., Favorov, A., Frith, M., Fu, Y., Kent, W., Makeev, V., Mironov, A., Noble, W., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23:137–144.

[van Kampen, 1985] van Kampen, N. (1985). *Stochastic Processes in Physics and Chemistry*. North-Holland, Amsterdam.

[Walsh, 2004] Walsh, B. (2004). Markov chain monte carlo and gibbs sampling. WWW document, http://nitro.biosci.arizona.edu/courses/EEB581-2004/EEB581.html.

[Watson and Crick, 1953] Watson, J. and Crick, F. (1953). A structure for deoxyribose nucleic acid. *Nature*, 171:737.

[Wightman et al., 1993] Wightman, B., Ha, I., and Ruvkun, G. (1993). Post-transcriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in c. elegans. *Cell*, 75:855–862.

[Wingender et al., 2001] Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhäuser, R., Prüß, M., Schacherer, F., Thiele, S., and Urbach, S. (2001). The transfac system on gene expression regulation. *Nucleic Acids Res.*, 29:281–283.