Statistical models of TF/DNA interaction

—

Licentiate Thesis

Aymeric Fouquier d'Hérouël

February 2008

Abstract

Gene expression is regulated in response to metabolic necessities and environmental changes throughout the life of a cell. A major part of this regulation is governed at the level of transcription, deciding whether messengers to specific genes are produced or not. This decision is triggered by the action of transcription factors, proteins which interact with specific sites on DNA and thus influence the rate of transcription of proximal genes. Mapping the organisation of these transcription factor binding sites sheds light on potential causal relations between genes and is the key to establishing networks of genetic interactions, which determine how the cell adapts to external changes.

In this work I review briefly the basics of genetics and summarise popular approaches to describe transcription factor binding sites, from the most straight forward to finally discuss a biophysically motivated representation based on the estimation of free energies of molecular interactions. Two articles on transcription factors are contained in this thesis, one published (Aurell, Fouquier d'Hérouël, Malmnäs and Vergassola, 2007) and one submitted (Fouquier d'Hérouël, 2008). Both rely strongly on the representation of binding sites by matrices accounting for the affinity of the proteins to specific nucleotides at the different positions of the binding sites.

The importance of non-specific binding of transcription factors to DNA is briefly addressed in the text and extensively discussed in the first appended article: In a study on the affinity of yeast transcription factors for their binding sites, we conclude that measured *in vivo* protein concentrations are marginally sufficient to guarantee the occupation of functional sites, as opposed to unspecific emplacements on the genomic sequence. A common task being the inference of binding site motifs, the most common statistical method is reviewed in detail, upon which I constructed an alternative biophysically motivated approach, exemplified in the second appended article.

Keywords: gene expression, regulation, transcription factor, binding motif, matrix representations, gibbs sampling, binding affinity, non-specific binding

ISBN 978-9171788740 • TRITA-CSC-A 2008:01 • ISSN 1653-5723 • ISRN KTH/CSC/A--08/01--SE

Acknowledgement

The accomplishment of this work would not have been possible without the support of people who I wish to acknowledge. My Ph.D. supervisor *Erik Aurell* at KTH and co-supervisor *Massimo Vergassola* at Institut Pasteur deserve all my thanks for constant encouragement, an immense amount of patience, and especially for offering me the opportunity to extend my research beyond the *rim* of theory.

To my roommate at KTH, *Maria Werner* goes this very special MILLE MERCIS! (cut out and carry with you) for cheering me up when things get stuck and for helping me to express a baseline of confidence. *Harriett Johannson* always offers fast and efficient solutions when it comes to administrative matters and I'd like to thank her together with all the staff in *Hus 16* for providing a very pleasant atmosphere, plenty of caffeinated research fuel and countless cakes.

Many thanks and regards also go to *Pak-Lee Chau* and *Christian Kothe* for their friendship and the good time we spend during my stays in Paris and Stockholm, respectively, and to my parents for all their support.

To *Anna*, finally, goes all of my heart. Plupp.

Contents

List of Figures

Part I.

Groundwork

Chapter 1

Introduction

1.1. Scope of this Work

This Licenciate thesis summarises my work of the past two years which began in the group of Theoretical Biological Physics at KTH by analysing the properties of DNA binding proteins involved in gene regulation. The aim has been to develop novel algorithms for the prediction of DNA binding sites for such regulatory proteins these methods to answer further specific biological questions. The most obvious questions are on the position and strengths of putative binding sites. More involved questions are on the properties of populations of DNA binding proteins and how they interact with a multitude of competing binding sites, random sequences, remains of former binding sites which have changed during evolution and strong sites, whether they have an immediate regulatory function or not.

The models considered here are based on physical arguments of molecular interactions and should be generally comprehensible to anyone with some background in physics or statistics. Simplifying descriptions are however provided where it seems necessary to make the ideas as clear as possible also to the non-specialist readers.

1.1.1. Structure & Overview

Following a statement on the articles which substantiate this thesis, I give a brief description of historical interactions between physics and biology with an emphasis on the impact of physics on biology. This concludes the introductory chapter which is followed by a presentation of the here relevant biological background on gene expression. Transcription and translation regulation are revisited as well as their susceptibility to different sources of randomness. Subsequently, the modelling approaches are discussed, focusing on the interaction between proteins and DNA, and different methods to predict DNA sequences

bound by regulatory proteins. Questions on the functionality of such binding sites at varying protein concentrations are also addressed, since such dependencies might play a major role in restraining the influence of randomness on genetic responses. Following a short summary, results from the articles are discussed. To round off this thesis I present further directions of my work as well as the path I will strike in the future.

1.1.2. Appended Papers

Two papers, as found in the second chapter, accompany this work. Coauthors of the first one were my supervisors Prof Erik Aurell (EA), KTH, Dr Massimo Vergasola (MV), Institut Pasteur, and my former colleague Claes Malmnäs (CM) at KTH.

Article I

The first article investigates the relationship between binding affinities of regulatory proteins in yeast to the organism's DNA and their experimentally determined concentrations.

MV suggested the project, and CM and I wrote a first draft of the article after some initial computations. I wrote parts of the subsequent manuscript versions together with MV and EA who both did the main work on rewriting. I performed all further computations and produced all illustrations.

The article has been published in *Physical Biology* **4**, 134–143 (2007).

Article II

The second article is about a biophysically inspired method of identifying protein binding sites on DNA by stochastically approximating probability distributions describing their expected emplacements.

EA and I discussed the feasibility of the project. I did the literature search, the analytical work, and implemented the algorithm. EA contributed with clarifying discussions, and finally I wrote the manuscript myself.

The implementation is publicly available under the GPL. The manuscript has been submitted and a preprint is available on `http://arxiv.org/` with the reference `arXiv:0802.0258v1 [q-bio.QM]`.

1.2. Physics and the Study of Life

What is life? – Aristotle addressed this question almost two and a half millenia ago (Aristotle, 1931), and when it was posed by Erwin Schrödinger (Schrödinger, 1944) in 1944 it was still far from being answered. The old Greek's account being today of rather philosophical value, Schrödinger's approach still surprises with clarity and wit, condensing the question to facets of its most eminent substance (Advice, 2008). One of today's popular definitions (Wikipedia, 2008) claims seven aspects to be required for something to be *alive*:

Homeostasis – ability to regulate internal functions to maintain an overall constant internal state

Organization – presence of specialized subunits performing different tasks

Metabolism – decomposition of external compounds and internal material into processable components (catabolism), and synthesis of necessary matter from external sources or catabolic products (anabolism)

Growth – maintenance of a higher rate of synthesis than decomposition

Stimulation – response to external signals on a short time-scale

Adaptation – ability to modify the above aspects in response to environmental changes on a long time-scale

Reproduction – autonomous production of *living* offspring

Individual aspects are certainly discussable. Still, if one accepts some of these statements to be valid, how are they orchestrated and controlled? Genetics, as is known, offers parts of the answer and I will give a short introduction to the regulation of genes. Before that, however, let me present a somewhat biased account on physicists and their contributions to biology in the past century, especially their impact on genetics.

Niels Bohr saw no reason why biology should not undergo the same revolution as atomic physics (Bohr, 1933). He hoped for the discovery of *fundamental constants* (Bohr, 1963), much as they had emerged in physics, and sought for "complementarity of mechanistic and teleological descriptions" (McKaughan, 2005) in biology. This was partially satisfied by Linus Pauling's "research into the nature of the chemical bond and its application to the elucidation of the structure of complex substances", awarded the 1954 Nobel Prize in Chemistry, which led to a more detailed understanding of molecular structures, in particular also of biological molecules. James Watson and Francis Crick eventually published their famous work Watson and Crick (1953) on DNA in 1956, for which they were awarded the 1962 Nobel Prize in Physiology or Medicine together with Maurice Wilkins "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material". Bohr's strong influence and support eventually led Max Delbrück to switch from physics to genetics, where he made fundamental "discoveries concerning the replication mechanism and the genetic structure of viruses", amongst others definitely proving DNA to be the main carrier (Rassoulzadegan *et al.*, 2006) of genetic information. Delbrücks work was later awarded the 1969 Nobel Prize in Physiology or Medicine along with Alfred Hershey and Salvador Luria. Erwin Schrödinger, on his part motivated by the developments in physics and admiring the work of Delbrück, endeavoured to popularise the core questions of genetics during his years in Dublin. Today, the properties of biological networks of arbitrary scale have emerged as important research topics, affecting subjects from gene regulation (Alon, 2007; Maslov and Sneppen, 2002) to the understanding of neuronal circuits in the brain (Eguíluz *et al.*, 2005).

Not forgetting formal requisites of this present manuscript and aiming to assure some degree of consistency, the next chapter begins with a recapitulation of some biological knowledge needed to motivate the specific questions which I address.

Chapter 2

Biological Background

Genetics, one of the fundamental disciplines in biology, began with Gregor Mendel's work on the heredity of distinct characters in peas (Mendel, 1866). Towards the end of the 19th century, almost half a century after Mendel's experiments and nearly a decade after his death in 1884, botanists Hugo de Vries and Carl Correns, as well as agronomist Erich Tschermak-Seysenegg independently rediscovered the rules of heredity, today well-known as the laws of *Mendelian inheritance*:

Uniformity – offspring from parents being homozygous in a particular genetic trait is uniformly heterozygous in that trait

Segregation – the fraction of heterozygous and homozygous offspring from parents being heterozygous in a particular gene is equal

Independent Assortment – offspring from parents being homozygous in two particular genetic traits inherits both traits independently from each other

The name *genetics* was coined in the early 20^{th} century by Danish botanist and pharmacist Wilhelm Ludvig Johannsen introducing *genes* as carriers of hereditary information, later defined as being coded by a specific region on a chromosome. Zygotism in Mendel's laws refers to diploid organisms, carrying each chromosome in two copies, one from each parent. A trait is thus called homozygous if the corresponding chromosomes carry the same gene sequence and heterozygous if the sequences are different. Applications for the rediscovered principles were obvious: manipulation of culture and breed, and possible lessons even more tempting: understand life. With discoveries on the structure and function of the DNA, genetic mechanisms became more and more important.

2.1. Regulation of Gene Expression

The expression of genes can be controlled on several levels. Being coded by specific sequences of DNA, they are first *transcribed* to messenger (m)RNA and thereafter *translated* to proteins. Each of these steps characteristic regulatory mechanisms take effect, which are briefly resumed here.

2.1.1. Transcriptional Control

The most important regulation of mRNA synthesis is performed by transcription factor (TF) proteins which bind in vicinity of the transcription start site, in the so called promoter region. A simple model is illustrated in figure 2.1. The template strand of DNA – complementary to the usually annotated coding strand – is transcribed to mRNA by RNA polymerases (RNAP) which binds the promoter regions. Transcription thus means unwinding of DNA, separation of the strands

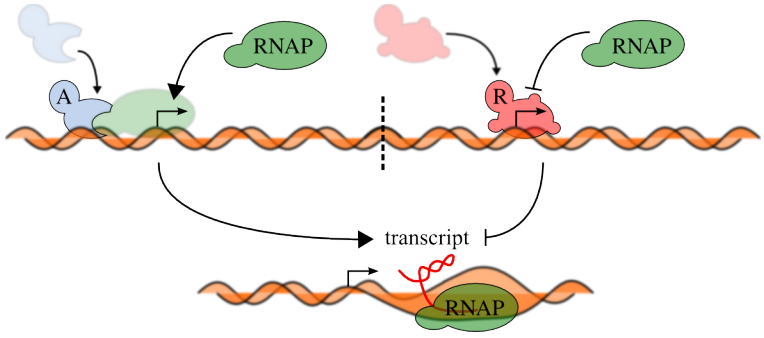


Figure 2.1. *Schematic representation of activation (A) and repression (R) of transcription by DNA binding proteins. Transcription start sites are depicted by arrows and the unwinding of DNA by RNA polymerase is suggested.*

and covalently binding free nucleotides to a polymer of RNA.

The recognition of the binding sites for the polymerase is usually under control by a set of TFs which may either enhance or reduce transcription by facilitating binding of RNAP enzymes to DNA or by blocking the binding site for RNAP. Transcribed mRNAs may then be processed by ribosomes, translating them into proteins, or interact with other molecules in the cell, e.g. metabolites leading to changes in the secondary structure of the mRNA and most prominently other, so called regulatory non-coding (nc)RNA (c.f. Prasanth and Spector (2007) or Eddy (2001)). *Thermus aquaticus* RNAP bound to an essential transcription factor σ^{A} (Murakami *et al.*, 2002) is pictured in figure 2.2. All illustrated structures are, if not otherwise mentionned, retrieved from the Protein Data Bank `http://www.rcsb.org/` (Berman *et al.*, 2003) with access numbers stated in figure captions. The images made with the QuteMol software (Tarini *et al.*, 2006).

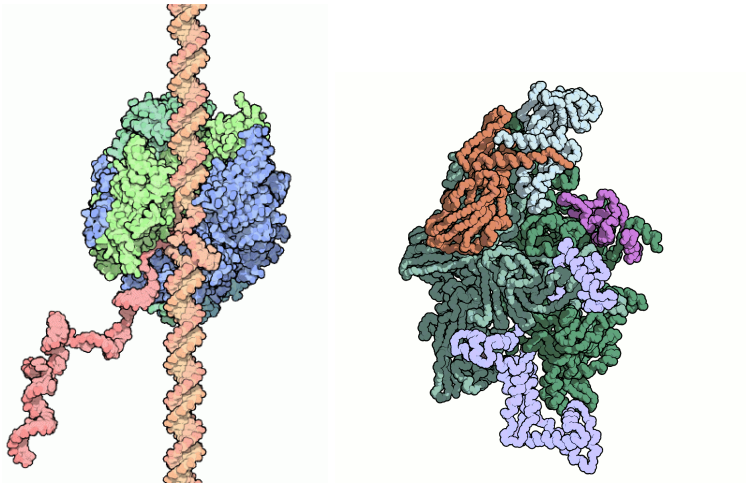


Figure 2.2. left: *Cartoon representation of Saccharomyces cerevisiae RNA polymerase II (PDB ID 1I6H) elongation complex travelling along DNA and synthesising single stranded RNA (composition by David S. Goodsell of The Scripps Research Institute).* right: Thermus aquaticus *(Taq) RNA polymerase (PDB ID 1L9U) holoenzyme with attached DNA binding σ^{A} initiation factor (lilac, shadeless, bottom chain). Protein subunits are displayed in different colours.*

TF proteins can be classified as *basal* and *effecting* factors. The former are usually associated to the RNAP enzyme, then referred to as *holoenzyme*, before it binds to the promoter and are thus relevant for the default behaviour of the holoenzyme. The latter bind independently to the promoter region and either further improve the affinity of the holoenzyme for its binding site, or reduce it, usually by covering that site. Figure 2.3 depicts *Mus musculus* TF Stat3B (signal transducer and activator of transcription 3, isoform B) bound to a short stretch of DNA and *Escherichia coli* Crp (cyclic AMP receptor protein) strongly bending its target seqeunce.

2.1.2. Post-transcriptional Control

Until 1993, gene regulation was believed to be mainly transcriptional. Transcribed functional mRNA was supposed to lead to the synthesis of a corresponding protein and mRNA degradation by RNase proteins was rather seen as passive *garbage collection*. No major post-transcriptional mechanism had been observed in animals until small fragments of RNA, dubbed micro (mi)RNA, with surprising properties were identified in the roundworm *Caenorhabditis elegans* (Lee *et al.*, 1993). These short fragments of ~ 22 nt were processed from parts of the lin-4 mRNA as it was 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 reduced the translation of lin-14 possibly 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 point out 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. Regulation by *short* RNA appears to be a fundamental mechanism in eukaryotes and several different types of such regulators have emerged. Eventually, the 2006 Nobel Prize in Physiology or Medicine was awarded to Andrew Fire and Craig Mello "for their discovery of RNA interference - gene silencing by double-stranded RNA" in the nematode *Caenorhabditis elegans*.

By now, a multitude of ncRNAs have emerged as important post-transcriptional regulators. In prokaryotes, large functional ncRNAs have been identified with the ability of modifying mRNA secondary structures upon hybridisation (see Storz and Haas (2007) for a review), while the short miRNAs and their associated mechanisms have not been observed to date. Furthermore, specific regions of mRNAs have been identified as post-transcriptional regulators of the encoded gene. *Riboswitches* and *thermosensors* denote such regions, which can en- or disable their mRNA's translation by modifying its affinity for the ribosomes in response to the binding of metabolite molecules or a change in environmental temperature, respectively. In the simplest case, this is done by changing the secondary structure such that the ribosome binding site is being exposed or made unavailable.

Summarising, several schemes of post transcriptional regulation have been observed so far

mRNA Modification – ncRNA, thermosensors and riboswitches modify the secondary structure of mRNA, regulating the access of ribosomes to their binding sites. This being mostly observed in prokaryotes, some hints point towards the existence of such mechanisms in higher organisms (Prasanth and Spector, 2007).

mRNA Degradation – miRNA interfers with mRNA, leading to a degradation of the latter. This ability of miRNA and small interfering (si)RNA is frequently observed in plants (Rhoades *et al.*, 2002).

Translation 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.

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 genetic machinery. TF proteins still play a major role in this game but yet unknown participants may be undisclosed any time, enlarging the necessary set of rules – on the other hand enabling us to understand the implications of these very rules.

2.2. Noise in Gene Expression

Based on chemical reactions, gene expression is an intrinsically noisy process (Spudich and Koshland, 1976). Intrinsically not only in the sense that a large network of biochemical reactions may tend to show chaotic behaviour under certain circumstances (see e.g. Aldana and Cluzel (2003) or Pécou (2005)), but also in that translation of mRNA sequences is counteracted by degradation enzymes, a stochastic process at the molecular level. Further, folding of proteins into their functional form is also commonly regarded as stochastic (Gō, 1983). Extrinsic noise may further play a role, stemming from environmental variations of temperature, pressure, radiation or distributions of cellular components, thus rather acting on the level of populations than on single cells (Elowitz *et al.*, 2002).

It appears plausible that regulatory systems need some degree of robustness against intrinsic and, to some degree, extrinsic noise. An important aspect here is the affinity of TF proteins for non-specific binding sites. Such sites may play the role of a *thresholding pool* which has to be filled to a certain extent before functional sites can be bound efficiently. This issue is addressed in more detail in *Article I*, where parts of the results can be interpreted as noise filtering features of transcription regulation.

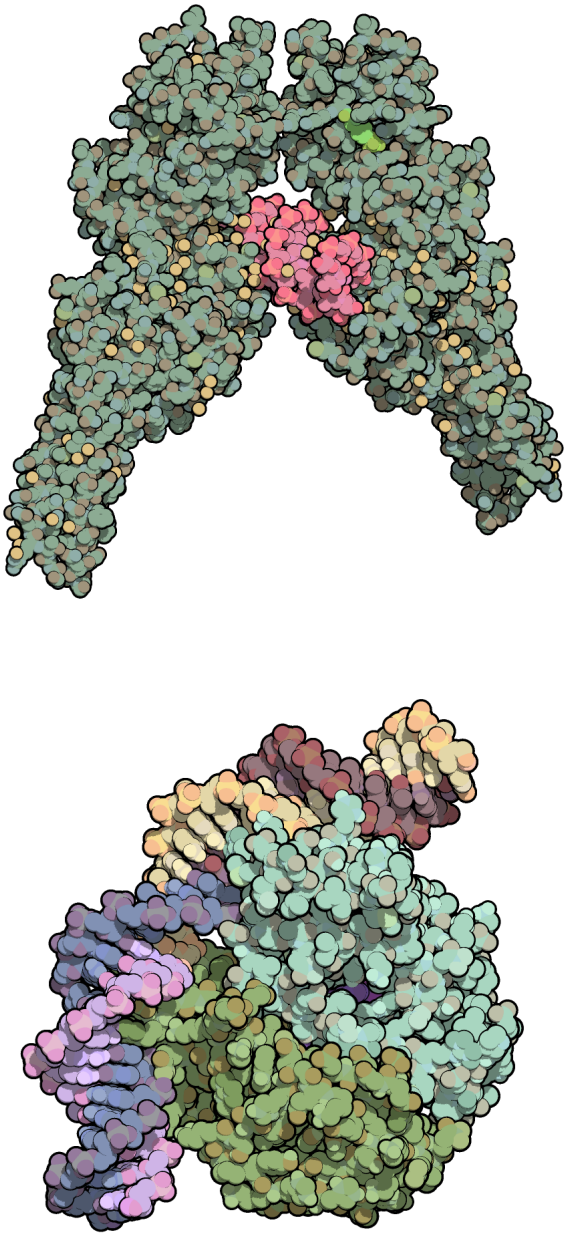


Figure 2.3. *Cartoon representations of the TF proteins Stat3B and Crp*

Chapter 3

Methodological Approach

This chapter deals with the presentation of basic models of TF/DNA interaction. The classic theory is briefly addressed, followed by the fundamentals on binding site motifs and their identification. Thereafter I discuss a biophysical method of motif inference, followed by a detailed presentation of a generic stochastic algorithm and two basic implementations.

3.1. Classic Theory of TF/DNA Interaction

The classic statistical-mechanical theory for the evolution of DNA sequences to TF binding sites (Berg and von Hippel, 1987) is based on three assumptions

Specificity – Binding sites evolve to carry out a specific regulatory task and differ thus from random sequences.

Homogeneity – All sites on a genome satisfying such a task are equally likely to evolve to binding sites.

Additivity – Individual nucleotides of the binding site contribute independently to the total binding energy of a TF.

A set $\mathcal{M} = \{M_1, M_2, \dots\}$ of possible binding sites corresponding to a specific factor is defined by a limited range around the discrimination energy $\bar{E}$,

$$\mathcal{M} = \left\{ S \,:\, |E(S) - \bar{E}| \leq \Delta E \right\} . \tag{3.1}$$

Finding the expected 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

$$f_{\nu i}(\bar{E}) = \frac{\exp(-\lambda \varepsilon_{\nu i})}{\Omega^i_{\bar{E}}(\lambda)} \tag{3.2}$$

with the energy matrix entries $\varepsilon_{\nu i}$ accounting for the energetic contribution of nucleotide ν at position i of the binding site, and the selection parameter λ which plays the role of an inverse temperature. The *canonoical partition function* $\Omega^i_{\bar{E}}$ accounts for the number of accessible states – binding sites in an ensemble of random sequences – around the energy $\bar{E}$ as function of the selection parameter

$$\Omega^i_{\bar{E}}(\lambda) = \sum_{\eta} \exp(-\lambda \varepsilon_{\eta i})\,. \tag{3.3}$$

In principle, λ is not related to the temperature 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, 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

$$\bar{E} = \sum_{i=1}^{l} \sum_{\alpha} \varepsilon_{\nu i} f_{\nu i}\,. \tag{3.4}$$

We will revisit the fundamentals of this theory in the following when describing a method of biomodelling to estimate energy matrices which represent TF binding sites. Especially the ensemble interpretation will be of great use, allowing the computation of statistical quantities from the free energy of binding.

3.1.1. Representation of Binding Motifs

Put very simply, motifs are diffuse patterns on nucleotide sequences sharing a certain degree of similarity. Different models of representation have been elaborated. I describe here the most common ones as well as some terms which are useful to quantify the descriptions. The most straightforward representation using *consensus sequences* is easily established from a sample of nucleotide sequences containing the motif. A majority count of single nucleotides at each position in the sample then defines the corresponding entry in the consensus. Variability between nucleotides at specific positions can be considered by us-

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

Figure 3.1. *IUPAC symbols for nucleic acids*

ing ambiguous IUPAC symbols as listed in figure 3.1 to describe the motif. The simple representation obviously dispenses with relevant information such as

positional importance of the nucleotides in a motif and possible dependencies between nucleotide pairs and triplets.

A more visual representation of a motif is given by its *sequence logo*, introduced by Schneider and Stephens (1990), in which the importance of a nucleotide at each position is related to the size of the corresponding entry. In figure 3.2, two different sequence logos are displayed, along with a weight matrix representation explained in the following paragraph.

A more sensitive representation is found by counting the relative occurrences $f_{\nu i}$ of each nucleotide ν at all positions i of the motif in our collection of N sequences. Doing so, we calculate the entries of a frequency matrix

$$f_{\nu i} = \frac{c_{\nu i} + 1}{N + 4} \tag{3.5}$$

considering a *pseudocount* (Berg and von Hippel, 1987) of one for each nucleotide. With $c_{\nu i}$ we denote the occurrence count of ν at position i of the alignment. Relating the frequency matrix to some nucleotide occurrence probability p_ν, we reach the concept of weight matrices

$$w_{\nu i} = \log \frac{f_{\nu i}}{p_\nu} \tag{3.6}$$

which has been successfully applied by (Stormo and Hartzell, 1989) for the description of protein binding sites.

It is worthwhile to note the relationship between these weight matrices and the energy matrices from above. With the canonical partition function $\Omega^i_{\bar{E}}$, we have

$$\begin{aligned} \varepsilon_{\nu i} &= -\lambda^{-1} \log \left(f_{\nu i} \Omega^i_{\bar{E}} \right) && (3.7) \\ &= -\lambda^{-1} \left[w_{\nu i} + \log \left(p_\nu \Omega^i_{\bar{E}} \right) \right] && (3.8) \\ &= -\lambda^{-1} \left(w_{\nu i} + H_{\nu i} \right) , && (3.9) \end{aligned}$$

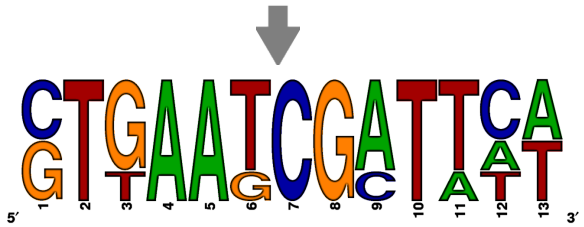
where the shifting term can be identified with the *local entropy* $H_{\nu i}$ of the set $\mathcal{M}$.

Assuming typical nucleotide occurrence probabilities for the gene upstream regions of the *Escherichia coli* genome, *i.e.* $p_{\mathrm{A}} \approx p_{\mathrm{T}} \approx p_{\mathrm{G}} \approx p_{\mathrm{C}} \approx 0.25$, the weight matrix for the FruR alignment is shown in figure 3.2. It is worth to note the usefulness of pseudocounts at this point. As the set of sequences from which we construct $w_{\nu i}$ can be rather small, we might not observe a certain base at a certain position, simply due to possible undersampling. The logarithm in (3.6) would then lead to entries of $-\infty$ in the weight matrix, making the description useless.

Associated to both weight and frequency matrix is the *information score*, measuring how unlikely an alignment is to occur by chance. Formally it is related to the probability of observing the set of positional occurrences $\{c_{\nu i}\}$ in 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

$$P(\{c_{\nu i}\} \mid \{p_\nu\}) = \prod_{k=1}^{L} N! \prod_{\eta} \frac{(p_\eta)^{c_{\eta k}}}{c_{\eta k}!} . \tag{3.10}$$

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



	w_{A}	w_{C}	w_{G}	w_{T}
1	−0.88	0.63	0.63	−0.88
2	−0.88	−0.47	−0.47	0.73
3	−0.88	−0.47	0.92	−0.18
4	0.73	−0.47	−0.47	−0.88
5	0.73	−0.47	−0.47	−0.88
6	−0.88	−0.47	0.22	0.51
7	−0.88	1.14	−0.47	−0.88
8	−0.88	−0.47	1.14	−0.88
9	0.51	0.22	−0.47	−0.88
10	−0.88	−0.47	−0.47	0.73
11	−0.18	−0.47	−0.47	0.51
12	−0.18	0.63	−0.47	−0.18
13	0.22	−0.47	−0.47	0.22

Figure 3.2. *From consensus sequence to weight matrix representation of a binding motif for the FruR TF protein of* Escherichia coli. *The first sequence logo corresponds to the occurrence counts and the second is scaled by the positional information score.*

The product over η is apparently the generation probability for an unordered set of nucleotides at position k, the columns of the alignment, and to get the probability of generating the whole alignment, we have to evaluate the product over all such columns. Writing the occurrences as $c_{\nu i} \approx N \cdot f_{\nu i}$ – with equality when omitting pseudocounts – and making use of the Stirling approximation $\log(n!) \approx n \log n - n$, we find

$$\begin{aligned} P(\{c_{\nu i}\} \mid \{p_\nu\}) &= \prod_{k=1}^{L} \exp(\log N!) \prod_{\eta} \exp\left(N f_{\eta k} \log \frac{p_\eta}{(N f_{\eta k})!} \right) \\ &\approx \exp\left(-N \sum_{\eta k} f_{\eta k} \log \frac{f_{\eta k}}{p_\eta} \right) \equiv \exp(-nI)\,, \end{aligned} \tag{3.11}$$

with the information score (Schneider *et al.*, 1986) of an alignment described by the frequencies $f_{\nu i}$ and weights as $w_{\nu i}$

$$I = \sum_{k=1}^{L} I_k = \sum_{\eta k} f_{\eta k} w_{\eta k}\,. \tag{3.12}$$

The positional information score was implicitly defined as I_k. Due to the sign in (3.11), these quantities increases as the probability for the alignment to occur by chance decreases. The information score thus gives a possibility to assess the quality of an alignment by judging how unexpected it is.

3.1.2. Identification of TF Binding Sites

After building a motif model the search for binding sites on the whole genome can begin. Sites similar to a given consensus sequences are easily found by pattern matching algorithms (Rice *et al.*, 2000), allowing for specified numbers of mismatches. Matrix representations $m_{\nu i}$ of a motif of length L, are readily evaluated on the genomic sequence $S = (\alpha_1, \alpha_1, \dots)$ by calculating the positional score

$$R(a) = \sum_{k=1}^{L} m_{\alpha_k k}\,, \tag{3.13}$$

where a is the first position of the assumed motif on S. Depending on the matrix type, a threshold score for the acceptance of a putative instance of the motif has to be defined. In the case of frequency and weight matrices this task is not obvious *a priori* and a top-down approach eventually analysing the best scoring matches first often seems to be the best approach when no further biological information is available. Such information can be, for instance, a detailed model of the sequence topology around some of the relatively high-scoring matches or experimental data pointing towards a regulatory relationship between nearby sequences. We will see below that energy matrices are free from this issue.

3.2. Inference of Binding Motifs

3.2.1. Basic Variants

Lexical Analysis

Given the abundance of genomic material, attempts have been made to identify 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 with limited success. Still further attempts are made to deduce *context-free grammars* from small sequence samples (Dyrka and Nebel, 2007). A more promising algorithm (Bussemaker *et al.*, 2001) fitted a lexical motif model to the experimental data of gene expression levels, thus deducing a descriptive pattern.

Others (see *e.g.* Pavesi *et al.* (2004) for an overview) try to find functional patterns using stochastic methods often based on the sampling algorithm which I describe below. Still simplicity and pragmatism often beats realism, as was most strikingly noted in a recent assessment of motif inference tools (Tompa *et al.*, 2005) on a variety of eukaryotic datasets ranging from yeast to human sequences. The pattern alignment algorithm by Pavesi *et al.* (2007) is constructed on a heuristic set of penalty rules for mismatches and gaps and performed better than its competitors on most datasets.

Biomodelling

A different approach to the inference of motif representations is to consider the *free energy* of binding from a transcription factor bound to a DNA molecule. This quantity can be expanded in interaction terms of different order, all depending on the sequence S of length L

$$E(S) = \sum_{i}^{L}\sum_{\alpha}^{4} \varepsilon_{\nu i} S_\nu^i + \sum_{ij}^{L}\sum_{\nu\eta}^{4} J_{ij}^{\nu\eta} S_\nu^i S_\eta^j + \sum_{ijk}^{L}\sum_{\nu\eta\kappa}^{4} Q_{ijk}^{\nu\eta\kappa} S_\nu^i S_\eta^j S_\kappa^k + \dots \tag{3.14}$$

where all subscripts are integers starting at one and the greek letters are control variables for the four base types at arbitrary convention. We already encountered the basics of this concept in section 3.1, where I omitted the precise definition of the free energy $E(S)$.

The nucleotide sequence $S = (\alpha_1, \alpha_2 \dots, \alpha_N)$ is represented by its *indicators*, defined as

$$S_\nu^i = \begin{cases} 1 & \text{if } \alpha_i = \nu \\ 0 & \text{otherwise} \end{cases} . \tag{3.15}$$

Since the corrections by the higher order terms can be assumed to be small, only the linear approximation is kept for modelling. Additionally, high order corrections implicate the introduction of high-dimensional objects as $J_{ij}^{\nu\eta}$, $Q_{ijk}^{\nu\eta\kappa}$,

etc, circumventing the use of the simplifications which I detail below. It is convenient to think in terms of matrices

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{A1} & \dots & \varepsilon_{AL} \\ \varepsilon_{C1} & \dots & \varepsilon_{CL} \\ \varepsilon_{G1} & \dots & \varepsilon_{GL} \\ \varepsilon_{T1} & \dots & \varepsilon_{TL} \end{pmatrix} \quad \text{and} \quad \mathbf{s} = \begin{pmatrix} S_A^1 & \dots & S_A^L \\ S_C^1 & \dots & S_C^L \\ S_G^1 & \dots & S_G^L \\ S_T^1 & \dots & S_T^L \end{pmatrix}, \tag{3.16}$$

entailing the energy to be given by the Frobenius inner product

$$E(S) \approx \boldsymbol{\varepsilon} : \mathbf{s} = \mathrm{tr}(\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{s}), \tag{3.17}$$

or equivalently as vectors

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

and

$$\mathbf{s} = (S_A^1, S_G^1, S_T^1, S_C^1; \dots; S_A^L, S_G^L, S_T^L, S_C^L)^{\mathrm{T}}. \tag{3.18}$$

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. Let the probability to generate the sequence S be P_S. Upon equilibration, N TFs are extracted with their associated sequence. The likelihood of observing a set $\mathcal{M} = \{S_1, S_2, \dots, S_N\}$ of N sequences and *no other* sequence is then

$$\mathcal{P} = \prod_{S \in \mathcal{M}} [\gamma P_S f(E(S) - \mu)] \prod_{S' \notin \mathcal{M}} [1 - \gamma P_{S'} f(E(S') - \mu)] \tag{3.19}$$

where $f(E - \mu)$ is the *Fermi-Dirac distribution* and μ is the *chemical potential* relating the TF concentration and its rate of binding to DNA. Both terms are discussed in more detail below. 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 $e^{-x} \approx 1 - x$, equation (3.19) can be simplified to

$$\mathcal{P} \approx \prod_{S \in \mathcal{M}} [\gamma P_S f(E(S) - \mu)] \exp\left(\sum_{S' \notin \mathcal{M}} [-\gamma P_{S'} f(E(S') - \mu)] \right) \tag{3.20}$$

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 (3.20) we hence get

$$\mathcal{L} = N \log \gamma + \sum_{S \in \mathcal{M}} \log[P_S f(E(S) - \mu)] - \gamma \sum_{S' \notin \mathcal{M}} [P_{S'} f(E(S') - \mu)] \tag{3.21}$$

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

A short excursion shows why the binding probability can be assumed to be Fermi-Dirac distributed. One considers the simple kinetic reaction describing the binding of a TF do DNA with reaction constants k_a and k_d

$$\mathrm{TF} \; + \; \mathrm{DNA} \underset{k_\mathrm{d}}{\overset{k_\mathrm{a}}{\rightleftharpoons}} \; \mathrm{TF} \circ \mathrm{DNA} \tag{3.22}$$

This represents a pair of coupled ordinary first-order differential equations and the system can be interpreted as being in 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

$$\partial_t [\mathrm{TF} \circ \mathrm{DNA}] = k_\mathrm{a} [\mathrm{TF}][\mathrm{DNA}] - k_\mathrm{d} [\mathrm{TF} \circ \mathrm{DNA}] \equiv 0 \tag{3.23}$$

leading to

$$\frac{[\mathrm{TF} \circ \mathrm{DNA}]}{[\mathrm{TF}][\mathrm{DNA}]} = \frac{k_\mathrm{a}}{k_\mathrm{d}} = K \cdot \exp(-\beta E)\,, \tag{3.24}$$

where equality to the right comes from the two-state model. β is the inverse temperature $(k_\mathrm{B} T)^{-1}$ in units of the Boltzmann constant and E stands for the free energy of binding, while K is an inverse equilibrium concentration.

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

$$P_\mathrm{b}(S) = \frac{[\mathrm{TF} \circ \mathrm{DNA}]}{[\mathrm{TF} \circ \mathrm{DNA}] + [\mathrm{DNA}]} \tag{3.25}$$

into which one can insert the statement of equation (3.24), obtaining

$$\begin{aligned} P_\mathrm{b}(S) &= \left(1 + \frac{[\mathrm{DNA}]}{[\mathrm{TF} \circ \mathrm{DNA}]}\right)^{-1} \\ &= \left(1 + \frac{\exp(\beta E(S)}{K \cdot [\mathrm{TF}]}\right)^{-1} \\ &= \frac{1}{1 + \exp[\beta(E(S) - \mu)]}\,, \end{aligned} \tag{3.26}$$

with the chemical potential $\mu = k_\mathrm{B} T \log(K \cdot [\mathrm{TF}])$, yielding the Fermi-Dirac distribution. Note that μ has the same form as in an ideal gas with particle concentration $[\mathrm{TF}]$.

To further simplify equation (3.20), one considers the border case of all TFs being bound, claiming $T \to 0$ or equivalently $\beta \to \infty$. In this limit, the Fermi-Dirac distribution turns into the Heaviside step 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

$$\sum_{S' \notin \mathcal{M}} P_{S'} f(E(S') - \mu) = \int_{-\infty}^{\infty} dE \rho_\varepsilon(E) f(E - \mu) \xrightarrow{T \to 0} \int_{-\infty}^{\mu} dE \rho_\varepsilon(E)\,. \tag{3.27}$$

The continuous density function ρ_ε may be approximated by a Gaussian distribution as long as E is close to the mean energy. This is a central assumption

in Djordjevic *et al.* (2003) and is assumed to hold for the set of unobserved sequences. Initially keeping the temperature finite, this yields a simplified likelihood function

$$\mathcal{L} = N \log\gamma + \sum_{S\in\mathcal{M}} \log[P_S f(E(S)-\mu)] - \gamma \int dE \rho_\varepsilon(E) f(E-\mu)\,, \tag{3.28}$$

to be maximised in the $(\varepsilon, \mu, \gamma)$ space. Requiring the variations to vanish gives

$$\partial_{\varepsilon_{\nu i}}\mathcal{L} = -N\log\gamma \sum_{S\in\mathcal{M}} [1-f(E(S)-\mu)]\cdot\beta S^i_\nu + \gamma\int dE f(E-\mu)\partial_{\varepsilon_{\nu i}}\rho_\varepsilon(E) \equiv 0 \tag{3.29}$$

$$\partial_\mu\mathcal{L} = N\log\gamma \sum_{S\in\mathcal{M}} [-f(E(S)-\mu)]\cdot\beta - \gamma\beta\int dE\rho_\varepsilon(E) f(E-\mu)[1-f(E-\mu)] \equiv 0 \tag{3.30}$$

$$\partial_\gamma\mathcal{L} = \frac{N}{\gamma} - \int dE\rho_\varepsilon(E) f(E-\mu) \equiv 0\,. \tag{3.31}$$

The extraction factor γ turns out to be the only non-transcendent parameter

$$\gamma = \frac{N}{\int dE\rho_\varepsilon(E) f(E-\mu)}\,. \tag{3.32}$$

Using this result and the zero temperature approximation in (3.28) yields a simplified maximisation problem

$$\max_{\boldsymbol{\varepsilon},\mu}\left\{ N\left(\log N - \int_{-\infty}^{\mu} dE\rho_\varepsilon(E)\right)\right\}\,, \tag{3.33}$$

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

$$\min_{\boldsymbol{\varepsilon},\mu}\left\{ \int_{-\infty}^{\mu} dE\rho_\varepsilon(E) = \mathrm{erf}\left(\frac{\mu-\bar{E}}{\sigma}\right)\middle|\, \boldsymbol{\varepsilon}\in\mathbb{R}^4\times\mathbb{R}^L,\ \mu\in\mathbb{R},\ E(S)\le\mu\ \forall\ S\in\mathcal{M}\right\} \tag{3.34}$$

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

$$\mu = \max_{S\in\mathcal{M}} E(S)\,, \tag{3.35}$$

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

$$\left\{\boldsymbol{\varepsilon}\in\mathbb{R}^4\times\mathbb{R}^L\ :\ \sigma^2(\boldsymbol{\varepsilon}) = \min_{\boldsymbol{\varepsilon}'}\sigma^2(\boldsymbol{\varepsilon}')\right\} \tag{3.36}$$

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

$$\begin{cases} \underset{\boldsymbol{\varepsilon}}{\text{minimise}} & \sigma^2(\boldsymbol{\varepsilon}) = \sum_{\nu\eta}\sum_{ij}\varepsilon_{\nu i}P(\nu,i;\eta,j)\varepsilon_{\eta j} \\ \text{subject to} & E(S) = \boldsymbol{\varepsilon}:\mathbf{s}\le -1\ \forall\ S\in\mathcal{M} \end{cases}\,. \tag{3.37}$$

The $P(\nu, i; \eta, j)$ are derived from a statistical model of the genomic background, describing the probabilities of observing the nucleotide ν at position i in the motif, while η is observed at position $j > i$. To construct the elements $P(\nu, i; \eta, j)$, consider the occurrence probabilities p_ν of nucleotide ν and the stochastic matrix $\mathbf{T}$ of probabilities of observing ν followed by η

$$(\mathbf{T})_{\nu\eta} = P(\nu \mid \eta)\,. \tag{3.38}$$

Both p_ν and $\mathbf{T}$ are readily constructed from genomic sequences. Further introducing the vectors $\mathbf{P}_{,\nu}$ and $\mathbf{P}_{\eta,}$ with elements

$$(\mathbf{P}_{,\nu})_\alpha = P(\alpha \mid \nu) \quad \text{and} \quad (\mathbf{P}_{\eta,})_\alpha = P(\eta \mid \alpha)\,, \tag{3.39}$$

we can explicitely calculate the genomic background model

$$P(\nu, i; \eta, j) = p_\nu \left[(1 - \delta_{ij})\mathbf{P}_{\eta,}\mathbf{T}^{j-i-1}\mathbf{P}_{,\nu} + \delta_{\nu\eta}\delta_{ij}\right]\,, \tag{3.40}$$

with the discrete delta function on the indices a and b defined as

$$\delta_{ab} = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases}\,. \tag{3.41}$$

The problem of estimating the $\varepsilon_{\nu i}$ is thus reduced to a quadratic optimisation problem, which can be handled by standard *quadratic programming*. Let us first summarise the system of equations of (3.37) as

$$\begin{cases} \underset{\boldsymbol{\varepsilon}}{\text{minimise}} & \frac{1}{2}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{P}\boldsymbol{\varepsilon} \\ \text{subject to} & \boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{S} + \mathbf{1} \leq \mathbf{0} \end{cases} \tag{3.42}$$

with the vector $\mathbf{S}$ of sequence matrices from each of the N sequence extractions,

$$\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N]\,. \tag{3.43}$$

$\mathbf{1}$ and $\mathbf{0}$ are the N dimensional vectors of ones and zeros, respectively. $\mathbf{P}$ can be interpreted as the Hessian of the variance. Its elements being probabilities, $\mathbf{P}$ is guaranteed to be positive semi-definite, leading to a convex optimisation problem. Consequently, there exists only a global optimum. Lagrange optimisation with constraints requires the system – introducing the vector of multipliers $\boldsymbol{\lambda}$ – to be

$$\begin{cases} \underset{\boldsymbol{\varepsilon},\boldsymbol{\lambda}}{\text{minimise}} & \frac{1}{2}\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{P}\boldsymbol{\varepsilon} + \boldsymbol{\lambda}\left(\boldsymbol{\varepsilon}^{\mathrm{T}}\mathbf{S} + \mathbf{1}\right)^{\mathrm{T}} \\ \text{subject to} & \mathbf{P}\boldsymbol{\varepsilon} + \mathbf{S}\boldsymbol{\lambda} = \mathbf{0} \wedge \boldsymbol{\lambda} \geq 0 \end{cases}\,. \tag{3.44}$$

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

$$\begin{cases} \underset{\boldsymbol{\lambda}}{\text{maximise}} & -\frac{1}{2}\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{P}^{-1}\mathbf{S}\boldsymbol{\lambda} + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{1} \\ \text{subject to} & \boldsymbol{\lambda} \geq \mathbf{0} \end{cases}\,. \tag{3.45}$$

This dual problem is equivalent to the primal one (Nash and Sofer, 1996) and has only $\boldsymbol{\lambda}$ left as free parameter.

Zero entries in the sequence matrix $\mathbf{S}$ lead inevitably to the loss of some information, "flattening" the manifold on which to find an optimum. To counter this, we perform a shifting operation on the sequences, defining for the entries of each sequence matrix $\mathbf{s}_j$

$$(\hat{\mathbf{s}}_j)_{\nu i} \equiv (\mathbf{s}_j)_{\nu i} - p_\nu \,. \tag{3.46}$$

This transformation does not affect the optimisation problem but rescales the chemical potential, which was arbitrarily set to unity, as in equation (3.44).

To reconstruct the estimated energy matrix we have to evaluate

$$\varepsilon = -\mathbf{P}^{-1}\hat{\mathbf{S}}\boldsymbol{\lambda}\,, \tag{3.47}$$

obtaining the energy matrix ε in terms of the chemical potential μ.

Solving the dual problem yields a computational benefit when optimising via quadratic programming. The matrix of the dual quadratic form is $\mathbf{S}\mathbf{P}^{-1}\mathbf{S}$ and therewith of rank N instead of $4L$ for P in the primal problem. In general we have $N \ll 4L$, and one has just to perform one matrix inversion and two multiplications, while the optimisation procedure evaluates the quadratic form numerous times. The smaller the problems dimension, the faster the evaluation.

Figure 3.3 visualises the idea behind the likelihood maximisation at finite temperature. For $T \to 0$ the sigmoidal distribution becomes a sharp step. The

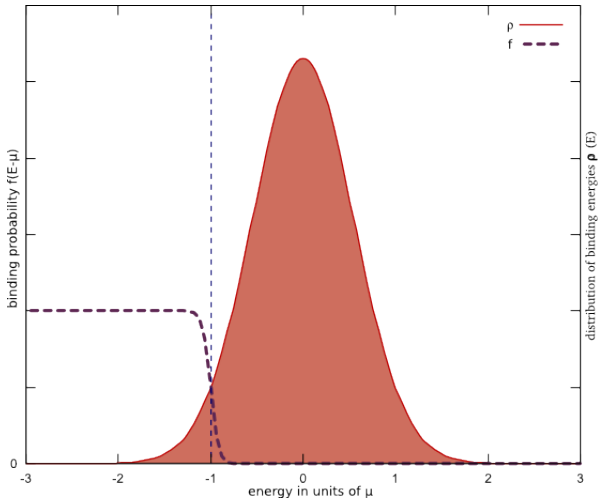


Figure 3.3. *Distribution of binding energies and sequence binding probability*

domain of overlapping Fermi-Dirac and Gauss distributions is a measure for the amount of binding sites that are wrongly not being considered in the model (3.19). By minimising the variance of the Gaussian, the probability diminishes,

maximising the likelihood of only finding the observed set of sequences in the thought experiment.

A calibration can be done respective to a known set of binding sequences, in order to find an estimate for the binding-free-energies ε. 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.

3.2.2. Stochastic Method

Suppose we do not have exact binding sites given from which to infer a motif representation, but instead a set $\mathcal{S} = \{S_i \mid i = 1 \dots N, \|S_i\| = L_i\}$ of N nucleotide sequences of lengths L_i, each containing zero or more binding sites for the same TF. What we would like to set up is the joint probability distribution

$$p_{[N]}(a_1, a_2, \dots, a_N) \tag{3.48}$$

of the positions of binding sites on each sequence and analyse the most probable configurations given $\mathcal{S}$. However, this task turns out to be hard to perform in general, and approximative methods are used instead. The *Gibbs sampler* is such a method, based on the Metropolis-Hastings algorithm (Hastings, 1970), which allows the estimation of statistical quantities from a set of random variables when detailed information about their distributions is missing.

Let me start by describing this method in a more general way before coming back to the problem of sequence alignment. We are interested in the properties of random variables X and $Y_1, \dots, Y_n$. This nomenclature is just to skip the index of the first random variable whose properties we will estimate. Such properties can be for instance the *joint* probability distribution of $n + 1$ variables

$$p_{[n+1]}(x, y_1, y_2, \dots, y_n) \tag{3.49}$$

or any *marginal* distribution from which we can extract information on the corresponding variable independently of the others, i.e.

$$p_{[1]}(x) = \int dy_1 \int dy_2 \dots \int dy_n \; p_{[n+1]}(x, y_1, y_2, \dots, y_n) \tag{3.50}$$

Note that when the random variables X and Y_i are generated from a Markovian process, meaning the conditionals distributions depend only on their predeceding neighbours, fulfilling $p_{[1|n]}(y_i \mid x, y_1, y_2, \dots, y_{i-1}, y_{i+1}, \dots, y_n) = p_{[1|1]}(y_i \mid y_{i-1})$, it holds for the marginal distribution that

$$p_{[1]}(x) = \frac{p_{[n]}(x, y_1, y_2, \dots, y_n)}{p_{[1|1]}(y_n \mid y_{n-1}) \cdots p_{[1|1]}(y_2 \mid y_1) \cdot p_{[1|1]}(y_1 \mid x)}\,, \tag{3.51}$$

with arbitrary realizations y_i of the random variables Y_i. In such cases, the marginal distribution of x is accessible without performing any integration over the set $\{y_i\}$.

The generic scheme is to draw a sequence $X^{(0)}, X^{(1)}, \ldots, X^{(k)}$ of the random variable $X \sim p_{[1]}(x)$ and to use this sequence to approximate specific properties of the marginal distribution $p_{[1]}(x)$ or even the distribution itself without knowing it explicitly. The sequence is drawn from the conditional probabilities $p(x|y_1, y_2, \ldots, y_n)$ and $p(y_i|x, y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$ in an iterative procedure starting with randomly chosen realizations $\hat{y}_1, \ldots, \hat{y}_n$, sequentially updated as new realisations of Y_i become available

$$\begin{array}{lcl} X^{(0)} & \sim & p_{[1|n]}(x \mid \{Y_k = \hat{y}_k\,,\ k = 1 \ldots n\})\,, \\ Y_i^{(0)} & \sim & p_{[1|n]}(y_i \mid X = x^{(0)}, \{Y_{j<i} = y_j^{(0)}, Y_{j>i} = \hat{y}_i\})\,, \\ & \cdots & \\ X^{(t)} & \sim & p_{[1|n]}(x \mid \{Y_k = y_k^{(t-1)}\,,\ k = 1 \ldots n\})\,. \end{array} \tag{3.52}$$

Each drawing of $X^{(k)}$ is thus preceded by n drawings $Y_1^{(k-1)}, \ldots, Y_n^{(k-1)}$, and as $k \to \infty$, it is possible to reach arbitrary precision in the approximation. The marginal density can now be estimated from conditionals

$$\hat{p}_{[1]}(x) = k^{-1} \sum_{i=0}^{k} p_{[1|n]}(x) \ \stackrel{k\to\infty}{\longrightarrow}\ p_{[1]}(x)\,, \tag{3.53}$$

while for instance the expectation value of X can be approximated by

$$\mathrm{E}_k[X] = k^{-1} \sum_{i=0}^{k} X^{(i)} \ \stackrel{k\to\infty}{\longrightarrow}\ \mathrm{E}[X]\,, \tag{3.54}$$

both consequences of the Rao-Blackwell theorem (Barton, 1961). I refer to more specialised literature for a rigorous mathematical treatment (Tanner, 1998) and a variety of examples (Casella and George, 1992).

In the context of motif inference on the $\mathcal{S} = \{S_i \mid i = 1 \ldots N\,,\ \|S_i\| = L_i\}$, let the set of alignment boxes be $\mathcal{A}_\mathcal{S} = \{(a_i, w_i) \mid i = 1 \ldots N\,,\ a_i \in [1, L_i - w_i + 1]\,,\ w_i \in [1, L_i]\}$, with specific positions and widths on each S_i, as illustrated in figure 3.4. The random variables are now discrete alignment positions A_i and widths W_i on each sequence S_i with realizations a_i and w_i, respectively. Density functions representing the probabilities of observing a given alignment are thus replaced by discrete probability distributions of the form

$$p_{[2N]} : \prod_{i=1}^{N} ([1, L_i - w_i + 1] \times [1, L_i]) \ \to\ (0, 1) \quad | \quad \|p_{[2N]}\| = 1\,, \tag{3.55}$$

and of interest are the marginal distributions $p_{[2]}(A_i = a_i, W_i = w_i)$, describing the probabilities to find a binding site of width w_i at position a_i on sequence S_i.

To illustrate the probabilistic framework and clarify the relationship between the marginals and the actual presence of binding sites, let the alignment widths w_i be given and fixed to the same value w, thus pruning the set of random variables. This reduces the dimensionality of the space of random variables in order to simplify the notation and leaves us with $A_1, \ldots, A_N$ or alternatively the N-dimensional random variable $\mathcal{A}$. To this aim, let also the sequence lengths

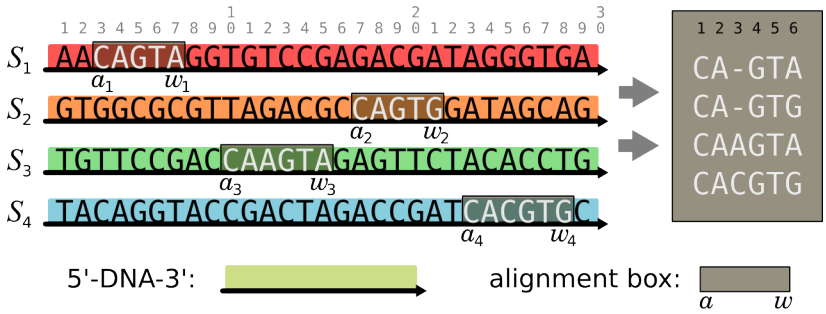


Figure 3.4. *Alignment boxes of a set of four sample sequences with* $a_1 = 3$, $a_2 = 17$, $a_3 = 10$ *and* $a_4 = 23$. *Different alignment widths* $w_1 = w_2 = 5$ *and* $w_3 = w_4 = 6$ *correspond to a gapped alignment, as shown to the right.*

L_i be fixed to L, constraining the alignment positions to $a_i \in [1, \hat{L}]$, with the rightmost alignment position $\hat{L} = L - w + 1$, leading to distributions

$$p_{[N]} : [1, \hat{L}]^N \;\to\; (0, 1) \quad | \quad \|p_{[N]}\| = 1\,. \tag{3.56}$$

Marginals of interest are now $p_{[1]}(a_i)$ and conditional probabilities for finding a binding site at a_i on sequence S_i, depending on the alignments on the other sequences $S_{j\neq i}$, can be explicitly written as

$$p_{[1|N-1]}(A_i = a_i \mid \{A_{j\neq i} = a_j\}) = \frac{p_{[N]}(\{A_k = a_k \mid k = 1 \dots N\})}{\sum_{b_i=1}^{\hat{L}} p_{[N]}(A_i = b_i, \{A_{j\neq i} = a_j\})}\,, \tag{3.57}$$

which will be evaluated in the following paragraphs on the *Gibbs Motif Sampler*. The definition of conditional probabilities further yields the expression for distributions with reversed dependence of the alignments on the $S_{j\neq i}$,

$$p_{[N-1|1]}(\{A_{j\neq i} = a_j\} \mid A_i = a_i) = \frac{p_{[N]}(\{A_k = a_k \mid k = 1 \dots N\})}{\sum_{\{b_{j\neq i}=1\}}^{\hat{L}} p_{[N]}(A_i = a_i, \{A_{j\neq i} = b_j\})}\,. \tag{3.58}$$

In principle, the conditional distributions are easily calculated for the actual problem, as illustrated below, and upon construction, they can be directly applied to iteratively sample a sequences of alignments $\mathcal{A}^{(0)}, \mathcal{A}^{(1)}, \dots, \mathcal{A}^{(k)}$. Each $\mathcal{A}^{(i)}$ represents a set of N random alignments $A_1^{(i)}, \dots, A_N^{(i)}$ on the i^{th} nucleotide sequence from $\mathcal{S}$. Before focusing on an actual implementation, however, let me make plausible why the marginal distribution is recovered from from the sampled alignments. Using the conditional distributions, the transition probabilities for the alignment on a specific nucleotide sequence to go from $A_i \to A_i'$ with realizations $a_i \to a_i'$ in any iteration can be written as

$$\hat{p}_{[1|1]}(A_i' \mid A_i) = \sum_{\{a_{j\neq i}=1\}}^{\hat{L}} p_{[1|N-1]}(A_i' \mid \{A_{j\neq i}'\}) \cdot p_{[N-1|1]}(\{A_{j\neq i}\} \mid A_i)\,, \tag{3.59}$$

where the notation was shortened by the attribution of realisations to the random variables. These conditional probabilities define a *homogeneous* Markov process (Papoulis, 1991), since they are independent of the actual iteration. They can be expressed as $\hat{L} \times \hat{L}$ transfer matrix $\hat{\mathbf{P}}_{[i]}$ for the alignment transition on sequence S_i with $\hat{p}_{[1|1]}$ as elements

$$\hat{P}_{[i],a_i,a_i'} = \hat{p}_{[1|1]}(A_i' = a_i' \mid A_i = a_i)\,, \tag{3.60}$$

with

$$\sum_{a_i} \hat{P}_{[i],a_i,a_i'} = 1\,. \tag{3.61}$$

In a sequence of sampled alignments $\mathcal{A}^{(0)}, \mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(k)}$, let us now assume that we know the marginal probability distribution for the specific random variable $A_i^{(k)}$ after the k^{th} iteration. In fact, we could have estimated this distribution by creating a great number of independent sequences of alignment, each with k iterations, and averaging over the realizations obtained for each final $A_i^{(k)}$, again as consequence of the Rao-Blackwell theorem (Barton, 1961).

The marginal distribution of $A_i^{(k)}$ can on its part be expressed as $\hat{L}$ dimensional vector $\mathbf{p}_{[i]}^{(k)}$ with elements

$$p_{[i],a_i}^{(k)} = p_{[1]}(A_i^k = a_i)\,, \tag{3.62}$$

and we can readily verify that the marginal distribution could as well have been calculated from any earlier iteration using powers of the transfer matrix

$$\mathbf{p}_{[i]}^{(k)} = \hat{\mathbf{P}}_{[i]}\mathbf{p}_{[i]}^{(k-1)} = \hat{\mathbf{P}}_{[i]}^2\mathbf{p}_{[i]}^{(k-2)} = \cdots = \hat{\mathbf{P}}_{[i]}^k\mathbf{p}_{[i]}^{(0)}\,. \tag{3.63}$$

Further, the Perron-Frobenius theorem (Graham, 1987) guarantees the existence of a stationary distribution $\mathbf{p}_{[i]}$ satisfying the eigenvalue equation

$$\mathbf{p}_{[i]} = \hat{\mathbf{P}}_{[i]}\mathbf{p}_{[i]}\,. \tag{3.64}$$

even for non trivial transfer matrices, and it is clear from the above that this distribution is reached by infinite sampling

$$\lim_{k\to\infty} \mathbf{p}_{[i]}^{(k)} = \mathbf{p}_{[i]}\,. \tag{3.65}$$

We can hence define simple criteria for the convergence of $\mathbf{p}_{[i]}^{(k)}$ by monitoring its evolution and deciding at which iteration to stop the sampling. Hereafter, we may decide what to learn from the process. Estimating marginal distributions was the initial task, but in practice just one of the alignments $\mathcal{A}^{(i)}$, *e.g.* the most common, is often of greater interest, thus representing a snapshot of the joint distribution of alignments on all sequences.

Gibbs Motif Sampler

To give a concrete example of the estimation of optimal alignments, let me summarise the basic version of the stochastic algorithm by Lawrence *et al.* (1993). In figure 3.5, the quantities computed during the sampling iterations are displayed. The algorithm itself proceeds as follows:

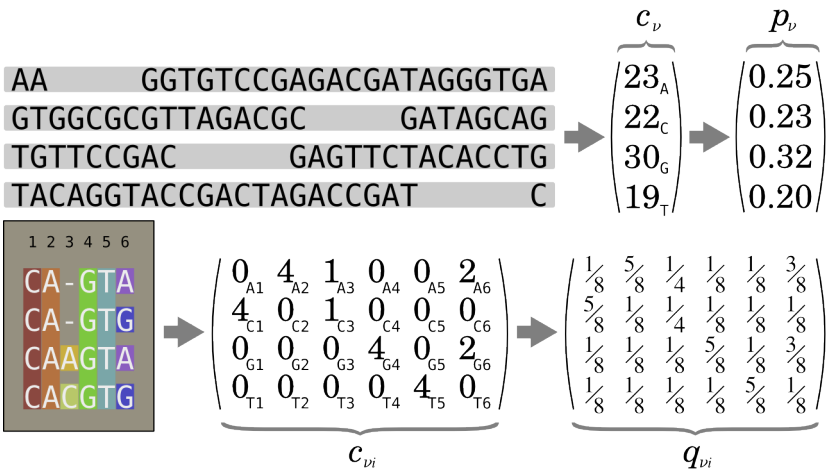


Figure 3.5. *Evaluation of the model matrices $c_{\nu i}$ and $q_{\nu i}$, and the background vector p_ν on the* complete *set of sequences from the previous example in figure 3.4.*

Initialization – Random alignment positions a_k are chosen on each sequence.

Predictive update step – The sequence S_k is removed from $\mathcal{S}$, where k can be chosen arbitrarily or in a specific order covering all members of $\mathcal{S}$. From the remaining set we calculate the p_ν and the $q_{\nu i}$ according to

$$p_\eta = \frac{c_\nu + b_\nu}{N - 1 + B} \quad \text{and} \quad q_{\nu i} = \frac{c_{\nu i} + b_{\nu i}}{N - 1 + B} \tag{3.66}$$

where occurrences of ν (at position i) are denoted by c, and *pseudocounts* by b and B, respectively (Berg and von Hippel, 1987). Optimal values for these regularisers are intuitively $b_\nu = b_{\nu i} = 1$ and $B = 4$, which can be shown with rigour (Karplus, 1995).

Sampling step – Every alignment box (a, w) in $S_k = (\alpha_1, \ldots, \alpha_L)$ is being considered as instance of the pattern and we calculate the background and model probabilities $P_w(a)$ and $Q_w(a)$ for the alignment box from

$$P_w(a) = \prod_{i=a}^{a+w} p_{\alpha_i} \quad \text{and} \quad Q_w(a) = \prod_{i=a}^{a+w} q_{\alpha_i i}\,. \tag{3.67}$$

From those we can directly construct the conditional probability for the alignment on S_k given the other alignments as

$$p_{[1|N-1]}(A_k = a \mid \{A_{j\neq k} = a_j\}) \sim \frac{Q_w(a)}{P_w(a)}\,, \tag{3.68}$$

and draw a new alignment position on S_k before iterating with the following predictive update step.

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. Formally, the fixed point of equation (3.65) is approximated.

Quadratic Programming Sampler

Instead of building the probability distribution of new alignment positions from motif frequency matrices $f_{\nu i}$ and background probabilities p_ν, we can directly construct it from the probabilities for a TF protein to bind to the nucleotide sequences in $\mathcal{S}$. Let us first introduce the alignment probability distribution $\hat{P}_\varepsilon(x)$ for a binding motif at position x on sequence $S = (\alpha_1, \alpha_2, \ldots, \alpha_L)$, which is constructed from the binding probabilities P_b from equation (3.26) by setting

$$\hat{P}_\varepsilon(x) = \frac{P_\mathrm{b}((\alpha_x, \alpha_{x+1}, \ldots, \alpha_{x+w}))}{\sum_{x'=1}^{L-w+1} P_\mathrm{b}((\alpha_{x'}, \alpha_{x'+1}, \ldots, \alpha_{x'+w}))} \,. \tag{3.69}$$

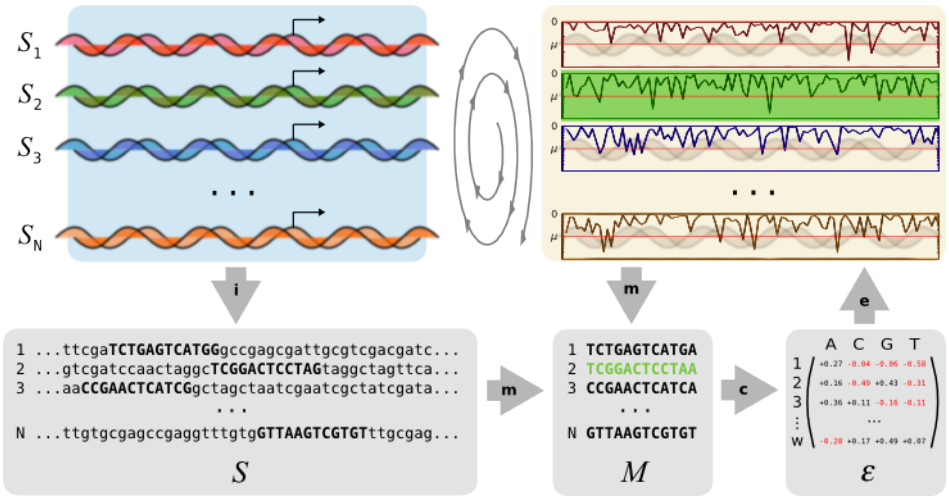


Figure 3.6. *Schematics of the QPS algorithm taking N input sequences S_1 to S_N with initialisation* **i**, *motif extraction* **m**, *matrix computation* **c**, *and evaluation steps* **e**. *S_2 is highlighted as being retained for the updating of a_2.*

The optimal local alignment is then approximated upon convergence of the following procedure:

Initialisation – Random alignment positions a_i, $i = 1, \ldots, N$ are set on each of the input sequence.

Model construction – The sequence motif $\mathcal{M}$ of $N - 1$ sequences is extracted, excluding sequence S_i.

Matrix computation – The energy matrix ε of the motif $\mathcal{M}$ is computed.

Sequence evaluation – ε is evaluated on the excluded sequence S_i using R_ε.

Model construction' – A new alignment position a_i is drawn on S_i from the alignment probability distribution $\hat{P}_\varepsilon(a)$, exclude a new sequence $S_{i'}$ and iterate with a subsequent **matrix computation**.

Figure 3.6 illustrates the procedure in which each iteration draws a new alignment on the skipped sequence S_i. The algorithm is described in some more detail in *Article II*.

Chapter 4

Results & Discussion

4.1. Article I – Numbers and Affinity

TFs exert their regulatory functions by binding to nucleotide sequences on DNA, but they do not solely bind to sequences related to a specific function. Indeed TFs are thought to bind mostly non-specifically to DNA. We report in this article on the properties of such non-specific binding described by an effective background free energy F_{b} describing the affinity of a *single* TF for a random stretch of DNA. Other quantities of interest in this study were the chemical potential μ of the type of considered TFs as well as their binding energy E^* to the optimal binding site.

We analysed the relationship between F_{b}, E^* and the average amount n_{obs} of TF proteins measured in exponentially growing yeast. As demonstrated in the article, the relationship between chemical potential, protein amount and background free energy can be written as

$$\mu \simeq \beta^{-1} \log n + F_{\mathrm{b}} \,. \tag{4.1}$$

We argue that the chemical potential, representing an energetic threshold for strong binding, is on average of the order of the best binding energy E^*, thus requiring the background free energy to satisfy

$$F_{\mathrm{b}} \approx E^* - \beta^{-1} \log n_{\mathrm{t}} \,, \tag{4.2}$$

with a thresholding amount n_{t} guaranteeing occupation of the specific binding sites. This statement is in contrast to earlier assertions in the literature (Gerland *et al.*, 2002) which we discuss, claiming the free background energy to neglect the $\log n_{\mathrm{t}}$ term, thus non-specific binding to be much weaker, i.e.

$$F_{\mathrm{b}} \approx E^* \,. \tag{4.3}$$

We calculated the energies in question by adopting both weight and energy matrix representations of 63 TFs of *Saccharomyces cerevisiae*. Figure 4.1 shows

our assumption of a strong binding background to be satisfied on average, while the claim of a weaker F_{b} appears to hold only for a small number of TFs. This on average behaviour observed using experimentally measured amounts further suggests that *in vivo* TF concentrations as measured in exponential growth are close to the threshold amounts n_{T}. A direct consequence of this conjecture is that the *background pool* of non-specific binding sites is filled up before specific binding sites can be occupied with high probability, thus presenting a natural barrier for genetic responses to spurious TF productions.

4.2. Article II – Quadratic Programming Sampling

In this article I describe the implementation of a Gibbs sampler procedure making use of the energy matrices described previously. The idea behind the development of this sampler was to make direct use of the promising biophysical description of TF binding sites by such matrices and present an alternative to the widely used *Gibbs Motif Sampler* which is implicitely based on a frequency matrix description of motif sequences.

I validated the functionality of QPS on a small set of coregulated promotors in in *Escherichia coli* consisting of aceBAKp, icdAp, pckAp, and ptsHp. Each region contains an experimentally known binding site for the fructose repressor protein FruR, which I tried to infer. Figure 4.2 shows the evolution of expected alignment positions on the small set of promoter sequences. The heat maps illustrate the probability distributions of alignment positions which was assumed to be stationary if it did not change within five iterations. The principal functionality of the sampler procedure has been verified. A larger-scale assessment on realistic biological data is necessary to argue for or against the quality of predictions made by QPS.

The benchmark proposed by Tompa *et al.* (2005) has attracted some attention as test of popular inference tools. It is composed of 52 individual data sets of gene upstream regions of yeast (8), fruitfly (6), mouse (12) and human (26). Each data set is made of one to 35 sequences of lengths varying between 500 and 3000 nucleotides and contains zero to 76 experimentally known binding sites (Wingender *et al.*, 1996).

The very construction of the data sets, however, makes the *bona fide* evaluation of QPS hard, since no recognition of multiple binding sites or the discrimination of uninformative sequences (not containing binding sites) has yet been implemented. Extensive preprocessing of the sequences is necessary to apply QPS and draw conclusions on its applicability. Still, lacking extensive knowledge on the content of the sequences in the data sets makes it difficult to differ between wrong motif predictions and possibly unknown relationships. A recent discussion of the benchmark by Sandve *et al.* (2007) addresses these issues and tries to reduce the dependence on preprocessing and the bias due to incomplete knowledge and results of QPS on the reviewed benchmark are pending.

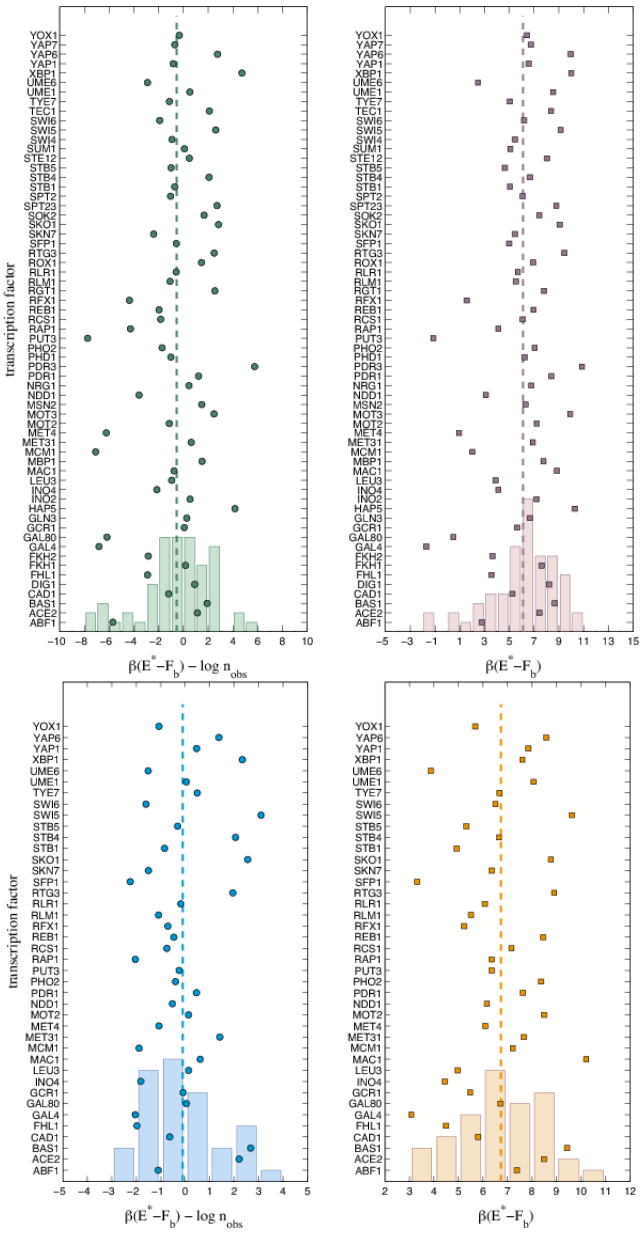


Figure 4.1. *Comparison of the relation between the background energy F_b and the abundance for a set of S. cerevisiae transcription factors. Values of the difference between the consensus energy E^* and the background energy F_b are reported as squares. Their values shifted by the logarithm of the TF abundance (as measured experimentally) are reported as circles. Vertical dashed lines correspond to the average values for the two sets of points. Points have a sizeable scatter but circles are clearly centered around zero. No relation has been found between the deviation of the points around zero and the functional role of the corresponding TFs (upper: results for log-odds ratio matrices; lower: results for energy matrices). Histograms give better visual access to the distribution widths.*

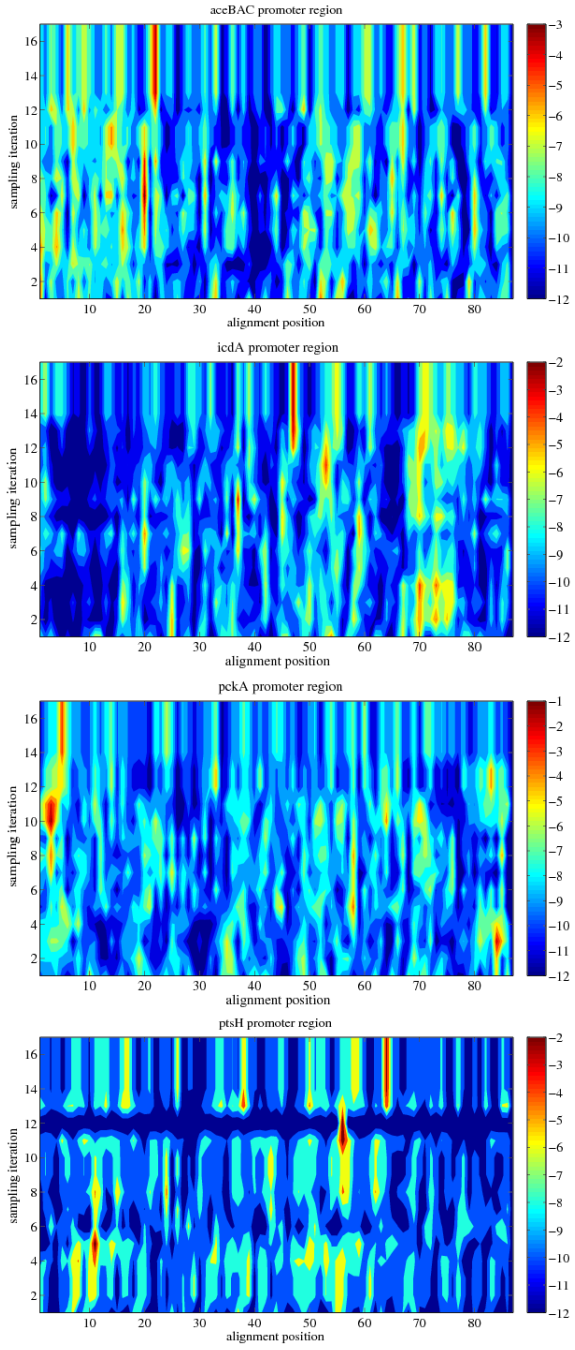


Figure 4.2. *Logarithmic heat maps of the evolution of alignment position probability distributions on the promoter regions of different operons in* Escherichia coli. *The regions contain one known TF binding site for FruR each and were aligned by QPS. At iteration 14, the distributions have reached their stationary form.*

Chapter 5

Outlook & Perspectives

The topics I introduced here represent my entry point into the physics of regulatory systems, and are yet largely of algorithmic nature. Concluding this initial work on transcription regulation by TF proteins, I will pursue my research in this vast and to large parts still unexplored domain, reorienting on a different kind of regulator: non-coding RNA. An immense layer of regulation by non-coding RNA of a wide range of sizes, from tenths to several thousands of nucleotides, and functions is about to emerge, and it seems opportune to focus on its mechanistic description. Moreover, RNA being a rather simple polymer compared to proteins, fundamental research on its properties may be more fruitful than on the latter. Concretely, I will emphasise the two following aspects in my future research...

5.1. Transcription Regulation by non-coding RNA

The broad range of regulatory interactions between ncRNA and mRNA as briefly mentioned in a few examples in chapter 2.1.2 is just a small part of the picture. ncRNA has the ability to exert regulatory functions by interacting not only with other RNAs, but also with regulatory proteins, competing with TF binding sites on DNA. Another recently discovered surprising functions is for instance the mimicking of open promoter structures on DNA, as prominently performed by 6S-RNA in *Escherichia coli*.
This part of my research will be partially experimental since I had the opportunity to establish ties with biology labs at the French National Institute for Agricultural Research. The common interest on the *ab initio* identification of ncRNA will hopefully lead to exciting results. We currently work on a genomic map of ncRNA in the opportunistic lactic ferment *Enterococcus faecalis*.

5.2. Dynamics of RNA Secondary Stucture Formation

Detailed understanding of the time scales of RNA folding, unfolding, interaction with hybridisation targets, separation from those and eventually refolding into a stable form is becoming more and more necessary as the importance of RNA-RNA interactions emerges. This will answer questions on principal functioning and efficiency of ncRNA in presence of degrading enzymes *in vivo*. A choice of questions I tend to look into are the following. How fast is the overall folding process? How does it depend on local alterations of the sequence? How do the interaction dynamics depend on the target sequence? Can we design stable ncRNAs with higher efficiency than wildtype sequences? How about target genes with varied susceptibility? Which *minimal mutations* optimise or disable the interaction? Can we quantify the requirements for an efficient ncRNA regulator?

Bibliography

ADVICE, A., Continuously check Citations, This Thesis, **2008**.

ALDANA, M. AND CLUZEL, P., A natural class of robust networks, *Proc Natl Acad Sci U S A*, 100:8710–8714, **2003**.

ALON, U., Network motifs: theory and experimental approaches, *Nature Reviews Genetics*, 8:450–461, **2007**.

ARISTOTLE, *De Anima*, Clerendon Press, Oxford, **1931**, translated into English by J.A. Smith, see also *Historia Animalium* and *De Generatione Animalium* in *The complete works of Aristotle : the revised Oxford translation* edited by Jonathan Barnes.

BARTON, D., Unbiased Estimation of a Set of Probabilities, *Biometrika*, 48:227–229, **1961**.

BERG, O. AND VON HIPPEL, P., Selection of DNA binding sites by regulatory proteins. I. Statistical-mechanical theory and application to operators and promoters, *J. Mol. Biol.*, 193:723–750, **1987**.

BERMAN, H., HENRICK, K. AND NAKAMURA, H., Announcing the worldwide Protein Data Bank, *Nature Structural Biology*, 10:980, **2003**.

BOHR, N., Licht und Leben, *Die Naturwissenschaften*, 21:245–250, **1933**.

BOHR, N., Licht und Leben – noch einmal, *Die Naturwissenschaften*, 50:725–727, **1963**.

BUSSEMAKER, H., LI, H. AND SIGGIA, E., Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical anaysis, *Proc. Natl. Acad. Sci.*, 97:10098–10100, **2000**.

BUSSEMAKER, H., LI, H. AND SIGGIA, E., Regulatory Element Detection using Correlation with Expression, *Nature Genetics*, 27:167–171, **2001**.

CASELLA, G. AND GEORGE, E., Explaining the Gibbs sampler, *The American Statistician*, 46:167–174, **1992**.

DJORDJEVIC, M., SENGUPTA, A. AND SHRAIMAN, B., A Biophysical Approach to Transcription Factor Binding Site Discovery, *Genome Research*, 13:2381–2390, **2003**.

DYRKA, W. AND NEBEL, J.-C., A probabilistic context-free grammar for the detection of binding sites from a protein sequence, *BMC Systems Biology*, 1(P78), **2007**, poster.

EDDY, S., Non-coding RNA genes and the modern RNA world, *Nature Reviews Genetics*, 2:919–929, **2001**.

EGUÍLUZ, V., CHIALVO, D., CECCHI, G., BALIKI, M. AND APKARIAN, A., Scale-Free Brain Functional Networks, *Physical Review Letters*, 94:018102, **2005**.

ELOWITZ, M., LEVINE, A., SIGGIA, E. AND SWAIN, P., Stochastic Gene Expression in a Single Cell, *Science*, 297:1183–1186, **2002**.

GERLAND, U., MOROZ, J. AND HWA, T., Physical constraints and functional characteristics of transcription factor-DNA interaction, *Proceedings of the National Academy of Sciences*, 99(19):12015–12020, **2002**.

GŌ, N., Protein folding as a stochastic process, *Journal of Statistical Physics*, 30:413–423, **1983**.

GRAHAM, A., *Nonnegative Matrices and Applicable Topics in Linear Algebra*, Ellis Horwood, Chichester, **1987**.

HASTINGS, W., Monte Carlo Sampling Methods Using Markov Chains and Their Applications, *Biometrika*, 57:97–109, **1970**.

KARPLUS, K., Regularizers for Estimating Distributions of Amino Acids from Small Samples, in *ISMB-95*, **1995**.

LAGOS-QUINTANA, M., RAUHUT, R., LENDECKEL, W. AND TUSCHL, T., Identification of novel genes coding for small expressed RNAs, *Science*, 294:853–858, **2001**.

LAWRENCE, C., ALTSCHUL, S., BOGUSKI, M., LIU, J., NEUWALD, A. AND WOOTTON, J., Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment, *Science*, 262:208–216, **1993**.

LEE, R., FEINBAUM, R. AND AMBROS, V., The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14, *Cell*, 75:843–854, **1993**.

MASLOV, S. AND SNEPPEN, K., Specificity and Stability in Topology of Protein Networks, *Science*, 296:910–913, **2002**.

MCKAUGHAN, D., The Influence of Niels Bohr on Max Delbrück, *Isis*, 96:507–529, **2005**.

MENDEL, G., Versuche über Pflanzen-Hybriden, *Verhandlungen des Naturforscher-Vereins Brünn*, 4:3–47, **1866**.

MURAKAMI, K., MASUDA, S. AND DARST, S., Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 A resolution, *Science*, 296:1280–1284, **2002**.

NASH, S. AND SOFER, A., *Linear and Nonlinear Programming*, pp. 464–475, McGraw-Hill Int., **1996**.

PAPOULIS, A., *Probability Random Variables, and Stochastic Processes*, McGraw-Hill College, **1991**.

PAVESI, G., MAURI, G. AND PESOLE, G., *In silico* representation and discovery of transcription factor binding sites, *Briefings in Bioinformatics*, 5:217–236, **2004**.

PAVESI, G., ZAMBELLI, F. AND PESOLE, G., WeederH: an algorithm for finding conserved regulatory motifs and regions in homologous sequences, *BMC Bioinformatics*, 8(46), **2007**.

PÉCOU, E., Splitting the dynamics of large biochemical interaction networks, *Journal of Theoretical Biology*, 232:375–384, **2005**.

PRASANTH, K. AND SPECTOR, D., Eukaryotic regulatory RNAs: an answer to the *genome complexity* conundrum, *Genes & Development*, 21:11–42, **2007**.

RASSOULZADEGAN, M., GRANDJEAN, V., GOUNON, P., VINCENT, S., GILLOT, I. AND CUZIN, F., RNA-mediated non-mendelian inheritance of an epigenetic change in the mouse, *Nature*, 441:469–474, **2006**.

RHOADES, M., REINHART, B., LIM, L., BURGE, C., BARTEL, B. AND BARTEL, D., Prediction of plant microRNA targets, *Cell*, 110:513–520, **2002**.

RICE, P., LONGDEN, I. AND BLEASBY, A., EMBOSS: The European Molecular Biology Open Software Suite, *Trends in Genetics*, 16:276–277, **2000**.

SANDVE, G., ABUL, O., WALSENG, V. AND DRABLOS, F., Improved benchmarks for computational motif discovery, *BMC Bioinformatics*, 8:193, **2007**.

SCHNEIDER, T. AND STEPHENS, R., Sequence logos: a new way to display consensus sequences, *Nucleic Acids Research*, 18:6097–6100, **1990**.

SCHNEIDER, T., STORMO, G., GOLD, I. AND EHRENFEUCHT., A., Information content of binding sites on nucleotide sequences, *J. Mol. Biol.*, 188:415–431, **1986**.

SCHRÖDINGER, E., *What is Life? The Physical aspects of the Living Cell*, Cambridge University Press, Cambridge, **1944**.

SPUDICH, J. AND KOSHLAND, D., Non-genetic individuality: chance in the single cell, *Nature*, 262:467–471, **1976**.

STORMO, G. AND HARTZELL, G., Identifying Protein-Binding Sites from Unaligned DNA Fragments, *Proc. Natl. Acad. Sci.*, 86:1183–1187, **1989**.

STORZ, G. AND HAAS, D., Cell regulation, *Current Opinion in Microbiology*, 10(2), **2007**, RNA special issue.

TANNER, M., *Tools for Statistical Inference*, Springer, **1998**.

TARINI, M., CIGNONI, P. AND MONTANI, C., Ambient Occlusion and Edge Cueing for Enhancing Real Time Molecular Visualization, *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1237–1244, **2006**, ISSN 1077-2626.

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., Assessing computational tools for the discovery of transcription factor binding sites, *Nature Biotechnology*, 23:137–144, **2005**.

WATSON, J. AND CRICK, F., A structure for Deoxyribose Nucleic Acid, *Nature*, 171:737, **1953**.

WIGHTMAN, B., HA, I. AND RUVKUN, G., Post-transcriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans, *Cell*, 75:855–862, **1993**.

WIKIPEDIA, Life — Wikipedia, The Free Encyclopedia, **2008**, URL `http://en.wikipedia.org/Life`, [Online; accessed 10-January-2008].

WINGENDER, E., DIETZE, P., KARAS, H. AND KNÜPPEL, R., TRANSFAC®: A database on transcription factors and their DNA binding sites, *Nucleic Acids Research*, 24:238–241, **1996**.

Part II.

Publications