Evolution of Genetic Networks

I n a u g u r a l - D i s s e r t a t i o n

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Stana Willmann

aus

Belgrad

Köln

2005

.

Berichterstatter: Prof. Dr. Michael Lässig
Prof. Dr. Dietrich Stauffer

Tag der mündlichen Prüfung: 12. Januar 2005

Contents

Summary / Zusammenfassung

Abstract

In this work we investigate structure and evolution of regulatory sequences. The regulation of a gene depends on the binding of transcription factors to specific sites located in the regulatory region of the gene. The generation of these binding sites and of cooperativity between them are essential building blocks in the evolution of complex regulatory networks. We study a theoretical model for the sequence evolution of binding sites, which includes point mutations, selection, and genetic drift. Using empirically grounded fitness landscapes, we demonstrate the possibility of selective sweeps generating a new binding site for a given transcription factor.
We develop a code based on position weight matrices to detect transcription factor binding sites. In parallel we use other scoring schemes for comparison. We search computationally for coregulated genes at an inter-species (human), and intra-species (human and mouse) level. This part of the thesis was carried out in collaboration with Prof. Nikolaus Rajewsky at NYU.
Knowing that HIV-1 promoter is mutating fast, we also tried to determine its transcription factor binding site structure in order to detect possible sites for which there is no experimental evidence. In doing so, we detected all well known binding sites, and discovered a novel site whose role is to be checked experimentally.
Using the HIV-1 regulatory sequences, we have constructed a consistent phylogeny. Furthermore we observed many point mutations but few indels. Long branches of the tree are seen to correspond to emergence of new subtypes. We have carried out a differential analysis of mutations along the branches of the tree for the subsequence containing putative binding sites and the background. This study has shown an enhanced rate of substitutions per nucleotide for the site regions along the inter-subtype branches of the phylogeny consistent with positive selection for changes within the binding sites. Reasonable gauge of selective versus random mutations ratio is presented.
We predict the fitness of different subtypes in specific well defined environments from their sequence. The relationship between the fitness and the binding probability for individual factors is quantified. This part of the thesis was carried out in close collaboration with Professor Nikolaus Rajewsky.

Zusammenfassung

Gegenstand dieser Arbeit ist die Untersuchung der Struktur und Evolution genetischer regulatorischer Sequenzen. Die Regulierung eines Gens ist abhängig von der Bindung von Transkriptionsfaktoren an spezifische

Bindungsstellen in regulatorischen Regionen der Gene. Die Entstehung solcher Bindungsstellen sowie kooperative Effekte zwischen Ihnen sind die grundlegenden Komponenten der Evolution komplexer regulatorischer Netzwerke.
Wir untersuchen ein theoretisches Modell für die Evolution der Sequenzen von Bindungsstellen, das punktweise Mutationen, Selektion und genetische Drift beinhaltet. Unter verwendung von Fitness-Landschaften, die auf empirischen Daten basieren demonstrieren wir, daß selektive sweeps neue Bindungsstellen für einen gegebenen Transkriptionsfaktor generieren können. Dieser Teil der Arbeit wurde in Zusammenarbeit mit Professor Nikolaus Rajewsky an der New York University durchgeführt.
Wir entwickeln einen Algorithmus, der auf positionsabhängigen Gewichtsmatrizen basiert, um Bindungsstellen für Transkriptionsfaktoren zu identifizieren. Parallel dazu verwenden wir andere Bewertungsschemata zu Vergleichszwecken. Wir führen eine rechnergestütze Suche nach korregulierten Genen sowohl auf Inter- (Mensch) als auch Intra-Spezies Ebene (Mensch und Maus) durch.
Es ist bekannt, daß der HIV-1 Promoter schnell mutiert. Deshalb versuchen wir, die Sktruktur der Transkriptionsfaktor-Bindungsstellen zu bestimmen, mit dem Ziel, mögliche Bindungsstellen zu identifizieren, für die noch keine experimentellen Hinweise bekannt sind. Die Aufklärung dieser Struktur führt nicht nur zum Auffinden aller bisher bekannten Bindungsstellen, sondern liefert Hinweise auf eine neue Bindungsstelle, nach der nun experimentell gesucht wird.
Unter Verwendung regulatorischer HIV-1 Sequenzen konstruieren wir einen konsistenten phylogenetischen Stammbaum. Wir beobachten hier eine grosse Anzahl von Punktmutationen, aber wenige insertions und deletions. Lange Zweige des phylogenetischen Stammbaums werden als Entstehung neuer Subtypen gesehen. Wir führen eine differentielle Analyse der Mutationen entlang des Stammbaums durch, und zwar sowohl für solche Teilsequenzen, die mögliche Bindungsstellen enthalten, als auch für den Hintergrund. Diese Studie hat eine erhöhte Rate von Substitutionen pro Nukleotid für die Regionen entlang der Inter-Subtyp Zweige der Phylogenie gezeigt, was konsistent ist mit positiver Selektion für Anderungen innerhalb der Bindungsstellen. Eine plausible Abschätzung des Verhältnisses von selektiven zu zufälligen Mutationen wird präsentiert.
Innerhalb von definierten Umgebungen sagen wir die Fitness verschiedener Subtypen anhand der Sequenz vorher. Das Verhätnis zwischen Fitness und der Bindungswahrscheinlichkeit wird für einzelne Faktoren quantifiziert. Dieser Teil der Arbeit wurde in Zusammenarbeit mit Professor Nikolaus Rajewsky erstellt.

Chapter 1

Preface

"What is life?" is the title of Schrödinger's book [2], a physicist's book on biology. Physicists have played an important role in life science research ever since, trying to answer major questions on the origin of life and its development.

It is well known that species adapt to their environment in order to survive [3]. This adjustment occurs via internal changes resulting in improved phenotype (an organism's body with all its characteristics). A physicist, Max Delbrück was one of the first [4, 5] to distinguish between random and non random variations. During the 1940's, Delbrück and Luria performed experiments with the bacterium *E. coli* and viruses infecting bacteria - phages. They conjectured that changes to phage resistence are rare and occur spontaneously.

Some ten years later, the physicist Francis Crick wrote a letter to Delbrück informing him about his work with James Watson, on the discovery of a double-helix molecule - the famous DNA. They also proposed that DNA could specify characteristics of an organism. Soon it became clear that inner changes bringing about phenotype adaptation, turned out to be mutations within the DNA molecule [6].

DNA or 'the book of life' is universal, it encodes the organism's building instructions in humans, animals, plants, bacteria and viruses, thus representing its genome. With the advance of technology, it became possible to identify all letters in the 'book of life' in human, mouse, fly, worm, rice etc ([97–99]). Projects of 'writing down' the sequence of letters in different species are taking place with an immense enthusiasm. Still, we are ignorant about many parts of the sequence forming the DNA. The challenge is to decipher the 'text' in the book. What is known is that there are coding and noncoding text. Coding text describes how proteins, building blocks of all organisms, must be produced. Proteins can be divided into a constitutive and a regulatory group. While constitutive proteins build up an organism, regulatory proteins (also known as Transcription Factors) play 'guardians' that control how much of the protein should be produced. Noncoding text is still a mystery. Some parts of it are responsible for controlling protein production, and these 'chapters' in the 'book of life' are known as regulatory sequences or promoters. They contain certain words (letter motifs), known as Transcription Factor Binding Sites (TFBS) that bind Transcription Factors (TF). Once TFs are bound to corresponding binding sites, a gene gets expressed and protein production starts. This preparatory process of TFs attaching to TFBSs is called transcription initiation (see Fig 1.1). After that, a portion of

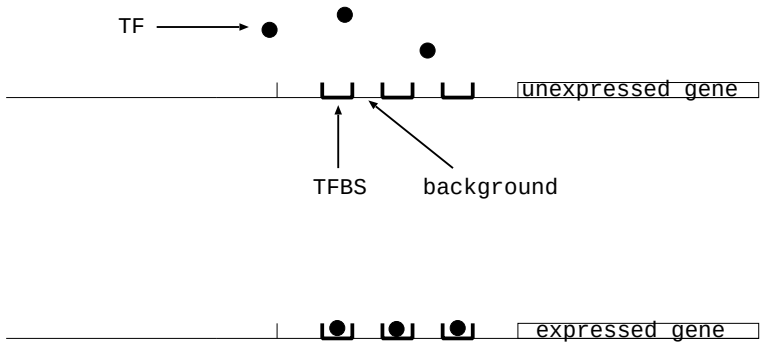

Figure 1.1: Transcription initiation

DNA is transcribed (in a process of transcription) and the transcript is then transformed into a protein (in a process called translation) (see Fig. 2.2).

In our study of the evolution of genetic networks, we want to gauge mutational events by using powerful concepts of statistical physics and bioinformatics.

In the chapter 2, we try to generalize and quantify empirical evidence on transcriptional regulation via general statistical methods. We show how thermodynamic concepts may be applied to TF-TFBS interactions. Thermodynamics of a single and several TFBSs is discussed, together with bioinformatical approaches for specific TFBS motif detection. We try to link the binding energy of a TFBS to a bioinformatical analog, motif score, and discuss possible interpretations of fitness in terms of energy.

In chapter 3, we investigate how the regulation of a gene depends on the binding of transcription factors to specific sites located in the regulatory region of the gene. The generation of these binding sites and cooperativity between them are essential building blocks in the evolution of complex regulatory networks. We study a theoretical model for the sequence evolution of binding sites driven by point mutations.

The approach is based on biophysical models for the binding of transcription factors to DNA. Hence we derive empirically grounded fitness landscapes, which enter a population genetics model including mutations, genetic drift, and selection.

We show that the selection for factor binding generically leads to specific correlations between nucleotide frequencies at different positions of a binding

site. We demonstrate the possibility of rapid adaptive evolution generating a new binding site for a given transcription factor by point mutations. Experimental tests of this picture involving the statistics of polymorphisms and phylogenies of sites are discussed.

In chapter 4, we give a detailed description of three different methods for TFBS detection. After studying the sequence evolution of binding sites by point mutations, we want to focus on TFBS detection on real-life sequences from human and mouse. By performing comparative sequence analysis in search for TFBS motifs, we want to predict coregulated genes within human based on promoter sequence similarity between human and mouse. For this purpose we use a well known gene in vertrebates, the TNF-α gene, playing an important role in the immune system.

In chapters 5,6,7,8, equipped with general conjectures on transcriptional regulation and fitness on one side, and practical bioinformatical tools for genomic sequence analysis on the other side, we try to understand evolutionary mechanisms in the HIV-1 promoter. Using the HIV promoter sequences, we have constructed a consistent phylogeny, detecting many substitutions but few insertions or deletions. The phylogenetic tree connects HIV-1 promoters based on their sequence similarity. More similar sequences are closer, and less similar sequences are further apart i.e. tree branches linking them are longer. Long branches of the tree are seen to correspond to the emergence of new subtypes. We have identified bioinformatically a number of TF binding sites. We have carried out a differential analysis of mutations along the branches of the tree for the sequence part containing putative binding sites and the background sequence. The distribution of mutations per base (summed over all branches) follows an approximate Poisson distribution, consistent with the concept of an approximately random evolution of the background. The regions containing binding sites are not significantly conserved. Rather, we find an enhanced rate of substitutions per base for the site regions along the inter-subtype branches of the phylogeny, consistent with positive selection for change on the binding sites. This finding is also consistent with the empirical data pointing at significant fitness differences in specific environment between constructs differing in the promoter sequence.

Chapter 2

Introduction to genomic sequence analysis

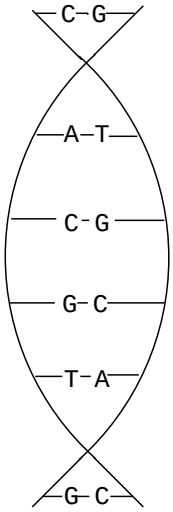

Figure 2.1: DNA double helix

2.1 Gene Expression

The responsibility of ensuring the transmission of the traits of an eukaryote individual through generations is taken up by the nucleus of the organism's cell, which contains the Deoxyribonucleic acid or DNA. [6, 94] DNA is also present in prokaryotes that do not possess the nucleus, and carries out the same task, transmission of the genetic information heritage. In some cases, like in some types of viruses, the genetic information is encoded by RNA. One way or the other, **total description of an organism's build up is written in each cell, with the four letter alphabet (A, C, G, T for DNA or A, C, G, U for RNA)**.

Precisely, the alphabet is given by bases that are grouped as Purines and Pyridines. The Purines are represented by Adenine and Guanine (A,G), while the Pyridines include Cytosine and Thymine (C,T). As, Cs, Gs and Ts are attached to sugar phosphate molecules, which are the building blocks of the DNA backbone. Each strand is a helix, wound around each other, in an opposite direction. As always pair with Ts, and Gs with Cs, forming a hydrogen bond, holding the two strands to make the double helix. The nature of bonding between the bases results in two complementary strands wound opposite to each other. **The order of the bases codes the genetic build up of an organism. It specifies its cells' function and the way of communication with the environment.**

2.2 Transcription and Translation of a Gene

In order to survive, a cell needs proteins, that play the role of molecular machines.

- Proteins are made up of amino-acids. A protein is a line of amino-acids, like a "string of beads".
- Amino-acids are lined up into a protein in "protein factory"- organelles (cellular "organs") called ribosomes. There are 20 different amino-acids.
- Since each "string" (protein) has its own combination of "beads" (amino-acids), the question arises:
  In which order should amino-acids be strung together?

Gene expression is a two-stage process, involving transcription and translation, by which proteins are produced [87, 94]. Transcription is the initial step where genetic information in the DNA is copied to a single-stranded molecule mRNA (messenger RNA). It is carried out in the following manner:

1. The DNA helix un-twists a portion of its length which contains the information for a protein that is needed. Protein regulators turn on a gene i.e. a gene gets expressed. There are different levels of "being turned on", and these are called expression levels.
2. RNA polymerase ("a sack of letters $\mathcal{A}, \mathcal{G}, \mathcal{U}, \mathcal{C}$") , **with the help of regulators**, attaches to one DNA strand.
3. For each letter on the DNA strand, RNA polymerase will donate a letter, in the antiparallel way: for $\mathcal{A}$ it will give $\mathcal{U}$, for $\mathcal{G} - \mathcal{C}$ etc. These letters from "the sack" will lign up into a strand, called m(essenger)RNA.

This new single-stranded molecule mRNA is the carrier of information which is required in ribosomes. Once it is finished, it will leave the nucleus and make its way into the "protein factory".

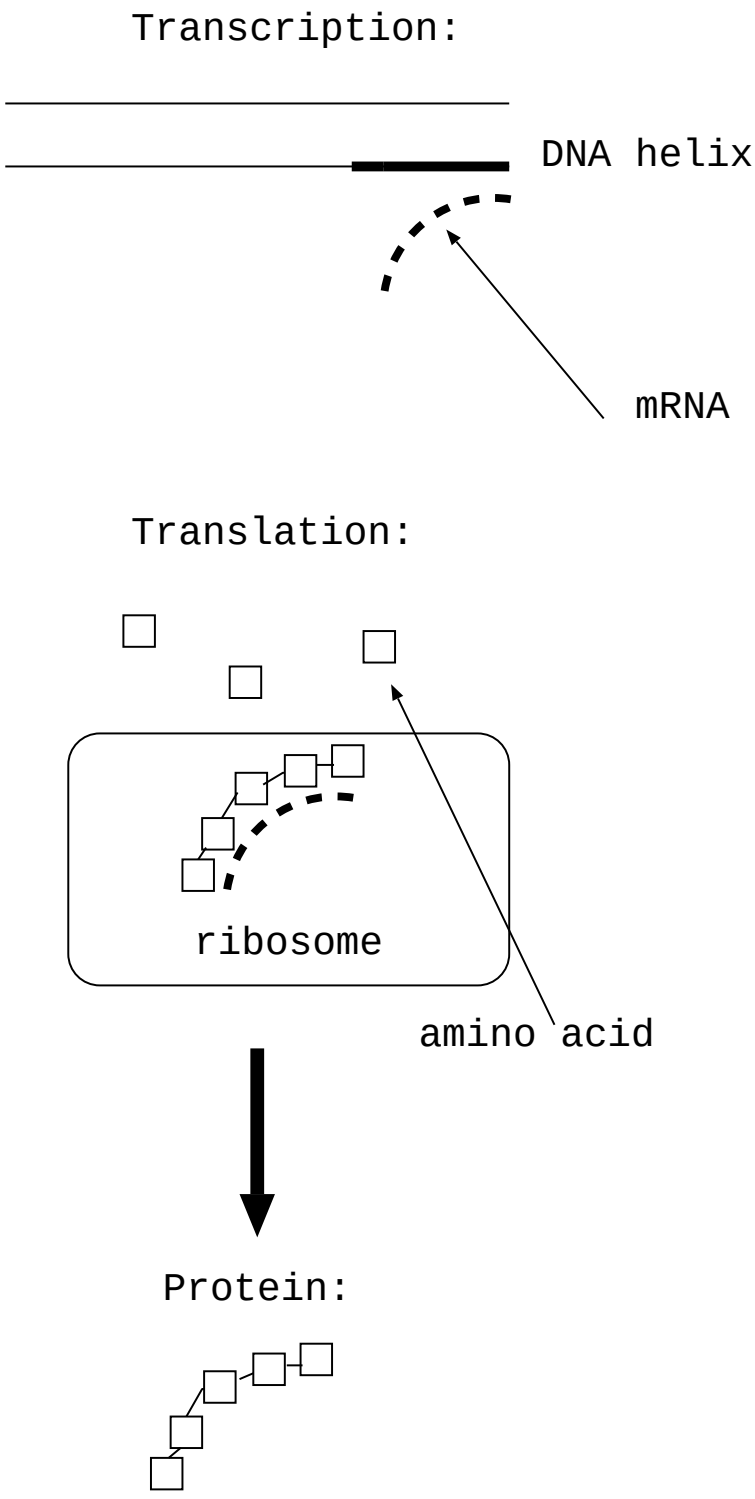

Figure 2.2: Transcription and Translation

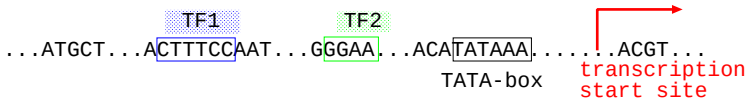

Figure 2.3: Transcription Initiation

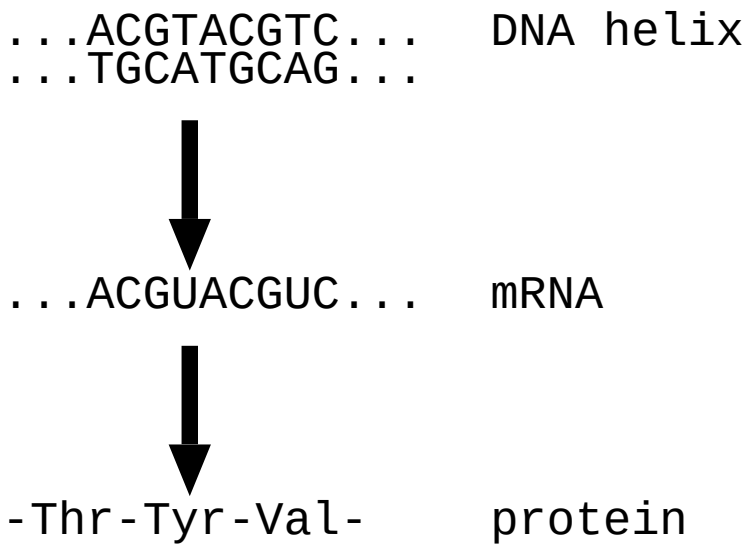

Figure 2.4: Transcribed mRNA goes to ribosomes

Note the second step: **'with the help of regulators '**. Transcriptional regulation takes place in the sequence, which is located upstream (in front of) of the protein coding DNA and is called promoter region. The promoter region contains several short subsequences (also known as motifs) that bind regulatory proteins called transcription factors (TFs) . Depending on which motifs (or transcription factor binding sites TFBSs) are occupied (see Fig. 2.3) the gene can be turned on i.e. expressed.

2.3 The Evolution of Transcriptional Regulation

Understanding transcriptional regulation is one of the most challenging tasks of genome analysis. It comprises different fields of science, resulting in an exciting interplay of biophysics, bioinformatics, evolutionary and molecular biology.

The process of gene transcription is controlled by a complex machinery of TFs. A cell must respond to extra and intra-cellular signals correctly, in order to function properly. These responses are in the form of protein levels produced, which are under tight control of transcriptional regulation. Recent reviews [59] have argued that transcriptional regulation plays an important role in evolution. Empirical evidence show that the complexity of different organisms relies heavily on regulatory reorganization and development. Although number of genes in higher organisms also increases, quantitative difference can not account for immense difference in complexity. Fruit fly is estimated of having about 13000 genes, nematode (a worm) has about 18000 genes, mouse about 30000 and human is said to have about 35000 genes, according to the Human Genome Project. Still rice *indica* [96] has more genes than humans. Also, if we compare mouse and human [97–99], a comparable number of genes cannot explain the differences between these two species.

It turned out that **changes in gene expression play a crucial role in genotype-phenotype relationship** in all organisms. Many regulatory genes are common for different species, which raises a question: where do distinct features come from? Since studies have shown that there are strong correlations between gene expression and anatomy, it seems that reorganization of complex TF interactions leads to formation of different organisms (since TFs themselves are proteins, thus being encoded by genes, we can regard this TF interplay as a **genetic network**). For instance, domestication of maize is partially due to changes in the promoter region of a gene encoding a protein called teosinte-branched [49]. Another example is the HIV-1 virus: one of its subtypes' increased aggressiveness is due to changes in the promoter [56, 57] i.e. gain in an additional binding site motif for the TF called NFκB.

Loss or gain of TFBSs happens due to mutations. Mutations can affect a single base, in which case they are called point mutations, or sequence segments, known as segmental mutations. We are concerned with point mutations, since they seem to be the major cause of promoter sequence alteration. **Point mutations** can occur in different ways: as substitutions (replacement of one base by another), as deletions (deletion of one or more bases) and as insertions (insertion of one or more bases) [82].

Mutations may or may not affect the organism's phenotype, which in turn may influence the individual's ability to survive or reproduce. This capability of survival and reproduction is measured by the **fitness** of a genotype. Now, fitness may be improved by a mutation (like a TFBS gain), and these are

advantageous mutations. On the other hand, fitness may be reduced due to a mutation (like a TFBS loss), and such mutations are denoted as deleterious.

A basic mechanism of the **evolution of DNA sequences** is comprised of point mutations during evolutionary time [82, 87].

2.4 Information Integration

Not all of the genes in a eukaryotic cell are expressed at any given time point. Eukaryotic genomes comprise $10^3 - 10^4$ genes, and only some genes are expressed at certain time point, therefore controlling this differential gene expression demands an extraordinary complex set of specific interactions among transcription factors. It includes processes like transcriptional initiation, mRNA and protein stability, intracellular trafficking etc. *For every eukaryotic gene encoding relevant information, transcriptional intitation appears to be one of the primary determinant, if not the only one, of the overall gene expression profile.*
All proteins that regulate transcription directly or indirectly influence the frequency with which the RNA polymerase complex assembles onto the basal promoter. Genes encoding transcription factors possess some of the most complex expression profiles, while those of constitutive ones are much simpler [59].

At its most fundamental level, the function of a promoter is to integrate information about the state of the cell, and to alter the rate of transcriptional initiation of a single gene according to the cell's needs. The inputs that a promoter integrates are diverse, eventually reaching the promoter in the form of TFs, that bind certain sequence motifs (TFBSs) of the DNA strand almost always in front of a gene, altering rates of transcriptional initiation.

2.5 Thermodynamics of Transcription Factor Binding

A TFBS is 10-15 bases long in procaryotes, and l=5-8 bases long in eukaryotes ($\vec{a} = (a_1, a_2, ...a_l)$) . Experiments have revealed that TF-binding to TFBS is specific, thus it can be quantified via specific binding energies [14], [20], [17], [103]:

$$E_s(\vec{a}) = \sum_{i=1}^{l} \epsilon_{a_i,i}, \tag{2.1}$$

The energy matrix has been determined from single-base substitution experiments on TFs:

	1	2	.	.	.	l
A	$\epsilon_{A,1}$	$\epsilon_{A,2}$	.	.	.	$\epsilon_{A,l}$
C	$\epsilon_{C,1}$	$\epsilon_{C,2}$	.	.	.	$\epsilon_{C,l}$
G	$\epsilon_{G,1}$	$\epsilon_{G,2}$	.	.	.	$\epsilon_{G,l}$
T	$\epsilon_{T,1}$	$\epsilon_{T,2}$	.	.	.	$\epsilon_{T,l}$

A few points could be generalized:

1. Positions i=1,2,...l contribute independently to E_s.
2. There is typically one preferred base $a^*{}_i$ at each position i.
3. Mismatch energies $\epsilon_{a_i,i} - \epsilon_{a^*{}_i,i}$ are typically in the range 1-3 k_BT.
4. The energy difference between optimal specific and unspecific binding is approx. 15 k_BT.

The authors of ref. [19] introduced a two-state approximation for individual base energy contributions to the specific energy:

$$\begin{aligned} \epsilon_{a_i,i} - \epsilon_{a^*{}_i,i} &= \epsilon, a_i, i \neq a^*{}_i, i \\ \epsilon_{a_i,i} - \epsilon_{a^*{}_i,i} &= 0, a_i, i = a^*{}_i, i \end{aligned} \tag{2.2}$$

which leads to

$$E(\vec{a}) = E^* + d\epsilon \tag{2.3}$$

where d stands for the number of mismatches between $\vec{a}$ and $\vec{a}*$, so called Hamming distance. If the optimal binding site looks like $TTTTCC$, motif 1 $CTTTCC$ ($d = 1$) will have lower energy than the motif 2 $CTAACC$ ($d = 3$), thus binding the TF much better. The model is known in physics as the Potts model.
Given the energies, we would like to determine the corresponding probabilities. For that purpose, we write the likelihood of a state α

$$L_\alpha = \exp\left(\frac{-E_\alpha}{k_BT}\right), \tag{2.4}$$

the partition function

$$Z = \sum_\alpha \exp\left(\frac{-E_\alpha}{k_BT}\right), \tag{2.5}$$

and arrive at the probability of state α:

$$p_\alpha = \frac{\frac{-E_\alpha}{k_BT}}{Z}. \tag{2.6}$$

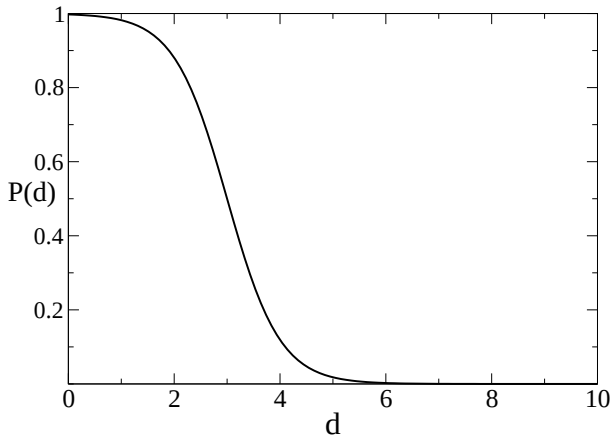

Figure 2.5: Probability for specific binding

It turns out that the probability for a specific binding is a Fermi function of the (specific binding) energy (see figure 2.5).

The Fermi step appears for a threshold energy, which depends on the rest of the sequence, whether it is random or it offers many motifs of the TFBS-type under investigation. Therefore, an efficiency criterion must be introduced, that the TF will bind an optimal binding site with a finite probability. Thanks to genome sequencing, we know that the length N (number of bases) of the viral genome is of the order 10^3, that of bacteria 10^6 and of humans 10^9. Using the efficiency criterion, we can estimate the energy shift between optimal and unspecific binding (E_0) [103]:

$$\begin{aligned} \exp\left(\frac{-\beta E^*}{k_B T}\right) &\geq N \exp\left(\frac{-\beta E_0}{k_B T}\right) \\ \frac{E_0 - E^*}{k_B T} &\geq \log N \end{aligned} \tag{2.7}$$

$$\begin{aligned} \exp\left(\frac{-\beta E^*}{k_B T}\right) &\geq l \exp\left(\frac{-\beta (E^* + \epsilon)}{k_B T}\right) \\ \frac{\epsilon}{k_B T} &\geq \log l \end{aligned} \tag{2.8}$$

The first condition refers to the background (rest of the DNA sequence) and estimates the minimal length of a TFBS. The second condition refers to mutations of an optimally binding TFBS. So far, we focused on thermodynamics of a single TFBS. Yet, promoters contain several binding sites, and a collection of these sites is called a module.

A module is defined as a cluster of binding sites that influences the total transcription profile. A single module typically contains about 6 to 15 bases and binds 4 to 8 different TFs.

Deletion of a single module eliminates a specific aspect of the expression profile without disrupting the remainder. Also, predictable artificial expression profiles can be obtained by experimentally combining modules from different promoters [15], [59].

Thermodynamics of several TFBSs must be taken into account if we want to describe real-life transcription initiation.

In a cell, there are $n = 10^3$ of different TF floating around. If we observe specific binding probability as a function of Hamming distance d, we see that it has a sigmoid form [17], [103], with a threshold ρ which divides binding sites into two groups, functional and disfunctional.

The threshold value follows from the condition:

$$n \exp \frac{\rho\epsilon}{k_B T} \approx 1 \tag{2.9}$$

yielding:

$$\rho = \frac{k_B T}{\epsilon} \log n \tag{2.10}$$

Knowing that the assembly of TFs bound to their corresponding TFBS results in transcription, and therefore in enhanced levels of gene expression, we can try to estimate expression levels as a function of the logarithm of the number of TFs n (see figure 2.6):

$$\log n_{threshold} = \frac{\rho\epsilon}{k_B T} \tag{2.11}$$

Even within experimentally well-studied promoters, we should assume that some binding sites remain uncharacterized. Yet, a few generalizations can be made, such as TFBS typically comprise a minority of the bases within a promoter region. Bases that do not fall into TF binding motifs are generally assumed to be nonfunctional with respect to transcription and are denoted as background.

Most binding sites can tolerate at least one, and often more, specific base substitutions and still bind the same TFs. All sequences that are reported to bind a particular TF **with much higher specificity than random DNA** are often described by a **position weight matrix (PWM)**.

Sequence comparisons (or the average sequence of multiple binding sites of the same TF), yield a so called consensus sequence that captures most of the weight matrix. PWMs are constructed from count matrices, that simply

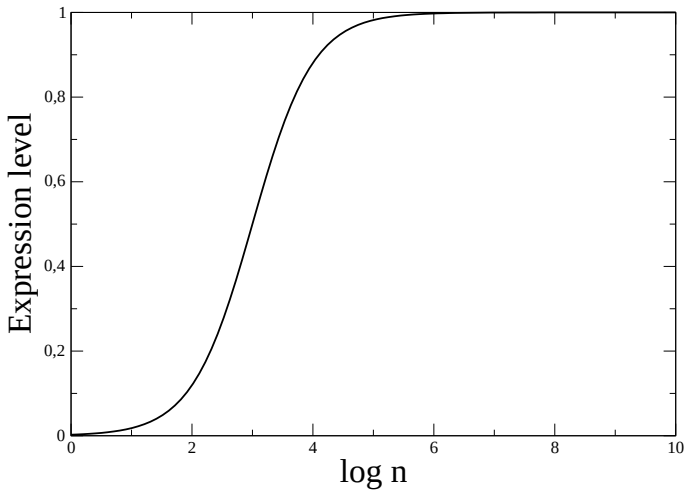

Figure 2.6: Expression level

count how many bases were detected at each position of a TFBS for all trials.

	A	C	G	T
1	40	0	0	9
2	35	2	2	1
.	.	.	.	.
l	0	10	25	5

Here is an example where at the first position, out of 40 experiments, an A was detected all 40 times, at the second position an A was found 35 times, a C twice, etc.
Matrices of this sort can be found in the TRANSFAC Database at http://www.gene-regulation.com/ and they are starting point for bioinformatical detection of binding sites on DNA strand.

In order to detect a TFBS signal, we need a Null-model for comparison. Therefore, a Markov model for background sequences is introduced:

$$P_0(a_1, a_2, ...a_N) = \prod_k p_0(a_k), \quad a_k = A, C, G, T, \tag{2.12}$$

where $p_0(a_k)$ is background frequency of letter a at k^{th} position.

On the other hand, the Markov model for a TFBS is of the form

$$Q(a_1, a_2, ...a_l) = \prod_k q(a_k), \quad a_k = A, C, G, T, \tag{2.13}$$

where $q(a_k)$ is frequency of letter a at position k, obtained from count matrices for a particular TF .
Of course, the probability model for a TFBS is different from the background distribution. The difference between P_0 and Q is of exponential form:

$$Q(a) = P_0(a) \exp(S(a)), \tag{2.14}$$

where $S(a)$ is score matrix $s_{a_k,k}$ which looks very much like an energy. The score of a motif is given by the sum of individual base scores [58], [103], [104]:

$$S(\vec{a}) = \sum_{k=1}^{l} s(a_k, k). \tag{2.15}$$

The question arises what is the correct interpretation of the score $S(\vec{a})$. Is it related to the fitness under a population genetics model? Although the score resembles an energy, is its connection to the concept of energy straightforward? We will try to answer some of these questions in the following chapters.

There is no general framework for promoter evolution, due to the lack of a reading frame, the low density of functionally important bases, and the ability of many binding sites to operate in a position-independent manner. It seems that the background located between binding sites should be free to vary.

Assumed that the background mutates with neutral evolutionary rates, two bases (say a and b) are replaced by each other with rates that are linked to the stationary distribution (Fig 2.7):

$$p_a \mu_{a \to b} = p_b \mu_{b \to a} \tag{2.16}$$

Or in other words, a detailed balance condition is fulfilled in nature [103]. The rates $\{\mu\}$ and equilibrium distributions $\{p\}$ are very well known. For example, it is a well known fact that different parts of the genome mutate according to different rates. Basically, neutral evolution is a random process, a process evolving without any selection.

As opposed to the background, TFBS are expected to mutate under selection

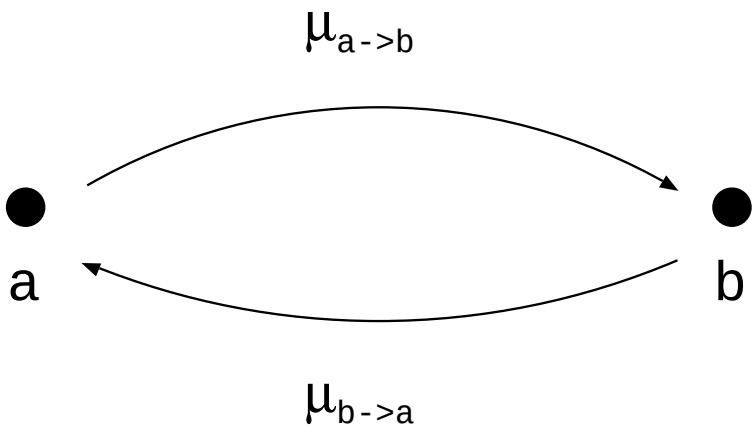

Figure 2.7: Detailed balance rates

pressure in order to preserve their functionality. Fitness F is defined as the growth rate of a subpopulation M_i compared to the rest of the population

$$FM_i = \frac{dM_i}{dt} \tag{2.17}$$

therefore it is proportional to the population size M, $F \sim \log M$, and it is a function of a state $\vec{a}$, $F(\vec{a})$, where $\vec{a}$ describes a sequence of genomic bases (A, C, G, T). According to Kimura-Ohta theory [41], a state $\vec{a}$ will transform into a state $\vec{b}$ and vice versa, with a substitution rate $u_{\vec{a}\to\vec{b}}$, obeying detailed balance.

$$u_{\vec{a}\to\vec{b}} = \mu_{\vec{a}\to\vec{b}} M \frac{1 - \exp\left[-4(F(\vec{b}) - F(\vec{a}))\right]}{1 - \exp\left[-4M(F(\vec{b}) - F(\vec{a}))\right]} \tag{2.18}$$

The formula has a clear interpretation: the mutation rate of an event that results in a fitness advantage will be higher than μ. The enhancement factor $\frac{u}{\mu}$ is 1 if there is no fitness difference $\Delta F = 0$. There is an obvious scaling variable contained in the above formula: $4M\Delta F$. Fitness difference between the final and initial state is scaled. Now, the stationary distribution can be computed from the equilibrium condition:

$$Q(\vec{a})u_{\vec{a}\to\vec{b}} = Q(\vec{b})u_{\vec{b}\to\vec{a}}, \tag{2.19}$$

yielding deviation from the background distribution:

$$Q(\vec{a}) = p_0(\vec{a}) \exp(-4MF(\vec{a})) \tag{2.20}$$

So, how is bioinformatical score related to fitness? Since fitness depends on the quality of genetic material, we can make an assumption that score $S(\vec{a})$ is directly related to the fitness advantage [103], [104], recalling eq. 2.14.

$$S(\vec{a}) = 4MF(\vec{a}) \tag{2.21}$$

On the other hand, fitness can be regarded in evolutionary theory as energy in mechanics [38]. Fitness is defined as a nonnegative quantity, furthermore up to an additive constant, therefore we can construct a **fitness landscape**, in which evolutionary processes take place, very much like some energy landscape. The space where our sequences reside has its own metrics, called Hamming distance. The difference would be that sequences tend to climb up the fitness hills, while physical bodies tend to roll down into the energy landscape valleys.

Understanding the dynamic link between genotype and phenotype is still a central challenge in evolutionary biology. Phenotypic evolution is governed heavily by transcriptional regulation [16]. It seems that promoters are more 'evolvable' than coding regions. As it can be seen in TNF-α promoter, there is evidence that natural selection acts on regulatory sequences from cases of obvious evolutionary conservation among distantly related species. **The naive approach would be that new structures require formation of new genes, while the more sophisticated view is that new structures are built by reorganizing the interactions among existing genes.** Although it is believed that mutations within promoter regions constitute the most 'relevant' source of genetic variation, the fraction of TFBS changes versus background within regulatory sequences, is not known, even to a rough approximation. Estimating this ratio represents an important challenge in molecular evolution.

Chapter 3

On the evolution of gene regulation

The results of this part were obtained in collaboration with Johannes Berg and published in ref. [103].

3.1 Introduction to Quasispecies

In the early 1950's Miller conducted experiments [90] showing that organic substances, such as amino acids and nucleotides (DNA letters), could emerge spontaneously from the nitrogen-hydrogen-carbon soup when electric discharges were pumped in. These conditions are thought to represent a realistic picture of the atmosphere on Earth just before life occurred.
Newly formed short molecular chains are believed to be capable of polymerizing, giving RNA and DNA strands. It is assumed that these polymers started evolving under two 'forces', called Darwinian selection and random mutation. An attempt to describe a process of this kind mathematically was introduced by [89]. The idea is that at the beginning there is a pool of n sequences of length L, which we denote as the set $\{\sigma^k = (\sigma^k{}_1, ..., \sigma^k{}_L)\}$ ([91]). Assuming that $n,\ L \gg 1$ and that the sequences reproduce at a rate $F(\sigma^k)$. The reproduction rate $F(\sigma^k)$ is maximal for the 'best' sequence, so called, master sequence σ^m. The closer some sequence σ^k (in terms of letter-structure) to the master sequence σ^m, the higher its reproduction rate, also called fitness. Thus, we can say that $F(\sigma^k)$ depends on σ^k's Hamming distance relative to σ^m. The pool of $\{\sigma^k\}$ is described by a time-dependent frequency distribution $P(\sigma^k)$. Its evolution is described by the Schroedinger equation for the population state

$$\partial_t P = HP \tag{3.1}$$

where $F(\sigma^k)$ is a scalar potential. Random mutations are captured by the kinetic term in the Hamiltonian H ([91]).
The quasispecies theory [89] is widely applicable to a population of sequences that undergo mutations and frequency-dependent selection ([92]). In the case of a four-letter alphabet $\{A, C, G, T(U)\}$, the letter strings are called genotypes, playing the star role in the dynamics of a viral population.

3.2 Adaptive evolution of transcription factor binding sites

The expression of a gene is controlled by other genes expressed at the same time and by external signals, a process called *gene regulation* [51]. Due to the combinatorial complexity of regulation, a large number of functional tasks

can be performed by a limited number of genes. Differences in gene regulation are believed to be a major source of diversity in higher eukaryotes.

To a large extent, gene regulation is the control of transcription. It is accomplished by a number of regulatory proteins called *transcription factors* that bind to specific sites on DNA. These binding sites contain about $10-15$ base pairs relevant for binding and are mostly located in the cis-regulatory promoter region of a gene. A cis-regulatory region in *E. coli* is about 300 base pairs long and contains a few transcription factor binding sites [11]. There may be two or more sites for the same factor in one promoter region. At the same time, the sequences of binding sites are *fuzzy*, that is, different sites for the same factor differ by about $20-30$ percent of the bases relevant for binding [11]. This makes the identification of sites a difficult bioinformatics problem [12–14]. Frequently, the simultaneous binding at two nearby sites is energetically favoured. This so-called *binding cooperativity* can be related to various functions. In a *genetic switch* such as the famous page lambda switch in *Escherichia coli* [52], it produces a sharp increase of the expression level at a certain threshold concentration of a transcription factor. A pair of sites for two different kinds of factors with cooperative binding can be a simple module for *signal integration*, leading to the expression of the downstream gene only when both kinds of factors are present simultaneously [51]. These examples are discussed in more detail below. Regulation in higher eukaryotes shares these features but is vastly more complicated [33]. A promoter region is typically a few thousand base pairs long and contains many different binding sites with often complex interactions. At the same time, individual sites are shorter, with about 5-8 relevant base pairs. The sites are sometimes organized in *modules* interspersed between regions containing no sites. In many known cases, the expression of a gene depends on the simultaneous presence of several factors. Well-studied examples of regulatory networks in eukaryotes include the sea urchin *Strongylocentrotus purpuratussea* [15] and the early developmental genes in *Drosophila* [16].

The sequence statistics of binding sites has been addressed in two recent theoretical studies [17,18]. Based on an empirical model of sequence-factor interaction [19,20], a *fitness landscape* for binding site sequences is constructed (see the discussion in the next section). The resulting mutation-selection equilibrium is analysed using a mean-field *quasispecies* approach [21]. This approach, which neglects the effects of genetic drift, is applicable in very large populations. In both studies [17, 18], fuzziness is attributed to *mutational entropy* as a possible reason: the single or few sequence states with optimal binding of the transcription factor can be outweighed by the vastly higher number of sub-optimal states at some mutational distance from the optimal binding sequence. This effect is similar to the fuzziness of amino

acid sequences in proteins discussed in [22].

From an evolutionary perspective, explaining the molecular programming of regulatory networks presents a striking problem. The diversification of higher eukaryotes, in particular, requires the efficient generation and alteration of regulatory binding interactions. One likely mode of evolution is gene duplications with subsequent complementary *losses of function* in both copies [23, 24]. However, the differentiation of regulation should also require complementary processes that generate *new functions* of genes as a response to specific demands. This task must be accomplished mainly by sequence evolution of regulatory DNA. There are examples of highly conserved regulatory sequences with a conserved function but binding sites can also appear, disappear, or alter their sequence even between relatively closely related species; see, e.g., refs. [25–29]. This turnover of binding sites has been argued to follow an approximate molecular clock in *Drosophila* [30]. The transcription factors themselves are known to remain more conserved, especially if they are involved in the regulation of more than one gene.

The modes of regulatory sequence evolution and their relative importance remain largely to be explored. Contributions may arise from point mutations, slippage processes [31], and larger rearrangements of promoter regions [32]. The latter processes may lead to the shuffling of entire modules of binding sites between different genes. In this chapter, we are more interested in the local sequence evolution within a module, which has been argued to contribute most of the promoter sequence difference between species [49]. It is also the most promising starting point for a *quantitative* analysis of binding site evolution. We study a theoretical model that takes into account point mutations, selection, and genetic drift. The form of selection is inferred from the biophysics of the binding interactions between transcription factors and DNA.

We derive the stationary distribution of binding sites under selection, which shows specific correlations between nucleotide frequencies at different positions in a binding site. The non-stationary solutions of the model describe efficient adaptive pathways for the molecular evolution of regulatory networks by point mutations. This efficiency can be quantified in terms of the length of the binding motif, and the length of the promoter region, and the fitness landscape for factor binding, which is amenable to quite explicit modeling.

With the parameters found in natural systems, our model predicts that a new binding site for a given transcription factor can be generated by a fast series of adaptive substitutions, even if the expression of the corresponding gene bears even a modest fitness advantage. The evolutionary time required for site formation in response to a *newly arising* selection pressure is esti-

mated in terms of the characteristic time scales of mutations, selection, and drift. For *Drosophila*, it may be as short as 10^5 years even for moderate selection pressures. However, this pathway is found to depend crucially on the presence of selection. It would be too slow under neutral evolution, in contrast to the results of [33], see also the recent discussion in [8]. Cooperative interactions between binding sites can evolve adaptively on similar time scales, as we show for the two simple examples alluded to above, the genetic switch and the signal integration module. These results are discussed at the end of the chapter with particular emphasis on possible experimental tests.

3.3 Factor Binding and Selection

The binding energy (measured in units of k_BT) between a transcription factor and its binding site is, to a good approximation, the sum of independent contributions from a small number of important positions of the binding site sequence, $E/k_BT = \sum_{i=1}^{\ell} \epsilon_i$, with $\ell \approx 10 - 15$ [34, 35, 37]. The individual contributions ϵ_i depend on the position i and on the nucleotide a_i at that position. There is typically one particular nucleotide a_i^* preferred for binding; the sequence $(a_1^*, \ldots, a_\ell^*)$ is called the *target sequence*. The target sequence can be inferred as the consensus sequence of a sufficiently large number of equivalent sites. The so-called *energy matrix* $\epsilon_i(a)$ has been determined experimentally for some factors from *in vitro* measurements of the binding affinity for each single-nucleotide mutant of the target sequence. Typical values for the loss in binding energy are 1-3 k_BT per single-nucleotide mismatch away from the target sequence. In this chapter, we use the further approximation $\epsilon_i = \epsilon$ if $a_i = a_i^*$ and $\epsilon = 0$ otherwise, the so-called *two-state model* [19]. The binding energy of any sequence $(a_1, \ldots, a_\ell)$ is then, up to an irrelevant constant, simply given by its Hamming distance r to the target sequence: $E/k_BT = \epsilon r$. (The Hamming distance is defined as the number of positions with a mismatch $a_i \neq a_i^*$.)

It is important to note the status of this "minimal model" of binding energies for the discussion in this chapter. Both approximations underlying the model can be violated. Even though typical mismatch energies are of the same order of magnitude, there can be considerable differences between different substitutions at one position and between different nucleotide positions. Moreover, deviations from the approximate additivity of binding energies for the single nucleotide positions have also been observed. However, these complications do not affect the order-of-magnitude estimates for adaptive sequence evolution. As it will become clear, the efficiency of binding site formation depends only on the qualitative shape of the fitness landscapes de-

rived below. In these landscapes, the regime of weakly-binding sequences and of strongly-binding sequences are separated by only a few single nucleotide substitutions. The relative magnitude of the fitness increase of these substitutions does not matter in first approximation. Indeed, inhomogeneities in the values of the $\epsilon_i(a)$ tend to reduce the number of *crucial* steps in the adaptive process and thereby to further increase its speed.

Within the two-state model, the binding probability of the factor in thermodynamic equilibrium is

$$p = \frac{1}{1 + \exp[\epsilon(r - \rho)]}. \tag{3.2}$$

Here ϵ is the binding energy per nucleotide mismatch and $\epsilon\rho$ is the chemical potential measuring the factor concentration. Both parameters are expressed in units of $k_B T$ and hence dimensionless. Appropriate values for typical binding sites have been discussed extensively in refs. [17, 20]. It is found that ϵ should take values around 2, which is consistent with the measurements for known transcription factors mentioned above [34, 35, 37]. The chemical potential depends on the number of transcription factors present in the cell, on the binding probability to *background* sites elsewhere in the genome (which have a sequence similar to the target sequence by chance), and on the *functional* sites in the in the genome other than the binding site in question that may compete for the same protein. Binding to background sites does not significantly reduce the binding to a specific functional site [20]. This leads to values $\rho \approx (\log n_f)/\epsilon \approx 2 - 4$, given observed factor numbers n_f of about $50 - 5000$ [20]. Binding to other copies of the same functional sequence becomes only relevant at low factor concentrations and high number of copies, when sites compete for factors.

A *fitness landscape* quantifies the fitness $F(a_1, \ldots, a_\ell)$ of each sequence state at the binding site. Fitness differences arise due to different expression levels of the regulated gene, and these in turn depend on the binding of the transcription factors. Following the conceptual framework of ref. [17], we assume that the environment of the regulated gene can be described by a number of *cellular states* (labelled by the index α) with different transcription factor concentrations, i.e., with different chemical potentials ρ^α. These cellular states can be thought of as different stages within a cell cycle. In each state, the fitness depends on the expression level of the regulated gene in a specific way. This expression level is determined by the binding probability p^α of the transcription factor. Assuming that both dependencies are linear (this is not crucial) and that the cellular states contribute additively to the

overall fitness F, we obtain

$$F = \sum_{\alpha} s^{\alpha} p^{\alpha}. \tag{3.3}$$

Here the *selection coefficient* s^{α} is defined as the fitness difference (due to different expression of the downstream gene) between the cases of complete factor binding and no binding in the state α. Such fitness differences can now be measured directly in viral systems [57]. Inserting (3.2), the fitness becomes a function of the Hamming distance r only.

In a simple case, there are just two relevant cellular states. The *on* state favours expression of the gene, the *off* state disfavours it. It is then natural to assume selection coefficients of similar magnitude; here we take for simplicity $s = s^{\mathrm{on}} = -s^{\mathrm{off}} > 0$. We then obtain a *crater* landscape,

$$F(r) = \frac{s}{1 + \exp[\epsilon(r - \rho^{\mathrm{on}})]} - \frac{s}{1 + \exp[\epsilon(r - \rho^{\mathrm{off}})]}, \tag{3.4}$$

with a high-fitness rim between ρ^{off} and ρ^{on} flanked by two sigmoid thresholds; see fig. 3.7 (a). The generic features of this fitness landscape are easy to interpret: the two-state selection assumed here favors intermediate binding strength (i.e., intermediate Hamming distances r) where binding occurs and the gene is expressed in the *on state* but not in the *off* state. Sequences with large Hamming distance $r > \rho_{\mathrm{on}}$ can bind the factor neither in the *on* nor in the *off* state, while sequences with $r < \rho_{\mathrm{off}}$ lead to binding in the *on* and the *off* state. Both cases lead to misregulation of the downstream gene, and hence to a lower fitness.

An even simpler fitness landscape is obtained if only the *on* state contributes significantly to selection, i.e., if $s = s^{\mathrm{on}} > 0$ and $s^{\mathrm{off}} = 0$. The crater landscape then reduces to the *mesa* landscape discussed in [17, 39],

$$F(r) = \frac{s}{1 + \exp[\epsilon(r - \rho^{\mathrm{on}})]}, \tag{3.5}$$

which has a high-fitness plateau of radius ρ and one sigmoid threshold; see fig. 3.7 (b). In this case, all sequences with sufficiently small Hamming distance to the target sequence ($r < \rho^{\mathrm{on}}$) have a high fitness.

In both cases, the parameters of the binding model have a simple geometric interpretation: ϵ gives the slope and the ρ^{α} give the positions of the sigmoid thresholds in the fitness landscape. Eqs. (3.4) and (3.5) are again to be understood as minimal models of fitness landscapes for binding sites, representing target sequence selection for a given level of binding ($\rho^{\mathrm{off}} < r < \rho^{\mathrm{on}}$) and for sufficiently strong binding ($r < \rho^{\mathrm{on}}$), respectively. Despite its simplicity, this type of selection model based on biophysical binding affinities is

nontrivial from a population-genetic viewpoint since it leads to generic correlations between frequencies of nucleotides a_i and a_j within a site, see the Results section below. We will also study generalized models with correlations between two sites generated by cooperative binding. On the other hand, these models neglect the context dependence of the binding process through cofactors and chromatin structure. However, they are a good starting point for order-of magnitude estimates of the adaptive evolution of binding sites.

3.4 Mutations, Selections, and Genetic Drift

The rates of nucleotide point mutation show a great variation, ranging from $\mu \sim 10^{-4}$ per site and generation for RNA viruses to values several orders of magnitude lower in eukaryotes, e.g., $\mu \approx 2\times 10^{-9}$ in *Drosophila* [40]. (Here we model mutation as a single-parameter Markov process; we do not distinguish between transitions and transversions.) The evolution of a sufficiently large population under mutation and selection can be described in terms of the average fraction of the population with a given binding sequence. This so-called mean-field approach neglects the fluctuations due to finite population size (genetic drift). It leads to the so-called *quasispecies* theory [21]. For a population of sequences at a single binding site, the quasispecies population equation can be written for the fraction $n(r,t)$ of individuals at Hamming distance r from the target sequence at time t. Along with a generalisation for two binding sites, it has been analysed in detail in ref. [17]. For the mesa landscape, the stationary solution $n_{\rm stat}(r)$ has been found exactly [39]. It depends only on the ratio s/μ and describes a stable *polymorphic* population, i.e., several sequence states coexist. The mean-field approach is valid as long as the stochastic reproductive fluctuations are leveled out by mutations. This requires absolute population numbers $Nn_{\rm stat}(r) \gg 1/\mu$ for all relevant r, a stringent condition on the total population size N.

This chapter is concerned with a different regime of population dynamics, as described by the Kimura-Ohta theory for finite populations evolving by stochastic fluctuations (genetic drift) and selection [9, 36, 41]. According to this theory, a new mutant with a fitness difference ΔF relative to the pre-existing allele could spread to fixation in the population. This is a stochastic process, whose rate constant is given by

$$u = \mu N \frac{1-\exp(-2\Delta F)}{1-\exp(-2N\Delta F)} \tag{3.6}$$

in a diffusion approximation valid for $\Delta F \ll 1$ [42]. Here N is the *effective* population size (with an additional factor 2 for diploid populations).

Eq. (3.6) has three well-known regimes. For substantially *deleterious* mutations ($N\Delta F \lesssim -1$), substitutions are exponentially suppressed. *Nearly neutral* substitutions ($N|\Delta F| \ll 1$) occur at a rate $u \approx \mu$ approximately equal to the rate of mutations in an individual. For substantially *beneficial* mutations ($N\Delta F \gtrsim 1$), the substitution rate is enhanced, with $u \simeq 2\mu N\Delta F$ for $N\Delta F \gg 1$.

In this picture, a population has a monomorphic majority for most of the time and occasional coexistence of two sequence states while a substitution is going on. The time of coexistence is $T \sim N$ for nearly neutral and $T \sim 1/\Delta F$ for strongly beneficial substitutions. The picture is thus self-consistent for $Tu \ll 1$, i.e., for $\mu N \ll 1$. Asymptotically, it describes monomorphic populations moving through sequence space with hopping rates u.

Introducing an *ensemble* of independent populations, this stochastic evolution takes the form of a Master equation. For a single binding site, we obtain

$$\begin{aligned}\tfrac{\partial}{\partial t}P(r,t) = \\ &(c-1)(\ell-r+1)\,u_{r-1,r}P(r-1,t) + \\ &(r+1)\,u_{r+1,r}P(r+1,t) - \\ &[r\,u_{r,r-1} + (c-1)(\ell-r)\,u_{r,r+1}]P(r,t).\end{aligned} \tag{3.7}$$

Here $P(r,t)$ denotes the probability of finding a population at Hamming distance r from the target sequence, and $u_{r,r'}$ is given by (3.6) with $\Delta F = F(r') - F(r)$. The combinatorial coefficients arise since a sequence at Hamming distance r can mutate in $(c-1)(\ell-r)$ different ways that increase r, and in r ways that decrease r, where $c = 4$ is the number of different nucleotides. The stationary distribution is

$$P_{\rm stat}(r) \sim \exp[S(r) + 2NF(r)]. \tag{3.8}$$

Here $S(r) = \log[\binom{\ell}{r}(c-1)^r/c^\ell]$ is the *mutational entropy* (the log fraction of sequence states with Hamming distance r) [39] and we have used the exact result $u_{r+1,r}/u_{r,r+1} = e^{2(N-1)\Delta F}$. To derive (3.8), we then simply approximated $N-1$ by N. The form of $P_{\rm stat}(r)$ reflects the selection pressure, i.e., the scale s of fitness differences in the landscape $F(r)$. For neutral evolution ($2sN = 0$), the stationary distribution

$$P^0_{\rm stat}(r) \sim \sum \exp[S(r)] \tag{3.9}$$

is obtained from a flat distribution over all sequence states. For moderate selection ($2sN \sim 1$), $P_{\rm stat}(r)$ results from a nontrivial balance of stochasticity

and selection. For strong selection ($2sN \gg 1$), $P_{\rm stat}(r)$ takes appreciable values only at points of near-maximal fitness, where $F(r) \gtrsim F_{\rm max} - 1/2sN$. In this regime, the dynamics of a population consists of beneficial mutations only, i.e., the system moves uphill on its fitness landscape.

The Master equation (3.7) and the mean-field quasispecies equation thus describe opposite asymptotic regimes, $\mu N \ll 1$ and $\mu N \gg 1$, of the evolutionary dynamics. Effective population sizes show a large variation, from values of order 10^9 in viral systems to $N \sim 10^6$ in *Drosophila* and $N \sim 10^4 - 10^5$ in vertebrates. (These numbers bear some uncertainty; one reason is that N varies across the genome [43].) We conclude that the mean-field quasispecies is well suited for viral systems, while eukaryotes clearly show a stochastic dynamics of substitutions.

3.5 Results and Discussion

3.5.1 Stationary distributions and nucleotide frequency correlations

In the previous sections, we have expressed the fitness landscape and the resulting population distributions as a function of the Hamming distance r because it is a convenient parameterization of the binding energy in the two-state model. In order to compare this approach to standard population genetics, it is useful to recast eq. (3.8) for the elementary sequence states $(a_1, \ldots, a_l)$,

$$P^0_{\rm stat}(r) = \sum_{(a_1,\ldots,a_l)|r} \mathcal{P}_{\rm stat}(a_1, \ldots, a_l) \ , \tag{3.10}$$

where the sum runs over all sequence states at fixed r. At neutrality, the distribution over sequence states factorizes in the single nucleotide positions,

$$\mathcal{P}^0_{\rm stat}(a_1, \ldots, a_l) = \prod_{i=1}^{l} \nu_0(a_i). \tag{3.11}$$

In the specific case of the two-state model, $\nu_0(a_i)$ is simply a flat distribution over nucleotides but it is obvious how this form can be generalized to arbitrary nucleotide frequencies.

According to eq. (3.8), the stationary distribution under selection takes the form

$$\mathcal{P}_{\rm stat}(a_1, \ldots, a_l) = \mathcal{P}^0_{\rm stat}(a_1, \ldots, a_l) \exp[2NF(r)]. \tag{3.12}$$

The salient point is that $F(r)$ is generically a strongly nonlinear function of r due to the sigmoid dependence of the binding probability on r. An analogous

statement holds beyond the two-state approximation for the dependence of F on the binding energy E. Hence, even if $\mathcal{P}^0_{\rm stat}(a_1,\ldots,a_l)$ factorizes in the single nucleotide positions, $\mathcal{P}_{\rm stat}(a_1,\ldots,a_l)$ does not. The selection introduces specific correlations between the nucleotides: the fitness differences and, hence, the nucleotide frequencies at one position i depend on all other $l-1$ positions in the motif.

3.5.2 Adaptive generation of a binding site

We now apply the dynamics (3.7) to the problem of adaptively generating a binding site in response to a newly arising selection pressure. We study a case of strong selection ($sN = 100$) in the crater fitness landscape (3.4) with parameters $\ell = 10$, $\epsilon = 2$, $\rho^{\rm on} = 3$, $\rho^{\rm off} = 1$ (implying that the factor concentrations differ by a factor of 50), and a case of moderate selection ($sN = 7$) in the mesa landscape with parameters $\ell = 10$, $\epsilon = 1$, $\rho = 3.6$. (The mesa type may be most appropriate for factors with multiple binding sites such as the CRP repressor in *E. coli*, where binding to an individual site is negligible in the *off* state.) The fitness landscapes for both cases are shown in fig. 3.7 (a,b) in units of the selection pressure s. Substantially beneficial mutations occur only on their sigmoid slopes, i.e., in narrow ranges of r. The upper boundary of this region is given by $r_s = \rho^{\rm on} + \log[sN(\mathrm{e}^\epsilon - 1)]/\epsilon$, which takes typical values $r_s = 5-7$. In fig. 3.7 (c,d), we show a sample history of adaptive substitutions from $r = 5$ to lower values of r, which are close to the point $r_{\rm max}$ of maximal fitness. The statistics of this adaptation is governed by the ensemble $P(r,t)$; the average $\overline{r}(t)$ and the standard deviation $\delta r(t)$ appear also in fig. 3.7 (c,d). In the case of strong selection, the expected time of the adaptive process is readily estimated in terms of the uphill rates in (3.7),

$$T_s = \frac{1}{2\mu N} \sum_{r=r_{\rm max}+1}^{r_s} \frac{1}{r(F(r-1)-F(r))}, \tag{3.13}$$

and takes values of a few times $1/s\mu N$. We emphasize again that this simple form depends only on the qualitative form of the fitness landscape, namely, that weakly and strongly binding sequence states are separated only by few point mutations. The conclusions are thus largely independent of the details of the fitness landscape, which justifies using the two-state approximation.

Can such a selective process actually happen? This depends on the initial state of the promoter region in question *before* the selection pressure for a new site sets in. The region is approximated as an ensemble of $L_1 = L - \ell + 1$ candidate sites undergoing *independent* neutral evolution, i.e., the simultaneous updating of ℓ sites by one mutation is replaced by

independent mutations. The length of the promoter region is denoted by L. At stationarity, the Hamming distance at a random site then follows the distribution $P_{\rm stat}(r) \sim \exp[S(r)]$ shown as empty bars in fig. 3.7 (e,f). The minimal distance $r_{\rm min}$ in the entire region is given by the distribution $\mathcal{P}(r) = Q_{\rm stat}^{L_1}(r) - Q_{\rm stat}^{L_1}(r+1)$, where $Q_{\rm stat}(r) = \sum_{r' \geq r} P_{\rm stat}(r')$ is the cumulative distribution for a single site. $\mathcal{P}(r)$ is found to be strongly peaked, taking appreciable values only in the range $\overline{r_{\rm min}}(\ell, L) \pm 1$ around its average. We assume selective evolution sets in as soon as at least one site has a Hamming distance $r \leq r_s$. This is likely to happen spontaneously if $r_s \gtrsim \overline{r_{\rm min}}(\ell, L)$, leading to a joint condition on ℓ, L, and r_s. For $r_s \lesssim \overline{r_{\rm min}}(\ell, L) - 1$, there is a neutral waiting time before the onset of adaptation. Its expectation value

$$T_0 = \frac{1}{\mu} \frac{Q_{\rm stat}^{L_1+1}(r_s+1)}{L_1(r_s+1) P_{\rm stat}(r_s+1)} \tag{3.14}$$

is calculated in the appendix. It is generically much larger than the adaptation time T_s, rendering the effective generation of a new site less feasible.

The stationary distribution $P_{\rm stat}(r)$ under selection is given by (3.8) and shown as filled bars in fig. 3.7 (e,f). For strong selection, it is peaked at the point $r_{\rm max}$ of maximal fitness. For moderate selection, it takes appreciable values for $r = 0 - 4$: the binding site sequences are *fuzzy*. Assuming that the CRP sites at different positions in the genome of *E. coli* have to a certain extent evolved independently, we can fit $P_{\rm stat}(r)$ with their distance distribution (data taken from [17]). At the values of ϵ and $\rho^{\rm on}$ chosen, the two distributions fit well, see fig.3.7 (f). This finding is discussed in more detail below.

3.5.3 Adaptation of Binding Cooperativity

The cooperative binding of transcription factors involves protein-protein interactions which may be specific to the DNA substrate. These interactions often do not require conformational changes of either protein involved and depend only on few specific contact points. They result in a modest energy gain of order $3 - 4 k_B T$ [51]. Hence, it is a reasonable simplification to study the adaptive adjustment of binding affinities using a simple generalisation of the two-state binding model. We define the energies $E_1/k_BT = \epsilon r_1$ and $E_2/k_BT = \epsilon r_2$ for the binding of a single factor and $E_{\rm pair}/k_BT = \epsilon[r_1 + r_2 - 2(\gamma/\tilde{\ell})(\tilde{\ell} - \tilde{r})]$ for the simultaneous binding of both factors. The cooperativity gain is assumed to result from mutations at $\tilde{\ell}$ positions in the DNA sequences of the factors, which encode the amino acids at the protein-protein contact points. These mutations define a Hamming

distance $\tilde{r} = 0, \ldots, \tilde{\ell}$ from the target sequence for optimal protein-protein binding, and $2\gamma\epsilon/\tilde{\ell}$ is the binding energy per nucleotide. Here we use the values $\epsilon = 2$, $\tilde{\ell} = 6$ and $\gamma = 1$ but the qualitative patterns shown below are rather robust.

The resulting equilibrium probabilities for the four thermodynamic states $(--)$ (both factors unbound), $(+-)$ and $(-+)$ (one factor bound), and $(++)$ (both factors bound) are

$$\begin{aligned} &q_{--}, \\ &q_{+-} = q_{--} \exp[-\epsilon(r_1 - \rho_1)], \\ &q_{+-} = q_{--} \exp[-\epsilon(r_2 - \rho_2)], \\ &q_{++} = q_{--} \exp[-\epsilon(r_1 + r_2 - \rho_1 - \rho_2 - 2\gamma)], \end{aligned} \tag{3.15}$$

with the normalisation $q_{--} + q_{+-} + q_{-+} + q_{++} = 1$. The scaled chemical potentials ρ_1 and ρ_2 are independent variables if the two sites bind to different kinds of factors and are equal if they bind to the same kind. As before, the binding probabilities determine expression levels and, therefore, the fitness. Here we study only pairs of sites contributing additively to the expression level in each cellular state, where we have

$$F = \sum_{\alpha} s^{\alpha}(q^{\alpha}_{+-} + q^{\alpha}_{-+} + 2q^{\alpha}_{++}). \tag{3.16}$$

Other important cases include activator-repressor site pairs such as the famous *lac* operon [53], where the transcription-factor induced expression level is proportional to q_{+-}. The stochastic dynamics of substitutions is straightforward to generalise; it leads to a Master equation like (3.7) for the joint distribution $P(r_1, r_2, \tilde{r}, t)$. This higher-dimensional equation can again be solved exactly for its steady state

$$P_{\rm stat}(r_1, r_2, \tilde{r}) \sim \exp[S(r_1) + S(r_2) + S(\tilde{r}) + 2NF(r_1, r_2, \tilde{r})]. \tag{3.17}$$

Here we discuss two simple examples of fitness landscapes where binding cooperativity evolves by adaptation to specific functional demands. A *genetic switch* with a sharp expression threshold is favoured in a system with a single transcription factor having similar concentrations in its *on* and *off* cellular state. As can be seen from eq. (3.15), cooperative binding can sharpen the response of the binding probability to variations in factor concentration, $q_{++} \sim 1/[1 + \exp(-2\epsilon\rho + \ldots)]$ versus $p \sim 1/[1 + \exp(-\epsilon\rho + \ldots)]$ as given by (3.2) for individual binding. Figs. 3.7 (a,c) show the fitness landscape $F(r_1, r_2, \gamma)$ obtained from (3.15) and (3.16) for $\rho^{\rm on} = 2.5$, $\rho^{\rm off} = 1.5$, and $s = s^{\rm on} = -s^{\rm off}$. A simple *signal integration module* responds to two different factors in four different cellular states, (*on, on*), (*on, off*), (*off, on*), (*off, off*).

Individually weak but cooperative binding leads to expression of the gene only if both factors are present simultaneously. This case is favoured by a fitness function of the form (3.16) with selection coefficients $s = -s^{\mathrm{off,off}} = -s^{\mathrm{on,off}} = -s^{\mathrm{off,on}} = s^{\mathrm{on,on}}/2$. The resulting fitness landscape $F(r_1, r_2, \gamma)$ is shown in figs. 3.7 (b,d) for chemical potentials $\rho^{\mathrm{on}} = 3$, $\rho^{\mathrm{off}} = 1$ (for each factor).

In both cases, a pair of sites with weaker individual binding ($r_1, r_2 = 3-4$) and cooperativity ($\gamma = 1$) is seen to have a higher fitness than an optimal pair ($r_1 = r_2 = 2$) without cooperativity, as expected. Adaptive pathways $\overline{r_{1,2}}(t)$ and $\overline{\gamma}(t)$ for strong selection ($sN = 100$) are shown in fig. 3.7 (e,f). Typical adaptation times T_s are again a few times $1/(s\mu N)$. A closer look reveals that this fast adaptation sometimes leads to a *metastable* local fitness maximum with some degree of cooperativity. *Compensatory* mutations (see below) are then required to reach the global maximum, a process that may be considerably slower. The fuzziness $\delta r_{1,2}(t)$ and $\delta\gamma(t)$ observed in fig. 3.7 (e,f) decays on the larger time scale of compensatory mutations, reflecting the presence of such metastable states.

3.6 Conclusions

Transcription factors and their binding sites emerge as a suitable starting point for quantitative studies of gene regulation. Binding site sequences are short and their sequence space is simple. Moreover, the link between sequence, binding affinity, and fitness is experimentally accessible. For a single site, the simplest examples are of the *mesa* [17] or of the *crater* type, see fig. 3.7 (a,b). Landscapes for a pair of sites with cooperative binding interactions are of a similar kind as shown in fig. 3.7 (a-d). They can be used to predict the outcome of specific single-site mutation experiments to a certain extent.

Fast adaptation may generate or eliminate a new binding site

Despite this simplicity, the evolutionary dynamics of binding sites is far from trivial, since it is governed, in the generic case, by the interplay of three evolutionary forces: selection, mutation, and genetic drift. Here we have focused on the dynamical regime appropriate for eukaryotes, where the evolution can be approximated as a stochastic process of substitutions. We find the possibility of selective pathways generating a new site in response to a newly arising selection pressure, starting from a neutrally evolved initial state and progressing by point substitutions. Such a selective formation takes roughly

$T_s \approx \Delta r/(2s\mu N)$ generations, where Δr is the number of adaptive substitutions required. This number is given by the Hamming distance between the onset of selection and the point of optimal fitness, $\Delta r = r_s - r_{\max}$, and takes values $2-3$ for typical fitness landscapes; see fig. 1(a,b). For *Drosophila melanogaster*, with $\mu \approx 2 \times 10^{-9}$ [40] and $N \approx 10^6$, the resulting T_s is of the order of 10^6 generations or 10^5 years even for sites with a relatively small selection coefficient $s = 10^{-3}$. Such selective processes are faster than neutral evolution by a factor of about 1000 and would allow for independent generation of sites even after the split from its closest relative *Drosophila simulans* about 2.5×10^6 years ago. Notice that new sites are more readily generated in large populations. As discussed above, generating a new site may also require a neutral waiting time T_0 until at least one candidate site in the promoter region of the gene in question reaches the threshold distance r_s from the target sequence, where selection sets in. For site formation to be efficient, however, selection must be able to set in spontaneously, i.e., T_0 must not greatly exceed the adaptive time T_s. This places a bound on the relevant length ℓ of the binding motif that can readily form in a promoter region of length L. Given $L \approx 300$, for example, a motif with $\ell = 8$ and $r_s = 3$ could still allow for spontaneous adaptive site formation. (For longer motifs, corresponding to groups of sites with fixed relative distance, this pathway would require promoter regions of much larger L.) A more general case has recently been treated numerically in [8], where the dependence of the neutral waiting time on the G/C ratio of the initial sequence has been investigated. One may speculate that this adaptive dynamics is indeed one of the factors influencing the length of regulatory modules in higher eukaryotes.

Clearly, the present model also allows for pathways of *negative selection* leading to the elimination of spurious binding sites in regulatory or non-regulatory DNA where the binding has an adverse fitness effect. This is important since under neutral evolution, candidate sites with a distance of at most r_s from the target sequence occur frequently on a genome-wide scale. A recent study has indeed found evidence for such negative selection from the underrepresentation of binding site motifs over the entire genome [50].

Binding sites under selection have nucleotide frequency correlations

We have shown that under stationary selection the frequencies of nucleotides at any two positions of the binding sequence are correlated. For the two-state model, the correlations are the same for any pair of positions $i \neq j$ and can be computed exactly from the joint distribution (3.12). We emphasize that these correlations refer to an ensemble of independently evolving (monomorphic)

populations and are not to be confused with linkage disequilibria within one population. This finding limits the accuracy of bioinformatic weight matrices, which are often assumed to factorize in the nucleotide positions even in the presence of selection.

Experimental tests: Binding site polymorphisms and phylogenies

The predictions of our model lend themselves to a number of experimental tests. In the dynamical regime appropriate for eukaryotes ($\mu N \ll 1$), populations should be monomorphic at most positions of their binding site sequences and polymorphic at a few. On the other hand, the quasispecies model discussed in refs. [17, 18] (which assumes $\mu N \gg 1$) may be most appropriate in viral systems. The intermediate regime $\mu N \sim 1$ with frequent polymorphisms *and* genetic drift could be realized in some bacterial systems and presents a challenge for theory. Thus it would be very interesting to compare the statistics of single-nucleotide polymorphisms at binding sites in eukaryotes, bacteria, and viruses. Polymorphism data are expected to contain evidence for adaptive evolution. However, statistical tests of selection must be modified for promoter sequences [46, 50]. A recent study uses data on binding sites in three yeast species and deduces the rates of sequence evolution [10].

A complementary source of information are phylogenies of binding sites. Trees with functional differences between branches contain information on the generation of new sites or of interactions between sites and on the time scales involved. In a tree for a conserved site or group of sites with sufficiently long branches, the fuzziness of the sequences observed on different branches is given by the ensemble $P_{\rm stat}$ introduced above. For strong selection, $P_{\rm stat}$ lives on the *quasi-neutral* network of sequence states with maximal fitness, where two neighbouring sequence states are linked by neutral mutations or by pairs of *compensatory* mutations at two different positions. In the crater landscape for a single site, this quasi-neutral network consists of all sequences with a fixed distance $r = r_{\rm max}$ from the target sequence; see fig. 3.7 (a). Beyond the two-state approximation for binding energies, it will be smaller since only some of the positions are energetically equivalent. For a group of sites, however, quasi-neutral networks can be larger since compensatory mutations can also take place at positions on different sites as shown in fig. 3.7 (d) for the example of a signal integration module. This is consistent with experimental evidence that the sequence divergence between *Drosophila melanogaster* and *Drosophila pseudoobscura* involves compensatory mutations and stabilising selection between different binding sites [47].

For weaker selection, site fuzziness increases further since $P_{\rm stat}$ extends

beyond the sequence states of maximal fitness and is influenced by mutational entropy. As shown in fig. 3.7 (f), one can explain in this way the observed fuzziness in CRP sites of *E. coli*. It would then reflect different evolutionary histories of independent populations, rather than sampling in one polymorphic population as in the quasispecies picture of refs. [17, 18]. (In a mean-field quasispecies, appreciable fuzziness occurs only for selection coefficients $s \sim \mu$, minute in other than viral systems.) However, the data are also compatible with strong selection if the selection coefficients s^α, and hence the value of $r_{\max}$, vary between different genes. Clearly, comparing $P_{\rm stat}$ with the distribution of sites in a single genome requires the assumption that the evolutionary histories of sites at different positions are at least to some extent independent. Future data of orthologous sites in a sufficient number of species will be more informative. Thus, further experimental evidence is needed to clarify the role of mutational entropy in the observed fuzziness.

Evolvability of binding sites

The present work was aimed at obtaining some insight into the molecular mechanisms and constraints underlying the dynamics of complex regulatory networks, thereby quantifying the notion of their *evolvability*. The programming of binding sites and of cooperative interactions between them is found to provide efficient modes of adaptive evolution whose tempo can be quantified for the case of point mutations. The formation of complicated signal integration patterns and of multi-factor interactions in higher eukaryotes, however, requires generalizing our arguments in two ways. There are further modes of sequence evolution such as slippage events, insertions and deletions, large scale relocation of promoter regions, and recombination. Our ongoing work is aimed at quantifying their relative importance in terms of substitution rates. Moreover, there are also more general fitness landscapes describing, e.g., binding sites interacting via the expression level of the regulated gene (such as activator-repressor site pairs) and the coupled evolution of binding sites in different genes. The rapid evolution of networks hinges upon the existence of adaptive pathways for these formative steps with a characteristic time scale $T_s \sim 1/(s\mu N)$ much smaller than $T_0 \sim 1/\mu$, the time scale of neutral evolution. The presence of these two time scales has a further interesting consequence. If the selection pressure on an existing site ceases, that site will disappear on the larger time scale T_0. It is possible, therefore, that large existing networks have accumulated a considerable number of *redundant* regulatory interactions acquired by selection in their past. This may be one factor contributing to their robustness against perturbations.

3.7 Methods - Neutral evolution of binding sites

To estimate the average neutral waiting time T_0, we study the mutation dynamics in the restricted range $r = r_s + 1, ..., \ell$, allowing mutations from $r_s + 1$ to r_s but suppressing mutations from r_s back to $r_s + 1$. We evaluate the time-dependent solution $P(r,t)$ of the Master equation (3.7) with the initial condition $P(r,0) = P_{\rm stat}(r)$, and the resulting cumulative probability $Q(t) = \sum_{r \geq r_s+1} P(r,t)$. The current across the lower boundary, $J(t) = \mu(r_s+1)P(r_s+1,t) = -{\rm d}Q/{\rm d}t$, determines the waiting time for a single site,

$$T_0 = \int_0^\infty {\rm d}t\, t\, J(t) = \int_0^\infty {\rm d}t\, Q(t). \tag{3.18}$$

This is formally solved by expanding in eigenfunctions of the mutation operator.

In the case relevant here, the system remains close to equilibrium since the boundary current is much smaller than typical currents for $r \geq r_s$. Hence, $P(r,t) \approx P_{\rm stat}(r)\exp(-\lambda t)$ with $\lambda = J(0)/Q(0) = \mu(r_s+1)P_{\rm stat}(r_s+1)/Q_{\rm stat}(r_s+1)$. We conclude that the waiting time for a single site is positive with probability $Q_{\rm stat}(r_s+1)$, following a distribution $\sim \exp(-\lambda t)$, and 0 otherwise. The resulting expectation value is $T_0 = Q_{\rm stat}(r_s+1)/\lambda$. For L_1 independent sites, the distribution of positive waiting times is still exponential, and T_0 is given by an expression of the form (3.18) with a total boundary current $J(t,L_1) = {\rm d}Q^{L_1}(t)/{\rm d}t$. This yields $T_0 = Q^{L_1}_{\rm stat}(r_s+1)/L_1\lambda$ as given by (3.14). The average waiting time (in units of $1/\mu$) becomes large for values of r_s in the tail of the distribution $\mathcal{P}(r)$, where $Q^{L_1}_{\rm stat}(r_s+1) \approx 1$. This is the case for $r_s \lesssim \overline{r_{\rm min}}(\ell, L) - 1$.

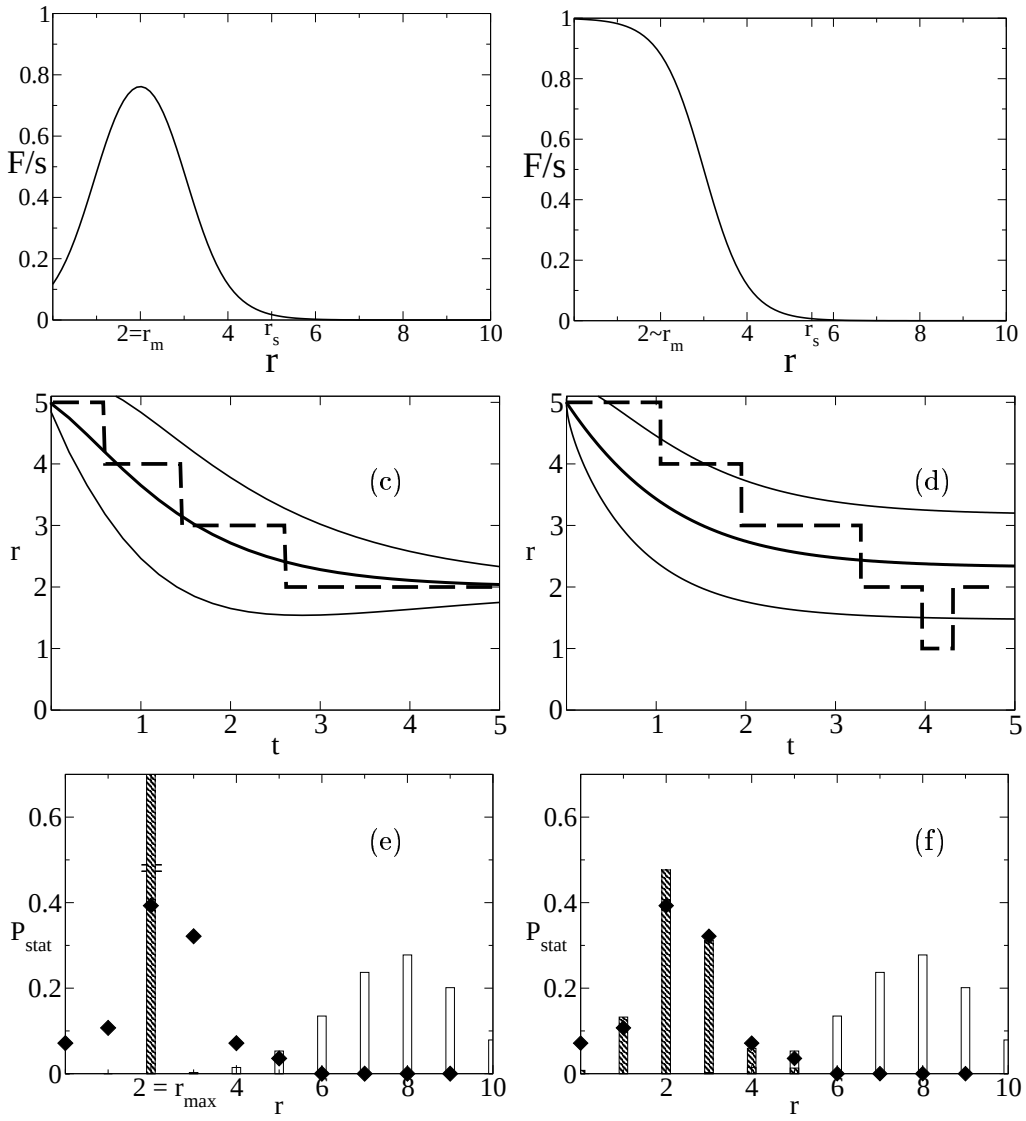

Fig. 3.7 Fitness landscapes and adaptive evolution for a single binding site. Strong selection ($sN = 100$, left column), moderate selection ($sN = 6.8$, right column). (a) *Crater* landscape (eq. 3.4) and (b) *Mesa* landscape (eq. 3.5), as a function of the Hamming distance r from the target sequence (within the approximation of the two-state model). (c,d) Adaptive dynamics as a function of time t measured in units of $1/2s\mu N$: Single history $r(t)$ (dashed lines), ensemble average $\bar{r}(t)$ (thick solid lines) and width given by the standard deviation curves $\bar{r}(t) \pm \delta r(t)$ (thin solid lines). (e,f) Stationary ensembles $P_{\rm stat}(r)$ of binding site sequences with selection (filled bars) and for neutral evolution (empty bars). Histogram of Hamming distances of CRP site sequences in *E. coli* from their consensus sequence (diamonds, from [17]).

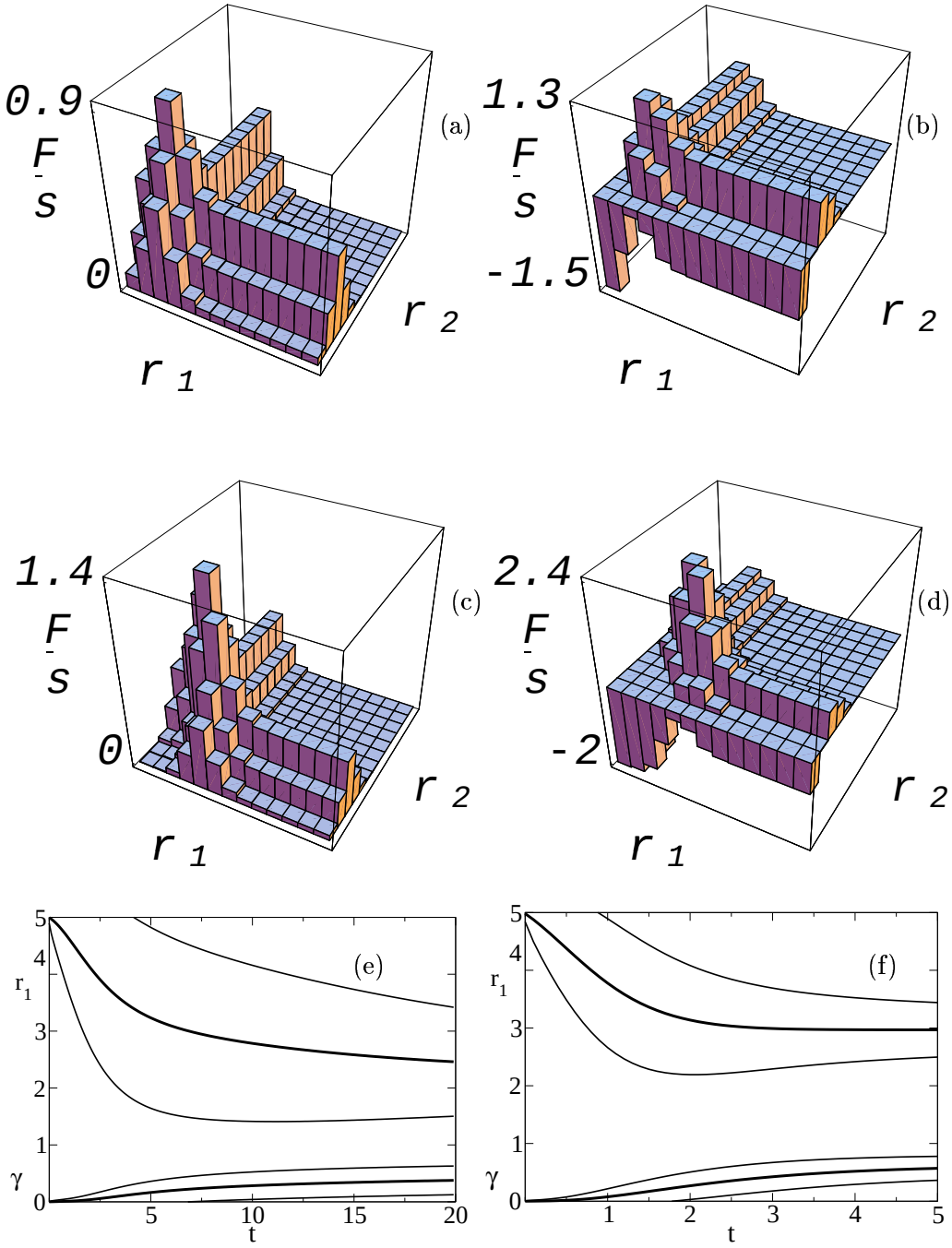

Fig. 3.7 Fitness landscapes and adaptive evolution for a pair of sites with cooperative binding. Genetic switch (left column), signal integration module (right column). (a,b) Fitness landscape $F(r_1, r_2)$ without cooperativity ($\gamma = 0$). (c,d) Fitness landscape $F(r_1, r_2)$ with cooperativity ($\gamma = 1$). Next-nearest neighbour states (r_1, r_2) and $(r_1' = r_1 \pm 1, r_2' = r_2 \pm 1)$ of similar fitness are linked by *compensatory* mutations if the intermediate states (r_1, r_2') and (r_1', r_2) have lower fitness. (e,f) Adaptive dynamics: ensemble averages $\overline{r}_1(t) = \overline{r}_2(t)$ and $\overline{\gamma}(t)$ (thick lines), ensemble width given by $\overline{r}_1(t) \pm \delta r_1(t)$ (same for r_2) and $\overline{\gamma}(t) \pm \delta\gamma(t)$ (thin lines);

Chapter 4

Search for Co-regulated Genes

The results of this part were obtained in collaboration with Prof. Nikolaus Rajewsky (New York University) and Prof. Anne Goldfeld (Harvard Medical School).

4.1 Gene Coregulation

So far, many promoters have been located experimentally. What is known is that a particular set of TFs binding TFBS in a promoter, forms an enhanceosome which determines how many transcripts per second will be produced. In short, *enhanceosomes are machines that use an input concentrations of TFs, do the mathematics, and as an output yield the rate of transcription* [7, 29]. The logic of regulatory sequences is still unknown and the question if there is a code controlling how TFBS are organized is still to be answered.

Cells know how to select the right cocktail of genes to express and produce TFs relevant for a tissue [59, 87]. It turns out that the function of genes is mainly transcriptional regulation (via TF production) over constitutive-protein production. TFs bind to motifs in a promoter, thus turning on transcription at the transcription start site (TSS). Not the whole of the transcribed sequence is translated. The region between the TSS and the translation start site is called 5'UTR (five prime untranslated region).

We assume that " similar " promoters control coregulated genes. If different genes are controlled by promoters containing same binding sites (in the first approximation, regardless of synteny or number of binding sites), they are likely to be expressed simultaneously, or at least in some correlated fashion.

4.2 Biology of $TNF-\alpha$

A protein called human *TUMOR NECROSIS FACTOR-*α, from the family of *cytokines* [88], induces *apoptosis*. Upon blood infection, **TNF-**α is produced, which in turn stimulates many cell types to kill themselves, in particular cells that line the blood vessels, thus causing circulatory failure or *septic shock*. Human TNF-α promoter sequence is shown below [63, 66]:

>human-TNF-α-promoter
ATGCTTGTGTGTCCCCAACTTTCCAAATCCCCGCCCCC
GCGATGGAGAAGAAACCGAGACAGAAGGTGCAGGGCC
CACTACCGCTTCCTCCAGATGAGCTCATGGGTTTCTCC
ACCAAGGAAGTTTTCCGCTGGTTGAATGATTCTTTCCC
CGCCCTCCTCTCGCCCCAGGGACATATAAAGGCAGTTG
TTGGCACACCC

The TNF-α gene activation is one of the few examples in human and mouse, where a detailed understanding of combinatorial transcriptional regulation and specificity exists. Experiments have shown that different sets of transcription factors bind to shared binding sites in the $TNF-\alpha$ promoter in response to different stimuli [63, 64, 66, 69]. Transcription of the $TNF-\alpha$ requires specific sets of TFs and architectural proteins to form an active higher-order complex called or enhanceosome. The Goldfeld lab investigated different enhanceosomes, each being formed as a consequence of a distinct stimulation.

stimulus	TF complex	protein transcribed
Ca^{++}	enhanceosome1	$TNF-\alpha$
$virus$	enhanceosome2	$TNF-\alpha$

TNF-α promoter comprise TFBS for the following TFs (see Fig. 4.1) :

1. CRE,
2. ETS,
3. NFAT,
4. SP1.

Upon Ca^{++} stimulation, only CRE and NFAT moleculues attach to the promoter (see Fig. 4.2), while virus stimulation causes recruitment of ETS, CRE, NFAT and SP1 (see Fig. 4.3).

Thus, activation of *TNF-α* gene transcription requires a unique combination of transcriptional activators and regulatory elements. In this fashion, a single gene may be controlled in response to different extracellular stimuli. Furthermore, the specificity of *TNF-α* transcriptional activation is achieved through the assembly of stimulus-specific enhancer complexes and through synergistic interactions among the distinct activators within these enhancer

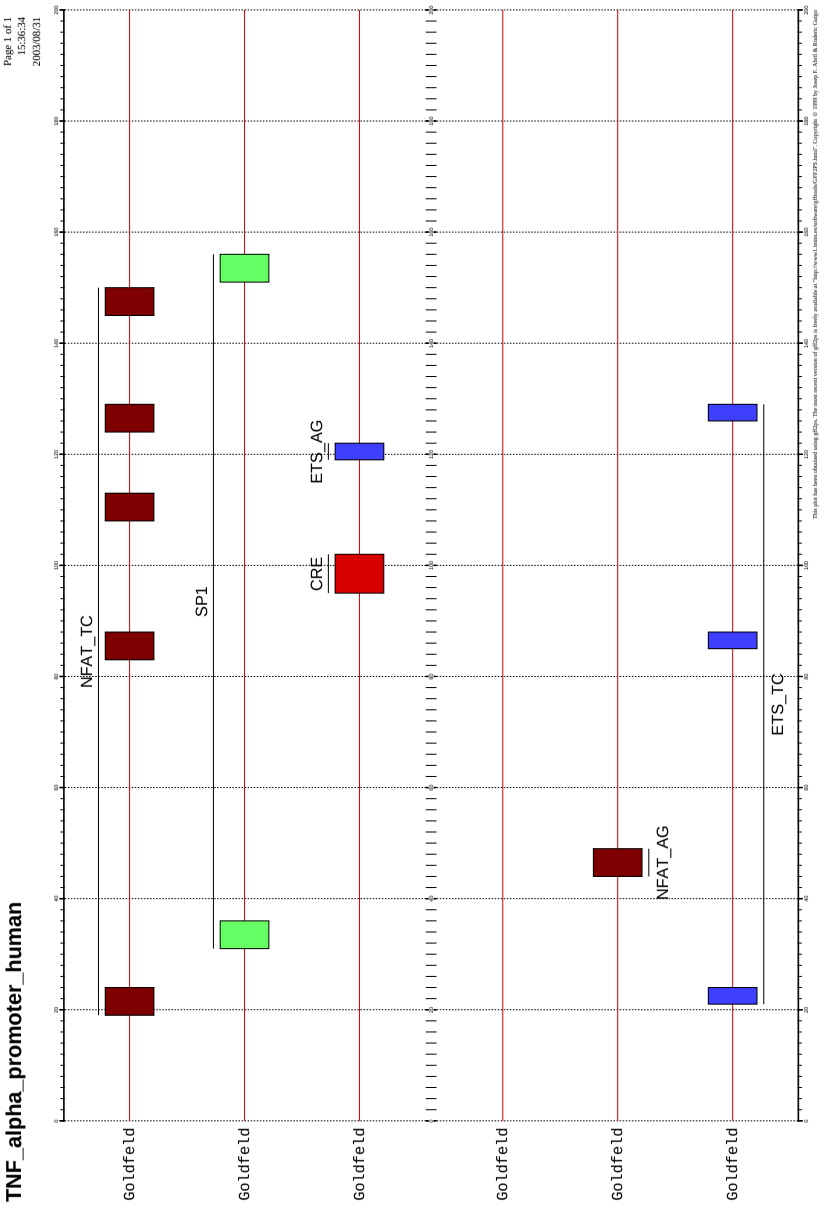

Figure 4.1: TNF-α promoter binding sites

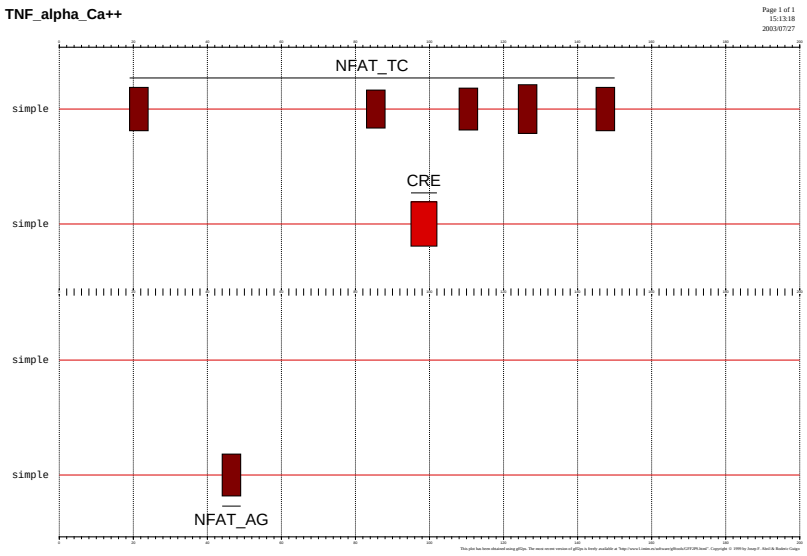

Figure 4.2: Ca^{++} stimulus-activated binding sites

complexes.
The table 4.1 clarifies synergy observed in experiments on viral stimulation. If only one type of TFs is available, for example NFAT, relative transcription will be much lower than if NFAT binds the promoter at the same time as SP1. If NFAT and SP1 both attach to their binding sites, relative transcription will increase by a factor of 17. It will not simply be the sum of NFAT-alone and SP1-alone relative transcriptions (here 4+18). This experiment, from ref. [69] , is a beautiful example of TF synergy.

TF(s)	relative transcription
NFAT	4
SP1	18
NFAT + SP1	70
CRE	2
CRE + NFAT	25
CRE + SP1	48
CRE + NFAT + SP1	720

Table 4.1: TF synergy from [69]

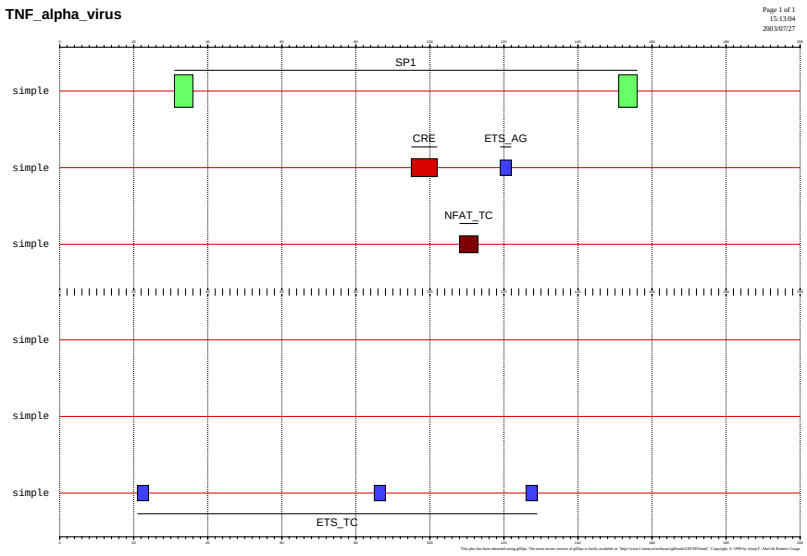

Figure 4.3: Virus stimulus-activated binding sites

4.3 Conservation of Promoter Contents

It is believed that evolutionary conserved regions are most likely to be functional [59]. On the other hand, regulation of genes, being combinatorial, allows for different architecture of enhanceosomes (loops, knots ...). Yet, TNF-α promoter lies right upstream from the gene. Its closeness and length (200 bp) set limits to spacial organization of binding sites into enhanceosomes. Therefore we expect binding sites and their synteny to be conserved, when we compare $TNF-\alpha$ regulatory sequences at the cross-species level. First, how can the degree to which two genomic sequences are similar be quantified? A common ancestor, an ancestral sequence has evolved into species-specific sequences via *substitutions* (point mutations) and *indels* (insertions and deletions). Therefore, their dissimilarity is due to presence of incorrect and missing bases. In order to see how much two different sequences of lengths l_1 and l_2 overlap, they should be placed one over another according to following rules:

- the original order of bases of the sequences must be unchanged,
- gaps may be inserted to either or to both sequences,
- the lengths of the sequences in the alignment must be equal.

There are many different ways to align two sequences, and the question is which alignment to choose, how to score the alignments [83–85]. In general a scoring scheme is based on a (match)/(mismatch) score $(D_{i,j})/(-|D_{i,j}|)$ and a gap penalty g, $i = 1, ...l_1$, $j + 1, ...l_2$. For instance,

$$\begin{aligned} D_{i,j} &= +1, match \\ D_{i,j} &= -1, mismatch \\ g &= -2, gap. \end{aligned} \tag{4.1}$$

Using a scoring scheme, the l_1xl_2 matrix is constructed. In the end, the score of the alignment is the sum of individual scores for each base. But, the task is to find the alignment which maximizes the score. Once an alignment-matrix-path is chosen, computing the score is trivial (simple summing up), yet there are many paths, and the best one should be picked out. The **transfer-matrix** method (reinvented by computer scientists (Needleman-Wunsch) as **dynamic programming** can solve the problem. The maximum score can be computed, and this score represents *similarity* of two sequences. More sophisticated algorithms have a gap extension penalty g_e in addition to gap (or gap open) penalty g_o.
In general, there are three crucial steps in the dynamic programming algorithm:

- Initialization.
- Matrix Fill.
- Traceback (Optimal Alignment).

If we need to compare more than two DNA strings, we must construct a multiple alignment. Even for the low number of sequences, finding the best multiple alignment is extremely time consuming. Therefore, fast heuristic methods are exploited [84]. The most efficient method, so far, is the restriction to aligning sequences pairwise and repetition of this procedure until all sequences to be aligned are exhausted. One way to do this is an *increment approach*, where two sequences are aligned, and then a third is added without affecting the first-pair-alignment etc. Another way would be a *divide-and-conquer* approach, where each sequence gets a partner, and their pair is then aligned. The aligned pairs are combined into aligned groups, until a final alignment is obtained.

The multiple alignment can be regarded as an *evolutionary history* of the sequences. The multiple alignment construction is equivalent to revealing the

evolutionary relationships among the aligned sequences. If they align very well, it is highly probable that they evolved recently from a common ancestor. Conversely, poorly aligned sequences are thought to have undergone a more complex and longer evolutionary path.

In order to align our 16 vertebrate (see 4.4) $TNF-\alpha$ promoters, we use a multiple alignment software ClustalW (http://www.ebi.ac.uk/clustalw/) with default parameters. In general, the major shortcoming of ClustalW program is the dependence of the final multiple alignment on the initial pairwise alignments, comprising the two most similar sequences. In order to minimize errors, these two sequences should be as close as possible. In our set of sequences we do not encounter this problem since we have, among other TNF-α promoters, the human and primate sequences, which are very closely related.

Indeed, ClustalW yielded multiple alignment (see Fig. 4.4) which shows that gaps do not fall into binding sites, as an evidence of TFBS conservation [65, 68]. Furthermore, the order of binding sites is conserved, too. It seems that the specificity of the TFBSs is roughly conserved, if not completely, then to a high degree.

4.4 Sequence Analysis of TNF-α promoter

In order to perform sequence analysis of the TNF-α promoter we focus on three TFBS search methods: PWM (position weight matrix), AHAB and SMASH. First, we give brief description of the three bioinformatical tools.

4.4.1 PWM

Experiments have shown that TFs can bind slightly different motifs, thus tolerating, to some extent, variability of TFBSs. Gapless alignment of those instances gives a count matrix f_{kj} for a particular TF.

In general, if Ω is a finite alphabet and $|\Omega|$ its cardinality ($|\Omega|$=4, Ω=$\{A, C, G, T\}$), a PWM for a particular TF that binds motif of length l, is an lx$|\Omega|$ matrix p_{kj} derived from the count matrix:

$$p_{kj} = \frac{f_{kj} + 1}{N + 4}, \tag{4.2}$$

where $k = 1, ...l$, $j \in \Omega$ and N is the total number of samples in the gapless alignment [58, 76, 83].

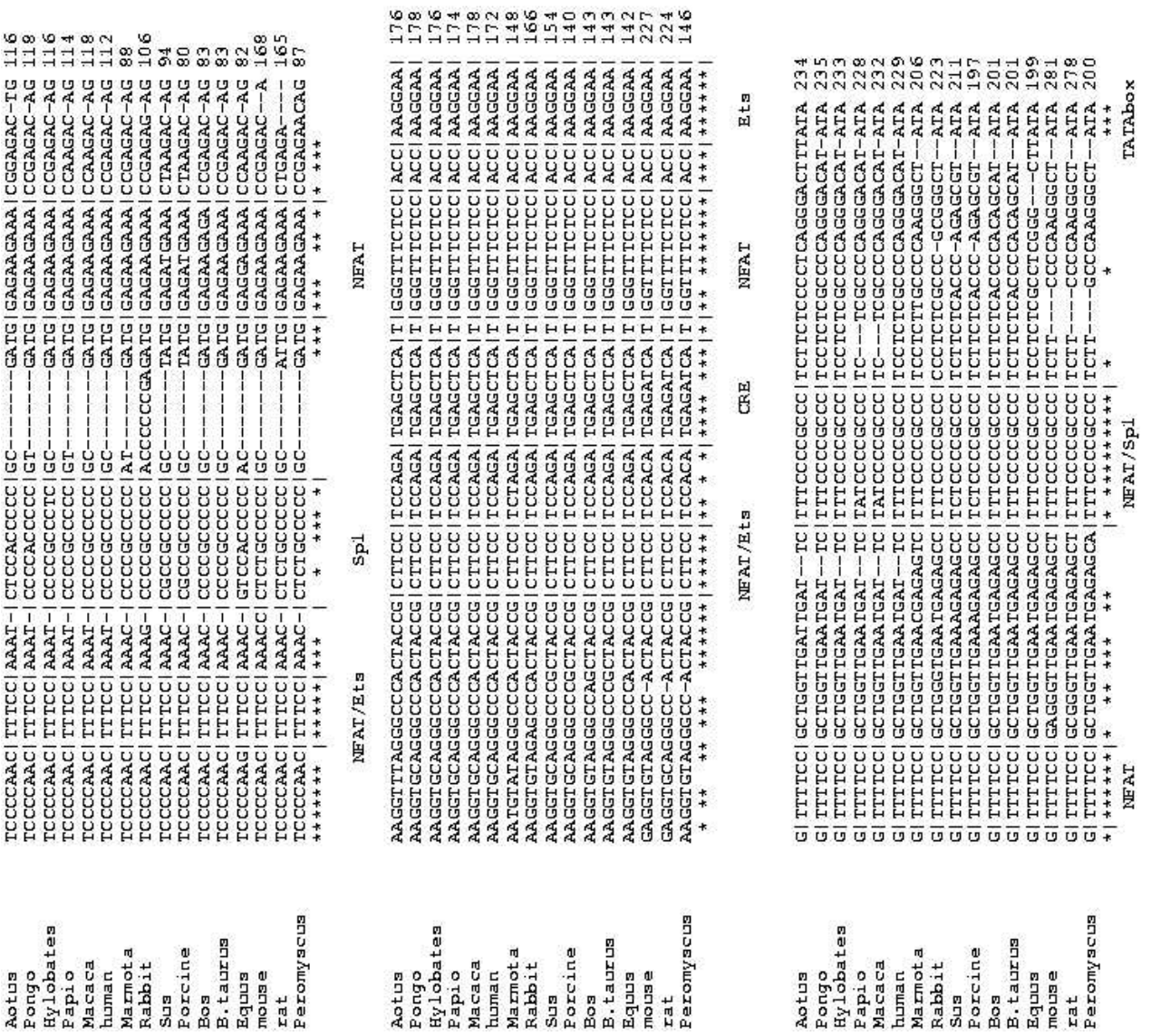

Figure 4.4: Multiple alignment of 16 vertrebate TNF-α promoters

Since $p_{kj} \geq 0$ and $\sum_j p_{kj} = 1$, p_{kj} represents a probabilistic description of a motif. This procedure is commonly called regularization and the numbers added to the count-matrix elements are called *pseudocounts*. Regularization, in a way, makes up for finite-size effects i.e. finite number of samples used to construct the count matrix. Here, we assume that, if 4 more samples were available, there is an equal chance for each of the four bases to occur at each position of the motif.

If too many pseudocounts are added when compared to real counts, motif probabilistic profile would become distorted and the search method using PWM would perform poorly.

Once probability matrices p_{kj} have been constructed, the position weight matrices can be calculated. PWM elements are defined to be the log probability of base occurrence. Since the search method should distinguish real TFBS patterns from background, a null model must be introduced p_{bkgd}. Thus, PWM elements are given by

$$w_{kj} = \log \frac{p_{kj}}{p_{bkgd}}, \tag{4.3}$$

representing log odds scores for each position $k = 1, ...l$. If $w_{kj} > 0$, the observed base j is more likely under p_{kj} than it is under p_{bkgd}. At this point we can mention another useful quality of pseudocounts. A count zero may not be converted to logarithms, and addition of pseudocounts solves the problem of singularities.

Score of a motif is the sum of individual position scores

$$S = \sum_{k=1}^{l} w_{kj}. \tag{4.4}$$

While searching a sequence with PWM, we slide a window of length l and add l logarithms to get a motif score. This is equivalent to multiplying the probabilities of the bases at each motif position. For a sequence of length L, there are $L - l + 1$ such windows, and their scores. In order to single out putative TFBSs, we need to introduce a threshold score S_t which will separate significant from insignificant scores. We rated all window scores for the $TNF - \alpha$ promoter and noticed a natural gap between 'good' and 'bad' values (see Fig. 4.5). 'Good' means 'experimentally verified'. We used this information to determine S_t, for the purpose of searching $TNF - \alpha$-like

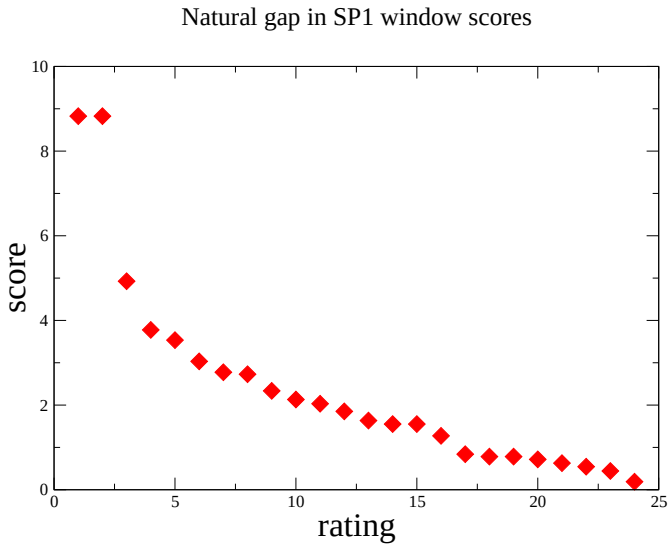

Figure 4.5: Rated SP1 window scores in TNF-α

promoters.

4.4.2 AHAB

If we know which set of TFs influences expression of a gene, we can build a TF motif-model for each TF on the basis of count matrices. As already described, count matrices are directly related to probabilities to observe certain bases at particular TFBSs. Each base position contributes to the free energy and it seems that this contribution is independent. Thus, the probability of a motif s to be a TFBS for some TF is:

$$Prob(s = TFBS) = \prod_{k=1}^{l} p_{s_k} = m(s|TF) \tag{4.5}$$

Assume there are three different TFs $\{X, Y, Z\}$ with position weight matrices $\{W_X, W_Y, W_Z\}$ and we want to know what is an optimal probabilistic tiling of a sequence s with TFBSs and background. Probability of one possible tiling T (or configuration) is given by:

$$P(T) = \prod_{TF=X,Y,Z} p_{TF} m(s|TF), \tag{4.6}$$

where p_{TF} is probability for attaching the TF to its motif and m is a measure of how close is the motif to the TFBS consensus [55]. There has never been reported that two TFs with overlapping TFBS bind simultaneously and the above formula agrees with this, say, exclusion principle. Furthermore, if there are multiple binding sites for the same TF, they are usually weak, another experimental fact included in the mentioned formula. Therefore, AHAB takes into account that quantity can make up for quality when multiple TFBS exist.

There are many possible configurations T, each with likelihood $P(T)$. The likelihood to observe s is then:

$$Z = \sum_{T} P(T), \tag{4.7}$$

in the physics language, the partition function of the system. Z can be computed recursively, using dynamic programming (transfer matrix method) in a time proportional to the sequence length and the number of W_{TF}s. By trying out all parameters, the set $\{p_{TF}\}$ which maximizes Z is chosen, applying *Maximum Likelihood Method*. Promoter score is:

$$\sigma = -\log \frac{Zmax}{Z_{bkgd}}. \tag{4.8}$$

Precisely, the free energy of the system $\Psi = -\log Z$ is minimized:

$$\frac{\delta \psi}{\delta p_{TF}} = 0 \tag{4.9}$$

Finally, the probability $P_k(W|s)$ to observe the start of PWM of length L_{TF} is:

$$P_k(W|s) = \frac{Z(1, k-1) p_{TF} m(s|W_{TF}) Z(k+l, L)}{Z(1, L)}, \tag{4.10}$$

where p_{TF} are from the converged set of probabilities $\{p_{TF}\}$.

4.4.3 SMASH

SMASH is an algorithm that can score sequences individually, as well as estimate conserved non coding regions from pairwise (e.g. human/mouse) sequence alignments and score them to detect putative TFBSs and background [54]. TFBS search is based on PWM method. In order to parse sequence pairs, SMASH takes as an input a set of TF weight matrices $\{W_{TF}\}$

and the sequences to be scored. When compared to AHAB, SMASH's major difference is that the probability of a configuration T is computed via uniformly chosen terms C_{TF} instead of optimized parameters $\{p_{TF}\}$:

$$P(T) = \prod_k C_{TF} m(s|TF). \tag{4.11}$$

The output comprises probabilities for each TF corresponding to $\{W_{TF}\}$ to bind at putative TFBSs in the sequence.

4.5 Search for human/mouse TNF-α Orthologs

Looking at the multiple alignment we can see that TNF-α promoters show high similarity, supporting the assumption that they evolved from a common ancestor. Such genes are called orthologs. Our data set comprises mouse cDNA sequences (cDNAs or complementary DNAs represent one-dimensional information which codes for a protein. Upon mRNA extraction, with the help of reverse transcriptase, a DNA copy (a cDNA) of the mRNA is created, and stored in the 'library'.)

We start with MouSDB cDNA RIKEN set ($\approx$ 19000 unique sequences, assembled by Mihaela Zavolan, Rockefeller University) of sequences. Each sequence is 50001 bases long, with an 5'UTR start in the middle. (see Fig. 4.6) Next, we extract 2001-bases-long subsequences with the 5'UTR start in the middle, i.e. the 5'UTR start flanked by 1000 bases on both sides. By using PWM, we scored all the sequences and collected a set of those sequences that contain CRE consensus 'TGAGCTCA' ($\approx$ 1000 sequences). Since TNF-α promoter contains a CRE TFBS in the middle, we decided to focus on DNA strings of bases that also contain a CRE binding motif in the center. Out of those CRE-containing sequences, we pulled out subsequences of 300-base -length with the CRE start in the center. Introducing a simple criterion that content-similar promoters will lead us to coregulated genes, we scored (PWM) the latter sequences to see how many of them contain each of the transcription factor binding sites (CRE, SP1,NFAT, ETS) and obtained 29 hits.

We repeat the same procedure starting with 2001-bp-long sequences centered around a letter 20 000 bases away from the UTR start, and detected 20 such CRE-keyed sequences.

Higher density of TNF-α-like sequences in the vicinity of the 5'UTR start, motivated us to continue the search for human orthologs using the set of

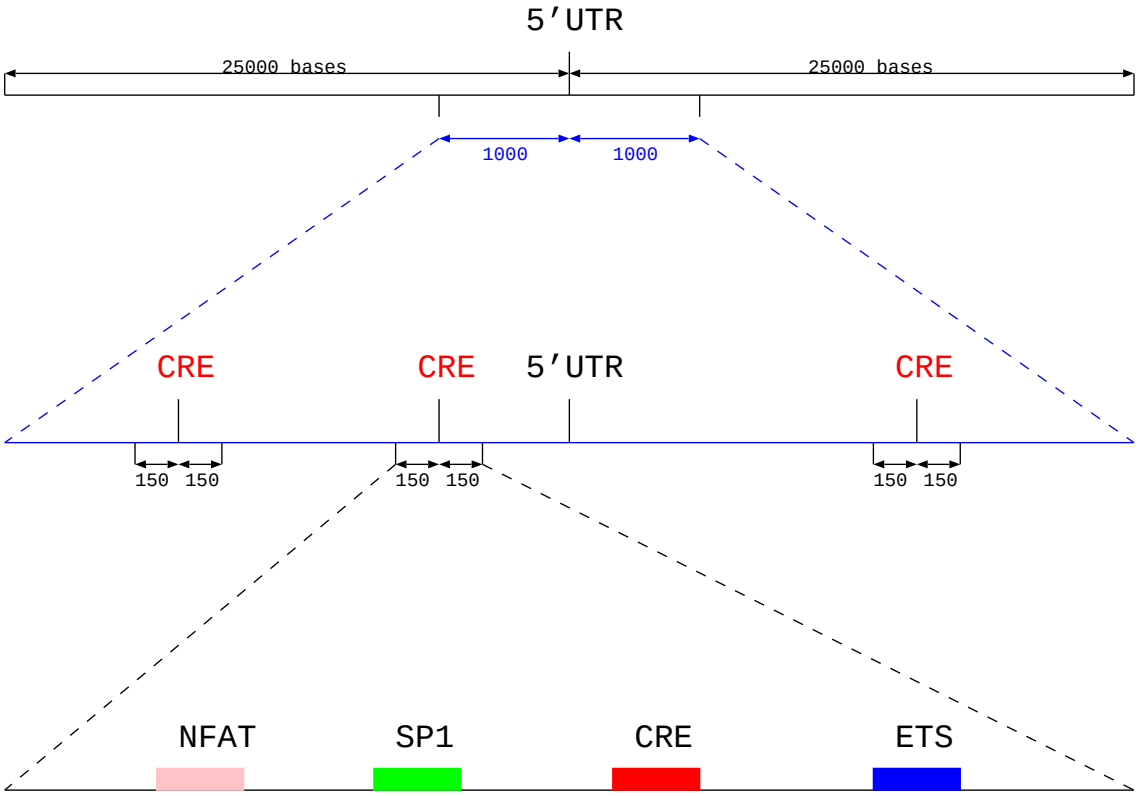

Figure 4.6: Search for human/mouse orthologs

CRE-keyed mouse sequences.

4.6 TNF-α Coregulated Genes in Humans

4.6.1 BLAST

So far, we looked for highly conserved substrings of bases (TFBS motifs) by constructing alignments of entire sequences (TNF-α promoters), based on a transfer-matrix application to computer science i.e. the Needleman-Wunsch algorithm which constructs *global alignments*. Since we are after TNF-α-similar regulatory sequences, we would like to align the human TNF-α promoter ($\approx$ 200 bases long) against substrings of the entire human genome ($\approx$ 3x10^9 bases long) and detect putative TNF-α co-regulated genes' promoters. Therefore we need to construct *local alignments*. A local alignment of two DNA strings is an alignment of any subsequence of the first and any subsequence of the second DNA string. Local similarity is quantified as the highest score yielded by any local alignment of the two sequences. The local alignment algorithm is a modification of the Needleman-Wunsch procedure,

$$\begin{aligned} D_{i,j} = max\{ & D_{i-1,j-1} + d(i,j), \\ & max_{x \leq 1}(D_{i-x,j-g_x}), \\ & max_{y \leq 1}(D_{ij-y-g_y}) \\ & \} \end{aligned} \tag{4.12}$$

called the Smith-Waterman algorithm:

$$\begin{aligned} H_{i,j} = max\{ & H_{i-1,j-1} + h(i,j), \\ & max_{x \leq 1}(H_{i-x,j-g_x}), \\ & max_{y \leq 1}(H_{ij-y-g_y}), \\ & 0 \\ & \}, \end{aligned} \tag{4.13}$$

where $h(i,j)$ is the alignment score of the bases at positions j and j, g_x is the penalty for a gap of length x in the first sequence, and g_y is the penalty for a gap of length y in the second sequence.

It is implemented into the heuristic similarity-searching program : **BLAST**, [83, 84].

BLAST anchors identical snippets of certain length (in BLAST terminology this option is called 'word length' denoted by '-W'), i.e. number of consecutive identities in order to detect a signal. For noncoding regions -W

should be the lowest possible. For example, if two sequences are very similar, but one has mutations every 7 bases, search with '-W 7' will not yield any signal at all. Identification of high-scoring word pairs leads to high-scoring alignments. Thus, BLAST heuristics is efficient in aligning sequences in a database (e.g. Human Genome) with a given query sequence to identify those that are most similar to the query. Besides -W, there are other options available to adjust similarity search to particular cases. For instance, chimp and human have genomes which are identical up to 98%, while mouse and human are known to be identical (in the DNA build-up, of course!) up to 70%. Meaning that, if we compare same-function sequences in mouse and human, it does not make sense to increase stringency level above 70%.

Suppose we found a score R by BLAST-ing a query sequence against the human genome. What is the probability to find R by chance, in other words, how many matches to a random string of bases can be found with some score $R_0 \leq R$? The quantity E, the 'E-value' of score R, for query length l_1 and database length l_2, gives a good approximation of score distribution. It is known that local ungapped scoring is Poisson distributed with mean $E(R) = Kl_1l_2\exp(-\lambda R)$:

$$Prob(R_0 \leq R) = 1 - \exp(-E(R)) \tag{4.14}$$

giving *extreme-value distribution* (also known as Gumbel distribution).

4.6.2 Results

Our data set comprises a set of mouse sequences that have a CRE binding motif in the middle and at least one of the other three TFBS, namely, ETS, NFAT and SP1. After masking repeats, we search among them for putative promoters (TNF-α-like, i.e. ≈ 200 bases long and of similar content). Besides TFBS probabilities, the AHAB scoring scheme yields scores for sequences as a whole Ψ, which we can call free energy. Therefore we begin with identifying those CRE-centered sequences that have best 60 scores. The score distribution of 962 CRE-keyed mouse sequences is shown in figure 4.7.

Before, we scored the CRE-keyed sequences applying the PWM scoring scheme which simply identifies motifs close to the consensus of corresponding TFs, without taking into account overlapping of binding sites. 29 sequences had at least one of all types of binding sites found in the TNF-α promoter.

We BLASTed well-scoring (60 obtained by AHAB and 29 by PWM scheme) CRE-centered mouse sequences against human genome with parameters:

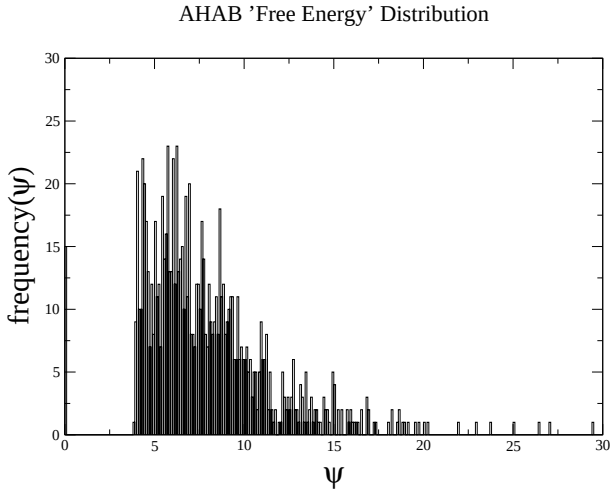

Figure 4.7: AHAB: Distribution of Ψ for 962 mouse CRE-centered sequences

-W 7 -e 1e-15 -r 3 -G 3

i.e. word size (-W) 7, E-value (-e) 10^{-15}, reward for base match (-r) 3 and cost to open gap (-G) 3.

This yielded (in all but one case) unique coordinates in the human genome which match the mouse sequences.

Out of those human sequences that aligned well we picked out 10 best AHAB-scored sequences, while we accepted all human sequences that aligned well with mouse sequences obtained by PWM (14 sequences).

Here we present the 24 human sequences that may be coregulated with the TNF-α. We searched them to identify binding sites with the following consensus motifs, using four different scoring schemes: AHAB, SMASH , SMASH on human/mouse pairs of sequences and PWM (Figures 4.8, 4.9, 4.10, 4.11):

1. NFAT, brown;
2. CRE, red;
3. SP1, green;
4. ETS, blue;

In total we ended up with 24 human sequences, whose annotation (if available at UCSC) is given in table 4.6.2. These sequences (labeled by a corresponding mouse cDNA together with the position of its alignment in the human genome) sit in front of the genes described in the second column.

Description *' chr#:start position-end position '* refers to the location of a human sequence in the Human Genome (version: April 2003) at:

http://genome.ucsc.edu/

seq	*human*	description
1	scl10809-matches-chr3:114746277-114745994	MAK3P; Molecular function: N-acetyltransferase activity (inferred from electronic annotation).
2	scl16439-matches-chr1:109396215-109395950	contains 1 SH3 domain: Epidermal growth factor receptor
3	scl16877-matches-chr9:126941746-126942058	hypothetical
4	scl17576-matches-chr20:32918601-32918858	C20orf144, CTE4-HUMAN Inhibits the interaction of APBA2 with beta-amyloid precursor protein (APP), and hence allows formation of beta-amyloid
5	scl17819-matches-chr20:63205873-63206141	predicted
6	scl18978-matches-chr4:152403723-152403391	nothing
7	scl2031-matches-chr17:9789642-9789461	GAS-7; may play a role in promoting maturation and morphological differentiation of cerebral neurons
8	scl22882-matches-chr19:59339888-59340229	hypothetical
9	scl2500-matches-chr17:40761180-40761516	EZH1-human may be involved in the regulation of gene transcription and chromatin structure
10	scl2739-matches-chr17:78371307-78371597	predicted, yet tissue: adenocarcinoma, colon
11	scl4423-matches-chr17:10684655-10684347	predicted
12	scl4877-matches-chr17:40894687-40894888	PSE3-HUMAN Implicated in immunoproteasome assembly and required for efficient antigen processing. The PA28 activator complex enhances the generation of class I binding peptides by altering the cleavage pattern of the proteasome; Synonym: Ki nuclear autoantigen A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1.
13	scl6111-matches-chr14:76256526-76256796	BC017459, predicted, found in Tissue: colon, adenocarcinoma
14	scl7533-matches-chr6:26187031-26186711	H2AL-HUMAN Molecular function: DNA binding
15	scl4915-matches-chr17:43513868-43514168	predicted
16	scl4915-matches-chr17:63187743-63188044	predicted
17	scl776-matches-chrX:97935471-97935771	sushi-repeat containing protein , SRPUL : biological process: oncogenesis
18	scl11629-matches-chr16:360876-360590	predicted
19	scl2738-matches-chr17:78332287-78332607	CBX8- involved in maintaining the transcriptionally repressive state of genes; modifies chromatin
20	scl14246-matches-chr2:160927277-160927563	ITGB6-the gene organization of the human beta 7 subunit the common beta subunit of the leukocyte integrins HML-1 and LPAM-1
21	scl14331-matches-chr2:178455289-178455578	predicted
22	scl22077-matches-chr7:155988313-155988653	hypothetical, sits in front of C7orf2
23	scl13686-matches-chr1:167340318-167340556	PMX-1 related to acute myeloid leukemia
24	scl20768-matches-chr1:47014360-47014632	MA17-HUMAN associated protein, may play an important role in tumor biology

This result may be important for tumor/stress biology, but it also allows us to study how regulatory control elements for the TNF-α coregulated genes have evolved.

The identifier *ri* # corresponds to mouse sequences in the FANTOM RIKEN database at

http://fantom2.gsc.riken.go.jp/db/search/

If you enter the ri identifier into the ID search engine, you will get the available information on the sequence you are interested in.
Those identifiers marked by a star are not annotated in the RIKEN set. They can be found by submitting the identifier to the Mouse Genome Server:

http://genome.ucsc.edu/

seq	*mouse sequence*	ri identifier
1	scl10809-matches-chr3:114746277-114745994	2600005K24
2	scl16439-matches-chr1:109396215-109395950	9630046N21
3	scl16877-matches-chr9:126941746-126942058	C630031E05
4	scl17576-matches-chr20:32918601-32918858	1700003F12
5	scl17819-matches-chr20:63205873-63206141	1810057D19
6	scl18978-matches-chr4:152403723-152403391	D330001C05
7	scl2031-matches-chr17:9789642-9789461	9530092J08
8	scl22882-matches-chr19:59339888-59340229	C730015N08
9	scl2500-matches-chr17:40761180-40761516	AB004817*
10	scl2739-matches-chr17:78371307-78371597	BC018304*
11	scl4423-matches-chr17:10684655-10684347	D130028M21
12	scl4877-matches-chr17:40894687-40894888	BC015255*
13	scl6111-matches-chr14:76256526-76256796	2610005A10
14	scl7533-matches-chr6:26187031-26186711	2610022J01
15	scl4915-matches-chr17:43513868-43514168	D630048O14
16	scl4915-matches-chr17:63187743-63188044	D630048O14
17	scl776-matches-chrX:97935471-97935771	BC028307*
18	scl11629-matches-chr16:360876-360590	1110015G04
19	scl2738-matches-chr17:78332287-78332607	AF180370*
20	scl14246-matches-chr2:160927277-160927563	D830026H09
21	scl14331-matches-chr2:178455289-178455578	D030036L15
22	scl22077-matches-chr7:155988313-155988653	4632411P08
23	scl13686-matches-chr1:167340318-167340556	C130069E24
24	scl20768-matches-chr1:47014360-47014632	2010015F12

Here, annotation of human sequences is shown. They can be searched for by submitting their RefSeq to the Human Genome at:

http://genome.ucsc.edu/

In some cases RefSeq is not available.

seq	*human sequence*	RefSeq
1	scl10809-matches-chr3:114746277-114745994	NM-025146
2	scl16439-matches-chr1:109396215-109395950	NM-024526
3	scl16877-matches-chr9:126941746-126942058	
4	scl17576-matches-chr20:32918601-32918858	NM-080825
5	scl17819-matches-chr20:63205873-63206141	
6	scl18978-matches-chr4:152403723-152403391	
7	scl2031-matches-chr17:9789642-9789461	NM-003644
8	scl22882-matches-chr19:59339888-59340229	NM-144686
9	scl2500-matches-chr17:40761180-40761516	NM-001991
10	scl2739-matches-chr17:78371307-78371597	
11	scl4423-matches-chr17:10684655-10684347	
12	scl4877-matches-chr17:40894687-40894888	NM-005789
13	scl6111-matches-chr14:76256526-76256796	
14	scl7533-matches-chr6:26187031-26186711	NM-003512
15	scl4915-matches-chr17:43513868-43514168	
16	scl4915-matches-chr17:63187743-63188044	
17	scl776-matches-chrX:97935471-97935771	NM-014467
18	scl11629-matches-chr16:360876-360590	
19	scl2738-matches-chr17:78332287-78332607	NM-020649
20	scl14246-matches-chr2:160927277-160927563	NM-000888
21	scl14331-matches-chr2:178455289-178455578	
22	scl22077-matches-chr7:155988313-155988653	NM-022458
23	scl13686-matches-chr1:167340318-167340556	NM-006902
24	scl20768-matches-chr1:47014360-47014632	

Our predictions remain to be checked experimentally. Wet lab results will not just tell us if the twenty predicted genes are coregulated with the TNF-α gene, but will clarify how successful our bioinformatical approaches are. If we look at the Fig. 4.7 where the 'free energy' distribution of about 1000 sequences is shown, we can see a Poisson-like shape. It seems that we may indeed find analogs in sequence analysis to thermodynamic concepts like

free energy. Thus, it is plausible to use (in physics a well-known method) minimization of free energy in order to detect putative binding sequences.

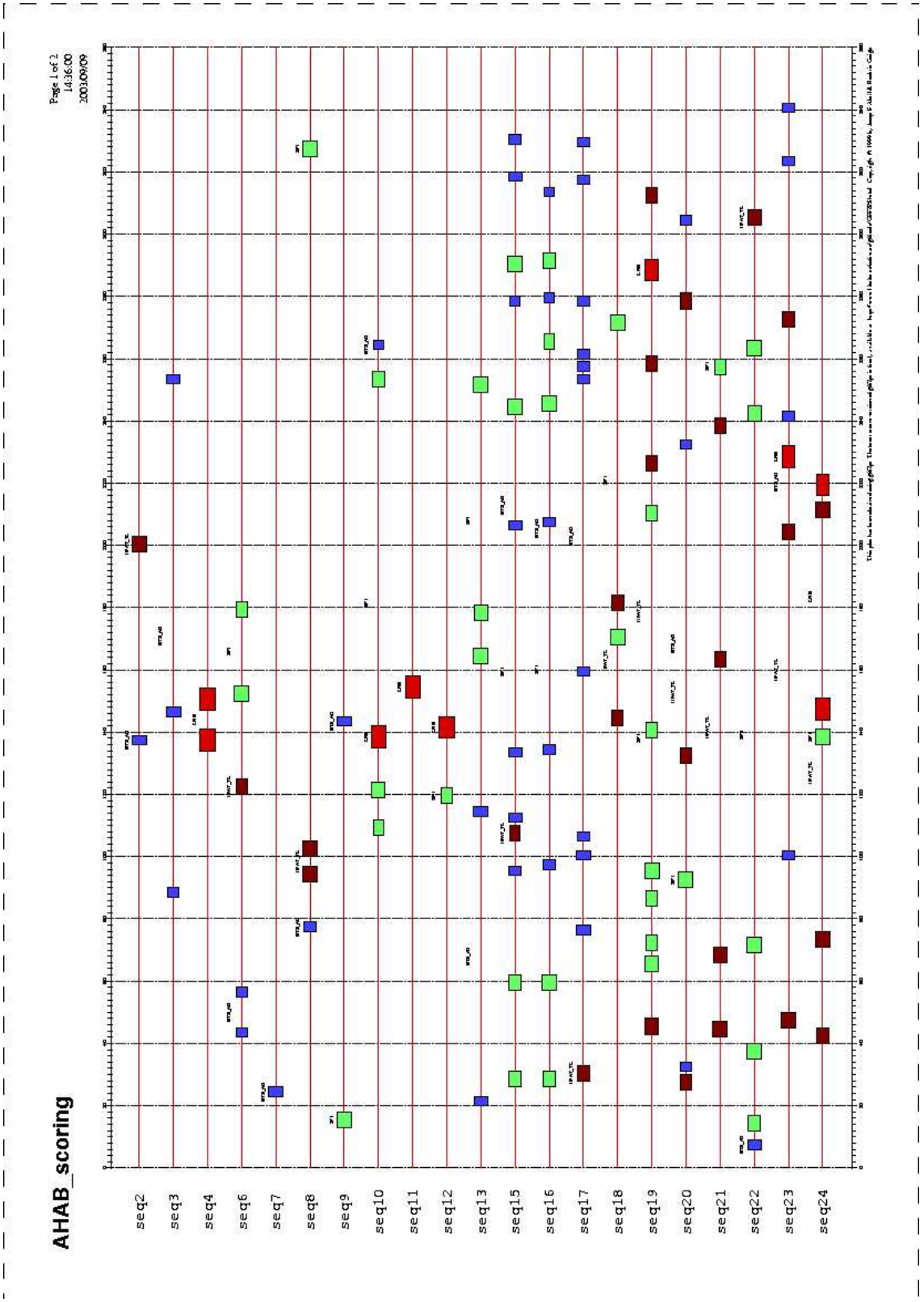

Figure 4.8: AHAB scoring method

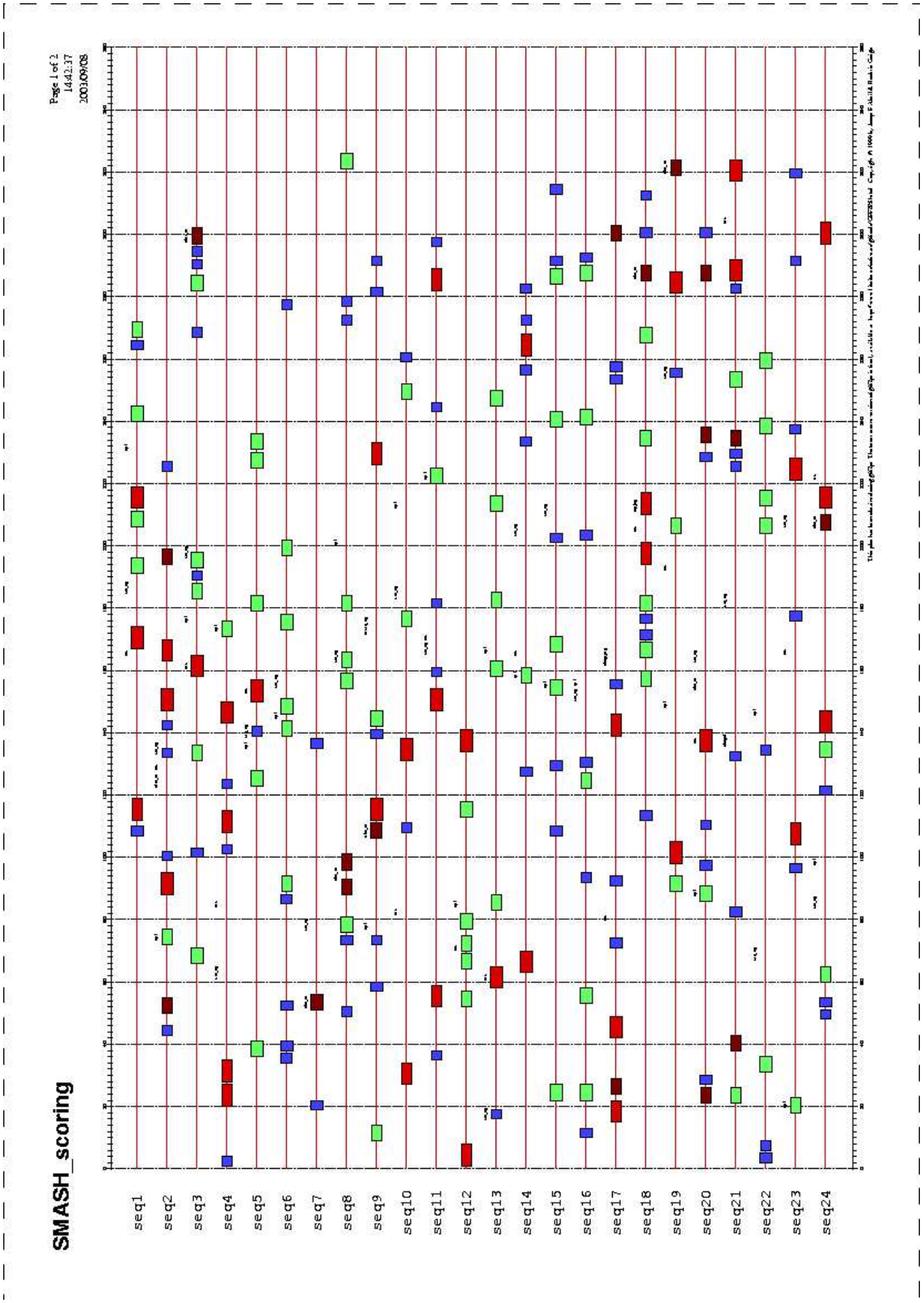

Figure 4.9: SMASH scoring method

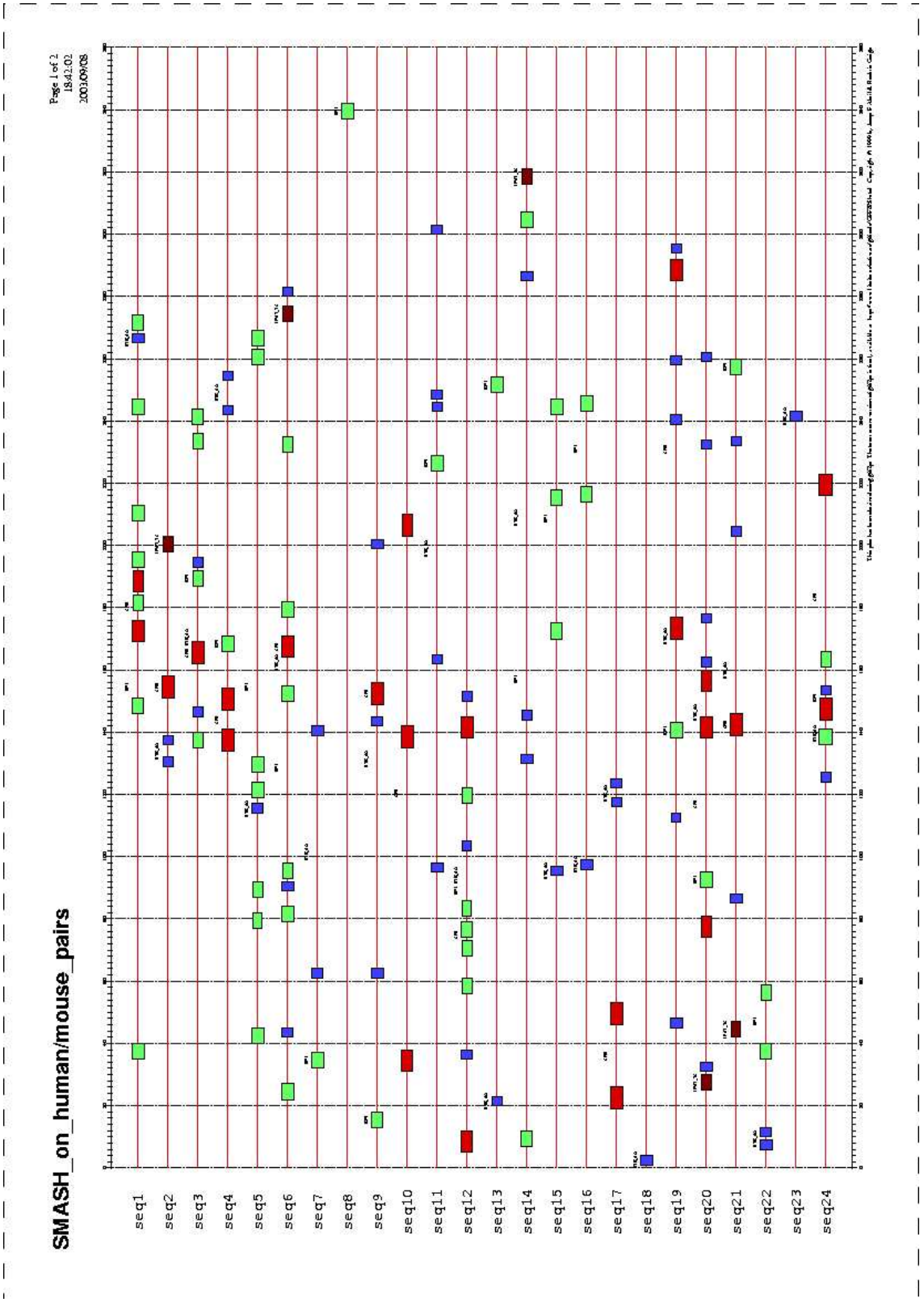

Figure 4.10: SMASH on human/mouse conserved pairs scoring method

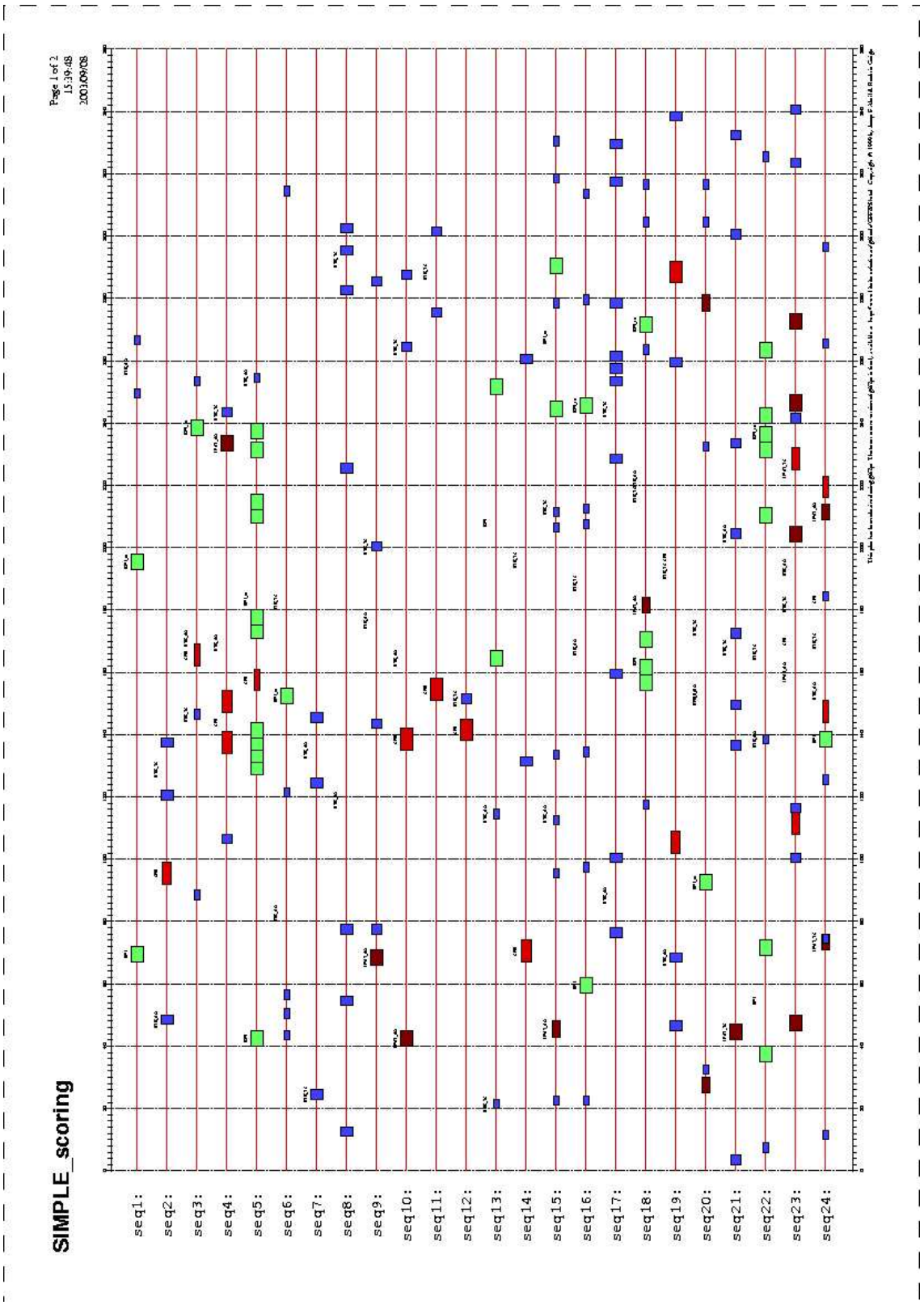

Figure 4.11: PWM scoring method

Chapter 5

Promoter Evolution

5.1 Neutral vs Adaptive Model of Promoter Evolution

One of the most prominent geneticists of the 20^{th} century Theodosius Dobzhansky said : **"Nothing in biology makes sense except in the light of evolution"**. Evolutionary changes not only enable organisms to survive despite newly arising conditions, but they allow life to conquer new surroundings, and to establish stronger control over old ones [1]. This striking adaptedness of life to its environments is due to mutations in the genetic material.

Kimura-Ohta's neutral model [41, 42] of molecular evolution supports the notion that the rate of evolution at a base position is inversely related to its functional relevance. Since TFBSs affect transcription, whereas background (nonTFBS-subsequences) should have no functional importance, it is assumed that overall point mutation and indel frequencies in the background are higher than in TFBSs. Furthermore, there is experimental evidence that many binding sites operate in a position-independent manner which should allow for unconstrained indel formation in the background. (Surprisingly, we discovered that HIV-1 promoter sequences prefer substitutions to insertions and deletions.) In short, sequences between binding sites should be free to vary. The problem is that some base positions assumed to be background may in fact be part of binding sites that have not yet been identified. Also, some TFBS point mutations may be functionally tolerated, resulting in either weak binders or in binding site turn over.

Not all binding sites are equally important for promoter function. Experimental comparisons usually reveal weak or almost no detectable influence of certain TFBSs (mostly a greater number), while some are literally indispensable (only a few) for transcription activation. Therefore, essential binding sites should evolve at a relatively lower rate. Binding sites with lower transcription impact are often multiply-represented. Functionally redundant sites belong to this category, too. In such cases, binding sites should mutate faster than unique ones, unless these multiple TFBSs act synergistically or have different functions [59]. On the other hand, neutral mutations should take place in TFBSs, too. Since slightly different motifs can be bound by the same TF, mutations resulting in sequence differences that do not alter the transcription profile should be acceptable. Such neutral substitutions causing silent changes will accumulate by drift. Non neutral mutations can

affect transcription rates, which in turn could influence fitness. That is why we might expect that specific variants with a fitness advantage are under positive selection. It is still not clear to which extent mutational rates are correlated with function. One possible scenario could be based on quantity versus quality. Promoters containing multiple binding sites should evolve relatively fast and vice versa. It might be that potential functional redundancy allows for weaker constraints.

Taking into account all these issues, we would like to gauge the impact of HIV-1 transcriptional regulation on its fitness. We want to use computational TFBS-identification methods to search for well known and potentially novel binding sites that may be involved in HIV-1 disease progression. Recovery of known sites would encourage the use of our bioinformatical tools, and the discovery of new binding sites would motivate experimentalists to check our predictions in a wet lab. Another task is to predict if viruses with similar make-up respond in similar disease dynamics. The assumption behind such study is that functional TFBS are well conserved during HIV-1 evolution, thus recurring in most of HIV-1 regulatory sequences. Our goal is to understand the evolution of the transcriptional regulation of the HIV-1, and to correlate the presence or absence of binding sites with infection dynamics.

5.2 TFBS in HIV-1 LTR

The results of this part were obtained in collaboration with Prof. Nikolaus Rajewsky (New York University) and Prof. Anne Goldfeld (Harvard Medical School).

5.2.1 HIV-1 Virus

HIV-1 virus is a retrovirus, referring to the backward flow of the genetic information i.e. DNA is synthesized using RNA as a template, which is carried out by an retroviral enzyme called *reverse transcriptase*. The life cycle of the HIV-1 virus includes the following steps (see figure 5.1):

1. Entry into a cell via proteins
   that enable the virus to bind to cells;

2. Two single-stranded RNA genomes enter a cell;
   Loss of envelope and of capsid;

3. The reverse transcriptase makes a cDNA strand
   to form a DNA/RNA hybrid double helix;

4. The RNA strand is removed;
   The reverse transcriptase (which can use either
   DNA or RNA as a template) synthesizes a complementary
   strand giving a DNA double helix;

5. This DNA double helix is inserted into a randomly selected
   site in the host genome;

6. **Activation of the HIV-1 transcription;**
   Transcription of the integrated viral DNA
   by the host cell RNA polymerase;

7. The newly produced RNA molecules are then translated by the
   host cell machinery to produce different proteins,
   building blocks of the virus;

8. Assembly of building blocks and the RNA genome
   into new virus particles; New progeny floating around;

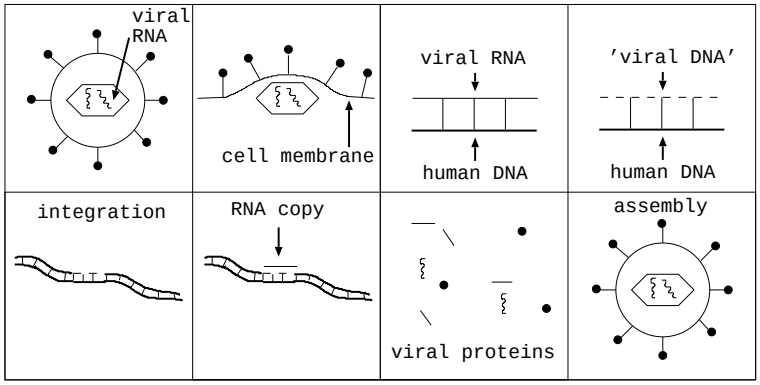

Figure 5.1: Life cycle of HIV-1. Description see text.

We are concerned with the step number 6 i.e. with transcriptional regulation of the HIV-1 virus. Transcription is controlled by a single promoter called LTR (for Long Terminal Repeats) [56, 57, 72]. LTR activates expression generating a 9-kb primary transcript that has the potential to encode all HIV-1 genes.
The crucial part is that the LTR sequence binds host TFs, playing a fifth columnist, with lethal consequences. LTR contains DNA binding sites for several *cellular* transcription factors (see Fig. 5.2 for AHAB scoring):

- NFκB
- SP1
- NFAT, AP1, ETS...
- TBP

Activation of the LTR by cellular TFs leads primarily to the generation of short transcripts. Some complete transcripts, however, are generated and allow the production of a viral protein, which then interacts with the TAR element to enhance the levels of transcription of viral RNAs. The TAR element is a short subsequence of LTR. Its 30-31 bases fold to form a stable stem-bulge-loop structure, necessary for the HIV-1 transcription (see Fig. 5.3 obtained by software MFOLD [77–79]).

As already mentioned, the HIV-1 virus packages two viral RNAs into each particle. Thus, if cells are infected by more than one viral individual, RNAs from different viruses can be packaged subsequently into offspring

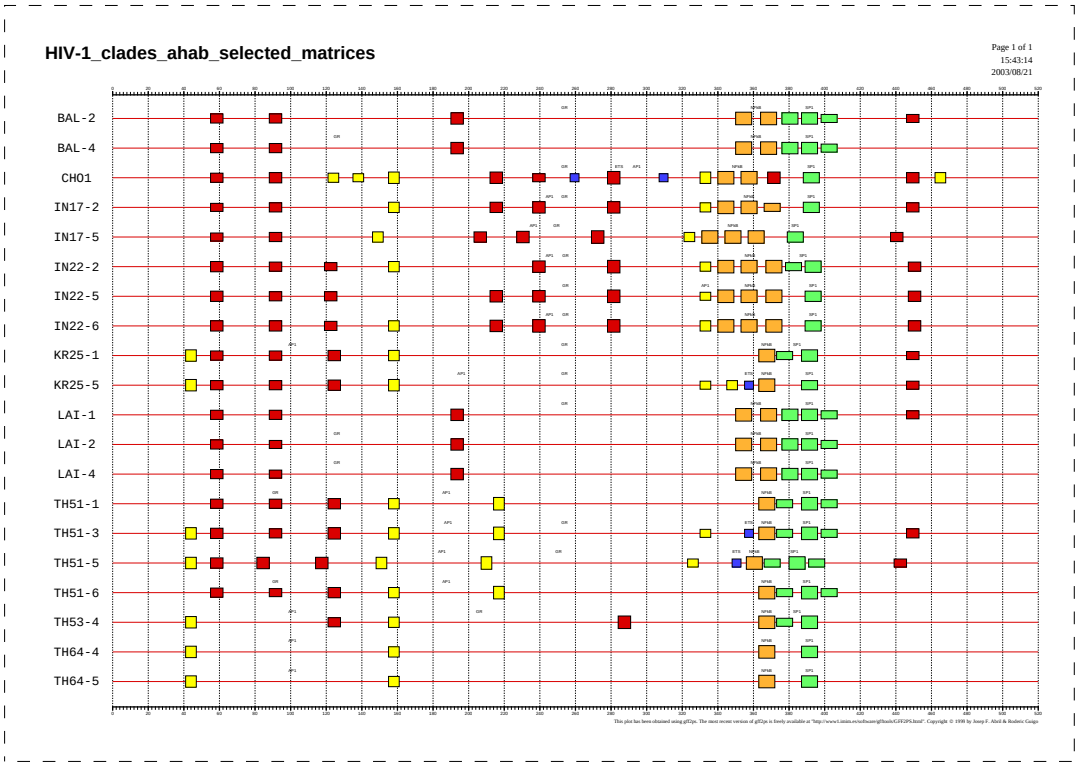

Figure 5.2: HIV-1 clades AHAB scoring for standard TFs

particles, resulting in recombination occurring subsequently during reverse transcription. Recombination provides a means of cleaning up the deleterious mutations from the viral genome. This mechanism seems to play an important role in generating the multiple drug-resistance phenotypes that occur during HIV-1 drug therapy.

5.2.2 TFBS Search in HIV-1 LTR

Patient samples. Sequences of HIV-1 U3 and TAR region of long terminal repeats (LTR) from patients having a B, a C, an E and an E/B/C recombinant HIV-1 infection. Virus annotation: TH51, E clade from Thailand, dual tropic, primary isolate; TH53, E clade from Thailand, T-tropic; TH64, E clade from Thailand, M-tropic; KR25, E/B/C recombinant clade from Cambodia, primary isolate, T-tropic; BAL, B clade from USA , lab adapted, M-tropic ; LAI, B clade from Europe, dual tropic; IN22, C clade from India,

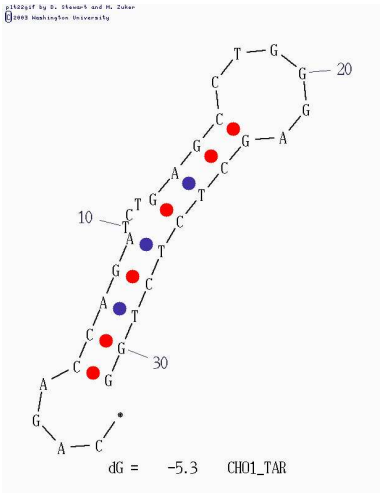

Figure 5.3: Stem-bulge-loop from CHO1 clade. Obtained by MFOLD.

primary isolate, M-tropic; IN17, C clade from India, dual tropic; CHO, C clade from China, M-tropic. Patient samples were obtained by [100].

We used 68 TF weight matrices from the Transfac set and scored the sequences with AHAB and SMASH. Our bioinformatical search for binding sites for 68 TFs produced putative sites for NFκB, SP1, NFAT, GR, AP1, MYB, E2F, ELK1, ATF6, with the criterion: at least one site in at least one clade. This list contains all factors known to bind in the LTR and predicts only one additional factor Evi-1 (Evi-1 is annotated as ectopic viral integration site 1 encoded factor). Interestingly, we see evidence for changes in binding-site composition among clades with the exception of NFκB and SP1 which are always present (and always in the same position). Detected binding sites are shown in Figs. 5.2, 5.4 and 5.5.

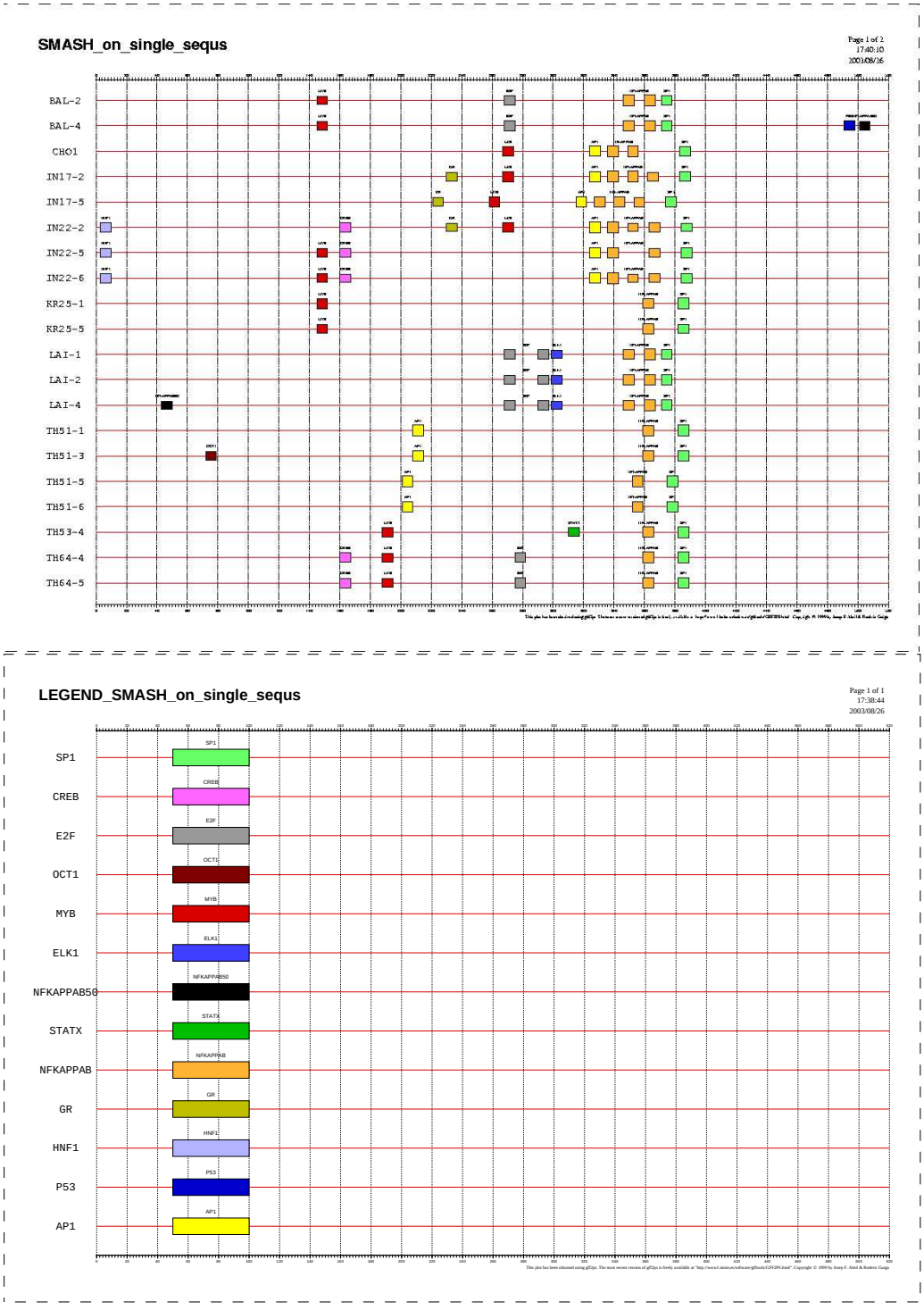

Figure 5.4: SMASH on single sequs, 30 TFs; lower: legend

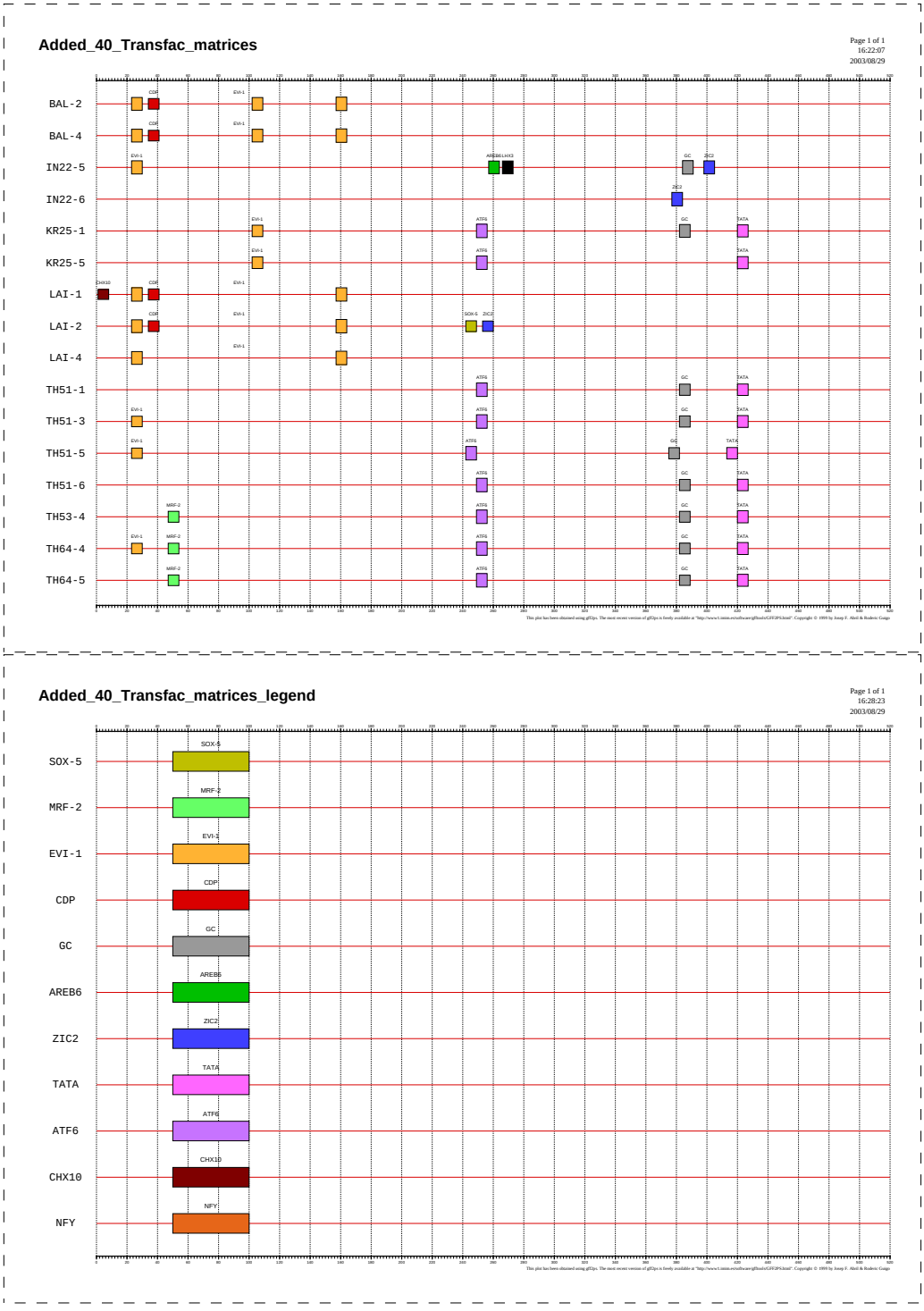

Figure 5.5: SMASH on single sequences, another 38 TFs; lower: legend

Chapter 6

HIV-1 Promoter Mutations

6.1 HIV-1 LTR Evolution

6.1.1 HIV-1 LTR Evolution

The HIV-1 core promoter (see Fig. 6.1), located in the long terminal repeat (LTR), is one of the most conserved parts of the HIV-1 genome. There is evidence that the major impact on HIV-1 replication is performed not by the whole LTR, but its part in front of the transcription start site (its core promoter). LTR region is about 500 bases long, while the core promoter represents some 160-base-long subsequence just upstream of the transcription start site. High conservation points at functional importance for transcrip-

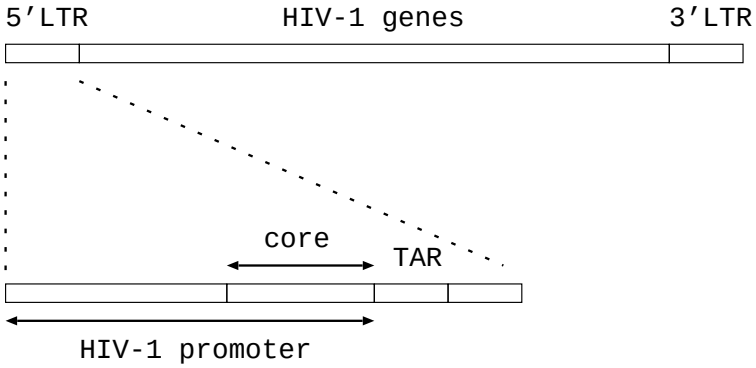

Figure 6.1: Sketch of HIV-1 genome and close-up on promoter region

tion. HIV-1 transcription is controlled by various TFBSs, and this variability is subtype-specific [56, 57, 73]. Up to date, at least nine subtypes are known and they seem to be of different geographical origin. The interesting part is that all subtypes require presence of SP1 and NFκB sites for basal and induced transcription. Other TFBSs, such as NFAT, AP1, GR, ETS have been detected, however, in a more irregular fashion than the essential sites. Distinct disease progression has been reported for different subtypes and it has been suggested that the reason for this behavior lies in the interaction between the virus subtype-specific genotype and the host environment. Experiments have shown that the core promoter composition has a strong influence on the viral fitness. For example, a C subtype with a high NFκB binding site load performs much better in NFκB rich environments than an E subtype, with weaker NFκB content. It seems that a mutation can have both positive and negative effect on fitness, depending on cellular environment in which the virus resides. Therefore, we have to focus on subtype-specific analysis of the data set from our collaborators' laboratory [100] in order to bring

A	C	G	T
-	-	-	-
1	2	1	32
0	3	0	33
0	4	0	32
0	35	0	1
1	33	0	2

Table 6.1: NFAT count matrix

A	C	G	T
-	-	-	-
0	0	40	0
0	0	40	0
1	0	39	0
15	5	15	5
3	17	3	17
3	16	3	18
1	2	2	35
1	2	2	35

Table 6.2: NFκB count matrix

correct conclusions about the HIV-1 LTR evolutionary mechanisms.

Position Weight Matrices

We used frequency matrices from the Transfac database in order to construct PWMs. Most of the matrices describe essential HIV-1 LTR TF motifs of length $\geq$ 10 positions. However, tails (i.e. left and right ends) are not well conserved. Therefore we focus on matrix cores, in other words, sharply defined part of the motif (see Tables 6.1,6.2,6.3). The NFAT-5-base consensus has been suggested by [101].

6.1.2 Phylogeny

Already at the dawn of evolutionary biology, scientists have tried to assign known organisms to correct positions in the 'tree of life'. Early classifications

A	C	G	T
-	-	-	-
8	103	1	0
2	102	5	3
10	0	83	19
0	110	2	0
0	112	0	0
1	93	1	17

Table 6.3: SP1 count matrix

were based on phenotypic features which, as we know today, cannot differentiate organisms as precisely as their genetic content can do. Nowadays, with the immense development of the genetic build up knowledge, evolutionary trees are constructed based on sequence similarity, more precisely, on the basis of a multiple sequence alignment [83–85].

Evolutionary relationship among sequences is usually shown by an evolutionary tree, which is simply a two dimensional graph. If we are ignorant of the oldest ancestry, sequences and their relationships are presented by an unrooted tree. An unrooted tree does not specify the direction of descent, still it shows how close pairs of sequences are. Fig. 6.2 is an example of an unrooted tree. Sequences under investigation are placed at leaf nodes (outer nodes) while their evolutionary connections are represented by branches and ancestral nodes. Ancestral nodes are imaginary sequences, that are thought to preceed real-life sequences in the course of time, and they can be constructed using different models (maximum likelihood, maximum parsimony and distance method). Branch lengths between two nodes gauge the amount of changes between the nodes.

The ultimate goal of phylogenetic analysis is to determine evolutionary relationship among sequences, together with branch lengths. Phylogenetic analysis is especially useful in following the alterations in rapidly changing species, like viruses. Analysis of a virus promoter can reveal which TFBSs are under selection.

First step of the phylogenetic analysis is the construction of a multiple sequence alignment. Each base position (column) is assumed to evolve in-

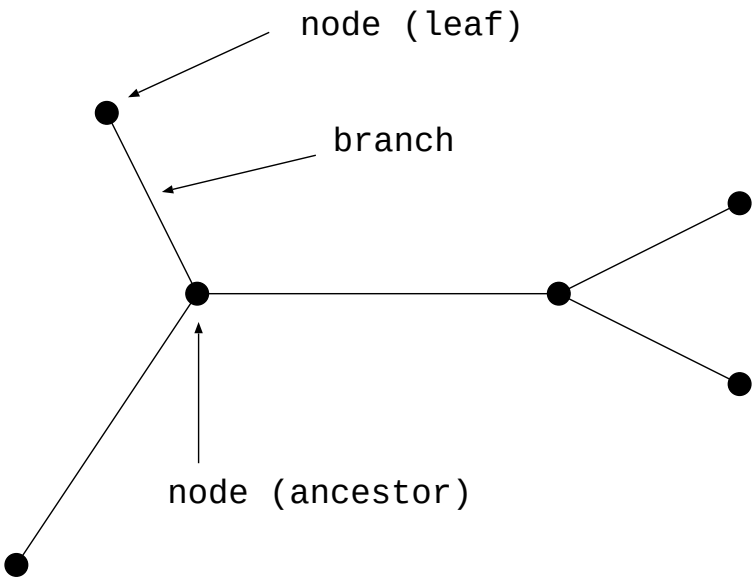

Figure 6.2: An Unrooted Tree

dependently of others. Depending on the variability of sequences different methods are used, but the maximum likelihood method is regarded as most general, and may be used for any set of related sequences.

The Maximum Likelihood Method

The maximum likelihood method is a probabilistic approach in search for a tree based on the variation among sequences from a data set [82, 83]. Mutations in each column of a multiple sequence alignment are considered in order to construct a tree. Each tree has a certain likelihood which is computed from the product of the mutation rates in each branch of the tree, which in turn is the product of the substitution rate in each branch and the branch length. The tree with the highest likelihood is then identified. Fig 6.3 depicts a maximum likelihood scenario for four sequences labeled 1, 2, 3, 4. Suppose we have aligned those four sequences and wish to estimate their phylogenetic tree. There are three possible unrooted trees $t = 1, 2, 3$. We focus on one of them. Then, we detect base changes along one column k in the alignment. Since we are interested in the evolutionary course of changes we draw a rooted tree j with the root S. Each node has its own probability $p_1, p_2, p_3, p_4, p_S, p_I, p_{II}$. There are 64 combinations (4 bases can turn into 4 bases) for one column k, so that the probability of the j^{th} tree is:

$$P_j = {p_1}^j {p_2}^j {p_3}^j {p_4}^j {p_S}^j {p_I}^j {p_{II}}^j \tag{6.1}$$

The likelihood of the tree is then:

$$L_k = \sum_{j=1}^{64} P_j \tag{6.2}$$

This routine is repeated for all columns k in the multiple alignment. Each of the three $t = 1, 2, 3$ possible unrooted trees is assigned a likelihood:

$$Ł_t = \sum_k L_k, \tag{6.3}$$

and the tree with the highest value is chosen as the representative of evolutionary relationships among the four sequences.

Bases of extinct ancestors (or simply ancestral node sequences) can be reconstructed by fitting known sequences (1,2,3,4). The maximum likelihood approach uses branch lengths and the substitution pattern for this purpose, yielding ancestral sequences (I,II,S).

6.1.3 Alignments

Our data set comprises 20 LTR HIV-1 primary clades from the Goldfeld Lab. Using ClustalW (with default parameters) we aligned all clades from the Goldfeld data set. Multiple alignment of the sequences reveals that mutational mechanism favors *substitutions* over *indels*. Based on the alignment of the 20 sequences, we construct a maximum-likelihood phylogenetic tree. For this purpose we use fastDNAml software [74, 75], a program that starts with an evolutionary model of sequence change that provides estimates of rates of substitution of one base for another in a set of nucleic acid sequences. In the end, the most likely tree given the data is then identified (see Fig. 6.14). In order to gain better statistics we decided to reconstruct ancestral nodes (artificially constructed sequences at internal nodes of the unrooted fastDNAml tree). We used software by Ziheng Yang PAML *Phylogenetic Analysis by Maximum Likelihood* [81]. Input includes HIV-1 LTR sequences and its tree file. In this fashion we obtained 18 more sequences, which are labeled as 'nodes'. We scored the enhanced set of 38 sequences in search for NFAT, NFkB and SP1 binding sites (see Fig. 6.13). Detected motifs are marked in the shown alignment (see Figs. 6.4 to 6.12).

6.1.4 Phylogeny of HIV-1 promoter sequences

HIV-1 is a virus with many different subtypes, each subtype being prevalent in a distinct part of the world. Goldfeld data set comprises sequences of HIV-1 promoter and TAR region of long terminal repeats (LTR) from patients having a B, a C, and an E HIV-1 infection.
Based on tree branch lengths and central position of the "S" (root) node, we could classify ancestral nodes into subtypes (See Fig. 6.14 and table 6.4). Crucial branch is defined as a branch linking two subtypes.

6.1.5 Mutation Profile

Entropy

Information theory is closely connected to probabilistic modeling [85]. Given a random variable "X" with probabilities $p(l_i)$ for discrete set of K events (here K=4 i.e. $l_1 = A, l_2 = C, l_3 = G, l_4 = T$) the Shannon entropy is defined by:

$$H(X) = -\sum_i^K p(l_i) \log p(l_i) \tag{6.4}$$

Since true distributions are usually not known exactly, entropies are calculated from the frequencies of events. The entropy is maximized when all events are equally probable $P(l_i) = \frac{1}{K}$. ($H_{max} = \log K$).

If we know for sure the outcome of a sample from the distribution i.e. $p(l_k) = 1$ for one k and the other $p(l_i) = 0, i \neq k$, the entropy is zero.

Relative Entropy

Relative entropy of two distributions p and q is defined as:

$$S_{rel}(p||q) = \sum_i p(l_i) \log \left(\frac{p(l_i)}{q(l_i)} \right) \tag{6.5}$$

The relative entropy of two probability distributions measures in a way the dissimilarity between them [86]. Relative entropy is sometimes called the distance between the two distributions. However, since $S_{rel}(p||q)$ is not equal to $S_{rel}(q||p)$, this is not appropriate, since relative entropy does not satisfy a basic requirement of a distance measure, $S_{rel}(p||q) = S_{rel}(q||p)$.

node	node	branch length
B1	B13	0.00389
B2	B13	0.00388
B13	B12	0.01372
B3	B12	0.00388
B12	B11	0.00000
B4	B11	0.00194
B11	B10	0.00194
B5	B10	0.00000
B10	**S**	0.12313
E1	E12	0.00194
E2	E13	0.00000
E3	E13	0.00584
E13	E12	0.00389
E12	E11	0.00195
E4	E11	0.00000
E11	E10	0.01878
E5	E16	0.00390
E6	E16	0.00193
E16	E15	0.00985
E7	E15	0.01990
E15	E14	0.00180
E8	E17	0.00584
E9	E17	0.00784
E17	E14	0.01989
E14	E10	0.00511
E10	**S**	0.17749
Cb1	Cb12	0.00376
Cb2	Cb12	0.02184
Cb12	Cb11	0.01094
Cb3	Cb11	0.01312
Cb11	**Ca11**	0.31601
Ca1	Ca12	0.00592
Ca2	Ca12	0.00000
Ca12	Ca11	0.00844
Ca11	Ca10	0.00654
Ca3	Ca10	0.01255
Ca10	**S**	0.14805

Table 6.4: ML-tree Branch Lengths

node	node	branch length
Cb11	**Ca11**	0.31601
Ca10	**S**	0.14805
B10	**S**	0.12313
E10	**S**	0.17749

Table 6.5: Crucial Branch Lengths

Position Activity and Entropy

In Fig. 6.15 (upper part) we display mutational activity of each position in the enhanced data set (along the tree) i.e. number of mutations as a function of position. Bars denote the core promoter. Lower part of Fig. 6.15 shows smoothed version of mutational activity. Here we slide a window of 10 positions, and compute the average number of mutations per window.

In Fig. 6.16 (upper) we display entropy of each position in the enhanced data-set-alignment. The lower part of Fig. 6.16 shows smoothed version of position entropy. As before, we slide a window of 10 positions, and compute the average entropy per window.
Spikes denote borders of the core-promoter region (loaded with TFBS).

As we can see, analysis of HIV-1 LTRs without paying attention at subtype-specificity cannot reveal much information on TFBS conservation. On the contrary, entropy analysis of the multiple alignment of the whole data set, points at the core promoter as the most variable part of the LTR sequence. This is misleading, because the major mutational movement takes place at subtype-subtype transitions. Within a subtype, the core promoter is the most conserved part of the LTR.

Tree Analysis of Mutations

After extracting mutationally active positions in the sequences based on the the PAML output file, we counted how many mutations occurred in each of the TFBS and background. Thus we were able to determine *average number of mutations per 100 bases* $< x >$ for each TF, as well as *mutation rate ratios* $r = \frac{<x>_{TF}}{<x>_{BKGD}}$ along branches of the tree (see table 6.6). Assuming that background obeys mechanisms of neutral evolution, ratios r can give us a good estimate of how selective mutations along particular branches are.

	NFAT	NFkB	SP1
Branch	$r \pm \Delta r$	$r \pm \Delta r$	$r \pm \Delta r$
B1-B13	0 +/- 0	0 +/- 0	0 +/- 0
B2-B13	* +/- 0	* +/- 0	* +/-0
B13-B12	2 +/- 1.5	2.7 +/- 1.7	1.9 +/- 1.5
B3-B12	* +/- 0	* +/- 0	* +/- 0
B12-B11	* +/- 0	* +/- 0	* +/- 0
B4-B11	* +/- 0	* +/- 0	* +/- 0
B11-B10	* +/- 0	* +/- 0	* +/- 0
B5-B10	* +/- 0	* +/- 0	* +/- 0
B10-S	1.7 +/- 0.5	0.7 +/- 0.3	0.6 +/- 0.3
E1-E12	* +/- 0	* +/- 0	* +/- 0
E2-E13	* +/- 0	* +/- 0	* +/- 0
E3-E13	* +/- 0	* +/- 0	* +/- 0
E13-E12	* +/- 0	* +/- 0	* +/- 0
E12-E11	* +/- 0	* +/- 0	* +/- 0
E4-E11	* +/- 0	* +/- 0	* +/- 0
E11-E10	0 +/- 0	0.5 +/- 0.5	0.7 +/- 0.6
E5-E16	6.0 +/- 4.5	4.0 +/- 3.7	0 +/- 0
E6-E16	* +/- 0	* +/- 0	* +/- 0
E16-E15	4.0 +/- 2.1	0 +/- 0	0 +/- 0
E7-E15	0.7 +/- 0.6	1.0 +/- 0.6	0 +/- 0
E15-E14	* +/- 0	* +/- 0	* +/- 0
E8-E17	* +/- 0	* +/- 0	* +/- 0
E9-E17	2 +/- 1.5	1.3 +/- 1.2	0 +/- 0
E17-E14	0 +/- 0	2.6 +/- 1.2	1.0 +/- 0.7
E14-E10	* +/- 0	* +/- 0	* +/- 0
E10-S	2.3 +/- 0.5	2.1 +/- 0.4	1.7 +/- 0.4
Cb1-Cb12	* +/- 0	* +/- 0	* +/- 0
Cb2-Cb12	0 +/- 0	2.0 +/- 1.1	2.9 +/- 1.3
Cb12-Cb11	* +/- 0	* +/- 0	* +/- 0
Cb3-Cb11	1.2 +/- 0.9	0 +/- 0	0 +/- 0
Cb11-Ca11	1.7 +/- 0.3	2.7 +/- 0.4	3.6 +/- 0.5
Ca1-Ca12	3.0 +/- 2.3	2.0 +/- 1.8	2.9 +/- 2.2
Ca2-Ca12	* +/- 0	* +/- 0	* +/- 0
Ca12-Ca11	3.0 +/- 2.3	6.0 +/- 3.2	2.9 +/- 2.2
Ca11-Ca10	* +/- 0	* +/- 0	* +/- 0
Ca3-Ca10	0 +/- 0	0 +/- 0	1.9 +/- 1.5
Ca10-S	7.6 +/- 1.1	7.7 +/- 1.1	6.3 +/- 1.0

Table 6.6: Mutations on individual branches

	NFAT	NFkB	SP1
Branch	$r \pm \Delta r$	$r \pm \Delta r$	$r \pm \Delta r$
Cb11-Ca11	1.7 +/- 0.3	2.7 +/- 0.4	3.6 +/- 0.5
Ca10-S	7.6 +/- 1.1	7.7 +/- 1.1	6.3 +/- 1.0
B10-S	1.7 +/- 0.5	0.7 +/- 0.3	0.6 +/- 0.3
E10-S	2.3 +/- 0.5	2.1 +/- 0.4	1.7 +/- 0.4

Table 6.7: Mutations on crucial branches

As we can see, significant values of ratios are along crucial branches, connecting different subtypes (see table 6.7). The C subtype has evolved into two subtypes, that we call Ca and Cb. At first glance, looking only at the maximum likelihood tree, it appears that a similar movement takes place in the B subtype, namely that it got split in two distinct subtypes. Careful examination of mutation rate ratios made us conclude that we are dealing with noise, and that there are no new subtypes emerging from the B clade.

Neutral Background

The mutations outside binding sites follow an overall molecular clock. The figure shows the histogram of the mutations per base (summed over all branches of the tree), which are approximately Poisson-distributed, with $< x >= 0.6199$.

The Poisson Distribution formula

$$P_0 = \exp(- < x >) \left[\frac{< x >^x}{x!} \right] \tag{6.6}$$

works for an infinite number of trials. Even if a set of events obeys Poisson distribution P_0 it may show certain discrepancies (within error bars) due to finite-size effects.

By drawing numbers from the Poisson distribution finite number of times, we simulate finite size effects. In order to determine a significance test for finite size effects we iterate this procedure N=1000 times. So each sample has a distribution P_{sample}, and its relative entropy $< S_{rel}^{sim} >$.

Therefore,

$$S_{rel}^{sim}(P_{sample}||P_0) = \sum_x P_{sample}(x) \log \frac{P_{sample}(x)}{P_0(x)} \tag{6.7}$$

and

$$< S_{rel}^{sim}(P_{sample}||P_0) >= \frac{1}{N} \sum\nolimits_i^N {S_{rel}}^i (P_{sample}||P_0) \tag{6.8}$$

$$< S_{rel}^{sim} >= 0.1274 \pm 0.005. \tag{6.9}$$

To answer the question how far is the data distribution from the exact Poisson-distribution we compute relative entropy:

$$S_{rel}(P_{measured}||P_0) = 0.1271. \tag{6.10}$$

We conclude that background indeed has Poisson-distributed mutations (see Fig. 6.17).

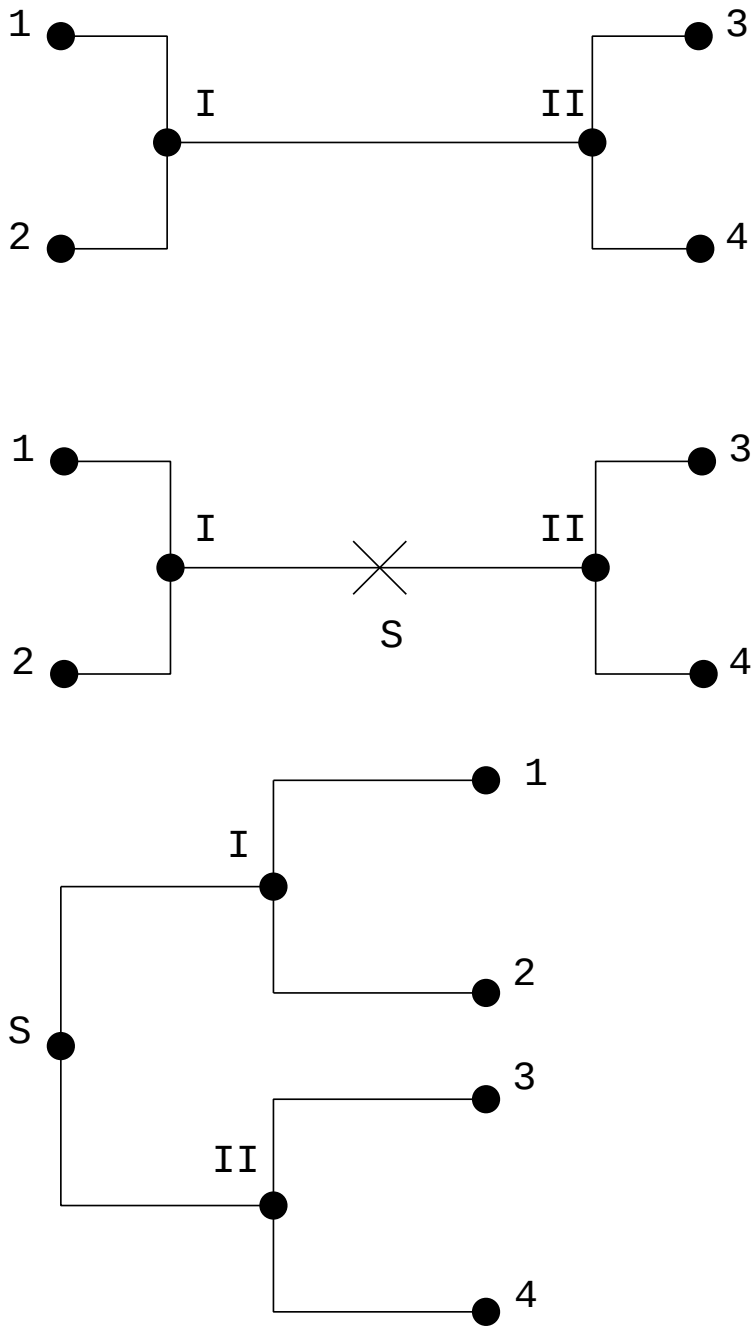

Figure 6.3: Maximum Likelihood Tree Construction

```
bal2          TGGAAGGGCTAATTCACTCCCAACGAAGACAAGATATCCTTGATCTGTGGATCTACCACA 60
bal4          TGGAAGGGCTAATTCACTCCCAACGAAGACAAGATATCCTTGATCTGTGGATCTACCACA 60
node21        TGGAAGGGCTAATTCACTCCCAACGAAGACAAGATATCCTTGATCTGTGGATCTACCACA 60
lai4          TGGAAGGGCTAATTCACTCCCAACGAAGACAAGATATCCTTGATCTGTGGATCCACCACA 60
node22        TGGAAGGGCTAATTCACTCCCAACGAAGACAAGATATCCTTGATCTGTGGATCTACCACA 60
lai2          TGGAAGGGCTAATTCACTCCCAACGAAGACAAGATATCCTTGATCTGTGGATCTACCACA 60
node23        TGGAAGGGCTAATTCACTCCCAACGAAGACAAGATATCCTTGATCTGTGGATCTACCACA 60
lai1          TGGAAGGGCTAATTCACTCCCAAAGAAGACAAGATATCCTTGATCTGTGGATCTACCACA 60
node24        TGGAAGGGCTAATTCACTCCCAAAGAAGACAAGATATCCTTGATCTGTGGATCTACCACA 60
th516         TGGAAGGGCTAATTTACTCCAAGAAAAGACAAGAGATCCTTGATTTATGGGTCTATAACA 60
th515         TGGAAGGGCTAATTTACTCCAAGAAAAGACAAGAGATCCTTGACTTATGGGTCTACAACA 60
th513         TGGAAGGGCTAATTTACTCCAAGAAAAGACAAGAGATCCTTGACTTATGGGTCTACAACA 60
node29        TGGAAGGGCTAATTTACTCCAAGAAAAGACAAGAGATCCTTGACTTATGGGTCTACAACA 60
node28        TGGAAGGGCTAATTTACTCCAAGAAAAGACAAGAGATCCTTGATTTATGGGTCTATAACA 60
th511         TGGAAGGGCTAATTTACTCCAAGAAAAGACAAGAGATCCTTGATTTATGGGTCTATAACA 60
node27        TGGAAGGGCTAATTTACTCCAAGAAAAGACAAGAGATCCTTGATTTATGGGTCTATAACA 60
th645         TGGATGGGCTAATTTACTCCAAGAGAAGACAAGAGGTCCTTGACTTATGGGTCTATAATA 60
th644         TGGATGGGCTAATTTACTCCAAGAGAAGACAAGAGATCCTTGACTTATGGGTCTATAATA 60
node33        TGGATGGGCTAATTTACTCCAAGAGAAGACAAGAGATCCTTGACTTATGGGTCTATAATA 60
th534         TGGATGGGCTAATTTACTCCAAGAGAAGACGAGAGATCCTTGACTTATGGGTCTATAATA 60
node32        TGGATGGGCTAATTTACTCCAAGAGAAGACAAGAGATCCTTGACTTATGGGTCTATAATA 60
kr255         TGGATGGGCTAATTCACTCCAAGAGAAGACAAGACATCCTTGACTTATGGGTCTATAACA 60
kr251         TGGATGGGCTAATTCACTCCAAGAGAAGACAAGACATCCTTGACTTATGGGTCTATAACA 60
node31        TGGATGGGCTAATTCACTCCAAGAGAAGACAAGACATCCTTGACTTATGGGTCTATAACA 60
node30        TGGATGGGCTAATTTACTCCAAGAGAAGACAAGAGATCCTTGACTTATGGGTCTATAACA 60
node26        TGGAAGGGCTAATTTACTCCAAGAGAAGACAAGAGATCCTTGATTTATGGGTCTATAACA 60
in226         TGGAAGGGTTAATTTACTCCAAGAAAAGGCAAGAGATCCTTGATTTGTGGGTCTACCACA 60
in225         TGGAAGGGTTAATTTACTCCAAGAAAAGACAAGAGATCCTTGATTTGTGGGTCTACCACA 60
node37        TGGAAGGGTTAATTTACTCCAAGAAAAGGCAAGAGATCCTTGATTTGTGGGTCTACCACA 60
in222         TGGAAGGGTTAATTTACTCCAAGAAAAGGCAAGAGATCCTTGATTTATGGGTCTACCACA 60
node36        TGGAAGGGTTAATTTACTCCAAGAAAAGGCAAGAGATCCTTGATTTGTGGGTCTACCACA 60
in175         TGGAAGGGTTAATTTACTCTAAGAAAAGGCAAGAGATCCTTGATTTGTGGGTCCATCACA 60
in172         TGGAAGGGTTAATTTACTCTAAGAAAAGGCAAGAGATCCTTGATTTGTGGGTCTATCACA 60
node38        TGGAAGGGTTAATTTACTCTAAGAAAAGGCAAGAGATCCTTGATTTGTGGGTCTATCACA 60
node35        TGGAAGGGTTAATTTACTCTAAGAAAAGGCAAGAGATCCTTGATTTGTGGGTCTATCACA 60
cho1          TGGAAGGGTTAATTTACTCTAAGAAAAGGCAAGAGATCCTTGATTTGTGGGTCTATCACA 60
node34        TGGAAGGGTTAATTTACTCTAAGAAAAGGCAAGAGATCCTTGATTTGTGGGTCTATCACA 60
node25        TGGAAGGGCTAATTTACTCCAAGAGAAGACAAGAGATCCTTGATTTGTGGGTCTATCACA 60
              **** *** ***** ****  *   *** * ***  *******  * *** ** *  * *
```

Figure 6.4: Multiple alignment of HIV-1 promoters, part1

```
bal2          CACAAGGCTACTTCCCTGATTGGACACACCAGGGCCAGGGATCAGATATCCACTGACCTT 120
bal4          CACAAGGCTACTTCCCTGATTGGACACACCAGGGCCAGGGATCAGATATCCACTGACCTT 120
node21        CACAAGGCTACTTCCCTGATTGGACACACCAGGGCCAGGGATCAGATATCCACTGACCTT 120
lai4          CACAAGGCTACTTCCCTGATTGGACACACCAGGGCCAGGGGTCAGATATCCACTGACCTT 120
node22        CACAAGGCTACTTCCCTGATTGGACACACCAGGGCCAGGGGTCAGATATCCACTGACCTT 120
lai2          CACAAGGCTACTTCCCTGATTGGACACACCAGGGCCAGGGGTCAGATATCCACTGACCTT 120
node23        CACAAGGCTACTTCCCTGATTGGACACACCAGGGCCAGGGGTCAGATATCCACTGACCTT 120
lai1          CACAAGGCTACTTCCCTGATTGGACACACCAGGGCCAGGGGTCAGATATCCACTGACCTT 120
node24        CACAAGGCTACTTCCCTGATTGGACACACCAGGGCCAGGGGTCAGATATCCACTGACCTT 120
th516         CACAAGGCTTCTTCCCTGATTGGACACACCAGGGCCAGGGACCAGATACCCACTGTGTTT 120
th515         CACAAGGCTTCTTCCCTGATTGGACACACCAGGGCCAGGGACCAGATACCCACTGTGTTT 120
th513         CACAAGGCTTCTTCCCTGATTGGACACACCAGGGCCAGGGACCAGATACCCACTGTGTTT 120
node29        CACAAGGCTTCTTCCCTGATTGGACACACCAGGGCCAGGGACCAGATACCCACTGTGTTT 120
node28        CACAAGGCTTCTTCCCTGATTGGACACACCAGGGCCAGGGACCAGATACCCACTGTGTTT 120
th511         CACAAGGCTTCTTCCCTGATTGGACACACCAGGGCCAGGGACCAGATACCCACTGTGTTT 120
node27        CACAAGGCTTCTTCCCTGATTGGACACACCAGGGCCAGGGACCAGATACCCACTGTGTTT 120
th645         CACAAGGCTTCTTCCCTGATTGGACACACCAGGACCAGGGATCAGATACCCACTATGTTT 120
th644         CACAAGGCTTCTTCCCTGATTGGACACACCAGGACCAGGGATCAGATACCCACTATGTTT 120
node33        CACAAGGCTTCTTCCCTGATTGGACACACCAGGACCAGGGATCAGATACCCACTATGTTT 120
th534         CACAAGGCTTCTTCCCTGATTGGACATACCAGGGCCAGGGATCAGATTCCCACTGTGTTT 120
node32        CACAAGGCTTCTTCCCTGATTGGACACACCAGGGCCAGGGATCAGATACCCACTGTGTTT 120
kr255         CACAAGGCTACTTCCCTGATTGGACACACCAGGGCCAGGGATCAGATATCCACTGTGTTT 120
kr251         CACAAGGCTACTTCCCTGATTGGACACACCAGGACCAGGGATCAGATATCCACTGTGTTT 120
node31        CACAAGGCTACTTCCCTGATTGGACACACCAGGGCCAGGGATCAGATATCCACTGTGTTT 120
node30        CACAAGGCTTCTTCCCTGATTGGACACACCAGGGCCAGGGATCAGATACCCACTGTGTTT 120
node26        CACAAGGCTTCTTCCCTGATTGGACACACCAGGGCCAGGGATCAGATACCCACTGTGTTT 120
in226         CACAAGGCTACTTCCCTGACTGGACACACCAGGACCAGGGGTCAGATACCCACTGACCTT 120
in225         CACAAGGCTACTTCCCTGACTGGACACACCAGGACCAGGGGTCAGATACCCACTGACCTT 120
node37        CACAAGGCTACTTCCCTGACTGGACACACCAGGACCAGGGGTCAGATACCCACTGACCTT 120
in222         CACAAGGCTACTTCCCTGATTGGACACACCAGGACCAGGGGTCAGATACCCACTGACCTT 120
node36        CACAAGGCTACTTCCCTGATTGGACACACCAGGACCAGGGGTCAGATACCCACTGACCTT 120
in175         CACAAGGCTACTTCCCTGATTGGACACACCAGGACCAGGGACCAGATTCCCACTGACTTT 120
in172         CACAAGGCTACTTCCCTGATTGGACACACCAGGACCAGGGACCAGATTCCCACTGACTTT 120
node38        CACAAGGCTACTTCCCTGATTGGACACACCAGGACCAGGGACCAGATTCCCACTGACTTT 120
node35        CACAAGGCTACTTCCCTGATTGGACACACCAGGACCAGGGGTCAGATTCCCACTGACTTT 120
cho1          CACAAGGCTACTTCCCTGATTGGACACACCAGGACCAGGGGTCAGATTCCCACTGACTTT 120
node34        CACAAGGCTACTTCCCTGATTGGACACACCAGGACCAGGGGTCAGATTCCCACTGACTTT 120
node25        CACAAGGCTACTTCCCTGATTGGACACACCAGGGCCAGGGGTCAGATACCCACTGACTTT 120
              ********* ********* ****** ****** ******  *****  *****    **
                        |  |                  |     | |   |
                       NFAT I                 NFkB I  NFAT II
```

Figure 6.5: Multiple alignment of HIV-1 promoters, part2

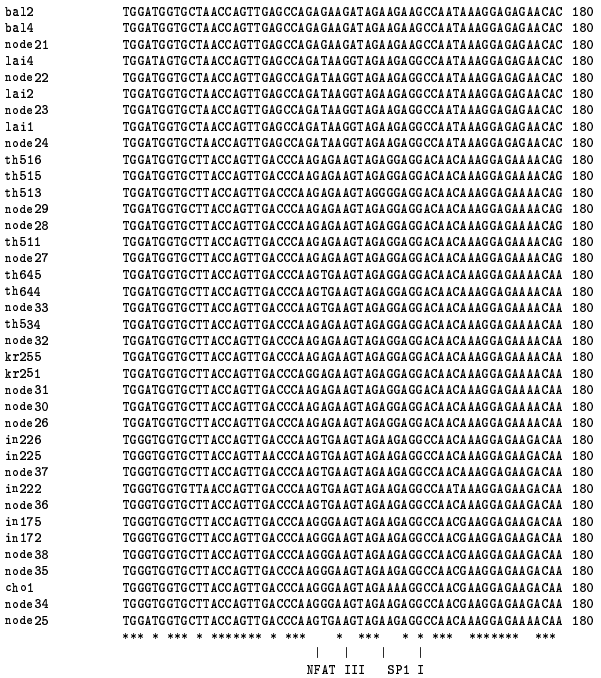

Figure 6.6: Multiple alignment of HIV-1 promoters, part3

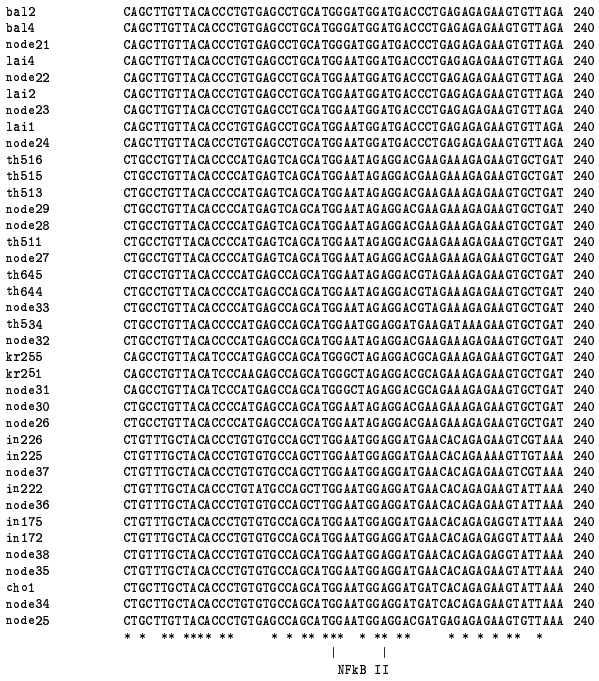

Figure 6.7: Multiple alignment of HIV-1 promoters, part4

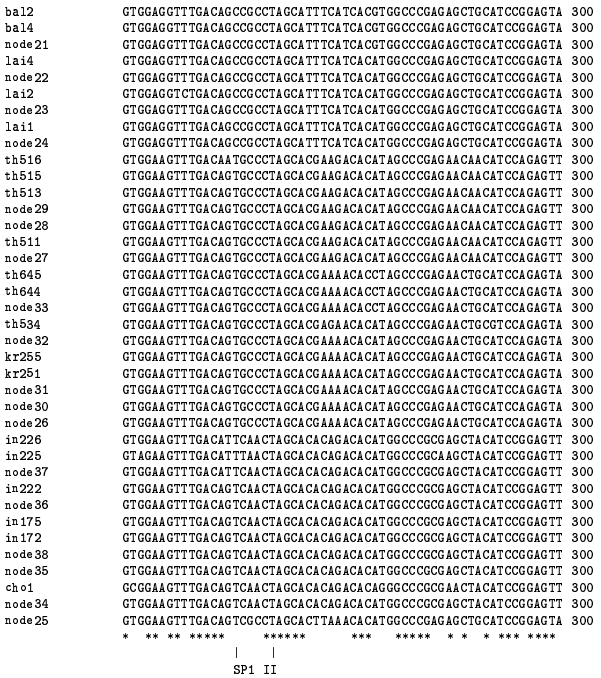

Figure 6.8: Multiple alignment of HIV-1 promoters, part5

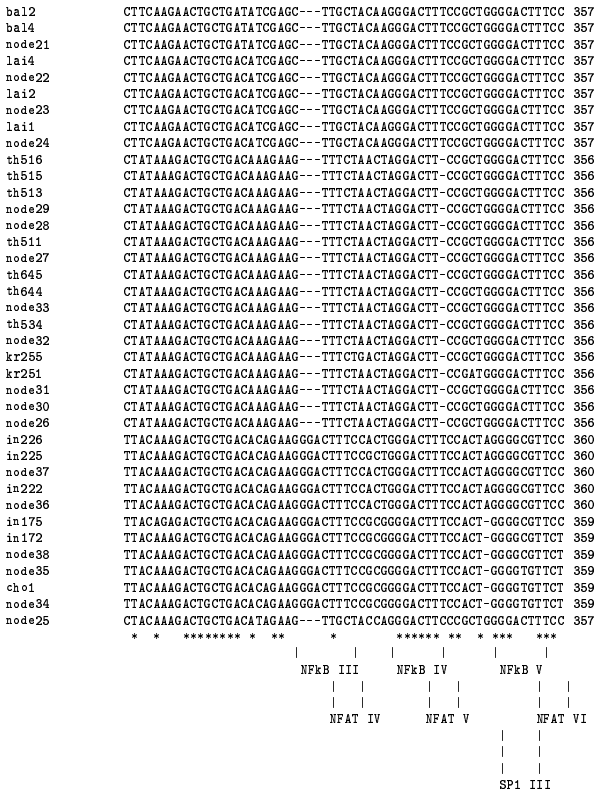

Figure 6.9: Multiple alignment of HIV-1 promoters, part6

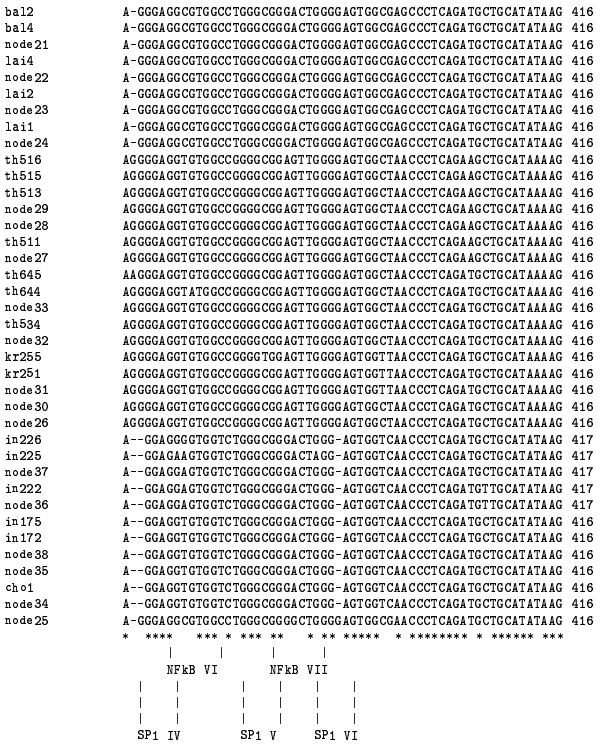

Figure 6.10: Multiple alignment of HIV-1 promoters, part7

```
bal2          CAGCTGCTTTTTGCCTGTGCTGGGTCTCTCTGGTTAGACCAGATTTGAGCCTGGGAGCTC 476
bal4          CAGCTGCTTTTTACCTGTACTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTC 476
node21        CAGCTGCTTTTTGCCTGTACTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTC 476
lai4          CAGCTGCTTTTTGCCTGTACTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTC 476
node22        CAGCTGCTTTTTGCCTGTACTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTC 476
lai2          CAGCTGCTTTTTGCCTGTACTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTC 476
node23        CAGCTGCTTTTTGCCTGTACTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTC 476
lai1          CAGCTGCTTTTTGCCTGTACTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTC 476
node24        CAGCTGCTTTTTGCCTGTACTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTC 476
th516         CAGCCGCTTCTCGCTTGTACTGGGTCTCTCTTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
th515         CAGCCGCTTCTCGCTTGTACTGGGTCTCTCTTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
th513         CAGCCGCTTCTCGCTTGTACTGGGTCTCTCCTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
node29        CAGCCGCTTCTCGCTTGTACTGGGTCTCTCTTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
node28        CAGCCGCTTCTCGCTTGTACTGGGTCTCTCTTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
th511         CAGCCGCTTCTCGCTTGTACTGGGTCTCTCTTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
node27        CAGCCGCTTCTCGCTTGTACTGGGTCTCTCTTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
th645         CAGCCGCTTTTCGCTTGTACTGGGTCTCTCTTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
th644         CAGCCGCTTTTCGCTTGTACTGGGTCTCTCTTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
node33        CAGCCGCTTTTCGCTTGTACTGGGTCTCTCTTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
th534         CAGCCGCTTTTCGCTTGTACTGGGTCTCCCTTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
node32        CAGCCGCTTTTCGCTTGTACTGGGTCTCTCTTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
kr255         CAGCCGCTTTTCGCTTGTACTGGGTCTCTCTTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
kr251         CAGCCGCTTTTCGCTTGTACTGGGTCTCTCTTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
node31        CAGCCGCTTTTCGCTTGTACTGGGTCTCTCTTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
node30        CAGCCGCTTTTCGCTTGTACTGGGTCTCTCTTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
node26        CAGCCGCTTTTCGCTTGTACTGGGTCTCTCTTGTTAGACCAGGTC-GAGCCCGGGAGCTC 475
in226         CAGCTGCTTTTCGCCTGTGCTGGGTCTCTCTTGGTAGACCAGATCTGAGCCTGGGAGCTC 477
in225         CAGCTGCTTTTCGCCTGTGCTGGGTCTCTCTTGGTAGACCAGATCTGAGCCTGGGAGCTC 477
node37        CAGCTGCTTTTCGCCTGTGCTGGGTCTCTCTTGGTAGACCAGATCTGAGCCTGGGAGCTC 477
in222         CAGCTGCTTTTCGCCTGTACTGGGTCTCTCTTGGTAGACCAGATCTGAGCCTGGGAGCTC 477
node36        CAGCTGCTTTTCGCCTGTGCTGGGTCTCTCTTGGTAGACCAGATCTGAGCCTGGGAGCTC 477
in175         CAGCTGCTTTTCGCCTGTACTGGGTCTCTCTAGGTAGACCAGATCTGAGCCTGGGAGCTC 476
in172         CAGCTGCTTTTCGCCTGTACTGGGTCTCTCTAGGTAGACCAGATCTGAGCCTGGGAGCTC 476
node38        CAGCTGCTTTTCGCCTGTACTGGGTCTCTCTAGGTAGACCAGATCTGAGCCTGGGAGCTC 476
node35        CAGCTGCTTTTCGCCTGTACTGGGTCTCTCTAGGTAGACCAGATCTGAGCCTGGGAGCTC 476
cho1          CAGCTGCTTTTCGCCTGTACTGGGTCTCTCTAGTCAGACCAGATCTGAGCCTGGGAGCTC 476
node34        CAGCTGCTTTTCGCCTGTACTGGGTCTCTCTAGTTAGACCAGATCTGAGCCTGGGAGCTC 476
node25        CAGCTGCTTTTCGCCTGTACTGGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTC 476
              **** **** *  * *** ********* *  *  ******* *  ***** ********
```

Figure 6.11: Multiple alignment of HIV-1 promoters, part8

```
bal2            TCTGGCTAACTAGGGAACCCACTG- 500
bal4            TCTGGCTAGCTAGGGAACCCACTG- 500
node21          TCTGGCTAACTAGGGAACCCACTG- 500
lai4            TCTGGCTAACTAGGGAACCCACTG- 500
node22          TCTGGCTAACTAGGGAACCCACTG- 500
lai2            TCTGGCTAACTAGGGAACCCACTG- 500
node23          TCTGGCTAACTAGGGAACCCACTG- 500
lai1            TCTGGCTAACTAGGGAACCCACTG- 500
node24          TCTGGCTAACTAGGGAACCCACTG- 500
th516           TCTGGCTAGCAAGGGAACCCACTGC 500
th515           TCTGGCTAGCAAGGGAACCCACTGC 500
th513           TCTGGCTAGCAAGGGAACCCGCTGC 500
node29          TCTGGCTAGCAAGGGAACCCACTGC 500
node28          TCTGGCTAGCAAGGGAACCCACTGC 500
th511           TCTGGCTAGCAAGGGAACCCACTGC 500
node27          TCTGGCTAGCAAGGGAACCCACTGC 500
th645           TCTGGCTAGCAAGGGAACCCACTGC 500
th644           TCTGGCTAGCAAGGGAACCCACTGC 500
node33          TCTGGCTAGCAAGGGAACCCACTGC 500
th534           TCTGGCTAGCAAGGGAACCCACTGC 500
node32          TCTGGCTAGCAAGGGAACCCACTGC 500
kr255           TCTGGCTAGCAAGGGAACCCACTGC 500
kr251           TCTGGCTAGCAAGGGAACCCACTGC 500
node31          TCTGGCTAGCAAGGGAACCCACTGC 500
node30          TCTGGCTAGCAAGGGAACCCACTGC 500
node26          TCTGGCTAGCAAGGGAACCCACTGC 500
in226           TCCGGCTATCTAGGGAACCCACT-- 500
in225           TCTGGCTATCTAGGGAACCCACT-- 500
node37          TCTGGCTATCTAGGGAACCCACT-- 500
in222           TCTGGCTATCTAGGGAACCCACT-- 500
node36          TCTGGCTATCTAGGGAACCCACT-- 500
in175           TCTGGCTATCTAGGGAACCCACTG- 500
in172           TCTGGCTATCTAGGGAACCCACTG- 500
node38          TCTGGCTATCTAGGGAACCCACTG- 500
node35          TCTGGCTAACTAGGGAACCCACTG- 500
cho1            TCTGGCTAACTAGGGAACCCACTG- 500
node34          TCTGGCTAACTAGGGAACCCACTG- 500
node25          TCTGGCTAACTAGGGAACCCACTG- 500
                ** ***** * ********* **
```

Figure 6.12: Multiple alignment of HIV-1 promoters, part9

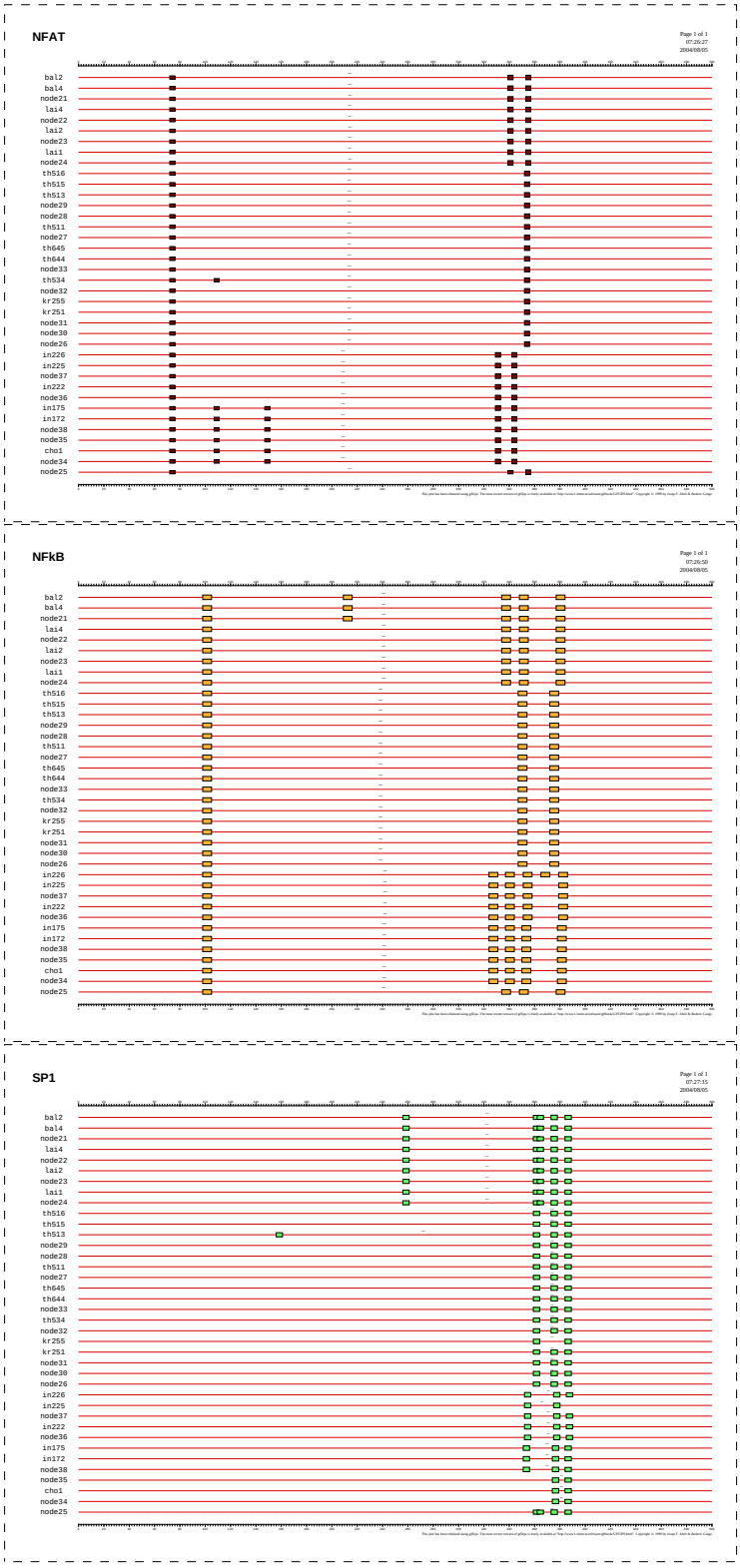

Figure 6.13: NFAT, NFkB and SP1 TFBS

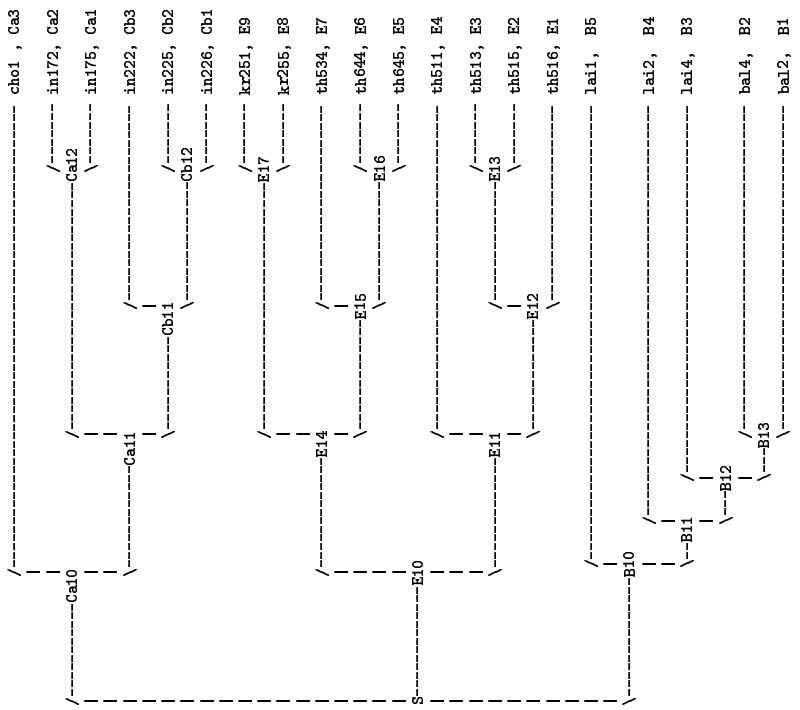

Figure 6.14: HIV-1 LTR Maximum-likelihood Tree

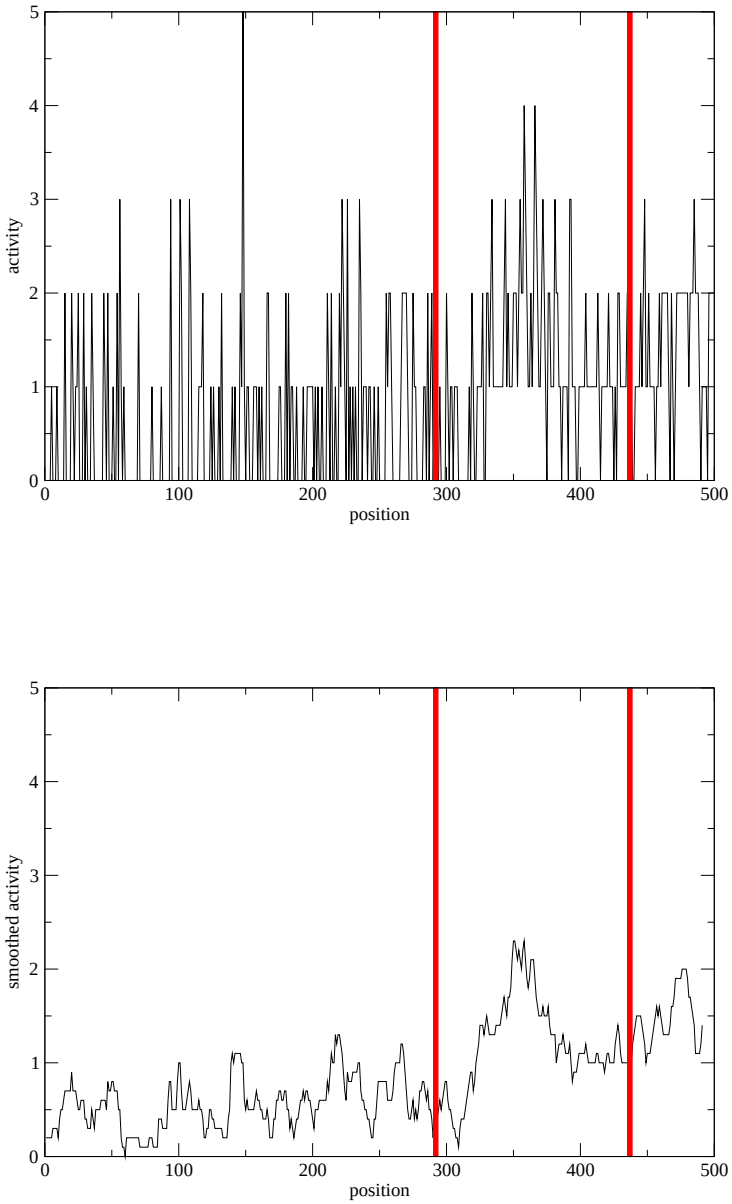

Figure 6.15: Position Activity of the alignment and smoothed activity of the alignment

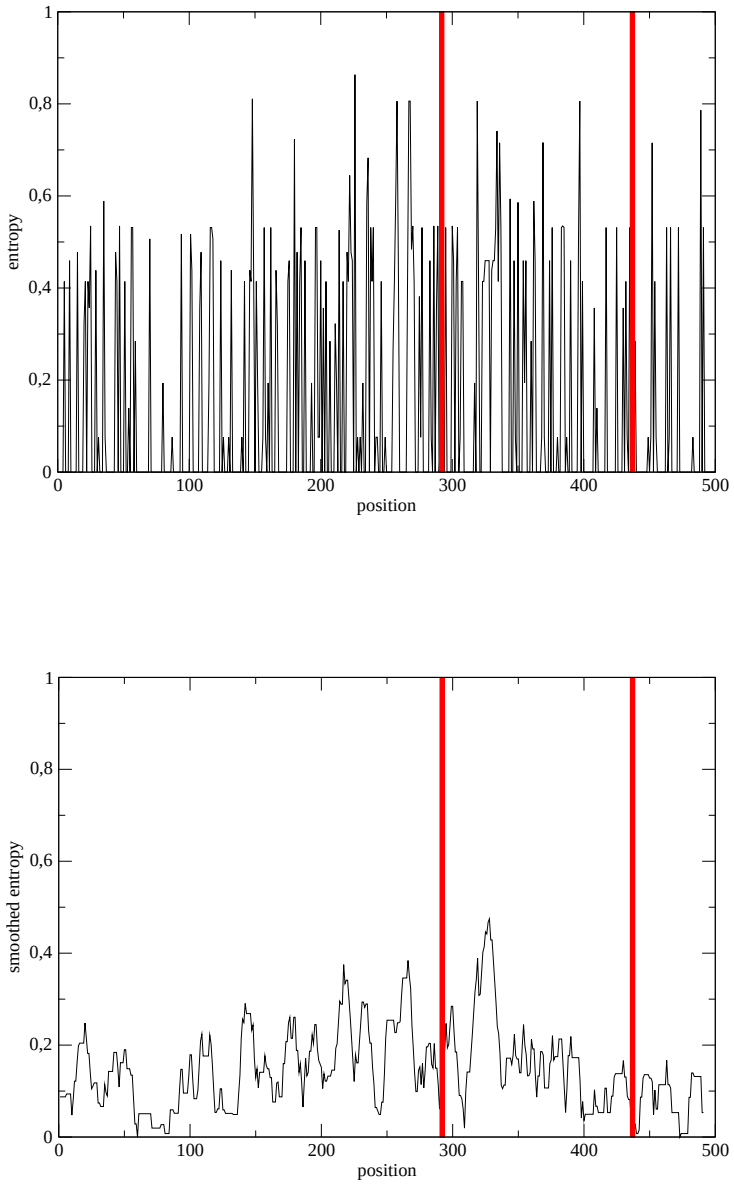

Figure 6.16: Entropy of the alignment and smoothed entropy of the alignment

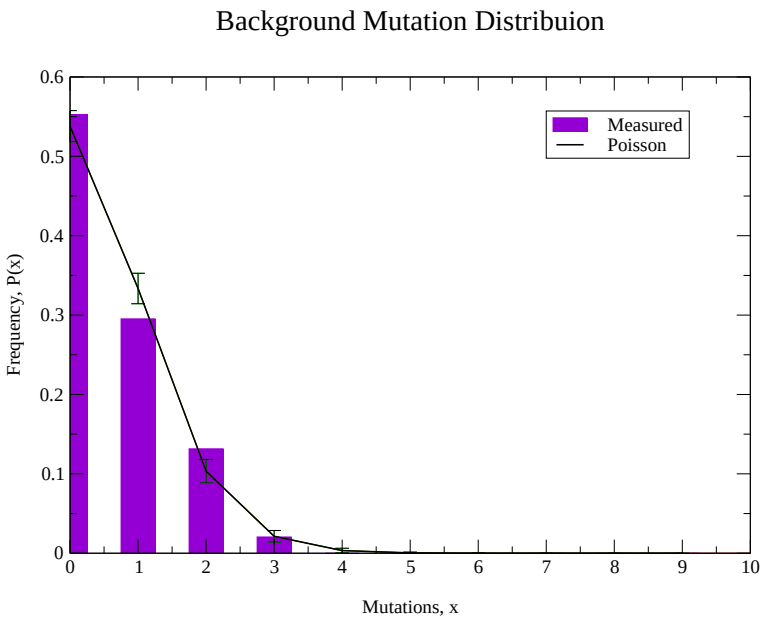

Figure 6.17: Neutral Background

Chapter 7

Experimental Check

7.1 Experimental Check of TFBS

The Goldfeld Lab [102] determined TFBSs in the HIV-1 LTR sequences we are investigating computaionally, using an experimental method called *DNAse I footprinting*. The basic experimental kit includes a DNA strand, a transcription factor and an enzyme. The DNA strand is assumed to have the binding site motif for a transcription factor. Thus, the TF will wrap around the DNA strand where this motif (or TFBS) is located. On the other hand, an enzyme is capable of cutting the DNA strand into pieces. Yet, if the DNA strand is protected by a protein like the TF being studied, the enzyme will fail to cut it anywhere along the TFBS motif. The segment of DNA shielded from enzyme digestion by a TF is called *TF footprint*. This footprint is typically wider than the actual binding site [59] i.e. footprints are typically 10-20 bases long, whereas TFBSs span 5 to 8 bases.

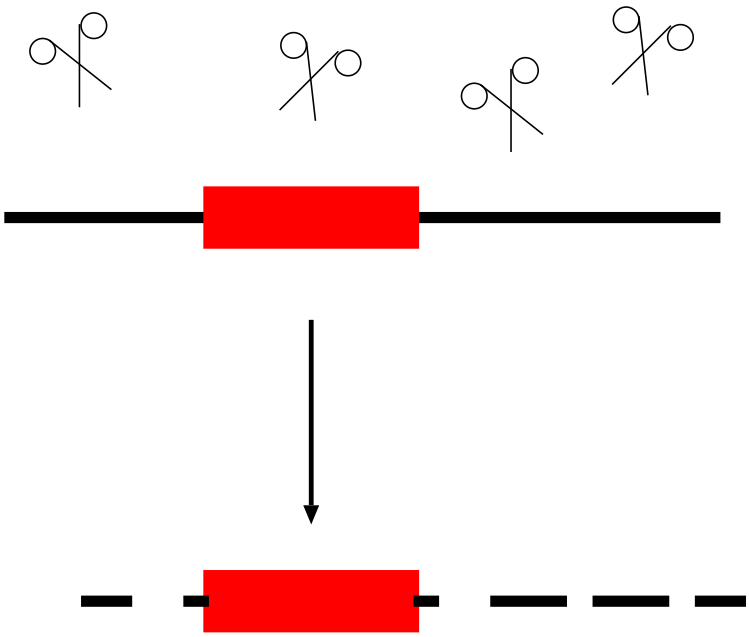

Figure 7.1: DNAse Footprinting Sketch

In Figures 7.2, 7.3, lower case letters indicate experimentally verified motifs. Computationally detected binding sites are marked with bars beneath alignments. It looks as if computational tools are able to map the interactions between TFs and the DNA.

Altogether, knowing that mathematical modeling of genomic sequences

is in search of short-cuts for TFBS identification, we think that our computational tools fulfilled this task.

```
                                                  nf-kb        nf-kb
in225          CTACATCCGGAGTTTTACAAAGACTGCTGACACAGAAgggactttccGCTgggactttcc 60
in226          CTACATCCGGAGTTTTACAAAGACTGCTGACACAGAAgggactttccACTgggactttcc 60
in222          CTACATCCGGAGTTTTACAAAGACTGCTGACACAGAAgggactttccACTgggactttcc 60
               ******************************************** ************
                                                  | NFkB |     | NFkB |

                    nf-kb          sp1      sp1       sp1
in225          ACTAggggcgttccAGgagaagtggtctgggcgggactaggagtggtCAACCCTCAGATG 120
in226          ACTAggggcgttccAGgaggggtggtctgggcgggactgggagtggtCAACCCTCAGATG 120
in222          ACTAggggcgttccAGgaggagtggtctgggcgggactgggagtggtCAACCCTCAGATG 120
               ******************  ***************** ********************
                   | NFkB |   | SP1|       | SP1|    | SP1 |


in225          CTGCATATAAGCAGCTGCTTTTCGCCTGTGCTGGGTCTCTCTTGGTAGACCAGATCTGAG 180
in226          CTGCATATAAGCAGCTGCTTTTCGCCTGTGCTGGGTCTCTCTTGGTAGACCAGATCTGAG 180
in222          TTGCATATAAGCAGCTGCTTTTCGCCTGTACTGGGTCTCTCTTGGTAGACCAGATCTGAG 180
                ************************** *****************************
```

Figure 7.2: Experimental check for TFBS in subtype C_b

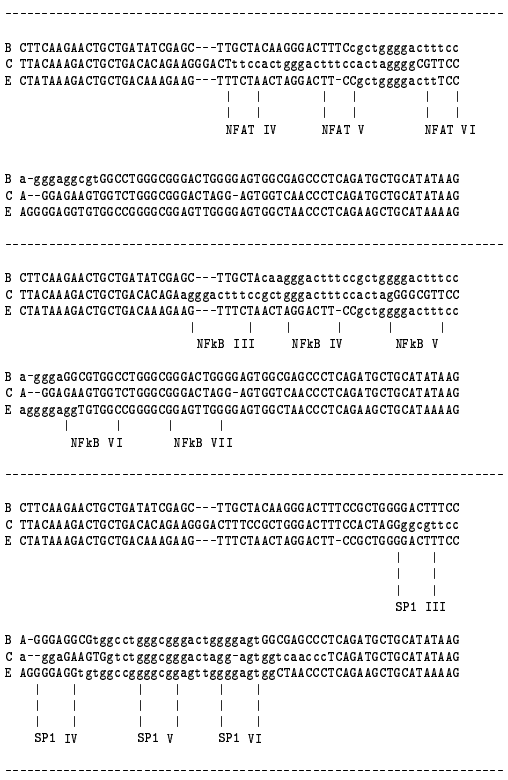

Figure 7.3: Experimental check of NFAT, NFκB and SP1 binding sites

Chapter 8

HIV-1 Fitness

8.1 Binding Site - Fitness Relationship

The results of this part were obtained in collaboration with Prof. Nikolaus Rajewsky (New York University) and Prof. Anne Goldfeld (Harvard Medical School).

In order to gauge the number of mutations we want to determine scores of individual binding sites for specific factors first, and then score differences along branches, for the same transcription factors.

As a motif score (S_i) we define the difference between the actual score and the best-binder-score. The energy of a TFBS i is defined as its negative score (S_i):

$$E_i = -S_i, \tag{8.1}$$

where i takes values: $1, 2...n_{slots}$, and n_{slots} is the number of slots for a particular transcription factor. Here, $n_{NFAT} = n_{SP1} = 6$ and $n_{NFkB} = 7$. The maximum-likelihood tree has three subtrees which correspond to B, C and E HIV-1 subtypes. All three subtrees have a common ancestor (S). S is the grandfather, from which we start going down the subtree toward the leaves when determining subtype of the father node subtype . Father subtype is defined as subtype of daughters. Each slot i is assigned an energy E_i, and for each sequence we calculate the total energy E_{tot} (last column).

$$E_{tot} = \sum\nolimits_i E_i. \tag{8.2}$$

Tables 9.1,9.4, 9.7 shows a very good subtype-specific classification on the basis of total energy. If we look carefully at the subtype column and the total energy column, we see a clear relationship for grouping nodes (HIV-1 promoter sequences) according to TFBS E_{tot}. This works as well for the sequences from [57] shown in table .

The energy change of a TFBS i is defined as the difference of its own and its father's energy:

$$D_i = E_i - E_i^{father}. \tag{8.3}$$

Again, i takes values: $1, 2...n_{slots}$, n_{slots}-number of slots for a particular transcription factor. Here, $n_{NFAT} = n_{SP1} = 6$ and $n_{NFkB} = 7$. Each slot i is assigned an energy change D_i, and for each sequence we calculate the total energy change D_{tot} .

$$D_{tot} = \sum\nolimits_i D_i. \tag{8.4}$$

In addition, we report compensations along branches. We sum up all positive energy differences, and all negative energy differences separately. As a compensation we report the sum with smaller absolute value.
Tables 9.2,9.5, 9.8 show that majority of movements takes place on crucial

subtype	SupT1	SupT1(TNF)	SupT1(PHA)	MT2	MT2(TNF)	MT2(PHA)
BA	-0.1	0.03	0	-0.06	-0.04	-0.06
BCa	-0.05	-0.02	0.01	-0.03	-0.03	-0.03
BCb	-0.13	-0.08	-0.06	-0.09	-0.09	-0.1
BD	-0.02	0.04	0.09	-0.02	0	-0.01
BE	-0.19	0.11	-0.06	-0.06	-0.07	-0.07
BF	0.02	0.08	0.05	-0.05	-0.03	-0.05
BG	-0.05	0.05	0.01	-0.03	-0.01	-0.03

Table 8.1: Fitness Differences from ref. [57]

branches connecting different subtypes (**B10-S**,**E10-S**,**Cb11-Ca11**,**Ca10-S**).

In order to visualize the tables from appendix, we show average binding site composition for B, C_a, C_b and E (see Fig. 8.1). Scores are averaged over all sequences in each subtype. These figures correspond to Tables 9.1,9.4, 9.7. **Binding site composition is subtype-specific.**

In Fig. 8.2 we display energy changes along crucial branches, where majority of movements happen. The changes determine subtype-specificity. Along branches within a subtype, changes are noise-like. The graphs visualize Tables 9.2,9.5, 9.8. **Emergence of subtypes is presented by energy changes along crucial branches.**

The authors of [57] performed replication studies with the set of A, B, C, D, E, F, G subtype HIV-1 viruses in six different cellular environments. Their study showed strong cellular environment effects on replication rates. By conducting pairwise competition experiments between subtypes, they demonstrated significant subtype-specific differences in the fitness values. We reproduce TABLE3 from [57] to see fitness differences relative to subtype B. We include their A, D, F and G sequences into the Goldfeld data set. Additional data comprise subtypes A, D, F and G, however, only the core promoter. Therefore the additional tables display slots 4, 5 and 6 for NFAT, slots 3, 4, 5, 6 and 7 for NFκB, slots 3, 4, 5 and 6 for SP1. In order to construct equations describing fitness dependence on motif scores, we use only core promoter slots from the Goldfeld data.

Analogously, we assign scores to multiple alignment slots for NFκB and SP1. Then we report sums of scores in each node, for each TF, in Goldfeld

subtype X	$< E_{NFAT} >$	$< E_{NFkB} >$	$< E_{SP1} >$
A	4.69	14.86	12.25
B	5.92	14.63	12.76
Ca	5.23	7.58	6.80
Cb	5.23	6.67	6.80
D	2.29	9.83	12.46
E	7.84	19.36	12.83
F	5.13	8.98	11.14
G	4.36	8.73	14.09

Table 8.2: Average transcription factor energies for each subtype.

full-length sequences 9.10. Again, we can see a partition according to sums of energies for TFs separately, which coincides with experimental subtype classification. Major movements take place on crucial branches.

8.2 Subtype Fitness Model

In the previous chapter we see that TFBS energy tables show subtype-specific pattern, in agreement with experimental data. This finding encouraged us to use the computed energies in order to model subtype fitness Φ. For a subtype X=A, B, C, D, E, F, G we determine average TF energy:

$$< E_{TF} >^X = \frac{\sum_{k=1}^{N_X} (\sum_{i=1}^{n_{slots}} E_i^{TF})}{N_X}. \tag{8.5}$$

where TF=NFAT, NFκB, SP1, and N_X is the number of X-subtype sequences.
We construct subtype fitness Φ_X (recall equation 3.3):

$$\Phi_X = -\sum\nolimits_{TF} f_{TF} < E_{TF} >^X \tag{8.6}$$

We want to make use of data from our collaborators' laboratory and that of [57], attempting to predict fitness differences from the TFBS energies. The fitness difference is given by:

$$\Delta\Phi_{X_1,X_2} = \Phi_{X_1} - \Phi_{X_2} \tag{8.7}$$

For the eight subtypes we determined average TF energies $< E_{TF} >$:

We arrive at 7 modeled fitness differences relative to subtype B:

$$\begin{aligned}
\Delta\Phi_{BA} &= 1.23f_{NFAT} - 0.23f_{NFkB} + 0.51f_{SP1} \\
\Delta\Phi_{BCa} &= 0.69f_{NFAT} + 7.05f_{NFkB} + 5.96f_{SP1} \\
\Delta\Phi_{BCb} &= 0.69f_{NFAT} + 7.96f_{NFkB} + 5.96f_{SP1} \\
\Delta\Phi_{BD} &= 3.63f_{NFAT} + 4.8f_{NFkB} + 0.29f_{SP1} \\
\Delta\Phi_{BE} &= -1.92f_{NFAT} - 4.73f_{NFkB} - 0.07f_{SP1} \\
\Delta\Phi_{BF} &= 0.79f_{NFAT} + 5.65f_{NFkB} + 1.62f_{SP1} \\
\Delta\Phi_{BG} &= 1.56f_{NFAT} + 5.90f_{NFkB} - 1.33f_{SP1}
\end{aligned} \tag{8.8}$$

On the other hand we have experimental fitness differences $\Delta\Phi$ in Table 8.1.

Fitted coefficients f_{NFAT}, f_{NFkB}, f_{SP1} and fitted fitness differences $\Delta\Phi^{fit}$ are obtained from the best least-square fit of:

$$\Delta\Phi = f_1\Delta < E_1 > + f_2\Delta < E_2 > + f_3\Delta < E_3 > . \tag{8.9}$$

Generalized fitting procedure includes k TFs labeled by λ, e environments labeled by α and s subtype pairs labeled by i. In our case, k=3 for NFAT, NFκB and SP1, e=6 for SupT1, SupT1+TNF, SupT1+PHA, MT2, MT2+TNF and MT2+PHA, and s=7 for BA, BCa, BCb, BD, BE, BF and BG. According to [57], each cellular environment represents a different mixture (in biological language- nuclear pool) of TFs. A change in the environment TF composition can have a strong effect on the fitness of HIV-1.
If we denote fitness of a subtype j in environment α as $\Phi_{j,\alpha}$, the fitness difference of j relative to an arbitrary chosen subtype with which all other subtypes are compared (in our case 'B') is given by:

$$\Delta\Phi_{Bj,\alpha} =: \Phi_{B,\alpha} - \Phi_{j,\alpha} \tag{8.10}$$

Score of a TF λ for a subtype j is denoted by $< E_{\lambda,j} >$. Thus, score difference between j and B is:

$$\Delta < E_{\lambda,Bj} > =: < E_{\lambda,B} > - < E_{\lambda,j} > . \tag{8.11}$$

We try a fit of the form:

$$\Delta\Phi_{i,\alpha} = \sum_{\lambda} f_{\lambda,\alpha}\Delta < E_{\lambda,i} > + C_i, \tag{8.12}$$

where i=1,2...s. To obtain the best fit, we first find in each environment the best fit with given C_i for:

$$(f_{1,\alpha}, ...f_{k,\alpha})(C_1, ..., C_s) \tag{8.13}$$

Subtype pair	$\Delta\Phi^{fit}$	$\Delta\Phi$
BA	-0.01	0
BC_a	0.01	0.01
BC_b	-0.06	-0.06
BD	0.08	0.09
BE	-0.09	-0.06
BF	0.02	0.05
BG	0.01	0.01

Table 8.3: Fitted and experimental fitness differences in SupT1+PHA environment

and

$$({\Delta\Phi_{1,\alpha}}^{fit}, ...{\Delta\Phi_{s,\alpha}}^{fit})(C_1, ..., C_s). \tag{8.14}$$

Fitting is performed using statistical software R which offers the $lsfit()$ function for least-square fitting.
This allows us to compute the mean square deviation

$$\sigma^2(C_1, ..., C_s) = \sum_{\alpha}^{e} \sum_{i}^{s} ({\Delta\Phi_{i,\alpha}}^{fit} - \Delta\Phi_{i,\alpha})^2. \tag{8.15}$$

Then, we find the best optimal values $(C_1, ..., C_s)$ by minimizing σ^2.

In doing so, we determined σ^2=0.075. This value corresponds to the optimal set of C_i: (-0.036,0.030,-0.038,0.002,-0.047,0.012,-0.035).

For the third environment SupT1(PHA) we got the best fit. The table clarifies data points (see table 8.3 and Fig. 8.2) .

In order to quantify the correlation between the two variables ($\Delta\Phi^{fit}$ and $\Delta\Phi$) we computed the Pearson correlation coefficient (8.16), a quantity which takes values between -1 and +1:

$$c_P = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_X} \right) \left(\frac{y_i - \bar{y}}{s_Y} \right), \tag{8.16}$$

where $\bar{x}, \bar{y}$ are the means of x and y, s_X, s_Y are their standard deviations and n is the number of (x, y) pairs [95].

Its sign indicates if two variables are positively (+) or negatively (-) associated, while its absolute value gives a measure of the strength of correlation

Subtype	Φ	Φ^{fit}
E	1.10	1.04
C_b	1.07	1.09
A	1.02	1.01
B	1.01	1.01
C_a	1.00	0.99
G	1.00	0.97
F	0.99	0.95
D	0.93	0.94

Table 8.4: Ranked fitted and experimental fitness in SupT1+PHA

of the two. To our delight, its value for $\Delta\Phi^{fit}$ and $\Delta\Phi$ turns out to be 0.97! Obviously, relative fitness ranking of HIV-1 in SupT1+PHA is almost perfectly conserved (see Table 8.4).

8.3 Concluding Remarks

We think that the above analysis has shown a very good agreement between theory and experiment, especially for one of the environments under investigation (SupT1 + PHA). We can conclude that NFAT, NFκB and SP1 play a crucial role in determining HIV-1 fitness in the SupT1+PHA collection of TFs. In our opinion, this is an important result, since it is little known about the differences in the transcription factor composition between different cellular environments. Although it is estimated that there are a few thousands of TFs encoded by the human genome, it is still poorly understood when and where the majority of these TFs are expressed [57, 59], in plain English, when and where are they floating around.
There may be different reasons why the fit is performing less successfully for other environments (see Figures 8.3,8.4,8.5,8.6,8.7) . One possibility would be that the total fitness is not a linear function of individual TF fitnesses, due to synergistic interactions between TFs that are absent in SupT1+PHA, yet present in other environments. Another possibility would be the existence of additional TFBSs that we are not aware of. A third scenario could include some interactions of certain TFs with the TAR sequence. Also, we could imagine that subregions of the core promoter are weighted unequally.

Our main interest was to find the relationship between the fitness and

the binding probability for individual transcription factors. We relied on the experiments indicating that binding sites play a significant role for growth rates in specific environments. Thus, we thought we should be able to pinpoint the genomic sequence origin of the fitness differences, which could be verified by experiments. The last modeling step in this study, where we assumed that transcription factor binding site scores contribute additively to fitness, turned out to be very encouraging. By bringing together theory and experiment, we were able to construct a model relating bioinformatical score to fitness with good predictive power.

Understanding the genotype-phenotype relationship (here, genomic sequence relationship to fitness) is still a largely uncharted territory in genetics. We believe that solutions to these problems will emerge from close interplay between computational methods and experimental cross-checks.

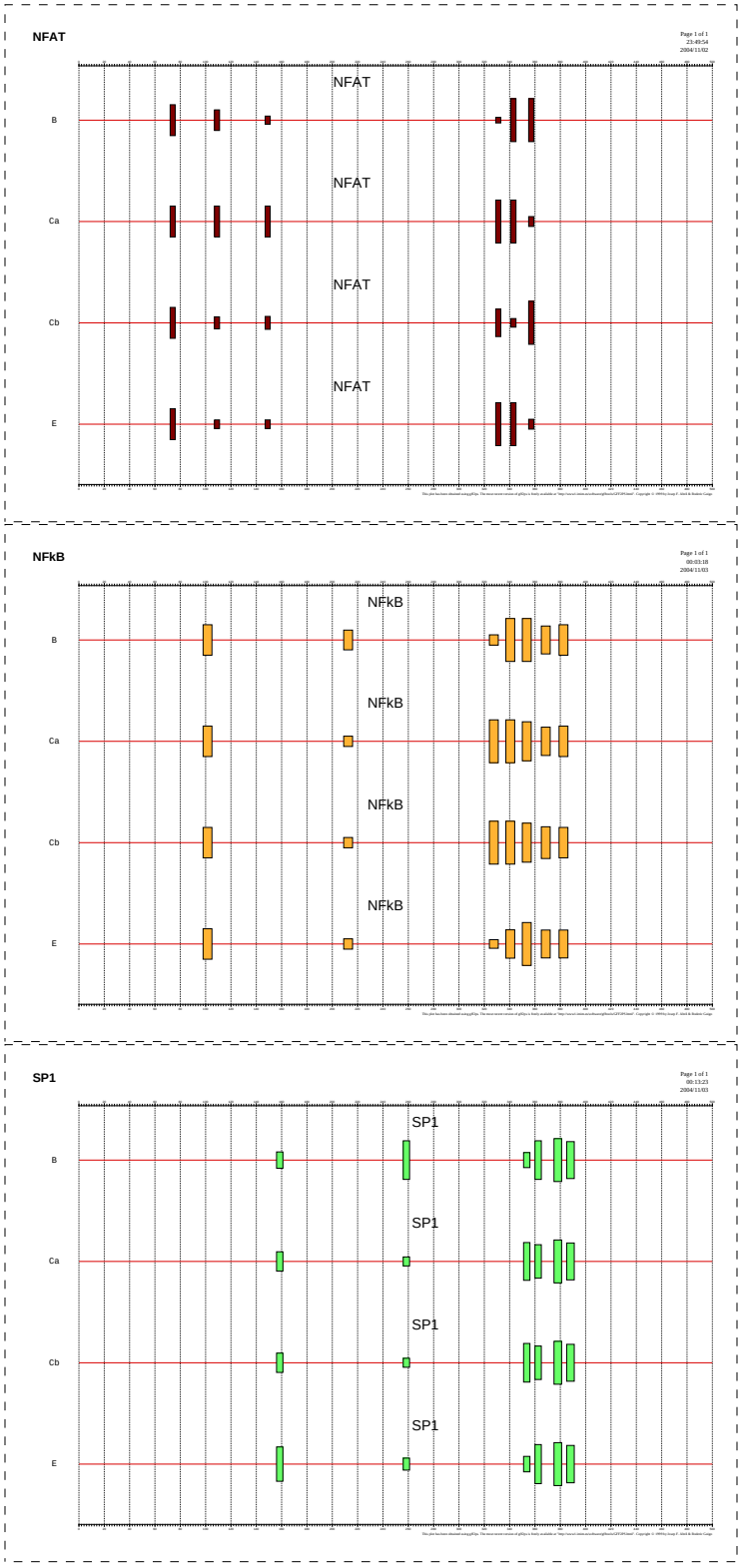

Figure 8.1: NFAT, NFkB and SP1 average score profiles

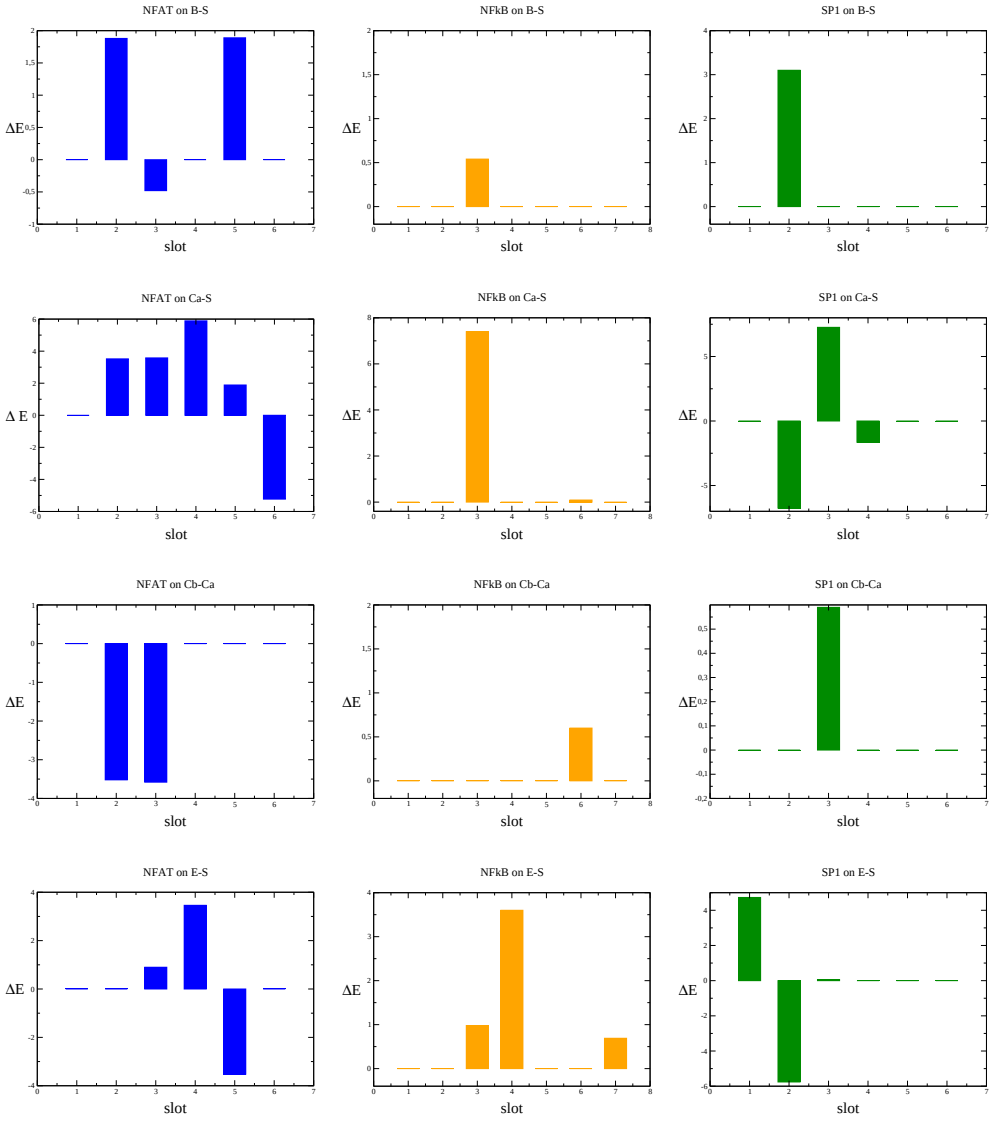

Fig. 8.2 Evolution of binding site composition along crucial branches.
Energy change ΔE in each slot is shown.
First column: NFAT. Second column: NFκB. Third column: SP1.

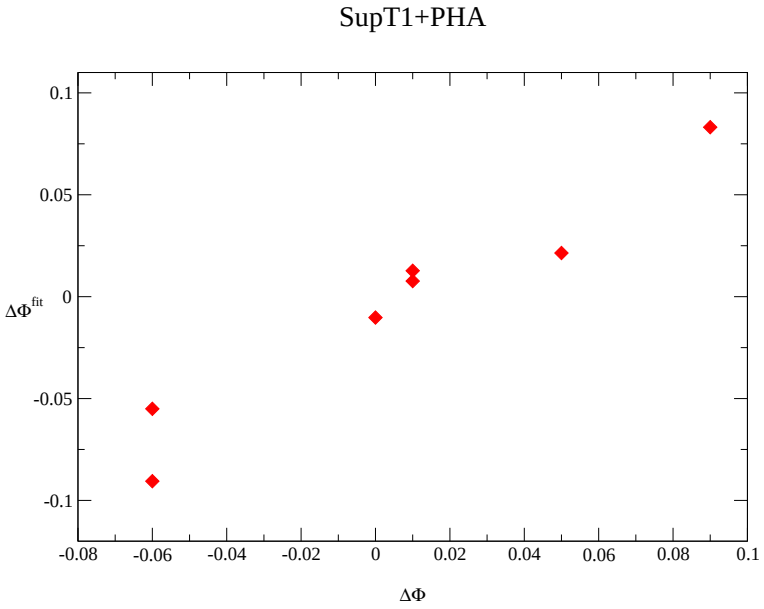

Figure 8.2: $\Delta\Phi$ versus $\Delta\Phi^{fit}$

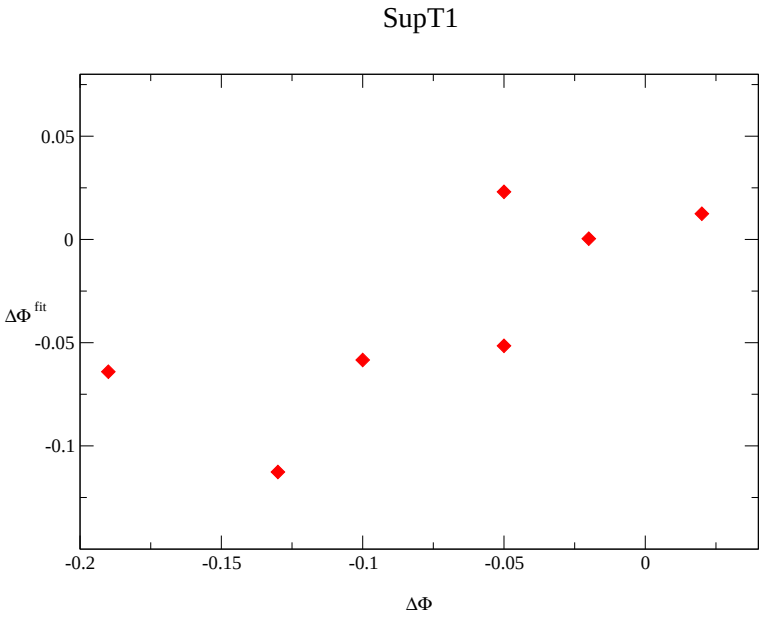

Figure 8.3: $\Delta\Phi$ versus $\Delta\Phi^{fit}$

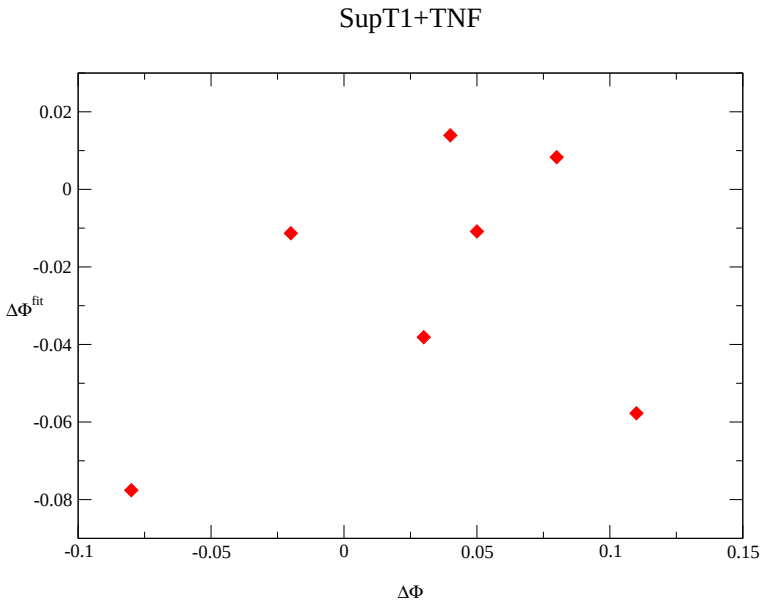

Figure 8.4: $\Delta\Phi$ versus $\Delta\Phi^{fit}$

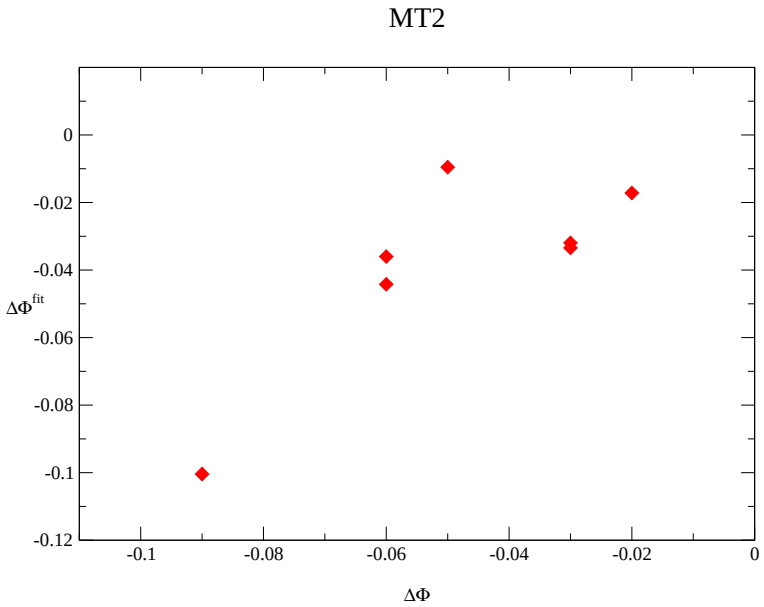

Figure 8.5: $\Delta\Phi$ versus $\Delta\Phi^{fit}$

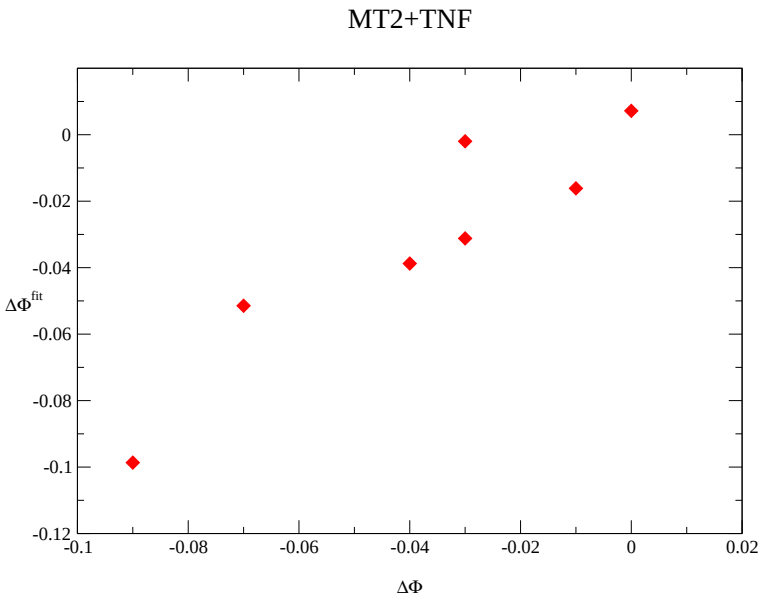

Figure 8.6: $\Delta\Phi$ versus $\Delta\Phi^{fit}$

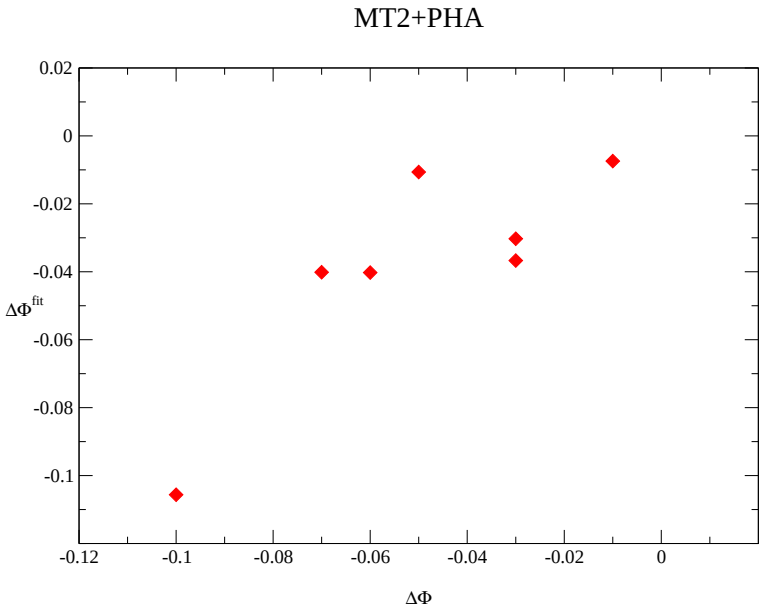

Figure 8.7: $\Delta\Phi$ versus $\Delta\Phi^{fit}$

Chapter 9

Appendix

sequence	slot1	slot2	slot3	slot4	slot5	slot6	E_{tot}
B1	1.89	3.53	4.57	5.92	0	0	15.91
B2	1.89	3.53	4.57	5.92	0	0	15.91
B13	1.89	3.53	4.57	5.92	0	0	15.91
B3	1.89	3.53	5.95	5.92	0	0	17.29
B12	1.89	3.53	5.95	5.92	0	0	17.29
B4	1.89	3.53	5.95	5.92	0	0	17.29
B11	1.89	3.53	5.95	5.92	0	0	17.29
B5	1.89	3.53	5.95	5.92	0	0	17.29
B10	1.89	3.53	5.95	5.92	0	0	17.29
E1	1.89	5.41	4.57	2.43	5.41	0	19.71
E2	1.89	5.41	4.57	2.43	5.41	0	19.71
E3	1.89	5.41	4.57	2.43	5.41	0	19.71
E13	1.89	5.41	4.57	2.43	5.41	0	19.71
E12	1.89	5.41	4.57	2.43	5.41	0	19.71
E4	1.89	5.41	4.57	2.43	5.41	0	19.71
E11	1.89	5.41	4.57	2.43	5.41	0	19.71
E5	1.89	5.41	5.47	2.43	5.41	0	20.61
E6	1.89	5.41	5.47	2.43	5.41	0	20.61
E16	1.89	5.41	5.47	2.43	5.41	0	20.61
E7	1.89	1.89	4.57	2.43	5.41	0	16.19
E15	1.89	5.41	4.57	2.43	5.41	0	19.71
E8	1.89	3.53	4.57	2.43	5.41	0	17.83
E9	1.89	3.53	4.57	2.43	5.41	0	17.83
E17	1.89	3.53	4.57	2.43	5.41	0	17.83
E14	1.89	5.41	4.57	2.43	5.41	0	19.71
E10	1.89	5.41	4.57	2.43	5.41	0	19.71
Cb1	1.89	5.41	5.47	0	0	5.23	18
Cb2	1.89	5.41	5.47	0	0	5.23	18
Cb12	1.89	5.41	5.47	0	0	5.23	18
Cb3	1.89	5.41	5.47	0	0	5.23	18
Cb11	1.89	5.41	5.47	0	0	5.23	18
Ca1	1.89	1.89	1.89	0	0	5.23	10.9
Ca2	1.89	1.89	1.89	0	0	5.23	10.9
Ca12	1.89	1.89	1.89	0	0	5.23	10.9
Ca11	1.89	1.89	1.89	0	0	5.23	10.9
Ca3	1.89	1.89	1.89	0	0	5.23	10.9
Ca10	1.89	1.89	1.89	0	0	5.23	10.9
S	1.89	5.41	5.47	5.89	1.89	0	20.55

Table 9.1: NFAT slots (Goldfeld data) in nodes

branch	slot1	slot2	slot3	slot4	slot5	slot6	D_{tot}	comp
B1-B13	0.00	0.00	1.38	0.00	0.00	0.00	1.38	0.00
B2-B13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B13-B12	0.00	0.00	1.38	0.00	0.00	0.00	1.38	0.00
B3-B12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B12-B11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B4-B11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B11-B10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B5-B10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B10-S	0.00	1.88	-0.48	-0.03	1.89	0.00	3.26	0.51
E1-E12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E2-E13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E3-E13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E13-E12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E12-E11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E4-E11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E11-E10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E5-E16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E6-E16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E16-E15	0.00	0.00	-0.90	0.00	0.00	0.00	-0.90	0.00
E7-E15	0.00	3.52	0.00	0.00	0.00	0.00	3.52	0.00
E15-E14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E8-E17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E9-E17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E17-E14	0.00	1.88	0.00	0.00	0.00	0.00	1.88	0.00
E14-E10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E10-S	0.00	0.00	0.90	3.46	-3.52	0.00	0.84	3.52
Cb1-Cb12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cb2-Cb12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cb12-Cb11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cb3-Cb11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cb11-Ca11	0.00	-3.52	-3.58	0.00	0.00	0.00	-7.10	0.00
Ca1-Ca12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ca2-Ca12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ca12-Ca11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ca11-Ca10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ca3-Ca10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ca10-S	0.00	3.52	3.58	5.89	1.89	-5.23	9.65	5.23
S-S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 9.2: NFAT slot energy difference in branches (Goldfeld data)

sequence	slot4	slot5	slot6	E_{tot}
A1	4.69	0	0	4.69
A2	4.69	0	0	4.69
A3	4.69	0	0	4.69
A4	4.69	0	0	4.69
A5	4.69	0	0	4.69
A6	4.69	0	0	4.69
A7	4.69	0	0	4.69
D1	1.89	0	0	1.89
D2	1.89	0	0	1.89
D3	1.89	0	0	1.89
D4	1.89	0	0	1.89
D5	1.89	0	0	1.89
D6	1.89	0	3.58	5.47
D7	1.89	0	0	1.89
D8	1.89	0	0	1.89
D9	1.89	0	0	1.89
F1	4.69	0	0	4.69
F2	4.57	0	0	4.57
F3	4.57	0	0	4.57
F4	4.69	0	0	4.69
F5	4.69	0	3.53	8.22
F6	4.57	0	0	4.57
F7	4.57	0	0	4.57
G1	4.69	0	0	4.69
G2	4.69	0	0	4.69
G3	4.03	0	0	4.03
G4	4.69	0	0	4.69
G5	4.69	0	0	4.69
G6	4.03	0	0	4.03
G7	4.03	0	0	4.03
G8	4.03	0	0	4.03

Table 9.3: NFAT scores slots from data from ref. [57]

sequence	slot1	slot2	slot3	slot4	slot5	slot6	slot7	E_{tot}
B1	3	0.98	7.94	0	0	3.69	3	18.61
B2	3	0.98	7.94	0	0	3.69	3	18.61
B13	3	0.98	7.94	0	0	3.69	3	18.61
B3	3	7.94	7.94	0	0	3.69	3	25.57
B12	3	7.94	7.94	0	0	3.69	3	25.57
B4	3	7.94	7.94	0	0	3.69	3	25.57
B11	3	7.94	7.94	0	0	3.69	3	25.57
B5	3	7.94	7.94	0	0	3.69	3	25.57
B10	3	7.94	7.94	0	0	3.69	3	25.57
E1	3	7.94	8.38	3.6	0	3.69	3.69	30.3
E2	3	7.94	8.38	3.6	0	3.69	3.69	30.3
E3	3	7.94	8.38	3.6	0	3.69	3.69	30.3
E13	3	7.94	8.38	3.6	0	3.69	3.69	30.3
E12	3	7.94	8.38	3.6	0	3.69	3.69	30.3
E4	3	7.94	8.38	3.6	0	3.69	3.69	30.3
E11	3	7.94	8.38	3.6	0	3.69	3.69	30.3
E5	3	7.94	8.38	3.6	0	3.69	3.69	30.3
E6	3	7.94	8.38	3.6	0	3.69	3.69	30.3
E16	3	7.94	8.38	3.6	0	3.69	3.69	30.3
E7	3	7.94	8.38	3.6	0	3.69	3.69	30.3
E15	3	7.94	8.38	3.6	0	3.69	3.69	30.3
E8	3	7.94	8.38	3.6	0	3.69	3.69	30.3
E9	3	7.94	8.38	3.6	0	3.69	3.69	30.3
E17	3	7.94	8.38	3.6	0	3.69	3.69	30.3
E14	3	7.94	8.38	3.6	0	3.69	3.69	30.3
E10	3	7.94	8.38	3.6	0	3.69	3.69	30.3
Cb1	3	7.94	0	0	0.98	0.98	3	15.9
Cb2	3	7.94	0	0	0.98	3.49	3	18.41
Cb12	3	7.94	0	0	0.98	3	3	17.92
Cb3	3	7.94	0	0	0.98	3	3	17.92
Cb11	3	7.94	0	0	0.98	3	3	17.92
Ca1	3	7.94	0	0	0.98	3.6	3	18.52
Ca2	3	7.94	0	0	0.98	3.6	3	18.52
Ca12	3	7.94	0	0	0.98	3.6	3	18.52
Ca11	3	7.94	0	0	0.98	3.6	3	18.52
Ca3	3	7.94	0	0	0.98	3.6	3	18.52
Ca10	3	7.94	0	0	0	3.6	3	17.54
S	3	7.94	7.4	0	0	3.69	3	25.03

Table 9.4: NFκB slots (Goldfeld data) in nodes

branch	slot1	slot2	slot3	slot4	slot5	slot6	slot7	D_{tot}	comp
B1-B13	0.00	6.96	0.00	0.00	0.00	0.00	0.00	6.96	0.00
B2-B13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B13-B12	0.00	6.96	0.00	0.00	0.00	0.00	0.00	6.96	0.00
B3-B12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B12-B11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B4-B11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B11-B10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B5-B10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B10-S	0.00	0.00	0.54	0.00	0.00	0.00	0.00	0.54	0.00
E1-E12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E2-E13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E3-E13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E13-E12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E12-E11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E4-E11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E11-E10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E5-E16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E6-E16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E16-E15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E7-E15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E15-E14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E8-E17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E9-E17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E17-E14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E14-E10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E10-S	0.00	0.00	0.98	3.60	0.00	0.00	0.69	5.27	0.00
Cb1-Cb12	0.00	0.00	0.00	0.00	0.00	2.02	0.00	2.02	0.00
Cb2-Cb12	0.00	0.00	0.00	0.00	0.00	0.49	0.00	0.49	0.00
Cb12-Cb11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cb3-Cb11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cb11-Ca11	0.00	0.00	0.00	0.00	0.00	0.60	0.00	0.60	0.00
Ca1-Ca12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ca2-Ca12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ca12-Ca11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ca11-Ca10	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.98	0.00
Ca3-Ca10	0.00	0.00	0.00	0.00	0.98	0.00	0.00	0.98	0.00
Ca10-S	0.00	0.00	7.40	0.00	0.00	0.09	0.00	7.49	0.00
S-S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 9.5: NFκB slot energy difference in branches (Goldfeld data)

sequence	slot3	slot4	slot5	slot6	slot7	E_{tot}
A1	7.4	0	0	4.5	5.22	17.12
A2	7.4	3.71	0	1.5	5.22	17.83
A3	7.4	0	0	0	5.22	12.62
A4	7.4	0	0	1.5	5.22	14.12
A5	7.4	0	0	1.5	5.22	14.12
A6	7.4	0	0	1.5	5.22	14.12
A7	7.4	0	0	1.5	5.22	14.12
D1	4.5	0	0	3.71	3	11.21
D2	4.5	0	0	3.71	3	11.21
D3	1.5	0	0	3.71	3	8.21
D4	1.5	0	0	3.71	3	8.21
D5	4.5	0	0	3.54	3	11.04
D6	3.6	0	0	2.48	3	9.08
D7	4.5	0	0	2.56	3	10.06
D8	1.5	0	0	3.71	3	8.21
D9	4.5	0	0	3.71	3	11.21
F1	2.48	0	0	2.48	3	7.96
F2	3.49	0	0	2.48	7.94	13.91
F3	2.48	0	0	2.48	3.6	8.56
F4	2.48	0	0	2.48	3	7.96
F5	2.48	0	0	2.48	3	7.96
F6	2.48	0	0	2.48	3.6	8.56
F7	2.48	0	0	2.48	3	7.96
G1	1.5	0	0	3.71	3	8.21
G2	1.5	0	0	3.71	3.6	8.81
G3	1.5	0	0	3.71	3.6	8.81
G4	1.5	0	0	3.71	3	8.21
G5	1.5	0	0	3.71	3	8.21
G6	4.5	0	0	3.69	3	11.19
G7	1.5	0	0	3.69	3	8.19
G8	1.5	0	0	3.69	3	8.19

Table 9.6: NFκB scores slots from data from ref. [57]

sequence	slot1	slot2	slot3	slot4	slot5	slot6	E_{tot}
B1	11.06	1.44	9.29	1.44	0	2.03	25.26
B2	11.06	1.44	9.29	1.44	0	2.03	25.26
B13	11.06	1.44	9.29	1.44	0	2.03	25.26
B3	7.82	1.44	9.29	1.44	0	2.03	22.02
B12	7.82	1.44	9.29	1.44	0	2.03	22.02
B4	7.82	1.44	9.29	1.44	0	2.03	22.02
B11	7.82	1.44	9.29	1.44	0	2.03	22.02
B5	7.82	1.44	9.29	1.44	0	2.03	22.02
B10	7.82	1.44	9.29	1.44	0	2.03	22.02
E1	3.09	10.3	9.23	1.44	0	2.03	26.09
E2	3.09	10.3	9.23	1.44	0	2.03	26.09
E3	1.44	10.3	9.23	1.44	0	2.03	24.44
E13	3.09	10.3	9.23	1.44	0	2.03	26.09
E12	3.09	10.3	9.23	1.44	0	2.03	26.09
E4	3.09	10.3	9.23	1.44	0	2.03	26.09
E11	3.09	10.3	9.23	1.44	0	2.03	26.09
E5	3.09	10.3	9.23	1.44	0	2.03	26.09
E6	3.09	10.3	9.23	1.44	0	2.03	26.09
E16	3.09	10.3	9.23	1.44	0	2.03	26.09
E7	3.09	10.3	9.23	1.44	0	2.03	26.09
E15	3.09	10.3	9.23	1.44	0	2.03	26.09
E8	3.09	10.3	9.23	1.44	2.03	2.03	28.12
E9	3.09	10.3	9.23	1.44	0	2.03	26.09
E17	3.09	10.3	9.23	1.44	0	2.03	26.09
E14	3.09	10.3	9.23	1.44	0	2.03	26.09
E10	3.09	10.3	9.23	1.44	0	2.03	26.09
Cb1	7.82	11.29	1.44	3.09	0	2.03	25.67
Cb2	7.82	11.29	1.44	3.09	0	2.03	25.67
Cb12	7.82	11.29	1.44	3.09	0	2.03	25.67
Cb3	7.82	11.29	1.44	3.09	0	2.03	25.67
Cb11	7.82	11.29	1.44	3.09	0	2.03	25.67
Ca1	7.82	11.29	1.44	3.09	0	2.03	25.67
Ca2	7.82	11.29	1.44	3.09	0	2.03	25.67
Ca12	7.82	11.29	1.44	3.09	0	2.03	25.67
Ca11	7.82	11.29	2.03	3.09	0	2.03	26.26
Ca3	12.52	11.29	2.03	3.09	0	2.03	30.96
Ca10	7.82	11.29	2.03	3.09	0	2.03	26.26
S	7.82	4.54	9.29	1.44	0	2.03	25.12

Table 9.7: SP1 slots (Goldfeld data) in nodes

branch	slot1	slot2	slot3	slot4	slot5	slot6	D_{tot}	comp
B1-B13	-3.24	0.00	0.00	0.00	0.00	0.00	-3.24	0.00
B2-B13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B13-B12	-3.24	0.00	0.00	0.00	0.00	0.00	-3.24	0.00
B3-B12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B12-B11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B4-B11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B11-B10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B5-B10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B10-S	0.00	3.10	0.00	0.00	0.00	0.00	3.10	0.00
E1-E12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E2-E13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E3-E13	1.65	0.00	0.00	0.00	0.00	0.00	1.65	0.00
E13-E12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E12-E11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E4-E11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E11-E10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E5-E16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E6-E16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E16-E15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E7-E15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E15-E14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E8-E17	0.00	0.00	0.00	0.00	-2.03	0.00	-2.03	0.00
E9-E17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E17-E14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E14-E10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E10-S	4.73	-5.76	0.06	0.00	0.00	0.00	-0.97	-4.79
Cb1-Cb12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cb2-Cb12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cb12-Cb11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cb3-Cb11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cb11-Ca11	0.00	0.00	0.59	0.00	0.00	0.00	0.59	0.00
Ca1-Ca12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ca2-Ca12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ca12-Ca11	0.00	0.00	0.59	0.00	0.00	0.00	0.59	0.00
Ca11-Ca10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ca3-Ca10	-4.70	0.00	0.00	0.00	0.00	0.00	-4.70	0.00
Ca10-S	0.00	-6.75	7.26	-1.65	0.00	0.00	-1.14	-7.26
S-S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 9.8: SP1 slot energy difference in branches (Goldfeld data)

sequence	slot3	slot4	slot5	slot6	E_{tot}
A1	9.29	1.44	0	1.44	12.17
A2	9.29	1.44	0	1.44	12.17
A3	9.29	2.03	0	1.44	12.76
A4	9.29	1.44	0	1.44	12.17
A5	9.29	1.44	0	1.44	12.17
A6	9.29	1.44	0	1.44	12.17
A7	9.29	1.44	0	1.44	12.17
D1	9.29	2.03	0	1.44	12.76
D2	9.29	1.44	0	1.44	12.17
D3	9.29	1.44	0	1.44	12.17
D4	9.29	1.44	0	1.44	12.17
D5	9.29	1.44	0	1.44	12.17
D6	9.29	1.44	2.03	1.44	14.2
D7	9.29	1.44	0	1.44	12.17
D8	9.29	1.44	0	1.44	12.17
D9	9.29	1.44	0	1.44	12.17
F1	9.29	1.44	0	1.44	12.17
F2	9.29	1.44	0	1.44	12.17
F3	9.29	0	0	1.44	10.73
F4	9.29	0	0	1.44	10.73
F5	9.29	0	0	1.44	10.73
F6	9.29	0	0	1.44	10.73
F7	9.29	0	0	1.44	10.73
G1	9.29	1.44	1.44	1.44	13.61
G2	9.29	1.44	1.44	1.44	13.61
G3	9.29	1.44	1.44	1.44	13.61
G4	9.29	1.44	1.44	1.44	13.61
G5	9.29	1.44	1.44	1.44	13.61
G6	9.29	5.29	1.44	1.44	17.46
G7	9.29	1.44	1.44	1.44	13.61
G8	9.29	1.44	1.44	1.44	13.61

Table 9.9: SP1 scores slots from data from ref. [57]

sequence	$\sum$ NFAT	$\sum$ NFkB	$\sum$ SP1
B1	15.91	18.61	25.26
B2	15.91	18.61	25.26
B13	15.91	18.61	25.26
B3	17.29	25.57	22.02
B12	17.29	25.57	22.02
B4	17.29	25.57	22.02
B11	17.29	25.57	22.02
B5	17.29	25.57	22.02
B10	17.29	25.57	22.02
E1	19.71	30.3	26.09
E2	19.71	30.3	26.09
E3	19.71	30.3	24.44
E13	19.71	30.3	26.09
E12	19.71	30.3	26.09
E4	19.71	30.3	26.09
E11	19.71	30.3	26.09
E5	20.61	30.3	26.09
E6	20.61	30.3	26.09
E16	20.61	30.3	26.09
E7	16.19	30.3	26.09
E15	19.71	30.3	26.09
E8	17.83	30.3	28.12
E9	17.83	30.3	26.09
E17	17.83	30.3	26.09
E14	19.71	30.3	26.09
E10	19.71	30.3	26.09
Cb1	18	15.9	25.67
Cb2	18	18.41	25.67
Cb12	18	17.92	25.67
Cb3	18	17.92	25.67
Cb11	18	17.92	25.67
Ca1	10.9	18.52	25.67
Ca2	10.9	18.52	25.67
Ca12	10.9	18.52	25.67
Ca11	10.9	18.52	26.26
Ca3	10.9	18.52	30.96
Ca10	10.9	17.54	26.26
S	20.55	25.03	25.12

Table 9.10: TF score sums in Goldfeld data sequences

branch	$\sum$ NFAT	$\sum$ NFkB	$\sum$ SP1	comp
B1-B13	1.38	6.96	-3.24	-3.24
B2-B13	0.00	0.00	0.00	0
B13-B12	1.38	6.96	-3.24	-3.24
B3-B12	0.00	0.00	0.00	0
B12-B11	0.00	0.00	0.00	0
B4-B11	0.00	0.00	0.00	0
B11-B10	0.00	0.00	0.00	0
B5-B10	0.00	0.00	0.00	0
B10-S	3.26	-0.54	3.10	-0.54
E1-E12	0.00	0.00	0.00	0
E2-E13	0.00	0.00	0.00	0
E3-E13	0.00	0.00	1.65	0
E13-E12	0.00	0.00	0.00	0
E12-E11	0.00	0.00	0.00	0
E4-E11	0.00	0.00	0.00	0
E11-E10	0.00	0.00	0.00	0
E5-E16	0.00	0.00	0.00	0
E6-E16	0.00	0.00	0.00	0
E16-E15	-0.90	0.00	0.00	0
E7-E15	3.52	0.00	0.00	0
E15-E14	0.00	0.00	0.00	0
E8-E17	0.00	0.00	-2.03	0
E9-E17	0.00	0.00	0.00	0
E17-E14	1.88	0.00	0.00	0
E14-E10	0.00	0.00	0.00	0
E10-S	0.84	-5.27	-0.97	0.84
Cb1-Cb12	0.00	2.02	0.00	0
Cb2-Cb12	0.00	-0.49	0.00	0
Cb12-Cb11	0.00	0.00	0.00	0
Cb3-Cb11	0.00	0.00	0.00	0
Cb11-Ca11	-7.10	0.60	0.59	1.19
Ca1-Ca12	0.00	0.00	0.00	0
Ca2-Ca12	0.00	0.00	0.00	0
Ca12-Ca11	0.00	0.00	0.59	0
Ca11-Ca10	0.00	-0.98	0.00	0
Ca3-Ca10	0.00	-0.98	-4.70	0
Ca10-S	9.65	7.49	-1.14	-1.14
S-S	0.00	0.00	0.00	0

Table 9.11: Sums of branch score differences of all TFs (Goldfeld data)

Bibliography

[1] Dobzhansky, T., 1951, *Genetics and the Origin of Species*. Columbia University Press.

[2] Schroedinger, E., 1944, *What is life?*. Cambridge University Press.

[3] Darwin, C., 1998, *On the origin of Species*. Harvard University Press.

[4] Cairns, J., Overbaugh, J. and Miller, S., 1988, The origin of mutants. Nature. **335:** 142-145.

[5] Luria, S. and Delbrück, M., 1943, Mutations of Bacteria from Virus Sensitivity to Virus Resistance. Genetics. **28:** 491-511.

[6] Watson, J. D. and Crick, F. H. C., 1953, Molecular Structure of Nucleic Acids. Nature. **171:** 737-738.

[7] Kulkarni, M. M., Arnosti, D. N., 2003, Information display by transcriptional enhancers. Development. **130(26):** 6569-6575.

[8] MacArthur, S., Brookfield, J., 2004 Expected rates and modes of evolution of enhancer sequences. Mol. Biol. Evol. **21**(6):1064-1073.

[9] Kingman, J., 1978 A simple model for the balance between selection and mutation. J. Appl. Prob. **15**:1-12.

[10] Moses, A., Chang, D., Kellis, M., Lander, E., Eisen, M., 2003 Position specific variation in the rate of evolution in transcription factor binding sites. BMC Evolutionary Biology. **3**:19.

[11] J. Collado-Vides and B. Magasanik and J. D. Gralla, 1991 Control site location and transcriptional regulation in *Escherichia coli*. Microbiol. Reviews **55:** 371-394.

[12] H. J. Bussemaker and H. Li and E. D. Siggia, 2000 Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. Proc. Nat. Acad. Sci. USA **97:** 10096-10100.

[13] G.Z. Hertz and G. D. Stormo, 1999 Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics **15:** 563-577.

[14] G. D Stormo and D. S. Fields, 1998 Specificity, energy and information in DNA-protein interactions. Trends Biochem. Sci. **23:** 109-113.

[15] E. H. Davidson, 1999 A view from the genome : Spatial control of transcription in sea urchin development. Current Opinion in Genetics & Development **9:** 530-541.

[16] D. Tautz, 2000 Evolution of transcriptional regulation. Current Opinion in Genetics & Development **10:** 575-579.

[17] U. Gerland and T. Hwa, 2002 On the Selection and Evolution of Regulatory DNA Motifs. J. Mol. Evol. **55:** 386-400.

[18] A. Sengupta and M. Djordjevic and B. Shraiman, 2002 Specificity and robustness in transcription control networks. Proc. Nat. Acad. Sci. USA **99:** 2072-2077.

[19] O.G. Berg and P.H. von Hippel, 1987 Selection of DNA binding sites by regulatory proteins. J. Mol. Biol. **193:** 723-750.

[20] U. Gerland and D. Moroz and T. Hwa, 2002 Physical constraints and functional characteristics of transcription factor-DNA interaction. Proc.Nat. Acad. Sci. USA. **99:** 12015-12020.

[21] M. Eigen and J. McCaskill and P. Schuster, 1989 The molecular Quasispecies. Adv. Chem. Phys. **75:** 149-263.

[22] R. A. Goldstein and Luthey-Schulten and P. G. Wolynes, 1992 Optimal Protein-Folding Codes from Spin-Glass Theory. Proc. Nat. Acad. Sci. USA. **89:** 4918-4922.

[23] A. Wagner, 2002 Selection after gene duplication: a view from the genome. Genome Biology **3:** 1012.1-1012.3.

[24] M. Lynch and M. O'Hely and B. Walsh and A. Force, 2001 The Probability of Preservation of a Newly Arisen Gene Duplicate. Genetics **159:** 1789-1804.

[25] M. Z. Ludwig and M. Kreitman, 1995 Evolutionary Dynamics of the Enhancer region of even-skipped in Drosophila. Mol. Biol. Evol. **12:** 1002-1011.

[26] M. Z. Ludwig and N. H. Patel and M. Kreitman, 1998 Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. Development **125:** 949-958.

[27] E. T. Dermitzakis and C. M. Bergman and A. G. Clark, 2002 Tracing the evolutionary history of *Drosophila* regulatory regions with models that identify transcription factor binding sites. Mol. Biol. Evol. **20:** 703-714.

[28] J. L. Scemama and M. Hunter and J. McCallum and V. Prince and E. Stellwag, 2002 Evolutionary divergence of vertebrate Hoxb2 expression patterns and transcriptional regulatory loci. J. Exp. Zool. **294:** 285-299.

[29] D. N. Arnosti, 2003 Analysis and function of transcriptional regulatory elements: Insights from *Drosophila*. Ann. Review Entymology **48:** 579-602.

[30] J. Costas and F. Casares and J. Vieira, 2003 Turnover of binding sites for transcription factors involved in early Drosophila development. Gene **310:** 215-220.

[31] A. P. McGregor *et al*, 2001 Rapid restructuring of bicoid-dependent hunchback promoters within and between Dipteran species: Implications for molecular evolution. Evolution and Development **3:** 397-407.

[32] J. A. Shapiro, 1999 Transposable elements as the key to a 21st century view of evolution. Genetica **107:** 171-179

[33] J. R. Stone and G. A. Wray, 2001 Rapid Evolution of cis-Regulatory Sequences via Local Point Mutations. Mol. Biol. Evol. **18:** 1764-1770.

[34] D. S. Fields and Y. He and A. Y. Al-Uzri and G. D. Stormo, 1997 Quantitative specificity of the mnt repression. J. Mol. Biol. **271:** 178-194.

[35] M. Oda and K. Furukawa and K. Ogata and A. Sarai and H. Nakamura, 1998 Thermodynamics of specific and non-specific DNA binding by the c-Myb DNA-binding domain. J. Mol. Biol. **276:** 571-590.

[36] Ohta, T. and Tachida, H., 1990 Theoretical study of near neutrality. I. Heterozygosity and rate of mutant substitution. Genetics **126:** 219-229.

[37] A. Sarai and Y. Takeda, 1989 RT Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. Proc. Nat. Acad. Sci. USA **86:** 6513-6517.

[38] L. Peliti, 1995, Fitness Landscapes and Evolution. Physics of Biomaterials: Fluctuations, Self-Assembly and Evolution, T. Riste and D. Sherrington, (Eds.) (Dordrecht: Kluwer, 1996) 287-308. cond-mat/9505003.

[39] L. Peliti, 2002 Quasispecies evolution in general mean-field landscapes. Europhys. Lett. **57:** 745-751.

[40] C. Schlötterer and M.-T. Hauser and A. v. Haeseler and D. Tautz, 1994 Comparative evolutionary analysis of rDNA ITS regions in *Drosophila*. Mol. Biol. Evol. **11:** 513-522.

[41] M. Kimura and T. Ohta, 1969 The average number of generations until fixation of a mutant gene in a finite population. Genetics **61:** 763-771.

[42] M. Kimura, 1962 On the probability of fixation of mutant genes in a population. Genetics **47:** 713-719.

[43] D. J. Begun and C. F. Aquadro, 1992 Levels of naturally occuring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356:** 519-520.

[44] H. Akashi, 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. Genetics **139:** 1067-1076.

[45] N. G. C. Smith and A. Eyre-Walker, 2002 Adaptive protein evolution in *Drosophila*. Nature **415:** 1022-1024.

[46] D. L. Jenkins and C. A. Ortori and J. F. Brookfield, 1995 A test for adaptive change in DNA sequences controlling transcription. Proc. R. Soc. Lond. **B 261:** 203-207.

[47] M. Z. Ludwig and C. Bergman and N. H. Patel and M. Kreitman, 2000 Evidence for stabilizing selection in a eukaryotic enhancer element. Nature **403:** 564-567.

[48] A. Carter and G. Wagner, 2002 Evolution of functionally conserved enhancers can be accelerated in large populations: a population-genetic model. Proc. R. Soc. Lond. **B 269:** 953-960.

[49] G. A. Wray *et al*, 2003 The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20:** 1377-1419.

[50] M. W. Hahn and J. E. Stajich and G. A. Wray, 2003 The effects of selection against spurious transcription factor binding sites. Mol. Biol. Evol. **20:** 901-906.

[51] M. Ptashne and A. Gann, 2002 *Genes and Signals.* Cold Spring Harbour Laboratory Press, Cold Spring Harbour, NY.

[52] M. Ptashne, 1992 *A genetic switch: Phage λ and higher organisms.* Blackwell Science, Malden, MA.

[53] B. Müller-Hill, 1996 *The lac operon.* deGruyter, Berlin.

[54] Zavolan, M., Rajewsky, N.,Socci, N. and Gaasterland, T., 2003 SMASHing regulatory sites in DNA by human-mouse sequence comparisons. Proceedings of the IEEE Conference on Computational Systems Bioinformatics.

[55] Rajewsky, N., Vergassola, M., Gaul, U., Siggia, E. D., 2002 Computational detection of genomic cis-regulatory modules, applied to body patterning in the early Drosophila embryo. BMC Bioinformatics. **3:30**

[56] Jeeninga, R. E., M. Hoogenkamp, M. Armand-Ugon, M. De Baar, K. Verhoef, and B. Berkhout. 2000. Functional differences between the long terminal repeat transcriptional promoters of human immunodeficiency virus type 1 subtypes A through G. J. Virol. **74:** 3740-3751.

[57] Opijnen, T. *et al* , 2004 Human Immunodeficiency Virus Type 1 Subtypes Have a Distinct Long Terminal Repeat That Determines the Replication Rate in a Host-Cell-Specific Manner. J. Virol. **78:** 3675-3683.

[58] Stormo, G. D. , 2000 DNA binding sites: representaion and discovery. BIOINFORMATICS. **716:** 16-23.

[59] Wray, G. A. *et al* , 2003 The Evolution of Transcriptional Regulation in Eukaryotes. Review Article. Mol. Biol. Evol. **20:** 1377-1419.

[60] Brown, C. T. and Callan, C. G. Jr. , 2004 Evolutionary comparisons suggest many novel cAMP response protein binding sites in *Escherichia coli.* PNAS. **101:** 2404-2409.

[61] De Bosscher, K. *et al*, 2000 Glucocorticoids repress NF-κB-driven genes by disturbing the interaction of p65 with the basal transcription machinery, irrespective of coactivator levels in the cell. PNAS. **97:** 3919-3924.

[62] Webster, C. J., Oakley R. H., Jewell, C. M., and Cidlowski J. A., 2001 Proinflammatory cytokines regulate human glucocorticoid receptor gene expression and lead to the accumulation of the dominant negative β isoform: A mechanism for the generation of glucocorticoid resistence. PNAS. **97:** 3919-3924.

[63] Barthel, R. *et al*, 2003 Regulation of Tumor Necrosis Factor Alpha Gene Expression by Mycobacteria Involves the assembly of a Unique Enhancesosome Dependent on the Coactivator Proteins CBP/p300. Mol. Cell. Biol. **23:** 526-533.

[64] Barthel, R. and Goldfeld, A. E., 2003 T Cell-Specific Expression of the Human TNF α Gene Involves a Functional and Highly Conserved Chromatin Signature in Intron 3. J. Immunol. **171:** 3612-3619.

[65] Baena, A., *et al*, 2002, TNF-α promoter single nucleotide polymorphisms are markers of human ancestry. Gen. Immun. **3:** 482-487.

[66] Tsytsykova, A. V. and Goldfeld, A. E., 2002, Inducer-specific Enhanceosome Formation Controls Tumor Necrosis Factor Alpha Gene expression in T Lymphocytes. Mol. Cell. Biol. **22:** 2620-2631.

[67] Tsai, E. Y., *et al*, 2000, A Lipopolysaccharide-Specific Enhancer Complex Involving Ets, Elk-1, Sp1, and CREB Binding Protein and p300 Is Recruited to the Tumor Necrosis Factor Alpha Promoter In Vivo. Mol. Cell. Biol. **20:** 6084-6094.

[68] Leung, J. Y., *et al*, 2000, Identification of phylogenetic footprints in primate tumor necrosis factor-α promoters. PNAS. **97:** 6614-6618.

[69] Falvo, J. V., *et al*, 2000, Stimulus-Specific assembly of Enhancer Complexes on the Tumor Necrosis Factor Alpha Gene Promoter. Mol. Cell. Biol. **20:** 2239-2247.

[70] Ilyinskii, P. O. and Desrosiers, R. C., 1996, Efficient Transcription and Replication of Simian Immunodeficiency Virus in the Absence of NF-κB and Sp1 Binding Elements. J. Virol. **70:** 3118-3126.

[71] Pohlmann, S., *et al*, 1998, Sequences Just Upstream of the Simmian Immunodeficiency Virus Core Enhancer Allow Efficient Replication in the Absence of NF-κB and Sp1 Binding Elements. J. Virol. **72:** 5589-5598.

[72] Hunt, G. M., Johnson, D., and Tiemessen, C. T., 2001, Characterisation of the Long Terminal Repeat Regions of South african Human Immunodeficiency Virus Type 1 Isolates. Virus Genes. **23:** 27-34.

[73] Opijnen, T. van, Kamoschinski, J., Jeeninga, R. E., and Berkhout, B., 2004, The Human Immunodeficiency Virus Type 1 Promoter Contains a CATA Box instead of a TATA Box for Optimal Transcription and Replication. **78:** 6883-6890.

[74] Olsen, G. J., Matsuda, H., Hagstrom, R., and Overbeek, R. 1994, fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comp. Appl. Biosci. **10:** 41-48.

[75] Felsenstein, J. 1981, Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. **17:** 368-376.

[76] Rahmann, S., Muller, T., Vingron, M., 2003, On the Power of Profiles for Transcription Factor Binding Site Detection. Statistical Applications in Genetics and Molecular Biology. *Volume 2, Issue 1, Article 7*

[77] M. Zuker, 1989, On Finding All Suboptimal Foldings of an RNA Molecule. Science, **244:** 48-52.

[78] J. A. Jaeger, D. H. Turner and M. Zuker, 1989, Improved Predictions of Secondary Structures for RNA. Proc. Natl. Acad. Sci. USA, BIOCHEMISTRY, **86:** 7706-7710.

[79] J. A. Jaeger, D. H. Turner and M. Zuker, 1989, Predicting Optimal and Suboptimal Secondary Structure for RNA. in "Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences", R. F. Doolittle ed. Methods in Enzymology, **183:** 281-306.

[80] R Development Core Team, 2003 R: A language and environment for statistical computing, http://www.R-project.org, Vienna, Austria.

[81] Yang, Z., S. Kumar, and M. Nei, 1995 A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641-1650.

[82] D. Graur and W.-H. Li, 2000 *Fundamentals of Molecular Evolution.* Sinauer, Sunderland, MA.

[83] David. W Mount, 2001 *Bioinformatics.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

[84] Rex A. Dwyer, 2003 *Genomic Perl.* Cambridge University Press.

[85] R. Durbin, S. Eddy, A. Krogh, G. Mitchison, 1998 *Biological sequence analysis*. Cambridge University Press.

[86] W. J. Ewans, G. R. Grant, 2002 *Statistical Methods in Bioinformatics*. Springer.

[87] T. A. Brown, 1999 *Genomes*. Wiley.

[88] B. Aggarwal and R. K. Puri, 1995 *Human cytokines: their role in disease and therapy*, Blackwell Science, Cambridge, Mass.

[89] M. Eigen and P. Schuster, 1979 *The Hypercycle: A Principle of Natural Self-Organization*. Springer, Berlin.

[90] N. A. Campbell, 1997 *Biologie*. Spektrum, Akad. Verl.

[91] M. Laessig, L. Peliti, F. Tria, 2003, Evolutionary games and quasispecies, Europhys. Lett., **62:** 446-451.

[92] A. Franciscus, 2004, HCV: Genotype & Quasispecies, hcpFACTsheet, Hepatitis C Support Project (www.hcvadvocate.org), 1-2.

[93] B. Alberts *et al*, 1998, *Essential Cell Biology: An Introduction to the Molecular Biology of the Cell*, Garland science Publishing.

[94] Roche *Genetics* Education Program. www.rochegenetics.com

[95] F. Daly *et al*, 1995, *Elements of Statistics*, Addison Wesley.

[96] Travis, J., 2002, Hot Cereal: Rice reveals bumper crop of genes. Science News. Vol. 161, No. 14, p. 211.

[97] E.W. Lander *et al*, 2001, Initial sequencing and analysis of the human genome, Nature, **409:**860-921.

[98] J.C. Venter *et al*, 2001, The sequence of the human genome, Science, **291:**1304-1351.

[99] Simon G. Gregory, *et al*, 2002, A physical map of the mouse genome ,Nature, **418:** 743-750.

[100] Ranjbar, Rajsbaum and Goldfeld, unpublished.

[101] Ranjbar, Rajsbaum, Tsytsykova and Goldfeld, unpublished.

[102] Ranjbar, Tsytsykova and Goldfeld, unpublished.

[103] J. Berg, S. Willmann and M. Lässig, 2004, Adaptive evolution of transcription factor binding sites, **4:42**. BMC Evolutionary Biology.

[104] Willmann, S., Goldfeld, A., Laessig, M., and Rajewsky, N. , *in preparation*

Acknowledgement

I am indebted to Prof. Michael Lässig for introducing me into interdisciplinary research and excellent supervision. Prof. Lässig's flexibility and openness to new ideas improved this research study. Special thanks go to Prof. Nikolaus Rajewsky who taught me bioinformatics and molecular biology. Professor Rajewsky's competent, patient and professional supervision together with warm hospitality in his lab, at New York University, made an important impact on the results of the thesis. Many thanks go to Dr. Johannes Berg for friendly support during our group's work. Last, but not least, I would like to thank to Prof. Anne Goldfeld at Harvard Medical School for providing us with excellent experimental data, indispensable for the research in our field.
I need to thank staff, postdocs and students in the physics department for the friendly working atmosphere. Also, I must thank to my husband Richard for constant support.

Erklärung

Ich versichere, daß ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; daß diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; daß sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, daß ich eine solche Veröffentlichung vor Abschluß des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Michael Lässig betreut worden.

Teilpublikationen

J. Berg, S. Willmann und M. Lässig
Adaptive evolution of transcription factor binding sites
BMC Evolutionary Biology, 4:42 (2004).