Uniwersytet Warszawski
Wydział Matematyki, Informatyki i Mechaniki

Aleksander Jankowski
Nr albumu: 219452

Predicting nucleosome binding sites in yeast genome

Praca magisterska
na kierunku MATEMATYKA

Praca wykonana pod kierunkiem
prof. Jerzego Tiuryna
Instytut Informatyki

Sierpień 2009

Oświadczenie kierującego pracą

Potwierdzam, że niniejsza praca została przygotowana pod moim kierunkiem i kwalifikuje się do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data Podpis kierującego pracą

Oświadczenie autora (autorów) pracy

Świadom odpowiedzialności prawnej oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami.

Oświadczam również, że przedstawiona praca nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Oświadczam ponadto, że niniejsza wersja pracy jest identyczna z załączoną wersją elektroniczną.

Data Podpis autora (autorów) pracy

Streszczenie

Nucleosomes form the fundamental repeating units of chromatin, which is used to pack large eukaryotic genomes into the nucleus while still ensuring appropriate access to it. Chromatin consists mostly of DNA and proteins, and its structure is inevitably related to the regulation of gene transcription.

In 2009, first genome-wide maps of nucleosome occupancy in vitro and in vivo were obtained for yeast [Kaplan *et al.*, Nature, vol. 458]. Relying on the experimental data, the authors devised a computational model of nucleosome sequence preferences. The model is based on thermodynamical equilibrium, and involves two free parameters, representing nucleosome concentration and inverse temperature.

My thesis is directed towards improvement of this model. I will analyse the impact of the model parameters to overall performance of prediction and compare different ways to estimate model parameters. I will also explain the influence of the individual components on model accuracy.

Słowa kluczowe

nucleosome, DNA binding, yeast, thermodynamical model, dynamic programming, parameter optimisation

Dziedzina pracy (kody wg programu Socrates-Erasmus)

11.1 Matematyka

Klasyfikacja tematyczna

92 Biology and other natural sciences
92C Physiological, cellular and medical topics
92C37 Cell biology

Tytuł pracy w języku angielskim

Predicting nucleosome binding sites in yeast genome

Contents

Chapter 1

Introduction

In my thesis, a computational model on nucleosome binding is being considered. Its application to the available experimental data gives the opportunity to asses its performance and the impact of the model parameters and components to overall accuracy. The detailed organisation of the thesis is described below.

The meaning and role of nucleosomes in eukaryotic cells is commonly known among biologists. However, for a mathematician or a computer scientist, there is a need to explain basic facts in this matter. Chapter 2 briefly presents the biological knowledge required to get the full understanding of my work.

The third and fourth chapter form the theoretical core of the thesis. In chapter 3, the nucleosome binding model formulated by Kaplan *et al.* [1] is introduced. It consists of two components: the position-specific dinucleotide component and the position-independent component. My contribution was to propose several variants of these two components. Although all of the variants have been inspired by [1], they depict diverse features, which were subject to further discussion.

Chapter 4 follows Field *et al.* [2] in describing a thermodynamical model of nucleosome binding. It makes use of scores calculated by the nucleosome binding model from the previous chapter. The thermodynamical model is based on thermodynamical equilibrium. It involves two free parameters, representing nucleosome concentration and inverse temperature.

To assess the accuracy of various model variants on the available data, it is useful to work with the optimal values of the two parameters mentioned above. Chapter 5 introduces two widely known optimisation algorithms, which were applied to this problem: the Newton-type algorithm and the Nelder and Mead algorithm.

Chapter 6 concerns the experimental results of my work. The data was collated from the publicly available *Saccharomyces* Genome Database [3] and from the novel experiments performed by Kaplan *et al.* [1]. The experiments provided genome-wide maps of *in vitro* nucleosome occupancy for yeast.

My contribution was to test performance of different model variants proposed in chapter 3. For each variant, I have preformed an experiment on a small scale, considering a single yeast chromosome. The experiment involved training the model on the data available from [1]. The optimal values of the two free parameters, maximising the correlation between the model prediction and the available nucleosome occupancy map, were found using the algorithms described in chapter 5. For the promising model variants, I have repeated the experiments on a large scale, taking the whole yeast genome.

In chapter 7, some final conclusions are given on the impact of the model variant and parameters to overall performance of prediction. Different ways to estimate model parameters and the influence of the individual components on model accuracy are also discussed. The concluding remarks propose possible improvements on the nucleosome binding model.

I am very grateful to Dr. Shyam Prabhakar from the Genome Institute of Singapore for his helpful suggestions. He proposed the relative (R) variant of the position-specific dinucleotide component, which happened to be very robust.

Last but not least, I am very thankful to my thesis advisor, Prof. Jerzy Tiuryn, not only for proposing more and more brave new concepts for my thesis, but for his assistance, invaluable discussions on results and helpful suggestions as well.

Chapter 2

Nucleosomes and their binding

2.1. Nucleosome structure and function

Nucleosomes form the fundamental repeating units of chromatin, which is used to pack large eukaryotic genomes into the nucleus while still ensuring appropriate access to it. Chromatin is the heterogeneous substance, consisting mostly of DNA and proteins, that makes up chromosomes in eukaryotic cells. Its structure is inevitably related to the regulation of gene transcription, allowing an easier access to regulatory regions.

The internal structure of chromosomes involves folding the DNA double helix several times to make it more compact, and nucleosomes are the first level of their structure. Alberts *et al.* [4, p. 208] explains the nucleosome organisation as following:

> The structural organization of nucleosomes was determined after first isolating them from unfolded chromatin by digestion with particular enzymes (called nucleases) that break down DNA by cutting between the nucleosomes. After digestion for a short period, the exposed DNA between the nucleosome core particles, the linker DNA, is degraded. Each individual nucleosome core particle consists of a complex of eight histone proteins – two molecules each of histones H2A, H2B, H3, and H4 – and double-stranded DNA that is 146 nucleotide pairs long. The histone octamer forms a protein core around which the double-stranded DNA is wound.

Some other sources, e.g. Lodish *et al.* [5], refer to 147 nucleotide pairs as the nucleosome length; it is the matter of convention. The DNA is wrapped around the histone core of the nucleosome in about $1\frac{2}{3}$ left-handed superhelical turns. The histone octamer enforces a very regular, repetitive and symmetric structure of nucleosomes, presented on Fig. 2.1.

The nucleosome structure is identical in all the known eukaryotic species living on Earth. Alberts *et al.* [4] points out that "...the histones are among the most highly conserved eucaryotic proteins. For example, the amino acid sequence of histone H4 from a pea and a cow differ at only at 2 of the 102 positions."

Repeating nucleosomes are interlaced with "linker" DNA fragments of length 5-80 base pairs (bp), usually about 50 bp. The structure formed by nucleosomes and linkers together is known descriptively as "beads on a string".

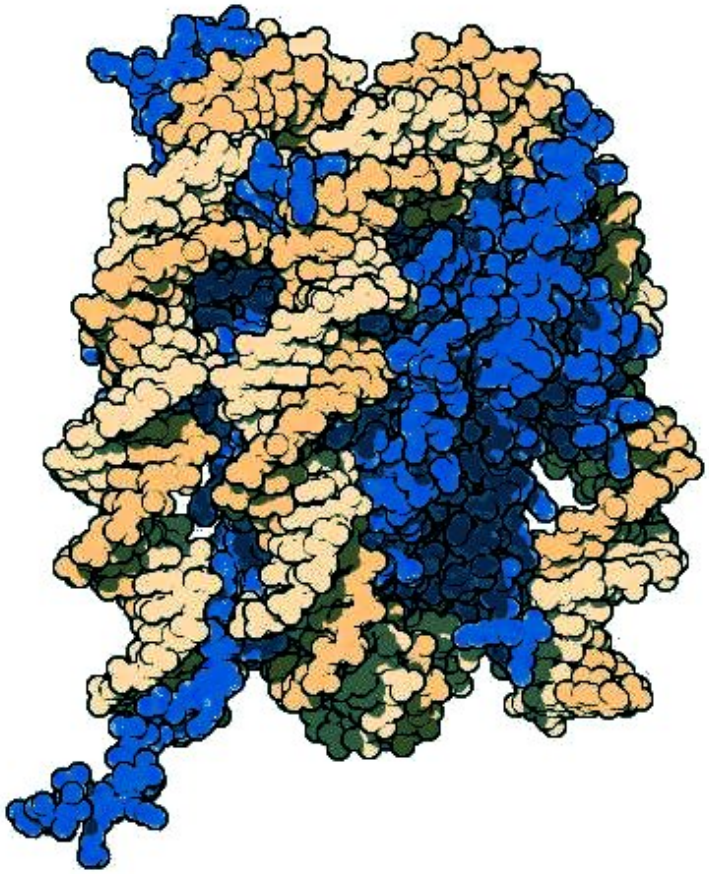


Figure 2.1: Nucleosome: DNA sequence (light orange) wrapped around histone core (dark blue). Illustration in public domain, by David S. Goodsell of The Scripps Research Institute.

2.2. Data sources on nucleosome binding

In 2009, first genome-wide maps of nucleosome occupancy *in vitro* and *in vivo* were obtained for yeast by Kaplan *et al.* [1].

A novel concept has been applied to obtain *in vitro* nucleosome occupancy, governed solely by nucleosome sequence preferences. To this end, purified chicken histone octamers were assembled on purified yeast genomic DNA in fluently changing biochemical environment.

In both the *in vitro* and *in vivo* cases, the nucleosomes were separated off the linkers using a DNA-cutting enzyme – micrococcal nuclease. The nucleosome-bound sequences have been unwound and sequenced.

In my experiments, I have used the *in vitro* nucleosome occupancy map, consisting of 9,313,383 reads of nucleosome-bound sequences of length 147, uniquely mapped to the yeast genome. The dataset involved 5,319,670 different sequence fragments.

It is noteworthy that the *in vitro* and *in vivo* maps are highly similar, but they contain definite differences in genomic regions related to gene transcription, such as transcription start sites.

Chapter 3

Nucleosome binding model

We represent the sequence binding preferences of nucleosomes by devising a probabilistic model. It will assign a numeric score to every genomic sequence of length 147 bp, which we assume to be the nucleosome length.

It is observed that the sequence binding preferences differ along the nucleosome, due to the periodic nature of DNA double helix. The first component will capture the periodic signal of nucleotide pairs (called dinucleotides) along the nucleosome.

The second component will capture the overall, position-independent nucleosome binding preferences. We will consider DNA sequences of length 5 (called 5-mers), and consider whether they are favoured or disfavoured by nucleosomes.

3.1. Position-specific dinucleotide component

We introduce this component in order to capture the periodic signal of dinucleotides along the nucleosome. Let us assume for a while that we have fixed statistical weights $N_i(pq)$ for each $1 \leq i \leq 146$, $p, q \in \Sigma = \{A, C, G, T\}$. We will consider $N_i(pq)$ as the probability of encountering dinucleotide pq at i-th position of the nucleosome. Thus $N_i(\cdot)$ is a probability distribution.

We can generalise the definition of N_i to include single nucleotides (also called mononucleotides). For each $1 \leq i \leq 146$ and $p \in \Sigma$, we put $N_i(p) = \sum_{q \in \Sigma} N_i(pq)$. Let us assume that S is a genomic sequence of length 147. Then we define the position-specific dinucleotide component $\mathcal{N}$ as follows:

$$\mathcal{N}(S) = N_1(S_1) \cdot \prod_{i=1}^{146} \frac{N_i(S_i S_{i+1})}{N_i(S_i)} = N_1(S_1 S_2) \cdot \prod_{i=2}^{146} \frac{N_i(S_i S_{i+1})}{N_i(S_i)} \tag{3.1}$$

The fractions in the preceding equation are natural representations of the conditional probability of observing S_{i+1} at the $(i+1)$-th position of the nucleosome, given the occurrence of S_i at the i-th position.

The weights $N_i(pq)$ have been estimated building on *in vitro* reads of nucleosome-bound sequences. Due to the two-fold symmetry in the nucleosome structure, each sequence has been included twice, once in its original form, and once in its reverse compliment form.

Table 3.1 explains how to make use of these data to estimate six variants of $N_i(pq)$, with abbreviated names N, NS, D, DS, R, RS. Moreover, a null variant, named Z, is defined such that $N_i(pq) = 1/16$ for each i and p, q.

For each $1 \leq i \leq 146$ and $p, q \in \Sigma = \{A, C, G, T\}$, let $\#_i(pq)$ be the empirical number of occurrences of dinucleotide pq at position i.

No smoothing	Smoothing (**S**)
	For each $2 \leq i \leq 145$ and $p, q \in \Sigma$, let $$\#'_i(pq) = \frac{\#_{i-1}(pq) + \#_i(pq) + \#_{i+1}(pq)}{3},$$ $\#'_1(pq) = (\#_1(pq) + \#_2(pq))/2$, $\#'_{146}(pq) = (\#_{145}(pq) + \#_{146}(pq))/2$. Now we substitute each $\#_i(pq)$ with $\#'_i(pq)$.

Natural variant (**N**)	Double-normalised variant (**D**)	Relative variant (**R**)
For $1 \leq i \leq 146$ and $p, q \in \Sigma$, we put $n_i(pq) = \#_i(pq)$.	For $1 \leq i \leq 146$ and $p, q \in \Sigma$, we put $$n_i(pq) = \frac{\#_i(pq)}{\sum_{j=1}^{146} \#_j(pq)}.$$	For $1 \leq i \leq 146$ and $p, q \in \Sigma$, we put $$n_i(pq) = \frac{\#_i(pq)}{\mathrm{Overall}(pq)},$$ where $\mathrm{Overall}(\cdot)$ is the genome-wide probability distribution of dinucleotides.

Normalise $n_i(pq)$ to a probability distribution:

$$N_i(pq) = \frac{n_i(pq)}{\sum_{r,s \in \Sigma} n_i(rs)}.$$

Table 3.1: Main workflow to estimate statistical weights $N_i(pq)$.

Following the above workflow, we are able to estimate six variants of $N_i(pq)$, denoted by abbreviations N, D, R (not smoothed) and NS, DS, RS (smoothed). Smoothing is done due to the observation that ± 1 bp shift of nucleosome positions may occur due to the experimental process.

In further analysis, only the central 127 bp of the nucleosomes will be used, and thus we put $N_i(pq) = 1/16$ for $i \in \{1, 2, \ldots, 10, 138, 139, \ldots, 146\}$ and each $p, q \in \Sigma$. In this way, we avoid the sequence biases occuring at the micrococcal nuclease cut sites.

Note that the explained procedure always generates a reverse complement symmetrical distribution, that is, assuming that sequences pq and rs are reverse complement, we have $N_i(pq) = N_i(rs)$ for each i. Therefore the position-specific dinucleotide component involves $(\frac{4}{2} + 4) \cdot (127 - 1) = (6 + 4) \cdot 126 = 1260$ parameters.

3.2. Position-independent component

We introduce this component to represent sequence fragments that are generally favoured or disfavoured by nucleosomes, regardless of their position within the nucleosome. In our model, we will assume that we have fixed statistical weights $L(p_1 \ldots p_5)$ for each 5-mer $p_1, \ldots, p_5 \in \Sigma$.

We will consider $L(p_1 \ldots p_5)$ as the probability of encountering a given 5-mer $p_1 \ldots p_5$ in the nucleosome-bound sequence. Thus $L(\cdot)$ is a probability distribution.

Following the idea from the previous section, we can generalise the definition of L to include 4-mers. For $p_1, \ldots, p_4 \in \Sigma$, we put $L(p_1 \ldots p_4) = \sum_{q \in \Sigma} L(p_1 \ldots p_4 q)$. Let us assume that S is a genomic sequence of length 147. Then we define the position-independent component $\mathcal{L}$ as follows:

$$\begin{aligned} \mathcal{L}(S) &= L(S_1) \cdot \frac{L(S_1S_2)}{L(S_1)} \cdot \frac{L(S_1S_2S_3)}{L(S_1S_2)} \cdot \frac{L(S_1 \ldots S_4)}{L(S_1S_2S_3)} \cdot \frac{L(S_1 \ldots S_5)}{L(S_1 \ldots S_4)} \cdot \prod_{i=2}^{143} \frac{L(S_i \ldots S_{i+4})}{L(S_i \ldots S_{i+3})} \\ &= L(S_1 \ldots S_5) \cdot \prod_{i=2}^{143} \frac{L(S_i \ldots S_{i+4})}{L(S_i \ldots S_{i+3})} \end{aligned} \tag{3.2}$$

The fraction in the preceding equation is a natural representation of the conditional probability of observing S_{i+4} at the $(i+4)$-th position of the nucleosome, given the occurrence of $S_i \ldots S_{i+3}$ at the i-th position.

The weights $L(p_1 \ldots p_5)$ have been estimated building on *in vitro* reads of nucleosome-bound sequences. For each $p_1 \ldots p_5 \in \Sigma$, let $\#_i(p_1 \ldots p_5)$ be the empirical number of occurrences of 5-mer $p_1 \ldots p_5$ in nucleosome reads. The following four variants of $L(p_1 \ldots p_5)$ have been tried:

1. Natural (**N**):
$$L(p_1 \ldots p_5) = \frac{\#(p_1 \ldots p_5)}{\sum_{q_1, \ldots, q_5 \in \Sigma} \#(q_1 \ldots q_5)}.$$

2. Literally as specified in Kaplan *et al.* [1] (**P**):
$$L(p_1 \ldots p_5) = \frac{\frac{1}{\#(p_1 \ldots p_5)}}{\sum_{q_1, \ldots, q_5 \in \Sigma} \frac{1}{\#(q_1 \ldots q_5)}}.$$

3. Relative (**R**):

$$L(p_1 \dots p_5) = \frac{\frac{\text{Overall}_{(p_1 \dots p_5)}}{\#(p_1 \dots p_5)}}{\sum_{q_1,\dots,q_5 \in \Sigma} \frac{\text{Overall}_{(q_1 \dots q_5)}}{\#(q_1 \dots q_5)}},$$

where Overall$(\cdot)$ is the genome-wide probability distribution of 5-mers.

4. Null (**Z**), defined such that $L(p_1 \dots p_5) = (\frac{1}{4})^5$.

Following the argument from the previous section, only central 127 bp of the nucleosomes will be used to calculate empirical number of 5-mer occurrences $\#_i(p_1 \dots p_5)$.
The position-independent component involves $4^5 = 1024$ parameters.

3.3. Overall binding score

Overall binding score has been calculated as

$$\text{Score}(S) = \ln \frac{\mathcal{N}(S)}{\mathcal{L}(S)}, \tag{3.3}$$

for each genomic sequence S of length 147.

Chapter 4

Thermodynamical model

In the previous chapter, we have explained how to assign nucleosome binding score to every genomic sequence of length 147. We then use these scores to compute the genome-wide nucleosome occupancy map, taking into account steric hindrance constraints between neighbouring nucleosomes.

Let us consider genomic sequence S of any length. We may think of S the whole chromosome. In fact, in our experiments we will exclude highly repetitive genomic regions, thus consider S as a maximal continuous chromosome fragment not including highly repetitive regions. To avoid possible bias caused by boundary effects, the length M of sequence S should be at least one order of magnitude larger than the fixed nucleosome length (147).

For the reasons of clarity, we simply assume that $M \geq 147$. If $M < 147$, then for sure there is no nucleosome bound to S.

Let $\mathcal{C}$ be the space of all legal configurations of nucleosomes on a sequence S, where a legal configuration is a set of 147 bp nucleosomes on S, represented by their start positions, such that no two nucleosomes overlap. For each configuration $c \in \mathcal{C}$, consisting of k nucleosomes of start positions $c[1], \ldots, c[k]$, we assign the statistical weight

$$W_c[S] = \prod_{i=1}^{k} \tau \cdot \exp(\beta \cdot \text{Score}(S[c[i] \ldots c[i] + 146])), \tag{4.1}$$

where τ and β are fixed parameters, and $S[k \ldots l]$ denotes subsequence of S from position k to position l, inclusively.

Parameter τ may be considered as nucleosome concentration, and β as inverse temperature. Assuming the Boltzmann distribution on $\mathcal{C}$, we can estimate the probability of each configuration $c \in \mathcal{C}$ in the following way:

$$P(c|S) = \frac{W_c[S]}{\sum_{c' \in \mathcal{C}} W_{c'}[S]}. \tag{4.2}$$

We may try to find out the configuration $c \in \mathcal{C}$ that maximises $P(c|S)$. However, it seems to be quite hard to avoid the exponential cost of looking through all the configurations from $\mathcal{C}$. Moreover, sticking to the most probable configuration may be misleading, because we may miss a subset of similar configurations having high cumulative probability.

Therefore we will take a different approach. Our goal will be to calculate, for each position on S, the average nucleosome occupancy of it, defined as the probability of covering this

position by any nucleosome. At first, we use a dynamic programming method to compute the probability of placing a nucleosome that starts at each position on S.

The most important observation is that the probability of placing a nucleosome starting at a particular position i is equal to the sum of the statistical weights W of all configurations in which a nucleosome starts at position i, divided by the sum $\sum_{c\in\mathcal{C}} W_c[S]$ of the statistical weights of all legal configurations of nucleosomes on S. Both of these sums can be computed efficiently in three effective steps:

1. **Forward step:** we compute a set of variables $F_1, \ldots, F_M$, where F_i represents the sum of the statistical weights of all legal configurations of the subsequence $S_1, \ldots, S_i$, as follows:

$$\begin{aligned} F_0 &:= 1 \quad \text{(for completeness)}, && (4.3)\\ F_i &:= F_{i-1} \quad \text{for } 1 \le i \le 146, && (4.4)\\ F_i &:= F_{i-1} + F_{i-147} \cdot \tau \cdot \exp(\beta \cdot \mathrm{Score}(S[i-146 \ldots i])) \\ &\qquad \text{for } 147 \le i \le M. && (4.5) \end{aligned}$$

   Due to the equation 4.1, for empty configuration of nucleosomes c_0, we have $W_{c_0}[S] = 1$ and thus $F_i = 1$ for $1 \le i \le 146$. The formula 4.5 incorporates the fact that every configuration c of nucleosomes on the subsequence $S_1, \ldots, S_i$ satisfies exactly one of the two following conditions:

   (a) c contains no nucleosome at $S[i]$, and thus can be considered as a configuration of nucleosomes on subsequence $S_1, \ldots, S_{i-1}$

   (b) c contains nucleosome at $S[i-146 \ldots i]$ (and on subsequence $S_1, \ldots, S_{i-147}$ any legal configuration of nucleosomes may occur).

2. **Reverse step:** we compute a set of variables $R_1, \ldots, R_M$, where R_i represents the sum of the statistical weights of all legal configurations of the subsequence $S_i, \ldots, S_M$, as follows:

$$\begin{aligned} R_{M+1} &:= 1 \quad \text{(for completeness)}, && (4.6)\\ R_i &:= R_{i+1} \quad \text{for } M-145 \le i \le M, && (4.7)\\ R_i &:= R_{i+1} + R_{i+147} \cdot \tau \cdot \exp(\beta \cdot \mathrm{Score}(S[i \ldots i+146])) \\ &\qquad \text{for } 1 \le i \le M-146. && (4.8) \end{aligned}$$

3. **Aggregation step:** First observe that by definition of F_i and R_i,

$$F_M = R_1 = \sum_{c\in\mathcal{C}} W_c[S]. \tag{4.9}$$

   We can now easily compute the probability $P_i[S]$ of placing a nucleosome starting at a particular position $1 \le i \le M-146$ of S:

$$P_i[S] = \frac{F_{i-1} \cdot \tau \cdot \exp(\beta \cdot \mathrm{Score}(S[i \ldots i+146])) \cdot R_{i+147}}{R_1}. \tag{4.10}$$

   The average nucleosome occupancy, defined as the probability of covering a particular position i by any nucleosome, may now be calculated as the sum of probabilities of starting a nucleosome at any of the positions from $i-146$ to i, that is

$$\sum_{k=0}^{146} P_{i-k}[S]. \tag{4.11}$$

Chapter 5

Optimisation algorithms

To estimate the parameters τ and β of the thermodynamical model, some sort of optimisation algorithm must be applied. In this chapter, the two used algorithms will be described: the Newton-type algorithm and the Nelder and Mead algorithm. Both of them are widely used in numerical computations.

Traditionally, the optimisation algorithms are formulated as minimisation problems, so in this chapter we will follow this point of view. It is obvious that by changing the sign of the objective function, they can be easily adopted to solve maximisation problems.

Both of the used algorithms are described in detail by Kincaid *et al.* [6].

5.1. Newton-type algorithm

The widely known Newton method of finding roots of a given differentiable function $f\colon \mathbb{R} \to \mathbb{R}$ starts with an initial guess $x^{(0)}$ and iteratively calculates

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})}. \tag{5.1}$$

If $x^{(n+1)} \neq x^{(n)}$, then

$$f'(x^{(n+1)}) = \frac{f(x^{(n)}) - 0}{x^{(n)} - x^{(n+1)}} \tag{5.2}$$

and it follows from the geometrical interpretation of the derivative that $x^{(n+1)}$ is the x-axis-intercept of the tangent line to f in $x^{(n)}$. If the process converges, that is $x^{(n)} \to x^*$, then the point x^* satisfies $f(x^*) = 0$.

Assuming that $f \in C^1$, we can apply the Newton method to find the local maxima and minima of f. In such a local extremum, the first derivative of f is zero, and we may iteratively find points satisfying $f'(x^*) = 0$:

$$x^{(n+1)} = x^{(n)} - \frac{f'(x^{(n)})}{f''(x^{(n)})}. \tag{5.3}$$

Let us consider a multi-dimensional problem of finding the local maxima and minima of $f\colon \mathbb{R}^k \to \mathbb{R}$ and assume that $f \in C^2$. We are looking for $x_1^*, \ldots, x_k^*$ satisfying the system of equations

$$\frac{\partial f(x_1^*, \ldots, x_k^*)}{\partial x_i} = 0 \qquad \text{for each } i = 1, \ldots, k. \tag{5.4}$$

To this end, start with an initial guess $x_1^{(0)}, \ldots, x_k^{(0)}$ and in the similar way as in Eq. 5.3, iteratively calculate

$$x_i^{(n+1)} = x_i^{(n)} - \frac{\frac{\partial f}{\partial x_i}(x_1^{(n)}, \ldots, x_k^{(n)})}{\frac{\partial^2 f}{\partial x_i^2}(x_1^{(n)}, \ldots, x_k^{(n)})} \qquad \text{for each } i = 1, \ldots, k. \tag{5.5}$$

If the process converges, that is $x_i^{(n)} \to x_i^*$ for each coordinate $i = 1, \ldots, k$, then the point $x^* = (x_1^*, \ldots, x_k^*)$ satisfies

$$\left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_k}\right)(x^*) = \nabla f(x^*) = 0. \tag{5.6}$$

It means that x^* may be *a good candidate* for a local maximum or a local minimum. Precisely speaking, if x^* is a local extremum of f, then $\nabla f(x^*) = 0$, but the opposite is not true.

In appliances, all the first and second derivatives are numerically approximated, so the method is computationally demanding. However, the underlying theory may give some reliability for the well-behaved functions.

5.2. Neder and Mead algorithm

The Nelder and Mead algorithm, originally described in [7], is designed to solve the optimisation problem of minimising a given function $f: \mathbb{R}^n \to \mathbb{R}^n$. This method belongs to the *direct search methods*, because it uses only the function values at some points in $\mathbb{R}^n$, and does not try to calculate an approximate gradient in any of these points.

The algorithm uses the concept of *simplex*, which is defined as a convex hull of $n+1$ vertices in n-dimensional space $\mathbb{R}^n$. The simplest simplexes are: a line segment in $\mathbb{R}$, a triangle in $\mathbb{R}^2$, a tetrahedron in $\mathbb{R}^3$.

The basic description of the algorithm is given below. In the following subsections more detail is given on the most important aspects of it.

The algorithm begins with a set of $n+1$ points $x_0, \ldots, x_n \in \mathbb{R}^n$ that are considered as the vertices of a working simplex S. The corresponding function values $f(x_i)$ are calculated for each initial or changed vertex and stored for further use. The initial working simplex S has to be non-degenerate, meaning that their points must not lie in the same hyperplane.

The main part of the algorithm is to perform a sequence of transformations of the working simplex S, aimed at decreasing the function values at its vertices. At each step, the transformation is determined by computing several test points, together with their function values. The function values at the test points are compared with those at the simplex vertices. To this end, a new set of simplex vertices is selected.

The process is successfully terminated when the working simplex S becomes sufficiently small in some sense, or when the function values at the simplex vertices are close enough.

5.2.1. Initial simplex

The frequent choice is to start with a given input point x_0 and to set up the remaining initial n vertices as

$$x_i = x_0 + h_i e_i \tag{5.7}$$

for each $i = 1, \ldots, n$, where e_i is a unit vector in i-th dimension in $\mathbb{R}^n$, and h_i are step sizes in the respective dimensions. Often $h_i = 1$ is assumed by default.

5.2.2. Simplex transformation

The following steps are repeated until one of the termination tests is satisfied. The steps 3-6 have been depicted on Fig. 5.1-5.5, with images acquired from Singer *et al.* [8].

1. **Ordering**: Determine the indices h, s, l of the *worst*, *second worst* and the *best* vertex in the working simplex S, respectively.

2. **Calculate centroid**: Calculate the centroid c of the *best side* of simplex S – that is, the side opposite *worst* vertex x_h:

$$c = \frac{1}{n} \sum_{i \neq h} x_i. \tag{5.8}$$

   In the next three steps, we will try to substitute the *worst* vertex x_h with a better point laying on the line defined by x_h and c. If all the three steps should fail, we will compute n new vertices of the simplex, shrinking it towards the *best* vertex x_l.

3. **Reflection**: Compute the reflection point $x_r = c + \alpha \cdot (c - x_h)$. Usually $\alpha = 1$ is assumed.

   If the reflected point is better than the second worst, but not better than the best ($f(x_l) \leq f(x_r) < f(x_s)$), substitute the vertex x_h with x_r and go to step 1.

4. **Expansion**: If the reflected point is better than the best ($f(x_r) < f(x_l)$), compute the expansion point $x_e = c + \gamma \cdot (x_r - c)$. Usually $\gamma = 2$ is assumed.

   If $f(x_e) < f(x_r)$, substitute the vertex x_h with the expanded point x_e, otherwise substitute it with the reflected point x_r. In both cases, go to step 1.

   The greedy approach to minimisation ensures that the better of the two points x_r, x_e is included in the new simplex. Moreover, the simplex is expanded if and only if $f(x_e) < f(x_r) < f(x_l)$, what helps to keep its size small.

5. **Contraction**: Now it is certain that $f(x_r) \geq f(x_s)$. Compute the contraction point x_c by using the better of two points x_h and x_r:

   - **Contraction outside**: if $f(x_r) < f(x_h)$, compute $x_c = c + \beta \cdot (x_r - c)$. If $f(x_c) < f(x_r)$, substitute the vertex x_h with the contracted point x_c and go to step 1.
   - **Contraction inside**: if $f(x_r) \geq f(x_h)$, compute $x_c = c + \beta \cdot (x_h - c)$. If $f(x_c) < f(x_h)$, substitute the vertex x_h with the contracted point x_c and go to step 1.

   Usually $\beta = \frac{1}{2}$ is assumed.

6. **Reduction**: When none of the above steps succeeded in substitution of the *worst* vertex x_h, we will substitute all but the *best* vertex. For each $i = 0, 1, \ldots, n$, satisfying $i \neq l$, substitute x_i with $x_i + \delta \cdot (x_i - x_l)$. Usually $\delta = \frac{1}{2}$ is assumed.

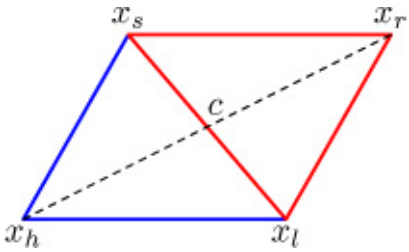


Figure 5.1: Reflection.

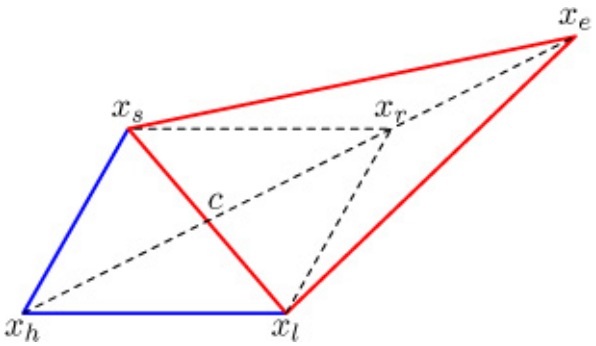


Figure 5.2: Expansion.

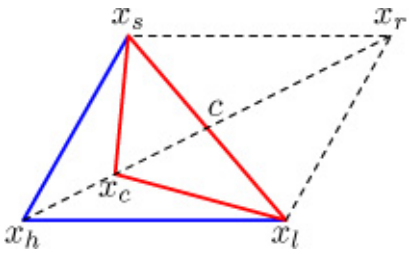


Figure 5.3: Contraction outside.

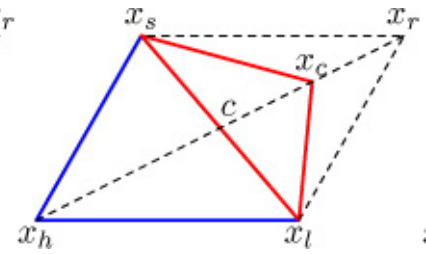


Figure 5.4: Contraction inside.

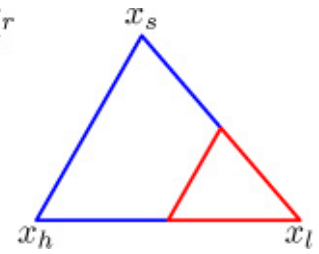


Figure 5.5: Reduction.

5.2.3. Termination tests

Three different termination tests are used concurrently. If at least one of them is satisfied, the algorithm will finish.

- **Domain**: if the working simplex S is sufficiently small in some sense, the algorithm terminates with success.
- **Function value**: if the function values $f(x_i)$ are close enough in some sense, the algorithm terminates with success.
- **No convergence**: if the number of iterations or function evaluations exceeds a given limit, the algorithm terminates with failure.

5.3. Comparison of optimisation algorithms

The main advantage of the Newton-type algorithm is the underlying theory, ensuring the existence of solutions for a wide class of functions. Nevertheless, while optimising the fitting of the model to the experimental data, we cannot expect the likelihood function to be well-behaved or even continuous.

In such a case the Nelder and Mead algorithm is more appropriate. It is designed to work reasonably well for non-differentiable functions. Because it uses only the function values, and does not try to estimate the derivatives, it is more robust.

It is important to recall that both of the described algorithms act locally. Therefore, they are finding local extrema, which need not to be global ones.

Chapter 6

Results

6.1. Performed experiments

We devised the nucleosome binding model using the *in vitro* nucleosome occupancy data for yeast, provided by Kaplan *et al.* [1] and described in section 2.2. The yeast genome data has been acquired from *Saccharomyces* Genome Database [3].

In the whole analysis, we have excluded highly repetitive genomic regions and their 150-bp vicinity, leaving out 10.7% of the whole yeast genome. The information on the repetitive regions was provided by Kaplan *et al.* [1],

Fig. 6.1 shows the amount of each nucleotide along the nucleosome reads. The anomalies at the ends of nucleosome-bound sequences are thought to be caused by the micrococcal nuclease specificity, and due to them, only the central 127 bp of the nucleosome reads will be used. Please note that the GC-content of yeast is about 38%.

The first stage was to estimate all the variants of position-specific dinucleotide component and position-independent component described in chapter 3.

Fig. 6.2-6.4 present the three unsmooothed position-specific nucleotide components (excluding the null component Z). For each of them, the joint weights for dinucleotides consisting only of adenine and thymine and for dinucleotides consisting only of cytosine and guanine were plotted. According to Lodish *et al.* [5], there are about 10.5 base pairs per DNA double helix turn, and this periodicity is clearly visible.

Table 6.1 shows the Pearson correlation between different variants of position-specific nucleotide component, considered as vectors of 1260 parameters. We may notice that smoothing does not introduce any significant change; the latter discussed results will confirm it.

	D	**DS**	**N**	**NS**	**R**	**RS**
D	**100.0%**	99.6%	**16.2%**	15.9%	**34.7%**	34.0%
DS	99.6%	100.0%	16.1%	16.0%	34.5%	34.1%
N	**16.2%**	16.1%	**100.0%**	100.0%	**-86.2%**	-86.5%
NS	15.9%	16.0%	100.0%	100.0%	-86.4%	-86.7%
R	**34.7%**	34.5%	**-86.2%**	-86.4%	**100.0%**	99.9%
RS	34.0%	34.1%	-86.5%	-86.7%	99.9%	100.0%

Table 6.1: Pearson correlation between different variants of position-specific nucleotide component, considered as vectors of 1260 parameters.

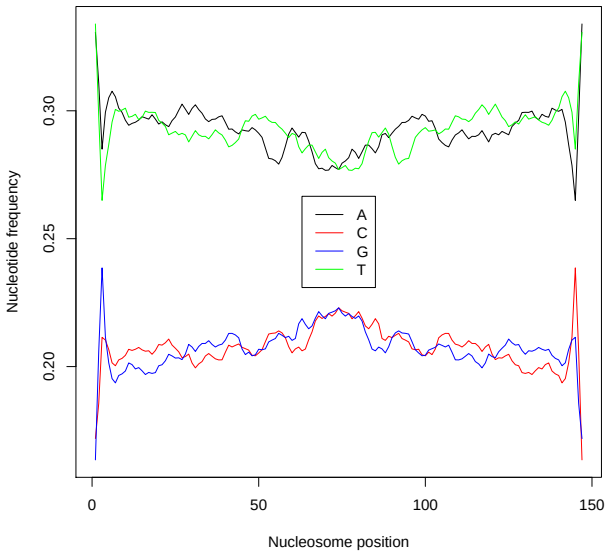


Figure 6.1: Nucleotide frequencies along the nucleosome.

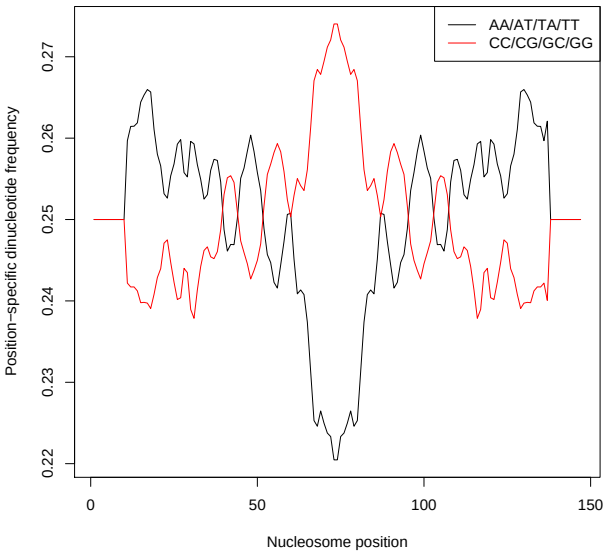


Figure 6.2: Position-specific dinucleotide frequency in component **D** (double-normalised).

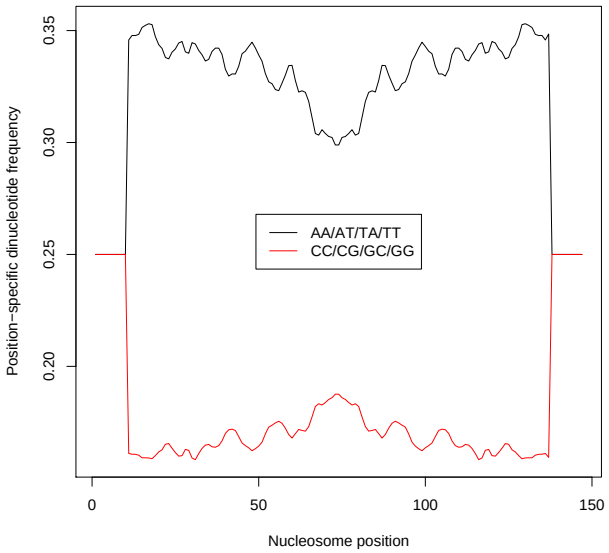


Figure 6.3: Position-specific dinucleotide frequency in component **N** (natural).

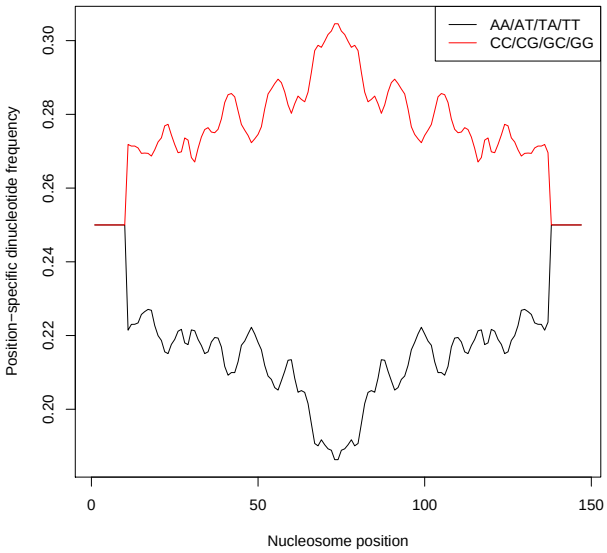


Figure 6.4: Position-specific dinucleotide frequency in component **R** (relative).

	N	P	R
N	100.0%	-80.2%	70.5%
P	-80.2%	100.0%	-50.7%
R	70.5%	-50.7%	100.0%

Table 6.2: Pearson correlation between different variants of position-independent component, considered as vectors of 1024 parameters.

Table 6.2 presents the analogous correlation for position-independent component. We may notice the outstandance of variant P. In fact, it will clearly become useless.

6.2. Experiment results for single chromosome

For all the combinations of component variants described in chapter 3, the experiments on a small scale were performed, using only yeast chromosome 1, i.e. 1.9% of the yeast genome. For fixed values of thermodynamical parameters τ and β, the predicted average nucleosome occupancy has been calculated.

Afterwards, the following log-transformation has been applied: for the vector of nucleosome occupancies, we took the binary logarithm of its values and set the vector mean to zero by adding a constant to all of them. It has been done to allow comparison to results achieved by Kaplan *et al.* [1]. The maximised function was the Pearson correlation coefficient between predicted log-transformed average nucleosome occupancy and the log-transformed *in vitro* data.

We have performed optimisation to estimate the optimal values of free parameters τ and β. Two optimisation procedures described in chapter 5 were used, namely a Newton-type algorithm and Nelder and Mead algorithm. The results are shown in table 6.3.

The results confirm that the variant N of the position-specific dinucleotide component and the variant P of position-independent component are rather worthless. Moreover, we should notice that the smoothing does not improve the results in the remaining cases, so in the whole-genome analysis we abandoned considering smoothing.

It is also important to point out that the two used optimisation algorithms gave very similar results in cases where the model performance is acceptable. It means that the maximised Pearson correlation is well-behaved in terms of free parameters τ and β.

6.3. Experiment results for the whole genome

The experiments, which gave acceptable results on a small scale, were repeated on a large scale, using the whole yeast genome. Due to the high similarity of results for the two optimisation algorithms used for single chromosome analysis, for the whole genome only the more robust Nelder and Mead algorithm has been used. Moreover, no smoothing has been used for position-specific component. The results are presented in table 6.4.

Moreover, Fig. 6.5-6.12 explain how the predictive capabilities of the model varies with changing its parameters. For each combination of variants considered, the two 3D plots and contour plot contain the same data. The optimal values of τ and β were marked on the contour plot.

The overall performance of the model agrees well with the results gained by Kaplan *et al.* [1]. He claims to have per-base-pair nucleosome occupancy correlation in the full model of 88.0%,

Position-independent component ($\mathcal{L}$)		Position-specific dinucleotide component ($\mathcal{N}$)			
		D	**N**	**R**	**Z**
	N	$\tau = 0.0107$ (0.0107) $\beta = 0.3237$ (0.3233) corr. **62.5%** (62.5%)	$\tau = 0.2823$ (0.2823) $\beta = 1.64 \cdot 10^{-9}$ ($1.64 \cdot 10^{-9}$) corr. **1.0%** (1.0%)	$\tau = 0.0277$ (0.0277) $\beta = 0.2913$ (0.2913) corr. **73.7%** (73.7%)	$\tau = 0.0102$ $\beta = 0.3269$ corr. **62.0%**
		$\tau = 0.0107$ (0.0106) $\beta = 0.3236$ (0.3233) corr. 62.5% (62.5%)	$\tau = 0.5811$ (0.5844) $\beta = 4.16 \cdot 10^{-7}$ ($4.02 \cdot 10^{-8}$) corr. 0.9% (0.9%)	$\tau = 0.0277$ (0.0277) $\beta = 0.2914$ (0.2913) corr. 73.7% (73.7%)	$\tau = 0.0102$ $\beta = 0.3269$ corr. 62.0%
	P	$\tau = 0.2038$ (0.0842) $\beta = 6.7007$ (6.7569) corr. **-5.4%** (-5.4%)	$\tau = 1.9705$ (1.9694) $\beta = 4.8304$ (4.8311) corr. **-8.9%** (-8.9%)	$\tau = 17.089$ (19.133) $\beta = 1.0958$ (1.0772) corr. **-4.0%** (-4.0%)	$\tau = 1.18 \cdot 10^{16}$ $\beta = 4.3289$ corr. **-3.2%**
		$\tau = 1.2968$ (1.3402) $\beta = 6.5883$ (6.5888) corr. -5.5% (-5.4%)	$\tau = 1.9692$ (1.9748) $\beta = 4.8303$ (4.8312) corr. -8.9% (-8.9%)	$\tau = 16.993$ (19.170) $\beta = 1.0964$ (1.0772) corr. -4.0% (-4.0%)	$\tau = 2.0842$ $\beta = 6.5919$ corr. -6.2%
	R	$\tau = 0.0123$ (0.0123) $\beta = 0.8301$ (0.8300) corr. **91.5%** (91.5%)	$\tau = 2.35 \cdot 10^{10}$ ($1.94 \cdot 10^{10}$) $\beta = 3.6483$ (3.6442) corr. **0.5%** (0.6%)	$\tau = 0.0289$ (0.0289) $\beta = 0.4862$ (0.4862) corr. **91.6%** (91.6%)	$\tau = 0.0118$ $\beta = 0.8075$ corr. **90.7%**
		$\tau = 0.0123$ (0.0123) $\beta = 0.8302$ (0.8301) corr. 91.5% (91.5%)	$\tau = 7274$ (7140) $\beta = 1.8628$ (1.8566) corr. -0.8% (-0.8%)	$\tau = 0.0289$ (0.0289) $\beta = 0.4863$ (0.4863) corr. 91.6% (91.6%)	$\tau = 0.0118$ $\beta = 0.8074$ corr. 90.7%
	Z	$\tau = 1.99 \cdot 10^{-9}$ ($3.46 \cdot 10^{-9}$) $\beta = 2.03 \cdot 10^{-7}$ ($3.63 \cdot 10^{-7}$) corr. **36.3%** (36.3%)	$\tau = 4136$ ($1.37 \cdot 10^{21}$) $\beta = 3.91 \cdot 10^{-3}$ (7.2311) corr. **-3.3%** (-5.8%)	$\tau = 0.0348$ (0.0347) $\beta = 0.9487$ (0.9480) corr. **87.5%** (87.5%)	–
		$\tau = 1.35 \cdot 10^{-9}$ ($1.05 \cdot 10^{-8}$) $\beta = 0.3621$ (0.7042) corr. 29.0% (21.2%)	$\tau = 7573$ (7329) $\beta = 1.5837$ (1.5791) corr. -10.6% (-10.6%)	$\tau = 0.0348$ (0.0347) $\beta = 0.9487$ (0.9479) corr. 87.5% (87.5%)	–

Table 6.3: Optimal τ and β values for yeast chromosome 1 (230,208 bp = 1.9% of the genome). The maximised function was the log-transformed Pearson correlation coefficient between predicted average nucleosome occupancy and the *in vitro* data. The optimal correlation value is also shown.
The results for smoothed position-specific component ($\mathcal{N}$) are shown in brackets, the other ones are for not smoothed component. The results of Nelder and Mead optimisation algorithm are shown on white background, the results of a Newton-type algorithm are presented on gray background.

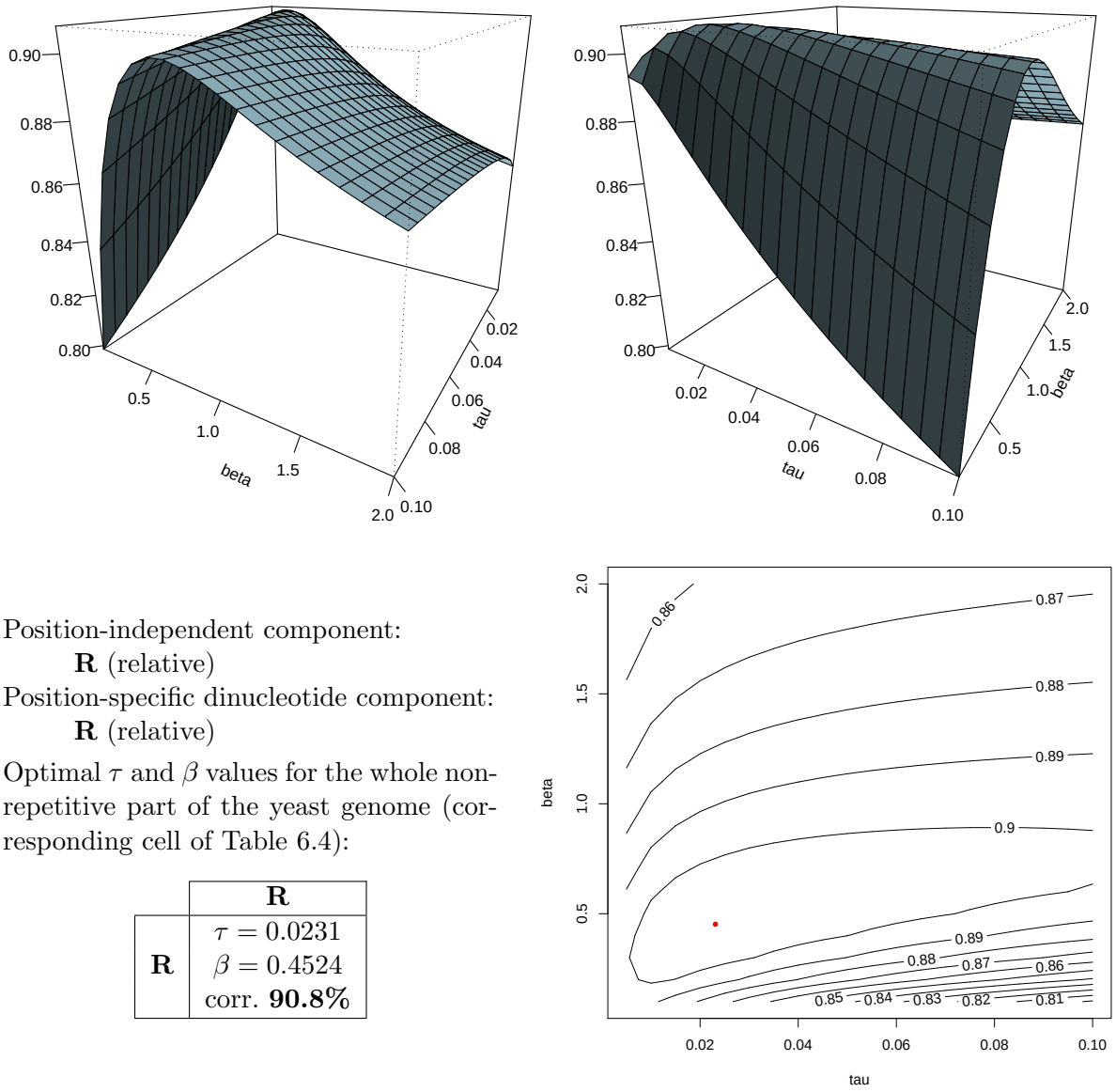
		Position-specific dinucleotide component ($\mathcal{N}$)		
		D	**R**	**Z**
Position-independent ($\mathcal{L}$)	**N**	$\tau = 0.0160$ $\beta = 0.3415$ corr. **59.9%**	$\tau = 0.0354$ $\beta = 0.2956$ corr. **71.7%**	$\tau = 0.0148$ $\beta = 0.3353$ corr. **59.2%**
	R	$\tau = 0.0105$ $\beta = 0.7918$ corr. **90.3%**	$\tau = 0.0231$ $\beta = 0.4524$ corr. **90.8%**	$\tau = 0.0100$ $\beta = 0.7593$ corr. **89.4%**
	Z	$\tau = 8.16 \cdot 10^{-13}$ $\beta = 3.50 \cdot 10^{-6}$ corr. **37.4%**	$\tau = 0.0306$ $\beta = 0.9160$ corr. **87.5%**	–

Table 6.4: Optimal τ and β values for the whole non-repetitive part of the yeast genome (10,774,972 bp = 89,3% of the genome).

compared with 87.6% taking only the position-independent component and 82.0% taking only the position-specific component. The model he refers to as a "full model" consists of position-independent component R and position-specific dinucleotide component D.

However, it is interesting to compare his good results gained for "P_N-only model" using only the position-specific dinucleotide component with mine. It seems that he might in fact be considering "full model" consisting of position-independent component R and position-specific dinucleotide component R.

The "relative" variants of the two components work well alone because they don't incorporate absolute probabilities of occuring a particular dinucleotide or 5-mer in nucleosome-bound sequences, but the relative chances of encountering it, taking into account the GC-content.



Position-independent component:

R (relative)

Position-specific dinucleotide component:

R (relative)

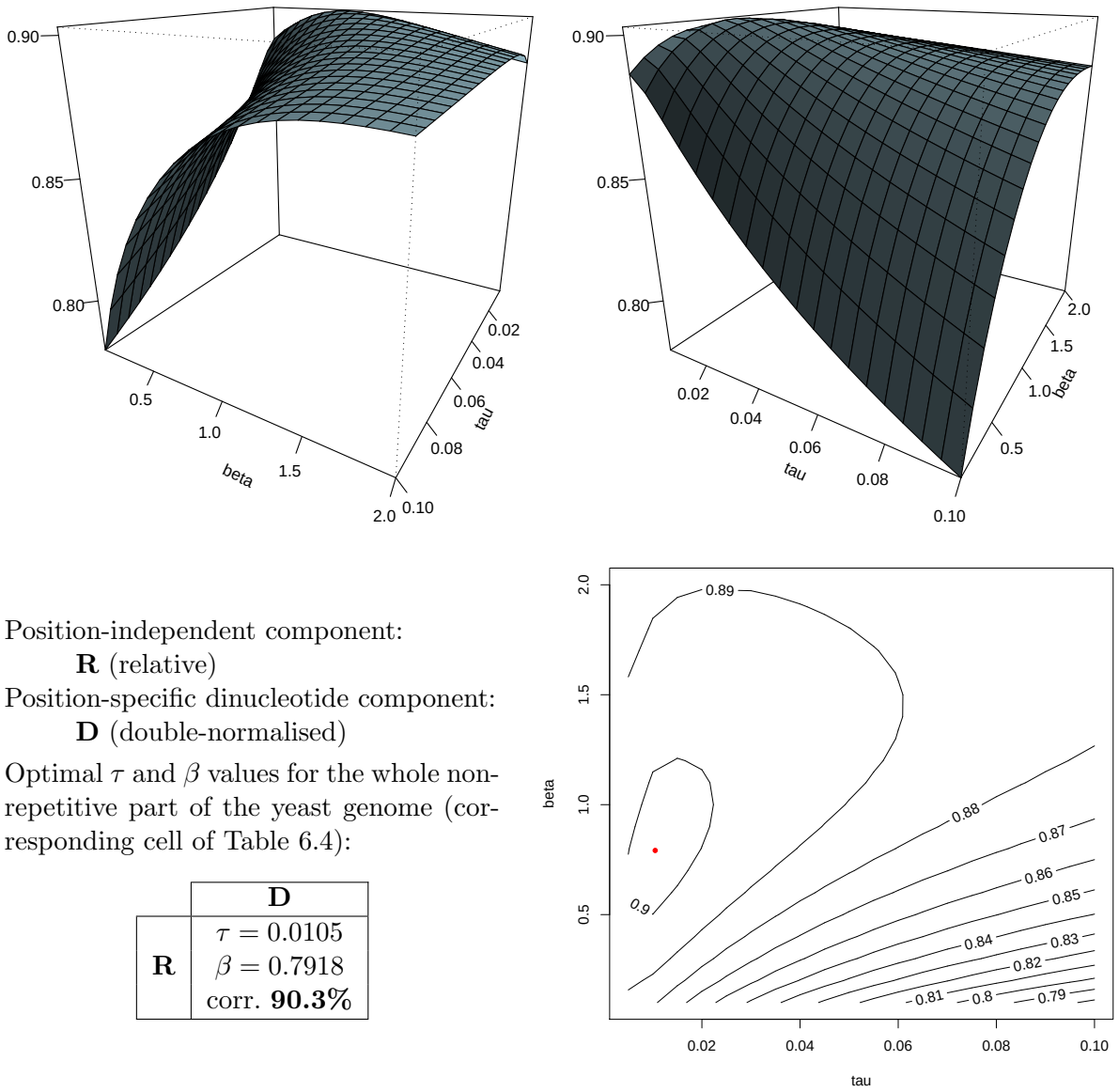
Optimal τ and β values for the whole non-repetitive part of the yeast genome (corresponding cell of Table 6.4):

	R
R	$\tau = 0.0231$ $\beta = 0.4524$ corr. **90.8%**

Figure 6.5: Nucleosome binding model consisting of position-independent component R and position-specific dinucleotide component R.

The above model has the best performance among all considered. In particular, the Pearson correlation between its prediction for optimal parameters and *in vitro* data is about four percent points better than 88.0%, which is the best result presented in [1].

For a grid of values of τ and β, the Pearson correlation coefficient between predicted average nucleosome occupancy and the *in vitro* data has been calculated. The correlation has been plotted on the two 3D plots and one contour plot above. Optimal values of τ and β, found using Nelder and Mead algorithm, has been marked by the red point on the contour plot.



Position-independent component:
R (relative)
Position-specific dinucleotide component:
D (double-normalised)

Optimal τ and β values for the whole non-repetitive part of the yeast genome (corresponding cell of Table 6.4):

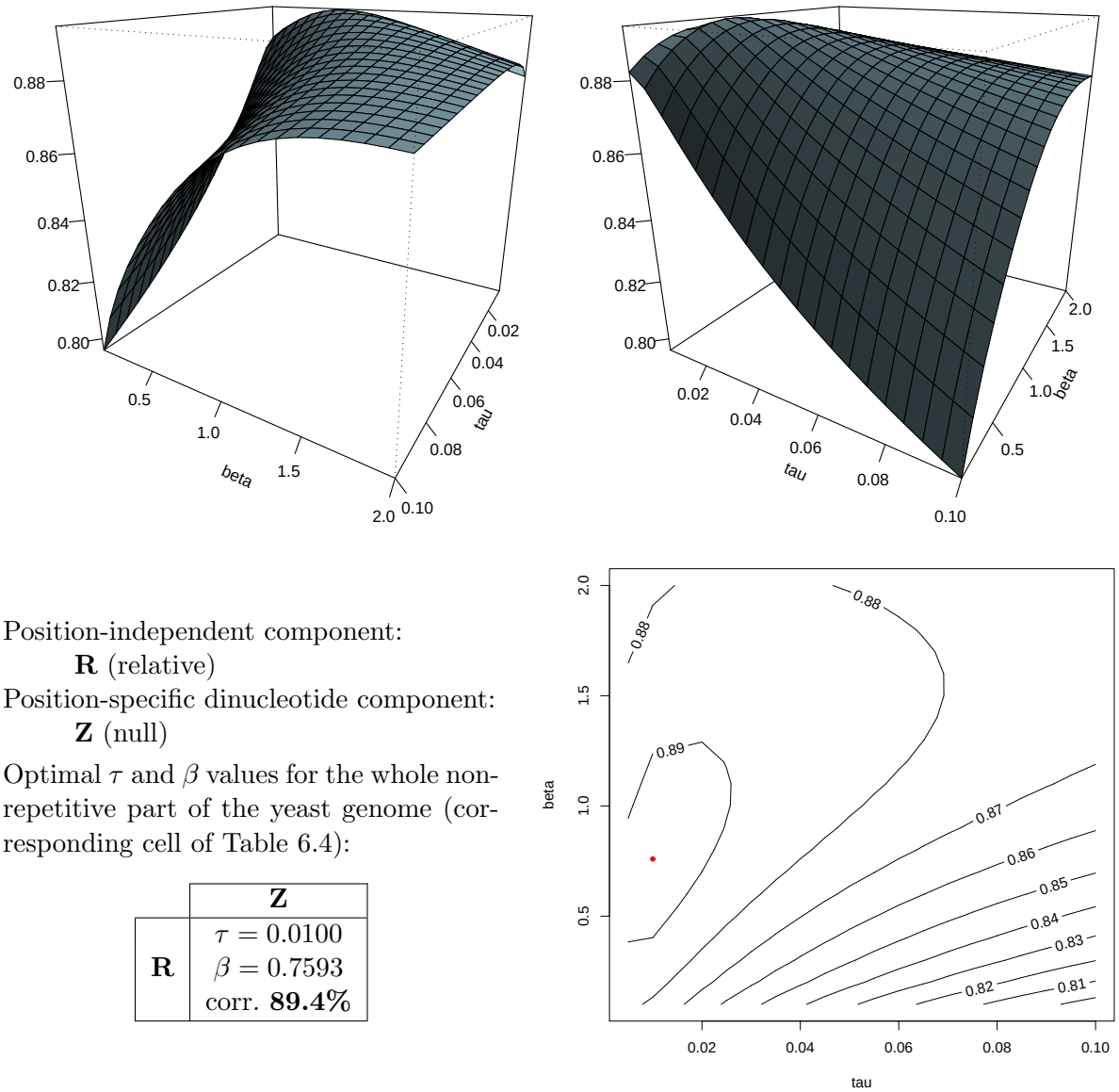
	D
R	$\tau = 0.0105$ $\beta = 0.7918$ corr. **90.3%**

Figure 6.6: Nucleosome binding model consisting of position-independent component R and position-specific dinucleotide component D.

The above model corresponds to the full model described in [1]. Kaplan *et al.* claims to have Pearson correlation between full model prediction and *in vitro* data of 88.0% for $\tau = 0.03$ and $\beta = 1$, which is about one and a half percent point worse than mine for the same parameter values.

However, I suggest that Kaplan *et al.* may in fact have considered "full model" consisting of position-independent component R and position-specific dinucleotide component R, like presented on Fig. 6.5.

The way of presenting the data is the same as on Fig. 6.5.



Position-independent component:

R (relative)

Position-specific dinucleotide component:

Z (null)

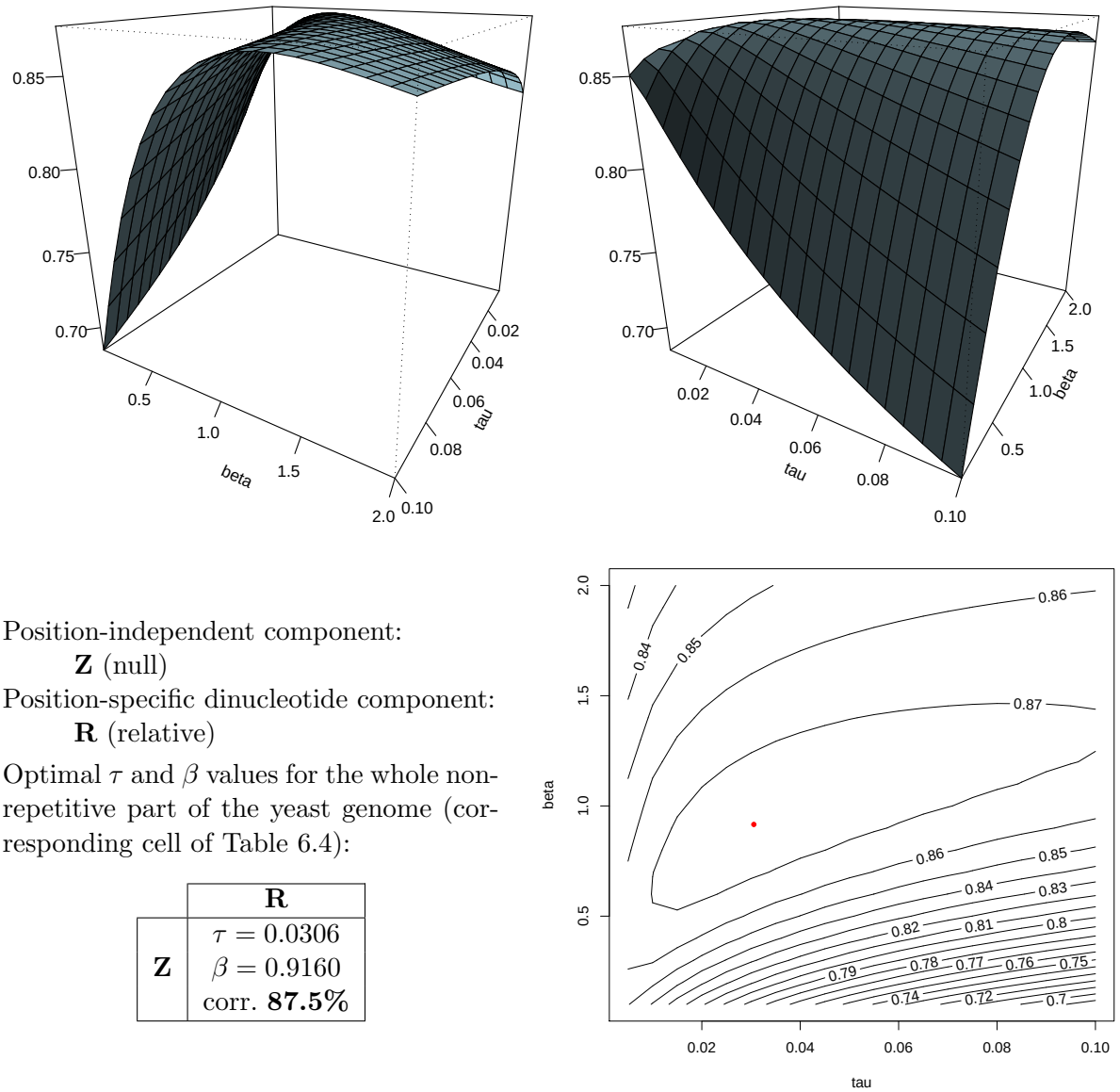
Optimal τ and β values for the whole non-repetitive part of the yeast genome (corresponding cell of Table 6.4):

	Z
R	$\tau = 0.0100$ $\beta = 0.7593$ corr. **89.4%**

Figure 6.7: Nucleosome binding model consisting of position-independent component R and position-specific dinucleotide component Z.

The above model corresponds to the P_L-only model in [1], and has the best performance among the models based solely on position-independent component. Kaplan *et al.* claims to have Pearson correlation between full model prediction and *in vitro* data of 87.6% for $\tau = 0.03$ and $\beta = 1$, which is about one percent point worse than mine for the same parameter values.

The way of presenting the data is the same as on Fig. 6.5.



Position-independent component:

Z (null)

Position-specific dinucleotide component:

R (relative)

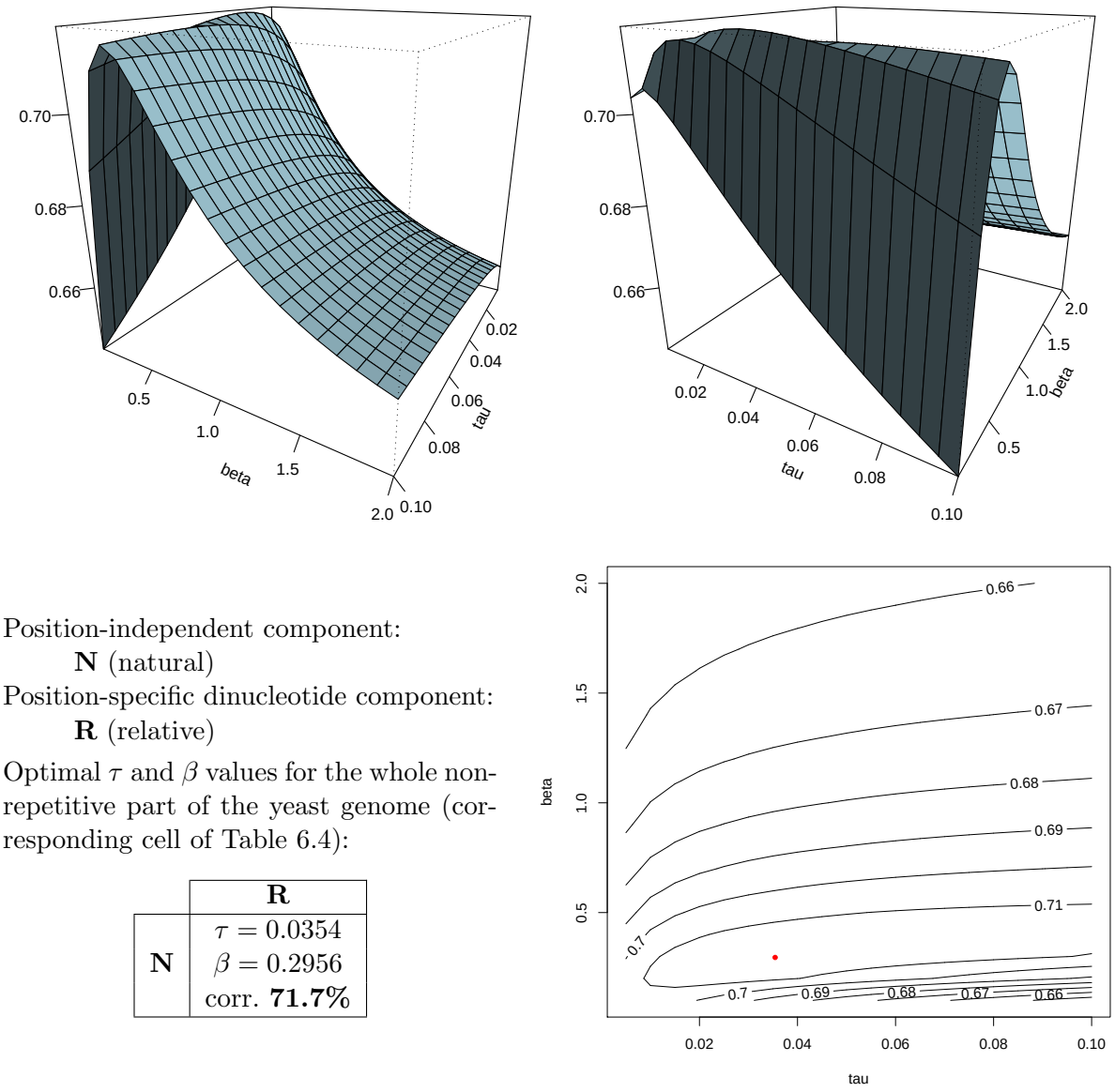
Optimal τ and β values for the whole non-repetitive part of the yeast genome (corresponding cell of Table 6.4):

	R
Z	$\tau = 0.0306$
	$\beta = 0.9160$
	corr. **87.5%**

Figure 6.8: Nucleosome binding model consisting of position-independent component Z and position-specific dinucleotide component R.

The above model has the best performance among the models based solely on position-specific dinucleotide component. In particular, the Pearson correlation between its prediction for optimal parameters and *in vitro* data is about four percent points better than 82.0%, which is the result for the "P_N-only" (based only on position-specific dinucleotide component) model presented in [1].

The way of presenting the data is the same as on Fig. 6.5.



Position-independent component:

N (natural)

Position-specific dinucleotide component:

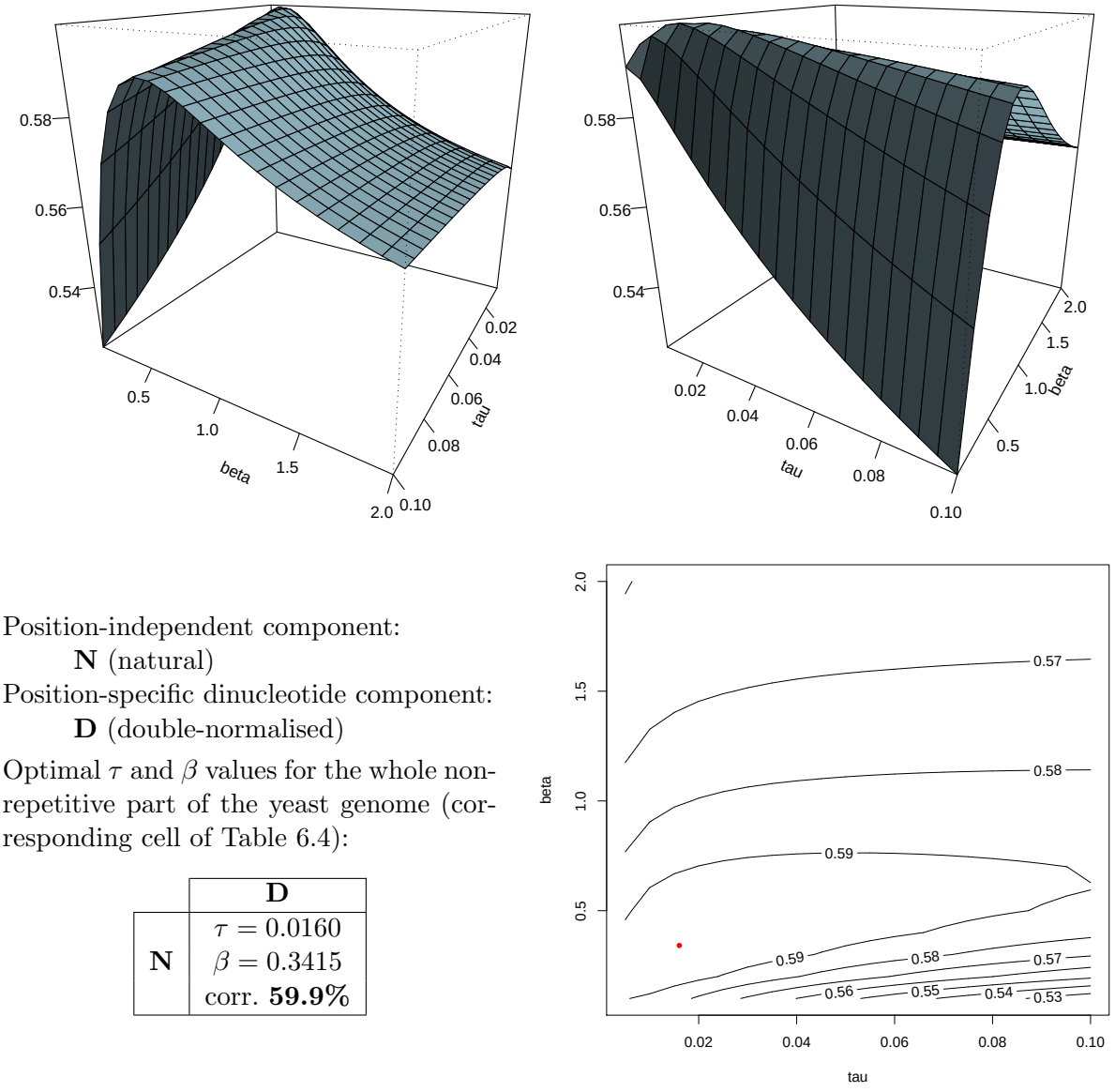
R (relative)

Optimal τ and β values for the whole non-repetitive part of the yeast genome (corresponding cell of Table 6.4):

	R
N	$\tau = 0.0354$ $\beta = 0.2956$ corr. **71.7%**

Figure 6.9: Nucleosome binding model consisting of position-independent component N and position-specific dinucleotide component R.

The way of presenting the data is the same as on Fig. 6.5.



Position-independent component:

N (natural)

Position-specific dinucleotide component:

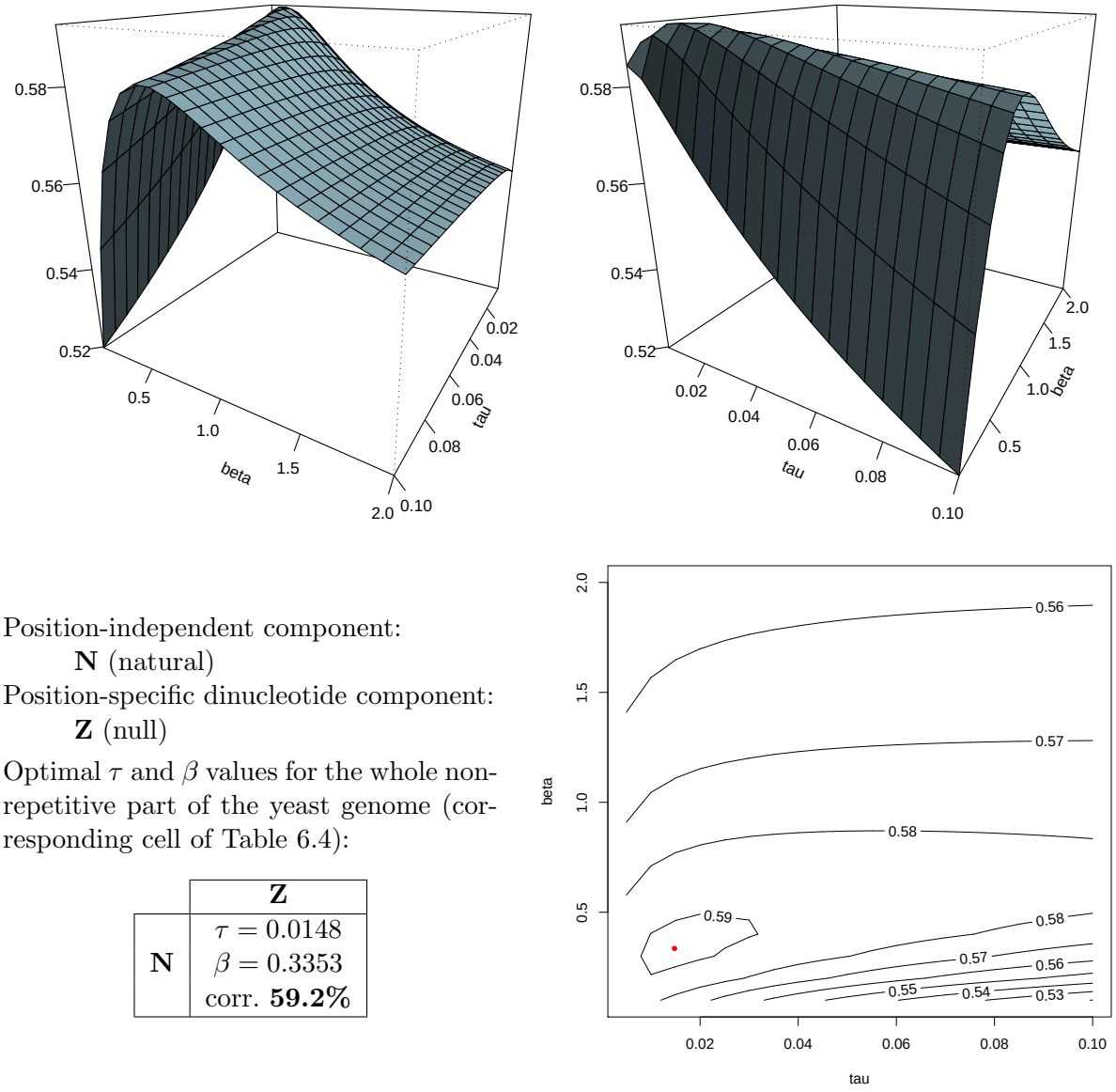
D (double-normalised)

Optimal τ and β values for the whole non-repetitive part of the yeast genome (corresponding cell of Table 6.4):

	D
N	$\tau = 0.0160$ $\beta = 0.3415$ corr. **59.9%**

Figure 6.10: Nucleosome binding model consisting of position-independent component N and position-specific dinucleotide component D.

The way of presenting the data is the same as on Fig. 6.5.



Position-independent component:

N (natural)

Position-specific dinucleotide component:

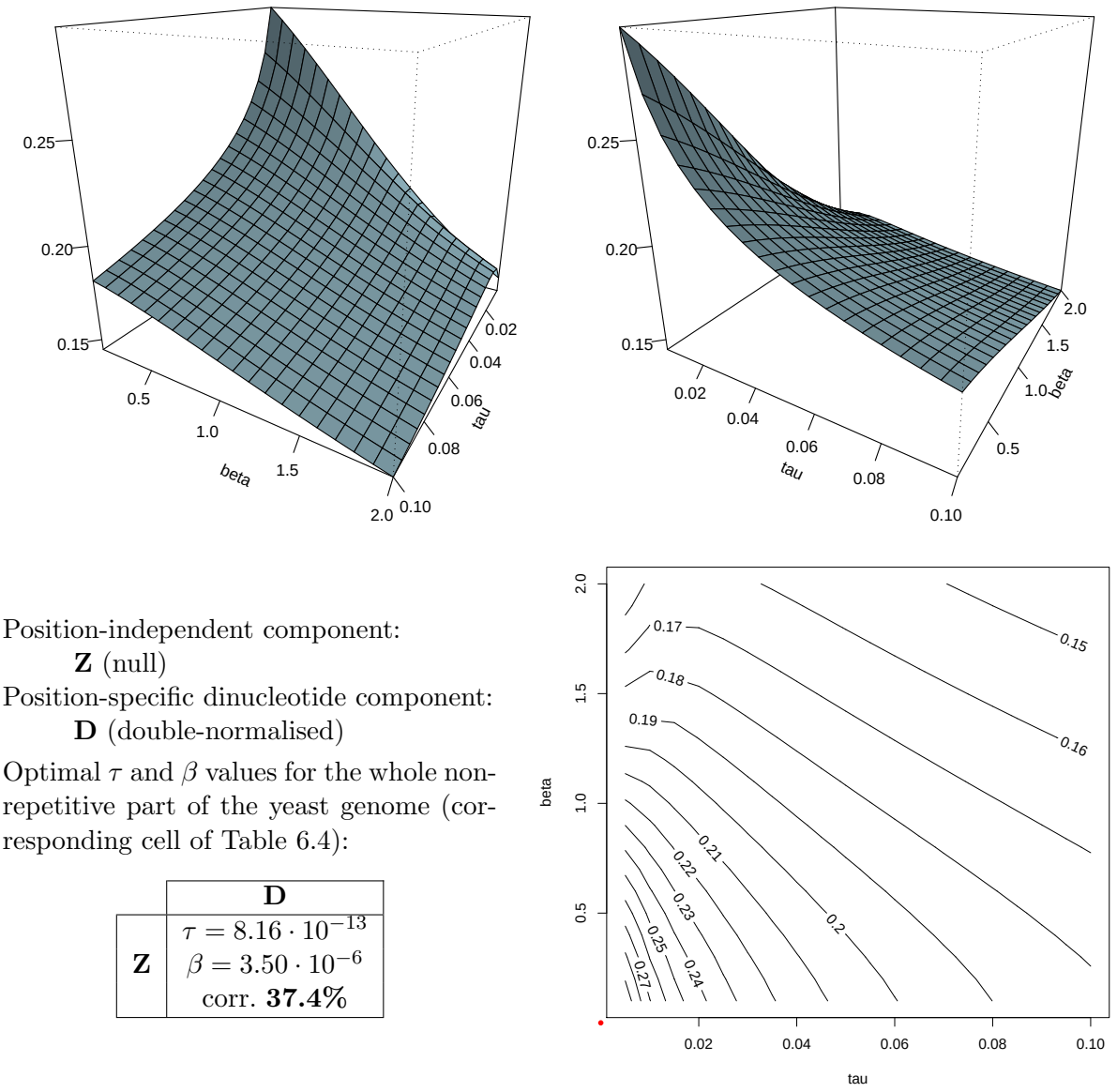
Z (null)

Optimal τ and β values for the whole non-repetitive part of the yeast genome (corresponding cell of Table 6.4):

	Z
N	$\tau = 0.0148$ $\beta = 0.3353$ corr. **59.2%**

Figure 6.11: Nucleosome binding model consisting of position-independent component N and position-specific dinucleotide component Z.

The way of presenting the data is the same as on Fig. 6.5.



Position-independent component:

Z (null)

Position-specific dinucleotide component:

D (double-normalised)

Optimal τ and β values for the whole non-repetitive part of the yeast genome (corresponding cell of Table 6.4):

	D
Z	$\tau = 8.16 \cdot 10^{-13}$ $\beta = 3.50 \cdot 10^{-6}$ corr. **37.4%**

Figure 6.12: Nucleosome binding model consisting of position-independent component Z and position-specific dinucleotide component D.

The above model corresponds to the P_N-only model in [1]. Kaplan *et al.* claims to have Pearson correlation between full model prediction and *in vitro* data of 82.0% for $\tau = 0.03$ and $\beta = 1$, which is about four times better than mine for the same parameter values.

However, I suggest that Kaplan *et al.* may in fact have considered "P_N-only model" consisting of position-independent component Z (null) and position-specific dinucleotide component R, like presented on Fig. 6.8.

The way of presenting the data is the same as on Fig. 6.5.

Chapter 7

Conclusion

The aim of my thesis was to repeat the experiments performed by Kaplan *et al.* [1], to investigate the performance of different model variants and to analyse the impact of the model parameters to overall performance of prediction.

The results of my experiments agree well with the ones obtained by Kaplan *et al.* [1]. Moreover, they are slightly but noticeably better, especially for the model based solely on position-specific dinucleotide component. My results confirm the Kaplan's *et al.* [1] observation that in practical applications, it is adequate to use only the position-independent component. Additionally, the smoothing of position-specific dinucleotide component seems to be unnecessary.

The issue not discussed by Kaplan *et al.* [1] was the choice of values for thermodynamical parameters τ and β; some values were assigned for them without justification. My study used two algorithms to estimate the optimal values of them. The impact of them has been also analysed; it is presented on Fig. 6.5-6.12. Fortunately, good results can be obtained for parameters from large intervals, i.e. the thermodynamical algorithm is to some extent not very sensitive in terms of its free parameters.

The nucleosome binding model presented in the thesis is obviously not perfect. The possible improvements, I am aware of, involve including in the model the physical interactions affecting nucleosome binding. There is a need to include the transcription factors, which bind directly to the DNA strand and thus are directly competing with nucleosomes.

Moreover, there are physical constrains on the mutual location of nucleosomes. For instance, there is a substantial periodic signal in the observed lengths of linkers between nucleosomes, due to the molecular interactions between them.

In other words, there is a need to further investigation and improvement of the model. It is also expected that the new experimental data that will eventually become available may give a new glance on the model accuracy.

Bibliography

[1] Noam Kaplan, Irene K. Moore, Yvonne Fondufe-Mittendorf, Andrea J. Gossett, Desiree Tillo, Yair Field, Emily M. LeProust, Timothy R. Hughes, Jason D. Lieb, Jonathan Widom and Eran Segal (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**, 362–366.

[2] Yair Field, Noam Kaplan, Yvonne Fondufe-Mittendorf, Irene K. Moore, Eilon Sharon, Yaniv Lubling, Jonathan Widom and Eran Segal (2008) Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Computational Biology*, **4**, e1000216.

[3] J. Michael Cherry, Caroline Adler, Catherine Ball, Stephen A. Chervitz, Selina S. Dwight, Erich T. Hester, Yankai Jia, Gail Juvik, TaiYun Roe, Mark Schroeder, Shuai Weng and David Botstein (1998) SGD: *Saccharomyces* Genome Database. *Nucleic Acids Research*, **26**, 73-79.

[4] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts and Peter Walter (2002) *Molecular Biology of the Cell*, 4th ed., Garland.

[5] Harvey Lodish, Arnold Berk, Paul Matsudaira, Chris A. Kaiser, Monty Krieger, Matthew P. Scott, Lawrence Zipursky and James Darnell (2003) *Molecular Cell Biology*, 5th ed., W. H. Freeman.

[6] David Kincaid and Ward Cheney (2009) *Numerical Analysis*, 3rd ed., American Mathematical Society.

[7] J. A. Nelder and R. Mead (1965) A simplex algorithm for function minimization. *Computer Journal* **7**, 308-313.

[8] Saša Singer and John Nelder (2009) Nelder-Mead algorithm. *Scholarpedia*, **4**, 2928.