BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

APPLICATIONS OF STATISTICAL MECHANICS TO BIOLOGICAL MACROMOLECULES

by

NIKOLAY V. DOKHOLYAN

B.S., Moscow Institute of Physics and Technology, 1992
M.S., Moscow Institute of Physics and Technology, 1994

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

1999

Approved by

First Reader ____

H. Eugene Stanley, Ph.D.
University Professor,
Professor of Physics

Second Reader ____

Rama Bansil, Ph.D.
Professor of Physics

Dedicated to my best friends:
my parents, my wife Rachel
and my sister Marina...

ACKNOWLEDGMENTS

I would like to acknowledge everybody who has helped and inspired me during my years in Boston. First, I express my sincere gratitude to my wife Rachel, my parents, and my sister Marina for their patience, support and encouragement. They helped me through the hardest times and taught me to be always optimistic. To them I dedicate my thesis.

I am grateful to my advisor, Prof. H. Eugene Stanley for making me part of his lively and fun group and for great support and encouragement during these years in Boston.

I thank all my collaborators for interesting and stimulating discussions, and without whom this work would not be possible: Prof. José S. Andrade Jr., Dr. Sergey Buldyrev, Prof. Shyamsunder Erramilli, Prof. Shlomo Havlin, Dr. Mi K. Hong, Prof. Peter R. King, Dr. Youngki Lee, Dr. Gerry Paul, Mr. Antonio Scala, Prof. Eugene I. Shakhnovich, Prof. H. Eugene Stanley, Ms. Rachel H. R. Stanley. Especially, I thank Dr. Sergey Buldyrev, Prof. Eugene I. Shakhnovich, and Prof. H. Eugene Stanley for their guidance and support in my research.

I must acknowledge all my friends who made my experience in Boston University fun and exciting. To mention a few, Luis Amaral, Palash Banerjee, Mark Barthélémy, Luis Cruz-Cruz, P. Gopikrishnan (Gopi), Ivo Grosse, Steve Harrington, Plamen Ivanov, Harald Kallabis, Emilia La Nave, Youngki Lee, Yanhui Liu, Hernán Makse, C.-K. Peng, Thadeu Penna, Puru Ray, Reza Sadr-Lahijany, Antonio Scala, Anna Skibinsky, Victor Spirin, Cristian Stagarescu, Francis Starr, Paul Trunfio, Anna Umansky, Brigita Urbanc, Gandhi Viswanathan, Masako Yamada, Stefano Zapperi, and many many others. I am also very thankful to my friends from outside Boston, from whom I have received enormous support: Yuri and Misha Dabaghian, Levon Esibov, Dmitry Sarkisov, Slava Stepanov, Armen and Sergey Stepanyants and Karen Tevosyan.

I would like to acknowledge my thesis committee, Prof. H. Eugene Stanley, Prof. Rama Bansil, Prof. Bernard Chasan, Prof. James Miller and Prof. Roscoe Giles for their patience and constant help. I also acknowledge NIH NRSA molecular biophysics predoctoral traineeship (1998-1999) for support and the Director of the molecular biophysics program, Prof. Kenneth Rothschild, for his interest in my research and many valuable discussions. I would like to specially thank Bob Tomposki and Jerry Morrow for their help and Chhaya Patel for her great artwork, which became part of my dissertation.

APPLICATIONS OF STATISTICAL MECHANICS TO BIOLOGICAL MACROMOLECULES

(Order No.)

NIKOLAY V. DOKHOLYAN

Boston University Graduate School of Arts and Sciences, 1999

Major Professor: H. Eugene Stanley, Professor of Physics

ABSTRACT

This thesis addresses questions concerning two types of biological macromolecules: DNA and proteins. In the case of DNA, the origin of the distribution of various repetitions of short fragments (oligomers) of DNA nucleotide sequences is addressed. The length distribution functions of dimeric tandem repeats are computed for the 16 possible distinct dimeric tandem repeats (corresponding to the combinations of 4 nucleotides A, C, G and T) in DNA sequences of various taxonomic partitions of the GenBank database. For those DNA sequences that code for proteins (*coding* DNA), it is found that all 16 distribution functions are exponential, which is an indication of strong evolutionary pressure against dimeric repeat expansion in active proteins. For *noncoding* DNA, the distribution functions for 12 dimers have long tails, which can be fit by power-law functions. The power-law distributions may be the result of different mutational mechanisms. A quantitative method is developed for analyzing repetitions of identical oligomers in DNA sequences. This method shows that some oligomers clump (group near each other) more than they would if randomly distributed, while others repel from one another. In the case of proteins, an algorithm for molecular dynamics simulations to study a protein-like chain is proposed. It is shown that this algorithm is an efficient way to study the thermodynamics and kinetics of protein folding. Study of the root mean square displacement of residues from their native state positions reveals the presence of the persistent part of the structure (core). Study of the kinetics of the folding transition reveals the nucleation scenario for protein folding.

Contents

List of Tables

List of Figures

List of Abbreviations

bp	base pair
DM	dystrophy myotonic
DNA	deoxyribonucleic acid
DRPLA	dentato-rubral pallidolusyian atrophy
DTR	dimeric tandem repeat
FRAXA, FRAXE	fragile X syndromes
HD	Huntington's disease
MC	Monte Carlo
MD	molecular dynamics
mRNA	messenger ribonucleic acid
NC	native contacts
NNC	non-native contacts

NS	native state
PFS	partially folded state
REM	random-energy model
RF	reading frame
rms	root mean square
SBMA	spinal and bulbar muscular atrophy
SCA1	spinocerebellar ataxia type 1
SOC	self organized criticality
SSM	slipped strand mispairing
SSR	simple sequence repeat
TS	transition state
tu	time units

Chapter 1

Introduction

Many questions remain open in the field of biology. Most of these questions are so complex that they can be accessed only through the cooperative effort of scientists from a variety of fields, such as chemistry, engineering, computer science, and physics. Physicists offer a set of models representing a simplified approach to complex problems. These models, despite their simplicity, serve as powerful tools for understanding the foundations of the phenomena under study. Among the vast variety of examples are the models of *(1)* helix – coil transition in polypeptide chains [1]; *(2)* electron transfer between biological molecules [2]; *(3)* neural networks [3]; *(4)* cell amoeboid motion [4]; *(5)* the random-energy model (REM) [5], which was designed to study disordered systems and was later applied to the problem of protein folding; *(6)* the viscous fingering of HCl through gastric mucin [6]; *(7)* the Gō model of protein folding [7, 8].

Many of these complex phenomena can be understood by focusing on the interplay of only a few of their most important processes, or *driving* forces. The remaining forces only slightly affect the system and may not be critical to understanding the phenomena. Physics models utilize this idea, by grasping only the driving forces underlying phenomena and, thus, eliminating "unimportant" degrees of freedom. The complex phenomenon becomes a simple robust model which is not affected by a small perturbation of the initial state of the system.

This thesis focuses on statistical mechanics models that have been developed to

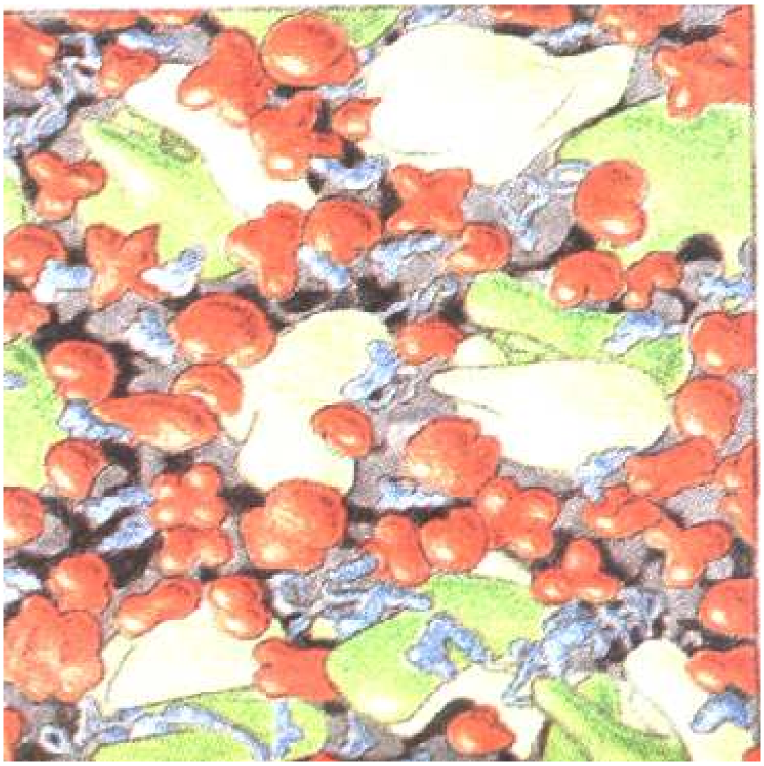

Figure 1.1: Macromolecules in cell cytoplasm.

A drawing of the 100 x 100 nm cross-section of the cell cytoplasm of *Escherichia coli* with only macromolecules shown. Reprinted from [9] with permission from Elsevier Science.

understand phenomena related to two macromolecules that serve as building blocks of life. These macromolecules are deoxyribonucleic acid (DNA) and proteins and both play a vital role in the cell [10, 11]. For instance, DNA stores genetic information in a form that is copied and transmitted from each cell to all of its progeny. Proteins carry out the transport and storage of cellular particles ranging in size from electrons to macromolecules and they carry out enzymatic activity in the cell, inhibiting or enhancing reactions. Proteins also control gene expression by switching genes on and off and they play a structural role as filamentous architecture (actin filaments and microtubules) within the cell. Both DNA and protein molecules carry specific information that is incorporated into their structure and can be understood by their interactions with other molecules (Fig. 1.1).

Due to the wide variety of functions of DNA and proteins, there are many complex phenomena that are not well understood. This thesis touches upon only a small fraction of them. The structure of the thesis is as follows: the second chapter of the thesis is dedicated to the statistical properties of DNA sequences and to the models developed to understand these properties. Protein folding studies are presented in the third chapter of the thesis. Thermodynamics and kinetics of the folding transition are described from the perspective of a simple model of inter-residue interactions.

Chapter 2

Patterns of nucleotides in DNA

2.1 Introduction

2.1.1 DNA: Structure and Function

The genetic information of organisms is stored in deoxyribonucleic acids (DNA) [10, 12]. DNA is composed of deoxyribonucleotides, which, in turn, consist of deoxyribose (a sugar), phosphate and a base. The deoxyribose and phosphate build the backbone of DNA molecules, while the bases are responsible for the "genetic alphabet" (Fig. 2.1). There are four different bases: adenine A and guanine G (purines), cytosin C and thymine T (pyrimidines). DNA is a double-helix molecule, in which two strands of deoxyribonucleotides are connected by pairing of the bases A-T and C-G (these pairs are called base pairs (bp)). The nucleotides are linked together by covalent phosphodiester bonds that join the $5'$ carbon of one deoxyribose group to the $3'$ carbon of the next. A-T are coupled by 2 hydrogen bonds ("weakly bonded pair"), while C-G are coupled by 3 hydrogen bonds ("strongly bonded pair"). Each DNA molecule is packaged in a chromosome, which varies in length from 10^5 bp (yeast) to 10^8 bp (human).

In the genomes of high eukaryotic organisms only a small portion of the total genome length is used for protein coding (as low as 3% in the human genome). The segments of the chromosomal DNA that are spliced out during the formation

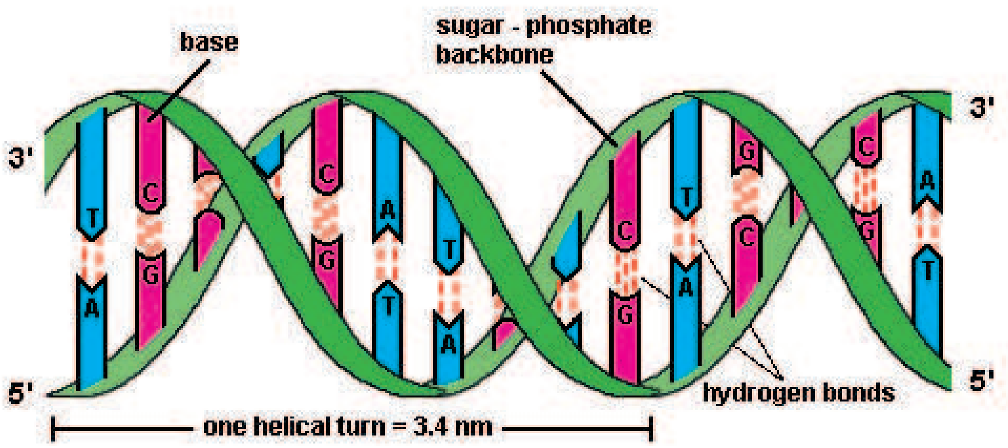

Figure 2.1: A schematic representation of DNA molecule.
DNA is a double-helix molecule, in which two strands of deoxyribonucleotides are connected by pairing of the bases A-T and C-G. Adapted from [10].

of a mature messenger ribonucleic acid (mRNA) are called *introns* (for intervening sequences). The coding sequences are called *exons* (for expressive sequences). Exons are translated into proteins by combinations of three nucleotides coding for amino acids. Even though there are $4^3 = 64$ possible combinations, only 20 amino acids, the building blocks of proteins, exist in nature. Combination of the base pairs A, C, G, and T provides the enormous informational content of DNA.

Definition 2.1 *mRNA is the molecule that specifies the amino acid sequence of a protein. mRNA is translated into protein in a process catalyzed by* ribosomes.

The role of introns and intergenic sequences that constitute large portions of the genome remains unknown. Furthermore, only a few quantitative methods are currently available for analyzing information which is possibly encrypted in the non-coding part of the genome. The statistical properties of nucleotides in *coding* DNA are different from those in *noncoding* DNA. This difference is especially pronounced for the simple sequence repeats (SSR).

2.1.2 Simple Sequence Repeats

The origin, evolution, and biological role of tandem repeats in DNA, also known as microsatellites, simple sequence repeats, or oligomers, are currently some of the most intriguing puzzles of molecular biology. The abundance of such SSR, also referred to as *SSR expansion*, is of great interest due to their role in genome organization and evolutionary processes [13–25]. It is known that SSR constitute a large fraction of noncoding DNA and are relatively rare in protein coding sequences.

Definition 2.2 *SSR can be represented as* $(X_1 X_2 \cdots X_N)_\ell$*, where* $X_{i=1,\ldots,N}$ *is one of* A*,* C*,* G *or* T*, and* ℓ *is the number of copies, which can, in some cases, be of the order of 100 for dimers* ($N = 2$)*.*

These observations lead to the concept that the bulk part of noncoding DNA is a "junkyard" of molecular evolution with the exception of a small fraction of functional sites [10]. However, if this is the case, there should be some physical and chemical restrictions on the size and structure of the noncoding DNA. Thus, an understanding of the statistical properties of SSR in noncoding DNA in comparison with coding DNA might shed light on the evolution of organisms, mutations and other hereditary phenomena.

SSR are of considerable practical and theoretical interest due to their high polymorphism [20]. The formation of a hairpin structure during replication [26, 27] is believed to be the cause of the $(CNG)_\ell$ repeat expansions ($N = \{A, C, G, T\}$), which are associated with a broad variety of genetic diseases. Among such diseases are fragile X syndrome (FRAXA and FRAXE) [28] (see Fig. 2.2), myotonic dystrophy (DM), Huntington's disease (HD) [17, 29], spinal and bulbar muscular atrophy (SBMA), spinocerebellar ataxia type I (SCA1), dentato-rubral pallidolusyian atrophy (DRPLA) and Machado-Joseph disease. For reviews on the role of trinucleotide repeats in neurological diseases, see Ref. [16, 30–33].

SSR of the type $(CA)_\ell$ are known to expand due to slippage in the replication process. These errors are usually eliminated by the mismatch-repair enzyme MSH2. However, a mutation in the MSH2 gene leads to an uncontrolled expansion

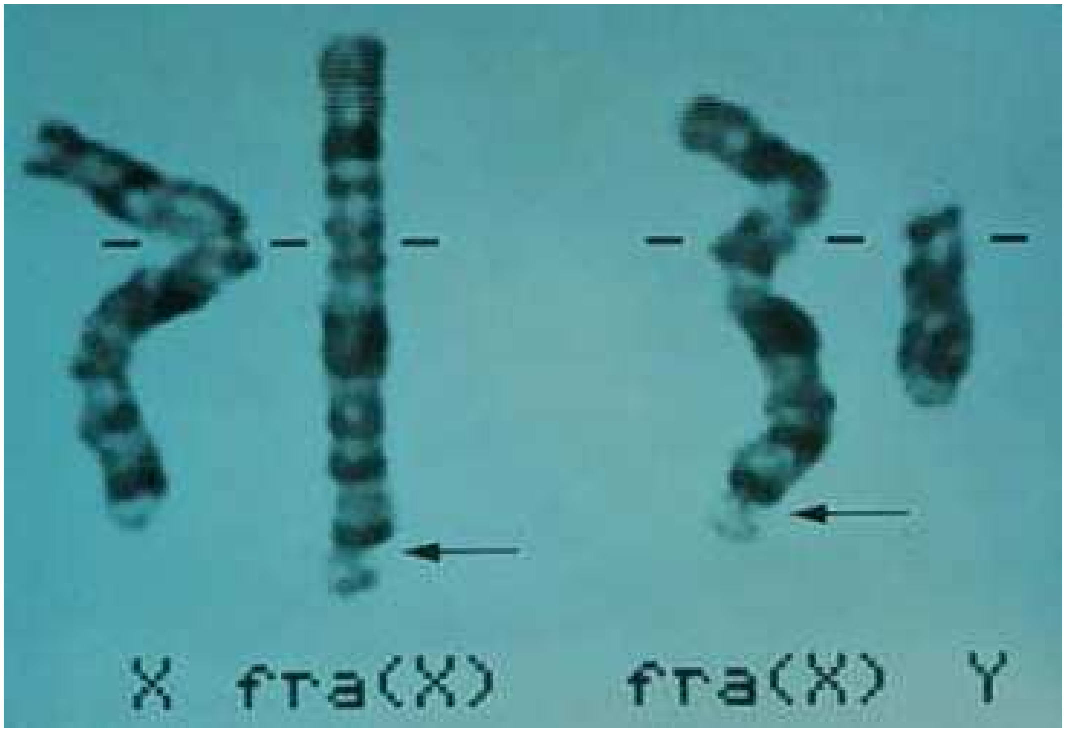

Figure 2.2: Fragile X syndrome.

A micrograph of two X chromosomes of a female (left) and an X and Y chromosome of a male (right). The two arrows point to the region known as a fragile site which was the original means of diagnosis for fragile X syndrome. Note that the indentation at the bottom of each looks as if it is ready to break. In this disease the triplet sequence *CGG* persistently repeats a few hundred to a few thousand times. The micrograph is a courtesy of the National Fragile X Foundation (http://www.nfxf.org).

of repeats—a common cause of ovarian cancers [34]. Similar mechanisms are attributable for other types of cancer [19, 35–41]. Telomeric SSR, which control DNA sequence size during replication, illustrate another crucial role of tandem repeats [10]. Other studies have noticed long tandem repeats of identical dinucleotides in noncoding regions of the genome [16, 42, 43].

Since microsatellites were first shown to aid genetic mapping [44], they have become primary genetic markers [45]. More recently, studies have used microsatellites to compare evolution among different species [46]. At some loci microsatellites are so polymorphic that they can be used for DNA fingerprinting [47].

Quantitative studies of microsatellites include analysis of runs of single nucleotides [48, 49]. Dinucleotides have been studied in terms of nearest neighbors [50] and relative frequencies [51]. Trinucleotides have been examined in terms of biased distributions [52]. A study of SSR from primates, emphasizing their abundance, length polymorphism, and overall tendency to expand in different sequence contexts, was reported recently by Jurka and Pethiyagoda [22]. The probability distribution functions for the length of special classes of repeats have been studied in recent publications [53–55]. Bell and Jurka [56] studied the length distributions of dimeric tandem repeats (DTR) of rodent and primate DNA. Other studies have been carried out on the periodicities of various SSR in introns [57, 58].

Systematic analysis of SSR distributions has yet to be done and is the goal of this Chapter of the thesis. Specifically, the distribution of the most simple case of SSR — repeats of identical dimers (DTR) — is considered (Section 2.2).

In addition, a quantitative method for studying the repetitions of oligomers in coding and noncoding DNA is developed (Section 2.7). The mean length (defined below) of repetitions of oligomers and the control, which is the expected length of repetitions if the oligomers were randomly distributed, are calculated. By taking the ratio between the actual value to the control value, it is possible to recognize whether oligomers "clump" (repeat more than they would if their order was randomly scrambled) or "repel" (repeat less than they would if their order was randomly scrambled).

Using this method, systematic comparison of clumping in coding and noncoding DNA is done for different organisms: primate, mammal, vertebrate, rodent, invertebrate and plant taxonomic partitions of GenBank. It is possible that any differences in the patterns of repetitions in coding and noncoding DNA can furnish ways to classify unknown sequences as coding or noncoding.

2.1.3 Dimeric tandem repeats

Definition 2.3 *DTR can be represented as* $(X_1X_2)_\ell$*, where* $X_{i=1,\ldots,N}$ *is one of* A*,* C*,* G *or* T*, and* ℓ *is the number of copies, e. g.* $ATATAT = (AT)_{\ell=3}$ *is DTR of length* $\ell = 3$*,* $GGGG \equiv (GG)_{\ell=2}$ *is DTR of length* $\ell = 2$*.*

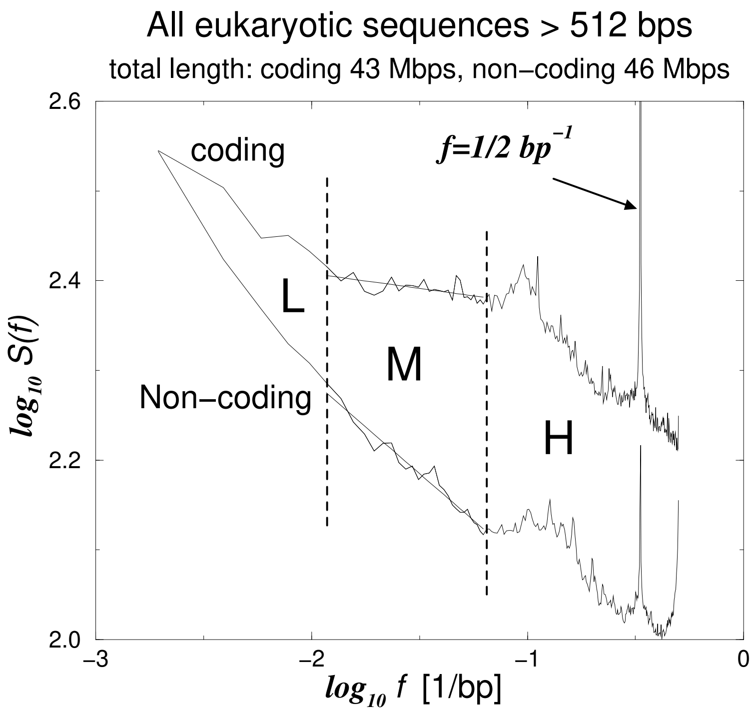

Figure 2.3: The average power spectrum of eukaryotic DNA sequences. Power spectra averaged over all eukaryotic sequences longer than 512 bp, obtained by FFT with window size 512. Upper curve is average over 29 453 coding sequences; lower curve is average over 33 301 noncoding sequences. Note, that there is a pronounced peak corresponding to the frequency $f = 1/2$ bp^{-1} for the noncoding eukaryotic DNA sequences. This peak is less pronounced for the coding eukaryotic DNA sequences. Courtesy of [59].

DTR are so abundant in noncoding DNA that their presence can even be observed by global statistical methods such as the power spectrum: e. g., the data of Buldyrev et al. [59] indicate the presence of a peak at $(1/2)$bp^{-1} in the power spectrum of *noncoding* DNA (corresponding to repetition of dimers) and significantly less pronounced peak at $(1/2)$bp^{-1} in *coding* DNA (here bp denotes base pair) (see Fig. 2.3). Also, it has long been known (e. g., see [60]) that the power spectrum has a peak at $(1/3)$bp^{-1} (corresponding to 3-tuples or trimers), which is more pronounced for coding sequences, and therefore suggests the presence of codon structure that can be used as a strong indicator of coding DNA. This difference in the abundance of DTR in coding DNA and noncoding DNA suggests that these repeats may play a special role in the organization and evolution of noncoding DNA. Section 2.2 is dedicated to the length distributions of DTR in coding and noncoding DNA. Models based on evolutionary processes that reproduce DTR length distributions are developed in Sections 2.3, 2.4, and 2.5.

2.2 Distribution of DTR in coding and noncoding DNA

2.2.1 Markovian DTR

Consider a random uncorrelated sequence which is a mixture of all possible dimers xy, each with a given probability p_{xy}, where x and y denote one of the four bp adenine (A), cytosine (C), guanine (G), and thymine (T). The probability that a given dimer belongs to a DTR $(xy)_\ell$ of length ℓ can be written as [43]

$$P_{xy}(\ell) = p_{xy}^{\ell} \cdot (1 - p_{xy})^2 \cdot \ell \, , \tag{2.1}$$

where $(1 - p_{xy})$ is the terminating factor responsible for not producing an additional unit xy at the beginning (or end) of the repeating sequences.

If $2\ell \ll 2L$, the total length of the DNA sequence, then the number of DTR $N(\ell)$ of ℓ dimers is

$$N(\ell) = p_{xy}^{\ell} \cdot (1 - p_{xy})^2 \cdot L \sim 10^{-\ell \cdot |\log_{10} p_{xy}|} \, , \tag{2.2}$$

which decreases exponentially with the length of the tandem repeat.

In order to compare the prediction of this simple model with real DNA data, p_{xy} is estimated for the real DNA as follows: *(i)* the DNA sequence is divided into L non-overlapping dimers, *(ii)* the total number of occurrences N_{xy} of dimer xy is counted in this sequence, and, then, the ratio

$$p_{xy} \equiv \frac{N_{xy}}{L} \tag{2.3}$$

is computed.

Since it is assumed that there are no correlations between dimers, the semilogarithmic plot of the number of occurrences of DTR of length ℓ is a straight line with slope

$$k_{\mathrm{unc}} = \log_{10}(p_{xy})\,. \tag{2.4}$$

In the extreme case of an uncorrelated sequence with equal probability for each bp, one must find $k_{\mathrm{unc}} = \log_{10}(1/4)^2 \approx -1.20$ for any of the 16 dimers.

It is found [42, 43] (see discussion below) that most 2-tuple tandem repeats in *coding* DNA produce linear semilogarithmic plots, but with slopes significantly different from those predicted by (2.4). Whether the deviations of the slope from prediction (2.4) can be explained by the short order Markov correlations [61] is tested below. A dimeric sequence generated by Markov process is described by 16×16 matrix Π whose coefficients $\Pi_{(xy)(zw)}$ are the conditional probabilities of finding a dimer zw after dimer xy. For an ergodic Markov process, the probability p_{xy} of finding a dimer xy is given by the final distribution which is an eigenvector of the matrix Π corresponding to the largest eigenvalue $\lambda = 1$. In this case, the number of DTR $(xy)_\ell$ of length ℓ in DNA of length $2L \gg 2\ell$ is given by an equation of the same form as (2.2):

$$N(\ell) = \Pi^{\ell-1}_{(xy)(xy)} \cdot \left(1 - \Pi_{(xy)(xy)}\right)^2 \cdot p_{xy} \cdot L \sim 10^{-|k_M| \cdot \ell}\,, \tag{2.5}$$

where $k_M \equiv \log_{10} \Pi_{(xy)(xy)}$. In cases when the semilogarithmic plots of $N(\ell)$ are straight lines, the slopes k from the actual plots are compared with k_M. The comparison of Eq. (2.5) with data on coding and noncoding DNA is presented in the Subsection 2.2.3.

2.2.2 Non-Markovian DTR

On the other hand, semilogarithmic plots for noncoding parts (Fig. 2.4) are usually not straight, but display negative slope with constantly decreasing absolute value, which indicates that their probability decays less rapidly than exponentially. This suggests that their distribution can be better described by a power law:

$$N(\ell) \sim \ell^{-\mu}\,. \tag{2.6}$$

In this case, these distributions should be straight lines on a double logarithmic plot. Therefore, for noncoding regions a double logarithmic plot of the distributions is shown in Fig. 2.5. In addition, the exponent μ is computed.

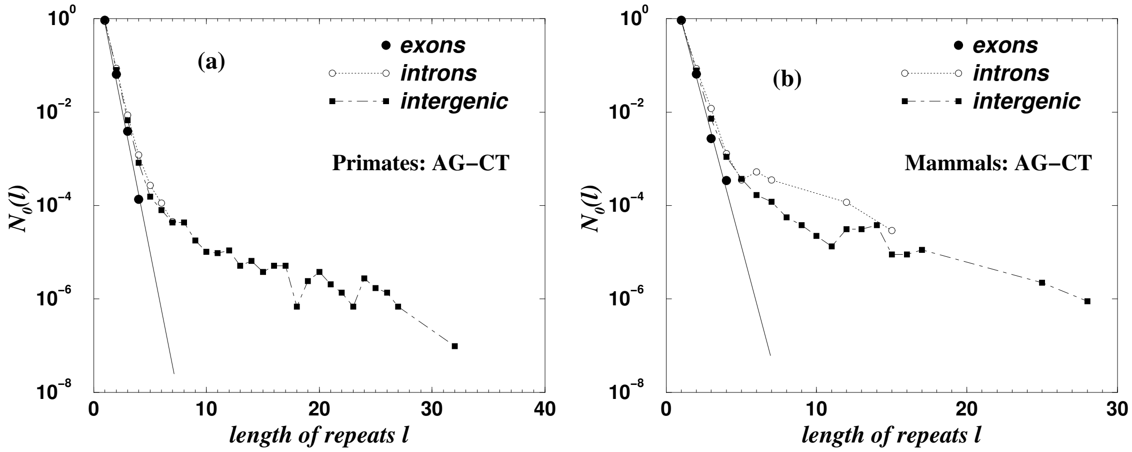

Figure 2.4: Semilogarithmic plot of $N_0(\ell)$ for $(AG)_\ell$ and $(CT)_\ell$ repeats in primates and mammals.
Semilogarithmic plot of average normalized number of repeats $(AG)_\ell$ and $(CT)_\ell$ of length ℓ versus ℓ for (a) primates and (b) mammals. Average plots for $(AG)_\ell$ and $(TC)_\ell$ are presented since they exhibit very similar behavior. Similar behavior is found for $(AA)_\ell$ and $(TT)_\ell$, and $(GT)_\ell$ and $(AC)_\ell$. The normalized number distribution for *coding* DNA is a straight line, while that for *noncoding* DNA significantly deviates from a straight line. For the special case of large $(dA) \cdot (dT)$ homopolymer tracts in *D. discoideum* Marx et al. [53] found quantitative differences between *coding* and *noncoding* DNA.

Figs. 2.4, 2.5, 2.6 and 2.7 are plotted as follows: *(i)* calculate the normalized number of repeats $N_0(\ell) = N(\ell)/N(1)$ of length ℓ , *(ii)* if there are no data points for one or more consecutive values of ℓ between points ℓ' and ℓ'', $N_0(\ell'')$ is substituted by $N_0(\ell'')/(\ell'' - \ell')$. To compare the goodness of fits for semilogarithmic and double logarithmic plots, their standard regression coefficients R are computed. The result of this analysis is presented in the Subsection 2.2.3.

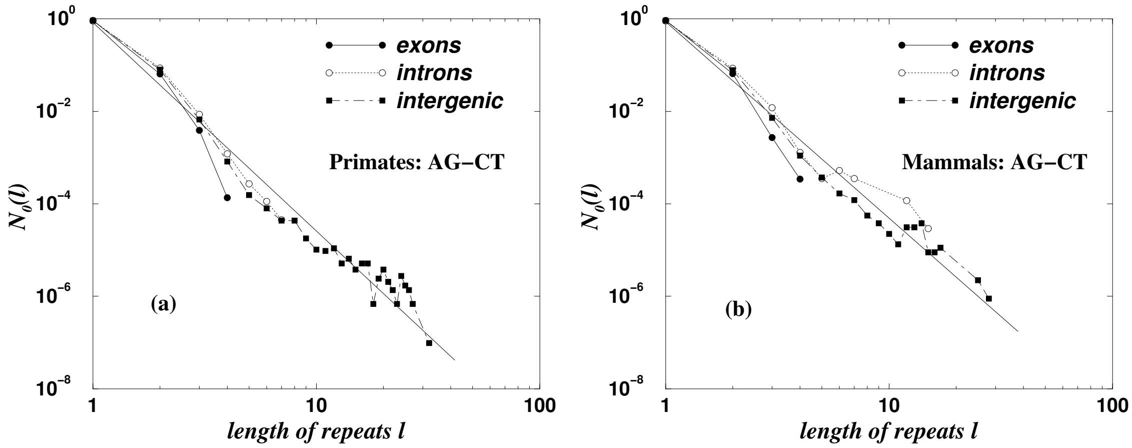

Figure 2.5: Double-logarithmic plot of $N_0(\ell)$ for $(AG)_\ell$ and $(CT)_\ell$ repeats in (a) primates and (b) mammals.
The same data as in Fig. 2.4 presented on a double logarithmic plot. Note that the normalized number distributions for *noncoding* DNA can be fit by a straight line. This means that these distributions can be approximated by a power-law (2.6).

Definition 2.4 *The linear correlation coefficient R for set of points (x_i, y_i), $i = 1, \ldots, N$, is defined as [62]*

$$R = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}}\,, \tag{2.7}$$

where $\bar{x} = N^{-1}\sum_{i=1}^{i=N} x_i$ and $\bar{y} = N^{-1}\sum_{i=1}^{i=N} y_i$. The R-value ranges between -1 and 1. When $R = 1$, x_i are completely positively correlated with y_i. When $R = -1$, x_i are completely negatively correlated with y_i. When $R = 0$, x_i are completely uncorrelated with y_i. If all the points (x_i, y_i) lie on the straight line then the absolute value of R is equal to 1 and the sign of R is $sign(R) = 1$ if the slope is positive and $sign(R) = -1$ if the slope is negative.

For semilogarithmic plots, the regression coefficients are computed by taking $x_i = \ell$ and $y_i = \ln(N_0(\ell))$ in Eq. (2.7). For double-logarithmic plots, the regression coefficients are computed by taking $x_i = \ln(\ell)$ and $y_i = \ln(N_0(\ell))$ in Eq. (2.7).

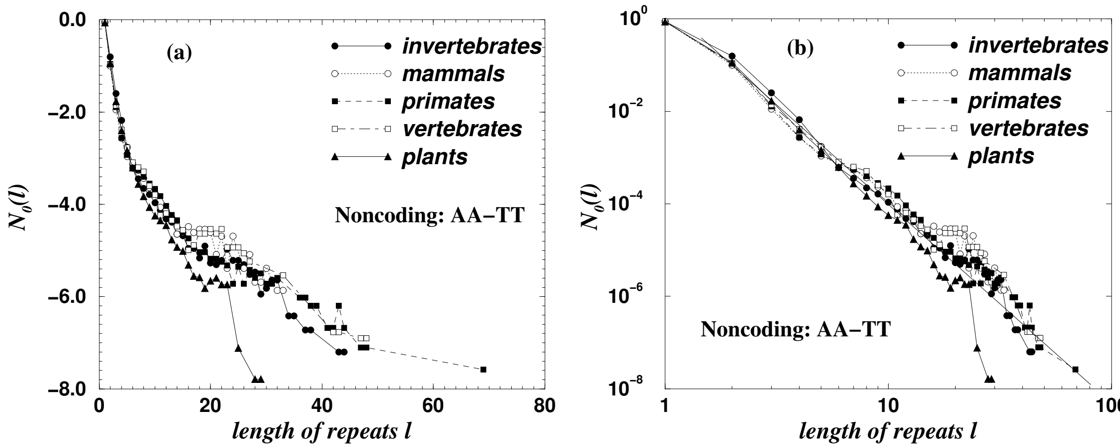

Figure 2.6: $N_0(\ell)$ for $(AA)_\ell$ and $(TT)_\ell$ repeats on (a) semilogarithmic and (b) double logarithmic scales.
The average normalized number distributions $N_0(\ell)$ of $(AA)_\ell$ and $(TT)_\ell$ dimers of length ℓ versus ℓ in (a) semilogarithmic and (b) double logarithmic scales for invertebrates, mammals, primates, vertebrates and plants. Note that all the plots collapse on a straight line on a double logarithmic plot.

2.2.3 DTR Distributions: Results and Analysis

Using data from GenBank at the NIH Internet site (*ncbi.nlm.nih.gov:/genbank*), $k_{\rm unc}$, k, k_M and μ are calculated for coding and noncoding regions of five types of organisms: primates, mammals (excluding primates and rodents), vertebrates (excluding mammals), invertebrates and plants. The results are shown in Tables 2.1, 2.2, 2.3, 2.4 and 2.5, and the goodness of fit is shown in Tables 2.6, 2.7 and 2.8.

The repeats xy and yx have almost identical distributions since repeat $(xy)_\ell$ must become $(yx)_{\ell\pm1}$ if one shifts the reading frame by one bp. For this reason, data for xy and yx are combined and presented on the same lines of Tables 2.1, 2.2, 2.3, 2.4 and 2.5.

One finds that semilogarithmic plots for coding regions are straight lines (see Fig. 2.4) and $k_{\rm unc}$ and the actual slope k for coding regions are close to each other; their relative difference lies in most cases between 5% to 18%. The $|r|$-values for the linear interpolation on semilogarithmic plots are close to 1. Thus, most dimers in

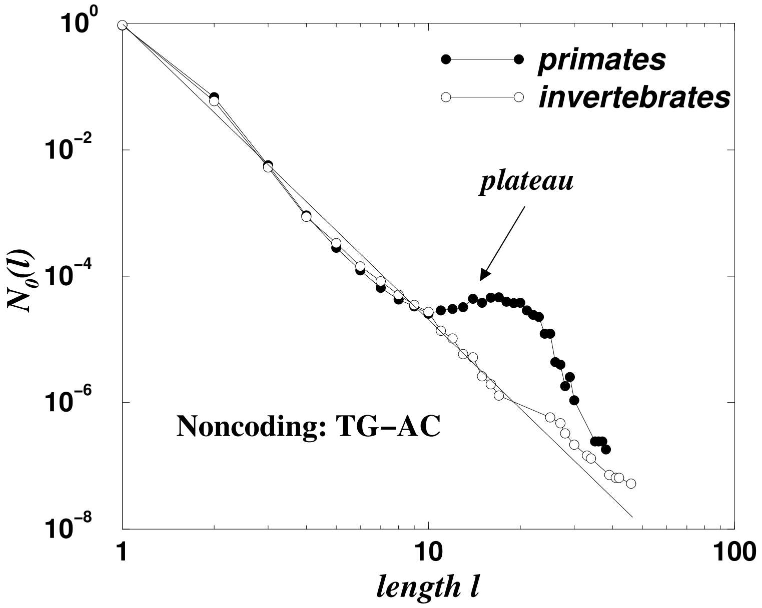

Figure 2.7: "Plateau" effect — $N_0(\ell)$ of $(AC)_\ell$ and $(TG)_\ell$ for primates and invertebrates.
To illustrate the plateau effect, a double logarithmic plot of the average normalized number distributions of $(AC)_\ell$ and $(TG)_\ell$ repeats is shown for primates and, in contrast, a plot for invertebrates to show the absence of plateau is provided.

exons are uncorrelated and, for this reason the μ values for the *coding* regions are not included.

The relative difference between k and k_{unc} is usually much greater for introns: it often exceeds a factor of 2. The difference is even greater for intergenic sequences where one finds long chains, up to 100 repetitions, of tandem repeats like "$TGTG$..." (in the intergenic sequences of vertebrates). In this case, values of k are still computed, although the regression coefficients of such fits are much smaller than for exons. It is notable that in most cases the $|r|$ values for the least square linear fits on semilogarithmic plots for the noncoding data are smaller than that on double logarithmic plots (see Tables 2.6, 2.7 and 2.8). The $|r|$ values for the least square linear fits on double logarithmic plots are close to 1, which indicates that power-law function is a better approximation for the *noncoding* data.

Note that if one assumes that a DNA sequence is uncorrelated or short-range correlated, then the estimated probability of a repeat $(TG)_{\ell=100}$ is $10^{k_{\mathrm{unc}}\cdot 100} \approx 10^{-118}$, which makes it highly improbable to find in any organism during the entire evolution! However, this argument fails when Eq. (2.6) is used. The probability of finding a hundred repetitions of TG is of the order of 10^{-8}, which is of the order of the inverse size of the vertebrate genome bank[1]. This is the reason why any model resulting in exponential length distributions is unable to explain long repeats (see critical review by Li [63]).

The difference between the statistical properties of DTR in coding and noncoding regions is demonstrated in typical examples in Figs. 2.4 and 2.5. One sees that the graphs for intergenic sequences and introns have long tails (see Fig. 2.6), while the graphs for exons have steep constant slopes; this difference between noncoding and coding regions is also apparent from Tables 2.1, 2.2, 2.3, 2.4 and 2.5 and Tables 2.6, 2.7 and 2.8.

The values of k_M indicate that the first order Markov process cannot *fully* explain the exponential behavior of the distribution functions for the coding regions. In some cases, (e. g. CA (AC) repeats for vertebrates, primates and mammals or TC (CT) for all organisms) k_M is very close to k.

In some cases (Fig. 2.7), the double logarithmic plots for noncoding regions have very long tails which are almost horizontal straight lines, and can not be approximated by a power-law function. Thus, for such sequences the values of μ are not estimated and the signs "plateau" are indicate on the corresponding lines of Tables 2.1, 2.2, 2.3, 2.4 and 2.5. These plateaus correspond to intergenic sequences of the dimers CA (AC), TG (GT) in vertebrates. This effect is not observed in invertebrates and plants. The presence of plateaus for both pairs TG (GT) and AC (AC) can be explained by the fact that sequence AC (CA) is complementary to TG (GT).

On the other hand, one can clearly see that all organisms lack CG (GC) repeats in their intergenic DNA. The exponent μ corresponding to CG (GC) ranges from -5.61 to -7.12 and k^{CG}_{unc} for introns is larger than k_{unc} for all other dimers (it is actually the largest among all 16 dimers). k^{CG} can, in some cases, be smaller than

[1]GenBank release 96.

k^{CG}_{unc} by as much as a factor of two.

It is hypothesized [43] that the existence of long TG (GT) repeats and a lack of CG (GC) repeats in vertebrates is possibly related to a process called *methylation* [10, 24]. During replication, the cytosine nucleotide C in the sequence CG can be covalently modified to *5-methylcytosine (5-methyl C)*. Later, the methylated C nucleotide undergoes accidental de-amination which cannot be repaired back to the previous state of C, but gets modified into a T nucleotide. Therefore, there is a strong tendency for methylated CG to mutate to TG. Three out of four CGs have been lost during evolution due to this mechanism [10]. The distribution of CG (GC) in DNA sequences is very inhomogeneous. In some regions there are CG (GC) rich "islands" of the length 1–2 kbp, in which the concentration of CGs (GCs) exceeds that of the entire genome by a factor of 10–20. The presence of CG-islands might be explained by the fact that the so-called *housekeeping genes*, encoding proteins essential for cell viability, remain unmethylated in the female germline. It is interesting to note that methylation does not occur in invertebrates and plants. Accordingly, plateaus of TG (GT) and AC (CA) dimers do not exist in these species (see Tables 2.1, 2.2, 2.3, 2.4 and 2.5).

Another mechanism possibly responsible for plateaus is a particular type of mutation called *slipped strand misspairing*, which is specifically strong for AC (CA) repeats in the case of MSH2 mutation [34, 64]. The presence of plateaus for the TG (GT) and AC (CA) repeat distributions is perhaps due to the cooperative effect of methylations and slipped strand misspairing.

Invertebrates											
sequence	exon			intron				intergenic			
	k_{unc}	k	k_M	k_{unc}	k	k_M	μ	k_{unc}	k	k_M	μ
AA	-1.13	-0.59	-0.91	-1.15	-0.14	-0.66	3.2	-0.90	-0.12	-0.72	4.0
CA (AC)	-1.26	-0.82	-1.26	-1.39	-0.36	-1.21	3.4	-1.24	-0.07	-1.18	3.2
CC	-1.29	-1.52	-1.50	-1.51	-0.75	-1.32	3.5	-1.45	-0.38	-1.34	3.1
GA (AG)	-1.26	-1.39	-1.16	-1.40	-0.46	-1.24	3.5	-1.24	-0.10	-1.16	3.5
GC (CG)	-1.32	-1.42	-1.41	-1.55	-1.35	-1.43	8.6	-1.46	-0.95	-1.40	6.3
GG	-1.25	-1.10	-1.68	-1.57	-0.58	-1.44	4.6	-1.45	-0.45	-1.34	3.8
TA (AT)	-1.27	-1.18	-1.20	-1.19	-0.13	-1.03	2.9	-1.10	-0.15	-1.06	3.0
TC (CT)	-1.31	-1.20	-1.23	-1.38	-0.30	-1.18	3.9	-1.24	-0.07	-1.16	3.3
TG (GT)	-1.27	-0.70	-1.26	-1.35	-0.59	-1.22	3.2	-1.24	-0.08	-1.17	3.8
TT	-1.25	-1.00	-1.09	-1.12	-0.27	-0.64	3.1	-0.90	-0.10	-0.72	4.0

Table 2.1: Invertebrates: k_{unc}, k, k_M and μ values.

The values of k_{unc}, k, k_M and μ for invertebrates for exons, introns and intergenic DNA. We study all loci larger than 100 bp. The size of DNA analyzed is 42,007,310 bp (total), 5,526,917 bp (introns), 10,895,865 bp (exons).

Mammals											
sequence	exon			intron				intergenic			
	k_{unc}	k	k_M	k_{unc}	k	k_M	μ	k_{unc}	k	k_M	μ
AA	-1.34	-1.10	-1.13	-1.31	-0.61	-0.92	4.4	-1.13	-0.12	-0.92	3.4
CA (AC)	-1.30	-1.20	-1.19	-1.37	-0.73	-1.13	4.3	-1.19	-0.06	-0.89	*plateau*
CC	-1.17	-1.27	-1.15	-1.33	-0.56	-0.96	4.4	-1.15	-0.36	-1.04	4.3
GA (AG)	-1.21	-1.28	-1.07	-1.29	-0.55	-1.00	4.2	-1.16	-0.14	-1.00	3.0
GC (CG)	-1.27	-1.33	-1.34	-1.56	-1.32	-1.48	5.5	-1.41	-0.90	-1.38	6.0
GG	-1.12	-1.22	-1.17	-1.29	-0.73	-0.94	4.1	-1.14	-0.47	-1.04	4.3
TA (AT)	-1.51	-1.27	-1.39	-1.40	-0.37	-1.13	4.4	-1.26	-0.17	-1.09	3.1
TC (CT)	-1.27	-1.15	-1.15	-1.32	-0.15	-0.98	3.0	-1.17	-0.18	-1.04	3.2
TG (GT)	-1.27	-1.21	-1.11	-1.29	-0.13	-1.04	2.6	-1.19	-0.08	-0.87	*plateau*
TT	-1.50	-1.12	-1.28	-1.29	-0.29	-0.83	3.5	-1.13	-0.28	-0.93	4.0

Table 2.2: Mammals: k_{unc}, k, k_M and μ values.

The values of k_{unc}, k, k_M and μ for mammals for exons, introns and intergenic DNA. The size of DNA analyzed is 2,458,786 bp (total), 484,569 bp (introns), 666,706 bp (exons).

Primates												
sequence	exon			intron				intergenic				
	k_{unc}	k	k_M	k_{unc}	k	k_M	μ	k_{unc}	k	k_M	μ
AA	-1.42	-1.12	-1.11	-1.35	-0.29	-0.85	3.4	-1.09	-0.06	-0.83	3.2
CA (AC)	-1.31	-1.25	-1.23	-1.36	-0.37	-1.14	3.1	-1.21	-0.10	-1.11	*plateau*
CC	-1.13	-1.16	-1.12	-1.28	-0.89	-1.06	6.9	-1.16	-0.52	-1.10	3.5
GA (AG)	-1.28	-1.27	-1.08	-1.30	-0.42	-1.03	4.6	-1.17	-0.18	-1.05	3.5
GC (CG)	-1.25	-0.87	-1.29	-1.55	-1.08	-1.53	6.8	-1.44	-0.56	-1.40	5.6
GG	-1.13	-1.18	-1.17	-1.26	-0.90	-1.04	6.7	-1.16	-0.35	-1.10	3.8
TA (AT)	-1.59	-1.59	-1.43	-1.43	-0.33	-1.15	4.2	-1.24	-0.15	-1.12	3.5
TC (CT)	-1.25	-1.19	-1.17	-1.30	-0.50	-1.00	4.6	-1.17	-0.14	-1.04	3.4
TG (GT)	-1.25	-1.06	-1.14	-1.30	-0.14	-1.07	3.1	-1.20	-0.12	-1.09	*plateau*
TT	-1.48	-1.42	-1.20	-1.30	-0.21	-0.80	3.6	-1.08	-0.17	-0.83	3.7

Table 2.3: Primates: k_{unc}, k, k_M and μ values.

The values of k_{unc}, k, k_M and μ for primates for exons, introns and intergenic DNA. The size of DNA analyzed is 21,285,464 bp (total), 3,574,597 bp (introns), 1,886,111 bp (exons).

Vertebrates												
sequence	exon			intron				intergenic				
	k_{unc}	k	k_M	k_{unc}	k	k_M	μ	k_{unc}	k	k_M	μ
AA	-1.26	-1.33	-1.12	-1.26	-0.60	-0.91	5.0	-1.08	-0.07	-0.89	3.0
CA (AC)	-1.24	-1.21	-1.23	-1.35	-0.37	-1.11	3.8	-1.19	-0.07	-1.01	*plateau*
CC	-1.23	-1.21	-1.17	-1.38	-0.32	-1.03	3.3	-1.22	-0.35	-1.02	4.0
GA (AG)	-1.23	-1.00	-1.04	-1.37	-1.21	-1.15	8.1	-1.20	-0.15	-0.89	3.2
GC (CG)	-1.30	-1.52	-1.42	-1.53	-1.37	-1.53	7.1	-1.37	-1.10	-1.31	7.1
GG	-1.19	-1.20	-1.24	-1.43	-0.46	-0.99	4.1	-1.21	-0.38	-1.01	3.6
TA (AT)	-1.40	-1.14	-1.31	-1.29	-0.53	-1.05	4.2	-1.21	-0.08	-1.09	2.7
TC (CT)	-1.25	-1.15	-1.18	-1.29	-0.19	-1.02	2.7	-1.20	-0.12	-1.08	3.3
TG (GT)	-1.22	-1.09	-1.15	-1.29	-0.77	-1.07	5.6	-1.18	-0.04	-1.02	*plateau*
TT	-1.35	-1.09	-1.24	-1.19	-0.36	-0.83	3.9	-1.07	-0.31	-0.85	4.3

Table 2.4: Vertebrates:k_{unc}, k, k_M and μ values.

The values of k_{unc}, k, k_M and μ for vertebrates (except *chloroplast* and *mitochondria*) for exons, introns and intergenic DNA. The size of DNA analyzed is 3,022,446 bp (total), 766,318 bp (introns), 859,929 bp (exons).

Plants											
sequence	exon			intron				intergenic			
	k_{unc}	k	k_M	k_{unc}	k	k_M	μ	k_{unc}	k	k_M	μ
AA	-0.99	-0.80	-0.95	-1.04	-0.28	-0.87	4.5	-0.99	-0.24	-0.84	4.6
CA (AC)	-1.19	-1.23	-1.30	-1.28	-0.30	-1.20	5.1	-1.25	-0.17	-1.25	4.7
CC	-1.33	-1.50	-1.40	-1.47	-0.33	-1.21	4.1	-1.38	-0.41	-1.34	5.3
GA (AG)	-1.16	-1.14	-1.15	-1.30	-0.38	-1.27	5.1	-1.24	-0.15	-1.19	4.8
GC (CG)	-1.38	-1.19	-1.48	-1.54	-0.87	-1.39	5.7	-1.42	-0.80	-1.36	7.1
GG	-1.30	-1.51	-1.55	-1.47	-0.62	-1.35	5.3	-1.37	-0.67	-1.31	6.1
TA (AT)	-1.17	-1.07	-1.19	-1.03	-0.15	-0.94	4.0	-1.09	-0.14	-1.00	4.2
TC (CT)	-1.23	-1.16	-1.22	-1.20	-0.23	-1.08	4.5	-1.22	-0.19	-1.16	4.7
TG (GT)	-1.22	-1.30	-1.34	-1.17	-0.16	-1.12	4.5	-1.23	-0.16	-1.23	4.7
TT	-1.08	-0.60	-1.07	-0.86	-0.23	-0.76	4.8	-0.96	-0.20	-0.82	4.7

Table 2.5: Plants: k_{unc}, k, k_M and μ values.

The values of k_{unc}, k, k_M and μ for plants for exons, introns and intergenic DNA. The size of DNA analyzed is 29,627,094 bp (total), 1,894,603 bp (introns), 17,397,439 bp (exons).

	Invertebrates					Mammals				
sequence	exon	intron		intergenic		exon	intron		intergenic	
	R_k	R_k	R_μ	R_k	R_μ	R_k	R_k	R_μ	R_k	R_μ
AA	-0.96	-0.75	-0.95	-0.90	-0.99	-1.00	-0.94	-0.91	-0.88	-0.99
CA (AC)	-1.00	-0.90	-0.99	-0.70	-0.95	-0.97	-0.92	-0.97	-0.74	-0.90
CC	-1.00	-0.91	-0.97	-0.93	-0.98	-1.00	-0.88	-0.96	-0.89	-0.98
GA (AG)	-1.00	-0.90	-0.98	-0.68	-0.95	-1.00	-0.95	-0.99	-0.81	-0.97
GC (CG)	-1.00	-1.00	-0.98	-0.99	-0.98	-0.98	-1.00	-0.99	-0.98	-0.98
GG	-1.00	-0.89	-0.98	-0.91	-0.99	-0.94	-0.92	-0.95	-0.90	-0.98
TA (AT)	-1.00	-0.77	-0.94	-0.88	-0.98	-1.00	-0.81	-0.96	-0.80	-0.97
TC (CT)	-0.99	-0.84	-0.97	-0.66	-0.94	-1.00	-0.67	-0.90	-0.79	-0.95
TG (GT)	-0.91	-0.92	-0.99	-0.64	-0.95	-1.00	-0.71	-0.91	-0.79	-0.90
TT	-1.00	-0.89	-0.98	-0.79	-0.98	-1.00	-0.84	-0.98	-0.93	-0.99

Table 2.6: Invertebrates and mammals: R_k and R_μ fit measures.
The r-values of the linear interpolation on semilogarithmic plot (R_k) and on double logarithmic plot (R_μ) for invertebrates and mammals for exons, introns and intergenic DNA. Note, that $|r_\mu|$-values for *noncoding* regions are closer to 1 than $|r_k|$, while we find the opposite for *coding* regions.

	Primates					Vertebrates				
AA	-0.97	-0.88	-0.98	-0.77	-0.98	-1.00	-0.94	-0.91	-0.88	-0.99
CA (AC)	-1.00	-0.81	-0.97	-0.83	-0.92	-0.91	-0.92	-0.97	-0.74	-0.93
CC	-1.00	-0.97	-0.98	-0.93	-0.99	-1.00	-0.88	-0.96	-0.89	-0.98
GA (AG)	-1.00	-0.88	-0.98	-0.83	-0.98	-1.00	-0.95	-0.99	-0.81	-0.97
GC (CG)	-0.99	-0.99	-0.99	-0.92	-0.99	-0.98	-1.00	-0.91	-0.98	-0.98
GG	-1.00	-0.97	-0.99	-0.85	-0.97	-0.94	-0.92	-0.95	-0.90	-0.98
TA (AT)	-1.00	-0.80	-0.99	-0.83	-0.98	-0.99	-0.81	-0.96	-0.80	-0.97
TC (CT)	-1.00	-0.93	-0.99	-0.80	-0.97	-1.00	-0.67	-0.90	-0.79	-0.95
TG (GT)	-0.95	-0.79	-0.95	-0.80	-0.92	-1.00	-0.71	-0.91	-0.79	-0.89
TT	-0.99	-0.84	-0.98	-0.88	-0.98	-0.99	-0.84	-0.98	-0.93	-0.99

Table 2.7: Primates and vertebrates: R_k and R_μ fit measures.

The r-values of the linear interpolation on semilogarithmic plot (R_k) and on double logarithmic plot (R_μ) for primates and vertebrates for exons, introns and intergenic DNA. Note, that $|r_\mu|$-values for *noncoding* regions are closer to 1 than $|r_k|$, while we find the opposite for *coding* regions.

	Plants				
sequence	exon	intron		intergenic	
	R_k	R_k	R_μ	R_k	R_μ
AA	-0.99	-0.91	-0.99	-0.94	-0.99
CA (AC)	-1.00	-0.88	-0.99	-0.86	-0.99
CC	-1.00	-0.88	-0.98	-0.90	-0.99
GA (AG)	-1.00	-0.87	-0.98	-0.86	-0.99
GC (CG)	-1.00	-0.97	-1.00	-0.97	-0.99
GG	-1.00	-0.91	-0.98	-0.94	-0.99
TA (AT)	-1.00	-0.87	-0.99	-0.91	-0.99
TC (CT)	-1.00	-0.89	-0.99	-0.86	-0.99
TG (GT)	-1.00	-0.81	-0.99	-0.82	-0.97
TT	-0.95	-0.92	-0.99	-0.92	-0.99

Table 2.8: Plants: R_k and R_μ fit measures.

The r-values of the linear interpolation on semilogarithmic plot (R_k) and on double logarithmic plot (R_μ) for plants for exons, introns and intergenic DNA. Note, that $|r_\mu|$-values for *noncoding* regions are closer to 1 than $|r_k|$, while we find the opposite for *coding* regions.

2.3 Models of the DTR distributions in coding and noncoding DNA

Several questions arise from the above findings: *(i)* Why are the distributions of most tandem repeats in coding regions exponential, while in noncoding regions they usually have long non-exponential tails? *(ii)* How is this phenomenon related to the different long-range correlation properties of coding and noncoding DNA? *(iii)* What is the relation of the distribution of DTR in coding DNA to the amino acid composition of proteins?

2.3.1 Relation to long-range correlations

In recent years long-range power-law correlations have been discovered in a wide variety of systems. Such long-range power-law correlations are a physical fact that in turn gives rise to the increasingly appreciated "fractal geometry of nature" [65–78]. Recognizing the ubiquity of long-range power-law correlations can aid efforts to understand nature, since as soon as one finds power-law correlations they can be quantified with a critical exponent. Quantification of this kind of scaling behavior for apparently unrelated systems allows for the recognition of similarities between different systems, leading to underlying unifications that might otherwise have gone unnoticed.

Traditionally, investigators in many fields characterize processes by assuming that correlations decay exponentially. However, there is one major exception: at the critical point, the exponential decay turns into a power law decay [79]. Many systems drive themselves spontaneously toward critical points [66, 80].

Recent studies [59, 81, 82] support earlier work [83–85] that long-range correlations in nucleotide composition are stronger in noncoding regions than in coding ones. It is plausible that long-range correlations are related to the specific features of repetitive elements [86], although the exact correspondence remains unclear.

From the theory of Lévy distributions [87, 88], it follows that simple tandem repeats randomly distributed along the sequence can produce long-range power-law correlations if, and only if, $\mu < 3$. However, in almost all real DNA sequences $\mu > 3$, which means that simple tandem repeats alone *cannot* explain long-range correlations. On the other hand, simple tandem repeats may be the primary source of the difference in long-range correlation properties of coding and noncoding sequences at relatively short length scales of $\ell \approx 100$ bp [86].

2.3.2 SOC models

Power-law long-range correlations are known to be characteristic of systems evolving near critical point. In recent years, many models have been proposed that have a tendency to self organization towards their critical points (SOC) without any tuning of external parameters [89]. These models give rise to scaling and produce sudden avalanche-like bursts of activity distributed according to a power law. Some SOC models are one dimensional systems and have been applied to biological systems [90–92]. Such models are of great interest and they might be relevant to the study of the power-law behavior of DTR distributions.

2.3.3 Relation to trimers

The absence of lengthy DTR in coding DNA is related to the 3 bp length of a codon. Since 3 is not a multiple of 2, the repetition in amino acid sequence does not correspond to the repetition of dimers. Only very specific pairs of amino acids, such as $HisThr$ ($CACACA$), $ArgGlu$ ($AGAGAG$), $ArgAla$ ($CGCGCG$), correspond to DTR. In this case, DTR of amino acids correspond to those of DNA. Even DTR, such as $(AA)_\ell$ and $(TT)_\ell$, are much more abundant in noncoding DNA. This observation probably indicates that there is strong evolutionary pressure against long stretches of amino acid repeats in proteins. It was proposed [93] that in order to fold into a native state, proteins should have statistical properties close to those of a random sequence. The above findings are in complete agreement with the absence of long amino acid repeats in globular proteins.

2.3.4 Can patchiness explain non-exponential length distributions of DTR?

It is known that DNA has a very complex mosaic structure with patches of different concentrations of various nucleotides [94], dinucleotides, etc. Can this heterogeneity account for the non-exponential distribution of DTR? It is assumed that DNA sequences consist of relatively long patches (each longer than the largest observed tandem repeat) of various dinucleotide contents which are characterized by different probabilities p_{xy} of finding dimer xy. The resulting distribution in the limit of infinitely long DNA will be

$$N(\ell) = \int_0^1 (1-p_{xy})^2 \cdot p_{xy}^{\ell} \cdot f(p_{xy}) dp_{xy} \,, \tag{2.8}$$

where $f(p_{xy})$ is the probability of finding a "patch" with dimer concentration p_{xy}.

In order to have power-law behavior of the distribution function

$$N(\ell) \sim \ell^{-\mu} \,, \tag{2.9}$$

one must assume that the density of patches of concentration p_{xy} behaves as a power-law when p_{xy} is close to 1,

$$f(p_{xy}) \sim (1-p_{xy})^{\mu-3} \,. \tag{2.10}$$

The integral (2.8), with the function $f(p_{xy})$ satisfying Eq. (2.10), is a Beta-function, which asymptotically behaves as $\ell^{-\mu}$ (e. g., [95]). The existence of the solution for (2.8) means that it is possible to construct a power-law function of the distribution of DTR out of the uncorrelated patches with exponentially distributed repeats. Since for real DNA data, $f(p_{xy})$ cannot have singularities at $p_{xy} = 1$, the exponent μ cannot be less than 3. This is consistent with our observations.

Indeed, analysis of real DNA shows that a significant fraction of windows of length 100 bp have an extremely high concentration of all asymmetric dimers, with the exception of CG (GC). It seems that for the asymmetric dimers (except CG and (GC)), $f(p_{xy})$ approaches a constant value when $p_{xy} \to 1$ (see Fig. 2.8). The absence of windows with a high concentration of strongly bonded repeats CC, GG,

CG (GC) is in agreement with the rapid decay of their length distributions. However, the absence of abnormally high concentrations of AA and TT repeats excludes the possibility that the power-law distributions of AA and TT in noncoding DNA are related to the AA and TT rich patches.

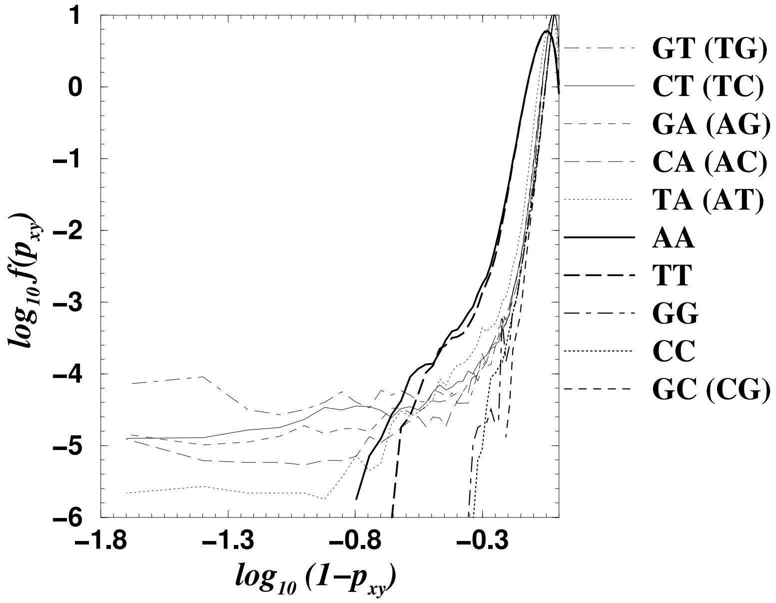

Figure 2.8: $f(p_{xy})$

A double logarithmic plot of probability density $f(p_{xy})$ of the windows with dimeric concentration p_{xy} for invertebrates (for different dimers xy) plotted versus $1 - p_{xy}$. The size of windows is 100 bp. One can see the difference between distributions $f(p_{xy})$ of asymmetric dimers AC, TG, TC, AT, AG (light lines) which appears to approach constant values when $p_{xy} \to 1$. There do not exist windows with extremely high concentrations of the symmetric dimers AA, TT, GG, and CC (heavy lines).

The fact that the function $f(p_{xy})$ approaches a constant value when $p_{xy} \to 1$ predicts $\mu = 3$ for asymmetric repeats. The measured values of μ for these repeats ranges between 3 and 4. The analysis of primate DNA, however, shows that long DTR $(xy)_\ell$ do not occur due to the abnormally high concentrations of the corresponding dimers xy in their neighborhood (see Table 2.9). Specifically, the concentration of xy dimers in the window of 100 bp adjacent to long repeats ($\ell \geq 10$) are calculated and it is found that this concentration never exceeds 0.21, and suggests that power-law expansions of dimers are not simply due to the presence of patches of abnormally high concentration. However, the average concentration in the adjacent windows is larger than that of a random sequence ($1/16 = 0.0625$) (see Table 2.9), a fact that might be explained by specific mutation mechanisms. According to Bell [96], the probability of repeat expansion is larger for longer repeats. However, the probability that point mutations occur is also proportional to the length of the sequence. Point mutations break the sequence, and prevent it from the unlimited expansions. Thus, smaller fragments can surround large DTR, increasing the concentration of corresponding dimers in the windows adjacent to the large DTR.

dimer	total number	downstream			upstream		
		window	average	ratio	window	average	ratio
AA	1161	0.12	0.08	1.60	0.06	0.08	0.80
TA	102	0.12	0.05	2.52	0.12	0.05	2.65
GA	78	0.13	0.06	2.14	0.12	0.06	2.02
CA	259	0.10	0.07	1.35	0.11	0.07	1.45
AT	102	0.12	0.06	2.03	0.13	0.06	2.13
TT	1253	0.06	0.08	0.75	0.13	0.08	1.60
GT	302	0.09	0.05	1.75	0.08	0.05	1.62
CT	97	0.16	0.07	2.23	0.18	0.07	2.45
AG	77	0.13	0.07	1.83	0.13	0.07	1.71
TG	304	0.11	0.08	1.47	0.10	0.08	1.34
GG	4	0.06	0.07	0.85	0.11	0.07	1.56
CG	1	0.04	0.02	2.06	0.02	0.02	1.03
AC	261	0.08	0.05	1.53	0.09	0.05	1.74
TC	103	0.16	0.06	2.71	0.17	0.06	2.84
GC	1	0.06	0.06	1.07	0.02	0.06	0.35
CC	5	0.21	0.07	2.97	0.20	0.07	2.91

Table 2.9: The concentrations of dimers (for primates) in a window of 100 bp located upstream and downstream from the repeat of length $\ell \geq 10$.

The concentrations of dimers (for primates) in a window of 100 bp located upstream and downstream from the repeat of length $\ell \geq 10$; the average concentrations of these dimers in the entire DNA and the ratios of the two above values. The total number of dimers found in all windows is shown in the second column. Note that almost all dimers exhibit abnormal concentration (ratio ≥ 1) in the windows surrounding long repeats, which might be a sign of mutation mechanisms discussed in Subsection 2.3.5

However, the above mechanism might not be the case for AA repeats. These repeats are usually located downstream from Alu sequences [22]. In agreement with this observation (Table 2.9), one finds that the concentration of AA dimers in the window of length 100 bp located upstream from a long AA repeat (more than 10 copies) constitutes only 80% of the average AA concentration (see Table 2.9). Indeed, the concentration of AA dimers in the Alu sequence is less than the average AA concentration. In contrast, the concentration of AA dimers in the window of the same length located downstream is 160% of the average overall AA concentration. One has the opposite picture for the concentration of TT dimers: the concentration of TT dimers is larger upstream than downstream from a long TT repeat.

In conclusion, that the presence of abnormally high concentrations of asymmetric dimers may, in some cases, be related to the power-law distribution of tandem repeats. It is unlikely, however, that the patchiness itself is the main biological reason for the power-law distribution of DTR. In fact, the expansion of DTR with subsequent mutations which result in the breaking of long DTR into shorter segments may result in the growth of patches with abnormally high concentrations of these dimers.

2.3.5 Can mutations explain non-exponential distributions of repeats?

Recently, several mechanisms of SSR expansion have been proposed [13, 15, 17, 19, 27, 56, 85, 96]. Most of the models can not produce non-exponential tails [56, 96] (see also a critical review by Li [63]). A mechanism based on known biological processes, has been proposed in [42, 43, 97], which can reproduce the observed non-exponential distribution of repeats. The increase or decrease of repeat length can occur due to unequal crossover [10, 98] or slippage during replication [15, 27, 99, 100]. Section 2.4 is dedicated to a model which has been developed [97] to understand the length distributions of DTR (and SSR, in general) resulting from the unequal chromosomal crossing over. This model is also applicable to the mutations due to slippage during replication (see Sec. 2.5 for discussion).

2.4 Model of unequal chromosomal crossing over

The crossing over of two homologous chromosomes (paternal and maternal) during meiosis is one of the major contributions to the reassortment of genetic information. During meiosis the paired chromosomes can both break due to the tension resulting from their mutual attraction. Further, the broken ends can be rejoined to the original parental chromosomes or can cross over to join the homologous parental chromosomes. It is sometimes possible that breakage occurs unequally in the two chromosomes. In the case that cross over occurs with such unequal breakage, each parental chromosome changes in length, one becomes longer, while the other becomes shorter. The next model is based on this mechanism of unequal chromosomal crossing over:

Model 2.1 *Consider a segment with a DTR of length ℓ (see Fig. 2.9).* Unequal crossing over *is defined to be a process when a segment of a repeat of length x on chromosome* **A** *(maternal) joins a segment of a repeat of length $\ell - y$ on chromosome* **B** *(paternal). As a result, two repeats of length $\ell - x + y$ (on chromosome* **A***) and $\ell + x - y$ (on chromosome* **B***) are obtained, i. e. the chromosome lengths change as follows*

$$\ell_{\mathbf{A}} \to \ell'_{\mathbf{A}} \equiv \ell_{\mathbf{A}} \cdot (1 - \xi_x + \xi_y)\,, \tag{2.11}$$

or

$$\ell_{\mathbf{B}} \to \ell'_{\mathbf{B}} \equiv \ell_{\mathbf{B}} \cdot (1 + \xi_x - \xi_y)\,, \tag{2.12}$$

where $\xi_x \equiv x/\ell$ and $\xi_y \equiv y/\ell$ range from 0 to 1.

Without loss of generality, one follows the evolution of chromosome **A**. Eq. (2.11) can be rewritten in terms of one parameter r, which is called here *expansion ratio*:

$$\ell \to \ell' = \ell \cdot r\,, \tag{2.13}$$

where $r \equiv 1 - \xi_x + \xi_y$ ranges from 0 to 2. Thus, if $0 < r < 1$, the repeat shrinks in length, while if $1 < r < 2$, the repeat grows.

Remark 2.1 *Note that in the mathematical sense, chromosomes* **A** *and* **B** *are equivalent upon the transformation $x \leftrightarrow y$.*

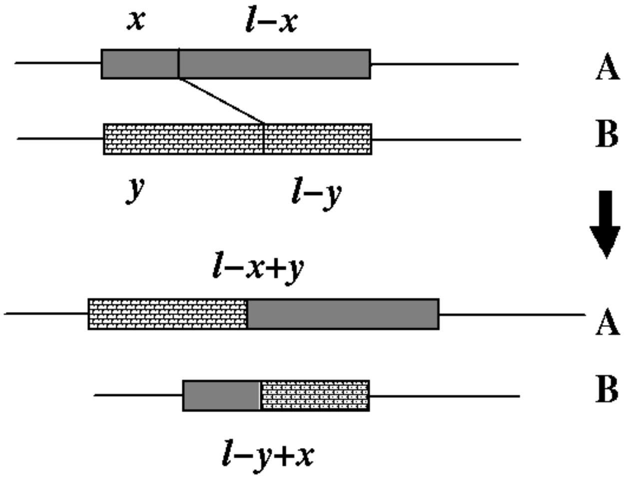

Figure 2.9: Unequal chromosomal crossing over.
Crossing over of two homologous chromosomes during meiosis. Due to unequal crossing over, the part containing DTR in the maternal chromosome **A** of length x joins the part containing the same type of DTR in the paternal chromosome **B** of length $\ell - y$. As the result, the length of the chromosome **A** changes to $\ell - x + y$. Similarly, the corresponding part of maternal chromosome **A** of length $\ell - x$ joins the the same type of DTR in the paternal chromosome **B** of length y. As the result, the length of the chromosome **B** changes to $\ell + x - y$

To derive the probability distribution function $C(r)$ of variable r one needs to integrate over all possible configurations of ξ_x and ξ_y to get a particular value of r:

$$C(r) = \int_0^1 d\xi_x f_1(\xi_x) \int_0^1 d\xi_y f_2(\xi_y) \cdot \delta(r - 1 - \xi_x + \xi_y)\,, \tag{2.14}$$

where f_1 and f_2 are distribution functions of x and y. In the most simple case, both x and y are distributed uniformly on the interval $[0, 1]$, so that f_1 and f_2 are equal to 1 on the interval $[0, 1]$ and equal to zero otherwise. In this case, after integration, Eq. (2.14) yields (see Fig. 2.10)

$$C(r) = \begin{cases} r, & 0 < r \leq 1 \\ 2 - r, & 1 < r \leq 2 \\ 0, & r \leq 0 \text{ or } r > 2\,. \end{cases} \tag{2.15}$$

Clearly, $C(r)$ is normalized for any f_1 and f_2 , i. e.

$$\int_0^\infty C(r)dr = 1\,. \tag{2.16}$$

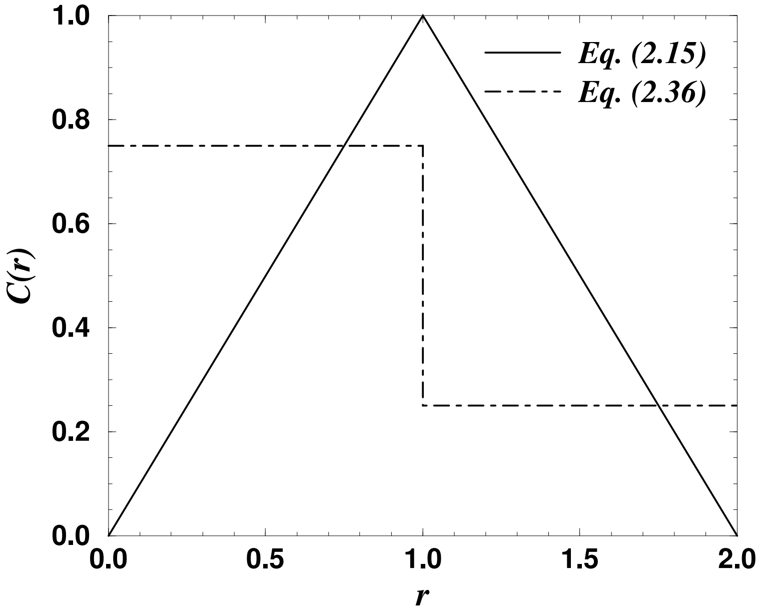

Figure 2.10: $C(r)$

Plot of distribution function $C(r)$ of variable r for the case of Eq. (2.15) (solid line) and Eq. (2.36) (dot-dashed line).

After t steps of evolution the length of the repeat is given by $\ell_t = \ell_0 \cdot \prod_{i=1}^{i=t} r_i$. Such a "random multiplicative process" in many cases leads in the long time limit ($t \to \infty$) to a stable distribution of repeat length $N(\ell)$. To avoid extinction of repeats, one can either *(a)* set a non-zero probability to reappear, or *(b)* set $\pi(r, \ell) = 0$ when $rl \leq 1$. Both ways are mathematically equivalent, since it is always possible to control the probability of creation of new repeats at $\ell = 1$ using condition *(a)* so that one reaches condition *(b)*. These conditions, *(a)* or *(b)*, might be biologically caused by point mutations [10] — the random substitution of a nucleotide by other ones, since they can create repeats of length $\ell = 1$.

If one changes variables to $z_t \equiv \ln \ell_t$, the dynamics of z_t becomes a random diffusion process in semi-infinite space $z > 0$ (due to the fact that $\ell \geq 1$) with a reflecting wall at $z = 0$ and an attractive potential. $\ell \geq 1$ corresponds to $z \geq 0$, which reflects the fact that the DTR cannot disappear. A classical example of such a process is Brownian motion in a potential field, which leads in the continuum limit to an exponential decrease of the atmosphere's density. Thus, in this case, the probability distribution function for the final outcome of this process can be written as

$$\bar{N}(z) \sim e^{-kz}\,, \tag{2.17}$$

where k is some constant. After transforming back to the original variables, Eq. (2.17) can be rewritten in the form of power law,

$$N(\ell) \sim \ell^{-\mu}\,, \tag{2.18}$$

where $\mu = k + 1$.

To derive this result, one can write the master equation for the z-variable in the continuous limit:

$$\frac{d\bar{N}(z,t)}{dt} = \int_{-\infty}^{+\infty} \bar{N}(z-x,t) \cdot C(x)dx - \bar{N}(z,t)\,. \tag{2.19}$$

If one assumes that there is a stable distribution, then one needs to require that in the long time limit ($t \to \infty$)

$$\frac{d\bar{N}(z,t)}{dt} = 0\,. \tag{2.20}$$

Thus, the Eq. (2.19) can be rewritten in the following form:

$$\int_{-\infty}^{+\infty} \bar{N}(z-x,t) \cdot C(x)dx - \bar{N}(z,t) = 0\,. \tag{2.21}$$

If one assumes that $\bar{N}(z,t)$ has a solution in the form of (2.17), then one obtains an additional condition on $C(x)$, which can be written in the following form:

$$\int_{-\infty}^{+\infty} C(x)\exp{(kx)}dx - 1 = 0\,. \tag{2.22}$$

Therefore, after transforming back to the original variables, Eq. (2.22) can be rewritten

$$\int_0^2 C(r)r^{\mu-1}dr = 1\,. \tag{2.23}$$

After integration, with the density $C(r)$ specified in Eq. (2.15), Eq. (2.23) yields two solutions $\mu = 1$ and $\mu = 2$. The first solution does not produce a normalizable distribution, so one can disregard it.

Remark 2.2 *Note that the reflecting boundary condition is crucial for this model. In the case when there are no boundaries, diffusion in the presence of the constant field yields*

$$\bar{N}(z,t) = \frac{1}{\sqrt{2\pi Dt}} \exp\left(\frac{-[z-vt]^2}{2Dt}\right), \tag{2.24}$$

where D and v are parameters of the field. The reflecting boundary condition is a special condition, which produces the exponential solution (2.17).

Remark 2.3 *In fact, the distribution function $C(r)$ can have a more general shape than that given by Eq. (2.15).*

Let us prove a useful theorem.

Theorem 2.1 *If $C(r)$ is normalized (i. e. Eq. (2.16) holds) and symmetric about point $r = 1$, i. e.*

$$C(2-r) = C(r)\,, \tag{2.25}$$

then $C(r)$ satisfies Eq. (2.23) with $\mu = 2$.

Proof: First of all, note that if Eq. (2.16) holds then

$$\int_0^2 C(2-r)dr = -\int_2^0 C(r')dr' = \int_0^2 C(r')dr' = 1\,. \tag{2.26}$$

If one denotes

$$I = \int_0^2 C(r)r dr\,, \tag{2.27}$$

then Eq. (2.25) yields

$$I = \int_0^2 C(2-r)r dr = \int_0^2 C(r')(2-r')dr' = 2 - I\,. \tag{2.28}$$

From the above equation one can conclude that $I = 1$. *QED.*

From this theorem one can conclude that any normalized distribution function $C(r)$, which is symmetric relative to $r = 1$, yields a power law DTR length distribution function. Therefore, one can easily guess the solution of Eq. (2.23) if one assumes a more simple form of $C(r)$ than that given by Eq. (2.15). Two examples have been used in [42, 97],

(i)

$$C(r) = \begin{cases} 1/2, & 0 < r \leq 2 \\ 0, & r \leq 0 \text{ or } r > 2\,. \end{cases} \tag{2.29}$$

and *(ii)*

$$C(r) = \pi_1 \cdot \delta(r - 1/2) + \pi_2 \cdot \delta(r - 2)\,, \tag{2.30}$$

where $\pi_1 + \pi_2 = 1$ and $\delta(r)$ is the Dirac delta-function. The first example *(i)* yields exponent $\mu = 2$, the second *(ii)* yields

$$\mu = 1 + \log_2(\pi_1/\pi_2)\,. \tag{2.31}$$

Remark 2.4 *Note that in the example (ii) μ ranges from 2 to 4.5 if π_1/π_2 ranges from 2 to 11.3. The fact that π_1 is greater than π_2 means that the probability for a repeat to shrink is larger than the probability to expand, which is biologically plausible since the repeats are preserved from unlimited expansion — i. e. the average repeat length does not diverge. In mathematical terms, the restriction $\pi_1/\pi_2 > 1$ is a necessary condition to obtain a stable probability distribution function, because otherwise condition (2.20) is not satisfied.*

Remark 2.5 *Indeed, in order to obtain exponent $\mu = 2$ one could impose less restrictive requirements for $C(r)$ than Eq. (2.25). In fact, one does not even need to assume that the distributions $f_1(\xi_x)$ and $f_2(\xi_y)$ in Eq. (2.14) are uniform. The only restriction is that the mathematical expectations of both ξ_x and ξ_y are the same*

$$\langle \xi_x \rangle = \langle \xi_y \rangle\,, \tag{2.32}$$

Theorem 2.2 *If variables ξ_x and ξ_y, distributed according to the functions $f_1(\xi_x)$ and $f_2(\xi_y)$ correspondingly, and their mathematical expectations are equal, i. e. Eq. (2.32) holds, then $C(r)$ satisfies Eq. (2.23) with $\mu = 2$.*

Proof: For simplicity, let us change variables from r to $\tilde{r} \to r - 1$. After this transformation, Eq. (2.14) can be rewritten as

$$\tilde{C}(\tilde{r}) = \int_0^1 d\xi_x f_1(\xi_x) \int_0^1 d\xi_y f_2(\xi_y) \delta(\tilde{r} - \xi_x + \xi_y)\,, \tag{2.33}$$

where $\tilde{r}$ ranges from $[-1,1]$. Also, Eq. (2.23) in terms of $\tilde{r}$ means that

$$\int_{-1}^{1} \tilde{C}(\tilde{r}')\tilde{r}'d\tilde{r}' = 0\,. \tag{2.34}$$

Combining Eqs. (2.33), (2.34) and (2.16) one finds

$$\begin{aligned} 0 &= \textstyle\int_{-1}^{1} \tilde{C}(\tilde{r}')\tilde{r}'d\tilde{r}' = \int_0^1 d\xi_x f_1(\xi_x) \int_0^1 d\xi_y f_2(\xi_y) \int_{-1}^{1} d\tilde{r}'\tilde{r}'\delta(\tilde{r}' - \xi_x + \xi_y) \\ &= \textstyle\int_0^1 d\xi_x f_1(\xi_x) \int_0^1 d\xi_y f_2(\xi_y)(\xi_x - \xi_y) = \langle \xi_x \rangle - \langle \xi_y \rangle\,, \end{aligned} \tag{2.35}$$

Thus, if Eq. (2.32) is satisfied, then Eq. (2.23) holds for $\mu = 2$. *QED.*

2.4.1 Discussion

The above model is able to produce power law behavior of DTR length distribution functions. The exponent $\mu = 2$ is different for real DNA sequences, where μ ranges from 2.5 to 4.5 for different taxonomic classes and different DTR. This difference can be attributed to processes, other than unequal crossover, that take place in nature.

(i) One possible scenario for $\mu > 2$ is that it can be due to the selective pressure against fixations of the longer repeats in the population, so that the probability of chromosomal growth is larger than that of chromosomal shrinkage. Thus, the function $C(r)$ becomes asymmetric and skewed to the left from the symmetry point $r = 1$. In this case, the solution to Eq. (2.23) yields power law function with exponent $\mu > 2$. For example, if instead of Eq. (2.29) function $C(r)$ is asymmetric relative to the point $r = 1$ (see Fig. 2.10):

$$C(r) = \begin{cases} 3/4, & 0 < r \leq 1 \\ 1/4, & 1 < r \leq 2 \\ 0, & r \leq 0 \text{ or } r > 2\,, \end{cases} \tag{2.36}$$

then the solution of Eq. (2.23) yields power law function DTR length distribution function $N(\ell)$ with the exponent $\mu \approx 3.7$. This example shows that with slight asymmetry of function $C(r)$ one can achieve any exponent $\mu > 2$ by varying the skewness of the function $C(r)$. Another example is set by Eq. (2.30) where the skewness of the function $C(r)$ is controlled by the ratio π_1/π_2 (see Eq. (2.31)).

(ii) One can suspect, however, that the higher observed exponent μ comes from point substitutions of one nucleotide by another, which are more likely to target longer repeats [96]. Such mutations would destroy long repeats and create more shorter repeats. This would result in a decrease of values of the DTR length distribution functions at the tail of distributions and, therefore, an increase of the exponent μ.

2.4.2 The dependence of mutation rates on repeat length

The assumption that $\pi(r, \ell) \equiv C(r)$ is independent of ℓ is an approximation. Due to the specific biochemical mechanisms, mutation rates may depend on the length ℓ of the DTR [17, 18, 27, 64, 101]. There might be other mechanisms which in particular target longer repeats [15, 17, 18, 37, 102, 103] (see also Sec. 2.5). For example, the power-law behavior usually starts from repeats of length $\ell \geq 5$. This may indicate that short repeats are distributed exponentially, as in random sequence. Thus, it is plausible to conclude that the mutation processes target repeats of length above $\ell \geq 5$.

To further test the model, the DTR length distributions for the artificial sequences are computed numerically in the case when $\pi(r, \ell)$ depends both on r and ℓ. These artificial sequences are generated to reproduce the actual DTR distributions of the complete yeast genome (total length 6,325,440 bp (see Fig. 2.11)).

The distribution of DTR in coding DNA (see Fig. 2.11a) of yeast can be fit by a straight line on a semilogarithmic scale, indicating that DTR in coding DNA are either uncorrelated or short-range correlated. Such behavior of DTR in coding DNA is consistent with Eqs. (2.2) and (2.5).

The DTR distributions in noncoding DNA (see Fig. 2.11b) for three groups of DTR: *(i)* AA, (TT); *(ii)* TA, (AT) and *(iii)* CA, (AC), TG, (GT) can be fit by a straight line on a double-logarithmic plot, so Eq. (2.6) approximately holds. The slopes for these three groups of DTR are $\mu = 4.7$, 4.2, and 4.8 (see Remark 2.6). The rest of DTR can not be fit by power law functions (see Remark 2.7).

Remark 2.6 *The correlation coefficients for the first three groups of DTR are* $R =$

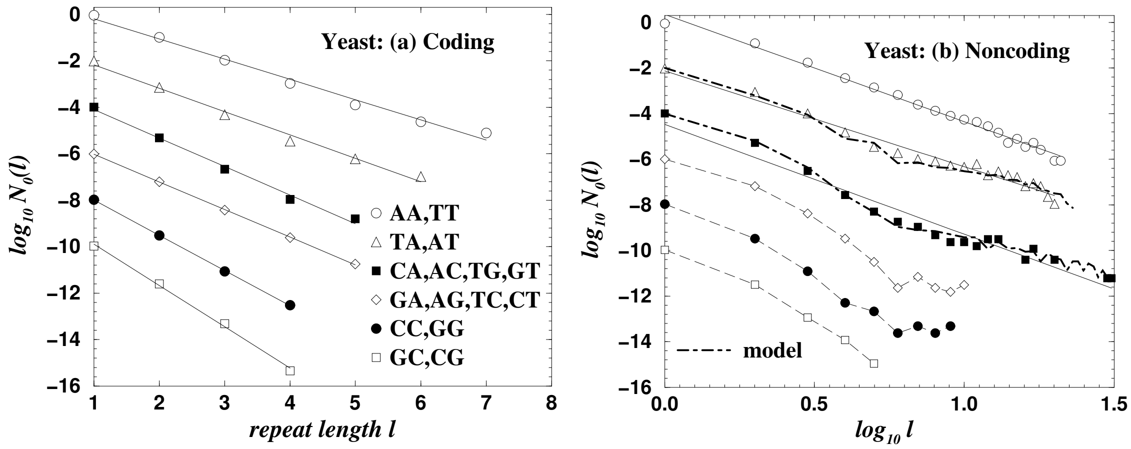

Figure 2.11: The dependence of mutational rates on repeat length.

The combined results for six groups of DTR: AA, TT (AA or TT) (◯); TA, AT (△); CA, AC, TG, GT (■); CC, GG (●); GA, AG, TC, CT (◇); and GC, CG (□) are shown for the complete yeast genome (total length 6,325,440 bp; total length of the *coding* DNA 3,549,593 bp). This classification is used because A is complementary to T, and C is complementary to G; also, it is averaged over two possible directions of reading of DNA sequences. For clarity plots for these six groups are separated by shifting them by $\log_{10}(100) = 2$. (a) Semilogarithmic plot of $N_0(\ell) \equiv N(\ell)/N(1)$ for *coding* DNA. Note that the data fall on straight lines, so Eqs. (2.2) and (2.5) hold. In all 16 yeast chromosomes there are 10 occurrences of DTR of length greater or equal to 10, which are not shown on the graph because they belong to putative genes (they are denoted by "not experimental" in the GenBank data base). (b) Double logarithmic plot of $N_0(\ell)$ for the *noncoding* DNA. Now, the data for the first three groups can be fit by a straight line, so Eq. (2.6) approximately holds. One finds that the slope $\mu = 4.7$, 4.2, and 4.8 for the first three groups of DTR 2.6. The other three groups of DTR can not be fit by power law functions 2.7. As an example of $\pi(r, \ell)$ being a function of both r and ℓ, the results of simulations (dot-dashed bold line) fitting the second and the third groups of repeats are included (see Fig. 2.13).

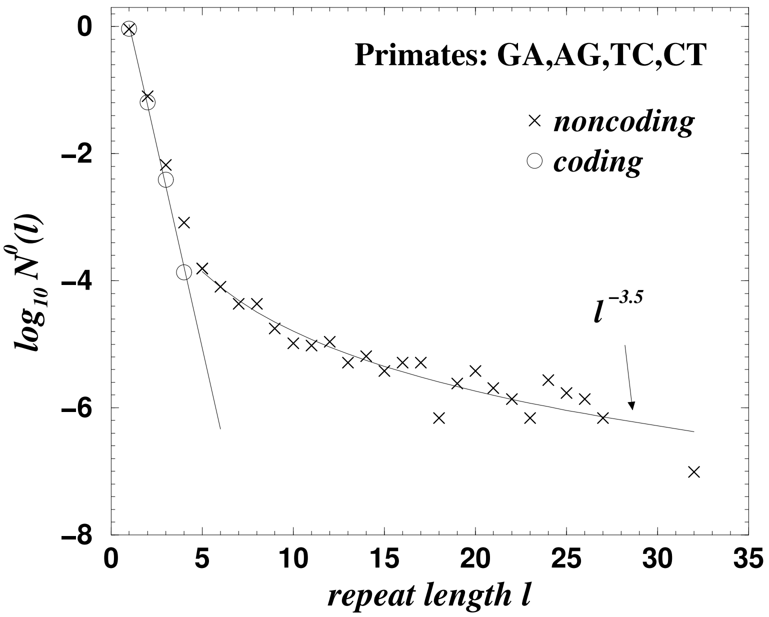

Figure 2.12: Power-law tail of the length distribution of GA, AG, TC, CT DTR. Semilogarithmic plot of $N_0(\ell)$ for all primate sequences in GenBank; as an example the combined results for GA, AG, TC, CT are shown. Note that the data for *coding* DNA fall on a straight line, so Eqs. (2.2) and (2.5) hold. The data for *noncoding* DNA can be approximately fit by a power law function, so Eq. (2.6) holds. One finds that the exponent $\mu = 3.5$ for fitting range $\ell \in [5, 35]$, with confidence value $R = 0.98$. The exponent is sensitive to the fitting range, e. g. if one concentrates on the tail of the distribution $\ell \in [10, 35]$ the exponent $\mu = 3$ is obtained. The difference of length distribution functions for *coding* and *noncoding* DNA is dramatic; one can observe a DTR of length of 10 dimers in *noncoding* DNA (with probability, roughly, $p \approx 10^{-5}$), while it is seven orders of magnitude less probable to find such DTR in *coding* DNA ($p \approx 10^{-12}$)!

0.99, 0.99, and 0.98. For comparison, the fit of DTR length distribution functions to an exponential (on semilogarithmic plot) is not nearly as good, yielding $R = 0.93$, *0.90, and 0.80 for these groups of DTR. The fitting range is* $\ell \in [5, 31]$ *for both methods of fitting.*

Remark 2.7 *Note that for the noncoding yeast DNA one finds that the distributions of* GA, AG, TC, CT *do not have power law tails, while for vertebrates (see Fig. 2.12) they do.*

In order to reproduce the deviations of the second *(ii)* and the third *(iii)* groups of repeats from the power-law behavior, the artificial sequences are produced from a random sequence with equal concentration of all dimers $1/16 = 0.0625$. The dimeric repeats in such sequences are expanded or shrunk in length by a multiplication on a random number r, which ranges from 0 to 2. Thus, if $0 < r < 1$ the repeat is shrunk, and if $1 < r < 2$, the repeat is expanded (see Eq. (2.13)). 10^6 iterations of such random multiplicative process are produced. The distribution $\pi(r, \ell)$ of the expansion ratios r is also a function of length ℓ and set as follows (see Fig. 2.13):

(i) For TA, (AT) repeats, $\pi(r, \ell)$ depends on ℓ as a step function: for $1 < r \leq 2$: $\pi(r, \ell) = 0.15$, when $0 < \ell < 6$; 0.60, when $6 \leq \ell < 13$; and 0, when $\ell \geq 13$. For $0 < r \leq 1$: $\pi(r, \ell) = 0.85$, when $0 < \ell < 6$; 0.40, when $6 \leq \ell < 13$; and 1, when $\ell \geq 13$.

(ii) For CA, (AC), TG, (GT) repeats, $\pi(r, \ell)$ is also a step function of ℓ: for $1 < r \leq 2$: $\pi(r, \ell) = 0.016$, when $0 < \ell < 5$; 0.48, when $5 \leq \ell < 8$; 0.32, when $8 \leq \ell < 18$; and 0, when $\ell \geq 18$. For $0 < r \leq 1$: $\pi(r, \ell) = 0.984$, when $0 < \ell < 5$; 0.52, when $5 \leq \ell < 8$; 0.68, when $8 \leq \ell < 18$; and 1, when $\ell \geq 18$.

Remark 2.8 *Even though the dependence of the mutational rates of the repeat length* ℓ *causes deviation of the model distributions from power law behavior, the power law functions often provide satisfactory approximations to the experimental data. For instance, the first group of repeats* (i) AA *and* (TT) *are best fit by a power-law function.*

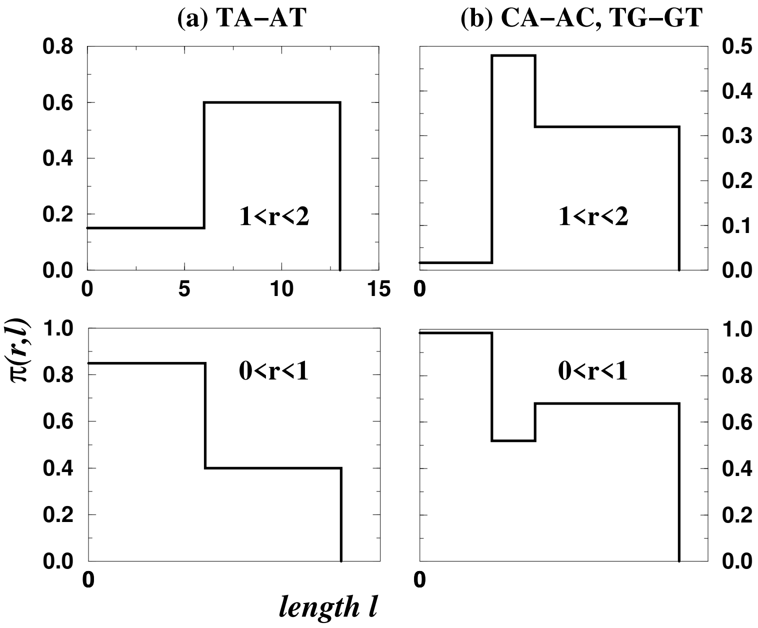

Figure 2.13: Model $\pi(r, \ell)$ functions.

As an example of $\pi(r, \ell)$ being a function of both r and ℓ, functions $\pi(r, \ell)$ are chosen in the following way to fit TA, AT CA, AC, TG, GT repeats in noncoding yeast DNA (see Fig. 2.11). (a) For TA, AT repeats, $\pi(r, \ell)$ depends on ℓ as a step function: for $1 < r \leq 2$: $\pi(r, \ell) = 0.15$, when $0 < \ell < 6$; 0.60, when $6 \leq \ell < 13$; and 0, when $\ell \geq 13$. For $0 < r \leq 1$: $\pi(r, \ell) = 0.85$, when $0 < \ell < 6$; 0.40, when $6 \leq \ell < 13$; and 1, when $\ell \geq 13$. For $r > 2$ it is assumed that $\pi(r, \ell) = 0$, which means that a repeat cannot expand by more than a factor of 2 in a single mutation. (b) For CA, AC, TG, GT repeats, $\pi(r, \ell)$ is also a step function of ℓ: for $1 < r \leq 2$: $\pi(r, \ell) = 0.016$, when $0 < \ell < 5$; 0.48, when $5 \leq \ell < 8$; 0.32, when $8 \leq \ell < 18$; and 0, when $\ell \geq 18$. For $0 < r \leq 1$: $\pi(r, \ell) = 0.984$, when $0 < \ell < 5$; 0.52, when $5 \leq \ell < 8$; 0.68, when $8 \leq \ell < 18$; and 1, when $\ell \geq 18$. In the case of both groups of repeats, simulations start from a random sequence with equal concentration of all dimers $1/16 = 0.0625$ and run for 10^6 iterations of the random multiplicative process.

Varying $\pi(r, \ell)$, one finds even better agreement with experimental data (Fig. 2.11) for each particular case. The non-universality of $\pi(r, \ell)$ is biologically plausible since the mutation rates may strongly depend on the organism as well as on the repeat type. However, such mutation rates are presently not known. Therefore, the model can serve as a tool for determining and testing the mutation rates by fitting the experimental data.

2.5 Slipped Strand Misspairing and other mutations

Slipped-strand misspairing (SSM) has been proposed in Refs. [15, 17, 18, 27, 37] as the mechanism to explain the expansion of trimeric tandem repeats in coding DNA that results in various neurological diseases (see Sec. 2.1). It was proposed that SSM arise [27, 104] during replications (Fig. 2.14), when multiple slippages are promoted by an "idling polymerase" that is caused by a strong block. Such an idling block, which can be either a protein or some DNA structure, causes continuous primer realignment (slippage). If slippage occurs on the leading strand of DNA, the expanded repeat is generated. If slippage occurs on the lagging strand of DNA, the deletion of the repeat occurs. Such a mechanism favors longer repeats, since the probability to slip grows with the length of the repeat.

It has also been pointed out that different repeats slip with different probability. The reason for such difference is due to the different binding strength of repeats. For example, it has been shown that among triplets the CTG repeats are expanded preferentially in *E. coli* [105].

Other models of triplet repeat expansion are examined in [27]. It is plausible to assume that such models for triplets can also be applicable to dimeric repeats.

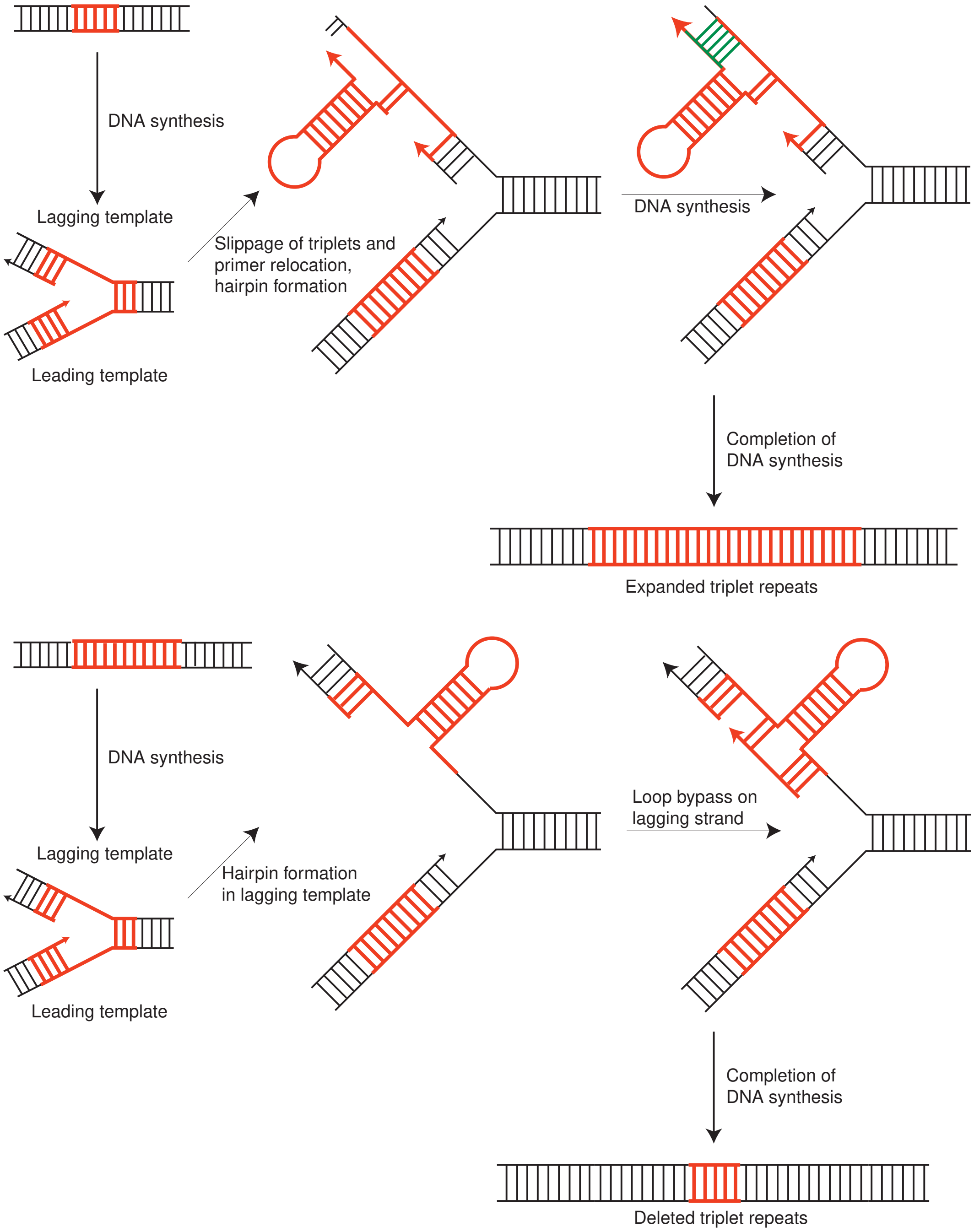

Figure 2.14: Model for insertions and deletions in SSM.

Model for insertions (upper diagram) and deletions (lower diagram) in SSM. Adapted from [27].

2.6 Does bias of the GenBank affect the distributions of DTR?

In order to understand if the bias in the GenBank affects the findings, the known completed genomes of the *Saccharomyces cerevisiae* (yeast) [42] (see Fig. 2.12) and *Caenorhabditis elegans* (see Fig. 2.15a) and a single complete human gene *HUMTCRB* [106] (Fig. 2.15b) have been studied. If the findings of the long tails in the length distributions of some dimeric tandem repeats were the result of bias, then one would find the distributions to be drastically different from ones for the complete genomes. On the contrary, for the yeast and *Caenorhabditis elegans*, it has been found that some dimeric tandem repeats have long tails. Even though the maximal repeat size is of the order of 20 in a single human gene *HUMTCRB*, DTR distributions also show power-law trend.

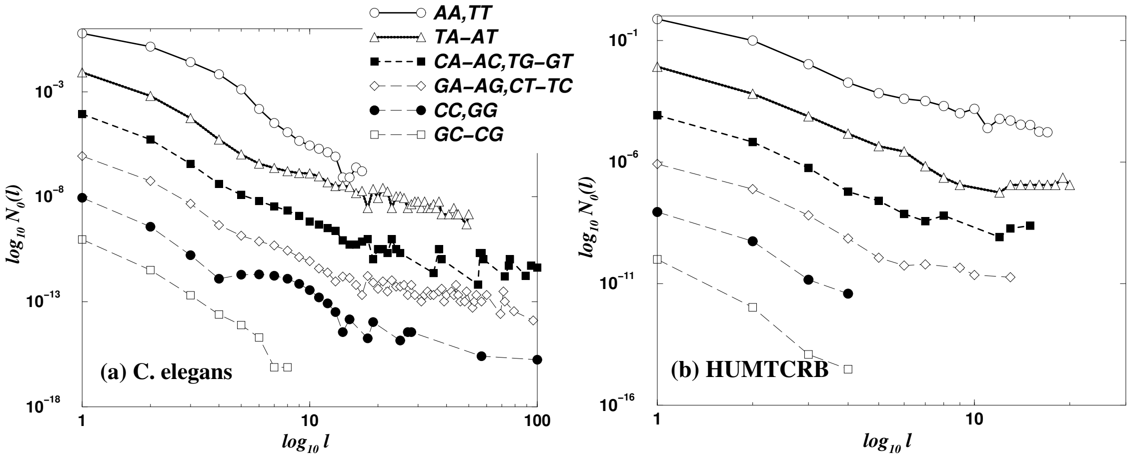

Figure 2.15: Distributions of DTR in noncoding DNA of C. elegans and a single human gene *HUMTCRB*.

The combined results for six groups of DTR: AA, TT (AA or TT) (◯); TA, AT (△); CA, AC, TG, GT (■); CC, GG (●); GA, AG, TC, CT (◇); and GC, CG (□) are shown for the noncoding part of the (a) complete genome of *Caenorhabditis elegans* (6 chromosomes; total length 48,381,234 bp); and (b) a single human gene *HUMTCRB* (total length 657,333 bp).

In order to answer the question of how a biased Genbank can affect the distributions of tandem repeats, let us denote by $F_i(\ell)$ the distributions of the tandem repeats in a GenBank sequence of length n_i, which has been submitted d_i times. Let the total length of the GenBank be L nucleotides and the actual distribution of the repeats be $P(\ell)$. Then,

$$P(\ell) = \frac{1}{L}\sum_i F_i(\ell) n_i d_i = \sum_i F_i(\ell) w_i\,, \tag{2.37}$$

where $w_i = n_i d_i / L$ and $\sum_i w_i = 1$.

For large ℓ, one finds that $P(\ell) \sim \ell^{-\mu}$ (Eq. (2.6)). Next, the assumption is made that the first and the second moments of the w_i distribution $W(w_i)$ exist. Then, one can express w_i via the average $\langle w \rangle$ and the deviations δw_i: $w_i = \langle w \rangle + \delta w_i$. Then, Eq. (2.37) can be written

$$\ell^{-\mu} \sim \langle w \rangle \sum_i F_i(\ell) + \sum_i F_i(\ell)\delta w_i \tag{2.38}$$

Since $\langle \delta w_i \rangle = 0$ and it is assumed that δw_i and $F_i(\ell)$ are not correlated, the above equation yields

$$\ell^{-\mu} \sim \langle w \rangle \langle F(\ell) \rangle_w. \tag{2.39}$$

Thus,

$$\langle F(\ell) \rangle_w \sim \ell^{-\mu}\,. \tag{2.40}$$

Therefore, the functional form of $F(\ell)$ is the same as that of the $P(\ell)$ if the assumptions of *(i)* the presence of the first and the second moments and *(ii)* distributions $W(w_i)$ and $F(\ell)$ are not correlated.

2.7 Clumping of oligonucleotides in coding and noncoding DNA

2.7.1 Clumping measures

Clumping measures have been implemented in [107] to quantify the repetitions of oligomers in all possible reading frames (RF). Coding DNA has three RF, one of which is the correct one. Noncoding DNA, however, does not have fixed RF, so there is no *a priori* reason to select a given one. The results for noncoding DNA remain roughly the same independent of RF, suggesting that the choice of RF is insignificant for noncoding DNA. Thus, only results using the correct RF for coding DNA and a single RF arbitrarily chosen for noncoding DNA are reported here.

Clumping measures are defined as follows. First, the actual number of occurrences of the repetitions for each oligomer is counted. Then, the control value is calculated, where the control is obtained by scrambling the order of the oligomers. If all the oligomers were evenly represented, each would have a concentration of $1/4^n$, where n is the size of the oligomers, e. g. $n = 1$ for monomers, $n = 2$ for dimers, etc. Since the concentrations of the oligomers vary, it is necessary to calculate the actual concentration of each trinucleotide based on its concentration in the GenBank sequence.

To calculate the control, the probability P_ℓ that a given oligomer belongs to a cluster of *exactly* ℓ repetitions in a random sequence is computed:

$$P_\ell = \ell p_i^\ell (1 - p_i)^2 \tag{2.41}$$

where p_i is the concentration of a particular oligomer. The number of oligomers that belong to clusters of size ℓ is obtained by multiplying each of these probabilities (a distinct probability for each of the oligomers) by L, the total number of oligomers in the GenBank sequence. Thus, the control value of the mean number of clusters $N_i(\ell)$ of size ℓ is given by

$$N_i(\ell) = L p_i^\ell (1 - p_i)^2 \,. \tag{2.42}$$

Next, the number of repeats of length ℓ of a given repeat in different partitions of the GenBank is computed: $N_i(\ell)$, where $i = 1, \ldots, M$ is the index of an oligomer and M is the total number of types of oligomers of size n: $M = 4^n$. Note that according to definition,

$$\sum_{\ell=1}^{\infty} \sum_{i=1}^{M} N_i(\ell)\ell = L \,. \tag{2.43}$$

Here, the the effect of the finite sequence length is neglected. Coding sequences, for which exons are concatenated together, and noncoding sequences are analyzed separately. The latter are identified as those that are not explicitly specified as exons or introns in the GenBank data base. In order to deal with the bias in the GenBank database due to multiple entries of short copies of some fragments of the larger DNA sequences, only sequences of total length at least 10^4bp are selected. This significantly reduces the redundancy of the GenBank. The total length and the number of sequences analyzed in coding and noncoding regions is reported in Table 2.10.

Next, two measures of length are introduced: *1)* the length average

$$\langle \ell \rangle_n = \frac{L}{N}, \tag{2.44}$$

where

$$N \equiv \sum_{\ell=1}^{\infty} \sum_{i=1}^{M} N_i(\ell) \tag{2.45}$$

is the total number of repeat occurrences; *2)* the "weight" average (see e. g. [108]):

$$\langle \ell \rangle_w \equiv \frac{\sum_{\ell=1}^{\infty} \sum_{i=1}^{M} \ell^2 N_i(\ell)}{\sum_{\ell=1}^{\infty} \sum_{i=1}^{M} \ell N_i(\ell)} = \frac{\sum_{\ell=1}^{\infty} \sum_{i=1}^{M} \ell^2 N_i(\ell)}{L}. \tag{2.46}$$

The second definition gives larger weight to long repeats, since the probability that a random oligomer belongs to a repeat of length ℓ is ℓ times larger than the probability of finding a repeat of length ℓ by counting all the repeats in the sequence.

It is possible to calculate theoretical predictions for both measures of repeat lengths for a control sequence in which the order of oligomers is randomly scrambled (see Appendix A for the derivation). Such an uncorrelated random sequence $\langle \ell \rangle_n^{\mathrm{th}}$ can be expressed as

$$\langle \ell \rangle_n^{\mathrm{th}} = \frac{1}{1 - \sum_{i=1} p_i^2}, \tag{2.47}$$

where p_i is the concentration of each oligomer, and

$$\langle \ell \rangle_w^{\mathrm{th}} = 1 + 2 \sum_{i=1} \frac{p_i^2}{1 - p_i}. \tag{2.48}$$

Next, to quantify relative clumping strength, two ratios are introduced

$$R_n \equiv \frac{\langle \ell \rangle_n - 1}{\langle \ell \rangle_n^{\mathrm{th}} - 1} \qquad \text{and} \qquad R_w \equiv \frac{\langle \ell \rangle_w - 1}{\langle \ell \rangle_w^{\mathrm{th}} - 1}. \tag{2.49}$$

The clumping ratios compare actual repeat length with the control case in which, by definition, no clumping occurs beyond the clustering that occurs in uncorrelated random process.

2.7.2 Results and Discussion

The ratios of the observed values of the average length of oligomers[2] (monomers, dimers, trimers, and tetramers) and their length weight average to the theoretically predicted for a randomly scrambled sequence are studied for primate, vertebrate, invertebrate, mammal, rodent, and plant taxonomic partitions of GenBank release 104.0. The complete results for clumping ratio values and for the error bars of these values[3] are presented in Tables 2.11 and 2.12. To compute error bars, GenBank data sets are partitioned into 10 subsets of size of 10% of the GenBank data sets. The clumping ratios are computed for each set and from the distribution of the clumping ratios values the mean and the standard deviation presented in Tables 2.11 and 2.12 are determined. The probability that these distributions for coding and noncoding DNA belong to the same distribution is characterized by the p-value of Kolmogorov - Smirnov test (see [62]). If p is close to 1, then the two distributions under consideration, are drawn from the same distribution with the probability 1. If p is close to 0, then these distribution are taken from two different distributions with the probability $(1-p) \to 1$. p-values are presented in Tables 2.11 and 2.12 for each of the clumping ratio values. It is found that:

(i) There is a significant difference between the clumping of monomers (excluding plants), dimers, and tetramers in coding versus noncoding DNA. The p-values for all the distributions of ratio value sets of above mentioned groups of repeats do not exceed $2 \cdot 10^{-5}$.

(ii) The clumping ratios for the monomers in coding DNA for all the taxonomic partitions except plants are close to 1 (within 9%), which means that they are almost randomly distributed. For noncoding DNA, however, these values are consistently greater than 1 which indicates the slight clumping of monomers.

(iii) The clumping ratios for the dimers in coding DNA are also close to 1 (within 7%). However, these values are consistently smaller than 1 which indicates the slight repulsion of dimers in coding DNA.

---

[2] Those oligomers for which statistically significant results can not be obtained (pentamers, hexamers, etc.) are omitted in this study.

[3] The error bars are one standard deviation.

organism	noncoding	coding
Vertebrates	206,757	91,323
Primates	5,161,953	692,634
Invertebrates	5,322,555	6,993,572
Plants	738,506	4,946,293
Mammals	121,556	56,994
Rodents	1,131,628	253,200

Table 2.10: Clumping measure: the total length of the analyzed coding and noncoding regions.
The total length in bp of the analyzed coding and noncoding regions. The protein coding sequences are constructed by concatenating sequences denoted as *CDS* in the GenBank. The noncoding sequences are constructed by concatenating sequences not denoted as *CDS* in the GenBank

(iv) The clumping ratios for the trimers for all organisms show strong clumping of the trimers in both coding and noncoding DNA.

(v) The clumping ratio values for the tetramers in coding DNA are consistently and significantly smaller than 1 (up to 32%) which indicates the repulsion of tetramers.

Interestingly, the difference between the clumping of trimers in coding versus noncoding DNA is less pronounced. For vertebrates and mammals the Kolmogorov-Smirnov p-values for the R_n ratio are of the order of 1, which indicates that one cannot distinguish between coding and noncoding DNA based only on R_n ratios.

Monomers						
organism	R_n			R_w		
	coding	noncoding	p-value	coding	noncoding	p-value
Primates	1.09 ± 0.01	1.26 ± 0.01	$< 2 \cdot 10^{-5}$	1.08 ± 0.01	1.43 ± 0.01	$< 2 \cdot 10^{-5}$
Vertebrates	1.03 ± 0.01	1.14 ± 0.01	$< 2 \cdot 10^{-5}$	1.02 ± 0.01	1.24 ± 0.01	$< 2 \cdot 10^{-5}$
Inverebrates	1.09 ± 0.01	1.40 ± 0.01	$< 2 \cdot 10^{-5}$	1.08 ± 0.01	1.58 ± 0.01	$< 2 \cdot 10^{-5}$
Mammals	1.06 ± 0.01	1.28 ± 0.02	$< 2 \cdot 10^{-5}$	1.04 ± 0.02	1.43 ± 0.04	$< 2 \cdot 10^{-5}$
Rodents	1.06 ± 0.01	1.18 ± 0.01	$< 2 \cdot 10^{-5}$	1.03 ± 0.01	1.30 ± 0.01	$< 2 \cdot 10^{-5}$
Plants	1.13 ± 0.01	1.10 ± 0.01	$< 2 \cdot 10^{-5}$	1.12 ± 0.01	1.17 ± 0.01	$< 2 \cdot 10^{-5}$
Dimers						
	R_n			R_w		
	coding	noncoding	p-value	coding	noncoding	p-value
Primates	0.96 ± 0.01	1.39 ± 0.01	$< 2 \cdot 10^{-5}$	0.95 ± 0.01	1.73 ± 0.01	$< 2 \cdot 10^{-5}$
Vertebrates	1.00 ± 0.02	1.32 ± 0.02	$< 2 \cdot 10^{-5}$	1.00 ± 0.02	1.81 ± 0.13	$< 2 \cdot 10^{-5}$
Inverebrates	0.95 ± 0.01	1.39 ± 0.01	$< 2 \cdot 10^{-5}$	0.97 ± 0.01	1.49 ± 0.01	$< 2 \cdot 10^{-5}$
Mammals	0.94 ± 0.02	1.43 ± 0.04	$< 2 \cdot 10^{-5}$	0.94 ± 0.02	1.74 ± 0.09	$< 2 \cdot 10^{-5}$
Rodents	0.93 ± 0.01	1.47 ± 0.01	$< 2 \cdot 10^{-5}$	0.93 ± 0.01	2.33 ± 0.01	$< 2 \cdot 10^{-5}$
Plants	0.95 ± 0.01	1.21 ± 0.01	$< 2 \cdot 10^{-5}$	0.95 ± 0.01	1.36 ± 0.03	$< 2 \cdot 10^{-5}$

Table 2.11: Clumping measures for monomers and dimers.

The average clumping ratio values R_n and R_w along with the error bars are shown for monomers and dimers in coding and noncoding DNA of primate, vertebrate, invertebrate, mammal, rodent, and plant taxonomic partitions of the GenBank. The mean values and the error bars are computed as follows. First, the GenBank data sets are partitioned into 10 subsets of size of 10% of the GenBank data sets. R_n and R_w are computed for each subset independently. Then, from the distributions of the values of R_n and R_w for coding and noncoding DNA and p-values for the Kolmogorov - Smirnov test are computed. p-values are the probability that R_n and R_w values (for coding and for noncoding DNA) are drawn from the same distribution. If p is close to 1, then the two distributions under consideration are drawn from the same distribution with the probability 1. If p is close to 0, then these distribution are taken from two different distributions with the probability $(1-p) \to 1$. These results suggest: *(i)* the evolutionary pressure against clumping of repeats (except trimeric) in coding DNA; *(ii)* the clumping ratios for all organisms show strong clumping of the trimers; *(iii)* the difference between the clumping of trimers in coding versus noncoding DNA is less pronounced.

Observations *(i)* – *(v)* might arise from the evolutionary pressure against clumping of repeats (except trimeric) in coding DNA. These observations are in agreement with recent work [43] in which the dimeric tandem repeats (DTR) were studied and it was found that DTR are abundant in noncoding DNA, but rare in coding DNA. The mechanism of expansion based on the different mutational mechanisms was proposed [42, 97] and the difference in length distributions of DTR in coding and noncoding DNA was attributed to the fact that noncoding DNA is more tolerant to evolutionary mutational alterations than coding DNA.

For coding DNA, the observed clumping of trinucleotides could be due to specific protein structures in which amino acids cluster together (such as an alpha helix). Another possibility is that clumping of amino acids is alloted to the general problem of the stability of a native state of the folded proteins [93, 109–111]. However, if clumping was due to amino acids, amino acids should clump more — or at least to the same degree — as trinucleotides. Instead, amino acids clump much less than trinucleotides, suggesting that clumping is the result of some property of DNA and not of proteins. For example, the clumping ratio values for the amino acids for the primates $R_n = 1.30 \pm 0.02$ and $R_n = 1.46 \pm 0.03$ are lower than the clumping ratios for the trimers $R_n = 1.51 \pm 0.02$ and $R_n = 1.63 \pm 0.03$.

The strength of clumping of trimers in coding DNA relative to dimers and tetramers can be explained by the fact that insertion or deletion of a dimer or a tetramer would lead to a frame shift. Such a shift in the RF leads in most cases to an unfunctional protein, which can be lethal for the organism. On the contrary, the insertion or deletion of a trimer is equivalent to insertion or deletion of an amino acid in the protein sequence. Such insertion or deletion, if it happens away from the functionally or structurally important sites of the protein (see [112, 113]), would not affect the protein function, and hence would be tolerated by natural selection.

The source of clumping of oligomers in noncoding DNA could be the result of various duplication processes or simple repeat expansion processes [16, 42], indicating that some of the neighboring oligomers evolved from the same single copy.

2.7.3 Clumping of individual dimers

In addition, the clumping measures are calculated for each individual oligomer. The clumping measures in this case are defined the same way as in Eqs. (2.44) – (2.49)[4], i. e.

$$R_{n,i} \equiv \frac{\langle \ell \rangle_{n,i} - 1}{\langle \ell \rangle_{n,i}^{\rm th} - 1} = \frac{\sum_{\ell=1}^{\infty} \ell_i N_i(\ell_i) / \sum_{\ell=1}^{\infty} N_i(\ell_i) - 1}{1/(1-p_i) - 1}\,, \tag{2.50}$$

and

$$R_{w,i} \equiv \frac{\langle \ell \rangle_{w,i} - 1}{\langle \ell \rangle_{w,i}^{\rm th} - 1} = \frac{\sum_{\ell=1}^{\infty} \ell_i^2 N_i(\ell_i) / \sum_{\ell=1}^{\infty} \ell_i N_i(\ell_i) - 1}{(1+p_i)/(1-p_i) - 1}\,. \tag{2.51}$$

The theoretical values for $\langle \ell \rangle_{n,i}^{\rm th} = 1/(1-p_i)$ and $\langle \ell \rangle_{w,i}^{\rm th} = (1+p_i)/(1-p_i)$ (see [108]) are computed similarly to Eq. (2.47) and (2.48).

Interestingly, the clumping ratios for each individual oligomer are conserved for each organism, i. e. the standard deviation of the clumping measures is around a few per cent. To illustrate this observation the clumping ratio values for dimers in primates are presented in Table 2.13[5]. This observation indicates that the clumping ratio can characterize the unique property of the DNA sequences to clump and can be utilized in further studies of aggregates of oligomers in DNA sequences. For example, the different clumping ratios of various dimers can suggest different specific mutation rates for each dimer and organism.

2.8 Summary

The statistical patterns of nucleotides in DNA are complex. These patterns vary from organism to organism and from region to region. The statistical properties of DTR differ in coding and noncoding parts of DNA. The length distributions of DTR in coding DNA are exponential and, thus, are consistent with uncorrelated or short-range correlated amino acid sequences. On the other hand, the length distributions of dimeric repeats in introns and intergenic regions usually deviate strongly from an

[4] The summation over all types of oligomers ($i = 1, 2, ..., M$) is omitted in calculations of number and weight average values, and clumping ratios.

[5] The data for the clumping ratio values for other oligomers and taxonomic partitions of the GenBank are consistent with this statement.

exponential function and can, in some cases, be fit by a power-law function. It has been shown that the distribution of AC-CA, TG-GT repeats in vertebrates have plateaus in the range $10 < \ell < 30$ and that the distributions of CC, CG, GC, and GG repeats decay much faster than other repeats that include weakly bonded bp.

The power-law distribution of SSR can be explained if one assumes a random multiplicative process for the mutation of the repeat length, i. e., each mutation leads to a change of repeat length by a random factor with a certain distribution [42]. Such a process may take place due to errors in replication [27] or unequal chromosomal crossing over (see [42] and references therein).

SSR expansion in coding regions leads to a loss of protein functionality (as in Huntington's disease [27]) and to the extinction of the organism. Thus, SSR distribution in protein coding sequences remain exponential due to Darwinian evolutionary pressure.

Moreover, the conservation of protein coding sequences [114] and, thus, lack of expansion dynamics of SSR in the coding regions of DNA, are probably related to the problem of protein folding. Monte-Carlo simulations of protein folding on the cubic lattice suggest that the statistical properties of the sequences that fold into a native state resemble those of random sequences [63, 93], i. e., they do not have long amino acid repeats.

The higher tolerance of the noncoding DNA to various mutations, especially to mutations involving the change of DNA length — e. g., duplication, insertion of transposable elements, and SSR expansion — lead to the statistical features that are distinct from the coding DNA.

In addition, a quantitative method for analyzing repetitions of identical short oligomers has been developed to study the favorable aggregates of the oligomers in coding and noncoding DNA sequences. From the analysis of sequences presently available in the GenBank separately for primate, mammal, vertebrate, rodent, invertebrate and plant taxonomic partitions, it is found that some oligomers clump more than they would if randomly distributed, while others repel. Clumping measures are defined to quantify this degree of clumping. It is found that *(i)* clumping significantly differs in coding and noncoding DNA; *(ii)* in most cases, monomers, dimers and tetramers clump in noncoding DNA, while they repel from each other in coding DNA. *(iii)* The degree of clumping for different sources (primates, invertebrates, and plants) is more conserved among these sources in the case of coding DNA than in the case of noncoding DNA. *(iv)* In contrast to other oligomers, trimers always prefer to clump.

Trimers						
	R_n			R_w		
	coding	noncoding	p-value	coding	noncoding	p-value
Primates	1.51 ± 0.02	1.49 ± 0.01	0.31	1.63 ± 0.03	1.91 ± 0.02	$1.7 \cdot 10^{-4}$
Vertebrates	1.54 ± 0.05	1.40 ± 0.04	$1.7 \cdot 10^{-4}$	1.68 ± 0.08	1.56 ± 0.06	$3.1 \cdot 10^{-2}$
Inverebrates	1.49 ± 0.01	1.21 ± 0.01	$1.7 \cdot 10^{-4}$	1.56 ± 0.01	1.27 ± 0.01	$1.7 \cdot 10^{-4}$
Mammals	1.53 ± 0.04	1.51 ± 0.04	0.97	1.63 ± 0.06	1.87 ± 0.10	$1.7 \cdot 10^{-4}$
Rodents	1.42 ± 0.02	1.40 ± 0.01	$6.9 \cdot 10^{-3}$	1.52 ± 0.02	2.13 ± 0.07	$< 2 \cdot 10^{-5}$
Plants	1.42 ± 0.01	1.29 ± 0.02	$1.7 \cdot 10^{-4}$	1.50 ± 0.01	1.45 ± 0.02	$1.7 \cdot 10^{-4}$
Tetramers						
	R_n			R_w		
	coding	noncoding	p-value	coding	noncoding	p-value
Primates	0.85 ± 0.02	2.85 ± 0.03	$< 2 \cdot 10^{-5}$	0.86 ± 0.02	4.61 ± 0.07	$< 2 \cdot 10^{-5}$
Vertebrates	0.89 ± 0.04	2.57 ± 0.19	$< 2 \cdot 10^{-5}$	0.89 ± 0.04	5.71 ± 1.17	$< 2 \cdot 10^{-5}$
Inverebrates	0.83 ± 0.01	1.31 ± 0.01	$< 2 \cdot 10^{-5}$	0.96 ± 0.02	1.53 ± 0.02	$< 2 \cdot 10^{-5}$
Mammals	0.68 ± 0.03	2.96 ± 0.29	$< 2 \cdot 10^{-5}$	0.69 ± 0.03	3.84 ± 0.46	$< 2 \cdot 10^{-5}$
Rodents	0.79 ± 0.02	4.57 ± 0.06	$< 2 \cdot 10^{-5}$	0.80 ± 0.02	11.32 ± 0.23	$< 2 \cdot 10^{-5}$
Plants	0.91 ± 0.01	1.85 ± 0.03	$< 2 \cdot 10^{-5}$	0.92 ± 0.01	2.27 ± 0.15	$< 2 \cdot 10^{-5}$

Table 2.12: Clumping measures for trimers and tetramers.
Clumping measures for trimers and tetramers (see caption to Table 2.11).

Dimers: Primates				
dimer	R_n		R_w	
	coding	noncoding	coding	noncoding
AA	1.36 ± 0.04	2.07 ± 0.04	1.35 ± 0.04	2.92 ± 0.09
AT	0.77 ± 0.02	1.17 ± 0.01	0.78 ± 0.02	1.39 ± 0.02
AG	0.93 ± 0.01	1.13 ± 0.01	0.93 ± 0.01	1.18 ± 0.01
AC	1.05 ± 0.01	1.61 ± 0.02	1.05 ± 0.01	2.11 ± 0.05
TA	1.70 ± 0.03	1.48 ± 0.02	1.70 ± 0.03	1.80 ± 0.04
TT	1.50 ± 0.02	1.95 ± 0.03	1.50 ± 0.03	2.70 ± 0.07
TG	0.89 ± 0.01	1.04 ± 0.01	0.89 ± 0.02	1.37 ± 0.03
TC	1.31 ± 0.02	1.72 ± 0.01	1.29 ± 0.02	1.78 ± 0.02
GA	1.22 ± 0.02	1.68 ± 0.01	1.20 ± 0.02	1.73 ± 0.01
GT	1.60 ± 0.04	1.55 ± 0.01	1.58 ± 0.04	2.27 ± 0.07
GG	0.80 ± 0.01	1.18 ± 0.01	0.77 ± 0.01	1.15 ± 0.01
GC	0.46 ± 0.01	0.34 ± 0.01	0.47 ± 0.01	0.35 ± 0.01
CA	0.71 ± 0.02	1.05 ± 0.01	0.72 ± 0.02	1.34 ± 0.03
CT	0.81 ± 0.01	1.14 ± 0.01	0.80 ± 0.01	1.20 ± 0.02
CG	1.25 ± 0.03	2.05 ± 0.06	1.26 ± 0.03	2.11 ± 0.06
CC	0.95 ± 0.01	1.18 ± 0.01	0.91 ± 0.01	1.15 ± 0.01

Table 2.13: Clumping measures for 16 dimers in primate DNA.

The average clumping ratio values R_n and R_w along with the error bars are shown for 16 dimers in primate taxonomic partition of the GenBank for coding and noncoding DNA. Note that the standard deviation of the clumping measures is about a few per cent, indicating conservation of the clumping measures for each particular dimer within the same organism.

Chapter 3

The protein folding problem

3.1 Introduction

Proteins [11, 115, 116] are among the most important building blocks of life (Fig. 3.1). They are responsible for many functions in cell organization, reproduction, signal transduction, and cell death (apoptosis). Proteins carry out transport and storage in living cells. Proteins inhibit or catalize chemical reactions. The most intriguing fact about proteins is that their functions in living cells are determined by their three-dimensional structure.

Proteins are linear heteropolymers built by a combination of twenty amino acids (residues), joined by peptide bonds (see Fig. 3.2). The linear polymer formed by amino acid residues linked by peptide bonds, $H-(NH-C_{\alpha}HR_i-CO-)_nOH$ is also called a *polypeptide chain* (R_i denotes an amino acid). The number of amino acids in proteins ranges from 40 to 3000.

Only the primary structure (sequence of amino acids) of proteins is encoded by DNA, i. e. only the primary structure of proteins is genetically transfered by DNA. Thus, proteins must themselves find their unique native configuration from the primary structure. However, if one takes a small protein of 100 residues and assumes that each residue can be positioned, on average, in 6 different ways relative to its chain neighbors[1], then the number of possible three-dimensional conformations

[1] This number is surely an underestimation. It is chosen just to make a lower-bound estimation

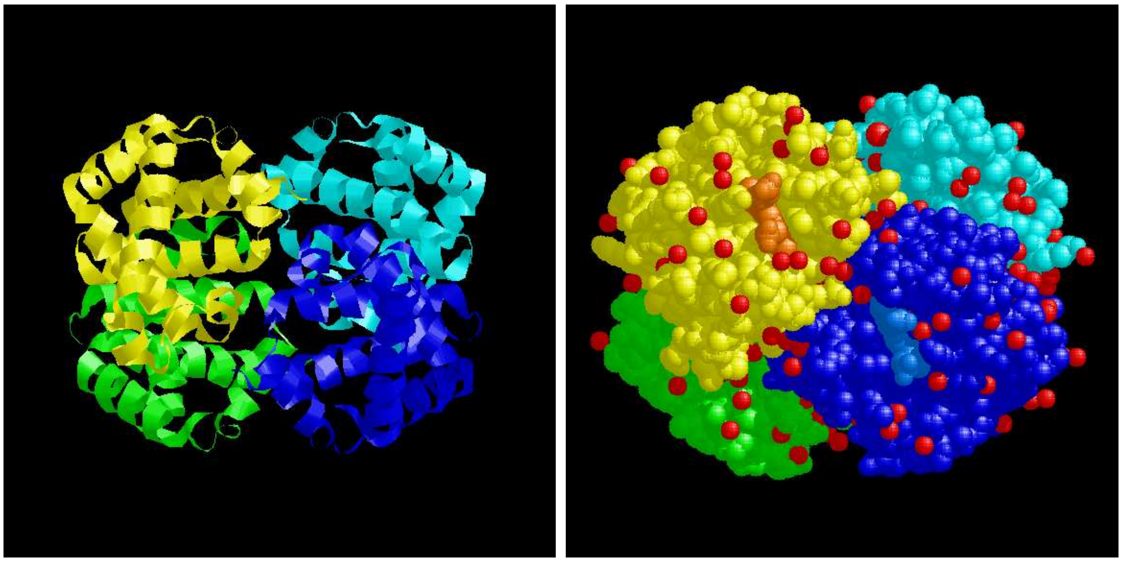

Figure 3.1: Hemoglobin (deoxy) molecule

Hemoglobin molecule (Brookhaven Protein Data Bank [117, 118] accession code $2HHB$) is responsible for oxygen transport. The left image is the ribbon representation; the right image is the spacefilling representation.

of such a protein will be $6^{100} \approx 6 \cdot 10^{77}$. The vibrational mode of proteins is of the order of picoseconds, so even if one takes 1 ps per conformation of a residue, the folding by random search will take $6 \cdot 10^{64}$ s, which is roughly $2 \cdot 10^{55}$ years. However, the folding time is of the order of miliseconds to seconds. This paradox was first described by Levinthal in 1968 [119]. Therefore, there should be some conformational "information" stored in a sequence of amino acids, which drives proteins to their native conformation. This information is a superposition of quantum mechanical properties of amino acids.

3.2 Protein folding problem

A simplified view of proteins is that they are molecules with specific three-dimensional structures of specific functionalities that are determined by the one-dimensional sequences of amino acids. There are two principal questions that arise from this view:

for the "random search" folding time.

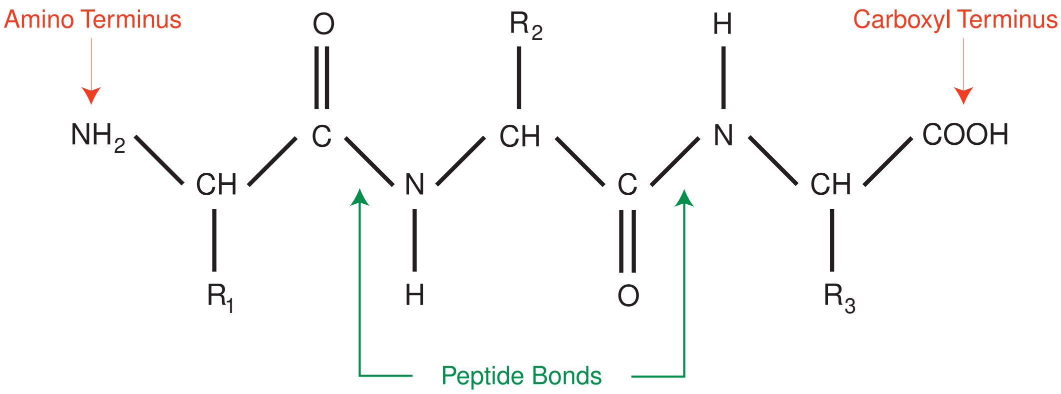

Figure 3.2: Schematic representation of the polypeptide chain.
R_1, R_2, etc. are the side groups attached to α-carbons (C_α) of the amino acids.

(i) The "direct" folding problem: How can one predict the three-dimensional conformation of a protein based on its sequence of amino acids? *(ii) The "inverse" or design folding problem:* How can one predict a sequence of amino acids based on the three dimensional structure of the protein?

3.2.1 Inverse (design) protein folding problem

There has been only limited success in experimental approaches to protein design, providing polypetides that could fold into compact but mostly disordered conformations [120, 121]. The bottleneck in the protein design problem is that the number of sequences grows exponentially with the sequence length of the protein. For example, there are $20^{100} \approx 10^{130}$ possible 100 amino acid sequences. The fraction of sequences that fold into protein-like structures is negligible. Thus, the probability that one will select at random a sequence that will fold into a protein-like structure is vanishingly low. Convincing success in the protein design problem may come from reliable theoretical approaches that make it possible to find a sequence that folds to a unique stable native structure.

The problem is particularly rich since peptide bonds cause "frustrations" [122] which prohibit monomers from completely optimizing the interaction energy between monomers in space. Thus, the plot of energy versus structural similarity of different conformations is very ragged, and has many relative (local) minima. However, proteins are able to fold to the *absolute minima*. These properties were first described analytically by Wolynes [123, 124], who related the Random Energy Model (REM) [5] to the protein folding problem by introducing a phenomenological Hamiltonian and deducing the nature of folding for random sequences.

Garel and Orland [125] examined the folding transition from a more conventional Hamiltonian within the replica approach. Later, Shakhnovich and Gutin [93, 126] derived the REM from a microscopic Hamiltonian. The REM was also studied numerically by enumeration of all conformations of short polymers on the lattice in order to calculate thermodynamic quantities exactly. REM became popular due to its simplicity and exact solvability. The existence of an exact solution enormously reduces computational effort required for the study of the folding process.

It has been pointed out [121] that *experimental protein design* is limited due to the poor corroboration of theory and experiment. Recent success of Dahiyat and Mayo [127] in the design of a small protein is based on the synergism of theory and experiment. This work demonstrates the importance of theoretical approaches in protein design. The limitation of this work comes from the protein sequence length. Their approach requires complete enumeration of all sequence candidates, which becomes exponentially difficult as sequence length increases. For example, Dahiyat and Mayo [127] used a library of sequences of the size of 1.9×10^{27} for the target 28 amino acid $\beta\beta\alpha$-motif, structurally similar to the second zinc finger module of the DNA binding protein Zif268. The success of Dahiyat and Mayo [127] calls for further refinements of the theoretical approaches to the design problem.

Because the number of possible sequences is enormous and the fraction of sequences which fold into a given structure [128] is negligibly small, the probability that one will find a sequence at random is also negligible. Even clever experimental approaches, such as *phage display* [129], which bias experimental sequence searches towards targets, suffer from technical limitations on the total number of sequences that can be surveyed.

Several design algorithms have been proposed (see e. g. [130, 131] and also review [121]). A Monte Carlo algorithm [132, 133] seems to be the most advantageous [121] among all other algorithms since it converges to the canonical distribution and its results can be understood from the statistical mechanics viewpoint [134–137]. Thus, the tools of statistical mechanics become relevant to the protein design.

3.2.2 Direct protein folding problem

The three-dimensional structure of globular proteins in cells is determined by their amino acid composition. Understanding the relevance of the interactions between amino acids to the folding process of a protein is a complex task that has been the subject of a number of theoretical and experimental studies in the past few decades [7, 112, 113, 115, 124, 138–148]. Many studies have been dedicated to identification of the dominant interactions between amino acids, i. e. forces that drive protein to the native state (NS) (see e. g. Refs. [143, 149–151]).

The transition of short proteins from the unfolded to the folded state is accompanied by a drastic reduction of the entropy (see Fig. 3.3). It was proposed that the folding transition for short proteins is analogous to the nucleation process at a first order transition [112, 113, 139, 140, 148]. In this scenario, there is competition between two minima of the free energy, the folded state with low energy and entropy and the unfolded state corresponding to high energy and entropy. Levinthal pointed out that the time scale for a random search in the space of protein configurations would be incompatible with biological folding times [119]. Specific scenarios, in which a fraction of the contacts play an important role in reducing the time scales for folding, have been proposed by different authors [112, 113, 139, 140] and confirmed experimentally [152–154]. Thus, the determination of this set of contacts is a step forward towards the solution of the direct protein folding problem.

3.2.3 Why it is hard to solve the protein folding problem

The vast dimensionality of the protein conformational space [119] makes the folding time too long to be reachable by direct computational approaches [115, 139, 140]. While the folding of the real protein occurs on the time scale of 1ms - 1s, the traditional molecular dynamics (MD) algorithms are able to resolve only the nanoscale

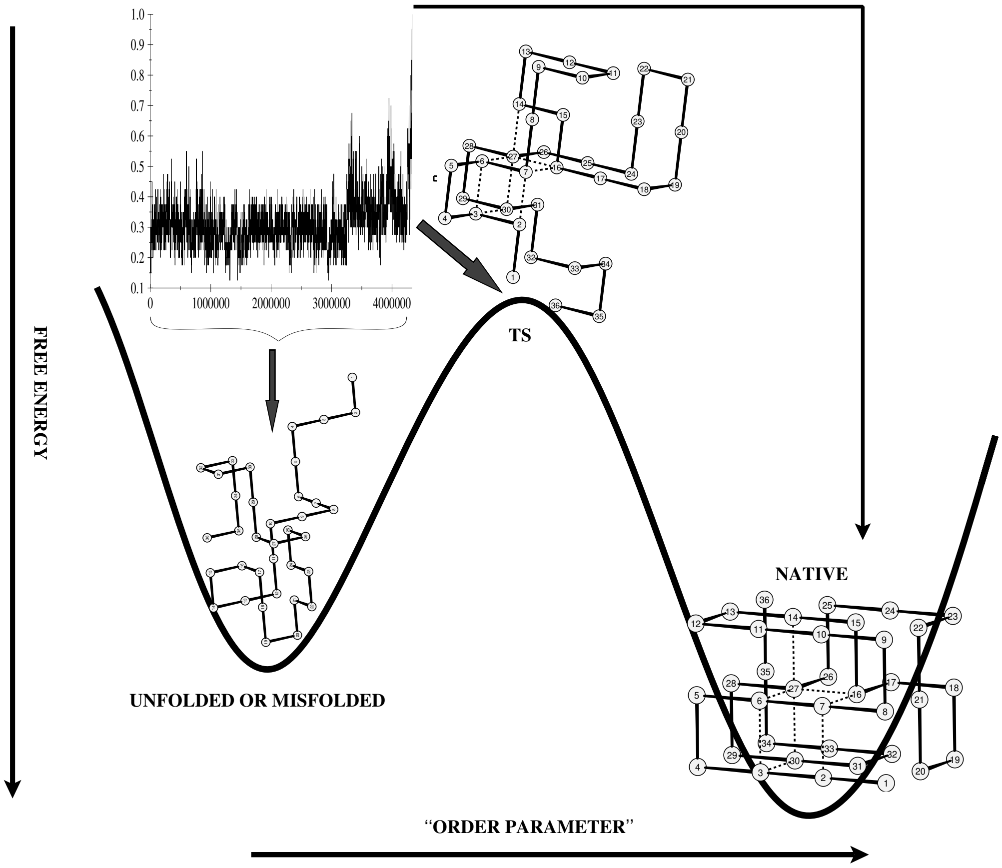

Figure 3.3: Schematic representation of the protein free energy landscape. Two wells of the free energy correspond to the folded (native state) and unfolded (or misfolded) states. The barrier separating folded and unfolded states is called transition state. Courtesy of [140].

time region. More realistic MD, based on the quantum mechanical calculations of the states of all atoms, limit the study only to polypeptides of five amino acids in length. However, the stability condition requires that the chain contains at least 20 amino acids, since the stabilization energy is of the order of 0.1 k_BT per amino acid [150], where k_BT is the average energy of the environment.

Simplified models [7, 8, 112, 124, 134, 141, 142, 150, 155, 156] became popular due to their ability to reach reasonable time scales and to reproduce the basic thermodynamic and kinetic properties of real proteins [11, 115, 157]: *(i)* unique native state,

i. e. there should exist a single conformation with the lowest potential energy; *(ii)* cooperative folding transition (resembling first order transition); *(iii)* thermodynamical stability of the native state; *(iv)* kinetic accessibility, i. e. the native state should be reachable in a biologically reasonable time [144, 156].

Monte Carlo (MC) simulations on lattices (see, e. g., [7, 8, 140, 141] and references therein) appear to be useful for studying theoretical aspects of protein folding. The Monte Carlo algorithm is based on a set of rules for the transition from one conformation to another. These transitions are weighted by some transition matrix, which reflects the phenomena under study. The simplicity of the algorithm and the significantly small conformational space of the protein models (due to the lattice constraints) make MC on-lattice simulations a powerful tool for studying the equilibrium dynamics of the protein models.

However, lattice models impose strong constraints on the angles between the covalent bonds, thereby greatly restricting the conformational space of the protein-like model. The additional drawback of this restriction lies in the poor capability of these models to discern the geometrical properties of the proteins.

Time in MC algorithms is estimated as the average number of moves (over an ensemble of the folding $\rightleftharpoons$ unfolding transitions) made by a model protein. It was pointed out [158] that MC simulations are equivalent to the solution of the master equation for the dynamics, so there is a relationship between physical time and computer time, which is counted as the number of MC steps. However, a number of delicate issues — such as the dependence of the dynamics on the set of allowed MC moves — remain outstanding, so an independent test of the dynamics using the MD approach is needed.

To address the questions with sensitivity to geometrical details, it is useful to study off-lattice models of protein folding. Thus far, several off-lattice simulations have been performed [159–161], which demonstrate the ability of the simplified models to study protein folding.

3.2.4 A rapid tool to perform off-lattice simulations

Recently, a new approach for simulations of model proteins, discrete molecular dynamics (DMD), has been implemented [113, 148, 162, 163] to study the dynamics of homo- and heteropolymers (proteins). This approach permits the rapid testing of the folding properties of proteins within reasonable processor time. This MD algorithm has proven to be a powerful tool to study the thermodynamics and kinetics of the folding $\rightleftharpoons$ unfolding transition of simplified models of proteins [7, 8].

The potential of interaction between pairs of residues is modeled by a "square-well", which allows an increase in the speed of the simulations [113, 148, 162, 163]. This approach is based on the tracing of collision events in the simulated system, such that it is not necessary to follow the dynamics of the system during times when there are no events (i.e., collisions). This is a crucial difference between traditional MD and DMD. In addition to increasing the speed of the simulation, the DMD algorithm allows for the tracking of "realistic" (not discretized) time. It has been shown that the DMD algorithm is able to resolve in time the collapse transition for a wide range of temperatures.

In addition, such an algorithm can be a useful compromise between computationally heavy traditional MD and fast, but restrictive MC. It has been demonstrated in [113] that a model protein reproduces the principal features of folding phenomena *(i)* – *(iv)* described in Section 3.2.3. Detailed description of the DMD algorithm is presented in Section 3.3.2.

3.2.5 A path to solution of the protein folding problem

There are two questions that might shed light on the protein folding problem and are the subject of this dissertation.

1. What are the thermodynamic properties of proteins and what are the structural properties of the protein near the folding transition? For example, it has been shown [113] that more than 50% of all of the inter-residue contacts are present near the folding transition temperature T_f. This set of contacts is called the *core* of the protein.

Definition 3.1 The *core* is a small subset of the residues, which maintains the backbone of the structure at temperatures close to the folding transition temperature.

Identification of the protein core is a challenging task, both in real experiments and computer simulations, and is the subject of the Sec. 3.5.

2. What are the kinetic properties of proteins at the folding transition? This question is specifically important since it is directly connected to the mechanism of protein folding. Several scenarios of protein folding have been proposed [121, 139, 148, 155, 164–168]. Simulations of the off-lattice model 46-mer [148] reveal the nucleation scenario that is presented in Sec. 3.8.

   Definition 3.2 The *nucleus* is the structure (few contacts $< 1\%$) that appears in the transition state and results in rapid folding. If the nucleus is disrupted in the native state, then the structure is unfolded.

Remark 3.1 *The difference between the core and the nucleus of a protein is crucial: while the* core *is a persistent part of the structure at equilibrium, the* nucleus *is a fragment of the protein structure, which is assembled in the transition state (TS) — the folding $\rightleftharpoons$ unfolding barrier (see Fig. 3.3).*

In addition, the nucleus is defined kinetically, and has "physical" meaning – like bubbles that appear in liquid during boiling. On the contrary, the core is defined thermodynamically, and refers to the geometrical properties of the protein which persist near the folding transition temperature.

3.3 Molecular dynamics studies of the folding of protein-like model

3.3.1 The model of inter-residue interactions

The "beads on a string" representation is used to model proteins. The residues are hard spheres of unit mass. The potential of interaction between residues is "square-well". The inter-residue interaction follows the Gō model [7, 8, 141], where an attractive potential between residues is assigned to the pairs that are in contact (Δ_{ij}, defined below) in the native state and a repulsive potential is assigned to the pairs that are not in contact in the native state. Thus, the potential energy

$$\mathcal{E} = \frac{1}{2}\sum_{i,j=1}^{N} U_{i,j} \tag{3.1}$$

where i and j denote residues i and j. $U_{i,j}$ is the matrix of pair interactions

$$U_{i,j} = \begin{cases} +\infty, & |r_i - r_j| \leq a_0 \\ -\mathrm{sign}(\Delta_{ij})\epsilon, & a_0 < |r_i - r_j| \leq a_1 \\ 0, & |r_i - r_j| > a_1 \,. \end{cases} \tag{3.2}$$

Here, $a_0/2$ is the radius of the hard sphere, and $a_1/2$ is the radius of the attractive sphere (Fig. 3.4a) and ϵ sets the energy scale. $||\Delta||$ is a matrix of contacts with elements

$$\Delta_{ij} \equiv \begin{cases} 1, & |r_i^{NS} - r_j^{NS}| \leq a_1 \\ -1, & |r_i^{NS} - r_j^{NS}| > a_1 \,, \end{cases} \tag{3.3}$$

where r_i^{NS} is the position of the i^{th} residue when the protein is in the native conformation. Note that the non-native contacts are penalized by imposing $\Delta_{ij} < 0$. The parameters are chosen as follows: $\epsilon = 1$, $a_0 = 9.8$ and $a_1 = 19.5$. The covalent bonds are also modeled by a square-well potential (Bellemans' bonds):

$$V_{i,i+1} = \begin{cases} 0, & b_0 < |r_i - r_{i+1}| < b_1 \\ +\infty, & |r_i - r_{i+1}| \leq b_0, \text{ or } |r_i - r_{i+1}| \geq b_1 \,. \end{cases} \tag{3.4}$$

The values of $b_0 = 9.9$ and $b_1 = 10.1$ are chosen so that the average covalent bond length is equal to 10 (See Fig. 3.4b). The original configuration of the protein ($N = 65$ residues) was designed by collapse of a homopolymer at low temperature [135, 160, 169]. It contains $n^* = 328$ native contacts, so $\mathcal{E}_{NS} = -328$. The 65 by 65 matrix of contacts of the globule in the native state is shown in Fig. 3.5a. Note that the large number of native contacts ($328/65 \approx 5$ contacts per residue) is due to the choice of the parameter: $a_1 \approx 2a_0$ — so that residues are able to establish contacts with the residues in the second neighboring shell.

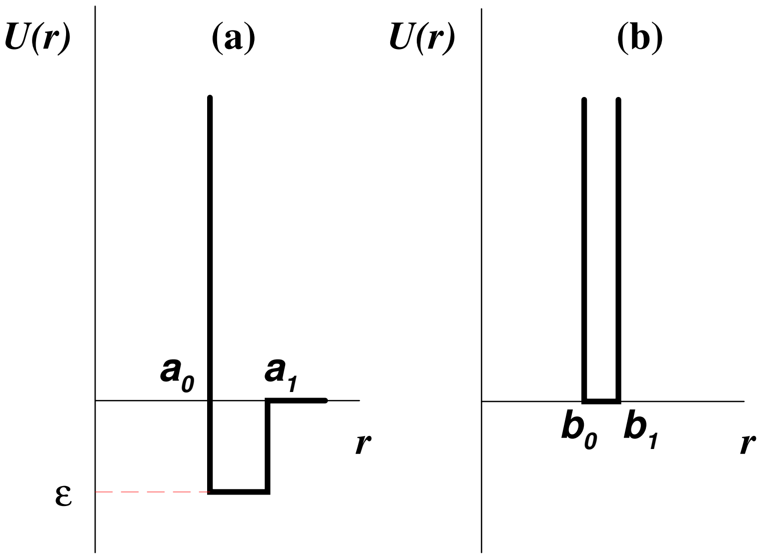

Figure 3.4: The potential of interaction of model proteins.
The potential of interaction between (a) specific residues; (b) neighboring residues (covalent bond). a_0 is the diameter of the hard sphere and a_1 is the diameter of the attractive sphere. $[b_0, b_1]$ is the interval where residues that are neighbors on the chain can move freely.

Definition 3.3 *The* radius of gyration R_G *of a polymer which consists of N monomers is defined as follows:*

$$R_G^2 = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\langle(\vec{r}_i - \vec{r}_j)^2\rangle \,, \tag{3.5}$$

where $\vec{r}_i$ and $\vec{r}_j$ are the position of the i^{th} and j^{th} monomers.

The radius of gyration of the globule in the native state is $R_G \approx 22.7$. The snapshot of the globule in the native state is shown in Fig. 3.5b.

3.3.2 DMD algorithm

The DMD program employs the discrete MD algorithm, which is based on the collision list, and is similar to one recently used by Zhou et al [162] to study the equilibrium thermodynamics of homopolymers; by Zhou and Karplus [163] to study the equilibrium thermodynamics of folding of model of *Staphylococcus aureus* protein A;

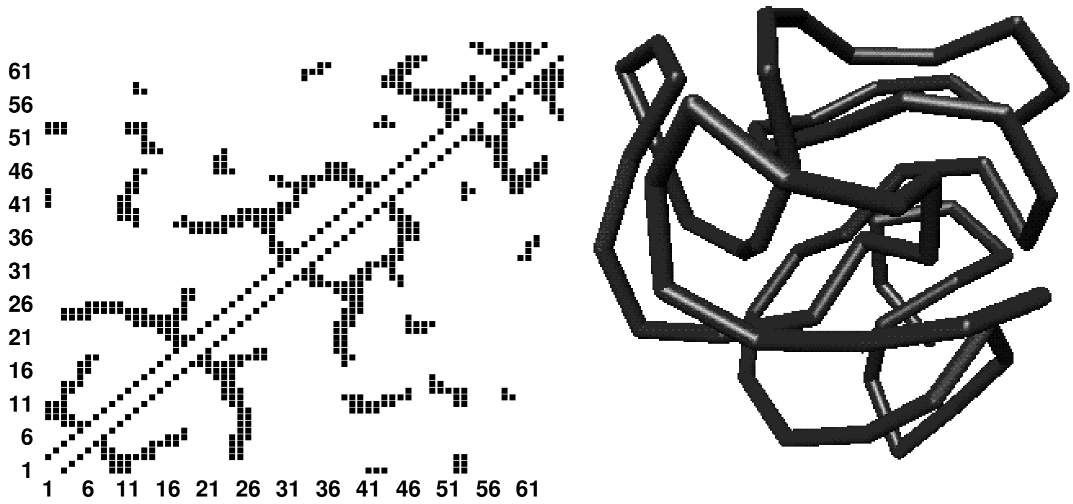

Figure 3.5: Contact map and snapshot of 65-mer.
(a) 65 x 65 contact matrix of the model protein in the native state. Black boxes indicate the matrix elements of those residue pairs which have a contact (their relative distance is between a_0 and a_1). (b) The snapshot of the protein of 65 residues in the native state obtained at temperature $T = 0.1$.

and by [113, 148]. A detailed description of the algorithm can be found in [132, 170–172]. To control the temperature of the protein, 935 particles are introduced. These particles do not interact with the protein or with each other in any way but via regular collisions, serving as a heat bath. Thus, by changing the kinetic energy of those "ghost" particles it is possible to control the temperature of the environment. The "ghost" particles are hard spheres of the same radii as the chain residues and have unit mass. Temperature is measured in units of ϵ/k_B. The time unit (tu) is estimated from the shortest time between two consequent collisions in the system between any two particles.

3.4 Thermodynamics of protein folding

In order to study the thermodynamics, MD simulations of the chain are performed at various temperatures. The globule is initially in the native state at temperature

$T = 0.1$ and the temperature of the heat bath is then raised to the desired one. The system is then allowed to equilibrate. At the final temperature the protein is relaxed for 10^6 time units. The typical behavior of the energy $\mathcal{E}$ and the radius of gyration R_G as functions of time is shown in Fig. 3.6 for three different temperatures.

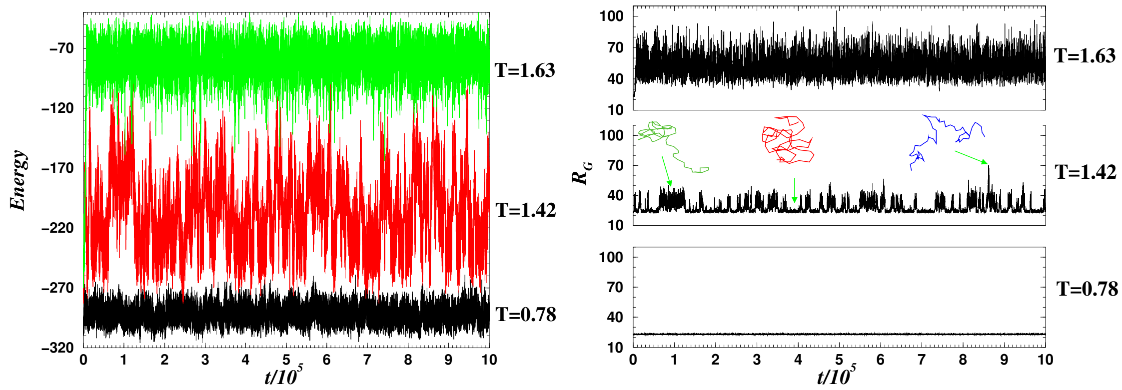

Figure 3.6: Energy and radius of gyration as a function of time.
The dependence on time of (a) energy $\mathcal{E}$ and (b) radius of gyration R_G. The globule is maintained at three different temperatures $T = 0.78 < T_f$, $T = 1.42$, and $T = 1.63 > T_f$ for 10^6 tu. For $T = 0.78$, the fluctuations of both energy $\mathcal{E}$ and R_G are small, i.e. the globule is found in one folded configuration. At high temperatures ($T = 1.63$) the fluctuations of $\mathcal{E}$ and R_G are large; the globule is mostly found in the unfolded state. At temperature $T = 1.42$, which is close to T_f, the globule is mostly present in two states. The lower energy configuration corresponds to the folded state: the globule is compact (see (b)). The other configuration has large fluctuations: the globule is in the PFS. There is an additional state – the unfolded state (see (b)). At $T = 1.42$ the protein model is rarely present in the unfolded state. The behavior of the globule at the temperatures close to T_f indicates the presence of three distinct states: folded, unfolded and PFS.

In the present model the non-native contacts (NNC) are penalized (i.e., the pairwise interaction between NNC is repulsive[2]), so their number increases as the temperature increases. At high temperatures (above T_θ), however, the number of NNC

[2]This corresponds to $g = 2$ of Ref. [163].

varies only due to the random motion of the ideal chain and, thus, on average their number should be constant at different temperatures. The maximal number of NNC occurs at T_θ and does not exceed 35, which is roughly 10% of the total number of native contacts (NC).

Remark 3.2 *The* Θ*-temperature is determined below (Fig. 3.10b) by the peak in the specific heat dependence on temperature. A rigorous definition of the* Θ*-temperature can be found in [173, 174].*

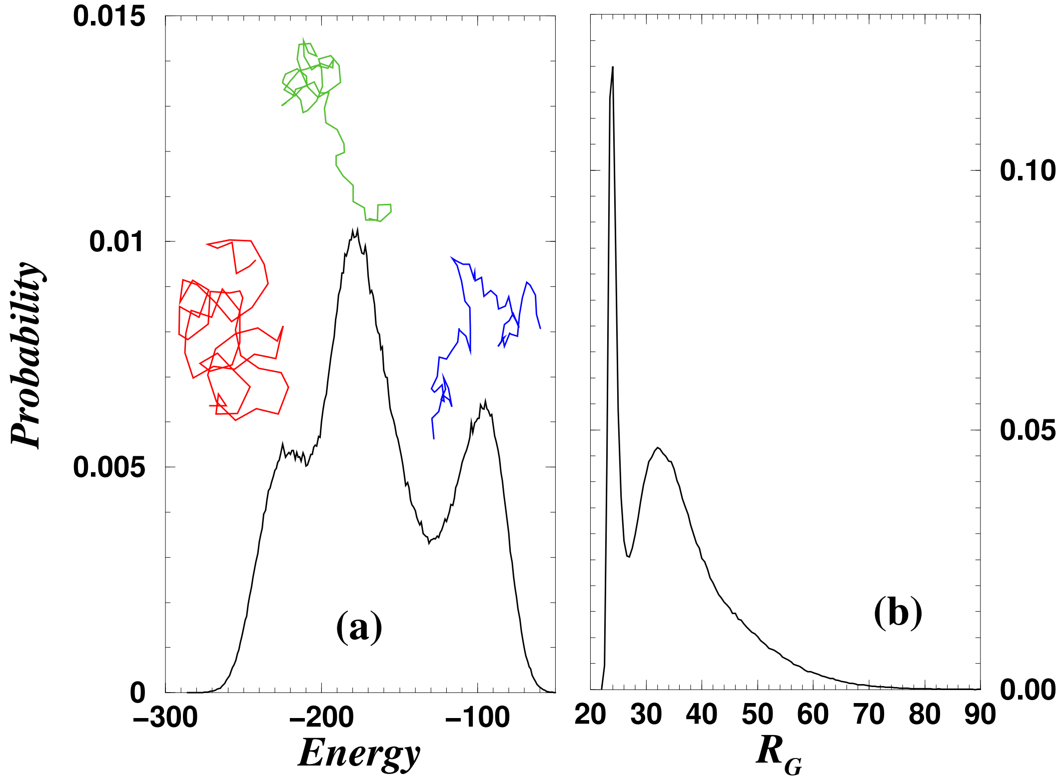

Figure 3.7: Probability distributions of the energy states and radii of gyration. The probability distribution of (a) the energy states $\mathcal{E}$ and (b) the radius of gyration R_G of the globule maintained at $T_f = 1.46$ for 10^6 tu. The trimodal distributions indicate the presence of three states: the folded state, PFS, and the unfolded state.

3.4.1 Partially folded states

The simulations reveal that the protein undergoes a folding $\rightleftharpoons$ unfolding transition as the temperature is increased to the proximity of the Θ-temperature T_θ, which in this model is $T_\theta \equiv T_f \approx 1.46$. At T_θ the distribution of energy has three peaks (Fig. 3.7a). The left peak corresponds to the folded state, the right peak corresponds to the unfolded state, and the middle one corresponds to the partially folded state (PFS), with a 19-residue unfolded tail. This trimodality of the energy distribution is also seen in Fig. 3.6b. The energy profile at temperature $T = 1.42$ (close to T_θ) also reflects these three states. Since $T < T_\theta$, only two states are mostly present on Fig. 3.6b. Thus, the energy distribution has only two peaks (Fig. 3.8), corresponding to the folded state and the PFS. Above T_θ, the globule starts to explore energetic wells other than the native well (see Fig. 13 in [111]).

To show that PFS is the cause of the middle peak in energy distribution (Fig. 3.7a), the 19-residue tail was eliminated and the energy distribution was plotted for the 46-mer at its Θ-temperature $T_\theta^* = 1.44$ (Fig. 3.9). One would expect to see only two states: folded and unfolded, since the 19-residue tail, which is the cause of PFS, is eliminated. Fig. 3.9 confirms these expectations.

3.4.2 Folding transition

The folding $\rightleftharpoons$ unfolding transition is further quantified in Fig. 3.10. The energy and the radius of gyration increase most rapidly near $T_f = T_\theta$ resembling the order parameter jump in a phase transition (see discussion below). This rapid increase of $\mathcal{E}$ and R_g reaches its maximum at Θ-point, where the potential of interaction is compensated by the thermal motion of the particles. Above T_θ, interactions between residues no longer hold them together any more and the chain becomes unfolded (see Fig. 3.11a). Note, that since all the attractive interactions are specific, the transition is described by one temperature T_f.

The presence of the PFS is observed in the temperature range between 1.40 and 1.48, in which the collapse transition occurs. Thus, in this particular region the folding temperature and the θ-temperature are indistinguishable within the accuracy

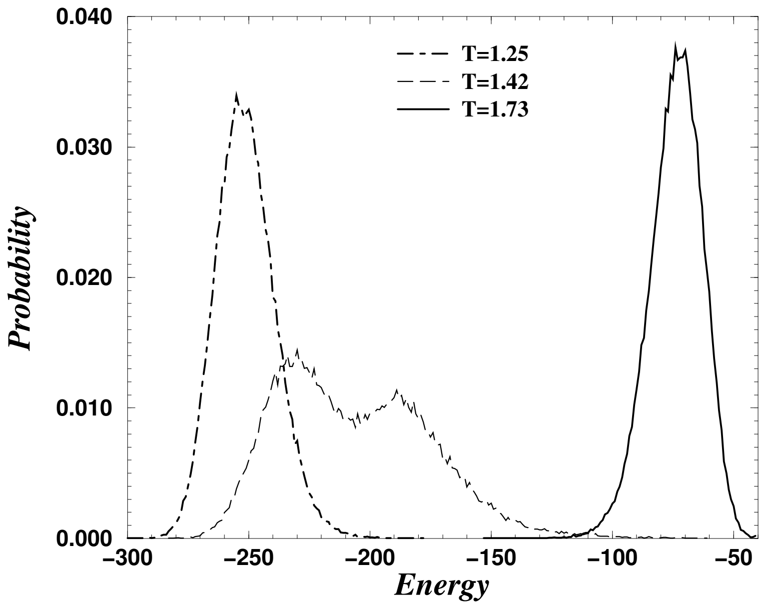

Figure 3.8: Probability distribution of energy states of the 65-mer.
The probability distribution of the energy states $\mathcal{E}$ of the globule maintained at three different temperatures: $T = 1.25$, 1.42, and 1.73. Note, that at $T = 1.42 \approx T_f$ the distribution has two expressed peaks. The right peak of this ($T = 1.42$) distribution corresponds to the PFS, while the left peak corresponds to the energetic well of the native state.

of their definitions.

Remarkably, a simple Flory type model of an excluded volume chain predicts T_θ within 20%. To demonstrate this, let us write the probability that the end-to-end distance of the chain is R [174]:

$$P(R) \propto p(R) \exp(-\frac{N^2 v}{2R^3} - \frac{\mathcal{E}(R)}{T}) \,, \tag{3.6}$$

where $v = (4\pi/3)(a_0/2)^3$ is the volume of the monomer and

$$p(R) \propto R^2 \exp(-3R^2/(2N(a_0/2)^2)) \tag{3.7}$$

is the probability that the end-to-end distance of the chain is R for the random walk model. For $T = T_f$, the repulsive excluded volume term $-(N^2 v)/(2R^3)$ balances the

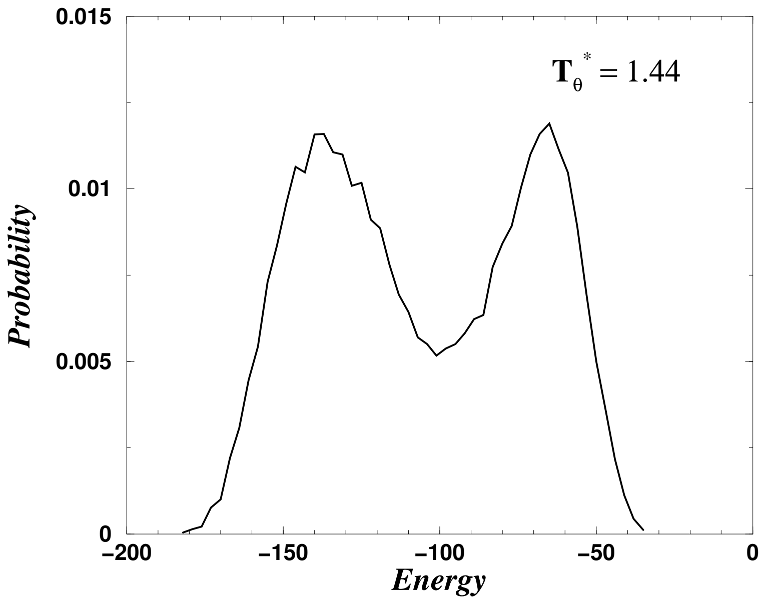

Figure 3.9: Probability distribution of energy states of the 46-mer.
The probability distribution of the energy states $\mathcal{E}$ of the 46-residue globule maintained at $T_f^* = 1.44$ during 10^6 tu. The bimodal distributions of energy indicates that the 19-residue tail is responsible for the PFS of the 65-residue globule: after eliminating the 19 residue tail the trimodal energy distribution of the 65-residue globule becomes the bimodal energy distribution of the 46-residue globule.

attractive term $-\mathcal{E}(R)/T > 0$. Thus,

$$T_f = \frac{2R^3|\mathcal{E}|}{N^2 v} \approx 1.7\,, \tag{3.8}$$

where $E \simeq -130$ and $R \simeq 24$ are taken for a certain configuration at the Θ-point.

The specific heat C_V is computed from the relation [175]:

$$C_V = \frac{\langle(\delta\mathcal{E})^2\rangle}{T^2}\,, \tag{3.9}$$

where $\langle(\delta\mathcal{E})^2\rangle \equiv \langle\mathcal{E}^2\rangle - \langle\mathcal{E}\rangle^2$ and $\langle ... \rangle$ denotes a time average. The time average is computed over 10^6 tu of equilibration at a fixed temperature. The dependence of the specific heat on temperature is shown in Fig. 3.10b. There is a pronounced peak of $C_V(T)$ for $T = T_f$.

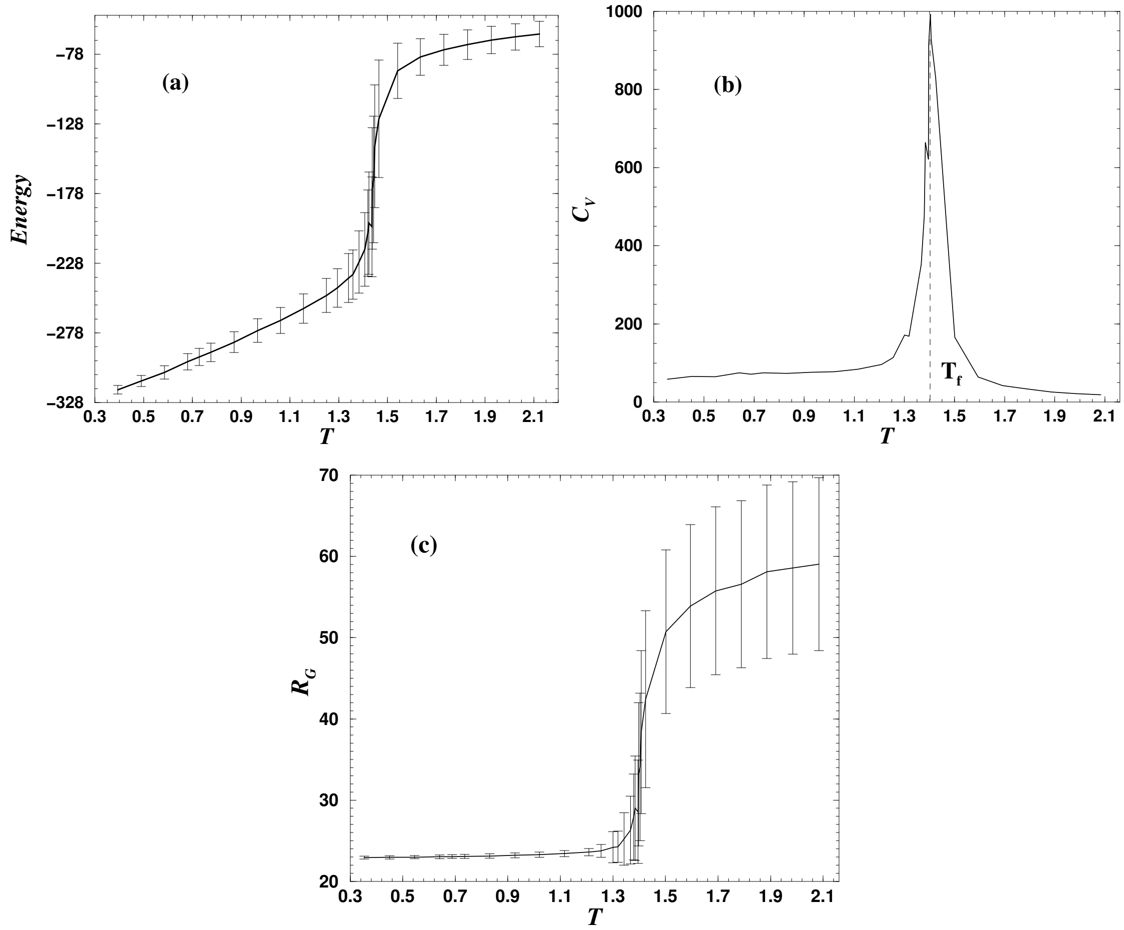

Figure 3.10: Energy, specific heat, and radius of gyration as a function of time. The dependence on temperature of (a) the energy $\mathcal{E}$, (b) the specific heat C_V, and (c) the radius of gyration R_G. The error bars are the standard deviation of fluctuations. The rapid increase of energy as well as the sharp peak in specific heat at $T = T_f$ indicates a first order phase transition.

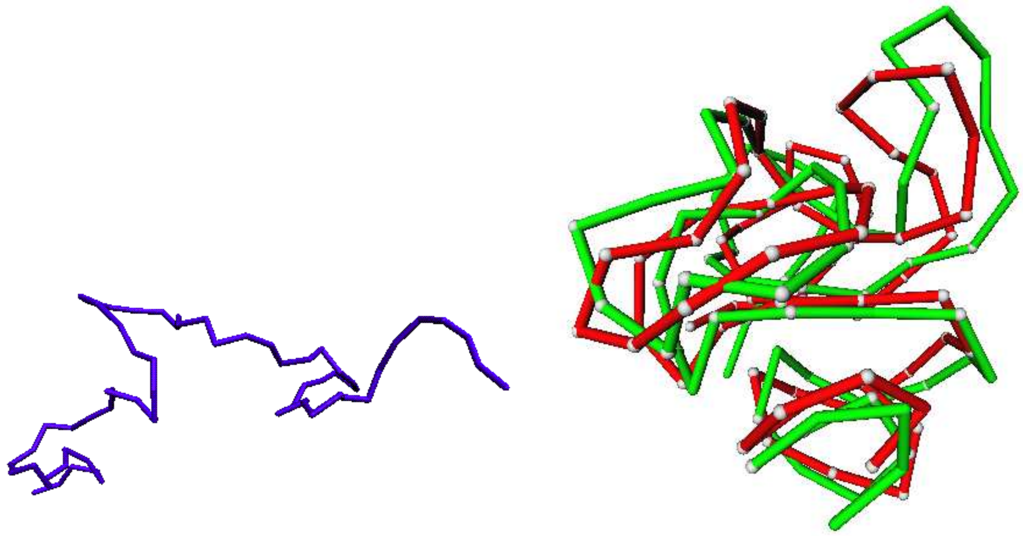

Figure 3.11: Snapshots of the 65-mer at the unfolded and folded states. The snapshot of the protein in (a) the unfolded stated, obtained at high temperature $T = 1.8$; and (b) the transition state, obtained at folding transition temperature $T_f = 1.46$ (light grey), overlapped with the globule at low temperature $T = 0.4$ (dark grey). Note that the TS globule has a close visual similarity to those maintained at low temperature and in the native state (see also Fig. 3.5b). It is more dispersed, however, which makes all the NC easily breakable. To compare the globule at the TS with the one maintained at temperature $T = 0.4$, the transformation proposed by Kabsch [176] is performed to minimize the relative distance between the residues in the TS and the state at $T = 0.4$. The "cold" residues (light grey spheres) denote residues whose rms displacement are smaller than a_1.

3.5 The core

3.5.1 Observation

It is noted that below the folding temperature T_f the globule (Fig. 3.11b) spends time in a state structurally similar to the native state (Fig. 3.5b). However, one can see that even though the globule maintains approximately the same structure, i.e. the same set of NC, the distances between residues are much larger than in the native state. Due to the fact that the potential of interaction between like residues is a square-well, there is no penalty for these residues to be maximally separated, yet they remain within the range of attractive interaction. This allows the globule to have more NNC and, thus, still maintain its similar-to-native structure, yet to have energy larger than the energy of the native state. This structure can be identified as the highest in energy that still maintains its core. As temperature increases, the ratio $|R_G - R_G^{NS}|/R_G^{NS}$ increases until temperature reaches $T = T_f$, where the ratio becomes roughly 0.87.

To confirm the presence of the core, $f \equiv N_{NC}/N_C$ is calculated at temperatures below $T = T_f$. The attractive inter-residue interaction term $-\mathcal{E}/T$ dominates the excluded volume repulsion term $-N^2 v/(2R^3)$ (see Eq. (3.6)), so

$$-\frac{\mathcal{E}}{T_f} - \frac{N^2 v}{2R^3} > 0 \,. \tag{3.10}$$

The total energy $\mathcal{E}$ has contributions from both NC and NNC contacts, so

$$\mathcal{E} = -\epsilon(N_{NC} - N_{NNC}) = -[2f - 1]\epsilon N_C \,. \tag{3.11}$$

At a temperature slightly below T_f, $|T - T_f|/T_f \approx 0.3$, the residues are maximally separated within their potential wells, yet they still maintain contacts. Therefore, the volume $\tilde{v}$ spanned by one residue is roughly $\tilde{v} \approx (4\pi/3)(a_1/2)^3 = 8v$. N_C is the product of the probability $\tilde{v}/R^3$ of having a bond (NC or NNC) and the total number of possible arrangements of the pair contacts between N residues, $N(N-1)/2 \approx N^2/2$. Thus,

$$N_C = \frac{N^2}{2R^3}\tilde{v} \,. \tag{3.12}$$

From Eqs. (3.10) – (3.12) we can estimate f, the fraction of N_C at the temperature $T \approx 1.42 < T_f$:

$$f > \frac{1}{2} + \frac{v}{\tilde{v}}\frac{T}{\epsilon} \approx 0.68\,. \tag{3.13}$$

Due to the fact that the globule maintains roughly the same volume at temperatures slightly below Θ-point, Eq. (3.13) implies that approximately 70% of all native contacts stay intact in the folded phase (see Fig. 3.8). This result is supported by simulations: at $T \approx 1.42$ the number of NNC is roughly $N_{NNC} \approx 28$, and the energy $\mathcal{E}$ is $\mathcal{E} = -206$. Therefore, the number of NC is $N_{NC} \approx 234$, and the fraction of NC is $f \approx 0.89$, which is even higher than the lower limit set by Eq. (3.13). Note that at a temperature higher than T_f, the fraction of native contacts becomes small due to the fact that in this regime the interactions are dominated by the excluded volume repulsion. This change in the number of NC from 70% to close to zero indicates the presence of the core structure maintained by this 70% of NC (see Fig. 3.11b and discussion below). Above the Θ-point the globule is completely unfolded (Fig. 3.11a).

The formation of a specific nucleus during the folding transition was suggested by many theoretical [112, 139, 140, 155, 164, 166, 177, 178] and experimental works [138, 153, 179–181]. The presence of the core at T_f may arise from a nucleation process driving the system from the unfolded state to the native state. Indication of a first order transition is found. Theoretical reasoning is also offered for the presence of a core (Eq. (3.13)) that might indicate the presence of a nucleus. Next, the core is identified.

3.5.2 Identification

The mean square displacement $\sigma(T)$ of the globule from its native conformation is calculated for various temperatures:

$$\sigma(T) \equiv \langle [\frac{1}{N}\sum_{i=1}^{N}(\vec{r}_i^{\,NS} - \overset{\leftrightarrow}{\mathcal{R}}\,(\overset{\leftrightarrow}{\mathcal{T}}\,\vec{r}_i))^2]^{1/2}\rangle = \langle [\frac{1}{N}\sum_{i=1}^{N}\sigma_i^2(T)]^{1/2}\rangle\,, \tag{3.14}$$

where $\vec{r}_i$ and $\vec{r}_i^{\,NS}$ are the coordinates of the residues of the globules at two conformations: some conformation at the temperature T, and the native conformation, respectively. $\overleftrightarrow{\mathcal{T}}$ is a translation matrix, which sets the centers of mass of these configurations to the same point in space. $\overleftrightarrow{\mathcal{R}}$ is a rotation matrix that minimizes the relative distance between the residues of two configurations (for details see [176, 182–184]). The $\sigma_i(T)$ in Eq. (3.14) are the rms displacements for each individual residue.

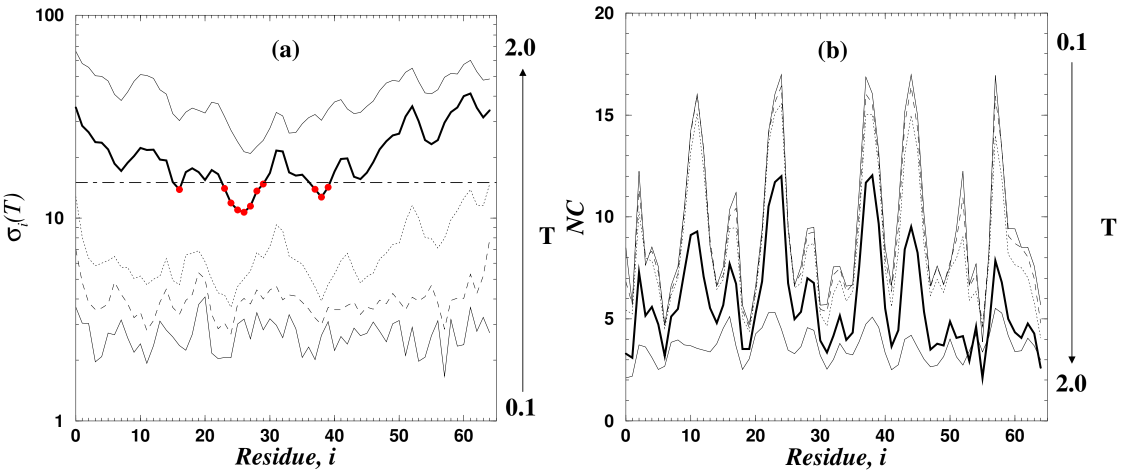

Figure 3.12: RMS displacements and NC for each residue as a function of temperature.
(a) The contour plot of the rms displacement $\sigma_i(T)$ for each residue $i = 0, 1, ..., 64$ at temperatures $T = 0.3$, 0.97, 1.34, 1.46 (bold line) and 1.54, averaged over 10^6 tu. Note that there is a distinct difference between the "cold" (small values of $\sigma_i(T)$) and "hot" residues (large values of $\sigma_i(T)$). The dashed dotted line indicates the breaking point of the NC, i.e. when $\sigma_i(T)$ is of the size of the average relative position between pairs of residues, i.e. $\sigma_i(T) = (a_0 + a_1)/2 \approx 15$. The bold lines (on both (a) and (b)) indicate the folding transition temperature line T_f. It is worth noting that 11 residues are still in contact (marked by circles on (a)): 16, 23, 24, 25, 26, 27, 28, 29, 37, 38, 39. (b) The analogous to (a) plot of the average number of NC for each residue. Note that the number of NC is strongly correlated with the rms: the local minima of the $\langle\sigma_i(T)\rangle$ plots correspond to the local maxima of the number of NC.

The plot of $\langle\sigma_i(T)\rangle$ is presented in Fig. 3.12a. From the roughness of the "landscape" in Fig. 3.12a, one can select a group of residues whose rms displacements are significantly smaller than the rms displacements of the other group of residues. The former group is denoted as "cold" residues and the latter group as "hot" residues. The rms displacement strongly depends on the temperature near the folding transition and grows slowly below T_f. Note that the average number of NC of the residues is correlated with the average rms displacement of these residues, i.e. the peaks on the $N_{NC,i}$ isothermal lines of Fig. 3.12b correspond to the "cold" residues.

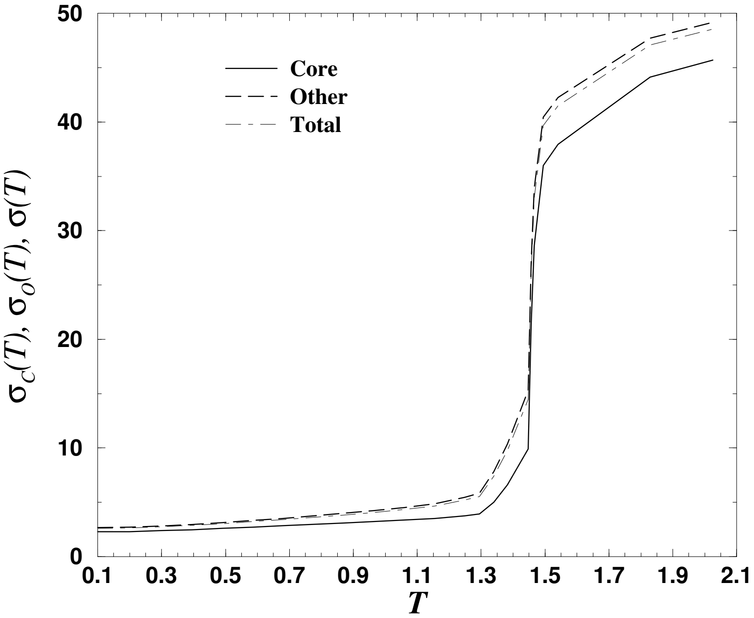

Figure 3.13: RMS displacement of the globule as a function of temperature. The dependence of the rms displacement of the core residues $\sigma_C(T)$ (solid line), the rest of the residues $\sigma_O(T)$ (dashed line) and all the residues $\sigma(T)$ on temperature. The above quantities are averaged over 10^6 tu. Note that for the ideal first order phase transition one would expect $\sigma_C(T)$ to be a step function. However, since a transition that would be first order in the limit of the infinite size is considered, $\sigma_C(T)$ exhibits only step-function like behavior. The difference between core residues and other residues is that at T_f the average rms displacement of the core residues is smaller than 15, which indicates that they are in contact (see caption to Fig. 3.12). On the contrary, the average rms displacement of the non-core residues is greater than 15, indicating that these residues are not in contact.

Next, the rms displacement $\sigma_C(T)$ is calculated for the selected 25% coldest residues (the core) and $\sigma_O(T)$ for the rest of the residues. Their dependence on temperature, as well as the dependence on temperature of the rms displacement calculated for *all* residues $\sigma(T)$ are shown in Fig. 3.13. There is a pronounced difference in the behavior of the rms displacement of the core residues and the rest of the residues below T_f. At T_f their behavior is the same, due to the fact that all the attractive interactions are balanced by the repulsion of the excluded volume. Above T_f the difference between $\sigma_C(T)$ and $\sigma_O(T)$ is only due to the fact that the core residues have most of the NC and, therefore, are more likely to spend time together even at $T > T_f$.

To study the behavior of the globule at T_f, the probability distribution of the energy states $\mathcal{E}$ of the globule maintained at $T_f = 1.46$ during 10^6 tu is subdivided into five regions: A, B, C, D, and E (see Fig. 3.14a). Region A corresponds to the folded state; region B corresponds to the transitional state between the folded state and PFS; region C corresponds to the PFS; region D corresponds to the transitional state between PFS and the completely unfolded state. Next, the rms displacement for each residue for each of the above regions is plotted (see Fig. 3.14b). Note that in region A all residues stay in contact; in region C both N- and C-termini tails break away, forming PFS; in region D, there are only a few core residues still intact; and in region E none of the residues is in contact. In region B, one can observe that part of the C-terminus tail residues are not in contact, indicating the formation of a PFS. Next, the dependence of the selected 11 core residues (see caption to Fig. 3.12) on the average energy of the window of the corresponding region is plotted (see Fig. 3.15). One observes that core residues remain close to one another even in the second transitional state D between the PFS and the completely unfolded state.

3.6 Thermal properties of the heat bath

The system is also studied by cooling it from a high temperature state. This technique corresponds to the simulated annealing, due to the fact that the temperature control is governed by the ghost particles that are present in the system. One finds that if the target temperature is above 1.1, the globule always reaches the state cor-

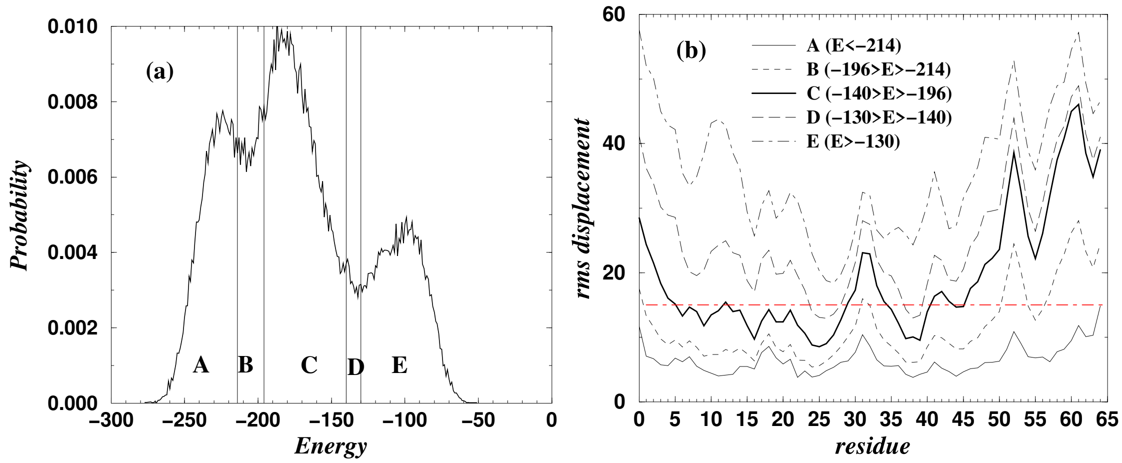

Figure 3.14: Energy states of the 65-mer at T_f.
(a) The probability distribution of the energy states $\mathcal{E}$ of the globule maintained at $T_f = 1.46$ during 10^6 tu. The probability distribution is divided into five regions: A, B, C, D, and E. Region A corresponds to the folded state; region B corresponds to the transitional state between the folded state and PFS; region C corresponds to the PFS; region D corresponds to the transitional state between PFS and the completely unfolded state. (b) The plot of the rms displacement $\sigma_i(T)$ for each residue $i = 0, 1, ..., 64$ for various regions A, B, C, D, and E of the plot (a) averaged over 10^6 tu. Note that in region A all residues stay in contact; in region C both N- and C-terminus tails break away, forming PFS; in region D only a few of the core residues are still intact; and in region E none of the residues are in contact. The dashed dotted line in (b) indicates the breaking point of the NC (see Fig 3.12).

responding to native state. However, if the target temperature is 0.96, the globule reaches the state corresponding to the native state in only $\approx 70\%$ of the cases, in the time interval of 10^5 time units. As an example, the cooling of the model protein from the high temperature state $T = 3.0$ to the low temperature state $T = 0.1$ is demonstrated on Fig. 3.16. The model protein collapses after 1200 tu.

What is particularly remarkable about Fig. 3.16 is that one can follow the kinetics of the collapse. First, the globule gets trapped in some misfolded conformation, where it stays for about 1000 tu (see Fig. 3.16a), and then it collapses to the native state.

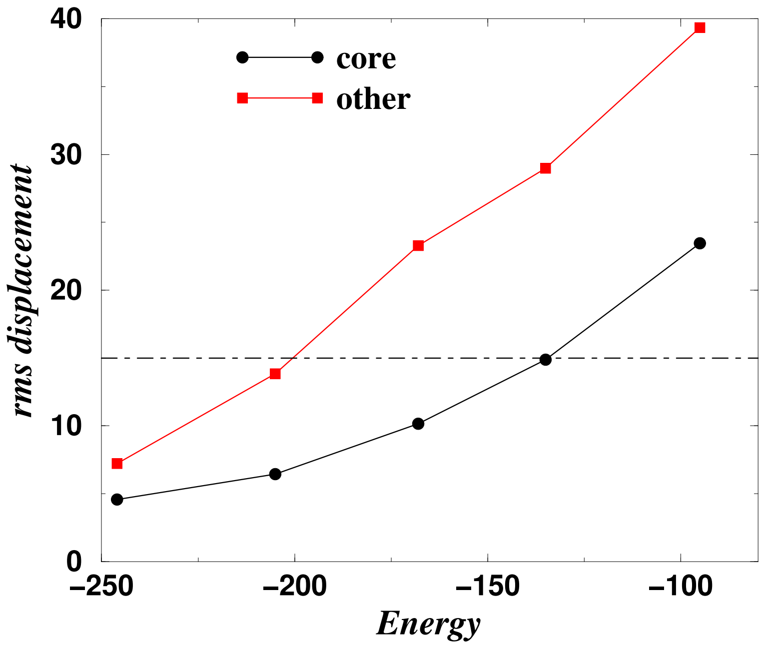

Figure 3.15: Rms displacement of the core residues.
The dependence of the rms displacement of the core residues (circles) and the rest of the residues (squares) on the average energy $\mathcal{E}$ of the window of the corresponding region. Note that core residues stay intact even in the second transitional state D between the PFS and the completely unfolded state. The dashed dotted line indicates the breaking point of the NC (see Fig 3.12).

The time behavior of the energy, however, can look a bit puzzling. After the rms displacement drops to close to 0, indicating the native state, the energy is still higher than that of the native state for about 10^4 tu (see Fig. 3.16b). The key to resolving this puzzle is the fact that after the collapse of the model protein, its potential energy transforms to kinetic energy, which slowly decreases by thermal equilibration with the bath of the ghost particles.

3.7 Discussion

One finds that the classical model of the self-avoiding chain with excluded volume shows good agreement with the simulations. From simple arguments and simulations

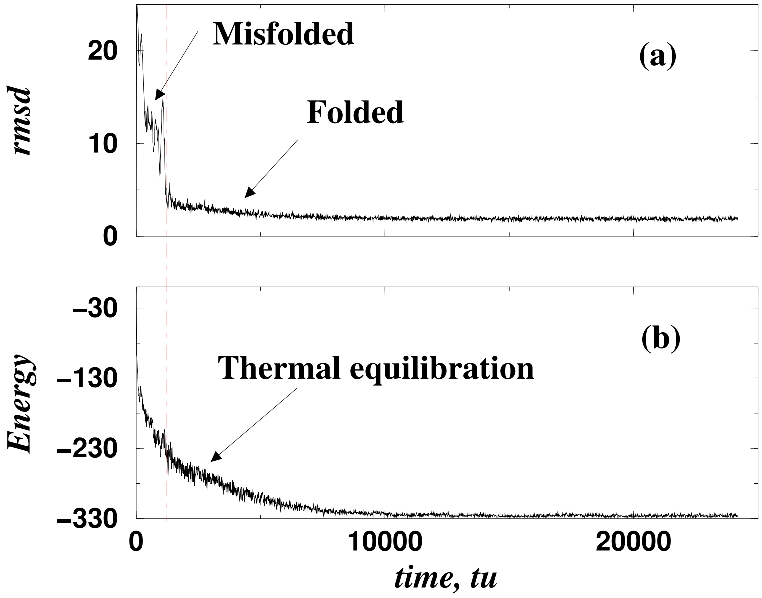

Figure 3.16: Cooling kinetics of the 65-mer.

The time evolution of the (a) rms displacement per residue of the globule from its native state and (b) energy when the system is cooled from the high temperature ($T = 3.0$) unfolded state down to the low temperature ($T = 0.1$) state. The model protein gets trapped in the misfolded conformation after 200 tu and then proceeds to its native state after 1000 tu. Although the rms displacement of the globule from its native state is close to 0 after 1200 tu, the energy of the globule is higher than the energy of the native state by about 10^4 tu. This effect is due to the thermal bath ghost particles which thermally equilibrate the system during $t_{relax.} \approx 10^4$ tu relaxation time.

it is shown that the fraction of NC at the folding temperature T_f is more than 70%, consistent with the presence of the core. The nucleus forms in the unstable transition state. From the transition state the globule jumps either to the completely unfolded conformation or to the folded conformation.

These simulations are in agreement with the recent work of Zhou and Karplus [163]. They performed discrete molecular dynamics simulations of *Staphylococcus aureus* protein A, the inter-residue interactions of which were modeled based on the Gō model [7, 8, 141]. The pair residues of the model protein, which form native contacts, had "square-well" potential of interaction with the depth of the well equal to $B_N\epsilon$, while all other pair residues had "square-well" potential of interaction with the depth of the well equal to $B_O\epsilon$. They characterized the difference between NC and NNC by the *"bias gap"*, g:

$$g = 1 - B_O/B_N\,. \tag{3.15}$$

Zhou and Karplus found that when $g = 1.3$, i. e. when the interaction between NC is of the opposite sign to the interaction between NNC, there is a strong first-order-like transition from the random coil to the ordered globule. The case with the present globule corresponds to $g = 2$, where, according to the work of Zhou and Karplus there should exist a strong first-order-like transition from the random coil to the ordered globule without intermediate.

The core residues are also selected and it is shown that their rms displacement behaves significantly differently than the rms displacement of the rest of the residues and exhibits step-function-like behavior upon temperature change. These findings are in agreement with the recent experimental study of the equilibrium hydrogen exchange behavior of *cytochrome c* of Bai et al. [180], who investigated the exposure of the amide hydrogens (NH) in *cytochrome c* to solvent (due to local and global unfolding fluctuations). The experiments were based on the properties of the amide hydrogens that are involved in hydrogen-bonded structure and can exchange with solvent hydrogens. Bai et al. demonstrated that proteins undergo folding $\rightleftharpoons$ unfolding transition "...through intermediate forms". They also selected these intermediate forms (cooperative units), which are 15 to 25 residues in size. The presence of PFS in the present simulations is thus in agreement with the findings of Bai et al. of the intermediate forms in *cytochrome c*.

The relation that is found between core residues and the nucleus is hard to establish due to the fact that TS is very unstable. Recent amide hydrogen exchange experiments on the CheY protein from *Escherichia coli* of Lacroix et al. [181] provided evidence for the residues involved in the folding nucleus. Furthermore, the lattice MC simulations of Abkevich et al. [155] also demonstrate that the presence of the nucleus is a necessary and sufficient condition for subsequent rapid folding to the native state. The crucial difference between the nucleus and the rest of the structure is in dynamics, which is manifest also in equilibrium fluctuations. All local unfolding fluctuations (i.e. the ones after which the chain returns rapidly back to the native state) keep the nucleus intact, while fluctuations that disrupt the nucleus lead to global unfolding: "descend" to the "unfolded" free energy minimum [140, 155]. This view is consistent with the hydrogen exchange experiments [180, 181]. Such behavior of the globule is consistent with a possible first order phase transition in a system of finite size.

3.8 Kinetics of the folding process

Thermodynamically, the folding transition in small proteins is analogous to a first-order transition whereby two thermodynamic states [185] (folded and unfolded) are free energy minima while intermediate states are unstable. The kinetic mechanism of transitions from the unfolded state to the folded state is nucleation [140, 186–189]. Folding nuclei can be defined as the minimal stable element of structure whose existence results in subsequent rapid assembly of the native state. This definition corresponds to a "post-critical nucleus" related to the first stable structures that appear immediately after the transition state is overcome [155]. The thermal probability of a transition state conformation is low compared to the folded and unfolded states, which are both accessible at the folding transition temperature T_f (see Fig. 3.17).

Kinetic analyses [112, 140, 155, 189–192] for a number of lattice model chains of different lengths and degrees of sequence design (optimization) point to a specific

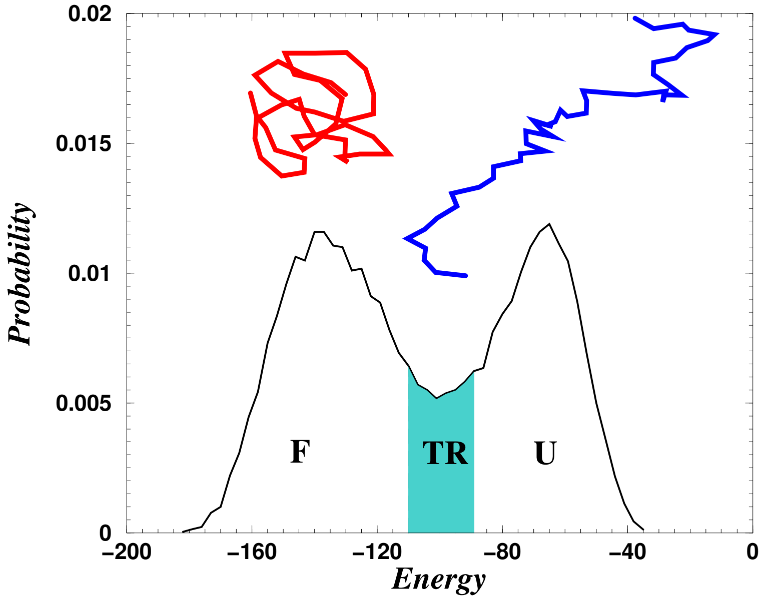

Figure 3.17: The distribution of the energy states of the 46-mer at T_f.
The probability distribution of the energy states E of the 46-mer maintained at the folding transition temperature $T_f = 1.44$. The bimodal distribution indicates the presence of two dominant states: the folded (region F) and the unfolded (region U) states. The transition state ensemble belongs to region TR of the histogram $\{-110 < E < -90\}$. The insets show typical conformations in the folded and unfolded regions.

protein folding nucleus scenario. Passing through the transition state with subsequent rapid assembly of the native conformation requires the formation of some (small) number of specific obligatory contacts (protein folding nucleus). This result has been verified [112] for sequences designed in the lattice model using different sets of potentials, where it is shown that nucleus location was identical for two different sequences designed with different potentials to fold into the same structure of a lattice 48-mer. This finding and related results [189, 193] suggest that the folding nucleus location depends more on the topology of the native structure than on a particular sequence that folds into that structure.

The dominance of geometrical/topological factors in the determination of the folding nucleus is a remarkable property that has evolutionary implications (see below). It is important to understand the physical origin of this property of folding proteins and to assess its generality. To this end, it is important to study other than lattice models and other than Monte-Carlo dynamic algorithms. To search for the nucleus in a continuous off-lattice model, the discrete molecular dynamics (MD) simulation technique [113, 162, 163] (the Gō model [7, 8, 141] with the square-well potential of the inter-residue interaction) is employed.

3.8.1 The transition region

The proposed method to search for a folding nucleus is based on the observation [155] that equilibrium fluctuations around the native conformation can be separated into "local" unfolding (followed by immediate refolding) and "global" unfolding that leads to a transition into an unfolded state and requires more time to refold. Local unfolding fluctuations are the ones that do not reach the top of the free energy barrier and, hence, are committed to moving quickly back to the native state. In contrast, global unfolding fluctuations are the ones that overcome the barrier and are committed to descend further to the unfolded state. Similarly, the fluctuations from the unfolded state can be separated into those that descend back to the unfolded state and those that result in productive folding. The difference between the two modes of fluctuation is whether or not the major free energy barrier is overcome. This means that the nucleation contacts (i. e. the ones that are formed on the "top" of the free energy barrier as the chain passes it upon folding) should be identified as contacts that are present in the "maximally locally unfolded" conformations but are lost in the globally unfolded conformations of comparable energy.

Thus, in order to identify the folding nucleus, the conformations of the 46-mer that appear in various kinds of folding $\rightleftharpoons$ unfolding fluctuations are studied. First, consider the time behavior of the potential energy at T_f (see Fig. 3.19). The transition state conformations belong to the transition region TR from the folded state to the unfolded state that lies in the energy range $\{-110 < E < -90\}$ (see Fig. 3.17). Region TR corresponds to the minimum of the histogram of the energy distribution. If one knows the past and the future of a certain conformation that belongs to the

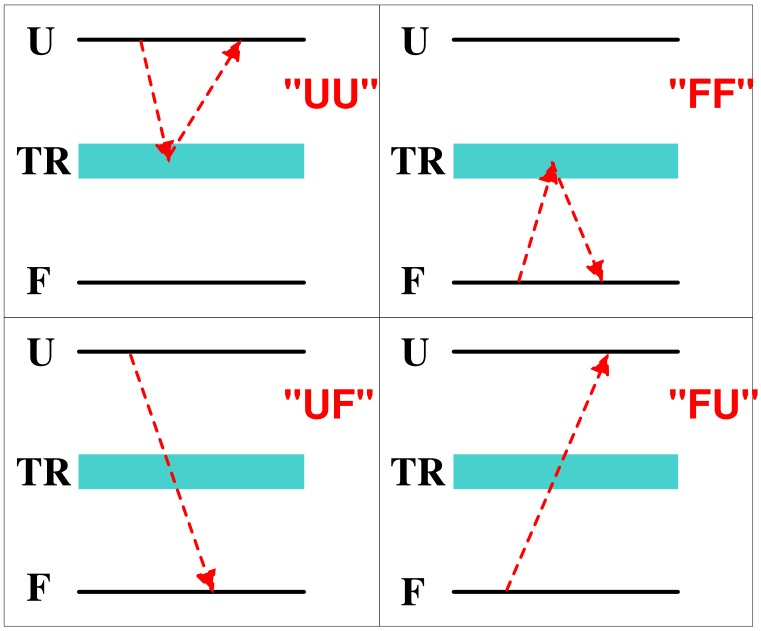

Figure 3.18: Four types of conformations.
The schematic definition of the four types of conformations: FF, UU, UF, and FU.

TR, then the four types of such conformations are distinguishable (see Figs. 3.18 and 3.19): *(i)* UU conformations that originate in and return to the unfolded region without ascending to the folded region; *(ii)* FF conformations that originate in and return to the folded region without descending to the unfolded region; *(iii)* UF conformations that originate in the unfolded region and descend to the folded region; and *(iv)* FU conformations that originate in the folded region and descend to the unfolded region.

If the nucleus exists, then the UF, FU, FF, and UU conformations must have different properties depending on their history. For example, the difference between UF, FU, FF, and UU conformations is pronounced for the rms displacements from the native state of the residues in the vicinity of the residues 10 and 40 and is illustrated in Fig. 3.20. One difference between the FF conformations and UU conformations is that the protein folding nucleus is more likely to be retained in the FF conformations than in the UU conformations. The contacts belonging to the critical nucleus ("nucleation contacts") start appearing in the UF conformations, and start disappearing

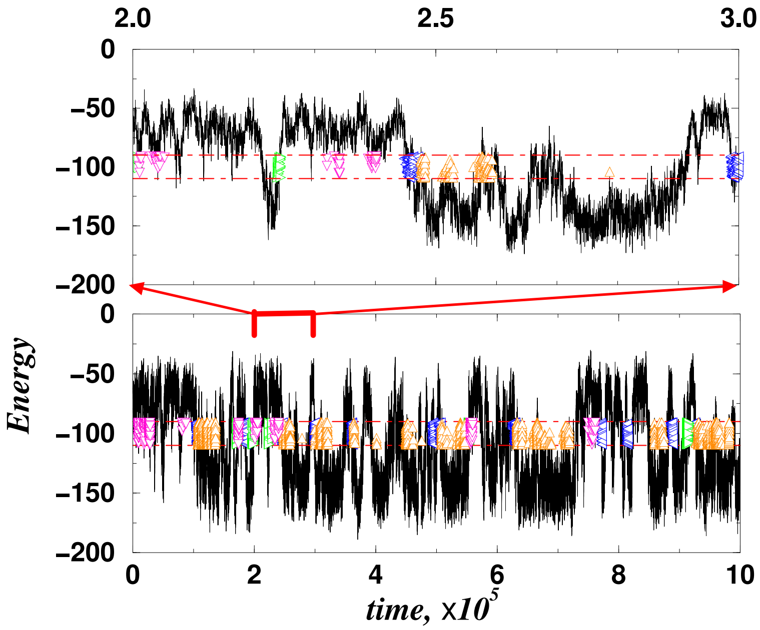

Figure 3.19: The time evolution of energy E of the 46-mer maintained at T_f. The time evolution of the energy E of the 46-mer maintained at the folding temperature $T_f = 1.44$. The focus is on region TR: $\{-110 < E < -90\}$, which is denoted by two dash-dotted lines. The triangles with arrows pointing up, down, left and right denote the conformations FF, UU, UF and FU. The upper figure magnifies the region $2 \cdot 10^5 \leq t \leq 3 \cdot 10^5$.

in the FU conformations, so that the frequencies of nucleation contacts in UF and FU conformations should be in between FF and UU.

The goal is to select the contacts that are crucial for the folding $\rightleftharpoons$ unfolding transition. To this end, the contacts that appear much more often in the FF conformations than in the UU conformations are selected. One discovers that if the threshold for the difference in contact frequencies between FF and UU conformations is set to 0.2, then there are only five contacts that persist: (residue 11, residue 39), $(10, 40)$, $(11, 40)$, $(10, 41)$, and $(11, 41)$ (see Figs. 3.21 and 3.22). These contacts can serve as evidence for the protein folding nucleus in the folding $\rightleftharpoons$ unfolding transition in the model.

Next, it is demonstrated that these five selected contacts belong to the protein

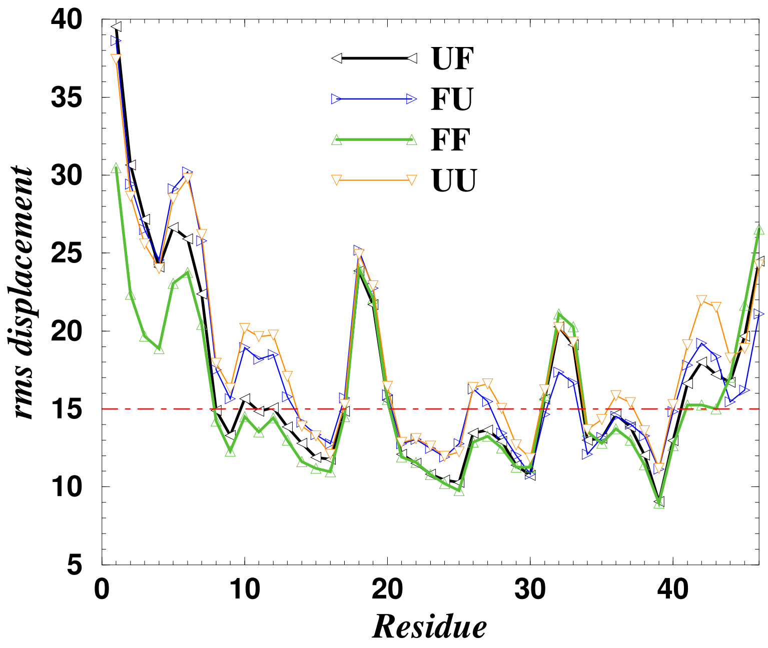

Figure 3.20: RMS displacement for FF, UU, UF and FU conformations

The average rms displacement of each residue from its native state for each of the four types of TR conformations: FF, UU, UF and FU. Note that there is a pronounced difference in rms displacement of the residues from their native state in the vicinity of the residues 10 and 40. The dash-dotted line indicates the breaking point of the native contacts, i. e. when the rms displacement σ is approximately the size of the average relative distance between pairs of residues, i. e. $\sigma = (a_0 + a_1)/2 \approx 15$ (see Sec. 3.3 for the definition of a_0 and a_1).

folding nucleus. Suppose just one of them is fixed, e. g. $(10, 40)$, i. e. a covalent ("permanent") link is imposed between residue 10 and residue 40. If this contact belongs to the protein folding nucleus, its fixation by a covalent bond would eliminate the barrier between the folded and unfolded states, i. e. only the native basin of attraction will remain. Hence, it is hypothesized that the cooperative transition between the unfolded and folded state will be eliminated and the energy histogram (Fig. 3.17) should change qualitatively from bimodal to unimodal. The MD simulations support this hypothesis (Fig. 3.23): fixation of only *one* nucleation contact, $(10, 40)$, gives rise to a qualitative change in the energy distribution from bimodal

Figure 3.21: The contact map of the 46-mer.

The darker the shade of grey, the larger is the frequency of a contact. Above the diagonal of the square matrix shows the *native contacts* (see Sec. 3.3) of the FF conformations (if the *native* contact frequency is larger than 0.2). Below the diagonal of the square matrix shows the *difference between the frequencies of the native contacts* in FF and UU conformations (if this difference is larger than 0.2). Five contacts that persist in the FF conformations — $(11, 39)$, $(10, 40)$, $(11, 40)$, $(10, 41)$, and $(11, 41)$ — are marked by crosses. The figure shows that identification of the protein folding nucleus is facilitated by the method used to construct the region of the matrix below the diagonal.

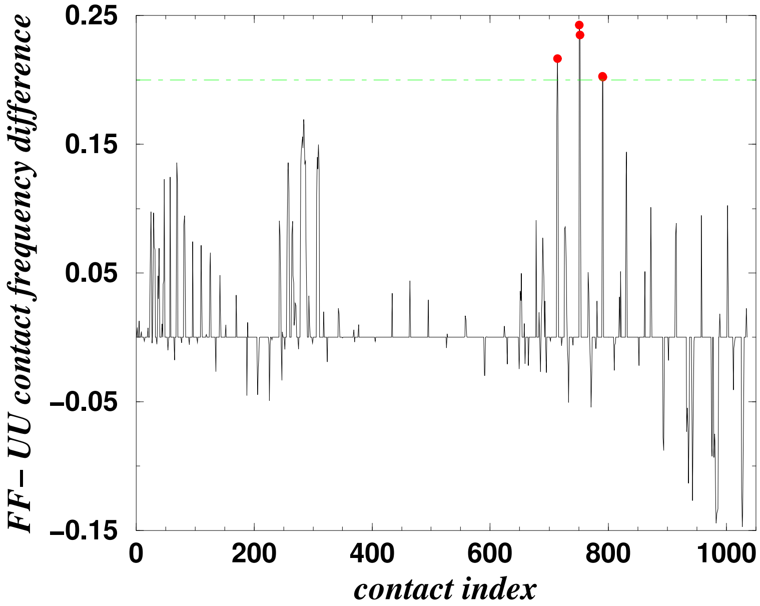

Figure 3.22: The contact frequency difference of the FF versus UU conformations. The contact frequency difference of the FF conformations versus UU conformations plotted versus the contact index $k = i(i-1)/2 + j$, where $0 \leq j < i = 1, 2, \ldots 45$. The dashed-dotted line indicates the threshold value for the frequency differences. The circles indicate the contacts whose frequency differences are larger than the threshold. (Due to the scale used, only four contacts can be seen in the figure.)

to unimodal. Indeed, the probability of finding an unfolded state with a fixed link, $(10, 40)$, belonging to the protein folding nucleus is drastically reduced compared to the probability of the unfolded state of the original 46-mer, indicating the importance of the selected contact $(10, 40)$.

To provide a "control" that a *specific* contact plays such a dramatic role in changing the character of the energy landscape, a randomly-chosen contact is fixed, $(19, 37)$, which is not predicted by the analysis, to belong to the critical nucleus. The present hypothesis predicts no qualitative change in the energy distribution histogram since the barrier, determined primarily by nucleation contacts, should not change dramatically for this control. Fig. 3.23 shows that this is indeed the case. (The stability of the folded state is somewhat increased for the control because any

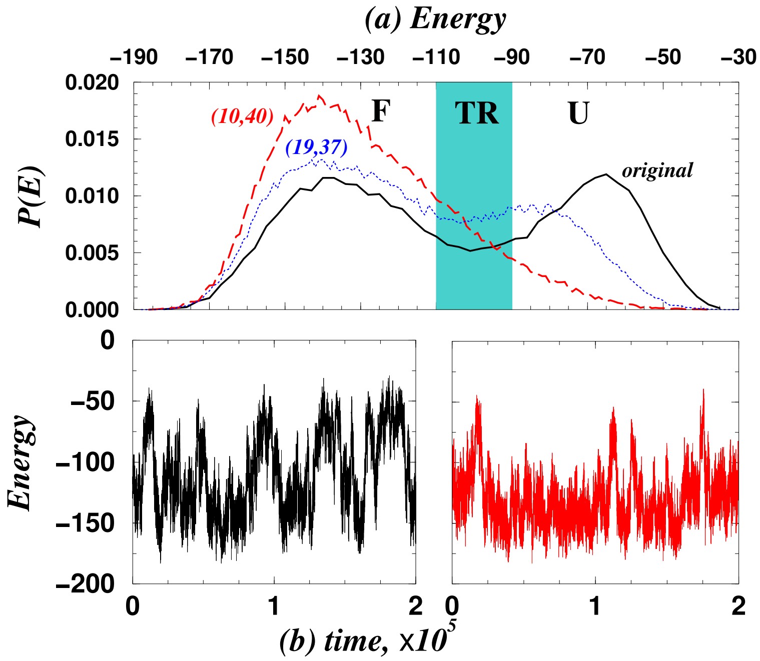

Figure 3.23: The probability distribution of the energy states of the 46-mer with fixed nucleic and control contacts.

(a) The probability distribution of the energy states E of the: *(i)* original 46-mer (at $T_f = 1.44$), *(ii)* 46-mer (at $T = 1.46$) with a fixed contact belonging to the protein folding nucleus, $(10, 40)$, and *(iii)* 46-mer (at $T = 1.46$) with fixed randomly-chosen control contact $(19, 37)$ that does not belong to the protein folding nucleus. Note that the probability of the unfolded state of the 46-mer with a fixed contact belonging to the protein folding nucleus is suppressed compared to that of the original 46-mer. (b) The time evolution of the energy E of *(i)* original (left) and *(ii)* fixed $(10, 40)$ contact (right). Case *(iii)* fixed $(19, 37)$ contact is similar to *(i)*, so we do not show it. For case *(i)*, the fluctuations are mostly between two extreme values of energy, corresponding to the folded and unfolded states. In contrast, for case *(ii)*, the fluctuations are mostly around one energy value, corresponding to the folded state.

preformed native contacts decrease the entropy of the unfolded state — i. e. they stabilize the folded state).

It is also found that for the UF conformation, the rms displacements of the residues from their native positions are smaller than those for the FU conformations (Fig. 3.20). This observation is consistent with the fact that the nucleation contacts are formed first upon entering into productive folding and are destroyed last upon unfolding.

3.8.2 Discussion

The main conclusion is that the existence of a few (≈ 5) specific contacts is signature of the transition state conformations. Those contacts can be defined as the protein folding nucleus. Other contacts may also be present in transition state conformations. However, they are optional and vary from conformation to conformation, while nucleation contacts are present in transition state conformations with high probability. Formation of nucleation contacts can be considered an obligatory step in the folding process: after they are formed the major barrier is overcome and subsequent folding proceeds "downhill" in the free energy landscape without encountering any further major free energy barriers. This is illustrated by the results that show that even one nucleation contact eliminates the free energy barrier and, hence, leads to fast "downhill" motion to the folded state. As a control, these results show that fixation of an arbitrary non-nucleation contact does not result in a similar effect.

The protein folding nucleus scenario of the transition state was initially derived from Monte-Carlo studies of lattice models [112, 155, 189] and was consistent with protein engineering experiments with several small proteins [152–154, 194]. Here, for the first time, this scenario is confirmed in off-lattice MD simulations. The consistency between conclusions made in different simulations [112, 155, 189] and in experiments [152–154, 194] is remarkable, and supports the possibility that the protein folding nucleus formation is a generic scenario to describe the protein folding transition state.

The present study buttresses the point that the location of a protein folding nucleus is determined by the geometry of the native state rather than the energetics of interactions in the native state (the two factors are not entirely independent, since native contacts must be generally more stable to provide stability to the native conformation). In the present study, the Gō model (where all native contacts have the same energy) was used. Nonetheless, it turns out that some contacts (nucleation) are "more equal than others" in terms of their role in shaping the free energy landscape of the chain and determining folding kinetics. This fact has implications for protein evolution, raising the possibility that proteins that have similar structures but different sequences may have similarly located protein folding nuclei. This prediction was verified for SH3 domains [152, 195] and for cold-shock proteins [196].

In terms of the evolutionary selection of protein sequences, the robustness of the folding nucleus suggests that any additional evolutionary pressure that controls the folding rate may have been applied selectively to nucleus residues, so that nucleation positions may have been under double (stability + kinetics) pressure in all proteins that fold into a given structure. Such additional evolutionary pressure has indeed been found in the analysis of several protein superfamilies [191].

3.9 The significance of the protein folding problem

The solution of the direct and inverse protein folding problems will impact modern medicine, biology, and biomolecular engineering [11]. One important clinical outcome would be effective drug design. For example, Benjamin and McMillan [197] underline the crucial significance of the heat shock family of stress proteins (HSPs) in cardiovascular conditions such as cardiac hypertrophy, vascular wall injury, ischemic preconditioning, and aging. HSPs serve as part of the defense strategy to ensure survival of the cell. They are engaged in response to various physiological stresses (e. g., heat, hemodynamics, mutant proteins, and oxidative injury) and in decisions to repair or degrade misfunctional, damaged proteins. Therefore, the development of proteins similar in function to the HSPs may aid the protection of essential cellular proteins.

In addition, solving the protein design problem would shed light on the relation between protein structure and function. For example, the mechanism of muscle contraction rests upon several proteins, namely myosin and actin filaments [198] and titin [199, 200]. Though the physical mechanisms of these proteins are thought to be known, the global picture of muscle contraction is still unresolved.

Specific interest has been generated by the special class of transmembrane proteins, *integrins*, responsible for the transduction of the mechanical signals between cell and extracellular matrix (ECM), cell-cell adhesion, and also involved in a vast variety of intercellular phenomena (e. g. *"tensegrity"*) [201]. The distinguished feature of integrins and titins is the strong binding affinity due to the special region called *type III repeat* [202]. This motif constitutes 2% of all proteins in the cell. *Type III repeat* fold will be used as a target model in my design procedure.

Important breakthroughs may also come from the understanding of *(i)* the expression mechanisms of growth factors, e. g. vascular endothelial growth factor (VEGF) [203], which are synthesized and secreted by differentiated cells in response to stimuli including hypoxia; *(ii)* the function of the low density lipoprotein (LDL) receptor [204], which mediates cellular uptake of lipoprotein particles. The LDL receptor plays a key role in cholesterol homeostasis.

The role of proteins in the cell is essential. Proteins are involved in most cellular and intercellular functions. The above mentioned proteins constitute just a small subset of all the proteins involved in various structural, regulatory, and defense processes. The knowledge of the thermodynamics and kinetics of protein folding process and basic mechanisms that govern folding process might aid us to solve the protein folding problem. The ability to predict the structure of a protein from its sequence and the ability to predict the sequence of a protein from its structure will not only allow us to understand protein function and develop new drugs, but also to open the door to yet unforeseen discoveries.

Appendix A

DNA Clumping: Derivation of Eqs. (2.47) and (2.48)

To derive Eq. (2.47) let us start from the definition of $\langle \ell \rangle_n$:

$$\langle \ell \rangle_n = \frac{L}{\sum_\ell^\infty \sum_{i=1}^M N_i(\ell)} = \frac{1}{\sum_\ell^\infty \sum_{i=1}^M N_i(\ell)/L} \,. \tag{A.1}$$

On the other hand, the number of i^{th} oligomers in a sequence of length L is determined by Eq. (2.42). Thus,

$$\langle \ell \rangle_n = \frac{1}{\sum_{i=1}^M \sum_\ell^\infty p_i^\ell (1-p_i)^2} = \frac{1}{\sum_{i=1}^M p_i(1-p_i)} = \frac{1}{1-\sum_{i=1}^M p_i^2} \,. \tag{A.2}$$

derive Eq. (2.48) is derived analogously:

$$\langle \ell \rangle_w = \frac{1}{L}\sum_\ell^\infty \sum_{i=1}^M \ell^2 N_i(\ell) = \sum_\ell^\infty \sum_{i=1}^M \ell^2 p_i^\ell (1-p_i)^2 = \sum_{i=1}^M (1-p_i)^2 \sum_\ell^\infty (\frac{d}{d \ln p_i})^2 p_i^\ell \,. \tag{A.3}$$

Thus,

$$\langle \ell \rangle_w = \sum_{i=1}^M (1-p_i)^2 (p_i \frac{d}{dp_i})^2 \sum_\ell^\infty p_i^\ell = \sum_{i=1}^M (1-p_i)^2 (p_i \frac{d}{dp_i})^2 \frac{p_i}{1-p_i} = 1 + 2\sum_{i=1}^M \frac{p_i^2}{1-p_i} \,. \tag{A.4}$$

Bibliography

[1] B. H. Zimm and J. K. Bragg, The theory of the phase transition between helix and random coil in polypeptide chains, *J. Chem. Phys.* **31**, 526–535 (1959)

[2] J. J. Hopfield, Electron transfer between biological molecules by thermally activated tunneling, *Proc. Natl. Acad. Sci. U. S. A.* **71**, 3640–3644 (1974)

[3] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, *Proc. Natl. Acad. Sci. U. S. A.* **79**, 2554–2558 (1982)

[4] E. Evans, New physical concepts for cell amoeboid motion, *Biophys. J.* **64**, 1306–1322 (1993)

[5] B. Derrida, Random-energy model: limit of a family of disordered models, *Phys. Rev. Lett.* **45**, 79–82 (1980)

[6] K. R. Bhaskar, P. Garik, B. S. Turner, J. D. Bradley, R. Bansil, H. E. Stanley and J. T. LaMont, Viscous fingering of HCl through gastric mucin, *Nature* **360**, 458–461 (1992)

[7] N. Gō and H. Abe, Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation, *Biopolymers* **20**, 991–1011 (1981)

[8] H. Abe and N. Gō , Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins, *Biopolymers* **20**, 1013–1031 (1981)

[9] D. S. Goodsell, Inside a living cell, *Trends in Biochem. Sci.* **16**, 203–206 (1991)

[10] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts and J. D. Watson, *Molecular biology of the cell*, Garland Publishing, New York, 1994

[11] T. Creighton, *Proteins: structures and molecular properties*, W. H. Freeman and Co., New York, ii edition, 1993

[12] M. D. Frank-Kamenetskii, *Unraveling DNA*, VCH Publishers, Inc., New York, ii edition, 1993

[13] G. I. Bell, Roles of repetitive sequences, *Computers and Chemistry* **16**, 135–143 (1992)

[14] G. I. Bell, Repetitive DNA sequences: some considerations for simple sequence repeats, *Computers and Chemistry* **17**, 185–190 (1993)

[15] R. I. Richards and G. R. Sutherland, Simple repeat DNA is not replicated simply, *Nature Genetics* **6**, 114–116 (1994)

[16] G. R. Sutherland and R. I. Richards, Simple tandem DNA repeats and human genetic disease, *Proc. Natl. Acad. Sci. U. S. A.* **92**, 3636–3641 (1995)

[17] X. Chen, S. V. Mariappan, P. Catasti, R. Ratliff, R. K. Moyzis, A. Laayoun, S. S. Smith, E. M. Bradbury and G. Gupta, Hairpins are formed by the single DNA strands of the fragile X triplet repeats: structure and biological implications, *Proc. Natl. Acad. Sci. U. S. A.* **92**, 5199–5203 (1995)

[18] A. M. Gacy, G. Goellner, N. Juramic, S. Macura and C. T. McMurray, Trinucleotide repeats that expand in human disease form hairpin structures in vitro, *Cell* **81**, 533–540 (1995)

[19] K. Orth, J. Hung, A. Gazdar, A. Bowcock, J. M. Mathis and J. Sambrook, Genetic instability in human ovarian cancer cell lines, *Proc. Natl. Acad. Sci. U. S. A.* **91**, 9495–9499 (1994)

[20] A. M. Bowcock, A. Ruiz-Linares, J. Tomfohrde, E. Minch, J. R. Kidd and L. L. Cavalli-Sforza, High resolution of human evolutionary trees with polymorphic microsatellites, *Nature* **368**, 455–457 (1994)

[21] B. Olaisen, M. Bekkemoen, P. Hoff-Olsen and P. Gill. Human VNTR mutation and sex. in S. D. J. Pena, R. Chakraborty, J. T. Epplen and A. J. Jeffreys, eds., *DNA Fingerprinting: State of the Science*. Springer-Verlag, Basel 1993.

[22] J. Jurka and C. Pethiyagoda, Simple repetitive DNA sequences from primates: compilation and analysis, *J. Mol. Evol.* **40**, 120–126 (1995)

[23] C. Burge, A. M. Campbell and S. Karlin, Over- and under- representation of short oligonucleotides in DNA sequences, *Proc. Natl. Acad. Sci. U. S. A.* **89**, 1358–1362 (1992)

[24] S. Karlin and J. Mrázek, What drives codon choices in human genes, *J. Mol. Biol.* **262**, 459–472 (1996)

[25] J. S. Beckmann and J. L. Weber, Survey of human and rat microsatellites, *Genetics* **12**, 627–631 (1992)

[26] R. L. Stallings, Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: implications for human genetic diseases, *Genetics* **21**, 116–121 (1994)

[27] R. D. Wells, Molecular basis of genetic instability of triplet repeats, *J. Biol. Chem.* **271**, 2875–2878 (1996)

[28] E. Kremer, M. Pritchard, M. Lynch, S. Yu, K. Holman, E. Baker, S. T. Warren, D. Schlessinger, G. R. Sutherland and R. I. Richards, Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence $p(\mathrm{CCG})n$, *Science* **252**, 1711–1714 (1991)

[29] Huntington's Disease Collaborative Research Group, A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes, *Cell* **72**, 971–983 (1993)

[30] K. Davies and S. Warren, eds., *Genome analysis,* Vol. 7, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1993

[31] A. R. La Spada, H. L. Paulson and K. H. Fischbeck, Trinucleotide repeat expansion in neurological disease, *Ann. Neurol.* **36**, 814–822 (1994)

[32] G. Bates and H. Lehrach, Trinucleotide repeat expansions and human genetic disease, *BioEssays* **16**, 277–284 (1994)

[33] S. Panzer, D. P. Kuhl and C. T. Caskey, Unstable triplet repeat sequences: a source of cancer mutations?, *Stem Cells* **13**, 146–157 (1995)

[34] Y. Ionov, M. A. Peinado, S. Malkhosyan, D. Shibata and M. Perucho, Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for clonic carcinogenesis, *Nature* **363**, 558–561 (1993)

[35] T. A. Kunkel, Slippery DNA and diseases, *Nature* **365**, 207–208 (1993)

[36] R. Wooster, A.-M. Cleton-Jansen, N. Collins, J. Mangion, R. S. Cornelis, C. S. Cooper, B. A. Gusterson, B. A. J. Ponder, A. von Deimling, O. D. Wiestler, C. J. Cornelisse, P. Devilee and M. R. Stratton, Instability of short tandem repeats (microsatellites) in human cancers, *Nature Genetics* **6**, 152–156 (1994)

[37] M. Strand, T. A. Prolla, R. M. Liskay and T. D. Petes, Destabilization of tracts of simple repeatative DNA in yeast by mutations affecting DNA, *Nature* **365**, 274–276 (1993)

[38] L. A. Aaltonen, P. Peltomäki, F. S. Leach, P. Sistonen, L. Pylkkänen, J. P. Mecklin, H. Järvinen, S. M. Powell, J. Jen, S. R. Hamilton, G. M. Petersen, K. W. Kinzler, B. Vogelstein and A. de la Chapelle, Clues to the pathogenesis of familial colorectal cancer, *Science* **260**, 812–816 (1993)

[39] S. N. Thibodeau, G. Bren and D. Schaid, Microsatellite instability in cancer of the proximal cancer, *Science* **260**, 816–819 (1993)

[40] D. Shibata, M. A. Peinado, Y. Ionov, S. Malkhosyan and M. Perucho, Genomic instability in repeated sequences is an early somatic event in colorectal tumorigenesis that persists after transformation, *Nature Genetics* **6**, 273–281 (1994)

[41] S. M. Baker, C. E. Bronner, L. Zhang, A. W. Plug, M. Robatzek, G. Warren, E. A. Elliott, T. Ashley J. Yu, N. Arnheim, R. A. Flavell and R. M. Liskay, Male mice defective in the DNA mismatch repair gene PMS2 exhibit abnormal chromosome synapsis in meiosis, *Cell* **82**, 309–319 (1995)

[42] N. V. Dokholyan, S. V. Buldyrev, S. Havlin and H. E. Stanley, Distribution of base pair repeats in coding and noncoding DNA sequences, *Phys. Rev. Lett.* **79**, 5182–5185 (1997)

[43] N. V. Dokholyan, S. V. Buldyrev, S. Havlin and H. E. Stanley, Non-randomness of dimeric tandem repeats in noncoding DNA sequences, *J. Theor. Biol.* **submitted** (1998)

[44] J. L. Weber and P. E. May, Abundant class of human DNA polymorphisms which can be typed using the Polymerase Chain Reaction, *Am. J. of Hum. Gen.* **44**, 388–396 (1989)

[45] L. M. Silver, Bouncing off microsatellites, *Nature Genetics* **2**, 8–9 (1992)

[46] D. C. Rubinsztein, W. Amos, J. Leggo, S. Goodburn, S. Jain, S. H. Li, R. L. Margolis, C. A. Ross and M. A. Ferguson-Smith, Microsatellite evolution: evidence for directionality and variation in rate between species, *Nature Genetics* **10**, 337–343 (1995)

[47] L. Jin and R. Chakraborty, Population dynamics of DNA fingerprint patterns within and between populations, *Gen. Res.* **63**, 1–9 (1995)

[48] R. Nussinov, Strong adenine clustering in nucleotide sequences, *J. Theor. Biol.* **85**, 285–291 (1980)

[49] Y. A. Sprizhitsky, Y. D. Nechipurenko, A. A. Alexandrov and M. V. Volkenstein, Statistical analysis of nucleotide runs in coding and noncoding DNA sequences, *J. Biomol. Str. & Dyn.* **6**, 345–358 (1988)

[50] R. Nussinov, Compositional variations in DNA sequences, *CABIOS* **7**, 287–293 (1991)

[51] R. Nussinov, Doublet frequencies in evolutionary distinct groups, *Nucl. Acid Res.* **12**, 1749–1763 (1984)

[52] J. Mrazek and K. Kypr, Biased distribution of adenine and thymine in gene nucleotide sequences, *J. Mol. Evol.* **39**, 439–447 (1994)

[53] K. A. Marx, S. T. Hess and R. D. Blake, Characteristics of the large $(dA)\cdot(dT)$ homopolymer tracts in *D. discoideum* gene flanking and intron sequences, *J. Biomol. Str. & Dyn.* **11**, 57–66 (1993)

[54] G. Yagil, The frequency of two-base tracts in eukaryotic genomes, *J. Mol. Evol.* **37**, 123–130 (1993)

[55] R. L. Stallings, A. F. Ford, D. Nelson, D. C. Torney, C. E. Hildebrand and R. K. Moyzis, Evolution and distribution of $(GT)_n$ repetitive sequences in mammalian genomes, *Genetics* **10**, 807–815 (1991)

[56] G. I. Bell and J. Jurka, The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single step mutation process, *J. Mol. Evol.* **44**, 414–421 (1997)

[57] D. G. Arqués and C. J. Michel, Periodicities in introns, *Nucl. Acid Res.* **15**, 7581–7592 (1987)

[58] A. K. Konopka, G. W. Smythers, J. Owens and J. V. Maizel Jr., Distance analysis helps to establish characteristic motifs in intron sequences, *Gene Anal. Techn.* **4**, 63–74 (1987)

[59] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C.-K. Peng, M. Simons and H. E. Stanley, Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis, *Phys. Rev. E* **51**, 5084–5091 (1995)

[60] J. W. Fickett and C. S. Tung, Assessment of Protein Coding Measures, *Nucl. Acid Res.* **20**, 6441–6450 (1992)

[61] B. V. Gnedenko, *The theory of probability*, Chelsea Publishing Company, New York, 1968

[62] W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, *Numerical recipes*, Cambridge University Press, Cambridge, 1989

[63] W.-H. Li, The study of correlation structure of DNA sequences: a critical review, *Computers and Chemistry* **21**, 848–851 (19)

[64] R. A. Farber, T. D. Petes, M. Dominska, S. S. Hudgens and R. M. Liskey, Instability of simple sequence repeats in a mammalian cell line, *Hum. Molec. Gen.* **3**, 253–256 (1994)

[65] B. B. Mandelbrot, *The fractal geometry of nature*, W.H. Freeman, San Francisco, 1982

[66] A. Bunde and S. Havlin, eds., *Fractals and disordered systems*, Springer-Verlag, Berlin, 1991

[67] A. Bunde and S. Havlin, eds., *Fractals in science*, Springer-Verlag, Berlin, 1994

[68] T. Vicsek, M. Shlesinger and M. Matsushita, eds., *Fractals in natural sciences*, World Scientific, Singapore, 1994

[69] J. M. Garcia-Ruiz, E. Louis, P. Meakin and L. Sander, eds., *Growth patterns in physical sciences and biology*, [Proc. 1991 NATO Advanced Research Workshop, Granada, Spain, October 1991]. Plenum, New York, 1993

[70] A. Yu. Grosberg and A. R. Khokhlov, *Statistical physics of macromolecules*, AIP Press, New York, 1994

[71] J. B. Bassingthwaighte, L.S. Liebovitch and B.J. West, *Fractal physiology*, Oxford University Press, New York, 1994

[72] A.-L. Barabási and H. E. Stanley, *Fractal concepts in surface growth*, Cambridge University Press, Cambridge, 1995

[73] S. Miyazima, ed., *Proceedings of the international conference "Future of Fractals", Nagoya, Japan, 22-27 July 1995*, World Scientific, London, 1997

[74] T. Vicsek, *Fractal growth phenomena*, World Scientific, Singapore, ii edition, 1992

[75] J. Feder, *Fractals*, Plenum, New York, 1988

[76] D. Stauffer and H. E. Stanley, *From Newton to Mandelbrot: a primer in theoretical physics*, Springer-Verlag, New York, 1990

[77] H. E. Stanley and N. Ostrowsky, eds., *Random fluctuations and pattern growth: experiments and models*, [Proceedings 1988 Cargèse NATO ASI], Kluwer Academic Publishers, Dordrecht, 1985

[78] S. V. Buldyrev, N. V. Dokholyan, A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley and G. M. Viswanathan, Analysis of DNA Sequences Using Methods of Statistical Physics, *Physica A* **249**, 430–438 (1998)

[79] H. E. Stanley, *Introduction to phase transitions and critical phenomena*, Oxford University Press, London, 1971

[80] H. E. Stanley and N. Ostrowsky, eds., *Correlations and connectivity: geometric aspects of physics, chemistry and biology*, [Proceedings 1990 Cargèse Nato ASI, Series E: Applied Sciences], Kluwer, Dordrecht, 1990

[81] S. M.Ossadnik, S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, C.-K. Peng, M. Simons and H. E. Stanley, Correlation Approach to Identify Coding Regions in DNA Sequences, *Biophys. J.* **67**, 64–70 (1994)

[82] A. Arneodo, E. Bacry, P. V. Graves and J. F. Muzy, Characterizing Long-Range Correlations in DNA Sequences from Wavelet Analysis, *Phys. Rev. Lett.* **74**, 3293–3296 (1995)

[83] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons and H. E. Stanley, Long-Range Correlations in Nucleotide Sequences, *Nature* **356**, 168–171 (1992)

[84] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley and A. L. Goldberger, Mosaic organization of DNA nucleotides, *Phys. Rev. E* **49**, 1685–1689 (1994)

[85] W. Li and K. Kaneko, Long-range correlations and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence, *Europhys. Lett.* **17**, 655–660 (1992)

[86] G. M. Viswanathan, S. V. Buldyrev, S. Havlin and H. E. Stanley, Quantification of DNA patchiness using long-range correlation measures, *Biophys. J.* **72**, 866–875 (1997)

[87] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons and H. E. Stanley, Generalized Lévy Walk Model for DNA Nucleotide Sequences, *Phys. Rev. E* **47**, 4514–4523 (1993)

[88] J. Klafter, G. Zumofen and M. F. Shlesinger. Lévy Description of Anomalous Diffusion in Dynamical Systems. in M. F. Shlesinger, G. M. Zaslavsky and U. Frisch, eds., *Lévy Flights and Related Topics in Physics. [Proceedings of the International Workshop Held at Nice, France, 27–30 June, 1994]*. Springer-Verlag, Berlin 1995.

[89] P. Bak, C. Tang and K. Wiesenfeld, Self-organised criticality: an explanation of $1/f$ noise, *Phys. Rev. Lett.* **59**, 381–384 (1987)

[90] P. Bak and K. Sneppen, Punctuated equilibrium and criticality in a simple model of evolution, *Phys. Rev. Lett.* **71**, 4083–4086 (1993)

[91] M. Paczuski, S. Maslov and P. Bak, Avalanche dynamics in evolution, growth and depinning models, *Phys. Rev. E* **53**, 414–443 (1996)

[92] B. Jovanovic, S. V. Buldyrev, S. Havlin and H. E. Stanley, Punctuated equilibrium and history-dependent percolation, *Phys. Rev. E* **50**, R2403–R2406 (1994)

[93] E. I. Shakhnovich and A. M. Gutin, Implications of thermodynamics of protein folding for evolution of primary sequences, *Nature* **346**, 773–775 (1990)

[94] S. Karlin and V. Brendel, Patchiness and correlations in DNA sequences, *Science* **259**, 677–680 (1993)

[95] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions with formulas, graphs and mathematical tables. p. 258.*, National Bureau of Standards, Washington, D. C., 1964

[96] G. I. Bell, Evolution of simple sequence repeats, *Computers and Chemistry* **20**, 41–48 (1996)

[97] N. V. Dokholyan, S. V. Buldyrev, S. Havlin and H. E. Stanley, Model of unequal chromosomal crossing over in DNA sequences, *Physica A* **249**, 594–599 (1998)

[98] B. Charlesworth, P. Sniegowski and W. Stephan, The evolutionary dynamics of repetative DNA in eukaryotes, *Nature* **371**, 215–220 (1994)

[99] G. Levinson and G. A. Gutman, Slipped-strand mispairing: a major mechanism for DNA sequence evolution, *Mol. Biol. Evol.* **4**, 203–221 (1987)

[100] C. Schlötterer and D. Tautz, Slippage synthesis of simple sequence DNA, *Nucl. Acid Res.* **20**, 211–215 (1992)

[101] A. Jaworski W. A. Rosche, R. Gellibolian, S. Kang, M. Shimizu, R. P. Bowater RP, R. R. Sinden and R. D. Wells, Mismatch repair in Escherichia coli enhances instability of $(CTG)_n$ triplet repeats from human hereditary diseases, *Proc. Natl. Acad. Sci. U. S. A.* **92**, 11019–11023 (1995)

[102] P. Modrich, Mismatch repair, genetic stability, and cancer, *Science* **266**, 1959–1960 (1994)

[103] A. P. Wolffe, Genetic effect of DNA packaging, *Sci. Am. Sci. Med.* **2**, 68–77 (1995)

[104] G. K. Smith, J. Jie, G. E. Fox and X. L. Gao, DNA CTG triplet repeats involved in dynamic mutations of neurologically related gene-sequences form stable duplexes, *Nucl. Acid Res.* **23**, 4303–4311 (1995)

[105] K. Ohshima, S. Kang and R. D. Wells, *CTG* triplet repeats from human hereditary diseases are dominant genetic expansion products in Escherichia coli, *J. Biol. Chem.* **271**, 1853–1856 (1996)

[106] J. L. Slightom, D. R. Siemieniak, L. C. Sieu, B. F. Koop and L. Hood, Nucleotide sequence analysis of 77.7 kb of the human V β T-cell receptor gene locus: direct primer-walking using cosmid template DNAs, *Genomics* **20**, 149–168 (1994)

[107] R. H. R. Stanley, N. V. Dokholyan, S. V. Buldyrev, S. Havlin and H. E. Stanley, Clumping of Identical Oligonucleotides in Coding and Noncoding DNA Sequences, *J. Biomol. Str. & Dyn.* **submitted** (1999)

[108] P. J. Reynolds, H. E. Stanley and W. Klein, Ghost fields, pair connectedness, and scaling: exact results in one-dimensional percolation, *J. Phys. A: Math. Gen.* **10**, L203–210 (1977)

[109] V. I. Abkevich, A. M. Gutin and E. I. Shakhnovich, Impact of local and non-local interactions on thermodynamics and kinetics of protein folding, *J. Mol. Biol.* **252**, 460–471 (1995)

[110] H. Herzel, E. N. Trifonov, O. Weiss and I. Große, Interpreting correlations in biosequences, *Physica A* **248**, 449–459 (1998)

[111] L. A. Mirny, V. I. Abkevich and E. I. Shakhnovich, Universality and diversity of the protein folding scenarios: a comprehensive analysis with the aid of a lattice model, *Folding & Design* **1**, 103–116 (1996)

[112] E. I. Shakhnovich, V. I. Abkevich and O. Ptitsyn, Conserved residues and the mechanism of protein folding, *Nature* **379**, 96–98 (1996)

[113] N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley and E. I. Shakhnovich, Molecular dynamics studies of folding of a protein-like model, *Folding & Design* **3**, 577–587 (1998)

[114] V. S. Pande, A. Yu. Grosberg and T. Tanaka, Non-Randomness in Protein Sequences: Evidence for a Physically Driven Stage of Evolution?, *Proc. Natl. Acad. Sci. U. S. A.* **91**, 12972 (1994)

[115] M. Karplus and E. I. Shakhnovich. Protein folding: theoretical studies of thermodynamics and dynamics. in T. Creighton, ed., *Protein Folding.* W. H. Freeman and Company, New York 1994.

[116] O. B. Ptitsyn. The molten globule state. in T. Creighton, ed., *Protein Folding.* W. H. Freeman and Company, New York 1994.

[117] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, The Protein Data Bank: a computer-based archival file for macromolecular structures, *J. Mol. Biol.* **112**, 535–542 (1977)

[118] E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle and J. Weng. Protein Data Bank. in F. H. Allen, G. Bergerhoff and R. Sievers, eds., *Crystallographic Databases-Information Content, Software Systems, Scientific Applications*, pp. 107–132. Data Commission of the International Union of Crystallography, Cambridge 1987.

[119] C. Levinthal, Are there pathways for protein folding?, *J. Chim. Phys.* **65**, 44 (1968)

[120] T. Quinn, N. Tweedy, R. Williams, R. Richardson and D. Richardson, Betadoublet: de novo design, synthesis and characterization of a β-sandwich protein, *Proc. Natl. Acad. Sci. U. S. A.* **91**, 8747–8751 (1994)

[121] E. I. Shakhnovich, Protein design: a perspective from simple tractable models, *Folding & Design* **3**, R45–R58 (1998)

[122] P. W. Anderson, The concept of frustration in spin glasses, *J. Less-Common Metals* **62**, 291–294 (1978)

[123] J. D. Bryngelson and P. G. Wolynes, Spin glasses and the statistical mechanics of protein folding, *Proc. Natl. Acad. Sci. U. S. A.* **84**, 7524–7528 (1987)

[124] J. D. Bryngelson and P. G. Wolynes, Intermediates and barrier crossing in a random energy model (with applications to protein folding), *J. Phys. Chem.* **93**, 6902–6915 (1989)

[125] T. Garel and H. Orland, Mean-field model for protein folding, *Europhys. Lett.* **6**, 307–309 (1988)

[126] E. I. Shakhnovich and A. M. Gutin, Formation of unique structure in polypeptide chains: Theoretical investigation with the aid of a replica approach, *Biophys. Chem.* **34**, 187–199 (1989)

[127] B. I. Dahiyat and L. Mayo, De novo protein design: fully automated sequence selection, *Science* **278**, 82–87 (1997)

[128] A. V. Finkelstein, A. Gutin and A. Badretdinov, Why are the same protein folds used to perform different functions?, *Proteins* **23**, 142–149 (1995)

[129] D. S. Riddle, J. V. Santiago, S. T. Bray-Hall, N. Doshi, V. P. Grantcharova, Q. Yi and D. Baker, Functional rapidly folding proteins from simplified amino acid sequences, *Nature Struct. Biol.* **4**, 805–809 (1997)

[130] D. T. Jones, Theoretical approaches to designing novel sequences to fit a given fold, *Curr. Opinion Biotechnol.* **6**, 452–459 (1995)

[131] P. Koehl and M. Delarue, Mean-field minimization methods for biological macromolecules, *Curr. Opinion Struc. Biol.* **6**, 222–226 (1996)

[132] M. P. Allen and D. J. Tildesley, *Computer simulation of liquids*, Clarendon Press, Oxford, 1987

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

[134] E. I. Shakhnovich, Proteins with selected sequences fold into unique native conformation, *Phys. Rev. Lett.* **72**, 3907–3910 (1994)

[135] E. I. Shakhnovich and A. M. Gutin, Engineering of stable and fast folding sequences of model proteins, *Proc. Natl. Acad. Sci. U. S. A.* **90**, 7195–7199 (1993)

[136] E. I. Shakhnovich and A. M. Gutin, A novel approach to design of stable proteins, *Protein Eng.* **6**, 793–800 (1993)

[137] J. Saven and P. Wolynes, Statistical mechanics of the combinatorial synthesis and analysis of folding molecules, *J. Chem. Phys.* **101**, 8375–8389 (1997)

[138] C. B. Anifsen, Principles that govern the folding of the protein chains, *Science* **181**, 223–230 (1973)

[139] N. Gō , Theoretical studies of protein folding, *Ann. Rev. Biophys. Bioeng.* **12**, 183–210 (1983)

[140] E. I. Shakhnovich, Theoretical studies of protein-folding thermodynamics and kinetics, *Curr. Opinion Struc. Biol.* **7**, 29–40 (1997)

[141] H. Taketomi, Y. Ueda and N. Gō , Studies on protein folding, unfolding and fluctuations by computer simulations, *Int. J. Peptide Protein Res.* **7**, 445 (1975)

[142] K. A. Dill, Theory for the folding and stability of globular proteins, *Biochemistry* **24**, 1501–1509 (1985)

[143] H. Li, C. Tang and N. S. Wingreen, Nature of driving force for protein folding: A result from analyzing the statistical potential, *Phys. Rev. Lett.* **79**, 765–768 (1997)

[144] D. K. Klimov and D. Thirumalai, Criterion that determines the foldability of proteins, *Phys. Rev. Lett.* **76**, 4070–4073 (1996)

[145] D. Hoffmann and E.-W. Knapp, Polypeptide folding with off-lattice Monte Carlo dynamics: the method, *Eur. Biophys. J.* **24**, 387–403 (1996)

[146] M. Vásquez, Modeling side-chain conformation, *Curr. Opinion Struc. Biol.* **6**, 217–221 (1996)

[147] C. Micheletti, J. R. Banavar, A. Maritan and F. Seno, Protein structures and optimal folding emerging from a geometrical variational principle, *preprint* **cond-mat/9811246** (1998)

[148] N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley and E. I. Shakhnovich, Identifying the protein folding nucleus using molecular dynamics, *Nature Struct. Biol.* **submitted, cond-mat/9812284** (1999)

[149] S. Miyazawa and R. L. Jernigan, Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation, *Macromolecules* **18**, 534–552 (1985)

[150] K. A. Dill, Dominant forces in protein folding, *Biochemistry* **29**, 7133–7155 (1990)

[151] R. L. Jernigan and I. Bahar, Structure-derived potentials and protein simulations, *Curr. Opinion Struc. Biol.* **6**, 195–209 (1996)

[152] J. C. Martinez, M. T. Pissabarro and L. Serrano, Obligatory steps in protein folding and the conformational diversity of the transition state, *Nature Struct. Biol.* **5**, 721–729 (1998)

[153] J. L. Neira, P. Sevilla, M. Menendez, M. Bruix and M. Rico, Hydrogen exchange in ribonuclease A and ribonuclease S: Evidence for residual structure in the unfolded state under native conditions, *J. Mol. Biol.* **285**, 627–643 (1999)

[154] D. E. Callihan and T. M. Logan, Conformations of peptide fragments from the FK506 binding protein: Comparison with the native and urea-unfolded states, *J. Mol. Biol.* **285**, 2161–2175 (1999)

[155] V. I. Abkevich, A. M. Gutin and E. I. Shakhnovich, Specific nucleus as the transition state for protein folding: evidence from the lattice model, *Biochemistry* **33**, 10026–10036 (1994)

[156] A. M. Gutin, V. I. Abkevich and E. I. Shakhnovich, Evolution-like selection of fast-folding model proteins, *Proc. Natl. Acad. Sci. U. S. A.* **92**, 1282–1286 (1995)

[157] P. L. Privalov, Thermodynamic problems of protein structure, *Ann. Rev. Biophys. Biophys. Chem.* **18**, 47–69 (1989)

[158] A. Baumgartner, ed., *Applications of the Monte-Carlo simulations in statistical physics*, Springer-Verlag, New York, 1987

[159] A. Irbäck and H. Schwarze, Sequence dependence of self-interacting random chains, *J. Phys. A: Math. Gen.* **28**, 2121–2132 (1995)

[160] G. F. Berriz, A. M. Gutin and E. I. Shakhnovich, Cooperativity and stability in a Langevin model of proteinlike folding, *J. Chem. Phys.* **106**, 9276–9285 (1997)

[161] Z. Guo and C. L. Brooks III, Thermodynamics of protein folding: a statistical mechanical study of a small all-β protein, *Biopolymers* **42**, 745–757 (1997)

[162] Y. Zhou, M. Karplus, J. M. Wichert and C. K. Hall, Equilibrium thermodynamics of homopolymers and clusters: molecular dynamics and Monte Carlo simulations of system with square-well interactions, *J. Chem. Phys.* **107**, 10691–10708 (1997)

[163] Y. Zhou and M. Karplus, Folding thermodynamics of a three-helix-bundle protein, *Proc. Natl. Acad. Sci. U. S. A.* **94**, 14429–14432 (1997)

[164] M. Karplus and D. L. Weaver, Protein-folding dynamics, *Nature* **260**, 404–406 (1979)

[165] H. A. Scheraga, Influence of interatomic interactions on the structure and stability of polypeptides and proteins, *Biopolymers* **20**, 1877–1899 (1981)

[166] D. B. Wetlaufer, Nucleation, rapid folding, and globular interchain regions in proteins, *Proc. Natl. Acad. Sci. U. S. A.* **70**, 691–701 (1973)

[167] P. Wolynes, The concept of nucleation, *Folding & Design* **3**, R107 (1998)

[168] D. Thirumalai and D. K. Klimov, Fishing for folding nuclei in lattice models and proteins, *Folding & Design* **3**, R112–R118 (1998)

[169] V. I. Abkevich, A. M. Gutin and E. I. Shakhnovich, Improved design of stable and fast-folding model proteins, *Folding & Design* **1**, 221–230 (1996)

[170] B. J. Alder and T. E. Wainwright, Studies in molecular dynamics. I. General method, *J. Chem. Phys.* **31**, 459–466 (1959)

[171] A. Yu. Grosberg and A. R. Khokhlov, *Giant molecules*, Academic Press, Boston, 1997

[172] D. C. Rapaport, *The art of molecular dynamics simulation*, Cambridge University Press, Cambridge, 1997

[173] P.-G. de Gennes, ed., *Scaling concepts in polymer physics*, Cornell University Press, Ithaca, 1979

[174] M. Doi, ed., *Introduction to polymer physics*, Clarendon Press, Oxford, 1996

[175] L. Landau and E. M. Lifshitz, *Statistical mechanics*, Pergamon Press, Oxford, 1980

[176] W. Kabsch, A discussion of the solution for the best rotation to relate two sets of vectors, *Acta Cryst.* **A34**, 827–828 (1978)

[177] V. I. Abkevich, A. M. Gutin and E. I. Shakhnovich, Domains in folding of model proteins, *Protein Science* **4**, 1167–1177 (1995)

[178] T. Lazaridis and M. Karplus, "New view" of protein folding reconciled with the old through multiple unfolding simulations, *Science* **278**, 1928–1931 (1997)

[179] T. Y. Tsong and R. L. Baldwin, Effects of solvent viscosity and different guanidine salts on the kinetics of ribonuclease A chain folding, *Biopolymers* **17**, 1669–1678 (1978)

[180] Y. Bai, T. R. Sosnick, L. Mayne and S. W. Englander, Protein folding intermediates: native-state hydrogen exchange, *Science* **269**, 192–197 (1995)

[181] E. Lacroix, M. Bruix, E. López-Hernández, L. Serrano and M. Rico, Amide hydrogen exchange and internal dynamics the chemotactic protein CheY from *Escherichia coli*, *J. Mol. Biol.* **271**, 472–487 (1997)

[182] C. L. Brooks III, Characterization of "native" apolyoglobin by molecular dynamics simulation, *J. Mol. Biol.* **227**, 375–380 (1992)

[183] V. Daggett and M. Levitt, A model of the molten globule state from molecular dynamics simulations, *Proc. Natl. Acad. Sci. U. S. A.* **89**, 5142–5146 (1992)

[184] F. B. Sheinerman and C. L. Brooks III, A molecular dynamics simulation study of segment B1 of protein G, *Proteins: Struc. Func. Genet.* **29**, 193–202 (1997)

[185] S. E. Jackson, How do small single-domain proteins fold?, *Folding & Design* **3**, R81–R91 (1998)

[186] V. G. Karpov and D. W. Oxtoby, Nucleation in disordered systems, *Phys. Rev. B* **54**, 9734–9745 (1996)

[187] E. M. Lifshitz and L. P. Pitaevskii, *Physical kinetics*, Pergamon Press, Oxford, 1981

[188] A. R. Fersht, Nucleation mechanisms in protein folding, *Curr. Opinion Struc. Biol.* **7**, 3–9 (1997)

[189] V. S. Pande, A. Yu. Grosberg, D. S. Rokshar and T. Tanaka, Pathways for protein folding: is a new view needed?, *Curr. Opinion Struc. Biol.* **8**, 68–79 (1998)

[190] E. I. Shakhnovich, Folding nucleus: specific of multiple? Insights from lattice models and experiments, *Folding & Design* **3**, R108–R111 (1998)

[191] L. A. Mirny, V. I. Abkevich and E. I. Shakhnovich, How evolution makes proteins fold quickly, *Proc. Natl. Acad. Sci. U. S. A.* **95**, 4976–4981 (1998)

[192] A. M. Gutin, V. I. Abkevich and E. I. Shakhnovich, A protein engineering analysis of the transition state for protein folding: simulation in the lattice model, *Folding & Design* **3**, 183–194 (1998)

[193] D. K. Klimov and D. Thirumalai, Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism, *J. Mol. Biol.* **282**, 459–480 (1998)

[194] L. S. Itzhaki, D. E. Otzen and A. R. Fersht, The structure of the transition-state for folding of chymotrypsin inhibitor-2 analyzed by protein engineering methods — evidence for a nucleation-condensation mechanism for protein-folding, *J. Mol. Biol.* **254**, 260–288 (1995)

[195] V. P. Grantchanova, D. S. Riddle, J. V. Santiago and D. Baker, Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain, *Nature Struct. Biol.* **5**, 714–720 (1998)

[196] D. Perl, C. Welker, T. Schindler, K. Shroder, M. A. Marahiel, R. Jaenicke and F. X. Schmid, Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins, *Nature Struct. Biol.* **5**, 229–235 (1998)

[197] I. J. Benjamin and D. R. McMillan, Stress (heat shock) proteins: molecular chaperones in cardiovascular biology and disease, *Circ. Res.* **83**, 117–132 (1998)

[198] R. Mendelson and E. P. Morris, The structure of the acto-myosin subfragment 1 complex: results of searches using data from electron microscopy and x-ray crystallography, *Proc. Natl. Acad. Sci. U. S. A.* **94**, 8533–8538 (1997)

[199] L. Tskhovrebova, J. Trinick, J. A. Sleep and R. M. Simmons, Elasticity and unfolding of single molecules of the giant muscle protein titin, *Nature* **387**, 308–312 (1997)

[200] M. S. Z. Kellermayer, S. B. Smith, H. L. Granzier and C. Bustamante, Folding-unfolding transitions in single titin molecules characterized with laser tweezers, *Science* **276**, 1112–1116 (1997)

[201] M. E. Chicurel, C. S. Chen and D. E. Ingber, Cellular control lies in the balance of forces, *Curr. Opin. Cell Biol.* **10**, 232–239 (1998)

[202] C. D. Dickinson, B. Veerapandian, X.-P. Dai, R. C. Hamlin, N. h. Xuong, E. Ruoslahti and K. R. Ely, Crystal structure of the tenth type III cell adhesion module of human fibronectin, *J. Mol. Biol.* **236**, 1079–1092 (1994)

[203] B. A. Keyt, L. T. Berleau, H. V. Nguyen, H. Chen, H. Heinsohn, R. Vandlen and N. Ferrara, The carboxyl-terminal domain $(111 - 165)$ of vascular endothelial growth factor is critical for its mitogenic potency, *J. Mol. Biol.* **271**, 7788–7795 (1996)

[204] T. Simmons, Y. M. Newhouse, K. S. Arnold, T. L. Innerarity and K. H. Weisgraber, Human low density lipoprotein receptor fragment. Successful refolding of a functionally active ligand-binding domain produced in Escherichia coli, *J. Biol. Chem.* **272**, 25531–25536 (1996)

Curriculum Vitæ
Nikolay V. Dokholyan

Boston University, Physics Department
Center For Polymer Studies
590 Commonwealth Avenue
Boston, Massachusetts 02215 USA

Telephone: 617/353-8936
Facsimile: 617/353-9393 or 617/353-3783
E-mail: dokh@polymer.bu.edu
WWW: http://polymer.bu.edu/dokh

EDUCATION

- Ph.D., 1999 - Physics; Boston University, Boston, MA
- M.S. honors, 1994 - Physics; Moscow Institute of Physics and Technology, Moscow, Russia
- B.S. honors, 1992 - Physics; Moscow Institute of Physics and Technology, Moscow, Russia

EMPLOYMENT

- Research Assistant, Physics Department, Boston University. Spring 1997 - present.
- Teaching Assistant, Physics Department, Boston University. Fall 1994 - Spring 1997.
- Research Assistant, Physics Department, Boston University. Summer 1994.
- Research Assistant, Institute for High Energy Physics, Russia. Fall 1991-Spring 1994.
- Teacher, Moscow Institute of Physics and Technology, Special Physics and Mathematics High School. Fall 1988 - Spring 1990.

HONORS AND GRANTS AWARDED

- Recipient of Honorary Stipend, Moscow Institute of Physics and Technology; 1990-1994.
- "Red Diploma" 1994 (in the former USSR "Red Diploma" is awarded for outstanding achievements).
- NSF Student Award for STATPHYS XIX, 1995.
- British Petroleum, $60,000. (I was a principal author of this proposal.) 1998.
- NIH molecular biophysics predoctoral traineeship (GM08291-09). 1998 (current).
- NSF Student Award for STATPHYS XX, 1998.
- British Petroleum Amoco, $60,000. 1999.

PROGRAMMING EXPERIENCE

- **Platforms** - VAX, SGI Indy/Indigo/Onyx II, IBM RS6000, DEC MIPS/Alpha, PC/ PowerPC.
- **OS** - DOS, Windows 3.x/95/NT, MacOS, VAX VMS, UNIX, LINUX.
- **Languages** - PC 8086/8088 Assembly, C/C++, Fortran/Fortran90, OpenGL, Pascal/ Delphi, Basic, Java/JavaScript, ActiveX, HTML, Perl, Awk, Inventor, VRML, Mathematica, FORM.

CONFERENCE PRESENTATIONS

- 1993 - Talk presented at Physics Institute, Tbilisi, Georgia - Single Top Quark Production and V-tb CKM Matrix Element measurement in High Energy e+e- Collisions.
- 1995 - Poster presented at Statphys XIX, Xiamen, China - Distribution of Repeats in DNA.
- 1995, 1996 - Patterns in Nature '95, and '96. Two-week summer Institutes to introduce teachers to the use of advanced technology materials, Boston University. [with colleagues]

- 1996 - Poster presented at Enrico Fermi Summer School, Varena, Italy - Mechanism of Expansion of Simple Repeats in DNA.
- 1997 - Poster presented at Conference on Complex Systems, Bar Ilan University, Israel - Model of Unequal Chromosomal Crossing Over in DNA Sequences.
- 1997 - Talk presented at Conference on Statistical Mechanics, Rutgers University - Distribution of Dimeric Repeats in DNA.
- 1998 - Talk presented at Biophysics Seminar, Boston University - Folding of 'Go' protein.
- 1998 - Talk presented at Biophysics Seminar, Harvard University - Molecular dynamics simulations of protein models.
- 1998 - Talk presented at Conference on Protein folding and structure prediction, Torino, Italy - Molecular dynamics simulations of protein-like model.
- 1998 - Poster presented at Conference on Percolation and disordered systems: Theory and applications, Gießen, Germany - Scaling of the distribution of shortest paths in percolation.

ARTICLES ON MY RESEARCH:

- T. McNeil, Bostonia, *"Percolating Texas Tea"*, p. 39, December 1998.

REFERENCES

- H. Eugene Stanley: University Professor of Physics, Professor of Physiology, Director of Center for Polymer Studies; Department of Physics, Boston University, 590 Commonwealth Avenue, Boston, MA 02215. **Tel:** (617) 353 2617, **FAX:** (617) 353 3783,

  Email: hes@bu.edu.

- Shyamsunder Erramilli: Professor of Physics, Department of Physics; Boston University, 590 Commonwealth Avenue, Boston, MA 02215. **Tel:** (617) 353 6114, **FAX:** (617) 353 9393,

  Email: shyam@buphy.bu.edu.

- Peter R. King: Professor of Physics, Department of Engineering, Cambridge University, Cambridge, UK; Senior Researcher, BP Exploration Operating Company Ltd., Sunbury-on-Thames, Middx., TW16 7LN, UK. **Tel:** +44 (1932) 763492, **FAX:** +44 (1932) 764639,

  Email: kingpr@eu1.bp.com.

- Shlomo Havlin: Professor of Physics, President of Israeli Physical Society, Director of Minerva Center for Fractals; Gonda-Goldschmied Center and Department of Physics, Bar-Ilan University, Ramat Gan, 52900 Israel. **Tel:** +972-3-5318436, **FAX:** 972-3-5353298,

  Email: havlin@ophir.ph.biu.ac.il

List of Publications

I. High Energy Physics:

[1] N. V. Dokholyan & G. V. Jikia, preprint: IHEP 93111, Protvino, Russia (1993).

[2] N. V. Dokholyan & G. V. Jikia, "Single Top Quark Production and V-tb CKM Matrix Element measurement in High Energy e+e- Collisions." *Phys. Lett. B* **336**: 251–256 (1994).

[3] N. V. Dokholyan, "Cabbibo - Kabayashi - Maskawa mixing matrix in the case of four generations." *preprint* (1994).

II. DNA:

[4] N. V. Dokholyan, S. V. Buldyrev, S. Havlin & H. E. Stanley, "Distribution of base pair repeats in coding and noncoding DNA sequences." *Phys. Rev. Lett.* **79:** 5182–5185 (1997).

[5] N. V. Dokholyan, S. V. Buldyrev, S. Havlin & H. E. Stanley, "Model of unequal chromosomal crossing over in DNA sequences." *Physica A* **249:** 594–599 (1998).

[6] S. V. Buldyrev, N. V. Dokholyan, A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley & G. M. Viswanathan, "Analysis of DNA Sequences Using Methods of Statistical Physics." *Physica A* **249:** 430–438 (1998).

[7] N. V. Dokholyan, S. V. Buldyrev, S. Havlin & H. E. Stanley, "Non-randomness of dimeric tandem repeats in noncoding DNA sequences." *J. Theor. Biol.* submitted (1998).

[8] R. H. R. Stanley, N. V. Dokholyan, S. V. Buldyrev, S. Havlin & H. E. Stanley, "Clumping of Identical Oligonucleotides in Coding and Noncoding DNA Sequences." *J. Biomol. Struct. & Dyn.* submitted (1998).

III. Protein folding and design:

[9] N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley & E. I. Shakhnovich, "Molecular dynamics studies of folding of a protein-like model." *Folding & Design*, **3:** 577–587 (1998).

[10] N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley & E. I. Shakhnovich, "The protein folding nucleus: a molecular dynamics observation." *Nature Struct. Biol.* submitted (1998). (preprint **cond-mat/9812284**).

[11] N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley & E. I. Shakhnovich, "The molecular dynamics studies of folding of BPTI protein: the role of disulfide bonds." in preparation (1999).

IV. Random systems:

[12] N. V. Dokholyan, Y. Lee, S. V. Buldyrev, S. Havlin, H. E. Stanley & P. King, "Scaling of the distribution of shortest paths in percolation." *J. Stat. Phys.* **93:** 603-613 (1998).

[13] P. King, J. S. Adrade Jr., S. V. Buldyrev, N. V. Dokholyan, S. Havlin, Y. Lee & H. E. Stanley, "Predicting oil recovery using precolation." *Physica A* in press (1998).

[14] N. V. Dokholyan, S. V. Buldyrev, S. Havlin, P. King, Y. Lee, & H. E. Stanley, "Distribution of shortest paths in percolation." *Physica A* in press (1998).

[15] Y. Lee, J. S. Andrade Jr., S. V. Buldyrev, N. V. Dokholyan, S. Havlin, P. R. King, G. Paul, & H. E. Stanley, "Traveling time and traveling length for flow in porous media." *PRL* submitted (1999).

V. Cell motility:

[16] M. Hong, A. Jeung, N. V. Dokholyan, T. I. Smith, H. A. Schwettman, P. Huie & S. Erramilli, "Imaging single living cells with a scanning near-field infrared microscope based on a free electron laser." *Nucl. Instr. & Methods in Physics Research* **B144**: 246–255, (1998).

List of Abstracts

[1] N. V. Dokholyan, "Power law distribution of random repeats in DNA sequences." STATPHYS XIX, The 19th IUPAP International Conference on Statistical Physics, Xiamen, China, July 31 – August 4, 1995. H. Bailin (Ed.) World Scientific, London (1995).

[2] N. V. Dokholyan, S. V. Buldyrev, S. Havlin, and H. E. Stanley, "Power law distribution of dimeric tandem repeats in DNA sequences." Proceedings of the International School of Physics "Enrico Fermi", Course CXXXIV. Varenna on Lake Como, Villa Monastero, Italy, July 9 – 19, (1996). *"The Physics of Complex Systems."* IOS Press, Oxford (1997).

[3] N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, "Discrete molecular dynamics studies of folding of proteinlike model." STATPHYS XX, The 20th IUPAP International Conference on Statistical Physics, Paris, France, July 20 – 24, 1998. A. Gervois, M. Gingold, D. Iagolnitzer (Eds.) UNESCO, Sorbone.