Tel Aviv University
Raymond and Beverly Sackler Faculty of Exact Sciences
School of Chemistry

Proteins: Unraveling Universality in a Realm of Specificity

by

Shlomi Reuveni

under the supervision of

Prof. Joseph Klafter

A thesis submitted for the
degree of M.Sc in Chemistry

Acknowledgments

This thesis summarizes four wonderful years I spent at Tel Aviv University. Four years ago, I was a first year student with vague ideas about interdisciplinary science. I wanted to explore the magnificent realm of biology through the eyes of a physicists/mathematician but didn't quite know how. Today, I have certainly got what I had hoped for, the taste of genuine interdisciplinary work: proteins and fractals, theoretical physics side by side with bioinformatics and more. For me this thesis has a little bit of everything.

Looking back I would like to thank the people who helped me follow my dream. First and foremost, I had the pleasure of being supervised by professor Yossi Klafter. Yossi took me under his wings, teaching me what science, research and independence are all about. Together we had numerous talks regarding this project and other issues. As one topic led to another, it was almost forgotten I started working on this project almost by accident. I would also like to thank Rony Granek that worked on this project together with me and Yossi. His original view of things, his willingness to help and the constructive criticism he brought into the triad were of utmost importance. Last but not least, I owe a great debt to the interdisciplinary program for outstanding students named after Adi Lautman and to its wonderful staff. Being a student of the interdisciplinary program was a unique experience, words are too dull to describe. Studying freely and without borders, wrapped up with endless support and encouragement is all one could ask for.

Contents

1 Abstract

Two seemingly conflicting properties of native proteins, such as enzymes and antibodies, are known to coexist. While proteins need to keep their specific native fold structure thermally stable, the native fold displays the ability to perform large amplitude motions that allow proper function[1, 2]. This conflict cannot be bridged by compact objects which are characterized by small amplitude vibrations and by a Debye density of low frequency modes. Recently, however, it became clear that proteins can be described as fractals; namely, geometrical objects that possess self similarity[3, 4, 5]. Adopting the fractal point of view to proteins makes it possible to describe within the same framework essential information regarding topology and dynamics[6, 7] using three parameters: the number of amino acids along the protein backbone N, the spectral dimension d_s and the fractal dimension d_f. The fractal character implies large amplitude vibrations of the protein that could have led to unfolding. We show that by selecting a thermodynamic state that is "close" to the edge of stability against unfolding, nature has solved the thermostability conflict. Starting off from a thermal marginal stability criterion we reach a universal equation describing the relation between the spectral and fractal dimensions of a protein and the number of amino acids. This equation is obeyed by a large class of proteins regardless of their source or function. We suggest that deviations from this "equation of state" for protein topology may render a protein unfolded. Nature's solution might be incorporated when planning biologically inspired catalysts.

Based on a generalization[8] of the Peierls instability criterion[9], we derive a general relation between the spectral dimension d_s, the fractal dimension d_f and the number of amino acids along the protein backbone:

$$\frac{2}{d_s} + \frac{1}{d_f} = 1 + \frac{b}{ln(N)} \ . \tag{1}$$

The spectral dimension d_s governs the density of low frequency normal modes of a fractal/protein. More precisely, denoting the density of modes $g(\omega)$, the scaling relation $g(\omega) \sim \omega^{d_s - 1}$ holds for low frequencies. Describing the mass fractal dimension d_f is most convenient using a three dimensional example. Draw a sphere of radius r enclosing some lattice points in space and calculate their mass $M(r)$, increase r and calculate again. Do this several times and if $M(r)$ scales as r^{d_f} the exponent d_f is called the fractal dimension. For a regular 3D lattice both d_s and d_f coincide with the usual dimension of 3. For proteins however, it is usually found that $d_s < 2$ and $2 < d_f < 3$, leading to an excess of low frequency modes and a more sparse fill of space[3, 4].

In order to test the validity of equation (1), we calculated the spectral and fractal dimensions for a data set of 543 proteins. Calculations were preformed on known protein structures, all structures were downloaded from the Protein Data Bank (PDB)[10]. The proteins that were chosen differ in function and/or source organism and represent a wide length scale ranging from 100 to 3000

residues. Statistical analysis of the data gathered reveals satisfying agreement with equation (1).

2 Thesis Outline

Due to the interdisciplinary nature of the work, the theoretical background presented in chapters 3 - 5 and 7, engulfs several presumably unrelated topics that merge together in later chapters. Striving towards a self contained document I couldn't avoid a brief introduction on proteins, fractals, the Gaussian Network Model (GNM) and thermal stability. One who is familiar with these topics can skip the relevant chapters without further due. One who chooses to read them must keep in mind that the theoretical background chapters are concise and mainly deal with topics that are relevant to later chapters. One must also keep in mind that these chapters don't contain any original material and are strictly a summary of the discussed topics. While some paragraphs were completely written by me in order to present things from the point of view I found appropriate, other paragraphs were selected from various sources and put together with only minor changes.

Chapters six eight and nine, may be considered the heart of this thesis. They describe our original contribution to the subject: From building a data base containing the spectral and fractal dimensions of 543 proteins through the derivation and validation of relation (1) and suggestions for further research. The more technical part of the research, dealing with calculation techniques of the spectral and fractal dimensions is also described in these chapters. The following chapters are actually appendixes. The first two are computer programs written in MATLAB. The first program calculates the fractal dimension and radius of gyration of a protein, given its PDB code. The second program calculates the eigenfrequencies of a protein within the Gaussian Network Model framework. The third appendix describes an alternative route to equation (1). The forth appendix is a raw data table that displays fractal and spectral dimension calculations preformed on 543 proteins. The computer code and data may be of help for researchers that choose to repeat the work done here and/or continue it.

3 Proteins

The word protein comes from the Greek word "prota", meaning "of primary importance". These molecules were first described and named by the Swedish chemist Jöns Jakob Berzelius in 1838. However, proteins central role in living organisms was not fully appreciated until 1926, when James B. Sumner showed that the enzyme urease was a protein[11]. Proteins are of prime importance in organisms and participate in numerous processes within cells. Many proteins are enzymes that catalyze biochemical reactions, and are vital to metabolism. Proteins also have structural or mechanical functions, such as actin and myosin in muscles, and the proteins in the cytoskeleton, which forms a system of scaffolding that maintains the cell shape. Other proteins are important in cell signaling, immune responses, cell adhesion, and the cell cycle.

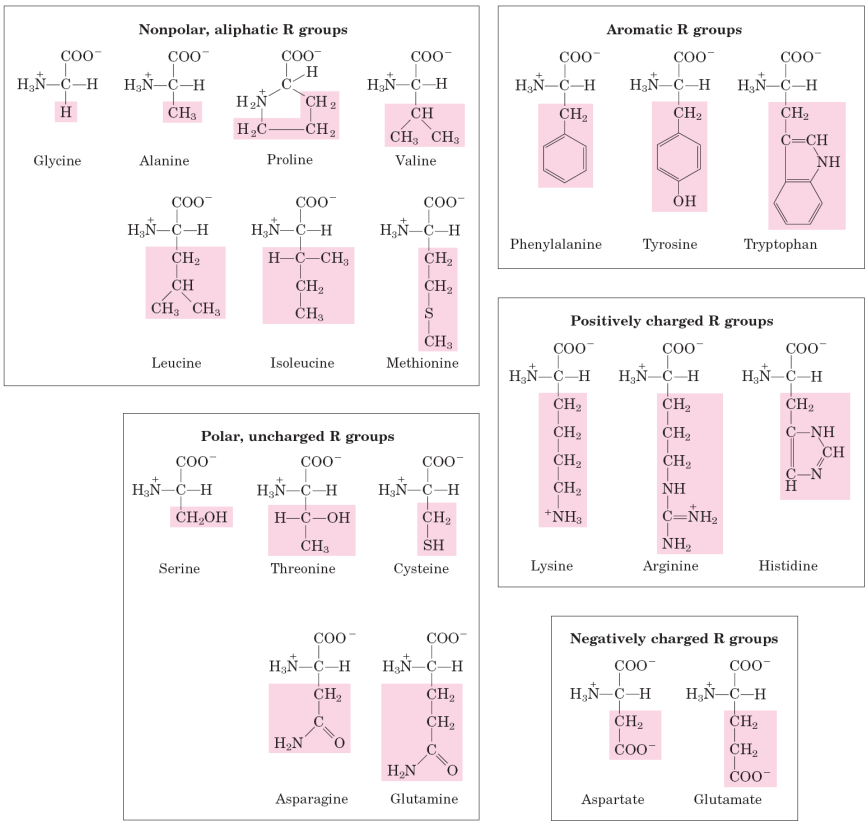


Figure 1: The 20 common amino acids of proteins. The structural formulae show the state of ionization that would predominate at pH 7.0. The unshaded portions are those common to all the amino acids. The portions shaded in red are the R groups, unique chemical groups that characterize the 20 common amino acids.

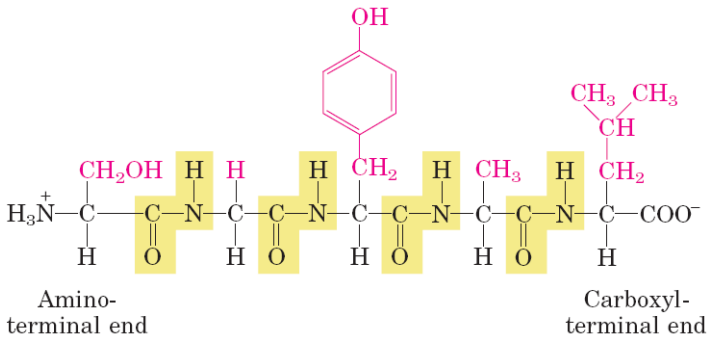


Figure 2: The pentapeptide serylglycyltyrosylalanylleucine, or Ser–Gly–Tyr–Ala–Leu. Peptides are named beginning with the amino terminal residue, which by convention is placed at the left. The peptide bonds are shaded in yellow; the R groups are in red.

3.1 Biochemistry

Proteins are large organic compounds, made of amino acids. The standard twenty amino acids that comprise proteins are shown in figure (1)[12]. Proteins are arranged in a linear chain and joined together between the carboxyl atom of one amino acid and the amine nitrogen of another. The amino acids in a protein are linked by peptide bonds formed in a dehydration reaction. Once linked in the protein chain, an individual amino acid is called a residue and the linked series of carbon, nitrogen, and oxygen atoms are known as the protein backbone. Due to the chemical structure of the individual amino acids, the protein chain has directionality. The end of the protein with a free carboxyl group is known as the C-terminus or carboxyl terminus, while the end with a free amino group is known as the N-terminus or amino terminus. An example of a short protein, which is actually a polypeptide, is depicted in figure (2)[12].

3.2 Biosynthesis

The sequence of amino acids in a protein is defined by a gene and encoded in the genetic code. Genes are actually segments of DNA that are first transcribed into pre-messenger RNA and then translated into proteins. This idea is part of the central dogma of molecular biology and is depicted in figure (3)[12]. The unique amino acid sequence is specified by the nucleotide sequence of the gene encoding the protein. The genetic code is a set of three-nucleotide sets called codons and each three-nucleotide combination stands for an amino acid; for instance CAG stands for Glutamine and CCC stands for Proline. Because DNA contains four nucleotides, the total number of possible codons is 64; hence, there is some redundancy in the genetic code and some amino acids are specified by more than one codon. The amino acid sequence of a protein is also referred to as the primary structure of the protein. The first protein to be sequenced was insulin,

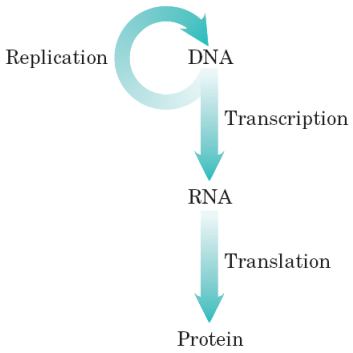


Figure 3: The central dogma of molecular biology, showing the general pathways of information flow via replication, transcription, and translation. The term "dogma" is a misnomer. Introduced by Francis Crick at a time when little evidence supported these ideas, the dogma has become a well-established principle.

by Frederick Sanger, who won the Nobel Prize for this achievement in 1958.

3.3 Structure and Function

Protein are known to naturally fold into native three dimensional structures, the folding of proteins into this native form is usually favored under physiological conditions. The native protein structure usually has functional relevance, a protein that has not folded into its native structure may not function properly. Biochemists often refer to four distinct aspects of a protein's structure:

- Primary structure: the amino acid sequence.
- Secondary structure: regularly repeating local structures stabilized by hydrogen bonds. The most common examples are the alpha helix and beta sheet, also see figure (4). Because secondary structures are local, many regions of different secondary structure can be present in the same protein molecule.
- Tertiary structure: the overall shape of a single protein molecule; the spatial relationship of the secondary structures to one another. Tertiary structure is generally stabilized by non local interactions, most commonly the formation of a hydrophobic core, but also through salt bridges, hydrogen bonds, disulphide bonds, and even post-translational modifications.
- Quaternary structure: the shape or structure that results from the interaction of more than one protein molecule, usually called protein subunits in this context, which function as part of the larger assembly or protein complex.

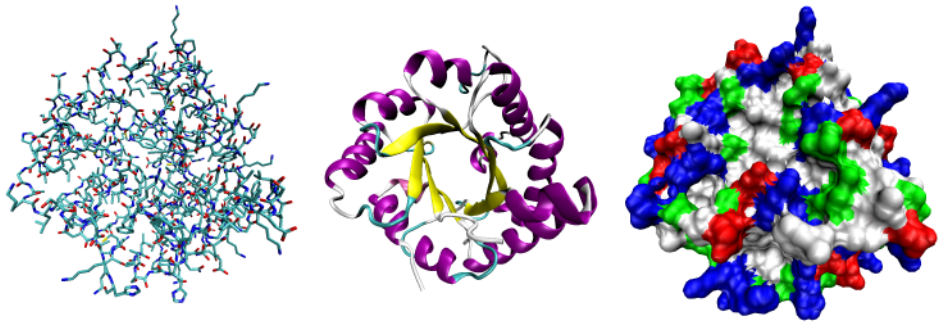


Figure 4: Three views of one monomer of the protein triose phosphate isomerase (PDB code 1TIM). Left, an all-atom view colored by atom type. Middle, a cartoon view colored by secondary structure. Alpha helices are colored magenta, beta sheets are colored yellow. Right, a solvent-accessible surface view colored by residue type (acidic residues red, basic residues blue, polar residues green, non polar residues white).

Proteins are not entirely rigid molecules. In addition to these levels of structure, proteins may shift between several related structures while they perform their biological function. In the context of these functional rearrangements, these tertiary or quaternary structures are usually referred to as "conformations," and transitions between them are called conformational changes. Such changes are often induced by the binding of a substrate molecule to an enzyme's active site, or the physical region of the protein that participates in chemical catalysis. In solution all proteins also undergo variation in structure through thermal vibration and the collision with other molecules.

The native conformation is lost, as a result of denaturation, at extreme pH values, at high temperatures, and in the presence of organic solvents, detergents, and other denaturing substances, such as urea. The fact that a denatured protein can spontaneously return to its native conformation was demonstrated for the first time with ribonuclease, a digestive enzyme. An illustration of the folding and denaturation of rebunuclease is shown in figure (5)[13]. Discovering the native structure of a protein can provide important clues about how a protein performs its function. Common experimental methods of structure determination include X-ray crystallography and NMR spectroscopy, both of which can produce information at the atomic resolution. The first protein structures to be solved included hemoglobin and myoglobin, by Max Perutz and Sir John Cowdery Kendrew, respectively, in 1958[14, 15]. Both proteins three-dimensional structures were first determined by X-ray diffraction analysis. The structures of myoglobin and hemoglobin won the 1962 Nobel Prize in Chemistry for their discoverers. Today there are more than 45,000 known protein structures in the

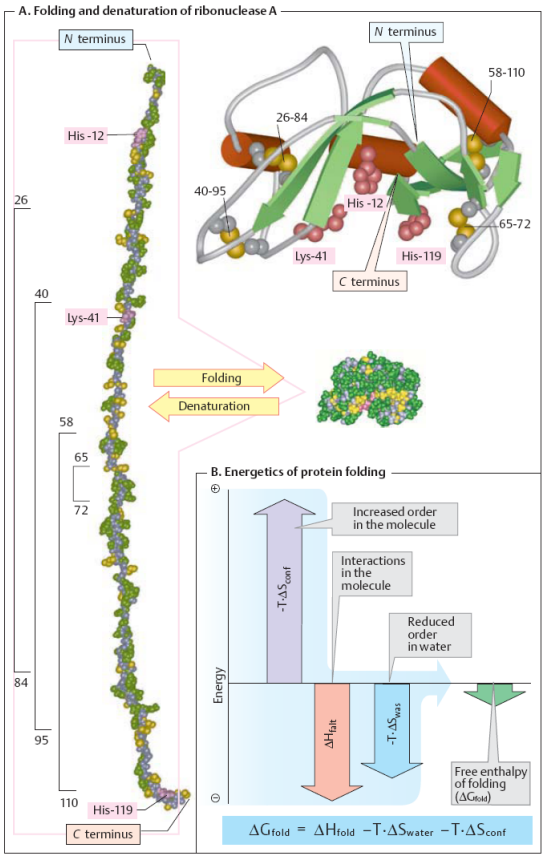


Figure 5: In the native form (top right), there are extensive pleated sheet structures and three α helices. The eight cysteine residues of the protein are forming four disulphide bonds. Residues His-12, Lys-41 and His-119 (pink) are particularly important for catalysis. Together with additional amino acids, they form the enzyme's active center. The disulphide bonds can be reductively cleaved by thiols (e.g., mercaptoethanol, HO-CH2-CH2-SH). If urea at a high concentration is also added, the protein unfolds completely. In this form (left), it is up to 35 nm long. Polar (green) and apolar (yellow) side chains are distributed randomly. The denatured enzyme is completely inactive, because the catalytically important amino acids (pink) are too far away from each other to be able to interact with each other and with the substrate. When urea and thiols are removed, secondary and tertiary structures develop again spontaneously. The cysteine residues thus return to a sufficiently close spatial vicinity that disulphide bonds can once again form under the oxidative effect of atmospheric oxygen. The active center also reestablishes itself. In the folded state, the apolar side chains (yellow) predominate in the interior of the protein, while the polar residues are mainly found on the surface. This distribution is due to the "hydrophobic effect", and it makes a vital contribution to the stability of the native conformation.

Protein Data Bank (PDB)[10], figure (4) shows computer generated models for one of them.

4 Fractals

This chapter will provide a short and informal introduction to fractals. The main aim is to provide the reader with a general idea about fractals and fractal dimensions. A well known fractal, the Sierpinski gasket, will serve as an example.

4.1 Introduction

- Fractal - "A rough or fragmented geometric shape that can be subdivided in parts, each of which is (at least approximately) a reduced size copy of the whole" Benoît Mandelbrot[16].

The term fractal was coined by Benoît Mandelbrot in 1975 and was derived from the Latin word fractus, meaning "broken" or "fractured". A fractal as a geometric object generally has the following features[17]:

1. Simple and recursive definition.
2. Fine structure at arbitrarily small scales.
3. Self similarity, at least approximately or stochastically.

These definitions may be rather puzzling encountered for the first time. Instead of trying to explain their meaning in the general case, I will discuss them through a simple example.

4.2 The Sierpinski Gasket

The Sierpinski gasket, also called the Sierpinski triangle, is a fractal, named after Waclaw Sierpinski who described it in 1915. The Sierpinski gasket is one of the basic examples of self-similar sets. Figure (6) illustrates the recursive construction of the Sierpinski gasket. Following the steps described in the caption

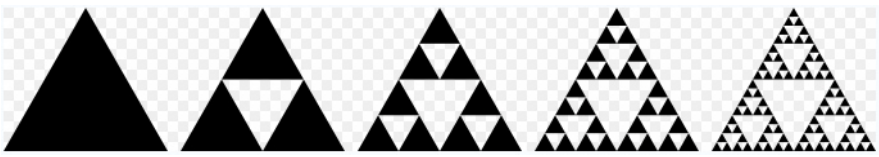


Figure 6: One can construct the Sierpinski gasket recursively as follows. Start with any triangle in a plane, the canonical Sierpinski gasket uses an equilateral triangle with a base parallel to the horizontal axis (left image). Shrink the triangle by half, make two copies, and position the three shrunken triangles so that each triangle touches the two other triangles at a corner (second image from the left). Repeat the second step with each of the smaller triangles an infinite number of times (third image from the left and so on).

of figure (6) an infinite number of times, it is clear that the Sierpinski gasket exhibits fine structure at arbitrarily small scales. The Sierpinski gasket is clearly self similar since it is constructed from three smaller Sierpinski gaskets that are exact miniature versions of the original gasket. The three smaller Sierpinski gaskets are in turn constructed from three smaller Sierpinski gaskets each and so on.

4.3 The Fractal Dimension

Fractals as I shall exemplify shortly, push the intuitive notion of dimension beyond its naive use. Consider for example an equilateral triangle, which we often think of as two dimensional, with side a. The area of such a triangle is $S = \frac{\sqrt{3}a^2}{4}$, the length of its perimeter is $L = 3a$. The properties of non zero area and finite perimeter length appear very natural. Indeed, many two dimensional geometrical objects such as the disk and square exhibit them, the Sierpinski gasket on the other hand does not.

Recall the construction of the Sierpinski gasket and assume we start with an equilateral triangle of side a. The area of the geometrical object obtained after the n-th recursive stage is $S_n = \frac{\sqrt{3}a^2}{4} \cdot \left[\frac{3}{4}\right]^n$. The area of the Sierpinski gasket is the limit of S_n as n tends to infinity, we hence conclude that this area is zero ! The perimeter length of the geometrical object obtained after the n-th recursive stage is $L_n = 3a \cdot \left[\frac{3}{2}\right]^n$. The perimeter length of the Sierpinski gasket is the limit of L_n as n tends to infinity, we hence conclude that this length is infinite ! What is the dimensionality of the Sierpinski gasket then ? Is it a two dimensional object with zero area or a bounded one dimensional object with infinite length ? The answer to this question depends on the definition of the ambiguous term: "dimension".

Suppose that a fractal consists of N identical parts that are similar to the entire fractal with a scale factor of L, is there a relation between N and L ? Figure (7) shows that for the Sierpinski gasket such a relation exists, indeed for this fractal object: $N = L^{ln(3)/ln(2)}$. Similarly for a general fractal, we define the fractal dimension d_f to be the exponent of L in the power law relation between N and L. From this definition it follows that the fractal dimension of the Sierpinski gasket is $d_f = \frac{ln(3)}{ln(2)} \simeq 1.585$. For a regular two dimensional object $d_f = 2$, as the example in the caption of figure (7) illustrates, which coincides with the regular dimension. Thedefinition of the fractal dimension may be generalized as follows: In a two dimensional example, draw a circle of radius r enclosing some lattice points in space and calculate the number of points enclosed by this circle $n(r)$, increase r and calculate again. Do this several times and if $n(r)$ scales as r^{d_f} the exponent d_f is called the fractal dimension. For a regular lattice d_f coincide with the usual dimension of 2 in our example. The fractal dimension of the Sierpinski gasket is independent of the way we calculate it and remains $\frac{ln(3)}{ln(2)}$.

Figure 7: Starting with the small Sierpinski gasket on the left and treating it as a basic unit, N denotes the number of basic units required to construct each of the gaskets above. L denotes the scale factor between the side of the basic building block and the side of the gaskets above. Noting that in the n-th stage $N = 3^n$ and $L = 2^n$, one may verify that $N = L^{ln(3)/ln(2)}$ by taking the natural logarithm of both sides. Note that if we would have done so with a regular equilateral triangle instead of a Sierpinski gasket, we would come to the conclusion that in the n-th stage $N = 4^n$ and $L = 2^n$. In this case it is clear that $N = L^2$.

4.4 The Spectral Dimension

Consider a general elastic network of masses and harmonic springs, it is well known that one of the characteristics of such a network is a set of normal modes and a corresponding set of eigenfrequencies. A normal mode of an oscillating system is a pattern of motion in which all parts of the system move sinusoidally with the same frequency. The frequencies of the normal modes of a system are known as its natural frequencies or eigenfrequencies. Figure (8) describes two different elastic networks of masses and springs.

The spectral dimension d_s governs the density of low frequency normal modes on a fractal. More precisely, denoting the density of modes with frequency ω : $g(\omega)$, the scaling relation $g(\omega) \sim \omega^{d_s - 1}$ holds for low frequencies[1]. The spectral dimension d_s is hence a quantity directly related to the vibrational dynamics of a fractal. It is well known that for an infinite regular square lattice the spectral dimension coincide with the regular dimension of 2. However, the spectral dimension of the Sierpinski gasket is smaller than 2 and is given by $\frac{2ln(3)}{ln(5)} \simeq 1.365$[18, 19]. Figure (9) describes the spectral analysis of a finite Sierpinski gasket with 9842 nodes (green) and a 99×99 regular square lattice (magenta). These finite networks serve as numeric approximations for the infinite networks.

[1] A related quantity is $G(\omega)$, the cumulative density of states defined as $G(\omega) = \int_0^\omega g(\omega')d\omega'$. $G(\omega)$ counts the number of modes with frequency less than ω. The scaling relation $G(\omega) \sim \omega^{d_s}$ holds for low frequencies.

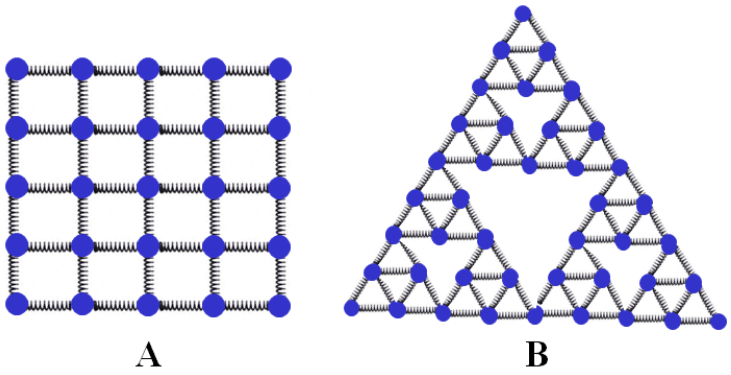


Figure 8: Two different elastic networks of masses and springs. Left - A regular square lattice as an elastic network, every node is connected with springs to its nearest neighbors. Right - The Sierpinski gasket as an elastic network, the nodes here are the vertices of the original gasket at each recursive step.

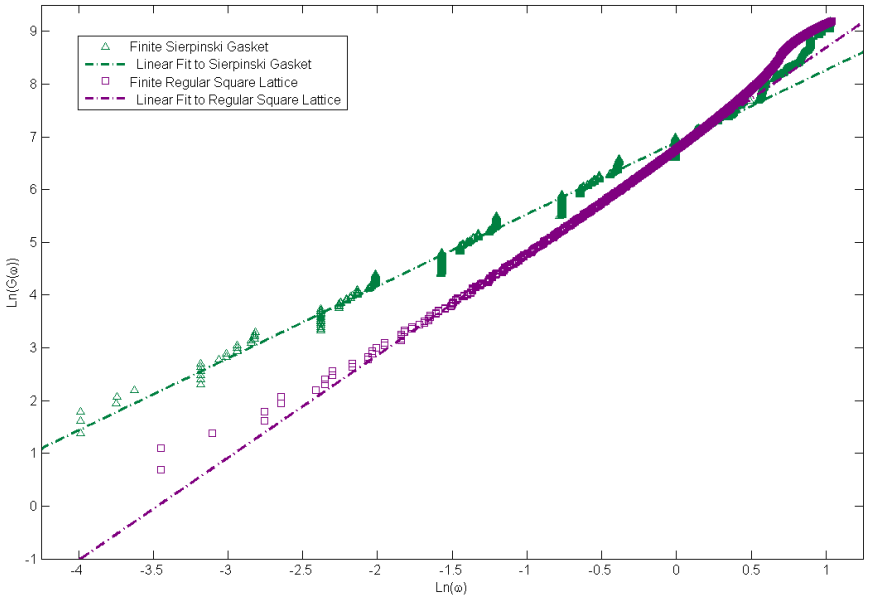


Figure 9: Spectral analysis of two finite elastic networks. For each elastic network, we found the set of vibrational eigenfrequencies $\{\omega_0, \omega_1, ..., \omega_{N-1}\}$ that characterize it and then plotted $Ln(G(\omega))$ vs. $Ln(\omega)$. Low frequency regions of the cumulative density of modes $G(\omega)$ clearly exhibit a power law behavior. Dashed lines indicate best fits to these regions, the slopes correspond to the spectral dimension, it is clearly visible that $d_s^{square} > d_s^{gasket}$. Numerical values are $d_s^{square} = 1.943$ and $d_s^{gasket} = 1.366$, these values agree with the theoretical values for the corresponding infinite networks.

5 The Gaussian Network Model

With recent advances in sequencing genomes, it has become clear that the canonical sequence-to-function paradigm is far from being sufficient. Structure has emerged as an important source of additional information required for understanding the molecular basis of observed biological activities. Yet, advances in structural genomics have now demonstrated that structural knowledge is not sufficient for understanding the molecular mechanisms of biological function either. The connection between structure and function presumably lies in dynamics, suggesting an encoding paradigm of sequence to structure to dynamics to function.

Not surprisingly, a major endeavor in recent years has been to develop models and methods for simulating the dynamics of proteins, and relating the observed behavior to experimental data. These efforts have been largely impeded, however, by the memory and time cost of molecular dynamics simulations. These limitations are particularly prohibitive when simulating the dynamics of large structures or supramolecular assemblies.

While accurate sampling of the conformational space is a challenge for macromolecular systems, the study of protein dynamics benefits from a great simplification. Proteins have uniquely defined native structures under physiological conditions and they are functional only when folded into their native conformation. Therefore, while the motions of macromolecules in solution are quite complex and involve transitions between an astronomical number of conformations, those of proteins near native state conditions are much simpler, as they are confined to a subset of conformations near the folded state. These conformations usually share the same overall fold, secondary structural elements, and even tertiary contacts within individual domains. Typical examples are the open and closed forms of enzymes, usually adopted in the unliganded and liganded states, respectively.

Exploring the fluctuation dynamics of proteins near native state conditions is a first step toward gaining insights about the molecular basis and mechanisms of their function; and fluctuation dynamics can be treated to a good approximation by linear models such as Normal Mode Analysis (NMA)[20]. Another distinguishable property of protein dynamics, in addition to confinement to a small subspace of conformations, is the collective nature of residue fluctuations. The fluctuations are indeed far from random, involving the correlated motions of large groups of atoms, residues, or even entire domains or molecules whose concerted movements underlie biological function. An analytical approach that takes account of the collective coupling between all residues is needed, and again NMA emerges as a reasonable first approximation.

5.1 The Model

Most analytical treatments of protein dynamics entail a compromise between physical realism and mathematical simplicity. The challenge is to identify a simple, yet physically plausible, model that retains properties of interest and experimental relevance[21]. As follows from the previous section, the dynamics of proteins revolves around and near their native state. As in numerous other cases in physics, the dynamics of a system near its energetic minima may be studied using an harmonic approximation. The Gaussian Network Model (GNM) proposed by Bahar et al[22], utilizes such an approximation and is widely applied because it yields results in agreement with X-ray spectroscopy experiments. Its main aim is to help explore the role and contribution of purely topological constraints, defined by the 3D structure, on the collective dynamics of proteins. The GNM considers proteins to be elastic networks whose nodes correspond to the positions of the alpha-carbons in the native structure and the interactions among nodes are modeled as harmonic springs taken to be homogeneous. An interaction between two nodes exists if the nodes are separated by a distance less than R_c that is known as the interaction cutoff. The cutoff distance is usually taken in the range $6\overset{\circ}{A} - 7\overset{\circ}{A}$, based on the radius of the first coordination shell around residues observed in PDB structures[23, 24]. The only information required to implement the method is knowledge of the native structure. Figure (10)[20] is an illustrative description of the GNM.

The GNM is defined by the quadratic Hamiltonian equation

$$H_{GNM}=\sum_i \frac{(\overrightarrow{P_i})^2}{2M} + \frac{\gamma}{2}\sum_{i,j>i} \Delta_{ij}(\triangle\overrightarrow{R_i} - \triangle\overrightarrow{R_j})^2. \tag{2}$$

The first term represents the kinetic energy of the system, γ is the spring force constant which is assumed to be homogeneous, $\overrightarrow{R_i}$ and $\triangle\overrightarrow{R_i}$ are the instantaneous position and the displacement with respect to $\overrightarrow{R_i^0}$ of the i-th C_α atom respectively. Δ is the network connectivity matrix with the following entries: $\Delta_{ij} = 1$ if $i \neq j$ and the distance $\left|\overrightarrow{R_i} - \overrightarrow{R_j}\right|$ between the two C_α atoms, in the native conformation, is below the cutoff R_c, $\Delta_{ij} = 0$ otherwise. Physically, this means that in addition to changes in inter-residue distances, any change in the direction of the inter-residue vector is also being resisted or penalized in the GNM potential. The GNM makes two assumptions, as implied by the Hamiltonian, fluctuations are isotropic and Gaussian.

5.2 Classical Mechanics

As in every mechanical system of masses and harmonic springs also in the GNM elastic network, the normal modes and eigenfrequencies are of prime interest. A simple analysis of the elastic network created with the GNM is readily preformed using Lagrangian mechanics (Hamiltonian mechanics will of course yield the

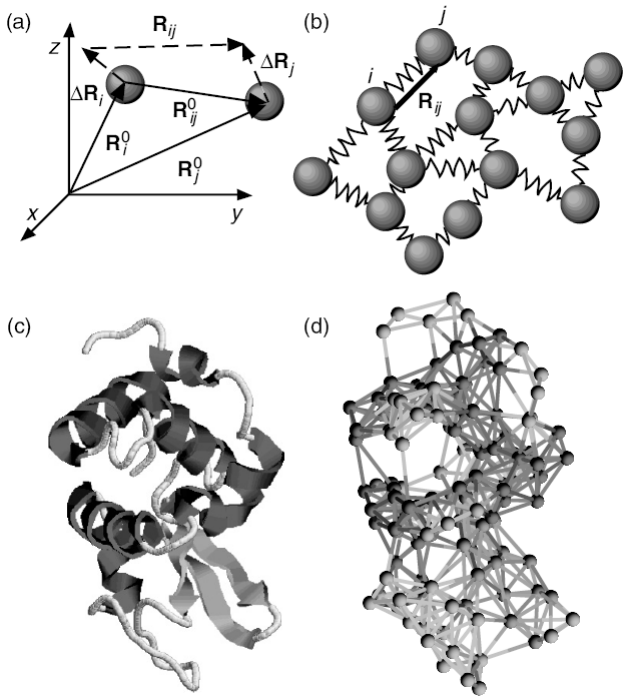


Figure 10: Description of the GNM. (a) Schematic representation of the equilibrium positions of the i-th and j-th nodes, $\vec{R}^0_i$ and $\vec{R}^0_j$ with respect to a laboratory-fixed coordinate system (xyz). $\vec{R}^0_{ij}$ is the equilibrium separation vector between nodes i and j. The instantaneous fluctuation vectors, $\Delta\vec{R}_i$ and $\Delta\vec{R}_j$, are shown by the dashed arrows, along with the instantaneous separation vector $\vec{R}_{ij}$ between the positions of the two residues. (b) In the elastic network of the GNM every residue is represented by a node and connected to spatial neighbors by uniform springs. These springs determine the degrees of freedom in the network and the structure's modes of vibration. (c) Three dimensional image of hen egg white lysozyme (PDB code 1hel) showing the C-alpha trace, secondary structure features are also indicated. (d) Using a cutoff value of choice, all connections between C-alpha nodes are drawn for the same lysozyme structure to indicate the nature of the elastic network analyzed by the GNM.

same results). The classical Lagrangian of the GNM elastic network can be written as:

$$\mathcal{L} = \sum_i \frac{M(\dot{\overrightarrow{R_i}})^2}{2} - \frac{\gamma}{2}\sum_{i,j>i}\Delta_{ij}(\triangle\overrightarrow{R_i} - \triangle\overrightarrow{R_j})^2 \ . \tag{3}$$

Written in a less compact form:

$$\mathcal{L} = \sum_i \frac{M\triangle\dot{X_i}^2}{2} + \frac{M\triangle\dot{Y_i}^2}{2} + \frac{M\triangle\dot{Z_i}^2}{2} - \frac{\gamma}{2}\sum_{i,j>i}\Delta_{ij}\left[(\triangle X_i - \triangle X_j)^2 + (\triangle Y_i - \triangle Y_j)^2 + (\triangle Z_i - \triangle Z_j)^2\right] \ , \tag{4}$$

it is clear that this Lagrangian is three times degenerate and separable into the variables: $\triangle X_i$,$\triangle Y_i$,$\triangle Z_i$. It is hence suffice to concentrate on:

$$\mathcal{L} = \sum_i \frac{M\triangle\dot{X_i}^2}{2} - \frac{\gamma}{2}\sum_{i,j>i}\Delta_{ij}(\triangle X_i - \triangle X_j)^2 \ . \tag{5}$$

To continue further we first note that:

$$\sum_{i,j>i}\Delta_{ij}(\triangle X_i - \triangle X_j)^2 = \sum_{i,j>i}\Delta_{ij}(\triangle X_i^2 - 2\triangle X_i\triangle X_j + \triangle X_j^2) = \sum_{i,j}\Gamma_{ij}\triangle X_i\triangle X_j \ , \tag{6}$$

where the matrix Γ is defined as follows :

$$\Gamma_{ij} = \begin{cases} -1 & if\ i \neq j\ and\, R_{ij}^0 \leq R_c \\ 0 & if\ i \neq j\ and\, R_{ij}^0 > R_c \\ \sum_k \Delta_{ik} & if\ i = j \end{cases} \ . \tag{7}$$

We can hence write the Lagrangian in the following form:

$$\mathcal{L} = \sum_i \frac{M\triangle\dot{X_i}^2}{2} - \frac{\gamma}{2}\sum_{i,j}\Gamma_{ij}\triangle X_i\triangle X_j \ . \tag{8}$$

Recalling the Euler Lagrange equations of motion ($\frac{\partial\mathcal{L}}{\partial\triangle X_i} = \frac{\partial}{\partial t}\frac{\partial\mathcal{L}}{\partial\triangle\dot{X_i}}$) we get an equation for every index i:

$$M\triangle\ddot{X_i} = -\gamma\sum_j\Gamma_{ij}\triangle X_j \ . \tag{9}$$

This set of N equations can be written in matrix form:

$$M\triangle\ddot{\vec{X}} = -\gamma\Gamma\triangle\vec{X} \ . \tag{10}$$

Substituting an oscillatory solution, $\triangle\vec{X} = \vec{A}e^{i\omega t}$, we get an eigenvalue equation for the matrix Γ:

$$\Gamma\vec{A} = \frac{M\omega^2}{\gamma}\vec{A} \ . \tag{11}$$

We may conclude that the eigenfrequencies of the this elastic network are, up to a proportionality factor, the square root of the eigenvalues of the matrix Γ.

5.3 Relevance

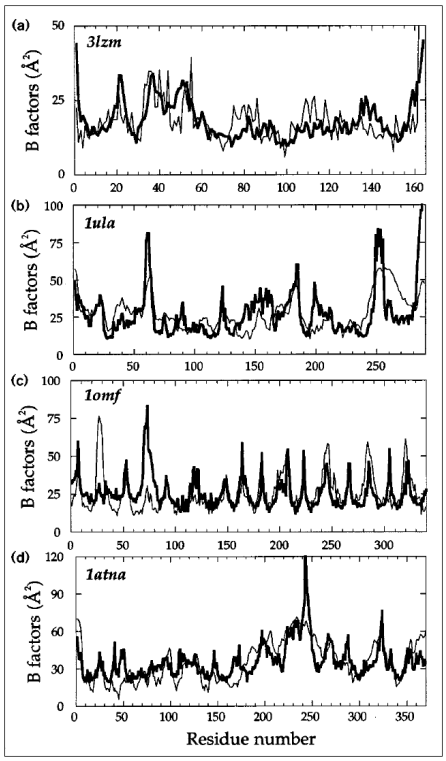
One may wonder, to what extent is the GNM a reliable description of protein dynamics ? The mean square fluctuation of a residue from its equilibrium position is experimentally measurable. In X-ray crystallography this quantity is directly related to measurable B-factors and in NMR experiments this quantity is simply the root mean-square difference between different NMR models.

GNM allows us to theoretically calculate the mean square fluctuation of every residue. A thermodynamic analysis of GNM[22] provides us with a formula for the theoretical value of the experimentally measurable B-factors mentioned above:

$$B_i \equiv \frac{8\pi^2}{3}\left\langle(\Delta\overrightarrow{R_i})^2\right\rangle = \frac{8\pi^2 k_B T}{\gamma}\left[\Gamma^{-1}\right]_{ii} \ ,$$

where the subscript i stands for the $i-th$ residue. It is thus possible to compare between theoretical and experimental results.

Figure 11: Theoretically calculated vs. experimentally measured B-factors for four different proteins.



Starting with the paper that introduced the GNM[22], several studies have demonstrated that the fluctuations predicted by the GNM are in good agreement with experimental B-factors. Figure (11)[22] shows a comparison between theoretically calculated and experimentally measured B-factors for four proteins, PDB codes (a) 3lzm, (b) 1ula, (c) 1omf and (d) 1atna. Curves shown in bold were obtained utilizing the GNM, curves drawn as thin lines represent experimental data. It is interesting to note that in a recent study[25] conducted on a set of 64 nonhomologous proteins, each containing a structure solved by NMR and X-ray crystallography. The GNM predictions for mean square fluctuation yielded a correlation of 0.59 with X-ray data and a distinctively better correlation (0.75) with NMR data. The higher correlation between GNM and NMR data, compared to that between GNM and X-ray B factors, was shown to arise from the differences in the spectrum of modes accessible in solution and in the crystal environment. Mainly, large amplitude motions sampled in solution are restricted, if not inaccessible, in the crystalline environment of X-rays.

6 Proteins as Fractal Entities

Proteins are not the abstract mathematical objects, described in chapter 4, known as fractals. Recently however, it turned out that proteins may be regarded, at least in some aspects, as fractals[3, 4, 5]. As recent studies show, one may attribute at least two "fractal" dimensions to proteins, the mass fractal dimension d_f and the spectral dimension d_s. These parameters, in addition to the number of amino acids along the protein backbone, sum up essential information regarding the topology of a protein and its basic dynamics. From reasons to become clear in chapters 7 and 8, we were required to calculate d_f and d_s simultaneously for a large number of proteins. The task presented a challenge since such a large scale analysis has never been conducted before and because it requires the development and implementation of "home made" calculation methods.

In this chapter I define and describe the terms: mass fractal dimension and spectral dimension. I also describe practical calculation methods for these dimensions given the 3D structure of a protein. By implementing the methods described in this chapter we were able to calculate the spectral and fractal dimensions for a data set of 543 protein structures all taken from the Protein Data Bank (PDB). The proteins that were chosen differ in function and/or source organism and represent a wide length scale ranging from 100 to 3000 residues. The results of our calculations are summarized in appendix D.

6.1 The Mass Fractal Dimension

The mass fractal dimension d_f, gives us an indication of how completely a fractal/protein fill space. Describing the mass fractal dimension d_f is most convenient using a three dimensional example. Draw a sphere of radius r enclosing some lattice points in space and calculate their mass $M(r)$, increase r and calculate again. Do this several times and if $M(r)$ scales as r^{d_f} the exponent d_f is called the fractal dimension. Leitner et al preformed a large scale analysis of the mass fractal dimension in proteins, calculating the fractal dimension for 200 proteins[3]. It has been found that proteins can be described as mass fractals whose mass fractal dimension d_f is close to 2.5 (with a statistical standard deviation of about 0.2). For a regular 3D lattice d_f coincide with the usual dimension of 3, for proteins however $2 < d_f < 3$, leading to sparser fill of space. Our calculations on a larger set of proteins assert these conclusions.

Although the definition of the fractal dimension is rather simple, while trying to develop a numerical method for calculating the fractal dimension in proteins, we encountered a few questions: How does one choose the origin for the sphere mentioned above ? Should one average the calculations over several origins ? Since proteins are finite objects, unlike exact fractals, how does one choose lower and upper cutoffs for r ? Through a process of trial and error we have devised a reasonable algorithm for calculating the fractal dimension of a protein. The

algorithm was implemented in MATLAB[2]; the MATLAB code appears in appendix A, the algorithm is described below.

We used the following algorithm in order to calculate the fractal dimension of a protein:

1. Find the protein's center of mass.

2. Compute the protein's radius of gyration R_g[26], a characteristic length scale, given by:

$$R_g = \sqrt{\frac{\sum_i m_i r_i^2}{\sum_i m_i}}$$

where the sum is over each atom i of mass m_i and distance r_i from the center of mass.

3. Find the ten C-Alpha atoms closest to the center of mass.

4. For every C-Alpha in (3), set it as origin and linearly fit $log(M(r))$ against $log(r)$ when $r = 1\overset{\circ}{A}, 2\overset{\circ}{A}, ..., round(R_g)\overset{\circ}{A}$.

5. The mass fractal dimension d_f is the average over the ten slopes obtained in step 4, Figure (12) illustrates the procedure.

6.2 The Spectral Dimension

The GNM models a protein as an elastic network. As was mentioned in chapter 5, the normal modes and eigenfrequencies of such a system are of prime interest. The spectral dimension d_s governs the density of low frequency normal modes on a fractal/protein, we will focus our attention on the behavior of this subset. More precisely, denoting the density of modes with frequency ω : $g(\omega)$, the scaling relation $g(\omega) \sim \omega^{d_s-1}$ holds for low frequencies. The spectral dimension d_s is hence a quantity directly related to the vibrational dynamics of a fractal/protein. Vulpiani et al[4] computed the spectral dimension, for a set of 57 proteins, within the GNM framework. It was found that low frequency regions clearly exhibit a power-law behavior and that usually $d_s < 2$ although higher values of the interaction cutoff R_c lead to higher values of d_s. For a regular 3D lattice d_s coincide with the usual dimension of 3, for proteins however $d_s < 2$, leading to an excess of low frequency modes. Our calculations on a much larger set of proteins assert these conclusions.

In order to calculate d_s we first built the matrix Γ defined in chapter 5 for every protein we studied. We have done so for two different values of R_c:

[2]I used MATLAB 2007a with the corresponding bioinformatics toolbox.

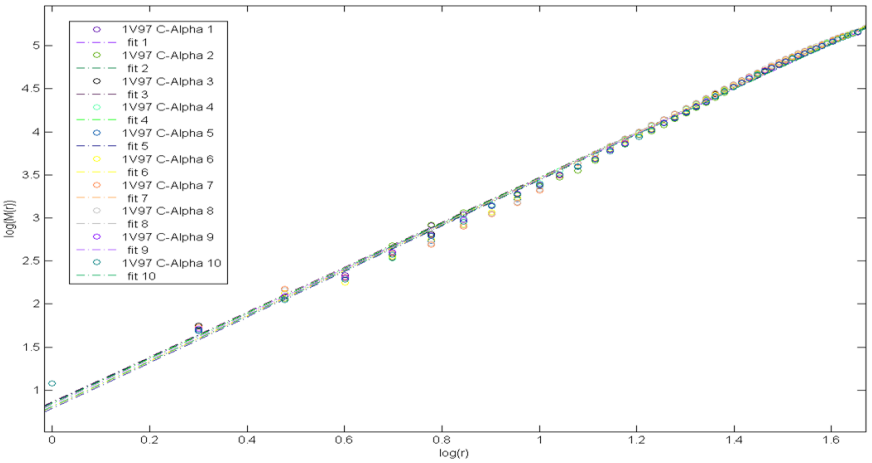


Figure 12: Calculating the mass fractal dimension d_f for PDB code 1V97 ($N = 2594$,$d_f = 2.64$), d_f was taken to be the average mass fractal dimension obtained by choosing the origin to be each and every one of the ten C-Alpha atoms closest the protein's center of mass. For a given origin, d_f was estimated via a power law fitting to $M(r)$, dashed lines indicate best fits, the average slope correspond to the fractal dimension.

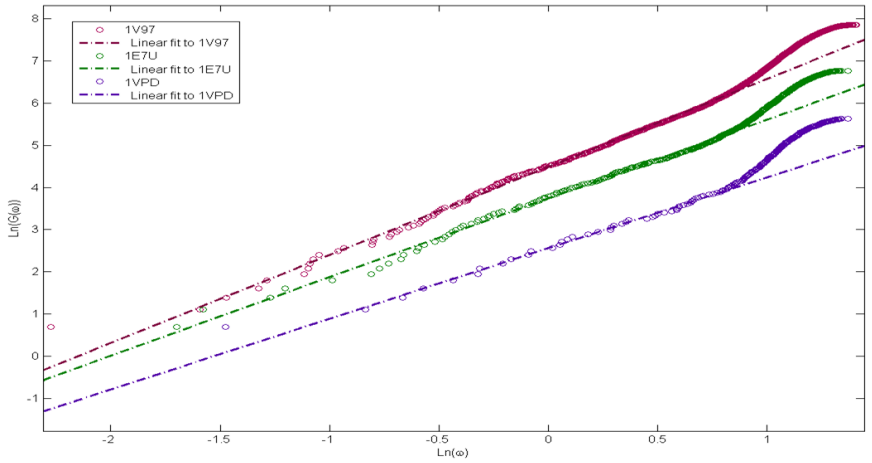


Figure 13: Calculating the spectral dimension d_s for three proteins with different sizes, PDB codes: 1V97 ($N = 2594$,$d_s = 2.09$), 1E7U ($N = 872$,$d_s = 1.86$) and 1VPD ($N = 279$,$d_s = 1.68$). For each protein, we found the set of vibrational eigenfrequencies $\{\omega_0, \omega_1, ..., \omega_{N-1}\}$ that characterize the elastic network it forms when modeled by the GNM and plotted $Ln(G(\omega))$ vs. $Ln(\omega)$. In this example $R_c = 7$Å. Low frequency regions of $G(\omega)$ clearly exhibit a power law behavior, i.e the scaling relation $G(\omega) \sim \omega^{d_s}$ holds for low frequencies. Dashed lines indicate best fits to these regions, the slopes correspond to the spectral dimension.

$6\overset{\circ}{A}$ and $7\overset{\circ}{A}$, as in Vulpiani et al[4]. The set of vibrational eigenfrequencies $\{\omega_0, \omega_1, ..., \omega_{N-1}\}$ was then obtained by diagonalizing the matrix Γ[3]. These two steps were implemented in MATLAB[4] and hence automated, the MATLAB code appears in appendix B. Creating a plot of $Ln(G(\omega))$ vs. $Ln(\omega)$, where $G(\omega)$ is the cumulative density of states, one finds that the low frequency section behaves linearly. Manual inspection of the plot is required in order to determine the boundaries of the linear section. The slop of this line is actually the spectral dimension d_s and is determined via fitting. Figure (13) illustrates the procedure by showing a plot of $Ln(G(\omega))$ vs. $Ln(\omega)$ for three different proteins.

6.2.1 The Boson Peak in Proteins

The Boson Peak is formally defined as the low frequency peak in the function $g(\omega)/g_D(\omega)$, where $g(\omega)$ is the vibrational density of states and $g_D(\omega) \propto \omega^2$ is the Debye behavior for low frequencies. A peak in this function would then represent an excess of low vibrational modes with respect to a perfect harmonic crystal. The Boson Peak is one of the most striking properties of glasses[27], recent experimental and theoretical studies imply that the Boson Peak appears in proteins as well[4, 5, 28]. Since in proteins we find that $d_s - 1 < 2$, our study asserts this conclusion.

[3] The eigenfrequencies of the the elastic network are, up to a scale factor, the square root of the eigenvalues of the matrix Γ. One should not be bothered with the proportionality factor since we are only interested in the scaling law of $g(\omega)$ with ω.

[4] I used MATLAB 2007a with the corresponding bioinformatics toolbox.

7 Thermal Instability

A classical result obtained by Peierls in 1934[9] provides a thermodynamic explanation for the instability of low dimensional crystalline structures. The argument, based on harmonic vibrational dynamics, shows that the mean square displacement of any structural unit at finite temperature diverges in the thermodynamic limit when the lattice dimension is 1 or 2. Indeed, when such a quantity exceeds the order of magnitude of the lattice spacing, the structure behaves as a liquid and the crystalline order makes no longer sense. Even if anharmonic terms are usually present in real structures, the result of Peierls still holds since the instability is present at any finite temperature and in particular in the low-temperature regime where the harmonic approximation is correct. On the other hand, on real finite structures far from the thermodynamic limit, the crystalline order is stable if the mean-square displacement does not exceed the lattice spacing. The maximum stability size at room temperature is so small for $d = 1$ that it make no sense speaking of 1-dimensional crystals, while for $d = 2$ the logarithmic divergence is slow enough to allow the existence of small finite 2-dimensional crystals.

In this section I will describe a generalization of the Peierls result for a generic fractal elastic network described by the GNM[8]. The general result extends the Peierls theorem for non crystalline structures, proving that stability in the thermodynamic limit is possible if and only if $d_s > 2$. In particular, in light of chapter six, this generalization applies for proteins. It will be shown in chapter eight how this result combined with an appropriate melting criterion leads to a general equation relating the spectral and fractal dimensions of a protein to the number of amino acids along the protein backbone.

7.1 The GNM in Normal Coordinates

As was described in chapter 5, in GNM the dynamics of an elastic network is fully described by the following set of equations:

$$M\triangle\ddot{X}_i = -\gamma\sum_j \Gamma_{ij}\triangle X_j \ , \tag{12}$$

or in matrix form:

$$M\triangle\ddot{\vec{X}} = -\gamma\Gamma\triangle\vec{X} \ . \tag{13}$$

A brief look at these equations reveals that they are coupled. The dynamics of the i-th node, described by the deviation from equilibrium $\triangle X_i$, depends not only on $\triangle X_i$ itself but also on other nodes $\{\triangle X_j\}$. Although the above description is very natural, since the coordinates used are the actual deviations from equilibrium, it leads to a rather complicated set of equations. It is sometimes beneficial to describe the system differently using a special set of coordinates called normal coordinates. Describing the system with normal coordinates leads

to N uncoupled equations of motion and great mathematical simplicity.

The matrix Γ defined in chapter 5 is real and symmetric[5] by definition. One of the basic theorems concerning such matrices is the finite-dimensional spectral theorem, which says that any symmetric matrix whose entries are real can be diagonalized by an orthogonal matrix[6]. More explicitly: to every symmetric real matrix Γ there exists a real orthogonal matrix A such that $D = A^{-1}\Gamma A \equiv A^T\Gamma A$ is a diagonal matrix. Every symmetric matrix is thus, up to choice of an orthonormal basis, a diagonal matrix. Another way of stating the real spectral theorem is that the eigenvectors of a symmetric matrix are orthogonal. More precisely, a matrix is symmetric if and only if it has an orthonormal basis of eigenvectors.

Letting A be the real orthogonal matrix that diagonalizes the matrix Γ. We define a new set of coordinates $\{\triangle U_i\}$ using the old set of coordinates $\{\triangle X_i\}$ by the orthogonal transformation:

$$\triangle\vec{X} = A\triangle\vec{U} \ . \tag{14}$$

We are now able to obtain the equations of motion for the new coordinates $\{\triangle U_i\}$, since $A\triangle\ddot{\vec{U}} = \triangle\ddot{\vec{X}} = -\frac{\gamma}{M}\Gamma\triangle\vec{X} = -\frac{\gamma}{M}\Gamma A\triangle\vec{U}$ we get:

$$\triangle\ddot{\vec{U}} = -\frac{\gamma}{M}A^{-1}\Gamma A\triangle\vec{U} = -\frac{\gamma}{M}D\triangle\vec{U} \ , \tag{15}$$

where D is a diagonal matrix whose entries are the eigenvalues of the matrix Γ, i.e the set $\left\{\frac{M}{\gamma}\omega_i^2\right\}$. As promised we got a set of N uncoupled equations, more explicitly the equation of motion for $\triangle U_i$ is given by:

$$\triangle\ddot{U}_i = -\omega_i^2\triangle U_i. \tag{16}$$

$\triangle U_i$ thus obey the equation of motion for a simple harmonic oscillator with angular frequency ω_i.

7.2 Thermal Averages

In order to discuss stability, we would first like to obtain an expression for the mean square displacement of the elastic network which defined by:

$$< \triangle\vec{R}^2 >= \frac{\sum\limits_i < \triangle X_i^2 > + < \triangle Y_i^2 > + < \triangle Z_i^2 >}{N} \ , \tag{17}$$

[5] In linear algebra, a symmetric matrix is a square matrix Γ, that is equal to its transpose $\Gamma = \Gamma^T$. The entries of a symmetric matrix are symmetric with respect to the main diagonal (top left to bottom right), so if the entries are written as a_{ij}, then $a_{ij} = a_{ji}$.

[6] In matrix theory, a real orthogonal matrix is a square matrix A whose transpose is its inverse: $A^TA = AA^T = I$. A real square matrix is orthogonal if and only if its columns form an orthonormal basis of the Euclidean space R_n with the ordinary Euclidean dot product, which is the case if and only if its rows form an orthonormal basis of R_n.

here the pointy brackets denote the thermal average. The GNM however is three times degenerate and hence in the GNM:

$$< \triangle \vec{R}^2 >= \frac{3\sum\limits_i < \triangle X_i^2 >}{N} \ . \tag{18}$$

As was mentioned in chapter 5, it is possible to calculate the thermal averages written on the r.h.s directly and obtain: $< \triangle \vec{R}^2 >= \frac{3k_BT\sum\limits_i \left[\Gamma^{-1}\right]_{ii}}{\gamma N}$. Here however, we follow an indirect path leading to an equivalent result that will be of great use to us.

We first note that:

$$\sum_i < \triangle X_i^2 >= \sum_i < \left[\sum_j A_{ij}\triangle U_j\right]^2 >=< \sum_i\sum_j\sum_k A_{ij}A_{ik}\triangle U_j\triangle U_k > \tag{19}$$

$$=< \sum_j\sum_k\sum_i A_{ij}A_{ik}\triangle U_j\triangle U_k >=< \sum_j\sum_k \delta_{jk}\triangle U_j\triangle U_k >= \sum_j < \triangle U_j^2 > \ ,$$

where we have used the fact that the columns of the matrix A are orthonormal. Calculating the thermal average $< \triangle U_i^2 >$ is easy, since it is nothing but the mean square displacement of a simple harmonic oscillator in thermal equilibrium. After factoring out the integration over momenta, $< \triangle U_i^2 >$ is given by:

$$< \triangle U_i^2 >= \frac{\int\limits_{-\infty}^{\infty} \triangle U_i^2 e^{-\frac{M\omega_i^2\triangle U_i^2}{2k_BT}} d\triangle U_i}{\int\limits_{-\infty}^{\infty} e^{-\frac{M\omega_i^2\triangle U_i^2}{2k_BT}} d\triangle U_i} = \frac{k_BT}{M\omega_i^2} = \frac{k_BT}{\gamma l_i} \ , \tag{20}$$

where l_i denotes the i-th eigenvalue of the matrix Γ and γ is the GNM spring constant. Denoting the density of eigenvalues of the matrix Γ: $g(l)$, we replace summation with integration and obtain:

$$< \triangle \vec{R}^2 >= \frac{3k_BT}{N\gamma} \int\limits_{l_{min}>0}^{l_{max}} \frac{g(l)}{l} dl \ , \tag{21}$$

where l_{min} denotes the smallest positive eigenvalue and l_{max} the largest.

7.3 Stability and Instability

In order to evaluate $< \triangle \vec{R}^2 >$ in the thermodynamic limit, $N \to \infty$, we must first say something about the scaling properties of $g(l)$ and l_{min} in this limit. First, we note that $g(l)$ is an extensive quantity and hence scales with N, this can also be understood from the normalization condition $\int_{l_{min}>0}^{l_{max}} g(l)dl = N$. Second, as was mentioned in chapter 6 the spectral dimension d_s governs the density of low frequency normal modes on a fractal/protein. More precisely, denoting the density of modes with frequency ω : $g(\omega)$, the scaling relation $g(\omega) \sim \omega^{d_s-1}$ holds for low frequencies. Since $l \backsim \omega^2$ we deduce that the scaling relation $g(l) \sim l^{\frac{d_s-2}{2}}$ also holds for low eigenvalues. Third, the lowest positive eigenfrequency ω_{min} corresponds to the lowest wave number k_{min}. The lowest wave number possible is limited by the size of the elastic network $k_{min} \sim \frac{1}{R_g} \sim \frac{1}{N^{1/d_f}}$[7], using the known dispersion relation for fractals[18] $\omega \sim k^{\frac{d_f}{d_s}}$, we conclude that $l_{min} \sim \omega_{min}^2 \sim k_{min}^{\frac{2d_f}{d_s}} \sim N^{-\frac{2}{d_s}}$.

The above scaling relations allow us to compute the value of $< \triangle \vec{R}^2 >$ and find that it is indeed related to the spectral dimension:

$$< \triangle \vec{R}^2 > \sim \frac{k_B T}{\gamma} N^{\frac{2}{d_s}-1} + const \ , \tag{22}$$

which for $N \to \infty$ is finite when $d_s > 2$ and diverges when $d_s \leq 2$.[8] As promised this result extends the Peierls theorem for non crystalline structures, proving that stability in the thermodynamic limit is possible if and only if $d_s > 2$.

[7] The radius of gyration R_g, is a characteristic length scale defined in chapter 6.

[8] For $d_s = 2$ one must recalculate the integral and find that the divergence is actually logarithmic.

8 Unraveling Universality

Two seemingly conflicting properties of native proteins, such as enzymes and antibodies, are known to coexist. While proteins need to keep their specific native fold structure thermally stable, the native fold displays the ability to perform large amplitude motions that allow proper function[1, 2]. This conflict cannot be bridged by compact objects which are characterized by small amplitude vibrations and by a Debye density of low frequency modes. Recently, however, it became clear that proteins can be described as fractals; namely, geometrical objects that possess self similarity[3, 4, 5]. Adopting the fractal point of view to proteins makes it possible to describe within the same framework essential information regarding topology and dynamics[6, 7] using three parameters: the number of amino acids along the protein backbone N, the spectral dimension d_s and the fractal dimension d_f. The fractal character implies large amplitude vibrations of the protein that could have led to unfolding. We show that by selecting a thermodynamic state that is "close" to the edge of stability against unfolding, nature has solved the thermostability conflict. Starting off from a thermal marginal stability criterion we reach a universal equation describing the relation between the spectral and fractal dimensions of a protein and the number of amino acids. This equation is obeyed by a large class of proteins regardless of their source or function. We suggest that deviations from this "equation of state" for protein topology may render a protein unfolded. Nature's solution might be incorporated when planning biologically inspired catalysts.

Based on a generalization[8] of the Peierls instability criterion[9], we derive a general relation between the spectral dimension d_s, the fractal dimension d_f and the number of amino acids along the protein backbone:

$$\frac{2}{d_s} + \frac{1}{d_f} = 1 + \frac{b}{ln(N)} \ . \tag{23}$$

The spectral dimension d_s governs the density of low frequency normal modes of a fractal/protein. More precisely, denoting the density of modes $g(\omega)$, the scaling relation $g(\omega) \sim \omega^{d_s - 1}$ holds for low frequencies. Describing the mass fractal dimension d_f is most convenient using a three dimensional example. Draw a sphere of radius r enclosing some lattice points in space and calculate their mass $M(r)$, increase r and calculate again. Do this several times and if $M(r)$ scales as r^{d_f} the exponent d_f is called the fractal dimension. For a regular 3D lattice both d_s and d_f coincide with the usual dimension of 3. For proteins however, it is usually found that $d_s < 2$ and $2 < d_f < 3$, leading to an excess of low frequency modes and a more sparse fill of space[3, 4].

8.1 The Underlying Physics

As was described in chapter 7, thermodynamic instability appears in inhomogeneous structures and is determined by the spectral dimension d_s. It was

demonstrated that for $d_s \leq 2$, the mean square displacement $< \triangle \vec{R}^2 >$ of a structural unit (in the GNM a single amino acid) in a system composed of N elements, diverges in the limit $N \to \infty$. In particular the mean square displacement in proteins, where we usually find $d_s \leq 2$, diverges in the thermodynamic limit. Here it will be shown that as a result of this divergence, the topological parameters describing a native protein fold are forced to obey a certain relation.

Using T to represent the temperature of the solvent, k_B the Boltzmann constant and γ the spring constant in the GNM, the divergence is given by the asymptotic law

$$< \triangle \vec{R}^2 > \sim \frac{k_B T}{\gamma} N^{\frac{2}{d_s}-1} \ . \tag{24}$$

Letting p be the ratio between the number of surface residues and the total number of residues in a protein and $q = 1 - p$ we write:

$$< \triangle \vec{R}^2 >_{total} = p < \triangle \vec{R}^2 >_{Surface} + q < \triangle \vec{R}^2 >_{Bulk} \ . \tag{25}$$

At very low temperatures the Mean Square Displacements (MSD) of surface residues and of bulk residues are of the same order of magnitude. As temperature increases, MSD values grow and since surface residues are the ones prone to interactions with the solvent, it is reasonable to assume that melting starts when MSD values of surface residues reach a certain threshold to be denoted: $< \triangle \vec{R}^2_{melting} >_{Surface}$.

Equation (25) holds for every N, hence both terms on the r.h.s must scale as the l.h.s, i.e as in equation (24). We note that the mass enclosed by a sphere of radius r is approximately proportional to the number of residues $n(r)$ enclosed by the same sphere and hence $n(r) \propto M(r) \propto r^{d_f}$. Now, since by definition p is directly proportional to the surface to volume ratio of a protein we obtain:

$$p \propto \frac{S}{V} \propto \frac{1}{R_g} \propto \frac{1}{N^{1/d_f}} \ , \tag{26}$$

where R_g is the gyration radius of the protein[26]. Letting T_m represent the melting temperature and utilizing the scaling law:

$$\frac{k_B T}{\gamma} N^{\frac{2}{d_s}-1} \sim p < \triangle \vec{R}^2_{melting} >_{Surface} \sim N^{-1/d_f} < \triangle \vec{R}^2_{melting} >_{Surface} \ , \tag{27}$$

we obtain the following approximation:

$$< \triangle \vec{R}^2_{melting} >_{Surface} \sim \frac{k_B T_m}{\gamma} N^{\frac{2}{d_s}+\frac{1}{d_f}-1} \ . \tag{28}$$

Rearrangement leads to equation (23), where the constant b depends on the proportionality constant between p and $\frac{S}{V}$, the spring elastic constant γ, the MSD melting threshold $< \triangle \vec{R}^2_{melting} >_{Surface}$ and the melting temperature T_m. This dependence, however, is logarithmic and thus very weak, allowing comparison between different proteins without computation of the specific parameters.

8.2 Validity and Robustness

8.2.1 Validation

In order to test the validity of equation (23), we used our calculations of the spectral and fractal dimensions for a data set of 543 proteins. The results of our calculations are summarized in appendix D, calculation methods for d_f and d_s were described in chapter 6. Calculations were preformed on known protein structures, all structures were downloaded from the Protein Data Bank (PDB). The proteins that were chosen differ in function and/or source organism and represent a wide length scale ranging from 100 to 3000 residues.

Statistical analysis of the data gathered reveals satisfying agreement with equation (23). Fitting our data with equation (23) yields the following best-fit parameters: $b = 2.80$ for $R_c = 7\overset{\circ}{A}$ and $b = 3.97$ for a slightly different cutoff $R_c = 6\overset{\circ}{A}$. Despite the diversity in the sample data both cases yield significant correlation coefficients, 0.64 for $R_c = 7\overset{\circ}{A}$ and 0.55 for $R_c = 6\overset{\circ}{A}$. Testing the validity of our predictions further, we tried fitting the data with the equation: $\frac{2}{d_s} + \frac{1}{d_f} = a + \frac{b}{ln(N)}$. The results for $R_c = 7\overset{\circ}{A}$ and $R_c = 6\overset{\circ}{A}$ are shown in figure (14) and (15) respectively. Allowing a free constant fitting parameter enabled us to confront theory with practice since our prediction was $a = 1$. Finding the observed value of the constant fitting parameter was also important, since in a previous study a similar relation between N and d_s was suggested and tested on a small set of proteins[4]. In that study a peculiar offset in the observed value of a constant fitting parameter predicted to be exactly 1 was reported. We believe to have explained the reason for this offset and by doing so we were led to equation (23).The results shown in figures (14) and (15) indicate that the observed value of the fitting parameter $'a'$ is indeed close to one.

We also checked the validity of equation (23) for proteins all originated from the same creature. We thus "sliced" the data according to various sources (Human, E.coli etc...) in order to gain further insight on the relation between the source organism and the fitting parameters. The results of this analysis are summarized in table (1). Of special interest are proteins originating in hyperthermophiles. Surprisingly, such proteins that were included in the analyzed data appear to fulfill equation (23). We shall return and discuss hyperthermostable proteins in chapter 9.

8.2.2 Robustness

The vigilant reader might have noticed that the spectral dimension we have calculated depends on the interaction cutoff R_c. For instance take PDB code: 1RFZ, a protein originated in Bacillus stearothermophilus. We have estimated the spectral dimension of this protein to be 1.87 for $R_c = 7\overset{\circ}{A}$ and 1.68 for $R_c = 6\overset{\circ}{A}$. Moreover, in some cases the spectral dimension calculated for

Source	Proteins	R_c	a	b	c.c
All	543	$7\dot{A}$	0.80 ± 0.06	4 ± 0.37	0.67
Mesophiles	432	$7\dot{A}$	0.80 ± 0.06	3.98 ± 0.39	0.70
E.coli	40	$7\dot{A}$	0.79 ± 0.17	3.95 ± 1.02	0.78
Bacillus subtilis	40	$7\dot{A}$	0.71 ± 0.28	4.51 ± 1.66	0.66
Bos taurus (Cow)	36	$7\dot{A}$	0.98 ± 0.18	2.90 ± 0.69	0.69
Homo sapiens (Human)	44	$7\dot{A}$	0.94 ± 0.28	3.18 ± 1.66	0.51
Mus musculus (Mouse)	37	$7\dot{A}$	0.88 ± 0.28	3.52 ± 1.57	0.61
Rattus norvegicus (Rat)	36	$7\dot{A}$	0.92 ± 0.34	3.38 ± 1.95	0.51
Saccharomyces cerevisiae	38	$7\dot{A}$	0.77 ± 0.31	4.29 ± 1.83	0.65
Salmonella typhimurium	28	$7\dot{A}$	0.62 ± 0.22	5.17 ± 1.35	0.84
Hyperthermophiles	111	$7\dot{A}$	0.80 ± 0.17	4.01 ± 1.02	0.60
Pyrococcus	44	$7\dot{A}$	0.97 ± 0.25	2.84 ± 1.55	0.50
T.maritima	47	$7\dot{A}$	0.75 ± 0.31	4.24 ± 1.81	0.57
A.aeolicus	20	$7\dot{A}$	0.66 ± 0.35	5.04 ± 2.07	0.77
All	543	$6\dot{A}$	0.90 ± 0.09	4.53 ± 0.57	0.55
Mesophiles	432	$6\dot{A}$	0.91 ± 0.1	4.45 ± 0.61	0.57
E.coli	40	$6\dot{A}$	1.05 ± 0.25	3.66 ± 1.53	0.62
Bacillus subtilis	40	$6\dot{A}$	0.66 ± 0.42	6.01 ± 2.49	0.62
Bos taurus (Cow)	36	$6\dot{A}$	1.00 ± 0.30	3.71 ± 1.79	0.59
Homo sapiens (Human)	44	$6\dot{A}$	1.13 ± 0.43	3.21 ± 2.54	0.36
Mus musculus (Mouse)	37	$6\dot{A}$	1.18 ± 0.40	3.11 ± 2.26	0.43
Rattus norvegicus (Rat)	36	$6\dot{A}$	0.86 ± 0.47	5.12 ± 2.71	0.55
Saccharomyces cerevisiae	38	$6\dot{A}$	0.81 ± 0.47	5.05 ± 2.78	0.55
Salmonella typhimurium	28	$6\dot{A}$	0.59 ± 0.50	6.45 ± 3.08	0.64
Hyperthermophiles	111	$6\dot{A}$	0.87 ± 0.25	4.80 ± 1.52	0.51
Pyrococcus	44	$6\dot{A}$	0.99 ± 0.42	3.90 ± 2.57	0.42
T.maritima	49	$6\dot{A}$	0.95 ± 0.46	4.46 ± 2.70	0.44
A.aeolicus	20	$6\dot{A}$	0.73 ± 0.40	5.84 ± 2.36	0.77

Table 1: Fitting the data from various creatures with the equation : $\frac{2}{d_s} + \frac{1}{d_f} = a + \frac{b}{ln(N)}$. It is apparent from table (1) that when allowing a constant fitting parameter it's value remains close to one, this is true for both the set as a whole and for the overwhelming majority of creatures we analyzed.

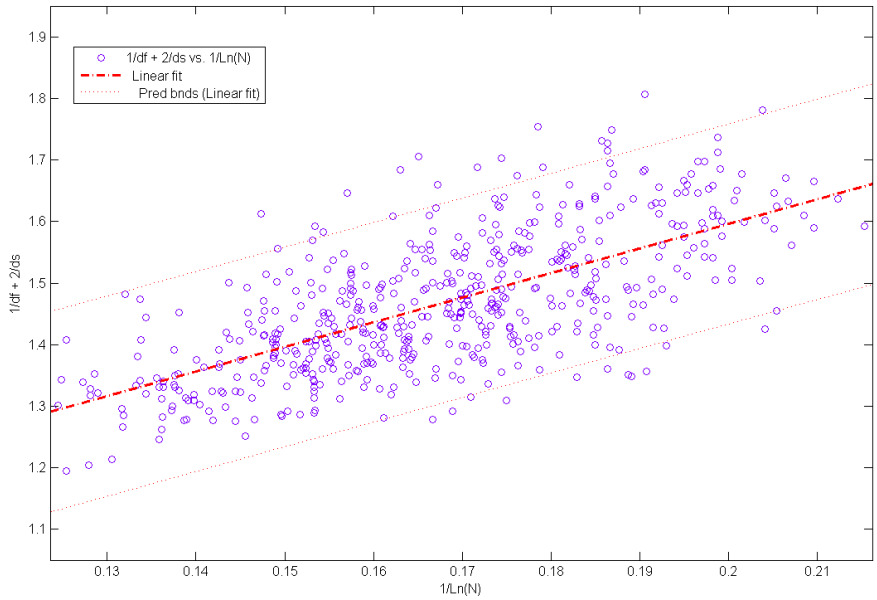


Figure 14: Fitting the data calculated for $R_c = 7.0\mathring{A}$ with the equation: $\frac{2}{d_s} + \frac{1}{d_f} = a + \frac{b}{ln(N)}$. The best fit parameters are $a = 0.8$ and $b = 4$, the correlation coefficient is 0.67. Prediction bounds are for a confidence level of 95%. As expected the observed value of the fitting parameter $'a'$ is indeed close to one.

$R_c = 7\mathring{A}$ is higher than 2, yet the spectral dimension calculated for $R_c = 6\mathring{A}$ is lower than 2. Since equation (23) was derived for proteins with $d_s < 2$, one may be bothered with this issue.

One way to solve the problem is to note that when choosing $R_c = 6\mathring{A}$, $d_s > 2$ only for handful of proteins. It is certainly possible that this choice of the interaction cutoff is one more suitable for proteins. Another way is to keep in mind that the GNM is a simple model, choosing reasonable values for R_c we have used this model in order to approximate the spectral dimension for a large set of proteins. It is certainly possible that our estimation for the spectral dimension d_s is different form the real/experimental spectral dimension d_s^0. Yet since : $\frac{1}{d_s} = \frac{1}{d_s^0 + \delta d_s} \simeq \frac{1}{d_s^0} - \frac{\delta d_s}{[d_s^0]^2}$, our errors translate to variance around the main trend line and equation (23) still holds. It is also possible that although proteins with a spectral dimension higher than 2, are not constrained by equation (23) they nevertheless obey it. Given that the majority of proteins are characterized by a spectral dimension lower than 2 and were hence designed by nature to fold into native structures that obey equation (23). Equation (23) may have become a "guideline" for the design of other proteins as well.

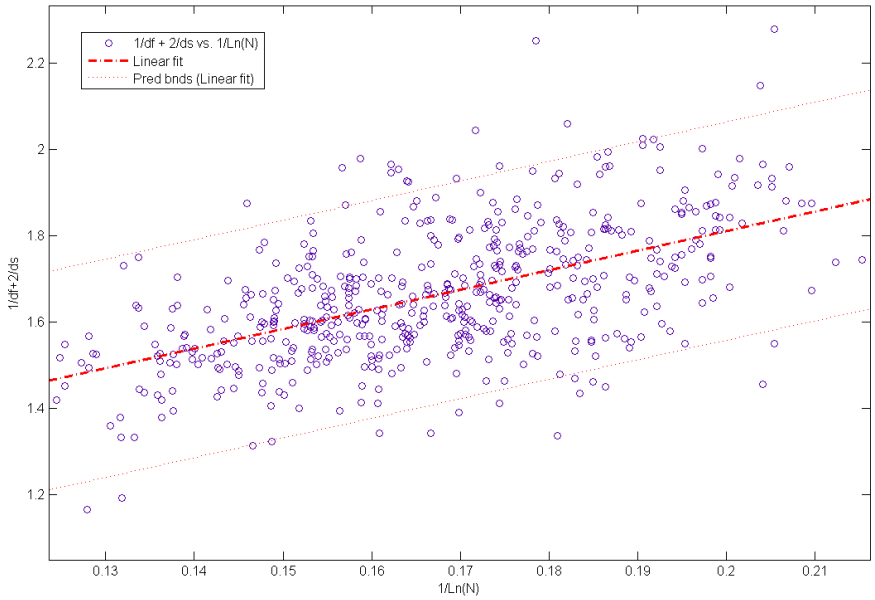


Figure 15: Fitting the data calculated for $R_c = 6.0\mathring{A}$ with the equation: $\frac{2}{d_s} + \frac{1}{d_f} = a + \frac{b}{ln(N)}$. The best fit parameters are $a = 0.9$ and $b = 4.53$, the correlation coefficient is 0.55. Prediction bounds are for a confidence level of 95%. As expected the observed value of the fitting parameter $'a'$ is indeed close to one.

Last but not least, here is the place to mention that another approach leading to relation (23) is described in appendix C. This approach introduces a non-Lindeman melting criterion used mainly for polymers and a bond bending Hamiltonian rather than the GNM Hamiltonian. The bond bending Hamiltonian, previously studied in percolation, describes the harmonic energy penalty associated with changing the bond angles between nodes on the network in addition to the stretch-compress penalty described by harmonic springs. In this approach the derivation of equation (23) is not limited by the demand for $d_s < 2$.

9 Towards Biological Relevance

Striving towards biological relevance, we sought for links between the fractal description of proteins and conventional biology. This part of the work has been more than challenging and I can humbly say that despite relentless efforts our success here has been limited. In the following subsections I will review aspects of our work that were not reviewed in previous chapters. I will present some questions but unfortunately much fewer answers. I believe that some of the work described here may lead to further discoveries and provide a deeper insight into the protein folded state.

9.1 Mutants

A mutant protein is a protein with an amino acids sequence that is not identical to the amino acids sequence of the wild type protein. Mutations are known to affect proteins in many different ways. A mutation in the catalytic site of an enzyme may affect, usually in a negative manner, a protein's ability to perform catalysis. Other mutations may affect biological properties such as the ability to biochemically recognize and interact with other proteins or the ability to adhere to the cell membrane. Some mutations have but minor effect, others may render a protein completely unfolded and hence useless in terms of functionality.

Since it is well known that some mutations are able to alter thermodynamic properties of proteins and since equation (23) is based on a thermostability criterion, It has only been natural to look for mutant-wild type pairs with known 3D structures, such that the mutant has a different melting temperature than the wild type. We have been interested in the affect of such mutations on the fractal properties of a protein. Do they change the spectral and fractal dimensions? If so, is there any correlation between this change and the change in melting temperature? We addressed these questions in a research conducted on 75 wild type – mutant pairs. Thermodynamic data was obtained using the Internet database ProTherm[29], a thermodynamic database for proteins and mutants. Protein structures were downloaded from the protein data bank (PDB).

Instead of presenting all the data we have gathered, I chose to display a representative case study in figure (16). This example illustrates the major difficulty we have encountered while trying to answer the questions presented above. The majority of mutant - wild type pairs we have studied, although sequentially different, were structurally very similar. The close structural similarity led to the fact that in most cases the change in fractal/spectral dimension was too subtle and hence within experimental error. This difficulty should have been expected since structure is more conserved, by evolution, than sequence[30]. Detectable levels of sequence similarity usually imply significant structural similarity. Significant structural similarities are known between proteins with sequential identity as low as 30%, let alone wild type–mutant pairs that differ in only a few amino acids.

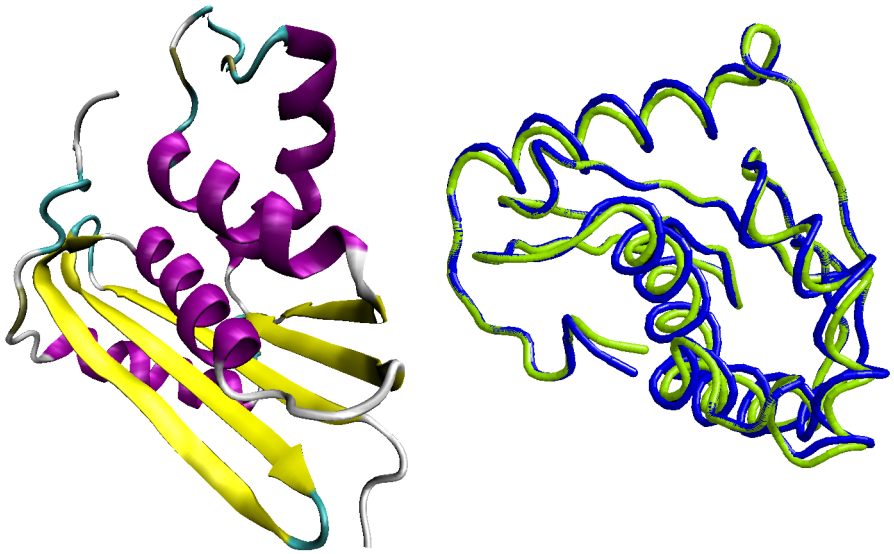


Figure 16: Left: A cartoon depicting the 3D structure of a wild type protein, ribonuclease h from E.coli (PDB code: 2rn2). Right: The back bone of the wild type protein superimposed on the backbone of the mutant protein (PDB code: 1kva). The alignment was produced with C-alpha Match at http://bioinfo3d.cs.tau.ac.il/c_alpha_match/, the server was developed here at T.A.U. The difference between the two proteins is in position 134 where the polar amino acid Aspartate in the wild type was replaced with the non polar amino acid Alanine in the mutant. The melting temperature of the mutant is 3 to 7 degrees higher, depending on the experimental conditions, than the melting temperature of the wild type. Nevertheless, it is clearly visible from the backbone alignment that the structural differences are negligible. It is due to this fact that the spectral and fractal dimensions of the two structures are practically identical.

9.2 Hyperthermophiles

What makes proteins originating in hyperthermophile creatures hyper thermostable ? How can these proteins withstand temperatures as high as 121 °C, allowing survival and reproduction, without unfolding[31]? The literature is full with mechanism that are alleged responsible for hyperthermostability, the most frequently reported include increased van der Waals interactions[32], higher core hydrophobicity[33], additional networks of hydrogen bonds[34], enhanced secondary structure propensity[35], ionic interactions[36], increased packing density[37] and decreased length of surface loops[38]. It was shown recently that proteins use various combinations of these mechanisms. However, no general physical mechanism for increased thermostability was found.

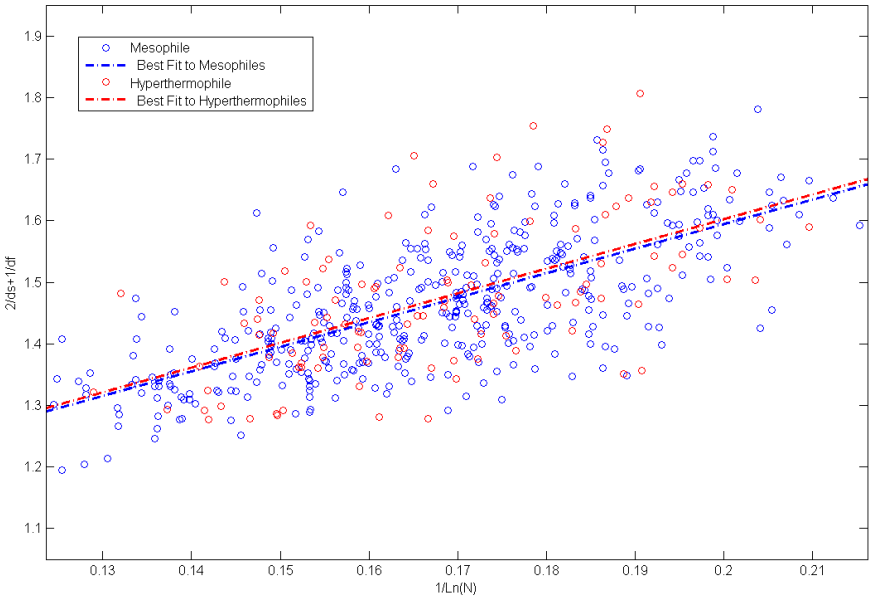


Figure 17: Separately fitting the data for hyperthermophiles and mesophiles with the equation: $\frac{2}{d_s} + \frac{1}{d_f} = a + \frac{b}{ln(N)}$. Each open circle represent a protein, thermostability is color coded, red indicates hyperthermophile, blue indicates mesophile. It is clear from the plots and fits that hyperthermophiles behave similarly to mesophiles.

During the course of our research, proteins originating in hyperthermophiles naturally raised special interest, since equation (23) was derived from fundamental principles all concerning thermal stability. We wondered whether hyperthermostable proteins would reveal their special nature either in the way they fit equation (23) or in any other way that concern their fractal properties. It turned out that we couldn't relate any abnormal behavior with proteins originating in hyperthermophiles. On the contrary regarding their fractal properties we found them to be rather similar to proteins originating in mesophiles. Figures (17) and (18) illustrate our findings.

9.3 GroEL

In biology, chaperones are proteins that assist the non-covalent folding/unfolding and the assembly/disassembly of other macromolecular structures[39]. GroEL belongs to the chaperonin family of molecular chaperones and is found in the bacteria E.coli. It is required for the proper folding of many proteins in the bacteria[40, 41]. To function properly, GroEL requires the lid-like cochaperonin protein complex GroES. The structure of GroEL is depicted in figure (19), the action mechanism is described in figure (20).

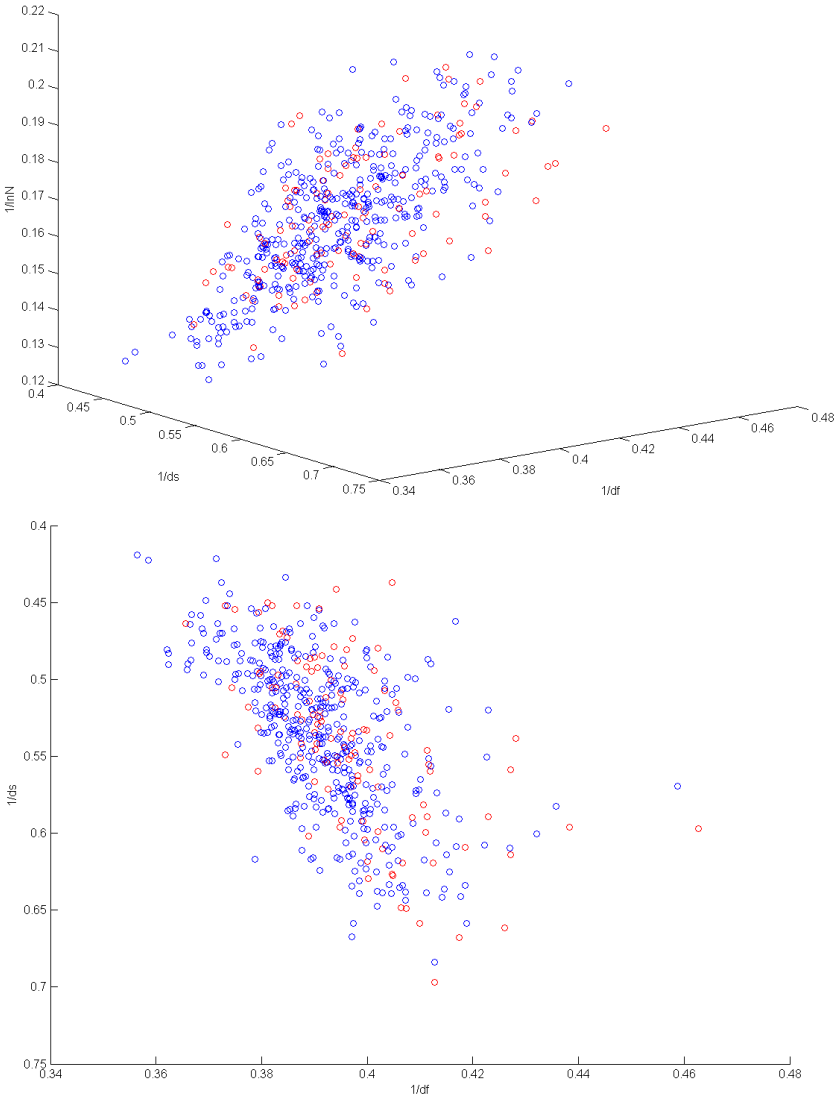


Figure 18: Upper figure - 3D plot of topological parameters for hyperthermophiles and mesophiles. Lower figure - a view of the same plot along the $1/\ln(N)$ axis. Each open circle represents a protein, thermostability is color coded, red indicates hyperthermophile, blue indicates mesophile. Although only two perspective angles are presented above, other angles assert that this plot doesn't reveal any difference between proteins originated in hyperthermophiles and proteins originated in mesophiles..

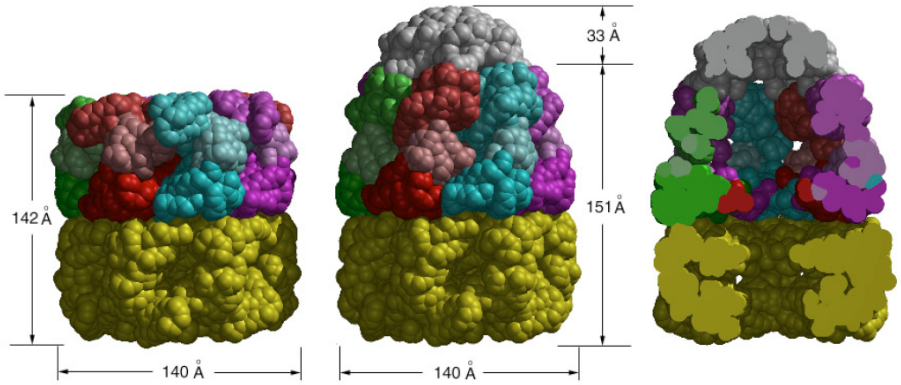


Figure 19: Three computer generated figures depicting the structure of GroEL (left) and GroEL complexed with the lid-like protein GroES (center and right). Structurally, GroEL is a dual-ringed tetradecamer, with both the cis and trans rings consisting of seven subunits each. Various colors are used to distinguish the subunits of GroEL in the upper ring, the lower GroEL ring is uniformly yellow. GroES is uniformly gray. The apical section of GroEL contains a large number of hydrophobic binding sites for "native" (unfolded) protein substrates. Many globular proteins won't bind to the apical domain because their hydrophobic parts are clustered inside, away from the aqueous medium since this is the thermodynamically optimal conformation. Thus, these "substrate sites" will only bind to proteins which are not optimally folded. The apical domain also has binding sites for GroES.

One may wonder what will happen if a protein is forced to strongly deviate from equation (23) and how artificial deformations of the protein fold may lead to a breakdown of criterion (23). Strong deformations of the protein fold may actually happen in vivo as part of a natural process. A possible example is GroEL, recent molecular dynamics simulations demonstrate the unfolding action of GroEL on a protein substrate[42, 43]. Figure (21)[42] illustrates this process. Our work provides a theoretical framework that may help understand GroEL induced unfolding.

9.4 Biotechnology

Enzymes are proteins that catalyze (i.e. accelerate) chemical reactions. In enzymatic reactions, the molecules at the beginning of the process are called substrates, and the enzyme converts them into different molecules, the products. Almost all processes in a biological cell need enzymes in order to occur at significant rates. Enzymes are characterized by a specific native fold that is thermally stable. On the other hand, experimental evidence demonstrate

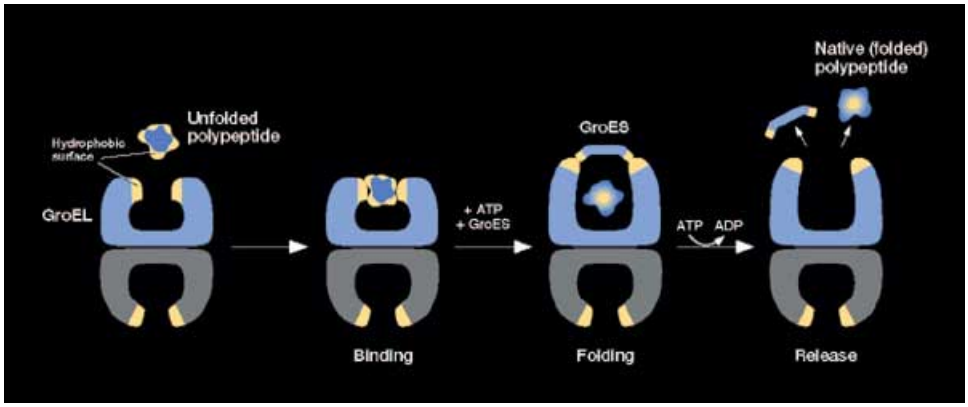


Figure 20: Within the cell, the process of GroEL/ES mediated protein folding involves multiple rounds of binding, encapsulation, and release of the substrate protein. Unfolded substrate proteins bind to a hydrophobic binding patch on the interior rim of the open cavity of GroEL, forming a binary complex with the chaperonin. Binding of substrate protein in this manner, in addition to binding of ATP, induces a conformational change that allows association of the binary complex with a separate lid structure, GroES. Binding of GroES to the open cavity of the chaperonin induces the individual subunits of the chaperonin to rotate such that the hydrophobic substrate binding site is removed from the interior of the cavity, causing the substrate protein to be ejected from the rim into the now largely hydrophilic chamber. The hydrophilic environment of the chamber favors the burying of hydrophobic residues of the substrate, inducing substrate folding. Hydrolysis of ATP and binding of a new substrate protein to the opposite cavity sends an allosteric signal causing GroES and the encapsulated protein to be released into the cytosol. A given protein will undergo multiple rounds of folding, returning each time to its original unfolded state, until the native conformation or an intermediate structure committed to reaching the native state is achieved. Alternatively, the substrate may succumb to a competing reaction, such as misfolding and aggregation with other misfolded proteins.

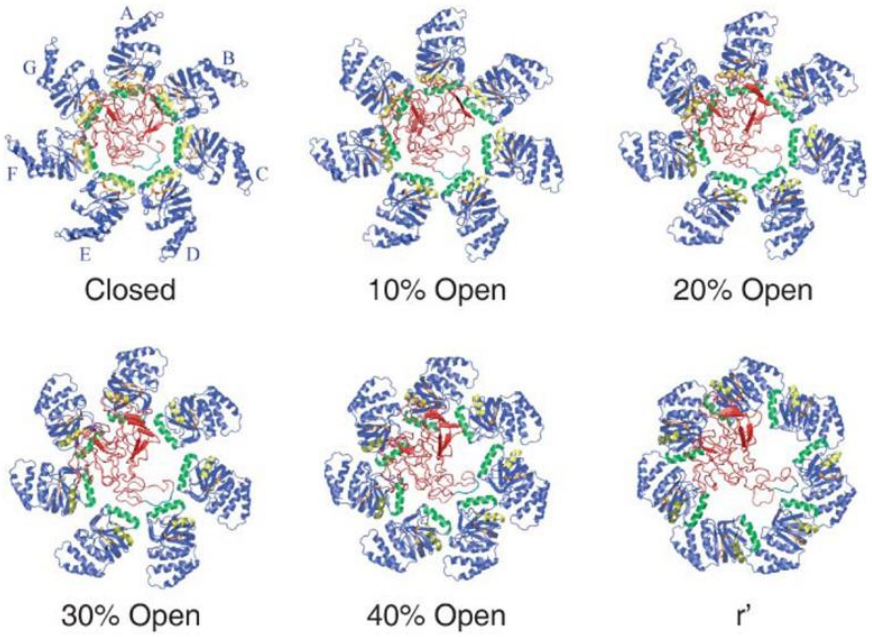


Figure 21: A molecular dynamics simulation of the active unfolding of denatured rhodanese by the chaperone GroEL. The compact denatured protein is bound initially to the cis cavity and forms stable contacts with several of the subunits. As the cis ring apical domains of GroEL undergo the transition from the closed to the more open (ATP-bound) state, they exert a force on rhodanese that leads to the increased unfolding of certain loops. The figure shows snapshots of the unfolding simulation.

that motion is an important ingredient in the ability of an enzyme to function properly[1, 2]. We have shown that enzymes may be described as fractal objects characterized by an excess of low vibrational modes. Furthermore, enzymes obey equation (23) that presumably provides a balance between thermal stability and internal motion. The technological implications hasn't escaped our eyes. Our work opens new possibilities for nanoscale and biologically inspired engineering of catalysts, emphasizing the importance of internal motion. In an attempt to mimic nature, one should consider designing artificial enzymes and other nano devices in a manner that comply with equation (23).

10 Appendix A

A Matlab computer program that calculates the fractal dimension and the radius of gyration of a protein given its PDB code.

```
function[dftenRg,Rg] = dfmasstenRg(PDBid)

% This function calculates :
% 1. The radius of gyration of a protein.
% 2. The df vector of a protein is calculated by taking the ten C-Alpha
%    atoms that are closest to the center of mass as centers. The fractal
%    dimension is then calculated for each of these atoms separately.
%    The upper cutoff was taken to be the radius of gyration.
% 3. dftenRg is the average of the ten df's that were calculated in (2)
% Usage example : [dftenRg,Rg] = dfmasstenRg('9rnt')

% First we get the PDB from the data bank

Protein = getpdb(PDBid);
Run_Time=cputime;

% Then we find the center of mass, and extract the C-alpha atoms positions.

Atom_Number=length(Protein.Model.Atom);
C_Alpha_Count=0;
Total_Mass=0;
meanX=0;
meanY=0;
meanZ=0;
for loop=1:1:Atom_Number
 if Protein.Model.Atom(loop).AtomName(1)=='C'
  if length(Protein.Model.Atom(loop).AtomName)==2
   if Protein.Model.Atom(loop).AtomName=='CA'
    C_Alpha_Count=C_Alpha_Count+1;
    X_C_Alpha_pos(C_Alpha_Count)=Protein.Model.Atom(loop).X;
    Y_C_Alpha_pos(C_Alpha_Count)=Protein.Model.Atom(loop).Y;
    Z_C_Alpha_pos(C_Alpha_Count)=Protein.Model.Atom(loop).Z;
    meanX=meanX+12.0107*X_C_Alpha_pos(C_Alpha_Count);
    meanY=meanY+12.0107*Y_C_Alpha_pos(C_Alpha_Count);
    meanZ=meanZ+12.0107*Z_C_Alpha_pos(C_Alpha_Count);
   else
    X_pos(loop-C_Alpha_Count)=Protein.Model.Atom(loop).X;
    Y_pos(loop-C_Alpha_Count)=Protein.Model.Atom(loop).Y;
    Z_pos(loop-C_Alpha_Count)=Protein.Model.Atom(loop).Z;
    meanX=meanX+12.0107*X_C_Alpha_pos(C_Alpha_Count);
    meanY=meanY+12.0107*Y_C_Alpha_pos(C_Alpha_Count);
```

```
    meanZ=meanZ+12.0107*Z_C_Alpha_pos(C_Alpha_Count);
    temp_Mass(loop-C_Alpha_Count)=12.0107;
   end
  else
   X_pos(loop-C_Alpha_Count)=Protein.Model.Atom(loop).X;
   Y_pos(loop-C_Alpha_Count)=Protein.Model.Atom(loop).Y;
   Z_pos(loop-C_Alpha_Count)=Protein.Model.Atom(loop).Z;
   meanX=meanX+12.0107*X_C_Alpha_pos(C_Alpha_Count);
   meanY=meanY+12.0107*Y_C_Alpha_pos(C_Alpha_Count);
   meanZ=meanZ+12.0107*Z_C_Alpha_pos(C_Alpha_Count);
   temp_Mass(loop-C_Alpha_Count)=12.0107;
  end
 else
  X_pos(loop-C_Alpha_Count)=Protein.Model.Atom(loop).X;
  Y_pos(loop-C_Alpha_Count)=Protein.Model.Atom(loop).Y;
  Z_pos(loop-C_Alpha_Count)=Protein.Model.Atom(loop).Z;
  if Protein.Model.Atom(loop).AtomName(1)=='N'
   meanX=meanX+14.0067*X_pos(loop-C_Alpha_Count);
   meanY=meanY+14.0067*Y_pos(loop-C_Alpha_Count);
   meanZ=meanZ+14.0067*Z_pos(loop-C_Alpha_Count);
   temp_Mass(loop-C_Alpha_Count)=14.0067;
  elseif Protein.Model.Atom(loop).AtomName(1)=='O'
   meanX=meanX+15.9994*X_pos(loop-C_Alpha_Count);
   meanY=meanY+15.9994*Y_pos(loop-C_Alpha_Count);
   meanZ=meanZ+15.9994*Z_pos(loop-C_Alpha_Count);
   temp_Mass(loop-C_Alpha_Count)=15.9994;
  elseif Protein.Model.Atom(loop).AtomName(1)=='S'
    meanX=meanX+32.065*X_pos(loop-C_Alpha_Count);
    meanY=meanY+32.065*Y_pos(loop-C_Alpha_Count);
    meanZ=meanZ+32.065*Z_pos(loop-C_Alpha_Count);
    temp_Mass(loop-C_Alpha_Count)=32.065;
  else
   meanX=meanX+1.00794*X_pos(loop-C_Alpha_Count);
   meanY=meanY+1.00794*Y_pos(loop-C_Alpha_Count);
   meanZ=meanZ+1.00794*Z_pos(loop-C_Alpha_Count);
   temp_Mass(loop-C_Alpha_Count)=1.00794;
  end
 end
end
C_Alpha_Number=C_Alpha_Count;
C_Alpha_Mass=12.0107*ones(1,C_Alpha_Number);
Mass=[ C_Alpha_Mass temp_Mass];
Total_Mass=sum(Mass);
meanX=meanX/Total_Mass;
meanY=meanY/Total_Mass;
meanZ=meanZ/Total_Mass;
```

```
clear protein;

% Then we find the distance of the C-alpha atoms from the Center of mass

distCA(:,1)=1:C_Alpha_Number;
distCA(:,2)=0;
dist_square_CA=(X_C_Alpha_pos-meanX).^2+(Y_C_Alpha_pos-meanY).^2;
dist_square_CA=dist_square_CA+(Z_C_Alpha_pos-meanZ).^2;
distCA(:,2)=sqrt( dist_square_CA );
distCA=sortrows(distCA,2);
dist_square_other=(X_pos-meanX).^2+(Y_pos-meanY).^2+(Z_pos-meanZ).^2;
Rg=sqrt((sum(dist_square_CA*12.0107)+sum(dist_square_other.*temp_Mass))/Total_Mass);

% Then we create the distance array on a log10 scale and sort it.

Distance_Array=(zeros(Atom_Number-1,10,'single'));
Mass_Array=(zeros(Atom_Number-1,10,'single'));
for loop=1:1:10
 a=(X_pos-X_C_Alpha_pos(distCA(loop,1))).^2+(Y_pos-Y_C_Alpha_pos(distCA(loop,1))).^2;
 a=a+(Z_pos-Z_C_Alpha_pos(distCA(loop,1))).^2;
 b=(X_C_Alpha_pos-X_C_Alpha_pos(distCA(loop,1))).^2;
 b=b+(Y_C_Alpha_pos-Y_C_Alpha_pos(distCA(loop,1))).^2;
 b=b+(Z_C_Alpha_pos-Z_C_Alpha_pos(distCA(loop,1))).^2;
 Temp_distance_vector= [a b];
 Temp_Mass_vector=Mass;
 Temp_Mass_vector(find(Temp_distance_vector==0))=[];
 Temp_distance_vector(find(Temp_distance_vector==0))=[];
 Combined_Matrix(:,1)=Temp_distance_vector;
 Combined_Matrix(:,2)=Temp_Mass_vector';
 Combined_Matrix=sortrows(Combined_Matrix,1);
 Distance_Array(:,loop)=Combined_Matrix(:,1);
 Mass_Array(:,loop)=Combined_Matrix(:,2);
 clear Combined_Matrix
end
clear Temp_vector a b distCA meanX meanY meanZ X_pos Y_pos Z_pos
 clear X_C_Alpha_pos Y_C_Alpha_pos Z_C_Alpha_pos
Distance_Array=sqrt(Distance_Array);
Distance_Array=log10(Distance_Array);
forloop=2:length(Mass_Array)
 Mass_Array(loop,:)=Mass_Array(loop,:)+Mass_Array(loop-1,:);
end
Mass_Array=Mass_Array+12;
Mass_Array=log10(Mass_Array);

% Then we create the df vector and calculate the fractal dimension
```

```
for loop=1:1:10
 r=Distance_Array(1:end,loop);
 const=log10(12);
 upper_cutoff=round(Rg);
 X=[ ones(upper_cutoff,1) log10(1:upper_cutoff)' ];
 y(1)=const;
 for second_loop=2:upper_cutoff
  Stop_Point=find(r>log10( second_loop),1,'first')-1;
  y(second_loop)=Mass_Array(Stop_Point,loop);
 end
 fit=regress(y' , X);
 dfvector(loop)=fit(2);
end

dftenRg=mean(dfvector);
Run_Time=cputime-Run_Time
```

11 Appendix B

A Matlab computer program that calculates the eigenfrequencies of a protein within the Gaussian Network Model framework.

```
function [logeigen_PDBid,logN_PDBid] = gnmeig(PDBid,Rc)

% This function calculates the log of the eigen frequencies of
% a protein according to the GNM. The interaction cutoff Rc is
% determened by the user. It also returns a log(1:N) vector required
% for a possibble log log plot.
% Usage example : [logeigen_9rnt,logN_9rnt] = gnmeig('9rnt',7)

% First we get the pdb file from the data bank

Protein = getpdb(PDBid);

% Then we extract the C-alpha Atoms to a seperate variable called Calpha

Calpha=[];
for i=1:1:length(Protein.Model.Atom)
 if length(Protein.Model.Atom(i).AtomName)==2
  if Protein.Model.Atom(i).AtomName=='CA'
   Calpha=[Calpha Protein.Model.Atom(i)];
  end
 end
end

% Then we build the Γ matrix

Conectivity=zeros(length(Calpha));
for i=1:1:(length(Calpha)-1)
 for j=i+1:1:length(Calpha)
  if ( (Calpha(i).X-Calpha(j).X)^2 + (Calpha(i).Y-Calpha(j).Y)^2 + (Calpha(i).Z-
Calpha(j).Z)^2 ) <= Rc^2
   Conectivity(i,j)=-1;
   Conectivity(j,i)=-1;
   Conectivity(i,i)=Conectivity(i,i)+1;
   Conectivity(j,j)=Conectivity(j,j)+1;
  end
 end
end

% Then we diagonolize the Γ matrix and discard the
% first eigenvalue that correspond to translatory movment.
```

```
eigen=sqrt(eig(Conectivity));
n=length(Calpha);
N=1:n;
logeigen_PDBid=log(eigen);
logeigen_PDBid(1)=[];
logN_PDBid=log(N)';
logN_PDBid(1)=[]';
```

12 Appendix C

A different route to equation (23) is to start with a tensorial elasticity model rather than the scalar elasticity (Born) model described by the GNM. Here we use the bond-bending Hamiltonian, previously studied in percolation[44, 45]. It describes the harmonic energy penalty associated with changing the bond angles between nodes on the network in addition to the stretch-compress penalty described by harmonic springs (namely, the GNM). Assuming bond bending potentials to be effectively softer than stretch potentials, a very likely situation in proteins, the vibrational density of states is dominated at low frequencies by bond-bending rather than bond stretch-compress behavior. In this case the cumulative density of states, similar to the scalar elasticity model, behaves as $G(\omega) \sim \omega^{d_E}$ where d_E is the bond-bending spectral dimension equivalent to the spectral dimension d_s. For percolation clusters $d_E < 1$, and this is expected also for other fractals.

Next consider the variance of fluctuations in the distance between two tagged points on the protein that are R_g apart. This may be evaluated in a similar way to the one described in Refs.[6, 46] as $< \overrightarrow{x}^2(R_g) > \sim N^{\frac{2}{d_E}-1}$. Importantly, if $d_E < 1$ and $d_f > 2$, this diverges with increasing N faster than $R_g^2 \sim N^{2/d_f}$. We postulate that melting occurs when the magnitude of these fluctuations reaches the protein size, namely when $< \overrightarrow{x}^2(R_g) > \sim R_g^2$. This leads to :

$$\frac{2}{d_E} - 1 - \frac{2}{d_f} = \frac{const}{lnN} \ . \tag{29}$$

In order to find d_E one has to solve for the eigenfrequencies of the of the bond bending Hamiltonian. To circumvent this difficulty, we use relations that have been derived for percolation clusters, assuming that they hold for other fractals and therefore at least approximately for protein networks[44, 45]. With these relations we find

$$\frac{2}{d_E} - 1 - \frac{2}{d_f} \propto \frac{2}{d_s} - 1 + \frac{1}{d_f} \tag{30}$$

which leads again to equation (23).

13 Appendix D

A raw data table that displays fractal and spectral dimension calculations preformed on 543 proteins.

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
1RZP	Achromobacter cycloclastes (Bacteria)	988	27.43	2.00	1.82	2.67
1RMM	Aequorea victoria (Fungi)	224	17.10	1.97	1.73	2.51
1RJP	Alcaligenes faecalis (Bacteria)	474	21.66	2.10	1.81	2.64
1J9T	Alcaligenes faecalis (Bacteria)	1008	27.62	2.05	1.78	2.69
1JLW	Anopheles dirus b (Insecta)	434	20.70	1.77	1.59	2.59
1R12	Aplysia californica (California sea hare)	502	24.81	1.95	1.54	2.49
1R4V	Aquifex aeolicus (hyperthermophile bacteria)	145	16.02	1.62	1.31	2.42
1ORY	Aquifex aeolicus (hyperthermophile bacteria)	159	15.72	1.70	1.26	2.45
1YE8	Aquifex aeolicus (hyperthermophile bacteria)	167	15.48	1.59	1.33	2.47
1M1H	Aquifex aeolicus (hyperthermophile bacteria)	182	19.73	1.63	1.56	2.34
1ZJR	Aquifex aeolicus (hyperthermophile bacteria)	197	17.05	1.62	1.42	2.50
1T6T	Aquifex aeolicus (hyperthermophile bacteria)	214	19.43	1.52	1.41	2.44
2NYV	Aquifex aeolicus (hyperthermophile bacteria)	215	17.65	1.86	1.53	2.52
1Q77	Aquifex aeolicus (hyperthermophile bacteria)	274	21.36	1.67	1.46	2.49
1L8Q	Aquifex aeolicus (hyperthermophile bacteria)	317	25.06	1.64	1.37	2.39
1C3P	Aquifex aeolicus (hyperthermophile bacteria)	372	19.55	1.79	1.54	2.64
2AU3	Aquifex aeolicus (hyperthermophile bacteria)	403	24.52	1.69	1.39	2.51
1MZH	Aquifex aeolicus (hyperthermophile bacteria)	446	22.88	1.99	1.74	2.59
1VQV	Aquifex aeolicus (hyperthermophile bacteria)	577	26.59	1.92	1.63	2.56
1WY5	Aquifex aeolicus (hyperthermophile bacteria)	622	28.57	1.76	1.57	2.52
2NUB	Aquifex aeolicus (hyperthermophile bacteria)	690	28.27	1.81	1.47	2.55
1NY5	Aquifex aeolicus (hyperthermophile bacteria)	769	29.49	1.79	1.54	2.50
2P3E	Aquifex aeolicus (hyperthermophile bacteria)	801	27.75	2.21	1.89	2.62
2GKS	Aquifex aeolicus (hyperthermophile bacteria)	1051	35.25	1.81	1.81	2.52
2ARK	Aquifex aeolicus (hyperthermophile bacteria)	1074	35.68	2.20	1.72	2.56
2EHH	Aquifex aeolicus (Hyperthermophile bacteria)	1174	31.22	2.21	1.61	2.59
1IVX	Arthrobacter globiformis (Bacteria)	1238	31.68	2.14	1.77	2.71

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
1CF3	Aspergillus niger	581	23.27	2.02	1.73	2.63
9RNT	Aspergillus oryzae	104	12.45	1.69	1.5	2.43
1DE0	Azotobacter vinelandii (Bacteria)	578	24.08	2.07	1.70	2.61
1A3H	Bacillus agaradherans (Bacteria)	300	17.64	1.92	1.70	2.59
1CDG	Bacillus circulans	686	25.37	2.06	1.77	2.61
1R0R	Bacillus licheniformis (Bacteria)	325	18.12	2.10	1.85	2.60
1RFZ	Bacillus stearothermophilus (Bacteria)	637	23.50	1.87	1.68	2.63
1RJW	Bacillus stearothermophilus (Bacteria)	1356	33.04	2.21	1.72	2.68
2GU3	Bacillus subtilis (Bacteria)	130	15.26	1.65	1.07	2.42
1R0U	Bacillus subtilis (Bacteria)	142	15.52	1.69	1.42	2.40
1ISP	Bacillus subtilis (Bacteria)	179	14.69	1.94	1.66	2.53
1SVI	Bacillus subtilis (Bacteria)	182	15.91	1.65	1.46	2.50
1P3J	Bacillus subtilis (Bacteria)	214	16.59	1.52	1.28	2.52
1QGQ	Bacillus subtilis (Bacteria)	238	17.49	1.76	1.48	2.54
1XDZ	Bacillus subtilis (Bacteria)	238	17.36	1.79	1.53	2.52
1COZ	Bacillus subtilis (Bacteria)	252	20.23	1.68	1.40	2.45
1VHX	Bacillus subtilis (Bacteria)	266	21.52	1.57	1.41	2.41
1I60	Bacillus subtilis (Bacteria)	272	18.01	1.82	1.66	2.53
1NRW	Bacillus subtilis (Bacteria)	284	18.17	1.68	1.46	2.58
1T9H	Bacillus subtilis (Bacteria)	287	22.59	1.75	1.46	2.44
1UV4	Bacillus subtilis (Bacteria)	291	17.57	1.78	1.49	2.58
1LF1	Bacillus subtilis (Bacteria)	296	17.67	1.97	1.69	2.57
2GKO	Bacillus subtilis (Bacteria)	309	17.40	2.07	1.96	2.57
1WKQ	Bacillus subtilis (Bacteria)	313	18.91	1.79	1.49	2.59
1S99	Bacillus subtilis (Bacteria)	352	19.51	1.98	1.63	2.59
1VI0	Bacillus subtilis (Bacteria)	367	21.46	1.74	1.52	2.53
1EX2	Bacillus subtilis (Bacteria)	370	25.46	1.75	1.40	2.48
1QD9	Bacillus subtilis (Bacteria)	372	19.19	2.21	1.76	2.57
1DBF	Bacillus subtilis (Bacteria)	381	20.33	1.88	1.60	2.59
1RKT	Bacillus subtilis (Bacteria)	409	23.45	1.64	1.38	2.58
1OYG	Bacillus subtilis (Bacteria)	440	20.81	2.01	1.69	2.61

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
1QE3	Bacillus subtilis (Bacteria)	467	21.90	1.91	1.66	2.61
1rty	Bacillus subtilis (Bacteria)	479	21.61	2.14	1.76	2.60
1GSK	Bacillus subtilis (Bacteria)	502	22.03	1.99	1.77	2.62
1KQP	Bacillus subtilis (Bacteria)	542	23.41	1.91	1.94	2.63
1BKP	Bacillus subtilis (Bacteria)	557	23.83	1.87	1.55	2.61
1MKI	Bacillus subtilis (Bacteria)	598	24.56	2.09	1.68	2.61
1RLI	Bacillus subtilis (Bacteria)	648	25.03	1.88	1.64	2.60
1RXQ	Bacillus subtilis (Bacteria)	680	33.82	2.16	1.50	2.40
1KAM	Bacillus subtilis (Bacteria)	708	27.23	2.07	1.64	2.58
1HQS	Bacillus subtilis (Bacteria)	871	27.74	1.94	1.70	2.62
1F9N	Bacillus subtilis (Bacteria)	887	29.52	1.62	1.44	2.64
1XM3	Bacillus subtilis (Bacteria)	965	30.09	2.31	1.59	2.60
1NSL	Bacillus subtilis (Bacteria)	1050	33.35	2.00	1.56	2.44
1IY9	Bacillus subtilis (Bacteria)	1096	35.08	2.16	1.83	2.51
2B3Z	Bacillus subtilis (Bacteria)	1437	41.25	2.04	1.64	2.43
1AO0	Bacillus subtilis (Bacteria)	1820	38.12	2.12	2.12	2.57
1YIF	Bacillus subtilis (Bacteria)	2128	38.59	2.37	2.02	2.69
1IJG	Bacteriophage phi-29 (Virus)	3084	50.79	2.20	1.95	2.56
1A0I	Bacteriophage t7 (Virus)	332	23.34	1.66	1.53	2.48
4GCR	Bos taurus (Bovine)	185	16.75	1.94	1.57	2.52
2B4Z	Bos taurus (Bovine)	111	12.70	1.65	1.52	2.37
1AGI	Bos taurus (Bovine)	125	14.75	1.74	1.29	2.42
1RIE	Bos taurus (Bovine)	127	14.00	1.58	1.42	2.47
1NEP	Bos taurus (Bovine)	130	14.50	1.93	1.76	2.41
1PRW	Bos taurus (Bovine)	147	14.78	1.82	1.41	2.37
1AMM	Bos taurus (Bovine)	174	16.64	1.86	1.50	2.50
1KT6	Bos taurus (Bovine)	175	15.99	1.68	1.70	2.48
1A44	Bos taurus (Bovine)	185	15.27	1.84	1.48	2.52
2DCB	Bos taurus (Bovine)	251	17.10	2.01	2.10	2.59
1B8O	Bos taurus (Bovine)	280	18.11	1.96	1.69	2.58
1OKC	Bos taurus (Bovine)	292	20.14	1.58	1.38	2.46

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
1RHS	Bos taurus (Bovine)	292	18.64	1.72	1.50	2.56
1M4L	Bos taurus (Bovine)	308	18.26	1.78	1.48	2.58
1ANN	Bos taurus (Bovine)	315	21.57	1.73	1.44	2.49
2CI1	Bos taurus (Bovine)	315	17.61	1.86	1.58	2.57
1GT1	Bos taurus (Bovine)	316	21.27	1.67	1.45	2.50
1VFL	Bos taurus (Bovine)	349	19.01	1.85	1.60	2.60
2ESC	Bos taurus (Bovine)	361	20.12	1.88	1.71	2.57
1G0W	Bos taurus (Bovine)	381	22.24	1.80	1.57	2.55
1F5A	Bos taurus (Bovine)	455	24.10	1.93	1.60	2.55
1FON	Bos taurus (Bovine)	464	22.15	1.94	1.74	2.57
1AKN	Bos taurus (Bovine)	547	23.36	1.86	1.68	2.63
1G08	Bos taurus (Bovine)	572	23.69	1.79	1.53	2.54
2GJ1	Bos taurus (Bovine)	584	23.75	1.85	1.54	2.61
1AVC	Bos taurus (Bovine)	642	28.44	1.99	1.59	2.51
1U19	Bos taurus (Bovine)	696	27.92	1.74	1.50	2.54
1F6R	Bos taurus (Bovine)	727	26.21	2.20	1.55	2.65
1FGH	Bos taurus (Bovine)	753	25.46	2.02	1.76	2.66
4NSE	Bos taurus (Bovine)	832	29.21	2.00	1.81	2.62
1TU5	Bos taurus (Bovine)	1264	33.60	1.99	1.71	2.69
1K8K	Bos taurus (Bovine)	1709	43.86	1.91	1.67	2.53
1AG8	Bos taurus (Bovine)	1972	35.84	2.18	2.42	2.71
4BLC	Bos taurus (Bovine)	1996	36.09	2.15	1.98	2.73
1V97	Bos taurus (Bovine)	2594	45.44	2.09	1.78	2.63
1HWX	Bos taurus (Bovine)	3006	43.05	2.04	1.73	2.76
3COX	Brevibacterium sterolicum	500	21.99	1.97	2.08	2.60
1R9H	Caenorhabditis elegans	118	13.46	1.60	1.59	2.41
1DXJ	Canavalia ensiformis (Jack bean)	242	17.03	1.84	1.59	2.53
1CNV	Canavalia ensiformis (Jack bean)	283	18.22	1.91	1.80	2.56
1CJK	Canis familiaris (Dog)	709	31.73	2.01	1.46	2.45
1J9Y	Cellvibrio japonicus (Bacteria)	337	18.97	1.98	1.56	2.59
1KKO	Citrobacter amalonaticus (Bacteria)	802	26.37	2.02	1.76	2.65

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
1RGY	Citrobacter freundii	360	19.71	1.82	1.61	2.61
1KEV	Clostridium beijerinckii	1404	32.80	2.05	1.78	2.65
129L	Coliphage t4	162	16.50	1.56	1.39	2.41
1A65	Coprinus cinereus (Inky cap fungus)	504	21.72	2.11	1.94	2.61
1BJK	Cyanobacterium anabaena	295	19.54	1.71	1.56	2.55
1IPE	Datura stramonium (Jimsonweed)	518	24.10	1.87	1.78	2.59
1B0P	Desulfovibrio africanus (Bacteria)	2462	38.56	2.09	1.77	2.76
1KEK	Desulfovibrio africanus (Bacteria)	2462	38.64	2.07	1.66	2.76
1LVK	Dictyostelium discoideum	743	29.38	1.82	1.53	2.59
1R18	Drosophila melanogaster (Fruit fly)	223	16.88	1.62	1.38	2.51
1JNE	Drosophila melanogaster (Fruit fly)	400	20.91	2.08	1.60	2.59
1HTY	Drosophila melanogaster (Fruit fly)	1014	29.82	2.22	1.70	2.66
1RI1	Encephalitozoon cuniculi (Fungus)	252	18.06	1.65	1.46	2.53
1RGZ	Enterobacter cloacae	370	19.48	1.88	1.63	2.62
1YIV	Equus caballus (Horse)	131	14.05	1.64	1.35	2.34
2FRF	Equus caballus (Horse)	152	15.29	1.58	1.31	2.39
1GJN	Equus caballus (Horse)	153	15.19	1.52	1.37	2.39
1GVZ	Equus caballus (Horse)	237	16.45	2.08	1.85	2.59
1B1X	Equus caballus (Horse)	689	29.47	1.86	1.66	2.49
2JHF	Equus caballus (Horse)	780	30.04	2.01	1.93	2.53
1GRJ	Escherichia coli (Bacteria)	151	21.61	1.72	1.42	2.29
1R67	Escherichia coli (Bacteria)	151	14.31	1.70	1.42	2.50
1HZT	Escherichia coli (Bacteria)	153	14.48	1.66	1.42	2.49
1RDA	Escherichia coli (Bacteria)	155	15.52	1.65	1.48	2.48
2DXA	Escherichia coli (Bacteria)	155	14.70	1.76	1.54	2.47
1RA9	Escherichia coli (Bacteria)	159	15.57	1.72	1.61	2.49
1RF7	Escherichia coli (Bacteria)	159	15.41	1.76	1.55	2.48
1K4N	Escherichia coli (Bacteria)	180	16.13	1.63	1.25	2.50
1FM0	Escherichia coli (Bacteria)	223	19.10	1.80	1.54	2.48
2GZS	Escherichia coli (Bacteria)	245	17.23	1.73	1.43	2.56
1R9L	Escherichia coli (Bacteria)	309	20.57	1.91	1.60	2.48

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
1A40	Escherichia coli (Bacteria)	321	19.93	1.88	1.59	2.57
1A54	Escherichia coli (Bacteria)	321	20.03	1.85	1.57	2.56
1MSK	Escherichia coli (Bacteria)	327	21.28	1.99	1.41	2.54
2HQ2	Escherichia coli (Bacteria)	331	20.61	2.05	1.54	2.53
1RI6	Escherichia coli (Bacteria)	333	18.67	1.92	1.54	2.57
1C4Q	Escherichia coli (Bacteria)	345	21.07	2.03	1.73	2.50
1USG	Escherichia coli (Bacteria)	345	22.31	1.92	1.67	2.46
1KHZ	Escherichia coli (Bacteria)	407	21.46	1.72	1.55	2.58
1R61	Escherichia coli (Bacteria)	415	22.03	1.97	1.65	2.57
2EX2	Escherichia coli (Bacteria)	456	24.84	2.08	1.67	2.50
1AOP	Escherichia coli (Bacteria)	460	21.73	1.90	1.60	2.59
1AYL	Escherichia coli (Bacteria)	532	23.40	2.08	1.63	2.61
1GVF	Escherichia coli (Bacteria)	548	25.76	1.91	1.62	2.55
2JG0	Escherichia coli (Bacteria)	554	22.66	1.84	1.53	2.66
1RQI	Escherichia coli (Bacteria)	598	24.37	1.94	1.98	2.60
1QOR	Escherichia coli (Bacteria)	652	28.84	1.90	1.72	2.49
1MXR	Escherichia coli (Bacteria)	678	25.90	2.19	1.44	2.64
1R65	Escherichia coli (Bacteria)	680	25.99	2.10	1.54	2.63
1RIB	Escherichia coli (Bacteria)	680	26.06	2.14	1.64	2.63
1RSV	Escherichia coli (Bacteria)	681	25.91	2.20	1.51	2.63
1DKG	Escherichia coli (Bacteria)	685	31.34	1.71	1.40	2.50
2DQ6	Escherichia coli (Bacteria)	865	28.07	2.03	1.65	2.63
1NEN	Escherichia coli (Bacteria)	1068	33.63	1.92	1.63	2.64
1JRQ	Escherichia coli (Bacteria)	1437	33.10	2.07	1.87	2.71
1CS1	Escherichia coli (Bacteria)	1532	33.43	2.09	1.92	2.66
1D8W	Escherichia coli (Bacteria)	1575	33.29	2.08	1.74	2.71
1RP7	Escherichia coli (Bacteria)	1602	33.30	2.04	1.69	2.73
1JR3	Escherichia coli (Bacteria)	1769	39.42	1.85	1.47	2.56
1GGJ	Escherichia coli (Bacteria)	2908	41.29	2.39	1.83	2.81
1CB8	Flavobacterium heparinum	674	27.50	1.97	1.66	2.61
1FOK	Flavobacterium okeanokoites	568	26.99	1.81	1.45	2.62

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
1CUS	Fusarium solani f. sp. pisi	197	15.25	1.86	1.72	2.55
1A26	Gallus gallus (Chicken)	351	20.90	1.86	1.70	2.52
1ROV	Glycine max (Soybean)	834	28.45	1.94	1.69	2.64
1IPJ	Glycine max (Soybean)	1088	32.97	2.15	1.90	2.55
1G12	Grifola frondosa (Fungi)	167	14.88	1.79	1.73	2.54
1NLZ	Helicobacter pylori 26695 (Bacteria)	1804	39.96	2.01	1.59	2.60
1R1K	Heliothis virescens (Noctuid moth)	477	22.78	1.80	1.27	2.54
1E2N	Herpes simplex virus	619	24.16	1.91	1.57	2.61
1BV7	Hiv-1	198	17.43	1.92	1.54	2.48
2CIA	Homo sapiens (Human)	121	13.27	1.66	1.36	2.46
1R2I	Homo sapiens (Human)	143	14.23	1.57	1.27	2.46
1LF7	Homo sapiens (Human)	164	15.34	1.74	1.47	2.44
1RM8	Homo sapiens (Human)	169	15.16	1.68	1.48	2.50
1IAP	Homo sapiens (Human)	190	17.47	1.57	1.25	2.46
1HDO	Homo sapiens (Human)	206	15.83	1.97	1.67	2.64
1ZD8	Homo sapiens (Human)	212	19.39	1.55	1.26	2.46
1REI	Homo sapiens (Human)	214	17.20	2.08	1.90	2.50
1A7S	Homo sapiens (Human)	221	16.42	1.93	1.77	2.56
1AE5	Homo sapiens (Human)	223	16.50	1.98	1.86	2.56
1SMO	Homo sapiens (Human)	223	18.71	1.62	1.38	2.46
1R5l	Homo sapiens (Human)	251	17.87	1.76	1.55	2.52
1RAY	Homo sapiens (Human)	258	17.49	1.94	1.55	2.55
1BKZ	Homo sapiens (Human)	270	20.22	1.98	1.72	2.48
1DWD	Homo sapiens (Human)	286	18.14	1.72	1.44	2.59
1RJB	Homo sapiens (Human)	298	19.20	1.81	1.45	2.56
2H14	Homo sapiens (Human)	303	17.83	2.17	1.86	2.56
1RKP	Homo sapiens (Human)	311	19.28	1.62	1.36	2.57
1ADS	Homo sapiens (Human)	315	18.94	1.82	1.49	2.57
2REN	Homo sapiens (Human)	320	19.73	1.86	1.68	2.57
1RYO	Homo sapiens (Human)	325	19.41	1.88	1.69	2.53
2NZL	Homo sapiens (Human)	351	19.07	1.82	1.67	2.61

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
1SO7	Homo sapiens (Human)	361	19.38	1.84	1.59	2.60
1AGD	Homo sapiens (Human)	375	22.94	1.91	1.57	2.49
1K3Y	Homo sapiens (Human)	442	21.69	1.71	1.30	2.55
2NW2	Homo sapiens (Human)	453	23.80	1.98	1.71	2.52
1R9O	Homo sapiens (Human)	455	22.45	2.00	1.55	2.58
1CPU	Homo sapiens (Human)	495	22.96	1.94	1.64	2.62
2O07	Homo sapiens (Human)	559	24.29	2.05	1.62	2.61
1RQ4	Homo sapiens (Human)	572	23.69	1.78	1.65	2.52
1HBA	Homo sapiens (Human)	574	23.57	1.84	1.56	2.53
1R1Y	Homo sapiens (Human)	574	23.41	1.81	1.61	2.53
1RPS	Homo sapiens (Human)	574	23.68	1.83	1.57	2.53
1RQ3	Homo sapiens (Human)	574	23.63	1.78	1.59	2.53
1E7E	Homo sapiens (Human)	582	27.83	1.60	1.36	2.50
1EER	Homo sapiens (Human)	592	28.45	1.74	1.28	2.51
1R4l	Homo sapiens (Human)	655	25.05	1.81	1.60	2.60
1DMT	Homo sapiens (Human)	696	26.38	1.98	1.82	2.54
1H2V	Homo sapiens (Human)	815	29.85	1.71	1.48	2.60
1KCW	Homo sapiens (Human)	1017	28.34	2.11	1.87	2.64
1KR2	Homo sapiens (Human)	1395	34.89	1.88	1.52	2.56
1KQO	Homo sapiens (Human)	1398	34.95	2.14	1.60	2.55
1IVH	Homo sapiens (Human)	1548	34.30	2.25	1.76	2.67
1RX0	Homo sapiens (Human)	1573	34.15	2.29	1.89	2.68
1EX1	Hordeum vulgare (Barley)	602	24.91	2.14	1.74	2.64
1BVW	Humicola insolens	360	19.19	1.88	1.73	2.58
1A39	Humicola insolens	410	20.74	1.94	1.78	2.60
2AYH	Hybrid	214	16.07	1.90	1.71	2.55
1EPX	Leishmania mexicana	1428	34.17	2.09	1.96	2.67
1GYP	Leishmania mexicana	1432	32.50	2.05	1.74	2.68
1RL9	Limulus polyphemus (Crab)	356	20.07	1.86	1.63	2.58
1GBE	Lysobacter enzymogenes	198	15.01	2.09	1.74	2.56
1EB8	Manihot esculenta (Cassava)	520	24.51	1.87	1.54	2.56

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
1R9C	Mesorhizobium loti (Bacteria)	243	17.68	1.59	1.20	2.51
1E6Y	Methanosarcina barkeri (Archaea)	2485	37.38	2.37	2.48	2.79
1NCJ	Mus musculus (Mouse)	214	26.79	1.68	1.48	2.16
1EJO	Mus musculus (Mouse)	443	24.32	2.06	1.69	2.48
1DQG	Mus musculus (Mouse)	134	13.44	1.96	1.90	2.47
1MD6	Mus musculus (Mouse)	154	14.96	1.82	1.60	2.46
2CZT	Mus musculus (Mouse)	155	14.94	1.66	1.49	2.46
1RUT	Mus musculus (Mouse)	160	23.62	1.76	1.51	2.18
1Z06	Mus musculus (Mouse)	165	14.89	1.56	1.36	2.51
1X1R	Mus musculus (Mouse)	169	15.38	1.70	1.38	2.50
1YZL	Mus musculus (Mouse)	172	15.02	1.67	1.36	2.53
1KN3	Mus musculus (Mouse)	180	15.26	1.83	1.49	2.53
1PQ1	Mus musculus (Mouse)	180	15.83	1.71	1.28	2.55
2E6M	Mus musculus (Mouse)	186	15.87	1.63	1.44	2.52
2ATF	Mus musculus (Mouse)	195	16.44	1.83	1.54	2.50
1Z0J	Mus musculus (Mouse)	220	17.24	1.66	1.26	2.52
1WNH	Mus musculus (Mouse)	224	18.24	1.85	1.42	2.51
2DTC	Mus musculus (Mouse)	227	18.99	1.84	1.58	2.50
1U2C	Mus musculus (Mouse)	228	20.75	1.77	1.50	2.44
1VET	Mus musculus (Mouse)	240	18.25	1.72	1.42	2.52
1IJY	Mus musculus (Mouse)	244	21.10	1.80	1.52	2.43
2J0A	Mus musculus (Mouse)	246	17.92	1.73	1.48	2.54
2GFH	Mus musculus (Mouse)	248	20.27	1.64	1.43	2.48
2HUO	Mus musculus (Mouse)	258	17.54	1.86	1.45	2.56
1RGX	Mus musculus (Mouse)	272	22.85	1.63	1.43	2.41
1KSH	Mus musculus (Mouse)	306	20.57	1.76	1.52	2.50
1ZCB	Mus musculus (Mouse)	318	21.29	1.73	1.54	2.52
1VJ1	Mus musculus (Mouse)	333	21.54	1.82	1.50	2.51
1RE8	Mus musculus (Mouse)	337	20.02	1.79	1.48	2.53
2GDG	Mus musculus (Mouse)	342	18.66	2.03	1.66	2.55
1RDQ	Mus musculus (Mouse)	360	19.91	1.88	1.42	2.58

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
2TNF	Mus musculus (Mouse)	444	21.01	2.06	1.69	2.62
1DKF	Mus musculus (Mouse)	449	22.36	1.75	1.39	2.53
1SH5	Mus musculus (Mouse)	464	26.21	1.80	1.47	2.51
1IG3	Mus musculus (Mouse)	497	23.88	1.93	1.56	2.57
2FR5	Mus musculus (Mouse)	540	22.32	2.02	1.65	2.64
2H3B	Mus musculus (Mouse)	922	28.90	2.07	1.70	2.62
1VM8	Mus musculus (Mouse)	955	32.67	1.82	1.56	2.55

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
2CXN	Mus musculus (Mouse)	1114	28.61	2.01	1.73	2.71
1MOP	Mycobacterium tuberculosis (Bacteria)	581	27.14	2.00	1.48	2.48
1K4Y	Oryctolagus cuniculus (Rabbit)	501	22.49	1.90	1.75	2.62
1LOX	Oryctolagus cuniculus (Rabbit)	647	27.49	1.95	1.54	2.60
1J0X	Oryctolagus cuniculus (Rabbit)	1324	31.88	2.14	1.72	2.67
1ADO	Oryctolagus cuniculus (Rabbit)	1452	35.16	2.16	1.65	2.69
1FIW	Ovis aries (Sheep)	274	17.52	1.98	1.75	2.60
1LHP	Ovis aries (Sheep)	615	25.53	2.00	1.60	2.62
1RFV	Ovis aries (Sheep)	615	25.58	1.93	1.71	2.62
1Q4G	Ovis aries (Sheep)	1110	31.35	1.98	1.91	2.64
1BXS	Ovis aries (Sheep)	1976	36.02	2.23	2.08	2.71
1X2I	Pyrococcus furiosus (hyperthermophile Archaea)	136	14.22	1.83	1.33	2.43
1SJ1	Pyrococcus furiosus (hyperthermophile Archaea)	147	16.21	1.86	1.60	2.34
1TWL	Pyrococcus furiosus (hyperthermophile Archaea)	172	15.06	1.78	1.51	2.51
1IM5	Pyrococcus furiosus (hyperthermophile Archaea)	182	15.38	1.75	1.62	2.49
1JG1	Pyrococcus furiosus (hyperthermophile Archaea)	215	16.33	1.77	1.41	2.51
1PRY	Pyrococcus furiosus (hyperthermophile Archaea)	226	17.44	1.82	1.46	2.52
1XI6	Pyrococcus furiosus (hyperthermophile Archaea)	234	16.88	1.68	1.48	2.53
1ELT	Pyrococcus furiosus (hyperthermophile Archaea)	236	16.37	1.91	1.71	2.57
1G3Q	Pyrococcus furiosus (hyperthermophile Archaea)	237	16.81	1.95	1.76	2.53
1F2T	Pyrococcus furiosus (hyperthermophile Archaea)	288	20.50	2.02	1.47	2.49
1XEW	Pyrococcus furiosus (hyperthermophile Archaea)	307	20.37	1.86	1.60	2.47

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
2IA0	Pyrococcus furiosus (hyperthermophile Archaea)	309	21.47	1.54	1.29	2.46
1NNQ	Pyrococcus furiosus (hyperthermophile Archaea)	340	20.02	1.82	1.50	2.54
1ELJ	Pyrococcus furiosus (hyperthermophile Archaea)	379	21.15	1.87	1.63	2.56
2P4W	Pyrococcus furiosus (hyperthermophile Archaea)	395	25.14	1.59	1.34	2.50
1XI3	Pyrococcus furiosus (hyperthermophile Archaea)	404	24.31	2.29	2.13	2.47
1DQ3	Pyrococcus furiosus (hyperthermophile Archaea)	454	24.85	1.77	1.54	2.56
1UA4	Pyrococcus furiosus (hyperthermophile Archaea)	454	21.13	1.98	1.66	2.62
1DQI	Pyrococcus furiosus (hyperthermophile Archaea)	496	21.10	2.22	1.70	2.62
1NNW	Pyrococcus furiosus (hyperthermophile Archaea)	502	24.20	1.83	1.83	2.51
1YQT	Pyrococcus furiosus (hyperthermophile Archaea)	507	23.64	1.82	1.61	2.58
1XI8	Pyrococcus furiosus (hyperthermophile Archaea)	534	24.32	1.81	1.70	2.55
1MJF	Pyrococcus furiosus (hyperthermophile Archaea)	540	24.40	2.12	1.87	2.60
2CFM	Pyrococcus furiosus (hyperthermophile Archaea)	545	24.97	2.00	1.26	2.54
2DFI	Pyrococcus furiosus (hyperthermophile Archaea)	602	25.52	1.85	1.65	2.58
1E19	Pyrococcus furiosus (hyperthermophile Archaea)	626	25.58	1.88	1.69	2.64
2CB0	Pyrococcus furiosus (hyperthermophile Archaea)	639	23.92	2.01	1.73	2.61
1PV9	Pyrococcus furiosus (hyperthermophile Archaea)	655	25.55	2.06	1.65	2.56
1II7	Pyrococcus furiosus (hyperthermophile Archaea)	665	30.98	1.94	1.54	2.47
1B43	Pyrococcus furiosus (hyperthermophile Archaea)	678	27.16	1.66	1.41	2.57
1Z26	Pyrococcus furiosus (hyperthermophile Archaea)	716	28.21	2.06	1.55	2.57
1AJ8	Pyrococcus furiosus (hyperthermophile Archaea)	741	26.42	1.98	1.63	2.67
1IOF	Pyrococcus furiosus (hyperthermophile Archaea)	832	29.09	2.02	2.14	2.57
1IQ8	Pyrococcus furiosus (hyperthermophile Archaea)	1154	33.46	2.13	1.67	2.61
1AOR	Pyrococcus furiosus (hyperthermophile Archaea)	1210	35.23	2.05	1.77	2.59
1IQP	Pyrococcus furiosus (hyperthermophile Archaea)	1949	42.63	1.83	1.49	2.56
2I14	Pyrococcus furiosus (hyperthermophile Archaea)	2334	42.15	2.13	1.75	2.60
1ION	Pyrococcus horikoshii (hyperthermophile Archaea)	234	17.06	1.87	1.64	2.52
1IU8	Pyrococcus horikoshii (hyperthermophile Archaea)	313	22.24	1.89	1.58	2.56
1JFL	Pyrococcus horikoshii (hyperthermophile Archaea)	456	23.24	2.04	1.82	2.53
1GDE	Pyrococcus horikoshii (hyperthermophile Archaea)	776	27.45	2.19	1.70	2.64
1LK5	Pyrococcus horikoshii (hyperthermophile Archaea)	916	28.32	2.21	2.12	2.68

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
1B8A	Pyrococcus kodakaraensis (hyperthermophile Archaea)	876	29.44	1.93	1.89	2.65
1AIS	Pyrococcus woesei (hyperthermophile Archaea)	374	25.32	1.80	1.39	2.54
1MXG	Pyrococcus woesei (hyperthermophile Archaea)	440	23.55	1.91	1.71	2.59
1HMU	Pedobacter heparinus (Bacteria)	674	27.49	1.95	1.67	2.61
1BVC	Physeter catodon	153	15.29	1.56	1.36	2.39
1KSI	Pivum sativum (Pea seedling)	1282	32.40	2.13	1.72	2.71
1D5C	Plasmodium falciparum	159	15.14	1.54	1.42	2.49
1Z6G	Plasmodium falciparum	191	16.91	1.57	1.36	2.48
2C0D	Plasmodium falciparum	353	20.56	1.72	1.55	2.52
1TV5	Plasmodium falciparum	373	19.39	2.07	1.63	2.61
1OKT	Plasmodium falciparum	422	25.34	1.75	1.41	2.44
2B4R	Plasmodium falciparum	1336	31.65	2.21	1.67	2.68
1CVR	Porphyromonas gingivalis (Bacteria)	432	21.17	2.09	1.68	2.62
3REQ	Propionibacterium freudenreichii shermanii	1345	33.46	2.15	1.69	2.62
1AQH	Pseudoalteromonas haloplanctis (Bacteria)	448	22.55	1.97	1.69	2.58
1IX1	Pseudomonas aeruginosa (Bacteria)	338	24.77	1.57	1.23	2.43
1R7O	Pseudomonas cellulosa (Bacteria)	362	19.25	1.90	1.69	2.57
1R7I	Pseudomonas mevalonii (Bacteria)	747	25.74	1.98	1.73	2.62
1R31	Pseudomonas mevalonii (Bacteria)	751	25.87	1.89	1.86	2.62
1RZ5	Pseudomonas nautica (Bacteria)	309	20.98	1.84	1.50	2.48
1RE5	Pseudomonas putida kt2440 (Bacteria)	1767	35.66	2.05	1.58	2.73
1OH0	Pseudomonas putida (Bacteria)	253	18.41	1.93	1.69	2.54
1X7D	Pseudomonas putida (Bacteria)	663	27.26	2.10	1.76	2.61
1ZOI	Pseudomonas putida (Bacteria)	824	26.99	1.93	1.76	2.63
1UWK	Pseudomonas putida (Bacteria)	1107	29.81	2.01	1.79	2.67
1J2T	Pseudomonas putida (Bacteria)	1542	33.14	2.18	1.80	2.73
1IQQ	Pyrus pyrifolia (Japanese pear)	200	16.72	1.83	1.58	2.52
1LFO	Rattus norvegicus (Rat)	126	14.04	1.67	1.38	2.31
2BWQ	Rattus norvegicus (Rat)	131	15.09	1.70	1.32	2.41
1RSY	Rattus norvegicus (Rat)	135	15.83	1.46	1.15	2.42
3RAB	Rattus norvegicus (Rat)	169	15.27	1.58	1.38	2.52

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
1ZIR	Rattus norvegicus (Rat)	178	16.99	2.00	1.53	2.52
2B5H	Rattus norvegicus (Rat)	186	16.26	1.81	1.56	2.51
1KLK	Rattus norvegicus (Rat)	203	16.83	1.76	1.42	2.52
1J02	Rattus norvegicus (Rat)	212	16.77	2.00	1.41	2.52
1XCL	Rattus norvegicus (Rat)	229	16.33	1.83	1.54	2.55
2J7Y	Rattus norvegicus (Rat)	235	17.76	1.63	1.31	2.52
1RJK	Rattus norvegicus (Rat)	250	17.93	1.75	1.29	2.53
1RK3	Rattus norvegicus (Rat)	250	18.02	1.73	1.40	2.52
1RKH	Rattus norvegicus (Rat)	253	17.99	1.75	1.30	2.53
1S50	Rattus norvegicus (Rat)	258	18.67	1.76	1.51	2.52
1F7Z	Rattus norvegicus (Rat)	269	18.54	2.06	1.77	2.57
1T27	Rattus norvegicus (Rat)	269	18.72	1.64	1.36	2.47
1A06	Rattus norvegicus (Rat)	279	20.02	1.70	1.43	2.48
1LC0	Rattus norvegicus (Rat)	290	19.94	1.69	1.35	2.51
1QHW	Rattus norvegicus (Rat)	300	17.90	1.91	1.74	2.62
2OVY	Rattus norvegicus (Rat)	306	19.19	1.60	1.39	2.56
1GFI	Rattus norvegicus (Rat)	313	20.78	1.62	1.41	2.54
1F3L	Rattus norvegicus (Rat)	321	21.48	1.87	1.46	2.51
2G8J	Rattus norvegicus (Rat)	322	20.82	1.93	1.79	2.54
1BD7	Rattus norvegicus (Rat)	347	20.20	2.16	1.75	2.58
1PD2	Rattus norvegicus (Rat)	398	21.07	1.62	1.43	2.57
1F20	Rattus norvegicus (Rat)	435	25.98	1.88	1.60	2.45
1ZCK	Rattus norvegicus (Rat)	445	23.42	1.97	1.31	2.50
1BU8	Rattus norvegicus (Rat)	446	25.05	1.97	1.72	2.56
1ZCJ	Rattus norvegicus (Rat)	459	22.44	1.87	1.60	2.55
1BCH	Rattus norvegicus (Rat)	462	26.52	1.56	1.29	2.47
1T1U	Rattus norvegicus (Rat)	597	25.15	1.81	1.51	2.62
1SQI	Rattus norvegicus (Rat)	685	27.88	1.99	1.66	2.61
1OM4	Rattus norvegicus (Rat)	818	29.57	1.79	1.56	2.58
1R4A	Rattus norvegicus (Rat)	864	35.87	1.81	1.46	2.43
1MJ3	Rattus norvegicus (Rat)	1548	31.93	2.13	1.73	2.69

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
1KY4	Rattus norvegicus (Rat)	1712	35.53	2.10	1.87	2.73
1BOL	Rhizopus niveus (Fungi)	222	17.38	1.81	1.63	2.50
1E18	Rhodobacter capsulatus	779	26.06	2.10	1.78	2.63
1RY5	Rhodobacter capsulatus	822	28.86	1.99	1.64	2.55
1RZH	Rhodobacter capsulatus	822	28.76	1.97	1.66	2.53
1RGN	Rhodobacter capsulatus	823	28.91	2.01	1.67	2.55
1RQK	Rhodobacter capsulatus	824	29.11	2.00	1.66	2.54
2FTX	Saccharomyces cerevisiae (Yeast)	146	15.38	1.64	1.34	2.40
1FUK	Saccharomyces cerevisiae (Yeast)	157	14.92	1.60	1.36	2.49
1KY3	Saccharomyces cerevisiae (Yeast)	162	15.29	1.61	1.44	2.47
1EK0	Saccharomyces cerevisiae (Yeast)	168	15.32	1.60	1.35	2.51
1JR8	Saccharomyces cerevisiae (Yeast)	210	16.78	1.57	1.42	2.49
1AKY	Saccharomyces cerevisiae (Yeast)	218	17.00	1.50	1.29	2.52
1G62	Saccharomyces cerevisiae (Yeast)	224	16.01	1.88	1.72	2.55
2AGK	Saccharomyces cerevisiae (Yeast)	233	17.14	1.71	1.93	2.50
1AEB	Saccharomyces cerevisiae (Yeast)	291	18.61	2.06	1.60	2.55
2EUT	Saccharomyces cerevisiae (Yeast)	291	18.45	1.97	1.54	2.56
2UY2	Saccharomyces cerevisiae (Yeast)	295	17.78	1.95	1.73	2.56
1A48	Saccharomyces cerevisiae (Yeast)	298	19.83	1.73	1.64	2.52
1RB7	Saccharomyces cerevisiae (Yeast)	304	18.90	1.85	1.72	2.55
1P6O	Saccharomyces cerevisiae (Yeast)	317	19.09	1.86	1.65	2.58
1I4W	Saccharomyces cerevisiae (Yeast)	322	21.54	1.71	1.40	2.52
1C02	Saccharomyces cerevisiae (Yeast)	332	22.87	1.69	1.33	2.52
1KA1	Saccharomyces cerevisiae (Yeast)	354	19.57	1.82	1.65	2.58
1G2Q	Saccharomyces cerevisiae (Yeast)	356	19.69	1.74	1.61	2.57
1DPJ	Saccharomyces cerevisiae (Yeast)	358	19.58	1.97	1.68	2.61
1T0I	Saccharomyces cerevisiae (Yeast)	362	21.04	1.77	1.42	2.58
1SQ9	Saccharomyces cerevisiae (Yeast)	378	20.11	1.97	1.54	2.58
1CI0	Saccharomyces cerevisiae (Yeast)	409	21.92	1.70	1.54	2.59
1AC5	Saccharomyces cerevisiae (Yeast)	483	22.14	2.00	1.56	2.62
1NEY	Saccharomyces cerevisiae (Yeast)	492	24.38	1.99	1.60	2.63

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
1VKH	Saccharomyces cerevisiae (Yeast)	498	24.69	1.81	1.37	2.57
1TXN	Saccharomyces cerevisiae (Yeast)	503	23.10	1.84	1.89	2.58
1R5T	Saccharomyces cerevisiae (Yeast)	554	23.06	2.05	1.52	2.61
1IG0	Saccharomyces cerevisiae (Yeast)	636	25.62	1.90	1.63	2.60
2DH4	Saccharomyces cerevisiae (Yeast)	652	27.22	1.68	1.70	2.56
1P43	Saccharomyces cerevisiae (Yeast)	872	26.54	2.04	1.86	2.63
1RSG	Saccharomyces cerevisiae (Yeast)	950	33.11	1.88	1.58	2.59
1M0W	Saccharomyces cerevisiae (Yeast)	993	33.04	1.96	1.69	2.60
1M2O	Saccharomyces cerevisiae (Yeast)	1758	46.45	1.98	1.91	2.53
1JIO	Saccharopolyspora erythraea	403	21.02	1.75	1.49	2.55
1JIP	Saccharopolyspora erythraea (Bacteria)	403	21.06	1.81	1.42	2.55
1FYE	Salmonella typhimurium (Bacteria)	220	16.48	1.70	1.56	2.54
1LST	Salmonella typhimurium (Bacteria)	239	17.69	1.78	1.59	2.52
2PKW	Salmonella typhimurium (Bacteria)	246	18.39	1.72	1.41	2.50
1AF7	Salmonella typhimurium (Bacteria)	274	20.87	1.84	1.29	2.48
1VPD	Salmonella typhimurium (Bacteria)	279	20.24	1.68	1.49	2.47
2OBW	Salmonella typhimurium (Bacteria)	280	18.35	1.71	1.45	2.59
1SBP	Salmonella typhimurium (Bacteria)	309	19.39	1.75	1.61	2.56
2AP1	Salmonella typhimurium (Bacteria)	312	20.92	1.75	1.45	2.49
1MDO	Salmonella typhimurium (Bacteria)	355	20.71	1.73	1.53	2.56
1DZR	Salmonella typhimurium (Bacteria)	367	21.62	1.93	1.74	2.55
1LH0	Salmonella typhimurium (Bacteria)	419	23.12	1.82	1.49	2.57
1T33	Salmonella typhimurium (Bacteria)	434	22.67	1.71	1.35	2.59
1K38	Salmonella typhimurium (Bacteria)	475	23.56	1.83	1.52	2.57
1JXH	Salmonella typhimurium (Bacteria)	496	23.20	1.91	1.62	2.61
2RKM	Salmonella typhimurium (Bacteria)	517	23.23	1.97	1.80	2.62
1B4Z	Salmonella typhimurium (Bacteria)	520	22.94	2.03	1.75	2.63
1CBU	Salmonella typhimurium (Bacteria)	539	24.93	1.76	1.52	2.57
2R2F	Salmonella typhimurium (Bacteria)	571	25.56	1.90	1.52	2.58
1OAS	Salmonella typhimurium (Bacteria)	630	24.60	2.03	1.76	2.60
2HI1	Salmonella typhimurium (Bacteria)	635	30.21	2.06	1.62	2.50

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
1CNZ	Salmonella typhimurium (Bacteria)	726	27.22	2.06	1.97	2.59
1JYO	Salmonella typhimurium (Bacteria)	726	37.56	1.92	1.54	2.36
1Z5G	Salmonella typhimurium (Bacteria)	835	30.93	1.98	1.97	2.54
1XR4	Salmonella typhimurium (Bacteria)	988	31.37	2.13	1.81	2.60
2FUV	Salmonella typhimurium (Bacteria)	1088	34.59	2.16	1.76	2.54
1T3T	Salmonella typhimurium (Bacteria)	1283	30.35	2.11	1.75	2.69
2AFA	Salmonella typhimurium (Bacteria)	2394	40.07	2.03	1.72	2.73
1RKM	Salmonella typhimurium	519	23.85	1.95	1.72	2.58
1ENF	Staphylococcus aureus (Bacteria)	222	17.41	1.86	1.65	2.54
1QWZ	Staphylococcus aureus (Bacteria)	235	17.99	1.62	1.50	2.53
2CCJ	Staphylococcus aureus (Bacteria)	391	25.24	1.83	1.45	2.50
1GQ6	Streptomyces clavuligerus (Bacteria)	888	27.59	2.10	1.80	2.61
3PTE	Streptomyces sp.	347	18.93	1.80	1.69	2.60
1NUY	Sus scrofa (Pig)	328	19.54	1.93	1.67	2.58
1ALV	Sus scrofa (Pig)	346	20.62	1.79	1.41	2.54
2GSR	Sus scrofa (Pig)	416	20.98	1.75	1.64	2.53
1QD1	Sus scrofa (Pig)	668	29.45	1.98	1.82	2.47
1EVI	Sus scrofa (Pig)	680	29.85	2.13	1.74	2.59
1EUD	Sus scrofa (Pig)	698	27.59	2.02	1.67	2.60
1B0J	Sus scrofa (Pig)	753	25.67	2.05	1.73	2.66
1FP3	Sus scrofa (Pig)	804	28.61	1.92	1.59	2.62
1LWD	Sus scrofa (Pig)	826	28.15	1.87	1.58	2.59
1E7U	Sus scrofa (Pig)	872	29.36	1.87	1.56	2.61
2BKH	Sus scrofa (Pig)	924	34.24	1.94	1.57	2.55
1JS3	Sus scrofa (Pig)	928	27.56	2.13	1.62	2.68
1ORV	Sus scrofa (Pig)	2912	53.17	1.98	1.74	2.52
1IFH	Synthetic construct	436	25.27	2.05	1.65	2.43
1CRL	Synthetic construct	534	22.11	2.00	1.79	2.62
1TMY	Thermotoga maritima (hyperthermophile bacteria)	118	12.95	1.70	1.37	2.43
1B8Z	Thermotoga maritima (hyperthermophile bacteria)	134	14.13	1.70	1.30	2.36
1GUI	Thermotoga maritima (hyperthermophile bacteria)	155	14.89	1.60	1.36	2.47

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
1J5Y	Thermotoga maritima (hyperthermophile bacteria)	167	18.90	1.79	1.58	2.34
1DMG	Thermotoga maritima (hyperthermophile bacteria)	172	16.48	1.61	1.49	2.46
1DD5	Thermotoga maritima (hyperthermophile bacteria)	184	22.19	1.68	1.26	2.28
1I82	Thermotoga maritima (hyperthermophile bacteria)	189	15.34	2.08	1.63	2.52
1I5D	Thermotoga maritima (hyperthermophile bacteria)	190	17.55	1.44	1.24	2.42
1L9G	Thermotoga maritima (hyperthermophile bacteria)	191	16.08	1.88	1.71	2.50
1K9V	Thermotoga maritima (hyperthermophile bacteria)	200	15.81	2.09	1.59	2.54
1XKR	Thermotoga maritima (hyperthermophile bacteria)	205	16.96	1.64	1.42	2.48
1KGS	Thermotoga maritima (hyperthermophile bacteria)	211	21.18	1.51	1.30	2.35
1P2F	Thermotoga maritima (hyperthermophile bacteria)	212	19.45	1.67	1.40	2.43
1O4T	Thermotoga maritima (hyperthermophile bacteria)	230	17.09	1.84	1.59	2.53
1O0X	Thermotoga maritima (hyperthermophile bacteria)	249	17.18	1.87	1.85	2.53
1J6O	Thermotoga maritima (hyperthermophile bacteria)	259	17.34	1.85	1.51	2.56
2ETH	Thermotoga maritima (hyperthermophile bacteria)	271	20.27	1.50	1.09	2.40
1VHN	Thermotoga maritima (hyperthermophile bacteria)	299	19.85	1.97	1.59	2.53
1J5X	Thermotoga maritima (hyperthermophile bacteria)	312	19.29	1.69	1.46	2.53
1VHO	Thermotoga maritima (hyperthermophile bacteria)	313	21.81	1.88	1.61	2.50
1GXJ	Thermotoga maritima (hyperthermophile bacteria)	323	21.48	1.79	1.41	2.43
1D1G	Thermotoga maritima (hyperthermophile bacteria)	328	21.91	1.96	1.56	2.55
1CZ3	Thermotoga maritima (hyperthermophile bacteria)	332	21.84	2.03	1.56	2.56
1JCF	Thermotoga maritima (hyperthermophile bacteria)	336	20.77	1.91	1.67	2.56
1WOS	Thermotoga maritima (hyperthermophile bacteria)	361	20.49	2.11	2.01	2.52
1I58	Thermotoga maritima (hyperthermophile bacteria)	364	23.47	1.72	1.32	2.44
1KQ3	Thermotoga maritima (hyperthermophile bacteria)	364	20.70	2.03	1.83	2.57
1E4F	Thermotoga maritima (hyperthermophile bacteria)	378	23.28	1.86	1.55	2.54
1VPE	Thermotoga maritima (hyperthermophile bacteria)	398	22.45	2.06	1.57	2.55
1O20	Thermotoga maritima (hyperthermophile bacteria)	412	24.67	1.92	1.59	2.46
1J6U	Thermotoga maritima (hyperthermophile bacteria)	422	22.27	1.90	1.67	2.56
1KUT	Thermotoga maritima (hyperthermophile bacteria)	427	26.80	1.54	1.36	2.45
1O0W	Thermotoga maritima (hyperthermophile bacteria)	477	27.30	1.66	1.29	2.50
1ZY9	Thermotoga maritima (hyperthermophile bacteria)	522	22.88	2.02	1.68	2.63

Protein	Source	N	Rg	d_s 7	d_s 6	d_f
1MRZ	Thermotoga maritima (hyperthermophile bacteria)	536	27.63	1.97	1.50	2.48
1J5W	Thermotoga maritima (hyperthermophile bacteria)	542	22.69	1.93	1.42	2.61
1GJW	Thermotoga maritima (hyperthermophile bacteria)	636	26.11	1.98	1.78	2.61
1O14	Thermotoga maritima (hyperthermophile bacteria)	638	30.55	1.80	1.62	2.43
1EG5	Thermotoga maritima (hyperthermophile bacteria)	707	27.15	2.02	1.66	2.63
1O12	Thermotoga maritima (hyperthermophile bacteria)	708	30.00	2.09	1.67	2.49
2ORD	Thermotoga maritima (hyperthermophile bacteria)	797	25.96	2.20	1.77	2.67
1ZOR	Thermotoga maritima (hyperthermophile bacteria)	814	29.28	1.95	1.62	2.57
1O26	Thermotoga maritima (hyperthermophile bacteria)	876	28.91	1.82	1.63	2.68
1LWJ	Thermotoga maritima (hyperthermophile bacteria)	882	30.87	1.90	1.66	2.58
1VLH	Thermotoga maritima (hyperthermophile bacteria)	947	28.35	1.91	1.34	2.60
1INL	Thermotoga maritima (hyperthermophile bacteria)	1151	34.79	2.27	1.78	2.54
1VJ0	Thermotoga maritima (hyperthermophile bacteria)	1459	33.18	2.16	1.75	2.74
1A47	Thermoanaerobacterium thermosulfurigenes	683	25.53	2.08	1.74	2.62
1D2M	Thermus thermophilus	552	27.17	1.86	1.66	2.53
1KOR	Thermus thermophilus	1538	34.06	2.06	1.98	2.70
1E3Q	Torpedo californica (Pacific electric ray)	532	22.83	2.01	1.63	2.66
1KUF	Trimeresurus mucrosquamatus	201	16.10	1.82	1.71	2.54
16PK	Trypanosoma brucei	415	23.14	1.82	1.62	2.54
1I13	Trypanosoma cruzi	380	20.06	1.77	1.57	2.59
1MZ5	Trypanosoma rangeli	622	27.10	2.06	1.66	2.55
1NAR	Vicia narbonensis	289	18.35	1.88	1.50	2.55

References

[1] Steinbach, P. J., Ansari, A., Berendzen, J., Braunstein, D., Chu, K., Cowen, B. R., Ehrenstein, D., Frauenfelder, H., Johnson J. B., Lamb D. C. Ligand binding to heme proteins: connection between dynamics and function. Biochemistry, 30: 3988–4001, (1991).

[2] Rasmussen, B. F., Stock, A. M., Ringe, D., Petsko, G. A. Crystalline ribonuclease A loses function below the dynamical transition at 220 K. Nature 357, 423–424 (1992).

[3] Enright, M. B., Leitner, D. M. Mass fractal dimension and the compactness of proteins. Phys. Rev. E 71, 011912 (2005).

[4] Burioni, R., Cassi, D., Cecconi, F., Vulpiani, A. Topological thermal instability and length of proteins. Proteins 55, pp. 529–535 (2004).

[5] For experimental evidence, see: Lushnikov, S., Svanidze, A., Sashin, I. Vibrational density of states of hen egg white lysozyme. Journal of Experimental and Theoretical Physics Letters 0021-3640, vol: 82 (1) p:30-33 (2005); Stapleton, H. J., Allen, J. P., Flynn, C. P., Stinson, D. G., Kurtz, S. R., Fractal form of proteins. Phys Rev Lett; 45: 1456-1459 (1980).

[6] Granek, R., Klafter, J. Fractons in Proteins: Can They Lead to Anomalously Decaying Time Autocorrelations? Phys. Rev. Lett. 95, 098106(1)-098106(4) (2005).

[7] Molecular dynamics: Fixed on fractals, Nature Research Highlights 437, 172-173 (8 September 2005)

[8] Burioni, R., Cassi D., Fontana, M. P., Vulpiani, A. Vibrational thermodynamic instability of recursive networks. Europhysics Lett; 58: 06-810, (2002).

[9] Peierls R., Helv. Phys. Acta, 2 81, (1934).

[10] Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M. The protein data bank: a computer-based archival file for macro-molecular structure. J. Mol. Biol., 112, 535–542 (1977).

[11] Sumner, J. B. The Isolation and Crystallization of the Enzyme Urease. J Biol Chem 69: 435-41 (1926).

[12] Lehninger Principles of Biochemistry, Fourth Edition by David L. Nelson and Michael M. Cox

[13] Koolman, Color Atlas of Biochemistry, 2nd edition © 2005 Thiem

[14] Muirhead, H., Perutz, M. Structure of hemoglobin. A three-dimensional fourier synthesis of reduced human hemoglobin at 5.5A resolution. Nature 199 (4894): 633-8 (1963).

[15] Kendrew, J., Bodo, G., Dintzis, H., Parrish, R., Wyckoff, H., Phillips, D. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 181 (4610): 662-6 (1958).

[16] Mandelbrot, B. B. The Fractal Geometry of Nature. W. H. Freeman and Company (1982).

[17] Fractal Geometry: Mathematical Foundations and Application. Second Edition Kenneth Falconer 2003 John Wiley & Sons, Ltd.

[18] Alexander, S. and Orbach, R. Density of states of fractals: 'fractons'. J. Phys. (Paris) Lett. 43, L625 (1982).

[19] Rammal, R. and Toulouse, G. Random walks on fractal structures and percolation clusters. J. Phys. (Paris) Lett. 44, L13 (1983)

[20] Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems. Qiang Cui (Editor) and Ivet Bahar (Editor) 2005 Chapman & Hall/CRC.

[21] For a general review see: Tozzini, V. Coarse-grained models for proteins. Curr Opin Struct Biol 2005, 15, 144.

[22] Bahar, I., Atilgan, A. R., Erman, B. Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential. Fold. Des. ,2,173, (1997).

[23] Miyazawa, S. and Jernigan, R. L. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules, 18, 534, (1985).

[24] Bahar, I. and Jernigan, R. L. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J. Mol. Biol., 266, 195, (1997).

[25] Yang, L., Eyal, E., Chennubhotla, C., Jee, J., Gronenborn, A., Bahar, I. Insights into Equilibrium Dynamics of Proteins from Comparison of NMR and X-ray Data with Computational Predictions. Structure, Volume 15, Issue 6, Pages 741-749, (2007).

[26] De Gennes, P. G. Scaling Concepts in Polymer Physics (Cornell Univ. Press, Ithaca, New York, 1979).

[27] Amorphous Solids, Low Temperature Properties. Edited by Phillips W. A. (Springer-Verlag, Berlin, 1981).

[28] Ciliberti, S., De Los Rios, P. and Piazza, F. Glasslike Structure of Globular Proteins and the Boson Peak. Phys. Rev. Lett. 96 , 198103 (2006).

[29] Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajimi, K., Sarai A. ProTherm version 4.0: thermodynamic database for proteins and mutants. Nucleic Acids Res. 32:D120–D121, (2004).

[30] Marti-Renom, M. A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., Sali, A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29: 291-325. (2000)

[31] Madigan, M., Martinko, J. (editors). Brock Biology of Microorganisms, 11th ed., Prentice Hall (2005).

[32] Berezovsky, I. N., Tumanyan, V. G. and Esipova, N. G. Representation of amino acid sequences in terms of interaction energy in protein globules. FEBS Lett. 418, 43–46. (1997)

[33] Schumann, J., Bohm, G., Schumacher, G., Rudolph, R. and Jaenicke, R. Stabilization of creatinase from Pseudomonas putida by random mutagenesis. Protein Sci. 2, 1612–1620. (1993)

[34] Jaenicke, R. and Bohm, G. The stability of proteins in extreme environments. Curr. Opin. Struct. Biol. 8, 738–748. (1998)

[35] Querol, E., Perez-Pons, J. A. and Mozo-Villarias A. Analysis of protein conformational characteristics related to thermostability. Protein Eng. 9, 265–271. (1996)

[36] Vetriani, C., Maeder, D. L., Tolliday, N., Yip, K. S., Stillman, T. J., Britton, K. L., Rice, D. W., Klump, H. H. and Robb, F. T. Protein thermostability above 100°C: A key role for ionic interactions. Proc. Natl. Acad. Sci. USA 95, 12300–12305. (1998)

[37] Hurley, J. H., Baase, W. A. and Matthews, B. W. Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. J. Mol. Biol. 224, 1143–1159. (1992)

[38] Thompson, M. J. and Eisenberg, D. Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. J. Mol. Biol. 290, 595–604. (1999)

[39] Hartl, F. U. Molecular chaperones in cellular protein folding. Nature 381, 571–580 (1996).

[40] Hartl F. U. and Mayer-Hartl M. Molecular chaperones in the cytosol: from nascent chain to folded protein. Science 295, 1852−1858 (2002).

[41] Sigler, P. B., Xu, Z. H., Rye, H. S., Burston, S. G., Fenton, W. A., Horwich, A. L. Structure and function in GroEL-mediated protein folding. Annu., Rev. Biochem. 67, 581–608 (1998).

[42] A. van der Vaart, Ma, J., Karplus, M. The unfolding action of GroEL on a protein substrate, Biophys. J. 87, pp. 562–573 (2004).

[43] Stan, G., Lorimer, G. H., Thirumalai, D., Brooks, B. R. Coupling between allosteric transitions in GroEL and assisted folding of a substrate protein. PNAS ;104(21):8803-8, 2007 May 22.

[44] Feng, S., Crossover in spectral dimensionality of elastic percolation systems. Phys. Rev. B 32,5793-5797, (1985).

[45] Webman, I., Grest, G. S., Dynamical behavior of fractal structures. Phys. Rev. B 31, 1689 - 1692, (1985).

[46] Alexander, S., Vibrations of fractals and scattering of light from aerogels. Phys. Rev. B 40, 7953 - 7965, (1989).