BAR ILAN UNIVERSITY

Computational studies of protein chaperone mediated folding interactions

ETAI JACOB

Submitted in partial fulfillment of the requirements for the Master's Degree in the Mina and Everard Goodman Faculty of Life Sciences

Ramat Gan, Israel **2007**

This work was carried out under the supervision of

Prof. Ron Unger

Mina and Everard Goodman faculty of life sciences
Bar Ilan University.

ACKNOWLEDGEMENTS

First, I would like to thank Prof. Ron Unger, my supervisor, for his support, understanding, great efforts and advices; for enabling me the best conditions to learn, research and to accomplish; for teaching me scientific criticism and substantial research; for placing borders and opening new horizons; for his wonderful openness and honesty; I have learnt a lot from my supervisor, from values of scientific integrity to persistence and for that I am deeply grateful.

Next, I would like to thank Prof. Amnon Horovitz, for great ideas and optimism; for valuable advices and for introducing me to new perspectives.

Also, I would like to thank Inna Myslyuk for the beautiful art work in the papers and this thesis and for her support.

Last but not least, I would like to thank my wonderful parents for their support and my beloved wife Michal, for her great support, endless patience and understanding.

PREFACE

This thesis is based on the following two papers. I was invited to talk about the second paper in the fifth European Conference on Computational Biology (ECCB 2006) in Eilat.

Etai Jacob, Amnon Horovitz and Ron Unger (2007) Different mechanistic requirements for prokaryotic and eukaryotic chaperonins: a lattice study. Submitted to ISMB/ECCB 2007, ***Bioinformatics***.

Etai Jacob and Ron Unger (2007) A tail of two tails: Why are terminal residues of proteins exposed? ***Bioinformatics*** **23**, 225-230.

Contents

IV. Structural Features of Protein Termini

V. References

VI. Appendix A - GRID report

Appendix B - *Different mechanistic requirements for prokaryotic and eukaryotic chaperonins: a lattice study.*
Etai Jacob, Amnon Horovitz and Ron Unger
Submitted to ISMB/ECCB 2007, *Bioinformatics*.

Appendix C - *A tail of two tails: Why are terminal residues of proteins exposed?*
Etai Jacob and Ron Unger (2007)
Bioinformatics **23**, 225-230.

ABSTRACT

Understanding how proteins fold is a major challenge in modern biology. The transformation of one dimensional polypeptide chain into three dimensional functional structure depends on the efficiency and accuracy of the process of protein folding. Polypeptide chains carry all the information required to fold to their native functional three dimensional structure, and do not require any additional molecules to direct the folding process. However, in vivo, inside the complex milieu of a living cell, protein folding is facilitated by proteins called chaperones whose role is especially essential under stress conditions such as heat shock. In my thesis I approached the folding process of proteins on these two sides; the folding of a protein by itself and the folding of a protein with the aid of molecular chaperones.

Studying the fundamental questions underlying the phenomenon of protein folding has been facilitated by the introduction of simple folding models simulations. Simple, or even abstract, models of protein folding, while ignoring many of the small details of this process, are very useful for elucidating general principles regarding protein folding. Examples are given in the two parts of this thesis. Clearly, simple model experiments can not be used to definitively prove specific folding phenomena, but conclusions from simple models can certainly be used to promote and critically assess ideas about protein folding mechanisms. In this research I used one of the major simple folding models - the lattice model. In this model, the polypeptide chains in the simulations are modeled as a linear sequence of residues on a 2D lattice, where a reduced number of residues types are used and interaction potentials between residues reflect the average strength of interactions in empirical mean force potentials.

In the first part of my thesis I investigated the folding process of proteins with the aid of two models of molecular chaperones; prokaryotic and eukaryotic. One major class of chaperones, called chaperonins, comprises ATP-dependent proteins that facilitate folding by binding the assisted protein in a cavity formed at each end of their double ring structure. The prokaryotic GroEL/GroES complex in E. coli is the best characterized chaperonin complex. GroEL consists of two rings each formed by seven identical protein subunits. GroES is a single-ring heptamer that binds to GroEL in the presence of ATP and serves as the cap of the cavity formed by each ring structure. Each GroEL subunit

can rotate and thus turn a different surface towards the inner cavity. The homologue eukaryotic chaperonin to the prokaryotic chaperonin GroEL is called CCT (also called TCP-1 ring complex) which is composed of eight similar (but not identical) subunits. Even though chaperonins are investigated for many years, how exactly they facilitate folding is still unclear. In this research, I approach and try to clarify this issue.

One major difference between the prokaryotic and eukaryotic chaperonins is the coordination between the surface changes of the subunits. While in the GroEL/GroES system the change is concerted, i.e. all subunits switch simultaneously, it was recently shown that in CCT the change is sequential, i.e. the subunits switch conformation one after the other. Concerning the substrates, approximately 70% of proteins in eukaryotic cells are multi-domain whereas in prokaryotes single-domain proteins are more common. Thus, it was suggested that the different modes of action of prokaryotic and eukaryotic chaperonins can be explained by the need of eukaryotic chaperones to facilitate folding of multi-domain proteins. Through the course of my thesis, I examined possible implications of these different allosteric mechanisms for the assistance ability of these chaperonins to folding process of proteins.

In order to enable the simulations of the protein folding process and the protein chaperon interactions in a simple and abstract model, I have developed a computational engine that can be used to simulate many types of interactions between a folding protein and a chaperone on a lattice. Using this engine, it is possible to specify many parameters of the system including the chaperonin's shape, size, surface composition, the way the chaperonin changes its surface, the strength of interactions between amino acids etc. In addition, this engine can generate different substrates; single-domain proteins, double-domain proteins, with different length and with various kinetic, structural or content features.

Using this software simulation engine, I found that the folding yields of single-domain protein substrates are greater when the chaperonin undergoes concerted and not sequential conformational changes. In contrast, the folding yields of double-domain proteins are greater in the presence of a chaperonin that undergoes sequential conformational changes. These results are consistent with the observation that large multi-domains proteins are more common in eukaryotes compared with prokaryotes.

Hence, they support the suggestion that the different allosteric mechanisms of prokaryotic and eukaryotic chaperones can be explained by the need of eukaryotic chaperones to facilitate folding of proteins with a multi-domain structure. A manuscript of this research and its conclusions was submitted to a leading scientific journal (*Bioinformatics*).

In the second part of my thesis I approached the folding process of a protein by itself and its linkage to the structural features of protein termini. It is widely known that terminal residues of proteins (i.e. the N and C termini) are predominantly located on the surface of proteins and exposed to the solvent. However, there is no good explanation as to the forces driving this phenomenon. The common explanation that terminal residues are charged, and charged residues prefer to be on the surface, can not explain the magnitude of the phenomenon. Using structural bioinformatics methods, I surveyed a large number of proteins from the protein data bank (PDB) in order to explore, quantitatively, the extend of this phenomenon, and then I used the lattice model to study the mechanisms involved by demonstrating that a series of constraints that affect proteins, had lead to the preference of terminal residues to be located on the surface. I suggested (by using same software engine platform) that the tendency of terminal residues of proteins to be located on the surface is a result of thermodynamic and kinetic evolutionary selection processes. This part of my thesis was published in *Bioinformatics* and I was invited to orally present it in the 5th European Conference on Computational Biology.

Jointly, the results of this study - two manuscripts comprising this thesis - form a valuable contribution to the current knowledge in the field of protein-chaperone interactions and structural features of proteins. Furthermore, this thesis introduces an innovative method of investigation of multi-domain protein folding with lattice model simulations.

CHAPTER I

Introduction

1.1. PROTEIN FOLDING

The Protein

A protein is synthesized on the ribosome as a linear sequence of amino acid residues. To guarantee its function, the protein must fold during and following synthesis to take up its native conformation which gives it the ability to function. In-spite of the fact that a protein spends most of its life in its native conformation, the native conformation is only marginally stable. Modest changes in the protein's environment, like increasing or decreasing the acidity level, heat stress, heavy metals or organic stress, can cause denaturation which results in protein functionality reduction and even aggregation. Figure 1(a) describes the effect of slow heating on a protein. The graph shows that the protein structure remains intact while heating, until an abrupt deterioration in structure occurs. Similar results appear when increasing the acidity of the environment as shown in figure 1(b). It should be mentioned that both graphs curves are sigmoids, which demonstrates the cooperatively in the fast denaturation process. This cooperative behavior can be explained by the fact that the loss of structure in one part of the protein may cause destabilization in other parts.

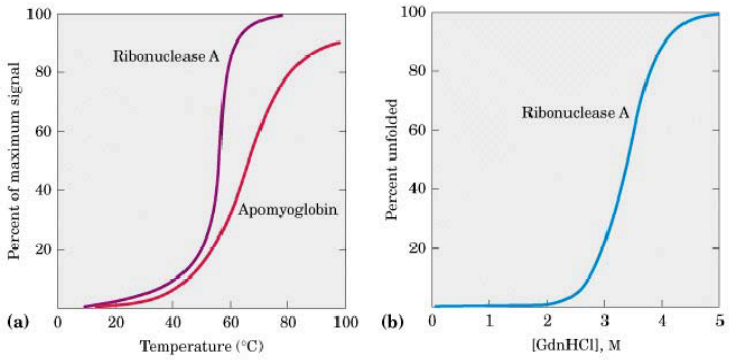

Fig. 1 Protein denaturation. (a) Thermal denaturation of horse apomyoglobin and ribonuclease A. **(b)** Denaturation of disulfide-intact ribonuclease A by guanidine hydrochloride.[1]

During the 1950's, the future Nobel prize winner, Christian Anfinsen, was the first to provide evidence to the fact that all the information needed to fold a protein into its native tertiary structure is contained in its amino acid sequence (Anfinsen, 1973). Figure 2 on the right, describes the classic experiment of a denaturation-renaturation which illustrated this concept.

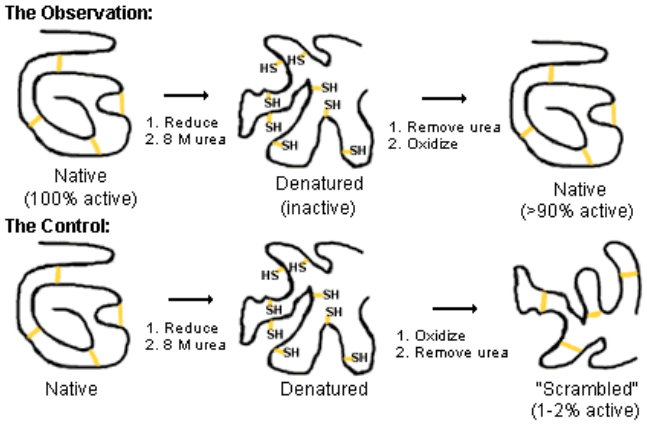

Fig. 2 Renaturation of unfolded denatured ribonuclease.

Protein Folding

The protein folding is the process by which the newly synthesized sequence of amino acids can find its way by folding to the ultimate functional conformation – its native structure. Let us assume a simple reduced model of a protein as a molecule of 100 amino acid residues and conservatively define that each residue can take up 2 different conformations. We find that there are still 2^{100} possible chain conformations even though the calculations are based on the simple reduced model. Hence, if we assume that conformational transitions within the chain can occur as rapidly as laws of physics (~10^{11} s^{-1}), it would take 10^{11} years for a protein to systematically search all its possible conformations (Dinner et al., 2000). However, folding of a protein of longer size than 100 amino acid residues to the lowest energy (native) structure in nature, takes a short period of time. This contradiction has become known as the 'Levinthal Paradox' (Karplus, 1997; Dobson et al., 1999). To understand how a protein can find its native state in a reasonable time, it was suggested that the folding process can be considered as a free energy funnel as follows: The protein commence in one of the unfolded states which are characterized by a high degree of conformational entropy. As the protein proceeds down the funnel, where both entropy and free energy decrease, it can fall into traps which are semi-stable intermediates (local minima) that can slow down the folding process. The protein ends its folding journey in the native state which sits on the bottom of the funnel beneath few folding intermediates (Lehninger; Wolynes, 1995) (see figure 3).

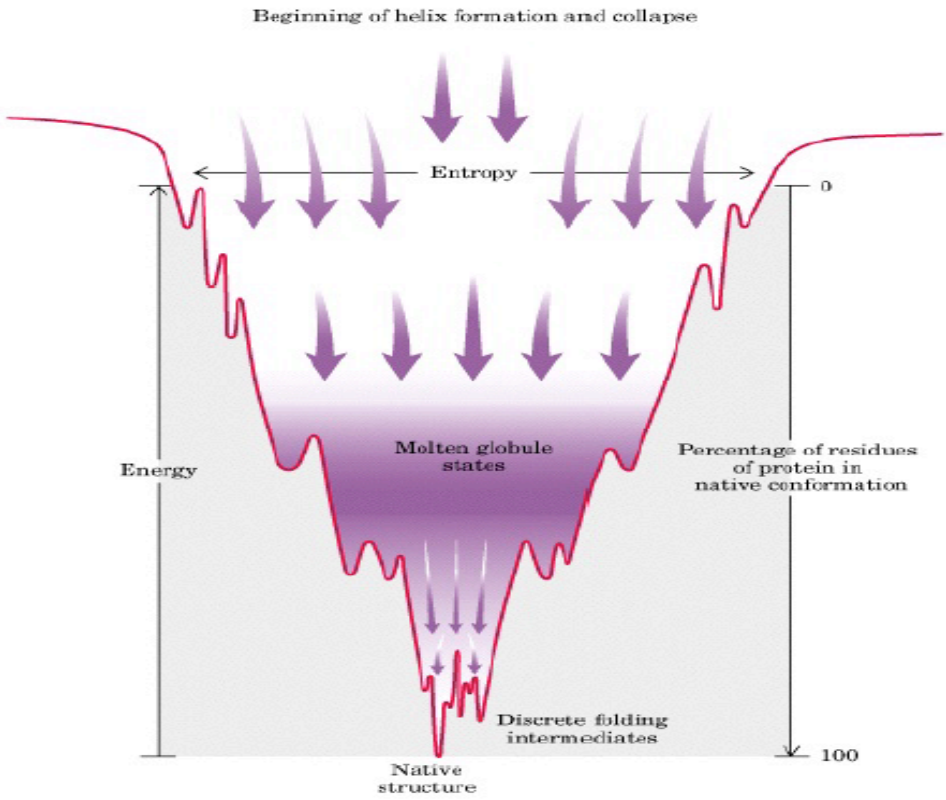

Fig. 3 Schematic of the folding funnel. The width of the funnel represents entropy, and its depth represents the energy. Depressions on the side of the funnel represent semi stable folding intermediates which could slow down the folding process (Fig. is taken from Lehninger).

1.2. PROTEINS AND CHAPERONES

Aiding the folding process

Anfinsen's classical experiment of denaturing and renaturing a ribonuclease was a result of spontaneous folding process of a protein (Anfinsen, 1973). Many other proteins face difficulties such as environmental stress (to be specified later) or low folding ability (RuBisCo and rhodanese - Yokota et al., 2000) which cause destabilization of the native conformation or inhibition of folding. To overcome these problems, the folding of many proteins is facilitated in vivo by the activity of molecular chaperons. The term 'molecular chaperones' covers a broad range of proteins families whose main role is to facilitate protein folding. This common attribute is accomplished by the ability of chaperones to recognize non-native proteins and proceed with actively folding or unfolding (using ATPase) non-native proteins by a range of mechanisms (Bukau et al., 1998). Chaperones are present in nearly all organisms and in all cases explored, they are essential for viability.

The best understood chaperones are the prokaryotic GroEL/S chaperonins, whose structure can be described as a double-doughnut shape, with enough space at the center to hold a compact collapsed protein as illustrated in figure 6.

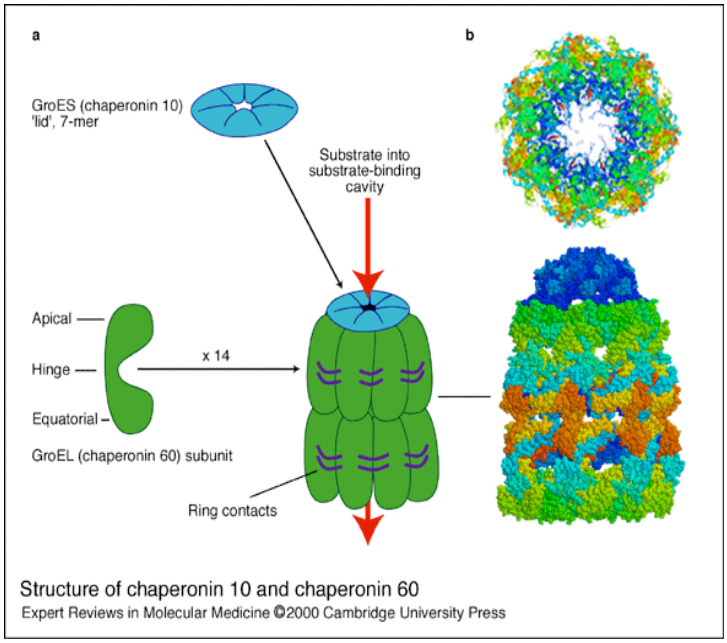

Fig. 6. Structure of chaperonin 10 and chaperonin 60. (a) The complex that is formed between GroEL (chaperonin 60, in green) and GroES (chaperonin 10, in blue). It comprises the two heptameric rings of GroEL, which have a characteristic 'double doughnut' structure, and the attached GroES heptameric 'lid'. (b) The central substrate-binding cavity can be seen on this diagram.

Chaperonin complexes provide folding chambers formed of flexible subunits with ability to alternate surface behavior following a protein binding which can be of different sizes. At first, chaperones were considered to be a passive isolating cage which holds the substrate protein confined from its crowded cell solution. In the passing years, when a more complex behavior of chaperones was revealed, a subject of heated debate was the question whether an active mechanism of cycles of protein binding and release from the chaperone ('iterative annealing'), or the former mentioned, isolating passive cage mechanism ('Anfinsen cage') assists folding. The chaperonin activity comprises several functional subsequent phases which will be defined as the chaperonin functional cycle. Figure 7 illustrates the functional phases which form the chaperonin cycle as follow: (1) High affinity of chaperonin apical domain (hydrophobic surface of the chaperonin chamber) for non native polypeptide substrate attracts a non-native protein substrate. (2) Encapsulating the substrate with the GroES bound ring while sequestering the hydrophobic binding sites. (3) The substrate folds inside the chamber and ATP is hydrolyzed. (4) ATP binding to the opposite ring primes the release of GroES and the trapped substrate protein. A new non-native substrate protein may bind to the other GroEL binding site and the process repeats (Saibil et al., 2002; Betancourt, 1999).

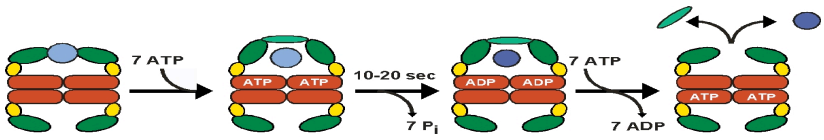

Fig. 7. The chaperonin cycle.

The Eukaryotic Chaperonin containing TCP-1 (CCT)

The chaperonin containing TCP-1 (CCT) is a molecular chaperone consisting of eight different subunit species and assists uniquely (cannot be replaced) in the folding of actin, tubulin and some other cytosolic and non cytosleletal proteins (Yokota et al., 2000) which are essential to the cell (Spiess et al., 2004). CCT and other heat-inducible chaperonins (GroEL/S, HSP60 of mitochondria) are thought to have evolved from a common ancestor as judged by amino acid sequence homology, oligomeric structure and chaperone activity. The CCT, like the GroEL/S, consists of two oligomeric rings, stacked back-to-back, with a cavity at each end that provides a protective environment for protein folding. Even though both have approximately similar structures, GroEL/GroES dynamic features are characterized by a concerted allosteric switch of GroEL, where the ATP-induced conformational changes in CCT are found to spread around the ring-like structure in a sequential fashion (Rivenzon-Segal et al., 2005).

Stress response

All cells, both prokaryotic and eukaryotic, immediately respond to environmental stress, such as heat or increase in acidity, by enhanced synthesis of few proteins termed heat shock proteins (Hsp) (Milton et al., 1990; Munchbach et al., 1999) which are to be found in the cell in different sizes and functioning. Hsps with chaperone activity are classified into six conserved families: Hsp100, Hsp90, Hsp70, Hsp60, Hsp40 and the small Hsp – sHsp. The Hsp60 is considered as the bacterial GroEL/S (Bukau et al., 1998) which is described above. Figure 8 summarizes the topology of binding and action of different chaperones (Hsps) in prokaryotic and eukaryotic cells.

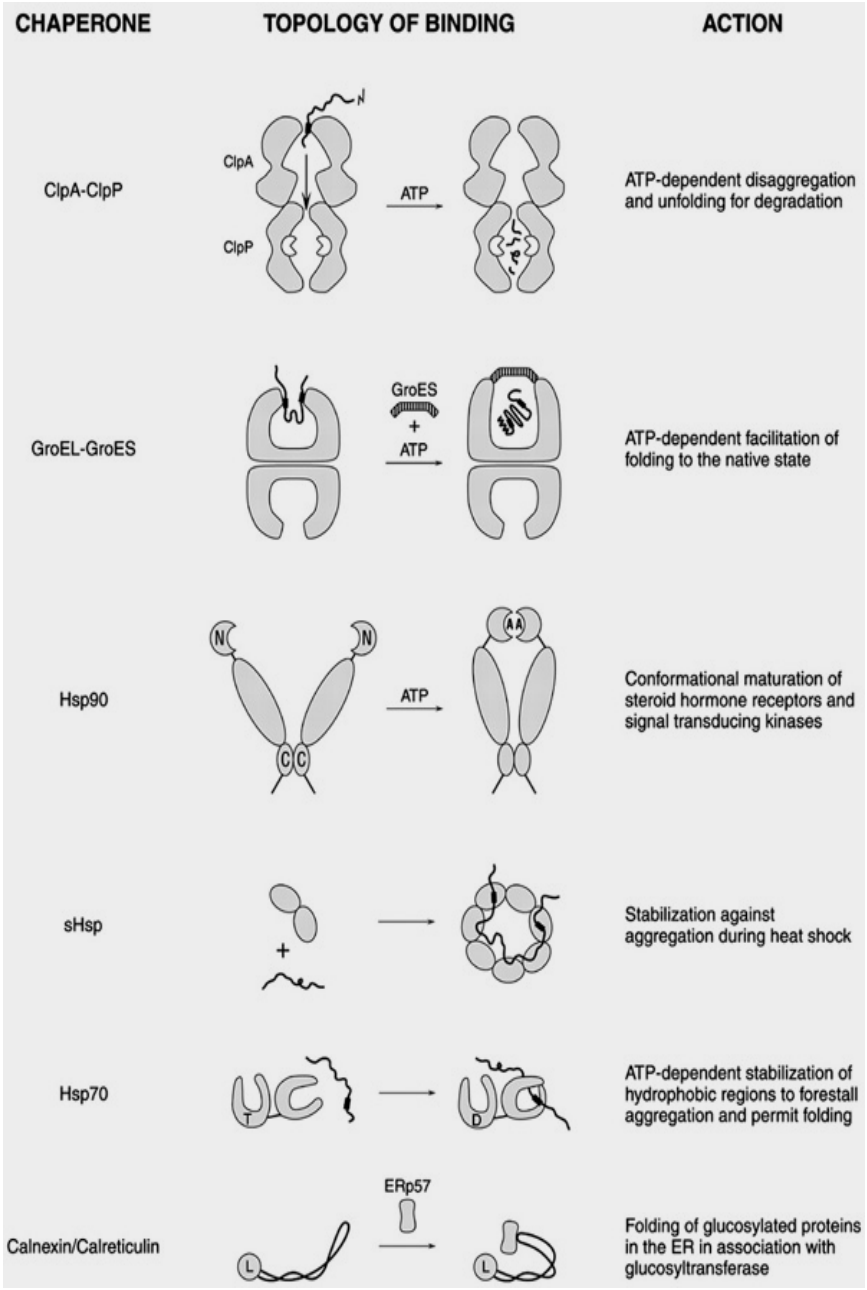

Fig. 8. Variety of different known chaperons. The drawings are not to scale. The solid, wavy lines indicate substrate polypeptides, with typically hydrophobic patches involved in binding to chaperones denoted by the thickened segments. N and C refer to the N- and C-terminal domains. In the Hsp90 drawing, A indicates adenine nucleotide; in the Hsp70 drawing, T and D indicate ATP and ADP, respectively. In the calnexin/calreticulin drawing, L indicates the lectin domain and ERp57 the associated PDI. (Fig. and text fig. are taken from Horwich, 2001).

1.3. THEORETICAL MODELS

Several types of theoretical models are commonly used:

1) *Molecular dynamics (MD)* - Simulation of motion of all atoms in a molecular system (in this case a protein). Current MD simulations can calculate properties of the folding molecules during a time period which is on the order of nanoseconds. Because protein folding takes place in a time scale of milliseconds to seconds, the simulation of protein folding with molecular dynamics is not within reach of present days computers.
2) *Simplified protein models* (Dinner et al., 2000) - simplified protein models (lattice or off-lattice) are able to reproduce many features of real protein structures while keeping the number of degrees of freedom within a tractable range. The main drawback of these models is what makes them computationally feasible - the reduction in the details of the systems. Thus, any result illustrated by one of these simplified models is only an estimation of real nature physics.

   - *Go-like models* - The interaction potential is constructed so that the native structure minimizes the potential energy. This preference of native interactions constructs a folding pathway of the linear sequence towards its native conformation, which makes the process of folding easier for the protein. In these models, the folding process of a protein is biased by a built-in potential energy, which is minimized specifically to a special native structure, which means that each protein is folded using a unique potential energy. Because the process of folding itself is what concerns us when exploring the protein-chaperone interactions, we chose to use a different method which is characterized with more degrees of freedom in the process of folding.
   - *Empirical potential models* - In contrast to Go-like models, in these models a similar pre-defined potential energy is used for all different sequences, in purpose to mimic the actual folding process of a protein in nature.

   <u>Sampling techniques</u>

   As mentioned above, it is not possible to calculate by natural resolution the movements (folding) of proteins. Therefore a much lower resolution is accomplished by sampling the energy surface of the protein conformation. Because of the rugged

nature of the energy surface of proteins, it is unavoidable to use one of the following techniques:

- *Monte Carlo by Metropolis* based protein folding simulation model (MC). This method is simple enough to be a feasible computational tool in our current computers. It produces a Markov chain of conformations which, for a sufficiently large number of iterations, approximates the canonical distribution of conformations a protein can adopt according to its energy (Boltzmann factor).
- *Genetic Algorithms (GA)* - This technique utilizes the same optimization procedures as natural genetic evolution, where a population is gradually improved by selection. GA can be a fast optimization tool but it uses unnatural operations (like crossover) which make it less suitable for studies which concentrate on the folding process itself. Thus, MC method will be used in our study.

1.4. THE CHOSEN THEORETICAL MODEL

One commonly used class of *Monte Carlo by Metropolis* based models is the HP 2 dimensions lattice model (an extended model of that kind is used in this research, see *methods* section) (Chan et al., 1996; Dill et al., 1995). In this model, a protein is represented as a chain of beads which forms a specific sequence of two types of monomers, hydrophobic and polar and contact interaction between non polar (HH) units is considered as favorite by energy smaller than zero. The motions of the model protein are simulated with a dynamic Monte Carlo algorithm. Small random 2 dimensions conformation changes of one, two, three or more beads of the chain are performed repeatedly and accepted or rejected according to a rule which is based on the free energy change (see *methods).* Figure 9 describes an example of a protein as a chain of beads on 2 dimensional lattice space.

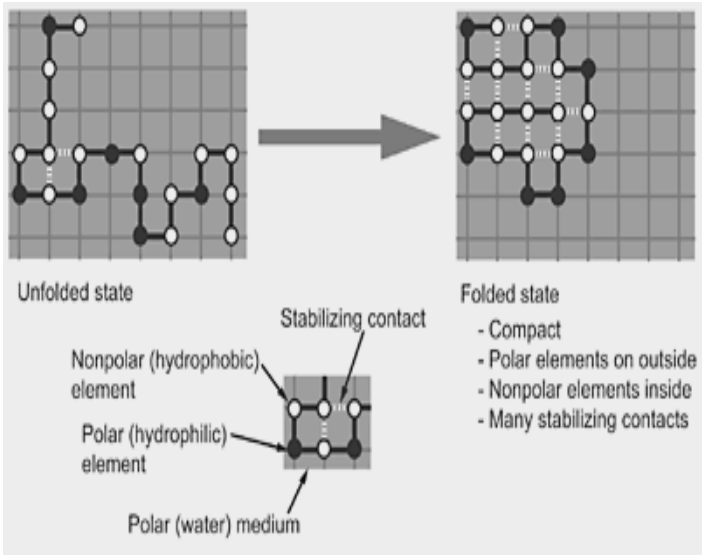

Fig. 9. The 2 dimensional configuration of the unfolded and folded states of a protein chain comprises of H and P beads.

Despite the simplification of the model, it has been shown to capture generic properties of real proteins such as: collapse transitions, mutational properties, development of secondary and tertiary structure and folding kinetics (Dill et al., 1995). For instance, by the definition of the accept/reject rule, a model protein is more likely to move to conformations of lower rather than higher free energy (i.e. The Boltzman probability), which brings the chain of beads characteristics of real proteins: native-like interactions are present, on the average, more than non native ones. Referring back to the 'Levinthal Paradox', lattice simulations (of 3 dimensions and 2 dimensions) have been one of the tools to reveal that because an individual molecule needs to sample only a very small fraction of total conformations during its folding (due to the nature of the guiding restrictions of the effecting energy surface) and because there are many trajectories a protein can pass through on its folding to the native state, the folding time is much shorter than the time of systematical search of all possible conformations. Figure 10 illustrates the folding path with a 3 dimensions lattice model. Lattice models of chaperone and protein interactions have been used for many years in exploring the GroEL-GroES system, specially concerning the encapsulation of the protein, dynamic changes in the chaperone surface and constructing a conceptual framework for understanding how it assists folding (Betancourt et al., 1999; Thirumalai et al., 2001; Brinker et al., 2001; Chan et al., 1996).

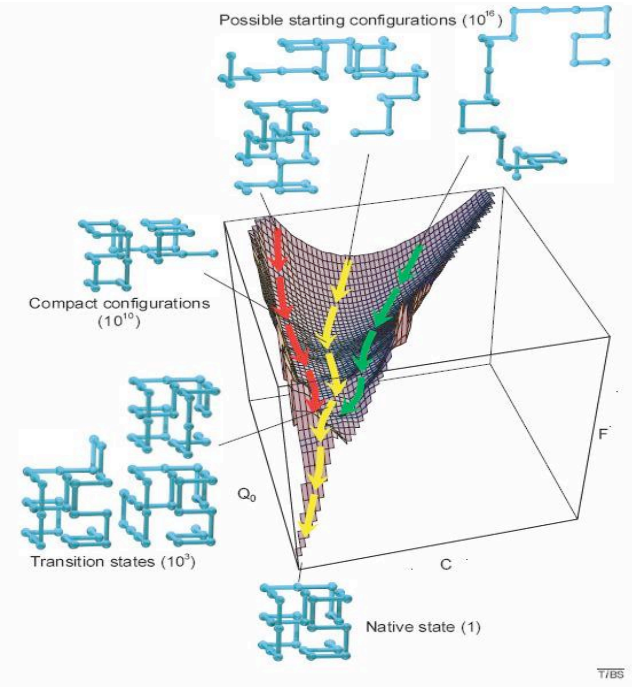

Fig. 10. Free-energy (F) surface of a 27-mer as a function of the number of native contacts (Q_0) and the total number of (native and non-native) contacts (C) obtained by sampling the accessible configuration space with Monte Carlo simulations. A fully extended chain has C = 0 (right-hand edge of surface), and a maximally compact 3x3x3 cube has C = 28 (left-hand edge of surface). The native state is a 3x3x3 cube (front left) with Q0 = 28 (100%). The yellow trajectory shows the average path traced by the last structure sampled at each value of Q0 [<C(Q0)>] for 1000 independent trials that each began in a different random conformation. The other two trajectories (green and red) show a range of two standard deviations around the average and are thus expected to include ~95% of the trajectories. The structures illustrate the various stages of the reaction. From one of the 10^{16} possible random starting conformations, a folding chain collapses rapidly to a disordered globule. It then makes a slow, non-directed search among the 10^{10} semi- compact conformations for one of approximately 10^{3} transition states that lead rapidly to the unique native state (fig and text are taken from [3]).

1.4.1. LATTICE MODEL OF PROTEINS

The polypeptide chains in the simulations are modeled as a linear sequence of residues on a 2D lattice. In order to increase the spectrum of interactions relevant to our study, four different types of residues are used, instead of the common HP model with only two types of interactions (Dill et al 1995; Chan and Dill, 1996). These are: Hydrophobic (H), Neutral Polar (P), positively charged (+) and negatively charged (-). Interactions are considered only between residues in neighboring lattice points (diagonal points are not considered neighboring). Interactions between consecutive residues in the sequence are not considered since they are always present and are independent of the conformation. The energy of sequence S of length N in conformation C is given by:

$$E(S) = \sum_{j>i+2}^{N} P_{ij}\delta(|r_i - r_j|)$$

Where

$$\delta(x) = \begin{cases} 1 & x = 1 \\ 0 & o.w. \end{cases}$$

and P_{ij} represents the energy of the contact interactions which are given in the following table:

	H	P	+	-
H	-1	0	0	0
P	0	-0.75	-0.25	-0.25
+	0	-0.25	+1	-1.25
-	0	-0.25	-1.25	+1

Those values were chosen to reflect the average strength of interactions in empirical mean force potentials (Miyazawa and Jernigan, 1993), where HH interactions are stronger than PP interactions, HP and H(+)/(-) interactions are neutral, P(+)/(-) interactions are weakly favorable, (+)(+) or (-)(-) are repulsive and (+)(-) interactions are the strongest attractor. Repeating the experiments described here with variations of this potential yielded similar results.

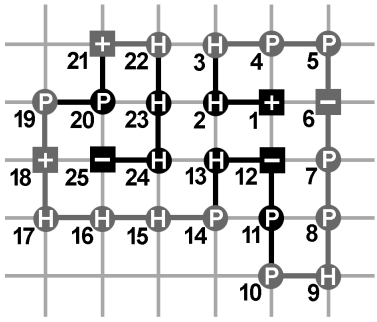

Fig. 11. An example of a model sequence structure. The structure is the native conformation of the 25 residues sequence +HHPP-PPHPP-HPHHH+PP+HHH- with minimum energy of -11. Buried residues appear in Black and exposed residues in Gray.

1.4.2. SIMULATION THECHNIQUE

Folding dynamics is simulated using the Monte Carlo (MC) method with the Metropolis criterion (Metropolis et al, 1953). A chain starts as a random conformation and folds by the following algorithm: From a conformation S_1 with energy E_1, a random change (a move) of conformation to S_2 is performed and the energy E_2 is evaluated. If $E_1 \geq E_2$, then the move to conformation S_2 is accepted, otherwise acceptance of the move depends on the following non-deterministic criterion:

$$Rnd < Exp\left[\frac{E_1 - E_2}{C_k T_f}\right]$$

Where *Rnd* is a random number between 0 and 1. C_k of 1 and T_f=0.5 were used for all sequences and simulations. If the move was not accepted, the former conformation S_1 is retained. Two types of moves are considered: a tail move, which is a random left or right turn of the first or last residue of the chain, and an internal move which is performed as follows: (a) Two residues are randomly selected in the conformation, with a sequences separation up to L residues. Then, the trajectory between the two residues is replaced by another trajectory, taken from a pre defined library of trajectories of the same length and the same relative translocation between their end points (see Figure 12). Only trajectories that do not collide with other part of the chain are considered. L is a parameter that in our simulations was varied between 3 and 11, for size L, trajectories of length <= L are considered. This notion of *local moves* is a generalization of the standard local moves, for example corner flip and crankshafts for L=3 (see review in Skolnick and Kolinski, 1991). The ratio between tail moves and internal moves was varied in the simulation, with no significant change in the results.

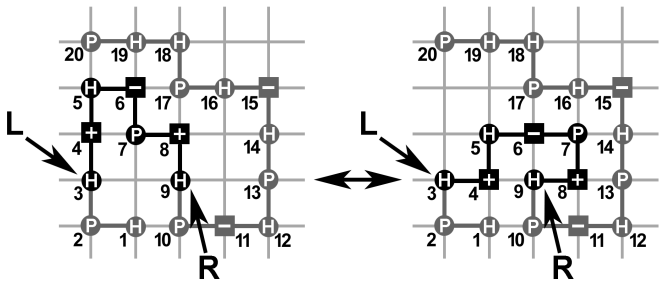

Fig. 12. An example of a local move. The trajectory of length 6 between two residues (3 and 9) is replaced by another valid trajectory of the same length between these points. The rest of the structure is unchanged.

CHAPTER II

Advanced Computational Technology

2.1. COMPUTING REQUIRMENTS OF THE PROJECT

The thesis project requires many repeated simulations of protein folding and chaperone protein interactions. For instance, one module contains population of 1000 different model proteins which interact with 5 different chaperones, where each such simulated interaction takes approximately 3 minutes on a regular modern PC (i.e. Intel Pentium 4 CPU). To perform a reliable statistical analysis, one needs to execute at least 50 repeated simulations for each interaction which will take ([3min] x [50 simulations] x [1000 proteins] x [5 chaperones]) 750,000 minutes which are 520 days of one CPU computing time. Such long duration is not feasible even when one has few more CPUs dedicated to these simulations and considering the fact that many more modules are required, this projects exceeds the time limits of a reasonable research. In this spirit, all project software components were integrated into a powerful computing platform which can perform these tasks in a reasonable time - the GRID (The mentioned module execution time on the GRID is approximately 2.5 days).

2.2. THE GRID PLATFORM

Grid computing is a computing model which provides the ability to perform higher throughput computing by taking advantage of many networked computers. Grids use the resources of many separate computers connected by a network as the internet, to solve large-scale computation problems. It provide the ability to perform computations on large data sets, by breaking them down into many smaller ones, or provide the ability to perform many more computations than a single computer of a cluster. It must be emphasized that all tasks which are relevant to be executed on the GRID in this thesis project concern with how many computing operations per month (or general one can say per year) can be extracted from the computing environment rather than the number of such operations the environment can provide per second or minute (High Performance Computing).

2.3. HIGH PERFORMANCE COMPUTING OF THE EGEE PROJECT GRID PLATFORM

EGEE

The EGEE (Enabling Grid for E-science) project brings together experts from over 27 countries with the common aim of building on recent advances in Grid technology and developing a service Grid infrastructure in Europe which is available to scientists 24 hours-a-day.

The project aims to provide researchers in academia and industry with access to major computing resources, independent of their geographic location. The EGEE project will also focus on attracting a wide range of new users to the Grid. With funding of over 30 million Euro from the European Commission, the project is one of the largest of its kind. EGEE is a two-year project conceived as part of a four-year programme, where the results of the first two years will provide the basis for assessing subsequent objectives and funding needs.

EGEE project aims to make Grid technology available on a regular and reliable basis to all European science, as well as Research and Development. Like the World Wide Web, which was initially developed for specialized scientific purposes, the impact of the emerging Grid technology on European society is difficult to predict at this stage but is likely to be huge.

Performance

The following graphs describe the computation time gained by using the GRID platform on an organization (called SEE) with moderate resources of about few hundred computers organized in clusters on different sites.

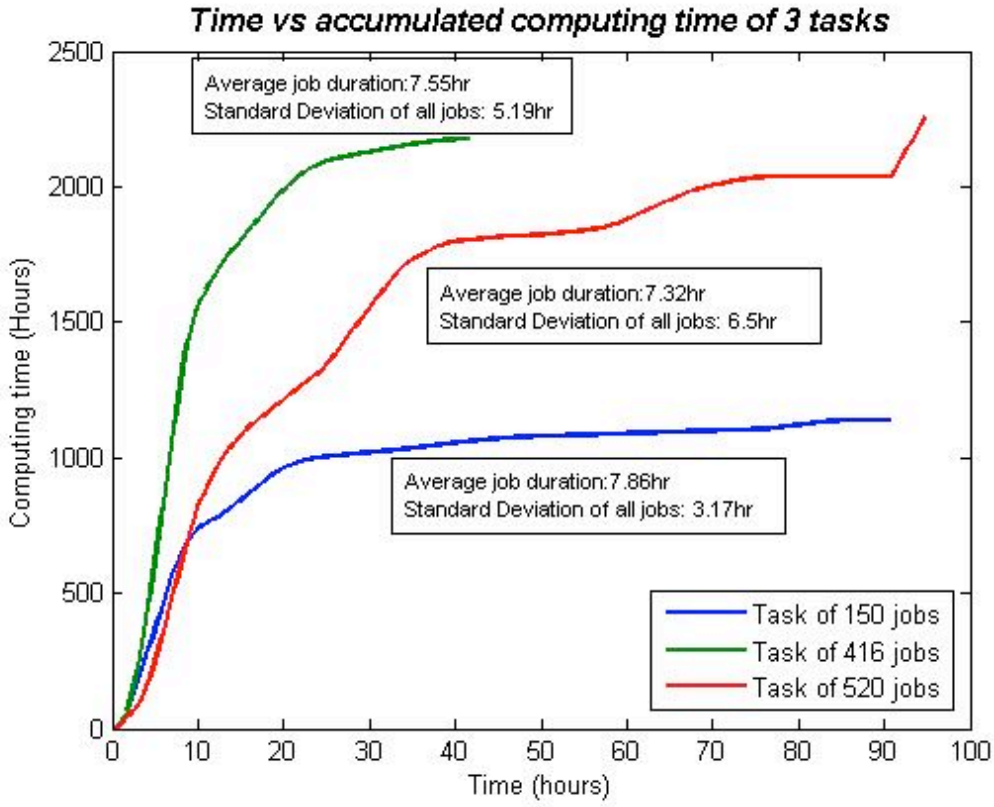

Fig. 13. A description of three different tasks each consist different number of jobs but the average duration of a job in different tasks is about the same. The maximum computing time gained, more than 2000 hours (83 days) was by the 416 jobs task (green). It should be mentioned that the tasks were submitted on different time so the GRID might not have been with the same load.

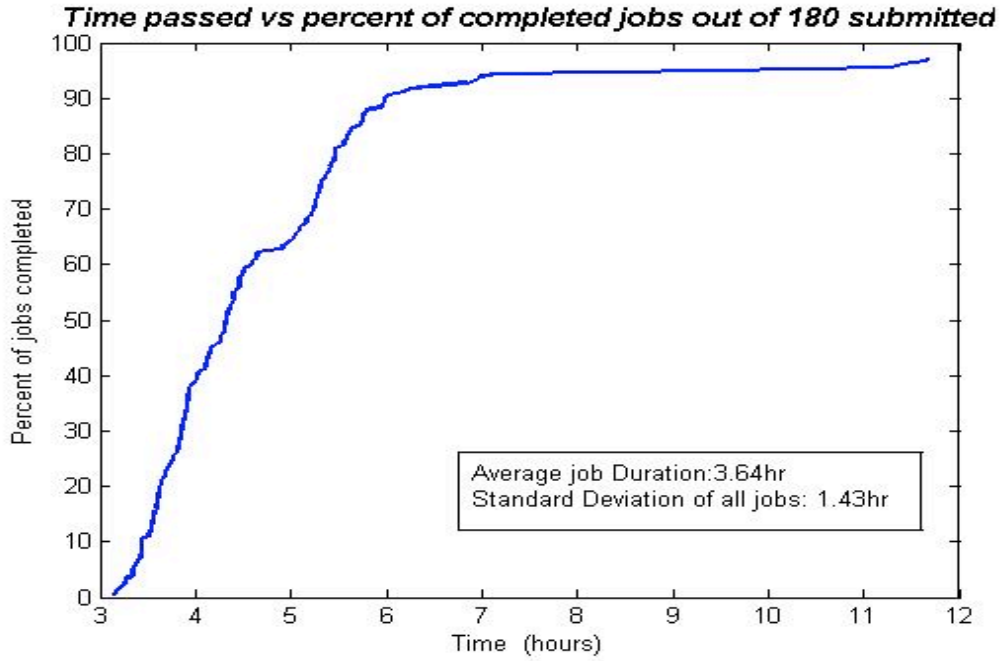

Fig. 14. A description of the percent of completed jobs on the same 180 jobs task as in figure 1. It can be seen that after 6 hours most of the jobs are successfully completed, where each job takes on the average 3.64 hours.

See appendix A for the GRID report.

CHAPTER III

Protein-Chaperone interactions

3.1. INTRODUCTION

Polypeptide chains carry all the information required to fold to their native functional three dimensional structure, and do not require any additional molecules to direct the folding process (Anfinsen 1973). However, in vivo, inside the complex milieu of a living cell, protein folding is facilitated by proteins called chaperones, whose role is especially essential under stress conditions such as heat shock. One major class of chaperones, called chaperonins, comprises ATP-dependent proteins that facilitate folding by binding the assisted protein (i.e. the substrate) in a cavity formed at each end of their double ring structure. A recent study in E. coli, suggested (Kerner et al., 2005) that close to a hundred of proteins in this bacterium require the chaparonin system (called GroEL/GroES, see below) in order to fold properly.

The prokaryotic GroEL/GroES complex in E. coli is the best characterized chaperonin complex. GroEL consists of two rings each formed by seven identical protein subunits. GroES is a single-ring heptamer that binds to GroEL in the presence of ATP and serves as the cap of the cavity formed by each ring structure. Each GroEL subunit can rotate and thus turn a different surface towards the inner cavity.

CCT (also called TCP-1 ring complex) is an eukaryotic chaperonin that is composed of eight similar (but not identical) subunits. CCT does not have a GroES-like cap; instead it contains a "built-in" lid that closes in an ATP-dependent manner to encapsulate its substrates, a process that is required for the folding process. How exactly chaperonins facilitate folding is still unclear. The “passive Anfinsen cage” model suggests that the main effect of chaperonins is to supply each folding molecule a safe environment that protects it from aggregation with other folding molecules or protease digestion. However, the fact that chaperonins undergo coordinated ATP-dependent allosteric transitions during the process, suggests that they play a more active role. Thus, they may provide an environment that is able to guide the substrate towards structures with the required characteristics, for example, towards structures that have their polar residues on the surface.

Active involvement of chaperonins in the folding of the substrate proteins may involve two alternative mechanisms: (i) iterative annealing (see for example, Todd et.al. 1996).where the protein binds several times to the chaperonin during its folding process and thus is offered multiple chances to reach the native state and (ii) confinement (or caging)..

The currently accepted model (Horovitz and Willison, 2005) suggests that chaperonin rings can be in either a T (tense) or R (relaxed) state. In the T state of GroEL, the subunits exhibit a hydrophobic surface towards the cavity; this is an acceptor state for non folded proteins which have many exposed hydrophobic residues. In the R state, chaperonins display polar residues towards the cavity, thereby enabling proteins to be released from the cavity surface and to continue folding either within the cavity volume or in bulk solution. This switch is mediated by ATP binding, since in the T state, the subunits have a low affinity for ATP and in the R state the subunits have high affinity for ATP. As the ATP concentration increases there is a cooperative change of all subunits from the T to the R state.

One major difference between the prokaryotic and eukaryotic chaperonins is the coordination between the surface change of the subunits. While in the GroEL/GroES system the change is concerted, i.e. all subunits switch simultaneously, it was recently shown (Rivenzon-Segal et. al., 2005) that in CCT the change is sequential, i.e. the subunits switch conformation one after the other.

For relatively short, single domain proteins, a concerted switch of the entire system is necessary since switching one subunit (i.e. one surface) to the release state is not effective if other surfaces are still in the T state and remain attached to other parts of the protein. However, it was suggested that a sequential change might be beneficial to eukaryotic proteins that tend to be larger and multi domain as it may enable one domain of these larger proteins to detach from the cavity surface and fold while the other domain(s) is still attached to the surface. In a recent study (Kipnis et. al. PNAS 2007, in Press), it was shown that a GroEL mutant that is defective in its ability to perform the concerted switch (Danziger, et al. 2003) and thus behaves more like the CCT sequential chaperonin, can release a protein in a domain-by-domain manner.

In this study, we used a simple lattice model of the chapaeonin-protein system to explore the implications of the concerted versus sequential conformational switching. Are longer, multidomain proteins more likely to benefit from a sequential mechanism of chaperonin transitions? We show here that our simulations are compatible with this hypothesis, and thus support the idea that the different switching mechanism of prokaryotic versus eukaryotic chaperonins is related to the requirement of eukaryotic cells to fold multi-domain proteins.

3.2. LATTICE MODEL OF PROTEIN CHAPERONIN INTERACTIONS

In our model, chaperonins are modeled as proteins with static conformations (octagonal or square) whose sequence is composed of the same set of four amino acid types as our model proteins (although in the current study, only H and P residues were used for the chaperonins). A variant of the table of interactions (Table 1) was used to evaluate the interactions between protein (substrate) residues and chaperonin residues (See appendix B for more details). Each chaperonin has a cavity that can contain a semi or fully compact collapsed protein. In accordance with current thinking on the role of allosteric switching in chaperonin function (Horovitz & Willison, 2005), our chaperonins have the ability to dynamically alter their cavity surface residues (e.g..from a sequence of successive Ps to one of successive Hs) during the course of the simulation. We consider two fundamentally different classes of chaperonin surface behavior: a concerted surface change (figure 15) and a sequential surface change (figure 16). In the former, all the surfaces that form the cavity of the chaperonin are changed simultaneously from hydrophobic to polar. In the latter, the surfaces that form the cavity are changed sequentially, one after the other, from hydrophobic to polar.

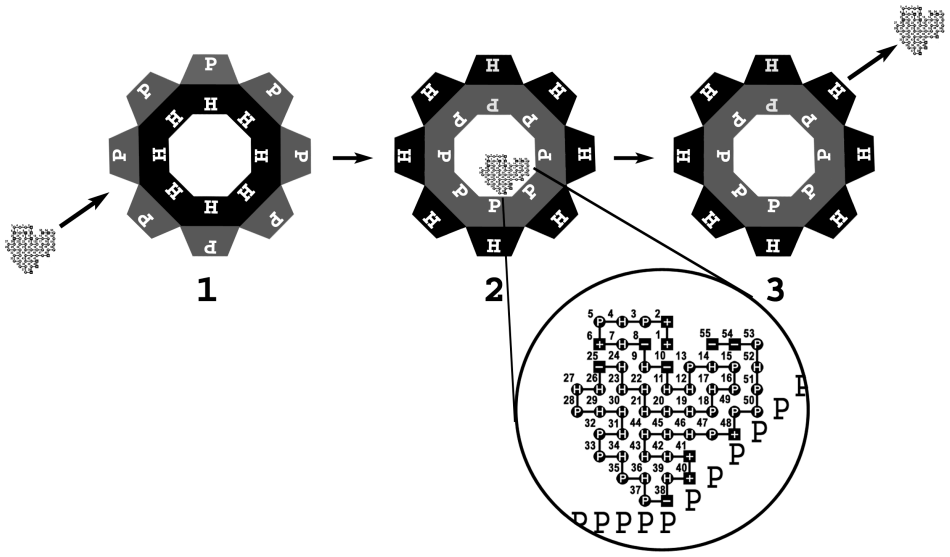

Fig. 15. Schematic view of a chaperonin that undergoes a concerted surface change. In this figure, an octameric chaperonin is depicted. (1) The chaperonin binds a protein substrate in its hydrophobic cavity. (2) The cavity surface switches from fully hydrophobic to fully polar. (3) After a predefined period that a protein substrate spends inside the chaperonin cavity, it is ejected outside. The magnified region shows a 55 residuesingle-domain protein interacting with a polar surface of a chaperonin cavity. Three charged-polar interactions are present between the surfaces of the chaperonin and the protein substrate. This concerted mode of surface change characterizes the GroEL/S prokaryotic chaperonin.

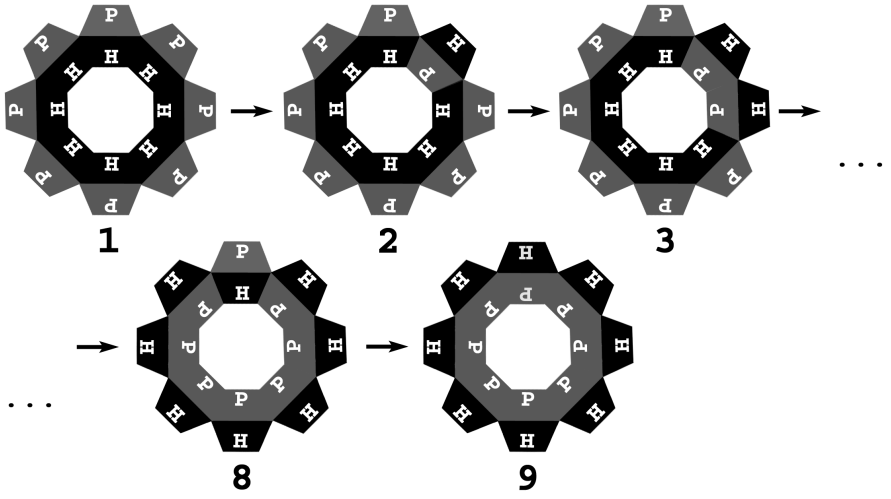

Fig. 16. A schematic view of a chaperonin that undergoes a sequential surface change. In this figure, an octameric chaperonin is depicted. (1) The chaperonin cavity in the substrate protein-bound state is initially fully hydrophobic. (2) The first change occurs when one side of the octagonal structure "flips" to a polar conformation. (3) In the subsequent stages, the adjacent subunits also switch to a polar conformation. (8,9) The process of sequential alteration of the cavity sides from hydrophobic to polar ends when all eight of them are polar. The time between each additional change is predefined for each simulation and is constant during the whole simulation. This model represents the eukaryotic chaperonin (CCT).

Two different mechanisms of the way substrate proteins interact with chaperonins were simulated:

(a) Binding and release - substrate protein commences its folding process in an open environment, where it can make any movement without colliding with a lattice boundary. After a predefined duration, the protein binds to a chaperonin which confines it in its cavity for an additional predefined duration (e.g 100,000 Monte Carlo Steps (MCS)). After this time, the protein is released from the chaperonin cavity back to the open environment. This binding-release cycle may be repeated several times during the folding process.

(b) Caging – The protein spends the entire simulation inside the cavity of the chaperonin.

3.3. CHAPERONIN SUBSTRATES

Approximately 70% of proteins in eukaryotic cells are multi-domain whereas in prokaryotes single-domain proteins are more common. We wanted to examine whether this fundamental difference between eukaryotic and prokaryotic cells may have had a selective effect on the mechanism of allosteric switching of their respective chaperonins. Hence, three types of substrate proteins interacting with chaperonins were studied: (i) proteins of 25 residues in length. (those are single-domain proteins as proteins of that size cannot form two domains); (ii) single-domains of 55 residues; and (iii) and double-domains of 55 residues (total length). For each type, about 100 different sequences were tested (123 sequences of 25-mer single-domains, 117 sequences of 55-mer single-domains and 104 sequences of 55-mer double-domain proteins). All model sequences were generated by random selection of 25 or 55 residue sequences, drawn from a distribution of 45% H, 30% P, 12.5% (+) and 12.5% (-). This composition is based on the composition of amino acid groups in the PDB (http://us.expasy.org/sprot/relnotes/relstat.html, PFB release 49.1) that is: 44.3% neutral, 30.7% polar, 12% positively charged amino acids and 13% negatively charged amino acids. To reflect the fact that protein termini (amino and carboxyl groups) are charged, oppositely charged amino acids (+ / -) were assigned to both termini.

3.3.1. SEQUENCE OF 25 RESIDUE LONG SINGLE DOMAIN SUBSTRATES

The generation of the 25-mer sequences was based on a thermodynamic selection criterion followed by a kinetic selection.

3.3.1.1. THERMODYNAMIC SELECTION

We created sequences with the amino acid composition mentioned above. For each sequence, we considered the lowest energy conformation amongst all possible 9,646,215 conformations that fit into a compact 6X6 square. If there was more than one conformation with the same minimum, one was arbitrarily chosen as the native conformation. Conformations for which the simulation (to be described below) demonstrated that the minimal energy is not a compact structure (i.e. the simulations found a minimal energy conformation that could not be contained within a 6X6 square) were excluded retroactively from consideration. We encountered only very few (less than 1%) such cases. Using this procedure, 1084 sequences were analyzed.

There is a large variance of the spectrum of energy values of the conformational space of different proteins. As suggested in (Sali et al., 1994), a significant energy gap is important in order to ensure kinetic accessibility of the native structure. Thus, for each sequence, we measured the difference between the minimal energy (i.e. the native conformation) and the average energy of all conformations, and expressed these in units of standard deviations of the average energy. The larger the difference between these two numbers, the more pronounced the energy gap. We selected approximately half of the sequences (out of the 1084), with the largest energy gap for further analysis.

3.3.1.2. KINETIC SELECTION

For each one of the 542 sequences, 100 independent Monte Carlo simulations were run, each comprising 10^6 Monte Carlo Steps (MCS). The simulation process was terminated once the native conformation was found or after 10^6 MCS. We considered a given sequence to be kinetically foldable if the simulation identified the native structure in more than 90% of the runs. Some flexibility was allowed in finding the native conformation. We considered the native conformation as "found" if the simulation reached a conformation within a distance of less than 0.5 Root Mean Square Distance (RMSD) from the native conformation. (This distance is roughly equal to two out of the 25 residues being off by one lattice point from the corresponding position in the native conformation.) This criterion left us with a total of 123 unique sequences. Examples of 25 residue- long structures are shown in figure 6 (A).

3.3.2. 55 RESIDUES LONG SEQUENCES

With current computational resources, it was not possible to computationally enumerate all the compact 2D conformations of the 55 residue-long sequences. Thus, in order to get the sequences we need for this study we had to adopt the following strategies.

3.3.2.1. 55 RESIDUE LONG DOUBLE DOMAIN SEQUENCES

To form homo-double-domains we duplicated each 25 residue-long sequence and added a linker of 5 P residues (see figure 17(A)). The native structure of the longer sequence can be either a double-domain with two cores, each with a structure quite similar to the native structure of the original 25 residue sequence, or a single-domain with one large core. To select for the former type of sequences, we needed to look for structures whose energy would be roughly twice the energy of the native conformation of the 25-mer (in addition to the energy gained by the interface between the two domains), with the structure of each domain similar to the structure of the 25-mer. All 123 25-mer sequences were used in the creation of the homo-double-domain substrates as follows. A 55-residue-long sequence was created by the duplication of a 25-mer sequence connected by a polar linker of 5 P residues.

For each such sequence, 200 independent, long (10^7 MCS) simulations runs were performed. If in any one of these 200 simulations, the simulation found a non double-domain conformation that that had a significantly lower energy than that of the double-domain structure, then the sequence was excluded from further analysis. A total of 104 homo- sequences were selected under these criteria. figure 17(B) illustrates an example of a double-domain structure.

3.3.2.2. 55 RESIDUE LONG SINGLE DOMAIN SEQUENCES

Since we can not enumerate all possible conformations for sequences of length 55 in order to identify a sequence with a native conformation which is kinetically accessible, we selected sequences for which long (10^7 MCS) simulations converged to a similar structure within a distance of 0.9 RMSD (i.e. found the same structure as minimal) in

more than 5% of 200 runs. A total of 117 out of 1000 randomly chosen sequences containing the residues composition described above, passed this criterion and were included in the set of 55-mer single-domains (see figure 18).

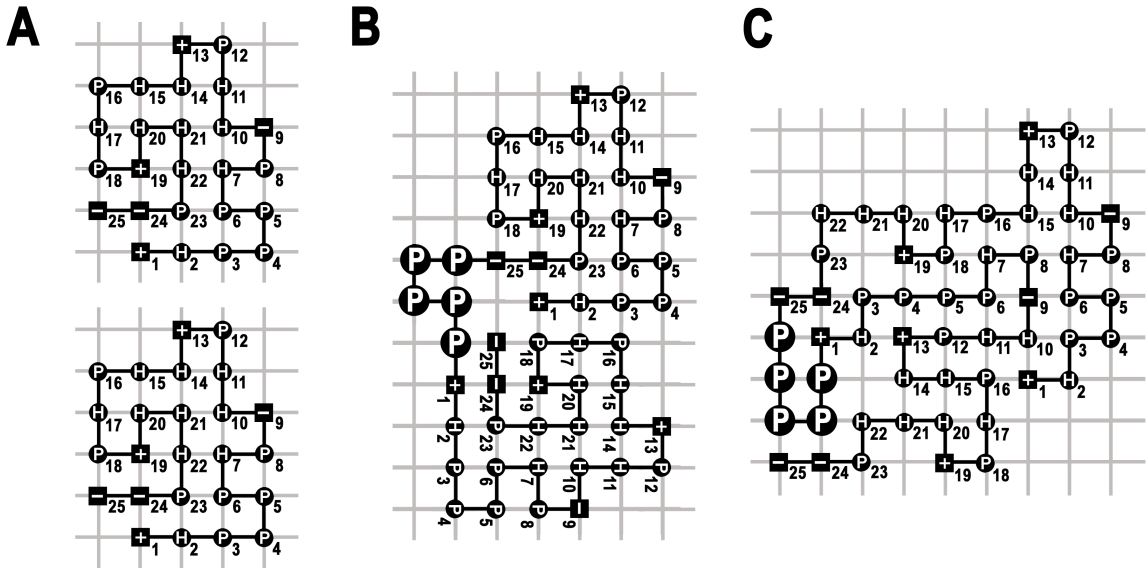

Fig. 17. An example of a dimer protein model. The figure shows the dimer structure of the sequence +HPPPPHP-HHP+HHPHP+HHHP--PPPPP+HPPPPHP-HHP+HHPHP+HHHP--. (A) The two identical 25 residue-long structures that form the 55 residue homo-double-domain. (B) An example of a double-domain native structure. The interface between the two cores and the polar linker position are not considered in the calculation of the root mean square distance (RMSD) of the native two core structures. (C) An example of a monomer structure formed by this sequence. The two cores ??? of this structure in C? are in a totally different conformation than the two cores in A.

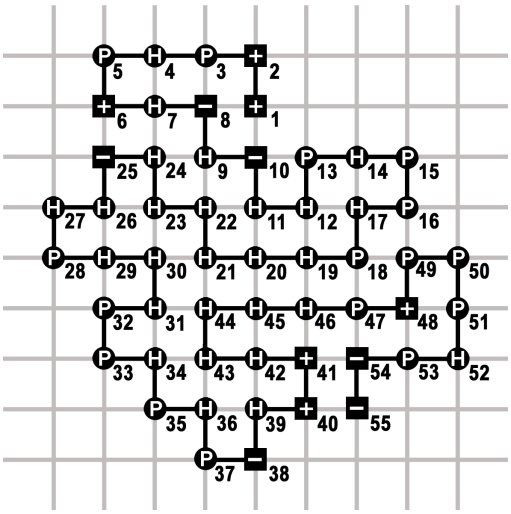

Fig. 18. An example of a model sequence single domain structure. The structure described above is the native minimal conformation (with minimum energy of -28.625) of the 55 residue sequence ++PHP+H-H-HHPHPPHPHHHHHH-HHPHHHPPHPHP-H++HHHHHP+PPPHP--.

3.4. RESULTS

We used a lattice model to investigate the effects of chaperonins on the folding of different substrate proteins. In particular, we wished to determine the effect of the chaperonin cavity surface, which interacts with the substrate, on the yield of successfully folded substrate proteins. The fundamental measure used in this study is the improvement in the percentage of "successful" simulations (simulations that yield the native structure) for a given protein substrate. All protein substrates were first subjected to a few hundred simulations in the absence of chaperonins. We considered a run to be successful if during the folding process of a pre-defined duration (e.g. a simulation of 10^6 MCS) the native minimal structure was found (or a structure within RMSD < 0.5 for a 25 residue-long sequence and RMSD < 0.9 for a 55 residue-long sequence). The percent of successful simulations, out of the total few hundred simulations executed for each protein, was defined as the folding yield of a protein. The same number of simulations under the exact same conditions (e.g. temperature, interaction potential etc) were then executed for each protein in the presence of a chaperonin. The difference between the yield with and without the chaperonin is defined as the improvement in the folding yield of the protein. The ratio between the yields with versus without the chaperonin is defined as the improvement factor. For instance, for a protein with a yield of 20% in the absence of a chaperonin, and yield of 30% with a chaperonin, the improvement factor is 1.5. All analyses presented here were tested by paried t-test and were found statistically significant.

We start by exploring several basic aspects of protein substrate-chaperonin interactions in the case of 25 residue single-domains, and then continue with more sophisticated models of 55 residue-long single- and double-domains.

3.4.1. ANALYSIS OF BASIC ASPECTS OF PROTEIN SUBSTRATE-CHAPERONIN INTERACTIONS IN THE CASE OF 25 RESIDUE LONG SINGLE DOMAIN

The simulations for the 123 25 residue-long sequences were more than 90% successful at an ideal simulation temperature of T=0.5. As was observed before (Betancourt and Thirumalai, 1999) in lattice simulations (and also in several experimental studies), the effect of the chaperonin on the folding yield under ideal folding conditions is minimal, and sometimes even adverse. Thus, the effect of chaperonins must be evaluated under conditions that are less than ideal for folding. Hence, we simulated folding of these

sequences at a higher temperature of 0.75 (Tf=0.75). The increase in temperature yielded, as expected, a lower percent of successful simulations. In the presence of a chaperonin, we observed a significant improvement in the folding yield when folding of the substrate was simulated inside a chaperonin that undergoes concerted surface changes and by invoking a binding-release mechanism (figure 19). In contrast, no significant improvement was obtained when substrate sequences were folded inside a chaperonin that undergoes sequential surface changes,.

Another fundamental aspect we investigated was the effect of chaperonin cavity size on the folding yield of substrate proteins. We expected that the action of caging a protein would reduce its entropy and thereby improve the process of finding the structure with the minimal energy which is usually also compact. Figure 20 illustrates the effect of the size of the chaperonin cavity on the folding yield of substrate proteins. As expected, smaller cavities of chaperonins result in better folding yield improvements. The cavity structure was a square and only the binding and release mechanism was considered for this analysis.

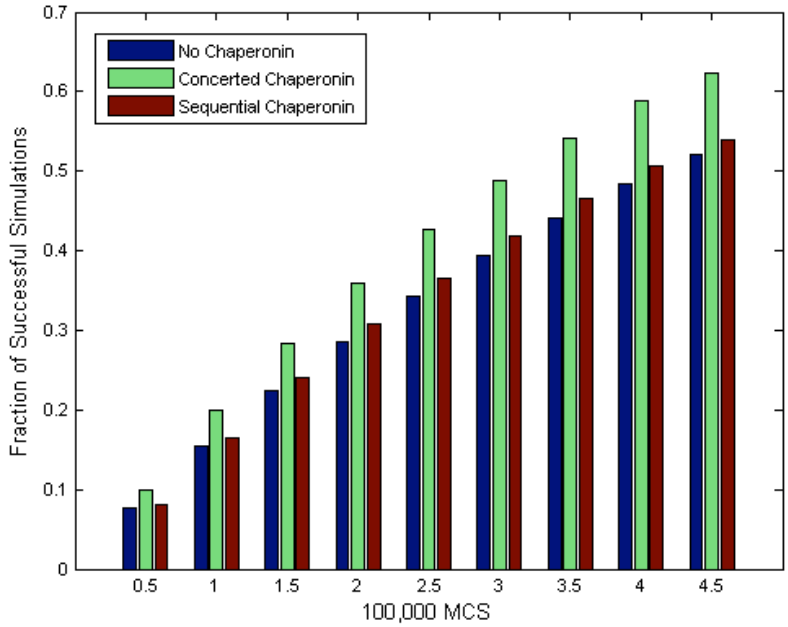

Fig. 19. The effect of a chaperonins on 25 residue-long sequences. Each of the 123 sequences was simulated 250 times at a temperature of 0.75 under three different conditions: in the absence of a chaperone (brown), in the presence of a chaperonin that undergoes sequential surface changes (green), and in the presence of a chaperonin that undergoes concerted surface changes (switching from hydrophobic to polar residues) (dark blue). The binding and release mechanism is used in these simulations and the chaperonin structure was modeled as an octagon with sides of 5 residues. The graph describes the fraction of successful simulations of all 123 sequences as a function of number of Monte Carlo steps. For instance, the three bars representing 200 KMCS show the fraction of successful simulations achieved until 200,000 Monte Carlo steps.

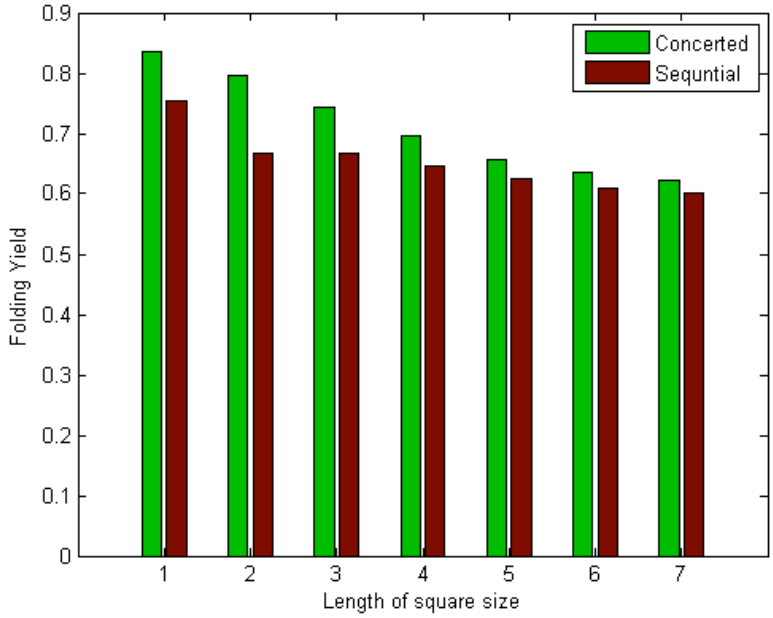

Fig. 20. The effect of chaperonin cavity size on the folding yield of 25 residue-long substrate sequences. In these simulations, the cavity has a square shape with sides that vary between 7 and 13 residues and a binding and release mechanism was invoked . Each of the 123 sequences was simulated 250 times for 106 MCS. The graph describes the fraction of successful simulations for all 123 sequences as a function of cavity size.

3.4.2. ANALYSIS OF PROTEIN-SUBSTRATE CHAPERONIN INTERACTIONS ON 55 RESIDUE LONG SINGLE-DOMAIN SUBSTRATES

In lattice simulations, longer sequences require longer MC runs to fold. Hence, the 117 55 residue-long sequences were simulated using 107 MCS runs.

In order to include in this analysis the standard error of the result of a series of simulations for a given sequence, we performed the following approximation. The standard error for each monomer is estimated from the binomial distribution. Where:

$$s = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

$\hat{p}$ is the estimated probability of a successful simulations (yield) out of the N independent simulations (390 or 360 observations) and σ is the estimated standard error. Figures 21 and 22 present the results of this analysis.

The same effect of a chaperonin on 25 residue-long substrates was observed as on 55 residue long sequences with a single-domain structure. Figure 9 demonstrates the effect

of chaperonins that undergo sequential and concerted surface changes on each of the 55 residue sequences. It can be seen that a concerted change has a major effect on the yield while a sequential change of the surface has only a minor effect.

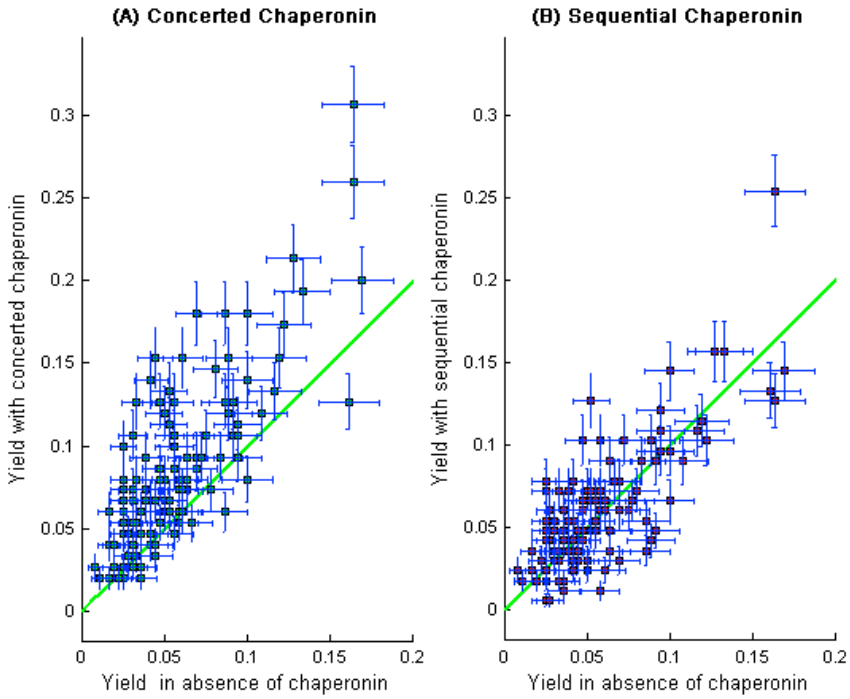

Fig. 21. Folding yield in presence of a chaperonin versus the yield in absence of a chaperonin, for each of the 117 55 residue-long single-domains. The folding yields (fraction of successful simulations) of different single-domain 55-mers in the presence of a chaperonin that undergoes concerted (A) or sequential (B) cavity surface changes are plotted against the respective yields in the absence of a chaperonin. The crosses indicate the estimated binomial distribution standard error bars for the 360 simulations for each sequence It may be seen (A) that the concerted chaperonin significantly improves the yield of these single-domain substrates as most of the data points are above the green line with a slope of one. In contrast, the sequential chaperone (B) does not improve significantly the folding yield of these substrates. The paired t-test value of the distribution for the concerted chaperonin distribution against that in the absence of chaperonin is <1.0E-16, demonstrating that the effect of a concerted chaperonin is statistically significant. The paired t-test value of the distribution for the sequential chaperonin distribution against that in the absence of chaperonin is 0.087651 showing that the effect of a sequential chaperonin is not significant. Progress speed for the sequential chaperonin is 5000 MCS and for the concerted chaperone the switch from hydrophobic to polar is made after 1000 MCS. The binding and release mechanism was invoked in these simulations and the chaperonin structure was modeled as an octagon with sides of size 7. Each of the 117 sequences was simulated 360 times with simulation duration of 107 MCS, at a temperature of 0.75.

3.4.3. ANALYSIS OF PROTEIN-SUBSTRATE CHAPERONIN INTERACTIONS ON 55 RESIDUE LONG DOUBLE-DOMAIN SUBSTRATES

As described earlier, eukaryotic chaperonins must often fold large multi-domain proteins. It was observed that eukaryotic chaparonins undergo a sequential mode of surface changes (Rivenzon-Segal et al., 2005) and it was, therefore, suggested that these two phenomena are linked. Thus, we next investigated whether the effect of a chaperone with sequential surface changes is stronger on dimeric versus monomeric structures. To this end, 104 55 residue long sequences with homodimeric native structure were simulated for 107 MCS. The results in Figure 10 show that, contrary to the results obtained for the 25 and 55 monomer substrate sequences, the effect of chaperone with sequential surface behavior on the homo-dimer substrates was very strong as the concerted chaperone effect.

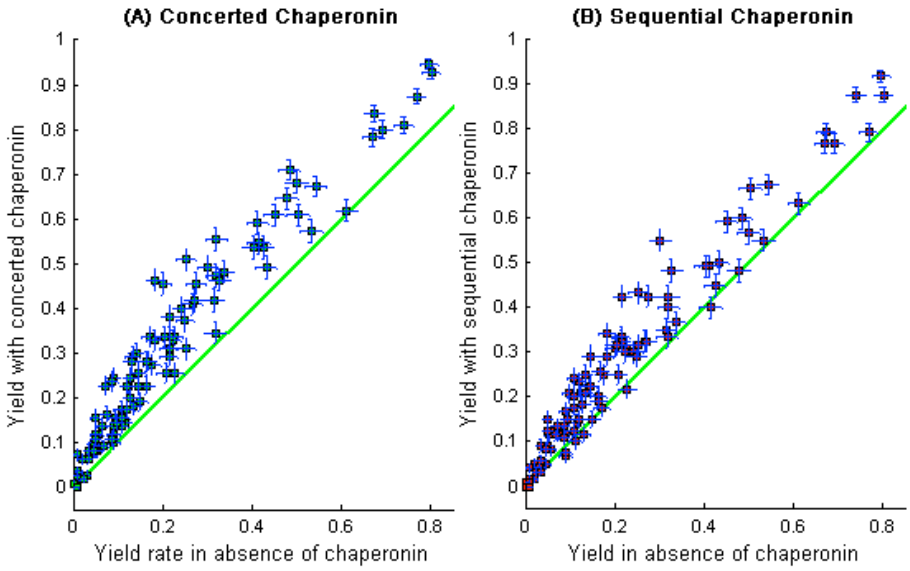

Fig. 22. A representation of the yield in presence of chaperone versus in absence of chaperone, for each of the 104 55 residue long dimers. See figure 21 for explanation on the figure. Each of the 104 sequences was simulated 390 times with simulation duration of 107 MCS, at a temperature of 0.75. The bind and release mechanism was used in these simulations (Similar results were found using the caging mechanism) and the chaperone structure was modeled as an octagon with sides of size 7. Progress speed for the sequential chaperone is 5000 MCS and for the concerted chaperone the switch from hydrophobic to polar is made after 1000 MCS. It is obvious that the effect of chaperone with sequential surface changes is almost as strong as the effect of chaperone with concerted surface changes. Paired t-test value of yield in presence of concerted chaperone versus in absence of chaperone distribution equals 8.38782E-28, and for sequential distribution against no chaperone distribution is < 1.0E-19. In both cases the null hypothesis is rejected; thus, the observed effects of both sequential and concerted chaperons are statistically significant.

3.4.4. COMPARISON BETWEEN THE EFFECT OF CHAPERONIN WITH SEQUENTIAL SURFACE CHANGES ON MONOMER AND DIMER SUBSTRATES

In order to compare the distinctive averaged effects on monomer and dimer substrate sets on a single scale, we needed to normalize the results. We define the factor of change as the ratio of yield (the fraction of successful simulations) with a chaperone versus the yield without a chaperone. Figure 23 shows the normalized results. It is clear that chaperones with sequential surface behaviour yield a significant improvement on dimer substrates and only a slight effect on the monomer substrates. In contrast, the effect of a chaperone with concerted surface behaviour, is significantly better for the monomer substrates than the dimers.

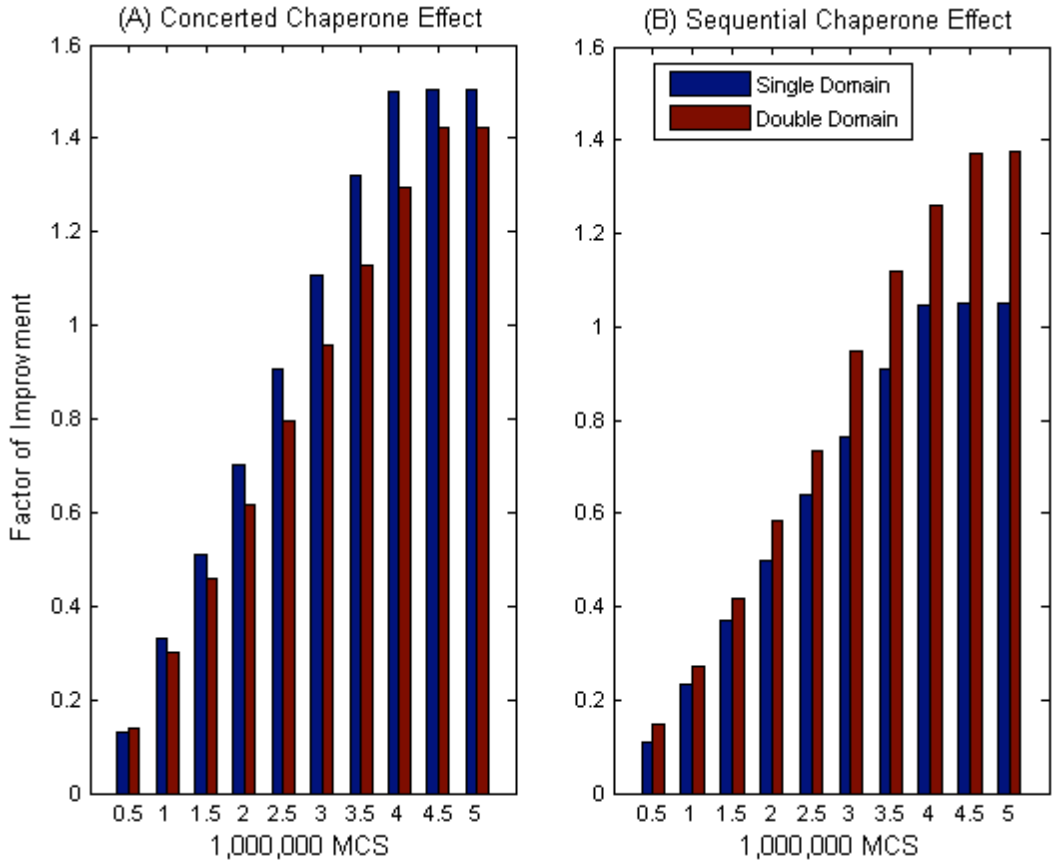

Fig. 23. The effect of chaperone with sequential surface changes on monomer and dimer substrates. The graph plots the improvement in folding yield as a function of MCS. The factor is normalized to the maximal yield of successful simulations of each group since each group has different average folding yield. A structure of RMSD < 0.9 from the native structure was considered as a success for both groups. In addition, these results apply to the same octagonal chaperons with side of size 7 using a bind and release mechanism. These data show that monomer folding is better facilitated by a concerted surface change chaperone, while dimer folding is better facilitated by a sequential surface change chaperone. All differences shown here are statistically significant by paired t-test analysis.

3.5. DISCUSSION

The prokaryotic chaperonin GroEL undergoes ATP-driven concerted conformational switching between a protein acceptor (T) state and a protein-release (R) state (Horovitz & Willison, 2005). A major difference between these states is that the surface of the chaperonin's folding chamber is hydrophobic in the T state (thus favoring nonfolded protein substrate binding) (Braig et al., 1994) and hydrophilic in the R state (thus favoring protein substrate release)(Ranson et al., 2002). In contrast, it has been shown that the eukaryotic chaperonin CCT undergoes ATP-driven conformational changes that are sequential (Rivenzon-Segal et al., 2005). In this work, we tested possible implications of these different allosteric mechanisms for the folding function of these chaperonins. We found that the folding yields of single-domain protein substrates are greater when the chaperonin undergoes concerted and not sequential conformational changes. In contrast, the folding yields of double-domain proteins are greater in the presence of a chaperonin that undergoes sequential conformational changes (Figure 23). These results are consistent with findings that indicate that large multi-domains proteins are more common in eukaryotes compared with prokaryotes. Hence, they support the suggestion (Rivenzon-Segal et al., 2005) that the different allosteric mechanisms of prokaryotic and eukaryotic chaperones can be explained by the need of eukaryotic chaperones to facilitate folding of proteins with a multi-domain structure.

CHAPTER IV

Structural Features of Protein Termini

4.1. INTRODUCTION

Quite a few studies have been devoted to understanding the structural features of the first and last protein residues (i.e. termini). Two lines of investigations were taken; one is the question whether the two termini of proteins tend to be closer to each other than would be expected for random distances distribution. The other question is whether the properties of the N-terminal are different than those of the C- terminal. This is an important question since it has bearing on the controversial issue of sequential folding, i.e. is folding, for example on the ribosome, a sequential process that proceeds from the N-terminal to the C-terminal. In pioneering work, Thornton and Sibanda, (1983) evaluated the distances between termini in 52 proteins and concluded that the distances between termini are smaller than expected for random chains. Christopher and Baldwin (1996) examined a much larger set of proteins and reached a different conclusion, that the distance between termini is not statistically different than the random expectation. A recent study (Krishna and Englander, 2005) has contributed an interesting observation, that proteins which fold in a two-state kinetics have their termini close together, while proteins that fold in a non two state kinetics have their termini separated.

The different environment of the termini was first studied in (Thornton and Chakauya, 1982) where it was observed that for proteins which exist at that time in the PDB, the N-terminal region tend to adopt an extended beta-strand conformation while C-terminal regions are often helical. In (Alexandrov, 1993) it was argued that N-terminal residues tend to have more intra-molecular contacts than the C-terminal, suggesting that the N-terminal folds before the C-terminal. Laio and Micheletti (2006) have re-examined the data, and did not see this tendency. They did find, however, that the C-terminal is significantly more compact and locally organized than the N-terminal, although they argue that the bias is not due to sequential folding.

All these studies are based on the observation that protein termini tend to be on the surface of proteins and not buried in the core. This fact is critical for all these studies since it supplies the background against which calculations are tested. For example, when comparing the expected distance between termini, it is critical to consider the fact that

termini are mostly on the surface, since the average distance of random points on a surface of a sphere is very different from the expected distance between random points found anywhere within its volume.

Surprisingly the tendency of termini to be located on the surface is commonly taken as a postulate without a sufficient explanation. For example, Christopher and Baldwin (1996) paper starts with the following statement: "The terminal regions of proteins differ in several ways from more internal segments. The termini are often surface exposed and flexible ".

We are not aware of studies aiming to explore this issue and explain how are the terminal residues get to be overwhelmingly located on the surface of proteins. At least for some proteins there is a need to bring the terminal residues to the surface to allow them to participate in post translational processes (e.g. in N-terminal acetylation or methylation). However, many proteins do not undergo such modifications, and in any case this functional reason does not supply a mechanism to support the tendency of terminal residues to be located on the surface of folded proteins.

A common explanation often given for this tendency is that terminal residues are charged: The first amino group (which is not bonded to a carboxyl group) is positively charged, and likewise the last carboxyl group which is not paired with an amino group is negatively charged. Charged residues would tend to be on the surface of proteins because of their favorable interactions with water which is a polar solvent. However, this argument is valid also for charged amino acids like Lysine, Arginine, Aspartic acid and Glutamic acid. While these residues tend indeed to be located on the surface of proteins, we show here that terminal residues are much more exposed than charged amino acids.

In our study we first use the large collection of protein structures that currently exist in the PDB to measure, by various methods, the extent to which termini are indeed located on the surface and exposed to the solvent. Next, we want to understand what are the mechanisms leading to this behavior.

Using a lattice, we generate a large population of model proteins and study their properties by selecting proteins on three levels: structural selection of compact structures; thermodynamic selection of conformations with strong energy preferences, and kinetic selection of fast folding proteins using Monte-Carlo simulations. We show how, progressively, each selection raises the proportions of proteins with termini on the surface, resulting in very similar proportions to what is measured for real proteins.

4.2. EXPOSURE ANALYSIS OF RESIDUES OF PROTEINS

PDB entries were taken from the non-redundant PDB set (http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html) using the non redundant threshold of p-value of 10^{-40}. From this list we took only monomeric structures of length between 50-200 amino acids that were solved by X-ray crystallography and for which no missing residues were reported. A total of 425 structures were considered.

Two methods were used to determine the extent to which termini are located on the surface of proteins. The first measure is based on the exposure of termini residues to solvent, and the second on the distance of the termini from the center of mass of each protein.

Exposure calculations

The corresponding DSSP files for the PDB entries were downloaded from ftp://ftp.cmbi.ru.nl/pub/molbio/data/dssp/. We used the solvent accessibility value in the DSSP as the exposure measurement as described in (Kabsch and Sander, 1983). The relative solvent accessibility of each residue was calculated by normalizing its solvent accessibility to the maximum possible value for that amino acid (Shrake and Rupley, 1973).

Distance from center of mass

While solvent exposure is a very common way to measure the extent to which amino acids are on the surface of proteins, there might be a problem in using it for terminal residues. Some of the protection from the solvent is supplied by the main chain and the side chain of the two immediate neighbors of each amino acid. However, terminal residues are truncated and have only one neighboring residue. Thus, to enable independent assessment of the location of terminal residues we suggest measuring the distance of each amino acid to the center of mass of its protein. Residues with the highest distance will be on the surface. Since proteins are of different sizes, and hence expected distances, we normalized this measure for each protein in terms of standard deviation according to:

$$RED = \frac{[D - Avg(Ds)]}{SDV}$$

Where *RED* is the relative distance of a residue (C_α only), *D* is the absolute distance from the center of mass, *Avg(Ds)* is the average distances of all residues from the center of mass, and *SDV* is the standard deviation of this average.

4.3. ANALYSIS OF PDB STRUCTURES

We start by calculating the exposure of the termini in a dataset of 425 non-redundant monomeric proteins from the PDB. The averaged normalized solvent accessibility of termini residues is 87.1% compared with 49.2% of charged residues and 35.9% of all residues. We consider a residue with solvent accessibility of more than 50% of its maximal surface area as exposed. Figure 24 shows the exposure of residues in the N and C terminal region, i.e. the first and last 10 residues of each protein. It is clearly seen that the terminal residues are highly exposed (80.3% and 86.1% for N and C terminal residues respectively), there is a much smaller effect on the residues adjacent to the termini. When the analysis is done based on amino acid type (Figure 25) we see, as expected, that charged residues are more exposed than hydrophobic and polar residues but that terminal residues are much more exposed than charged residues.

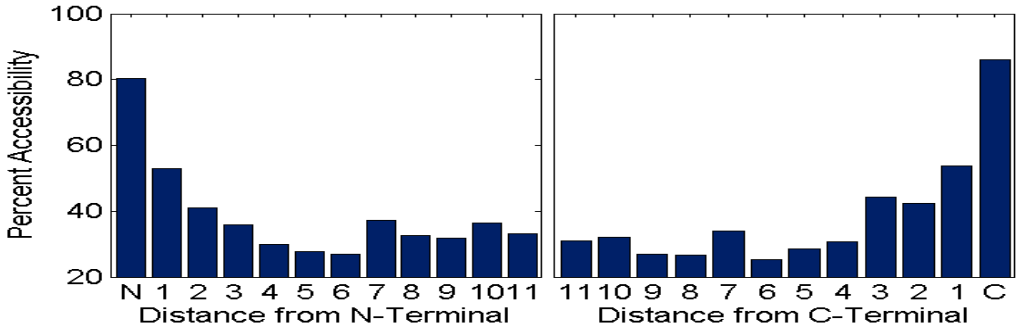

Fig. 24. Exposure of terminal residues in PDB. The percent of residues, averaged over 425 proteins, that have more than 50% of their surface area accessible to solvent. Ten residues from the N-terminal (Left) and C-terminal (right) are shown. The tendency of the terminal residues to be exposed is evident.

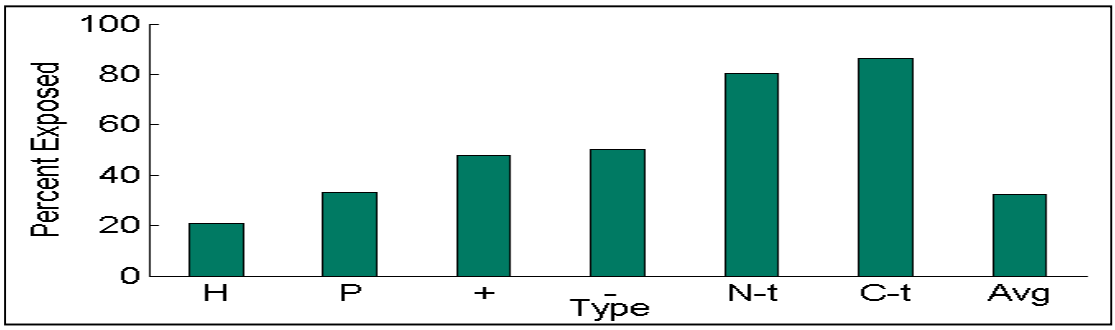

Fig. 25. Exposure of residues by type in PDB. For the dataset of 425 proteins, the percent of residues with more than 50% accessible surface area is shown by residue type. It is clear that terminal residues are much more exposed than charged residues.

It might be argued that solvent accessibility of terminal residues is large because they are missing one of their neighboring residues that could have provided additional shield from the solvent. Thus, in order to probe directly the location of the terminal residues we measured the distance of the terminal residues and all other residues from the center of mass of their proteins. The distance was normalized, in units of standard deviation, to the average distance of residues to the center of mass for each protein. The results, shown in Figure 26, indicate that indeed terminal residues are found much more on the exterior of proteins as compared to any other type of residues.

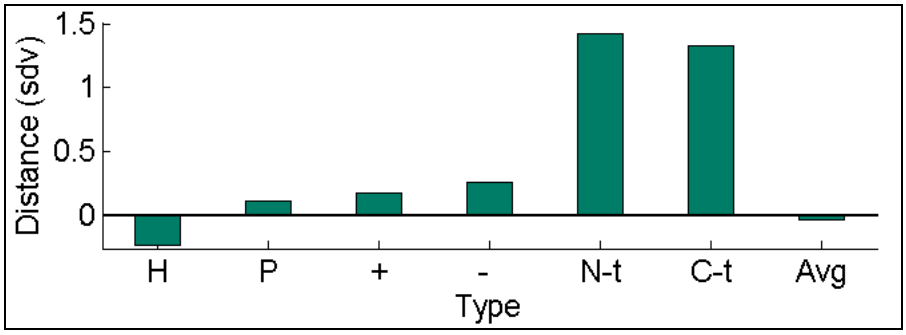

Fig. 26. The distance of residues from the center of PDB proteins. For the dataset of 425 proteins, the distance of residues to the center of mass of their proteins is presented. The average distances, in units of standard deviations of distances in each protein, are grouped by residue type. It is evident that terminal residues are most distant from the center of their proteins.

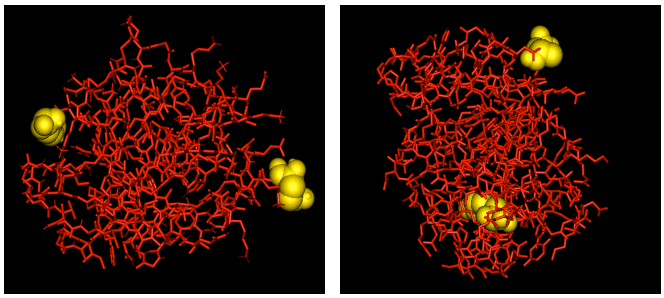

Fig. 27. Terminal residues of proteins. For most proteins both termini are exposed to the solvent, as in cytochrome c552 (PDB code 1C52) (left, where terminal residues are shown as yellow space filling objects). Only in very few cases, termini residues are buried as in (right) staphopain (code 1CV8) a cysteine proteinase where the C-terminal tyrosine is totally buried.

Thus, we can say that indeed protein termini are predominantly located on the surface. Out of the 425 proteins only 132 have one termini buried (i.e. less than 50% exposure), and 13 with both termini buried. If we use a cutoff of 25% exposure then there are only 38 proteins with one buried termini and 2 with both termini buried. With a 10% exposure cutoff, only 14 proteins have one terminal buried and none has both. An example of one of the 14 cases, staphopain, is shown in Figure 27.

4.4. LATTICE MODEL ANALYSIS

4.4.1. ANALYSIS OF MODEL PROTEINS

For extended conformations of model proteins, most residues are exposed. We collected data from 42,450 extended conformations produced by MC simulations and observed (Figure 7) that all residues are exposed in more than 80% of the extended structures. For the three terminal residues on each side, more than 90% are exposed and the very terminal residues are more than 95% exposed. To gather statistics about compact conformations, 3,342 unique random sequences of 25 residues were created. For each sequence, all possible 9,646,215 two dimensional compact non-symmetric conformations that fit into a 6X6 lattice were generated. For these compact structures, the exposure profile of the proteins is quite flat along the structure and all residues have about 70% exposure (Figure 28).

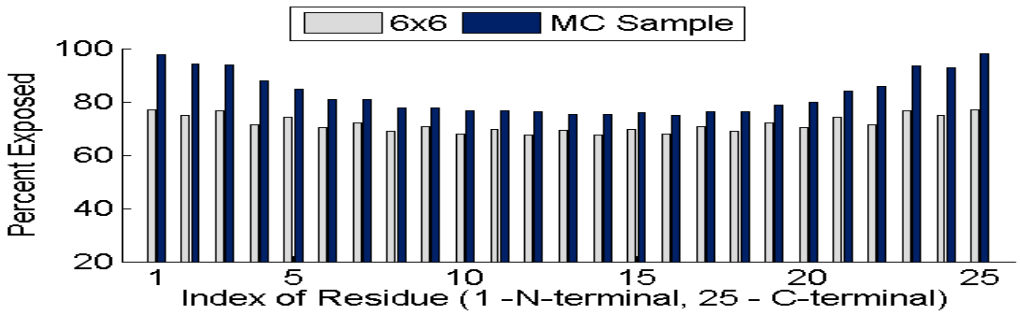

Fig. 28. Exposure profile of extended and compact conformations. For extended (dark) conformation, all residues are more than 80% exposed, with terminal residues reaching more than 95% exposure. For compact (light) conformations the exposure profile is quite flat with all residues having about 70% exposure.

Next we turn to analyze the exposure profile of native structures (i.e. minimal energy structure). We used enumeration of compact structures of the 3,342 sequences composed of an alphabet of 4 types: (H) Hydrophobic, (P) Polar, (+) positively charged and (-) negatively charged, in proportion similar to what is found in the PDB. For each sequence, using a table of mean force potential (reflecting an average of the strength of interactions between the corresponding amino acids (Miyazawa and Jernigan, 1993)) the energy of every compact conformation was evaluated. The conformation with the lowest energy was considered the native conformation. The percent of exposed residues was calculated for all the native structures.

Figure 29 shows the exposure by residue type and demonstrates that for native structures in our model, terminal residues are more exposed than other types of residues. While these exposures are higher than observed for real proteins (see Figure 25), they do show the same rank between residues type as in real proteins.

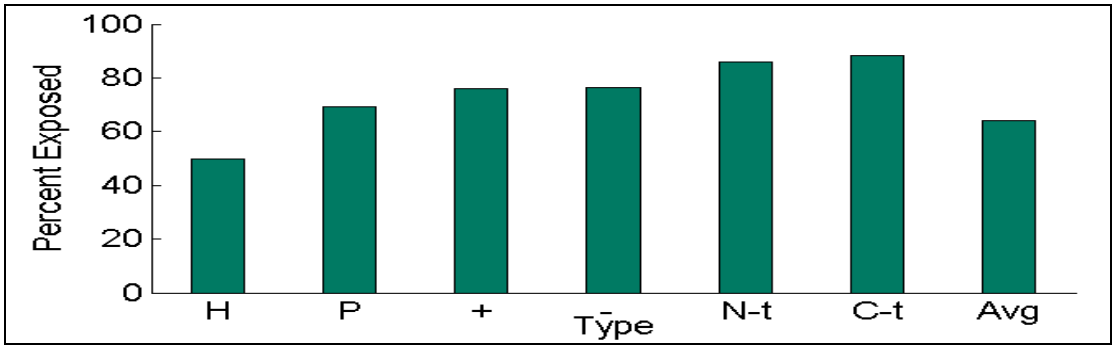

Fig. 29. Exposure of residues by type for compact structures. For the set of 3,342 native conformations of model proteins, the percent of exposure is shown by residue type.

The percent of exposed residues were calculated for the entire set and for the 800 proteins for which the native structure has the largest gap in energy from the averaged energy value. The tendency of the terminal residues to be exposed is slightly higher (89.2%) for those proteins than for the entire set (87%). If we use the top 200 sequences, the tendency goes slightly further higher to 90.8%.

4.4.2. ANALYSIS OF KINETIC FOLDING

4.4.2.1. KINETIC ACCESSIBILITY

In order to examine and characterize the kinetic accessibility of a model sequence to its pre calculated native structure, each of the 800 sequences with the largest energy gaps was simulated and analyzed by the following protocol: *A single simulation* of a model sequence consists of 10^6 Monte Carlo Steps (MCS). The simulation process is terminated once the native conformation is found or after 10^6 MCS. Some flexibility is allowed in reaching the native conformation. We considered the native conformation as found if the simulation reached a conformation within a distance of less than 0.5 Root mean Square Distance from the native conformation. (This distance is roughly equal to two out of the 25 residues being off by one lattice point from the corresponding position in the native conformation.) The number of MCS taken to find the structure is considered as the First Passage Time (FPT). For each sequence, 50 independent simulations were run with the same folding parameters (simulation temperature, local moves library size and tail moves probability). If a model sequence was folded successfully more than a defined percent threshold (e.g. 80%, 40 out of 50 runs), it is considered a *fast folder*; otherwise, it is considered as a *slow folder*. This threshold parameter, as well as other simulation parameters; like tail moves probability and local moves size (L) were varied in our simulations.

4.4.2.2. ANALYSIS

Proteins were divided into two groups, *fast folders* and *slow folders*. The separation was based on the ability of sequences to fold to their native conformation in a Monte Carlo (MC) simulation of 10^6 steps. Each sequence was run 50 times and proteins that were able to find the native conformation in more than a threshold percentage of the simulations were considered *fast folders*, and proteins that found their native structure in less than that threshold percentage of runs were considered *slow folders*. A threshold of 80% (which was used in most simulations) yielded 355 fast folders and 445 slow folders.

A comparison of the percent of exposed residues for fast and slow folding proteins is shown in Figure 30, showing a significant difference. The exposure by residue type for the 355 fast folders is shown in Figure 31.

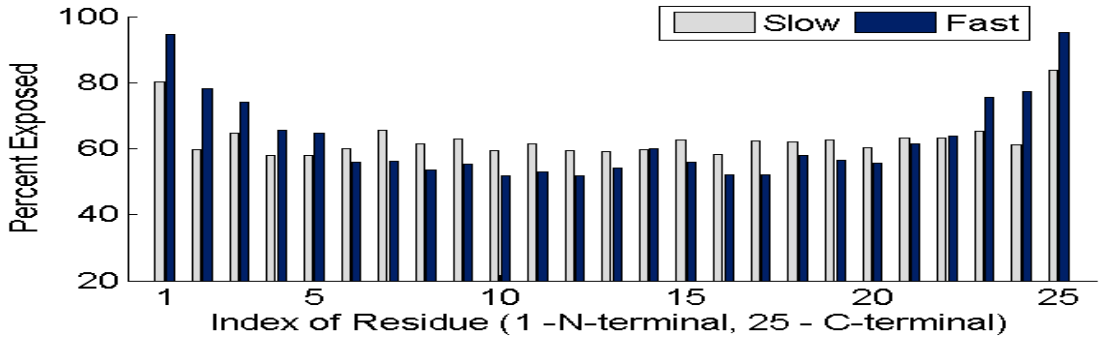

Fig. 30. Percent of exposed residues of fast and slow folders. A probability of 0.15 was used for tail moves and L = 7 of maximum local moves size.

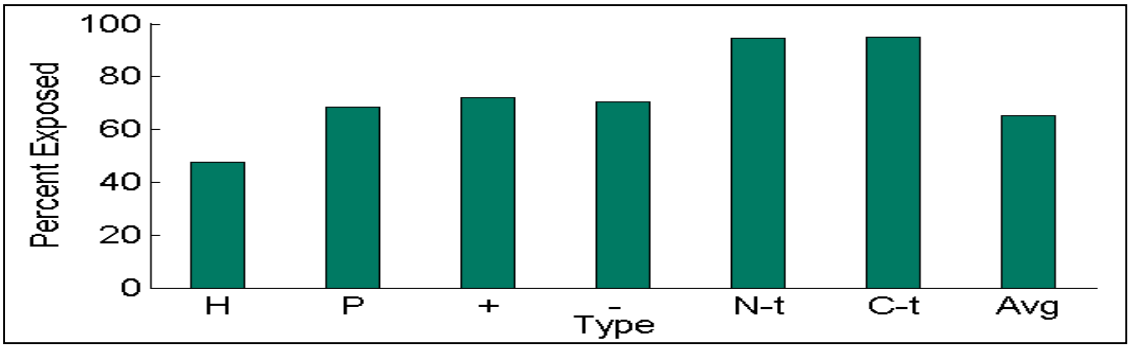

Fig. 31. Exposure of residues by type for fast folders. For 355 fast folding proteins, the percent of exposure is shown. As in real proteins, hydrophobic residues are most buried, followed by polar residues. Terminal residues are more exposed than charged residues.

The simulations were performed using different parameters of local move set, percent of tail moves, threshold between fast and slow folders and in all cases the conclusion was similar: In all simulations proteins that fold fast have a higher percentage of their termini exposed than slow folding proteins, see Figure 32.

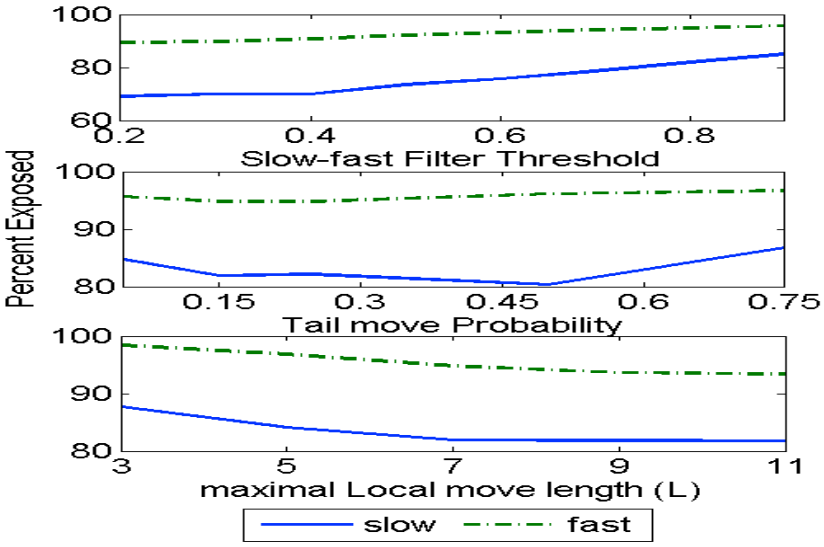

Fig. 32. Termini exposure of fast and slow folding proteins as a function of different simulation parameters. The percent of exposed termini for fast folders is shown in dashed green line and slow folders are shown in solid blue. (**Top**) Changing the threshold separating slow and fast folders from success in 20% of runs to success in 90%. (Tail move probability is fixed to 0.15 and library move size L = 7); (**Middle**) The percent of tail moves compared with internal moves is varied from 0.05 to 0.75 (library move size is fixed to 7 and threshold is 0.8); (**Bottom**) Library move size (L) is varied from 3 to 11 (threshold is fixed to 0.8 and tail move probability equals 0.15). In all cases the fast folding proteins have significantly higher tendency to have their terminal residues exposed.

Furthermore, we performed longer simulations of $6*10^6$ MC moves for two groups of proteins**.** 78 proteins for which the native conformation has the two termini on the surface, and 78 proteins for which in the native structure at least one of the termini was not exposed. Again we saw that proteins with exposed termini fold faster: The average folding time (First passage time) for proteins with exposed termini was 204,000 MCS compared with 404,000 MCS for proteins with at least one buried termini.

4.5. DISCUSSION

We set out to explain why terminal residues of proteins tend to be located on the surface. We first measured the location of the terminal residues in a dataset of 425 monomeric short proteins. We used two different measurements; first we checked the solvent accessibility of these residues and second we checked the distance of these residues from the center of mass of their proteins. Taken together, the results clearly indicate that indeed terminal residues are overwhelmingly located on the surface on proteins.

Based on this finding, we want to understand the mechanisms that force terminal residue to be on the surface. It is clear that many proteins need to have their terminal exposed in order to make them accessible to post translational modifications which are common for both termini (Dixon, 1984; Chung et al 2002). Thus, it can be argued that the location of terminal residues on the surface is a desirable feature that can be selected for by evolution. This feature could have been selected for directly, or, as is common in evolutionary processes, could have been incorporated into other considerations that would have preferred this feature. We suggest that the latter is true, i.e. thermodynamic and kinetic considerations that are known to have an effect on proteins could lead to such a preference.

Using a simple lattice model, we demonstrate that a series of constraints that affect proteins will lead to the preference of terminal residues to be located on the surface. Clearly, for extended conformations of protein, all residues tend to be exposed (Figure 28). But even for compact conformations, our analysis shows that the exposure profile is quite flat, and all residues tend to be equally exposed (Figure 28). When only conformations with minimal energy (i.e. native conformations) are considered, terminal residues start to prefer to be located on the surface. When native conformations with a profound energy gap are considered then this tendency increases. If we look at proteins that can fold fast in kinetic simulations, then we see that the tendency of terminal residues to be exposed is increased further (Figure 30,31). Proteins that require that terminal residues will be tucked inside the core may be prohibitively complicated to fold. To conclude, we suggest that the tendency of terminal residues of proteins to be located on the surface is a result of thermodynamic and kinetic selection processes. Indeed, model proteins that have been selected using these considerations (Figure 31) exhibit similar exposure profile to real proteins (Figure 25).

CHAPTER V

References

Alexandrov N. (1993) Structural argument for N-terminal initiation of protein folding. *Protein Sci.*, **2**, 1989-91.

Anfinsen C.B. (1973) Principles that govern the folding of protein chains, *Science* **181**, 223-230.

Betancourt M.R. and Thirumalai, D. (1999) Exploring the kinetic requirements for enhancement of protein folding rates in the GroEL cavity, *J. Mol. Biol*., **287**, 627-644.

Braig K., Otwinowski Z., Hegde R., Boisvert D.C., Joachimiak A., Horwich A.L. and Sigler P.B. (1994) The crystal structure of the bacterial chaperonin GroEL at 2.8 A, *Nature*, **371**, 578-586.

Brinker A., Pfeifer G., Kerner M.J., Naylor D.J., Hartl F.U., Hayer-Hartl M. (2001) Dual function of protein confinement in chaperonin-assisted protein folding, *Cell*, **107**, 223-33.

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

Bukau B., Horwich A.L. (1998) The Hsp70 and Hsp60 chaperone machines, *Cell*, **92**, 351-66.

Chan H.S. and Dill K.A. (1996) A Simple Model of Chaperonin-Mediated Protein folding. *PROTEINS* **24**, 345-351.

Christoph S. et al (2004) Mechnism of the eukaryotic chaperonin: protein folding in the chamber of secrets, *TRENDS in Cell Biology*, **14**.

Christopher, J.A. and Baldwin, T.O. (1996) Implications of N and C-terminal proximity for protein folding. *J. Mol. Biol*., **257**, 175-187.

Chung J.J., Shikano S., Hanyu Y., Li M. (2002) Functional diversity of protein C-termini: more than zipcoding? *Trends Cell Biol*., **12**, 146-50.

Danziger O., Rivenzon-Segal D., Wolf S.G. and Horovitz A. (2003) Conversion of the allosteric transition of GroEL from concerted to sequential by the single mutation Asp-155 -Ala. *Proc. Natl. Acad. Sci. U S A.,* **100**, 13797-13802.

Dill K.A., Bromberg S., Yue K., Fiebig K.M., Yee D.P., Thomas P.D. and Chan H.S. (1995) Principles of protein folding - a perspective from simple exact models, *Protein Sci.* **4**, 561-602.

Dinner A.R., Abkevich V., Shakhnovich E. and Karplus M. (1999) Factors That Affect the Folding Ability of Proteins, *PROTEINS.* **35**, 34-40.

Dinner A.R., Sali A., Smith L.J., Dobson C.M. and Karplus M. (2000) Understanding protein folding via free-energy surfaces from theory and experiment, *Trends Biochem. Sci.* **25,** 331-339.

Dixon H.B.F. (1984) N-terminal modification of proteins. *Journal of Protein Chemistry*, **3**, 99-108.

Dobson C.M., Karplus M. (1999) The fundamentals of protein folding: bringing together theory and experiment, *Curr. Opin. Struct. Biol.*. **9**, 92-101.

Gutin A.M., Abkevich V.I. and Shakhnovich E.I. (1995) Is burst hydrophobic collapse necessary for protein folding? *Biochemistry*, **34**, 3066-76.

Horovitz A.L, Willison K.R. (2005) Allosteric regulation of chaperonins. *Curr. Opin. Struct. Biol.*, **15**, 646-51.

Horwich A.L., Fenton W.A. and Rapoport T.A. (2001) Protein folding taking shape. Workshop on molecular chaperones, *EMBO Rep,* **2**, 1068-73.

Jacob Etai and Unger Ron (2007) A tale of two tails: why are terminal residues of proteins exposed? *Bioinformatics* **23**, 225-230.

Kabsch W. and Sander C. (1983) Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*, **22**, 2577-2637.

Karplus M. (1997) The Levinthal Paradox: yesterday and today. *Fold. Des*. 2, 69-76.

Kerner M.J., Naylor D.J., Ishihama Y., Maier T., Chang H.C., Stines A.P., Georgopoulos C., Frishman D., Hayer-Hartl M., Mann M., Hartl F.U. (2005) Proteome-wide analysis of chaperonin-dependent protein folding in Escherichia coli. *Cell*. **122**, 209-20.

Krishna M.M., Englander S.W. (2005) The N-terminal to C-terminal motif in protein folding and function. *Proc. Natl. Acad. Sci. U S A*., **102**, 1053-8.

Laio A, Micheletti C. (2006) Are structural biases at protein termini a signature of vectorial folding? *Proteins*, **62**, 17-23.

Lehninger Principles of biochemistry, Nelson & Cox, Third Edition.

Metropolis N., Rosenbluth A.W., Rosenbluth M. N., Teller A. H., and Teller E. (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087-1091.

Milton J. Schlesinger (1990) Heat Shock Proteins, *J. Biol. Chem.*, **265**, 21.

Miyazawa S., Jernigan R.L. (1993) A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. *Protein Eng.*, **6**, 267-278.

Munchbach M., Dainese P., Staudenmann W., Narberhaus F., James P. (1999) Proteome analysis of heat shock protein expression in Bradyrhizobium japonicum, *Eur J Biochem*., **264**, 39-48.

Ranson N.A., Farr G.W., Roseman A.M., Gowen B., Fenton W.A., Horwich A.L., Saibil H.R. (2001) ATP-bound states of GroEL captured by cryo-electron microscopy. *Cell*, **107**, 869-79.

Rivenzon-Segal D., Wolf S.G., Shimon L., Willison K.R. and Horovitz A. (2005) Sequential ATP-induced allosteric transitions of the cytoplasmic chaperonin containing TCP-1 revealed by EM analysis, *Nature structural and molecular biology*, **12**, 233-237.

Rutherford S.L. (2003) Between genotype and phenotype: protein chaperones and evolvability, *Nat Rev Genet.,* **4**, 263-74.

Saibil H.R. and Ranson N.A. (2002) The chaperonin folding machine, *TRENDS in Biochem. Sci.*, **27**, 12.

Sali, A., Shakhnovich, E., Karplus, M. (1994) How does a protein fold? *Nature*, **369**, 248–251.

Shakhnovich E.I. (1994) Proteins with selected Sequences Fold into Unique Native Conformation, *Physical review letters*, **72**, 3907-3910.

Shin-ichi Yokota, Hideki Yanagi, Takashi Yura and Hiroshi Kubota (2000) Upregulation of cytosolic chaperonin CCT subunits during recovery from chemical stress that causes accumulation of unfolded proteins, *European Journal of Biochemistry*, **267**, 1658.

Shrake A., Rupley J.A. (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.*, **79**, 351-371.

Skolnick J., Kolinski A. (1991) Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J Mol Biol.*, **221**, 499-531.

Stan G, Brooks B.R., Thirumalai D. (2005) Probing the "annealing" mechanism of GroEL minichaperone using molecular dynamics simulations, *J Mol Biol*., **22**, 817-29.

Takagi F., Koga N., Takada S.. (2003) How protein thermodynamics and folding mechanisms are altered by the chaperonin cage: molecular simulations, *Proc. Natl. Acad. Sci. U S A*.,**100**, 11367-72.

Tapan K. Chaudhuri et al, GroEL/GroES-Mediated Folding of a protein too large to be encapsulated, Cell, Vol. 107, 235-246 (2001).

Thirumalai D., Lorimer G.H. (2001) Chaperonin-mediated protein folding, *Annu Rev Biophys Biomol Struct*., **30**, 245-69.

Thornton J.M., Chakauya B.L. (1982) Conformation of terminal regions in proteins. *Nature*, **298**, 296-7.

Thornton J.M., Sibanda B.L. (1983) Amino and carboxy-terminal regions in globular proteins. *Mol Biol.*, **167**, 443-60.

Unger R., Moult J. (1993) Genetic algorithms for protein folding simulations, *J. Mol. Biol.,* **231**, 75-81.

Unger R., Moult J. (1996) Local interactions dominate folding in a simple protein model. *J. Mol. Biol.*, **259**, 988-994.

Valpuesta J.M., Martin-Benito J., Gomez-Puertas P., Carrascosa J.L., Willison K.R. (2002) Structure and function of a protein folding machine: the eukaryotic cytosolic chaperonin CCT, *FEBS Lett.,* **529**, 11-6.

Van der Vaart A., Ma J. and Karplus M. (2004) The unfolding action of GroEL on a protein substrate, *Biophys. J.,* **87**, 562-573.

Wang J.D., Herman C., Tipton K.A.., Gross C.A., Weissman J.S. (2003) Directed evolution of substrate-optimized GroEL/S chaperonins, *Cell*, **111**, 1027-39.

Wolynes Peter G., Onuchic Jose N. and Thirumalai D. (1995) Navigating the Folding Routes, *Science* **267**, 1619–1620.

CHAPTER VI

Appendixes

Appendix A

Grid Report

2410.2006.

High Throughput Computing Performance of the EGEE project GRID platform

Written by Etai Jacob

Abstract

The aim of this document is to (1) introduce the GRID technology and the EGEE project, (2) describe and illustrate the performance and usage of the GRID and its benefits.

Preface

EGEE

The EGEE (Enabling Grid for E-science) project brings together experts from over 27 countries with the common aim of building on recent advances in Grid technology and developing a service Grid infrastructure in Europe which is available to scientists 24 hours-a-day.

The project aims to provide researchers in academia and industry with access to major computing resources, independent of their geographic location. The EGEE project will also focus on attracting a wide range of new users to the Grid.

With funding of over 30 million Euro from the European Commission, the project is one of the largest of its kind. EGEE is a two-year project conceived as part of a four-year programme, where the results of the first two years will provide the basis for assessing subsequent objectives and funding needs.

EGEE will make Grid technology available on a regular and reliable basis to all European science, as well as Research and Development. Like the World Wide Web, which was initially developed for specialized scientific purposes, the impact of the emerging Grid technology on European society is difficult to predict at this stage but is likely to be huge.

GRID in Israel

IAG (Israeli Academic Grid) is an independent body within the IUCC founded by all Israeli universities. Its vision is to create a computational grid encompassing computer facilities in all participating Israeli universities, serve as a facility for performing tasks that necessitate large computational and data-intensive capabilities and become a meeting ground for users, developers and industry, all interested in promoting grid technology in Israel.
Active GRID sites within the IAG are already in Weizmann Institute, Tel Aviv University, the Technion and the Open University.

GRID description and performance

Environment

Developing on the GRID was done remotely through a Linux server in Tel Aviv University. A specific interface and job description language (JDL) are used to execute jobs on the GRID platform. Programs to be execute on the GRID are implemented by any standard programming language supported by basic Linux personal computer (i.e. C, C++, Perl, C shell, FORTRAN).

High Throughput Computing

It must be emphasized that all tasks which are relevant to be executed on the GRID concern with how many computing operations per month or per year can extract from the computing environment rather than the number of such operations the environment can provide per second or minute.

Qualifications needed

A developer need to have moderate UNIX developing experience with knowledge of Shell scripts programming and basic system familiarity. One has to learn job description language (JDL) and the GRID user interface commands to be able performing relevant tasks.

Registering

The GRID is comprised of virtual organizations where each organization has many sites of computer clusters (usually linux/unix PC or servers). There are organizations with different number of sites from few hundred computers to few thousands. Each organization has its own policy for accepting new members concerning the type of tasks they want to run (i.e. Bio-computing, physics, Medical-computing). Usually, one can join an organization only through a site which is part of the organization he wants to join. Registering as an active user on the GRID takes about two to five weeks depending on the organization and is done through a certificated person in Israel.

Performance Examination

In order to examine the real capability of the GRID, we performed tasks with different loads. The tasks were applications of lattice model simulations in the field of computational biology, implemented in C++ on Linux. It is important to mention that these sorts of applications mostly need CPU time for computations and almost no I/O time (i.e. Writing or reading from files). Simulations with different durations were executed from few thousand to hundred of thousand times (1000- 200,000) in tasks.

The following graphs describe the computation time gained by using the GRID platform on an organization (called SEE) with moderate resources of about few hundred computers organized in clusters on different sites. The SEE virtual organization is not considered to be a heavily loaded. It should be mentioned that there are other organizations much bigger than the SEE organization of few thousands computers.

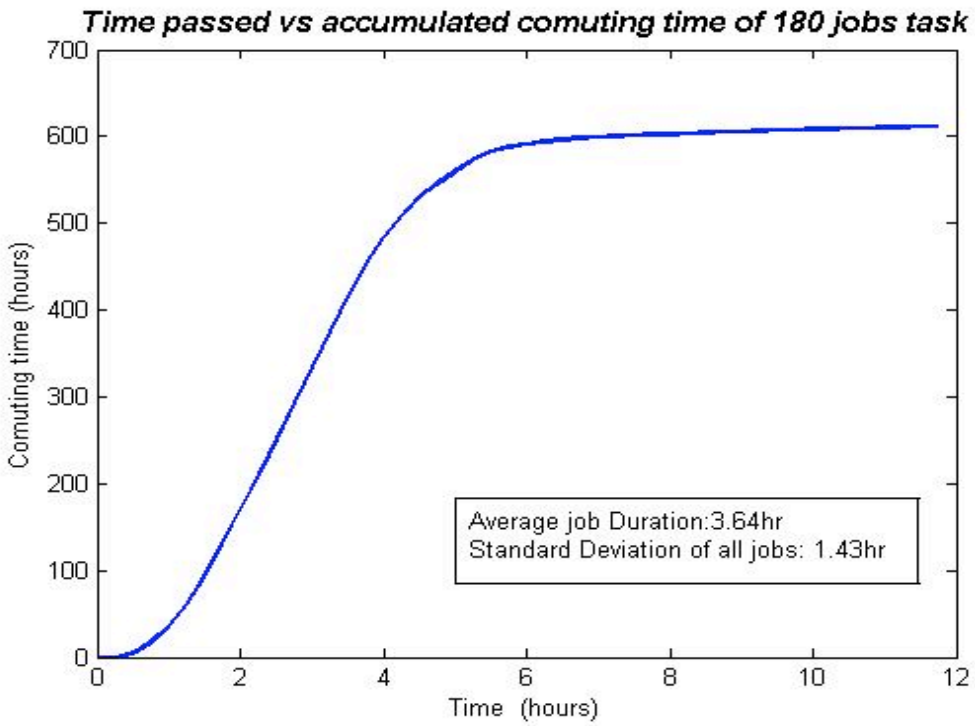

Fig. 1. A description of accumulated computing time in resolution of quarter of an hour of a task of 180 jobs submitted to the GRID. The average job duration was approximately 3 and a half hours. It is obvious to see that 5 hours of execution on the GRID platform equal more than 600 one CPU time (25 days).

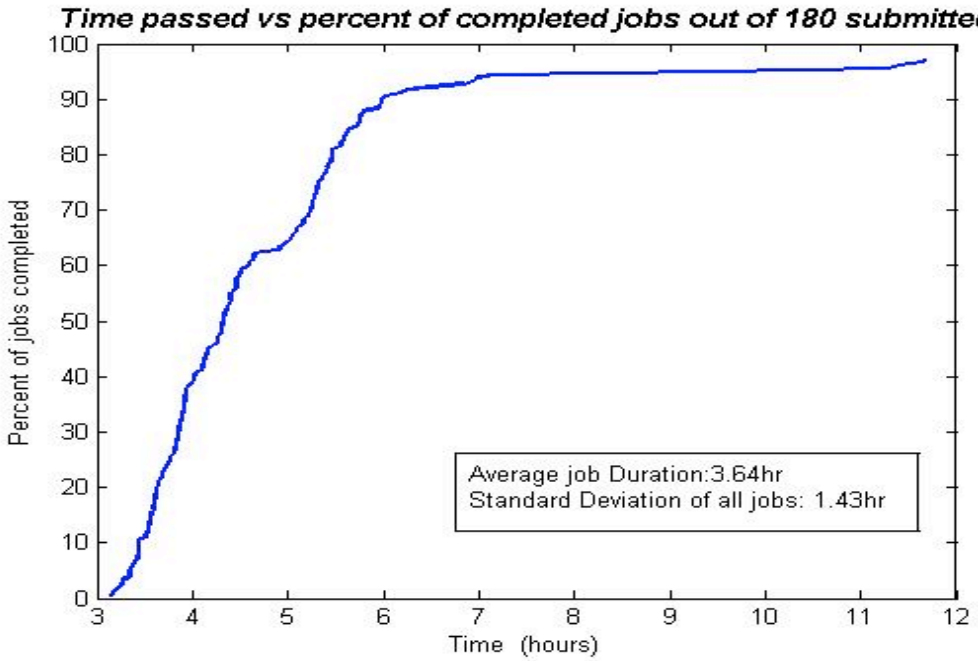

Fig. 2. A description of the percent of completed jobs on the same 180 jobs task as in figure 1. It can be seen that after 6 hours most of the jobs are successfully completed, where each job takes on the average 3.64 hours.

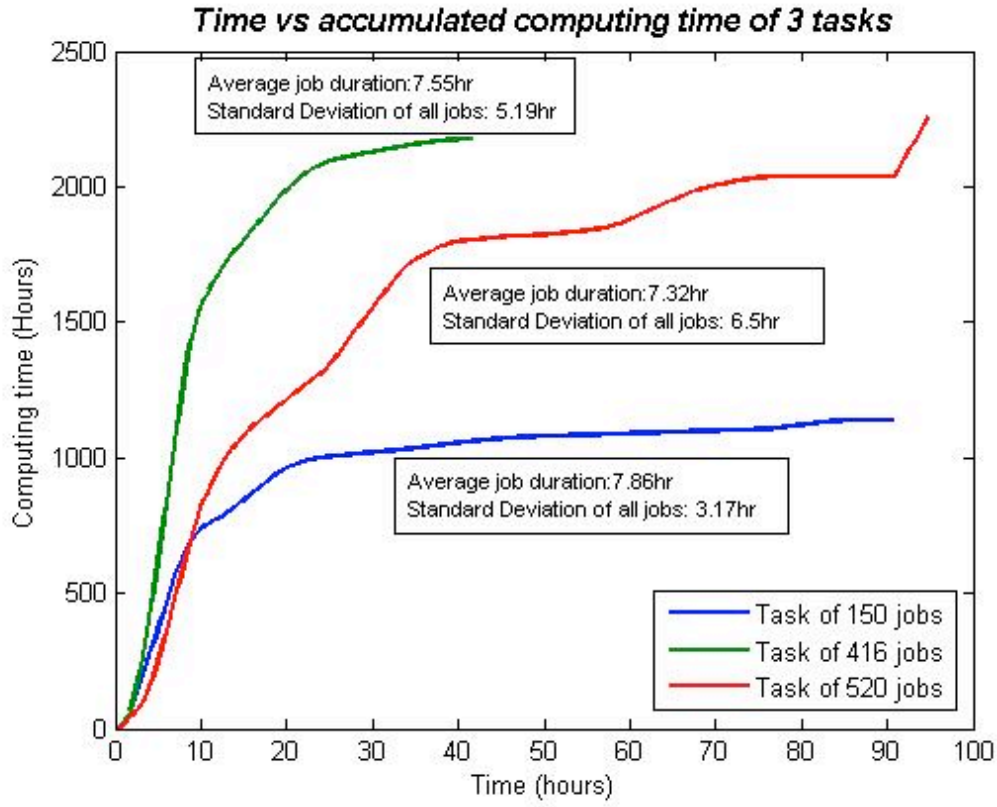

Fig. 3. A description of three different tasks each consist different number of jobs but the average duration of a job in different tasks is about the same. The maximum computing time gained, more than 2000 hours (83 days) was by the 416 jobs task (green). It should be mentioned that the tasks were submitted on different time so the GRID might not have been with the same load.

Task	Average time to begin running	SDV	Percent of Successful jobs
416 jobs of avg 7.5hr duration	3.07	5.19	288/416
180 jobs of avg 3.64hr duration	0.23	0.45	168/180
150 jobs of avg 7.86hr duration	7.73	16.23	145/150
416 jobs of avg 2.64hr duration	1.85	2.84	288/416
520 jobs of avg 7.32hr duration	27.13	32.57	307/520

Table 1. The above table describes the conducting of the GRID interface. It is obvious that the bigger the task, the greater the number of none completed jobs in a time window of less than 4 days. The maximum number of successful jobs was about 300. It can be concluded that it is better to submit tasks of maximum 300 jobs. Again, it should be emphasized that the sort of tasks submitted on the GRID are high throughput oriented (i.e. jobs within a task are not depended on each other) , hence the fact that fraction of the jobs are not completed is negligible and one can submit the same task twice or submit in advance an over size task.

Conclusions

The GRID platform is very powerful in the means of high throughput computing. With the GRID technology one can afford performing research he could never think of doing on a single or few clusters of computer on the campus. In a day time one can gain more than a month computing time. The GRID computing power enhances research quality (better statistics) and experiment flexibility. It takes few weeks to acquire the capability and understanding of the GRID technology, but after a single GRID-computing day all this time is gained back.

Appendix C

Different mechanistic requirements for prokaryotic and eukaryotic chaperonins: a lattice study.

Etai Jacob, Amnon Horovitz and Ron Unger
*Submitted to **ISMB/ECCB 2007**, Bioinformatics.*

Appendix C

A tail of two tails:
Why are terminal residues of proteins exposed?

Etai Jacob and Ron Unger (2007)
Bioinformatics **23**, 225-230.