# *Protein Folding Networks and Levinthal's Paradox*

Virtually all biological processes in a living cell involve proteins. These long amino-acid chains fold into well-defined three dimensional structures whose physical (chemical) properties are responsible for the protein's function. Thus, understanding and predicting protein structure is of great interest in both unveiling the way cells work, and designing new, desired protein functions.

Shortly after the first spatial structure of a protein has been experimentally determined, the question of how freshly assembled amino-acid chains can fold into these structures in biologically relevant time has been raised. In a lecture in 1969 C. Levinthal pointed out that a protein made of 150 amino-acids has around $10^{300}$ different conformations. A complete sampling of this whole space would take several times the age of the universe even for chains as short as 40 monomers [1]. Levinthal conjectured that proteins have to follow well-determined *folding pathways* that guide them to their final, stable conformations. One consensus in protein folding today is that most functional proteins have energy landscapes with funnel structure: they fold by hopping through small barriers separating lower and lower local minima. However, the question remains: how did Nature evolve proteins that have favorable energy landscapes? Sampling all possible amino-acid sequences while selecting ones capable of fast and reliable folding poses similar problems as Levinthal's original paradox, only stretched out to evolutionary time-scales.

We are currently working towards demonstrating that configuration spaces of large physical systems (such as peptide chains, proteins or atomic clusters) have a few generic properties that hold the answer to when and how funneled landscapes emerge. A useful way to define discrete configuration spaces is to look at them as complex networks: configurations are nodes of the network, while an elementary step the system can take from one configuration to another is a link. Our aim is to study, characterize and model these very large networks in order to gain insight into properties of the dynamics of motion in configuration space.

A few recent papers provided an initial insight on the topology of these networks. Scala *et al* [2] constructed the configuration network of a 15 component model polymer chain on a lattice, using a few allowed moves as links. They have found that the graph has the *small-world* property: in a network of $N$ nodes the shortest paths between configurations are of the order of $\log N$. More importantly, the also found that the connectivity or degree distribution of the network is binomial: all configurations have a similar number of nearest neighbors.

Interested in the generality of the configuration network properties found by Scala *et. al.*, we constructed a simple robot arm model system. This arm is made of equal length rods connected through joints that allow for three distinct angles (Fig. 1). Two configurations are connected if the change of only one joint angle (elementary step) to a neighboring position can inter-convert them (blue lines on 1). This configuration network has the same general properties as the network studied by Scala *et.al.*, with a characteristic value for the degrees and short paths. Recently, Rao and Caflisch [3] mapped out the conformational space of a small, 20-residue designer protein, *beta3s*, using implicit solvent molecular dynamics simulations (Fig. 2). In order to allow the protein to sample the configuration space efficiently, they used a temperature at which the chain goes back and forth between native and denaturated states. They have found that the emerging configuration network is scale-free, with a power-law degree distribution of exponent $\gamma = -2$. Thus, this landscape is dominated by a few hubs: configurations with a significantly higher number of neighbors than average. An intriguing finding of the paper is that the scale-free nature of the configuration space persists even for a chain obtained via random reshuffling of the original amino-acid sequence. This seems to indicate that the scale-free topology of the space is not specific to *beta3s*, rather it is generic to amino acid chains. Moreover, non-homogeneous configuration networks are also reported by Doye and Mason [4]. They map the energy landscapes of small Lennard-Jones atomic clusters into networks using the local minima of the

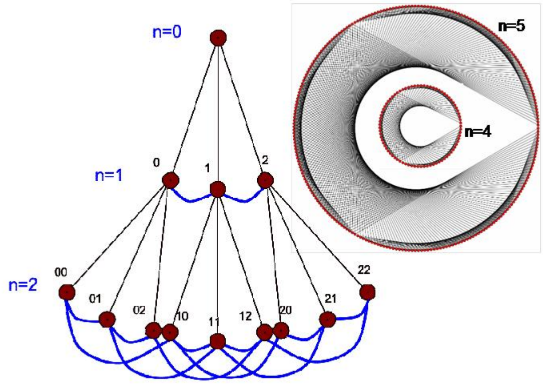

Figure 1: A simple robot arm model: $n$ is the number of joints, blue links indicate the configuration network for $n = 1$ and 2. The inset shows the network for $n = 4, 5$.

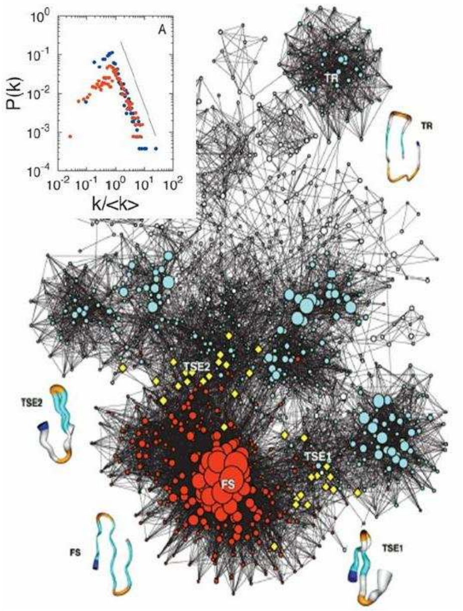

Figure 2: *Beta3s* conformation space network. From [3]. Inset: degree distribution of the network.

system as nodes and saddle-points between to minima as links: a coarse-graining of the complete configuration network. They also find graphs with power-law degree distributions.

These results seem to contradict our observation that configuration networks are more or less homogeneous. The scale-free nature of the network mapped out by Rao and Caflisch, however, would resolve the Levinthal paradox. Scale-free networks are known to have an average shortest path on the order of $\log(\log N)$, thus even a system with $10^{300}$ conformations is typically less than 10 hops away from it's minimum (*ultra-small*)! The questions is thus how to marry the two contradicting observations.

Recent results on gradient networks provide a simple framework to understand these seemingly contradicting results. Toroczkai *et. al.* [5] studied network flows, where flows are generated by gradients of a scalar field distributed on nodes of a network. A gradient network can be defined on any substrate graphs with a scalar assigned to it's nodes: the links of the gradient network indicate the direction of the largest local gradient at each node *on the substrate network*. They showed that for a random distribution of the scalars among the nodes, a variety of relatively homogeneous substrate networks (Erdős–Rényi, Small–World, etc.) give rise to gradient networks with scale-free in-degree distribution. Thus, scale-free flow structures arise naturally even on homogeneous substrate networks via selection of links supporting significant amounts of flow.

The discrepancy between homogeneous configuration networks (Scala *et al*, robot arm model) and the results of Rao and Caflisch can be easily resolved in light of this result [6]. The topology of configuration spaces is complemented by the energies of the configurations, together they determine the landscape under configurational motion. Changes in the protein conformation are due to the energy differences along the links of the substrate (i.e., configuration) network, thus a molecular dynamics simulation could sample the whole space in an unbiased manner at infinite temperatures only. At lower temperatures the network uncovered by the system would resemble the gradient network. Thus the scale-free network of Rao and Caflisch is not the actual configuration network of the protein, it's a biased sampling of its links. It nonetheless characterizes the landscape responsible for the folding. What is a bit harder, is to find the correct correlations between the topology of the configuration network and the energy values associated to the nodes. We have just solved this problem [6], and this has lead us to the recovery of the $\gamma = -2$ exponent found by the molecular simulations of Rao and Calfish.

A more difficult problem is to understand how shortcuts of configuration networks are distributed and how they arise. We hope that once the nature of shortcuts is understood, small alterations to a complex biological system (such as docking of a small molecule) could alter the configuration space with new shortcuts to facilitate more desired conformational dynamics, such as faster folding. Aside from the challenging task of proposing such an experiment, perhaps shortcuts could be used as computational tricks in lengthy simulations. The slightly altered system could quickly move close to a desired configuration and evolve from there after the change is undone.

## References

[1] C. Levinthal, *J. Chim. Phys.* **65**, 44 (1968); D.B. Wetlaufer, *PNAS* **70**, 691 (1973).

[2] A. Scala, *et. al. Europhys.Lett.* **55**, 594 (2001).

[3] F. Rao and A. Caflisch, *JMB* **342**, 299 (2004).

[4] J.P.K. Doye and C.P. Massen, [cond-mat/0411144] (2004).

[5] Z. Toroczkai, K.E. Bassler, *Nature*, **428**, 716 (2004); Z. Toroczkai, *et. al.* Gradient Networks, *submitted* (2005); cond-mat/0408262;

[6] E. Ravasz and Z. Toroczkai, *under publication* (2005).

Contact Information: Erzsébet Ravasz — Center for Nonlinear Studies, Los Alamos National Laboratory, MS-B258 T-03, Bl. 1690, 125, Los Alamos, NM, 87545. Phone: (505) 667-9467, email: eravasz@lanl.gov.