Characterizing the statistical properties of protein surfaces

Ji Hyun Bak[1,2], Anne-Florence Bitbol[1,2] and William Bialek[1,2]

***Short Abstract* — In order to ensure the accuracy as well as the specificity of biological signaling, it is crucial that proteins recognize their correct interaction partners. An important ingredient of recognition is shape complementarity. Not only does shape complementarity allow the short-ranged chemical attractions to be at work, but it also provides additional degrees of freedom in the space of interactions. Here we aim to characterize the statistical properties of the ensemble of protein surface shapes. Specifically, we evaluate the intrinsic dimensionality of the space of protein surfaces, and show how it is related to the characteristic length scale of the surfaces.**

***Keywords* — protein-protein interaction, specificity, shape complementarity, space of protein surfaces**

I. Introduction

PROTEINS and their interactions form the body of the signaling transduction pathways in many living systems. In order to ensure the accuracy as well as the specificity of signaling, it is crucial that proteins recognize their correct interaction partners. How difficult, then, is it for a protein to discriminate its correct interaction partner(s) from the possibly large set of other proteins it may encounter in the cell?

An important ingredient of recognition is shape complementarity. While there has been much attention to the determinants of protein-protein recognition [1], most efforts were directed to the role of chemical compositions, and we are still a long way from a system-level understanding of the role of shape. In fact, shape complementarity is a prerequisite for the recognition process, because of the short-ranged nature of chemical interactions.

The ensemble of protein shapes should be constrained by the need for maintaining functional interactions while avoiding spurious ones. There must be enough degrees of freedom to accommodate the whole proteome while maintaining the specificity of interactions, while too many effective degrees of freedom would make the recognition problem difficult. To address this aspect of protein recognition, we start by investigating the dimensionality of the space of protein shapes.

II. Methods

We consider the ensemble of proteins in terms of their three-dimensional shapes, more precisely in terms of their solvent-excluded surfaces. We take into account all complete high-resolution X-ray crystalized structures from *E. coli* non-DNA-binding cytoplasmic proteins that can be retrieved from the Protein Data Bank, resulting in a database of 397 proteins.

[1]Joseph Henry Laboratory of Physics, [2]Lewis-Sigler Institute of for Integrative Gemonics, Princeton University, Princeton, NJ 08544 USA
Email: jhbak@princeton.edu

In order to measure the intrinsic dimension of the dataset, we apply a statistics that was first developed in chaotic theory, called the correlation dimension [2], to the high-dimensional space where each point corresponds to a shape object. The space of surfaces is the set of geodesic-disk patches with a fixed surface area, sampled from the protein surfaces in the dataset. The space of curves is the set of geodesic curves with a fixed length, also sampled from the surfaces.

III. Results

The dimension D_2 of the space of surface patches turns out to be high, about $D_2 = 30$ for patches with area 1000 $\mathring{A}^2$, typical size of reported interfaces [3]. However, it is known that the effect of systematic error in calculating the correlation dimension of a finite dataset aggravates as the true dimension increases [4]. On the other hand, if we consider the dimension D_1 of the space of geodesic curves sampled from the surfaces, generically we can expect the dimensions to scale as $D_2 \sim (D_1)^2$. Our statistics can therefore be more reliable with this lower dimensional dataset.

We find that D_1 grows with the length of the curves, and that there is a clear transition between two regimes of growth: there is an initial steep growth, followed by a less steep and linear growth regime at larger curve lengths. We argue that this pattern of growth can be explained by a single length scale, which corresponds to the characteristic scale of the protein surface (that can actually be measured). Beyond the characteristic scale, there is roughly one extra dimension per characteristic scale, representing an extra degree of freedom; below this scale the steep growth reflects the smaller-scale "noisy" features, such as the roughness of the surface at the atomic scale. We test this idea by generating synthetic curves characterized by a single correlation length and calculating the dimensionality of the generated dataset.

Taking this argument further, we also discuss how these results may be connected back to the question of interaction specificity through a simple model with harmonic interactions, where each surface is modeled as a set of independent points.

References

[1] Janin J, Bahadur RP, Chakrabarti P (2008) Protein-protein interaction and quaternary structure, *Quart. Rev. Biophys.* **41**, 133-180.
[2] Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors, *Physica* **9D**, 189-208.
[3] Levy ED (2010) A simple definition of structural regions in proteins and its use in analyzing interface evolution, *J. Mol. Biol.* **403**, 660-670.
[4] Nerenberg MAH, Essex C (1990) Correlation dimension and systematic geometric effects, *Phys. Rev. A* **42**, 7065-7074.