

# Protein evolution and divergence: the role of adjacent structures

Yanlong O Xu<sup>1</sup>, Randall Hall<sup>2</sup>, Richard Goldstein<sup>3</sup>, David D. Pollock<sup>4</sup>

**Keywords:** divergence, homologous recombination, lattice model, thermodynamics.

## 1 Introduction.

We conducted a study on the process of divergence with regards to structural designability and thermodynamic competition with adjacent structures in modeled protein evolution system. This system includes both thermodynamic features of protein stability and population dynamics. We refer to this approach as *ab initio* evolution to emphasize that the equilibrium details of variant fitnesses arise from the physical principles of the system, and not from any pre-conceived notions or arbitrary mathematical distributions. We measured how context changes as divergence proceeds by determining the fitness of recombinants between divergent proteins. The distance distribution of adjacent structures is known to play a key role in determining designability, and we therefore considered the occupancy of adjacent structures in mutants and recombinants. We also co-selected for pairs of structures to analyze the otherwise inaccessible boundary space between them. We use the results of our analyses to guide improved methods for accurately approximating folding probabilities in more complex systems that would otherwise be beyond computational feasibility.

## 2 Model and Methods.

We modeled proteins as compact structures on a two-dimensional compact 5x5 lattice, with folding energies obtained from Miyazawa-Jernigan amino acid interaction potentials. Thus, the equilibrium probability that a protein sequence will fold to a particular conformation is given by Boltzmann statistic. We modeled evolution in constant-size haploid populations of 1000 individuals with mutation rates of 0.05 mutations per protein per generation. Individual fitness was based on the probability of folding into the “native” or functional structure(s). Populations were allowed to evolve to equilibrium, then duplicated. Each population was then allowed to evolve independently under identical conditions, and samples were taken every 500 generations. For each sample, the most frequent sequences in each population were recombined at all 24 possible sites. The structural occupancy of each structure in evolution was defined as the probability that a sequence would fold to that structure, averaged over all recombinants and over time. Structures were classified based on 98% designability, defined as the fraction of random sequences that will fold to the structure with a probability of at 98%. All structures were grouped into high-, medium-, and low-designable categories, and representatives from each category were randomly chosen and simulated separately.

## 3 Results and Discussion.

---

<sup>1</sup> Dept. of Chemistry, Louisiana State University, Baton Rouge, LA, USA. Email: yxu3@lsu.edu

<sup>2</sup> Dept. of Chemistry, Louisiana State University, Baton Rouge, LA, USA. Email: rhall@lsu.edu

<sup>3</sup> National Institute for Medical Research, Mill Hill, London, rgoldst@nimr.mrc.ac.uk

<sup>4</sup> Dept. of Biology, Department of Biological Sciences, and Biological Computation and Visualization Center, Email: dpollock@lsu.edu

Structural designability, the fraction of sequence space that folds to a structure, strongly affects divergence and recombinant function. We found the same log-linear relationship between structural distance and occupancy in both mutants and recombinants (Figure 1A), and a linear relationship between structural distance and fitness in co-selected populations (Figure 1B). Surprisingly, there was no difference between high- and low-designable structures in occupancy or co-selected fitness when comparing structure pairs of comparable distance (i.e., with the same number of shared contact pairs). We also did not find any relationship between equilibrium fitness and the designability of the structures in the pair (data not shown). This indicates that designability is determined by the number and distances of adjacent structures, and not by mutational or fitness biases. The results clarify the extent to which it is necessary to incorporate alternative structures when trying to understand evolutionary trajectories of real proteins. They also suggest that a simplified and effective way to approximate protein folding probability is to consider the closest structures, or to sample structure space based on structural distance when calculating the partition function for Boltzmann statistics. Figure 1C shows that the correlation is good between the probability of folding to the native structure and the results from an approximate partition function (based on sampling structures according to contact pair distance from the native structure). This work is supported in whole or in part by the National Science Foundation under Grant Number EPS-0346411 and the State of Louisiana Board of Regents Support Fund.

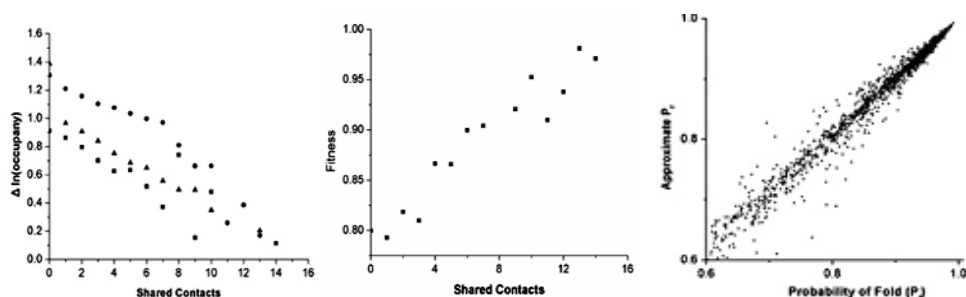


Figure 1: (A) Difference in structural log occupancy for parental and recombinant sequences as a function of distance to the native structure (shared contacts). (B) Fitness of co-selected structure pairs, as a function of shared contacts. (C) Accuracy of fitness approximation.

## References

- [1] Goldstein, R. A. 2003. Evolutionary perspectives on protein folding, structure, and thermodynamics. Abstracts of Papers, 226th ACS National Meeting, New York, NY, United States, September 7-11, 2003:HYS.
- [2] Shakhnovich, B. E., E. Deeds, C. Delisi, and E. Shakhnovich. 2004. Protein structure and evolutionary history determine sequence space topology. Los Alamos National Laboratory, Preprint Archive, Quantitative Biology, pp 78-97.
- [3] Williams, P. D., D. D. Pollock, and R. A. Goldstein. 2001. Evolution of functionality in lattice proteins. *Journal of Molecular Graphics & Modelling* 19: pp 150