Monoclonal evolution of protein-protein interactions in two cubic lattice proteins

Rostam M. Razban
Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138

(Dated: 13 May 2016)

A kinetic Monte Carlo method is developed to study the evolution of protein-protein interactions in two cubic lattice proteins expressed in organisms of a monoclonal population. A positive correlation between increase in probability of two native proteins to be in the interaction mode due to a surface amino acid mutating and the polarity of that surface amino acid are seen for 7 out of the 9 surface resides for each of the two proteins. The presence of a more polar surface throughout the course of evolution is consistent with experimental observations on protein-protein interactions, validating the model and encouraging future applications.

I. INTRODUCTION

Protein-protein interactions (PPIs) are responsible for various functions in the cell, from enzyme catalysis to signal transduction. The wide range of functions is explained by the fact that protein-protein interactions vary greatly among themselves[1]. However, numerous general results have been experimentally observed for monomers participating in PPIs. One particular result is that interaction surfaces are composed of polar and charged residues and their presence is positively correlated with association rate and binding[2]. This is not a trivial finding, as one might reasonably believe that two protein surfaces in contact could be described as the core of a monomeric protein, in which hydrophobic residues dominate.

Thus, any model accurately simulating PPIs should be in agreement this with fundamental result. In this study, the model developed employs two stable lattice protein monomers. By selecting for an increase in binding through a fitness function that also includes the stability of the monomer, it is investigated whether this kinetic Monte Carlo based method for lattice proteins can evolve an interaction surface at an interaction mode by having charged and polar surface residues.

II. THEORY

A. Lattice protein folding and interactions

The energy in lattice proteins is expressed as

$$E = \sum_{ij} \Delta_{ij} V(i,j) \tag{1}$$

Δ is the contact matrix, whose elements are one if residues i and j are nearest neighbors, or otherwise 0. V describes the interaction potentials between amino acid residues and are from Miyazawa and Jernigan (MJ 96)[3]. E_i^{mon}, energy of the monomer folding in a specific configuration i, has 28 nearest neighbor pairs in a cubic lattice protein. E_j^{dim}, energy of two monomers in specific configurations interacting at a specific interaction mode j, has 9 nearest neighbor pairs from surface residues interacting between two cubic lattice proteins.

The $3X3X3$ lattice protein contains 103,346 well-defined compact structures[4]. For computational efficiency, a representative subset of 10,000 structures is considered. Because the space is fully characterized, the partition function can be exactly calculated and hence the

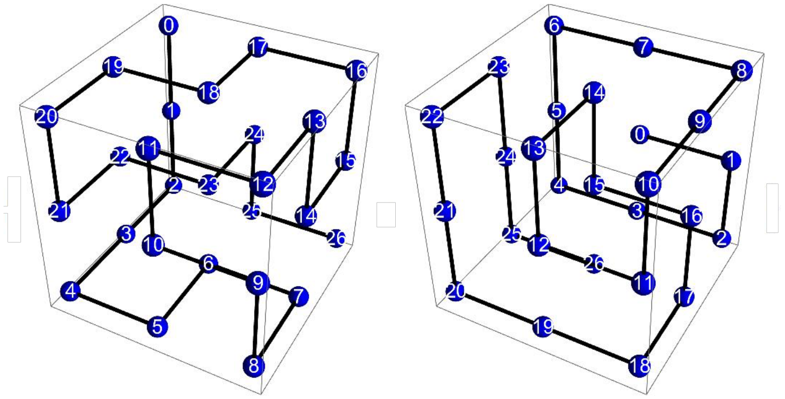

FIG. 1. Predefined native structures and specific interaction mode of the two proteins studied. The compact Hamilton path in each individual protein represents its native state fold. The orientation of the overlaying faces (protein 1: central residue of face is labeled 14; protein 2: 24) represents the interaction mode.

probability to be in the native fold is

$$P^{nat} = \frac{e^{-\beta E^{nat}}}{\sum_{i}^{10,000} e^{-\beta E_i^{mon}}} \qquad (2)$$

Analogously, a partition function for all possible interactions between two lattice proteins in the native fold can be tabulated by considering only rigid binding. There are 6*6*4 = 144 possible interaction modes. The model disallows mis-folded protein to interact; only native fold proteins can interact. The probability to be in the specific interaction mode only applies to proteins both in their native fold, and reads

$$P^{int} = \frac{e^{-\beta E^{int}}}{\sum_{j}^{144} e^{-\beta E_j^{dim}}} \qquad (3)$$

B. Kinetic Monte Carlo method

A fitness function for organisms in the population is defined as follows

$$f = P_1^{nat} P_2^{nat} P^{int} \qquad (4)$$

At each Monte Carlo step one random mutation occurs randomly along the genome. By assuming a Fisher-Wright population that is monoclonal and remains monoclonal, the probability of fixation of the mutation (Π) can be derived at steady state by the backward's Kolmogoroff Equation[5]

$$\Pi = \frac{1 - e^{-2s}}{1 - e^{-2Ns}} \qquad (5)$$

s is the selection coefficient and $s = \dfrac{f^{'} - f}{f}$ where $f^{'}$ is the fitness of the mutant. N is the population size. If $\Pi > R$, where R is a random number between 0 and 1, the mutation fixates and hence the genotype of the entire monoclonal population changes.

III. COMPUTATIONAL METHOD

The native state folds and the interaction mode are illustrated in Figure 1. For ease of reference, the protein on the left is referred to as protein 1; the right, protein 2 throughout this paper.

A stable genetic sequence of 81 nucleotides is generated for each protein from random sequences undergoing a kinetic Monte Carlo procedure which accepts potential random mutations if $\Delta P_{nat} > 0$. These two stable sequences are then fused to form a gene of 162 nucleotides that proceeds through the kinetic Monte Carlo procedure outlined in Section II B. Each simulation is ran for 10^5 Monte Carlo steps. Multiple simulations that are run with the same parameters are seeded with different random numbers. Parameters values in the model are listed in Table I.

TABLE I. Parameters in the model and their corresponding values. RT stands for room temperature.

Parameter	Abbr.	Value
Interaction potential	$V(i,j)$	in k_bT_{RT}[3]
Thermal energy	β	$1.0/k_bT_{RT}$
Population size	N	50

IV. RESULTS

Figure 2 demonstrates the dynamics of the thermodynamic properties in one simulation, where discontinuous steps represent a change in the genome of the monoclonal population. To answer the fundamental question: whether two lattice proteins develop charged/polar surfaces, Pearson correlations and corresponding p-values between ΔP^{int} due to a mutation at a certain position and the identity of that mutation which causes P^{int} to change, are calculated. For statistical significance, data points from a thousand simulations are compiled together. Results are summarized in Table II, where positive correlations signify that increases in ΔP^{int} correlate with increases in hydrophilicity of the mutated residue; negative correlation, increases in ΔP^{int} correlate with increases in hydrophobicity. Note that only two residues from protein 1 and two residues from protein 2 have negative correlations, although some positive correlations have relatively weak p-values (protein 1: positions 8 and 7; protein 2: positions 23, 4, and 20). Also note that those 2 residues in each protein are not nearest neighbors.

To aid in understanding how mutations lead to more polar residues, the transition matrix for surface residues are displayed in Figure 3, normalized by the total mutation events in the respective protein. Note that although some residues are increasing in hydrophilicity, the final residue still lies on the more hydrophobic side of the spectrum, i.e. protein 2: positions 24 and 21. Oppositely, residues with negative correlations, with the exception of protein 1; residue 25, have final amino acid identities that lie on the more hydrophilic side. Another interesting observation is the presence of vertical lines in most of the transition matrices. There is a certain end-point for residues in the course of evolution, however the path leading up to the end-point is less constrained. The model can be said to be 'funneling' sequence space as time progresses.

TABLE II. Correlation coefficients (r) and corresponding two-tailed p-values for surface residues labeled by positions (pos) across 1,000 simulations. Residues among the two columns that are adjacent are nearest neighbors in the specific interaction mode.

Protein 1			Protein 2		
pos	r	p-value	pos	r	p-value
12	0.21	4.58E-12	22	-0.10	2.58E-02
13	-0.15	1.91E-08	23	0.10	1.26E-03
16	0.46	2.45E-31	6	0.34	3.18E-31
9	0.34	1.17E-33	21	0.38	9.20E-44
14	-0.17	3.49E-19	24	0.47	5.70E-34
15	0.15	1.30E-09	5	0.22	4.20E-16
8	0.13	6.85E-03	20	0.16	1.60E-06
7	0.22	4.00E-06	25	-0.19	9.20E-10
26	0.29	1.37E-14	4	0.09	2.92E-03

V. DISCUSSION

The model and procedure carried out here relates to proteins that are classified as permanent, non-obligate, hetero-dimers because the

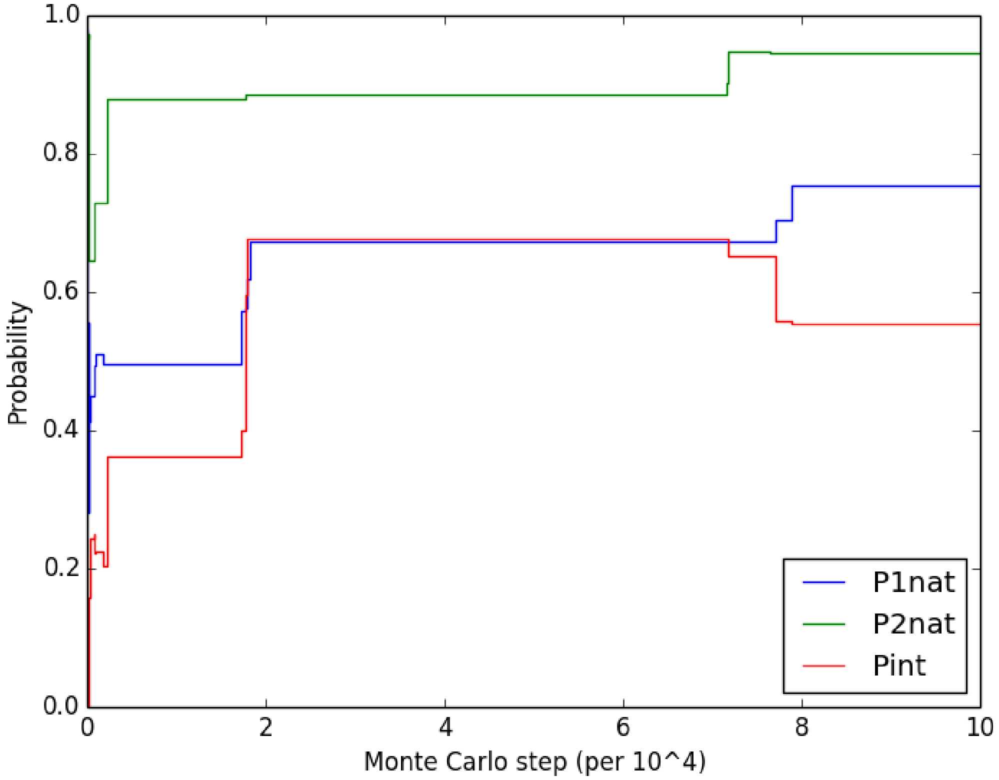

FIG. 2. Evolution of biophysical parameters with respect to the number of Monte Carlo steps taken. Note that the dynamics evolve by step functions.

fitness function rewards proteins which form stronger binding mode interactions, the initial sequence leads to stable monomers, and two different folds are chosen as native for the proteins. In principle, other PPI types between proteins can be studied by slightly tweaking the model. Still, notable classes of proteins belong to this category, including antibody-antigen interactions^1.

A positive correlation - between ΔP^{int} and the surface residue mutation that causes P^{int} to change exists - among 7 out of the 9 surface residues in each protein validates the model with general experimental results on the role of electrostatics in PPIs. This hints that population diversity and concentration effects are not the main factors in surface residues being polar since the model assumes monoclonal populations and fixed concentrations. Also, protein-protein interactions from mis-folded protein seem to be negligible since that also is not accounted for in the model.

It would be interesting to further apply the model to gauge other properties such as as P_{int} and P_{nat} trade-offs and mean passage time until a mutation occurs in the population, comparing results to bioinformatics. Because evolutionary data in reality is incomplete, this model can supplement experimental data to fill in the gaps between initial and final time points. Also,

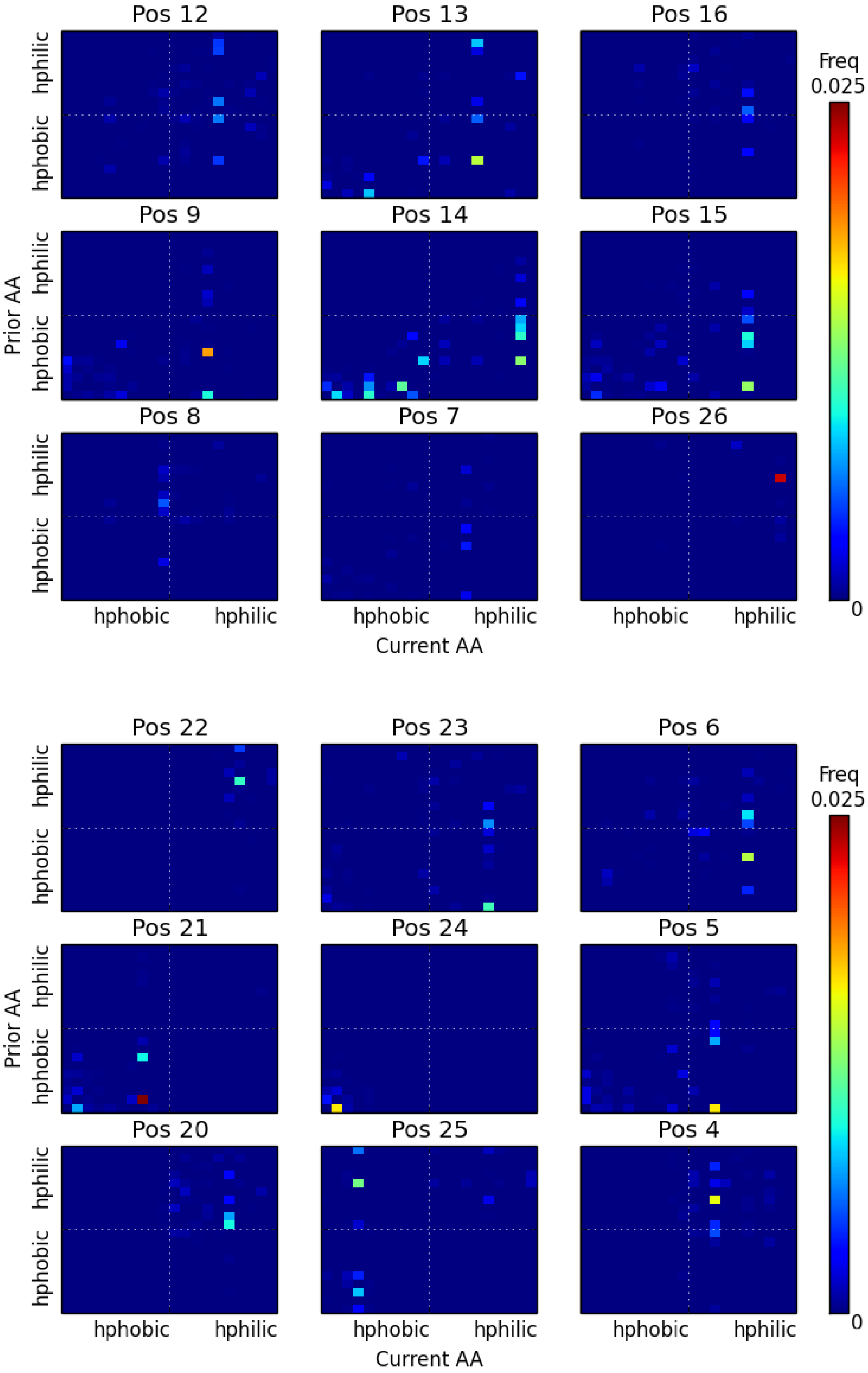

FIG. 3. Visualizing the normalized transition matrix of surface residues from protein 1 (top) and 2 (bottom) throughout 1,000 simulations. Relative hydrophobicity of amino acids are determined according to the MJ 96 potential. The order from left/bottom to right/top: L, F, I, M, V, W, C, Y, H, A, T, G, P, R, E, S, N, Q, D, K.

how robust evolutionary trajectories are to parameter values and structural conformations of the native fold and binding mode would be an interesting future study.

VI. ACKNOWLEDGMENTS

The computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. Many thanks to Eugene and the entire Shakhnovich group.

[1] I. M. Nooren, EMBO J. **22**, 3486 (2003).
[2] F. B. Sheinerman, R. Norel, and B. Honig, Curr. Opin. Struct. Biol. **10**, 153 (2000).
[3] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).
[4] E. Shakhnovich and A. Gutin, J. Chem. Phys. **93**, 5967 (1990).
[5] A. Kolmogoroff, Math. Ann. **104**, 415 (1931).