Journal of Theoretical Biology 443 (2018) 18–27

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

Cooperative "folding transition" in the sequence space facilitates function-driven evolution of protein families

Akira R. Kinjo

Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565–0871, Japan

ARTICLE INFO

Article history:
Received 27 August 2017
Revised 16 January 2018
Accepted 17 January 2018

Keywords:
Protein folding
Molecular evolution
Protein design
Sequence analysis
Monte Carlo simulation

ABSTRACT

In the protein sequence space, natural proteins form clusters of families which are characterized by their unique native folds whereas the great majority of random polypeptides are neither clustered nor foldable to unique structures. Since a given polypeptide can be either foldable or unfoldable, a kind of "folding transition" is expected at the boundary of a protein family in the sequence space. By Monte Carlo simulations of a statistical mechanical model of protein sequence alignment that coherently incorporates both short-range and long-range interactions as well as variable-length insertions to reproduce the statistics of the multiple sequence alignment of a given protein family, we demonstrate the existence of such transition between natural-like sequences and random sequences in the sequence subspaces for 15 domain families of various folds. The transition was found to be highly cooperative and two-state-like. Furthermore, enforcing or suppressing consensus residues on a few of the well-conserved sites enhanced or diminished, respectively, the natural-like pattern formation over the entire sequence. In most families, the key sites included ligand binding sites. These results suggest some selective pressure on the key residues, such as ligand binding activity, may cooperatively facilitate the emergence of a protein family during evolution. From a more practical aspect, the present results highlight an essential role of long-range effects in precisely defining protein families, which are absent in conventional sequence models.

1. Introduction

Natural proteins can be classified into families based on their sequence similarity (Finn et al., 2014). This is considered to be primarily a consequence of molecular evolution: proteins evolved from a common ancestral protein share similar sequences. However, evolution alone does not account for the existence of relatively well-defined (domain) families that are distributed rather discretely than continuously in the sequence space (Goldstein, 2008; Maynard Smith, 1970; Nishikawa, 1993, 2002). A key to understanding the family distribution is protein folding. As observed in protein structure classification databases (Cheng et al., 2014; Murzin et al., 1995; Orengo et al., 1997), each protein family corresponds to a unique three-dimensional fold, suggesting the existence of physical constraints imposed on protein sequences during the evolutionary process to maintain the fold (Morcos et al., 2014). While protein structures can tolerate great many mutations to the extent that proteins with little sequence similarity can share the same fold, residue conservation patterns reflect the structural context of protein sequences. This fact has long been exploited in protein structure prediction in the form of position-specific scoring matrices (Altschul et al., 1997; Gribskov and Eisenberg, 1987; Kinjo and Nakamura, 2008; Taylor, 1986) and, more recently, direct-coupling analysis and related methods (Balakrishnan et al., 2011; Ekeberg et al., 2013; Jones et al., 2012; Kinjo, 2015; Levy et al., 2017; Miyazawa, 2013; Morcos et al., 2011; Taylor et al., 2012).

Under a given physiological condition, a polypeptide is either able or unable to fold into some unique structure. This suggests the existence of a "folding transition" at the border between an "island" of a protein family and the "sea" of random polypeptide, that is analogous to the folding transition of a protein molecule in the conformational space (Nishikawa, 1993, 2002; Shakhnovich and Gutin, 1993b). It should be noted, however, that there are many families in the sequence space so that a sequence moving in the sequence space may fall into any one of these families. This is in contrast to protein folding in the conformational space where there is usually only one unique native structure for a given protein sequence. Furthermore, a (structural) domain, rather than a whole protein sequence, should be considered as a unit of folding as a particular domain may be found in different proteins in combination with other, different, domains. Therefore, the system in which the analogy of protein folding holds should be limited to the vicinity of each protein domain family rather than the entire sequence space. In the following, we focus on the folding transi-

E-mail address: akinjo@protein.osaka-u.ac.jp

https://doi.org/10.1016/j.jtbi.2018.01.019
0022-5193/

tion in a sequence subspace around a given protein domain family. Although biologically important, intrinsically disordered proteins (Dunker et al., 2001; Minezaki et al., 2006; Tompa, 2012) are excluded from the present study for the following two reasons. First, the analogy of the folding transition may not apply to those proteins. Second, it is difficult to obtain reliable and comprehensive multiple sequence alignments for this class of proteins (Lange et al., 2015), which are required for parameter estimation of the statistical model employed in the present study.

There have been a number of theoretical and computational studies on subjects related to the sequence space such as foldability and design (Govindarajan and Goldstein, 1995; Morcos et al., 2014; Shakhnovich and Gutin, 1993a, 1993b), molecular evolution within an island (Bastolla et al., 1999; Bornberg-Bauer and Chan, 1999; Wroe et al., 2005) and between islands (Holzgräfe and Wallin, 2014; Sikosek et al., 2016; Wroe et al., 2007), or the size and/or distribution of islands in the sequence space (Bornberg-Bauer, 1997; Govindarajan and Goldstein, 1996; Koehl and Levitt, 2002; Kuhlman and Baker, 2000; Li et al., 1996). On the contrary, relatively little attention has been paid to the transition between an island (a set of sequences belonging to the same family) and the sea (the set of sequences that do not belong to the family) apart from a few exceptions. In the context of protein design, Shakhnovich and Gutin (1993b) theoretically predicted the existence of a "folding transition". In a study of hierarchical evolution of protein fold families based on a simple model of the evolutionary selection by native stability using a generic contact potential (Miyazawa and Jernigan, 1985), Dokholyan and Shakhnovich (2001) observed a sharp transition at a certain design temperature. In neither of these studies, however, the nature of the transition was investigated further. Characterizing the folding transition in the sequence space may help understand essential features that constitute a protein family and possible evolutionary trajectories that may have led to the emergence of a protein family. It also has practical importance in identifying new family members and designing new proteins.

In the following, we investigate the folding transition in the sequence subspaces for 15 protein domain families including all-α, all-β, α/β and other folds by performing extensive Monte Carlo (MC) simulations of the modified lattice gas model (LGM) of protein sequence alignment (Kinjo, 2016; 2017) that coherently integrates long-range interactions and variable-length insertions. Using the LGM, the existence of a sharp two-state transition between natural-like sequences and random sequences in the sequence subspace is demonstrated. Furthermore, the nature of the transition is examined in detail by analyzing residue distribution of each site along the transition as well as by performing virtual "mutation" experiments.

2. Theory

We briefly summarize the theory to the extent that is necessary for understanding the present study. For more details about the formulation of the LGM as well as the algorithms for MC simulations and parameter optimization, refer to the previous papers (Kinjo, 2016; 2017). The LGM for a given Pfam (Finn et al., 2014) family consists of N "core" sites and $N-1$ "insert" sites which respectively correspond to the "match" states and "insert" states of the Pfam profile hidden Markov model (HMM) of length N, excluding the N- and C-terminal insert states. Exactly one of the 21 residue types (including the "delete" symbol) can exist on each core site whereas arbitrarily many (including zero) residues out of the standard 20 residue types can reside at each insert site. The array of core and insert sites are connected via "bonds" (solid arrows in Fig. 1) that reflect the linear polypeptide structure. A pair of model sites connected via a bond are called a "bonded pair" in the following. Between core sites more than 2 residues apart along the sequence, there may be interactions (dashed lines in Fig. 1) based on a representative native structure of the family. Two sites are defined to be interacting if the residues aligned to those sites are in contact in the corresponding representative native structure. Two residues are defined to be in contact if any non-hydrogen atoms in those residues are within 5Å. Interactions are defined only between core sites for simplicity. Interacting core sites are referred to as "non-bonded" pairs in the following.

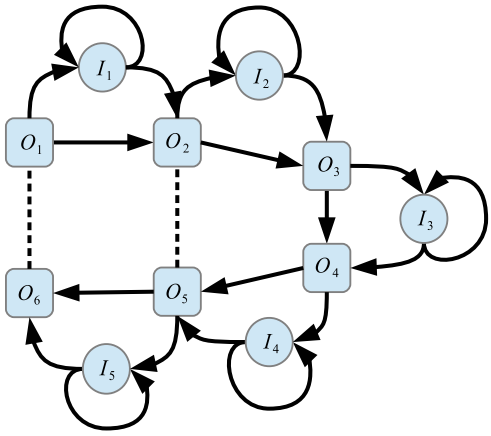

Fig. 1. An example model structure with model length $N = 6$. There are N core sites $O_1, \cdots, O_N$ and $N-1$ insert sites $I_1, \cdots, I_{N-1}$. These model sites are bonded via "bonds" (solid arrows). Some pairs of non-bonded core sites may be "interacting" (dashed lines).

Let $\mathbf{a} = a_1 \cdots a_L$ be an amino acid sequence of L residues and an LGM $\mathcal{M}$ of length N consist of core sites $O_1, \ldots, O_N$ and insert sites $I_1, \ldots, I_{N-1}$. An alignment between the sequence $\mathbf{a}$ and the model $\mathcal{M}$ is represented as a sequence of pairs of a model site (core or insert) and a residue: $\mathbf{X} = X_1 \ldots X_{L_\mathbf{X}}$ where $L_\mathbf{X}$ is the length of the alignment and each X_i is a pair such as $(S,\ a)$ with $S \in \{O_i\}_{i=1,\ldots,N} \cup \{I_i\}_{i=1,\ldots,N-1}$ and $a \in \{a_1, \ldots, a_L\}$. For example, given an amino acid sequence, say KCFPDGVW, and a model of length $N = 6$ (Fig. 1), one of many possible alignments is represented as $\mathbf{X} = X_1 \ldots X_9 = (O_1, \mathtt{K})(O_2, \mathtt{C})(O_3, \mathtt{F})(I_3, \mathtt{P})(I_3, \mathtt{D})(I_3, \mathtt{G})(O_4, -)(O_5, \mathtt{V})(O_6, \mathtt{W})$. Note there are multiple occurrence of the insert site I_3 whereas other insert sites are completely absent in this particular alignment. Since there may be any number of residues at each insert site, the alignment length is variable.

Based on this representation of sequence alignment, the energy function of alignment $\mathbf{X}$ is defined as

$$E(\mathbf{X}) = -\sum_{k=1}^{L_\mathbf{X}-1} J(X_k, X_{k+1}) - \sum_{(k,l)\in\mathcal{T}} K(X_k, X_l) - \sum_{k=1}^{L_\mathbf{X}} \mu(X_k) \tag{1}$$

where J and K are short-range and long-range interaction parameters, respectively, μ's are chemical potentials, and $\mathcal{T}$ indicates the set of all the interacting non-bonded pairs. The short-range interactions act only between bonded pairs of residues that are consecutive in the alignment (i.e., between X_k and X_{k+1} in Eq. (1)). The long-range interactions act between residues that are aligned to interacting non-bonded pairs. The chemical potentials are so called because they are used to control the residue densities of each site. Only J and K parameters constitute intrinsic energy, and they are to be determined from a given (observed) multiple sequence alignment (MSA) of the family sequences (see below).