Journal of Theoretical Biology 433 (2017) 21–38

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Selection originating from protein stability/foldability: Relationships between protein folding free energy, sequence ensemble, and fitness

Sanzo Miyazawa

6-5-607 Miyanodai, Sakura, Chiba 285–0857, Japan

ARTICLE INFO

Article history:
Received 27 March 2017
Revised 27 July 2017
Accepted 21 August 2017
Available online 24 August 2017

Keywords:
Folding free energy change
Inverse statistical potential
Boltzmann distribution
Selective temperature
Positive selection

ABSTRACT

Assuming that mutation and fixation processes are reversible Markov processes, we prove that the equilibrium ensemble of sequences obeys a Boltzmann distribution with $\exp(4N_e m(1-1/(2N)))$, where m is Malthusian fitness and N_e and N are effective and actual population sizes. On the other hand, the probability distribution of sequences with maximum entropy that satisfies a given amino acid composition at each site and a given pairwise amino acid frequency at each site pair is a Boltzmann distribution with $\exp(-\psi_N)$, where the evolutionary statistical energy ψ_N is represented as the sum of one body (h) (compositional) and pairwise (J) (covariational) interactions over all sites and site pairs. A protein folding theory based on the random energy model (REM) indicates that the equilibrium ensemble of natural protein sequences is well represented by a canonical ensemble characterized by $\exp(-\Delta G_{ND}/k_B T_s)$ or by $\exp(-G_N/k_B T_s)$ if an amino acid composition is kept constant, where $\Delta G_{ND} \equiv G_N - G_D$, G_N and G_D are the native and denatured free energies, and T_s is the effective temperature representing the strength of selection pressure. Thus, $4N_e m(1-1/(2N))$, $-\Delta\psi_{ND} (\equiv -\psi_N + \psi_D)$, and $-\Delta G_{ND}/k_B T_s$ must be equivalent to each other. With h and J estimated by the DCA program, the changes ($\Delta\psi_N$) of ψ_N due to single nucleotide nonsynonymous substitutions are analyzed. The results indicate that the standard deviation of $\Delta G_N (= k_B T_s \Delta\psi_N)$ is approximately constant irrespective of protein families, and therefore can be used to estimate the relative value of T_s. Glass transition temperature T_g and ΔG_{ND} are estimated from estimated T_s and experimental melting temperature (T_m) for 14 protein domains. The estimates of ΔG_{ND} agree with their experimental values for 5 proteins, and those of T_s and T_g are all within a reasonable range. In addition, approximating the probability density function (PDF) of $\Delta\psi_N$ by a log-normal distribution, PDFs of $\Delta\psi_N$ and K_a/K_s, which is the ratio of nonsynonymous to synonymous substitution rate per site, in all and in fixed mutants are estimated. The equilibrium values of ψ_N, at which the average of $\Delta\psi$ in fixed mutants is equal to zero, well match ψ_N averaged over homologous sequences, confirming that the present methods for a fixation process of mutations and for the equilibrium ensemble of ψ_N give a consistent result with each other. The PDFs of K_a/K_s at equilibrium confirm that T_s negatively correlates with the amino acid substitution rate (the mean of K_a/K_s) of protein. Interestingly, stabilizing mutations are significantly fixed by positive selection, and balance with destabilizing mutations fixed by random drift, although most of them are removed from population. Supporting the nearly neutral theory, neutral selection is not significant even in fixed mutants.

1. Introduction

Natural proteins can fold their sequences into unique structures. Protein's stability and foldability result from natural selection and are not typical characteristics of random polymers (Bryngelson and Wolynes, 1987; Pande et al., 1997; Ramanathan and Shakhnovich, 1994; Shakhnovich and Gutin, 1993a; 1993b). Natural selection maintains protein's stability and foldability over evolutionary timescales. On the basis of the random energy model (REM) for protein folding, it was discussed (Ramanathan and Shakhnovich, 1994; Shakhnovich and Gutin, 1993a; 1993b) that the equilibrium ensemble of natural protein sequences in sequence space is well represented by a canonical ensemble characterized by a Boltzmann factor $\exp(-\Delta G_{ND}(\boldsymbol{\sigma})/k_B T_s)$, where $\Delta G_{ND}(\boldsymbol{\sigma}) (\equiv G_N(\boldsymbol{\sigma}) - G_D(\boldsymbol{\sigma}))$ is the folding free energy of sequence $\boldsymbol{\sigma}$, G_N and G_D are the free energies of the native and denatured states, k_B is the Boltzmann constant, and T_s is the effective temperature representing the strength of selection pressure and must satisfy $T_s < T_g < T_m$ for natural proteins to fold into unique native structures; T_g is glass transition temperature and T_m is melting tem-

E-mail address: sanzo.miyazawa@gmail.com

http://dx.doi.org/10.1016/j.jtbi.2017.08.018
0022-5193/© 2017 Elsevier Ltd. All rights reserved.

perature. The REM also indicates that the free energy of denatured conformations (G_D) is a function of amino acid frequencies only and does not depend on amino acid order, and therefore the Boltzmann factor will be taken as $\exp(-G_N(\boldsymbol{\sigma})/k_B T_s)$, if amino acid frequencies are kept constant. It was shown by lattice Monte Carlo simulations (Shakhnovich, 1994) that lattice protein sequences selected with this Boltzmann factor were not trapped by competing structures but could fold into unique native structures. Selective temperatures were also estimated (Dokholyan and Shakhnovich, 2001) for actual proteins to yield good correlations of sequence entropy between actual protein families and sequences designed with this type of Boltzmann factor.

On the other hand, the maximum entropy principle insists that the probability distribution of sequences in sequence space, which satisfies constraints on amino acid compositions at all sites and on amino acid pairwise frequencies for all site pairs, is a Boltzmann distribution with the Boltzmann factor $\exp(-\psi_N(\boldsymbol{\sigma}))$, where the total interaction $\psi_N(\boldsymbol{\sigma})$ of a sequence $\boldsymbol{\sigma}$ is represented as the sum of one-body (h) (compositional) and pairwise (J) (covariational) interactions between residues in the sequence; $\psi_N(\boldsymbol{\sigma})$ is called the evolutionary statistical energy by Hopf et al. (2017). The inverse statistical potentials, the one-body (h) and pairwise (J) interactions, that satisfy those constraints for homologous sequences have been estimated (Ekeberg et al., 2014; 2013; Marks et al., 2011; Morcos et al., 2011) as one of inverse Potts problems and successfully employed to predict contacting residue pairs in protein structures (Ekeberg et al., 2014; 2013; Hopf et al., 2012; Marks et al., 2011; Miyazawa, 2013; Morcos et al., 2011; Sułkowska et al., 2012). Morcos et al. (2014) noticed that the ψ_N in the Boltzmann factor is the dimensionless energy corresponding to $G_N/k_B T_s$, and estimated selective temperatures (T_s) for several protein families by comparing the difference ($\Delta\psi_{ND}$) of ψ between the native and the molten globule states with folding free energies (ΔG_{ND}) estimated with associative-memory, water-mediated, structure, and energy model (AWSEM) (Davtyan et al., 2012).

A purpose of the present study is to establish relationships between protein foldability/stability, sequence distribution, and protein fitness. First, we prove that if mutation and fixation processes in protein evolution are reversible Markov processes, the equilibrium ensemble of genes will obey a Boltzmann distribution with the Boltzmann factor $\exp(4N_e m(1-1/(2N)))$, where N_e and N are effective and actual population sizes, and m is the Malthusian fitness of a gene. In other words, correspondences between $-\Delta G_{ND}/k_B T_s$, $-\Delta\psi_{ND}(\equiv \psi_N - \psi_D)$ and $4N_e m(1-1/(2N))$ are obtained by equating these three Boltzmann distributions with each other; $\psi_D \simeq G_D/k_B T_s + \text{constant}$.

The second purpose is to analyze the effects ($\Delta\psi_N$) of single amino acid substitutions on the evolutionary statistical energy of a protein, and to estimate from the distribution of $\Delta\psi_N$ the effective temperature of natural selection (T_s) and then glass transition temperature (T_g) and folding free energy (ΔG_{ND}) of protein. We estimate the one-body (h) and pairwise (J) interactions with the DCA program, which is available at “http://dca.rice.edu/portal/dca/home”, and then analyze the changes ($\Delta\psi_N$) of the evolutionary statistical energy (ψ_N) of a natural sequence due to single amino acid substitutions caused by single nucleotide changes. The data of $\Delta\psi_N$ due to single nucleotide nonsynonymous substitutions for 14 protein domains show that the standard deviation of $\Delta\psi_N$ over all the substitutions at all sites hardly depends on the evolutionary statistical energy (ψ_N) of each homologous sequence and is nearly constant for each protein family, indicating that the standard deviation of $\Delta G_N \simeq k_B T_s \Delta\psi_N$ is nearly constant irrespective of protein families. From this finding, T_s for each protein family has been estimated in relative to T_s for the PDZ family, which is determined by directly comparing $\Delta\Delta\psi_{ND}(\equiv \Delta(\psi_N - \psi_D) \simeq \Delta\psi_N)$ with the experimental values of folding free energy changes, $\Delta\Delta G_{ND}$, due to single amino acid substitutions. Also T_g and ΔG_{ND} for each protein family are estimated on the basis of the REM from the estimated T_s and an experimental melting temperature T_m. The estimates of T_s and T_g are all within a reasonable range, and those of ΔG_{ND} are well compared with experimental ΔG_{ND} for 5 protein families. The present method for estimating T_s is simpler than the method (Morcos et al., 2014) using AWSEM, and also is useful for the prediction of ΔG_{ND}, because the experimental data of ΔG_{ND} are limited in comparison with T_m, and also experimental conditions such as temperature and pH tend to be different among them. In addition, it has been revealed that $\Delta\psi_N$ averaged over all single nucleotide nonsynonymous substitutions is a linear function of ψ_N/L of each homologous sequence, where L is sequence length; the average of $\Delta\psi_N$ decreases as ψ_N/L increases. This characteristic is required for homologous proteins to stay at the equilibrium state of the native conformational energy $G_N \simeq k_B T_s \psi_N$, and indicates a weak dependency (Miyazawa, 2016; Serohijos et al., 2012) of $\Delta\Delta G_{ND}$ on $\Delta G_{ND}/L$ of protein across protein families.

The third purpose is to study an amino acid substitution process in protein evolution, which is characterized by the fitness, $m = -\Delta\psi_{ND}/(4N_e(1-1/(2N)))$. We employ a monoclonal approximation for mutation and fixation processes of genes, in which protein evolution proceeds with single amino acid substitutions fixed at a time in a population. In this approximation, ψ_N of a protein gene attains the equilibrium, $\psi_N = \psi_N^{eq}$, when the average of $\Delta\psi_N(\simeq \Delta\Delta\psi_{ND})$ over singe nucleotide nonsynonymous mutations fixed in a population is equal to zero. Approximating the distribution of $\Delta\psi_N$ due to singe nucleotide nonsynonymous mutations by a log-normal distribution, their distribution for fixed mutants is numerically calculated and used to calculate the averages of various quantities and also the probability density functions (PDF) of K_a/K_s in all arising mutants and also in fixed mutants only; K_a/K_s is defined as the ratio of nonsynonymous to synonymous substitution rate per site. There is a good agreement between the time average (ψ_N^{eq}) and ensemble average ($\langle\psi_N\rangle_{\boldsymbol{\sigma}}$), which is equal to the sample average, $\overline{\psi_N}$, of ψ_N over homologous sequences, supporting the constancy of the standard deviation of $\Delta\psi_N$ assumed in the monoclonal approximation.

We also study protein evolution at equilibrium, $\psi_N = \psi_N^{eq}$. The common understanding of protein evolution has been that amino acid substitutions observed in homologous proteins are neutral (Kimura, 1968; 1969; Kimura and Ohta, 1971; 1974) or slightly deleterious (Ohta, 1973; 1992), and random drift is a primary force to fix amino acid substitutions in population. The PDFs of K_a/K_s in all arising mutations and in their fixed mutations are examined to see how significant each of positive, neutral, slightly negative,and negative selections is. Interestingly, stabilizing mutations are significantly fixed in population by positive selection, and balance with destabilizing mutations that are also significantly fixed by random drift, although most negative mutations are removed from population. Contrary to the neutral theory (Kimura, 1968; 1969; Kimura and Ohta, 1971; 1974) and supporting the nearly neutral theory (Ohta, 1973; 1992; 2002), the proportion of neutral selection is not large even in fixed mutants. It is also confirmed that the effective temperature (T_s) of selection negatively correlates with the amino acid substitution rate (K_a/K_s) of protein at equilibrium.

2. Methods

2.1. Knowledge of protein folding

A protein folding theory (Pande et al., 1997; Ramanathan and Shakhnovich, 1994; Shakhnovich and Gutin, 1993a; 1993b), which is based on a random energy model (REM), indicates that the equilibrium ensemble of amino acid sequences, $\boldsymbol{\sigma} \equiv (\sigma_1, \ldots, \sigma_L)$ where σ_i is the type of amino acid at site i and L is sequence length,