Vol. 13, pp. 45–62 (2016)
doi: 10.2142/biophysico.13.0_45

Biophysics and Physicobiology

https://www.jstage.jst.go.jp/browse/biophysico/

Regular Article

A unified statistical model of protein multiple sequence alignment integrating direct coupling and insertions

Akira R. Kinjo[1]

[1]*Institute for Protein Research, Osaka University, Suita, Osaka 565-0871, Japan*

Received October 21, 2015; accepted March 18, 2016

The multiple sequence alignment (MSA) of a protein family provides a wealth of information in terms of the conservation pattern of amino acid residues not only at each alignment site but also between distant sites. In order to statistically model the MSA incorporating both short-range and long-range correlations as well as insertions, I have derived a lattice gas model of the MSA based on the principle of maximum entropy. The partition function, obtained by the transfer matrix method with a mean-field approximation, accounts for all possible alignments with all possible sequences. The model parameters for short-range and long-range interactions were determined by a self-consistent condition and by a Gaussian approximation, respectively. Using this model with and without long-range interactions, I analyzed the globin and V-set domains by increasing the "temperature" and by "mutating" a site. The correlations between residue conservation and various measures of the system's stability indicate that the long-range interactions make the conservation pattern more specific to the structure, and increasingly stabilize better conserved residues.

Key words: long-range interactions, short-range interactions, molecular evolution, protein structure, sequence conservation

A multiple sequence alignment (MSA) of a family of proteins provides us with valuable information to characterize the protein family in terms of patterns of amino acid residues at alignment sites [1]. The usefulness of analyzing the residue compositions in the MSA has led to the development of a class of sequence profile methods [1–3] such as PSI-BLAST [4] and profile hidden Markov models (HMM) [5], which can be used to detect distantly related proteins, to obtain high-quality alignments, and to improve structure prediction [6] as well as to characterize functional and structural roles of the conservation pattern [7]. In the sequence profile methods, it is assumed that the residue composition of each site is independent of other sites. With this crude assumption, the conservation of residues are explained in terms of their functional and structural roles. However, to further understand the mechanism of these roles in the context of protein sequences, one needs to drop the assumption of site independence. In fact, there seems to be no way for a residue to "know" that it is in a particular position in the sequence to play a particular functional or structural role other than by its interactions with other residues in the sequence (or with other molecules in the biological system). Therefore, to understand what makes particular residues important at each site, one needs to study the correlations between different sites.

Correlations between distant sites in a MSA can be quantified by identifying correlated substitutions. They have been exploited to gain further insights of structures and functions of proteins [8–10]. However, the apparent correlations observed in a MSA are only a result of intricate interactions between residues as in the underlying native structures of proteins. Recently, there have been a number of successful attempts to extract direct correlations [9,10] which are in fact found to be in excellent agreement with the residue-residue contacts in native structures [11–13] to the extent that the three-dimensional structures can be actually (re)constructed [14,15].

One drawback of the direct-coupling analysis (as well as other direct correlation methods) is that it takes into account only those alignment sites that are well aligned (the "core"

Corresponding auther: Akira R. Kinjo, Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan.
e-mail: akinjo@protein.osaka-u.ac.jp

sites), and ignores insertions. The primary difficulty in the treatmentof insertion is that they are of variable lengths, which makes the system size variable and hence greatly complicates the problem. When one is interested in some universal properties of a protein family such as their approximate three-dimensional fold, insertions may be irrelevant. However, when one is interested in a particular member of the family, the existence of some insertions may be important. In fact, insertions, which may be regarded as "embellishments" to a conserved structural core, are deemed to be an effective strategy for proteins to diversify and specialize their functions [16]. Some insertions are also known to play critical roles in protein oligomerization [17,18]. Of more fundamental concern is that ignoring insertions in a MSA means ignoring the polypeptide chain structure, which implies theoretical as well as practical consequences. Theoretically, it is questionable to ignore such a strong interaction as the peptide bond in order to accurately describe the sequence and structure of proteins. Practically, in order to identify new members of a family by aligning their sequences to some MSA-derived model incorporating direct correlations, a consistent treatment of polypeptide sequences is necessary.

In this paper, I present a new statistical model of the MSA that incorporates both direct correlations and insertions. The main objective of this model is to incorporate long-range correlations into multiple-sequence alignment, rather than to improve contact prediction by incorporating insertions. As will be apparent from the formulation, this model is a generalization of the direct-coupling analysis that is based on the principle of maximum entropy [11,19]. This model can be regarded as a finite, quasi-one-dimensional, multicomponent, and heterogeneous lattice gas model where the "particles" are amino acid residues. In the following, the "lattice gas model" refers to this model. The lattice system consists of two kinds of lattice sites, corresponding to the core (matching or deletion) or the insert, that are connected in a similar, but distinctively different, manner as in the profile HMM model. While long-range interactions are treated by using a mean-field approximation, short-range interactions are treated rigorously so that the partition function is obtained analytically by a transfer matrix method. One notable feature of this model is that its partition function literally accounts for all the possible alignments with all the possible protein sequences, including infinitely long ones. Based on this model, various virtual experiments can be performed by changing the "temperature" of the system or by manipulating the "chemical potentials" associated with the particles (residues) at each site. In addition, it is possible to align new sequences against a lattice gas model so that it can be used for remote homology detection (in much the same way as the profile HMMs), but with long-range correlations included (unlike the profile HMMs).

The paper is organized as follows. In Section 1, some basic quantities are defined and the lattice gas model of the multiple sequence alignment is formulated. Section 2 provides the details of numerical methods and data preparation. Section 3 gives the results of virtual experiments by increasing the temperature or by introducing alanine point mutants. In Section 4, limitations, implications as well as possible extensions of the present model are discussed.

1. Theory

1.1. Representing multiple sequence alignment as lattice gas system

A MSA may be regarded as a matrix of symbols in which each row is a protein sequence possibly with gaps and each column is an alignment site. Some columns may contain few gaps so the residues in such positions may be relatively important for the protein family. Here, I informally define a "core" (matching/deletion) site as an alignment site which are relatively well aligned. The remaining sites are defined to be insert sites. Core sites are ordered from the N-terminal to the C-terminal, and denoted as $O_1,O_2,...O_N$ with N being the number of core sites. For convenience, the terminal core sites O_0 and O_{N+1} are appended to indicate the start and end of the alignment, as in the profile HMM [1]. To each core site, either one of 20 amino acid residues or a gap (deletion) may be assigned,and the latter is treated as the 21-st type of residue. An insert site between two core sites O_i and O_{i+1} is denoted as I_i. All the gap symbols are ignored at an insert site. In the following, the (ordered) sets of core and insert sites are denoted as $\boldsymbol{O}=\{O_0...,O_{N+1}\}$ and $\boldsymbol{I}=\{I_0,...,I_N\}$, respectively, and their union as $\boldsymbol{S}=\boldsymbol{O}\cup\boldsymbol{I}$. In addition, let us define a set of amino acid residues allowed for an insert site I_i as $\boldsymbol{A}_{I_i}=\{\mathrm{A},...,\mathrm{Y}\}$ (20 amino acid residue types), and that for a core site O_i as $\boldsymbol{A}_{O_i}=\{\mathrm{A},...,\mathrm{Y},–\}$ (20 amino acid residues and deletion) for $i=1,...,N$ and $\boldsymbol{A}_{O_0}=\boldsymbol{A}_{O_{N+1}}=\{–\}$ (deletion only) for the terminal sites.

For one protein sequence in the MSA, at most one residue may correspond to each core site O_i whereas any number of residues may correspond to an insert site I_i. In this sense, residues behave like fermions on core sites and like bosons on insert sites. The set of core and insert sites comprise a quasi-one-dimensional lattice structure as shown in Figure 1. In this lattice structure, two sites are connected if two consecutive residues in a protein sequence (possibly including gap symbols) can be assigned. If two sites are directly connected, they are defined to be a bonded or short-range pair. The self-connecting loop in each insert site indicates that it makes a bonded pair with itself. Thus, an insertion may be indefinitely long, manifesting its boson-like character.

Based on this lattice system, an alignment $\mathbf{X}$ of a particular protein sequence $\boldsymbol{a}=a_1a_2...a_L$ in the MSA may be represented as a sequence of length L_X consisting of ordered pairs of a lattice site and a residue of $\boldsymbol{a}$: $\mathbf{X}=X_0X_1...X_{L_x}X_{L_x+1}$ ("matchings" to the terminal sites are also included). Here, each $X_k=(S,a)$ with $S\in\boldsymbol{S}$ and $a\in\boldsymbol{A}_S$. A whole MSA consisting of M sequences is a set of such aligned sequences:

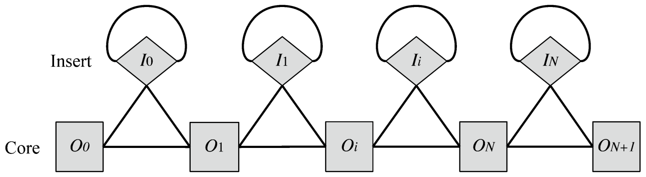

Figure 1 The lattice structure of the model. The squares marked with O_i (i=0,...,N+1) correspond to core (matching/deletion) sites, the diamonds marked with I_i (i=0,...,N) correspond to insert sites. The edges between sites indicates bonded interactions. See Figure 2 for concrete examples.

```
MVGA--HAGEY-   (S1)
-V----NVDEV-   (S2)
-VEA--DVAGH-   (S3)
-VKG------DG   (S4)
-VYS--TYETS-   (S5)
-FNA--NIPKH-   (S6)
-IAGADNGAGV-   (S7)
-------------------
IOOOIIOOOOOI
012333456788
```

Figure 2 Example of a multiple sequence alignment (based on [1]). Each row corresponds to a protein sequence (S1,...,S7) and each column to an alignment site. Below the horizontal line, each alignment site is annotated as to whether it corresponds to a core (matching or deletion) site ("O") or an insert site ("I"). Indicated below these "O"/ "I" symbols are the position of lattice sites. (c.f. Fig. 1) The size of the lattice model based on this MSA is N=8. Insert sites other than I_0, I_3 and I_8 are not explicit in this MSA. For example, the alignment of the sequence S2 in this figure is represented as $\mathbf{X}^{S2}=X_0...X_9=(O_0,-)(O_1,\mathrm{V})(O_2,-)(O_3,-)(O_4,\mathrm{N})(O_5,\mathrm{V})(O_6,\mathrm{D})(O_7,\mathrm{E})(O_8,\mathrm{V})(O_9,-)$ where the first and last pairs represent the start and end of the alignment, respectively. As another example, the alignment of sequence S7 is $\mathbf{X}^{S7}=X_0 \ldots X_{11}=(O_0,-)(O_1,\mathrm{I})(O_2,\mathrm{A})(O_3,\mathrm{G})(I_3,\mathrm{A})(I_3,\mathrm{D})(O_4,\mathrm{N})(O_5,\mathrm{G})\ (O_6,\mathrm{A})(O_7,\mathrm{G})(O_8,\mathrm{V})(O_9,-)$.

$\{\mathbf{X}^t\}_{t=1,...,M}$. Figure 2 shows some concrete examples of this representation of alignment.

1.2. Variables to characterize alignments

Using the above representation, let us define some quantities that characterize an alignment in a given MSA. For a given lattice model and its alignment $\boldsymbol{X}$ with a protein sequence, the number of the residue type $a\in\boldsymbol{A}_S$ at the lattice site $S\in\boldsymbol{S}$ is defined as

$$n_S(a|\mathbf{X}) = \sum_{k=0}^{L_\mathbf{X}} \delta_{(S,a),X_k}. \tag{1}$$

This quantity is referred to as the single-site count. Similarly, the number of a pair of residue types $a\in\boldsymbol{A}_S$ and $b\in\boldsymbol{A}_{S'}$ on a bonded pair of lattice sites S and S' occupied by two consecutive alignment sites is defined as

$$n^b_{SS'}(a,b|\mathbf{X}) = \sum_{k=0}^{L_\mathbf{X}} \delta_{(S,a),X_k}\delta_{(S',b),X_{k+1}}, \tag{2}$$

which is referred to as the bonded pair count. The single-site counts and bonded pair counts are the two fundamental stochastic variables in the present theory. For later convenience, let us define the non-bonded pair counts as

$$n^{nb}_{SS'}(a,b|\mathbf{X}) = n_S(a|\mathbf{X})n_{S'}(b|\mathbf{X}) \tag{3}$$

for $S,S'\in\boldsymbol{S}$. Note that the non-bonded pair counts may be defined for residues residing on neighboring lattice sites as well as on the same (S=S') site. The terms "bonded" and "non-bonded" here are meant to describe the connectivity along the polypeptide sequence rather than that along the lattice system (A pair of residues in neighboring lattice sites may be either bonded or non-bonded depending on the given alignment). From these definitions, several relations follow. First, by the fermion-like character of the core site, we have for each $O_i\in\boldsymbol{O}$

$$\sum_{a\in\boldsymbol{A}_{O_i}} n_{O_i}(a|\mathbf{X}) = 1. \tag{4}$$

Between bonded pair counts and single-site count, we have

$$\sum_{b\in\boldsymbol{A}_{O_{i+1}}} n^b_{SO_{i+1}}(a,b|\mathbf{X}) + \sum_{b\in\boldsymbol{A}_{I_i}} n^b_{SI_i}(a,b|\mathbf{X}) = n_S(a|\mathbf{X}), \tag{5}$$

$$\sum_{a\in\boldsymbol{A}_{O_i}} n^b_{O_iS'}(a,b|\mathbf{X}) + \sum_{a\in\boldsymbol{A}_{I_i}} n^b_{I_iS'}(a,b|\mathbf{X}) = n_{S'}(b|\mathbf{X}) \tag{6}$$

where S=O_i,I_i and S'=O_{i+1},I_i. Lastly, between non-bonded pair counts and single-site count, we have

$$\sum_{b\in\boldsymbol{A}_{O_j}} n^{nb}_{SO_j}(a,b|\mathbf{X}) = n_S(a|\mathbf{X}), \tag{7}$$

$$\sum_{a\in\boldsymbol{A}_{O_i}} n^{nb}_{O_iS'}(a,b|\mathbf{X}) = n_{S'}(b|\mathbf{X}) \tag{8}$$

where $S,S'\in\boldsymbol{S}$.

1.3. Probability distribution of alignments

I would like to statistically characterize the given MSA in terms of the above quantities. To do so, suppose that the probability $P(\mathbf{X})$ of an alignment $\mathbf{X}$ is known for the lattice model. Then, the expectation values of these numbers are defined as follows:

$$n_S(a) = \sum_{\mathbf{X}} P(\mathbf{X})n_S(a|\mathbf{X}), \tag{9}$$

$$n^b_{SS'}(a,b) = \sum_{\mathbf{X}} P(\mathbf{X})n^b_{SS'}(a,b|\mathbf{X}), \tag{10}$$

$$n^{nb}_{SS'}(a,b) = \sum_{\mathbf{X}} P(\mathbf{X})n^{nb}_{SS'}(a,b|\mathbf{X}) \tag{11}$$

which are referred to as single-site (number) densities, bonded pair (number) densities, and non-bonded pair (number) densities, respectively. These number densities naturally satisfy the relations analogous to Eqs. (4)–(8).

To determine the form of $P(\mathbf{X})$, the principle of maximum entropy is employed with the constraints that the densities are equal to those observed in the given MSA. The entropy is given as

$$S = -\sum_{\mathbf{X}} P(\mathbf{X}) \ln P(\mathbf{X}) . \tag{12}$$

Let us denote the densities estimated from the given MSA as $\bar{n}_S(a)$, $\bar{n}^b_{SS'}(a,b)$, and $\bar{n}^{nb}_{SS'}(a,b)$ (see Section 2 for the method to obtain these quantities). The following Lagrangian, consisting of the entropy (Eq. 12) and the constraints for the densities, is maximized:

$$\begin{aligned} L = & -T \sum_{\mathbf{X}} P(\mathbf{X}) \ln P(\mathbf{X}) \\ & + \alpha\Big(\sum_{\mathbf{X}} P(\mathbf{X}) - 1\Big) \\ & + \sum_{(S,S')}^{\text{b.p.}} \sum_{a,b} J_{SS'}(a,b)\, [n^b_{SS'}(a,b) - \bar{n}^b_{SS'}(a,b)] \\ & + \frac{1}{2} \sum_{S,S'} \sum_{a,b} K_{SS'}(a,b)\, [n^{nb}_{SS'}(a,b) - \bar{n}^{nb}_{SS'}(a,b)] \\ & + \sum_{S,a} \mu_S(a)\, [n_S(a) - \bar{n}_S(a)] \end{aligned} \tag{13}$$

where α, $\mu_S(a)$, $J_{SS'}(a,b)$ and $K_{SS'}(a,b)$ are undetermined multipliers, and the summation $\sum^{\text{b.p.}}_{(S,S')}$ is over bonded pairs. We have also introduced the "temperature" parameter T. Solving $\delta \boldsymbol{L}/\delta P(\mathbf{X})=0$ leads to the Boltzmann distribution:

$$P(\mathbf{X}) = \frac{\exp[-E(\mathbf{X})/T]}{\Xi} , \tag{14}$$

where Ξ is the normalization constant or the partition function defined by

$$\Xi = \sum_{\mathbf{X}} \exp[-E(\mathbf{X})/T] , \tag{15}$$

and $E(\mathbf{X})$ is the "energy" of the system given as

$$\begin{aligned} E(\mathbf{X}) = & -\sum_{(S,S')}^{\text{b.p.}} \sum_{a,b} J_{SS'}(a,b) n^b_{SS'}(a,b|\mathbf{X}) \\ & - \frac{1}{2} \sum_{S,S'} \sum_{a,b} K_{SS'}(a,b) n_S(a|\mathbf{X}) n_{S'}(b|\mathbf{X}) \\ & - \sum_{S,a} \mu_S(a) n_S(a|\mathbf{X}) . \end{aligned} \tag{16}$$

From this expression of the energy function, we can interpret $\mu_S(a)$ as the chemical potential imposed on the particle (amino acid residue) a at site S, and J and K as bonded and non-bonded coupling parameters, respectively. The problem of obtaining the probability distribution $P(\mathbf{X})$ is thus reduced to computing the partition function Ξ. In the following, the non-bonded interactions are considered only between core sites (i.e., core-insert and insert-insert pairs are discarded) for a technical reason (see the subsection "Determining the K matrix" below).

1.4. Partition function

In this subsection, I assume that the parameters μ, J and K are fixed. To treat the long-range interactions, a mean-field approximation is applied. Then, the partition function can be computed by a transfer matrix method. Let us define the mean field $K_S(a)$ acting on the residue type a on site S:

$$\tilde{K}_S(a) = \sum_{S',b} K_{SS'}(a,b)\, [n_{S'}(\mathrm{b}) - \bar{n}_{S'}(b)] \tag{17}$$

where $\bar{n}_{S'}(b)$ is subtracted for convenience, but this does not essentially change the system's behavior (it simply shifts the chemical potential $\mu_S(a)$ which can be compensated for by J; see Eq. 18 and Section 1.7). Next, let us define the transfer matrices between a bonded pair of sites $S=O_i$, I_i and $S'=O_{i+1}$, I_i as

$$T_{SS'}(a,b) = \exp[\{J_{SS'}(a,b) + \mu_{S'}(b) + \tilde{K}_{S'}(b)\}/T] . \tag{18}$$

To alleviate the expressions for the partial partition functions, a bracket notation is introduced. First, define a set of standard basis vectors: $\langle a|$ and $|a\rangle$ corresponding to each residue type a on each site. These vectors satisfy the following orthonormal properties:

$$\langle a|b\rangle = \delta_{a,b} , \tag{19}$$

$$\sum_{a \in A_S} |a\rangle \langle a| = \mathbf{I}_{|A_S|} \text{ (identity matrix)} \tag{20}$$

where $\mathbf{I}_{|AS|}$ is the $|A_S|$-dimensional identity matrix. For each site i, I define the partial partition functions $\langle O_i|$ and $\langle I_i|$ that count the statistical weight of all possible alignments starting from the start site O_0 and terminating at O_i and I_i, respectively. Similarly, partial partition functions $|O_i\rangle$ and $|I_i\rangle$ account for all possible alignments "starting" from the end site O_{N+1} and "terminating" at O_i and I_i. Any (complete) alignment starts at the start site O_0 and ends at the end site O_{N+1}, and these sites are formally treated as "deletion (–)." Therefore, the boundary conditions are given as

$$\langle O_0| \;=\; \langle -| = (0,\ldots,0,1), \tag{21}$$

$$|O_{N+1}\rangle \;=\; |-\rangle = (0,\ldots,0,1)^t . \tag{22}$$

Based on this setting, the recursion formulae for partial partition functions are given as

$$\langle O_{i+1}| \;=\; \langle O_i|T_{O_iO_{i+1}} + \langle I_i|T_{I_iO_{i+1}} , \tag{23}$$

$$\langle I_i| \;=\; \langle O_i|T_{O_iI_i} + \langle I_i|T_{I_iI_i} \tag{24}$$

in the forward (N- to C-terminal) direction, and

$$|O_i\rangle \;=\; T_{O_iO_{i+1}}|O_{i+1}\rangle + T_{O_iI_i}|I_i\rangle , \tag{25}$$

$$|I_i\rangle \;=\; T_{I_iO_{i+1}}|O_{i+1}\rangle + T_{I_iI_i}|I_i\rangle \tag{26}$$

in the backward (C- to N-terminal) direction. Here, each transfer matrix $T_{SS'}$ is viewed as a $|A_s|\times|A_{s'}|$ matrix with $\langle a|T_{SS'}|b\rangle = T_{SS'}(a,b)$. By expanding Eq. (24), we have

$$\langle I_i| = \langle O_i|T_{O_iI_i}(\mathbf{I} + T_{I_iI_i} + T^2_{I_iI_i} + \cdots) \quad (27)$$

$$= \langle O_i|T_{O_iI_i}(\mathbf{I} - T_{I_iI_i})^{-1} \quad (28)$$

where $\mathbf{I}=\mathbf{I}_{20}$ (the 20-dimensional identity matrix). Similarly, we have

$$|I_i\rangle = (\mathbf{I} - T_{I_iI_i})^{-1}\,T_{I_iO_{i+1}}|O_{i+1}\rangle\,. \quad (29)$$

Thus, $\langle I_i|$ and $|I_i\rangle$ indeed include contributions from infinitely long insertions. The inverse matrix $(\mathbf{I}-T_{I_iI_i})^{-1}$ exists if the spectral radius of $T_{I_iI_i}$ is less than 1.

Using Eqs. (28) and (29), the recursions can be explicitly solved as

$$\langle O_{i+1}| = \langle O_0|\prod_{k=0}^{i} U_{k,k+1}\,, \quad (30)$$

$$|O_i\rangle = \prod_{k=i}^{N} U_{k,k+1}|O_{N+1}\rangle \quad (31)$$

where

$$U_{i,i+1} = T_{O_iO_{i+1}} + T_{O_iI_i}(\mathbf{I} - T_{I_iI_i})^{-1}\,T_{I_iO_{i+1}}\,. \quad (32)$$

Finally, the total partition function is obtained as

$$\Xi = \langle O_0|\prod_{k=0}^{N} U_{k,k+1}|O_{N+1}\rangle\,. \quad (33)$$

1.5. Expected densities

Let us now compute the expected densities. From the definition of the partition function (Eq. 15), the following equalities hold for single-site and bonded pair densities:

$$T\frac{\partial \ln \Xi}{\partial \mu_S(a)} = n_S(a)\,, \quad (34)$$

$$T\frac{\partial \ln \Xi}{\partial J_{SS'}(a,b)} = n^b_{SS'}(a,b)\,. \quad (35)$$

By explicitly calculating the left-hand sides of these equations using Eq. (33), we have, for $S = O_i, I_i$ and $S' = O_{i+1}, I_i$,

$$n_s(a) = \frac{\langle S|a\rangle\;\langle a|S\rangle}{\Xi}\,, \quad (36)$$

$$n^b_{SS'}(a,b) = \frac{\langle S|a\rangle\;\langle a|T_{SS'}|b\rangle\;\langle b|S'\rangle}{\Xi}\,. \quad (37)$$

It is readily proved that these expressions satisfy the relations between bonded pair and single-site densities (Eqs. 5–6).

It is also possible to derive an analytical expression for the expected non-bonded pair densities from

$$T^2\frac{\partial^2 \ln \Xi}{\partial \mu_S(a)\,\partial \mu_{S'}(b)} = n^{nb}_{SS'}(a,b) - n_S(a)n_{S'}(b)\,. \quad (38)$$

That is,

$$n^{nb}_{SS'}(a,b) = \frac{\langle S|a\rangle\;\langle a|\Xi_{SS'}|b\rangle\;\langle b|S'\rangle}{\Xi} \quad (39)$$

where

$$\Xi_{O_iO_j} = \prod_{k=i}^{j-1} U_{k,k+1}\,, \quad (40)$$

$$\Xi_{O_iI_j} = \Xi_{O_iO_j}T_{OjIj}(\mathbf{I} - T_{I_jI_j})^{-1}\,, \quad (41)$$

$$\Xi_{I_iO_j} = (\mathbf{I} - T_{I_iI_i})^{-1}\,T_{I_iO_{i+1}}\Xi_{O_{i+1}O_j}\,, \quad (42)$$

$$\Xi_{I_iI_j} = (\mathbf{I} - T_{I_iI_i})^{-1}\,T_{I_iO_{i+1}}\Xi_{O_{i+1}I_j}\,. \quad (43)$$

However, Eq. (39) is not used in practice for the reason described below (Section 2.4). This expression should be considered as an artifact of the present approximation on the one-dimensional lattice system. In fact, under the mean-field approximation, one should have $n^{nb}_{SS'}(a,b)=n_S(a)n_{S'}(b)$, but this does not hold for Eq. (39).

1.6. Thermodynamic functions

Several "thermodynamic functions" are defined for quantifying the stability of the system under perturbations. First, the free energy function

$$\Omega = -T\ln\Xi \quad (44)$$

should be regarded as a grand potential because alignments of varying lengths are considered in the ensemble. This free energy is a measure of the likelihood of alignments expressed in terms of the number densities. By rearranging Eq. (14) and averaging over all alignments, the free energy can be decomposed as

$$\Omega = U - TS - G \quad (45)$$

where U, S and G are the internal energy, entropy and Gibbs energy of the system. The internal energy of the system is given as

$$U = U_b + U_{nb} \quad (46)$$

Where U_b and U_{nb} are bonded and non-bonded energies, respectively, defined (under the mean-field approximation) by

$$U_b = -\sum_{(S,S')}^{\text{b.p.}}\sum_{a,b} J_{SS'}(a,b)n^b_{SS'}(a,b)\,, \quad (47)$$

$$U_{nb} = -\frac{1}{2}\sum_{S}\sum_{a}\tilde{K}_S(a)[n_S(a) - \bar{n}_S(a)]\,. \quad (48)$$

These correspond to the first two terms on the right-hand side of Eq. (16). The internal energy represents the mean "direct" interactions (bonded and non-bonded) between sites. The Gibbs energy is defined as

$$G = \sum_{S,a}\mu_S(a)n_S(a)\,, \quad (49)$$

and this quantity represents the work exerted by the chemical potential to maintain the single-site densities. Finally, the entropy is given as

$$S = (\Omega - U + G)/T \tag{50}$$

which is equivalent to the entropy in Eq. (12) and thus is a measure of randomness of the alignments.

The temperature T is set to 1 and the chemical potentials are set to 0 for all $S \in \boldsymbol{S}$, $a \in \boldsymbol{A}_S$ when the parameters J (and K) are determined. This state is referred to as the reference state in the following.

1.7. Gauge fixing

The relations among the densities (Eqs. 4–8) indicate that not all the parameters, μ, J, and K, are independent. When determining or changing the model parameters, we may therefore fix some of them to arbitrary values without losing generality. From the normalization condition (Eq. 4) of core sites, it is always possible to set

$$\mu_{O_i}(-) = 0 \tag{51}$$

for all the sites $O_i \in \boldsymbol{O}$ ("–" stands for the deletion). From this and the relations Eqs. (5) and (6), it is always possible to set

$$J_{O_i O_{i+1}}(-,-) = 0\,. \tag{52}$$

Although there are other degrees of freedom that can be also fixed, they are not relevant to the present study so I will not fix them.

Furthermore, at the reference state, I set all $\mu_S(a)$ to zero. This is possible because any values of $\mu_{S'}(b)$ may be absorbed into $J_{SS'}(a,b)$ when determining the parameters (c.f., Eq. 18). Following the convention of Morcos *et al.* [11], I also set

$$K_{O_i O_j}(-,b) = K_{O_i O_j}(a,-) = 0\,, \tag{53}$$

for all $a \in \boldsymbol{A}_{O_i}$ and $b \in \boldsymbol{A}_{O_j}$.

2. Materials and Methods

2.1. Data preparation and determining lattice structure

I have downloaded the MSA's and profile HMM's for the globin (PF00042) and (immunoglobulin) V-set (PF07686) domains from the Pfam database (version 28) [20]. For the globin domain, the full alignment of 17,947 amino acid sequences were used. For the V-set domain, the full alignment of of 23,976 sequences was used. In addition, I have downloaded 17 families from the top 20 largest Pfam families with the model length of less than 300 sites. For these 17 families, the representative set of alignments (with 75% sequence identity cutoff) were used due to the large size of the alignments.

In the present study, the lattice structure of a MSA was derived from the corresponding Pfam model. That is, each core site corresponds to a profile HMM match state, and each insert site to a profile HMM insert state.

2.2. Observed densities

The simplest way to estimate the single-site, bonded and non-bonded pair densities from a MSA of M sequences is to approximate $P(\mathbf{X}) = 1/M$ for all the M sequences. In practice, I used pseudo-counts as well as sequence weights as in Morcos *et al.* [11] to improve the robustness of the estimates. Let there be M aligned sequences, $\{\mathbf{X}^t\}_{t=1,\ldots,M}$, in a given MSA and suppose the structure of the lattice system has been set. The observed densities are defined as follows:

$$\bar{n}_S(a) = C\left[\frac{\gamma}{q_s} + \sum_{t=1}^{M}\frac{n_S(a|\mathbf{X}^t)}{m_t}\right], \tag{54}$$

$$\bar{n}^b_{SS'}(a,b) = C\left[\frac{\gamma}{2q_s q_{s'}} + \sum_{t=1}^{M}\frac{n^b_{SS'}(a,b|\mathbf{X}^t)}{m_t}\right], \tag{55}$$

$$\bar{n}^{nb}_{SS'}(a,b) = C\left[\frac{\gamma}{q_s q_{s'}} + \sum_{t=1}^{M}\frac{n^{nb}_{SS'}(a,b|\mathbf{X}^t)}{m_t}\right] \tag{56}$$

where $S \in \boldsymbol{S}$, $q_S = |\boldsymbol{A}_S|$, γ is the pseudo-count, m_t is the number of sequences in the MSA that are highly homologous (>80% sequence identity) to the sequence t, and $C = 1/(\gamma + \sum_t 1/m_t)$ with $\gamma = 0.1\sum_t 1/m_t$. Note that these estimated densities satisfy the relations analogous to Eqs. (4)–(8).

2.3. Determining the *J* matrices

As mentioned above, the temperature is set to unity ($T = 1$) in the process of parameter determination. To determine J, Eq. (37) is rearranged to

$$J_{SS'}(a,b) = \log\left[\frac{n^b_{SS'}(a,b)\Xi}{\langle S|a\rangle\,\langle b|S'\rangle}\right] \tag{57}$$

where it is assumed $\mu_{S'}(b) = 0$ and $\tilde{K}_{S'}(b) = 0$ for all $S' \in \boldsymbol{S}$ and $b \in \boldsymbol{A}_{S'}$ (see Section 1.7). Setting $\tilde{K}_{S'}(b) = 0$ is possible because the expected number densities are set to the observed values (see Eq. 17). By replacing $n^b_{SS'}(a,b)$ with the observed value $\bar{n}^b_{SS'}(a,b)$, one can iteratively update the values of J and compute the partition function until this equation actually holds. In practice, a relaxation parameter α is introduced to improve the stability of convergence. Thus, from the ν-th step of iteration, the next updated value is obtained by the following scheme.

$$J'_{SS'}(a,b) = \log\left[\frac{\bar{n}^b_{SS'}(a,b)\Xi^{(\nu)}}{\langle S^{(\nu)}|a\rangle\,\langle b|S'^{(\nu)}\rangle}\right], \tag{58}$$

$$J^{(\nu+1)}_{SS'}(a,b) = (1-\alpha)J^{(\nu)}_{SS'}(a,b) + \alpha J'_{SS'}(a,b)\,. \tag{59}$$

I found the values α=0.1~0.3 were effective. The initial values of the J matrices are set to 0. Note that the concavity of the free energy function with respect to the parameters (J) (with fixed mean fields) guarantees that the optimized values of J are unique and do not depend on the initial values (see Goldenfeld [21], for example).

Determining $J_{I_i I_i}$ necessitates a special treatment due to the requirement that the spectral radius of the transfer matrix $T_{I_i I_i}$ must be less than 1 (see Eq. 28). In order to force $\mathbf{I} - T_{I_i I_i}$ to be invertible, a parameter $\lambda_i > 0$ is introduced such that $\|T_{I_i I_i}/\lambda_i\| < 1$. Then Eq. (37) for $S = S' = I_i$ becomes

$$n^b_{I_iI_i}(a,b) = \frac{\langle I_i|a\rangle \langle a|T_{I_iI_i}|b\rangle \langle b|I_i\rangle}{\Xi\lambda_i}. \quad (60)$$

Let us define the "loop length" l_i as

$$l_i = \sum_{a,b\in A_{I_i}} n^b_{I_iI_i}(a,b) \quad (61)$$

and denote its observed counterpart by $\bar{l}_i$. By imposing $l_i=\bar{l}_i$ we have

$$\lambda_i = \frac{\langle I_i|T_{I_iI_i}|I_i\rangle}{\Xi\bar{l}_i} \quad (62)$$

which is a self-consistent equation for λ_i. Thus, first λ_i is set to a sufficiently large value (the Frobenius norm [22] of the matrix $J_{I_iI_i}$ is used in practice) and compute the partition function and expected densities. Then, λ_i is updated by Eq. (62), and by using the updated value of λ_i, we again compute the partition function and expected densities. This process is repeated until the value of λ_i converges. After the convergence of λ_i for all i, $J_{I_iI_i}$ is updated as in Eq. (37) without including λ_i. In this way, the contribution of λ_i is incorporated into the updated value of $J_{I_iI_i}$, and λ_i will eventually converge to 1, and hence may be omitted in later calculations.

The overall procedure for determining the J matrix is shown in Figure 3. In this procedure, the given data are the observed densities and initial values for J and λ_i. After the partition function and expected densities are computed, λ_i is iteratively updated. After λ_i has converged, J is updated. Convergence is checked based on the difference of the expected bonded pair densities from their observed values: when the root mean square difference between the two densities is less than 10^{-12}, the iteration is stopped.

2.4. Determining the *K* matrix

In this study, only those between core sites are taken into account for non-bonded interactions. Including non-bonded interactions with insert sites was found to be numerically unstable because the spectral radius of $T_{I_iI_i}$ may easily exceed 1. Noting the gauge fixing (Eq. 53), we first determine $K_{O_iO_j}(a,b)$ viewed as a $20N\times 20N$ matrix (consisting of $N\times N$ blocks of 20×20 submatrices) by discarding the rows and columns including deletion. Then, by fixing the values of K, we determine the J matrices.

Let the observed covariance matrix of single-site counts be C:

$$C_{O_iO_j}(a,b) = \bar{n}^{nb}_{O_iO_j}(a,b) - \bar{n}_{O_i}(a)\bar{n}_{O_j}(b). \quad (63)$$

In a similar manner as in Morcos *et al.* [11], one could apply the Plefka expansion [11,23,24] to the grand potential (Eq. 44) with $K=0$ as the reference state. However, I found that K thus obtained made the system unstable under very weak perturbations. This behavior is perhaps due to the incompatibility of the mean-field approximation with the one-dimensional system (see the remark at the end of Section 1.5).

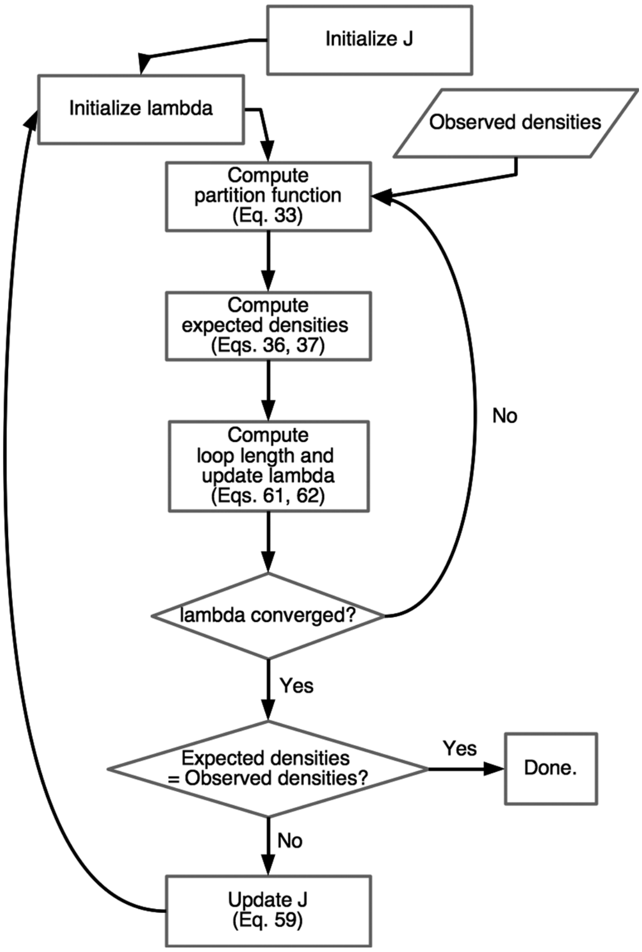
Initialize J
Initialize lambda
Observed densities
Compute partition function (Eq. 33)
Compute expected densities (Eqs. 36, 37)
Compute loop length and update lambda (Eqs. 61, 62)
lambda converged?
No
Yes
Expected densities = Observed densities?
Yes
Done.
No
Update J (Eq. 59)

Figure 3 Flow chart for determining the *J* matrix parameters.

In order to cope with this problem, I employ the following Gaussian (harmonic) approximation. By assuming the single-site densities are Gaussian random variables yielding the observed covariance, the non-bonded coupling is given as

$$K = -C^{-1}, \quad (64)$$

which is identical to that derived by Morcos *et al.* [11] except for the diagonal blocks (i.e., $K_{O_iO_i}$). Unlike their case (where the diagonal blocks are defined to be zero), I use the expression for K as in Eq. (64) including the diagonal blocks. The system was again found to be unstable when the diagonal blocks (and those for bonded pairs) of K were set to zero. This approximation makes the K matrix negative semi-definite so that the observed single-site densities are the most stable ones and there are no other optima as far as non-bonded pairs are concerned.

2.5. Self-consistent solutions with fixed parameters

To obtain a self-consistent solution for the recursion equation (Eqs. 23–26) with a given set of parameters μ, J and K, we first set the mean-field $\tilde{K}_S(a)=0$ for all S and a. Then compute the partition function and the expected densities

$n_S(a)$ and update $\tilde{K}_S(a)$ by Eq. (17). This process is repeated until convergence. In practice, however, I do not use this self-consistent solution (see below).

2.6. Self-consistent solutions with fixed sequence length

Note that our partition function is that of a grand canonical ensemble so the total number of particles (residues) can vary. In practice, however, it is preferable to fix the sequence length for comparing different conditions to be meaningful. This can beachieved by adjusting the chemical potentials. First, let us define the sequence length as the number of particles in the system:

$$L = \sum_{S\in \boldsymbol{S}} \sum_{a=1}^{20} n_S(a). \tag{65}$$

Note that the deletion (a=21 when S=O_i) is not included here. Let $\bar{L}$ denote the target sequence length (a constant) which is computed using Eq. (65) with the observed densities. For example, the globin domain (see below) consisting of 110 core sites and 111 insert sites has the target sequence length of ≈109.2 (this is less than 110 due to the presence of deletions at core sites). At every step of self-consistent calculation, update the chemical potential of each residue (except for deletion) by

$$\mu_S(a) = \mu_S(a) + \epsilon(\bar{L} - L) \tag{66}$$

where ϵ is a small positive constant (ϵ≈0.001). The iteration is terminated when the largest difference of $\mu_S(a)$ becomes less than 10^{-12}. The flow chart of this procedure is shown in Figure 4.

2.7. Self-consistent solutions with fixed single-site densities

In virtual alanine scanning experiments, the single-site densities of particular sites is specified. Given densities $\hat{n}_S(a)$ for all $a\in\boldsymbol{A}_S$ for a particular site S can be specified by adjusting the chemical potentials at every iteration of the self-consistent calculations:

$$\mu_S(a) = \mu_S(a) + \epsilon'[\hat{n}_S(a) - n_S(a)] \tag{67}$$

where ϵ' is a positive constant (ϵ'≈10). For the case of core sites, it is always possible to set $\mu_S(-)$=0 by subtracting this value from those of other residue types of the same site. When the sequence length is to be fixed as well, both Eqs. (66) and (67) are applied (Fig. 4).

2.8. Measures of site conservation and difference

A measure of site conservation is the site entropy [25] defined by

$$H_{O_i} = -\sum_a \bar{n}_{O_i}(a) \ln \bar{n}_{O_i}(a) \tag{68}$$

for the reference state. The more well-conserved a site, the lower the value of the site entropy. The difference between the reference state and a perturbed state is measured by the Kullback-Leibler divergence [25]:

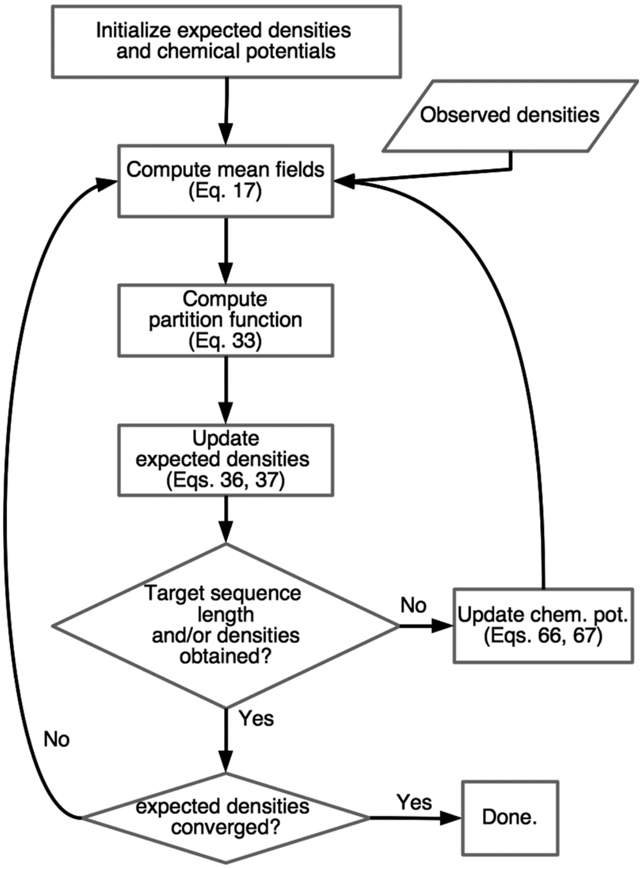
Initialize expected densities and chemical potentials
Observed densities
Compute mean fields (Eq. 17)
Compute partition function (Eq. 33)
Update expected densities (Eqs. 36, 37)
Target sequence length and/or densities obtained?
No
Update chem. pot. (Eqs. 66, 67)
Yes
No
expected densities converged?
Yes
Done.

Figure 4 Flow chart for obtaining the self-consistent solution with fixed sequence length (and fixed single-site densities). This process assumes that the parameters J (see Fig. 3) and K (Eq. 64) have been already determined. Initial expected densities and chemical potentials are set to the observed ones and 0, respectively. The constraint for the target sequence length is always imposed in the results given in this work. In addition, the constraint for the target densities is imposed on mutated sites in the case of alanine scanning. The convergence of target sequence length and target densities is reached if the chemical potentials does not change more than 10^{-12}. The convergence of expected densities is reached if the free energy does not change by more than 10^{-12} after an update.

$$D_{O_i} = \sum_a n_{O_i}(a) \ln \frac{n_{O_i}(a)}{\bar{n}_{O_i}(a)}, \tag{69}$$

and the total divergence is defined by

$$D = \sum_{i=1}^{N} D_{O_i}. \tag{70}$$

3. Results

I now study the behavior of the lattice gas model of multiple sequence alignment by varying temperature or by "mutating" a site. I mostly focus on the effect of non-bonded interactions in the following. For this purpose, I compare the system including both the bonded and non-bonded interactions (referred to as the "J+K" system in the following) with that including only the bonded interactions (the "J-only" system). The calculations for the J-only system were performed by simply discarding the mean-field, which is justified due to the present definition of the mean-field (Eq. 17).

All the calculations in the following are based on the "fixed-length" solution, and the sequence length (Eq. 65) was constrained to that of the reference state.

3.1. Parameter determination

The parameters are determined as described in the previous section. For all the MSA's tested below and in both the $J+K$ and J-only systems, the expected bonded pair densities matched precisely with the observed densities (RMSD$<10^{-12}$). Iterative updates of J (Eq. 59) usually converged within 200 steps. This observation confirms the validity of the present procedure.

3.2. Temperature scanning

Note that the present model does not exhibit phase transition due to the Gaussian approximation of the non-bonded pair interactions. That is, the K matrix is negative semidefinite so that there exists one and only one minimum for the non-bonded interactions (i.e., at the observed single-site densities). Nevertheless, solving the self-consistent equation with varying temperatures helps to understand the behaviors of interactions. At high temperatures, all the interactions are effectively weakened. This can be regarded as an idealization of uniform random mutations along the protein sequences of the given family. By observing the residue compositions perturbed by increased temperature, we can see which sites are more robust under the perturbations.

3.2.1. Globin domain

The globin domains are found in a wide variety of organisms ranging from bacteria to higher eukaryotes. Two of the most famous family members are myoglobins and hemoglobins both of which bind the heme prosthetic group. Structurally, globins belong to the class of all-α proteins, The lattice gas model of the globin domain consisted of 110 core sites (excluding the termini) and 111 insert sites.

The self-consistent equation was solved for temperature ranging from $T=1.0$ to $T=1.7$. Above the latter temperature, the solution could not be obtained stably because the spectral radius of some $T_{I_iI_i}$ exceeded 1.

As the temperature increases, the free energy (grand potential, Eq. 44) increases up to around $T=1.15$ and then it starts to decrease (Fig. 5A). Decomposing the free energy (Eq. 45) shows that both the internal energy (Fig. 5B) and entropy (Fig. 5C) increase with temperature. On the other hand, the Gibbs energy (Eq. 49) monotonically decreases with increasing temperature (Fig. 5D), indicating that the sequence length tends to be longer for higher temperature. This can be understood from the definition of the transfer matrix $T_{I_iI_i}$. Since $\|T_{I_iI_i}\|<1$ is required, $J_{I_iI_i}(a,b)<0$ holds for all $a,b\in\boldsymbol{A}_{I_i}$ $(I_i\in\boldsymbol{I})$ so the increased temperature potentially allows a larger number of residues to reside at insert sites. In order to fix the sequence length, the chemical potential must be negative, and hence the negative Gibbs energy.

The behaviors of the $J+K$ and J-only systems appear similar regarding the free energy, internal energy, entropy and Gibbs energy. To see the effect of non-bonded interactions more closely, the internal energy was decomposed into bonded interactions and non-bonded interactions for the $J+K$ system (Fig. 5E). It appears that the increase in non-bonded energy is more than an order of magnitude smaller (Fig. 5E, blue line) compared to that of bonded energy (Fig. 5E, magenta line). Furthermore, the divergence (difference of residue distributions from the reference state) shows a relatively large difference between the $J+K$ and J-only systems (Fig. 5F). Thus, the non-bonded interactions are very stable under increased temperatures, and they greatly stabilize the residue composition.

A closer examination of each site (at $T=1.2$) shows that the magnitude of the divergence of the J-only system is about three times as large as that of the $J+K$ system (Fig. 6). The broad peaks of the divergence roughly correspond to regions of α-helices. Furthermore, with non-bonded interactions, finer peaks match the periodicity of the helices (3 to 4 residues) whereas such periodicity is not observed with the J-only system. Thus, non-bonded interactions seem not only to stabilize the residue composition, but to make the composition more specific to the structure of the domain.

3.2.2. V-set domain

The V-set domains are found in many proteins the representative members of which are immunoglobulin variable domains. The lattice gas model of this domain consists of 114 core sites (excluding the termini) and 115 insert sites. Structurally, they belong to the all-β class having a β-sandwich structure.

The same procedures were applied to the V-set domain as the globin domain. In this case, however, self-consistent solutions could be obtained only for temperatures $T\leq1.25$. This may be due to a long insertion allowed at the insert site I_9 (average length of 23.5 residues). Other than this limitation, the results were found to be qualitatively similar to the case of globins (Fig. 7A–D). However, the free energy decrease is more pronounced for the $J+K$ system, compared to the case of the globin. Again, while the increase in temperature hardly changes the non-bonded energy (Fig. 7E), the difference of the total divergence between the $J+K$ and J-only systems is significant.

A close examination of individual sites at $T=1.2$ also indicates that inclusion of non-bonded interactions greatly suppresses the divergence, and broad peaks roughly correspond to secondary structure elements (in this case, β-strands). With the non-bonded interactions, finer peaks appear to match to the periodicity of β-strands (2 residues). Therefore, the conclusion drawn for the globin domain applies also to the V-set domain. That is, the non-bonded interactions act to

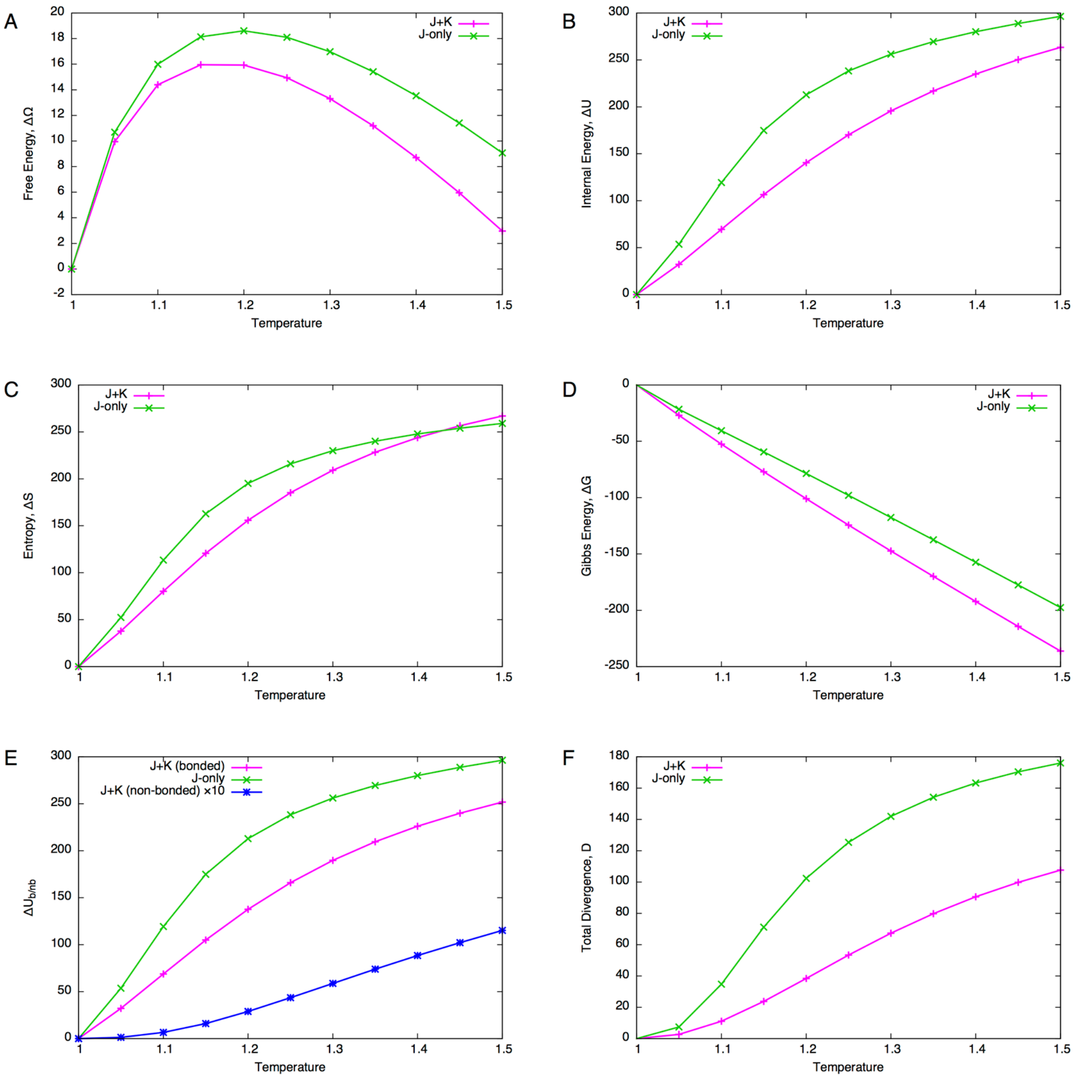

Figure 5 Temperature scanning of the globin domain. (A) Free energy difference ΔΩ from the reference state (T=1). (B) Internal energy difference ΔU. (C) Entropy difference ΔS. (D) Gibbs energy difference ΔG. (E) Decomposition of internal energy difference into bonded and non-bonded energy differences. The value of non-bonded energy difference (blue line) is multiplied by 10. (F) Total divergence of the core site compositions from the reference state (c.f., Eq. 70).

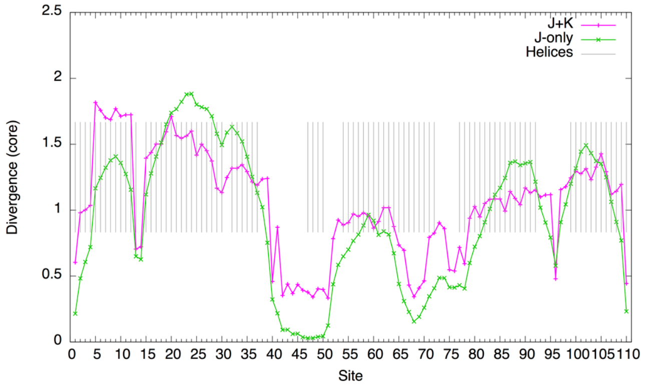

Figure 6 Divergence of core sites of the globin domain at T=1.2 (c.f., Eq. 69). Gray bars indicate sites annotated as helices (α-helix, "H" or 3_{10}-helix, "G") according to the Pfam model annotation (PF00042). The values for the J+K system are multiplied by 3.

stabilize the residue composition as well as to make composition more specific to the structure of the domain.

3.3. Alanine scanning

As opposed to global perturbations such as increased temperature, local perturbations helps us to examine the contribution of individual sites. Local perturbations can be imposed by biasing the residue composition at a site of interest. In this subsection, the composition of a particular core site was biased in such a way that single-site density was set to 0.95 for alanine and to 0.0025 for all other residue types (including the "deletion" residue type). This residue composition can be achieved by adjusting the chemical potential $\mu_{O_i}(a)$. When the site O_i is constrained in this way, the corresponding equilibrium state is referred to as the "A$_i$ mutant" in the following.

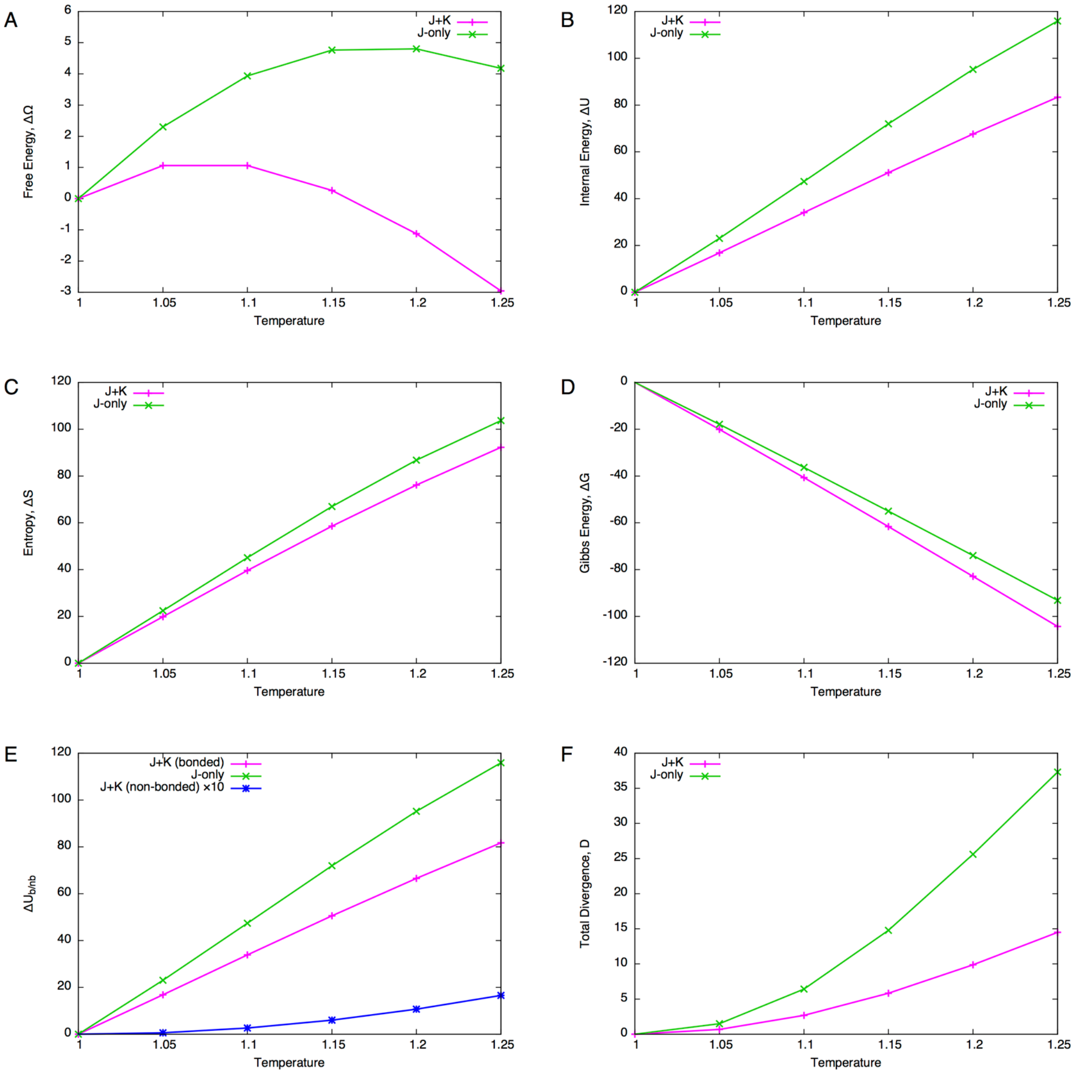

Figure 7 Temperature scanning of the V-set domain. (A) Free energy difference ΔΩ from the reference state (T=1). (B) Internal energy difference ΔU. (C) Entropy difference ΔS. (D) Gibbs energy difference ΔG. (E) Decomposition of internal energy difference into bonded and non-bonded energy differences. The value of non-bonded energy difference (blue line) is multiplied by 10. (F) Total divergence of the core site compositions from the reference state (c.f., Eq. 70).

3.3.1. Globin domain

Comparing the free energy difference between the $J+K$ and J-only systems, it is immediately noticed that the ranges of ΔΩ are very different between the two; the former being an order of magnitude larger than the latter. While a large number of alanine mutants for both the $J+K$ and J-only systems (82 and 101, respectively, out of 110) exhibit ΔΩ<0 (i.e., favorable mutants), the former ($J+K$) shows a larger number of unfavorable (ΔΩ>0) alanine mutants. Apart from the absolute values, the two systems appear to be correlated except for the region from the site 40 to 50 where secondary structures are sparse (c.f., Fig. 6). In addition, they seem to be negatively correlated with site entropy (Fig. 10A): Highly conserved sites tend to have high ΔΩ values (correlation coefficients, CC, were –0.60 and –0.57 for the $J+K$ and J-only systems, respectively). Thus, despite the great differ-

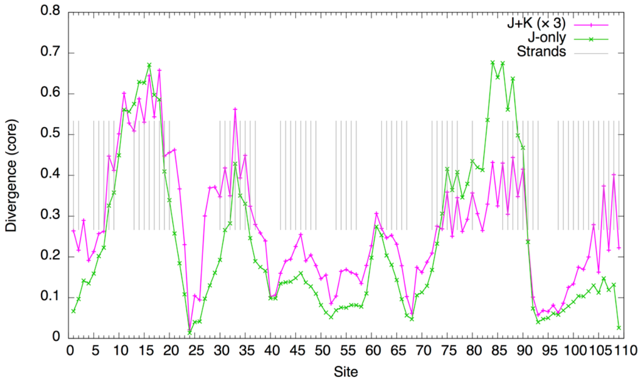

Figure 8 Divergence of core sites of the V-set domain at T=1.2 (c.f., Eq. 69). Gray bars indicate sites annotated as extended strand, "E" according to the Pfam model annotation (PF07686). The values for the $J+K$ system are multiplied by 3.

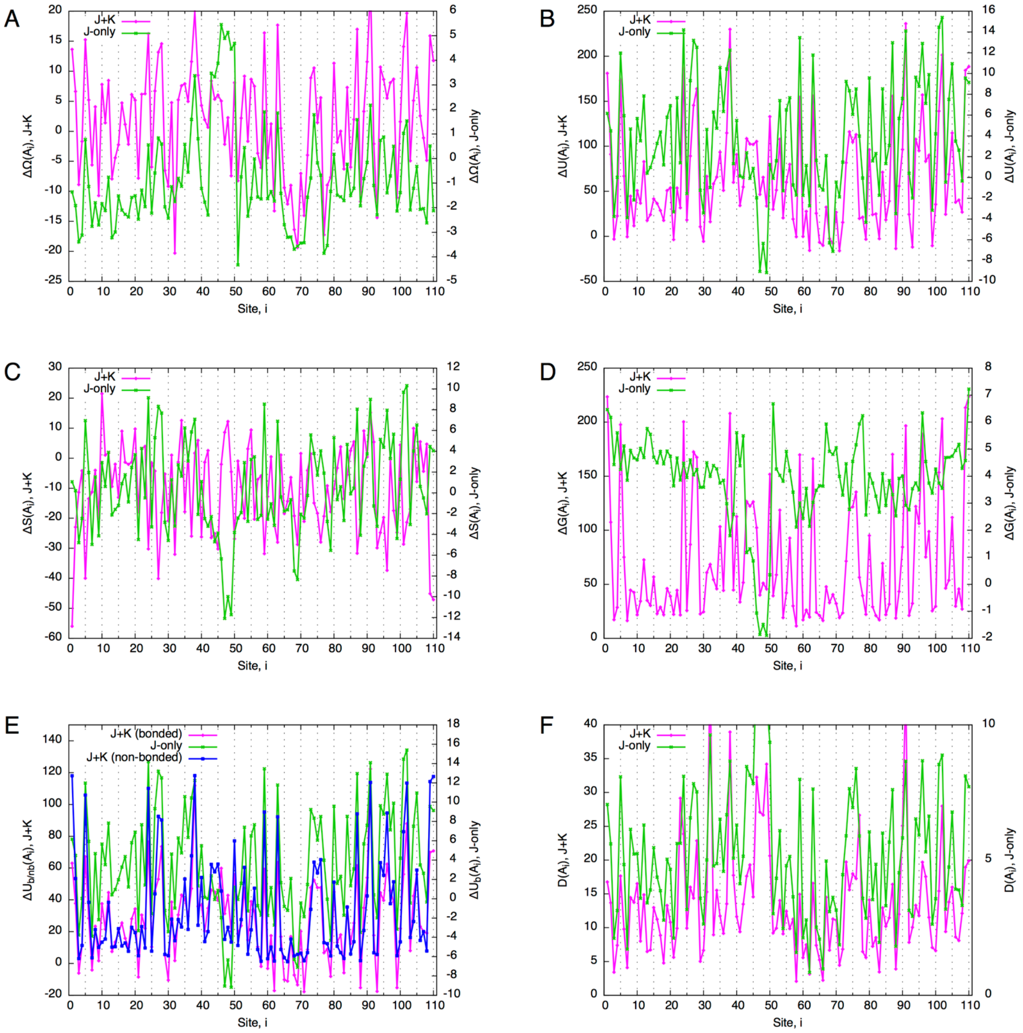

Figure 9 "Alanine scanning" of the globin domain. The horizontal axis indicates the site at which the single-site density of a core site was set to 0.95 for alanine, and to 0.0025 for other residue types; the vertical axes indicate associated values (A)–(F), with the *J*+*K* system on the left axis, and *J*-only system on the right. (A) Free energy difference of "alanine point mutants" from the reference state. (B) Internal energy difference. (C) Entropy difference. (D) Gibbs energy difference. (E) Decomposed internal energy difference. (F) Total divergence of core sites (Eq. 70).

ence in magnitudes, the *J*+*K* system and *J*-only system appear to be similar in terms of free energy difference. Behind this apparent similarity, however, exist different mechanisms, as we shall see in the following.

While internal energy difference, ΔU, also shows a similar correlation as $\Delta\Omega$ (Fig. 9B), entropy difference exhibits different, somewhat opposite, trends (Fig. 9C). In fact, the relations between the internal energy and entropy are completely different between the *J*+*K* and *J*-only systems (Fig. 10B). While ΔU and ΔS are linearly and positively correlated (CC=0.99) for the *J*-only system, their relation is more complicated for the *J*+*K* system: a positive correlation for $\Delta U<20$ (CC=0.65) and a negative correlation for $\Delta U>30$ (CC=–0.69). The region $\Delta U<20$ corresponds to that spanned by the *J*-only system, and therefore is considered to be the region where local (bonded) interactions are dominant in ΔU. This in turn indicates that a large increase in nonlocal (non-bonded) interactions greatly restricts the residue composition throughout the globin domain. In fact, unlike the case for temperature scanning (Fig. 5), the perturbation by a point mutation induces a large increase in non-bonded energy that is comparable with that of bonded energy in the *J*+*K* system (Fig. 9E).

The Gibbs energy difference, ΔG, reveals a sharp contrast between the two systems (Figs. 9D and 10C). The Gibbs energy differences of the *J*+*K* system are clustered below $\Delta G<50$, but has a long tail towards higher values (skewness was 1.1). On the other hand, those for the *J*-only system are more or less symmetrically distributed around $\Delta G=5$ (skewness was –1.4). The correlation between ΔG and site entropy is evident for the *J*+*K* system (CC=–0.71), but is nearly absent for the *J*-only system (CC=–0.18) (Fig. 10C).

The total divergence shows a trend similar to the Gibbs energy difference in that its values are clustered at lower val-

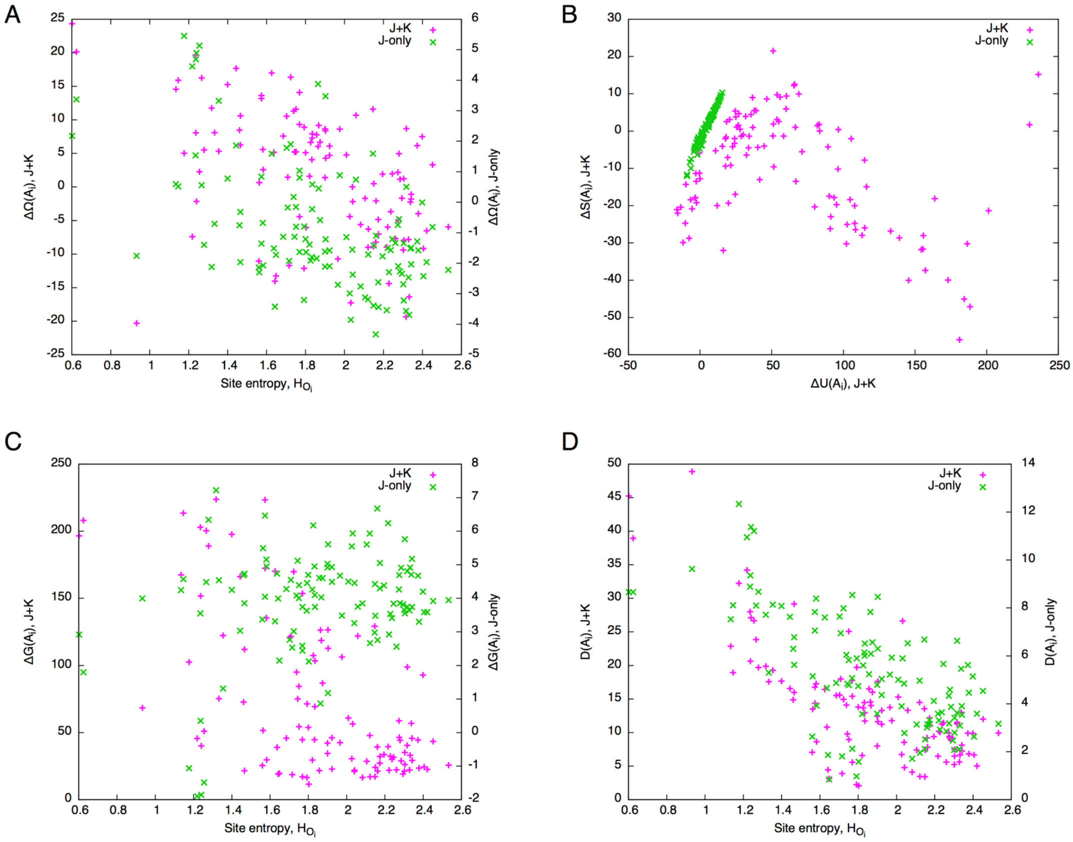

Figure 10 Correlations between various quantities for the globin domain. (A) Site entropy, H_{O_i}, vs. free energy change, ΔΩ (left vertical axis for the *J*+*K* system, right vertical axis for the *J* -only system). (B) Internal energy difference, Δ*U*, vs. entropy difference, Δ*S*. (C) Site entropy, H_{O_i}, vs. Gibbs energy change, Δ*G* (left vertical axis for the *J*+*K* system, right vertical axis for the *J* -only system). (D) Site entropy, H_{O_i}, vs. total divergence, *D* (left vertical axis for the *J*+*K* system, right vertical axis for the *J*-only system).

ues and has a long tail towards higher values for the *J*+*K* system, and that such is not the case for the *J*-only system (Fig. 9F). Although in both systems the total divergence is well correlated with site entropy, the correlation is higher for the *J*+*K* system (CC=–0.78) than for the *J*-only system (CC=–0.71) (Fig. 10D). In the *J*-only system, each mutation perturbs the residue compositions only locally around the mutated site, whereas in the *J*+*K* system, a mutation at one site perturbs many sites across the entire domain. As a result, the contrast between the effects of mutations at highly conserved sites and less conserved sites is higher for the *J*+*K* system than for the *J*-only system.

In the globin domain, the two most highly conserved residues are phenylalanine (Phe) at site 38 (H_{O_i}=0.67) and histidine (His) at site 91 (H_{O_i}=0.64). The alanine mutants at these sites show large differences in ΔΩ (Fig. 10A), Δ*U* (the two points with the largest Δ*U* in Fig. 10B) and Δ*G* (Fig. 10C). According to a detailed study by Ota *et al.* [26], these two residue are conserved for different reasons: Phe at site 38 (“CD1” in [26]) is conserved for structural stability whereas His at site 91 (“F8”) is conserved for the heme-binding function at the cost of structural stability. While it is reasonable to observe that the A_{38} mutant of the structurally conserved Phe significantly disturbs the system, the present result suggests that the A_{91} mutant of the functionally conserved His is also maintained by a significant amount of interactions with other sites. This may indicate the importance of structural scaffold to maintain protein function.

3.3.2. V-set domain

The case for the V-set domain is mostly similar to that for the globin domain (Figs. 11 and 12). However, there are some marked differences to be noted. First, the free energy differences ΔΩ due to alanine mutations take both positive and negative values for the *J*+*K* system, but only negative values for the *J*-only system except for several residues near the C-terminus. The positive values for the former corresponds to relatively well-conserved sites, as can be seen in Figure 12A. In fact, the correlation between ΔΩ and site entropy is significantly higher for the *J*+*K* system (CC=–0.72) than for the *J*-only system (CC=–0.52). Second, while the correlation between internal energy and entropy differences is linear and positive for the *J*-only system (CC=0.96) as was the case with the globin, that for the *J*+*K* system of the V-set domain shows only a negative trend for the entire range of Δ*U* (CC=–0.92). Third, the contrast of the Gibbs energy dif-

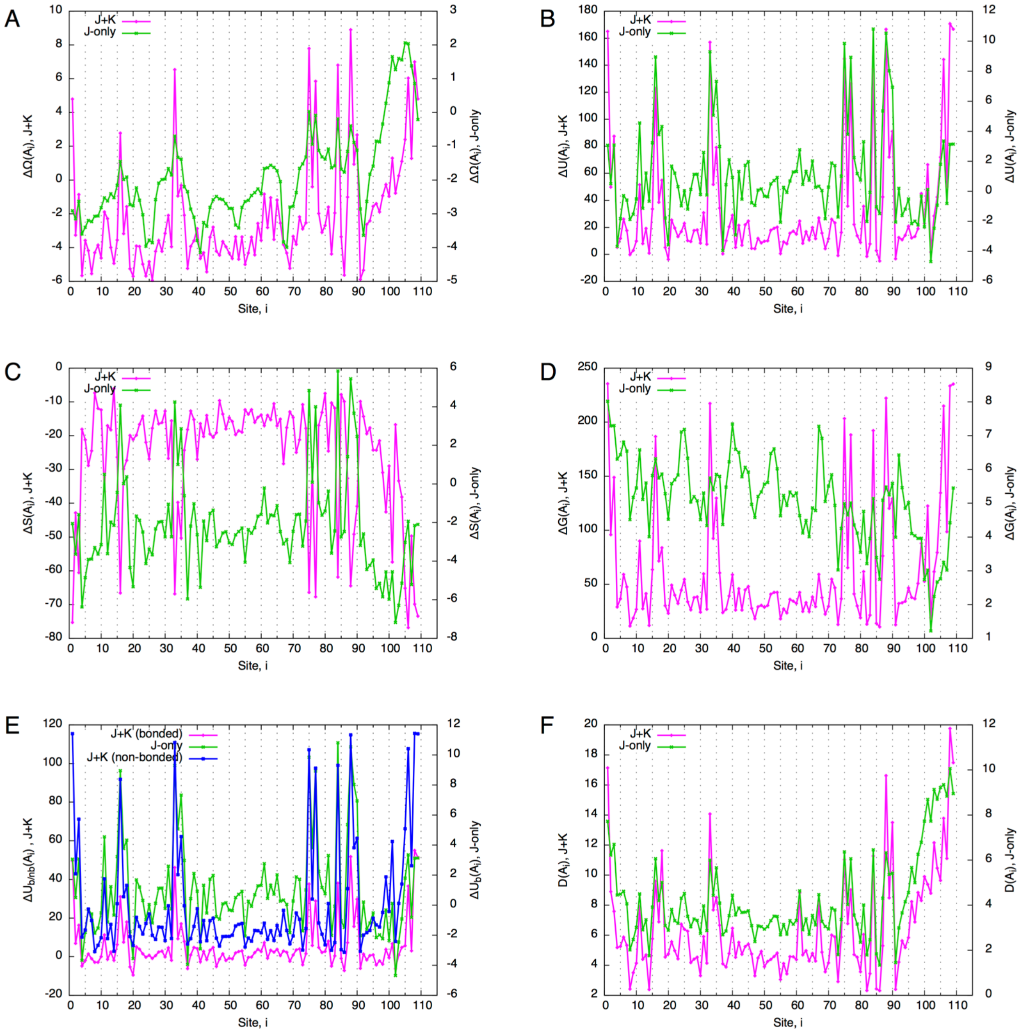

Figure 11 "Alanine scanning" of the V-set domain. The horizontal axis indicates the site at which the single-site density of a core site was set to 0.95 for alanine, and to 0.0025 for other residue types; the vertical axes indicate associated values (A)–(F), with the *J*+*K* system on the left axis, and *J* -only system on the right. (A) Free energy difference of "alanine point mutants" from the reference state. (B) Internal energy difference. (C) Entropy difference. (D) Gibbs energy difference. (E) Decomposed internal energy difference. (F) Total divergence of core sites (Eq. 70).

ference is far more pronounced (Fig. 11D, the skewness was 1.8 for *J*+*K* and –0.37 for *J*-only) and its correlation with site entropy is very high for the *J*+*K* system (CC=–0.80) whereas it is negligible for the *J*-only system (CC=–0.08) (Fig. 12C). Similarly, as for total divergence, the *J*+*K* system shows sharper contrast (Fig. 11F) and higher correlation with site entropy (CC=–0.80, Fig. 12D) than the *J*-only system (CC=–0.67).

Thus, compared to the case with the globin, the differences between the *J*+*K* and *J*-only systems are more pronounced. This may be due to the difference in the structures of these domains. The globin domain has an all-α fold in which local interactions in α-helices are prominent, whereas the V-set domain has an all-β fold in which nonlocal interactions between β-strands are prominent. This difference may be reflected in the non-bonded interactions of the lattice gas model, hence the pronounced difference between the *J*+*K* and *J*-only systems.

3.3.3. Other protein families

To confirm the observations made above, alanine scanning was performed for 17 Pfam families that are the largest in the number of family members and are of model length of less than 300 sites. The free energy difference, $\Delta\Omega(A_i)$, tends to have more positive values for the *J*+*K* system than for the *J*-only system (Fig. 13A, cf. Figs. 9A and 11C). The skewness (i.e., the standardized third moment) of $\Delta G(A_i)$ consistently have positive values for the *J*+*K* system whereas it can be either positive or negative for the *J*-only system (Fig. 13B). The negative correlation between site entropy and $\Delta G(A_i)$ was also clear for the *J*+*K* system whereas such was not the case for the *J*-only system (Fig. 13C). Thus, the trend that the non-bonded interaction enhances correlation with sequence conservation seem to hold generally.

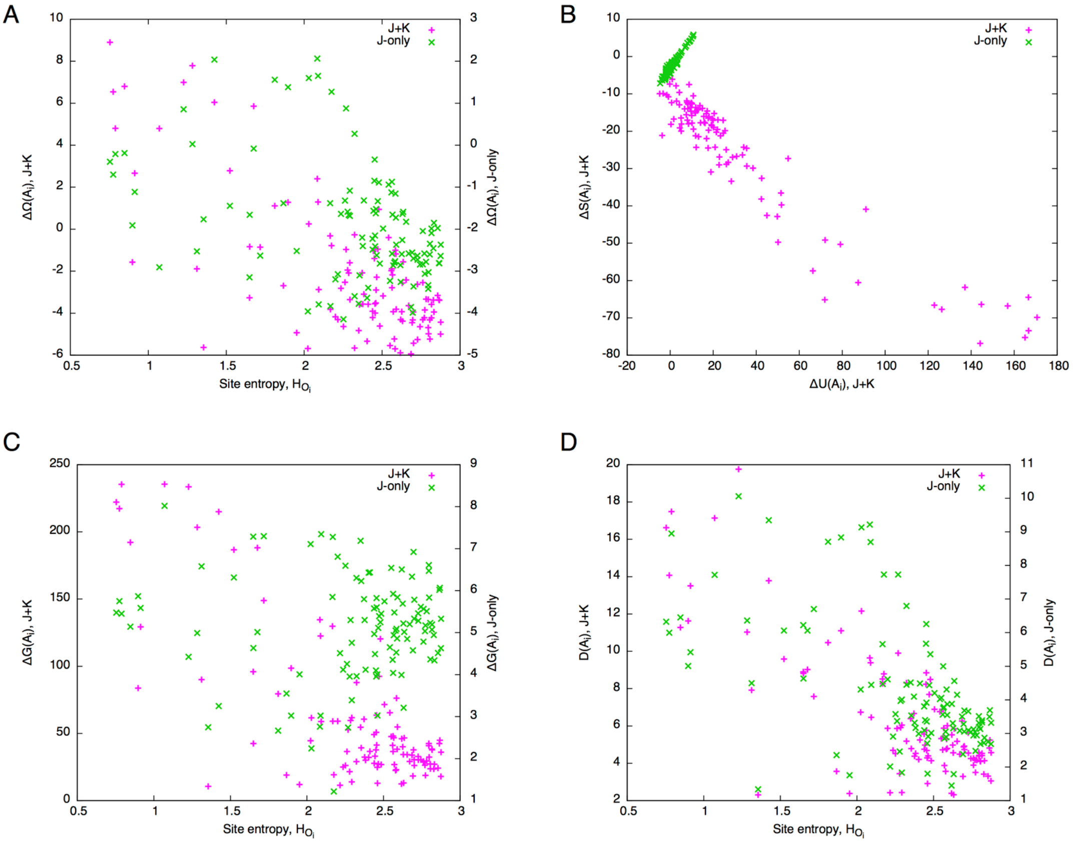

Figure 12 Correlations between various quantities for the V-set domain. (A) Site entropy, H_{O_i}, vs. free energy change, $\Delta\Omega$ (left vertical axis for the $J+K$ system, right vertical axis for the J -only system). (B) Internal energy difference, ΔU, vs. entropy difference, ΔS. (C) Site entropy, H_{O_i}, vs. Gibbs energy change, ΔG (left vertical axis for the $J+K$ system, right vertical axis for the J -only system). (D) Site entropy, H_{O_i}, vs. total divergence, D (left vertical axis for the $J+K$ system, right vertical axis for the J -only system).

4. Discussion

One of the fundamental assumptions of the present lattice gas model is that alignment sites can be classified into core sites and insert sites. Although this classification may be ambiguous to some extent, once the classification is made, the lattice structure is uniquely determined. While the lattice structure reflects the chemical structure of polypeptide chains, interactions between the lattice sites are not limited to those that are local along the chain. The principle of maximum entropy allows the model to treat bonded (local) and non-bonded (nonlocal) interactions in a coherent manner. In comparison, the profile HMM [1] shares a similar lattice structure as the lattice gas model, but it cannot treat nonlocal interactions due to its assumption of the Markov process along the lattice structure. On the other hand, the direct-coupling analysis (as applied to contact prediction) [11], which casts a MSA as a Potts model [27], simply ignores insert sites so that it cannot faithfully represent polypeptide chains. Threading methods [28] or conditional random field models [29] can combine the polypeptide structure with nonlocal interactions, but such integration is often *ad hoc* because there are no well-defined rules or principles for determining the relative contributions of various interactions. It is possible to treat a MSA without classifying its columns into cores and inserts if one ignores the possibility of adding new sequences in the future. In fact, this approach is adopted by the GREMLIN method by Balakrishnan *et al.* [30] that is based on the Markov random fields (the present lattice gas model also belongs to this class of statistical models). In practice, however, they discarded columns with excessive gaps. Such a preprocessing seems to be required because alignments within an insertion are often meaningless. This does not necessarily mean, however, that the existence of the insertion is meaningless. In any case, discarding columns of a MSA will lose the information about the linear chain structure of protein sequences as well as the possibility of adding new sequences without changing the core structure of the MSA. The present lattice gas model resolves the shortcomings of these previous models as both bonded and non-bonded interactions as well as insertions naturally emerge from a single framework. This feature of the model makes it possible to align new sequences against the model by using the dynamic programming technique (combined with iterations for self-consistency). The main tricks here are the classification of core and insert sites and the use of residue

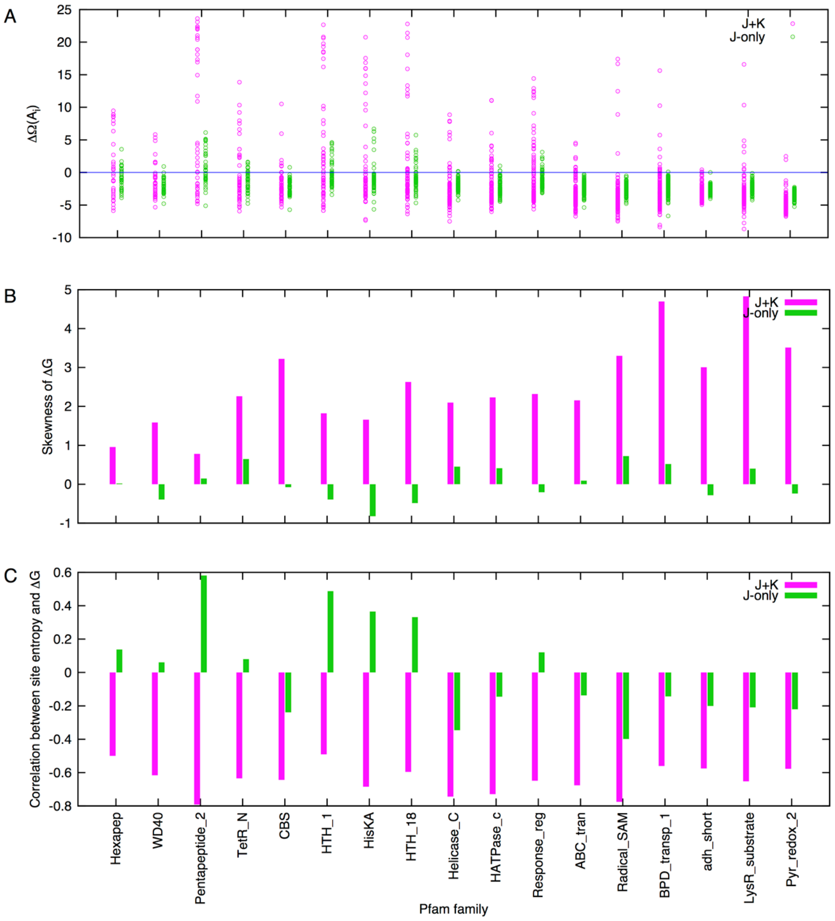

Figure 13 Alanine scanning of 17 Pfam families. (A) Free energy difference (cf. Figs. 9A and 11A). (B) Skewness (standardized) of ΔG (cf. Figs. 9D and 11D). (C) Correlation coefficient between site entropy and Gibbs energy ΔG (cf. Figs. 10C and 12C).

counts, $n_S(a|\mathbf{X})$ and $n^b_{SS'}(a,b|\mathbf{X})$, as fundamental variables rather than the raw alignment sequences (**X**). These are especially important for treating insert sites where any number of residues are allowed to exist. The lattice gas model can compute the probability of an entire alignment, and what has been conventionally regarded as the probability of residue occurrence at sites should be regarded as the expected number of residues at the sites.

From a theoretical point of view, the present formulation of the lattice gas model offers an interesting perspective regarding the interplay between local and nonlocal interactions. As can be seen from the relations Eqs. (5)–(8), or more precisely, from the analogous relations that hold for the number densities, local and nonlocal interactions are not independent of each other, but are related via single-site densities. In this sense, local and nonlocal interactions must be consistent with each other [31], and the consistency is inherently embedded in a (well-curated) MSA. In the conventional formulation of the direct-coupling analysis, only the relations corresponding to Eqs. (7) and (8) are present because the chain structure is absent. Since the parameters conjugate to the single-site densities are external fields (chemical potentials in the present case) which are not intrinsic to the system, the relations Eqs. (7) and (8) alone do not address the consistency between local and nonlocal interactions.

In this study, I have adopted the Gaussian approximation for the non-bonded coupling parameters (Eq. 64) as well as the mean-field approximation (Eq. 17) for computing the partition function. This approach has its advantages and disadvantages. The advantages are that the parameters are readily obtained and that the partition function can be computed analytically and efficiently. These enable us to study the system under various perturbations relatively easily. A major disadvantage is that it is not possible to determine the K matrix self-consistently. The Gaussian approximation implicitly assumes that each site is independent of other sites, which is not fully consistent with the lattice structure of the system. The reason for this inconsistency is likely to be that the assumption for the mean-field approximation (i.e., non-bonded interactions are relatively weak; see references [11,23]) does not actually hold in the present case. Due to this approximation, the system does not exhibit a phase transition that might otherwise be induced by increased temperatures or by mutations at potentially important sites.

In addition, the Gaussian approximation required that the diagonal blocks of the K matrix, $K_{O_iO_i}(a,b)$, be used as in Eq. (64), otherwise the reference state was found to be unstable. The diagonal blocks represent self-interactions, and hence, are purely site-specific quantities. In this sense, they obscure the mechanism by which the interactions of each site with other sites induce the residue composition of that site. Overcoming these problems would require the direct maximization of the Lagrangian (Eq. 13) with respect to the parameters $K_{SS'}(a,b)$ without diagonal (and bonded pair) blocks. It is also possible to apply other approximate methods such as pseudo-likelihood maximization [30,32,33].

Despite these limitations in the treatment of non-bonded interactions, the present results already provided some interesting observations regarding the role of non-bonded interactions. An increased temperature exerts a global and unbiased perturbation on the system. In this case, it was found that non-bonded energy did not significantly change compared to the bonded energy (Figs. 5E and 7E). This implies that the residue compositions at each site adapt to the perturbation in a cooperative manner so that they stay stable. This in turn suggests, at least within the limitation of the approximations, that a protein family can accommodate a diverse variety of amino acid sequences as far as the pattern of correlations between sites is conserved. On the other hand, the virtual alanine scanning revealed a more conspicuous effect of non-bonded interactions. Alanine mutations at well-conserved sites disturbed the system to a greater extent as measured by free energy, Gibbs energy and total divergence (Figs. 10 and 12), and the relation between internal energy and entropy changes was completely different from those of *J*-only systems. In particular, the observation that many or most of the free energy changes were negative for the *J*-only system (Figs. 9A and 11A) suggests that residue conservation cannot be explained without considering nonlocal (non-bonded) effects.

The interactions in the lattice gas model originate solely from the statistics of a MSA. They are therefore not directly related to physical interactions. However, it has been demonstrated that the K matrix as used in this study is a good predictor of physical contacts in native protein structures [9,10]. To further confirm this, the present results showed that the effect of non-bonded (statistical) interactions was more pronounced in the V-set domain (an all-β fold, involving more nonlocal physical interactions) than in the globin domain (an all-α fold, involving less nonlocal interactions). In addition, the *J*-only system showed relatively better correlations with conservation for the globin than for the V-set domain, indicating that the bonded interactions also reflect physical local interactions to some extent. This point is also supported by the correlation, albeit weak, between divergence and secondary structures (Figs. 6 and 8). Thus, the lattice gas model provides a means to connect the information in amino acid sequence with the underlying three-dimensional structure of the domain. This connection cannot be addressed directly in conventional sequence analysis methods such as the profile HMM. In fact, the very existence of long-range correlations indicates that MSA's cannot be modeled as a purely one-dimensional system where long-range correlations simply cannot exist [21]. Considering this fact, it is surprising that conventional multiple sequence alignment methods, inherently based on the one-dimensional system, can produce MSA's with long-range correlations. This may be a manifestation of the consistency principle indicated above [31].

There are a few possible extensions and applications of the present lattice gas model. In the present form, the model is autonomous in the sense that it does not require an input or target sequence for computing various statistical quantities (once the observed statistical quantities are obtained). Nevertheless, it is readily possible to align the model with a particular amino acid sequence to compute a partition function and therefore other quantities conditioned on that input sequence. In this way, the lattice gas model may be used for detecting remote homologs. The present results (e.g., Figs. 6 and 8) suggest that inclusion of non-bonded interactions would increase the specificity of the alignment. Furthermore, the model can be aligned with a "sequence" of a given length with unspecified amino acid residues to compute the partition function that is conditioned on all the amino acid sequences of that length. In this way, one can enumerate those sequences that are compatible with the model. In other words, the model may be used for designing optimal sequences for a given protein family. Such applications may be pursued in the future to open new possibilities in protein sequence analysis.

Acknowledgments

The author thanks Kentaro Tomii and Sanzo Miyazawa for reading the initial manuscript and providing some references and comments, Haruki Nakamura for the support during the development of this work, and Nobuhiro Go for a critical advice on the consistencyprinciple. This work was supported in part by a grant-in-aid "platform for drug discovery, informatics, and structural life sciences" from the MEXT, Japan.

Conflict of Interest

None declared.

Author contribution

ARK did everything.

References

[1] Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. (Cambridge Univ. Press, Cambridge, UK., 1999).

[2] Taylor, W. R. Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.* **188**, 233–258 (1986).

[3] Gribskov, M., McLachlan, A. D. & Eisenberg, D. Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* **84**, 4355–4358 (1987).

[4] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., *et al.* Gapped blast and PSI-blast: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

[5] Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531 (1994).

[6] Rost, B. Prediction in 1D: secondary structure, membrane helices, and accessibility. *Methods Biochem. Anal.* **44**, 559–587 (2003).

[7] Kinjo, A. R. & Nakamura, H. Nature of protein family signatures: insights from singular value analysis of position-specific scoring matrices. *PLoS ONE* **3**, e1963 (2008).

[8] Toh, H. Bioinformatics for functional analyses of proteins. (Kodan-sha, Tokyo, Japan, 2004).

[9] de Juan, D., Pazos, F. & Valencia, A. Emerging methods in protein co-evolution. *Nat. Rev. Genet.* **14**, 249–261 (2013).

[10] Taylor, W. R., Hamilton, R. S. & Sadowski, M. I. Prediction of contacts from correlated sequence substitutions. *Curr. Opin. Struct. Biol.* **23**, 473–479 (2013).

[11] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA* **108**, E1293–E1301 (2011).

[12] Jones, D. T., Buchan, D. W., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184–190 (2012).

[13] Miyazawa, S. Prediction of contact residue pairs based on co-substitution between sites in protein structures. *PLoS ONE* **8**, e54252 (2013).

[14] Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766 (2011).

[15] Taylor, W. R., Jones, D. T. & Sadowski, M. I. Protein topology from predicted residue contacts. *Protein Sci.* **21**, 299–305 (2012).

[16] Dessailly, B. H., Redfern, O. C., Cuff, A. & Orengo, C. A. Exploiting structural classifications for function prediction: towards a domain grammar for protein function. *Curr. Opin. Struct. Biol.* **19**, 349–356 (2009).

[17] Hashimoto, K. & Panchenko, A. R. Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proc. Natl. Acad. Sci. USA* **107**, 20352–20357 (2010).

[18] Nishi, H., Koike, R. & Ota, M. Cover and spacer insertions: small nonhydrophobic accessories that assist protein oligomerization. *Proteins* **79**, 2372–2379 (2011).

[19] Lapedes, A. S., Giraud, B. G., Liu, L. & Stormo, G. D. Correlated mutations in models of protein sequences: phylogenetic and structural effects. in *Statistics in Molecular Biology and Genetics.* pp. 236–256 (Institute of Mathematical Statistics, Hayward, CA, USA, 1999).

[20] Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., *et al.* The pfam protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).

[21] Goldenfeld, N. Lectures on phase transitions and the renormalization group, volume 85 of *Frontiers in physics*. (Addison-Wesley, Reading, Massachusetts, 1992).

[22] Horn, R. A. & Johnson, C. R. Matrix analysis. (Cambridge University Press, Cambridge, UK, 1985).

[23] Plefka, T. Convergence condition of the TAP equation for the infiniterange Ising spin glass model. *J. Phys. A* **15**, 1971–1978 (1982).

[24] Kappen, H. & Rodríguez, F. Boltzmann machine learning using mean field theory and linear response correction. pp. 280–286 (The MIT Press, 1998).

[25] MacKay, D. J. C. Information Theory, Inference, and Learning Algorithms. (Cambridge University Press, Cambridge, UK, 2003).

[26] Ota, M., Isogai, Y. & Nishikawa, K. Structural requirement of highly-conserved residues in globins. *FEBS Lett.* **415**, 129–133 (1997).

[27] Wu, F.-Y. The Potts model. *Rev. Mod. Phys.* **54**, 235268 (1982).

[28] Eidhammer, I., Jonassen, I. & Taylor, W. R. Protein bioinformatics. (Wiley & Sons, Chichester, England, 2004).

[29] Kinjo, A. R. Profile conditional random fields for modeling protein families with structural information. *BIOPHYSICS* **5**, 37–44 (2009).

[30] Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S. I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins* **79**, 1061–1078 (2011).

[31] Gō, N. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **12**, 183–210 (1983).

[32] Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Phys. Rev. E* **87**, 012707 (2013).

[33] Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residueresidue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. USA* **110**, 15674–15679 (2013).