Available online at www.sciencedirect.com
ScienceDirect
Chemical Physics Letters 451 (2008) 132–135
CHEMICAL PHYSICS LETTERS
www.elsevier.com/locate/cplett

On the optimal contact potential of proteins

Akira R. Kinjo [a,*], Sanzo Miyazawa [b]

[a] Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan
[b] Faculty of Technology, Gunma University, Kiryu, Gunma 376-8515, Japan

Received 30 October 2007; in final form 3 December 2007
Available online 8 December 2007

Abstract

We analytically derive the lower bound of the total conformational energy of a protein structure by assuming that the total conformational energy is well approximated by the sum of sequence-dependent pairwise contact energies. The condition for the native structure achieving the lower bound leads to the contact energy matrix that is a scalar multiple of the native contact matrix, i.e., the so-called Gō potential. We also derive spectral relations between contact matrix and energy matrix, and approximations related to one-dimensional protein structures. Implications for protein structure prediction are discussed.

1. Introduction

Proteins' biological functions are made possible by their precise three-dimensional (3D) structures, and each 3D structure is determined by its amino acid sequence through the laws of thermodynamics [1]. Therefore, predicting protein structures from their amino acid sequences is important not only for inferring proteins' biological functions, but also for understanding how 3D structures are encoded in such one-dimensional information as amino acid sequence. The problem of protein structure prediction is naturally cast as an optimization problem where a potential function is minimized. Given an appropriate potential function, conformational optimization should yield the native structure as the unique global minimum conformation of the potential function. Thus, the problem has been traditionally divided into two sub-problems: one is to establish an appropriate potential function [2], and the other is to develop the methods to efficiently search the vast conformational space of a protein [3]. Among various forms of effective energy functions, statistical contact potentials [4,5] have been widely used. In this Letter, we exclusively treat a class of such contact potentials, neglecting other contributions such as electrostatics and local interactions. Accordingly, a protein conformation is represented as a contact matrix in which the (i, j) element is 1 if the residues i and j are in contact in space, otherwise it is 0. Although the contact matrix is a coarse-grained representation of protein conformation, it has been known that the contact matrix contains sufficient information to recover the three-dimensional (native) structure of proteins [6]. It is noted that, for the lattice model of proteins [7], these representations of protein conformation and energy function are exact.

2. Theory

2.1. Lower bound of contact energy

Our fundamental assumption is that the conformational energy of a protein can be somehow expressed in terms of a contact matrix. Now let us assume that the total energy of a protein can be well approximated by the sum of pairwise contact energies between amino acid residues, and that each pairwise contact energy can be decomposed into a sequence-dependent term and a conformation-dependent term. The sequence-dependent term is expressed as a matrix $\mathscr{E}(S) = (\mathscr{E}_{ij})$ which we call the contact energy matrix, or E-matrix for short. Each element $\mathscr{E}_{ij}$ of the

* Corresponding author. Fax: +81 6 6879 8636.
E-mail address: akinjo@protein.osaka-u.ac.jp (A.R. Kinjo).

0009-2614/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.cplett.2007.12.005

E-matrix represents the energy between the residues i and j when they are in contact. This form of the E-matrix is a very general one: Each element, $\mathscr{E}_{ij}$, may depend on the entire sequence, S, or it may depend only on the types of the interacting amino acid residues, i and j, as in the conventional contact potentials. The conformation-dependent term is expressed as another matrix $\Delta(C) = (\Delta_{ij})$ which we call the contact matrix, or C-matrix. Each element Δ_{ij} of the C-matrix assumes a value of either 1 or 0, depending on the residues i and j are in contact or not, respectively. Hence the total energy $E(C,S)$ of a protein of sequence S of N residues and having conformation C is given by

$$E(C,S) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathscr{E}_{ij}(S)\Delta_{ij}(C) \tag{1}$$

$$= \frac{1}{2}[\mathscr{E}(S), \Delta(C)] \tag{2}$$

where $[\cdot,\cdot]$ denotes the Frobenius inner product between two matrices [8,9]. Based on this assumption, we derive the lower bound for the conformational energy and the conditions for the native structure and E-matrix to achieve the bound.

The Frobenius inner product leads to the matrix l_2 norm defined as, for a matrix M, $\|M\| \equiv [M,M]^{1/2} = (\sum_{i,j}M_{ij}^2)^{1/2}$. In the case of C-matrix, since $\Delta_{ij} = 0$ or 1, we have

$$\|\Delta(C)\|^2 = 2N_c(C) \tag{3}$$

where $N_c \equiv (1/2)\sum_{i,j}\Delta_{ij}$ is the total number of contacts. As for any inner products, the Frobenius inner product satisfies the Cauchy–Schwarz inequality ($|[A,B]| \leqslant \|A\|\|B\|$) from which we have

$$[\mathscr{E}, \Delta] \geqslant -\|\mathscr{E}\|\|\Delta\| \tag{4}$$

where the equality holds if and only if

$$\mathscr{E} = \varepsilon\Delta \tag{5}$$

for some scalar $\varepsilon < 0$. Although the inequality (Eq. (4)) holds for any pair of matrices, we now regard it as the lower bound for conformational energy for a given E-matrix. For simplicity, we first consider the energy minimization problem for conformations with $\|\Delta(C)\|$ fixed to the value of the native conformation. It is desirable for the native conformation to satisfy the lower bound and hence its condition Eq. (5). If the native conformation indeed satisfies the condition Eq. (5), then the elements of the E-matrix is either 0 or ε so that only the contacts present in the native conformation are stabilizing. Thus, the native conformation satisfying Eq. (5) is actually a GMEC among any conformations with arbitrary values of $\|\Delta(C)\|$. An E-matrix that satisfies Eq. (5) for the native C-matrix is a kind of the so-called Gō potential [10,11] which has been essential for studying the protein folding problem. At this point, it is still possible that the native structure is not the unique GMEC. For example, if a conformation contains all the native contacts together with some other contacts, this conformation has the same energy as the native conformation. In order for a native conformation to be the unique GMEC, it is required that the total number of contacts of the native conformation is larger than that of any other conformations that contain all the native contacts. From the relation Eq. (3), maximizing the total number of contacts is equivalent to maximizing the norm of the C-matrix, which in turn implies the minimization of the right-hand side of Eq. (4). To summarize, for a given E-matrix, $\mathscr{E}(S)$, of a protein, its native conformation, C_n, achieves the lower bound in Eq. (4) if and only if $\mathscr{E}(S) = \varepsilon\Delta(C_n)$ for some $\varepsilon < 0$, and such native structure is the unique GMEC if and only if $\|\Delta(C_n)\|$ is the maximum of all possible conformations that contain all the native contacts. Note that the former condition is a relation between E-matrix and C-matrix whereas the latter is a condition for a native structure to satisfy. The magnitude of ε is not specified here, but it should be determined by other factors such as the folding temperature. It should be noted that a native structure can be the unique GMEC without achieving the lower bound of Eq. (4). Such a case is made possible either by the limitation of the conformational space imposed by other steric factors such as chain connectivity or excluded volumes, or by inherent inconsistencies of the E-matrix so that no plausible conformations are allowed to satisfy the lower bounds.

2.2. Spectral relations

To examine more closely how the lower bound can be achieved, we next derive a more generous lower bound in a more restricted case. First, the C-matrix is decomposed as

$$\Delta = \sum_{\alpha=1}^{N}\sigma_\alpha \mathbf{u}_\alpha \mathbf{v}_\alpha^{\mathrm{T}} \tag{6}$$

where σ_α is the αth singular value and $\mathbf{u}_\alpha$ and $\mathbf{v}_\alpha$ are the corresponding left and right singular vectors, respectively. $U = (\mathbf{u}_1, \ldots, \mathbf{u}_N)$ and $V = (\mathbf{v}_1, \ldots, \mathbf{v}_N)$ are orthogonal matrices. The singular components are sorted in decreasing order of the singular values: $\sigma_1 \geqslant \cdots \geqslant \sigma_N (\geqslant 0)$. Since Δ is real symmetric, the singular values are the absolute values of the eigenvalues of Δ, and the singular vectors are such that $\mathbf{u}_\alpha = \pm\mathbf{v}_\alpha$, where the sign corresponds to that of the respective eigenvalue. Next, the E-matrix is decomposed in the same manner as

$$\mathscr{E} = \sum_{\alpha=1}^{N}\tau_\alpha \mathbf{x}_\alpha \mathbf{y}_\alpha^{\mathrm{T}} \tag{7}$$

where τ_α are singular values, and $\mathbf{x}_\alpha$ and $\mathbf{y}_\alpha$ are left and right singular vectors, respectively. Since $\mathscr{E}$ is also real symmetric, the singular components have the same properties as the C-matrix Δ. Noting that $[\mathscr{E}, \Delta] = \mathrm{tr}(\mathscr{E}\Delta^{\mathrm{T}})$, von Neumann's trace theorem [9] leads to the following inequality:

$$[\mathscr{E}, \Delta] \geqslant -\sum_{\alpha=1}^{N}\sigma_\alpha\tau_\alpha \tag{8}$$

where the equality holds if and only if

$$(\mathbf{u}_\alpha^{\mathrm{T}}\mathbf{x}_\beta)(\mathbf{v}_\alpha^{\mathrm{T}}\mathbf{y}_\beta) = -\delta_{\alpha,\beta} \tag{9}$$

for all α and β with non-zero singular values σ_α and τ_β ($\delta_{\alpha,\beta}$ is Kronecker's delta). We now regard this inequality as a lower bound for the conformational energy for a given E-matrix. For a fixed set of the singular values σ_α ($\alpha = 1,\ldots,N$), if and only if there exists such a conformation that satisfies the condition in Eq. (9), then that conformation is the lowest possible energy conformation. Let λ_α and ε_α ($\alpha = 1,\ldots,N$) be the eigenvalues of the C-matrix and E-matrix, respectively, sorted in the decreasing order of their absolute values. Then $\sigma_\alpha = |\lambda_\alpha|$ and $\tau_\alpha = |\varepsilon_\alpha|$ for $\alpha = 1,\ldots,N$, and $\mathbf{u}_\alpha$ and $\mathbf{x}_\alpha$ are the eigenvectors of the corresponding matrices. Thus, in terms of eigenvalues and eigenvectors, the lower bound in Eq. (8) is equal to $\sum_\alpha \lambda_\alpha\varepsilon_\alpha$ with $\lambda_\alpha\varepsilon_\alpha \leqslant 0$ for $\alpha = 1,\ldots,N$. In addition to the condition Eq. (9) for the lower bound of Eq. (8), if Δ and $\mathscr{E}$ are of the same rank, then the numbers of positive, negative, and zero eigenvalues of Δ and $-\mathscr{E}$ are the same and $\mathbf{u}_\alpha = \pm\mathbf{x}_\alpha$. Thus, from Sylvester's law of inertia [8], there exists a real non-singular matrix S such that

$$\mathscr{E} = -S\Delta S^{\mathrm{T}}, \tag{10}$$

i.e., the E-matrix is *congruent to the C-matrix. If the conformation that satisfy the condition Eq. (10) is the native structure, the E-matrix is consistent in the sense that the contributions from all the eigencomponents are stabilizing the native structure ($\lambda_\alpha\varepsilon_\alpha \leqslant 0$). Since the matrix S is non-singular, we can 'predict' the native structure from the E matrix as $\Delta = -S^{-1}\mathscr{E}S^{-\mathrm{T}}$ (if we can construct the appropriate matrix S). At this point, however, the native structure may not be the GMEC since other conformations with a different set of singular values may have lower energies.

In order to compare the energies of conformations with different sets of singular values, we use another inequality [9]

$$-\sum_{\alpha=1}^{N}\sigma_\alpha\tau_\alpha \geqslant -\|\mathscr{E}\|\|\Delta\| \tag{11}$$

where the lower bound is the same as that in Eq. (4). We note that, in terms of singular values, the matrix norms are expressed as $\|\Delta\| = (\sum_\alpha\sigma_\alpha^2)^{1/2}$ and $\|\mathscr{E}\| = (\sum_\alpha\tau_\alpha^2)^{1/2}$. Hence, it is clear that the equality in Eq. (11) holds if and only if, in addition to the condition in Eq. (9), there exists a scalar constant c such that $\tau_\alpha = c\sigma_\alpha$ for all $\alpha = 1,\ldots,N$. These conditions are equivalent to Eq. (5).

2.3. One-dimensional approximations

To connect the present results with previous studies, we next introduce two approximations. First, we consider the case where the E-matrix is well approximated by its principal eigencomponent, that is, $\mathscr{E} \approx \varepsilon_1\mathbf{x}_1\mathbf{x}_1^{\mathrm{T}}$. This approximation is motivated by the eigenvalue analysis of the Miyazawa–Jernigan (MJ) contact potential [4] performed by Li et al. [12], and has been employed by others [13–15]. In this case, the lower bound Eq. (8) is achieved if and only if $\mathbf{x}_1 = \pm\mathbf{u}_1$ and $\varepsilon_1\lambda_1 < 0$. This result was previously derived by Cao et al. [13] who subsequently showed that the vector $\mathbf{x}_1$ constructed by using the components of the principal eigenvector of the MJ contact potential is indeed highly correlated with the principal eigenvector of the native contact matrices [14]. Bastolla et al. [15] obtained a similar result, but they also showed that taking the average of such $\mathbf{x}_1$ over evolutionarily related proteins greatly improved the correlation. Since the rank of the contact matrix is in general not 1, Eq. (10) does not hold and the equality in Eq. (4) cannot be satisfied. Consequently, there are attractive interactions between non-native contacts even when $\mathbf{x}_1 = \mathbf{u}_1$ holds exactly. Nevertheless, Porto et al. [16] have demonstrated that the knowledge of $\mathbf{u}_1$ alone is practically sufficient for reconstructing the native contact matrix of small single-domain proteins. Therefore, construction of effective rank-1 E-matrices is of great interest [17]. Based on the Porto et al.'s result, it is tempting to postulate that the satisfaction of the lower bound by a rank-1 E-matrix is sufficient for the native conformation to be the unique GMEC. At present, however, there is no clear connection between the present formulation (energy minimization) and the Porto et al.'s combinatorial algorithm.

Another approximation is a kind of mean-field approximations in which the matrix element $\mathscr{E}_{ij}$ is replaced by its average over column $\langle\mathscr{E}_{i\cdot}\rangle \equiv \sum_{j=1}^{N}\mathscr{E}_{ij}/N$. Let us define $\mathbf{e} = (\langle\mathscr{E}_{1\cdot}\rangle,\ldots,\langle\mathscr{E}_{N\cdot}\rangle)^{\mathrm{T}}$ and $\mathbf{n} = (n_1,\ldots,n_N)^{\mathrm{T}}$ where $n_i \equiv \sum_{j=1}^{N}\Delta_{ij}$ is the contact number of the ith residue. Then, we have the following approximation and the lower bound:

$$E(C,S) \approx \frac{1}{2}\mathbf{e}^{\mathrm{T}}\mathbf{n} \tag{12}$$

$$\geqslant -\frac{1}{2}\|\mathbf{e}\|\|\mathbf{n}\| \tag{13}$$

where the equality in (13) holds if and only if the column-averaged E-matrix is anti-parallel to the contact number vector, that is, $\mathbf{e} = \varepsilon\mathbf{n}$ for some $\varepsilon < 0$. This lower bound condition is analogous to Eq. (5), and can be regarded as another kind of the Gō potential for one-dimensional protein structure. It has been suggested that contact number vector can significantly constrain the conformational space [18]. Together with other one-dimensional structures, contact number vector is also used for recovering the native structures [19], and can be accurately predicted [20–24]. It has been pointed out that the contact number vector is highly correlated with the principal eigenvector of the C-matrix [16,19], which suggests that this mean-field approximation is qualitatively similar to the principal eigenvector approximation introduced above.

3. Discussion

Using a more restricted, but conventional, form of the E-matrix where each element $\mathscr{E}_{ij}$ depends only on the types

of *i*th and *j*th residues (e.g., the MJ potential), Vendruscolo et al. [25,26] have shown that it is impossible for such *E*-matrices to stabilize all the native structures in a database. The conventional *E*-matrices such as those they studied do not take into account the sequence-dependence beyond a summation of the contributions from residue pairs. In the present study, we assumed a more general form for the *E*-matrix, allowing each element $\mathscr{E}_{ij}$ to depend on the whole amino acid sequence. In practical situations of protein structure prediction, we want to optimize an energy function so that the native conformations of arbitrary proteins achieve the lower bound. Now let us impose this as a requisite for the *E*-matrix. Then, there should exist a function, namely $\mathscr{E}$, that maps each amino acid sequence to the corresponding optimal *E*-matrix, that is, the Gō potential. Thus, the problem of structure prediction becomes a trivial matter. Currently, most efforts for developing energy functions seem to be focused on accurate estimation of a fixed set of parameters for a given functional form [2]. The present analysis suggests that inferring the function $\mathscr{E}$ that can generate the Gō-like *E*-matrices from amino acid sequences is essential if a contact potential is used. The lower bound inequality (Eq. (4)) and its condition for the equality (Eq. (5)) will serve as the guiding principle for inferring such a function. This approach to structure prediction is apparently similar to machine-learning approaches to contact matrix prediction [17,27]. Although conventional machine-learning methods are not directly targeted at the optimization of the form of Eq. (4), their prediction accuracy should be indicative of the possibility for identifying the function $\mathscr{E}$.

In the preceding paragraph, we have assumed the existence of the function $\mathscr{E}$ to construct the optimal contact potential from a given amino acid sequence. What if, however, there is no such function? In fact, the limited success of current contact matrix prediction [28] strongly suggests that this is more likely the case. Such a case implies either that there are proteins for which the lower bound energy cannot be achieved, or that the total energy cannot be sufficiently accurately approximated by Eq. (1). The former case indicates that some proteins are inherently frustrated, but to a good approximation such proteins should be rather exceptional for natural proteins [10,11]. The latter case may indicate that multi-body contact interactions [29] and/or other energy components than contact energies are more important.

In summary, we have shown that the requirement for the native structure to achieve the lower bound naturally leads to the Gō potential and the requirement for such a conformation to be the unique GMEC leads to the native conformation being the most compact one among those containing all the native contacts. These results suggest that protein structure prediction should be possible simply by constructing the optimal energy matrices or that the contact potential alone is not suitable for the problem. Although not yet definitive, the current state of contact prediction [28] as well as recent studies on local interactions [30,31] suggest that the latter may be the case. Nevertheless, the present results may be useful for evaluating the optimality of potential functions in either case.

References

[1] C.B. Anfinsen, Science 181 (1973) 223.
[2] M.-H. Hao, H.A. Scheraga, Curr. Opin. Struct. Biol. 9 (1999) 184.
[3] A. Mitsutake, Y. Sugita, Y. Okamoto, Biopolymers 60 (2001) 96.
[4] S. Miyazawa, R.L. Jernigan, Macromolecules 18 (1985) 534.
[5] S. Miyazawa, R.L. Jernigan, J. Mol. Biol. 256 (1996) 623.
[6] M. Vendruscolo, E. Kussell, E. Domany, Fold. Des. 2 (1997) 295.
[7] H. Taketomi, Y. Ueda, N. Gō, Int. J. Pept. Prot. Res. 7 (1975) 445.
[8] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, UK, 1985.
[9] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, Cambridge, UK, 1991.
[10] N. Gō, Annu. Rev. Biophys. Bioeng. 12 (1983) 183.
[11] S. Takada, Proc. Natl. Acad. Sci. USA 96 (1999) 11698.
[12] H. Li, C. Tang, N.S. Wingreen, Phys. Rev. Lett. 79 (1997) 765.
[13] H.B. Cao, Y. Ihm, C.Z. Wang, M. Su, D. Dobbs, K.M. Ho, Polymers 45 (2004) 687.
[14] H.B. Cao, C.Z. Wang, D. Dobbs, Y. Ihm, K.M. Ho, Phys. Rev. E 74 (2006) 031921.
[15] U. Bastolla, M. Porto, H.E. Roman, M. Vendruscolo, Proteins 58 (2005) 22.
[16] M. Porto, U. Bastolla, H.E. Roman, M. Vendruscolo, Phys. Rev. Lett. 92 (2004) 218101.
[17] A. Vullo, I. Walsh, G. Pollastri, BMC Bioinformatics 7 (2006) 180.
[18] A. Kabakçioğlu, I. Kanter, M. Vendruscolo, E. Domany, Phys. Rev. E 65 (2002) 041904.
[19] A.R. Kinjo, K. Nishikawa, Bioinformatics 21 (2005) 2167, doi:10.1093/bioinformatics/bti330.
[20] A.R. Kinjo, K. Horimoto, K. Nishikawa, Proteins 58 (2005) 158, doi:10.1002/prot.20300.
[21] A.R. Kinjo, K. Nishikawa, Biophysics 1 (2005) 67, doi:10.2142/biophysics.1.67.
[22] Z. Yuan, BMC Bioinformatics 6 (2005) 248.
[23] T. Ishida, S. Nakamura, K. Shimizu, Proteins 64 (2006) 940.
[24] A.R. Kinjo, K. Nishikawa, BMC Bioinformatics 7 (2006) 401.
[25] M. Vendruscolo, E. Domany, J. Chem. Phys. 109 (1998) 11101.
[26] M. Vendruscolo, R. Najmanovich, E. Domany, Proteins 38 (2000) 134.
[27] J. Cheng, P. Baldi, BMC Bioinformatics 8 (2007) 113.
[28] O. Graña et al., Proteins 7 (Suppl.) (2005) 214.
[29] P.J. Munson, R.K. Singh, Protein Sci. 6 (1997) 1467.
[30] G. Chikenji, Y. Fujitsuka, S. Takada, Proc. Natl. Acad. Sci. USA 103 (2006) 3141.
[31] P.J. Fleming, H. Gong, G.D. Rose, Protein Sci. 15 (2006) 1829.