Coarse-Graining Protein Energetics in Sequence Variables

Fei Zhou,[1] Gevorg Grigoryan,[2] Steve R. Lustig,[3] Amy E. Keating,[2] Gerbrand Ceder,[4] and Dane Morgan[5]

[1]*Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[2]*Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[3]*DuPont Central Research and Development, Experimental Station, Wilmington, Delaware 19880, USA*
[4]*Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[5]*Department of Materials Science and Engineering, University of Wisconsin, Madison, Wisconsin 53706, USA*

(Received 17 March 2005; published 29 September 2005)

We show that cluster expansions (CE), previously used to model solid-state materials with binary or ternary configurational disorder, can be extended to the protein design problem. We present a generalized CE framework, in which properties such as energy can be unambiguously expanded in the amino-acid sequence space. The CE coarse grains over nonsequence degrees of freedom (e.g., side-chain conformations) and thereby simplifies the problem of designing proteins, or predicting the compatibility of a sequence with a given structure, by many orders of magnitude. The CE is physically transparent, and can be evaluated through linear regression on the energies of training sequences. We show, as example, that good prediction accuracy is obtained with up to pairwise interactions for a coiled-coil backbone, and that triplet interactions are important in the energetics of a more globular zinc-finger backbone.

DOI: 10.1103/PhysRevLett.95.148103 PACS numbers: 87.14.Ee, 87.15.Aa

Protein folding and protein design stand among the most formidable challenges in contemporary computational biology. The 3D structure of a protein is uniquely encoded in its 1D sequence of amino acids (AA), and enormous theoretical and computational research effort has been devoted to understanding this encoding [1–3]. The problem can be posed two ways: protein *folding* deals with predicting the final 3D structure of a protein given its AA sequence, whereas protein *design* is concerned with finding an optimal sequence to fold to a predefined structure. Protein design is useful both because it allows for the engineering of macromolecules with desired properties [4], and because the development of computational design methods deepens our general understanding of protein folding and stability. Scoring functions that indicate the ability of sequences to fold to any given structure are central to both the folding and design problems. These range from statistical knowledge-based functions derived from databases of known protein structures [5] to empirical functions mainly based on experimental measurements [6], to more physics-based functions that attempt to model protein free energy [6,7].

Physics-based energy functions have the potential of being the most accurate and interpretable. These express the energy of a protein sequence adopting a specified structure in terms of atomic coordinates, and account for energies arising from van der Waals (vdW) forces, electrostatics, and solvation. All atoms in a protein can be classified as either ''backbone'' or ''side chain.'' The backbone atoms are the same for each AA and represent the overall structure or ''fold'' of a protein, as shown for two examples in Fig. 1(a). The side-chain atoms are different for different AAs, and give rise to additional degrees of freedom termed ''side-chain conformations'' or ''rotamers'' [see Figs. 1(b) and 1(c)]. Even for a relatively small protein fold of 100 AAs there are roughly 10^{130} possible sequences. Accounting for a set of common rotamers expands the search space to $\sim 10^{230}$ structures. The computational complexity of high-quality physics-based scoring functions makes optimization over all sequences and rotamers infeasible. Because sequence determines the structure of a protein, a function should exist that maps sequence directly to energy. A sufficiently accurate and computationally tractable approximation of this function would find widespread use in computational studies of protein structure.

Mapping sequence to energy is similar to the configurational problem in alloy theory [8–10] where distributions of A and B atoms on a fixed topology of lattice sites specify the energy [11]. The cluster expansion (CE) [8,9] has proven extremely useful for rapidly expanding the energies of alloys and searching for low-energy configurations. In this Letter, we apply the CE to the protein design problem, deriving two structure-specific functions that can determine the energies of a sequence adopting either a coiled-coil or a zinc-finger geometry. Searches using these functions can be used in the future to identify low-energy sequences that adopt these folds. Further, CE can potentially be applied directly to the more challenging protein folding problem by deriving a function specific to each of the $\sim$1000 known protein folds. Rapid evaluation of a sequence with the full panel of functions could identify the best structure. This approach, termed ''threading'' or ''fold recognition,'' is widely used for structure prediction in combination with statistically derived energy functions.

While in alloys one typically treats binary distributions (two possible species per site) or on rare occasions ternaries [12,13], the general protein design problem requires

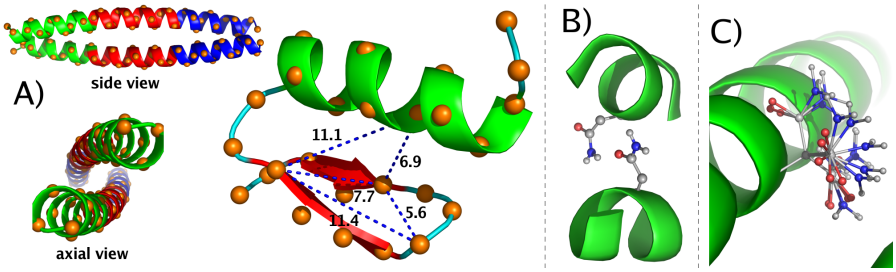

FIG. 1 (color online). (a) The coiled coil (left) and the zinc-finger (right) protein folds. Orange spheres are backbone atoms and the ribbons are a cartoon representation of the backbone geometry. The coiled-coil unit cells are highlighted. (b) The optimal rotamers for two AA's in an all-atom representation. (c) A set of common rotamers for one AA shown superimposed.

extension to all 20 possible AAs. For a protein of L residues let the variable $\sigma^i = 1 \ldots m$ indicate which of the m AAs is present at site i. A sequence is then expressed by $\vec{\sigma} = \{\sigma^1, \ldots, \sigma^L\}$. The energy of a protein $E[\vec{\sigma}, \vec{\tau}]$ depends on this sequence and on the other microscopic information $\vec{\tau}$ (e.g., positions of all atoms on the protein and solvent molecules). The important energy function in protein design, $E_{\min}[\vec{\sigma}]$, can be obtained by optimizing over $\vec{\tau}$:

$$E_{\min}[\vec{\sigma}] = \min_{\vec{\tau}} E[\vec{\sigma}, \vec{\tau}]. \tag{1}$$

The CE is a general approach to obtain $E_{\min}$ by expanding in a suitable set of independent basis functions. Let $i, j, k = 1 \ldots L$ denote sites and $\alpha, \beta, \gamma = 1 \ldots m - 1$ basis function indices. Because the point basis $\{1, \phi^i_1, \ldots, \phi^i_{m-1}\}$ fully describes the AA space at site i, the polynomials of these functions, $\{1, \phi^i_\alpha, \phi^i_\alpha \phi^j_\beta, \phi^i_\alpha \phi^j_\beta \phi^k_\gamma, \ldots\}$, form a complete basis in which to expand the properties of the sequence. The energy can be expanded as:

$$E_{\min}[\vec{\sigma}] = J_\varnothing + \sum_{i,\alpha} J^i_\alpha \phi^i_\alpha(\sigma^i) + \sum_{ij,\alpha\beta} J^{ij}_{\alpha\beta} \phi^i_\alpha(\sigma^i) \phi^j_\beta(\sigma^j) + \sum_{ijk,\alpha\beta\gamma} J^{ijk}_{\alpha\beta\gamma} \phi^i_\alpha(\sigma^i) \phi^j_\beta(\sigma^j) \phi^k_\gamma(\sigma^k) + \cdots, \tag{2}$$

where the J's are expansion coefficients. We leave it to a future paper to describe the mathematical properties of this basis set [14]. Equation (2) is in principle exact, though in practice the expansion has to be truncated. While the J's depend on the choice of basis functions, the sum over terms spanning a cluster of sites $\{i, \ldots, j\}$ has a physical interpretation, and can be defined as the effective interaction (EI) between the AA's on these sites:

$$\mathrm{EI}\,(\sigma^i \ldots \sigma^j) = \sum_{\alpha \ldots \beta} J^{i \ldots j}_{\alpha \ldots \beta} \phi^i_\alpha(\sigma^i) \ldots \phi^j_\beta(\sigma^j). \tag{3}$$

The choice of point basis functions ϕ_α is in principle arbitrary though we have found that previously proposed basis functions [8] have poor numerical stability for the high dimensional configuration spaces of proteins and make the expansion converge slowly. In this Letter we use $\phi_\alpha(\sigma) = \delta(\sigma - \alpha)$. Hence $\phi_\alpha(m) \equiv 0$ and the hypothetical sequence $\{m, \ldots, m\}$ has energy J_0. If we assign m to Alanine (Ala) any point $\mathrm{EI}\,(\sigma^i)$ equals the energetic contribution of σ^i relative to Ala. Therefore, pair $\mathrm{EI}\,(\sigma^i, \sigma^j)$ is the interaction of an AA pair, a measure well known to biochemists [15]. This concept can be taken beyond pairs—contributions purely from triplets can be measured similarly. Although this is difficult to do experimentally, the CE allows one to systematically analyze the importance of higher order interactions.

Given $E_{\min}$ for enough sequences, Js can be extracted by standard fitting procedures. Determining which Js to keep in the fit is not always obvious. While one may be guided by the idea that point terms are larger than pairs, which in turn are larger than triplets, this is not always true. We use a more systematic way for evaluating important Js based on the cross-validation (CV) score [16]. Essentially, the CV score is the average error with which each sequence is predicted when left out of the fitting, and as such is a good measure of the prediction power. Our procedure consists of fitting a selected set of candidate clusters and ordering them by the average $|J|$. Clusters for which the J value largely arises from numerical noise increase the CV score, and are excluded. When a cluster is included, so are all of its subclusters.

We demonstrate the power of the CE by testing it on two different protein folds, mimicking the protein design problem. The folding energy is defined as the energy difference between the folded and the unfolded states: $E_{\text{folding}} = E_{\text{folded}} - E_{\text{unfolded}}$. Although the CE can in principle be used with any energy model, we test it here with a physically meaningful but relatively simple expression similar to Hamiltonians commonly used in the design field [17]:

$$E[\vec{\sigma}, \vec{\tau}] = E^{\text{vdW}} + E^{\text{elec,wat}} + E^{\text{solv,sc}} + E^{\text{torsion}}, \tag{4}$$

where E^{vdW} is the vdW interaction modeled as a 6–12 Lennard-Jones potential, $E^{\text{elec,wat}}$ is the total electrostatic energy (excluding intra-side-chain interactions), $E^{\text{solv,sc}}$ is the solvation energy of all backbone and side-chain atoms [18], and E^{torsion} is the side-chain torsional energy. All energy terms are calculated using the CHARMM package

[19] with the *param19* parameters. The unfolded state is modeled by retaining only side-chain self-energies and local interactions between side chains and their surrounding penta-peptide backbone. Because $E[\vec{\sigma}, \vec{\tau}]$ in Eq. (4) is pairwise decomposable, we are able to apply the dead-end-elimination (DEE) algorithm [20] followed by a branch-and-bound search to arrive at the optimal rotamers corresponding to $E_{\min}$. Thus, in a CE derived from these $E_{\min}$, the Js and EIs parametrize optimized energies whereby all the side-chain degrees of freedom are coarse grained out. The EI, defined at the sequence level, may include higher order terms even though the initial energy expressions at the conformational level are pairwise decomposable. The advantage of this procedure is an enormous reduction in the search space, from $(20m)^L$ to m^L, where 20 is the average number of rotamers considered per AA.

In order to more accurately fit the important low energies, our fitting is weighted by $\max(e^{-(E-E_0)/K}, w_0)$, where E_0 is the lowest energy in the data set, K is approximately the range of interest above E_0, and w_0 is the minimal weight at large E to avoid numeric instability.

Our first case study involves the coiled coil, a common and well-characterized protein interaction interface [Fig. 1(a)]. An ideal coiled-coil backbone possesses a screw axis with a repeating unit every 7 residues (a heptad) as well as C2 symmetry about the coil axis [21]. Since a coiled-coil dimer can be of arbitrary length, we defined a unit cell as a fragment of 4 heptads [highlighted in Fig. 1(a)] and modeled it surrounded by unit cells with identical sequence to avoid end effects. The energy of the central unit cell plus half of its interaction with the rest of the system is calculated. Only 4 sites in each heptad are each modeled as one of 16 selected AA species (the 3 remaining sites are set to Ala). These 4 sites have been shown, in many cases, to be sufficient to determine coiled-coil dimerization preferences and other properties [22].

Our training set consists of 21 066 randomly chosen sequences weighted by $\max(e^{-(E+26)/120}, 0.01)$. Truncating the CE at the pair level is sufficient to accurately reproduce the energetics of the system. The structural symmetry reduces all 137 clusters up to pairs to 1 constant, 4 point, and 36 pair-level independent cluster (7741 independent Js). We are therefore able to include all of them as candidate clusters in the fitting. Figure 2 shows the weighted rms and CV scores of the least square fitting versus the number of included clusters (ordered by $\langle|J|\rangle$). Although the rms decreases monotonically as expected, the CV score reaches a minimum at 22 clusters, and fluctuates (mostly increases) slightly afterwards. We thus come to an "optimal" set of 22 clusters (3676 Js) for energy prediction, with weighted rms $= 1.0$ kcal/mol and CV $= 1.1$ kcal/mol. The most significant EIs are found to correspond to residues that mediate contacts between different helices, in agreement with biologists' intuition about the system.

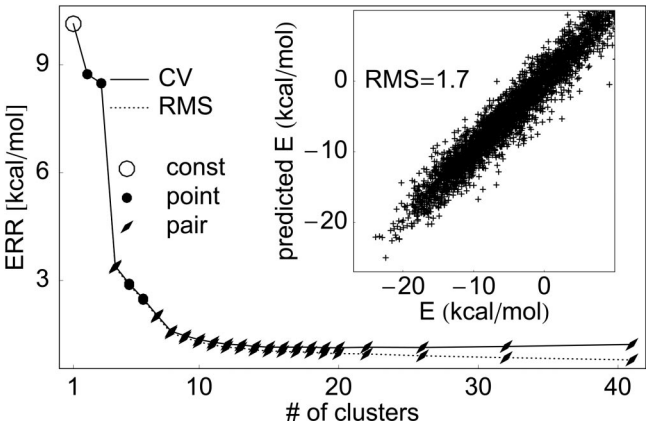

FIG. 2. rms and CV scores vs number of clusters included for coiled-coil fitting. Inset: CE predicted vs atomistic $E_{\min}$ for 3995 random sequences (only $E_{\min} \leq 10$ kcal/mol shown).

To test the predictive character of the CE we compare its prediction for 3995 random sequences not included in training to the directly calculated energy (Fig. 2 inset). The unweighted rms error is 2.4 kcal/mol for all energies and 1.7 kcal/mol for $-26 < E_{\min} < 10$ kcal/mol. The error is sufficiently small for such applications as sequence optimization, and is comparable with the accuracy of the underlying energy model. We trade such a small error for being able to predict the optimal energy of any sequence by summation of EIs for 22 clusters, as opposed to performing global optimization over an average of 5.9×10^{55} states. Even compared to the highly efficient DEE method for side-chain positioning, the time to calculate $E_{\min}$ of a sequence is reduced from $\sim$200 sec to ~ 1 μs with our coarse-grained Hamiltonian, a 2×10^8-fold acceleration.

As a second application we consider the zinc finger, a common DNA-binding fold [Fig. 1(a)]. The backbone of Zif268 (PDB ID 1ZAA) residues 33–60 is used as a model Zn-finger structure. Following Mayo *et al.* [23], we consider a sequence space in which 2 sites are fixed, 1 site has 7 candidate species, 18 sites have 10, and the other 7 sites have 16. The training set consists of 29 864 random sequences weighted by $\max(e^{-(E+35)/100}, 0.01)$. Because there are too many pairs (325 pairs, or 4×10^4 Js) to easily include in one single fitting, we start with constant and point terms and add pairs one by one to the existing clusters, retaining a pair if it decreases CV. We iterate until no new pair can be selected. However, truncation at pairs leads to an unsatisfactory fitting with CV > 6 kcal/mol. Instead of trying all 2600 triplets, we use an information theory based approach to determine two particularly important triplet clusters [14]. These triplets have significant 3-body EIs when the 4 constituent sites [see Fig. 1(a)], located in proximity to each other, are occupied by aromatic side chains W, H, Y, and F. We end up with one constant, 26 point, 24 pair, and 2 triplet clusters (5692 Js in total) for fitting. The rms and CV scores versus the number

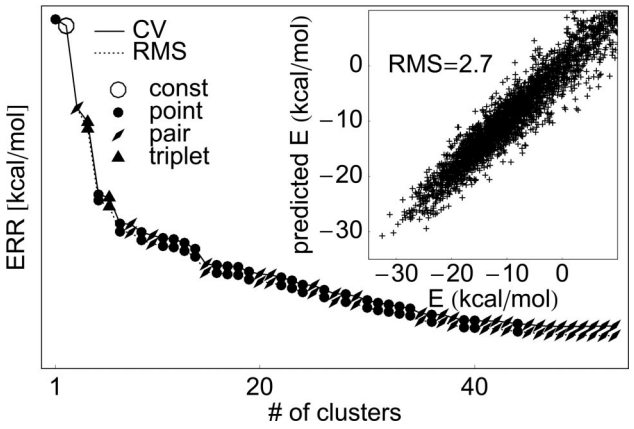

FIG. 3. rms and CV scores for Zn-finger fitting. Inset: CE predicted vs atomistic $E_{\min}$ for 4000 random sequences (only $E_{\min} \leq 10$ kcal/mol shown).

of clusters included are shown in Fig. 3. The two triplets are found to be indispensable in correctly reproducing the energies. This demonstrates the existence of complex correlations in a globular protein, and the CE provides a systematic, quantitative way of identifying such correlated sites. Prediction of 4000 random Zn-finger sequences is shown in Fig. 3 inset. Again a reasonably good accuracy of 2.7 kcal/mol for $-35 < E_{\min} < 10$ kcal/mol is obtained. Although a larger prediction error 15.4 kcal/mol is obtained with all energies, high energy sequences are correctly detected. Such error is traded for a remarkable reduction in search space: from 1.4×10^{60} to 1.9×10^{27} states.

In summary, we have demonstrated how the energetics of a protein with predefined backbone can be coarse grained to a function of sequence only. We have successfully applied the method to two distinct families of proteins, and found that two different types of interactions are important for representing the energy. The accuracy of the CE predictions, which can be systematically improved, implies that this much simpler expression can be used in place of traditional Hamiltonians, dramatically improving computational efficiency.

The CE methodology can be coupled with any energy model, e.g., more accurate Hamiltonians or experimentally determined energies, and properties other than energy are potentially expandable. Thus, it can be extended to treat any multispecies search problem for which an appropriate scoring scheme can be generated. In structural biology, this includes modeling not only protein stability, but protein interaction specificity, DNA and RNA structure, protein-DNA interactions, and potentially the interactions of small-molecule pharmaceuticals. We are optimistic that the method will find a wide range of practical applications in biology research.

This work is supported by funding from the DuPont-MIT Alliance to G. C. and NIH Grant No. GM67681 to A. K. F. Z. thanks M. Kardar for critical reading of the manuscript.

[1] C. M. Kraemer-Pecore, A. M. Wollacott, and J. R. Desjarlais, Curr. Opin. Chem. Biol. **5**, 690 (2001).
[2] N. Pokala and T. M. Handel, J. Struct. Biol. **134**, 269 (2001).
[3] M. Vendruscolo and E. Paci, Curr. Opin. Struct. Biol. **13**, 82 (2003).
[4] M. A. Dwyer, L. L. Looger, and H. W. Hellinga, Science **304**, 1967 (2004); M. Allert, S. S. Rizk, L. L. Looger, and H. W. Hellinga, Proc. Natl. Acad. Sci. U.S.A. **101**, 7907 (2004); B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, Science **302**, 1364 (2003).
[5] W. P. Russ and R. Ranganathan, Curr. Opin. Struct. Biol. **12**, 447 (2002).
[6] R. Guerois and L. Serrano, Curr. Opin. Struct. Biol. **11**, 101 (2001).
[7] J. Mendes, R. Guerois, and L. Serrano, Curr. Opin. Struct. Biol. **12**, 441 (2002).
[8] J. M. Sanchez, F. Ducastelle, and D. Gratias, Physica A (Amsterdam) **128**, 334 (1984).
[9] D. de Fontaine, *Solid State Physics* (Academic, New York, 1994), Vol. 47.
[10] G. Ceder, Comput. Mater. Sci. **1**, 144 (1993).
[11] G. D. Garbulsky and G. Ceder, Phys. Rev. B **51**, 67 (1995); V. Blum and A. Zunger, Phys. Rev. B **69**, 020103(R) (2004); C. Wolverton, X. Y. Yan, R. Vijayaraghavan, and V. Ozolins, Acta Mater. **50**, 2187 (2002); G. Ceder, P. D. Tepesch, A. F. Kohan, and A. Van der Ven, J. Electroceram. **1**, 15 (1997).
[12] J. D. Althoff, D. D. Johnson, and F. J. Pinski, Phys. Rev. Lett. **74**, 138 (1995).
[13] R. McCormack, D. de Fontaine, C. Wolverton, and G. Ceder, Phys. Rev. B **51**, 15808 (1995).
[14] F. Zhou *et al.* (to be published).
[15] D. Krylov, I. Mikhailenko, and C. Vinson, EMBO J. **13**, 2849 (1994); A. Acharya, S. B. Ruvinov, J. Gal, J. R. Moll, and C. Vinson, Biochemistry **41**, 14122 (2002).
[16] A. van de Walle and G. Ceder, J. Phase Equilib. **23**, 348 (2002).
[17] M. H. Ali, C. M. Taylor, G. Grigoryan, K. N. Allen, B. Imperiali, and A. E. Keating, Structure **13**, 225 (2005).
[18] T. Lazaridis and M. Karplus, Proteins: Struct., Funct., Genet. **35**, 133 (1999).
[19] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[20] J. Desmet, M. Demaeyer, B. Hazes, and I. Lasters, Nature (London) **356**, 539 (1992); R. F. Goldstein, Biophys. J. **66**, 1335 (1994).
[21] F. H. Crick, Acta Crystallogr. **6**, 689 (1953).
[22] E. K. Oshea, R. Rutkowski, and P. S. Kim, Cell **68**, 699 (1992); J. H. Fong, A. E. Keating, and M. Singh, Genome Biol. **5**, R11 (2004).
[23] B. I. Dahiyat and S. L. Mayo, Science **278**, 82 (1997).