A many-body term improves the accuracy of effective potentials based on protein coevolutionary data

A. Contini and G. Tiana

Articles you may be interested in

Essential roles of protein-solvent many-body correlation in solvent-entropy effect on protein folding and denaturation: Comparison between hard-sphere solvent and water
J. Chem. Phys. **142**, 145103 (2015); 10.1063/1.4917075

Myths and verities in protein folding theories: From Frank and Evans iceberg-conjecture to explanation of the hydrophobic effect
J. Chem. Phys. **139**, 165105 (2013); 10.1063/1.4827086

Potential of mean force between a large solute and a biomolecular complex: A model analysis on protein flux through chaperonin system
J. Chem. Phys. **135**, 185101 (2011); 10.1063/1.3657856

Dewetting-induced globule-coil transitions of model polymers and possible implications high-temperature and low-pressure unfolding of proteins
J. Chem. Phys. **132**, 165101 (2010); 10.1063/1.3394864

A many-body model to study proteins. II. Incidence of many-body polarization effects on the interaction of the calmodulin protein with four Ca 2+ dications and with a target enzyme peptide
J. Chem. Phys. **119**, 1874 (2003); 10.1063/1.1579479
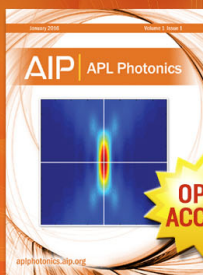
A many-body term improves the accuracy of effective potentials based on protein coevolutionary data

A. Contini[1] and G. Tiana[2,a)]
[1]*Department of Physics, Università degli Studi di Milano, via Celoria 16, 20133 Milano, Italy*
[2]*Department of Physics, Università degli Studi di Milano, and INFN, via Celoria 16, 20133 Milano, Italy*

The study of correlated mutations in alignments of homologous proteins proved to be successful not only in the prediction of their native conformation but also in the development of a two-body effective potential between pairs of amino acids. In the present work, we extend the effective potential, introducing a many-body term based on the same theoretical framework, making use of a principle of maximum entropy. The extended potential performs better than the two-body one in predicting the energetic effect of 308 mutations in 14 proteins (including membrane proteins). The average value of the parameters of the many-body term correlates with the degree of hydrophobicity of the corresponding residues, suggesting that this term partly reflects the effect of the solvent. © *2015 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4926665]

I. INTRODUCTION

The availability of simplified protein models with reduced degrees of freedom is useful for studying several biophysics problems. For example, the study of conformational changes in large protein systems is still unfeasible even on the fastest computers.[1] Conversely, with a reduced model it could be possible to study the thermodynamics of a 341-residues protein in a crowded environment.[2] Free-energy differences upon mutation can be calculated *ab initio* only for small systems, while in more challenging cases one must resort to *ad-hoc* potentials.[3] The elimination of solvent molecules is a standard example in which the use of a simplified model allows to study large and complex systems.[4] Anyway, the main problem associated with the reduction of the number of degrees of freedom in physical systems is the design of an effective potential which depends on them.

A way which has been followed several times to obtain effective potentials for proteins is the statistical approach.[5–7] The input data are the distribution of residues–residues contacts between the different types of amino acids in a selected set of proteins. Based on these, one has to solve an inverse statistical-mechanics problem. The goal is to find the potential which generated, during natural evolution, the observed frequencies of contacts as equilibrium distribution, assuming a Boltzmann relation between contact frequency and contact energy.[8–10]

A variation of this approach is the calculation of contact energies based on the observed correlations between mutations in homologous proteins, using the same framework as that described in Ref. 11 for a different problem, namely, that of predicting the native conformation of a protein from sequence information only. Here, pairs of residues which mutate in a correlated way in homologous sequence are regarded as in spatial contact, and from the full set of spatial contacts, it could

be possible to reconstruct the three-dimensional structure of several proteins. An inverse Ising-model formalism was used to subtract the effect of indirect correlations from the experimental data.

The same formalism was then used in Ref. 12 to design an effective, non-portable two-body contact potential, assuming that the native conformation of the protein is known. This potential proved successful in back-calculating residue–residue interactions in families of proteins generated by simulated evolution. It was also used to calculate the thermodynamic effect of mutations in four well-known proteins, giving correlation coefficients ranging between 0.65 and 0.89 between the experimental and the calculated $\Delta\Delta G$.

The formalism at the basis of Refs. 11 and 12 is meant to find the numerical values of the parameters of the effective energy,

$$\mathcal{U}(\{\sigma_i\}) = \sum_{i<j} e_{ij}(\sigma_i,\sigma_j)\Delta(|r_i - r_j|) + \sum_i h_i(\sigma_i), \quad (1)$$

from the knowledge of the observed frequencies $f_i(\sigma)$ of appearance of amino acid σ at site i and of the observed correlations $f_{ij}(\sigma,\tau)$ obtained in a set of M aligned homologous sequences of length L. In Eq. (1), σ_i is the type of residue at position i of the protein, $\Delta(|r_i - r_j|)$ is a contact function which takes the value 1 if residues i and j are close in space (i.e., they contain a pair of heavy atom closer than a distance d_r) and zero otherwise, $e_{ij}(\sigma_i,\sigma_j)$ is the interaction energy between residues σ_i at position i and σ_j at position j, and $h_i(\sigma_i)$ is a one-body potential acting on each residue.

Once the numerical parameters entering Eq. (1) are calculated, the two-body energy $U(\{r_i\}) = \sum_{i<j} e_{ij}(\sigma_i,\sigma_j)\Delta(|r_i - r_j|)$ can be applied for describing the conformational space of the protein. In Ref. 12, for example, besides the calculation of mutational $\Delta\Delta G$, it was used to identify the frustrated regions of the protein. In doing so, the fields $h_i(\sigma)$ were regarded just as chemical potential meant to fix the average concentration

a)Electronic mail: guido.tiana@unimi.it

of the twenty types of amino acids. Consequently, they were considered relevant only to control the underlying evolution of the set of homologous proteins, but not for the characterization of the conformational space of a well-defined sequence, of fixed amino-acid composition. Hence, they were neglected in the calculation of the $\Delta\Delta G$.

However, one can think that the fields $h_i(\sigma)$ contain not only a chemical potential but also a real interaction contribution associated with the position of a specific amino acid within the native conformation of the protein, not encoded in the two-body terms $e_{ij}(\sigma, \tau)$, and thus controlled by the $f_i(\sigma)$ rather than by the $f_{ij}(\sigma, \tau)$. This could be the case, for example, of the hydrophobic interaction, which depends, in first approximation, on the degree of burial of the ith site into the protein conformation, and not on the sum of two-body terms.

In the present work, we want to disentangle the contribution to the potential which can be interpreted as an interaction term, from the one which is purely a chemical potential. We show that evolution of protein sequences onto a (fixed) native conformation can be described by an effective energy of the form

$$\mathcal{U}(\{\sigma_i\}) = \sum_{i<j} e_{ij}(\sigma_i, \sigma_j)\Delta(|r_i - r_j|)$$
$$+ \sum_i \eta_i(\sigma_i) + \sum_i \mu(\sigma_i), \qquad (2)$$

where $\eta_i(\sigma_i)$ is the associated energy and $\mu(\sigma_i)$ is the chemical potential. We regard the first two terms as an effective interaction potential

$$U(\{r_i\}) = \sum_{i<j} e_{ij}(\sigma_i, \sigma_j)\Delta(|r_i - r_j|) + \sum_i \eta_i(\sigma_i)\Theta_i(\{r_i\}), \qquad (3)$$

assigning a conformational dependence to its second term through a function $\Theta_i(\{r_i\})$ which measures the solvent-exposure of the ith residue. We show that this effective potential predicts the experimental $\Delta\Delta G$ better than what the model involving only the two-body terms did.[12]

II. DERIVATION OF THE POTENTIAL

Given an alignment of M homologous sequences, the input of the model is, as in the case of Ref. 11, the frequency $f_i(\sigma)$ of the amino acid of type σ at site i and the frequency $f_{ij}(\sigma, \tau)$ of the pair of types σ and τ at sites i and j, respectively, reweighted by the appropriate pseudocounts[13] as

$$f_i(\sigma) = \frac{1}{M_e(x + y + z + 1)}$$
$$\times \left[\tilde{f}_i(\sigma) + x\frac{M_e}{q} + y\frac{\sum_j \tilde{f}_j(\sigma)}{L} + z\tilde{f}_i(\sigma)\right],$$
$$f_{ij}(\sigma, \tau) = \frac{1}{M_e(x + y + z + 1)}\left[\tilde{f}_{ij}(\sigma, \tau) + x\frac{M_e}{q^2}\right.$$
$$\left. + \frac{y}{L^2 M_e}\sum_{kl}\tilde{f}_k(\sigma)\tilde{f}_l(\tau) + \frac{z}{M_e}\tilde{f}_i(\sigma)\tilde{f}_j(\tau)\right]. \qquad (4)$$

Here, $\tilde{f}_i(\sigma) \equiv \sum_s \delta(\sigma, \sigma_i^s)/m_s$ and $\tilde{f}_{ij}(\sigma, \tau) \equiv \sum_s \delta(\sigma, \sigma_i^s)\delta(\tau, \sigma_j^s)/m_s$ are the raw frequencies, where the index s runs over the sequences of the alignment, m_s is the number of sequences with similarity larger than 70% to sequence s, q is the number of residue types, and $M_e = \sum_s 1/m_s$ is an effective number of sequences. The parameters x, y, and z weight the pseudocounts depending, respectively, on the overall fraction of residue types, on the overall fraction of residue types in the specific alignment, and on the overall fraction of residue types in the specific pair of positions. These are *a priori* probabilities which complement the empirical frequencies of contacts when the statistics is poor (cf. Ref. 12 for a more complete discussion).

We shall search for a potential to generate a global distribution $p(\{\sigma_i\})$ for residue types in all the positions of the alignment that matches the empirical distributions. In particular, we shall require that

$$P(\tau) \equiv \sum_{\{\sigma_k\}} p(\{\sigma_i\})\sum_{i=1}^{L}\frac{\delta(\sigma_i, \tau)}{L} = \frac{1}{L}\sum_i f_i(\tau),$$
$$\Delta P_i(\tau) \equiv \sum_{\{\sigma_k\}} p(\{\sigma_i\})\left[\delta(\sigma_i, \tau) - \sum_{j=1}^{L}\frac{\delta(\sigma_j, \tau)}{L}\right]$$
$$= f_i(\tau) - \frac{1}{L}\sum_j f_j(\tau), \qquad (5)$$
$$P_{ij}(\tau, \rho) \equiv \sum_{\{\sigma_k\}} p(\{\sigma_k\})\delta(\sigma_i, \tau)\delta(\sigma_j, \rho) = f_{ij}(\tau, \rho).$$

The quantity $P(\tau)$ is the overall probability to find amino acid of type τ in any site, while $\Delta P_i(\tau)$ is the difference between the probability in a specific site and the overall one, defined in such a way to be uncorrelated to $P(\tau)$. We also define the connected correlation function $C_{ij}(\tau, \rho) = f_{ij}(\tau, \rho) - f_i(\tau)f_j(\rho)$.

Since we have no other knowledge of the potential but the frequencies defined above, it seems reasonable to use the principle of maximum entropy with the constrains given by Eq. (5) and the normalization condition of $p(\{\sigma_i\})$. Maximizing the entropy, we obtain

$$p(\{\sigma_i\}) = \frac{1}{Z}\exp\left[-\sum_{i<j} e_{ij}(\sigma_i, \sigma_j) - \sum_{i=1}^{L}\mu(\sigma_i)\right.$$
$$\left. - \sum_{i=1}^{L}\left(\tilde{h}_i(\sigma_i) - \frac{1}{L}\sum_{j=1}^{L}\tilde{h}_i(\sigma_j)\right)\right], \qquad (6)$$

where the quantities $e_{ij}(\sigma, \tau)$, $\tilde{h}_i(\sigma)$, and $\mu(\sigma)$ are Lagrange multipliers. Due to the formal similarity with Boltzmann's distribution, we regard these quantities as effective energies. In particular, μ is site-independent and we assign to it the meaning of chemical potential.

Assuming that there are q types of amino acids, Eq. (6) contains $q + Lq + q^2L(L - 1)/2$ parameters. The experimental input of Eq. (5) consists of $(q - 1) + (L - 1)(q - 1) + (q - 1)^2L(L - 1)/2$ independent equations. Consequently, one has $1 + (L + q - 1) + (2q - 1)L(L - 1)/2$ free parameters which can be used to set the zeros of the energies. We must thus choose some $\bar{\sigma}$, $\tilde{\sigma}$, and σ^* such that

$$\mu(\bar{\sigma}) = 0,$$
$$\widetilde{h}_i(\widetilde{\sigma}) = 0 \quad \forall i,$$
$$\widetilde{h}_{\widetilde{i}}(\sigma) = 0 \quad \forall \sigma, \qquad (7)$$
$$e_{ij}(\sigma^*, \sigma) = e_{ij}(\sigma, \sigma^*) = 0 \quad \forall i, j, \sigma.$$

In other words, one has to choose an amino-acid type $\bar{\sigma}$ as the zero of the chemical potential, a type $\widetilde{\sigma}$ as the zero for the field $\widetilde{h}_i$ in each site (which in principle could be different from site to site), and a site $\widetilde{i}$ (the *reference site*) in which the field $\widetilde{h}_{\widetilde{i}}(\sigma) = 0$ for any type of amino acid. The residue σ^* sets the zero of the two-body interaction.

For the purpose of determining the numerical values of the fields $\widetilde{h}$ and of the chemical potentials μ in Eq. (6), we follow the spirit of Ref. 11 and write the argument of its exponential as an effective energy,

$$\mathcal{U}_\alpha = \alpha \sum_{i<j} e_{ij}(\sigma_i, \sigma_j) + \sum_{i=1}^{L} \left[\mu(\sigma_i) + \widetilde{h}_i(\sigma_i) - \sum_{j=1}^{L} \frac{\widetilde{h}_i(\sigma_j)}{L} \right], \qquad (8)$$

depending on the parameter α which controls the ratio between the two-body energy and the other energy terms. The associated Helmoltz free energy is

$$\mathcal{F}_\alpha = -\ln(\mathcal{Z}) = \langle \mathcal{U}_\alpha \rangle - S, \qquad (9)$$

where temperature is immaterial in this derivation and is set to 1. The Gibbs free energy, obtained by a Legendre transform over the independent variables, is

$$\mathcal{G}_\alpha = \mathcal{F}_\alpha - L \sum_{\sigma=1}^{q-1} \mu(\sigma) \frac{\partial[-\ln(\mathcal{Z})]}{\partial \mu(\sigma)}$$
$$- \sum_{i=1}^{L-1} \sum_{\sigma=1}^{q-1} \widetilde{h}_i(\sigma) \frac{\partial[-\ln(\mathcal{Z})]}{\partial \widetilde{h}_i(\sigma)} \qquad (10)$$

in which the partial derivatives can be shown to be exactly $P(\sigma)$ and $\Delta P_i(\sigma)$, respectively. Consequently,

$$\mathcal{G}_\alpha = \mathcal{F}_\alpha - L \sum_{\sigma=1}^{q-1} \mu(\sigma) P(\sigma) - \sum_{i=1}^{L-1} \sum_{\sigma=1}^{q-1} \widetilde{h}_i(\sigma) \Delta P_i(\sigma). \qquad (11)$$

From Eq. (11), it follows that the values of the fields and of the chemical potentials can be obtained as

$$\mu(\sigma) = -\frac{1}{L} \frac{\partial \mathcal{G}_\alpha}{\partial P(\sigma)}, \qquad (12)$$

$$\widetilde{h}_i(\sigma) = -\frac{\partial \mathcal{G}_\alpha}{\partial \Delta P_i(\sigma)}. \qquad (13)$$

To find a manageable expression for $\mathcal{G}_\alpha$, this is expanded to the first order around $\alpha = 0$, that is

$$\mathcal{G}_\alpha = \mathcal{G}_0 + \left. \frac{d\mathcal{G}_\alpha}{d\alpha} \right|_{\alpha=0} \cdot \alpha. \qquad (14)$$

In the zeroth-order term, the two-body energy does not appear because is proportional to α, while the thermal average [cf. Eq. (9)] of the other three terms of the effective potential [cf. Eq. (8)] cancels out with the last two terms of Eq. (11), leaving only the opposite of the entropy. Writing it in terms of the independent probabilities only, one obtains (cf. Appendix A

in supplementary material[15])

$$\mathcal{G}_0 = \sum_{i=1}^{L-1} \sum_{\sigma=1}^{q-1} P_i(\sigma) \ln[P_i(\sigma)]$$
$$+ \sum_{i=1}^{L-1} \left[1 - \sum_{\sigma=1}^{q-1} P_i(\sigma) \right] \ln \left[1 - \sum_{\sigma=1}^{q-1} P_i(\sigma) \right]$$
$$+ \sum_{\sigma=1}^{q-1} \left[LP(\sigma) - \sum_{i=1}^{L-1} P_i(\sigma) \right] \ln \left[LP(\sigma) - \sum_{i=1}^{L-1} P_i(\sigma) \right]$$
$$+ \left[1 - \sum_{\sigma=1}^{q-1} \left(LP(\sigma) - \sum_{i=1}^{L-1} P_i(\sigma) \right) \right]$$
$$\times \ln \left[1 - \sum_{\sigma=1}^{q-1} \left(LP(\sigma) - \sum_{i=1}^{L-1} P_i(\sigma) \right) \right]. \qquad (15)$$

In the second, third, and fourth lines, the square brackets contain expressions for $P_i(\widetilde{\sigma})$, $P_{\widetilde{i}}(\sigma)$, and $P_{\widetilde{i}}(\widetilde{\sigma})$, respectively, which are not independent from the other probabilities [cf. Eq. (7)].

Remembering that $P(\sigma) + \Delta P_i(\sigma) = P_i(\sigma)$, the first-order term in Eq. (14) results identical to that of Ref. 11 and can be written as

$$\left. \frac{d\mathcal{G}_\alpha}{d\alpha} \right|_{\alpha=0} = \sum_{\sigma, \tau} \sum_{i<j} e_{ij}(\sigma, \tau) P_i(\sigma) P_j(\tau). \qquad (16)$$

Inserting into Eqs. (12) and (13) the expression of Eqs. (14)–(16), one obtains

$$\widetilde{h}_m(\sigma) = -\ln \left[\frac{P_m(\sigma)}{P_m(\widetilde{\sigma})} \right] + \ln \left[\frac{P_{\widetilde{i}}(\sigma)}{P_{\widetilde{i}}(\widetilde{\sigma})} \right]$$
$$- \alpha \sum_\tau \sum_{i|i\neq m} e_{mi}(\sigma, \tau) P_i(\tau) \qquad (17)$$

and

$$\mu(\sigma) = -\frac{1}{L} \sum_i^{L} \ln \left[\frac{P_m(\sigma)}{P_m(\bar{\sigma})} \right]$$
$$- \frac{\alpha}{L} \sum_\tau \sum_{i\neq m} e_{mi}(\sigma, \tau) P_m(\sigma) P_i(\tau). \qquad (18)$$

On the other hand, since the second term of Eq. (8) can be written as $\sum_i h_i(\sigma_i)$, the two-body interaction terms do not change with respect to Ref. 11 [cf. Eq. (1)], resulting in

$$e_{ij}(\sigma, \tau) = -C_{ij}^{-1}(\sigma, \tau). \qquad (19)$$

For sake of simplicity, we shall write the potential which controls the Boltzmann probability of Eq. (6) as

$$\mathcal{U} = \sum_{i<j} e_{ij}(\sigma_i, \sigma_j) \Delta(|r_i - r_j|) + \sum_{i=1}^{L} \eta_i(\sigma_i) + \sum_{i=1}^{L} \mu(\sigma_i), \qquad (20)$$

with $\eta_i(\sigma) = \widetilde{h}_i(\sigma_i) - L^{-1} \sum_{j=1}^{L} \widetilde{h}_i(\sigma_j)$. The function $\Delta(|r_i - r_j|)$, which is zero if $|r_i - r_j| > d_r$, is also inserted in the potential to reduce the noise in the calculation of the energy in the native conformation. In fact, pairs of residues which do not interact directly would have $e_{ij} = 0$ due to the procedure described above to suppress indirect correlations. Effects such as the limited statistics of counts or the approximation associated with the perturbative expansion of the potential

could result in non-zero energies even in absence of direct correlations. Since we expect correlations to drop with the distance between residues, we introduce the Δ function (the choice of d_r is discussed in detail in Sec. IV) to avoid spurious effects.

III. EFFECT OF THE MANY-BODY TERM ON THE PREDICTION OF THE EXPERIMENTAL ΔΔG

To test the validity of the potential defined by Eq. (20), we shall calculate the energetic effect ΔΔG of 308 point mutations on the stability of 14 proteins and compare them with the experimental values.

The quantity ΔΔG is the change in the difference between the free energies of the denatured and of the native state of the protein upon mutation. To calculate this quantity, we need therefore to define the free energy of the denatured state. We assume, as often done when interpreting experimental data,[14] that the mutation has no effect on the entropy of the chain, and that the interaction terms are zero in the denatured state (cf. Eq. (7)). Consequently, we shall make use of the interaction potential

$$U(\{r_i\}) = \sum_{i<j} e_{ij}(\sigma_i, \sigma_j)\Delta(|r_i - r_j|) + \sum_{i=1} \Theta_i(\{r_i\})\eta_i(\sigma_i),$$

(21)

where $\Theta_i(\{r_i\})$ is some function of the coordinates of the protein which is 1 in the native conformation and zero in the denatured state. This function is not simply the sum of two-body terms (accounted by the first term of Eq. (21)), and consequently should be regarded as a many-body interaction. The chemical potential has been dropped because it plays no role in configurational space in which the sequence $\{\sigma_i\}$ of the protein is fixed. The energetic effect of a point mutation is thus described by

$$\Delta\Delta G(\sigma_i \to \sigma_i') = \sum_j [e_{ij}(\sigma_i, \sigma_j) - e_{ij}(\sigma_i', \sigma_j)]\Delta(|r_i - r_j|)$$
$$+ \eta_i(\sigma_i) - \eta_i(\sigma_i').$$

(22)

The calculation of ΔΔG by Eq. (22) requires the knowledge of the structure of the protein to evaluate the function $\Delta(|r_i - r_j|)$ but, assuming that the system is two-state, that the residue is completely exposed to the solvent in the denatured state, not to evaluate $\Theta_i(\{r_i\})$. In principle, one could remove the structure-dependent function $\Delta(|r_i - r_j|)$ from Eq. (22), assuming that the long-range contribution to the interaction energy of each residue is anyway controlled by the mutational correlations in the sequence alignment. However, as shown in Section IV below, the cutoff introduced by the function $\Delta(|r_i - r_j|)$ improves strongly the quality of the results. Moreover, Eq. (22) assumes that the mutation does not induce large conformational changes in the native conformation of the protein. Specifically, we assume that only changes which do not alter the residue–residue contact network defined by the function $\Delta(|r_i - r_j|)$ can take place.

The protein-independent parameters of the model which gave the best results in terms of correlation coefficient between calculated and experimental ΔΔG are $d_r = 4.0$ Å, $\alpha = 0.15$,

TABLE I. The list of protein domains, with the associated PDB structure the id of the PFAM family, the number M of sequences in the family, the number M_{eff} of effective sequences after reweighting for similarity, and the number of mutations characterized experimentally. The last column indicates the number of mutations studied for each protein, along with the associated reference.

Protein/domain	Pdb	Family	M	M_{eff}	Mutations
BPTI	1BPI	00 014	4 915	1566	35[18]
Myoglobin	1BVC	00 042	6 000	688	7[19]
FKBP1	1FKJ	00 014	16 739	2284	26[20]
c-Src/SH3 domain	1FMK	00 018	10 749	1542	17[21]
Fibronectin/fnIII domain	1FNA	00 041	17 225	8102	21[22]
PTP-BL/PDZ domain	1GM1	00 595	26 099	2715	23[23]
α-Lactalbumin	1HMK	00 062	1 035	119	14[24]
ecDHFR	1RX4	00 186	5 237	956	29[25]
Staphiloc. nuclease	1STN	00 565	4 232	1144	39[26]
ACBP	2ABD	00 887	1 677	420	23[27]
Bacteriorhodopsin	2BRD	01 036	3 174	208	24[28]
Del1-9-G129R-hPRL	2Q98	00 103	1 608	97	9[29]
Tenascin/fnIII domain	2RB8	00 041	17 225	8054	26[22]
Azurin	5AZU	00 127	1 467	282	15[30]

$x = 0.5$, $y = 0.1$, and $z = 1.0$. As in Ref. 12, we consider a gap in the alignment as a "type" of amino acid, and assign σ^* to it. Consistently, we use as reference state $\widetilde{\sigma}$ for the field $\widetilde{h}_i$ at each site the gap as well, and as reference site $\widetilde{i}$ the most exposed site to the solvent occupied by polar or charged residues (for membrane proteins see below). The effect of variation of these parameters is described in Sec. IV.

For this study, we chose a set of protein domains with at least 1000 homologs in the Protein Families (PFAM) database, whose native structure is present in the Protein Data Bank (PDB) and on which the energetic effect of mutations has been characterized. This set is listed in Table I. The calculated values of ΔΔG is plotted versus their experimental values in Fig. 1. The overall correlation coefficient between predicted and experimental values, excluding 23 outliers, is $r = 0.77$. This should be compared with the value $r = 0.47$ obtained following the same procedure, but making use of a potential including only the two body term e_{ij}, without the term η_i (see Fig. S1 in the supplementary material[15]). The model does not set an overall energy scale, but returns all energy in arbitrary units. To transform these into kcal/mol, we have carried out a linear regression of the calculated versus the experimental ΔΔG.

A point is regarded as outlier if the difference between the calculated and experimental value is larger than 3σ, where σ is the error provided by the overall fit, also including the experimental error bars when available. Outliers can be classified into three categories (see Table S1 in the supplementary material[15]). 10 of them correspond to sites which are highly conserved, and consequently, there is little (or no) statistics for the mutated sequence; 2 outliers are in sites which were experimentally characterized as structured in the denatured state, thus invalidating Eq. (22). The remaining 11 outliers cannot be explained in a satisfactory way, or the denatured state of their protein is not precisely experimentally determined.
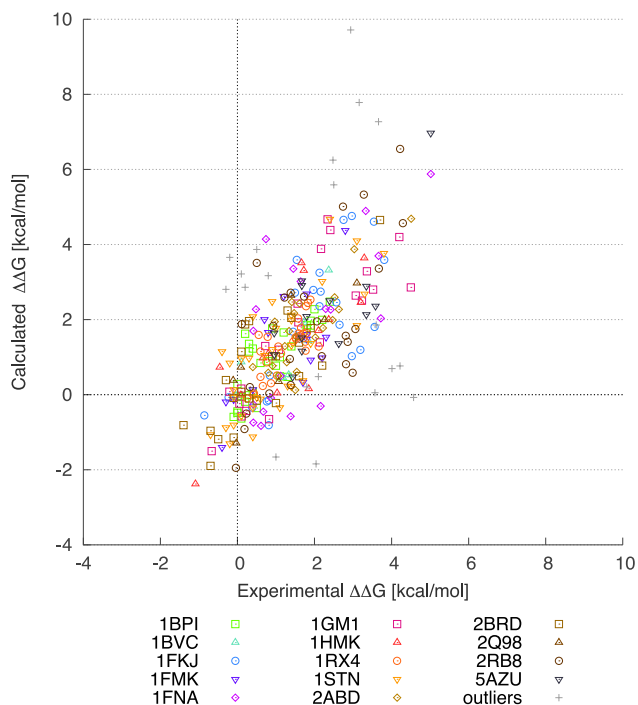
FIG. 1. The values of $\Delta\Delta G$ predicted by the model as a function of the corresponding experimental values.
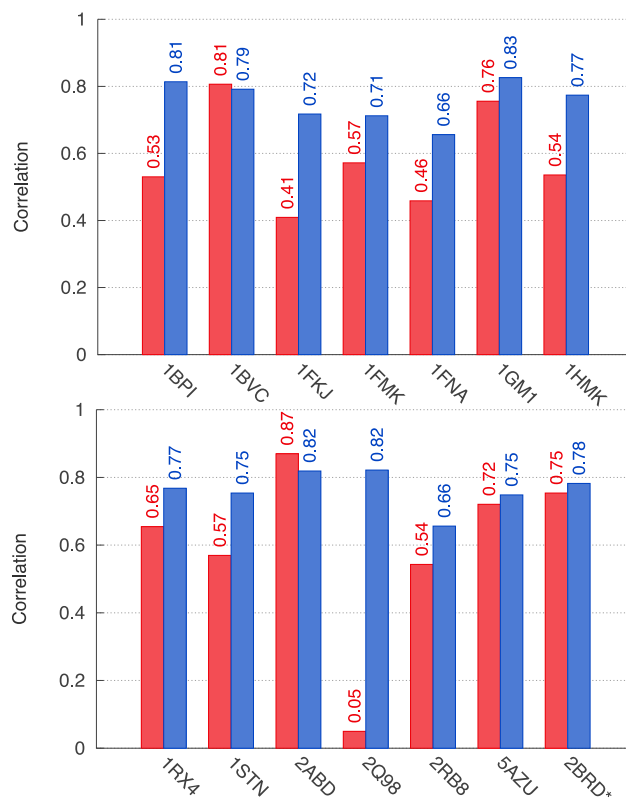

FIG. 2. The correlation coefficient between predicted and experimental $\Delta\Delta G$ for each protein. The red bars indicate the results obtained calculating the energies with the two-body term only, while the blue bars with the complete potential. The protein marked with an asterisk is a membrane protein.

The correlation coefficients between predicted and experimental data for each protein are displayed in Fig. 2 and are compared with those obtained without the term η_i (cf. Fig. S2 in the supplementary material[15] in which a detailed comparison of the $\Delta\Delta G$ is shown or each protein). We can see that including the new term η_i gives better correlation for most of the proteins (only 1BVC slightly decreases from 0.81 to 0.79 and 2ABD from 0.87 to 0.82).

In the set, we have also a membrane protein (Bacteriorhodopsin, pdb entry 2BRD) for which this method is successful in predicting $\Delta\Delta G$ for 24 mutations, without any outlier. To obtain this result, we used a different reference state $\tilde{i}$ than for cytosolic proteins, namely, the most exposed hydrophobic site. Not unexpectedly, using for bacteriorhodopsin the same reference state used for the other solution proteins (i.e., the most exposed polar/charged site) gave a poor correlation coefficient of 0.53.

IV. ROLE OF THE PARAMETERS OF THE MODEL

The model is defined by the values of d_r, x, y, z, and by the choice of the reference states in Eq. (7). Moreover, although the maximum-entropy principle is satisfied for $\alpha = 1$, we found a better agreement with the experimental data for $\alpha < 1$. Consequently, we regard α as a parameter of the model as well.

The dependence on the correlation coefficient r between predicted and experimental $\Delta\Delta G$ on the interaction range d_r of the two-body term is displayed in Fig. 3 for some of the proteins studied above (see also Fig. S3 in the supplementary material[15] for the other proteins). For all proteins, r is a decreasing function of d_r. Its maximum lies between 3 and

6 Å, depending on the protein. This is compatible with the size of the shells of other residues interacting with each residue in the native conformation. The best choice for d_r seems to be 4.0 Å, although small variations of this have little effect in the prediction of the $\Delta\Delta G$.

The results displayed in Fig. 3 indicate that the cut-off in the spatial dependence of the two-body term introduced by the function $\Delta(|r_i - r_j|)$ in Eq. (22) is important to reach good correlations between predicted and experimental energies. This fact suggests that the knowledge of the native structure of the protein, necessary for the calculation of $\Delta(|r_i - r_j|)$, is helpful in reducing the errors due to residual
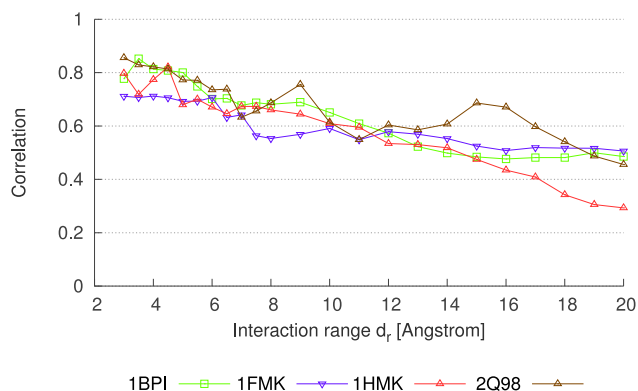

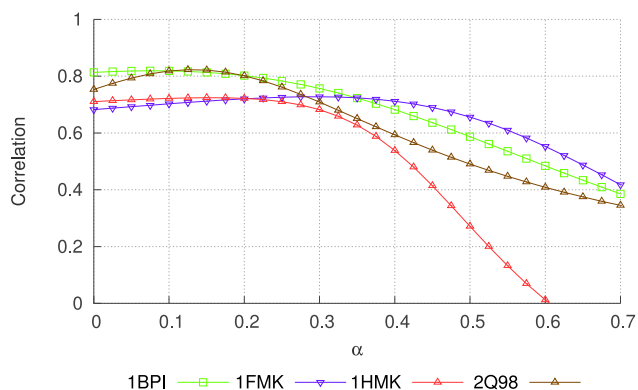FIG. 3. The correlation coefficient r as a function of the interaction range d_r of the two-body energy term.

FIG. 4. The correlation coefficient r as a function of the perturbation coefficient α.

indirect correlations in the alignment and to lack of statistics in spatially distant residues.

The correlations coefficients r as a function of α are displayed in Fig. 4 (cf. also S4 in the supplementary material[15]). Overall, they display a maximum at low values of α and decrease when α approaches 1. In few cases, the maximum is exactly at $\alpha = 0$, that is, when the terms $\widetilde{h}_i(\sigma)$ are decoupled from the terms $e_{ij}(\sigma, \tau)$ [see Eq. (17)]. In the production calculations, we chose $\alpha = 0.15$, although small variations of α have little effect if kept small, that is, in the range where the perturbation expansion of the Gibbs free energy holds.

The coefficients x, y, and z weight the pseudocounts, which are *a priori* probabilities meant to compensate the limited statistics in the alignments and make the correlation matrix invertible.[11–13] The dependence of r on these parameters is displayed in Fig. S5 in the supplementary material.[15] For most of the proteins, the best choice is $x = 0.5$, $y = 0.1$, and $z = 1.0$. Anyway, the quality of the results depends mainly on z, while the choice of x and y seems not critical.

While a natural and efficient choice for the reference amino acids σ^* and $\widetilde{\sigma}$ [see Eq. (7)] are the gaps in the alignment,[12] that for the reference site $\widetilde{i}$ of the terms $\widetilde{h}_i(\sigma)$ is not straightforward. For cytosolic proteins, a sensible choice seems to be to set the reference site at the position of the most exposed polar or charged residue. The degree of solvent-exposure of
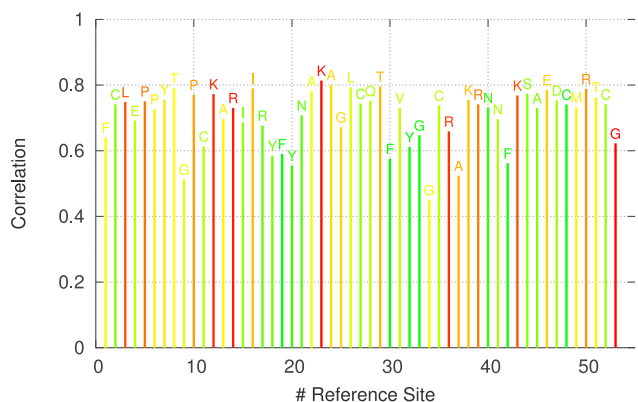
a residue is quantified by the occupancy factor S_{fact} defined in Ref. 3. This choice assures that the many-body effective energy associated with the reference site does not change upon folding, since in the denatured state ($\Theta = 0$), the sidechain is approximately as exposed as it is in the native state ($\Theta = 1$). The dependence of the correlation coefficient r on the choice of the reference site is displayed in Fig. 5. Suboptimal choices do not change dramatically the correlation coefficient, while the choice of hydrophobic sites significantly decreases it.

Bacteriorhodopsin, which is a membrane protein, behaves in the opposite way. Good results are obtained using as reference the most exposed hydrophobic site, which worsen choosing more hydrophilic sites.

V. PROPERTIES OF THE η-TERM

The term $\eta_i(\sigma)$ in the potential accounts for the contribution to the total energy which is not related to two-body interactions. As a result of the principle of maximum entropy, Eq. (6), it is formally a one-body term of the potential, that is, an external field. However, it is hard to justify an external field in the present context, and consequently $\eta_i(\sigma)$ must be regarded as the result of the combined effect of the surrounding residues, that is a many-body term.

The average value of η over all its occurrences in the proteins of Table I for each type of amino acid is displayed in Fig. 6. Except that for proline and tyrosine, the average of η has a good correlation ($r = 0.81$) with the hydrophobicity of the corresponding residue, as measured by the scale of Kyte and Doolittle.[16] This fact suggests that η represents, at least partially, the contribution of the solvent to the positioning of the amino acids in the native conformation of the proteins. In fact, it is known that effective interaction associated with the presence of the solvent is intrinsically many-body.[17]

While it is not completely unexpected that proline escapes the linear correlation between η and hydrophobicity, because of its peculiar, rigid chemical structure, the behavior of tyrosine is surprising. Anyway, it cannot be explained in terms of poor statistics, since tyrosine appears in the proteins studied above with a frequency comparable to that of the other residues.

For the calculation of the $\Delta\Delta G$, the conformational dependence of the η-term of the potential has been regarded as

FIG. 5. The correlation coefficient r as a function of the choice of the reference site $\widetilde{i}$ for $\widetilde{h}_i(\sigma)$ for 1BPI. The color code indicates the degree of solvent exposure. The color scale goes from red (exposed) to green (buried). Residue K23 (K26 according to the numbering of the pdb) is selected as the reference state.
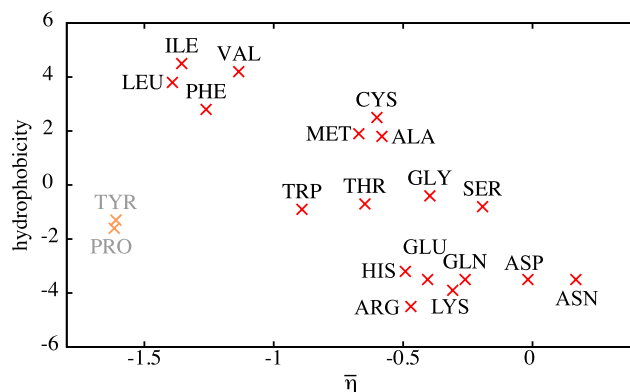

FIG. 6. The correlation between the average value of η associated with each type of amino acid and its hydrophobicity, defined by the scale of Kyte and Doolittle. Excluding proline and tyrosine, the correlation coefficient is 0.81.

two-state, in the sense that the only needed property of the function $\Theta_i(\{r_i\})$ in Eq. (21) was to be 1 in the native state and 0 in the denatured state. To extend the use of the effective potential U to characterize the conformational properties of a protein, one should define the full functional form of $\Theta_i(\{r_i\})$. The correlation of η-term with the hydrophobicity of the corresponding amino acids suggests that a reasonable assumption for $\Theta_i(\{r_i\})$ is the relative change in solvent exposure of the amino acid with respect to the native conformation, something which is indeed a many-body feature.

VI. CONCLUSIONS

While effective potentials based on *ab initio* calculations contain no more and no less than the physical terms which are used in the underlying calculations, statistical potentials have the virtue to summarize all possible physical effects, even unknown ones. As an example of their power, statistical potentials do not distinguish between globular and membrane proteins. Moreover, their functional form is usually simpler, and then computationally cheaper, than other kinds of force fields. Thus, statistical potentials are potentially a powerful tool to study the properties of proteins. In particular, those obtained from the analysis of mutational correlations proved efficient in predicting the native conformation of proteins[11] and the experimental $\Delta\Delta G$.[12]

In the present work, we have shown that the prediction of experimental $\Delta\Delta G$ can be further improved considering in the interaction potential a many-body term. This term arises naturally from a maximum-entropy principle and can be parametrized within the same theoretical framework used for the two-body interaction term. It partially describes the effective interaction due to the solvent, but probably also other effects which cannot be reduced to a two-body interaction. As typical for statistical potentials, the choice of the reference state, that is the zero of the energy terms, plays a critical role in the correctness of the results.

In the calculations of the $\Delta\Delta G$ described above we made use of the native structure of the protein. It is shown that a structure-based contact function in the two-body interaction helps to reduce the errors in the estimation of the energies and thus to reach good correlations between calculated and experimental data. In the case the protein structure is not known, a natural way to proceed would be to predict the structure of the protein within the same framework, using the alignment of homologs in the original sense,[11] and from this to predict the $\Delta\Delta G$.

Furthermore, the good results associated with the specific spatial dependence of the effective energy function allow, in principle, using it for conformational sampling, for example, to characterize the energy profile of proteins.

[1]S. Piana, K. Lindorff-Larsen, and D. E. Shaw, J. Phys. Chem. B **117**, 12935 (2013).

[2]D. Homouz, M. Perham, A. Samiotakis, M. S. Cheung, and P. Wittung–Stafshede, Proc. Natl. Acad. Sci. U. S. A. **105**, 11754 (2008).

[3]R. Guerois, J. E. Nielsen, and L. Serrano, J. Mol. Biol. **320**, 369 (2002).

[4]C. Arnarez, J. J. Uusitalo, M. F. Masman, H. I. Ingolfsson, D. H. de Jong, M. N. Melo, X. Periole, A. H. de Vries, and S. J. Marrink, J. Chem. Theory Comput. **11**, 260 (2015).

[5]S. Tanaka and H. A. Scheraga, Macromolecules **9**, 945 (1976).

[6]S. Miyazawa and R. Jernigan, Macromolecules **18**, 534 (1985).

[7]L. A. Mirny and E. I. Shakhnovich, J. Mol. Biol. **264**, 1164 (1996).

[8]E. I. Shakhnovich and A. Gutin, Protein Eng. **6**, 793 (1993).

[9]E. I. Shakhnovich and A. Gutin, Proc. Natl. Acad. Sci. U. S. A. **90**, 7195 (1993).

[10]G. Tiana, M. Colombo, D. Provasi, and R. A. Broglia, J. Phys.: Condens. Matter **16**, 2551 (2004).

[11]F. Morcos, A. Pagnani, B. Lunta, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc. Natl. Acad. Sci. U. S. A. **108**, E1293 (2011).

[12]S. Lui and G. Tiana, J. Chem. Phys. **139**, 155103 (2013).

[13]S. F. Altschul, E. M. Gertz, R. Agarawala, A. A. Schäffer, and Y.-K. Yu, Nucleic Acids Res. **37**, 815 (2009).

[14]A. Fersht, *Structure and Mechanism in Protein Science* (W. H. Freeman and Co., 1999).

[15]See supplementary material at http://dx.doi.org/10.1063/1.4926665 for figures showing correlations in single proteins and the effect of variation of the parameters of the model.

[16]J. Kyte and R. F. Doolittle, J. Mol. Biol. **157**, 105 (1982).

[17]P. De Los Rios and G. Caldarelli, Phys. Rev. E **62**, 8449 (2000).

[18]M.-H. Yu, J. S. Weissman, and P. S. Kim, J. Mol. Biol. **52**, 388 (1995).

[19]L. Lin, R. J. Pinker, and N. R. Kallenbach, Biochemistry **32**, 12638 (1993).

[20]E. R. G. Main, K. F. Fulton, and S. E. Jackson, Biochemistry **37**, 6145 (1998).

[21]V. P. Grantcharova, D. S. Riddle, J. V. Santiago, and D. Baker, Nat. Struct. Biol. **5**, 714 (1998).

[22]E. Cota, S. J. Hamill, S. B. Fowler, and J. Clarke, J. Mol. Biol. **302**, 713 (2000).

[23]S. Gianni, C. D. Geierhaas, N. Calosci, P. Jemth, G. W. Vuister, C. Travaglini-Allocatelli, M. Vendruscolo, and M. Brunori, Proc. Natl. Acad. Sci. U. S. A. **104**, 128 (2007).

[24]K. Saeki, M. Arai, T. Yoda, M. Nakao, and K. Kuwajima, J. Mol. Biol. **341**, 589 (2004).

[25]M. Arai and M. Iwakura, J. Mol. Biol. **347**, 337 (2005).

[26]A. K. Meeker, M. Garcia-Moreno, and D. Shortle, Biochemistry **35**, 6443 (1996).

[27]B. B. Kragelund, P. Osmark, T. B. Neergaard, J. Schiodt, K. Kristiansen, J. Knudsen, and F. M. Poulsen, Nat. Struct. Biol. **6**, 594 (1999).

[28]S. Faham, D. Yang, E. Bare, S. Yohannan, J. P. Whitelegge, and J. U. Bowie, J. Mol. Biol. **335**, 297 (2004).

[29]C. Keeler, M. C. Tettamanzi, S. Meshack, and M. E. Hodson, Protein Sci. **18**, 909 (2009).

[30]C. J. Wilson and P. Wittung-Stafshede, Biochemistry **44**, 10054 (2005).