PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Computational protein design with side-chain conformational entropy

Daniele Sciretti,[1,2] Pierpaolo Bruscolini,[1,2*] Alessandro Pelizzola,[3,4*]
Marco Pretti,[3,5] and Alfonso Jaramillo[6]

[1] Departamento de Física Teórica, Universidad de Zaragoza, c. Pedro Cerbuna 12, Zaragoza 50009, Spain
[2] Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza,
c. Corona de Aragón 42, Zaragoza 50009, Spain
[3] Dipartimento di Fisica and CNISM, Politecnico di Torino, c. Duca degli Abruzzi 24, Torino, Italy
[4] INFN, Sezione di Torino, Ecole Polytechnique, Palaiseau 91128, France
[5] Consiglio Nazionale delle Ricerche (INFM–CNR), Ecole Polytechnique, Palaiseau 91128, France
[6] Laboratoire de Biochimie, CNRS–UMR 7654, Ecole Polytechnique, Palaiseau 91128, France

ABSTRACT

Recent advances in modeling protein structures at the atomic level have made it possible to tackle "*de novo*" computational protein design. Most procedures are based on combinatorial optimization using a scoring function that estimates the folding free energy of a protein sequence on a given main-chain structure. However, the computation of the conformational entropy in the folded state is generally an intractable problem, and its contribution to the free energy is not properly evaluated. In this article, we propose a new automated protein design methodology that incorporates such conformational entropy based on statistical mechanics principles. We define the free energy of a protein sequence by the corresponding partition function over rotamer states. The free energy is written in variational form in a pairwise approximation and minimized using the Belief Propagation algorithm. In this way, a free energy is associated to each amino acid sequence: we use this insight to rescore the results obtained with a standard minimization method, with the energy as the cost function. Then, we set up a design method that directly uses the free energy as a cost function in combination with a stochastic search in the sequence space. We validate the methods on the design of three superficial sites of a small SH3 domain, and then apply them to the complete redesign of 27 proteins. Our results indicate that accounting for entropic contribution in the score function affects the outcome in a highly nontrivial way, and might improve current computational design techniques based on protein stability.

Proteins 2009; 74:176–191.

Key words: Computational protein design; side-chain entropy; protein stability.

INTRODUCTION

Protein design is a major goal in biochemistry and bioengineering, and reliable tools for rational design can help experimentalists to cut down the costs and time needed to produce sequences with the desired thermodynamical properties. Also, rational design represents a benchmark for theoretical understanding of the interactions stabilizing a protein native structure against folding and misfolding, and the theoretical tools thereby developed can be extended to apply to protein folding simulations and drug design.

At present, there are many different procedures for computational protein design,[1–9] which differ in the degree of coarse-graining of structure representation, in the energy function they adopt, and in the optimization procedure by which they explore the sequence space. Without going into the details of the various procedures and their performances, let us stress that the level of detail used in describing the structure is the most important factor determining the computational complexity of the task.

A common feature of protein design procedures based on physical atomic models is that they should rely on the evaluation of the folding free energy for each protein sequence, during their search of the sequence space. This requires evaluating the entropy of the unfolded and folded state. Usually it is assumed that the contributions coming from vibrational degrees of freedom in the two states cancel each other[10] and that only the conformational and solvation entropies change upon protein folding. The latter, which is related to the solvent degrees of freedom and the

Grant sponsor: Spanish Education and Science Ministry (MEC) FPI fellowship; Grant number: BFM 2003-08532-C03-01; Grant sponsor: MEC; Grant number: FIS2006-12781-C02-01; Grant sponsor: EU grant BioModular H_2; Grant number: FP6-NEST-043340; Grant sponsor: European Social Fund.

*Correspondence to: Pierpaolo Bruscolini, Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza, c. Corona de Aragón 42, Zaragoza 50009, Spain. E-mail: pier@unizar.es; or Alessandro Pelizzola, Dipartimento di Fisica and CNISM, Politecnico di Torino, c. Duca degli Abruzzi 24, Torino, Italy. E-mail: alessandro.pelizzola@polito.it

Received 9 November 2007; Revised 12 May 2008; Accepted 13 May 2008
Published online 10 July 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22145

solvent–protein interaction, is commonly accounted for in all protein design procedure, by including implicit solvation free energies[11] in the definition of the energy (which becomes, in fact, an effective energy[12]).

On the contrary, the estimation of conformational entropy presents serious computational problems. Although this entropy can be estimated in a reasonable way in the unfolded state,[1] its computation in the folded state remains an open problem, due to the computational intractability of evaluating the partition function on a folded protein, characterized by well-packed residues. Some investigations suggest that there could be a large number of equivalent side-chain repackings that stabilize a given fold,[13] and that an optimal side-chain packing is not strictly necessary for a sequence to present a tertiary structure in solution.[14] Moreover, the use of a conformational entropy has recently improved the recognition of native structures over a set of decoys.[15] These results show some of the advantages one might get by taking into account the conformational entropy in the evaluation of the folded state free energy.

So far, the energy functions used in automatic protein design did not incorporate conformational entropy, aside from the use of single residue terms either from mean-field techniques[5,16] or from empirical terms.[17] In the typical approach, common to most methods, the main-chain atoms of a protein native structure are kept fixed, while side-chains are replaced according to the proposed new sequence. Side-chains can assume several different rotamer states, which are selected from a precalculated library of discrete structures, sometimes adapted to the main-chain under consideration. It is also widespread in the use of a residue pair-wise scoring function to perform the combinatorial optimization because of the availability of efficient algorithms (Dead End Elimination, Branch and Bound) and the possibility of precomputing a score matrix incorporating all the time-consuming interaction energies.

This kind of protein design procedures usually yield a large number of low energy sequences,[3–6,9] that is, sequences that present at least one low-energy arrangement of the side chains. Nevertheless, not all these sequences would have the same number of low energy side-chain conformations, so that one may end up with the problem of choosing between sequences with a few low energy conformations and other sequences with higher energy but a larger number of available conformations. Indeed, in experiments a sequence with suboptimal energy, perhaps very low in rank, might perform better than the top ranking solution, if it possesses more conformations with reasonably low energy than the "best" one does.

The above observations call for a different approach, resorting to the conformational free energy as the cost function for the sequence optimization procedure, in order to take into account not only the energy of the best configuration, but also the entropy of alternative side-chain conformations. Therefore, in this article we propose a new scheme, which amounts to (i) writing down an approximate expression for the rotamer free-energy associated to each sequence, (ii) minimizing it by Belief Propagation (BP) to find the most likely distribution of rotamers and the corresponding conformational free-energy, and (iii) using the latter as a score function to search just the sequence space by means of Simulated Annealing (SA).

To test our method, we start by considering the design of just three surface positions of the N-terminal SH3 domain of C-Crk protein (in the following, we will refer to it as "1cka", according to its Protein Data Bank[18] code). For such a simple case, it is possible to characterize completely the low-energy part of the configuration space,[1] identifying the sequence of residue and rotamers having an energy below a given threshold. In this case, we are able to compute exactly the conformational free-energy, by summing over the states corresponding to different rotamers of the same sequence, each one considered with the appropriate statistical Boltzmann weight. For the same low-lying sequences, the conformational free-energy is calculated within our approximate procedure, and the results are compared, with special attention to the reliability of the sequence rankings obtained with the two different criteria.

Then, we shall repeat the procedure with the whole protein 1cka: in this case, an exhaustive evaluation of the low-lying configurations is ruled out by the magnitude of the configuration space, and one must resort to heuristic methods to characterize it. We perform several runs of SA, using the DESIGNER energy[1] as a cost function, and collect all the states below a given threshold. As a result, we get the list of the best sequences with their rotamers "frozen" in the configuration of minimum energy, which is the configuration they would adopt at zero temperature under the specified energy function. We refer to these sequences as the "SA sequences." Yet, in order to see whether, at higher temperature, the entropy of alternative rotamer conformation affects the ranking, we need a recipe to evaluate their free energy: we shall use the BP scheme to evaluate the (approximate) free energy of the best ranking SA sequences, at room temperature. Subsequently, we shall implement a combined SA/BP scheme, where the sequence space is explored through a SA procedure, characterized by a temperature T_{SA}, but each new sequence is accepted or rejected according to the value of the conformational free-energy (and not just the energy as before), which is found by minimizing the free-energy on the rotamers space through the BP algorithm. The resulting best ranking sequences ("SA/BP sequences") and their energies are compared with those coming from the standard SA method.

Finally, we repeat the design procedure with 26 other proteins to see whether the results obtained with 1cka generalize for different template structures.

MATERIALS AND METHODS

As mentioned in the Introduction, computational protein design procedures address the problem of searching for the sequence that optimizes a given energy function. Usually, the main-chain atoms of a protein native structure are kept fixed, whereas side-chains are replaced according to the proposed new sequence. Side-chains often assume several different rotamer states, which are selected from a precalculated library of discrete structures, adapted to the main-chain under consideration. Once the energy of the side-chain conformations is computed, automated protein design procedures use a combinatorial optimization algorithm to search for the optimal solutions. Those algorithms require using a residue pair-wise scoring function. We have chosen the DESIGNER methodology[19] to perform our calculations, but our results could also extend to other procedures that use the folding free energy as scoring function to design proteins.

Folding free energy

Following Refs. 1, 9, 11, 20 and 21, we use an approximation to the folding free energy to score a given structure-sequence. This is done by estimating the free energy in the unfolded and folded states using an atomic model. Given a protein sequence, we construct the atomic description by patching the corresponding side-chains to a protein main chain model. The side-chains are built using a rotamer library. Atomic energies are computed using CHARMM22 force-field[22] and an atomic surface term as implicit solvation.

The atomic energy terms are approximated by using residue–residue potentials and neglecting all three-body (and higher order) effects

$$V(x) = \sum_{i=1}^{N} V'_i(x_i) + \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} V''_{i,j}(x_i, x_j), \qquad (1)$$

where $x_i = (s_i, r_i)$ is the collective degree of freedom made up of the aminoacid and the rotamer species, respectively, at position i, and $x = \{x_i, i = 1, \ldots, N\}$, where N is the length of the protein. The $V_i^{'}(x_i)$ corresponds to the energy (nonbonded and solvation) of the rotamer x_i at position i interacting with the backbone and with itself. $V''_{ij}(x_i,x_j)$ correspond to the energy of interaction of rotamer x_i at position i with rotamer x_j at position j. The energy contains the standard CHARMM22 electrostatics (with a dielectric constant of 8) and van der Waals (with standard radii) terms. The solvation is based on a solvent-accessible area term with atomic hydration coefficients. This two-body approximation allows the use of powerful combinatorial optimization methods[1] and without this approximation computational protein design would not be feasible with current computers.

To describe the unfolded state, we use a dipeptide backbone model where neighboring residues only interact by burying 20% in average of their solvatable surface to account for residual 3D structure. The free energy of the unfolded state is computed using a partition function average over rotamer states. For the folded state we use a fixed protein backbone from a high-resolution structure; in previous work[19] alternative conformational states were neglected (i.e. no partition function was computed to account for alternative rotamer configurations over the same main-chain scaffold), so that the free energy of the native state was indeed represented by the energy of the best configuration of side-chain rotamers. We shall discuss later about how we introduce the conformational entropy contribution coming from alternative side-chain conformations.

According to the above mentioned scheme, for any given sequence, the free energy difference between a rotamer configuration of the folded protein and the unfolded state inherits the same structure of the one- and two-body terms as in Eq. (1):

$$E(x) = \sum_{i=1}^{N} E'_i(x_i) + \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} E''_{i,j}(x_i, x_j), \qquad (2)$$

In the following, we shall refer to the above (free) energy difference as the “DESIGNER” energy. The “conformational free energy” of a sequence $s = \{s_i, i = 1, \ldots, N\}$ will be defined as $F(s) = U - TS_r$, where U is the average of $E(x)$ over rotamer states and S_r is the conformational entropy of side-chains. The conformational free energy is the cost function which will be minimized in our approach, and, according to statistical mechanics, is related to $E(x)$ by

$$\exp[-\beta F(s)] = \sum_{\{r_i \in \mathcal{R}(s_i)\}} \Omega(x) \exp[-\beta E(x)], \qquad (3)$$

where $\beta = (RT)^{-1}$, R is the gas constant and T absolute temperature, $\mathcal{R}(s_i)$ denotes the set of rotamers available for the aminoacid s_i, and $\Omega(x)$ is a measure of the volume in the configuration space associated to any rotamer conformation of the residues in s: in replacing the continuous configuration space of the side-chains through a discrete set of rotamers, we are implicitly parting the configurations of each side chain into potential energy basins, and representing each whole basin with the discrete rotamer corresponding to the minimal energy.[23] In principle, the statistical weight $\Omega(x)$ (accounting for the vibrational free energy within the corresponding basin) would depend on the choice of the full rotamer set along the protein chain, but in the following we shall simplify the problem by assuming that all basins have an identical weight. Accordingly, the vibrational contribution turns out to be just an additive constant to the total free energy, so that the rotamer probability can be estimated from the Boltzmann factor of the only rotamer energy $E(x)$.[23] In other words, the only entropic contribution taken into account is the information entropy associated to uncertainty about the rotamer state. This approximation, which has been already used by other authors,[15,24] is not grounded on first principles, and is introduced here for the sake of simplicity, since any other choice of a priori probabilities would be arbitrary as well.

With these assumptions, we will take $F(s)$ to represent the stability of the native state against denaturation, within the hypotheses of the model (notice that the true stability should be calculated as the difference between free energy in the native state and that of all other configurations, and should resort to the sum over all atomic degrees of freedom, in the continuous space: this is of course beyond the scope of any simplified model).

When not specified, in the following we will also effectively take the gas constant $R = 1$ and report temperatures, as well as energies, in units of kcal/mol (e.g. $T = 0.6$ corresponds to a real temperature $T = 0.6$ kcal/mol/$R \approx 302$ K).

Sequence design by combinatorial optimization

We analyze four different schemes of sequence design: the first two apply just to the redesign of a few positions, and serve as a benchmark of the methods, while the others involves the extensive redesign of the protein.

In the first case, we take the wild type sequence of 1cka and, after performing a run of SA to optimize its rotamers, we consider all mutations of the three superficial sites 29, 30, and 31 (RDK in the wild type), while fixing the rotamers at all other positions. Because of the small number of mutating sites, the total configurations space is made by a few millions of states, so that we can characterize it completely, and calculate not only the combination of sequence and rotamers minimizing the DESIGNER energy, but also the conformational free energy, the entropy, and the average energy associated to each sequence. We do so simply by storing the energies of all the rotamer configurations corresponding to each sequence; then we sum the corresponding Boltzmann weights, in order to calculate the partition function associated to each sequence, and from this the free energy and the other thermodynamic functions. Finally, we calculate the side-chain free energies, over the rotamers at the three positions, for each sequence, resorting to the pair-approximation on probabilities and the BP algorithm (see Section "Pairwise Approximation for side-chain Conformational Entropy" later). We check the reliability of the BP scheme by comparing the latter results with the former, exact ones.

In the second scheme, we mutate exhaustively the sequence of 1cka at the same three sites, but we let relax the rotamers at all positions. Now the configuration space is much larger, and it is not possible to characterize it completely, to calculate exact free energies in a reliable way. Yet, we can still evaluate the free energy within the BP scheme over the rotamers at all positions, and study the role of the entropy.

In the last two approaches, we redesign entire proteins, letting all positions change, except the prolines. In these cases not only the space of the configurations, but even the sequence space cannot be explored exhaustively, and we have to resort to stochastic search methods.

In particular, in the third approach we perform some long runs of SA with the DESIGNER energy as the cost function, to find the sequence/rotamers configurations yielding the lowest energies; we also collect all solutions below a given threshold. Then, we use BP to calculate the free energy of the best ranking sequences at $T = 0.6$, to see whether the rank is affected by the introduction of the rotamer entropy. The SA protocol we use is the following: at each site i, the state is specified by the x_i variables introduced earlier. The elementary attempted move of the Monte Carlo (MC) algorithm, which is at the core of the SA, is the "flip" of the state at a randomly selected position to a new, randomly selected one. The acceptance rule is the usual Metropolis one: a flip is always accepted if it lowers the DESIGNER energy, while if energy change ΔE is positive, it is accepted with a probability $\exp(-\Delta E/T_{SA})$. The update of T_{SA} follows a geometric schedule, which is terminated when T_{SA} goes below a given threshold $T_{SA}^{(\min)}$ or the number of accepted moves goes below another threshold $n_{\min}$. Notice that the "temperature" T_{SA} is just a parameter in the algorithm and has no relation with the real temperature: the procedure is expected to provide the state that minimizes the DESIGNER energy, which can be seen as the conformational free energy when the real temperature is zero. We have tried different values of the number of MC steps per SA temperature, and the number and length of the runs we performed vary depending on the sequence, but typically we have performed at least four SA runs with 200 millions of MC steps per SA temperature, starting from $T_{SA} = 2$ and cooling down to $T_{SA} = 0.05$, dividing each time the temperature by a factor 1.2. In most cases, we have compared the results with those from shorter runs, with 10–30 millions of steps per temperature, but a cooling factor of 1.1. In all cases, we have stored the states in a range of 2 kcal/mol over the best one found, and subsequently grouped them according to the sequence, ranking the sequences according to the energy of their best rotamer configuration.

In the last approach to the design problem, we propose a combined SA/BP minimization scheme, where SA steps, in which a sequence point mutation is proposed, are accepted according to a free energy criterion, that resorts to BP to evaluate the free energy.

The details of BP and SA/BP minimization schemes are explained in the next sections.

Pairwise approximation for side-chain conformational entropy

The free energy $F(s)$ [Eq. (3)], representing the stability of the native state, cannot be computed exactly when designing more than a few residues (for instance, the design of 53 positions in protein 1cka, with an average of 7.35 rotamers per position, would require summing over around 10^{46} states). Hence, to make the problem tractable, we resort to the pairwise approximation of the cluster variation method[25–27] for the joint probabilities, as other authors have already done in similar contexts.[28–30] This approximation represents an improvement with respect to the standard mean-field theory,[16] that uses just single-site probabilities, as it allows to account precisely for the correlations between pairs of rotamers at different sites.

Using the variational formulation of equilibrium statistical physics we can write

$$F(s) = \min \sum_{\{r_i \in \mathcal{R}(s_i)\}} [p(r)E(x) + RTp(r)\ln p(r)], \qquad (4)$$

where $r = \{r_i,\ i = 1, \ldots, N\}$ and the minimum should be taken over all the distributions $p(r)$ satisfying the normalization constraint

$$\sum_{\{r_i \in \mathcal{R}(s_i)\}} p(r) = 1. \qquad (5)$$

Then, we use the pairwise approximation of the cluster variation method,[25–27] expressing the entropy in terms of just single-site and pair probabilities, and yielding, for the variational free energy:

$$\begin{aligned}\mathcal{F}(\{p_i\},\{p_{i,j}\}) = {} & \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sum_{r_i\in\mathcal{R}(s_i)}\sum_{r_j\in\mathcal{R}(s_j)}\big[p_{i,j}(r_i,r_j) \\ & \times \big(E_{i,j}(x_i,x_j) + RT\ln p_{i,j}(r_i,r_j)\big)\big] \\ & - (N-2)RT\sum_{i=1}^{N}\sum_{r_i\in\mathcal{R}(s_i)} p_i(r_i)\ln p_i(r_i), \qquad (6)\end{aligned}$$

to be minimized with respect to the single-site probabilities $\{p_i(r_i),\ 1 \le i \le N\}$ and the pair probabilities $\{p_{i,j}(r_i,r_j),\ 1 \le i < j \le N\}$ (whenever needed we shall assume $p_{j,i}(r_j,r_i) = p_{i,j}(r_i,r_j)$), subject to the obvious normalization constraints,

$$\begin{aligned}&\sum_{r_i\in\mathcal{R}(s_i)} p_i(r_i) = 1 \\ &\sum_{r_i\in\mathcal{R}(s_i),r_j\in\mathcal{R}(s_j)} p_{i,j}(r_i,r_j) = 1,\end{aligned} \qquad (7)$$

together with the compatibility constraints:

$$\begin{aligned}&\sum_{r_j\in\mathcal{R}(s_j)} p_{i,j}(r_i,r_j) = p_i(r_i) \\ &\sum_{r_i\in\mathcal{R}(s_i)} p_{i,j}(r_i,r_j) = p_j(r_j).\end{aligned} \qquad (8)$$

In this scheme $E_{i,j}(x_i,x_j)$ appearing in Eq. (6) needs to be defined, as a function of the corresponding quantities in Eq. (2), as

$$E_{i,j}(x_i,x_j) = E''_{i,j}(x_i,x_j) + \frac{1}{N-1}\Big[E'_i(x_i) + E'_j(x_j)\Big]. \qquad (9)$$

The minimization of $\mathcal{F}(\{p_i\},\{p_{i,j}\})$ is a well-studied problem and it is known that, if the problem under investigation is not too frustrated, a simple and fast technique which allows to find local minima is the so-called BP.[27,31,32] This is based on the introduction of a new set of variables, the so-called messages $m_{i\to j}(r_j)$ that a node i sends to a node j about the state r_j of the latter. The stationarity conditions of $\mathcal{F}$ with respect to the probabilities are solved by (leaving apart a normalization factor):

$$\begin{aligned}&p_i(r_i) \propto \prod_{j\neq i} m_{j\to i}(r_i) \\ &p_{i,j}(r_i,r_j) \propto \exp\big[-\beta E_{i,j}(x_i,x_j)\big]\prod_{k\neq i,j}\big[m_{k\to i}(r_i)m_{k\to j}(r_j)\big],\end{aligned} \qquad (10)$$

while the compatibility constraints lead to (again normalization is omitted):

$$m_{j\to i}(r_i) \propto \sum_{r_j\in\mathcal{R}(s_j)} \exp\big[-\beta E_{i,j}(x_i,x_j)\big]\prod_{k\neq i,j} m_{k\to j}(r_j), \qquad (11)$$

which can be solved by iteration (an asynchronous update scheme is important to achieve convergence) to determine the messages. From these, the probability and an approximate estimate of the sequence free energy $F(s)$ can be obtained by using Eqs. (10) and (6), respectively.

Free energy based sequence optimization

To determine the sequences of lowest free energy at a given temperature T, we use again SA. In this process the degrees of freedom are the s_i variables. The elementary attempted move of the MC algorithm at the core of the simulated annealing is then the "flip" of a randomly selected position to a new, randomly selected, aminoacid. The acceptance rule is the usual Metropolis one, where energies are replaced by sequence free energies computed as described in the previous section, and a fictitious temperature T_{SA} is introduced. Occasionally, we have seen that BP fails to converge, for some sequence, in a reasonable number of BP iterations: in these cases the

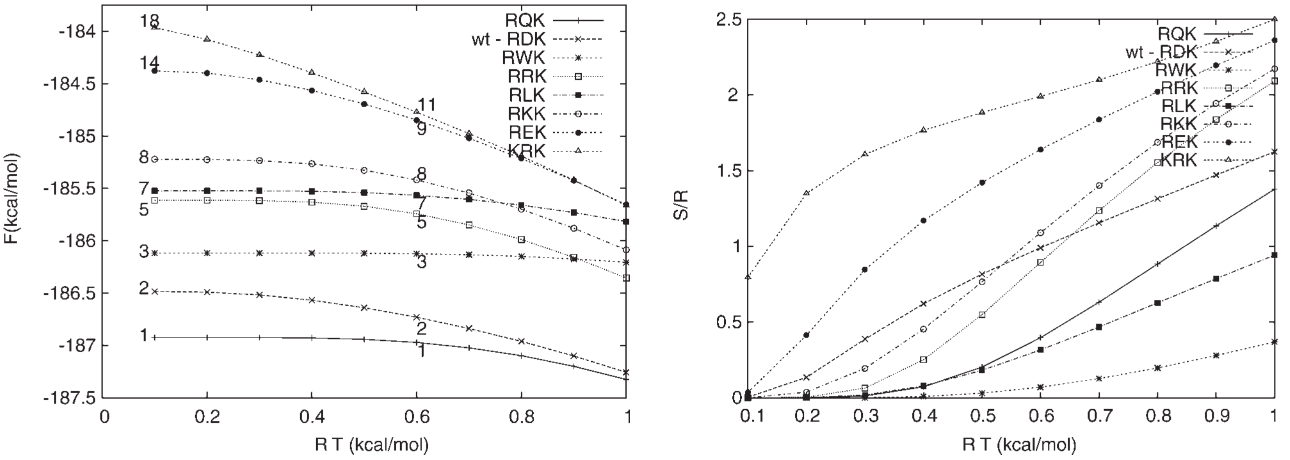

Figure 1

Free-energies (left panel) and entropies (right panel) as a function of temperature, for some of the best sequences found in the exhaustive procedure of exploring all possible mutations at positions 29, 30, and 31, while keeping all the rotamers of the other residues fixed (see text for more details). BP results are not reported because they overlap perfectly with exact ones. In the left panel: numbers on top of each line specify the ranking at the corresponding temperature. Notice that there are few crossings: low lying sequences at $T = 0.1$ kcal/mol keep their rank at $T = 0.6$ kcal/mol, while some bad scoring sequence gains some positions in the rank at higher temperatures. This is reflected in the right panel: different sequences display a different form of the entropy as a function of T, but in all cases the entropic contribution to the free energy is not sufficient to produce substantial changes in the sequence rankings.

attempted mutation is rejected and another one is attempted.

The update of T_{SA} follows a geometric schedule which is terminated when T_{SA} goes below a given threshold $T_{SA}^{(min)}$ or the number of accepted moves goes below another threshold n_{min}. During a simulated annealing run, the best 1000 sequences are stored for later analysis.

Our best results are obtained using the same temperature factor and initial and final T_{SA} as mentioned in Section "Sequence Design by Combinatorial Optimization", and with 50–200 thousands MC flips per temperature.

RESULTS

Design of three surface residues of 1cka

As mentioned in the Introduction, we test the pair approximation and BP algorithm on the design of just three residues, where also exact results can be calculated. We start by considering the wild-type 1cka sequence, and first perform a SA to optimize the rotamer configurations with respect to the crystal structure deposited in the Protein Data Bank. Then, we freeze the rotamers in all but the three positions 29, 30, and 31, and perform all mutations at those sites. The limited number of possible rotamer combinations (a few millions) for each sequence allows us to perform a complete characterization of the energy landscape, simply collecting the energy of all the states by exhaustive enumeration (see Section "Sequence Design by Combinational Optimization"), so that we can evaluate the partition function, and hence the free energy, associated to a fixed sequence. On the other hand, we can calculate the free-energies of each sequence also using BP, letting the rotamers of the three positions change, while keeping fixed the rotamers at the other positions, as above. Despite our main goal in the above protocol is to provide a reliable and controllable benchmark for comparison with BP results, we notice that the interest in such a "small" system is not purely academic: indeed, many protein interactions with small molecules involve just a few residues, so that, in general, we can learn valuable lessons from the design of just three surface positions.

Figure 1 reports the free-energies and entropies, as a function of the temperature, for the best sequences we have found in the exhaustive procedure: we do not plot the corresponding values obtained with BP, simply because they overlap perfectly, on the scale of the figure. Indeed, we have checked that the relative difference between the two is of order 10^{-7} at all investigated temperatures.

The numbers reported next to the lines, in the left panel, denote the corresponding sequence ranking: we see that low lying sequences at $T = 0.1$ kcal/mol (notice that, here and in the following, we write T instead of RT: 0.1 corresponds to roughly 50 K) keep their rank at $T = 0.6$ kcal/mol (roughly 302 K), while some bad scoring sequence gains some positions in the rank at higher temperatures. Notice that the wild type sequence, which is the second best at $T = 0.1$, preserves its ranking at higher temperatures.

So, we can conclude that, in the present case, the entropies associated to the different sequences give small contributions to the free energies, that, with some excep-

Table I
Correlation Between Exact and BP Results Over the 19^3 Sequences Resulting From All Possible Mutations (Excluding Pro) of Residues 29, 30, and 31

	EX vs BP ($T = 0.1$)		EX vs BP ($T = 0.6$)		EX: $T = 0.1$ vs $T = 0.6$	
	All	Pruned	All	Pruned	All	Pruned
Rank correl.	0.648890	0.999989	0.998904	1.000000	0.998902	0.998227
Correlation	0.528437	0.999994	0.997363	1.000000	0.999387	0.998546

Spearman's correlation of the sequence ranking obtained with the two methods, as well as the ordinary Pearson's correlation between the corresponding sequence free energies, are reported at $T = 0.1$ and $T = 0.6$. Also, the same correlations between exact results at $T = 0.1$ and $T = 0.6$ are reported in the rightmost column. The latter correlations confirm that the entropy is not enough to affect significantly the ranking of the sequences, and, as far as the three-residue case is concerned, information at low temperature allows to draw conclusions also for higher and more interesting temperatures. The correlation between exact and BP results at $T = 0.6$ is very high, confirming the goodness of the BP approximation in this simple case. At low temperatures, where convergence can be more problematic, BP may produce wrong results for some sequences, and the correlation taken over all the sequences gets much lower. However, eliminating problematic sequences (according to a simple thermodynamic criterion based on BP results themselves: see text) correlation becomes practically perfect at all considered temperatures. Notice that removal of sequences, at least in the present case, does not affect the top ranking sequences: "hard" sequences for BP are not the best ones, so that using BP we do not miss any potentially relevant, low energy sequence. Pearson's correlation between two series of values x_i and y_i is defined as $r = \sum_i (x_i - \bar{x})(y_i - \bar{y})/(\sum_i(x_i - \bar{x})^2 \sum_i(y_i - \bar{y})^2)^{0.5}$; Spearman's correlation is obtained from Pearson's substituting the values x_i, y_i by their ranks.

tions, hardly bend downwards in the relevant temperature range $T < 0.7$: basically, this means that if we keep all the rotamers of the remainder of the sequence fixed, the free energy associated to the various mutations are almost the same as the energies. As a consequence, there are just a few crossings within the best scoring sequences as the temperature increases: the sequence ranking obtained at $T = 0$, that is, minimizing the DESIGNER energy, is quite robust at higher temperatures, at least for the best sequences.

This was expected, of course, due to the strong structural constraints imposed by fixing the main chain and most of the rotamers. However, the important feature in the above figures is that there is an extremely good agreement between exact and BP results. These results are very encouraging, even if we do not expect a similar performance in the full-design case, because the approximated nature of the pairwise expression should be more evident when more interactions are present, as it happens for core residues.

To give a quantitative estimate of this agreement, without relying just on the trends of the best-scoring sequences, we have calculated the Spearman's rank correlation and the ordinary Pearson's correlation of free energies, between exact and BP results at $T = 0.1$ and $T = 0.6$, as well as between exact results at $T = 0.1$ and $T = 0.6$, for all the 19^3 sequences (Pro is excluded). The results are given in Table I. Notice how both the Spearman's correlation between the ranks and the ordinary Pearson's correlation between the free energy values reveal an extremely good correlation between exact results at $T = 0.1$ and $T = 0.6$ (basically, the rank is overall preserved), and between BP and exact results at $T = 0.6$, while the correlation is somewhat worse at $T = 0.1$, where BP may present some problems. Indeed, introducing a criterion for pruning sequences producing bad BP signals, correlations get almost perfect also at $T = 0.1$.

The "pruned" results refer to those obtained omitting the sequences that, in the BP procedure, present an anomalous behavior at low temperature: namely, their free energy increases in a small temperature range, to start decreasing normally at higher temperatures. In practice, we choose $F_{\mathrm{BP}}(T = 0.1) < F_{\mathrm{BP}}(T = 0.6)$ as a test to define a bad sequence: indeed, we have seen that this produces the smallest number of false positives and false negatives, when comparing $F_{\mathrm{BP}}(T = 0.1)$ with the exact $F_{\mathrm{ex}}(T = 0.1)$. Note that the pruning does not affect the true best ranking sequences: the first pruned sequence in the BP rank (YYK, position 10) is found at the 1774th position in the true ranking; the first pruned sequence according to the true ranking is RYA, at position 408. Therefore, the true low-lying sequences appear to be also "easy" sequences for the BP algorithm. The existence of bad sequences may be due to the existence of a strong network of interactions: the anomalous behavior could indicate the fact that there are sequences for which the minimization problem is frustrated and the BP approximation is not very accurate. In fact, omitting the sequences that present anomalous BP signals indeed improves correlation.

We plot in Figure 2 the BP signals of free energy, average energy, and entropy for one of such sequences (right panel), as well as the corresponding true signals, while in the left panel we report the corresponding curves for a "good" sequence. We can appreciate that the entropy is much bigger in the left panel.

Before moving on to the design of the complete protein, it is interesting to consider the case where we mutate the same three residues, but we allow the rest of the rotamers to relax accordingly, in a more realistic way than before. Now the configuration space is too big to attempt exact enumeration, but we can use BP to estimate the free energies: Figure 3 reports the results.

We can see from the bending of the free energies that the rotamer entropy now is more relevant, and also that rank crossing is more frequent, even if the best sequence remains the best one along all the temperature range. We may also

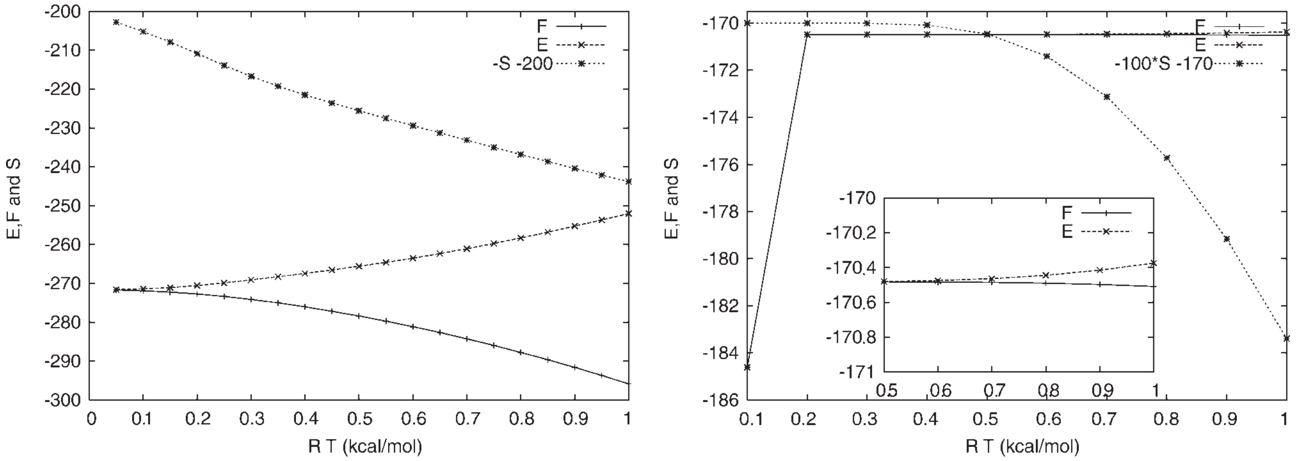

Figure 2

Left panel: BP curves of free energy, average energy, and entropy versus temperature for a "good" sequence. Right panel: BP curves for the same thermodynamic quantities of a problematic sequence: notice how free-energy and average energy, that appear to be indistinguishable, display an unphysical jump at low temperatures. In such a case, free-energy is irrealistically low at low temperatures and could induce wrong conclusions about the ranking of the sequence, so that one must check the behavior at higher temperatures to decide whether to trust BP results. Inset: the exact energy and free-energy signals, for the "bad" sequence. In both panels, entropies are rescaled for better display.

expect that, with the increased role of the entropy, the differences between BP and true free energy is enlarged.

Full design of protein 1cka

With the information on the substantial agreement between exact and BP free energies for the three-residue case, we move to the redesign of the complete protein. Here we will base our analysis on the reranking of sequences once we switch on the conformational entropy contribution. We performed SA, with the DESIGNER energy as the cost function, in the complete state space, looking for the best sequence and rotamers, and keeping all the sequences with at least a rotamer configuration below a threshold, as explained in Section "Sequence Design by Combinatorial Optimization". The state of minimal energy is also the state of minimal free energy at $T = 0$: in order to calculate the free energies of the best scoring sequences as a function of the temperature, we perform BP on these sequences. The best scoring sequences are reported in Table II, with the energy of their best side-chain configuration (i.e. the free energy at $T = 0$), as well as the BP free energy at $T = 0.6$. It is possible to see a number of crossovers between sequence rankings at low and high temperature.

Therefore, now the question is: are these the good sequences at $T = 0.6$, that is, those that minimize the conformational free-energy at that temperature? To answer this question, we perform a SA sequence minimization, but using the BP free energy (calculated at $T = 0.6$) of the sequence, and not simply the energy of the states, to accept or reject mutations (see Section "Free Energy Based Sequence Optimization" in Methods). Then, we also follow the best sequences down to $T = 0.05$, to see how they compare with the sequences previ-

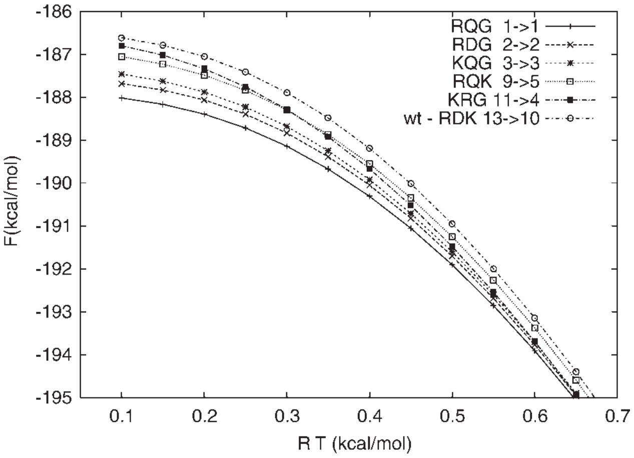

Figure 3

BP free energies versus temperature for some of the best sequences we have found in the exhaustive procedure of exploring all possible mutations at positions 29, 30, and 31, but now allowing also the rotamers of the rest of the residues to relax during the BP procedure: for each of the 19^3 mutants, the rotamers of all the residues now contribute to the side chain entropy and free energy, in a more realistic way than before. However, in this case the configuration space is too big to attempt exact enumeration, so a comparison of BP and exact results is not feasible any more. Notice, from the bending of the free energies, that the rotamer entropy now is more pronounced, and also that rank crossing are more frequent, even if the best sequence remains the best one along all the temperature range. The label reports, as "$x \rightarrow y$" next to the sequence, the ranking of the sequence at $T = 0.1$ ("x") and at $T = 0.6$ ("y").

Table II
The 20 Best Scoring Sequences, as Found by Ordinary SA (See Text)

$T = 0$	Sequence	E (best)	$F_{BP}(0.6)$	$T = 0.6$
01	DKKVRARDEFRGRQYQQLFYKGEELWIKQEDRYWKKAEVRYRYQGLIDRVEHG	−274.923	−279.319	09
02	DKKVRARDEFRGRQYQQLFYKGEELWIKQEDRYWKKAEVRYRRQGLIDRVEHG	−274.887	−279.970	05
03	DKKVRARDEFRGRQYQQLFYKGEELWIKQEDRYWKKAEVRRRQQGLIDRVEHG	−274.878	−280.006	04
04	DKKVRARDEFRGRKYEQLFYKGEELWIKQEDRYWKKAEVRYRYQGKIDRVEHG	−274.804	−278.717	15
05	DKKVRARDEFRGRKYEQLFYKGEELWIKQEDRYWKKAEVRYRRQGKIDRVEHG	−274.769	−279.368	08
06	DKKVRARDEFRGRKYEQLFYKGEELWIKQEDRYWKKAEVRRRQQGKIDRVEHG	−274.760	−279.405	07
07	DKKVRARDEFRGRQYQQLFYKGEELWIKQEDRYWKKAEVRYRQQGLIDRVEHG	−274.756	−279.648	06
08	DKKVRARDEFRGRKEDQLFYKGEELWIKQEDRYYKKAEVRYRYQGIIDRVEHG	−274.714	−278.484	19
09	DKKVRARDEFRGRQYQQLFYKGEELWIKQEDRYWKKAEVRRRWQGLIDRVEHG	−274.691	−279.197	10
10	DKKVRARDEFRGRKEDQLFYKGEELWIKQEDRYYKKAEVRYRRQGIIDRVEHG	−274.678	−279.135	13
11	DKKVRARDEFRGRQYQQLFYKGEELWIKQEDRYWKKAEVRYRWQGLIDRVEHG	−274.672	−279.172	11
12	DKKVRARDEFRGRKEDQLFYKGEELWIKQEDRYYKKAEVRRRQQGIIDRVEHG	−274.669	−279.172	12
13	DKKVRARDEFRGRKYEQLFYKGEELWIKQEDRYWKKAEVRYRQQGKIDRVEHG	−274.637	−279.046	14
14	DKKVRARDEFRGRKREQLFYKGEELWIKQEDRYWKKAEVRYRRQGKIDRVEHG	−274.614	−280.141	03
15	DKKVRARDEFRGRKREQLFYKGEELWIKQEDRYWKKAEVRRRQQGKIDRVEHG	−274.606	−280.179	01
16	DKKVRARDEFRGRREEQLFYKGEELWIKQEDRYWKKAEVRYRYQGKIDRVEHG	−274.582	−278.520	18
17	DKKVRARDEFRGRKYEQLFYKGEELWIKQEDRYWKKAEVRRRWQGKIDRVEHG	−274.573	−278.595	16
18	DKKVRARDEFRGRKEEQLFYKGEELWIKQEDRYWKKAEVRYRYQGKIDRVEHG	−274.569	−278.576	17
19	DKKVRARDEFRGRKWEQLFYKGEELWIKQEDRYWKKAEVRYRYQGKIDRVEHG	−274.566	−278.373	20
20	DKKVRARDEFRGRRRQQLFYKGEELWIKQEDRYWKKAEVRYRYQGLIDRVEHG	−274.553	−280.171	02

Only the 53 mutating positions are reported. The first column refers to their SA rank (that is, their free-energy rank at $T = 0$). The last column reports their relative rank at $T = 0.6$, and shows how entropy introduces rank-crossing between them. Notice that this is not their absolute rank at $T = 0.6$: see text for discussion. E (best) is the energy of the best rotamer configuration found, and corresponds to the free energy at $T = 0$. $F_{BP}(0.6)$ is their free-energy calculated with BP at $T = 0.6$. Energies and temperatures in kcal/mol.

ously found with SA, and validate the energy of these new sequences using SA with the ordinary energy.

The results for the SA are reported in Figure 4, that shows some of the best SA and SA/BP sequences, as well as their BP continuation in the range of temperatures 0–0.65.

First, we see how BP free energies converge to SA energies at low temperature, suggesting that the pair-wise approximation is reliable also at low temperatures. However, the most relevant feature is that SA/BP finds new and better sequences with respect to the former procedure. Moreover, we can see that this performance is not due to insufficient sampling with ordinary SA in the $T = 0$ case: at low temperature, these SA/BP sequences score low in the rank, due to the fact that their energy is substantially bigger than that of best scoring SA sequences. That is, both the ordinary SA procedure at $T = 0$ and SA/BP at $T = 0.6$ succeed in finding good (and possibly, the best) sequences, but the latter do not match, due to the important role of the rotamer entropy. The best scoring sequences found with SA/BP at $T = 0.6$ are reported in Table III, with their free energies, and energies at $T = 0$.

It is also important to notice that best SA and SA/BP sequences differ in 14 positions, and that each of them falls really low in the rank of the other, so low that they fall out of the energy range we keep from the best. We stress that this is not a trivial problem related only to the choice of the energy threshold for saving configurations during the simulations: the configuration space increases

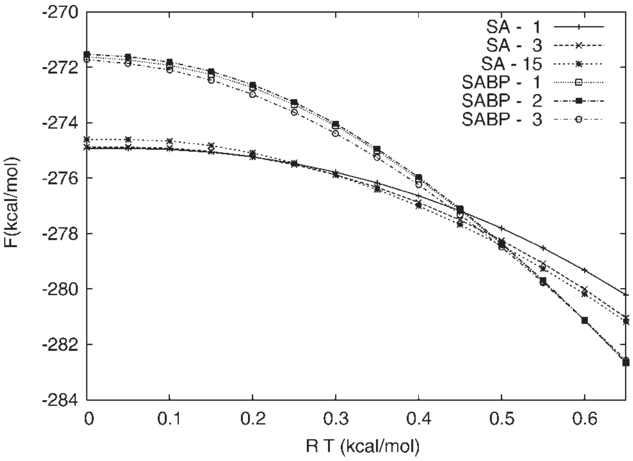

Figure 4
Free energies vs temperature for some of the best ranking sequences as found by SA (at $T = 0$) and SA/BP at $T = 0.6$, and then continued at all temperature with BP performed at fixed sequence. We have checked that BP free energies converge to SA energies at low temperature, suggesting that for these sequences the pair-wise approximation is reliable also at low temperatures. We report the 1st, 3rd, and 15th sequence in the ranking obtained by the SA procedure at $T = 0$, and the 1st, 2nd, and 3rd found by SA/BP at $T = 0.6$. Notice that not only there are crossings between the ranks in each group, but, most importantly, SA/BP finds new and better sequences at $T = 0.6$, and provides the only way to get information of the true ranking at $T = 0.6$: due to the role of conformational entropy, sequences that were extremely low in ranking at $T = 0$, and could be missed by the SA procedure, become the best scoring ones at $T = 0.6$. Therefore, collecting even a huge list of good sequences through ordinary energy minimization procedures is very likely to miss the sequences that will become important at high temperatures.

Table III
The 20 Best Scoring Sequences, as Found by Joint SA/BP Procedure (See Text for Details)

$T = 0.6$	Sequence	$F_{BP}(0.6)$	E (best)	$T = 0$
01	DKKVRARDEFRGRDRRQLFRKGEELEIRREDRYWKQAQDRRGRQGLIDRVEHQ	−281.138	−271.625	06
02	DKKVRARDEFRGRRRQQLFRKGEELEIRREDRYWKQAQDRRGRQGLIDRVEHQ	−281.121	−271.528	08
03	DKRVRARDEFRGRDRRQLFRKGEELEIQREDRYWKQAKDRRGRTGLIDRVEHQ	−281.120	−271.725	03
04	DKRVRARDEFRGRDRRQLFRKGEELEIRREDRYWKQAQDRRGRQGLIDRVEHQ	−281.098	−271.587	07
05	DKRVRARDEFRGRRRQQLFRKGEELEIQREDRYWKQAKDRRGRTGLIDRVEHQ	−281.090	−271.628	05
06	DKKVRARDEFRGRDRRQLFRKGEELEIQREDRYWKQAKDRRGRTGLIDRVEHQ	−281.081	−271.863	01
07	DKRVRARDEFRGRRRQQLFRKGEELEIRREDRYWKQAQDRRGRQGLIDRVEHQ	−281.081	−271.491	09
08	DKKVRARDEFRGRRRQQLFRKGEELEIQREDRYWKQAKDRRGRTGLIDRVEHQ	−281.052	−271.767	02
09	DKKVRARDEFRGRDRRQLFRKGEELEIRREDRRWKQAQDRRGRQGLIDRVEHQ	−281.017	−270.848	18
10	DKKVKARDEFRGRDRRQLFRKGEELEIRREDRYWKQAQDRRGRQGLIDRVEHR	−281.011	−271.478	10
11	DKKVRARDEFRGRRRQQLFRKGEELEIRREDRRWKQAQDRRGRQGLIDRVEHQ	−281.001	−270.752	20
12	DKRVRARDEFRGRDRRQLFRKGEELEIQREDRRWKQAKDRRGRTGLIDRVEHQ	−280.999	−270.948	16
13	DKKVKARDEFRGRRRQQLFRKGEELEIRREDRYWKQAQDRRGRQGLIDRVEHR	−280.995	−271.381	11
14	DKKVRARDEFRGRDRRQLFRKGEELEIRREDRYWKQAQDRRGRQGLIDRVEHA	−280.994	−271.331	12
15	DKRVRARDEFRGRDRRQLFRKGEELEIRREDRRWKQAQDRRGRQGLIDRVEHQ	−280.978	−270.811	19
16	DKKVRARDEFRGRRRQQLFRKGEELEIRREDRYWKQAQDRRGRQGLIDRVEHA	−280.978	−271.234	13
17	DKRVRARDEFRGRRRQQLFRKGEELEIQREDRRWKQAKDRRGRTGLIDRVEHQ	−280.970	−270.852	17
18	DKKVRARDEFRGRDRRQLFRKGEELEIQREDRRWKQAKDRRGRTGLIDRVEHQ	−280.961	−271.087	15
19	DKRVRARDEFRGRRRQQLFRKGEELEIRREDRRWKQAQDRRGRQGLIDRVEHQ	−280.961	−271.234	14
20	DKKVKARDEFRGRDRRQLFRKGEELEIQREDRYWKQAKDRRGRTGLIDRVEHR	−280.952	−271.716	04

Only the 53 mutating positions are reported. The first column reports their absolute rank at $T = 0.6$, where they are calculated. The last column refers to their SA relative rank (that is, their free-energy rank at $T = 0$): notice that this is not their absolute SA rank, since there are hundreds of sequences having an energy, at $T = 0$, between −274.92 (best SA sequence) and −271.62 (best SA/BP sequence). E(best) is the SA energy of the best rotamer configuration found, and corresponds to the free energy at $T = 0$. $F_{BP}(0.6)$ is their free-energy calculated with BP at $T = 0.6$. Energies and temperatures in kcal/mol.

enormously as we raise the energy from the "ground state" of minimal energy, so that it is already impossible to be sure to sample all the configurations within 1 kcal/mol from the best one. We find more than 1000 sequences within 1 kcal/mol from the best SA or SA/BP sequences, and this is probably a small fraction of the real number: the difference between the SA/BP and SA free energies is 3.3 kcal/mol at $T = 0$, and −1.8 kcal/mol at $T = 0.6$, so that the possibility of recovering SA/BP solutions by storing an appropriate number of best scoring SA sequences and then performing BP on each of them, to calculate their free energy at $T = 0.6$, is practically ruled out.

Full design of other 26 proteins, from seven protein families

To see how the results for 1cka extend to other proteins, we have redesigned completely other 26 proteins: the full list is reported in Table IV together with their wild type sequences. We followed the same steps as for 1cka, performing SA simulations with the DESIGNER energy as the cost function to find the best sequence of each protein, and then using BP to evaluate their free energy at $T = 0.6$, and finally comparing the results with those obtained through the combined SA/BP method. The resulting sequences are listed in Tables V and VI, while in Table VII we report the comparison between the free energies, at low temperatures and at $T = 0.6$.

From the latter Table, we see that the main results obtained for 1cka extend to all proteins: the best SA sequence, when studied at $T = 0.6$, is far from optimal: typical energy differences range between 0.3 and 3 kcal/mol, with an average around 2 kcal/mol. The same happens for the best SA/BP sequences that rank low with respect to the SA sequences, when studied at low temperatures (we stop at $T = 0.05$, since we cannot apply BP at exactly $T = 0$: the reported free energy at $T = 0.05$ is therefore expected to be lower than the true $T = 0$ one). These free-energy differences may appear small, but actually we have seen in our simulations that there can be tens of thousands of sequences within 1 kcal/mol from the minimum, and hundreds of thousands within 2 kcal/mol. And indeed, the best SA and SA/BP sequences differ often in many residues, as can be seen from the last column of Table VIII so that they cannot be "guessed" from one another.

The other columns of Table VIII report the alignments of the SA and SA/BP best sequences to the corresponding wild-type ones, obtained with the Needleman-Wunsch global alignment algorithm at the EBI server,[33] with the default BLOSUM62 matrix and the highest values of the gap parameters, to enforce that the sequences keep in register. It can be noticed that the results are similar for the two cases: actually, there is on average a slight increase from the new method, but the change is not statistically relevant. The same holds if we restrict the alignment to just the core positions.

In general, similarity with wild type species is around 35–55% for the whole proteins, and 55–85% for the cores, in both SA and SA/BP cases. This is not surprising, because it is well known that members of the same pro-

Table IV
Wild Type Sequences of the 27 Proteins Considered for Redesign, Grouped According to the Family They Belong To

PDB	WT sequence
	434 repressor
1per	SISSR**V**KSK**R**IQLGLNQAE**LA**QK**V**GTTQQS**I**EQ**LE**NGKTKR***P***R**FL***P*E**LA**SA**L**GVSVDW**L**LNGT
1r69	SISSR**V**KSK**R**IQLGLNQAE**LA**QK**V**GTTQQS**I**EQ**LE**NGKTKR***P***R**FL***P*E**LA**SA**L**GVSVDW**L**LNGT
2cro	TLSER**L**KKR**R**IALK**M**TQTE**LA**TK**A**GVKQQS**IQLIE**AGVTKR***P***RFLFE**IA**M**AL**NCD*P*VW**L**QYGT
	Antifreeze
1ame	NQA**SVVA**NQL**I***P*INTA**L**TLV**MM**RSEVVT*P*V**GI***P*AE**D**I*P*R**L**VSMQ**V**NRA**V***P*LGT**TL**M*P*D**MVKGY**AA
1b7i	NQA**SVVA**NQL**I***P*INTA**L**TLV**MM**RSEVVT*P*V**GI***P*AE**D**I*P*R**L**VSMQ**V**NRA**V***P*LGT**TL**M*P*D**MV**RG**Y**AA
1ekl	NQA**SVVA**NQL**I***P*INTA**L**TLV**MM**RSEVVT*P*V**GI***P*AK**D**I*P*R**L**VSMQ**V**NRA**V***P*LGT**TL**M*P*D**MV**KG**Y**AA
	CI2
1ypc	MKTE**W***P*E**L**VGKSVAA**A**KKV**I**LQDK*P*EAQIIVL*P*VGTIVTMEYRID**RVRLFV**DKLDN**I**AQV***P***R**VG**
2ci2	NLKTE**W***P*E**L**VGKS**V**EE**A**KKV**I**LQD**K***P*E**A**QIIVL*P*VGTIVTMEYRIDR**V**R**L**F**V**DKLDN**I**AEV***P***R**VG**
2sec	KS**F***P*E**V**VGKT**V**DQ**A**REY**F**TLHY*P*QYNVYFL*P*EGS*P*VTLDLRYN**RV**R**V**FYN*P*GTNV**V**NHV***P***H**VG**
	Homeobox
1au7	TTISIA**A**KDA**L**ERH**F**GEHSK***P***SSQE**I**MR**MA**EELNLEKEV**V**RVW**F**CNRRQREKR
1b72	TNFTTR**Q**LTE**L**EKE**F**HFNKYLSRARRVE**IA**AT**L**ELNETQ**V**KI**WF**QN**R**RMKQKK
1b8i	RQTYTRY**Q**TLE**L**EK**EF**HTNHY**L**TRRR**R**IE**MA**HA**L**S**L**TERQ**I**KIW**F**QNRRMKLKKEI
1du0	AFSSEQLAR**L**KR**EF**NENRYLTERRRQQ**L**SSE**L**GLNEAQ**I**KI**WF**ANKRAKIKK
	SH3
1abo	LFV**A**LYDFVASGDNT**L**S**I**TKGEK**L**RVLGYNHNGEW**C**E**A**Q**T**KNGQ**G**W**V*P***SNY**I**T*P*V
1bk2	EL**V**L**A**LYD**Y**QEKS*P*R**EV**T**M**KKGDI**L**T**L**LNSTNKDWWK**V**E**V**NGRQ**G**F**V*PA***AY**V**KKL
1cka	AEY**V**R**A**LFD**F**NGNDEE**DL***P***F**KKGDI**L**R**I**RDK*P*EEQW**W**N**A**EDSEGKR**GMI*P***V*P*Y**V**EKY
1ckb	AEY**V**R**A**LFD**F**NGNDEE**DL***P***F**KKGDI**L**R**I**RDK*P*EEQW**W**N**A**E**D**SEGKR**GMI*P***V*P*Y**V**EKY
1fmk	TT**F**V**A**LYD**Y**ESRTET**DL**S**F**KKG**E**R**L**Q**I**VNNTEGDWWL**A**H**S**LSTGQT**G**Y**I*P***SNY**V**A*P*S
1pwt	MGTGKEL**V**L**A**LYD**Y**QEKS*P*RE**V**T**M**KKGDI**L**T**L**LNSTNKDWWK**V**E**V**NDRQ**G**F**V*PA***AY**V**KKLD
1qcf	RII**V**V**A**LYD**Y**EAIHHED**L**S**F**QKGDQ**MVV**LEESGEWWK**A**R**S**LATRKE**G**Y**I*P***SNY**V**ARV
1sem	ETKF**V**Q**A**LFD**F**N*P*QESGE**L**A**F**KRGDV**I**T**L**INKDD*P*NWWE**G**QLNNRR**G**I**F*P***SNY**V**C*P*YN
1shf	VTLFV**A**LYDYEARTED**DL**S**F**HKG**E**K**F**Q**I**LNSSEGDWWE**A**R**S**LTTGET**G**Y**I*P***SNY**V**A*P*VD
1shg	KEL**V**L**A**LYD**Y**QEKS*P*R**EV**T**M**KKGDI**L**T**L**LNSTNKDWWK**V**E**V**NDRQ**G**F**V***P***A**AY**V**KKLD
2src	TTFV**A**LYD**Y**ESRTET**DL**S**F**KKGER**L**Q**I**VNNTEGDWWL**A**H**S**LSTGQT**G**Y**I*P***SNY**V**A*P*S
	Cold Shock
1c9o	MQR**G**K**V**KW**F**NNEK**G**Y**G**F**I**EVEGGSD**V**F**V**HFT**AI**QGEGFKT**L**EEGQE**V**S**F**EIVQGNRG*P*Q**A**AN**V**VKL
1csp	MLE**G**K**V**KW**F**NSEK**G**F**G**F**I**E**V**EGQDD**V**F**V**HFS**AI**QGEGFKT**L**EEGQ**AV**S**F**EIVEGNRG*P*Q**A**AN**V**TKEA
	Protein G
2igd	MT*P*AVTT**Y**K**L**V**I**N**G**KTLKGETTTK**A**VDAET**A**EKA**F**KQY**A**NDNG**V**DGVWTYDDATKT**F**T**V**TE

Boldface indicates core sites in the native structure, defined by $A^{(N)}/A^{(U)} < 0.1$, where $A^{(N)}$ is the accessible surface area of the residue in the native conformation and $A^{(U)}$ is the corresponding quantity in the model dipeptide representing the unfolded state. Positions in italic are not redesigned (all prolines and ALA4 in 2igd); underlining in 1cka sequence indicates the three residues that are exhaustively mutated in the first two approaches. The pairs 1cka/1ckb, 1per/1r69, 1fmk/2src correspond to different structures (in complex with different molecules) of the same sequences.

tein family can present very little sequence homology. Indeed, a low sequence identity with wild-type species has been reported also in other experiments of protein design.[14]

DISCUSSION

Typically, present methods to design protein sequences and improve their stability look for the best rotamer configuration of the best sequence that can fold on a given scaffold structure: that is, in the best case, they find the sequence that would be optimal if the protein lived and functioned at $T = 0$. However, it is not sure that the results (and in particular, the sequence ranking) at low T are the same at the interesting (room, physiological) temperatures, when alternative rotamer conformations may provide an entropy contribution able to change the scoring of the different sequences. To improve this scheme it is necessary to use the conformational free energy, instead of energy, as a cost function: however, due to the combinatorial number of configurations, even with discrete rotamers, it is not possible to calculate these free energies exactly, so that the new procedure necessarily involves the introduction of approximations.

In this work, we have proposed a new scheme and addressed the following important questions: is the approximated free energy reliable enough to be useful? Does the introduction of the new cost function change the sequence ranking obtained by the traditional scheme? A negative answer to either question would imply that one should go on with the ordinary protocol.

From our results we learn that:

1. For the majority of the sequences the pair approximation for free energy and the BP algorithm appear to be reliable at all investigated temperatures, even if for some sequences the free energy found is wrong (and,

Table V
The Best Scoring Sequences, as Found by Ordinary SA (See Text), For all the Proteins Considered for Complete Redesign

PDB	Best sequence at $T = 0$ (ordinary SA)
1per	DIADE**V**DKK**K**KQDRKQDDD**LA**KQ**V**GADYRK**L**QQ**LV**QKYKREE**FK**Y**LA**RA**L**RKDERQ**L**DRGY
1r69	DIADE**V**KKK**M**QQERKDERR**LA**QD**V**RVQEQE**I**RQ**LI**QQRKKYS**RL**D**LA**RA**L**RKRQQE**L**KYGQ
2cro	DIWRE**L**DEK**K**KQQR**R**RERD**LA**KQ**A**RVQDEE**I**KR**VV**QGYQQYRQIYQ**IA**Q**AI**RQERK**L**KDGK
1ame	DDR**ACKA**QKE**I**QDDR**I**RRD**KL**EQRQDRR**GI**GD**Q**VY**L**ESLK**V**QQK**I**QGK**VL**EK**FV**QY**L**YL
1b7i	DDR**ACKA**QQE**I**RYQE**I**QFR**FL**EQDWVQR**GI**AQ**K**ID**L**YRKR**V**QQK**I**QGR**VL**EK**FV**QS**F**YQ
1ekl	DTR**VCVA**QQE**I**RYDE**I**QRQ**KL**EQRNKKE**GI**AQ**K**ID**L**ERKR**V**QQK**V**QGQ**VL**FY**WV**QY**F**YQ
1ypc	DKQE**W**E**L**QGRRDED**A**KRY**I**QYDLYAEIRKQEGQRDQRQRQQY**KVKLFV**REQQQ**V**ARRK**KG**
2ci2	DRKDR**W**Q**L**QGRR**Y**TD**A**RRE**V**QKQ**K**D**A**RVEQRYGYRQQYQRRDQD**V**K**FFL**DSYYR**V**RRQQ**VG**
2sec	DE**F**E**V**QGRQ**K**DQ**A**ERL**F**KQRYYDRCQWKKYQEEHRKQRY**DV**K**I**FFDYRKQ**V**DWKK**KE**
1au7	DREDQR**K**KQE**L**QKK**F**QQQQQDERE**K**DK**LA**RRLRLKREL**I**DRK**I**RQERQKR
1b72	DRKDRR**Q**DEK**L**KQK**F**QQQQQLDEREKWK**LA**QQ**C**RKDRRQ**V**DE**EL**KR**Q**RRQR
1b8i	SYRQDER**K**KRE**L**QK**KF**YQRQD**L**DDRE**K**QK**LA**QY**L**R**E**DKRY**V**EQF**I**RQQKREQQKRQ
1du0	DEDERQKRE**I**RK**EL**RKYRDVQEQKKRQ**L**AQK**L**RLKREQ**I**ER**KI**EEEKRRQ
1abo	DYE**A**QYEFKQRGSYQ**L**D**I**KQGKE**M**KKEKQDRYDDK**A**E**V**R**C**DWGQ**G**W**V**NRY**I**RR
1bk2	DR**G**K**A**RRQ**F**RQRQW**QV**E**M**YQGEE**V**E**F**KQYRNRYYIE**I**Q**V**RRRR**G**Q**VN**QL**V**QKR
1cka	DKK**V**R**A**RDE**F**RGRQYQ**QLF**YKGEE**L**W**I**KQEDRYW**K**K**A**EVRYRYQ**GLI**DR**V**EHG
1ckb	DYW**V**K**G**RDD**F**RGQRRE**DLF**RRGEQ**L**K**I**EQKDEYW**L**K**A**R**D**RYGRE**GKI**WR**V**EQQ
1fmk	DQ**L**K**G**EYD**F**RKQQYW**QL**D**F**QRG**R**I**L**K**V**WRKDSQRYLF**A**E**K**QDDRRR**G**Y**I**QQQ**V**RR
1pwt	DQYGKQK**A**D**A**RED**F**KKRQYQ**V**E**M**RRGEE**V**E**L**RQRRERRFVQ**V**K**V**RYYQ**G**Q**VK**WK**I**REKQ
1qcf	DKE**V**K**G**KYE**R**RKQDQYE**L**Q**F**RQGER**LKK**KKEQQYWIE**A**E**K**DRNKER**G**K**I**WWL**V**ESV
1sem	SRQQ**V**K**G**RQE**F**KRYYGE**L**E**M**QQGRR**L**E**K**EEERNYWIR**C**KLQYYR**G**K**L**KDR**V**EDR
1shf	DQYYE**G**KYDFRRQYSY**QL**E**F**RQR**D**K**F**K**A**EQYREGRWQQ**C**R**M**EKRRWE**G**K**I**QWQ**V**QDR
1shg	YLE**A**L**A**RNQ**F**DRRQY**QV**R**M**QQGQR**V**E**V**RQYNQRYFVQ**V**K**L**QRYQ**G**Q**VN**WL**I**RQRR
2src	DKCE**G**KYQ**F**WKQQYW**QL**D**F**RQREE**L**K**E**QQYRYREWKQ**A**E**A**QWKKYR**G**Q**I**ARY**L**RK
1c9o	RQQ**G**L**C**KW**F**EKRQ**G**R**G**Q**I**QLWSYRE**I**Q**V**KQE**KI**EGYGSKQ**L**DQGRQ**V**E**L**EVKYESYGV**A**DR**V**KER
1csp	KYK**G**K**C**KD**F**RKQR**G**E**G**W**I**Q**V**SYKRE**I**K**V**ERE**EI**EGYGQKR**L**DKER**QV**K**L**QEKQEYRGK**A**DR**V**QEDQ
2igd	MTVQEF**E**L**Q**I**Q**G**Q**YDKGREKDRK**Q**DREKA**K**REF**Q**KRA**Q**QKRM**R**GWEEEEWRQKRV**KKY**L

Boldface indicates core sites in the native structure. Only the positions considered for redesign are shown.

unfortunately, lower than the true one) at sufficiently low temperatures. We have found this for the three-residue case, and the problem should be more relevant when mutating more residues.

However, in practice this fact does not appear to represent a serious problem: we have seen that a simple consistency test on the BP results allows to eliminate the problematic sequences; moreover, none of the latter would score among the true best ranking sequences. On top of this, the problem seem to be almost irrelevant at $T = 0.6$, as witnessed by the good correlations we get between F_{BP} and F_{ex} at that temperature (see Table I), also without pruning the bad sequences.

Moreover, when we move to the complete redesign case, we see that there is no hint of problems: the SA/BP scheme is useful precisely at room temperature ($T = 0.6$), that was a "safe" temperature in the three-residue case. The possibility of the existence of an incorrectly low scoring sequence, that would dominate over all the others in the SA/BP exploration of the sequence space, is ruled out by noticing that the BP free energies for the best ranking SA and SA/BP sequences, Tables II and III, at low temperatures join smoothly to the corresponding SA solution, representing the free energy at $T = 0$: the best SA/BP sequences appear to be well-behaved sequences in the BP calculations at low temperatures. From this we can conclude that, after performing some consistency test, we can rely on the BP scheme to evaluate the free energy of any given sequence.

2. When relying on the traditional scheme that neglects conformational entropy, as in the SA study we performed, a reasonable number of well scoring sequences should always be collected, and BP should be performed on each of them, to calculate the corresponding free-energy at the temperature of interest: we have seen in Table II and in Figure 4 that the introduction of rotamer entropy and free energy causes crossovers in the rankings, so that it is most likely that some of the less-than-optimal solutions at $T = 0$ (SA) will end up in a better rank, at higher temperatures, than the optimal SA sequence.
3. However, it is likely that even the above mentioned scheme will not yield the true best sequence. From the analysis of 27 proteins we see that sequences found with the ordinary SA scheme, with the energy as the cost function, are not the best scoring ones when we move to realistic, physiological, temperatures and use the free energy as the cost function: we have seen that the SA/BP scheme can find better sequences. Moreover, the gap in the free energies amounts roughly to a few kcal/mol: this is enough to guarantee that it is

Table VI
The Best Scoring SA/BP Sequences, According to Their Free Energy at $T = 0.6$, for All the Proteins Considered for Complete Redesign

PDB	Best sequence at $T = 0.6$ (SA/BP)
1per	DIADE**V**DKK**K**KQDRRQDDD**LA**KQ**V**GADYRQ**L**QQ**LV**QKYKRRR**FL**Y**LA**RA**L**RKDQRQ**L**WQGR
1r69	DIADE**V**KKK**M**QQERKDERR**LA**QD**V**RVQQRE**I**QD**LI**QQRKKRR**RK**D**LA**RA**L**RKREQE**L**DRGR
2cro	DLGQE**L**DQK**K**KQQR**R**RERD**LA**RQ**A**RVRDEE**I**KR**VV**QRRQRRRQIYQ**IA**R**AI**RQERK**L**KDGR
1ame	DDR**AVVA**QKE**I**QDDQ**L**RRD**KM**EQRQVRR**GV**GR**Q**LR**L**ERLK**V**RQK**I**QGK**VL**ER**FI**QY**L**RL
1b7i	DDR**VCKA**QQE**I**QDQQ**I**QRR**FL**EQRNERK**GV**AQ**K**ID**L**ERKK**V**QQK**V**QGQ**VL**FR**WI**QR**L**RL
1ekl	DRR**VCVA**QQE**I**RYDE**I**DRR**KL**EQRNKKE**GV**AQ**K**ID**L**ERKR**V**QQK**V**QGQ**VL**FR**WI**QY**L**RL
1ypc	DKQE**W**E**L**QGQQYRR**G**EQR**I**KRDLYAEIRRQEGQRDQRRRDQR**KVKLFL**DRYQR**I**ARRK**KG**
2ci2	DRKDR**W**Q**L**QGQQ**Y**RR**A**DQE**V**KKQ**K**D**A**RVEQRDGYRYQRRREDRK**V**K**F**F**L**DRYYR**V**RRQQ**VG**
2sec	DR**F**Q**V**QGRQ**K**RQ**A**EQL**F**RRYYYDRVQWQKRYIERRRQRY**DV**K**V**FFDRRQQ**V**DRQQ**KN**
1au7	DRIDRR**K**KQE**L**QKK**F**QQQQRDERE**K**DK**LA**RRLRLKREL**I**DRQ**L**RREWQKR
1b72	DRKDRR**Q**DER**L**KQI**F**QRRQQLDRREKWK**QA**RR**L**RKDQRQ**I**DQ**EI**RK**Q**RQRR
1b8i	DRRKDRR**K**QQE**L**DK**KF**RQRRD**L**DRRE**I**QK**LA**RY**L**R**E**REEE**V**KRM**I**EQKKREQQRRR
1du0	DEDERKKRY**L**RY**EL**RRRRDLQEQKKRQ**L**ARR**L**RDQRRL**Q**EQ**KA**EQEKRRQ
1abo	EYE**A**RREFRQRGRRQ**L**Q**I**KQREK**M**WKRGQDRRDDK**A**Q**V**Q**C**RRGQ**G**W**V**NRY**I**YQ
1bk2	DY**G**L**A**KRQ**F**RQRQW**QV**E**M**YQGER**V**E**F**KQYRNRRYIE**I**Q**V**RRRR**G**Q**VN**QL**V**QKQ
1cka	DKK**V**R**A**RDE**F**RGRDRR**QLF**RKGEE**L**E**I**RREDRYW**K**Q**A**QDRRGRQ**GLI**DR**V**EHQ
1ckb	DRK**V**K**G**RDD**F**RGRRRE**DLF**RRGEE**L**K**I**EQKDEYW**L**K**A**R**D**RYGRE**GKI**WR**V**ERQ
1fmk	DQ**L**K**G**ERD**F**RKRDRR**QL**D**F**QRG**RIL**Q**V**WRKDRQRYLF**A**Q**K**QDDRRR**G**Y**I**QRY**V**RR
1pwt	DRRGKQK**A**R**A**ERD**F**RQRQYQ**V**E**M**RQGEE**V**E**L**RQRRERRFVQ**V**K**V**RRRQ**G**Q**VK**WQ**I**RERQ
1qcf	DKE**V**K**G**KYQ**Q**RKRDQRE**L**Q**F**QQGER**LKK**KKEQQRWIE**A**E**K**DRNKER**G**K**I**WDK**V**ESV
1sem	SRRQ**VEG**QYE**F**RRYRGE**L**E**M**RQGEQ**L**E**L**EEERNRWIR**C**KKDRRR**G**K**L**KDR**V**KKR
1shf	DQYYE**G**KRDFRQRYRY**QL**Q**F**RQR**D**K**F**K**A**EQYREGRWQQ**C**R**V**EKRRQE**G**K**I**QWQ**V**QDR
1shg	YLE**A**L**A**QRQ**F**RRRQY**QV**Q**V**KQGER**V**E**V**RQYRQRYFVQ**I**K**V**QRWQ**G**Q**VN**WI**I**RQRR
2src	DKLE**G**KYQ**F**RKRQRW**QL**D**F**RQREE**L**K**V**EQRRYREWFQ**A**K**A**EWKRQQ**G**Q**I**ARY**V**RK
1c9o	DQQ**G**L**C**KW**F**EKRQ**G**R**G**Q**I**QLWRYRE**I**Q**V**KRE**KI**EGRGRKQ**L**DQRRQ**VKL**DVRYERYGV**A**DR**V**QER
1csp	KYK**G**K**C**DD**F**RKRR**G**E**G**W**I**Q**V**YRKRR**I**K**V**EQR**AI**EGRGRKQ**L**DRER**QV**K**L**QERQEREGK**A**ER**V**QEDQ
2igd	MTVQEF**ELEIR**G**Q**RDRGREKDRA**Q**QYQEA**K**KRF**E**QRC**R**QKRM**R**GWDELRRDQRKV**R**K**W**L

Boldface indicates core sites in the native structure. Only the positions considered for redesign are shown.

Table VII
Free Energies of the Best Ranking Sequences, at Different Temperatures

		Best seqs. at $T = 0$		Best seqs. at $T = 0.6$		Differences	
PDB code	aa.	$E(T = 0.0)$	$F(T = 0.6)$	$F(T = 0.6)$	$F(T = 0.05)$	Col.6-Col.3	Col.5-Col.4
1per	61	−319.284	−325.240	−326.829	−317.614	1.67	−1.588
1r69	61	−330.966	−338.399	−340.434	−330.189	0.78	−2.035
2cro	61	−317.902	−326.763	−327.871	−314.451	3.45	−1.108
1ame	59	−310.705	−314.393	−316.729	−307.656	3.05	−2.336
1b7i	59	−310.294	−313.632	−316.471	−308.194	2.10	−2.839
1ekl	59	−311.113	−313.541	−315.225	−308.550	2.56	−1.685
1ypc	60	−325.246	−330.633	−332.015	−321.047	4.20	−1.382
2ci2	61	−332.135	−337.301	−339.828	−329.865	2.27	−2.527
2sec	57	−321.012	−325.126	−327.815	−316.213	4.80	−2.688
1au7	50	−289.160	−297.139	−297.502	−287.397	1.76	−0.364
1b72	51	−294.026	−300.951	−302.467	−291.082	2.94	−1.516
1b8i	56	−320.844	−327.314	−328.490	−314.141	6.70	−1.177
1du0	50	−282.968	−289.608	−290.355	−278.602	4.37	−0.746
1abo	53	−273.124	−276.950	−280.003	−270.015	3.11	−3.053
1bk2	53	−286.485	−292.975	−293.479	−286.312	0.17	−0.504
1cka	49	−274.923	−279.319	−281.138	−271.625	3.30	−1.819
1ckb	53	−292.211	−297.951	−299.411	−291.681	0.53	−1.460
1fmk	55	−288.970	−294.062	−296.136	−285.961	3.01	−2.074
1pwt	59	−318.475	−324.762	−328.116	−316.611	1.86	−3.355
1qcf	56	−306.657	−311.044	−312.965	−306.508	0.15	−1.921
1sem	54	−295.547	−301.637	−303.783	−294.193	1.35	−2.146
1shf	57	−304.566	−310.103	−311.656	−302.631	1.94	−1.553
1shg	55	−282.082	−286.572	−288.847	−281.134	0.95	−2.275
2src	55	−298.756	−302.574	−305.187	−297.034	1.72	−2.613
1c9o	65	−339.988	−344.792	−346.975	−336.689	3.30	−2.184
1csp	66	−338.401	−342.871	−346.042	−335.731	2.67	−3.171
2igd	59	−329.106	−335.458	−337.854	−327.055	2.05	−2.397

Column 2 reports the number of redesigned sites. Column 3 contains the energy of the best rotamer configuration of the best sequence found in the standard SA approach with DESIGNER energy, while column 4 lists the free energy of that sequence calculated with BP at $T = 0.6$. Column 5 reports the free energies of the best sequences found by the SA/BP procedure at $T = 0.6$, while column 6 reports the free energies of the same sequences, calculated with BP at $T = 0.05$. The last two columns show the differences between the corresponding low and high temperature results. It can be noticed that SA sequences are always suboptimal at $T = 0.6$, and SA/BP sequences perform poorly at low temperatures. Energies and temperatures in kcal/mol.

Table VIII
Alignment to Wild Type Sequences of the Best Ranking Sequences Obtained by SA (Columns 3–5) and SA/BP (Columns 6–8)

		Best seqs. at $T = 0$			Best seqs. at $T = 0.6$			Differences			SA-SA/BP
PDB code	aa.	Ident.	Simil.	Score	Ident.	Simil.	Score	C.6-C.3	C.7-C.4	C.8-C.5	Ident.
1per	61	0.279	0.426	47	0.328	0.475	66	0.049	0.049	19	52/61
1r69	61	0.344	0.443	59	0.344	0.459	62	0.000	0.016	3	50/61
2cro	61	0.213	0.525	49	0.230	0.541	54	0.017	0.016	5	49/61
1ame	59	0.169	0.390	19	0.220	0.424	39	0.051	0.034	20	44/59
1b7i	59	0.220	0.407	35	0.186	0.356	19	−0.034	−0.051	−16	36/59
1ekl	59	0.237	0.373	23	0.203	0.373	20	−0.034	0.000	−3	50/59
1ypc	60	0.317	0.500	61	0.300	0.483	58	−0.017	−0.017	−3	39/60
2ci2	61	0.230	0.492	63	0.213	0.459	54	−0.017	−0.033	−9	46/61
2sec	57	0.211	0.368	26	0.211	0.421	35	0.000	0.053	9	36/57
1au7	50	0.200	0.440	37	0.200	0.440	38	0.000	0.000	1	43/50
1b72	51	0.176	0.431	24	0.196	0.431	27	0.020	0.000	3	34/51
1b8i	56	0.179	0.464	33	0.214	0.464	41	0.035	0.000	8	33/56
1du0	50	0.200	0.500	40	0.160	0.380	10	−0.040	−0.120	−30	33/50
1abo	53	0.283	0.491	62	0.302	0.472	67	0.019	−0.019	5	33/53
1bk2	53	0.208	0.509	38	0.226	0.528	46	0.018	0.019	8	47/53
1cka	49	0.340	0.509	66	0.396	0.585	102	0.056	0.076	36	39/53
1ckb	53	0.396	0.566	103	0.396	0.547	102	0.000	−0.019	−1	48/53
1fmk	55	0.200	0.345	38	0.236	0.345	48	0.036	0.000	10	45/55
1pwt	59	0.220	0.508	43	0.237	0.525	54	0.017	0.017	11	47/59
1qcf	56	0.214	0.411	30	0.232	0.411	41	0.018	0.000	11	48/56
1sem	54	0.204	0.407	23	0.222	0.481	52	0.018	0.074	29	37/54
1shf	57	0.228	0.351	35	0.228	0.351	34	0.000	0.000	−1	50/57
1shg	55	0.164	0.436	19	0.164	0.473	22	0.000	0.037	3	43/55
2src	55	0.182	0.345	24	0.236	0.400	54	0.054	0.055	30	41/55
1c9o	65	0.308	0.523	86	0.277	0.492	70	−0.031	−0.031	−16	54/65
1csp	66	0.288	0.545	78	0.273	0.485	67	−0.015	−0.060	−11	50/66
2igd	59	0.254	0.424	40	0.220	0.390	30	−0.034	−0.034	−10	35/59

Identities, similarities and scores are calculated aligning the mutating sites (whose number is reported in column 2) by the Needleman-Wunsch global alignment algorithm at the EBI server,[33] with the default BLOSUM62 matrix and the highest values of the gap parameters (open gap penalty = 100.0, gap extension penalty = 10), to enforce that the sequences keep in register. Columns 9–11 report the differences between corresponding values from the two methods; column 12 lists the fraction of identical residues in SA and SA/BP solutions. It can be noticed that the quality of alignment is similar for the two cases, and that the SA and SA/BP solution differ tipically in a significant number of residues.

very unlikely to find SA/BP solutions among the saved SA ones, and viceversa, also if one stores a very long list of well-scoring sequences. Simply, entropy plays a fundamental role in determining the sequence ranking, and a more correct thing to do is to use the SA/BP scheme.

These observations are somewhat different from (and, in a sense, complementary to) the conclusions drawn by Hu and Kuhlman, that used a different potential and a different optimization technique.[34] Because of the different choice of the energy function, we cannot compare directly our results to theirs, which suggest that conformational entropy is not the main determinant of the aminoacid preferences at different positions: a statement that substantially holds also in the present case. However, our results demonstrate that the introduction of conformational entropy within a pair approximation, where the residues at different positions are not considered independent and covariant motion is accounted for, affects the design results, in that it deeply changes the rank of the best scoring sequences, so that the sequence identity between SA and SA/BP best results can be as low as 60% of the mutating residues.

CONCLUSIONS

We have studied the effect of introducing side-chain entropy in automated sequence design procedures using in particular the DESIGNER potential and 1cka as a model protein. The DESIGNER potential resorts to a detailed atomistic model to provide an accurate estimate of the free energy difference, with respect to the unfolded state, of any possible side-chain arrangement on top of a given native main-chain geometry. At present, the DESIGNER methodology disregards conformational entropy coming from different side-chain arrangements.

The picture that emerges from this study is that side-chain entropy cannot be neglected in protein sequence design, if one is interested in finding the best sequences.

The method we use to introduce rotamer entropy (whose exact evaluation is out of reach for any realistic protein length) resorts to a pair-approximation expression for the rotamer free energy, and the related BP Algorithm to find self-consistent solutions for the rotamer probability distributions at each site. The method itself does not depend on the particular choice of the DESIGNER potential as the energy function, and can be applied to other potentials as well. Studying the mutations of just three positions (a case amenable to exact solution) we demonstrate that this procedure produces, in general, reliable results: the failures, that represent a small fraction of all cases, can be identified using a simple criterion, and affect unimportant sequences.

Calculating free energies for the sequences recorded through "ordinary" simulated annealing, with the energy as the cost function to be minimized, already changes their ranking, suggesting that the rotamer free energy should always be calculated, on top of the outcome of the usual energy minimization, if one wants to find better candidate sequences to test in experiments. However, if the aim is to find the best sequences, the rotamer free energy should be used from the very beginning, as the function to minimize instead of energy. This yields sequences with an improved stability of around 2 kcal/mol over the value obtained calculating the BP free-energy of the best-ranking energy-minimized sequences; such SA/BP sequences differ typically in many positions from the corresponding SA solutions, while they present on average the same degree of similarity to the wildtype sequences, both when the whole sequence or just the core positions are considered.

Our findings are in line with several results from the other authors, that underline the necessity of taking into account conformational entropy, not only to estimate protein stability[15] but also in predictions of protein interactions,[35,36] in structure refinement[15,37] and possibly in structure prediction.[38] Therefore, it is most likely that future refinements of the methods and potentials, including some flexibility of the main chain and a more realistic description of the unfolded reference state, will increase the role of side-chain entropy in protein design, so that the necessity of using the conformational free energy as the cost function will become even more compelling. We believe that it is worth devoting some effort to find more efficient search schemes to make the rotamer free energy a standard ingredient of a fast sequence optimization protocol.

ACKNOWLEDGMENTS

The author A. J. thanks M. Karplus, M. A. Santos, M. Suarez, P. Tortosa and S. Wodak for useful discussions.

REFERENCES

1. Wernisch L, Hery S, Wodak SJ. Automatic protein design with all atom force-fields by exact and heuristic optimization. J Mol Biol 2000;301:713–736.
2. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science 1997;278:82–87.
3. Street AG, Mayo SL. Computational protein design. Structure Fold Des 1999;7:R105–R109.
4. Koehl P, Levitt M. De novo protein design. I. In search of stability and specificity. J Mol Biol 1999;293:1161–1181.
5. Koehl P, Levitt M. De novo protein design. II. Plasticity in sequence space. J Mol Biol 1999;293:1183–1193.
6. Raha K, Wollacott AM, Italia MJ, Desjarlais J. Prediction of amino acid sequence from structure. Prot Sci 2000;9:1106–1119.
7. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 2000;97:10383–10388.
8. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science 2003;302:1364–1368.
9. Jaramillo A, Wernisch L, Hery S, Wodak SJ. Automatic procedures for protein design. Comb Chem and High Throughput Screen 2001;4:643–659.
10. Karplus M, Ichiye T, Petit BM. Configurational entropy of native proteins. Biophys J 1987;52:1083–1085.
11. Jaramillo A, Wodak SJ. Computational protein design is a challenge for implicit solvation models. Biophys J 2005;88:156–171.
12. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35:133–152.
13. Berezovsky IN, Chen WW, Choi PJ, Shakhnovich EI. Entropic stabilization of proteins and its proteomic consequences. PLoS Comput Biol 2005;1:e47.
14. Jin W, Kambara O, Sasakawa H, Tamura A, Takada S. De novo design of foldable proteins with smooth folding funnel: automated negative design and experimental verification. Structure 2003;11: 581–590.
15. Zhang J, Liu JS. On side-chain conformational entropy of proteins. PLoS Comput Biol 2006;2:e168.
16. Koehl P, Delarue M. Application of a self consistent mean field theory to predict protein side-chain conformations and estimate their conformational entropy. J Mol Biol 1994;239:249–275.
17. Lopez de la Paz M, Lacroix E, Ramirez-Alvarado M, Serrano L. Computer-aided design of beta-sheet peptides. J Mol Biol 2001; 312:229–246.
18. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
19. Jaramillo A, Wernisch L, Hery S, Wodak SJ. Folding free energy function selects native-like protein sequences in the core but not on the surface. Proc Natl Acad Sci USA 2002;99:13554–13559.
20. Ogata K, Jaramillo A, Cohen W, Briand JP, Connan F, Choppin J, Muller S, Wodak SJ. Automatic sequence design of major histocompatibility complex class I binding peptides impairing CD8+ T cell recognition. J Biol Chem 2003;278:1281–1290.
21. Tortosa P, Jaramillo A. Active sites by computational protein design. In: Clemente-Gallardo J, Moreno Y, Saenz-Lorenzo JF, Velazquez-Campoy A, editors. From physics to biology vol. 851. Melville, NY: AIP Conference Proceedings, 2006. pp 96–101. Proceedings of the II BIFI 2006 International Conference, Feb 8–11, Zaragoza Spain, 2006.
22. MacKerell AD, Jr, Bashford D, Bellott M, Dunbrack RL, Jr, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, III, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102: 3586–3616.
23. Nakagawa N, Peyrard M. The inherent structure landscape of a protein. Proc Natl Acad Sci USA. 2006;103:5279–5284.
24. Brady GP, Sharp KA. Entropy in protein folding and in protein–protein interactions. Curr Opin Struct Biol 1997;7:215–221.

25. Kikuchi R. A theory of cooperative phenomena. Phys Rev 1951;81: 988–1003.
26. An G. A note on the cluster variation method. J Stat Phys 1988; 52:727–734.
27. Pelizzola A. Cluster variation method in statistical physics and probabilistic graphical models. J Phys A 2005;38:R309–R339.
28. Moore GL, Maranas CD. Identifying residue–residue clashes in protein hybrids by using a second-order mean-field approach. Proc Natl Acad Sci USA 2003;100:5091–5096.
29. Yanover C, Weiss Y. Approximate inference and protein-folding. In: Becker S, Thrun S., Obermayer K, editors. Advances in neural information processing systems 15. Cambridge (MA): MIT Press; 2003. AP08. Proceedings of the Neural Information Processing Systems (NIPS2002) Conference, December 9–14; Vancouver, Whistler, (Canada), 2002.
30. Yanover C, Weiss Y. Finding the M most probable configurations in arbitrary graphical models. In: Thrun S, Saul LK, Schölkopf B, editors. Advances in neural information processing systems 16. Cambridge, MA: MIT Press; 2004. pp 289–296. Proceedings of the Neural Information Processing Systems (NIPS2003) Conference, 2003 Dec 8–13; Vancouver, Whistler (Canada).
31. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inferences. San Francisco (CA): Morgan Kaufmann; 1988. 552p.
32. Yedidia JS, Freeman WT, Weiss Y. Generalized belief propagation. In: Leen TK, Dietterich TG, Tresp V, editors. Advances in neural information processing systems 13. Cambridge, MA: MIT Press; 2001. pp 689–695. Proceedings of the Neural Information Processing Systems (NIPS2000) Conference, November 27–Dec 2, Denver, Beckenridge, USA, 2000.
33. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 2000;16:276–277.
34. Hu X, Kuhlman B. Protein design simulations suggest that side-chain conformational entropy is not a strong determinant of amino acid environmental preferences. Proteins: Struct Funct Bioinf 2006;62:739–748.
35. Ma B, Elkayam T, Wolfson H, Nussinov R. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci USA 2003;100:5772–5777.
36. Frederick KK, Marlow MS, Valentine KG, Wand J. Conformational entropy in molecular recognition by proteins. Nature 2007;448:325–330.
37. Shapovalov MV, Dunbrack RL, Jr. Statistical and conformational analysis of the electron density of protein side chains. Proteins 2007;66:279–303.
38. Lin MS, Fawzi NL, Head-Gordon T. Hydrophobic potential of mean force as a solvation function for protein structure prediction. Structure 2007;15:727–740.