WILEY InterScience®
DISCOVER SOMETHING GREAT

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

A coarse-grained protein force field for folding and structure prediction

Julien Maupetit,[1] P. Tuffery,[1] and Philippe Derreumaux[2*]

[1] Equipe de Bioinformatique Génomique et Moléculaire, INSERM E0346, Université Paris 7, Tour 53–54, 2 place Jussieu, 75251 Paris, Cedex 05, France

[2] Laboratoire de Biochimie Théorique, UPR 9080 CNRS, Institut de Biologie Physico-Chimique et Université Paris 7, 13 rue Pierre et Marie Curie, 75005 Paris, France

ABSTRACT

We have revisited the protein coarse-grained optimized potential for efficient structure prediction (OPEP). The training and validation sets consist of 13 and 16 protein targets. Because optimization depends on details of how the ensemble of decoys is sampled, trial conformations are generated by molecular dynamics, threading, greedy, and Monte Carlo simulations, or taken from publicly available databases. The OPEP parameters are varied by a genetic algorithm using a scoring function which requires that the native structure has the lowest energy, and the native-like structures have energy higher than the native structure but lower than the remote conformations. Overall, we find that OPEP correctly identifies 24 native or native-like states for 29 targets and has very similar capability to the all-atom discrete optimized protein energy model (DOPE), found recently to outperform five currently used energy models.

Proteins 2007; 69:394–408.
© 2007 Wiley-Liss, Inc.

Key words: protein; coarse-grained model; thermodynamics; kinetics; structure prediction; folding.

INTRODUCTION

Designing an accurate protein force field aimed at recognizing the native topology from decoys and predicting the equilibrium structures from solely amino acid sequences has been the subject of extensive efforts for decades.[1,2] The most expensive energy models for protein structure prediction and folding work with all-atom molecular mechanics[3–6] and knowledge-based force fields.[7,8] All-atom Monte-Carlo (MC), ensemble molecular dynamics (MD), and replica exchange molecular dynamics (REMD) simulations using both types of force fields have succeeded in folding a limited set of small proteins with 40 residues. It is unclear at this time whether a wide range of topologies can be reached to an acceptable level of accuracy.

Given the increase in the number of genomic sequences and the plasticity of proteins upon molecular association, it is desirable to reduce the number of degrees of freedom used by the all-atom models and to increase the space of conformations that can be explored using current computer resources. Reduced models offer the possibility to achieve both tasks[9,10] and several coarse-grained protein models have been developed.[11] The one-bead Cα model is currently used by the Gaussian network model to explore the conformational changes around the native state.[12] Two-bead Cα-Sc models[13,14] or three-bead Cα-Sc-Pep models, where Pep is an interaction center of the backbone, have been used for recognition,[1] folding in water[15] and lipid environments,[16] binding,[17] and assembly of membrane proteins.[18] Cα-Sc1-Sc2 models with two beads for the side-chain,[19–21] four-bead models consisting of N, Cα, C, and Sc atoms and six-bead models have been used for structure prediction and folding,[22–24] protein-protein docking,[21] and the study of aggregation mechanisms of amyloid-forming proteins.[25–28]

Finding a good compromise between energy accuracy, structural precision, number of degrees of freedom, and computational cost has been been widely investigated.[11,29] Mirny and Shakhnovich pointed to the limitation of lattice protein models and 210 pairwise contact potentials to achieve significant gaps (conditions for fast folding) between native and non-native conformations of 100 proteins.[30] Qiu and Elber[31] developed atomically detailed pairwise potentials for numbers of atom types varying from 32 to 46 and found similar recognition capacities to the potential of Lu and Skolnick[32] based on 167 atom types. Buchette *et al.* showed that the introduction of

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsors: Society Alzheimer of Canada; The Sixth European PCRD; The Centre National de la Recherche Scientifique; Université of Paris VII.

*Correspondence to: Philippe Derreumaux, Laboratoire de Biochimie Théorique, UPR 9080 CNRS, Institut de Biologie Physico-Chimique et Université Paris 7, 13 rue Pierre et Marie Curie, 75005 Paris, France. E-mail: Philippe.derreumaux@ibpc.fr

Received 13 December 2006; Revised 13 February 2007; Accepted 9 March 2007

Published online 28 June 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21505

explicit orientation dependence in a coarse-grained, residue-level model improves the ability of pairwise potentials to recognize the native state.[1] Similarly, the impact of multibody interactions ranging from 4-body[33,34] to higher orders[35] is still being examined.

This work revisits the coarse-grained OPEP energy function. Such a coarse graining, based on a six-bead model per amino acid, yields very high structural precision of the backbone and was found to discriminate native from *ab initio*-generated structures of peptides.[23] Coupled to MC simulations, OPEP predicted the native structure of the 46-residue three helix-bundle from protein A,[36] helical hairpins,[37] and the 56-residue domain B1 of protein G,[38] within 3.0 Å Cα root-mean square deviations (cRMSd) from experiments. OPEP was also used to study peptide folding and aggregation. Using the activation–relaxation technique (ART),[39,40] folding pathways of a β-hairpin model were consistent with MD and MC methods,[41] and aggregation of amyloid-forming peptides revealed reptation moves of the chains, in agreement with isotope-edited IR spectroscopy study.[42]

Despite these successes, there are two body of data indicating that the OPEP parameters are not optimal for proteins. Firstly, MC simulations suggested that a βαβ supersecondary motif with two parallel β-strands joined by a single α-helix could be a folding unit by itself,[36] but this was not confirmed by CD and NMR experiments.[43] Secondly, the side-chain parameters were not learned using realistic pair distance probability distributions; side-chain pairs including cysteine were not considered and all other pairs were poorly populated.[23]

Here, our goal is to test OPEP on a more significant number of proteins and to determine whether it can discriminate native or native-like from non-native conformations. This article is organized as follows. The OPEP force field and the protein chain are first reviewed with emphasis on the change with respect the original version. We then describe the optimization procedure used, and the training and validation sets which include 29 targets and 28,553 conformations. Finally, we analyze the recognition performances of OPEP and discuss the physical properties of the parameters.

METHODS

Reduced protein model

The coarse-grained off-lattice model we use has slightly evolved from the initial work.[23] It still consists in a detailed representation of the backbone, modelled by its N, H, Cα, C′, O atoms (denoted as M for main chain) and in one bead or centroid for all side-chains (Sc), except the proline amino acid which is represented by all heavy atoms (Fig. 1).

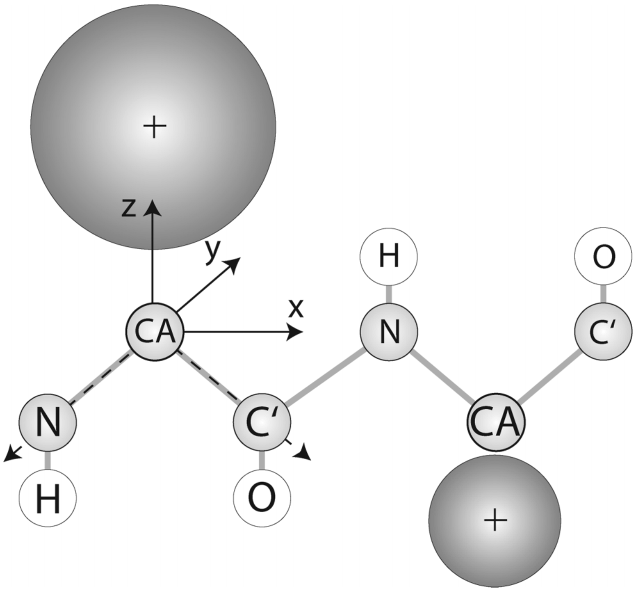


Figure 1

***OPEP coarse-grained representation**. All amino acid side-chains but proline are represented by one centroid in OPEP force field. Their positions, defined with respect to the backbone heavy atoms (N, CA, C′), and their van der Waals radii vary from one residue to another (see Table I).*

However, in contrast to the initial model, where each centroid was defined by its bond length, virtual bond angle, and improper dihedral angle with respect to the Cα atoms, each side-chain is now defined with respect to its N, Cα, and C′ atoms (see Table I). The positions and Van der Waals radii were calculated using a dataset of 2,248 PDB structures[44] sharing less than 30% sequence identity. The centroids were located at the mass center of the heavy atoms and calculated using the rotamer distribution of each side-chain. The Van der Waals radii were obtained by minimizing the sum over the 210 interactions of the squares of the difference between the mean inter centroid distance observed in the 2,248 PDB structures and the calculated value for each interaction type. Here, we have considered that two all-atom nonsequential side-chains are in contact if at least one distance is less than the sum of the van der Waals atomic radii +1 Å. Overall, the values of the radii agree well with previously published sets,[9] but such a procedure does not prevent the occurrence of severe collisions within our database of 2,248 structures (3.5% among all contacts).

OPEP energy function

The terms for modeling the short-range and long-range interactions in the OPEP function (version 3) are the same as in OPEP versions 1.0[23,36] and 2.0.[37] In particular, OPEP includes cooperative 4-body hydrogen-

Table I

Optimized Structural Parameters of the Side-Chain Centroids

Res	r_0 (Å)	r_{CA-SC} (Å)	N. CA. Sc	Sc. CA. C
ALA	2.287	1.52	116.60	111.10
CYS	2.426	1.95	108.50	117.65
ASP	2.707	2.14	110.74	119.16
GLU	2.968	2.77	110.95	120.87
PHE	3.33	2.62	110.72	124.15
GLY	—	—	—	—
HIS	3.078	2.60	109.93	124.39
ILE	2.939	2.27	109.16	118.96
LYS	3.143	3.11	112.92	121.57
LEU	2.928	2.40	112.72	124.94
MET	2.999	2.70	113.54	121.70
ASN	2.797	2.16	109.66	121.47
PRO	2.669	1.81	68.77	133.11
GLN	3.01	2.76	111.24	122.32
ARG	3.411	3.59	112.07	119.79
SER	2.334	1.77	106.78	107.98
THR	2.675	1.90	107.21	114.11
VAL	2.880	1.44	126.10	99.66
TRP	3.67	2.86	117.40	119.81
TYR	3.382	2.79	112.54	123.56

r_0 is the Van der Waals radius (Å). $r_{\mathrm{CA-Sc}}$ is the distance between the C_α and the centroid of the side-chain (Å). $N.\widehat{CA}.Sc$ and $Sc.\widehat{CA}.C'$ (degrees) are the valence angles involving respectively *N*, C_α, and the side-chain centroid, and the side-chain centroid, C_α and C' atoms. Values are given for each residue-type. For simplicity, we also give the one-bead parameters for the proline side-chain, although it is modelled by all heavy atoms in OPEP version 3.

bonding terms in an attempt to reproduce the strong co-operative nature of the amide H-bonds as demonstrated by quantum mechanical calculations,[45] and takes into account the propensities of the residues for α and β states, which are important in protein design or in predicting the aggregation rates of amyloid-forming proteins.[46] The analytic forms of the hydrogen-bonding and backbone torsional potentials have been modified since then to make them derivable, and have been used in many studies of protein folding and aggregation.[41,42,47,48] Only the parameters are therefore changed in the present exercise. OPEP (Optimized Potential for Efficient structure Prediction) version 3 is expressed as a sum of local, nonbonded and hydrogen-bond (H-bond) terms:

$$E = E_{\mathrm{local}} + E_{\mathrm{nonbonded}} + E_{\mathrm{H-bond}} \qquad (1)$$

Local potentials

The local potentials are expressed by:

$$E_{\mathrm{local}} = w_{\mathrm{b}} \sum_{\mathrm{bonds}} K_{\mathrm{b}}(r - r_{\mathrm{eq}})^2 + w_{\mathrm{a}} \sum_{\mathrm{angles}} K_{\alpha}(\alpha - \alpha_{\mathrm{eq}})^2$$
$$+ w_{\Omega} \sum_{\mathrm{imp-torsions}} k_{\Omega}(\Omega - \Omega_{\mathrm{eq}})^2 + w_{\phi,\psi}\left(\sum_{\phi} E_{\phi} + \sum_{\psi} E_{\psi}\right) \qquad (2)$$

The term E_{local} contains force constants associated with changes in bond lengths and bond angles of all particles as well as force constants related to changes in improper torsions of the side-chains and the peptide bonds. The force constants and equilibrium values associated with the main chain atoms are taken from AMBER,[49] the force constants associated with the side-chains are very similar to those in AMBER. The E_ϕ and E_ψ potentials attempt to generate Ramachandran plot of reduced protein structures in agreement with all-atom protein structures, and are expressed by the following quadratic polynomials:

$$E_{\phi} = k_{\phi\psi}(\phi - \phi_o)^2 \qquad (3)$$

$$E_{\psi} = k_{\phi\psi}(\psi - \psi_o)^2 \qquad (4)$$

where $\phi_0 = \phi$ within the interval $[\phi_{\mathrm{lower}},\phi_{\mathrm{upper}}]$ and $\phi_0 = \min(\phi - \phi_{\mathrm{lower}},\, \phi - \phi_{\mathrm{upper}})$, otherwise, with $\phi_{\mathrm{lower}} = -160°$, and $\phi_{\mathrm{upper}} = -60°$, respectively. Similarly, we use $\psi_{\mathrm{lower}} = -60°$ and $\psi_{\mathrm{upper}} = 160°$ in Eq. (4). We emphasize that these terms do not prevent sampling of conformations covering all values of ϕ and ψ.

Nonbonded potentials

The nonbonded potentials are expressed by:

$$E_{\mathrm{nonbonded}} = w_{1,4} \sum_{1,4} E_{\mathrm{VdW}} + w_{C\alpha,C\alpha} \sum_{C\alpha,C\alpha} E_{\mathrm{VdW}}$$
$$+ w_{1>4} \sum_{M',M'} E_{\mathrm{VdW}} + w_{1>4} \sum_{M',C\alpha} E_{\mathrm{VdW}} + w_{1>4} \sum_{M,\mathrm{Sc}} E_{\mathrm{VdW}}$$
$$+ \sum_{\mathrm{Sc,Sc}} w_{\mathrm{Sc,Sc}}\, E_{\mathrm{VdW}} \qquad (5)$$

with 1,4 the 1–4 interactions along each torsional degree of freedom, M' the N, C′, O, and H main chain atoms, and Sc the side-chain. As seen, we separate short-range from long-range $(j > i + 4)$ interactions, and the $C\alpha$ atom from the other main chain atoms.

E_{VdW} is defined as:

$$E_{\mathrm{VdW}} = \varepsilon_{ij}\left(\left(\frac{r_{ij}^0}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^0}{r_{ij}}\right)^{6}\right) H(\varepsilon_{ij}) - \varepsilon_{ij}\left(\frac{r_{ij}^0}{r_{ij}}\right)^{6} H(-\varepsilon_{ij}) \qquad (6)$$

Here the Heavyside function $H(x) = 1$ if $x \geq 0$ and 0 if $x < 0$, r_{ij} is the distance between particles *i* and *j*, $r_{ij}^0 = (r_i^0 + r_j^0)/2$ with r_i^0 the Van der Waals radius of particle *i*. The radius is given in Table I for the side-chains and is set to the following values for backbone atoms: $r_N^0 = 1.75$ Å, $r_H^0 = 1.00$ Å, $r_{C\alpha}^0 = 2.385$ Å, $r_{C'}^0 = 1.85$ Å, and $r_O^0 = 1.60$ Å. All nonbonded interactions are included (no cut-off distance).

In this work, $H(x)$ is set to 1 and a 12-6 potential is used for all interactions, except between two side-chains. For the side-chains, $H(x)$ is set to 1 if the interaction is hydrophobic in character or results from oppositely charged residues; otherwise $H(x) = 0$ and the repulsive

-6 potential is used. The initial ε_{ij}'s between (M′, M′), (M, Sc), (M′, Cα) and (Cα, Sc) particles are set to 0.005 kcal/mol, and those between (Cα, Cα) to 0.4 kcal/mol. The initial ε_{ij}'s between (Sc, Sc) are taken from the work of Betancourt and Thirumalai,[50] which refined the contact matrix first proposed by Skolnick *et al.* based on a total of 224 non homologous proteins.[51] Among a total of 210 interactions, 93 pairs are thus governed by the repulsive -6 potential.

Hydrogen-bonding potentials

The hydrogen-bonding potential ($E_{\text{H-bond}}$) consists of two-body (E_{HB1}) and four-body (E_{HB2}) terms.[28] Two-body H-bonds are defined by:

$$E_{\text{HB1}} = w_{\text{hb1}-4} \sum_{ij,j=i+4} \varepsilon_{\text{hb1}-4}\mu(r_{ij})\nu(\alpha_{ij}) + w_{\text{hb1}>4} \sum_{ij,j>i+4} \varepsilon_{\text{hb1}>4}\, \mu(r_{ij})\nu(\alpha_{ij}) \quad (7)$$

where,

$$\mu(r_{ij}) = 5\left(\frac{\sigma}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma}{r_{ij}}\right)^{10} \quad (8)$$

$$\nu(\alpha_{ij}) = \begin{cases} \cos^2\alpha_{ij}, & \alpha_{ij} > 90^\circ \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

The sum is over all residues i and j separated by $j = i+4$ and $j > i+4$ (helices 3_{10} are thus excluded), r_{ij} is the O..H distance between the carbonyl oxygen and amide hydrogen, α_{ij} the NHO angle and σ, set to 1.8 Å, the equilibrium value of the O..H distance. The initial parameters for $\varepsilon_{\text{hb1-4}}$ and $\varepsilon_{\text{hb1>4}}$ are set to 1.0 kcal/mol. Note that we distinguish short-range ($\varepsilon_{\text{hb1}-4}$) from long-range ($\varepsilon_{\text{hb1}>4}$) H-bond energies because the equilibrium Cα..Cα distances vary between α-helices (6.1 Å between 1–4 interactions) and β-sheets (4.5 Å).

Four-body effects, which represent cooperative energies between hydrogen bonds ij and kl, are defined by

$$E_{\text{HB2}} = \sum \varepsilon_\alpha^{\text{coop}} \exp(-(r_{ij}-\sigma)^2/2)\, \exp(-(r_{kl}-\sigma)^2/2) \times \Delta(ijkl) + \sum \varepsilon_\beta^{\text{coop}} \exp(-(r_{ij}-\sigma)^2/2) \times \exp(-(r_{kl}-\sigma)^2/2)\Delta'(ijkl) \quad (10)$$

The parameter $\Delta(ijkl)$ is set to 1 if residues $(k, l) = (i+1, j+1)$ and $(j = i+4,\ l = k+4)$, otherwise $\Delta(ijkl) = 0$. Thus helices Π are not stabilized. The parameter $\Delta'(ijkl) = 1$ if k and l satisfy either conditions: $(k, l) = (i+2, j-2)$ or $(i+2, j+2)$; otherwise $\Delta'(ijkl) = 0$. Together, these conditions help stabilize α-helices, antiparallel, and parallel β-sheets, independently of the (ϕ,ψ) dihedral angles, but also any segment satisfying the conditions on $ijkl$.

The parameters $\varepsilon_\alpha^{\text{coop}}$ and $\varepsilon_\beta^{\text{coop}}$ are expressed by:

$$\varepsilon_\alpha^{\text{coop}} = w_\alpha^{\text{coop}} E_\alpha^{\text{coop}} + \sum_R w_\alpha^R E_\alpha^R \quad (11)$$

$$\varepsilon_\beta^{\text{coop}} = w_\beta^{\text{coop}} E_\beta^{\text{coop}} + \sum_R w_\beta^R E_\beta^R \quad (12)$$

Here E_α^{coop} and E_β^{coop} are cooperative energies, independent of the amino acids involved and E_β^R and E_α^R are the propensity potentials of residue R for β-sheet and α-helix as defined in OPEP version 1.0.[23,36] The sum is over 4 residues for the first H-bond within a helix and over 1 residue for each additional H-bond, and the sum is generally over 2 residues associated with a H-bond in a sheet. Note that we verified that the propensity of any given residue is counted once within a long helix or β-sheet.

Overall, the interaction potential OPEP is expressed as a function of 261 w weights: 213 for $E_{\text{nonbonded}}$ with 210 for $E_{\text{Sc,Sc}}$; 4 for E_{local}, 4 for $E_{\text{H-bond}}$ and 40 for the propensities of the residues to be in α or β. This number of parameters to be optimized is much larger than in OPEP version 2 (47 parameters). In what follows, two sets will be discussed: version 3.1 with the parameters E_β^R and E_α^R included and version 3.2 with the same parameters set to zero.

Force field optimization

In protein structure prediction, an ideal force field must discriminate native from non-native structures for all natural sequences. This can be done by optimizing the ratio T_f/T_g, where T_f and T_g are the folding and glass transition temperatures, or maximizing the z-score between native and non-native structures.[22] Because the native basin consists of an ensemble of native-like conformations in equilibrium, it is desirable to learn also from these near-native structures.[52] Here, we follow the work of Qiu and Elber[31] and require that, for all protein targets, the native structure (N) is of lowest energy, and the native-like (NL) structures are simultaneously of higher energy than the native structure and of lower energy than the remote decoys (D). This double condition can be translated into a set of inequalities :

$$E(S_n, D_i) - E(S_n, N) > 0, \quad \text{for all i and n} \quad (13)$$

$$E(S_n, NL_j) - E(S_n, N) > 0, \quad \text{for all j and n} \quad (14)$$

$$E(S_n, D_i) - E(S_n, L_j) > 0, \quad \text{for all i, j, and n} \quad (15)$$

The energy of conformation C is $E(S,C)$, S_n is the sequence of the protein n, N is the native conformation,

DOI 10.1002/prot

D_i is the ith decoy conformation, and NL_j is the jth native-like conformation. Thus, for each protein, the scoring function (SF) is set to -1 for each satisfied inequality, otherwise 0.

To solve Eqs. (13–15) for all training proteins, the 261 weights are optimized using a genetic algorithm (GA) under two constraints. Firstly, the value of the Heavyside function is kept fixed for all van der Waals interaction types. This means that a given interaction type cannot varied from a 12$-$6 to $-$6 potential during optimization, and vice versa. Secondly, the magnitude of the Sc-Sc parameters are allowed to vary within specific ranges (see Validation Set). Here, GA uses standard mutation and single cross-over operators, with a mutation rate of 20% decreasing by 0.25% every 10 iterations and a cross-over rate of 14%. Double cross-over operators were tested, but identical results were obtained. We use a population size of 80 chromosomes and the elitism condition (13%) for selection. All these parameters are learnt from sensitivity analysis. Convergence in a single GA run is considered to be reached when the total number of inequalities solved does not change during 100 iterations, and the simulation is repeated 100 times using different random seeds. To guarantee that GA does not stop in a local minimum, we also perform 50 MC simulated annealing simulations of 50,000 steps starting from the best GA-determined parameters by varying one weight at each step. No improvement is observed.

Training set

The training set (TS) consists of the following 11 PDB entries: 1ABZ, 1DV0, 1E0M, 1ORC, 1PGB, 1QHK, 1SHG, 1SS1, 1VII, 2CI2, and 2CRO, and Betanova.[53] The last target 1PGBF spans the residues 41–56 of 1PGB. These proteins were selected by following three criteria. They are stable in solution without additional support, such as disulfide bonds and nonnatural amino acids. They cover a spectrum of structural classes (α, β, mixed α/β), and have medium chain lengths (16–65 amino acids) so as to facilitate conformation sampling.

The quality of the decoys employed is very important.[30–32] The decoys must cover the widest conformational space, be of lowest energy so as to compete with the native structures, be generated by various protocols so as to limit the influence of the algorithm used. In this work, we have combined four protocols to generate native-like and non-native structures for the training set.

i. We use molecular dynamics (MD) simulations using the all-atom GROMOS force field and the GROMOS package.[54] Starting from the experimental structures, explicit solvent MD is carried out for 1 ns at 600 K. All generated structures are clustered using a cRMSd cutoff of 2.5 Å. On average, 100 structures deviating by 1–12 Å from experiment are considered for each training protein.
ii. We use the simple gapless threading method, that is no gap is allowed within the template structures. The templates include the PDB entries 1PGB, 1SHG, 2CI2, 1CSP, 5FD1, 2ACY, 1CTF, 1UBQ, and 3CHY. The list is kept small because gapless threading performs poorly in fold recognition experiments, but the generated local mimina are interesting because they have native structure characteristics. On average, 200 decoys are generated for each training protein.
iii. We use the recently-designed greedy algorithm using a Go-based energy function and a cRMSd-based energy function.[55] Such a procedure provides us with both low-energy and high-energy conformations deviating by 0.05 to over 20 Å cRMSd from the experimental structures.
iv. We use the Decoys "R" Us publicly available fisa set for protein 2CRO, hereafter referred to as 2CRO-fisa, generated by assembling protein fragments using a simulated annealing procedure.[56]

Another important aspect in our optimization process is to assign each conformation with native-like or nonnative characters. It is common in the study of protein folding to use the cRMSd or the fraction of native contacts. Network and graph analyses of free energy surfaces have clearly shown the limitation of these parameters.[57] Here, we use the TM-align program, a fast and accurate protein structure alignment algorithm,[58,59] and the default TM-score of 0.5 for discriminating native-like (TM-score $>$ 0.5) from nonnative structures ($<$0.5). To this end, all decoys were optimally aligned with their corresponding native structure, and we selected five structures with TM-scores varying between 0.5 and 1.0 as native-like conformations. For 1SHG, the narrow distribution of TM-scores led us to select only four native-like structures. In what follows, the nonredundant training set of each target consists in the native structure, five (or four) native-like structures and the decoys with TM-scores $<$0.5, while the full training set includes all decoys, independently of the TM-scores.

Table II analyzes the structural characteristics of the full-training set in terms of cRMSd's and TM-scores with respect to the native state, and the percentage of residues in α and β states. Note that the extremes values of the parameters do not change between the nonredundant and the full training sets, and all structures were optimized using OPEP until the gradient norm $<$0.01 kcal/(mol Å). Overall the resulting density of states shows that the conformational space sampled is not limited to any specific basin of attraction.

Validation set

To test the effectiveness of OPEP, we have considered multiple decoy sets, namely 4-state reduced,[62,63] lmds

Table II
Description of the Full-Training and Full-Validating Sets

			Number of decoys				cRMSd (Å)		TM-score		Alpha-helix	Beta-strand
Set	*M*	*L*	MD	TH	GR	TOT	Min	Max	Min	Max	(%)	(%)
Full training set												
IABZ	1	38	49	281	278	608	2.6	11.9	0.14	0.58	46.9	5.82
Betanova	—	20	347	299	283	930	0.7	12.2	0.01	0.59	10.36	23.87
1DV0	1	45	118	209	275	602	1.9	13.6	0.16	0.76	35.81	4.91
1E0M	1	37	59	213	183	451	1.5	12.8	0.14	0.82	17.14	17.33
1ORC	1	64	79	89	262	430	0.9	13.6	0.22	0.92	26.69	20.00
1PGB	1	56	67	179	289	535	0.9	36.9	0.12	0.92	26.57	20.41
1PGBF	1	16	28	276	300	604	0.2	11.1	0.01	0.99	10.31	15.05
1QHK	1	47	59	189	277	525	3.3	31.1	0.13	0.78	24.22	10.03
1SHG	1	57	41	601	286	928	1.3	41.7	0.12	0.85	16.99	21.98
1SS1	1	60	90	128	212	430	2.2	13.8	0.19	0.85	39.77	4.49
1VII	1	36	45	250	270	565	1.6	10.9	0.13	0.68	37.41	4.33
2CI2	6	65	30	166	298	494	2.2	34.9	0.11	0.87	21.52	16.33
2CRO-fisa[a]	3	65	25	—	—	525	2.2	12.3	0.23	0.99	61.75	0.02
Full validation set												
1BBA-lmds[a]	1	36	—	—	—	500	0.9	9.2	0.41	0.82	47.33	0.00
1CTF-4state[a]	2	68	—	—	—	616	0.5	12.5	0.24	0.88	33.56	4.09
1CTF-lattice[a]	2	68	—	—	—	977	5.1	10.5	0.25	0.51	34.65	0.26
1CTF-lmds[a]	2	68	—	—	—	489	3.3	15.1	0.28	0.68	51.28	21.16
1CTF-semfold[a]	2	68	—	—	—	996	4.4	9.2	0.28	0.59	50.70	0.63
1F4I	1	45	11,519●	—	—	11,519	0.1	36.9	0.12	0.99	4.74	7.37
1FSD	1	28	84	319	290	693	1.4	10.7	0.13	0.80	34.49	4.68
1KHM-semfold[a]	1	73	—	—	—	998	3.9	7.9	0.32	0.58	44.31	2.22
1R69	1	63	72	99	281	452	0.1	27.7	0.23	0.99	46.81	3.51
1R69-4state[a]	1	63	—	—	—	670	0.8	11.4	0.25	0.94	36.63	0.36
1R69-rosetta[a]	1	61	—	—	—	998	0.2	14.9	0.27	0.99	63.13	0.00
1S04-CASP6[a]	1	110	27	—	—	213	0.6	56.5	0.15	0.98	28.19	15.66
1TE7-CASP6[a]	1	103	18	—	—	222	1.8	38.3	0.18	0.88	22.69	21.79
1UBI-lmdsv2[a]	1	76	61	—	—	361	2.0	15.4	0.22	0.81	15.54	30.44
2CRO-4state[a]	3	65	25	—	—	697	0.1	9.3	0.25	0.99	36.33	0.19
2CRO-lmds[a]	3	65	25	—	—	525	0.1	12.9	0.25	0.99	63.85	0.06

[a] *M* is the oligomeric state according to E-MSD (the European Bioinformatics Institute Macromolecular Structure Database[60]), *L* is the protein length (amino acids), *TOT* is the total number of decoys used per protein, generated either by molecular dynamics (MD), threading (TH) and greedy (GR) methods or taken from publicly available sets. This is followed by the maximal (max) and minimal (min) cRMSd's and TM-scores between the decoys and the native structure, as well as the average percentages of α-helix and β-strand conformations identified using the *Stride* program.[61] ●: decoys generated by ART-OPEP simulations as described in the text.

and lmdsv2,[63] semfold,[64] fisa,[56] lattice-ssfit,[65] Rosetta,[66] and CASP6.[67] Clearly, we cannot minimize ~120 targets and a total of ~210,000 conformations with a gradient norm of 0.01 kcal/(mol Å) using reasonable computer resources. In addition, some targets are in our training list (2CRO-fisa), are multi-domains (e.g., 1FC2), involve ligands such as calcium (e.g., 4ICB and 1E68) or iron (4RXN), are stabilized by disulfide bridges (2 among 7 proteins in 4-state reduced set, this requires to parametrize oxided Cys), or lack terminal residues in Rosetta.

In this work, we consider the following targets: lmds (1BBA, 1CTF, 2CRO), lmdsv2 (1UBI), 4-state reduced (1CTF, 1R69, 2CRO), lattice-ssfit (1CTF), semfold (1CTF, 1KHM), Rosetta (1R69), and CASP6 (1S04, 1TE7). For 1CTF semfold, 1CTF lattice-ssfit and 1KHM semfold, we did not consider the full set of 11399, 2000, and 21,079 decoys, respectively but randomly choose ~1,000 decoys covering the full TM-Score and C_α cRMSd distributions. Note also that 1CTF is the single target selected which is stabilized by a $SO4^{2-}$ ligand, and 1R69 has 63 amino acids in the experimental structure and the 4-state set, but 61 amino acids in the Rosetta set (the two C-terminal residues are missing). Finally, decoys were generated for 1FSD, 1R69, and 1CTF using the MD approach (see above), and decoys were also obtained for 1F4I using ART-OPEP[39,40] simulations. The ART-OPEP simulations started from fully extended conformations and were carried out for 20,000 steps using the parameters prior to optimization. This set of 10 distinct proteins (16 targets), consisting of 20,926 decoys, constitutes the full-validation set (VS). As for TS, we generate a nonredundant validation set. The structural properties of the full VS set are described in Table II. We see that the cRMSd's deviate by 0.1–36.9 Å and the TM-scores vary between 0.12 and 0.99.

Table III shows the total number of side-chain side-chain contacts and the fraction of each amino acid type in secondary structures for the VS plus TS sets. Although we use a total number of 22 proteins and 28,553 decoys, some Sc-Sc pairs are poorly populated in TS or VS.

Table III

Number of Side-Chain Side-Chain Contacts and Percentage of Residues in α and β States Within the Full-Training and Full-Validation Sets

VS/TS	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	3416 89,836	1532	2920	5692	5428	3516	204	5936	7222	13,478	3912	2994	702	4290	3102	1550	5612	6436	1176	3540
CYS	28,246	0 138	668	766	162	78	0	218	748	1378	246	892	466	376	98	44	416	910	776	622
ASP	32,084	2426	796 28,642	1698	1340	1750	108	448	5402	1384	586	2166	1490	1118	1590	1842	3794	1748	1042	2190
GLU	122,876	18,134	74,108	1154 73,224	2830	2496	94	4124	12,798	5174	2286	3440	1100	4664	6178	2686	7670	2550	1060	2374
PHE	136,940	29,442	34,466	57,020	1822 23,116	2254	8	3076	3284	8656	2260	2150	1900	2694	2850	1704	3284	3660	952	3184
GLY	47,274	706	8560	21,682	34,738	1104 9628	50	1166	5054	3534	870	3752	538	1434	3192	964	4148	4442	2000	4204
HIS	2194	38	1950	5686	4642	604	0 824	196	166	162	44	604	62	10	68	48	252	0	160	226
ILE	115,988	14,024	12,728	67,924	87,720	12,882	6612	1794 22,652	3686	8968	1336	816	466	3420	2384	1868	1954	3862	716	2884
LYS	79,474	4422	68,266	251,094	18,750	24,792	1156	39,906	2940 21,496	8064	1696	4238	1238	5612	4116	3054	8220	4548	2492	4380
LEU	358,120	20,358	29,590	81,190	272,606	87,196	2764	210,078	65,324	8602 259,172	4628	4184	3242	6990	6348	4290	5252	7624	3374	2424
MET	15,366	936	1592	12,602	10,744	5092	838	15,184	4182	30,226	178 2356	228	340	1162	862	920	2466	1126	486	716
ASN	65,442	4454	32,632	91,958	55,940	9326	330	7296	47,790	43,442	2630	2006 25,732	432	3132	1302	1896	3774	818	1932	2864
PRO	20,876	336	7348	13,810	6630	4872	10,524	13,748	6064	25,612	4234	3544	284 4850	936	2024	936	688	1826	1156	766
GLN	59,220	1064	9976	77,786	21,774	11,798	1134	45,194	52,296	60,014	8804	42,852	7766	2738 24,468	1642	3690	4272	2790	1118	2624
ARG	25,796	926	41,636	103,220	19,822	13,552	7222	31,148	37,720	51,934	7424	8686	17,682	20,140	1932 15,896	814	2944	2068	1268	1584
SER	47,080	3196	19,964	66,516	27,754	12,170	3872	38,052	31,264	66,456	1972	14,708	15,114	43,492	19,428	244 14,776	2224	1612	876	1366
THR	23,800	804	8564	23,334	13,764	18,580	5934	21,380	19,214	35,364	5382	7538	10,064	24,094	22,072	17,138	6400 11,406	4586	3272	5812
VAL	108,760	2840	24,072	44,636	50,290	42,562	2556	91,130	57,408	151,560	17,466	6038	15,804	10,052	38,516	36,876	17,462	1078 46,808	2498	3920
TRP	5856	1618	10,122	2160	2210	3472	548	3572	2524	16,834	3644	4542	1620	1922	5098	4482	2532	12,674	74 0	2242
TYR	46,228	2396	4878	16,264	32,274	9634	2532	24,698	8400	39,872	7044	7820	9130	11,470	10,596	5018	7028	22,162	2092	1930 2716
VS	18.453	0.025	5.643	11.502	0.890	2.904	0.332	5.761	15.230	13.956	0.452	1.234	1.038	5.429	5.304	3.213	1.926	5.440	0.600	0.661
TS	13.899	0.738	1.174	10.846	4.214	1.802	0.147	6.843	10.433	14.342	2.987	3.222	0.865	8.590	5.674	4.733	5.182	2.231	1.046	1.030
VS	1.071	0.006	0.246	0.636	0.281	0.335	0.000	0.503	0.441	0.662	0.005	0.058	0.034	0.102	0.080	0.122	0.249	0.989	0.007	0.021
TS	0.328	0.007	0.274	0.467	0.677	0.514	0.000	0.275	0.668	0.333	0.059	0.355	0.014	0.115	0.181	0.400	1.174	0.418	0.164	0.837

Upper (lower) values correspond to the number of Sc-Sc contacts within TS (VS) sets, excluding the threading generated decoys. VS α(β) and TS α(β) give the percentage of residues in α(β) within VS and TS. Number of contacts lower than 50 are in bold.

Table IV
Optimization Results for the Nonredundant Training Set

TARGET	D_{TOT}	NL	SF_{max}	SF_{start}	SF_{end}	Gain	Unsat
1ABZ	445	5	−2675	−2380	−2613	233	62
Betanova	618	5	−3713	−3658	−3649	−9	64
1DV0	378	5	−2273	−2153	−2256	103	17
1E0M	310	5	−1865	−1056	−1317	261	548
1ORC	175	5	−1043	−886	−943	57	100
1PGB	304	5	−1829	−1823	−1824	1	5
1PGBF	526	5	−3161	−3096	−3124	28	37
1QHK	416	5	−2501	−2442	−2445	3	56
1SHG	631	4	−3159	−3132	−3143	11	16
1SS1	162	5	−977	−536	−857	321	120
1VII	430	5	−2585	−2288	−2542	254	43
2CI2	189	5	−1139	−918	−936	18	203
2CRO-fisa	422	5	−2537	−2504	−2452	−52	85
Total	5006	64	−29,457	−26,872	−28,101	1229	1356

D_{TOT} is the total number of decoys. *NL* is the number of native-like structures used during the optimization process. SF_{max} is the maximal value of the scoring function, SF, for each target. SF_{start} and SF_{end} are the SF values before and after optimization. Gain is the SF gain after optimization ($= SF_{start} - SF_{end}$), and Unsat is the number of inequalities remaining unsatisfied after optimization ($= SF_{end} - SF_{max}$).

Among a total of 210 pairs, the number of contacts is less than 50 for 10 pairs. The low population for the Cys-Cys pair is expected since we do not consider proteins stabilized by disulfide bridges. The low population for the Cys-Ser, His-Phe, His-Gly, His-Met, His-Gln, His-Ser, and His-Val pairs is observed in TS, but not in VS. Conversely, the weak occurrence of the His-His pair is observed in VS only. This clearly indicates that our choice of targets is not free of any biases. In contrast, the Cys-His contact is rarely seen in both TS and VS, and in our 2248 PDB structures, indicating its weak occurrence in protein structures. To take account of the previous biases in our distribution of Sc-Sc contacts, the optimization of the matrix contact proposed by Betancourt and Thirumalai[50] is under control and the $w_{Sc,Sc}$ weights are restrained to vary between 0.7 and 1.3.

RESULTS

Overall performance of recognition

Table IV presents the results of the optimization of OPEP version 3.1 on the nonredundant training set. A total scoring function, SF, of −29,457 indicates that all inequalities in Eqs. (13–15) are satisfied, and thus all native structures are of lowest energy, and all selected native-like structures are of lower energy than the remote conformations. We find that SF decreases from −26,872 to −28,101 during the optimization. Although the improvement is not optimal, 48% of inequalities initially unsatisfied are satisfied. Refined OPEP improves recognition for 11 proteins (SF decrease of 1 for 1PGB and 321 for 1SS1) and appears less efficient for Betanova and 2CRO-fisa (SF increase of 9 and 52, respectively). The optimized Sc-Sc, H-bond, and α-helix and β-sheet propensities are given in Table 1S provided in Supplementary Materials.

We now consider the full-training and full-validation sets in Table V. Figure 1S provided in Supplementary Materials plots the energy versus TM-score for all proteins after optimization and the energy versus cRMSd for Betanova. 3,179,377 inequalities are to be satisfied. Prior to optimization 2,439,934 inequalities are satisfied (739,443 unsatisfied), and, after optimization, 2,748,877 inequalities are satisfied (430,500 unsatisfied). Thus, on the order of 42% of the unsatisfied inequalities have been solved. We also show in Table V the rank of the experimental structure and whether the lowest-energy structure is associated with a native-like or a non-native structure. The experimental structure is ranked first for four training proteins (1DV0, 1QHK, 1SS1, and 2CI2) and four validating targets (1CTF-4state, 1CTF-lattice, 1KHM-semfold and 1R69-4state). A native-like structure is

Table V
Detailed Performance of OPEP Version 3.1 on the Full Training and Full Validation Sets

TARGET	*E*(*N*)	*E*(D_{min})	*E*(NL_{min})	*Nrk*	*NLrk*	*Rc*
Full training set						
1ABZ	−87.3	−93.9	−91.4	30	11	No
Betanova	−29.0	−34.0	−29.0	26	29	No
1DV0	−107.2	−106.0	−106.0	1	2	Yes
1E0M	−50.1	−60.1	−54.0	34	14	No
1ORC	−135.5	−143.9	−143.9	5	1	Yes
1PGB	−141.8	−156.1	−156.1	21	1	Yes
1PGBF	−30.4	−30.7	−30.7	17	1	Yes
1QHK	−100.1	−93.1	−88.6	1	4	Yes
1SHG	−136.3	−139.4	−139.4	5	1	Yes
1SS1	−121.8	−115.5	−115.5	1	2	Yes
1VII	−71.2	−71.6	−71.6	2	1	Yes
2CI2	−173.3	−166.9	−166.9	1	2	Yes
2CRO-fisa	−141.9	−160.8	−160.1	115	2	No
Full validation set						
1BBA-lmds	−48.8	−57.7	−57.7	378	1	Yes
1CTF-4state	−180.3	−179.2	−179.2	1	2	Yes
1CTF-lattice	−180.3	−155.5	−145.5	1	11	Yes
1CTF-lmds	−180.3	−183.1	−177.8	2	3	No
1CTF-semfold	−180.3	−181.8	−181.8	3	1	Yes
1F4I	−78.4	−90.3	−90.3	6	1	Yes
1FSD	−55.3	−58.4	−58.4	12	1	Yes
1KHM-semfold	−161.8	−141.6	−141.1	1	4	Yes
1R69	−140.5	−140.6	−140.6	2	1	Yes
1R69-4state	−140.5	−136.2	−136.2	1	2	Yes
1R69-rosetta	−139.6	−140.0	−140.0	3	1	Yes
1S04-CASP6	−286.7	−309.1	−302.7	32	2	No
1TE7-CASP6	−254.2	−280.5	−280.5	24	1	Yes
1UBI-lmdsv2	−165.8	−179.2	−179.2	69	1	Yes
2CRO-4state	−141.9	−153.8	−153.8	28	1	Yes
2CRO-lmds	−141.9	−153.8	−153.8	37	1	Yes

E(*N*) is the energy of the native structure. *E*(D_{min}) is the lowest-energy obtained for the set excluding the native structure, and *E*(NL_{min}) is the lowest-energy native-like structure. *Nrk* is the rank of the experimental structure, *NLrk* is the rank of the lowest-energy native-like structure. Rc is the recognition status, "Yes" indicates that the lowest-energy structure is associated with the native or a near-native structure. Underline values denote cases where the recognition status changed during the optimization.

ranked first for five other TS proteins, and for 10 VS targets. Overall, OPEP version 3.1 is able to discriminate native or native-like structures from nonnative conformations for 23 of the 29 targets, that is 69% (9/13) of the TS targets and 87.5% (14/16) of the VS targets.

Problematic targets

It is of interest to examine the six problematic targets at the atomic level of detail. These include the four TS 1ABZ, Betanova, 1E0M, and 2CRO-fisa targets, and the VS 1CTF-lmds and 1S04-CASP6 targets. Figure 2 shows their experimental and lowest-energy structures.

For the 1ABZ and 1E0M training proteins, the minimal energy is associated with a near-native conformation. The lowest-energy conformation for 1ABZ has a TM-Score of 0.49 (see Figure 1S in Supplementary Materials), but deviates by 2.3 Å cRMSd from the NMR structure [Fig. 2(A)]. Similarly, the decoy of lowest-energy for the 34-residue 1E0M protein has a TM-score of 0.43 and deviates by 6.8 Å cRMSd from the NMR structure. However, its topology is native: excluding the extremities, where structural determination is very low by NMR,[70] the cRMSd using residues 6–31 is only 2.2 Å [Fig. 2(C)]. The energy gain in the decoy comes essentially from the extension of the third β-strand by 2 residues.

For the 65-residue 2CRO-fisa set, the energy gap between the lowest-energy decoy and native-like structures is marginal, on the order of k_BT. The best decoy (TM-score of 0.45) deviates by 6.0 Å cRMSd from the crystal structure [Fig. 2(D,E)]. The best native-like conformation (TM-score of 0.58) deviates by 4.2 and 2.7 Å cRMSd from the crystal structure using residues 1–65 and residues 3–61, respectively. While the crystal and native-like conformations are described by a five-helix bundle, the decoy is a four-helix bundle with the fifth helix unfolded and the first three helices deviating by 4.3 Å from their positions in the crystal structure [see Fig. 2(D,E)]. One must note, however, that 2CRO is one target which displays intermolecular interactions within the unit cell of the crystal structure (homotrimeric assembly, as described in the Macromolecular Structure Database[60]).

While Betanova is expected to adopt a three-stranded β-sheet in solution,[53] the minimal energy structure is a β-hairpin [Fig. 2(B)]. This decoy results from threading the sequence on the residues 1–20 of 1PGB. The hairpin is stabilized by 5 kcal/mol with respect to the NMR structure because the increase in two-body and four-body H-bond contributions outperforms the loss of side-chain–side-chain interactions. This preference for the β-hairpin is surprising, but there is a large body of experimental and theoretical studies indicating that the three-stranded β-sheet is not the dominant conformation in solution. In contrast to the first experimental study, the NMR β-sheet population has been reestimated to be ∼10% in water at 283 K.[71] All-atom MD simulations in explicit solvent showed that the ideal three-stranded β-sheet structure is in dynamic equilibrium with other disordered species within a 100-ns timescale.[72] Interestingly, for another peptide, namely the protein 1PGB C-terminal fragment, transition graph analysis of folding free energy surfaces suggests that the β-hairpin is slightly more populated than a triple stranded β-sheet,[57] and thus points to our two competiting conformations.

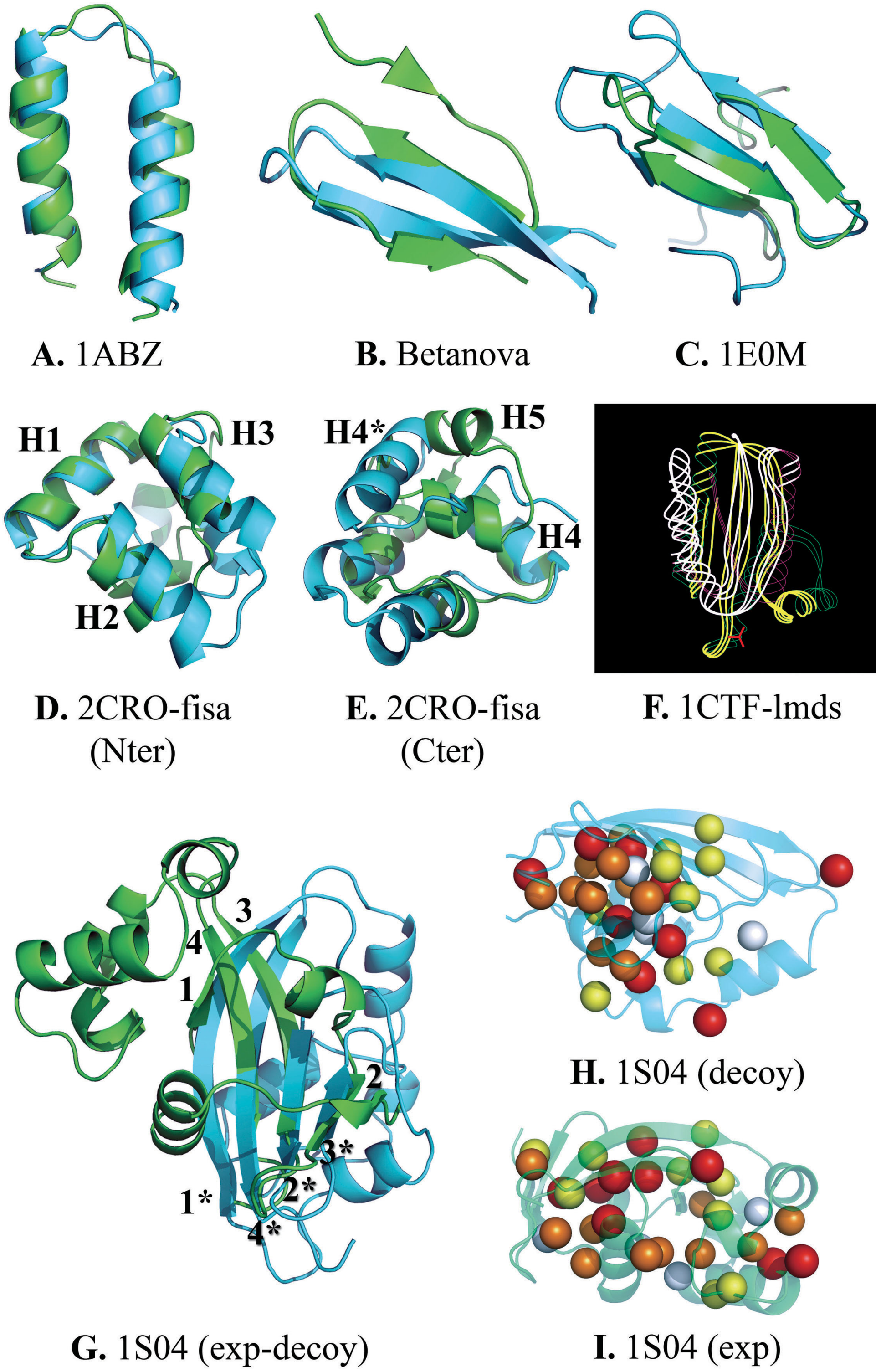
The lowest-energy lmds decoy of the 68-residue 1CTF protein has a TM-score of 0.42 and displays a nonnative topology. This decoy is stabilized by 3 and 5 kcal/mol with respect to the best near-native and native structures. The decoy and native structures differ by 8.7 Å cRMSd. As discussed earlier, the crystal structure has a SO_4^{2-} ligand hydrogen-bonded to the side-chain nitrogen of Lys65 and the amide nitrogen of Gly62 (in PDB numbering). It has been suggested that this sulfate environment may have functional implications.[73] In fact, there is no space for a SO_4^{2-} ligand in proximity of Lys65 and Gly62 in the decoy structure [see Fig. 2(F)].

Finally, the best decoy of 1S04 differs from the native and best native-like structures by 23 and 6 kcal/mol, respectively. As shown in Figure 2(G), the decoy, of nonnative topology, is characterized by a swap of the β-strands 2 and 3 and a destabilization of the α-helices (24% of helical residues vs. 39% in the native structure). In addition, the β-sheet is less twisted and the strands are extended (29% of β-strands vs. 25% in the native state). As seen in Figure 2(H,I), the topological change is stabilized by both the H-bonds and a more compact hydrophobic core.

In summary, we can consider that OPEP correctly recognizes 1ABZ (based on the cRMSd criterion), the calculation in solution might not be appropriate for 1CTF and 2CRO, and OPEP clearly fails for 1S04. This suggests that the balance between hydrogen-bonding and Sc-Sc interactions could be reevaluated, but before doing so, four questions must be addressed.

OPEP parameters are robust

The first question is whether a larger training set would impact the results.[30] To this end, we repeat the optimization process by training the parameters on the nonredundant TS and VS sets, starting from our previous optimal weights. The optimization, achieved after 100 GA runs, leads to a negligible gain of 0.05% of satisfied inequalities. Applied to the full TS and VS sets, the gain amounts to 2.2%. However, the correlation coefficient betwen the two OPEP vectors is 0.98, and the recognition performances remain identical. For instance, the rank of the native structure for the problematic targets is 28 versus 30 for 1ABZ, 33 versus 34 for 1E0M, 31 versus 32 for 1S04-CASP6, and 116 versus 115 for



A. 1ABZ **B.** Betanova **C.** 1E0M

D. 2CRO-fisa (Nter) **E.** 2CRO-fisa (Cter) **F.** 1CTF-lmds

G. 1S04 (exp-decoy) **H.** 1S04 (decoy) **I.** 1S04 (exp)

Figure 2

Problematic targets*. Superposition of the experimental structure and the decoys of minimal energy. Native structure is in green for all targets. For 2CRO-fisa (views D and E), the two views highlight the structural differences within the N-terminus and C-terminus regions. H# corresponds to the #th helix in the experimental structure, and H#* to the #th helix in the decoy. For 1CTF-lmds (view F), the red anchor shows the position of the $SO4^{2-}$ ligand. For the 1S04 target, three views are given: view G shows the swap of* β*-strands, and views H and I illustrate the difference in the packing of the non-polar side-chains, with isoleucine in red, valine in yellow, phenylalanine in white and leucine in orange. Pictures where generated using PyMol*[68] *for* ***A*** *to* ***E****, and* ***G*** *to* ***I****, and XmMol*[69] *for F.*

DOI 10.1002/prot

2CRO-fisa. This indicates that the OPEP parameters are robust and insensitive to the number of proteins used in the derivation.

Success rate: OPEP versus published force fields

It is also important to assess how OPEP compares to other published force fields in ranking native structures. This is not an easy task, since all studies differ in the list of targets, criteria and decoys used. However, Shen and Sali recently tested the atomic distance-dependent statistical DOPE potential[74,75] and compared its recognition performance with respect to five other scoring functions, namely Rosetta,[56,76,77] DFIRE,[78] and ModPipe.[79] DOPE correctly identified 28 native structures for 32 targets, versus 27 for DFIRE, 14 for Rosetta, and 19, 7 and 18 for the three ModPipe versions.

In this work, we have calculated the all-atom DOPE energies of our full list of decoys using the SCWRL3 program[80] to rebuild the all-atom side chains and the "Decoys" tool of the Protein Library available at http://protlib.uchicago.edu/.[74] Because such a side-chain reconstruction can affect DOPE performances with respect to the published results using the non-minimized decoys, we also report the DOPE results on the non-minimized sets. We emphasize that our goal is not to rank OPEP versus DOPE, because our analysis is not based on all the decoys used for evaluating DOPE. In addition, DOPE is less accurate on small proteins (less than 50 amino acids) and NMR-derived structures.[75]

Table VI compares OPEP and DOPE performances on each target using four criteria: the *z*-score calculated as the distance in standard deviations from the energy of the native structure E_N to the average energy of the decoys ($z\text{-score} = (E_N - \langle E_D \rangle)/\sigma_{E_D}$), the correlation coefficients between the energies and the cRMSd's or the TM-scores, and the rank of the native or native like structures. Figure 2S provided in Supplementary Materials plots the energy versus TM-score and energy versus cRMSd for all targets using OPEP and DOPE.

Using the minimized decoys, OPEP correctly identifies 23 native or native-like structures for 29 targets, while DOPE is successful for 19 targets. Both OPEP and DOPE energy functions display negative *z*-score values except for the 1BBA-lmds target using OPEP - but most decoys are in fact native like with cRMSd's of 2.2Å (see Supplementary Materials Figure 2S)-, and for the Betanova, 1BBA-lmds and 1FSD targets using DOPE. Overall, the mean *z*-score is slightly lower with OPEP (−2.07 versus −1.42 using DOPE), and the average *z*-score standard deviations are very similar (1.03 vs 1.11 using DOPE). Interestingly, the mean *z*-score values are very similar for OPEP on VS and TS. The correlations between the energies and TM-scores and between the energies and cRMSd's are almost identical. Using both TS and VS sets, the mean correlation coefficients are −0.574 versus −0.557 for *E* versus TM-scores and 0.543 versus 0.557 for *E* versus cRMSd's using OPEP and DOPE, respectively.

Using the publicly available 14 targets from Decoys 'R' Us, Rosetta and CASP6, that is without any minimisation and side-chain positioning, DOPE correctly identifies 10 targets, but fails for 1BBA-lmds, 1KHM-semfold, 1S04-CASP6, and 1TE7-CASP6, whereas it failed for 5 targets among 14 using the minimised structures. The average *z*-score is of −2.95, with extreme values of 15 for 1BBA-lmds and −10.1 for 1CTF-lattice. The corresponding val-

Table VI
Comparison of OPEP and DOPE Performances

	OPEP		DOPE				
TARGET	*Z*	*R*	*Z*	*R*	*Nrk*	*NLrk*	*Rc*
Full training set							
1ABZ	−1.87	−0.635	−0.56	−0.677	186	8	No
Betanova	−1.34	−0.555	1.12	−0.048	789	87	No
1DV0	−3.06	−0.604	−1.85	−0.762	23	1	Yes
1E0M	−1.69	−0.389	−1.39	−0.358	35	5	No
1ORC	−1.67	−0.753	−1.88	−0.830	2	1	Yes
1PGB	−1.88	−0.858	−1.23	−0.780	53	1	Yes
1PGBF	−2.32	−0.755	−1.41	−0.409	42	31	No
1QHK	−2.93	−0.509	−1.30	−0.584	13	1	Yes
1SHG	−2.07	−0.776	−1.79	−0.590	12	1	Yes
1SS1	−2.69	−0.601	−1.44	−0.751	29	1	Yes
1VII	−2.05	−0.574	−0.23	−0.583	247	6	No
2CI2	−1.90	−0.698	−1.18	−0.786	11	1	Yes
2CRO-fisa	−0.75	−0.493	−0.57	−0.451	152	4	No
Mean	−2.02	−0.631	−1.06	−0.585	—	—	—
SD	0.63	0.132	0.83	0.223	—	—	—
Full validation set							
1BBA-lmds	0.57	−0.212	0.22	−0.190	305	1	Yes
1CTF-4state	−2.36	−0.775	−2.32	−0.800	1	2	Yes
1CTF-lattice	−4.79	−0.251	−4.75	−0.281	1	4	Yes
1CTF-lmds	−2.54	−0.160	−3.05	−0.093	1	40	Yes
1CTF-semfold	−2.53	−0.274	−2.12	−0.326	6	1	Yes
1F4I	−2.14	−0.513	−2.13	−0.691	3	1	Yes
1FSD	−2.07	−0.718	0.62	−0.698	498	1	Yes
1KHM-semfold	−4.63	−0.356	−2.24	−0.361	13	3	No
1R69	−1.77	−0.787	−1.61	−0.795	4	1	Yes
1R69-4state	−2.58	−0.737	−2.20	−0.785	4	1	Yes
1R69-rosetta	−2.42	−0.381	−1.35	−0.417	89	1	Yes
1S04-CASP6	−1.24	−0.586	−1.54	−0.561	5	6	No
1TE7-CASP6	−1.28	−0.674	−1.20	−0.580	27	2	No
1UBI-lmdsv2	−1.01	−0.801	−1.64	−0.890	60	1	Yes
2CRO-4state	−1.86	−0.698	−1.58	−0.775	52	1	Yes
2CRO-lmds	−1.30	−0.514	−0.45	−0.289	164	4	No
Mean	−2.12	−0.527	−1.71	−0.533	—	—	—
SD	1.29	0.226	1.25	0.254	—	—	—
All targets							
Mean	−2.07	−0.574	−1.42	−0.557	—	—	—
SD	1.03	0.194	1.11	0.237	—	—	—

For each target of the full-training and full-validation sets, we report the z-score and the correlation coefficient, R, between the energies and TM-scores using OPEP 3.1 or DOPE.[74] The mean and standard deviation values are given for the training set, the validation set and both sets. *Nrk* is the rank of the native structure, *NLrk* is the rank of the lowest energy native-like structure. Rc reports the DOPE recognition status, "Yes" indicating that the lowest-energy structure is associated with the native or a near-native structure.

ues on the minimized structures are of −1.77, 0.62, and −4.75. On the same minimized targets, OPEP fails for 2CRO-fisa, 1CTF-lmds and 1S04-CASP6, that is correctly identifies 11 among 14 targets, and the average *z*-score value is −2.05. Plots of energy versus cRMSd are given in Figure 3S provided in Supplementary materials.

Impact of α-helix and β-sheet propensities on recognition

So far, we have examined the recognition performances of OPEP version 3.1. What is the impact of the α-helix and β-sheet propensity potentials set to zero? To this end, we repeat the optimization process on the non-redundant VS and analyse the results on the full TS+VS. We find that the number of inequalities satisfied after optimisation does not vary much from version 3.1 to version 3.2: 2,583,679 inequalities are now satisfied versus 2,748,877 in version 3.1. However, Table VII shows that the identity of the minimal energy structure remains identical for all targets, and the rank of the native structures vary substantially only for 1F4I (ranked 125 by version 3.2 vs. 6 by version 3.1). Overall, we see that version 3.1 and version 3.2 cannot be distinguished on the basis of the energy gap between native or native-like structures and non-native structures.

Table VII
Performance Comparison for OPEP 3.1 and OPEP 3.2 (Propensities Set to Zero)

TARGET	$E(N)$	$E(D_{\min})$	$E(NL_{\min})$	Nrk	$NLrk$	Rc
Full training set						
1ABZ	−75.8	−81.7	−78.8	31	11	No
Betanova	−29.1	−36.3	−29.0	33	36	No
1DV0	−103.3	−103.2	−103.2	1	2	Yes
1E0M	−49.2	−59.3	−52.5	38	16	No
1ORC	−137.1	−146.1	−146.1	5	1	Yes
1PGB	−142.4	−156.6	−156.6	19	1	Yes
1PGBF	−29.9	−30.2	−30.2	17	1	Yes
1QHK	−100.0	−97.1	−88.5	1	6	Yes
1SHG	−140.8	−144.0	−144.0	5	1	Yes
1SS1	−119.1	−111.5	−111.5	1	2	Yes
1VII	−71.1	−72.3	−72.3	5	1	Yes
2CI2	−175.0	−168.5	−168.5	1	2	Yes
2CRO-fisa	−136.6	−155.4	−154.7	122	3	No
Full validation set						
1BBA-lmds	−46.5	−55.4	−55.4	384	1	Yes
1CTF-4state	−174.3	−175.8	−175.8	2	1	Yes
1CTF-lattice	−174.3	−150.6	−138.5	1	12	Yes
1CTF-lmds	−174.3	−178.6	−173.7	2	3	No
1CTF-semfold	−174.3	−177.2	−177.2	3	1	Yes
1F4I	−75.0	−86.9	−86.9	125	1	Yes
1FSD	−53.7	−56.6	−56.6	11	1	Yes
1KHM-semfold	−162.6	−143.5	−143.0	1	3	Yes
1R69	−136.4	−136.4	−136.4	2	1	Yes
1R69-4state	−136.4	−1330	−133.0	1	2	Yes
1R69-rosetta	−135.9	−137.0	−136.3	3	1	Yes
1S04-CASP6	−284.3	−309.5	−301.9	33	2	No
1TE7-CASP6	−256.2	−281.6	−281.6	25	1	Yes
1UBI-lmdsv2	−169.0	−183.4	−183.4	70	1	Yes
2CRO-4state	−136.6	−148.9	−148.9	30	1	Yes
2CRO-lmds	−136.6	−148.9	−148.9	38	1	Yes

See legend of Table 5 for the definition of the variables.

Is OPEP accurate for thermodynamics and kinetics?

Although discriminating native from misfolded conformations is an important aspect in protein folding, that does not imply OPEP provides an accurate description of the dynamics around and outside of the native state.

In the vicinity of minima, the thermal atomic fluctuations are well described. This is supported by MD simulations using OPEP version 3 on three protein models which showed root mean square fluctuations comparable to those found by all-atom molecular mechanics force fields.[81] Outside of the minima, OPEP must generate energy barriers and transition states consistent with high-level quantum mechanics calculations and molecular mechanics. A detailed study of the folding of a 16-residue β-hairpin using ART-OPEP version 2 simulations identified two pathways following closely those observed by previous theoretical studies.[41] Additionally, MD-OPEP version 3 simulations revealed aggregation mechanisms for four chains of the amyloid-forming fragment Aβ(16–22) consistent with experiments.[81]

Similarly the present exercise does not guarantee that OPEP is accurate for thermodynamics. However, the results of four recent studies using OPEP version 3 are very encouraging. (i) MC simulations of the fragment Aβ(21–30) in solution provide a very good fit with NMR experiments.[82] (ii) MD simulations at 300 K show that a three-helix bundle and the B1 domain of protein G are stable within 50 ns timescale at room temperature.[81] (iii) REMD-OPEP simulations find that the melting temperature of the β-hairpin 1PGBF peptide is consistent with experiment (295 K calculated vs. 297 K experimentally) (unpublished results). (iv) Using eight replicas with T varying between 287 and 500 K, the REMD-OPEP derived free energy surface of the $A\beta_{16-22}$ dimer at 310 K is very similar to that generated by all-atom explicit solvent REMD simulations.[83]

CONCLUSION

We have revisited the OPEP energy function on 29 targets using a total of 28,553 decoys. These decoys are generated by various protocols or taken from publicly available sets so as to limit the impact of the method used. The OPEP parameters are varied by genetic and MC algorithms using a scoring function which requires that both native and native-like conformations are of lower energies than decoys.

Overall, OPEP correctly identifies 24 native or near-native conformations for 29 targets. This result is significant because the all-atom DOPE energy function, which was found to outperform five established force fields, is found to have comparable capabilities on our set of decoys. In fact, OPEP fails to recognize the native basin for five targets: 1E0M, 2CRO-fisa, 1CTF-lmds, Betanova, and 1S04-CASP6. However, the structure of 1CTF has been solved with a tightly bound sulfate ion and the structure of 2CRO displays crystal packing effects. This suggests that the ion and the contacts between subunits may be required in our potential to identify their native structures.[84] In contrast, there is no common features shared by 1E0M, Betanova, and 1S04, and two solutions can be proposed.

A more complex energy function may be required. Possible areas for improvement include higher-order ϕ-ψ angle pairs terms,[85] and many-body interactions between the side-chains.[86]

Alternatively, it is possible that all force fields will fail to discriminate native from nonnative states, because the experimental structure is not associated with the minimal effective energy, but with the global minimum of effective energy and conformational entropy.[57]

ACKNOWLEDGMENTS

P.D. thank Adrien Melquiond for the MD and threading-based structures, and Jean-Francois Saint-Pierre and Normand Mousseau for the ART-generated structures.

REFERENCES

1. Buchete NV, Straub JE, Thirumalai D. Development of novel statistical potentials for protein fold recognition. Curr Opin Struct Biol 2004;14:225–232.
2. Mackerell Jr, AD. Empirical force fields for biological macromolecules: overview and issues. J Comput Chem 2004;25:1584–1604.
3. Ulmschneider JP, Jorgensen WL. Polypeptide folding using Monte Carlo sampling, concerted rotation, and continuum solvation. J Am Chem Soc 2004;126:1849–1857.
4. Jang S, Kim E, Shin S, Pak Y. Ab initio folding of helix bundle proteins using molecular dynamics simulations. J Am Chem Soc 2003; 125:14841–14846.
5. Liu Y, Beveridge DL. Exploratory studies of ab initio protein structure prediction: multiple copy simulated annealing, AMBER energy functions, and a generalized born/solvent accessibility solvation model. Proteins 2002;46:128–146.
6. Zagrovic B, Snow CD, Shirts MR, Pande V. Simulation of folding of an alpha-helical protein in atomistic detail using worldwide distributed computing. J Mol Biol 2002;323:927–937.
7. Kussell E, Shimada J, Shakhnovich EI. A structure-based method for derivation of all-atom potentials for protein folding. Proc Natl Acad Sci USA 2002;99:5343–5348.
8. Irback A, Mohanty S. Folding thermodynamics of peptides. Biophys J 2005;88:1560–1569.
9. Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 1976;104:59–107.
10. Nielsen SO, Lopez CF, Srinivas G, Klein ML. Coarse grain models and the computer simulation of soft materials. J Phys Condens Matter 2004;16:R481–R512.
11. Tozzini V. Coarse-grained models for proteins. Curr Opin Struct Biol 2005;15:144–150.
12. Bahar I, Jernigan RL. Vibrational dynamics of transfer RNAs: comparison of the free and synthetase-bound forms. J Mol Biol 1998;281:871–884.
13. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213:859–883.
14. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
15. Oldziej S, Czaplewski C, Liwo A, Chinchio M, Nanias M, Vila JA, Khalili M, Arnautova YA, Jagielska A, Makowski M, Schafroth HD, Kazmierkiewicz R, Ripoll DR, Pillardy J, Saunders JA, Kang YK, Gibson KD, Scheraga HA. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests. Proc Natl Acad Sci USA 2005;102: 7547–7552.
16. Shih AY, Arkhipov A, Freddolino PL, Schulten K. Coarse grained protein-lipid model with application to lipoprotein particles. J Phys Chem B 2006;110:3674–3684.
17. Zhang C, Liu S, Zhou H, Zhou Y. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Sci 2004;13:400–411.
18. Bond PJ, Sansom MS. Insertion and assembly of membrane proteins via simulation. J Am Chem Soc 2006;128:2697–2704.
19. Srinivasan R, Fleming PJ, Rose GD. Ab initio protein folding using LINUS. Methods Enzymol 2004;383:48–66.
20. Wallqvist A, Ullner M. A simplified amino acid potential for use in structure predictions of proteins. Proteins 1994;18:267–280.
21. Zacharias M. Protein–protein docking with a reduced protein model accounting for side-chain flexibility. Protein Sci 2003;12: 1271–1282.
22. Fujitsuka Y, Takada S, Luthey-Schulten ZA, Wolynes PG. Optimizing physical energy functions for protein folding. Proteins 2004;54: 88–103.
23. Derreumaux P. From polypeptide sequences to structures using Monte Carlo simulations and an optimized potential. J Chem Phys 1999;111:2301–2310.
24. Gibbs N, Clarke AR, Sessions RB. Ab initio protein structure prediction using physicochemical potentials and a simplified off-lattice model. Proteins 2001;43:186–202.
25. Peng S, Ding F, Urbanc B, Buldyrev SV, Cruz L, Stanley HE, Dokholyan NV. Discrete molecular dynamics simulations of peptide aggregation. Phys Rev E Stat Nonlin Soft Matter Phys 2004;69(4 Part 1):041908.
26. Favrin G, Irback A, Mohanty S. Oligomerization of amyloid Abeta16-22 peptides using hydrogen bonds and hydrophobicity forces. Biophys J 2004;87:3657–3664.
27. Santini S, Wei G-H, Mousseau N, Derreumaux P. Pathway complexity of Alzheimer's β-amyloid $A\beta_{16-22}$ peptide assembly. Structure 2004;12:1245–1255.
28. Santini S, Mousseau N, Derreumaux P. In silico assembly of Alzheimer's Abeta16-22 peptide into beta-sheets. J Am Chem Soc 2004; 126:11509–11516.
29. Gatchell DW, Dennis S, Vajda S. Discrimination of near-native protein structures from misfolded models by empirical free energy functions. Proteins 2000;41:518–534.
30. Mirny LA, Shakhnovich EI. How to derive a protein folding potential? A new approach to an old problem. J Mol Biol 1996;264:1164–1179.
31. Qiu J, Elber R. Atomically detailed potentials to recognize native and approximate protein structures. Proteins 2005;61:44–55.
32. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins 2001;44: 223–232.
33. Kolinski A, Godzik A, Skolnick J. A general method for the prediction of the three dimensional structure and folding pathway of globular proteins: application to designed helical proteins. J Chem Phys 1993;98:7420–7433.

34. Gan HH, Tropsha A, Schlick T. Lattice protein folding with two and four-body statistical potentials. Proteins 2001;43:161–174.
35. Mayewski S. A multibody, whole-residue potential for protein structures, with testing by Monte Carlo simulated annealing. Proteins 2005;59:152–169.
36. Derreumaux P. Generating ensemble averages for small proteins from extended conformations by Monte Carlo simulations. Phys Rev Lett 2000;85:206–209.
37. Forcellino F, Derreumaux P. Computer simulations aimed at structure predictions of supersecondary motifs in proteins. Proteins 2001;45:159–166.
38. Derreumaux P. Insights into protein topology from Monte Carlo simulations. J Chem Phys 2002;117:3499–3503.
39. Malek R, Mousseau N. Dynamics of lennard-jones clusters: a characterization of the activation-relaxation technique Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 2000;62(6 Part A): 7723–7728.
40. Wei G, Mousseau N, Derreumaux P. Exploring the energy landscape of proteins: a characterization of the activation relaxation technique. J Chem Phys 2002;117:11379–11387.
41. Wei G, Mousseau N, Derreumaux P. Complex folding pathways in a β-hairpin. Proteins 2004;56:464–474.
42. Mousseau N, Derreumaux P. Exploring the early steps of amyloid peptide aggregation by computers. Acc Chem Res 2005;38:885–891.
43. Coincon M, Heitz A, Chiche L, Derreumaux P. The beta alpha beta alpha beta elementary supersecondary structure of the Rossmann fold from porcine lactate dehydrogenase exhibits characteristics of a molten globule. Proteins 2005;60:740–745.
44. 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.
45. Guo H, Salahub DR. Cooperative hydrogen bonding and enzyme catalysis. Angew Chem Int Ed Engl 1998;37:2985–2990.
46. Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM. Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature 2003;424:805–808.
47. Wei G, Derreumaux P, Mousseau N, Sampling the complex energy landscape of a simple β-hairpin. J Chem Phys 2003;119:6403–6406.
48. Santini S, Wei G, Mousseau N, Derreumaux P. Exploring the folding pathways of proteins through energy landscape sampling: application to Alzheimer's β-amyloid peptide. Internet Electron J Mol Des 2003;2:564–577.
49. Case DA, Cheatham, TE, III, Darden T, Gohlke H, Luo R, Merz KM, Jr, Onufriev A, Simmerling C, Wang B, Woods R. The Amber biomolecular simulation programs. J Comput Chem 2005;26:1668–1688.
50. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 1999;8:361–369.
51. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Sci 1997;6:676–688.
52. Zhang Y, Kolinski A, Skolnick J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J 2003; 85:1145–1164.
53. Kortemme T, Ramirez-Alvarado M, Serrano L. Design of a 20-amino acid, three-stranded β-sheet protein. Science 1998;281:253–256.
54. Berendsen HJC, van der Spoel D, van Drunen R. GROMACS: a message-passing parallel molecular dynamics implementation. Comp Phys Commun 1995;91:43–56.
55. Tuffery P, Guyon F, Derreumaux P. An improved greedy algorithm for protein structure reconstruction. J Comput Chem 2005;26:506–513.
56. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
57. Caflisch A. Network and graph analyses of folding free energy surfaces. Curr Opin Struct Biol 2006;16:71–78.
58. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
59. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57:702–710.
60. Boutselakis H, Dimitropoulos D, Fillon J, Golovin A, Henrick K, Hussain A, Ionides J, John M, Keller PA, Krissinel E, McNeil P, Naim A, Newman R, Oldfield T, Pineda J, Rachedi A, Copeland J, Sitnov A, Sobhany S, Suarez-Uruena A, Swaminathan J, Tagari M, Tate J, Tromm S, Velankar S, Vranken W. E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. Nucleic Acids Res 2003;31:458–462.
61. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.
62. Park B, Levitt M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 1996; 258:367–392.
63. Samudrala R, Levitt M. Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. Protein Sci 2000;9:1399–1401.
64. Samudrala R, Levitt M. A comprehensive analysis of 40 blind protein structure predictions. BMC Struct Biol 2002;2:1–16.
65. Xia Y, Huang ES, Levitt M, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. J Mol Biol 2000;300:171–185.
66. Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D. An improved protein decoy set for testing energy functions for protein structure prediction. Proteins 2003;53:76–87.
67. Moult J, Fidelis K, Rost B, Hubbard T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)–round 6. Proteins 2005;61(Suppl 7):3–7.
68. DeLano WL. The PyMOL Molecular Graphics System. San Carlos, CA, USA: DeLano Scientific; 2002.
69. Tuffery P. XmMol: an X11 and motif program for macromolecular visualization and modeling. J Mol Graph 1995;13:67–72, 62.
70. Macias MJ, Gervais V, Civera C, Oschkinat H. Structural analysis of WW domains and design of a WW prototype. Nat Struct Biol 2000;7:375–379.
71. 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.
72. Soto P, Colombo G. Characterization of the conformational space of a triple-stranded β-sheet forming peptide with molecular dynamics simulations. Proteins 2004;57:734–746.
73. Leijonmarck M, Liljas A. Structure of the C-terminal domain of the ribosomal protein L7/L12 from *Escherichia coli* at 1.7 A. J Mol Biol 1987;195:555–579.
74. Eramian D, Shen MY, Devos D, Melo F, Sali A, Marti-Renom MA A composite score for predicting errors in protein structure models. Protein Sci 2006;15:1653–1666.
75. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006;15:2507–2524.
76. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;Suppl 3:171–176.
77. Misura KM, Chivian D, Rohl CA, Kim DE, Baker D. Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci USA 2006;103:5361–5366.
78. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726. Erratum in: Protein Sci 2003;12(9):2121.
79. Melo F, Sanchez R, Sali A. Statistical potentials for fold assessment. Protein Sci 2002;11:430–448.

80. Canutescu AA, Shelenkov AA, Dunbrack RL Jr. A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 2003;12:2001–2014.
81. Derreumaux P, Mousseau N. Coarse-grained protein molecular dynamics simulations. J Chem Phys 2006;126:025101–025106.
82. Chen W, Mousseau N, Derreumaux P. The conformations of the amyloid-β (21–30) fragment can be described by three families in solution. J Chem Phys 2006;125:084911.
83. Wei GH, Mousseau N, Derreumaux P. Computational simulation of the early steps of protein aggregation. Prion 2007;1:3–8.
84. McConkey BJ, Sobolev V, Edelman M. Discrimination of native protein structures using atom-atom contact scoring. Proc Natl Acad Sci USA 2003;100:3215–3220.
85. Sims GE, Kim SH. A method for evaluating the structural quality of protein models by using higher-order phi-psi pairs scoring. Proc Natl Acad Sci USA 2006;103:4428–4432.
86. Shimizu S, Chan HS. Anti-cooperativity and cooperativity in hydrophobic interactions: three-body free energy landscapes and comparison with implicit-solvent potential functions for proteins. Proteins 2002;48:15–30. Erratum in: Proteins 2002;49(2):294.