WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Native atomic burials, supplemented by physically motivated hydrogen bond constraints, contain sufficient information to determine the tertiary structure of small globular proteins

Antônio F. Pereira de Araújo,[1]* Antonio L. C. Gomes,[1] Alexandre A. Bursztyn,[1] and Eugene I. Shakhnovich[2]

[1] Laboratório de Biologia Teórica, Departamento de Biologia Celular, Universidade de Brasília, Brasília-DF 70910-900, Brazil

[2] Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138

ABSTRACT

We investigate the possibility that atomic burials, as measured by their distances from the structural geometrical center, contain sufficient information to determine the tertiary structure of globular proteins. We report Monte Carlo simulated annealing results of all-atom hard-sphere models in continuous space for four small proteins: the all-β WW-domain 1E0L, the α/β protein-G 1IGD, the all-α engrailed homeo-domain 1ENH, and the α + β engineered monomeric form of the Cro protein 1ORC. We used as energy function the sum over all atoms, labeled by i, of $|R_i - R_i^*|$, where R_i is the atomic distance from the center of coordinates, or central distance, and R_i^* is the "ideal" central distance obtained from the native structure. Hydrogen bonds were taken into consideration by the assignment of two ideal distances for backbone atoms forming hydrogen bonds in the native structure depending on the formation of a geometrically defined bond, independently of bond partner. Lowest energy final conformations turned out to be very similar to the native structure for the four proteins under investigation and a strong correlation was observed between energy and distance root mean square deviation (DRMS) from the native in the case of all-β 1E0L and α/β 1IGD. For all α 1ENH and α + β 1ORC the overall correlation between energy and DRMS among final conformations was not as high because some trajectories resulted in high DRMS but low energy final conformations in which α-helices adopted a non-native mutual orientation. Comparison between central distances and actual accessible surface areas corroborated the implicit assumption of correlation between these two quantities. The Z-score obtained with this native-centric potential in the discrimination of native 1ORC from a set of random compact structures confirmed that it contains a much smaller amount of native information when compared to a traditional contact Go potential but indicated that simple sequence-dependent burial potentials still need some improvement in order to attain a similar discriminability. Taken together, our results suggest that central distances, in conjunction to physically motivated hydrogen bond constraints, contain sufficient information to determine the native conformation of these small proteins and that a solution to the folding problem for globular proteins could arise from sufficiently accurate burial predictions from sequence followed by minimization of a burial-dependent energy function.

Proteins 2008; 70:971–983.

© 2007 Wiley-Liss, Inc.

Key words: protein folding; structure prediction; computer simulation; hydrophobic potential; atomic burial.

INTRODUCTION

Protein folding and structure prediction can be approached from different, complementary, perspectives. Protein molecules can arguably be considered, in a certain sense, as simple analogs of biological systems in which genetic information (protein sequences), which is subject to natural selection and evolution, encodes the formation of individual organisms (protein structures). From this biological perspective protein evolution and folding would correspond, respectively, to phylogenesis and ontogenesis of cells, animals and plants. A crucial difference, however, is that proteins can be unfolded and

Grant sponsor: NIH; Grant number: R03 TW006859; Grant sponsors: CNPq; Millineium Institute for Structural Biology in Biomedicine and Biotechnology (MISB3).

*Correspondence to: Antônio F. Pereira de Araújo, Laboratório de Biologia Teórica, Departamento de Biologia Celular, Universidade de Brasília, Brasília-DF 70910-900, Brazil. E-mail: aaraujo@unb.br

Received 5 November 2006; Revised 28 March 2007; Accepted 16 April 2007

Published online 10 September 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21571

© 2007 WILEY-LISS, INC.

refolded reversibly while biological death is intrinsically irreversible. It is possible, therefore, to consider protein native structures as physical systems at thermodynamic equilibrium and protein folding as the relaxation towards these final most stable thermodynamic states. Thermodynamic stability implies that in protein models with explicit consideration of different chain conformations, but not of solvent degrees of freedom, the native structure should correspond to a sufficiently deep global energy minimum in order to be lower in free energy than the entropically very favorable unfolded state.[1,2] Since conformations that are close to each other in conformational space are expected to have correlated energies, kinetic accessibility should follow automatically as the result of the formation of a funnel-shaped free energy surface.[3–5] Native stability has also been stated in terms of the ratio T_f/T_g between the folding temperature, reflecting the bias in free energy towards the native structure, and a glass transition temperature, reflecting the roughness of the free energy surface.[6] The stability hypothesis has been corroborated by many studies with minimalist models, as reviewed in[7–12] but direct transfer of this simple idea to models with sufficient detail to simulate real proteins using information exclusively from the sequence is yet to be shown. Successful simulations of realistic models would actually provide unprecedented insight on the folding mechanism and, as corollary, an algorithm for structure prediction from amino acid sequence.

Ab initio folding simulations using physical potentials have been reported for quite some time.[13] Although significant developments have emerged during the last three decades, not only from increasing computing capacity but also on models, potentials, and efficient exploration of conformational space, it is apparent that our physical understanding of the system is still incomplete and available potentials need significant improvement in order to become reliable in general. In particular, α-helical proteins appear to be more easily predictable than proteins rich in β-sheets.[14] An illustrative example is provided by a recent study using an united residue (UNRES) model and force field, in which folding trajectories were simulated and native-like structures obtained for some members of a study group containing seven small globular proteins although the small all-β 1E0L tended to fold to α-helical non-native conformations.[15] Accordingly, methods based on structural homology of substructures maybe as small as a few residues, like ROSETTA,[16] are apparently displacing physics-based simulations in CASP (Critical Assessment of Structure Prediction) experiments even in the category of template-free modeling.[17,18]

An alternative approach has been the use of *a priori* structural information to bias the energy explicitly towards the native conformation. In so-called Go potentials, for example, pairwise native interactions are attractive while non-native interactions are either neutral or repulsive, and the native structure becomes a deep global energy minimum by construction.[19] Monte Carlo simulations with Go potentials have been used to fold small protein molecules in an all-heavy-atom representation. In addition to interesting results regarding the folding mechanism of specific proteins, including the suggestion of alternative interpretations of experimental results, these simulations have demonstrated that the ab initio folding of protein models with a realistic level of structural detail is feasible with present-day computers.[20–23] Attempts to transform Go potentials into less specific functions, which could be transferable between different proteins, have been performed through a reduction in the number of effective atomic types and some encouraging results have been reported.[24,25] Agreement, for some proteins, between the transition state ensemble observed in folding simulations using Go potentials and experimental ϕ-values have also been used as an evidence of the prevalence of the native structure over the stabilizing energy function on determining the folding mechanism,[26] as had been previously suggested by the experimental ϕ-value conservation between different sequencies with similar structures.[27–29] Limitations to this interesting hypothesis have also been reported, however.[30,31]

It is interesting to note that the main caveat of Go potentials for eventual structure prediction from sequence is not the use, in itself, of a priori information about the native structure but that the amount of this information, in the sense of Shannon information theory,[32,33] is too large to be encodable in the sequence. Estimating the number contact matrices for a chain of M monomers, $\Omega_c(M)$, as the number of $M \times M$ symmetric matrices with two types of nondiagonal elements corresponding to "contact" and "not contact", or $\Omega_c(M) = 2^{M(M-1)/2}$, the amount of information, in bits, required to specify one such matrix would be given by

$$H_c(M) = \log_2(\Omega_c(M)) = \frac{M^2 - M}{2} \qquad (1)$$

which grows with M^2. The amount of information transmitted by a sequence composed of twenty amino acid types, on the other hand, cannot be larger than

$$H_s(M) = \log_2(\Omega_s(M)) = \log_2(20^M) = 4.32M \qquad (2)$$

which is proportional to M. Since $H_c(M) > H_s(M)$ for $M \geq 10$, the native contact matrix required by the Go potential cannot be directly encoded even in small protein sequences. Note that H_c, but not H_s, would be even larger for contact matrices between atoms, instead of monomers, and that consideration of different frequencies for individual "letters" and/or possible correlations between them, although likely to improve the estimates for the two Shannon entropies, would not affect the essential difference in their dependence on protein size.

It would actually be extremely useful, therefore, to identify parameters in protein native structures containing an amount of information sufficiently large to uniquely determine the native state among all alternative conformations and, at the same time, sufficiently small to be encodable in protein sequences. In particular, it is reasonable to expect that the amount of this information should increase only linearly with protein size, like the information content of protein sequences, and not like the native contact matrix which increases with the square of protein size. Recent theoretical studies and simulations of minimalist lattice models have shown that a non-specific hydrophobic potential can reproduce many aspects of protein folding behavior, both in two[34–38] and three[31,39] dimensions, for appropriate, "segregated", native structures using only information about native contact vectors, i.e., the number of contacts made by each monomer in the native structure,[40,41] an amount of information that increases only linearly with sequence length and is directly encoded in a sequence of monomer "hydrophobicities". The energetic contribution of each monomer in a given conformation in these models is simply the negative product between the number of contacts it makes and its hydrophobicity. Since hydrophobic monomers (positive hydrophobicity) decrease the energy when forming a contact while the reverse is true for hydrophilic monomers (negative hydrophobicity) this function can be considered to mimic, in a very simplified way, the hydrophobic effect, which is believed to be the dominant factor in the stabilization of protein structures.[42,43] The number of contacts made by each monomer is taken, in this case, as an appropriate measure of its exposure to the solvent or its burial inside the globular structure.

In the present study, partially inspired by these previous results with minimalist models, we explore the possibility that atomic native burials, as measured by some convenient parameter whose information increases only linearly with sequence length, might uniquely determine the tertiary structure of small globular proteins. If we are able to fold small globular protein using only information about native burial vectors and if we are also able to predict the most probable atomic burials, or hydrophobicities, from the amino acid sequence alone then we would have a powerful approach to the eventual solution of the protein folding problem, as illustrated in the following scheme:

$$\begin{tikzcd} \text{sequence} \arrow[rr, "?"] \arrow[dr, "1"'] & & \text{3D structure} \\ & \text{atomic burials} \arrow[ur, "2"'] & \end{tikzcd}$$

The present study, therefore, deals mainly with step 2 above, i.e., the prediction of the tertiary structure from atomic burials. Some preliminary but instructive results regarding step 1 are also discussed, however. We have decided not to use the number of atomic contacts as a measure of atomic burial in the present continuous models because large changes in contact numbers can be expected to result from subtle conformational variations and many different conformations could be adjusted to the same contact vector. Accessible surface areas would be more discriminative in this respect but their calculation during folding simulations would be computationally expensive. We have chosen therefore to use atomic distances from the molecular geometrical center as a measure of atomic burial in globular proteins. Although less general, since correlation with solvent exposure cannot be expected for proteins with arbitrary shapes, these distances are more discriminative than the number of contacts and easier to compute than accessible surfaces. Additionally, it is crucial that the amount of information contained in the resulting atomic burials increases only linearly with protein size, since each atom corresponds to a single distance value. Numerical estimates for the corresponding Shannon entropy are beyond the scope of the present study but it is clear, on one hand, that the necessary burial correlation between covalently bonded atoms will play a significant role in decreasing the entropy with respect to hypothetical sequences of uncorrelated numbers while, on the other hand, the entropy should increase with the number of burial levels to be considered as distinguishable.

METHODS

We have performed Monte Carlo simulated annealing, with Metropolis algorithm,[44] for four small globular proteins in continuous space: the all-β WW-domain 1E0L,[45] the α/β protein-G 1IGD,[46] the all-α engrailed homeo-domain 1ENH,[47] and the α + β engineered monomeric form of the Cro protein 1ORC.[48] The computer program was adapted from the code that has been used in simulations with Go potentials.[20,21] All heavy atoms of the protein are represented explicitly as hard spheres connected by covalent bonds with rigid lengths and angles. Move attempts are generated by small changes in ϕ, ψ and χs dihedral angles of randomly chosen residues. The energy function in the present case, in addition to the implicit infinite term accounting for excluded volume, was initially taken simply as the sum of the absolute value of the difference between the distance of each atom i from the center of coordinates, $R_\mathrm{i} = \sqrt{x_\mathrm{i}^2 + y_\mathrm{i}^2 + z_\mathrm{i}^2}$, and its "ideal" value R_i^* corresponding to its distance from the molecular geometrical center in the native structure:

$$E(\{\vec{R}_\mathrm{i}\}) = E(\{R_\mathrm{i}\}) = K \sum_\mathrm{i} |R_\mathrm{i} - R_\mathrm{i}^*|, \qquad (3)$$

where K is a constant intended to properly convert the distance dimension, resulting from the sum of central distances, into some other convenient dimension. In the present study we have used $K = 1$ Å^{-1} in order to compute the energy in adimensional units. It should be stressed that this expression is referred to as energy only because of the scoring role it plays in the Metropolis algorithm and not because of any implied relation to physical energy. Simulations were also performed with two ideal distance values for backbone atoms participating in hydrogen bonds in the native structure, depending on whether they are forming or not a geometrically defined hydrogen bond, independently of bond partner. This modification results in a constraint that effectively couples ideal atomic burials of polar backbone atoms to hydrogen bond formation.

Hydrogen bonds were defined by the distance, d, between hydrogen donor and hydrogen acceptor and two angles, θ_1 and θ_2, involving the hydrogen bond direction determined by the straight line connecting these two atoms.[49,50] θ_1 is the angle between this hydrogen bond direction and the expected direction of the covalent bond between hydrogen donor and donated hydrogen while θ_2 is the angle between the hydrogen bond direction and the double bond connecting the hydrogen acceptor to the adjacent carbon atom. For the initial constrained simulations with the WW-domain, 1EN0, hydrogen bonds were considered to arise when the three following criteria were satisfied simultaneously for any pair of potential donor and acceptor: 2.6 Å $< d <$ 3.2 Å, $\theta_1 < 20^\circ$ and $\theta_2 < 60^\circ$. For the other three proteins we have used a somewhat more relaxed hydrogen bond definition: 2.5 Å $< d <$ 3.8 Å, $\theta_1 < 45^\circ$ and $\theta_2 < 60^\circ$. All backbone nitrogen and oxygen atoms were considered as potential hydrogen donors and acceptors, respectively, in all simulations. For potential donors and acceptors actually forming a hydrogen bond in the native structure the ideal distance R_i^* in Eq. (1) was the usual native distance from the geometrical center whenever the atom was forming a hydrogen bond, independently of bond partner, and the somewhat arbitrary value of 14 Å, which is larger than the native distance of most buried hydrogen bonds in these small proteins, when this was not the case.

Initial conformations were generated by unfolding at very high temperature, T = 100,000, during 10 million (10^7) time steps. Simulation temperature during folding was decreased very slowly from a high initial value at which essentially all move attempts are accepted to a very low temperature at which the chain effectively freezes. Annealing trajectories for 1EN0 without hydrogen bond constraints began at temperature T = 1000 which was slowly decreased by a factor of 0.99 every 2 million (2×10^6) time steps for a total of 2 billion (2×10^9) time steps. For constrained simulations of the same protein the initial temperature was T = 200 and it was decreased every 1 million (10^6) time steps by the factor 0.99 during 1 billion (10^9) time steps. For the other three proteins, for which only constrained simulations were performed, an initial temperature of 1000 was decreased by the same factor 0.99 every 2.5 million (2.5×10^6) time steps for 2 billion (2×10^9) time steps. Similarity to the native structure was measured by the orientation-independent C_α distance root mean square deviation (DRMS) defined as the root mean square deviation of the distances between all pairs of C_α atoms in an arbitrary structure from the same distances in the native conformation.

RESULTS

Figure 1 shows representative trajectories for the all-β WW domain 1E0L with and without hydrogen bond constraints. Along each trajectory, we show the energy as a function of time step (a,b), DRMS as function of time step (c,d) and DRMS as function of energy (e,f). Two trajectories are shown in each panel, corresponding to the best and worst final conformations, when compared to the native structure, in a set of 10 trajectories without hydrogen bond constraints (a,c,e) or 15 constrained trajectories (b,d,f). Different simulations with the same energy function display the same behavior at high temperatures, with DRMS and energy tending to decrease monotonically as temperature is slowly decreased, as seen in the initial parts of the trajectories. At low temperatures, however, each trajectory freezes at different points in energy and DRMS. The final, low temperature, portion of the constrained trajectories shown in Figure 1(b,d) are also shown in Figure 2(a,b) in nonlogarithmic scale.

For the unconstrained potential, final structures had moderately high DRMS values, ranging from around 2 to 4 Å. Final DRMS values decreased significantly with the addition of hydrogen bond constraints, ranging from as low as 0.7 Å to slightly above 2 Å. Very significantly, a strong correlation between final energies and DRMS is observed for the constrained potential (Pearson's correlation coefficient $\mathcal{C} = 0.76$ for 15 simulations) but not for the unconstrained potential ($\mathcal{C} = 0.37$ for 10 simulations), as seen in Figure 3(a), with energy values divided by the number of atoms in the protein. It is possible, therefore, that an improved conformational search could result in lower final energies for the unconstrained potential but lower DRMS values are unlikely to be obtained without consideration of hydrogen bonds. Representative final conformations for 1E0L are shown in the first row of Figure 4 together with the native conformation, which is shown in Figure 4(a). As illustrated by the conformation shown in Figure 4(d), final structures from unconstrained trajectories tend to reveal the general correct native topology although with poor secondary structure formation. Hydrogen bond constraints greatly improve secondary structure formation, even for the less successful trajectories, as illustrated in Figure 4(c), when compared to the lowest energy, essentially native, final conformation shown in Figure 4(b).

DOI 10.1002/prot

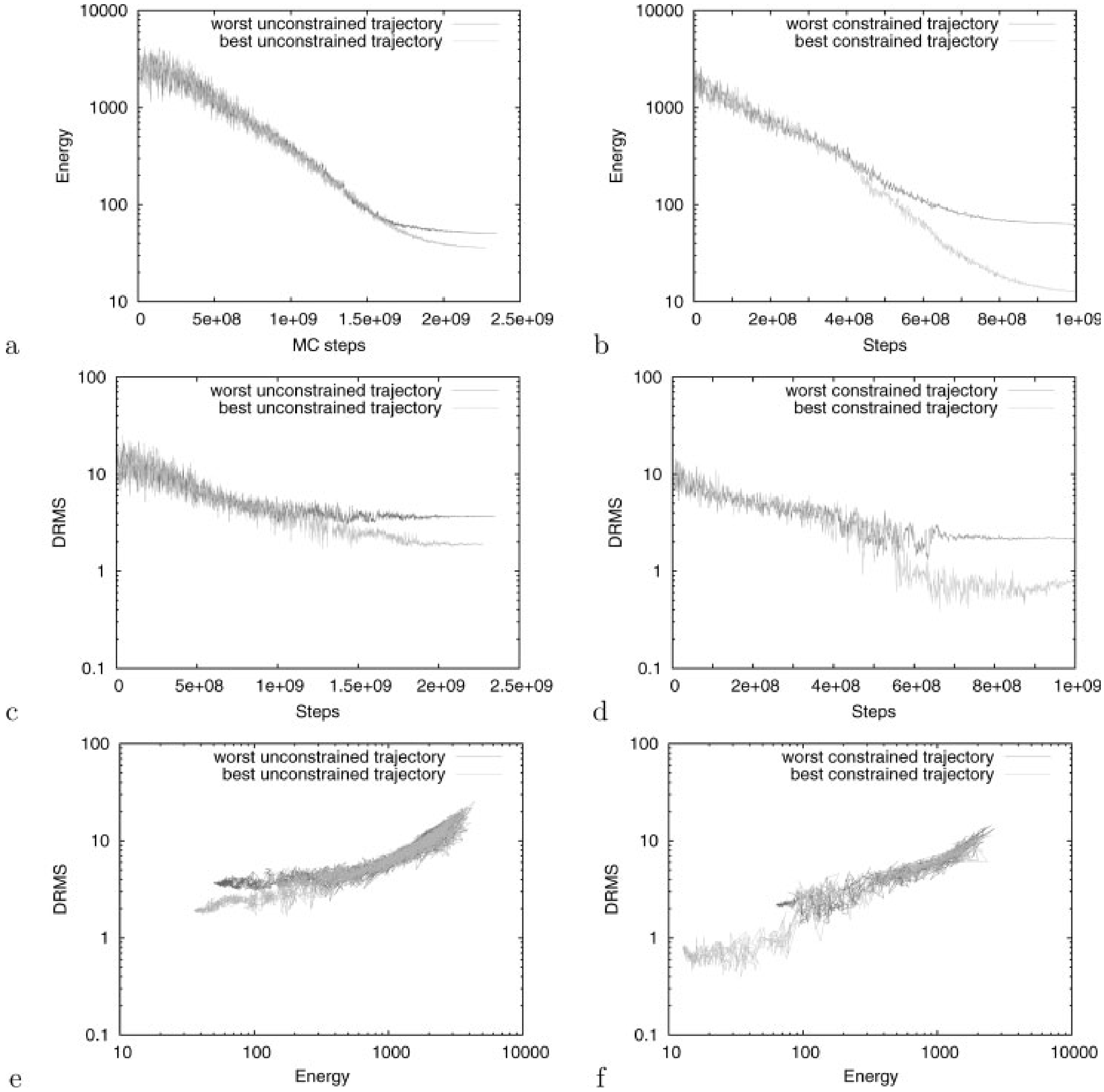

Figure 1

Annealed folding trajectories with the unconstrained energy function (left) and with hydrogen constraints (right) for the WW domain 1E0L. The two curves in each plot correspond to the worst (dark gray) and best (light gray) results obtained with each energy function. Temperature is slowly decreased as the number of move attempts increases. Structural dissimilarity to the native conformation as measured by C_α DRMS (top) and energy (middle) are plotted as a function of move attempts and as a function of each other (bottom).

For the other three proteins we report results only for the constrained potential. As shown in Figure 3(b), DRMS and energy values for 12 final trajectories of 1IGD are also strongly correlated, with $\mathcal{C} = 0.82$. The lowest energy conformation [Fig. 4(f)] is essentially native [Fig. 4(e)] with DRMS of 0.67 Å. We label this structure as "A" in Figure 3(b). A second group of conformations with DRMS between 2 and 3 Å is observed with energies per atom around 0.3. Conformations in this group, which are labeled by "B" in Figure 3(b), can be considered to be "almost folded" because they have the overall correct topology, with first and last strands of the β-sheet pairing together in the correct parallel orientation, but with slightly poorer secondary structure formation, as illustrated in Figure 4(g). Remaining high energy final conformations tend to also have high DRMS and they are all characterized either by not having the initial and final regions of the chain pairing to each other

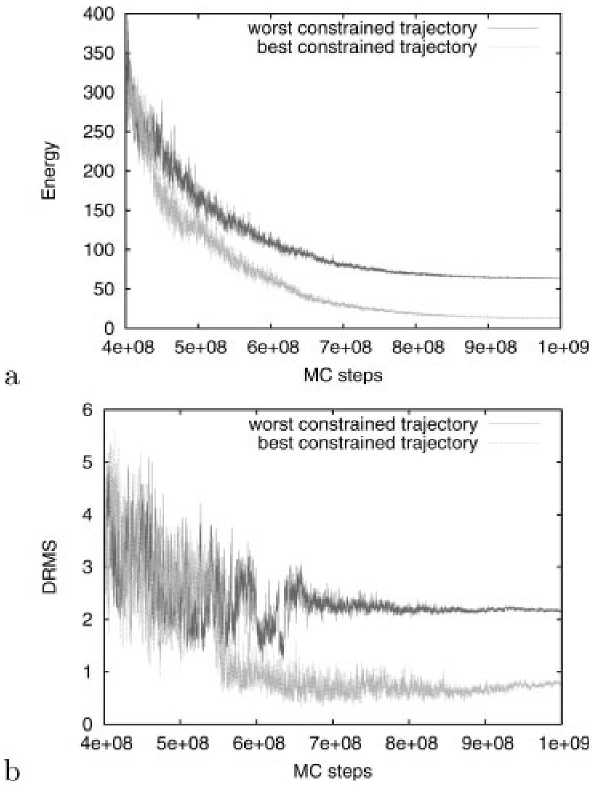

Figure 2

Final portion of trajectories shown in Figure 1(b, d) in nonlog scale.

as β-strands of a β-sheet [conformations labeled by "C" in Fig. 3(b)] or by having them interestingly pairing in the incorrect antiparallel orientation [conformations labeled by "D" in Fig. 3(b)], an example of which is shown in Figure 4(h).

For 1ENH, whose native structure can be seen in Figure 4(i), 11 out of 12 final conformations display the overall correct native topology, varying almost continuously from essentially native structures with DRMS between 1.5 and 2 Å and energy per atom around 0.15, labeled by "A" in Figure 3(c) and illustrated in Figure 4(j), through somewhat poorer conformations with DRMS between 2 and 2.5 Å and energy per atom around 0.2, labeled by "B" in Figure 3(c), to conformations with DRMS between 2.5 and 3 Å and energy per atom ranging from slightly below 0.2 to slightly above 0.3, labeled by "C" in Figure 3(c) and illustrated in Figure 4(k). The correlation coefficient between energy and DRMS for these eleven structures is $\mathcal{C} = 0.78$. The twelfth structure, however, which is labeled by "D" in Figure 3(c) and shown in Figure 4(l), has all helices properly formed but arranged in a non-native orientation, resulting in an energy per atom below 0.2 and high DRMS around 3.5 Å. The correlation between DRMS and energy for all twelve structures is therefore significantly lower, with $\mathcal{C} = 0.47$. It is relevant that even with all three helices well formed, this misfolded "D" conformation is higher in energy than the correctly folded "A" structures.

Final DRMS and energy values for 10 1ORC trajectories are shown in Figure 3(d). Two trajectories resulted in essentially native conformations, labeled by "A", with energy per atom lower than 0.2 and DRMS between 1.5 and 2 Å. All secondary structure elements are properly formed and adopt the correct global topology. The structure of one of them is shown in Figure 4(n) while the native structure is shown in Figure 4(m). Another final structure, labeled by "B", also has low DRMS around 2 Å but higher energy around 0.27, mainly because of poorer secondary structure formation. Three trajectories resulted in final conformations with high energy, above 0.3, and high DRMS, between 3 and 4 Å. The three strands in the native β-sheet that are adjacent along the sequence are not folded correctly in these structures, labeled by "C" in Figure 3(d), since corresponding segments are either not adopting a β-strand conformation or are not hydrogen bonding to the same strands as in the native structure, as illustrated in Figure 4(o). For four final conformations labeled by "D" in Figure 3(d), one of which can be seen in Figure 4(p), the three adjacent β-strands are folded correctly but the three helices of the α-helical region, although properly formed, adopt a non-native arrangement among themselves and prevent the correct positioning of the first β-strand in the β-sheet. The behavior of the α region of this α + β protein is therefore similar to the one already observed for all-α 1ENH and the correlation between energy and DRMS is accordingly increased, from $\mathcal{C} = 0.56$ with 10 structures to $\mathcal{C} = 0.92$ with six structures, when "D" conformations are excluded from consideration. The proportion of these conformations with non-native α-helical arrangement is however higher for 1ORC than for 1ENH (4/10 to 1/12) in these small sets of final structures.

As previously discussed, central distances in globular proteins are expected to be correlated to solvent exposure, in which case preferential distances could possibly be estimated from atomic hydrophobicities predicted from the sequence. For nonglobular structures, however, central distances could provide more detailed geometric information not available in actual burials. In this case, structure prediction from central distances would be facilitated but, on the other hand, eventual prediction from sequence would be more difficult since no correlation between central distances and hydrophobicities should be expected. In a recent statistical analysis of protein structures,[51] globular proteins were considered to be both sufficiently spherical, with $\mathcal{A} < 0.1$, where $\mathcal{A}$ is the "asphericity" structural parameter computed from the eigen vectors of the gyration (or shape) tensor,[52] and compact, with $\mathcal{B} = R_g/N^{1/3} < 2.9$, where $\mathcal{B}$ is the ratio between the radius of gyration R_g and the cubic root of the number of residues N. In the present case, 1ENH,

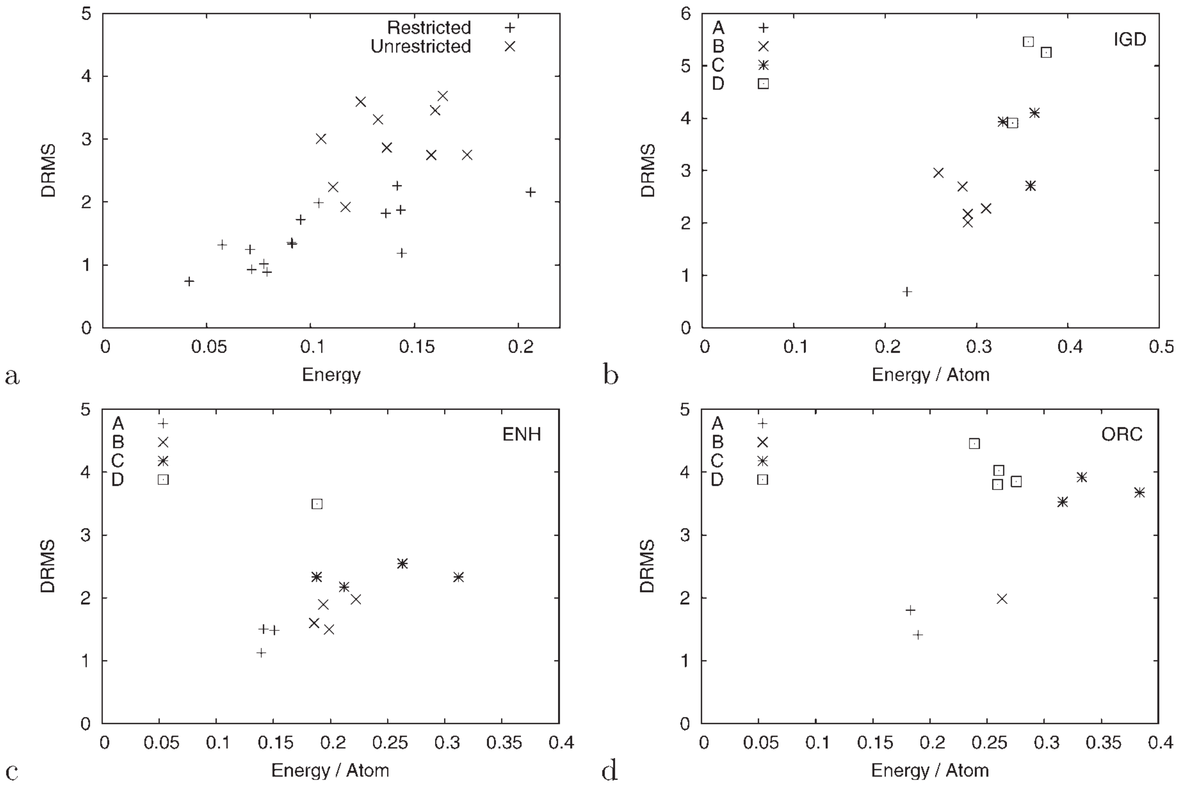

Figure 3

*Final energy and DRMS for 1E0L (**a**), 1IGD (**b**), 1ENH (**c**), and 1ORC (**d**). Each point corresponds to the final conformation of one trajectory. Two types of points are used in (a) to distinguish trajectories with and without hydrogen bond constraints. Different types of points are shown in the other three plots to label different classes of final conformations. "A" and "B" correspond to final conformations considered as correctly folded and almost folded, respectively, while "C" and "D" correspond to not folded or misfolded structures, as described in the text.*

with $\mathcal{A} = 0.005$ and $\mathcal{B} = 2.76$, and 1ORC, with $\mathcal{A} = 0.038$ and $\mathcal{B} = 2.80$, would be considered to be globular by these criteria but not 1IGD, with $\mathcal{A} = 0.155$ and $\mathcal{B} = 2.77$, or 1E0L, with $\mathcal{A} = 0.179$ and $\mathcal{B} = 2.95$. From visual inspection of the four native structures it is apparent that 1ENH and 1ORC are actually more spherical when compared to the somewhat elongated 1IGD and 1E0L. Direct computation of accessible surface areas using the SURFRACE program[53] confirms, however, that central distances are indeed correlated to actual exposure to the solvent for these four proteins. Figure 5 shows relative accessible surface area as a function of average central distance for all residues in the four proteins. Relative accessible surfaces were obtained from the division of the actual surface computed with SURFRACE by the maximal area for each residue[54] as reproduced in Table I of Ref. 55. For globular 1ENH and 1ORC (c and d), the Pearson's correlation coefficient between the two quantities are $\mathcal{C} = 0.89$ and $\mathcal{C} = 0.82$, respectively. For 1IGD (b) the correlation coefficient is somewhat lower, $\mathcal{C} = 0.72$, while for 1E0L (a), even though corresponding to least globular of the four proteins, $\mathcal{C} = 0.82$.

Finally, it must be stressed that a sufficiently accurate prediction of preferential atomic central distances from sequence alone is not expected to be trivial. In a recent study we have shown that central distance distributions of different atomic types in compact globular proteins can be described by simple mathematical expressions resembling the Fermi function of quantum statistical mechanics and depending on a small number of physically meaningful parameters.[51] Prediction of these parameters from the sequence of amino acids with an appropriate typing scheme[56] could provide a starting point for the development of burial potentials to be tested in ab initio simulations. It is instructive therefore, as an initial step in this crucial direction, to investigate the behavior of simple sequence-dependent functions, with parameters computed from the data bank of 321 structures used in ref. 51, in the discrimination of the native conformation of 1ORC from random compact decoys. For this purpose, 300 decoys were generated in long Monte Carlo simulations with an uniformly attractive contact potential at a temperature in which the average radius of gyration is similar to the native radius of gyration.

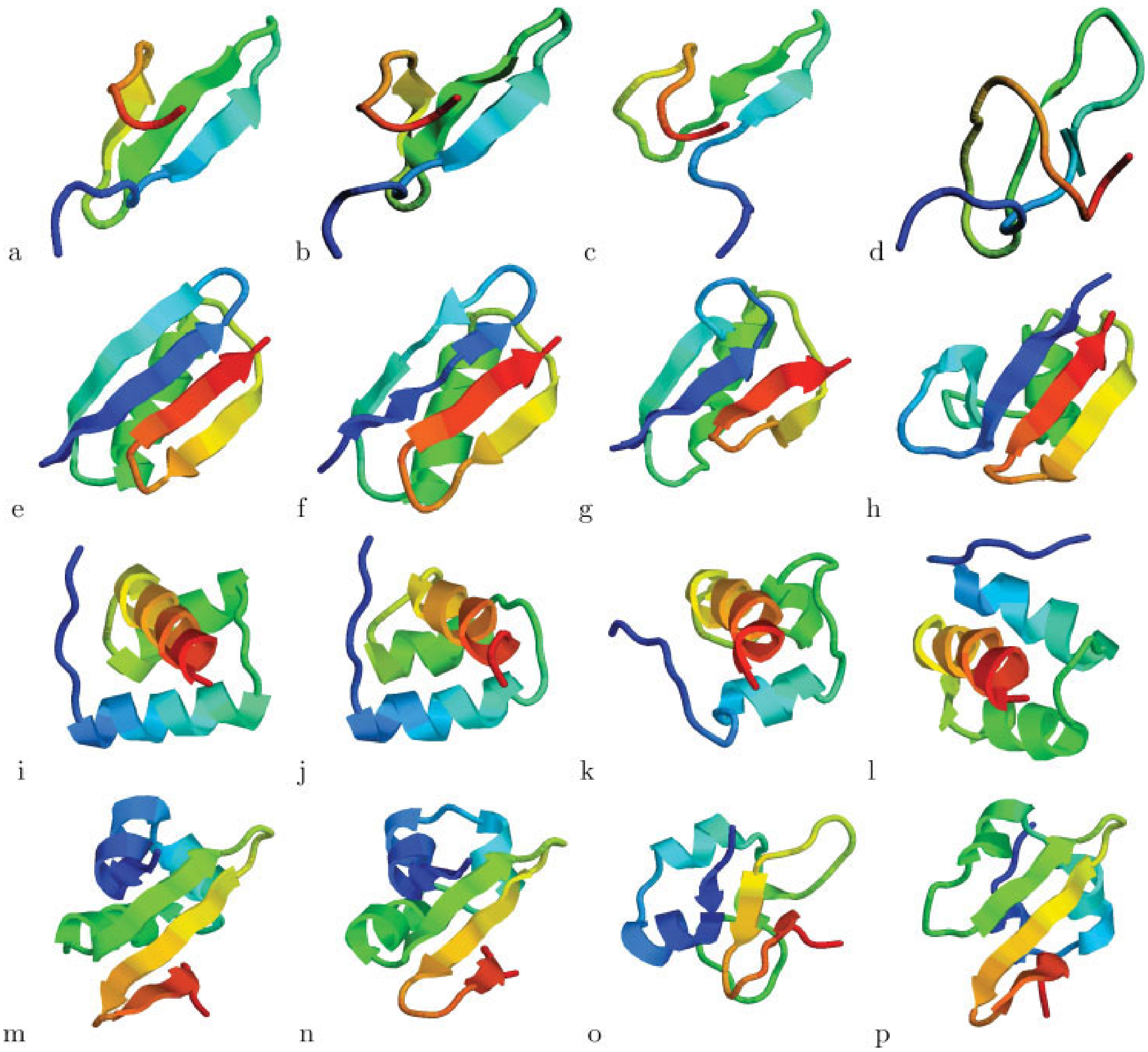

Figure 4

Native structure (first column), with representative final conformations discussed in the text for 1E0L (first row), 1IGD (second row), 1ENH (third row), and 1ORC (fourth row). For 1E0L, the best and worst final conformations with hydrogen bond constraints are shown and also a final conformation from an unconstrained trajectory. For the remaining proteins, the second column shows a final conformation labeled by "A" in Figure 3, the second column shows a conformation labeled in Figure 3 either by "B" (for 1IGD) or "C" (for 1ENH and 1ORC), while the fourth column shows a misfolded conformation labeled by "D" in Figure 3. The N-terminal region is shown in blue and the C-terminal region in red in all structures.

The sequence-dependent energy function for this last exercise is taken simply as

$$\begin{aligned} E_s(\{r_i\}) &= \sum_i E_\tau(r_i) \\ &= \sum_i \left(h_\tau r_i^{\alpha_\tau} - r_i + \Delta\mu_\tau\right), \end{aligned} \tag{4}$$

where $r_i = R_i/R_g$ is the reduced central distance of atom i, i.e. the ratio between the central distance and the radius of gyration, h_τ, α_τ, and $\Delta\mu_\tau$ are type-dependent parameters obtained from a data bank of native structures derived from PDBSELECT[57,58] from which nonglobular proteins have been excluded as described in Ref. 51, and the sum is over all atoms of the protein. We have simply used the native radius of gyration in the computation of r_i from R_i for the decoys, since they all have near-native radius of gyration. The s subscript in E_s is intended to avoid confusion between the sequence-dependent energy used in the decoy exercise, given by Eq. (4), and the native-dependent energy used in the Monte Carlo simulations, given by Eq. (3). Note that E_s is a function of adimensional reduced central distances and is itself adimensional. The form for the contribution from each

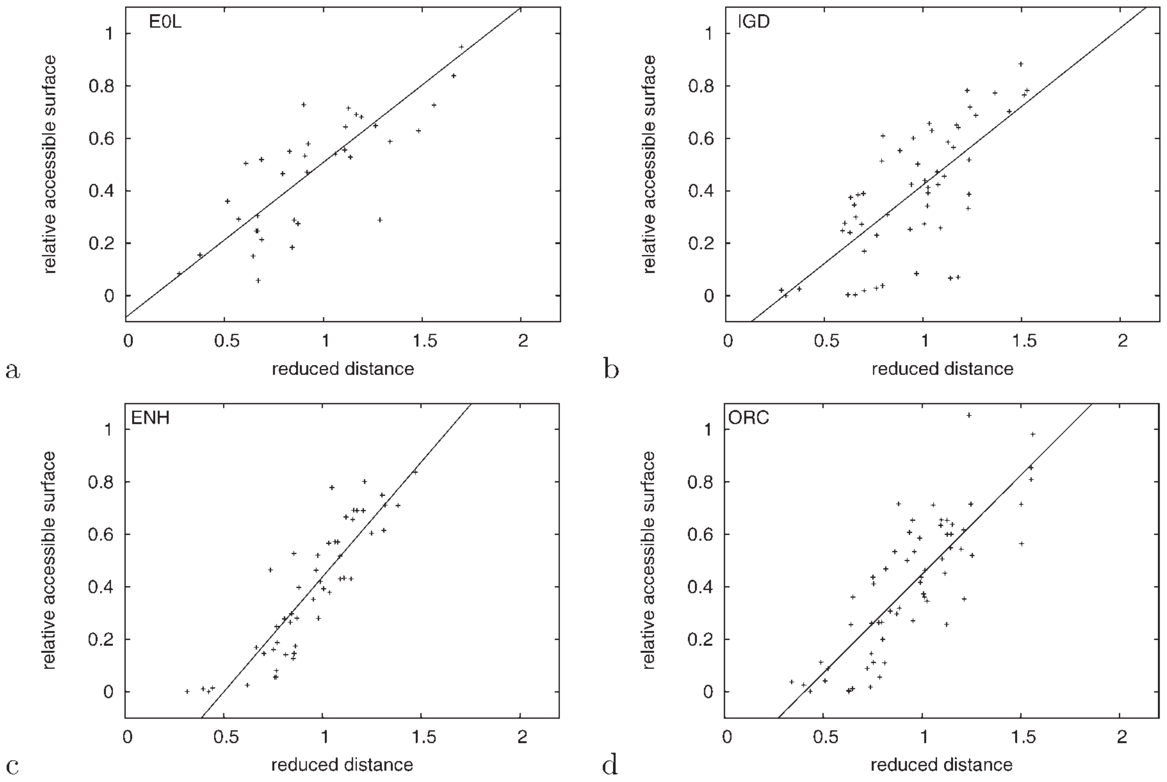

Figure 5

*Relation between fractional residue accessible surface areas and average atomic central distances for 1E0L (**a**), 1IGD (**b**), 1ENH (**c**), and 1ORC (**d**). Straight lines represent linear fits to the data.*

atom was obtained from a previous analysis which suggested that the probability density of an arbitrary atom *i* to be at a given reduced distance from the molecular geometrical center in globular proteins, independently of protein size, could be described by

$$p(r) = \frac{Ar^2 e^{-\beta(r-\mu)}}{1 + e^{-\beta(r-\mu)}}, \tag{5}$$

while the probability density for a given specific atomic type, τ, would be described by

$$p_\tau(r) = \frac{Ar^2 e^{-\beta(h_\tau r^{\alpha_\tau} - \mu_\tau)}}{1 + e^{-\beta(r-\mu)}}, \tag{6}$$

where A, β, and μ are type-independent adjustable parameters characterizing the data bank of globular structures while h_τ, α_τ, and μ_τ are additional type-specific adjustable parameters, with the product $h_\tau\alpha_\tau$ being strongly correlated to available hydrophobicity scales.[51] From the three preceding equations, with $\Delta\mu_\tau = \mu_\tau - \mu$, it follows that the ratio between $p_\tau(r)$ and $p(r)$ has the simple form of a Boltzmann factor with $E_\tau(r)$ playing the role of an effective energy of atomic type τ relative to an average type-independent atom, or

$$\frac{p_\tau(r)}{p(r)} = e^{-\beta E_\tau(r)}. \tag{7}$$

Positive and negative atomic energies, therefore, indicate, respectively, lower and higher probability than average. The adimensional "temperature" parameter β was previously found to be around 9.4.[51]

We have combined Eq. (4) to four specific typing schemes, which determine how the protein atoms are grouped into different atomic types, resulting in four sequence-dependent burial potentials: BUR-1, BUR-2, BUR-3, and BUR-4. In the first, simplest, scheme there are only two atomic types, hydrophobic and hydrophilic, whose parameters had already been computed in Ref. 51. The resulting two energy curves for BUR-1, as a function of reduced central distance, are shown in Figure 6(a). In BUR-2, atoms are grouped in 20 types corresponding to the 20 amino acid residues while in BUR-3 each specific atom of the 20 residues is considered as a different type. BUR-2 energy curves for lysine and phenylalanine are shown in Figure 6(b), labeled by "all" since they corre-

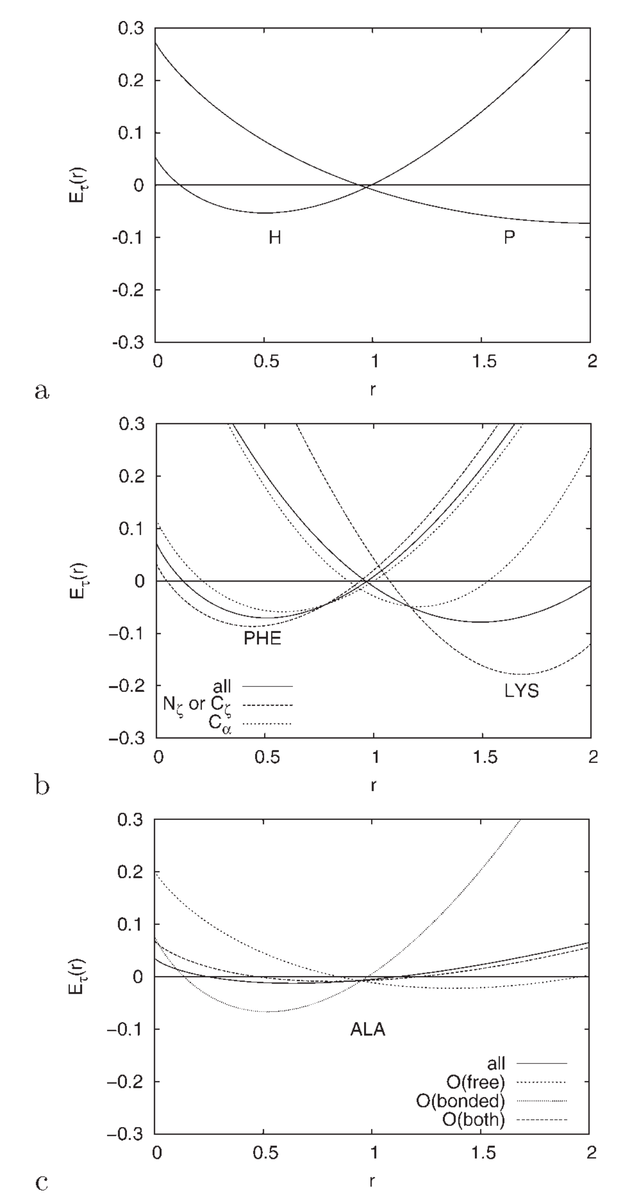

Figure 6

*Energy curves as a function of reduced central distance for representative atomic types according to different typing schemes: BUR-1 (**a**), BUR-2 and BUR-3 (**b**), BUR-2, BUR-3, and BUR-4 (**c**).*

spond to all atoms of each residue, while the curve for alanine is shown in Figure 6(c). BUR-2 curves for phenylalanine and lysine suggest that decoy discrimination should improve with the 20-letter typing scheme, since their minima at small and large distances, respectively, are more pronounced when compared to BUR-1 curves. The behavior of alanine, on the other hand, appears to be quite close to average. Differences in BUR-3 curves for individual atoms can be quite significant for large residues as also seen in Figure 6(b) for the C_α and C_ζ atoms of phenylalanine and, particularly, for the C_α and N_ζ of lysine. For small alanine, on the other hand, the BUR-3 curve for its backbone oxygen atom, labeled by "O (both)" in Figure 6(c), is similar to its BUR-2 curve. In the BUR-4 scheme, backbone oxygen and nitrogen atoms of each residue are additionally divided into two types, depending on the formation of a hydrogen bond. As also seen in Figure 6(c) the previously neutral backbone oxygen atom of alanine actually behaves as quite hydrophobic when forming a hydrogen bond and as hydrophilic when this is not the case.

According to all four sequence-dependent potentials, the native structure of 1ORC was found to be lower in energy than all compact decoys and the discrimination gradually improved, as measured by corresponding *Z*-scores of the native conformation, as the typing scheme became more detailed. $Z_N = (E_N - \bar{E})/\sigma_E$ values, or the difference between the native energy and the average energy among the decoys divided the corresponding standard deviation, were −3.0, −3.3, −3.4, and −4.1 for BUR-1, BUR-2, BUR-3, and BUR-4, in this order. Decoy energy distributions for BUR-1 and BUR-4 are shown in Figure 7(a), together with the distribution for the same decoys with the native burial potential, BUR-Nat, which has been used to successfully refold this protein. The improvement in *Z*-score from BUR-1 to BUR-4, therefore, from around −3 to around −4, might appear to be small when compared to the presently suggested goal provided by the *Z*-score of BUR-Nat, around −10. Figure 7(b) shows, however, that the goal itself, even though provided by a "native-centric" potential, is around 40 *Z*-score units closer to these simple sequence-dependent potentials than the *Z*-score around −50 computed for the same decoys with the traditional Go potential, according to which native contacts are uniformly attractive. Discrimination by the sequence-dependent burial potentials does not result from a trivial recognition of different radii of gyration, which could result, for example, from abnormally high reduced distances in the eventual case of non-compact decoys. As seen in Figure 7(c), there is no significant correlation between energy and radius of gyration and all decoys, with radii of gyration ranging from 10 to 12 Å, are distinguished from the native structure with radius of gyration around 11 Å.

DISCUSSION

Results for the WW domain 1E0L indicate that native central distances uniquely determine the tertiary structure of this small all-β protein. Conformational space was searched efficiently by the annealing protocol and all trajectories resulted in conformations with at least the correct topology. Hydrogen bond constraints, not very surprisingly, were shown to be important for the correct local orientation of backbone nitrogen and carbonyl groups in secondary structure formation. Results for the α/β protein 1IGD are also consistent with this conclu-

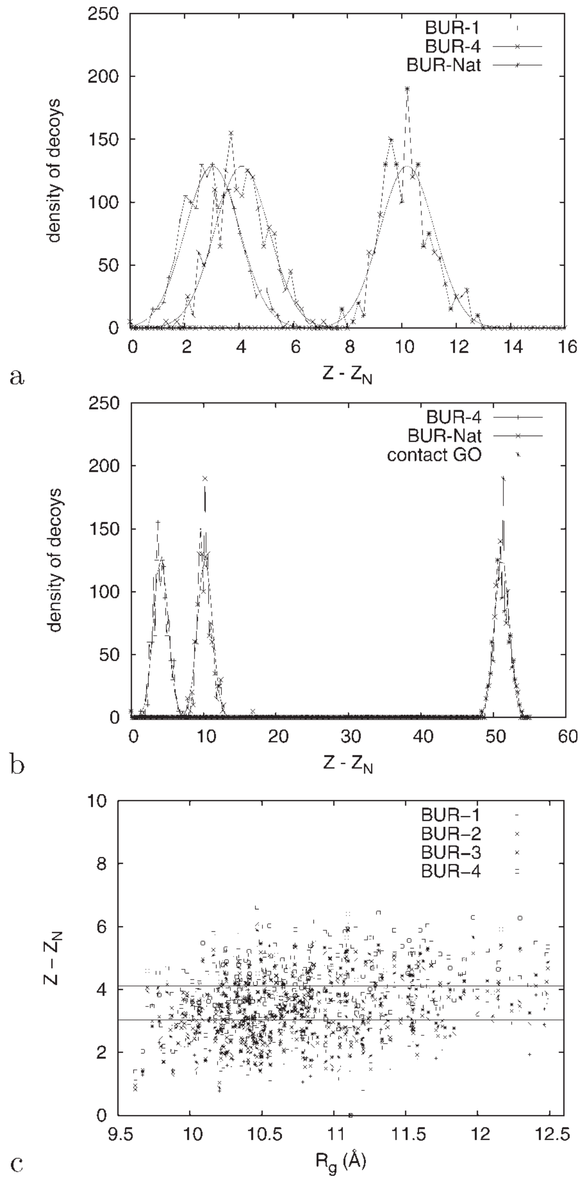

Figure 7

*Native energy discrimination from a set of compact random decoys by sequence-dependent BUR-1 (**a**) and BUR-4 (a and **b**) burial potentials and by the native-centric BUR-Nat (a and b) and Contact Go (b) potentials. Gaussian curves expected to describe the empirical frequencies are also shown in (a) and (b). The absence of significant correlation between sequence dependent energies and radii of gyration for these decoys is shown in (**c**). The point at Z − Z_N = 0 and R_g = 11.1 corresponds to the native structure. Lower and upper horizontal lines represent the average Z − Z_N value for BUR-1 and BUR-4 typing schemes, respectively.*

sion although conformational search does not appear to have been so efficient in this case, possibly because of its larger size and/or more complex topology. The energetic discrimination between folded, almost folded, and unfolded/misfolded final conformations, as indicated by the correlation between DRMS and energy [Fig. 3(b)], strongly suggests that a more efficient exploration of conformational space is likely to result in a larger number of folded conformations.

Native central distances also provided sufficient information for determination of all-α 1ENH, since lowest energy structures were essentially native while high energy structures had high DRMS. One relatively low energy conformation with the three helices well formed but arranged in a wrong topology was also observed, however [Fig. 4(l)]. It is likely therefore that a more efficient search of conformational space would result in a larger number of correctly folded conformations but it is also apparent that properly folded helices adopting non-native topologies would also appear in some trajectories. It is not clear at this point if these misfolded conformations will be always distinguishable from correctly folded structures by atomic burials alone. Careful observation of low energy non-native conformations might reveal possible approaches for distinguishing them. It is relevant, however, that misfolded helical conformations are not an exclusive problem of the present energy function and that topological "mirror images" have been observed in recent ab initio simulations with contact pairwise interactions governed by the μ-potential.[25]

This possible qualitative difference between more easily discriminated β-sheets and mirror-prone α-helical domains is nicely illustrated by the two regions of the α + β 1ORC. While the last three adjacent strands of the β-sheet, comprising approximately the second half of its sequence, folded correctly in 7 out of 10 trajectories, the three α-helices tended to adopt an incorrect topology, reminiscent of the misfolded structure observed for 1ENH. Since in this case we are considering two regions of a single, globular, structure it is apparent that the difference observed in the behavior of α and β proteins do indeed result from intrinsic properties of these two types of secondary structure and not from an eventual insufficient globularity of 1E0L and 1IGD. Despite this possible difference in behavior between secondary structures, the final conformation with lowest energy among independent folding trajectories did indeed correspond, for all four proteins under consideration, to an essentially native structure both in terms of global topology and secondary structure formation. Results from these simulations strongly suggest, therefore, that native central distances contain sufficient information to determine the structure of small proteins, with β-sheets possibly being more easily discriminated than α-helical regions.

The present approach is also quite original with respect to the effect of hydrogen bonds, whose definition is completely independent of structural class and/or eventual type of secondary structure being formed. Instead of simply decreasing conformational energy, hydrogen bond formation increases the burial propensity of initially hydrophilic atoms, as has been very recently investigated in the context of folding cooperativity in lattice models.[38] This dependence of hydrophobicity on local environment provides a simple mechanism for taking important multi-body effects into consideration. In particular, the effective energetic contribution of a hydrogen bond depends on its degree of exposure, which is physically reasonable. Note, however, that amino acid residues, with the notable exception of proline, are not expected to behave significantly different from each other in terms of backbone hydrogen bond formation. Not much information about these interactions, therefore, is encoded in the amino acid sequence. Backbone hydrogen bonds can actually be thought as very important sequence-independent constraints that can drastically reduce the number of relevant collapsed conformations,[59] just like the specific geometry of peptide bonds drastically restricts the range of backbone local structure.

From the observed correlation between central distances and accessible surface areas shown in Figure 5, it is apparent that much of the information provided by central distances, particularly for the most globular structures, is indeed shared by actual burials and could hopefully be at least partially predictable from the sequence through parameters reflecting burial propensities or hydrophobicities. It remains to be investigated, therefore, how accurate must be the information about atomic burials in order to be sufficient for native structure prediction and how close to this required limit are the available approaches to burial prediction from sequence. Inaccuracies in atomic burial predictions are unavoidable and they will gradually deform the original perfectly accurate effective energy surface into a random surface unable to distinguish the native structure from random conformations.[60–63] Results from the exercise with 1ORC decoys shown in Figure 7 demonstrate that even very simple sequence-dependent burial potentials are able to distinguish the native structure from random compact decoys and that discriminability improves for more detailed typing schemes. More importantly, however, they provide an upper estimate for the required discriminability for ab initio simulations drastically below the discriminability of traditional native-centric Go potentials, although still above the discriminability of the investigated sequence-dependent potentials. It is possible that the upper estimate can be decreased further, through simulations with native central distances combined to gradual noise addition, for example. Improvement in sequence-dependent potentials, which will almost certainly be required in order to provide an appropriate discriminability, might be attempted by even more detailed typing schemes. Consideration of the local sequence environment around each residue, for example, is likely to produce additional informative atomic types.

CONCLUSION

The present study demonstrates that a potential based on pairwise interactions is not the simplest functional form capable of folding small globular proteins. Correctly folded structures can also result from a sufficiently accurate burial potential, reflecting the tendency of different atoms to be buried or exposed to the solvent or, in globular proteins, to be near or far from the molecular geometrical center. A possible solution for the protein folding problem might arise, therefore, from sufficiently accurate burial predictions from sequence followed by computer simulation of the folding process with a burial-dependent potential.

REFERENCES

1. Shakhnovich EI, Gutin AM. Engineering of stable and fast-folding sequences of model proteins. Proc Natl Acad Sci USA 1993;90:7195–7199.
2. Abkevich VI, Gutin AM, Shakhnovich EI. Specific nucleus as the transition state for protein folding: Evidence from the lattice model. Biochemistry 1994;33:10026–10036.
3. Leopold PE, Montal M, Onuchic JN. Protein folding funnels: a kinetic approach to the sequence-structure relationship. Proc Natl Acad Sci USA 1992;89:8721–8725.
4. Bryngelson JD, Onuchic J, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins: Struct Funct Genet 1995;21:167–195.
5. Dill KA, Chan HS. From levinthal to pathways to funnels. Nat Struct Biol 1997;4:10–19.
6. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. Optimal protein-folding codes from spin-glass theory. Proc Natl Acad Sci USA 1992;89:4918–4922.
7. Shakhnovich EI. Theoretical studies of protein folding thermodynamics and kinetics. Curr Opin Struct Biol 1997;7:29–40.
8. Mirny L, Shakhnovich E. Protein folding theory: from lattice to all-atom models. Annu Rev Biophys Biomol Struct 2001;30:361–396.
9. Pande VS, Grosberg A, Tanaka T. Statistical mechanics of simple models of protein folding and design. Biophys J 1997;73:3192–3210.
10. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem 1997;48:545–600.
11. Onuchic JN, Wolynes PG. Theory of protein folding. Curr Opin Struct Biol 2004;14:70–75.
12. Dill KA, Bronberg S, Yue K, Fiebig KM, Yee D, Thomas PD, Chan HS. Principles of protein folding—a perspective from simple exact models. Prot Sci 1995;4:561–602.
13. Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 1976;104:59–107.
14. Oldziej S, Czaplewski C, Liwo A, Chinchio M, Nanias M, Vila JA, Khalili M, Aranautova YA, Jagielska A, Makowski M, Schafroth HD, Kaźmierkiewicz R, Ripoll DR, Pillardy J, Saunders JA, Kang YK, Gibson KD, Scheraga HA. Physics-based protein prediction using a hierarchical protocol bases on the unres force field: assessment in two blind tests. Proc Natl Acad Sci USA 2005;102:7547–7552.
15. Liwo A, Khalili M, Scheraga HA. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. Proc Natl Acad Sci USA 2005;102:2362–2367.
16. Rohl C, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004;383:66–93.
17. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 2005;15:285–289.
18. Aloy P, Stark A, Hadley C, Russell RB. Predictions without templates: new folds, secondary structure, and contacts in CASP5. Proteins: Struct, Funct Bioinf 2003;53:436–456.

19. Go N, Taketomi H. Respective roles of short- and long-range interactions in protein folding. Proc Natl Acad Sci USA 1978;75:559–563.
20. Shimada J, Kussell E, Shakhnovich EI. The folding thermodynamics and kinetics of crambin using an all-atom monte carlo simulation. J Mol Biol 2001;308:79–95.
21. Kussell E, Shimada J, Shakhnovich EI. Excluded volume in protein side-chain packing. J Mol Biol 2001;311:183–193.
22. Shimada J, Shakhnovich EI. The ensemble folding kinetics of protein g from an all-atom monte carlo simulation. Proc Natl Acad Sci USA 2002;99:11175–11180.
23. Hubner IA, Shimada J, Shakhnovich EI. Commitment and nucleation in the protein G transition state. J Mol Biol 2002;336:745–761.
24. 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.
25. Hubner IA, Deeds EJ, Shakhnovich EI. High-resolution protein folding with a transferable potential. Proc Natl Acad Sci USA 2005; 102:18914–18919.
26. Clementi C, Nymeyer H, Onuchic JN. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 2000;298:937–953.
27. Matinez JC, Serrano L. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. Nat Struct Biol 1999;6:1010–1016.
28. Riddle DS, Grantcharova VP, Santiago JV, Alm E, Ruczinski I, Baker D. Experiment and theory highlight role of native state topology in SH3 folding. Nat Struct Biol 1999;6:1016–1024.
29. Chiti F, Taddei N, White PM, Bucciantini M, Magherini F, Stefani M, Dobson CM. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. Nat Struct Biol 1999;6:1005–1009.
30. Koga N, Tanaka S. Roles of native topology and chain-length scaling in protein folding: a simulation with a go-like model. J Mol Biol 2001;313:171–180.
31. Garcia LG, Pereira de Araújo AF. Folding pathway dependence on energetic frustration and interaction heterogeneity for a three dimensional hydrophobic protein model. Proteins: Struct, Funct Bioinf 2006;62:46–63.
32. Pierce JR. An introduction to information theory—symbols, signals and noise. New York: Dover Publications; 1980.
33. Reza FM. An introduction to information theory. New York: Dover Publications; 1994.
34. Pereira de Araújo AF. Folding protein models with a simple hydrophobic energy function: the fundamental importance of monomer inside/outside segregation. Proc Natl Acad Sci USA 1999;96:12482–12487.
35. Pereira de Araújo AF. Sequence rotation in *N*-dimensional space and the folding of hydrophobic protein models: surpassing the diagonal unfolded state approximation. J Chem Phys 2001;114:570–578.
36. Treptow WL, Barbosa MAA, Garcia LG, Pereira de Araújo AF. Non-native interactions, effective contact order and protein folding: a mutational investigation with the hydrophobic model. Proteins: Struct, Funct Genet 2002;49:167–180.
37. Barbosa MAA, Pereira de Araújo AF. Relevance of structural segregation and chain compaction for the thermodynamics of folding of a hydrophobic exact protein model. Phys Rev E 2003;67:051919.
38. Barbosa MAA, Garcia LG, Pereira de Araújo AF. Entropy reduction effect imposed by hydrogen bond formation on protein folding cooperativity: evidence from a hydrophobic minimalist model. Phys Rev E 2005;72:051903.
39. Garcia LG, Treptow WL, Pereira de Araújo AF. Folding simulations of a three-dimensional protein model with a non-specific hydrophobic energy function. Phys Rev E 2001;64:011912.
40. Skorobogatiy M, Guo H, Zuckermann MJ. A deterministic approach to the protein design problem. Macromolecules 1997;30: 3403–3410.
41. Kabakçioĝlu A, Kanter I, Vendruscolo M, Domany E. Statistical properties of contact vectors. Phys Rev E 2002;65:041904.
42. Dill KA. Dominant forces in protein folding. Biochemistry 1990;29: 7133–7155.
43. Dyson HJ, Wright PE, Scheraga HA. The hydrophobic interactions in initiation and propagation of protein folding. Proc Natl Acad Sci USA 2006;103:13057–13061.
44. Metropolis N, Rosembluth A, Rosembluth M, Teller A. Equation of state calculations by fast computing machines. J Chem Phys 1953; 21:1087–1092.
45. Macias M, Gervais V, Civera C, Oschkinat H. Structural analysis of ww domains and design of a ww prototype. Nat Struct Biol 2000;7:375–379.
46. Derrick D, Wigley JP. The third igg-binding domain from streptococcal protein g. An analysis by X-ray crystallography of the structure alone and in a complex with fab. J Mol Biol 1994;243:906–918.
47. Clarke N, Kissinger C, Desjarlais J, Gilliland G, Pabo C. Structural studies of the engrailed homeodomain. Protein Sci 1994;3:1779–1787.
48. Albright R, Mossing M, Matthews B. High-resolution structure of an engineered cro monomer shows changes in conformation relative to the native dimer. Biochemistry 1996;35:735–742.
49. Baker EN, Hubbard RE. Hydrogen bonding in globular proteins. Prog Biophys Mol Biol 1984;44:97–179.
50. Richardson JS, Richardson DC. Principles and patterns of protein conformation. In: Fasman GD (editor). Prediction of protein structure and the principles of protein conformation. Plenum Press; 1989. Chapter 1, pp 1–98.
51. Gomes ALC, de Rezende JR, Pereira de Araújo AF, Shakhnovich EI. Description of atomic burials in compact globular proteins by Fermi-Dirac probability distributions. Proteins: Struct, Funct Bioinf 2007;66:304–320.
52. Baumgärtner A. Shapes of flexible vesicles at constant volume. J Chem Phys 1993;98:7496–7501.
53. Tsodikov OV, Record MT, Sergeev YV. Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. J Comput Chem 2002;23:600–609.
54. Rose GD, Geselowitz AE, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acids in globular proteins. Science 1985;229: 834–838.
55. Rose GD, Dworkin JE. The hydrophobicity profile. In: Fasman GD (editor). Prediction of protein structure and the principles of protein conformation. Plenum Press; 1989. Chapter 15, pp 625–633.
56. Chen WW, Shakhnovich EI. Lessons from the design of a novel atomic potential for protein folding. Protein Sci 2005;14:1741–1752.
57. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. Protein Sci 1992;1:409–417.
58. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.
59. Hoang TX, Trovato A, Seno F, Banavar JR, Maritan A. Geometry and symmetry presculpt the free-energy landscape of proteins. Proc Natl Acad Sci USA 2004;101:7960–7964.
60. Pereira de Araújo AF, Pochapsky TC. Monte Carlo simulations of protein folding using inexact potentials: how accurate must parameters be in order to preserve the essential features of the energy landscape? Folding Design 1996;1:299–314.
61. Pereira de Araújo AF, Pochapsky TC. Estimates for the potential accuracy required in realistic protein folding simulations and structure recognition experiments. Folding Design 1997;2:135–139.
62. Bryngelson JD. When is a potential accurate enough for structure prediction? Theory and application to a random heteropolymer model of protein folding. J Chem Phys 1994;100:6038–6045.
63. Pande VS, Grosberg A, Tanaka T. How accurate must potentials be for successful modeling of protein folding? J Chem Phys 1995;103: 9482–9491.