Exploiting Homology Information in Nontemplate Based Prediction of Protein Structures

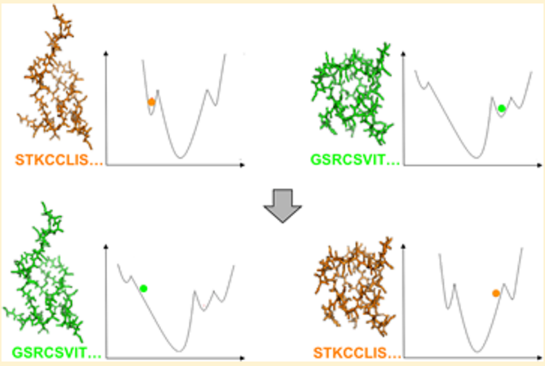
Alfredo Iacoangeli,[‡,§] Paolo Marcatili,[*,‡,†,§] and Anna Tramontano[*,§,∥]

[§]Department of Physics, Sapienza University of Rome, P.le A. Moro 4, 00185 Rome, Italy

[∥]Istituto Pasteur Fondazione Cenci Bolognetti, Sapienza University of Rome, P.le A. Moro 4, 00185 Rome, Italy

Supporting Information

ABSTRACT: In this paper we describe a novel strategy for exploring the conformational space of proteins and show that this leads to better models for proteins the structure of which is not amenable to template based methods. Our strategy is based on the assumption that the energy global minimum of homologous proteins must correspond to similar conformations, while the precise profiles of their energy landscape, and consequently the positions of the local minima, are likely to be different. In line with this hypothesis, we apply a replica exchange Monte Carlo simulation protocol that, rather than using different parameters for each parallel simulation, uses the sequences of homologous proteins. We show that our results are competitive with respect to alternative methods, including those producing the best model for each of the analyzed targets in the CASP10 (10th Critical Assessment of techniques for protein Structure Prediction) experiment free modeling category.



1. INTRODUCTION

Challenges in protein structure prediction can be roughly divided in two categories. When the target protein is homologous to a protein of known structure, template-based methods[1,2] can be applied. In these cases, the major challenges are the inference of the conformation of structurally divergent regions and of the flexible ones. Neither problem is minor as these regions are often related to the specificity, and thereby to the function, of the target protein.[3]

When a homologous protein of known structure is not available, the problem becomes more complex. Here the issues to be faced are the identification of an effective energy function, be it of physical, statistical, or heuristic nature, which can accurately describe the conformational space of the protein and of a method able to effectively explore the energy landscape to detect the conformation corresponding to the minimum representing the protein native structure.

Despite this, with the progress of structure resolution methods and the consequent enrichment of our data set of proteins of known structure, cases where template based methods cannot be applied become less and less frequent, and there are good reasons why these remain interesting. First is because from an intellectual point of view the understanding of the folding mechanism is an important goal; second is because often these proteins are the most challenging for the experimental approaches and therefore computational methods are called upon to help. Furthermore, template-free methods can be combined with template-based ones to address the problem of structurally divergent regions.

Template-based methods base their efficacy on a simple and almost universal rule: protein structure is more evolutionary conserved than protein sequence.[4,5] Homologous proteins, even if sharing very little sequence similarity, have similar native folds.[6] The question is whether this information can be exploited to improve the computational methods for non-template based protein structure prediction. We show here that this is indeed the case.

Our strategy consists in using a homology based Monte Carlo Replica Exchange method.[7] Monte Carlo Methods are very commonly used for exploring the conformational space described by an energy function[8−10] in order to find its global minimum. Many strategies have been developed to avoid local minima during the search. Among them the Replica Exchange Methods (REMs) seem to produce the most encouraging results.[11−14] In a REM, many folding simulations are run in parallel. Each of these simulations is called a replica and can be completely described at each step of the simulation by its current conformation, temperature, and energy function. At predetermined intervals during the simulations, the different replicas are compared and some of their characteristics are exchanged according to a given criterion (e.g., the Metropolis Criterion[15]). In this way, they are usually able to explore a larger number of conformations and possibly avoid being trapped in local minima.

In our work, we use a Monte Carlo REM, but with a relevant difference: we assign to the different replicas the sequences of

Received: April 20, 2015

proteins homologous to the target protein and let them exchange their sequence during the simulation. This is based on the rationale, schematically represented in Figure 1, that the energy global minimum of homologous proteins should correspond to similar conformations since they have similar native structures, but the shape of the energy landscape, and consequently the positions of the local minima, is likely to be different for each replica.

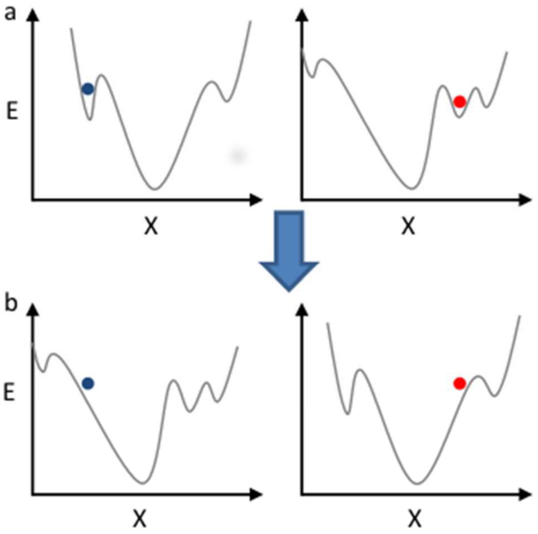


Figure 1. Schematic representation of the rationale of the *hom*REM method described here. The curves indicate an idealized 2-D energy landscape for two homologous proteins before (a) and after (b) a replica exchange move. While the local features and the location of the local minima of the two proteins might be different, their global minimum is expected to be located in a similar position in configuration space and consequently the exchange can help in escaping from local minima.

We compare our results with those of the ab initio simulation protocol implemented in the Rosetta low resolution protocol[16] and of the Hamiltonian Replica Exchange method described by Shmygelska and Levitt.[11] We will refer to these protocols as ROSETTA and HREM, respectively.

We focused here on the low-resolution mode of Rosetta, since its ability to produce native like conformations has been shown to be a major determinant of the overall accuracy of the method.[17]

As reported in Results, our method, named *hom*REM, performs better than both ROSETTA and HREM, and its performance is competitive with those obtained by groups participating to the CASP10[18] free modeling category.

We also compare the results of *hom*REM with those obtained by Bonneau et al.[19] where the authors also run the Rosetta folding simulations on homologous sequences but without applying the replica exchange protocol and select the final models using a clustering approach. We used the same target structures as in Bonneau et al.[19] trying also to reproduce the databases available at the time of their experiment and we obtained better models in the majority of the cases. This suggests that the improvements we obtain are due to the use both of homologue information and of the replica exchange approach. This is confirmed by an additional test, in which a modified version of our protocol (named *hom*MC) is run without using the replica-exchange approach. The results are subpar with respect to the *hom*REM ones. The detailed description of these tests can be found in the Supporting Information (Tables S1 and S3a,b and Figures S1 and S2).

2. RESULTS

In our Homologous Replica Exchange Method (*hom*REM), different sequences, derived from homologues of the target protein (see Materials and Methods for details), are assigned to different replicas. We simulate 100 replicas at the time, 30% of which have the original target sequence and 70% the sequences of homologous proteins selected as described in Materials and Methods.

Similarly to the ROSETTA low-resolution protocol, each amino acid side chain is described by its centroid, i.e., replaced by a single pseudoatom located at the center of mass of the side-chain atoms and with a radius proportional to the size of the side chain.

The energy is defined as

$$U_{tot} = U_{const} + \sum_i \lambda_i U_i \tag{1}$$

with $i = 1, ..., 8$ where U_{const} includes terms, such as the van der Waals hard sphere repulsion, the helix−helix pairing, contact order, and the Ramachandran torsion angle energy contributions, which do not vary across the energy functions. The U_i terms are the Rosetta energy terms with different weights: λ_{env} (environment), λ_{pair} pair potential, λ_{cb} (packing density), λ_{hs} (helix-strand pairing), λ_{ss} (strand−strand pairing), λ_{rsigma} (strand pair distance/register), λ_{sheet} (strand arrangement into sheets), and λ_{rg} (radius of gyration) described in Rohl et al.[16] During the different stages of the protocol, these terms are weighted differently following the same strategy used in ROSETTA and outlined in Materials and Methods.

The simulations follow the standard Rosetta protocol: we start with the target proteins in an extended conformation. First, we perform 50 of the so-called "fragment insertion" stages (each consisting of 2000 Monte Carlo steps) to produce a set of final low-resolution models. During each fragment insertion stage, the protein conformation is modified by iteratively replacing small portions of its structure (3 and 9 amino acids long) with fragments of similar length randomly extracted from a library derived from solved structures (see Materials and Methods). The new conformation is then retained or discarded according to the Metropolis criterion based on the energies calculated using a low-resolution Rosetta energy function that, as in the Rosetta protocol, depends on the specific "fragment insertion" stage:[16]

$$P_{accept} = \begin{cases} 1 & \text{if} \quad \Delta U \leq 0 \\ e^{-\Delta U/K_B T} & \text{if} \quad \Delta U > 0 \end{cases} \tag{2}$$

where P_{accept} is the probability of accepting the exchange, ΔU the system energy difference between the conformation before and after the move, T the temperature, and K_B the Boltzmann constant.

After each of the 50 fragment insertion stages, we perform a replica exchange: we order the N replicas according to their energy; since the rate of acceptance at this stage is expected to be rather low given the way the homologous sequences are selected (see Materials and Methods), we attempt exchanges also between replicas whose energy is not necessarily very close in order to achieve an acceptance rate similar to what was suggested in the literature.[20] Indeed, exchanges are attempted between the *i*th replica and the *j*th one for $j = i + 1, ..., N$ until

DOI: 10.1021/acs.jctc.5b00371

an exchange is accepted or $j = N$. For each attempt, keeping the backbone fixed, we mutate each residue of the two proteins to the corresponding residue (as derived from the multiple sequence alignment described in Materials and Methods) of the other protein while keeping the centroid position unaltered using the Rosetta "low-resolution" side-chain Packing mover.[21] Again, the exchange is accepted or rejected according to the Metropolis criterion.

After the 50 fragment insertion stages, we rank the final structures according to the so-called score4 Rosetta energy term (see Materials and Methods for details).

We tested the ROSETTA, HREM, and *hom*REM protocols on the 40 domain data set used in a previous work by Shmygelska and Levitt[11] (SCOP data set) and described in Materials and Methods. All simulations were ran using KT = 2.0. In all cases, we performed 100 000 Monte Carlo fragment insertions. We also applied our method to a subset of the CASP10 targets (see Materials and Methods) and compared our predictions with the best models submitted to the experiment for each target.

2.1. Comparison of *hom*REM with ROSETTA and HREM.

Figure 2 summarizes the results of the three methods (ROSETTA, HREM, and *hom*REM) on the SCOP data set. We evaluate the results using the same criteria adopted in the CASP experiment,[22] i.e., by analyzing, for each protein and each method, the model with the lowest energy (MinEn), the model with the best GDT_TS (see Materials and Methods) among the 5 models with the lowest energy (Min5), and the average GDT_TS of the 1%, 5%, and 10% of the models with the lowest energy (Mean1%, Mean5%, and Mean10%).

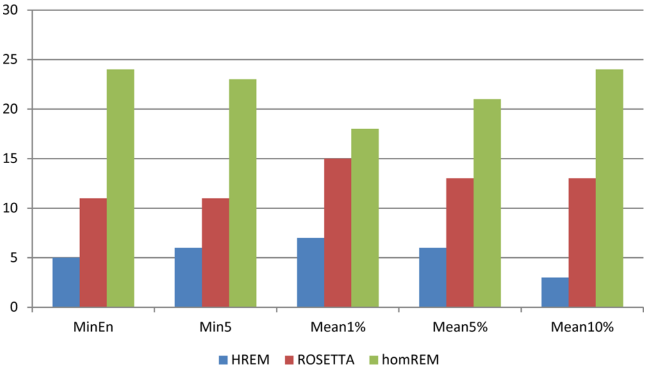


Figure 2. Comparison of the results of *hom*REM, Rosetta, and HREM on the SCOP data set. The histogram shows the number of times each of the tested methods obtains a result better than any of the other two. Data are shown for the model with the reported lowest energy (MinEn), for the best model (in terms of GDT_TS) among the five models with the lower energy (Min5), for the mean GDT_TS of the 1%, 5%, and 10% of the produced models with lower energy (Mean1%, Mean5%, and Mean10%, respectively).

As it can be seen from Figure 2, the *hom*REM method produces the best models in the majority of cases. Moreover, ROSETTA provides more accurate structures, according to our measures, than HREM; therefore, in the following we only describe the comparison of the results of *hom*REM with those of ROSETTA.

Figure 3 shows the GDT_TS difference between the models produced by ROSETTA and *hom*REM across the whole SCOP data set. The MinEn, Min5, and Mean10% values are used to evaluate the performance. *Hom*REM produces better results in 65% of cases (positive bars in the plot) with an average difference in GDT_TS of 2.11, 1.13, and 0.35 for MinEn, Min5, and Mean10%, respectively. Considering only the cases in which *hom*REM performs better, the average GDT_TS differences for the MinEn, Min5, and Mean10% sets are 4.46, 3.85, and 1.07, respectively. The complete table with all the results from this test can be found in the Supporting Information (Table S3a,b).

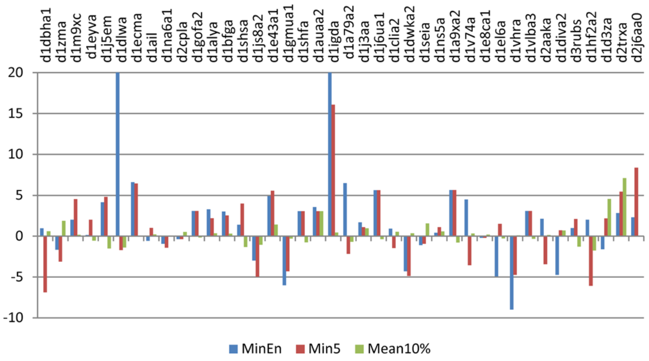


Figure 3. GDT_TS difference between the models produced by *hom*Rem and Rosetta for the SCOP data set. Data are shown for the model with the reported lowest energy (MinEn), for the best model (in terms of GDT_TS) among the five models with the lowest energy (Min5), and for the mean GDT_TS of the 50 lowest energy models, corresponding to 10% of the produced models (Mean10%). The scop identifier of each protein is shown in the upper part of the figure. Positive values indicate cases where *hom*REM produces a better model than Rosetta.

The better performance of *hom*REM seems to be mainly due to its ability to more efficiently explore the conformational space of the target protein. The plot in Figure 4 supports this hypothesis. It shows a typical energy landscape sampled from the extended conformation by ROSETTA, *hom*REM, and HREM. We plot the model energy as a function of its C_αRMSD (root mean square deviation) from the native structure for the 500 simulated structures of the first domain of UDP-N-acetylmuramate ligase Murc from *Thermotoga maritima* (scop Id: d1j6ua1, PDB code 1J6U residues 0 to 88). As it can be seen from the figure, the *hom*REM simulations explore more efficiently the lower left part of the plot, which corresponds to cases where a lower computed energy corresponds to a structure closer to the native one.

2.2. Results of *hom*REM on the CASP10 Target Set.

We also applied the *hom*REM method to a data set including the CASP10 Free Modeling targets and compared the results with all the models submitted to the experiment available on the CASP10 Web site.[22] For this test we modified our *hom*REM to include a simulated annealing temperature schedule[23] decaying exponentially from KT = 300 to KT = 0.28. For comparison, we also ran the simulations at constant temperature (KT = 2.0, see Supporting Information) and performed 5000 simulations per target.

For the models of this data set, we computed the GDT_TS for the modeled structures (see Materials and Methods).

It is not trivial to fully reproduce the official ranking for the CASP10 free modeling challenge, since the assessment included some structure visual inspection steps[24] particularly relevant for difficult cases where only low quality models were submitted. To compare our results with those submitted to CASP10 in an unbiased way, we can only use numerical indicators such as the GDT_TS values available on the CASP10 result web page. For

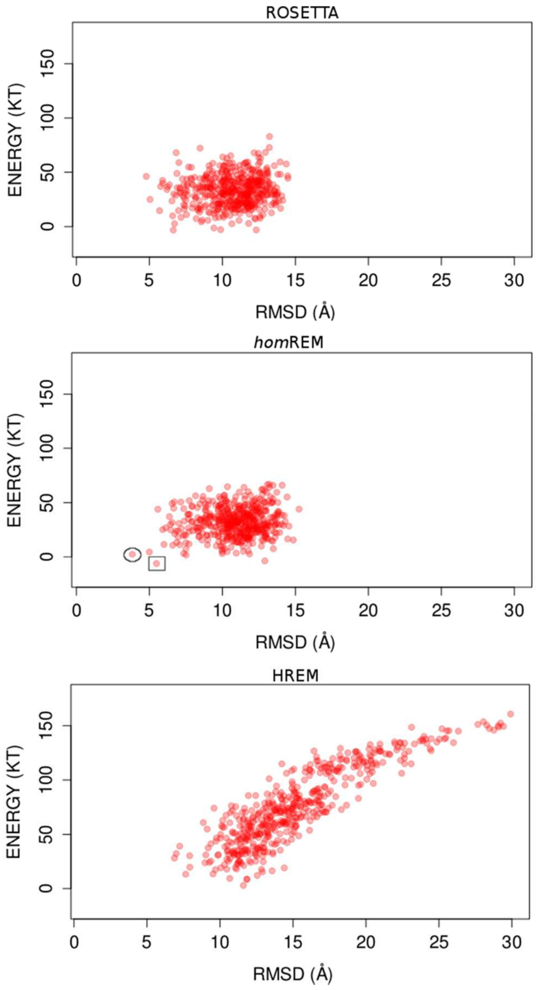


Figure 4. Scatter plot of the score4 energy term and the C_αRMSD of the models produced during the simulation for the d1j6ua1 (UDP-N-acetylmuramate ligase Murc from *Thermotoga maritima*) protein by ROSETTA, *hom*REM, and HREM. In an ideal case the model with the lowest energy (indicated by a square in the *hom*REM plot) should coincide with that with the lowest C_αRMSD from the native structure (circled in the same plot). As it can be seen, the *hom*REM simulation efficiently explores the low C_αRMSD region and its lowest energy model is close to the experimental structure.

this reason, we focus here only on the nine targets for which predicted structures with a GDT_TS higher than 30 were submitted by at least one group.

It is worth mentioning that, in the CASP experiment, neither the domain boundaries of the targets nor the best modeling strategy—either template based or template free—are known to the participants. These differences, which can certainly affect the accuracy of the results, are not taken into account in this paper. Nevertheless, by limiting our comparison only to targets for which good models have been submitted and to the performances of the best groups, we expect these two factors to have little or no influence on the results presented here.

In Table 1 we show the GDT_TS values obtained for these targets using *hom*REM and compare them with those obtained by the best performing group for each domain in the CASP experiment and by the Rosetta Server, for the model with the lowest energy (MinE) and, consistently with what is done in CASP in the free modeling category, for the best among the five models with the lowest energy (Min5). Considering that our *hom*REM uses the Rosetta energy functions and a similar overall prediction algorithm, comparing our results with the Rosetta Server ones is clearly of interest.

The complete table with the results on all the CASP10 targets is available in Supporting Information Table S4.

The performance of the CASP10 participants was very heterogeneous; indeed only a few groups out of 147 were able to submit good models for more than one or two targets. Our *hom*REM compares favorably with the best performing groups for at least five targets. Figure 5 shows the computed energy as a function of its GDT_TS from the native structure for the 5000 simulated structures of the targets for which we achieved the best results: T0693d1, T0735d2, T0739d1, and T0739d2. In particular, for T0739d1, we were able to predict and select a structure with a GDT_TS of 54.11, which is significantly more accurate than all other models submitted to CASP10 (the best model for this target had a GDT_TS of 35.88 and was produced by group 335). As an example, Figure 6 shows the best prediction submitted to CASP10 for T0739d1 together with its native structure and our best prediction. It can be seen that our model has a topology of the secondary structure elements that more closely resembles the native structure.

Table 1. Results of *hom*REM and Rosetta on the CASP Dataset[a]

	*hom*REM		Rosetta		CASP	
target	MinEn	Min5	MinEn	Min5	MinEn (group ID)	Min5 (group ID)
T0693d1	***36.00***	**36.00**	23.00	26.25	35.50 (315)	36.75 (237)
T0735d2	***42.04***	***42.04***	35.80	39.49	39.49 (315)	41.76 (237)
T0739d1	***31.76***	***54.11***	23.82	31.18	32.06 (294)	35.88 (335)
T0739d2	**32.06**	***38.62***	24.14	30.17	34.48 (315)	38.36 (237)
T0666d1	**23.33**	**24.88**	21.94	24.03	33.75 (475)	33.75 (475)
T0756d2	***40.46***	**42.32**	39.53	42.15	40.70 (081)	43.90 (114)
T0726d3	**26.44**	**32.88**	23.73	29.66	35.17 (172)	36.02 (172)
T0737d1	**27.69**	27.69	23.50	32.48	35.26 (335)	40.60 (045)
T0740d1	**20.90**	**23.22**	20.64	22.74	30.16 (350)	38.87 (358)

[a] The last two columns show the GDT_TS for the best models submitted across all the participating groups. Cases where *hom*REM outperforms Rosetta are displayed in bold and underlined if the difference is larger than 5 GDT_TS units, while a gray background is used to indicate cases where the *hom*REM method performs similarly or better than the best model submitted to CASP for the specific target. Notice that the best model submitted to CASP10 is not necessarily produced by the same group(s).

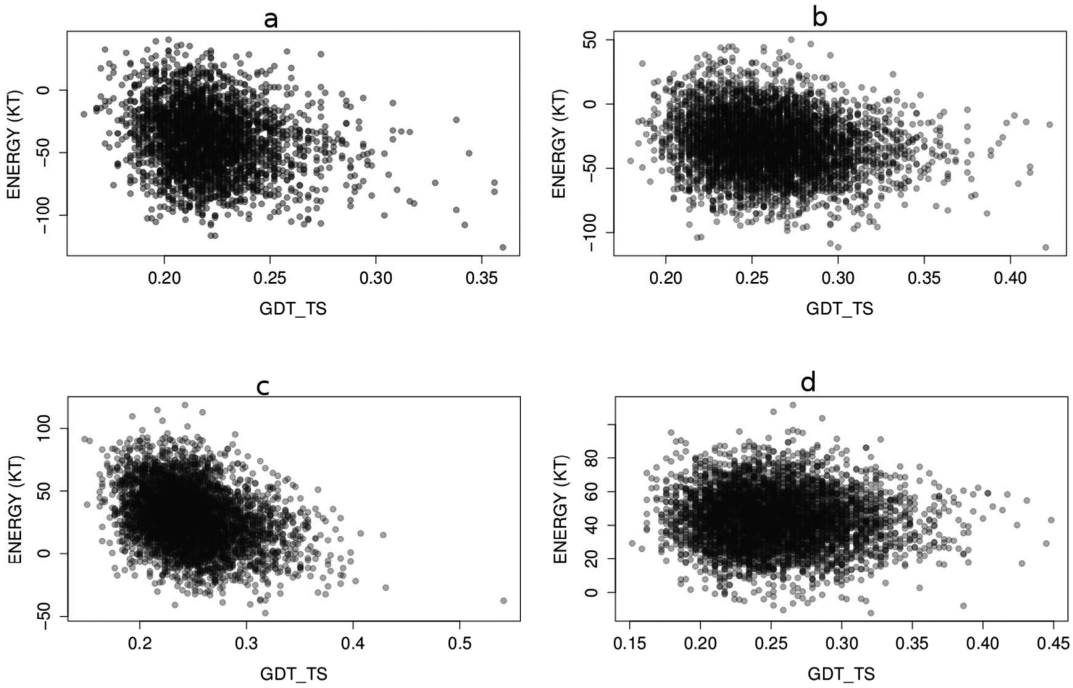


Figure 5. Plot of the computed score4 as a function of the GDT_TS values for the models produced by *hom*REM for some of the CASP targets. (a) T0693d1; (b) T0735d2; (c) T0739d1; and (d) T0739d2.

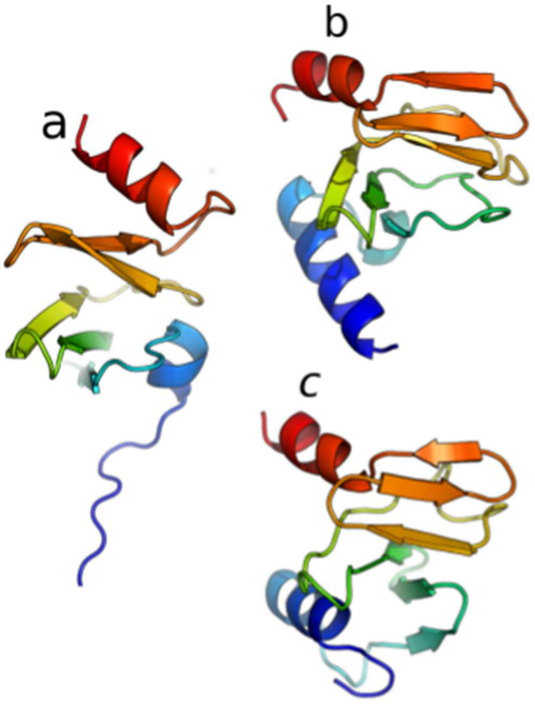


Figure 6. Result of the predictions of the CASP10 target T0739d1. Panel a represents the best model submitted to CASP10 (by the "Zhang" group, group Id: 335) for the region comprised between residues 12−96 of the protein (PDB code: 4oj6), the native structure of which is shown in panel b. The *hom*REM best model among the five with lower energy is shown in part C. The GDT-TS values for the two models are 35.9 (model a) and 54.1 (model c), respectively.

3. CONCLUSIONS

In this work we present a novel algorithm to improve the quality of the models generated by ab initio structure prediction protocols. The method exploits the fact that homologous proteins have similar native folds in order to improve the efficiency of Monte Carlo methods in exploring the protein conformation landscape and avoiding local minima. Local minima can be caused by a specific characteristic of the target sequence, by inaccuracies of the energy function or by a combination of these two factors. We believe that our method is effective in overcoming both problems. Conformations that are native-like are unlikely to be affected by our sequence exchange procedure; this can intuitively explain the observed ability of our method to improve the quality of the generated models: not only are we able to explore a larger fraction of the conformational space but most importantly we obtain models more similar to the native structure.

According to our tests, our strategy produces improved models with respect to the original Rosetta protocol and performs comparably to the best methods participating in the last CASP10 experiment.

We are not the first in using information derived from homology in de novo structure prediction,[19,25] but, to the best of our knowledge, our method uses this information in a novel way. We believe that by performing exchanges between of the homologous replica sequences during the simulation, our method can better explore the structure space by avoiding spurious or local minima. This is also supported by the fact that we observe a performance drop when we remove the replica exchange steps from our protocol (see the Supporting Information *hom*MC Supplementary Method, Figures S1 and S2, and Table S3a,b) and by the comparison of our results with those reported by Bonneau et al.[19] (Table S1).

In the past few years other methods have been published that, by exploiting the evolutionary information present in multiple sequence alignment, can predict nonlocal contact and improve ab initio prediction protocols.[26] One major advantage of our method is that it does not require an extremely large and diverse number of homologous sequences. In some of our test cases only a few homologous sequences were sufficient to successfully apply homREM. For example, in the case of the third IgG-binding domain from streptococcal protein G (scop id: d1igda, PDB code 1IDG, residues 1 to 61) as few as four homologous sequences were available after applying our sequence selection protocol. These were enough to achieve a reasonable improvement with respect to other methods (Figure 3).

4. MATERIALS AND METHODS

4.1. Data Sets. We use two data sets to test our procedure. The first (SCOP data set) consists of a set of 40 nonhomologous proteins (55−208 amino acid long) from the four structural classes (α, β, $\alpha + \beta$, α/β) of the SCOP database.[27] The second (CASP data set) consists of protein domains selected among the CASP10 Free Modeling targets.

The complete list with all the information on the data sets can be found in Supporting Information Table S2.

The fragment library is generated by the Robetta Server[28] using the option that excludes protein homologous to the target protein from the set of searched fragment sources.

4.2. Parameters Used for Comparison. To compare the results with those submitted to CASP, and to assess the performance of our *hom*REM method with ROSETTA and HREM, we use the GDT_TS defined as

$$\mathrm{GDT}_{\mathrm{TS}} = 100\frac{\sum_d \frac{\mathrm{GDT}_d}{N}}{4} \qquad d \in \{1, 2, 4, 8\} \tag{3}$$

where N is the number of residues and GDT_d is the number of corresponding C_α atoms within a distance of d Å between the model and the experimental structure.

4.3. Selection of the Homologous Proteins. To select the sequences of the homologous proteins to be assigned to the replicas in the simulations we first perform three iterations of psiBLAST[29] using the target sequence as query and select all the homologues with a sequence identity higher than 40% and an E-value lower than 1×10^{-4}. We then perform a global multiple sequence alignment of all the retrieved sequences using MAFFT.[30] The alignment is modified so that deletions and insertions with respect to the target protein are replaced by the corresponding amino acids of the target protein or removed, respectively.

CD-HIT[31] is used to cluster the resulting sequences using a 70% maximum identity threshold and a single representative from each cluster, excluding the target sequence cluster, is selected.

The procedure used for insertions and deletions ensures that all sequences have the same length, while the minimum and maximum similarity thresholds of 40% and 70% guarantee a sufficient diversity of the sequences while minimizing the risk of including nonhomologous or erroneously aligned proteins in the simulations.

4.4. Protocols Used for Comparison. As mentioned above, we compared our results with those obtained by the Rosetta low resolution protocol (ROSETTA) and by a more classic Hamiltonian Replica Exchange Method (HREM).[11]

For ROSETTA, we ran 100 000 Monte Carlo steps following the procedure described in Rohl et al.,[16] i.e., 78 000 steps of 9-residue fragment insertion with different combinations of the energy score functions and 22 000 steps of triple 3-residue fragment insertions. In the end we rescored the obtained structures using a different score (named score4 in Shmygelska and Levitt[11] and described in Rohl et al.[16]) that also takes into account short and long-range backbone−backbone hydrogen bonding energy terms.

For HREM, we followed the protocol adopted in Shmygelska and Levitt:[11] pairs of replicas are exchanged every 2000 Monte Carlo steps. Replicas are ordered according to their energy before applying the exchange algorithm, and exchanges are only attempted between the ith and the $(i + 1)$th replica for $i \epsilon (1, 2, 3, ..., N - 1)$ where N is the number of parallel simulations. The exchanges are accepted or rejected according to the Metropolis Criterion. It is to be noted that the HREM protocol does not precisely follow the ROSETTA fragment insertion stages described before since all the low-resolution energy functions are used at each fragment insertion stage; it performs nonetheless an equal number of overall steps.

All the protocols are implemented using PyRosetta,[32] an interactive Python-based interface to the Rosetta modeling suite.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.5b00371.

Data set specifics and complete results on both the data sets (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*(A.T.) Tel. +39 06 49914550, e-mail anna.tramontano@uniroma1.it.

*(P.M.) e-mail paolo.marcatili@gmail.com.

Present Address

†(P.M.) Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Anker Engelunds Vej 1, 2800 Lyngby, Denmark.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Author Contributions

‡(A.I. and P.M.) These authors contributed equally.

Funding

This work was supported by the King Abdullah University of Science and Technology (KAUST), Award Number KUK-I1-012-43 and PRIN 20108XYHJS.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful to all members of our group and to Dr. Ingemar André (Lund University) for their useful suggestions.

■ REFERENCES

(1) Qu, X.; Swanson, R.; Day, R.; Tsai, J. A guide to template based structure prediction. *Curr. Protein Pept. Sci.* **2009**, *10*, 270−85.

(2) Martí-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sánchez, R.; Melo, F.; Sali, A. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291−325.

(3) Espadaler, J.; Querol, E.; Aviles, F. X.; Oliva, B. Identification of Functional-associated loop motifs and application to protein function prediction. *Bioinformatics* **2006**, *22*, 2237−2243.

(4) Chothia, C.; Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **1986**, *5*, 823.

(5) Sander, C.; Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct., Funct., Genet.* **1991**, *9*, 56−68.

(6) Russell, R. B.; Saqi, M. A. S.; Sayle, R. A.; Bates, P. A.; Sternberg, M. J. E. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J. Mol. Biol.* **1997**, *269*, 423−439.

(7) Swendsen, R. H.; Wang, J. S. Replica Monte Carlo simulation of spin glasses. *Phys. Rev. Lett.* **1986**, *57*, 2607−2609.

(8) Zhang, Y.; Arakaki, A. K.; Skolnick, J. TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins: Struct., Funct., Genet.* **2005**, *61*, 91−8.

(9) Ortiz, A. R.; Kolinski, A.; Rotkiewicz, P.; Ilkowski, B.; Skolnick, J. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 177−185.

(10) Zhang, Y.; Kihara, D.; Skolnick, J. Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 192−201.

(11) Shmygelska, A.; Levitt, M. Generalized ensemble methods for de novo structure prediction. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 1415−20.

(12) Okamoto, Y. Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *J. Mol. Graphics Modell.* **2004**, *22*, 425−39.

(13) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. Replica exchange with solute tempering: a method for sampling biological systems in explicit water. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 13749−54.

(14) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: application to protein structure prediction. *J. Chem. Phys.* **2002**, *116*, 9058−9067.

(15) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.

(16) Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D. Protein structure prediction using Rosetta. *Methods Enzymol.* **2004**, *383*, 66−93.

(17) Bradley, P.; Misura, K. M.; Baker, D. Towards high-resolution de novo structure prediction for small proteins. *Science* **2005**, *309*, 1868−1871.

(18) Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)–round x. *Proteins: Struct., Funct., Genet.* **2014**, *82*, 1−6.

(19) Bonneau, R.; Strauss, C. E. M.; Baker, D. Improving the performance of rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 1−11.

(20) Rathore, N.; Chopra, M.; de Pablo, J. J. Optimal allocation of replicas in parallel tempering simulations. *J. Chem. Phys.* **2005**, *122*, 024111.

(21) Gray, J. J.; Chaudhury, S.; Lyskov, S. *The PyRosetta Interactive Platform for Protein Structure Prediction and Design*; Greylab Predictions: Miami, 2009; pp 27−32.

(22) Kryshtafovych, A.; Monastyrskyy, B.; Fidelis, K. CASP Prediction Center infrastructure and evaluation measures in CASP10 and CASP ROLL. *Proteins: Struct., Funct., Genet.* **2014**, *82*, 7−13.

(23) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P., Jr Optimization by simulated annealing. *Science* **1983**, *220*, 671−80.

(24) Tai, C. H.; Bai, H.; Taylor, T. J.; Lee, B. Assessment of template−free modeling in CASP10 and ROLL. *Proteins: Struct., Funct., Genet.* **2014**, *82*, 57−83.

(25) Keasar, C.; Tobi, D.; Elber, R.; Skolnick, J. Coupling the folding of homologous proteins. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 5880−5883.

(26) Morcos, F.; Pagnani, A.; Lunt, B.; Bertolino, A.; Marks, D. S.; Sander, C.; Zecchina, R.; Onuchic, J. N.; Hwa, T.; Weigt, M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (49), E1293−E1301.

(27) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536−40.

(28) Kim, D. E.; Chivian, D.; Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **2004**, *32*, W526−W531.

(29) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.

(30) Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **2002**, *30*, 3059−3066.

(31) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150−2.

(32) Chaudhury, S.; Lyskov, S.; Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **2010**, *26*, 689−691.

DOI: 10.1021/acs.jctc.5b00371