WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Separating true positive predicted residue contacts from false positive ones in mainly α proteins, using constrained Metropolis MC simulations

Spyridon Vicatos[1,2] and Yiannis N. Kaznessis[1,2]*

[1] Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, Minnesota

[2] Digital Technology Center, University of Minnesota, Minneapolis, Minnesota

ABSTRACT

We present a method that significantly improves the accuracy of predicted proximal residue pairs in protein molecules. Computational methods for predicting pairs of amino acids that are distant in the protein sequence but close in the protein 3D structure can benefit attempts to* in silico *recognize the fold of a protein molecule. Unfortunately, currently available methods suffer from low predictive accuracy. In this work, we use Monte Carlo simulations to fold protein molecules with proximal pair predictions used as additional energy constraints. To test our methods, we study molecules with known tertiary structures. With Monte Carlo, we generate ensembles of structures for each set of residues constraints. The distribution of the root mean square deviation of the folded structures from the known native structure reveals clear information about the accuracy of the constraint sets used. With recursive substitutions of constraints, false positive predictions are identified and filtered out and significant improvements in accuracy are observed.

Proteins 2008; 70:539–552.

Key words: proximal residues; fold prediction; contact map; residue constraints; Metropolis Monte Carlo.

INTRODUCTION

Prediction of an arbitrary protein's tertiary structure, given only its primary amino acid sequence, is one of the most important scientific problems that is yet to be solved. The three-dimensional structure of a protein largely defines its biological function. Therefore, knowledge of its structure is crucial for the understanding of its biological mechanisms.

Among the many existing computational protein structure prediction methods, there are those focusing on the prediction of residue pairs, which are distant in protein primary structure, but proximal in its native folded structure, also called distant residue contacts. Finding those pairs can be a very important first step for protein fold prediction. The main problem with current methods is that they typically predict a substantial number of pairs not proximal in reality. Those are characterized as false positive predicted contacts.

Although efforts have been reported for improving the predictive accuracy of theoretical methods,[1–4] not much effort has been expended to separate the correct predictions of a single prediction set from the incorrect ones, i.e. enrich this prediction set with true positive predicted pairs. Fariselli and Olmea[2,3] combined correlated mutations analysis (CMA)[1] with neural networks and other statistical information such as sequence separation along the chain, alignment stability, residue specific contact occupancy, and formation of contact networks. The results obtained had an accuracy (defined as the ratio of true positive predictions over all predictions) of no more than 30%. In our previous work,[4] although high accuracy predictions are obtained, this occurs at the expense of the number of predicted contacts. More recently, Fariseli *et al.*[5] created the CORNET neural network predictor, with an accuracy of around 26%. Singer *et al.*[6] have used structural information from the protein data bank (PDB)[7] to create a contact likelihood matrix, which can be used for contact map prediction, but the average accuracy obtained is still low, around 15%. Recent contact prediction methods, used in CASP6[8,9] such as those developed by MacCallum,[10] and Punta and Rost,[11] cannot exceed an accuracy threshold of 30%. Enriching the sets of predicted contacts requires a method to recognize the false from the true positive predictions, usually different from the initial method used to create the prediction set.

Grant sponsor: American Chemical Society (Petroleum Research Grant); Grant number: G7-38758; Grant sponsor: National Center for Supercomputing Applications; Grant number: TG-MCA04N033.
*Correspondence to: Yiannis Kaznessis, Department of Chemical Engineering and Materials Science, University of Minnesota, 421 Washington Ave. SE, Minneapolis, MN 55455. E-mail: yiannis@cems.umn.edu
Received 11 September 2006; Revised 28 February 2007; Accepted 27 March 2007
Published online 24 August 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21553

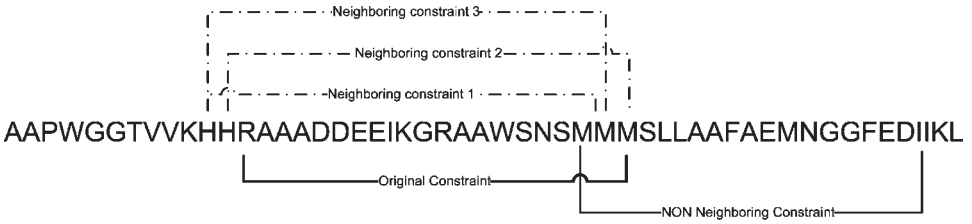

Figure 1

A protein sequence and its possible constraints. An original constraint is shown on the lower side of the sequence. Possible neighboring constraints examples, close to the original, are indicated on the upper side of the sequence. Those are the ones that are applied for the generation of replacement fold ensembles. A non-neighboring constraint is the one shown in the lower right corner, which fixes two residues very far away from those of the original one.

In this article, we report on the potential of constrained Metropolis Monte Carlo simulations of proteins, to separate true positive residue contacts (TP) from the false positive ones (FP). When performing Metropolis Monte Carlo simulation on a protein, its conformational phase space is explored. Protein conformations with lower energy are more preferable than those with higher ones, and eventually after many trials, a conformation with minimum energy can in principle be achieved.[12–14] The idea of using constraints during a MC simulation has been previously discussed.[15] To expedite the search toward minimum energy, residues that are considered to be proximal can be used as constraints, so that the total energy of the protein would be penalized with an additional energy function directly proportional to the squared distance of the constrained residues. Therefore, the simulation would be forced to explore conformations, which satisfy the constraints by minimizing not only the protein chain's energy, but also the penalty energy functions of the constraints. When the constraint is correct, i.e. the constrained residues are truly proximal, the simulation will, in principle, be directed toward protein conformations closer to its true native state. The opposite occurs when a constraint is incorrect, i.e. the constrained residues are not actually proximal. MC simulations usually involve large number of moves (trials). Typical simulations need more than 10^6 moves to obtain the necessary results. In protein simulations, however, the simulated chain is very often entrapped in local minima energy conformations before it reaches this amount of trials. In this work, the goal is to use a relatively small number of around 10^5 trials, avoid local minima entrapments, and create simulated protein folds, which can give us an indication of the true or false proximity of the used residue constraints.

We fold atomistically represented protein molecules with known tertiary structure, using a classical mechanical potential energy function. We use sets of predicted pairs of proximal residues, including both true and false positives, to generate ensembles of folded structures. Analysis of the distribution of the generated structures reveals a clear strategy for identifying proximal residues. In what follows, we describe the method for constraint-based protein-folding and detail the results that point to a novel method for filtering out false positive contacts.

MATERIALS AND METHODS

Fundamental hypothesis of the method

In a MC simulation with constraints based on predicted proximal residue pairs, the general case would be that not all pairs are proximal. It is expected that regardless of the length of the simulation, false positive pairs will tend to lead the simulation to non-native fold conformations. On the other hand, true positive pairs used will tend to fold the protein closer to its native, low energy structure. Therefore, when ensembles of protein conformations are generated using different sets of constrained residue pairs, we expect the ensemble created by using a set with more truly proximal residue pairs to have on average a fold closer to the protein's native state than an ensemble of folds created with a set of constraints with less truly proximal residue pairs. In other words, a heavy atom (nonhydrogen) root mean square deviation (RMSD) distribution of protein conformations from its native state will be toward lower RMSD values, if the constraints used to create those conformations are mostly true positive.

Consider one ensemble of protein conformations, created by using a small number of residue pair constraints. Suppose that one of the constrained pairs is truly proximal, contributing favorably toward native folds conformations during MC simulations. In this work, two residues are considered to be proximal when the minimum distance of their nonhydrogen atoms[4] is equal to or less than 6 Å, which is a widely used minimum distance.[4,16–19] If this pair is replaced by a different one, *close* to the original in terms of its position in the protein sequence, then the replacement pair is expected to usually favor less-than-native fold conformations. Examples of neighboring constraints, close to the original, are shown in Figure 1.

DOI 10.1002/prot

There is however the possibility that the new pair allows the protein to reach a better fold conformation, compared to the fold of the original pair constraint. The result of an unfavorable fold compared to the fold of the original can be explained as follows: The replacement pair constraint, possibly a nonproximal pair, drives the simulation to a path where the resulted fold has the segment of the protein containing the original TP pair constraint misfolded. The other possibility, an improved fold, may occur if the segment of the original true positive is a cluster of many residues, which are proximal. In this case, a replacement pair constraint, regardless of the proximity of its residues, may drive the simulation to a path where a larger number of residues contribute favorably in the final fold than with the original constraint pair. In both cases, and we give emphasis to this hypothesis, when a TP is replaced by a neighboring constraint, we expect a considerable difference in folding behavior of the simulated protein, either favorable or unfavorable, compared to that of the original TP pair constraint.

On the other hand, if a nonproximal pair is replaced by a neighboring constraint, it is expected that, like the original nonproximal pair, the new pair will not contribute favorably toward native fold conformations during simulations. In this case, the replacement would provide conformations during simulations similar to those of the original nonproximal one, and no dramatic changes will occur. This would be valid especially when the vicinity of residues consisting the original nonproximal constraint is separated by a large distance, more than 10 Å. In this case, the majority of the neighboring constraints would be, as the original, nonproximal ones, and their contribution during simulations would be again nonfavorable ones without any dramatic changes compared to the original constraint. Therefore, in the case of a nonproximal pair, we do not expect a significantly altered folding behavior when we exchange the original pair with the neighboring ones, as in the truly proximal pair case described previously. In other words, we expect the fold ensembles to have similar, far-from native, folds.

The question we want to answer is whether we can assess the nature of the original constraint, by comparing the quality of the fold of protein ensembles of the original constrain and its replacements. When we observe significant changes in the fold quality, between the simulated structures created using the original constraint, and those created by its replacement, then the original constraint can be classified as a truly proximal one, i.e. a true positive prediction. If the changes are minimal, then the original constraint can be characterized as nonproximal.

In all cases discussed before, the assumption has been made that all nonproximal residue pairs have the same or at least very similar nonfavorable contributions during simulations, regardless of their distance or the residues involved. Similarly, the assumption has been made that all proximal residues contribute favorably during simulations.

The algorithm of the TP–FP separation method, recognizing the truly proximal pairs from the nonproximal ones is outlined in the following steps:

- A proximal residue prediction method is employed, to predict proximal residue pairs for the protein of interest.
- An ensemble of simulated protein folds is created with MC, using a small subset of residue constraint pairs. This ensemble is called original pair ensemble.
- Initial constraint pairs are replaced one at the time with similar, neighboring constraint pairs and new ensembles of simulated protein folds are created by using the replacement pairs. Those are called replacement pair ensembles.
- Comparison of the average fold of the two ensembles, the original and the replacement, leads to the assessment of the original constraint.

Choice of the proteins, proximal residue pairs sets

We choose protein sequences with known structure, so that predictions can be evaluated, and generate a set of residue pairs predicted by some method to be proximal. From our previous work, proximal residue pairs in terms of their minimum distance have been predicted for a large number of protein architectures by using a correlated mutation analysis (CMA) algorithm.[4] Correlation coefficients (cc) between pairs have been calculated, with higher cc value indicating high probability of proximity. A brief overview of CMA is shown in Figure 2. From the original protein set used in our previous work, 10 proteins have been chosen, all from the mainly alpha class. These proteins also represent well studied PFAM families.[20] They are selected so that most α class architectures and sequence lengths up to 150 residues are represented. For each of those proteins, the 10 residue contact predictions with the highest correlation coefficients have been selected as the prediction sets. The proteins chosen, the prediction sets for each one of them, and the name of the PFAM family they represent are shown in Table I. In all predicted pair sets, we have two categories of pairs: those which are truly proximal, with a minimum distance to be less or equal to 6 Å, and those which are not. We have chosen this value of the definition of a residue contact, because it was shown in our previous work[4] that it is the minimum distance that CMA gives the highest accuracy. Also, 6 Å minimum distance is comparable, and in some cases stricter compared to the 8 Å distance of C-β carbon atoms, used in CASP6.[8] For the proteins selected, our CMA method has an average accu-

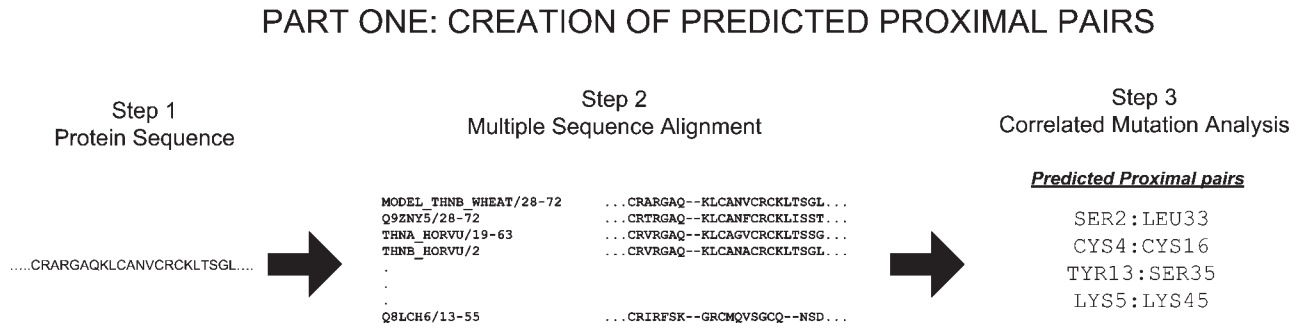

Figure 2

First part of the TP–FP separation algorithm. In this work, correlated mutation analysis (CMA) has been employed, to create a set of protein residue pairs predicted to be proximal. The pair sets are large, containing hundreds of proximal pair predictions. Only the pairs with very high correlation coefficients are taken for the creation of the proximal pairs set used for simulations. Any other method, which predicts proximal residues can be used.

Table I

Families, Proteins, and Residue Pairs Sets Chosen for the Method Evaluation

HTH_8[a] (1FIP[b], 41[c])		Arg_Repressor[a] (1AOY[b], 71[c])		FE_DEP_REPR_C[a] (2TDX[b], 71[c])		Phospholipid_A2_1[a] (1BUN[b], 119[c])		ATP-synt_DE[a] (1AQT[b], 45[c])	
N^d	Proximal	N^d	Proximal	N^d	Proximal	N^d	Proximal	N^d	Proximal
1	GLU59, LYS91	4	LEU10, MET49	2	ARG77, ASP110	1	CYS61, CYS86	3	GLU91, ILE131
4	GLN74, ARG85	7	SER24, MET65	5	MET76, ILE132	7	CYS79, CYS91	6	LYS98, ILE125
7	GLN74, ARG89	9	VAL56, CYS68	8	ILE90, TRP104	8	TYR11, LEU72	8	ALA101, LEU121
9	MET81, LEU88								
	Not Proximal		Not Proximal		Not Proximal		Not Proximal		Not Proximal
2	LEU63, GLY82	1	LEU10, SER44	1	ALA101, ASP110	2	CYS100, CYS119	1	LYS98, ALA115
3	MET65, ASN84	2	ILE29, ARG57	3	ARG77, ALA101	3	ASP39, CYS51	2	LYS98, VAL112
5	LEU62, ARG76	3	LEU10, LEU73	4	LEU92, ARG103	4	CYS86, ASP94	4	GLU91, GLU120
6	ASP64, GLY72	5	LYS45, LEU73	6	ALA82, ILE94	5	CYS79, CYS119	5	LEU89, LYS123
8	LEU62, ARG89	6	PHE14, LYS52	7	LEU85, GLU100	6	CYS86, CYS100	7	LYS100, ALA115
10	GLY72, ARG85	8	SER24, LYS52	9	GLY91, GLU105	9	TYR11, SER104	9	ALA94, ALA117
		10	ALA5, ALA55	10	ALA101, PRO133	10	CYS51, CYS61	10	LYS100, ILE125

Cher_N[a] (1AF7[b], 57[c])		Acyl_coa_dh[a] (1IVH[b], 150[c])		Chorismate_mut[a] (1ECM[b], 85[c])		RGS[a] (1AGR[b], 119[c])		ATP-synt_C[a] (1A91[b], 70[c])	
N^d	Proximal	N^d	Proximal	N^d	Proximal	N^d	Proximal	N^d	Proximal
3	ARG57, TYR69	4	ARG255, ALA375			2	TYR98, ILE114	2	MET16, GLY69
7	ARG29, LEU70	7	ILE332, GLY374			5	GLU96, GLN153	7	GLY32, THR51
10	TYR51, PHE66	8	ILE347, GLY362			8	ILE119, LEU129		
						10	LEU107, PRO144		
	Not Proximal		Not Proximal		Not Proximal		Not Proximal		Not Proximal
1	ILE40, MET49	1	LEU304, LYS326	1	ARG47, LEU86	1	LYS113, GLN122	1	ARG50, VAL60
2	GLN32, LEU70	2	VAL257, ILE332	2	LEU55, SER84	3	LYS125, ASN140	3	MET17, GLY32
4	SER23, GLY39	3	ARG280 MET297	3	TYR72, THR87	4	ILE93, GLU126	4	MET17, VAL60
5	ARG29, ARG56	5	GLY260, ILE347	4	LEU7, LEU17	6	GLU97, ILE114	5	ALA21, GLY32
6	ILE30, TYR51	6	GLU337, ILE347	5	VAL46, LEU55	7	TYR98, LEU129	6	ALA21, PHE35
8	LEU33, ARG56	9	LEU258, GLN291	6	ARG11, LEU25	9	LEU107, PHE118	8	LEU31, ILE66
9	VAL41, MET49	10	CYS349, GLY376	7	ASP18, ILE80			9	GLY32, GLN52
				8	GLU19, VAL46			10	LEU9, GLY32
				9	LEU7, GLU30				
				10	GLY63, LEU76				

[a]The PFAM family name.
[b]The PDB ID name of the protein.
[c]The sequence size, in number of residues.
[d]The ID number randomly given to the specific pair constraint.

DOI 10.1002/prot

Table II
The Random Shuffling of Constraints and the Resulting Subsets

	Subset 1			Subset 2			Subset 3			
Shuffle 1	5	6	9	2	8	10	1	3	4	7
Shuffle 2	7	8	9	2	4	10	1	3	5	6
Shuffle 3	1	2	10	6	7	9	3	4	5	8
Shuffle 4	2	3	10	4	5	8	1	6	7	9
Shuffle 5	2	5	7	8	9	10	1	3	4	6
Shuffle 6	1	4	6	2	5	9	3	7	8	10
Shuffle 7	3	8	9	1	6	7	2	4	5	10
Shuffle 8	3	7	10	2	5	9	1	4	6	8

In each subset, the ID numbers of the constraints shown in Table I have been used, so that random combinations of constraint pairs are created for the MC simulations.

racy of around 30%, defined as the ratio of TP over all 10 predictions.

Generation of random subgroups from the predicted pair set and replacement pairs

Each MC simulation uses up to four constraints, whether it is performed with original constraints or with neighboring replacements. We have found that this is an appropriate constraint number for the protein sizes studied. More constraints would make the protein very difficult to manipulate due to excessive constraining. We have performed simulations with medium sized proteins of around 75 residues, using up to nine constraints. We have found that when the system was using seven constraints and above the MC rotational moves rejection rate at the initial steps of the process was more than 95%. The result was that the initial structure at the end of the MC process was virtually unchanged, and a poor simulated structures ensemble was obtained. Reducing the number of constraint to five and six, considerably improved (reduced) the rejection rate at the initial stages of the process. Because we wanted to study small proteins as well, a smaller number of constraints compared to the medium-sized proteins had to be used, again for the same reason of excessive constraining. Therefore, we decided for all tested proteins to use up to four constraints.

For each predicted residue set shown in Table I, random subgroup sets of three and four pairs have been created, and used during simulations. Table II shows in detail all the constraint subsets used for all proteins. Since the ID number given to the constraints in Table I was random, the combinations of constraints, which were created and shown in Table II were also random.

Neighboring residue pairs are selected randomly, to replace each one of the original predicted pairs. The replacement pair is close to the original pair by a maximum delta radius[4,21] of two residues. For example, looking at protein 1BUN of family Phospholipid_A2_1, the process is as follows: First, the (5, 6, 9) set of constraints (as shown in Shuffle 1, subset 1 of Table II) is chosen and tested. Then a neighboring residue pair replacement for constraint 5 is chosen and the new set of constraints is tested again. The same procedure is repeated for constraints 6 and 9. Constraints 5, 6, and 9 for 1BUN are CYS79–CYS119, CYS86–CYS100, and TYR11–SER104 respectively, as shown in Table I. For a different protein, like for example protein 1AOY, the (5, 6, 9) set constraints are LYS45–LEU73, PHE14–LYS52, and the truly proximal constrain VAL56–CYS68. Ultimately, for each protein, 24 original constraints sets are tested (as shown in Table II), as well as 16 × 3 and 8 × 4 replacement sets. The total number of sets of constraints for each protein is 104.

Monte Carlo simulation

Once the protein and the constraints—original or replacement—have been chosen, Metropolis Monte Carlo simulation (MC) subject to residue constraints is conducted to generate an ensemble containing 100 simulated structures. We have thus generated 10,400 folded structures for each protein. The CHARMM[22] molecular force field has been used to analyze and calculate the total free energy of the protein in each MC step. A diagram of the entire process is shown in Figure 3.

The initial structure of each simulated protein, which serves as a starting point for the simulations, has been created by feeding its secondary structure information into the program ProteinShop.[23] ProteinShop creates an initial structure of the protein, with all its secondary structure elements positioned randomly in a tertiary structure. This structure is then relaxed by using 200 Metropolis MC steps with small rotational angle step and without any constraints. Since this work is testing the hypothesis of true positive pairs, we are using accurate secondary structure information, for the initial structure creation. It is not the intention of this work to test the hypothesis with radically nonaccurate secondary structure information, such as for example replacing a helical structure with an incorrectly predicted β sheet.

Metropolis MC is performed for 5000 moves subject to a maximum of four constraints. Moves consist of randomly rotating either one of the two Cα bonds between N and C, respectively. The main secondary structures of helices are kept intact, their backbone coordinates not being subject to MC moves. The protein side chains however are subject to MC moves, during the relaxation part of the process. The Monte Carlo algorithm is described in detail in Figure 4.

Not all constraints are applied at all times, and constraints do not have the same Young's modulus constant during the simulation. In the first steps of the simulation, only the constraints, which are not far apart in the protein sequence, are employed with a relatively small Young's modulus, and with the van der Waals forces

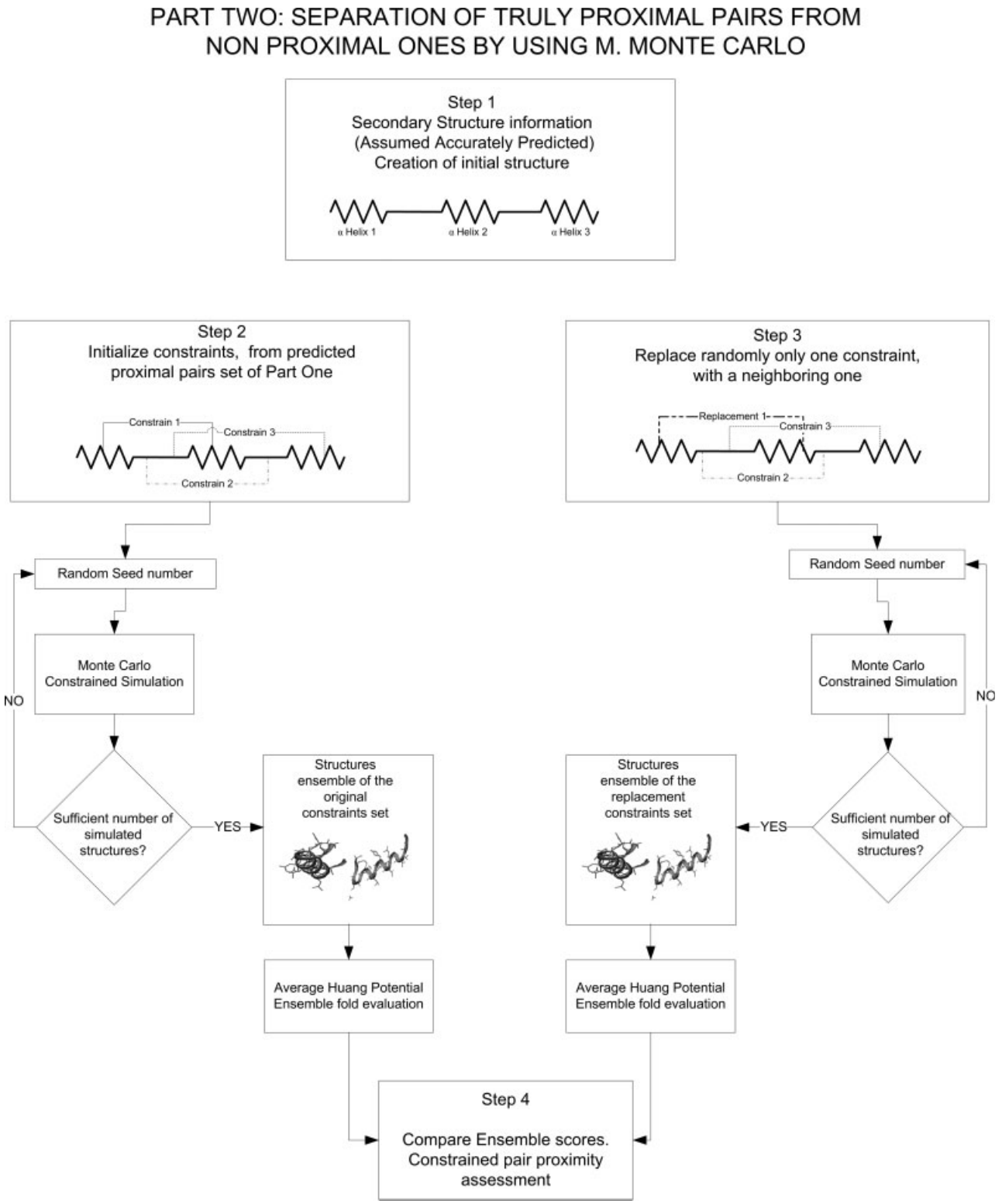

Figure 3
Diagram of the prediction–separation algorithm.

DOI 10.1002/prot

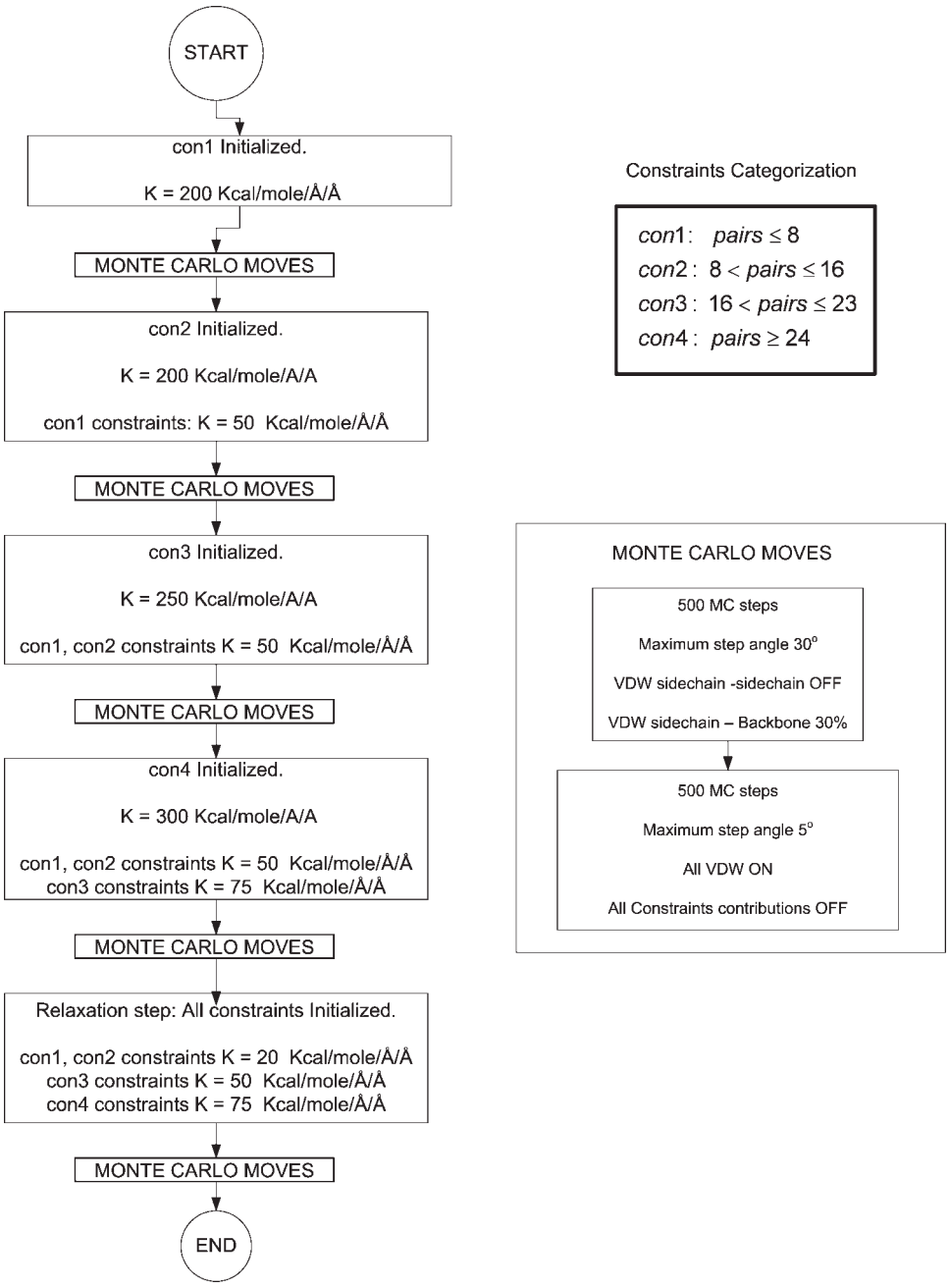

Figure 4

The Monte Carlo process for the creation of one simulated structure, followed in this work. Each "MONTE CARLO MOVES" block from the left diagram is described in detail in the lower right block with the same label. This figure describes in detail the individual Monte Carlo block of Figure 3.

(VDW) from the side chain interactions reduced to 30% of their original value, for better maneuverability of the protein's helices. The maximum rotational angle is 30°, and the constraints are employed for 500 MC steps. Then a period of energy relaxation follows, where the protein chain is allowed to undergo Cα bonds rotation without any constraints but with a maximum rotational angle of only 5°, and with all VDW forces included at full strength into the total energy of the protein chain. The reason behind this is that due to the brute force used to bring the constrained amino acids close, and by relaxing the side chain energy contributions, the protein may have been allowed to reach a conformation where atom collisions and secondary structure strong overlap-

ping may occur. The relaxation taking place in this part of the simulation consists of another 500 MC steps, and allows the protein to separate its parts which undergo collisions.

The same procedure is followed for the constraints, which are further apart in the protein sequence. The constraints used in the previous cycles still contribute to the total protein energy, but the newly used constraints have much larger Young's modulus. The gradual increase of the Young's modulus of the constraints, which are far apart in the primary sequence has been done under the assumption that it is energetically-speaking more difficult to force two residues far away in the primary sequence (for example 20 Å) to get and stay close. Intermediate residue collisions and backbone steric hindrances are much stronger when the constrained residues become more distant in the sequence. Therefore a stronger force (thus a higher Young's modulus) needs to be implemented to bring these distant residues closer into the final tertiary structure. That is the reason why residue contacts of up to 12 residues have up to one third reduced Young's modulus strength, compared to the contacts more than 17 residues apart.

The process is performed five times for different values of Young's modulus, as described in the left block diagram of Figure 4, for 1000 total MC steps each time as described in the lower right block of Figure 4. The whole process for the creation of one simulated structure needs 5000 MC steps to be completed. Changing the initial seed number for the random number generator results in a different final structure for each trial.

Use of hydrophobic potential for the evaluation of the simulated folds

We generate ensembles with a hundred simulated structures each. We have found this to be a sufficient number for comparison. Once the ensemble of simulated protein structures has been created for both the original and the replacement constraint sets, the quality of their fold has to be assessed. The simple, direct way to do that is by comparing all resulted folds with the native fold and calculating the coordinate's root mean square deviation (RMSD). We do this for all proteins tested, since the native structure is known. However, in the general case of protein structure prediction, the native fold will be unknown. It would then be useful to assess the simulated folds not by simple native structure comparison, but with the use of a scoring function, which takes into account the protein structural features and evaluates the correctness of the folds. The Huang hydrophobic contact potential[24,25] has been used in this work for fold evaluations, because of its simplicity and robustness, as well as the good performance for distinguishing a large margin of misfolded structures. Certainly other methods exist for evaluating protein fold, but exploring an optimum one is beyond the scope of this study.

For each ensemble, the quality of the fold of each individual protein structure is being assessed by using the Huang potential, and then the average score for the entire ensemble is calculated. This average Huang score is used for comparing the original and replacement ensembles.

$$\mathrm{EHS} = \frac{\sum_j^n \mathrm{HS}_j}{n} \tag{1}$$

where EHS is the average Huang score of the ensemble, n is the number of simulated structures in the ensemble (in our case 100 structures), and HS_j is the Huang score for each individual structure j of the ensemble.

The absolute difference between average Huang scores of the original constraint and replacement constraint ensembles DHS_i is calculated, and then averaged among the number k of original ensembles the constraint in question is used, to give the final average Huang score FHS for the constraint. We expect a high value of average Huang score for a proximal constraint, and a low value for a nonproximal constraint.

For example, for protein 1BUN, we look at the EHS of the eight original constraint ensembles that contain constraint 5, and the eight related ensembles that contain replacements for the same constraint, and calculate the absolute difference DHS_i.

$$\mathrm{DHS}_i = \left|\left(\mathrm{EHS}_{\mathrm{original}}\right)_i - \left(\mathrm{EHS}_{\mathrm{replacement}}\right)_i\right| \tag{2}$$

Finally, we average over the eight absolute differences to find the final Huang score for constraint 5.

$$\mathrm{FHS} = \frac{\sum_{i=1}^k \mathrm{DHS}_i}{k} \tag{3}$$

RESULTS AND DISCUSSION

General evaluation of the simulations and the method

Simulations have been performed for 10 main α proteins. Each set of predicted residue pairs of a tested protein, as shown in Table I, is shuffled as shown in Table II. Then, each shuffle of three or four residue pairs is used as constraints in MC simulations of the tested protein.

All simulated structure distributions, regardless of the quality of the constraints, have an average RMSD, which is not close to the native structure. Average RMSD is around 5–10 Å for small and 15–23 Å for large proteins.

DOI 10.1002/prot

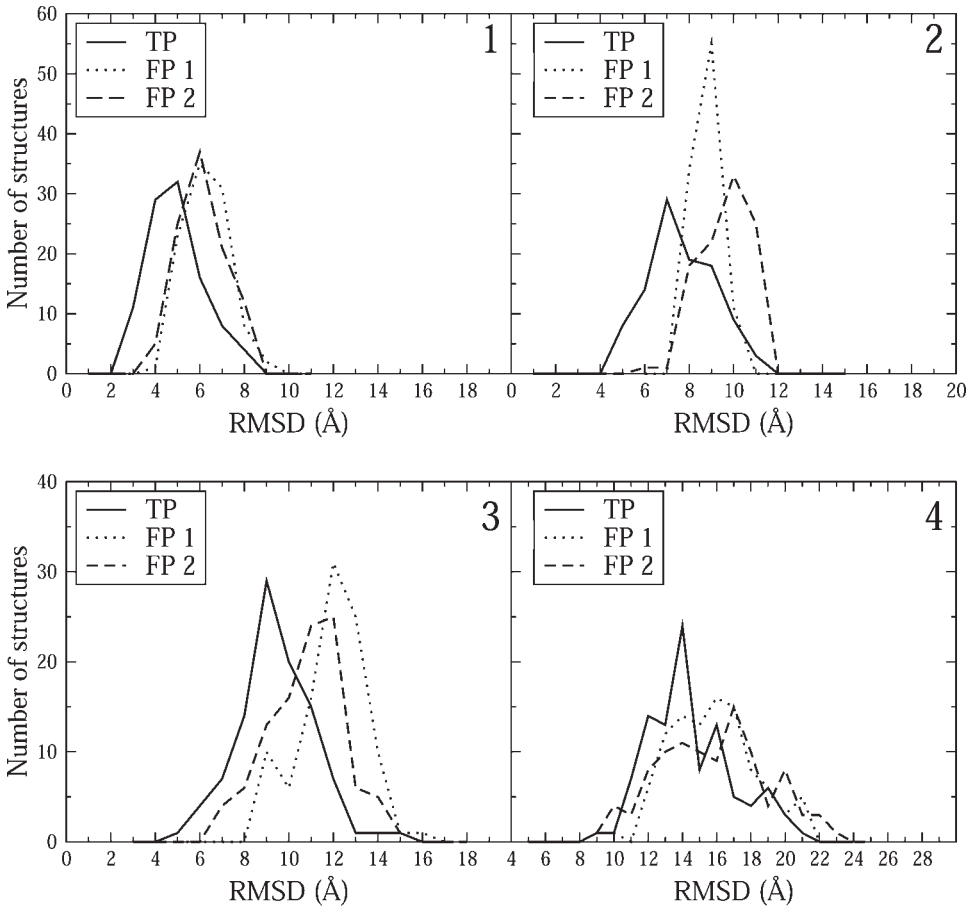

Figure 5

Comparison between RMSD distributions compared to the native structure of simulated structures for four main α proteins. Each curve represents the RMSD distribution from the native structure, for an ensemble created by using different sets of constraints. The solid line for all four cases represent the ensemble of the number of protein structures created by using a majority of truly proximal constraints (TP), while the dashed and the dotted lines represent ensembles created by using mainly nonproximal constraints (FP 1 and FP 2). In all four cases, we observe that the ensemble created by the TP constraints is always shifted to the left, compared to all the other ensembles of the same protein created by FP constraints. Therefore, ensembles created by using mostly TP constraints have an average RMSD closer to the native structure. Graph 1: Protein 1FIP of the HTH_8 family. Graph 2: Protein 1AF7 of the Cher_N family. Graph 3: Protein 1A91 of the ATP-synt_C family. Graph 4: Protein 1AGR of the RGS family. The constraint IDs used in MC simulations to create the ensembles are shown in Table III.

Variation depends upon the protein architecture and the initial structure from which the simulation starts. The initial structure used, created by Proteinshop software, is always very far from its native fold, around 20–24 Å for small and above 35 Å for large proteins. However, simulations with a large number of truly proximal constraints provide simulated structures with average RMSD much closer to the native structure than that of the initial structure, implying that MC simulation with more truly proximal constraints than nonproximal ones, can lead to the creation of favorable simulated folds.

First, we demonstrate the fact that simulations containing mainly truly proximal constraints provide ensembles with an average RMSD closer to the native structure, compared to those which use nonproximal constraints. In Figure 5, we show that ensembles created by using proximal constraints have their average RMSD shifted to the left of the RMSD axis (closer to native fold), while ensembles created by using mainly nonproximal constraints tend to obtain higher RMSD values i.e. shifted to the right. The average RMSD difference in the proteins shown in Figure 5 varies from small values of around 1 Å in the Cher_N 1AF7 protein, to larger values of around 2.5 Å in the Atp-synt_C 1A91 protein. Values of average RMSD differences for the entire protein test set are shown in Table III. Protein 1ECM of the Chorismate_mut family has a residue test set without truly proximal constraints (also shown in Table I), therefore average RMSD comparison has not been performed for this protein's proximal residues test set. In Table III, it is shown that differences up to around 4–7 Å for different ensembles can be achieved.

Table III
RMSD Differences for Protein Set

Protein family (Name)	Mainly proximal constraints (TP)		Mainly nonproximal constraints (FP 1)		Mainly nonproximal constraints (FP 2)	
	Constraints	Average RMSD (Å)	Constraints	Average RMSD (Å)	Constraints	Average RMSD (Å)
HTH_8[a] (1FIP[b])	1[c], 4[c], 9[c]	5.45	2, 8, 10	6.79	1[c], 3, 5, 6	6.65
Arg_Repressor[a] (1AOY[b])	6, 7[c], 9[c]	10.59	2, 3, 10	11.06	1, 3, 5, 7	11.2
FE_DEP_REPR_C[a] (2TDX[b])	2[c], 8[c], 10	9.31	6, 7, 9	14.25	1, 3, 4, 7	10.8
Phospholipid_A2_1[a] (1BUN[b])	7[c], 8[c], 9	18.71	2, 4, 10	25.11	1, 3, 5, 6	22.81
ATP-synt_DE[a] (1AQT[b])	3[c], 8[c], 9	5.86	2, 4, 10	6.94	1, 6, 7, 9	8.38
Cher_N[a] (1AF7[b])	3[c], 7[c], 10[c]	8.15	2, 5, 9	9.27	1, 4, 6, 8	10.10
Acyl_coa_dh[a] (1IVH[b])	4[c], 5, 8[c]	17.82	5, 6, 9	18.67	1, 2, 10	18.31
Chorismate_mut[a] (1ECM[b])	–	–	–	–	–	–
RGS[a] (1AGR[b])	2[c], 8[c], 10[c]	14.98	1, 4, 6	16.24	3, 7, 10[c]	16.26
ATP-synt_C[a] (1A91[b])	2[c], 5, 7[c]	9.92	8, 9, 10	12.42	1, 4, 6	11.27

[a]The PFAM family name.
[b]The PDB ID name of the protein.
[c]Proximal constraints.

It should be mentioned that it was only for protein 1IVH of the Acyl_Coa_dh family, the largest protein of 150 residues in our protein test set, that some ensembles created by mainly proximal constraints do not show average RMSD closer to native structure compared to ensembles created by mainly nonproximal constraints. This protein shows average RMSD of 21.69 Å for the ensemble created by proximal constraints 7 and 8 and nonproximal constraint 9, while two ensembles created by nonproximal constraints 5, 6, 9 and 1, 2, 10 show an average RMSD of 18.67 and 18.31 Å, respectively. Additionally, the ensemble created from nonproximal constraints 2, 3, and 10 show an average RMSD of 16.06 Å, which is far lower than many of the other ensembles for this protein. Therefore, we conclude that our method does not work well for proteins of about 150 residues and above, and we have not tested any protein fold above that sequence size.

In this work, we concentrate on trying to separate proximal residues predictions from nonproximal ones. Once an enriched set or proximal residues has been obtained, the important question of how to obtain simulated protein structures closer to the native state can then be addressed, although this is beyond the scope of this work.

Randomly chosen neighboring residue pairs have replaced the original ones shown in Table I. Because of this replacement, simulated structures with different fold quality compared to those of the original constraints have been obtained. According to the results, it has been found that the average RMSD of the simulated structures obtained using replacement constraints differs from the average RMSD of the simulated structures obtained from the original constraints. The extent of this difference depends upon the residue proximity of the original pair. There are single simulations where TP constraints are replaced with neighboring constraints, which give an RMSD difference of more than 1.7 Å. On the other hand, simulations with FP constraints do not show RMSD difference larger than 0.2 Å. The RMSD shift, due to the replacement of an original truly proximal constraint, could be either for better or for worse fold qualities, as shown in Graphs 1 and 2 of Figure 6. Certainly, cases where the shift of a truly proximal constraint replacement is negligible cannot be avoided (Figure 6, Graph 3). It is in this case that it becomes difficult to reliably discern a FP from a TP prediction.

It is because of this observation shown in Figure 6 that we investigate different shuffles of combinations for the 10 constraints of each tested protein, shown in Table II. Each constraint is used in a total of eight randomly chosen shuffles with the other nine constraints belonging to the same group of the prediction set. This is done to allow each constraint to interact with as many other constraints in MC simulations as possible. Some of these combinations would provide RMSD profiles similar to that of Graph 3 of Figure 6. However, the exploration of more combinations of the same constraint would eventually create ensembles where a replacement of the TP constraint with a neighboring one will show the desired fold quality change.

The Huang potential score distributions follow a similar shifting with the corresponding RMSD distributions. This is shown in Figure 7, where the corresponding Huang potential score distributions of protein 1AF7 of Figure 6 are shown.

Constraint proximity assessment

For all the simulations performed by using the constraint subsets shown in Table II, the simulated structures ensembles created were evaluated by the use of the

DOI 10.1002/prot

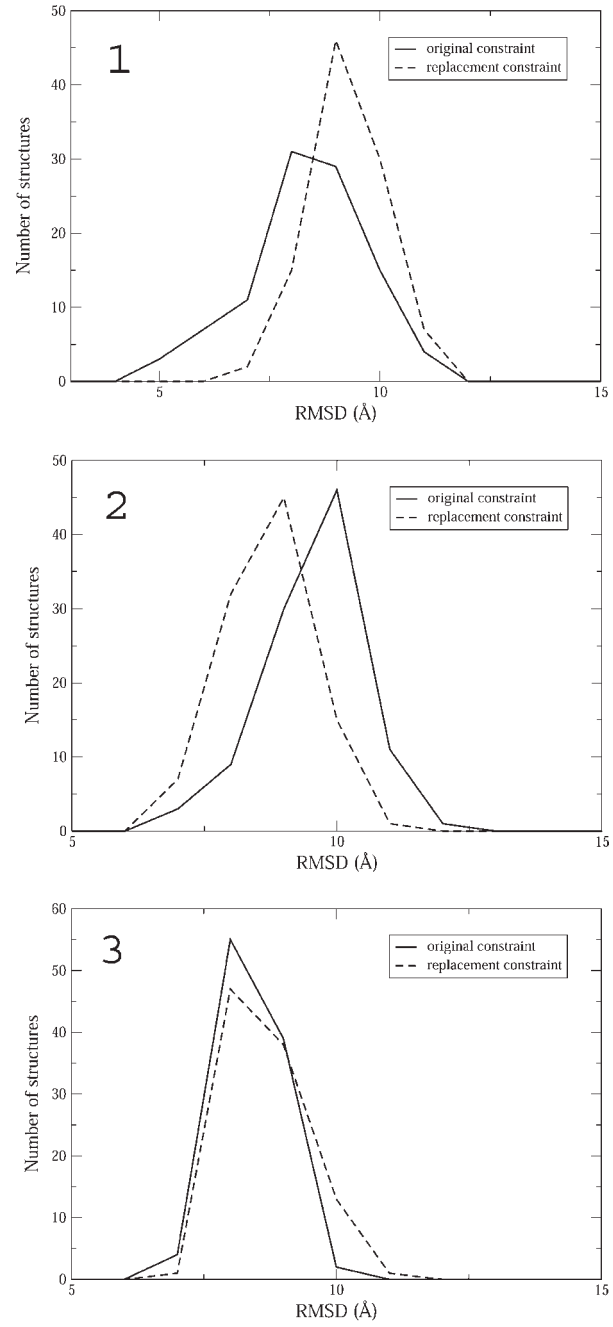

Figure 6

Comparison between RMSD distributions of simulated structures created by using truly proximal constraints and replacement constraints for protein 1AF7 family Cher_N. For all graphs, solid lines represent the original constraint and dashed lines the replacement. Graph1: TYR51-PHE66 true positive constraint pair 10 from Table I. The simulated structures distribution created by using the replacement is shifted to the right, indicating that the new folds are on average less favorable compared to the folds of the original simulated structures distribution. Graph 2: ARG57-TYR69 true positive constraint pair 3 from Table I. The simulated structures distribution created by using the replacement is shifted to the left, indicating that the new folds are on average more favorable compared to the folds of the original simulated structures distribution. Graph 3: ARG29:LEU70 true positive constraint pair 7 from Table I. The two distributions are very similar, therefore their average RMSD difference gives a nonaccurate prediction that constraint 7 is a nonproximal one.

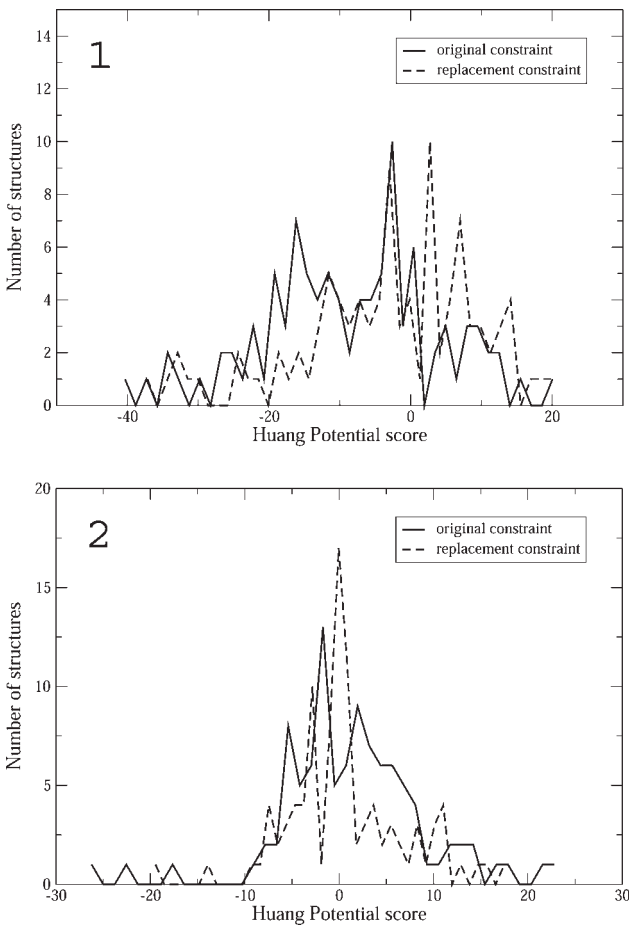

Figure 7

Comparison between Huang potential score distributions of simulated structures created by using proximal constraints of protein 1AF7 family Cher_N. For all graphs, solid lines represent the original constraint and dashed lines the replacement. The simulated structures used for this graph are the same with those used for Figure 6. Graph 1: ARG57-TYR69 constraint pair 3 from Table I. The simulated structures distribution created by using the replacement is shifted to the right, having an average Huang potential difference of 5.53, indicating that the new folds are on average different, compared to the folds of the original simulated structures distribution. Graph 2: ARG29:LEU70 constraint pair 7 from Table I. The two distributions are very similar, having an average Huang potential difference of only 0.52. In this case, the original constraint, although a true positive one is predicted to be a nonproximal one.

Huang potential, and an average Huang score HS for each constraint was obtained.

Average Huang scores FHS of all constraints over all shuffles for all the proteins used are shown in Figure 8. Following the example of our method involving CMA[4] and previous works,[1–3,26,27] an arbitrary positive cutoff value for the Huang score has to be established. Constraints having a value of FHS equal or larger to the cutoff value are predicted to be truly proximal.

It has been found that truly proximal constraints have a final Huang score far greater than nonproximal constraints. Also, low final Huang scores indicate low probability of

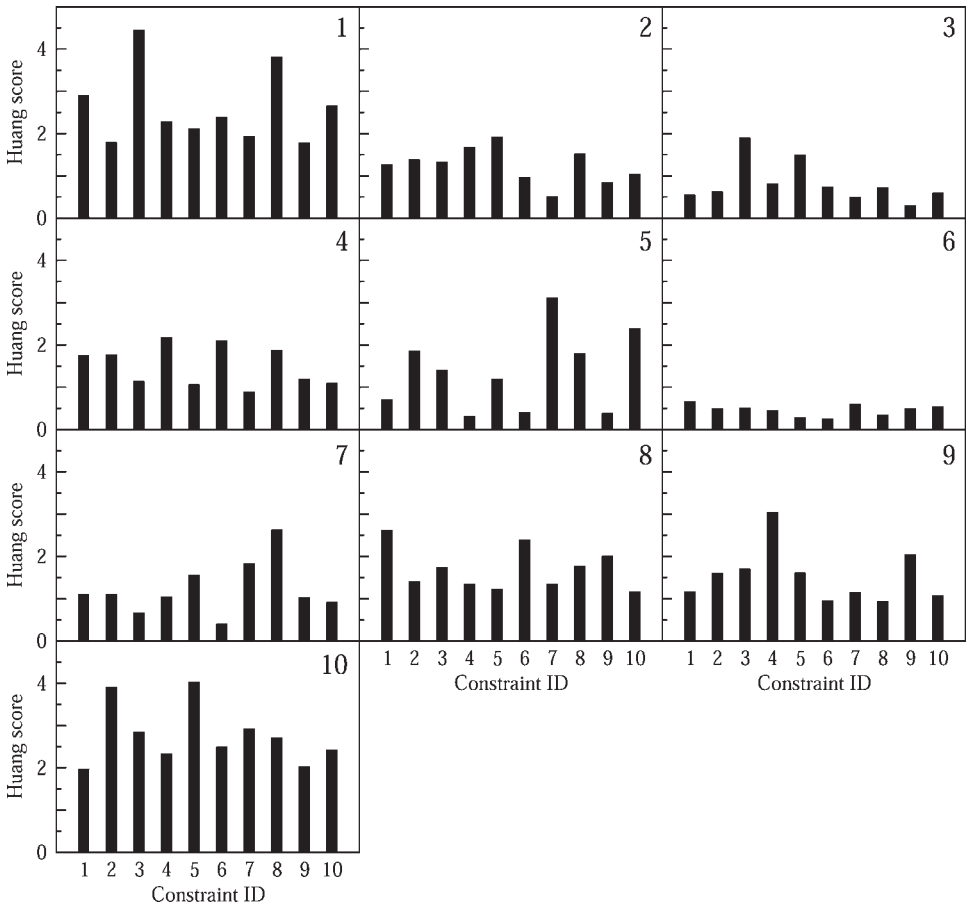

Figure 8

Average Huang scores for the constraints of all tested proteins. The score of each constraint is calculated for each shuffle, and then averaged overall eight shuffles. 1: Cher_N protein 1AF7, 2: RGS protein 1AGR, 3: ATP-synt_DE protein 1AQT, 4: ATP-synt_C protein 1A91, 5: Chorismate_mut protein 1ECM, 6: Acyl_Coa_dh protein 1IVH, 7: Phospholip_A2_1 protein 1BUN, 8: FE_DEP_REPR_C protein 2TDX, 9: Arg_repressor.

proximity. Specifically, investigation of all constraints used has shown that 14 out of the 34 constraints with FHS of 2 and above are truly proximal, showing an accuracy of 41%, and 22 out of the 27 constraints with FHS of 0.7 and lower, are nonproximal, showing an accuracy of 81%. Therefore, we conclude that a cutoff value of 2 for the proximal constraints, and 0.7 for the nonproximal ones, guarantees predictions of very high accuracy. The region however between HS of 0.7 and 2, comprising almost 40% of the constraints, contains both proximal and nonproximal constraints, indicating it as a "gray" region.

It has been observed that small proteins provide much higher values of HS for their tested constraints compared to those of large proteins. Specifically, the smallest proteins in our test set, Cher_n protein 1AF7, and HTH_8 protein 1FIP show all their constraints to have a FHS of 1.8 and above. Therefore, an increase in the threshold for the truly proximal constraints to 2.5 (instead of 2.0) significantly increases accuracy. Following the same empirical approach, we find that if we gradually lower the threshold with sequence size, we observe that the accuracy increases significantly, for both proximal and nonproximal constraints. A lower value of around 1.5 is more appropriate for the proteins of 70 residues and above. Finally, it gets the lowest value of only 0.7 for the 150-residue protein 1IVH of the Acyl_coa_dh family. The results for all tested proteins, using the empirical cutoffs are shown in Table IV.

For example, in Figure 8, protein 1BUN final Huang scores are shown in histogram 7. The truly proximal constraints are those with ID numbers: 1, 7, and 8, while the remaining constraints are nonproximal. Using a threshold value of 1.5 (Table IV), the method identifies constraints 7 and 8 as truly proximal, incorrectly predicts constraint 5 as proximal, fails to identify constraint 1 as proximal, and correctly identifies constraints 2, 3, 4, 6, 9, and 10 as nonproximal. The achieved predictive accuracy for this protein is then 80%.

DOI 10.1002/prot

Table IV
Accuracy for the Protein Families Test Set

Family	Sequence size	Threshold[a]	TP accuracy[b]	FP accuracy[c]	Accuracy[d] (%)
HTH_8	41	2.5	1/5	2/5	30
ATP-synt_DE	45	2.5	0/0	7/10	70
Cher_N	57	2.5	2/4	5/6	70
ATP-synt_C	70	1.5	1/5	4/5	50
FE_DEP_REPR_C	71	1.5	2/6	3/4	50
Arg_Repressor	71	1.5	2/5	4/5	60
Chorismate_mut	85	1.5	0/4	6/6	60
Phospholipid_A2_1	119	1.5	2/3	6/7	80
RGS	119	1.5	3/4	5/6	80
Acyl_coa_dh	150	0.7	0/0	7/10	70

[a]The average Huang score threshold used to predict proximal residues from the nonproximal ones.
[b]The denominator is the total number of constraints predicted to be proximal, and the nominator is the actual proximal ones.
[c]The denominator is the total number of constraints predicted to be nonproximal, and the nominator is the actual nonproximal ones.
[d]Total accuracy is the total amount of predictions, both proximal and nonproximal ones to be correct.

From Table IV, we conclude that overall the method gives very good predictions for most protein families tested. Only HTH_8 family gives very poor results, but we should mention that protein 1FIP belonging to HTH_8 family is the smallest protein tested, with the simplest of folds. Out of 100 constraints from all tested proteins, 13 out of 28 truly proximal constraints and 49 out of 72 nonproximal constraints were predicted correctly, giving an overall accuracy of 62%. Concerning the true proximal predictions, from Table IV, we see that from a total of 36 proximal pairs predictions, 13 were correctly identified, giving a proximal constraint accuracy of 36%. Similarly from the total of 64 nonproximal pairs predictions, 49 of them were correctly identified, giving a nonproximal constraint accuracy of 76.5%. Defining random accuracy for the entire constraint set as the fraction of the truly proximal constraints, we have an improvement over random for the proximal and the nonproximal constraints by a factor of 1.35 and 1.09, respectively. The value for random nonproximal constraint accuracy of 72% was already very large, therefore a small improvement was expected.

CONCLUSIONS

Metropolis Monte Carlo simulations subject to residue constraints can be used to distinguish the truly proximal residue pairs from a set of distant residue contact predictions. The method is relatively easy to implement. The computational cost however is not insignificant. For one protein studied, 10,400 protein structures were generated to evaluate 10 constraints. Also, for larger protein sequences, the computational cost becomes prohibitively large, unless extraordinary computational resources are expended.

The method can be applied to any set of predicted pairs, regardless of the original algorithm used to predict them in the first place. The results show that the method distinguishes well both the true positive and the false positive ones. According to results, the method is very specific, since it predicts almost four out of five nonproximal pairs. As far as the truly proximal pairs concerned, the method has a very good sensitivity, although not as high as that for the nonproximal ones, identifying about one out of three proximal pairs. Overall accuracy is 64%. We find that a Huang potential threshold value of 2.5 should be used for small proteins, and 1.5 for larger ones.

We have not studied proteins containing β strands and β sheets, because it is difficult to simulate β strands cooperativity into parallel or antiparallel sheets, using ab initio methods. The simulated structures of main β proteins we obtained by using Monte Carlo failed to show any indication of aligning the β strands into sheets in a way similar to that of the native structure. Therefore, simple Monte Carlo simulations of around 10^5 moves are certainly not adequate to generate meaningful β sheet protein ensembles. Further studies have to be made to study main β class proteins, perhaps with additional constraints for β sheets and strands, and this is not the focus of this work.

Protein 1IVH is the largest protein used in our test set and it has certain characteristics, which make it very difficult to simulate, compared to the other proteins. 1IVH contains two α helices consisting of 35 and 32 residues each, the largest helices used in the protein test set. This is twice the size of a helix tested compared to the next largest protein tested, protein 1ARG, and three times the size of medium size proteins such as 2TDX and 1ECM. The method in this case fails to achieve proper packing, and its phase space is not adequately sampled. Further investigation needs to be conducted for larger size proteins. The number of constrains and their strength should also be addressed for these cases.

An important remark concerning this work is that the results were obtained by assuming that we have accurate secondary structure prediction information. Since the

native structure was known, sufficiently accurate secondary information directly taken from PDB databank[7] has been used during the simulations. Also, test simulations have been performed by using slightly altered secondary structure information, up to two residues different from each major secondary structure such as helices, strands and coils (for example a helix consisting of residues 6–12 would also include residues 4, 5, and 13 in the test simulation), and it has been found that the results are nearly identical (results not shown). However, no simulations have been performed with the use of secondary structure information very different from the native structure (for example an α helix instead of an amorphous coil). Therefore, the influence of greatly misplaced and wrongly predicted secondary structure information has not been studied in this work.

The method is very general and can be applied to any proximal residue prediction set of any mainly alpha protein domain lower than 150 residues. In future work, we expect that by creating an automated and optimized version of our algorithm, and with more computational power in our hands, we can conduct studies between different families and protein architectures, and create very accurate residue contact maps, by using a very large pair prediction set for each tested protein. In this work, we concentrated on predicted pairs, which had the highest probability according to CMA to be proximal. Now that we have established the basic rules for our search, our future goal is to go into lower CMA correlation coefficient thresholds, create predictions test sets with lower probability of proximity, and test our method with these constraints. High quality contact map is our final goal.

Finally, when we obtain a highly accurate set of proximal residue pairs, we can address the important question of the influence of correct residue constraints into protein folding simulations.

ACKNOWLEDGMENTS

This work has been supported by the American Chemical Society Petroleum Research Grand (Award no. G7-38758). We also want to thank Dr. Yuk Sham and Dr. Patton Fast for their support, as well as the Minnesota Supercomputing Institute. This work was partially supported by the National Center for Supercomputing Applications under Grant TG-MCA04N033 and utilized by the NCSA Mercury Cluster.

REFERENCES

1. Goebel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18:309–317.
2. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Des 1997;2:S25–S32.
3. Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of contact maps with neural networks and correlated mutations. Protein Eng 2001;14:835–843.
4. Vicatos S, Reddy BV, Kaznessis Y. Prediction of distant residue contacts with the use of evolutionary information. Proteins 2005;58:935–949.
5. Fariselli P, Olmea O, Valencia A, Casadio R. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. Proteins 2002 (Suppl. 5):157–162.
6. Singer MS, Vriend G, Bywater RP. Prediction of protein residue contacts with a PDB-derived likelihood matrix. Protein Eng 2002; 15:721–725.
7. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Jr., Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
8. Grana O, Baker D, MacCallum RM, Meiler J, Punta M, Rost B, Tress ML, Valencia A. CASP6 assessment of contact prediction. Proteins 2005;61(Suppl. 7):214–224.
9. Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. Proteins 1995;23:ii–v.
10. MacCallum RM. Striped sheets and protein contact prediction. Bioinformatics 2004;20(Suppl. 1):I224–I231.
11. Punta M, Rost B. PROFcon: novel prediction of long-range contacts. Bioinformatics 2005;21:2960–2968.
12. Vasquez M, Nemethy G, Scheraga HA. Conformational energy calculations on polypeptides and proteins. Chem Rev 1994;94:2183–2239.
13. Frauenfelder H, Sligar SG, Wolynes PG. The energy landscapes and motions of proteins. Science 1991;254:1598–1603.
14. Kidera A. Enhanced conformational sampling in Monte Carlo simulations of proteins: application to a constrained peptide. Proc Natl Acad Sci USA 1995;92:9886–9889.
15. Ortiz AR, Kolinski A, Skolnick J. Tertiary structure prediction of the KIX domain of CBP using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. Proteins 1998;30:287–294.
16. Fariselli P, Casadio R. A neural network based predictor of residue contacts in proteins. Protein Eng 1999;12:15–21.
17. Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of contact maps with neural networks and correlated mutations. Protein Eng 2001;14:835–843.
18. Thomas DJ, Casari G, Sander C. The prediction of protein contacts from multiple sequence alignments. Protein Eng 1996;9:941–948.
19. Karlin S, Zuker M, Brocchieri L. Measuring residue association in protein structures possible implications for protein folding. J Mol Biol 1994;239:227–248.
20. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 1997;28:405–420.
21. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. Ab initio folding of proteins using restraints derived from evolutionary information. Proteins 1999 (Suppl. 3):177–185.
22. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4:187–217.
23. Crivelli S, Kreylos O, Hamann B, Max N, Bethel W. ProteinShop: a tool for interactive protein manipulation and steering. J Comput-Aided Mol Des 2004;18:271–285.
24. Huang ES, Subbiah S, Levitt M. Recognizing native folds by the arrangement of hydrophobic and polar residues. J Mol Biol 1995; 252:709–720.
25. Huang ES, Subbiah S, Tsai J, Levitt M. Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. J Mol Biol 1996;257:716–725.
26. Ortiz AR, Kolinski A, Skolnick J. Nativelike topology assembly of small proteins using predicted restraints in Monte Carlo folding simulations. Proc Natl Acad Sci USA 1998;95:1020–1025.
27. Ortiz AR, Kolinski A, Skolnick J. Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. J Mol Biol 1998;277:419–448.

 DOI 10.1002/prot