An On/Off Lattice Approach to Protein Structure Prediction from Contact Maps

Stefano Teso, Cristina Di Risio, Andrea Passerini, and Roberto Battiti

Dipartimento di Ingegneria e Scienza dell'Informazione
Università degli Studi di Trento, Italy
{teso,dirisio,passerini,battiti}@disi.unitn.it

Abstract. An important unsolved problem in structural bioinformatics is that of protein structure prediction (PSP), the reconstruction of a biologically plausible three-dimensional structure for a given protein given only its amino acid sequence. The PSP problem is of enormous interest, because the function of proteins is a direct consequence of their three-dimensional structure. Approaches to solve the PSP use protein models that range from very realistic (all-atom) to very simple (on a lattice). Finer representations usually generate better candidate structures, but are computationally more costly than the simpler on-lattice ones. In this work we propose a combined approach that makes use of a simple and fast lattice protein structure prediction algorithm, REMC-HPPFP, to compute a number of coarse candidate structures. These are later refined by 3Distill, an off-lattice, residue-level protein structure predictor. We prove that the lattice algorithm is able to bootstrap 3Distill, which consequently converges much faster, allowing for shorter execution times without noticeably degrading the quality of the predictions. This novel method allows us to generate a large set of decoys of quality comparable to those computed by the off-lattice method alone, but using a fraction of the computations. As a result, our method could be used to build large databases of predicted decoys for analysis, or for selecting the best candidate structures through reranking techniques. Furthermore our method is generic, in that it can be applied to other algorithms than 3Distill.

Keywords: Protein Structure Prediction, HP model, Contact Maps, Simulated Annealing, Replica Exchange Monte Carlo.

1 Introduction

Protein structure prediction (PSP) is the problem of inferring the tertiary structure of proteins given only information on their primary structure. This problem is of the highest importance for several reasons: the function of a protein is strictly tied to its three-dimensional structure, but the experimental determination of the tertiary structure is still a complex, time consuming and expensive process. In addition, in some cases it is impossible to obtain structural information with experimental techniques: many proteins are too large for NMR analysis and some classes of proteins such as membrane ones are very difficult

T.M.H. Dijkstra et al. (Eds.): PRIB 2010, LNBI 6282, pp. 368–379, 2010.

to crystallize for X-ray diffraction [4]. As a matter of fact, most of the known protein sequences are not yet assigned a corresponding structure: in spite of the long-standing community-wide effort, most known proteins still lack a resolved structure, and of the nearly two million protein sequences currently known, fewer than 2% have an associated structure.

Following the idea that similar sequences are bound to represent similar structures [6], at least at a local level, comparative modeling methods have been developed which exploit local homology information to compute the structure of novel query protein sequences. Remote homology techniques are employed for fold recognition when sequence conservation is lacking. However, when no close or remote homologues are available, these methods cannot be applied, and *de novo* structure prediction must be performed.

The current methods for *de novo* PSP can be split in roughly three groups. A first group accounts for all-atom molecular simulation methods, which try to mimic the physical folding process starting from first principles. They have huge computational requirements and have not been very successful for realistically sized proteins. A second group includes all those methods that search the space of atom- or residue-level conformations for a native-like fold using some more or less empirical energy function to assess the quality of the candidate structures. Meta-heuristic optimization algorithms are usually employed to perform the search. These methods have had much more success than the ones in the previous class, but they are still very computationally expensive. Finally, a third group includes methods that rely on residue-level (or coarser) structure representations and enforce them to lie on a regular lattice, embedded in purely synthetic force fields. Methods in this group can find a native-like decoy with relatively less computational effort than methods in the former groups, but the resulting structures are not as realistic. They are typically used as tools to analyze the statistical properties of the folding landscape, rather than to generate reliable structures.

In this work we present a novel method that combines the complementary strengths of off-lattice empirical models and on-lattice ones, and allows to generate a large number of comparatively good quality decoys with a fraction of the computational power required by standard methods. The underlying idea is to combine two existing *de novo* PSP algorithms: a modified version of REMC-HPPFP [15], a fast prediction method based on a very coarse structure representation, which is used to compute a first set of rough decoys; 3Distill [1], a more realistic method that uses a finer structure representation, which is employed to refine the decoys generated by REMC-HPPFP. Albeit based on a simple idea, we prove that our method is indeed able to combine the features of REMC-HPPFP and 3Distill: it generates competitive structures with much less effort. Furthermore, the idea underlying our method is generic, meaning that it is in no way restricted to 3Distill, and may prove useful to improve the efficiency of other fine-grained structure prediction algorithms.

This paper is structured as follows. In Section 2 we review some of the relevant methods for the protein structure problem. In Section 3 we describe our combined

PSP approach and the two methods on which it is based. In Section 4 we describe the experiments carried out to benchmark our method and compare it to the baseline. In Section 5 we discuss the results of the experiments and show that our method is ultimately successful in reducing the amount of computation required. Finally, in Section 6 we draw the conclusions on this work and describe some future research directions.

2 Related Work

De novo PSP methods include both off-lattice and on-lattice models and methods. In off-lattice models, the residues are free to be placed at arbitrary continuous coordinates in the three-dimensional Euclidean space. The simplest way to represent residues is as hard spheres of fixed radius centered on the C_α atom, but other more complex representations are available as well. In other, intermediate models, all the atoms of the backbone are modeled, but the side chain is represented as a hard sphere centered at the center of mass of the real side chain. It has been noted however that the lower computational demands of coarse-grained models does not necessarily come at the cost of inferior expressiveness [10].

In on-lattice models the protein conformation is restricted, such that each residue occupies a different vertex on a lattice. Consecutive residues in the primary structure are placed at adjacent positions, and the protein chain becomes itself a self-avoiding path on the lattice. Lattice models employ a variety of two- and three-dimensional lattice: square, triangular, cubic, face-centered cubic, diamond, and others with very high degrees of freedom. For the representational power of different common and less-common lattices, we refer the reader to [11]. On-lattice models have been chiefly used as tools for studying protein folding, because the simplified representation allows for an easier mathematical treatment [10].

Common approaches to the PSP problem include the aggregation of short structural fragments, for instance Rosetta [12], and the use of contact maps [16,1]. We focus on the latter approach. The idea is to split the prediction task into two simpler sub-tasks: first generate *de novo* an accurate, residue-by-residue contact map from the protein sequence, and then reconstruct the protein structure from the contact map. This is a sound approach, as contact maps can be shown to encode the same information as the structure they represent [16]. To date, a few contact map predictors have been proposed: SVMcon [2], Xxstout [1], and NNcon [14] among others. As for the reconstruction process itself, a popular approach is to use some form of stochastic optimization, as in the seminal paper by Vendruscolo et al. [16] and 3Distill [1].

3 Method

Our proposed method is based on two well known existing *de novo* PSP algorithms: in the next couple of sections we will introduce them and explain their pros and weaknesses.

3.1 3Distill

An often advocated approach to the PSP is to split the main *de novo* structure prediction problem into a set of simpler prediction tasks. Distill [1] is a hierarchy of state-of-the-art prediction servers that follows this approach. The Distill servers compute a number of one-dimensional features (such as secondary structure, solvent accessibility, and contact density) and two-dimensional features (such as fine and coarse contact maps, coarse protein topology). The main idea is that all servers make use of features predicted in the lower levels of the hierarchy, starting from the primary structure, to predict more complex features.

At the top of the hierarchy, the 3Distill server computes the protein tertiary structure, as a residue-level C_α trace, given predicted features from all the other servers. A preliminary implementation of 3Distill took part to the CASP 6 competition [8] and was ranked among the best 20 predictors out of 181 on Novel Fold hard targets and Near Novel Fold targets. 3Distill was chosen because it is simple and relatively fast when compared to other *de novo* algorithms.

The main feature input into 3Distill is a predicted (multi-class) contact map, which specifies a set of soft physical constraints for all pairwise inter-residue distances. For a detailed description of contact maps, see [16]. Other input features include a predicted per-residue secondary structure and a predicted coarse-grained contact maps, which defines the appropriate distances between pairs of secondary structure elements. To avoid the computational burden of all-atom models, 3Distill relies on a reduced backbone-only protein model. Furthermore, residues that are predicted to belong to an α-helix are modeled as rigid, ideal helices. This solves the problem of folding the helices during the optimization stage, and decreases the complexity of the conformational search. To mimic the minimal observed distance between atoms of different amino acids, the volume of each C_α is modeled as a hard sphere of radius 5.0 $\mathring{A}$, and the distance between consecutive residues is set to 3.8 $\mathring{A}$. These values were rigorously inferred from statistical analysis of real world data [1].

All candidate conformations have an associated pseudo-potential that is defined in terms of the input contact maps and secondary structure. The energy of a conformation estimates how much it violates the constraints imposed by the given fine and coarse contact maps, while at the same time penalizing non-physical configurations (i.e., overlapping or too far away residues). For an in detph description, see [1].

The mechanism used by 3Distill to search for the native conformation is Simulated Annealing (SA) [7]. SA is an iterative procedure: starting from a random candidate structure, at each iteration it perturbates the structure producing another candidate configuration. The newly generated configuration replaces the old one if it is better (has a lower energy), with probability one; or if it is worse (higher energy) with a probability that depends on the magnitude of the energy difference. This second condition is controlled by a so called temperature parameter: when the temperature is high, even very bad configurations have a high probability of having accepted; when it is low, almost all worsening configurations are rejected. In 3Distill the temperature decreases linearly with the

number of iterations, meaning that as the search proceeds the temperature moves towards zero and the probability of accepting worsening moves goes to zero as well. For further details, we refer the reader to [7].

In 3Distill, each iteration of SA traverses the whole structure, perturbating each residue in the order in which it appears in the protein chain. A perturbation amounts to displacing a residue according to the following rules: (1) If the residue is neither an endpoint nor in a helix, it is rotated by a random angle around the segment joining its two neighboring residues. (2) If the residue is an endpoint of the chain and not part of a helix, it is rotated at random around its only neighbor. (3) If the residue is part of a helix, the whole helix is rotated at random. This set of moves guarantees 3Distill to efficiently explore the conformational space. We note that each traversal of the protein structure amounts to h perturbations, where h is the overall number of free residues (not in a helix) and helices. The SA algorithm stops after a given amount of traversals.

3.2 REMC-HPPFP

The Hydrophobic-Polar model (HP model for short) [3] is a very basic model of protein folding based on a reduced, residue-level representation of the tertiary structure. In this model, proteins are represented as backbone-only configurations and the residues are forced to lie on a regular, typically cubical lattice, with no overlap. In the HP model, each residue is either hydrophobic (H) or polar (P). The HP model is designed to capture the fact that folding is mainly driven by hydrophobic interactions between the residues. Following this idea, the energy of a configuration $\mathbf{x}$ is defined empirically in terms of neighboring residues: two residues are called *topological neighbors* if they are not consecutive in the protein sequence and share an edge of the lattice. The energy associated to an HP configuration is the negated number of topological neighbors that are both hydrophobic. In other words, this energy function favors those configurations containing a densely packed core of hydrophobic residues. Solving an HP problem instance involves finding the native conformation, that is, the structure having the lowest possible associated energy.

Despite its simplicity, the HP model has been proven to be NP-complete in both two and three dimensions on the cubic lattice [7, 14], and NP-hard on a general lattice [21], including the face-centered cubic and triangular lattices. For this reason, HP model solvers usually resort to heuristic optimization algorithms to search the conformational space. REMC-HPPFP [15] is one of the state-of-the-art solvers of square and cubic lattice HP instances. It makes use of a very effective stochastic search procedure, named Replica Exchange Monte Carlo (REMC for short) that is especially geared towards high-dimensional optimization problems. REMC-HPPFP has been shown to lead to superior results with respect to competing methods, such as PERM [5] and ACO-HPPFP-3 [13] in a set of synthetic and on biologically-derived benchmark instances [15]. The core features are the REMC optimization heuristic and the set of moves used to perform the search itself. We briefly discuss them in the following, see [15] for details.

The REMC search heuristic is reminiscent of Simulated Annealing, in that a candidate protein structure is perturbated at each iteration, by applying a random move, and the resulting structure is accepted or rejected depending on the energy delta with respect to the old configuration. However in this case, multiple configurations, called replicas, are optimized concurrently. Each configuration has its own fixed temperature, which does *not* decrease with time. Replicas are indexed from 1 to m, and the temperature of each replica is a monotonically increasing function of its index. Once every k iterations, with k a fixed parameter, the energy of adjacent replicas is compared, and if certain energy conditions are met, the two replicas are exchanged, meaning that the ith replica will become the $(i+1)$th and vice versa. This way the replicas change temperature based on their energy level. The set of moves used by REMC-HPPFP to perturbate the candidate configurations comprises a set of standard residue by residue moves, termed VHSD moves, and the non-standard pull move [9]. This set of moves is the most complete and efficient set of moves available to date for the HP model on the square and cubic lattices.

3.3 On/Off Lattice Cascade

The main issue with 3Distill is that, even being one of the simplest *de novo* predictors proposed, the conformational space is huge and requires a large amount of computational power to find low energy configurations. This is a common problem for all fine-grained structure predictors. On the other hand, the REMC-HPPFP algorithm shows very good performances on HP instances. Our primary aim in this work is to combine the efficiency of on-lattice methods with the accuracy of off-lattice models. We do so by first using a suitably modified version of REMC-HPPFP to quickly produce a candidate on-lattice structure that (partially) satisfies a given residue-level contact map, and then refining the obtained structure by using 3Distill with the same contact map. The intermediate lattice structures generated by the modified REMC-HPPFP can be thought as bootstrapping 3Distill, by making it start its search from more favorable regions of the search space.

To obtain the best results from the cooperation of REMC-HPPFP and 3Distill, we had to implement a new lattice energy function. The new function defines the fitness of a configuration in terms of how much it satisfies a given multi-class contact map. The formal definition is as follows:

$$E(\mathbf{x}; C, p, k) = \sum_{i,j} E(d_{ij}; c_{ij}, p, k)$$

$$E(d_{ij}; c_{ij}, p, k) = \begin{cases} |d_{ij} - \tau_c|^p & \text{if } d_{ij} < \tau_c \\ |d_{ij} - \tau_{c+1}|^p & \text{if } d_{ij} > \tau_{c+1} \\ -k & \text{otherwise} \end{cases}$$

where $\mathbf{x}$ is a candidate protein structure, $C = [c_{ij}]$ is a multi-class contact map, with each class c having range $[\tau_c, \tau_{c+1}]$, and d_{ij} is the Euclidean distance between residues i and j. The pairwise energy potential is a polynomial of the

difference between the actual distance between residues i and j and the closest threshold of the predicted contact class. The two constants p and k are parameters used to adjust the energy function to the data at hand. In particular, k defines the net gain for a satisfied contact, and p controls the amount of penalty for an unsatisfied contact. In this new model, structures lie on a cubic three-dimensional lattice of fixed side 3.8 $\mathring{A}$, the same as the default inter-residue distance for 3Distill.

To summarize, our method consists of a modified REMC-HPPFP version that, by virtue of a new energy function, is able to find on-lattice configurations that best satisfy a given residue-level contact map. Aside from the new energy function, the REMC-HPPFP algorithm is unchanged. This novel method is used to generate one or more lattice configurations, which are then refined with 3Distill; both algorithms use the same predicted contact map. All in all, the new cascade method requires four additional parameters to be specified: p and k shape the energy function, the other two are T_1 and T_2, the number of iterations to run the on-lattice and off-lattice algorithms for, respectively.

To allow for a common measurement unit of computation, we define the concept of *big iteration* as a complete traversal of the protein structure by the search algorithm. For 3Distill a big iteration involves h structure perturbations, each requiring to compute the value of the energy function for the newly generated configuration, for a total of $h = O(n)$ energy computations. For REMC-HPPFP, a big iteration equates to $n \times m$ structure perturbations, where m is the number of replicas, again amounting to $O(mn) = O(n)$ energy updates. The computational complexity of the two algorithms is thus $O(n^2)$ per big iteration, as both require $O(n)$ pseudo-instructions for each energy function evaluation.

4 Experiments

The goal of the experiments is to assess the ability of our combined method to generate decoys of quality comparable to that of the original 3Distill algorithm, and to evaluate the amount of computation required to attain such decoys. The quality of the decoys is defined in terms of the TM-score [17] to the experimentally determined native fold. TM-score values range in $[0, 1]$, with all values larger than 0.4 suggesting a topologically correct prediction, and for all scores above 0.7 a good structural superposition between the predicted and the native folds.

The tests are based on a dataset of 171 proteins with no detected homology, with length between 50 and 200 residues. The contact maps were predicted by Xxstout [1] with threshold values $\tau_1 = 8$Å, $\tau_2 = 13$Å, and $\tau_3 = 19$Å using a recursive neural network while exploiting evolutionary information in the form of multiple alignment profiles, plus the contact map of the nearest template when available. All template-matching qualities and all relevant SCOP classes (all α, all β, α/β, $\alpha+\beta$, coiled-coil, and small) are represented in this data set. The data was kindly provided by the Distill team.

4.1 Selection of the Lattice Energy Function

The goal of the first batch of experiments was to tune the parameters p and k of the new lattice energy function to maximize the TM-score of the resulting decoys as expected. During previous experiments we observed that the quality of 3Distill results is positively correlated to the TM-score of the inputs structures, and the same holds for its convergence. The dataset is varied enough to guarantee that the parameters p^* and k^* found are generalizable to other data sets. During these experiments, for simplicity we kept the other parameters of the modified REMC-HPPFP fixed to values used in the original paper for the three-dimensional lattice [15]. In particular, the number of replicas is two.

For this set of experiments, we sampled the performance of the modified REMC-HPPFP for (p, k) values taken from a grid in the (p, k) parameter space. A preliminary set of runs was performed on a small subset of protein instances to determine the extents of the grid, for a total of 10 structures for each proteins, 100 iterations each. We found some reasonable values to be $p \in [0.25, 2.25]$ (at increments of 0.50) and $k = \{0, 10, 100, 1000\}$. Outside this range, the performance of our method degraded quickly. The grid itself is uniform in the p dimension and exponential in k: the reason is that p appears as an exponent in the energy function, while k is an additive linear term. Next we performed a thorough exhaustive search: for each (p, k) value in the grid, now with p increments of 0.25, we ran our modified REMC-HPPFP on all proteins in the dataset, 100 runs per protein, 100 iterations per run, and compared the average TM-score of the generated decoys. Using this method, the best parameters were found to be $p^* = 1.75$ and $k^* = 0$.

4.2 Behavior over Time

Given the optimal values p^* and k^*, we evaluated the number of big iterations (T_1, T_2) that our method needs to obtain results comparable to those of 3Distill alone. To compare the performance of our combined approach to 3Distill, we use the ratio between the TM-score reached by our algorithm and the best TM-score obtained by 3Distill alone. We defined a uniform grid in the (T_1, T_2) parameter space. The upper bound for $T_2 < 5000$ was determined experimentally by observing the number of big iterations needed to achieve pseudo-convergence with 3Distill. For T_1 we just used the same number of iterations defined in the original paper, $T_1 < 100$.

In all the runs, the lattice algorithm was run with the same parameters as in the previous set of experiments, together with the newly found p and k. The parameters of 3Distill were setup as in [1]. We ran the combined algorithm for all proteins in the dataset, 100 runs for each protein, with $T_1 < 100$ and $T_2 < 5000$, recording the intermediate candidate structures during the optimization procedures, so to properly fill in the (T_1, T_2) grid.

For every protein and (T_1, T_2) pair, we computed the average TM-score of the predicted folds and normalized it with respect to the average TM-score of the structures for the same protein found at $(T_1, T_2) = (0, 5000)$. We call this

quantity the "quality ratio", i.e., the ratio between the TM-score for proteins found by our method using (T_1, T_2) iterations, and the TM-score of the structures predicted by 3Distill. Then for each point in the (T_1, T_2) grid we computed the average of the structure quality ratio over all decoys and over all proteins.

We note that the number of energy evaluations per big iteration in 3Distill is equal to the number of control points h, whereas for REMC-HPPFP it is equal to the number of residues n for each replica. Despite being both asymptotically $O(n)$, in practice these two quantities are not identical. This makes it difficult to experimentally compare the values of T_1 and T_2, because h is a structural property depending on the predicted protein secondary structure. Hence h may be different between proteins of the same size. To account for this fact, we split the results by protein length in 3 different classes, with ranges from 50 to 200: the first class contains proteins of length from 50 to 99, the second those of length from 100 to 149, the third those of length from 150 to 200. For each class we computed the average number of hinges $\hat{h}$ and the average number of residues $\hat{n}$, and rescaled the T_2 axis by $\hat{n}/\hat{h}$. This results in 3 grids, shown in Figure 1.

5 Discussion

The main result of this paper is that, in all the plots, the combined algorithm is shown to be able to produce structures of quality comparable to that of 3Distill alone, but with a far smaller number of energy evaluations. Multiple combinations of T_1, T_2 show this behavior. Generally, it can be observed that: (a) To obtain structures of quality ratio at least 0.7, that is, structures whose quality is comparable to that of structures found by a full run ($T_2 = 5000$) of the costly off-lattice algorithm, it is sufficient to use $T_1 = 100$ and $T_2 \leq 500$. This amounts to one about tenth of the energy evaluations. (b) To obtain structures of quality ratio at least 0.9, that is, structures whose quality is indistinguishable from that of structures found with $(T_1 = 0, T_2 = 5000)$, roughly 100 on-lattice iterations followed by 2000 off-lattice iterations are sufficient. This amounts to less than one half of the energy evaluations. Thus employing an on-lattice search strategy to obtain initial candidate configurations actually improves the search speed of the off-lattice algorithm.

One surprising result, implicit in the previous discussion, is that the on-lattice algorithm can generate structures of good quality, with respect to those found by the off-lattice method. This can be seen by observing the curves at all grid points with $T_2 = 0$. It follows that despite its simplicity, the cubic lattice, when paired with our contact map driven energy function, is able to model topologically correct, even if coarse, decoys. This seem to support the idea that the on-lattice algorithm is able to bootstrap 3Distill in a region of the search space that contains native-like folds.

Finally, the plots show that the quality ratios reported at the curves with T_1=fixed improve monotonically with respect to T_1. This means that allowing for increasing amounts of on-lattice search, and consequently for better initial candidates to the off-lattice algorithm, helps the latter. This proves that it is

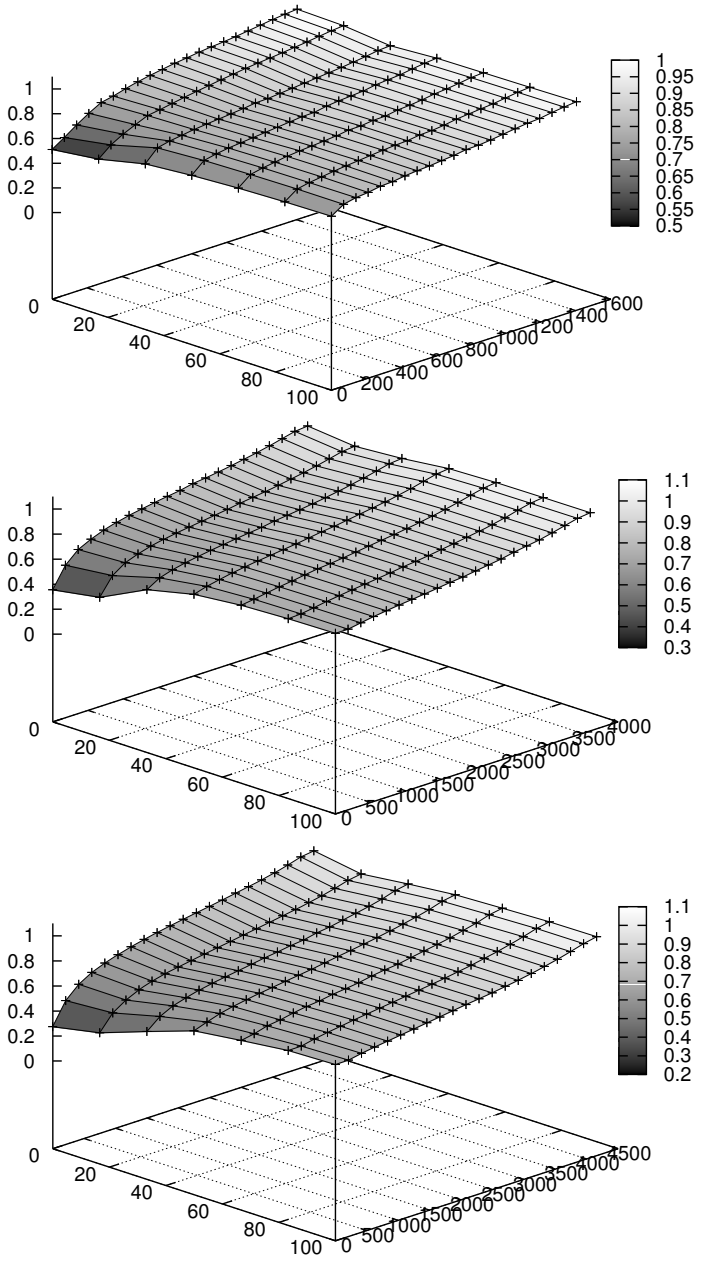

Fig. 1. Each plot represents the behavior over time of our combined method. The axes represent T_1 and T_2 and the height of each point represents the average solution quality ratio (over all decoys and all proteins in the dataset) described in Section 4.2. The upper plot refers to proteins of length 50 to 99 residues; the middle plot to proteins of length 100-149; the last one to proteins of length 150-200.

the on-lattice to be ultimately responsible for enhancing the convergence speed of 3Distill, and not some random external factor such as a different distribution of the initial configurations. The curves with T_2=fixed instead appear to reach convergence at $T_1 = 100$. This validates our choice of $T_1 \leq 100$, and shows that increasing its value would not improve the performance of the lattice algorithm any further.

We note, however, that running the combined algorithm with both T_1 and T_2 set to the maximum values does not significantly improve upon the solutions found by the off-lattice algorithm alone. A possible explanation is that, simply, 3Distill has already reached convergence and that it would be unable to do better than it actually is even when initialized with a good candidate structure.

Summarizing, the above results show that our novel combined on/off lattice approach to protein structure prediction indeed requires potentially fewer energy evaluations to generate good quality, low energy decoys for proteins of length less than 200 residues. This enables for reduced execution time and an increased throughput of structure prediction whenever a contact map is given. The key point is that the resulting pool of structures will probably contain some native-like folds. Ultimately, the higher throughput of our method can serve two purposes: firstly, producing a large population of decoys for statistical analysis; and secondly, to apply reranking techniques with an improved likelihood of finding native-like structures. The ranking approach is very interesting, because it is possible to tune our algorithm with small $(T1, T2)$ values and be able to select very good decoys with small computational effort.

6 Conclusions

In this work we presented a method that combines two existing state-of-the-art approaches to the Protein Structure Prediction problem in a novel way by exploiting the complementary strengths of the two. In particular, a lattice algorithm is used to quickly construct a number of coarse, yet relatively good quality, decoys from predicted contact maps; an off-lattice algorithm is later employed to refine the search. Thanks to the lower number of degrees of freedom, the on-lattice search effectively acts as a bootstrapping step for 3Distill, which converges much faster since the starting candidate conformation is already located in a favorable region of the search space. We proved experimentally that the proposed method allows to generate structures of quality comparable to those generated by 3Distill alone with a fraction of the computational effort. The improvement amounts to one order of magnitude less evaluations of the energy potential, which is the most computationally intensive part of most search algorithms. We stress that our approach is not restricted to 3Distill at all, and that other fine-grained *de novo* algorithms could benefit from it as well. The proposed method potentially allows to build large databases of decoys for analysis or for the later application of reranking techniques to determine the most plausible native folds.

Acknowledgments

The authors would like to thank Gianluca Pollastri for sharing the 3Distill source code.

References

1. Baú, D., Martin, A.J.M., Mooney, C., Vullo, A., Walsh, I., Pollastri, G.: Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins. BMC bioinformatics 7(1), 402 (2006)
2. Cheng, J., Baldi, P.: Improved residue contact prediction using support vector machines and a large feature set. BMC bioinformatics 8(1), 113 (2007)
3. Dill, K.A.: Theory for the folding and stability of globular proteins. Biochemistry 24(6), 1501–1509 (1985)
4. Garavito, R.M., Picot, D., Loll, P.J.: Strategies for crystallizing membrane proteins. Journal of bioenergetics and biomembranes 28(1), 13–27 (1996)
5. Hsu, H.P., Mehra, V., Nadler, W., Grassberger, P.: Growth-based optimization algorithm for lattice heteropolymers. Physical Review E 68(2), 21113 (2003)
6. Kaczanowski, S., Zielenkiewicz, P.: Why similar protein sequences encode similar three-dimensional structures? Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)
7. Kirkpatrick, S.: Optimization by simulated annealing: Quantitative studies. Journal of Statistical Physics 34(5), 975–986 (1984)
8. Kryshtafovych, A., Venclovas, C., Fidelis, K., Moult, J.: Progress over the first decade of CASP experiments. Proteins: Structure, Function, and Bioinformatics 61(S7), 225–236 (2005)
9. Lesh, N., Mitzenmacher, M., Whitesides, S.: A complete and effective move set for simplified protein folding. In: Proceedings of the seventh annual international conference on Research in computational molecular biology, p. 195. ACM, New York (2003)
10. Oakley, M.T., Barthel, D., Bykov, Y., Garibaldi, J.M., Burke, E.K., Krasnogor, N., Hirst, J.D.: Search strategies in structural bioinformatics. Current Protein and Peptide Science 9(3), 260–274 (2008)
11. Pierri, C.L., De Grassi, A., Turi, A.: Lattices for ab initio protein structure prediction. Proteins: Structure, Function, and Bioinformatics 73(2), 351–361 (2008)
12. Rohl, C.A., Strauss, C.E.M., Misura, K., Baker, D.: Protein structure prediction using Rosetta. Methods in enzymology, 66–93 (2004)
13. Shmygelska, A., Hoos, H.H.: An ant colony optimisation algorithm for the 2 D and 3 D hydrophobic polar protein folding problem. BMC bioinformatics 6(1), 30 (2005)
14. Tegge, A.N., Wang, Z., Eickholt, J., Cheng, J.: NNcon: Improved Protein Contact Map Prediction Using 2D-Recursive Neural Networks. Nucleic Acids Research (May 2009)
15. Thachuk, C., Shmygelska, A., Hoos, H.H.: A replica exchange Monte Carlo algorithm for protein folding in the HP model. BMC bioinformatics 8(1), 342 (2007)
16. Vendruscolo, M., Kussell, E., Domany, E.: Recovery of protein structure from contact maps. Folding and Design 2(5), 295–306 (1997)
17. Zhang, Y., Skolnick, J.: Scoring function for automated assessment of protein structure template quality. PROTEINS-NEW YORK- 68(4), 1020 (2007)