Datalogisk Institut
KØBENHAVNS UNIVERSITET

EBBA: Efficient Branch and Bound Algorithm for Protein Decoy Generation

Martin Paluszewski og Pawel Winter

Technical Report no. 08-08
ISSN: 0107-8283

Dept. of Computer Science
University of Copenhagen • Universitetsparken 1
DK-2100 Copenhagen • Denmark

EBBA: Efficient Branch and Bound Algorithm for Protein Decoy Generation

Martin Paluszewski and Pawel Winter

Department of Computer Science, University of Copenhagen, Universitetsparken 1, 2100 Copenhagen, Denmark

Abstract. We are faced with three major challenges when dealing with the problem of de novo protein structure prediction. One is to determine a suitable energy function having a global minimum near the native structure of the protein. The second challenge is to sample the conformational space such that some of the sampled decoys are near the native structure. The third challenge is to identify the native-like structures among the sampled decoys. Here we present a novel method for decoy generation and therefore attack the second of these challenges.
We propose a new discrete protein structure model (using a modified face-centered cubic lattice). A novel branch and bound algorithm for finding global minimum structures in this model is suggested. The objective energy function is very simple as it depends on the predicted half-sphere exposure numbers of C_α-atoms. Bounding and branching also exploit predicted secondary structures and expected radius of gyration. The algorithm is fast and is able to generate the decoy set in less than 48 hours on all proteins tested.
Despite the simplicity of the model and the energy function, many of the lowest energy structures, using exact measures, are near the native structures (in terms of RMSD). As expected, when using predicted measures, the fraction of good decoys decreases, but in all cases tested, we obtained structures within 6 Å RMSD in a set of low-energy decoys. To the best of our knowledge, this is the first *de novo* branch and bound algorithm for protein decoy generation that only depends on such one-dimensional predictable measures. Another important advantage of the branch and bound approach is that the algorithm searches through the entire conformational space. Contrary to search heuristics, like Monte Carlo simulation or tabu search, the problem of escaping local minima is indirectly solved by the branch and bound algorithm when good lower bounds can be obtained.

1 Background

Here we present our approach for protein decoy generation using the branch and bound paradigm. A shorter version of this paper appeared in [1]. The contact number (CN) is a very simple solvent exposure measure that only depends on the positions of C_α-atoms. Given a fixed backbone structure, the CN of a residue A_i is the number of other C_α-atoms in a sphere of radius r centered at the C_α-atom

of A_i. The CN of all residues of a given structure is called the *CN-vector*. A more information rich measure is called the *half-sphere-exposure* (HSE) measure [2]. Here, the sphere is divided into an upper and a lower hemisphere as illustrated in Figure 1. The up and down numbers of a residue therefore refer to the number of other C_α-atoms in the upper and lower hemispheres respectively. For a given fixed structure, the up and down numbers for all residues is called the *HSE-vector*. CN- and HSE-vectors therefore only depend on the radius of the spheres and the coordinates of C_α-atoms, which is very convenient when using simplified models.

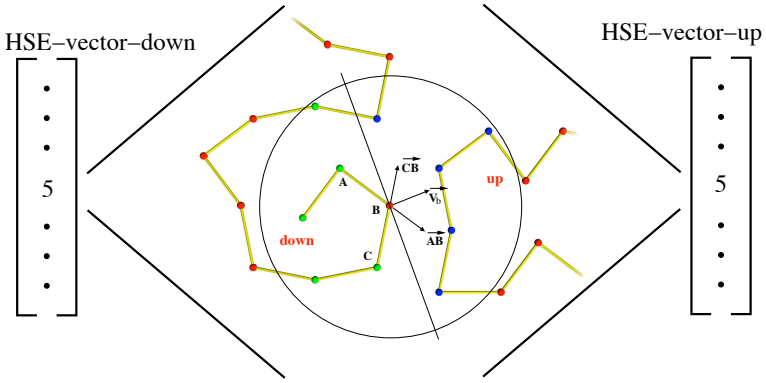


Fig. 1. Given the positions of 3 consecutive C_α-atoms (A, B, C), the approximate side-chain direction $\bar{V}_b$ can be computed as the sum of $\bar{AB}$ and $\bar{CB}$. The plane perpendicular to $\bar{V}_b$ cuts the sphere centered at B in an upper and a lower hemisphere.

Recently it was shown that it is possible to approximately reconstruct small protein structures from CN-vectors or HSE-vectors only [3]. These results showed that HSE-optimized structures in general have better coordinate RMSD with the native structure and more accurate orientations of the side-chains compared to CN-optimized structures. This is very interesting in regards to *de novo* protein structure prediction, because CN- and HSE-vectors can be predicted with reasonable accuracy [4, 5]. To use these results for *de novo* structure prediction, one could therefore first predict the HSE-vector from the amino acid sequence and then reconstruct the protein backbone from this vector. However, the results in [3] were only based on small proteins with up to 35 amino acids and it was conjectured that the reconstruction of larger proteins would require more information than what is contained in an HSE-vector [3]. Another difficulty is that HSE-based energy functions appear to have many local minima in the conformational space. This is often a problem for search heuristics like Monte Carlo simulation or tabu search, since they get trapped in these minima and must spend much time escaping them.

The problem of reconstructing protein structure from vectors of one-dimensional structural information has also been studied by Kinjo et al. [6]. They used exact vectors of secondary structure, CN and *residuewise contact order* (RWCO) together with refinement using the AMBER force field to reconstruct native-like structures. Their results indicated that secondary structure information and CN without the use of RWCO is *not* enough to reconstruct native-like structures. Unfortunately, RWCO is difficult to predict compared to CN, HSE and secondary structure [6] and it would therefore be difficult to use their method directly for de novo structure prediction.

Here we attack these problems by adding more predicted information to our model and use a thorough branch and bound algorithm for finding minimum energy structures. By adding more predicted information we expect to increase the probability of the energy function to have global minimum near the native structure. Furthermore, using a branch and bound approach we are able to implicitly search the whole conformational space and therefore avoid getting trapped in local minima. Besides using HSE vectors, we also use *secondary structure* (SS) and *radius of gyration* (Rg). These three measures, (HSE, SS and Rg), can all be predicted from the amino acid sequence only [4, 7, 8], and can therefore be used for *de novo* protein structure prediction. The energy function is simple, and we show how a good lower bound of the energy for a subset of the conformational space can be computed in polynomial time. This lower bound enables the branch and bound algorithm to bound large conformational subspaces and to find global minimum energy structures in a reasonable amount of time. Throughout the text our branch and bound algorithm is referred to as EBBA (Efficient Branch and Bound Algorithm).

The idea of using secondary structure elements in a discrete model has been suggested by others, i.e., Fain et al. [9] and Levitt et al. [10]. However, their models have a relatively small conformational space and it is therefore possible to completely enumerate all structures allowed by the model. Branch and bound algorithms and other algorithms for determining global minimum structures have been used for protein structure prediction earlier. Some of these algorithms work on very simplified models like the HP-lattice model [11, 12]. Even though these algorithms can solve most problems to optimality, the global minimum structures are often very far from the native structure. Another branch and bound algorithms, called αBB[13] uses more detailed potential energy functions which depend on several physical terms. In [13], the αBB is shown to be successful on small molecules. In [14], the αBB was improved and was used for prediction of real protein structures. Dal Palu et al.[15] use a constraint logic programming approach for protein structure prediction. They also use secondary structure segments in a simplified model. However, in their model, all C_α-atoms must be placed in a lattice (FCC). This differs from our approach, where we only demand lattice directions of the secondary structure segments. Dal Palu et al. use a standard solver (SICStus Prolog) which makes use of standard bounding techniques, while we have developed a much more efficient bounding algorithm specialized for this particular problem. Furthermore, the results published in [15, 14] are not

true *de novo* - the secondary structures are all derived from the native structure of the proteins. On the contrary, the results presented here are true *de novo*. All parts of the energy function are predicted from amino acid sequences only. EBBA is, to our knowledge, the first *de novo* branch and bound algorithm that only use one-dimensional predictable measures.

We use 6 benchmark proteins for evaluating EBBA. Our results show that EBBA is able to find global minimum energy structures for most of these proteins in less than 48 hours. We have evalutated EBBA using both exact values and predicted values to estimate the importance of prediction quality. The results show that predicted structures having global minimum energy are *not always* native-like, however among the 10.000 lowest energy structures we typically find many good decoys (less than 6 Å RMSD). Our algorithm therefore reduces the protein structure prediction problem to the problem of identifying a near-native structure in a relatively small set of decoys.

2 Methods

Each amino acid of a protein can be classified as belonging to a unique secondary structure. Here we consider three classes of secondary structures; helix, sheet and coil. Helices and sheets are distinguished by the unique geometric shape of the C_α atoms in their tertiary structure. Coil is the class of all other shapes that are neither helices nor sheets. C_α-atoms of a coil therefore have a large degree of freedom, compared to helices and sheets, since there are few geometric constraints on the tertiary structure of a coil.

A sequence of residues of the same secondary structure class is called a *segment*. Segments can be considered as rigid rods that describe the overall path of C_α-atoms belonging to the segment. Segments always have a start coordinate and a direction, and for helices and sheets their end coordinate can also be determined because of their constrained geometry. A segment is therefore an abstract representation of a sequence of residues and it does not explicitly contain the coordinates of internal C_α-atoms. We therefore define a *segment structure* to be the coordinates of all C_α-atoms of a segment. Note that a segment in principle allows for infinitely many different segment structures even though they are restricted to be of a specific secondary structure class. However, this model is discrete and therefore only a finite representative set of segment structures are generated. This is described in detail in Section *Segment structures*.

Any tertiary structure of a protein can be described in these terms; a list of segments and a segment structure for each segment. We call such a list of segments a *super structure* and a super structure with a segment structure for each segment is called a *complete structure*.

The tertiary structure of any protein can always be described by a complete structure. However, to discretize and reduce the conformational space of this model, we reduce the degree of freedom for segments. Segments are therefore only allowed to have a discrete set of predefined directions between the first and last C_α-atoms. Obviously, the more directions allowed, the more super structures can

be described by the model. This of course also increases the chance of describing a super structure similar to the native structure. Therefore, there is a trade-off between the number of directions allowed and the computational feasibility of the model. Ad-hoc experiments show that the 12 uniformly distributed directions acquired from the *face-centered cubic* (FCC) lattice is a good tradeoff (see the results section for further discussion). The direction of a segment therefore has one of the following 12 direction vectors: [1,1,0], [1,0,1], [1,-1,0], [1,0,-1], [-1,1,0], [-1,0,1], [-1,-1,0], [-1,0,-1], [0,1,1], [0,1,-1], [0,-1,1], [0,-1,-1]. Figure 2 shows an example of a super structure and a corresponding complete structure.

To further discretize the model, we set an upper limit (u) on the number of possible segment structures allowed by a segment. Given an amino acid sequence with m segments and u possible segment structures for each segment, the total number of complete structures, N, allowed by this model is

$$N = 4 \times 11^{m-2} \times u^m \tag{1}$$

One might think that this should be $N = 12^m \times u^m$ (a segment has 12 possible directions and u possible segment structures), but because of rotational symmetry of the energy function, many complete structures can be disregarded and therefore the first segment direction can be fixed. Also, the angle between two FCC vectors is 0°, 60°, 90° or 120°. Therefore, only 4 directions of the second segment need to be considered. The factors $(4 \times 11^{m-2})$ therefore describe the possible directions of segments in the super structure. Note that a segment only has 11 (not 12) possible directions, since a segment is not allowed to clash with the previous segment.

2.1 Segment Structures

Here we describe how the allowed segment structures of a given segment are computed. This computation depends on the secondary structure class of the segment.

Helix and Sheet Structures The right-handed helix is the most commonly observed secondary structure in proteins. In helices, the most observed angle between three consecutive C_α-atoms is $\phi \simeq 91°$ and the most observed dihedral angle of four consecutive C_α-atoms is $\tau \simeq 49°$. Given a helix segment, we generate one segment structure having these angle properties. Then the other $u-1$ segment structures are generated by rotating the first structure uniformly around the axis going through the first and last C_α-atoms (Figure 3).

Sheet structures are constructed in the same way as helices, but with other angle values. For sheets, the most observed angle between three consecutive C_α-atoms is $\phi \simeq 120°$ and the dihedral angle $\tau \simeq 163°$. The angle values were found by using P-SEA [16] to compute secondary structure of 3080 proteins from PDB Select (25) [17].

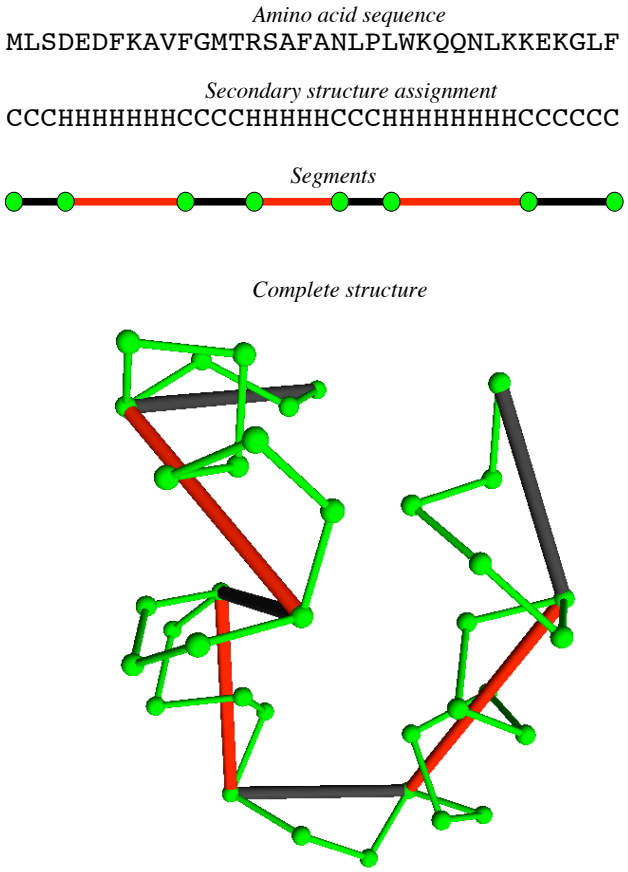


Fig. 2. The Figure shows an example of how an amino acid sequence (from Villin headpiece) can be described as a list of segments based on the secondary structure (H: helix, C: coil). The Figure also shows an example of a super structure and a corresponding complete structure (coordinates of internal C_α-atoms).

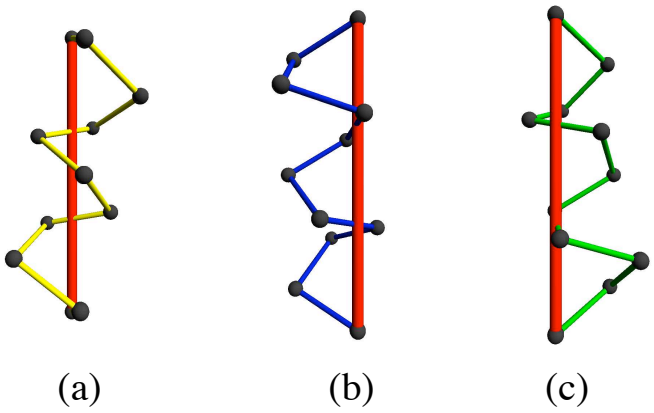


Fig. 3. (a) The first helix with angles $\phi \simeq 91^\circ$ and $\tau \simeq 49^\circ$. (b) and (c) Two other helices are generated (when $u = 3$) by uniformly rotating the first helix around the axis of the segment.

Coil Structures There are no simple geometric constraints that describe coil structures. However, experiments show that short sequences with similar amino acid sequences, so-called homologous sequences, often have similar tertiary structures [18]. Given a coil segment, we therefore query PDB Select (25) with protein sequences and their known structures and find the $\sqrt{u}$ best fragment matches in terms of amino acid similarity. Each of these structures is also rotated uniformly $\sqrt{u}$ times as for helices and sheets such that a total of u structures are obtained. The fragment database does of course not contain the proteins used in the experiments.

2.2 Energy

The structures allowed by the model always have the desired secondary structure (from a prediction), however the HSE-vector and radius of gyration of the structures varies. Therefore, we want to identify those structures having correct radius of gyration and HSE-vectors similar to the predicted HSE-vectors. The radius of gyration can be predicted from the number of residues n of the protein [8]:

$$R_g = 2.2n^{0.38} \tag{2}$$

This prediction is often accurate for globular proteins. We therefore assign infinite energy to structures having radius of gyration more than 5% away from the predicted R_g. We assign infinite energy to structures if their subchain of amino acids from the first amino acid to the l'th ($l < n$) amino acid is more than 5% away from the predicted R_g.

A structure is said to be *clashing* if the distance between two C_α-atoms is less than 3.5 Å. We also assign infinite energy to clashing structures.

Let $\mathcal{P}$ denote the conformational space of a protein with n residues $A_1, A_2, ..., A_n$. Let $P \in \mathcal{P}$. The total energy $Q(P)$ of P is defined as the sum of the residue energy contributions $Q_P(A_i)$, i.e.,

$$Q(P) = \sum_{i=1}^{n} Q_P(A_i) \tag{3}$$

with

$$Q_P(A_i) = \begin{cases} \Delta CN(A_i)^2 & \text{if } A_i \text{ is the first residue of a segment.} \\ \Delta HD(A_i)^2 + \Delta HU(A_i)^2 & \text{otherwise.} \end{cases} \tag{4}$$

where

- $\Delta CN(A_i)$ is the difference between the contact number of the i-th residue A_i in P and the desired (i.e., predicted) contact number of A_i.
- $\Delta HD(A_i)$ is the difference between the down half sphere exposure number of A_i in P and the desired down half sphere exposure number of A_i.
- $\Delta HU(A_i)$ is the similar difference for the up half sphere exposure.

The reason why CN instead of HSE is used for the first residue of a segment is that the HSE value depends on the position of the two neighbour residues as illustrated in Figure 1. For all residues of a segment structure except the first residue, the neighbour positions are always fixed and the upper and lower hemispheres can be computed. In the branch and bound algorithm we want to evaluate the energy of structures where not all segment structures are fixed which is described in detail in the next section. Instead of using HSE for these residues, we use CN which ultimately gives tighter bounds.

The radius of the contact sphere is set to 15 Å. This is known to give a good prediction quality [4] and it seems to capture both local and non-local contacts. The optimal radius has yet to be determined, both in terms of predictability and information content.

2.3 Branch and Bound

Searching for a structure with minimum global energy can be done by evaluating all structures allowed by the model. However, the number of allowed complete structures grows exponentially in terms of the number of segments m and the number of segment structures u (Equation 1). An explicit evaluation of all allowed structures is therefore only feasible for proteins with very few segments and segment structures. A standard approach for overcoming such combinatorial explosion is to use the branch and bound technique [19].

Branching The root of the branch and bound tree represents all complete structures allowed by the model. This is done by only fixing the direction of the first segment. Every other node s represents a smaller subset of complete

structures $\mathcal{P}_s$ than its parent. This is done by either fixing a segment direction or by fixing a segment structure. Therefore, when branching on a node, either 11 children with fixed segment directions are created or u children with fixed segment structures are created. A node at level $2 \times m$ has all segment directions and segment structures fixed and therefore represents a complete super structure. Nodes at level $2 \times m$ cannot be branched on further and are called leaves.

We branch the directions of segments in the order they occur in the protein. Experiments show that the total running time of the algorithm depends much on the order of how the segment directions and segment structures are fixed. The best performance is when the segment directions are fixed as early as possible and the segment structures are fixed as late as possible. The ideal case would therefore be to fix the directions in the first m levels and the segment structures in the next m levels. However, if a protein contains coil segments, it is not possible to fix all segment directions in the first m levels. This is because the end point of a coil segment depends on which coil structure is eventually chosen from the fragment database. Note that this is only a problem for coils, since all helix and sheet structures of a segment share the same end point once the start point and direction are fixed. An example of a branch and bound tree is shown in Figure 4. In the first two levels, the helix and coil segment directions are fixed. In the third level, the structure of the coil segment is fixed. This decision cannot be postponed, because the positions of the following segments depend on the chosen coil structure. At level 4 the direction of the last helix is fixed and at levels 5 and 6 the segment structures of the helices are fixed. In level 6 all directions and segment structures are fixed and the leaves therefore represent complete structures.

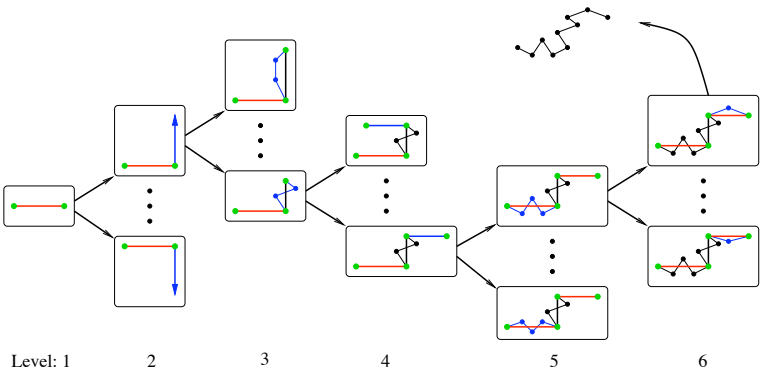


Fig. 4. The super structure consists of three segments: *helix, coil, helix*. For simplicity, in each level, only two nodes are shown and only one node is branched on.

Bounding In theory, one could simply construct the full tree, evaluate the energy function on all leaves and return the lowest energy structure. Unfortunately, because of the exponential number of leaves, this approach is computationally infeasible. Instead, we describe here a method for computing a lower bound of a non-leaf node. A lower bound is a value that is less than, or equal to the lowest energy of any leaf in the subtree of the node. Such a value can be used to disregard, or *bound*, the subtree of a node if the lower bound is larger than some observed energy (*an upper bound*). An upper bound of the energy can be found using some advanced heuristic or a simple depth first search as described in section *Searching*. Here we present a reasonable tight lower bound that can be computed fast. The use of this lower bound makes it possible to solve large problems as described in the results section.

Let $\mathcal{P}_S$ denote the subset of the conformational space $\mathcal{P}$ at any node of the branch and bound tree where some segments have fixed directions while others might have fixed segment structures (i.e., fixed coordinates of all C_α-atoms) as explained in the description of the branching strategy above. We are looking for a lower bound for $\min_{P \in \mathcal{P}_S}\{Q(P)\}$.

Consider the j-th segment S_j, $1 \leq j \leq m$, where m is the number of segments. Let

$$Q_P(S_j) = \sum_{A_i \in S_j} Q_P(A_i)$$

where $Q_P(A_i)$ is defined in Equation 4. Then the energy of a structure can be written as

$$Q(P) = \sum_{1 \leq j \leq m} Q_P(S_j)$$

Suppose that a lower bound for $\min_{P \in \mathcal{P}_S}\{Q_P(S_j)\}$ can be determined. Summing up these lower bounds for all m segments will therefore yield a lower bound for the energy of all conformations in $\mathcal{P}_S$. To compute such a lower bound for a segment S_j, the following problem is solved for all segment structures of S_j. For simplicity we only describe how a lower bound using CN vectors can be computed, however it is straightforward to use a similar approach for HSE vectors.

Given a segment structure for S_j, we determine for each of it's C_α-atom all possible values of CN when the super structure is fixed. This problem can clearly be solved in exponential time by complete enumeration (see Figure 5). However, using the following dynamic programming approach, this problem can be solved in polynomial time. The input to the dynamic programming algorithm is the table constructed as described in Figure 5(c). This table is in the following called $c_{a,b}$.

Let $c_{a,b}(i,r)$ where $(1 \leq i \leq m)$ and $(1 \leq r \leq u)$ be the number of contacts of residue a in segment b contributed by residues in segment i having segment structure r. Let (i,j) be an entry in the dynamic programming table and let $q_{a,b}(i,j) \in \{0,1\}$ represent whether or not residue a in segment b can have a total of j contacts contributed by residues in segments S_l, $(l < i)$. Then the

recursive equation of the dynamic programming algorithm is:

$$q_{a,b}(i,j) = \begin{cases} 1 \text{ if } i = 1 \text{ and } c_{a,b}(1,r) = j \text{ for some r} \\ 1 \text{ if } i > 1 \text{ and } q(i-1,k) = 1 \text{ and } c_{a,b}(i,r) = j - k \text{ for some r} \\ 0 \text{ otherwise} \end{cases} \tag{5}$$

Each row can be computed in $\mathcal{O}(n \times u)$ time using the values from the previous row, so the total running time of the algorithm is $\mathcal{O}(m \times n \times u)$. The last row in the table represents all possible contact numbers for residue a in segment b. The last row can therefore easily be used to find the minimum difference between the desired CN and one of the possible CNs. The dynamic programming problem is solved for all residues of the segment and the sum of the minimum differences for each residue is the lower bound of the segment energy.

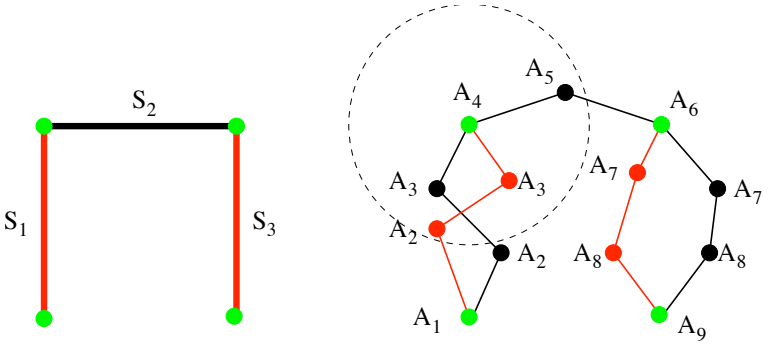
In the above discussion, it was assumed that all C_α-atoms in S_j have their coordinates fixed in $\mathcal{P}_S$. Lower bounds can also be computed if only the segment structure has not been fixed. The above lower bound computation is then merely repeated for each of the u possible segment structures, and the smallest one is selected as the overall lower bound of the segment.

Lower bounds can also be computed for nodes where a number of the last segment directions have not yet been fixed. Here, the input to the dynamic programming algorithm is only the first fixed segments. Then, the CN row for the last fixed segment is augmented by checking whether each C_α-atom on the free segments can possibly be in contact with the C_α-atom in question.

We also bound structures where two succeding segment structures have unlikely angle properties. Figure 7 shows a plot of (θ, τ) pairs from proteins in PDB. The regular angle between 3 consecutive C_α positions is θ and τ is the dihedral angle between 4 consecute C_α positions as illustrated in Figure 8. The plot shows that some regions in the (θ, τ)-plane are much more likely than others. We have marked what we think is a reasonable seperation between likely and unlikely points. Therefore structures with one ore more (θ, τ) points in the unlikely region are bounded.

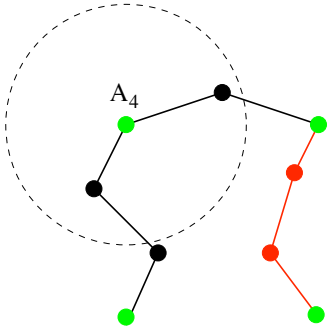
2.4 Searching

Searching the branch and bound tree is done using a combination of cost first and depth first search. The cost of a non-leaf node is the lower bound of the energy and the cost of a leaf node is the energy of the corresponding structure. We search the branch and bound tree by keeping a set of nodes for which the lower bound has been computed but not bounded. Initially the set contains only the root of the branch and bound tree. Iteratively the algorithm chooses the lowest cost node and replaces it with the children obtained by branching. When using this strategy, an optimal solution is found when the lowest cost node in the set is a leaf node. In practice the set of unbranched nodes becomes very large and difficult to store in memory. We therefore combine it with a depth first search, such that when the node set contains more than 50.000 nodes we shift to depth first search until the set is less than 50.000 again. This approach gives



(a) (b)

	Black	Red
S_1	1	2
S_2	1	–
S_3	0	0



(c) (d)

Fig. 5. (a) shows the directions of three segments (a super structure). In this example we want to compute all possible CN values for residue A_4 which is the first residue of segment S_2. The contact radius of residue A_4 is illustrated by the circle in (b). S_1 and S_3 both have two choices of segment structures (red and black), so $u = 2$. The table in Figure (c) shows the contribution of contacts to residue A_4 if either red or black segment structure is chosen. If the black structure of S_1 is chosen, S_1 only contributes with 1 contact to A_4 and if the red structure is chosen, S_1 contributes with 2 contacts. Computing all possible CN values for A_4 can be done by considering all combinations of segment structures for the other segments which is exponential. (d) shows one of these combinations which gives a CN value of 2 for A_4.

	Black	Red
S_1	1	2
S_2	1	–
S_3	0	0

	0	1	2	3	4	5	6	7	8
S_1		x	x						
S_2			x	x					
S_3			x	x					

Fig. 6. (a) shows the input to the dynamic programming algorithm as constructed in Figure 5. (b) shows Table $q_{a,b}$ where empty entries correspond to 0 and x correspond to 1. In the first row, only 1 or 2 contacts can be contributed to residue A_4 if either black or red structure of segment S_1 is chosen. Segment S_2 has a fixed segment structure and therefore always contributes with one contact as shown in row 2 and finally row 3 shows that segment S_3 does not contribute with any contacts to A_4. The last row is also the solution to the problem. It shows that from all combinations of segment structures, the CN value of residue A_4 can only be 2 or 3.

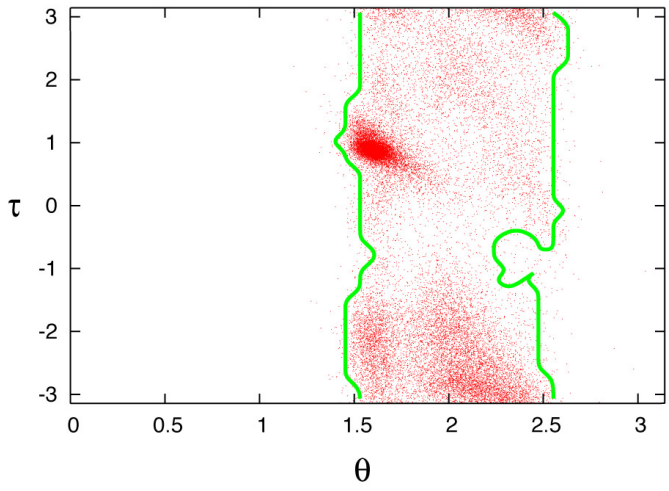


Fig. 7. A plot of (θ, τ) pairs from PDB

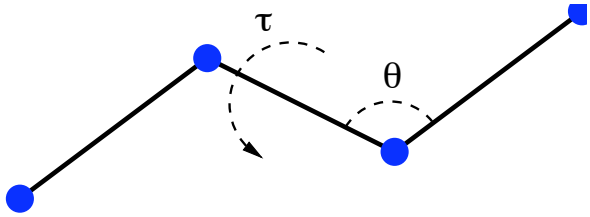


Fig. 8. θ is the normal angle between 3 consecutive C_α positions and τ is the dihedral angle between 4 consecute C_α positions.

a more memory efficient algorithm, but we might end up computing more lower bounds than in a pure cost first search.

3 Experiments

Here we predict the tertiary structures of 6 proteins. The tertiary structures of the proteins are known and we can therefore evaluate the quality of our results. These proteins have previously been used for benchmarks in the literature [20, 21].

The input to EBBA is a secondary structure assignment, HSE-vector and the radius of gyration. For each protein we obtain these values using prediction tools. Based on the amino acid sequence, we predict the secondary structure using PSIPRED [7] and we predict HSE-vectors using LAKI [4]. Note that PSIPRED and LAKI are neural networks trained on a selection of proteins from PDB. The 6 benchmark proteins used here also exist in PDB, so there is a slight chance that the training sets for PSIPRED and LAKI contain some of these proteins. However, the prediction quality of the 6 benchmark proteins is close to what should be expected from PSIPRED and LAKI. Here, the average Q_3 score of secondary structure prediction is 80.7% (compared to an average score of 80.6% on CASP targets). The average correlation of the HSE up and down values are respectively 0.74 and 0.66 (compared to the reported up and down correlations of 0.713 and 0.696 respectively). We do therefore not consider it to be a problem that the benchmark proteins exist in PDB. We predict the radius of gyration using Equation 2.

Branch and bound algorithms are typically used to find the global minimum solutions. However, our experiments show that the global minimum solutions in our models are not always native-like. Therefore, EBBA is modified such that the 10.000 best structures in terms of energy are found and not just the global minimum. This can be done by maintaining a queue fo 10.000 structures during the search. This number is still very small compared to the exponential size of the conformational space. For comparison and evaluation of the model and prediction quality, all experiments are also done using the exact secondary structure

and exact HSE-vectors obtained from the native structure of the proteins. All experiments were initially run with $u = 8$ (the number of segment structures). Some did not finish in 48 hours, and they were run with the highest value of u that could be solved in less than 48 hours. All computations were performed on a 2.4 GHz P4 with 512 RAM.

4 Results and Discussion

Table 1 shows the complexity of the model for different proteins and the running time of EBBA. Table 1 also shows the results of running EBBA on the 6 benchmark proteins. Figures 9 and 10 show 2D histograms of the energy vs. RMSD distribution for the 10.000 structures. For better comparison of the energies for the different proteins, the root-mean of the energies are reported in this section.

The maximum number of segment structures (u) that could be solved in less than 48 hours depend much on the number of segments of the protein. For the smallest proteins (1FC2 and 1ENH) the algorithm terminated in less than 48 hours using $u = 8$. Even though 2GB1 has relatively many segments the algorithm also terminated in less than 48 hours using $u = 8$. This is because of the efficiency of the bounding algorithm. In Figure 11 it is shown that for 2GB1 a large fraction of the search space can be bounded early. The most difficult protein in terms of bounding efficiency is 4ICB (predict), where it turns out that significant bounding first occurs in level 5 of the branch and bound tree. In all instances the conformational space is huge, and it clear that finding global minimum structures could not have been done in reasonable time without efficient bounding.

Figures 9 and 10 show that the exact energy vs. RMSD is well correlated for the three smallest proteins while this is not the case for the larger proteins. The larger proteins have a higher degree of freedom, and it therefore seems that secondary structure, radius of gyration and HSE do not contain enough information to identify the native structure of proteins with more than ~ 60 residues. However, among the 10.000 best structures, structures close to the native structure exists for the longer proteins also.

Table 4 shows that the set of 10.000 low energy structures for all 6 proteins contains good decoys (RMSD less than 6 Å). Also, for all proteins the lowest RMSD is smallest when using exact values compared to the predicted values. This is expected since the energy landscape should have a global minimum closer to the native structure when using exact values. However, it is surprising that for two of the proteins (1FC2 and 2GB1) the fraction of good decoys (< 6 Å RMSD) is better when using predicted values compared to exact values. The plots in Figures 10 show that for these two proteins, the structures are much more clustered when using the predicted values. This indicates that the energy landscapes described using the predicted values have fewer local minima and for 1FC2 and 2GB1 they are clustered closer to the native structure.

In Table 2 the energy span of the 10.000 structures is shown. The table also shows the energy of the native structure of the protein using the predicted energy

Type	m segments	u SS	N	T hours	< 6 Å RMSD	< 5 Å RMSD	< 4 Å RMSD	lowest RMSD	$Q(P^*)$	P^* RMSD
			Protein A (1FC2), 43 residues							
Exact	5	8 CHCHC	1.7×10^8	0.1	18.1	7.0	0.7	**2.8**	4.34	6.6
Predicted	7	8 CHCHCHC	1.4×10^{12}	6.9	33.0	13.8	0.0	**4.5**	5.26	8.4
			Homeodomain (1ENH), 54 residues							
Exact	6	8 CHCHCH	1.5×10^{10}	0.6	21.6	13.2	1.8	**3.1**	4.36	3.5
Predicted	7	8 CHCHCHC	1.4×10^{12}	6.1	4.1	0.8	0.0	**4.1**	5.70	10.2
			Protein G (2GB1), 56 residues							
Exact	9	8 SCSCHCSCS	1.0×10^{16}	18.2	60.8	36.6	13.7	**3.4**	4.22	4.3
Predicted	10	8 SCSCHCSCSC	9.2×10^{17}	4.7	73.1	0.0	0.0	**5.3**	6.22	7.8
			Cro repressor (2CRO), 65 residues							
Exact	11	4 CHCHCHCHCHC	4.0×10^{16}	24.1	5.7	1.4	0.0	**4.3**	6.49	9.2
Predicted	10	3 HCHCHCHCHC	5.1×10^{13}	7.4	1.5	0.0	0.0	**5.3**	5.89	9.4
			Protein L7/L12 (1CTF), 68 residues							
Exact	8	8 SCHCHCHC	1.2×10^{14}	5.6	5.1	1.9	0.0	**4.6**	7.19	11.0
Predicted	11	3 SCHSHCHCHCS	1.7×10^{15}	19.2	0.1	0.0	0.0	**5.4**	5.84	11.3
			Calbindin (4ICB), 76 residues							
Exact	11	2 CHCSHCHCHCH	1.9×10^{13}	3.56	4.5	0.7	0.0	**4.4**	6.18	7.4
Predicted	8	7 CHCHCHCH	4.1×10^{13}	31.4	0.5	0.0	0.0	**5.1**	6.79	6.4

Table 1. Column 2 shows the number of segments m and column 3 shows the number of segment structures u. Column 4 shows the order of helix, sheet and coil segments. Column 5 shows the size of the conformational space given by Equation 1 and column 6 shows the number of hours spent by the algorithm. Column 7 to 9 show the percentage of the 10.000 structures that fall below the given threshold. Column 10 shows the lowest RMSD of the 10.000 structures. Column 11 shows the energy of P^* which is the lowest energy structure. The last column shows the coordinate RMSD between the native structure and P^*. For each protein, there is an *exact* and a *predicted* row. Exact refers to HSE-vectors, radius of gyration and secondary structure obtained from the native structure. In the *predicted* rows, all input values are predicted from the amino acid sequence and the results can therefore be considered as *de novo*.

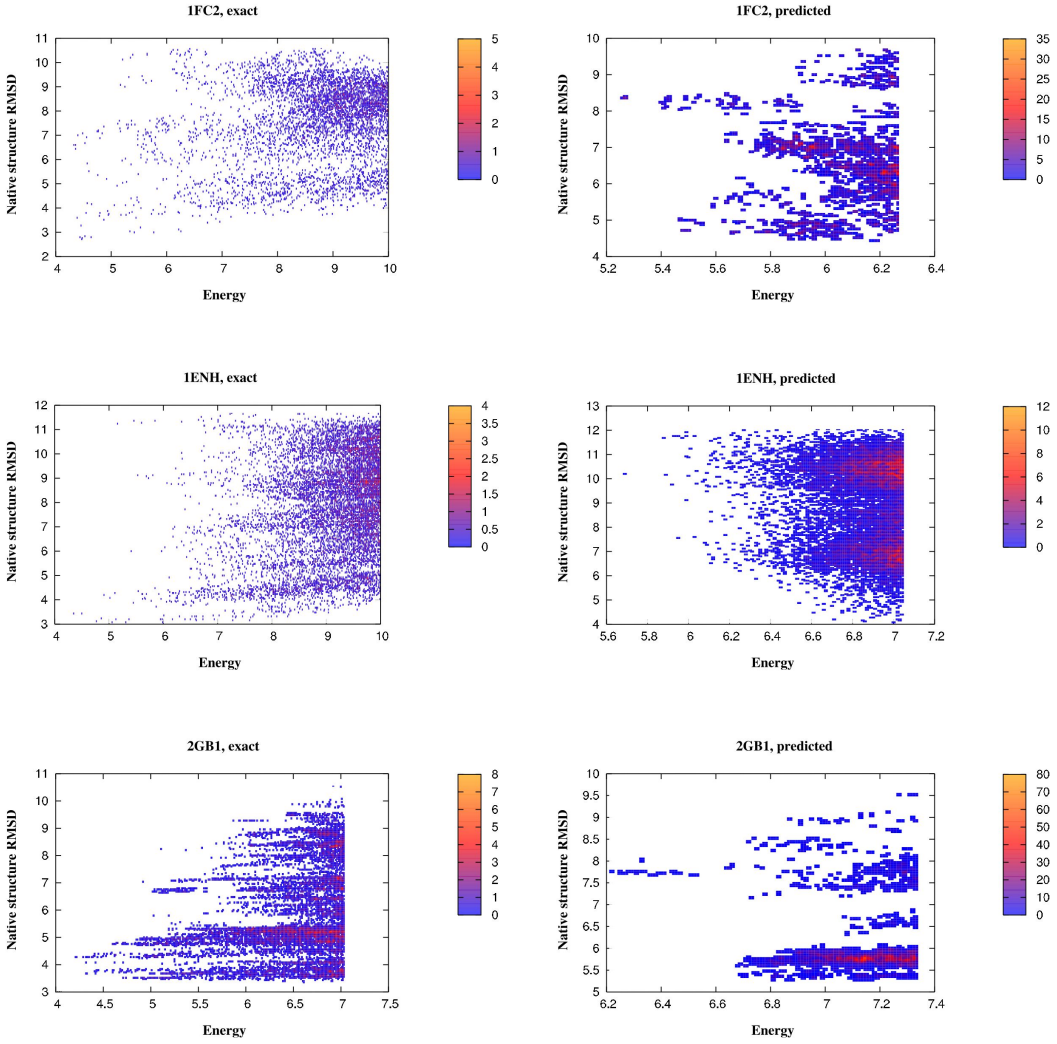


Fig. 9. Energy vs. RMSD histograms of 1FC2, 1ENH and 2GB1.

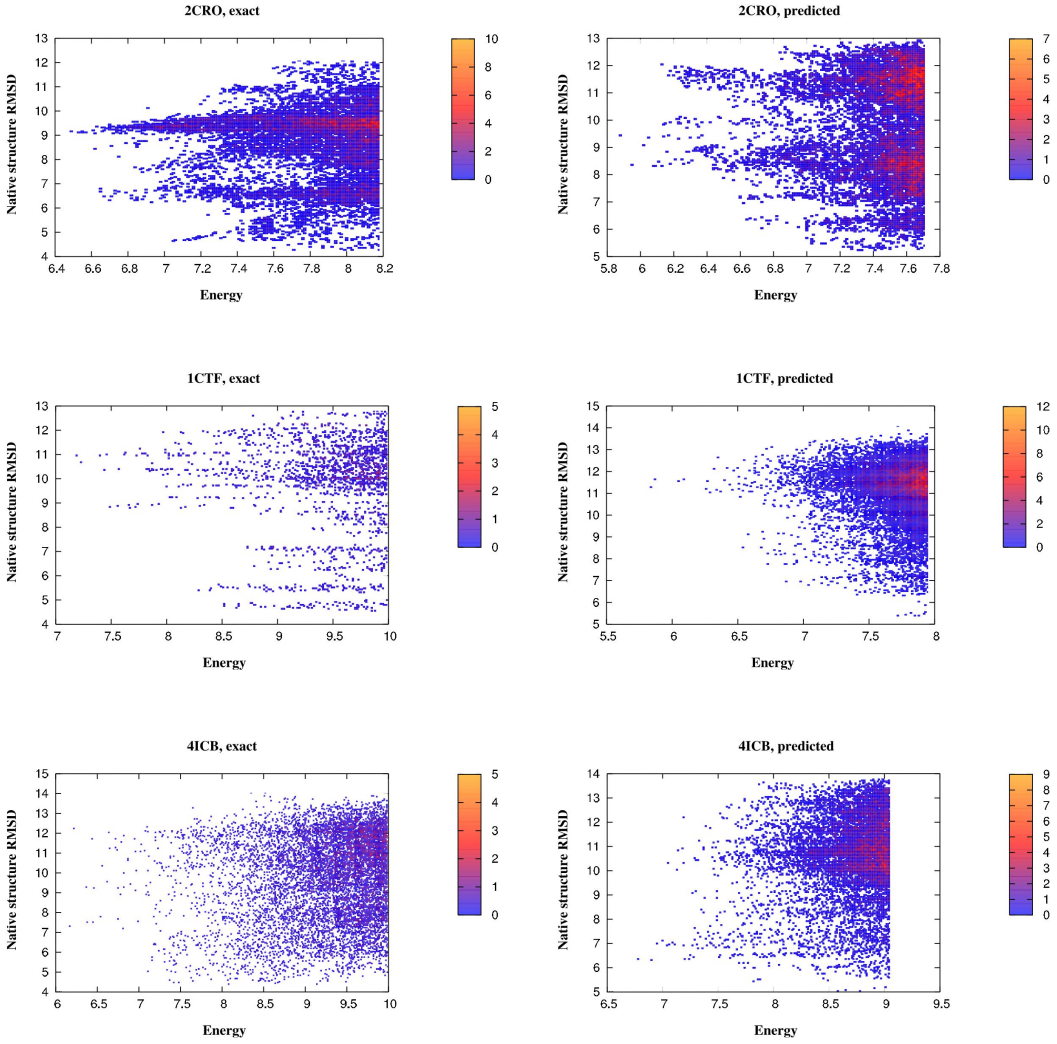


Fig. 10. Energy vs. RMSD histograms of 2CRO, 1CTF and 4ICB.

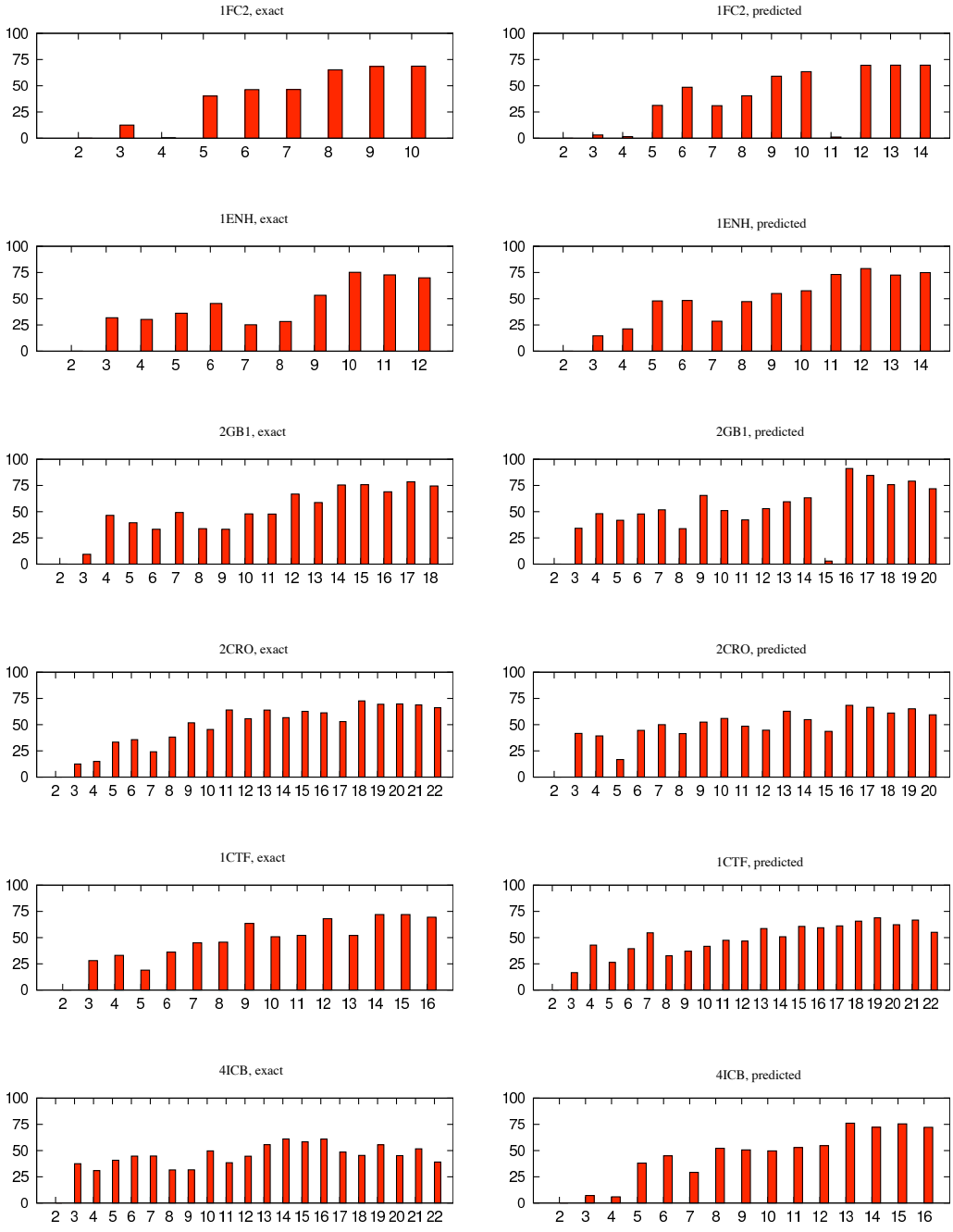


Fig. 11. The histograms show the bounding efficiency for each of the 12 runs of EBBA. The bars show the percentage of nodes in each level that was bounded. Level 1 is omitted, since the node in level 1 is never bounded (this would cause the whole search space to be bounded)

function. The values indicate that for most of the proteins (except 4ICB), the model is able to represent structures with lower energy than the native structure. Adding more degree of freedom in terms of segment directions and using more segment structures could consequently lower the energy of the 10.000 structures. However, since the energies of these structures are already comparable to the energy of the native structure, it should *not* be expected that more degree of freedom would improve the RMSD of the structures. Instead, improvements should come from adding more predictable information to the model or energy function or using more accurate predictions of HSE and secondary structure.

PDB	$Q(P^*)$	$Q(P^{10.000})$	Q^{native}
1FC2	5.26	6.28	6.46
1ENH	5.70	7.06	6.63
2GB1	6.22	7.34	7.53
2CRO	5.89	7.71	8.40
1CTF	5.84	7.96	7.58
4ICB	6.79	9.05	6.67

Table 2. For each protein the lowest energy of the 10.000 structures is $Q(P^*)$. The highest energy of the 10.000 structures is $Q(P^{10.000})$ and the energy of the native structure is Q^{native}.

The results have been compared directly with FB5-HMM [21] in Table 3. FB5-HMM is a successful method for conformational sampling. The method is based on a Hidden Markov Model and generates a large set of structures which usually contains many good decoys (< 6 Å RMSD) when enforcing compactness. The major difference between FB5-HMM and EBBA is that FB5-HMM does not use an energy function. FB5-HMM can also benefit from the secondary structure prediction and radius of gyration prediction. The results we have shown for FB5-HMM are therefore obtained using predicted secondary structure and using a greedy collapse scheme. The results for FB5-HMM are from [21] where 100.000 structures are generated. For all proteins, except 2GB1, FB5-HMM finds at least one structure with lower RMSD than EBBA. However, EBBA finds a better percentage of good decoys for most of the proteins (1FC2, 2GB1, 2CRO and 4ICB). Another advantage of the EBBA generated structures, is that the geometry of the secondary structure segments is perfect because they are constructed using the correct secondary structure geometry.

5 Conclusions

We have presented a branch and bound algorithm for finding the lowest energy structures in a large conformational search space. The energy function is based on HSE which is a simple predictable measure. This algorithm is the first ab initio

Protein	FB5-HMM		EBBA	
	< 6 Å	Min. RMSD	< 6 Å	Min. RMSD
Protein A (1FC2)	17.1	2.6	33.0	4.5
Homeodomain (1ENH)	12.2	3.8	4.1	4.1
Protein G (2GB1)	0.001	5.9	73.1	5.3
Cro repressor (2CRO)	1.0	4.1	1.5	5.3
Protein L7/L12 (1CTF)	0.3	4.1	0.1	5.4
Calbindin (4ICB)	0.4	4.5	0.5	5.1

Table 3. Comparison between FB5-HMM and EBBA. Column 2 and column 4 show the percentage of good decoys for FB5-HMM and EBBA respectively. Column 3 and column 5 show the lowest RMSD of a structure found by FB5-HMM and EBBA respectively. Both algorithms uses predicted secondary structure information and predicted radius of gyration.

branch and bound algorithm for prediction of protein structure using only one-dimensional predictable information. We have shown experimentally that good decoys always exist among the 10.000 lowest energy structures for the proteins used here. However, the energy function is not accurate enough to pinpoint the lowest RSMD structure in this set. An important future research direction is therefore to examine this set of low energy structures with a more detailed energy function and to identify the native-like structures. The largest protein considered have 76 residues. There is a problem using the branch and bound algorithm on larger proteins since then only a small fraction of the conformational space can be searched in reasonable time. However, we believe that exploiting how super secondary structures [22, 23] arrange in nature, might be a way to solve this problem. Better search heuristics for finding upper bounds on the energy can also be relevant since a good upper bound on the energy also improves the performance of the branch and bound algorithm. Using a more probabilistic approach might also improve the quality of the results. One idea is to compute probabilities from the (ϕ, ψ)-plot in Figure 7 instead of a simple threshold bound used here. It might also be possible to train a Bayesian network to predict the probability of a given HSE-vector given the amino acid sequence. This would be a more detailed usage of the HSE-vector compared to the simple energy function used here.

6 Acknowledgements

We would like to thank Thomas Hamelryck at the Bioinformatics Centre, University of Copenhagen for valuable contributions and insights. Martin Paluszewski and Pawel Winter are partially supported by a grant from the Danish Research Council (51-00-0336).

References

1. Paluszewski, M., Winter, P.: Protein decoy generation using branch and bound with efficient bounding. Lecture Notes in Bioinformatics, (to appear) (2008)
2. Hamelryck, T.: An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. Proteins **59**(1) (2005) 38–48
3. Paluszewski, M., Hamelryck, T., Winter, P.: Reconstructing protein structure from solvent exposure using tabu search. Algorithms for Molecular Biology **1** (2006)
4. Vilhjalmsson, B., Hamelryck, T.: Predicting a New Type of Solvent Exposure. ECCB, Computational Biology Madrid 05, P-C35, Poster (2005)
5. Pollastri, G., Baldi, P., Fariselli, P., Casadio, R.: Prediction of coordination number and relative solvent accessibility in proteins. Proteins **47**(2) (2002) 142–53
6. Kinjo, A. R., Nishikawa, K.: Recoverable one-dimensional encoding of three-dimensional protein structures. Bioinformatics **21**(10) (2005) 2167–70
7. McGuffin, L. J., Bryson, K., Jones, D. T.: The PSIPRED protein structure prediction server. Bioinformatics **16** (2000) 404–405
8. Skolnick, J., Kolinski, A., Ortiz, A. R.: MONSSTER: A method for folding globular proteins with a small number of distance restraints. J. Mol. Biol. **265** (1997) 217–241
9. Fain, B., Levitt, M.: A novel method for sampling alpha-helical protein backbones. Journal of Molecular Biology **305** (2001) 191–201
10. Kolodny, R., Levitt, M.: Protein decoy assembly using short fragments under geometric constraints. Biopolymers **68**(3) (March 2003) 278–285
11. Backofen, R.: The protein structure prediction problem: A constraint optimization approach using a new lower bound. Constraints **6** (2004) 223–255
12. Backofen, R., Will, S.: A constraint-based approach to fast and exact structure prediction in three-dimensional protein models. Constraints **11**(1) (January 2006) 5–30
13. Maranas, C. D., Floudas, C. A.: A deterministic global optimization approach for molecular structure determination. J. Chem. Phys. **100** (1994) 1247–1261
14. Standley, D. M., Eyrich, V. A., Felts, A. K., Friesner, R. A., McDermott, A. E.: A branch and bound algorithm for protein structure refinement from sparse nmr data sets. J. Mol. Biol. **285** (1999) 1961–1710
15. Palu, A. D., Dovier, A., Fogolari, F.: Constraint logic programming approach to protein structure prediction. BMC Bioinformatics **5**(186) (November 2004)
16. Labesse, G., Colloc'h, N., Pothier, J., Mornon, J.-P.: P-SEA: a new efficient assignment of secondary structure from calpha trace of proteins. Bioinformatics **13** (1997) 291–295
17. Hobohm, U., Sander, C.: Enlarged representative set of protein structures. Protein Science **3** (1994) 522–524
18. Chothia, C., Lesk, A. M.: The relation between the divergence of sequence and structure in proteins. The EMBO Journal **5** (1986) 823–826
19. Wolsey, L. A.: Integer Programming. Wiley-Interscience (1998)
20. Simons, K. T., Kooperberg, C., Huang, E., Baker, D.: Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol **268**(1) (1997) 209–25
21. Hamelryck, T., Kent, J. T., Krogh, A.: Sampling realistic protein conformations using local structural bias. PLoS Computational Biology **2**(9) (September 2006) 1121–1133

22. Sun, Z., Jiang, B.: Patterns and conformations of commonly occurring supersecondary structures (basic motifs) in protein data bank. J. Protein Chem. **15**(7) (October 1996) 675–690
23. Boutonnet, N. S., Kajava, A. V., Rooman, M. J.: Structural classification of alphabetabeta and betabetaalpha supersecondary structure units in proteins. Proteins **30**(2) (February 1998) 193–212