Artificial Intelligence in Medicine (2008) xxx, xxx–xxx

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

# A graph theoretic approach to protein structure selection

Marco Vassura [a,*], Luciano Margara [a], Piero Fariselli [b], Rita Casadio [b]

[a] Computer Science Department, University of Bologna, Via Mura Anteo Zamboni, 7, 40127 Bologna, Italy
[b] Biocomputing Group, Department of Biology, University of Bologna, Via Irnerio, 42, 40127 Bologna, Italy

Received 19 October 2007; received in revised form 25 July 2008; accepted 26 July 2008

**KEYWORDS**
Protein structure prediction;
Protein folding;
Protein structure selection;
Contact maps;
Graph algorithm

**Summary**

*Objective:* Protein structure prediction (PSP) aims to reconstruct the 3D structure of a given protein starting from its primary structure (chain of amino acidic residues). It is a well-known fact that the 3D structure of a protein only depends on its primary structure. PSP is one of the most important and still unsolved problems in computational biology. Protein structure selection (PSS), instead of reconstructing a 3D model for the given chain, aims to select among a given, possibly large, number of 3D structures (called decoys) those that are closer (according to a given notion of distance) to the original (unknown) one. In this paper we address PSS problem using graph theoretic techniques.
*Methods and materials:* Existing methods for solving PSS make use of suitably defined energy functions which heavily rely on the primary structure of the protein and on protein chemistry. In this paper we present a new approach to PSS which does not take advantage of the knowledge of the primary structure of the protein but only depends on the graph theoretic properties of the decoys graphs (vertices represent residues and edges represent pairs of residues whose Euclidean distance is less than or equal to a fixed threshold).
*Results:* Even if our methods only rely on approximate geometric information, experimental results show that some of the adopted graph properties score similarly to energy-based filtering functions in selecting the best decoys.
*Conclusion:* Our results highlight the principal role of geometric information in PSS, setting a new starting point and filtering method for existing energy function-based techniques.


## 1. Introduction

One of the most important and largely unsolved problems in bioinformatics is the so called *3D protein structure prediction* (PSP) [1–3]. It is a well

* Corresponding author. Tel.: +39 051 209 48 48; fax: +39 051 209 45 10.
*E-mail addresses:* vassura@cs.unibo.it (M. Vassura), margara@cs.unibo.it (L. Margara), piero@biocomp.unibo.it (P. Fariselli), casadio@alma.unibo.it (R. Casadio).

0933-3657/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2008.07.016

known fact that all protein molecules are uniquely identified by means of their *primary structure*, i.e., the sequence of amino acidic residues that forms the backbone of the protein. In other words, under physiological conditions, a chain of residues admits a unique compact and functionally active conformation called *native structure*. In principle, the 3D structure of a protein (called *tertiary structure*) and then its biological function might be deterministically computed once its primary structure is provided. PSP is the problem of computing the 3D structure of a protein, i.e., the spatial coordinates of all the residues, taking as input its primary structure.

Heuristics for tackling PSP can be divided into two broad classes: *ab initio* methods and *homology detection* techniques. *Ab initio* protein modeling methods seek to build 3D protein structure from scratch simulating physical and biological molecular interactions. Homology detection techniques try to predict a 3D model for a given chain of residues taking advantage of known 3D structures. The output of PSP methods usually consists of a possibly large set of candidates (decoys) that are expected to approximate the given protein conformation.

Protein structure selection problem (PSS), instead of reconstructing a 3D model for a given chain of residues, aims to select, among a given, possibly large, number of decoys those that are closer (according to a given notion of distance) to the original (unknown) protein 3D structure. A number of heuristics have been developed during the last few years for solving PSS (see for example [3–6]). All of them make use of the so called *energy functions*. Energy functions take as input the primary structure of the protein and the description of a decoy and yield a numerical value (score) which is expected to measure the quality of the decoy. The lowest is the energy of a decoy the closest to the 3D structure of the original protein it should be. Unfortunately, it is known that small intrinsic errors can lead to predict a high number of erroneous structures having a lower energy than the native structure [7]. Decreasing the number of erroneous structures among the predictions is therefore of great importance.

In this paper we present a new approach to face PSS based on the analysis of some selected graph properties on suitably defined decoys graphs. We represent each decoy as an undirected graph where vertices represent residues and edges represent pairs of residues whose Euclidean distance is less than or equal to a fixed threshold. Distances are actually computed between pairs of C-$\alpha$ atoms. Decoys graphs can be represented in an equivalent way by using contact maps. A contact map for any given pair of residues, yields 0 if their Euclidean distance is larger than a given threshold, 1 otherwise. Instead of evaluating the quality of a decoy on the basis of its exact geometric conformation and of the primary structure of the protein, we produce a ranking of the decoys according to some selected graph properties.

The main goal of this paper is to shed some light on the relations existing between decoys properties and graph properties. We wish to emphasize that our ranking techniques are completely independent of the primary structure of the proteins. In other words, we evaluate decoys quality according to (coarse) geometric information only. Such information is extracted from the decoy structure by the graph properties that we consider. In this way we compute a score indicating the quality of the decoy without any information on the corresponding native structure.

To test our methods and to make them comparable to other methods, we use one of the most widely accepted benchmark data [6], available at the Baker Laboratory web site[1]. Given the data set of protein and decoy structures, we consider seven graph properties, namely *Average Degree*, *Contact Order*, *Normalized Complexity*, *Network Flow*, *Connectivity*, and a weighted version of *Network Flow* and *Connectivity*. The ability of each graph property to distinguish between correct and incorrect 3D structures is then evaluated by computing the *Z* score and the Enrichment score [6]. Experimental results show that the above listed properties perform similarly (if not better) to previously described methods based on backbone energy functions [6,8].

In addition, we assess the quality our method comparing it with the results of the latest CASP7 experiment. The critical assessment of techniques for protein structure prediction (CASP) experiment is a blind test that provides an objective assessment of the effectiveness of PSP methods. The quality assessment (QA) category of the CASP7 evaluates capabilities of state of the art programs to distinguish near native structures among decoys. We show that our procedure has performances similar to the ones of the best Model Quality Assessment Programs at the CASP7 experiment.

The rest of this paper is organized as follows. Section 2.1 contains basic definitions, Section 2.2 describes the benchmark data set that we use in our experiments, Section 2.3 describes the selected graph properties, Section 3 contains the results

[1] ftp://ftp.bakerlab.org/pub/decoys/decoys_11-14-01.tar.gz.

Please cite this article in press as: Vassura M, et al. A graph theoretic approach to protein structure selection. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.07.016

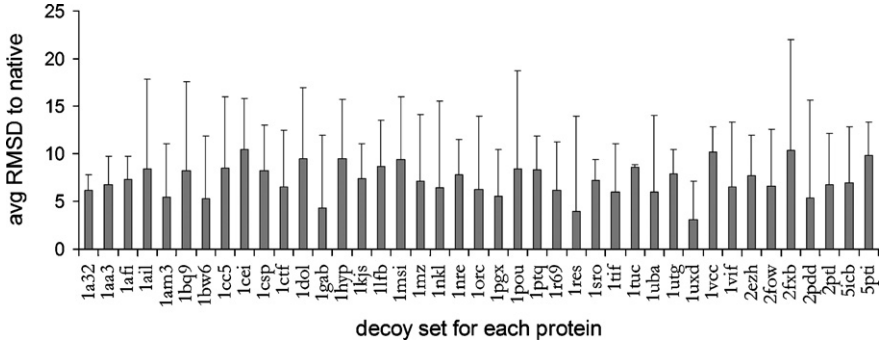


**Figure 1** Average C-$\alpha$ RMSD from native structures (in Å) computed over the Rosetta decoy set adopted for benchmarking our method. The set is constructed to be an objective and difficult test for PSS.

regarding the performances of our method on both the test data set and CASP7, and finally Section 4 contains conclusions and ideas for further works.

## 2. Materials and methods

### 2.1. From protein structure to adjacency matrix

Proteins constructed to test PSS are called *decoys*, while predicted proteins are *models*, and the real protein is the *native protein*. In general we will refer to them as proteins. For a given protein 3D structure we compute its C-$\alpha$ trace [2], so that each residue is identified with its C-$\alpha$ atom.

Accordingly, protein residues $i$ and $j$ are defined to be in contact when:

$$d(i, j) < s$$

where $d(i, j)$ is the Euclidean distance between 3D coordinates of C-$\alpha$ atoms and $s$ is the contact map threshold. For a search in the $s$ space, threshold is changed from 7 to 18 Ångstrom (Å) [9] by adopting a 1 Å step. By this each protein 3D structure generates 12 different graph representations. A contact map is the representation of a graph adjacency matrix where the vertices are the C-$\alpha$ atoms and the edges are the contacts between them. At increasing $s$ values, the number of edges for a given vertex increases. Given that the Euclidean distance is symmetric, the contact map is also symmetric. This implies that the corresponding graph is undirected ($i \to j = j \to i$).

The C-$\alpha$ root mean square deviation (C-$\alpha$ RMSD) is the common used distance measure between molecular structures. Given the set of coordinates $C, C' \in \mathbb{R}^{3\times n}$ it is defined as the minimum distance

$$D_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(C'[i] - C_k[i])^2}$$

over any rototranslation $C_k$ of the coordinate set $C$.

Graph edges can also be related to the primary protein structure by defining a sequence separation among adjacent residues as

$$\Delta_{ij} = \mathrm{abs}(i - j)$$

where $\mathrm{abs}(i - j)$ is the absolute value of the distance along the primary structure between residue positions $i$ and $j$.

### 2.2. Decoy set

The decoy set was downloaded from the Baker's Laboratory web site.[2] This set (Rosetta) was obtained with the Rosetta algorithm that routinely can generate reasonable low-resolution structures, but that cannot reliably identify the most native-like model [1,10,11]. We choose this decoys set because it is the most recent and complete in terms of number of proteins and decoys per protein.[3] In particular, the decoys were produced following four criteria: (1) containing conformations for a wide variety of different proteins; (2) containing conformations close ($<$0.4 nm) to the native structures; (3) consisting of conformations that are at least near local minimum of a reasonable scoring function; (4) being produced by a relatively unbiased procedure (see [6] for details). The Rosetta decoy set consists of 41 native proteins with about 1800 decoys per native protein, for a total of about 76,000 protein structures. Fig. 1 shows for each native protein in the set the average C-$\alpha$ RMSD of the decoys from the native protein together with the corresponding standard deviation.

### 2.3. Graph properties

We selected graph properties with polynomial time worst case computational complexity able to cap-

[2] ftp://ftp.bakerlab.org/pub/decoys/decoys_11-14-01.tar.gz.
[3] For other decoys set see Decoys'R'Us web site http://dd.compbio.washington.edu/ (Accessed: 23 July 2008).

Please cite this article in press as: Vassura M, et al. A graph theoretic approach to protein structure selection. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.07.016

ture some of the properties of the whole protein 3D structure. Each is computed on the native protein structure, all related decoys as downloaded from the Baker's Laboratory web site and models produced and ranked at the last CASP7. Each property is then considered a measure of the quality of the structure on which it has been computed, reasonably higher values correspond to better structures, and include those corresponding to the native structure.

*Average degree* (AvgDeg). This property is obtained as the ratio of the number of edges (contacts) in the protein structure (nedges) divided by the protein length ($n$). It is computed as:

$$\text{AvgDeg} = \frac{\text{nedges}}{n}$$

and its computational complexity, given the number of contacts nedges, is $O(1)$. The number of edges of a given residue depends on the protein 3D structure. The greater the number of edges the more contacts each residue makes.

*Contact order* (CO). CO measures the average contact (edge) distribution with respect to the residue sequence separation [12]. CO is computed as:

$$\text{CO} = \frac{\sum^{\text{nedges}} \Delta_{ij}/n}{\text{nedges}}$$

where the summation index runs only on the adjacent residue pairs $i$ and $j$, $n$ is the protein length, $\Delta_{ij} = |i - j|$ is the sequence separation between residues $i$ and $j$, and nedges is the number of edges in the graph associated to the protein contact map. Its computational complexity is $O$(nedges). A high contact order value implies that there are several adjacent residues that are far apart on the residue sequence but are close in the 3D structure.

*Normalized complexity* (Ncompl). The complexity is the number of spanning trees of the graph [13], namely the number of all the trees that link all the graph vertices. It is computed as previously described in [14] using GSL libraries [15]. Its computational complexity is $O(n^3)$, where $n$ is the number of residues of the protein. Values of this property are exponentially increasing at increasing number of edges per protein length (AvgDeg). We define Ncompl as the complexity of the graph divided by AvgDeg$^n$. Since AvgDeg is different for each structure, the information contained in Ncompl can therefore be regarded as a normalized complexity.

*Network flow* (Flow). Flow computes the maximum flow from the first (residue in position 1) to the last residue (residue in position $n$), i.e. the minimum number of contacts (edges) that have to be deleted in order to disconnect the first and the last residue [13]. This property is related to the protein connectivity with respect to the N and C terminus (the first and last residues) and is computed with HI_PR,[4] an efficient implementation of the push-relabel method [16,17]. Its computational complexity is $O(n^3)$.

*Weighted flow* (Wflow). Wflow is the maximum flow (Flow) considering the graph edges (contacts) weighted by the value of the sequence separation between adjacent residues $\Delta_{ij}$. As for Flow, its computational complexity is $O(n^3)$.

*Connectivity* (Conn). It is the edge connectivity: the minimum number of edges that have to be deleted to disconnect at least one residue from all the others [13].

It is computed as the minimum of each maximum flow from the first residue to each other residue, its computational complexity is $O(n^4)$.

*Weighted connectivity* (Wconn). Wconn is similar to Conn, however with weights associated to edges. Weights are values of sequence separation between adjacent residues $\Delta_{ij}$. As for connectivity its computational complexity is $O(n^4)$.

## 2.4. Scoring functions

In order to compare our results with those already published we computed the same accuracy scores previously described in [6]: the $Z$ and Enrichment score.

The $Z$ score accounts for the deviation from the average distribution in standard deviation units. More formally for a graph property (see Section 2.3) $m$:

$$Z = \frac{m_n - \text{avg}_m}{\text{var}_m}$$

where $m_n$ is the value of $m$ for the native protein 3D structure, avg and var are respectively the average and variance of $m$ for the corresponding protein models/decoys. To compute the $Z$ score of a set of decoys we take the average of $Z$ scores values for each native structure (with respect to corresponding decoys) in the set. The larger is the absolute $Z$ score value, the better a specific graph property sorts out the native structure among its decoys.

The Enrichment score, as introduced by [6], accounts for the correlation between the property under examination and the C-$\alpha$ RMSD between the decoys/models and the native structure. The Enrichment is computed as:

$$\text{Enrichment} = \frac{(\text{First}(k, m) \cap \text{First}(k, \text{C-}\alpha\text{RMSD}))}{(\text{First}(k, \text{Random}) \cap \text{First}(k, \text{C-}\alpha\text{RMSD}))}$$

[4] http://www.igsystems.com/hipr/download.html(Accessed: 23 July 2008).

Please cite this article in press as: Vassura M, et al. A graph theoretic approach to protein structure selection. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.07.016

**Table 1** Average time in seconds necessary to compute each graph property on a given decoy set

| Size [a] | AvgDeg [b] | CO [b] | Ncompl [b] | Flow [b] | Wflow [b] | Conn [b] | Wconn [b] |
|---|---|---|---|---|---|---|---|
| 50–99 (563) | 0.004 | 0.000 | 0.023 | 0.004 | 0.004 | 0.070 | 0.058 |
| 100–199 (1973) | 0.012 | 0.000 | 0.120 | 0.044 | 0.044 | 0.228 | 0.195 |
| 200–299 (1442) | 0.048 | 0.001 | 1.185 | 0.343 | 0.343 | 0.848 | 0.700 |
| 300–350 (440) | 0.077 | 0.001 | 3.151 | 1.131 | 1.132 | 2.873 | 2.411 |

Average values are shown for different decoy sizes. The computational feasibility allows to adopt the graph properties as filtering tools in wide-scale computing.

[a] Number of residues in the chain (within brackets: number of decoys in each set).

[b] Time in seconds, average computed for thresholds ranging from 7 to 18 Å; see Section 2.3 for abbreviations.

where First $(k, m)$ is the subset of the first $k$ decoys ranked according to the graph property $m$, at decreasing values; First $(k, \text{C-}\alpha\text{RMSD})$ is the subset of the first $k$ decoys ranked according to the C-$\alpha$ RMSD from to the native structure; First $(k, \text{Random})$ is the subset of the first $k$ decoys ranked according to the random assignment. To compare our results with those in [6], the number of decoys found in the intersection set between the top high scoring 15% decoys (as obtained according to a given graph property), and the top 15% decoys with the lowest C-$\alpha$ RMSD for a given native protein is divided by the number of random assignment (the random assignment value is equal to 15% $\times$15% $\times$ total number of decoys in the set). This is done to highlight the performance of the graph property at hand. An Enrichment value equal to one indicates that the graph property does not perform better than a random assignment (the higher the value the better is the performance of the property at hand).

## 3. Results

Several methods have been applied so far to address the problem of the selection of near native structures in a decoy set. In this paper we test our method using the Rosetta decoy set [6], computing the same accuracy scores previously described in [6]: the $Z$ and Enrichment score (see Section 2.4 for detailed description). In order to effectively evaluate our graph-based properties we compare our results also with methods that have been previously described, and have been proved to be very efficient (in [6,8]).

### 3.1. Computational complexity

As explained in Section 2.3 all the different graph properties used here have polynomial time complexity. In Table 1 we show the behavior of these graph properties in the real application. Computations were run on a system equipped with 2 Gb RAM and 2.40 GHz Intel(R) Xeon(TM) CPU. Note that average times are few seconds for any graph property on any protein structure, allowing the use of this method in wide-scale computing.

### 3.2. Decoy set

In Table 2 we report the results for protein graphs computed with a threshold of 8 Å. The results indicate that Ncompl, CO, AvgDeg and Wflow are graph properties satisfactory enough to obtain an average Enrichment value higher than one (random assignment) over the protein set. It is interesting that the best performing graph property Ncompl is

**Table 2** The Enrichment and $Z$ score values computed for the different graph-based properties on the Rosetta decoy set

| Property [a] | All | | NMR | | X-ray | |
|---|---|---|---|---|---|---|
| | Enrichment | $Z$ score | Enrichment | $Z$ score | Enrichment | $Z$ score |
| AvgDeg | 1.069 | 0.565 | 1.004 | 0.876 | 1.111 | 0.365 |
| CO | 1.179 | 0.809 | 1.072 | 0.481 | 1.248 | 1.019 |
| Ncompl | 1.457 | 0.378 | 1.322 | 0.252 | 1.543 | 0.458 |
| Flow | 0.852 | −0.083 | 0.820 | −0.175 | 0.873 | −0.024 |
| Conn | 0.584 | −0.104 | 0.638 | −0.229 | 0.548 | −0.024 |
| Wflow | 1.151 | 0.109 | 1.026 | 0.093 | 1.232 | 0.120 |
| Wconn | 0.687 | −0.089 | 0.706 | −0.150 | 0.675 | −0.050 |

Enrichment measures the number of times that the graph property is better than random assignment at PSS. $Z$ score measures the ability of the graph property to separate native structure from all other decoys. The best performing graph property (Ncompl) is related to the computation of trees spanning through the protein graph, extracting a global description of the protein structure.

[a] The graph properties (see Section 2.3 for the description) are evaluated with a threshold of 8Å.

Please cite this article in press as: Vassura M, et al. A graph theoretic approach to protein structure selection. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.07.016

related to the computation of trees spanning through the protein graph, extracting a global description of the protein structure. This finding supports the idea that native-like protein structures have peculiar contact networks that are really different from those obtained using regular or random graphs [18,19]. From the results shown in Table 2, it can also be concluded that the Enrichment scores are slightly better for protein structures that were resolved with X-ray diffraction methods rather than with NMR. This is expected if we consider the higher precision of X-ray experiments, allowing better distinction between native protein and decoys. The results shown in Table 3 are grouped as a function of the different secondary structure contents of proteins and decoys. Our results confirm [6] that all alpha decoys are more difficult to rank as compared to beta and mixed secondary structure types. The results may be due to the fact that in the all alpha proteins, the majority of edges are located at very short sequence separations with few edges among the different helices. This type of graphs may therefore cause more difficulties in discriminating different spatial dispositions of helical domains.

The density of edges of a decoy graph strictly depends on the threshold cutoff value of the radius adopted, increasing at increasing cutoff value. For this reason the different graph-based functions were computed using different thresholds, in the range from 7 to 18 Å with a step of 1 Å for each protein of the Rosetta data set. As shown in Fig. 2 it appears that the capability of discriminating close-to-the-native decoy structures with Ncompl, CO, and AvgDeg is above random (equal to 1) at all the different values of threshold cutoff adopted. Conn, Wconn, and Wflow have better than random performances at higher (lower for Wflow) thresholds. In Table 4 we compare our graph properties (see Section 2.3) with methods that have been previously described (in [6,8,20]). Benchmarking is done on the Rosetta decoy set [6](described in Section 3.2). From Table 4 emerges that the graph properties (last rows) perform quite similarly to existing functions, in terms of Enrichment. This is very surprising if we take into account that such accuracy is obtained without any knowledge of chemical information of the native structure, so that they are really complementary to other existing methods for PSS. On the contrary, when the Z score is considered, our graph-properties perform worse than some energy-based functions, indicating that a finer tuning is necessary to pick up the native structure among a set of very close-native decoys. These findings show that our approach, based only on the properties computed from the protein graphs can be adopted when addressing the problem of discriminating native-like conformations in decoy sets, at least as a pre-filtering procedure.

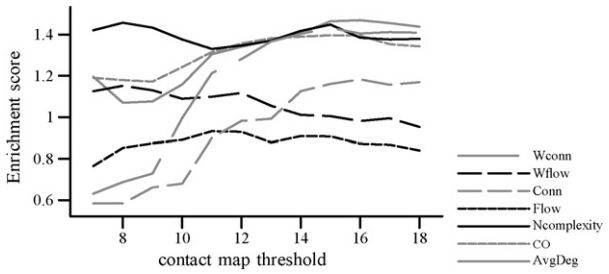


**Figure 2** The Enrichment of the graph properties at different threshold values. The capability of discriminating close-to-the-native decoy structures with Ncompl, CO, and AvgDeg is above random (equal to 1) at all the different values of threshold cut-off adopted, while Conn, Wconn, and Wflow have better than random performances at higher (lower for Wflow) threshold values.

## 3.3. CASP7

CASP experiments aim to evaluate the state of the art of protein 3D structure prediction methods. The last CASP experiment (CASP7) collected more than 60,000 models from 253 different research groups.

**Table 3** The Enrichment and *Z* score values grouped by secondary structure type on the protein set

| Property [a] | $\alpha$ | | $\alpha\beta$ | | $\beta$ | |
|---|---|---|---|---|---|---|
| | Enrichment | *Z* score | Enrichment | *Z* score | Enrichment | *Z* score |
| AvgDeg | 0.840 | −0.063 | 1.218 | 1.088 | 1.479 | 1.474 |
| CO | 0.863 | −0.120 | 1.537 | 1.481 | 1.464 | 2.351 |
| Ncompl | 1.187 | −0.104 | 1.576 | 0.764 | 2.043 | 1.108 |
| Flow | 0.803 | −0.239 | 0.939 | 0.061 | 0.840 | 0.118 |
| Conn | 0.691 | −0.271 | 0.555 | 0.079 | 0.316 | 0.060 |
| Wflow | 1.026 | −0.113 | 1.191 | 0.147 | 1.453 | 0.705 |
| Wconn | 0.819 | −0.097 | 0.521 | −0.097 | 0.600 | −0.050 |

Our results are in agreement with previous finding [6] and confirm that all-alpha decoys are slightly more difficult to rank as compared to beta and mixed secondary structure types.

[a] As in Table 2.

Please cite this article in press as: Vassura M, et al. A graph theoretic approach to protein structure selection. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.07.016

The quality assessment (QA) category of CASP7 is a comparison of methods to distinguish near-native structures from decoys.[5] A model quality assessment program (MQAP) is defined as a program that receives as input a 3D model and produces as output a real number representing the quality of the model. A blind test on the models predicted by the other prediction methods on 87 targets was used to independently assess the quality of the participating MQAPs [21]. We adopted GDT_TS [22] as evaluation measure of model correctness, as this is the standard measure used in CASP. Global distance test (GDT_TS) measure performs sequence-independent superposition of the model and the native structure and calculates the number of residues that are within a specified distance $d$. The GDT_TS score is obtained averaging the values computed with $d = 1$, 2, 4 and 8 Å and dividing by the number of residues of the native structure. We downloaded models, targets and MQAP predictions from the Prediction Center.[6] For performance comparison we considered the same set of 19,221 models and 87 targets used in previous works, so that all considered MQAPs have a prediction for each model [21].

For each of the 87 targets the best of the top five models ranked according to the MQAP is considered. The average GDT_TS score of this models is used to measure the ability of each MQAP to select good models. The model quality varies significantly among different targets, so MQAPs may have different performances for targets of different difficulty. The distribution of the GDT_TS score per target median was found to be bimodal with peaks around GDT_TS = 20 and 60. Targets are then divided into two sets of hard and easy targets using a cutoff of GDT_TS = 40 (as in [21]). Results for MQAPs, for OSP, and for our best graph property Ncompl are shown in Table 5. Models and details are available at the CASP7 web site.[7] In addition to CASP7 MQAPs we analyze the performances of Verify3D [23,24], ProsaII [25] and OSP [26–28]. The first two are included as reference, since are among the most frequently used MQAPs. The occluded surface packing (OSP) is a method to evaluate atomic packing of protein model structures. It is worth noticing that our method that exploits only geometric information of the C-$\alpha$ trace without knowledge of the primary structure, can select good models better than a score based on atom packing. GDT_TS is included as the perfect MQAP for reference, to understand what is the maximum score that

**Table 4** Comparison of graph-based functions with the state of the art functions on Rosetta decoy set

| Function/property | Rosetta | |
|---|---|---|
| | Enrich. | $Z$ score |
| All atom | | |
| RAPDF [a] | 1.23 | −6.71 |
| SOLV [b] | 0.84 | −2.96 |
| HYDB [b] | 1.33 | −6.29 |
| TORS [b] | 1.36 | −2.09 |
| FRST [b] | 1.41 | −3.72 |
| LJ attractive [c] | 1.40 | −1.48 |
| LJ attractive, side chain only [c] | 1.35 | −1.47 |
| LJ repulsive capped [c] | 0.85 | 4.37 |
| LJ repulsive linear [c] | 0.87 | 3.10 |
| LJ repulsive linear, side chain only [c] | 0.78 | −1.48 |
| LJ total, capped [c] | 0.92 | 4.38 |
| LJ total, linear [c] | 1.14 | −2.48 |
| LJ total, linear, side chain only [c] | 1.26 | −2.86 |
| Coulomb [c] | 1.14 | −1.52 |
| Screened Coulomb [c] | 0.87 | −0.96 |
| GB desolvation [c] | 0.63 | 1.51 |
| GB SA [c] | 1.61 | −1.29 |
| GB total [c] | 0.63 | 1.08 |
| SASA-ASP [c] | 1.53 | −1.60 |
| Effective solvent [c] | 0.93 | 1.77 |
| Main chain hydrogen bonding [c] | 1.01 | −1.16 |
| Side chain hydrogen bonding [c] | 0.97 | −2.05 |
| Centroid/backbone | | |
| Residue-environment (structural) [c] | 1.22 | 1.22 |
| Residue-residue (pair) [c] | 1.33 | 1.14 |
| Hard sphere repulsion [c] | 0.98 | −0.53 |
| Strand assembly in sheets [c] | 0.99 | −0.18 |
| Strand orientation [c] | 1.41 | −1.38 |
| Strand packing [c] | 1.38 | −0.98 |
| Helix-strand packing [c] | 1.04 | 0.45 |
| AvgDeg [d] | 1.07 | 0.56 |
| CO [d] | 1.18 | 0.81 |
| Ncompl [d] | 1.46 | 0.38 |
| Flow [d] | 0.85 | −0.08 |
| Conn [d] | 0.58 | −0.10 |
| flow [d] | 1.15 | 0.11 |
| Wconn [d] | 0.69 | −0.09 |

The graph properties (last seven rows) perform quite similarly to existing functions in terms of Enrichment. This is surprising if we take into account that such accuracy is obtained without any knowledge of the chemical information of the protein structure.

[a] From [20], computed using Victor/FRST software available at http://protein.cribi.unipd.it/frst/ (Accessed: 23 July 2008).

[b] From [8], computed as 1.

[c] From [6].

[d] This work, as in Table 2.

[5] http://predictioncenter.org/casp7/doc/categories.html (Accessed: 23 July 2008).

[6] http://predictioncenter.org/ (Accessed: 23 July 2008).

[7] http://www.predictioncenter.org/casp7/Casp7.html (Accessed: 23 July 2008).

Please cite this article in press as: Vassura M, et al. A graph theoretic approach to protein structure selection. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.07.016

**Table 5** Comparison of the best performing graph property Ncompl with the latest MQAPs and OSP on CASP7 data: average GDT_TS (Avg) for the best of the top five ranked models for each MQAP is shown, together with the corresponding standard deviation (std)

| | CASP7 | | | | | |
|---|---|---|---|---|---|---|
| | All | | Easy | | Hard | |
| | Avg | Std | Avg | Std | Avg | Std |
| Energy-based [a] | | | | | | |
| Pcons | 61.0 | 21.8 | 73.7 | 12.2 | 38.1 | 32.6 |
| ProQ | 61.2 | 21.3 | 73.5 | 11.8 | 38.9 | 31.9 |
| ProQprof | 54.3 | 22.9 | 68.0 | 12.9 | 29.5 | 34.3 |
| ProsaII | 61.3 | 21.2 | 73.4 | 12.1 | 39.6 | 31.6 |
| QA-ModFOLD | 60.2 | 21.1 | 72.3 | 11.7 | 38.4 | 31.6 |
| Verify3D | 61.1 | 20.8 | 73.1 | 11.6 | 39.5 | 31.2 |
| Graph-based [b] | | | | | | |
| Ncompl | 59.8 | 21.5 | 72.5 | 11.6 | 37.1 | 32.5 |
| Other | | | | | | |
| OSP [c] | 57.0 | 20.4 | 68.8 | 11.7 | 35.7 | 30.3 |
| GDT-TS | 64.5 | 19.9 | 75.9 | 11.3 | 43.9 | 29.7 |

GDT_TS shows the maximum score that a MQAP could obtain.
[a] From CASP7 except for Verify3D [23,24] and ProsaII [25].
[b] This work, threshold 9 Å.
[c] Described in [26—28].

a MQAP could obtain. It is interesting to note that Ncompl, which is computed from a single model without any knowledge of the chemical structure of the target protein, performs similarly to other methods using more information such as: primary structure, multiple sequence alignment, multiple structural alignment, consensus of the predictions (e.g. Pcons), neural networks (e.g. ProQ).

## 4. Conclusions and further work

In this paper we test several polynomial time computable graph properties to assess quality of predicted protein 3D structures. Our results indicate that it is possible to implement scoring functions capable of selecting near-native structures from a set of decoys or computed models by exploiting the graph representation of the 3D structure. Although the accuracy of some force fields is higher than our graph properties, it is very interesting to notice that our method achieves a comparable accuracy without having any knowledge of the chemistry of the protein chain, not even of the residue sequence. This finding further support the idea that simple backbone geometry is one of the most relevant piece of information, and suggests our method as a possible tool to improve performances of existing energy based functions. Furthermore the computational feasibility of the implemented graph properties makes them suitable filtering tools in wide-scale computing.

Further works may be done to test the ability of other, NP-hard, graph properties to act as MQAPs.

## Acknowledgements

We thank MIUR for the following grants: PNR 2001-2003 (FIRB art.8) and PNR 2003 projects (FIRB art.8) on Bioinformatics for Genomics and Proteomics and LIBI-Laboratorio Internazionale di BioInformatica, both delivered to RC. This work was also supported by the Biosapiens Network of Excellence project no LSHG-CT-2003-503265 (a grant of the European Unions VI Framework Programme).

## References

[1] Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CE, et al. Rosetta in CASP4: progress in ab initio protein structure prediction. Proteins 2001;45(Suppl. 5):119—26.
[2] Fariselli P, Casadio R. A neural network based predictor of residue contacts in proteins. Protein Eng 1999;12:15—21.
[3] Hardin C, Pogorelov TV, Luthey-Schulten Z. Ab initio protein structure prediction. Curr Opin Struct Biol 2002;12: 176—81.
[4] Park BH, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;2664:831—46.
[5] Felts AK, Gallicchio E, Wallqvist A, Levy RM. Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the Surface Generalized Born solvent model. Proteins 2002;48:404—22.

Please cite this article in press as: Vassura M, et al. A graph theoretic approach to protein structure selection. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.07.016

[6] Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D. An improved protein decoy set for testing energy functions for protein structure prediction. Proteins 2003;53:76—87.
[7] Finkelstein AV. Protein structure: what is it possible to predict now? Curr Opin Struct Biol 1997;7:60—71.
[8] Tosatto SCE. The Victor/FRST function for model quality estimation. J Comp Biol 2005;12:1316—27.
[9] Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002;47:142—53.
[10] Simons KT, 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 1997;268:209—25.
[11] Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;(Suppl. 3):171—6.
[12] Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985—94.
[13] Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to algorithms. London, UK: The MIT Press; 2001.
[14] Biggs N. Algebraic graph theory, vol. VIII. Cambridge, UK: Cambridge University Press; 1974.
[15] Galassi M, Davies J, Theiler J, Gough B, Jungman G, Booth M, et al. GNU Scientific Library Reference Manual, Revised Second Edition, ISBN 0954161734. Bristol, UK: Network Theory; 2006.
[16] Goldberg AV, Tarjan RE. A new approach to the maximum flow problem. J ACM 1988;35:921—40.
[17] Cherkassky BV, Goldberg AV. On implementing push-relabel method for the maximum flow problem. In: Balas E, Clausen J, editors. Proceedings of the 4th International IPCO Conference on Integer Programming and Combinatorial Optimization. London, UK: Springer-Verlag; 1995 . p. 157—71.
[18] Vendruscolo M, Dokholyan NV, Paci E, Karplus M. Phys Rev E 2002;65(6 Pt 1):061910.
[19] Greene LH, Higman VA. J Mol Biol 2003;334:781—9.
[20] Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895—916.
[21] Wallner B, Elofsson A. Prediction of global and local model quality in CASP7 using Pcons and ProQ. Proteins: Struct Funct Bioinform 2007;69(S8):184—93.
[22] Zemla A, Veclovas C, Moult J, Fidelis K. Processing and analysis of CASP3 protein structure predictions. Proteins Suppl 1999;3:22—9.
[23] Eisenberg D, Luethy R, Bowie J. VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol 1997;277:396—404.
[24] Luethy R, Bowie J, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature 1992;356:283—5.
[25] Sippl M. Recognition of errors in three-dimensional structures of proteins. Proteins 1993;17:355—62.
[26] Fleming PJ, Richards FM. Protein packing: dependence on protein size, secondary structure and amino acid composition. J Mol Biol 2000;299:487—98.
[27] Pattabiraman N, Ward KB, Fleming PJ. Occluded molecular surface: analysis of protein packing. J Mol Recogn 1995;8:334—44.
[28] Vorobjev YN, Hermans J. SIMS: computation of a smooth invariant molecular surface. Biophys J 1997;73:722—32.

Please cite this article in press as: Vassura M, et al. A graph theoretic approach to protein structure selection. Artif Intell Med (2008), doi:10.1016/j.artmed.2008.07.016