Bioinformatics, 31(21), 2015, 3499–3505
doi: 10.1093/bioinformatics/btv390
Advance Access Publication Date: 30 June 2015
Original Paper

Structural bioinformatics

GDFuzz3D: a method for protein 3D structure reconstruction from contact maps, based on a non-Euclidean distance function

Michal J. Pietal[1,2], Janusz M. Bujnicki[1,3,*] and Lukasz P. Kozlowski[1]

[1]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland, [2]Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, Warsaw, Poland and [3]Bioinformatics Laboratory, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University, Poznan, Poland

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on October 12, 2014; revised on June 7, 2015; accepted on June 23, 2015

Abstract

Motivation: To date, only a few distinct successful approaches have been introduced to reconstruct a protein 3D structure from a map of contacts between its amino acid residues (a 2D contact map). Current algorithms can infer structures from information-rich contact maps that contain a limited fraction of erroneous predictions. However, it is difficult to reconstruct 3D structures from predicted contact maps that usually contain a high fraction of false contacts.

Results: We describe a new, multi-step protocol that predicts protein 3D structures from the predicted contact maps. The method is based on a novel distance function acting on a fuzzy residue proximity graph, which predicts a 2D distance map from a 2D predicted contact map. The application of a Multi-Dimensional Scaling algorithm transforms that predicted 2D distance map into a coarse 3D model, which is further refined by typical modeling programs into an all-atom representation. We tested our approach on contact maps predicted *de novo* by MULTICOM, the top contact map predictor according to CASP10. We show that our method outperforms FT-COMAR, the state-of-the-art method for 3D structure reconstruction from 2D maps. For all predicted 2D contact maps of relatively low sensitivity (60–84%), GDFuzz3D generates more accurate 3D models, with the average improvement of 4.87 Å in terms of RMSD.

Availability and implementation: GDFuzz3D server and standalone version are freely available at http://iimcb.genesilico.pl/gdserver/GDFuzz3D/.

Contact: iamb@genesilico.pl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Predicting protein tertiary structure from amino acid sequence has been a major challenge in structural biology for over four decades. It was proposed by Anfinsen that in the native environment, all proteins assume their tertiary structure spontaneously and this process is solely determined by the amino acid sequence (Anfinsen *et al.*, 1961). Still, predicting the 3D structure accurately from its amino acid sequence remains a formidable challenge. Contemporary protein structure prediction protocols are built upon two distinct approaches: comparative modeling (typically based on the evolutionary relationship between the target sequence and a sequence of another protein with known structure, which can be used as a template) and folding simulation (typically based on a search of the conformational space, with a scoring function that attempts to identify 'native-like' conformations for the polypeptide chain). Both approaches can benefit from the use of additional data encoded as

spatial restraints, which can be derived from experimental analyses or from independent predictions.

One of the commonly considered types of additional data useful for protein 3D structure prediction are the spatial proximities between individual amino acid residues. Information about the proximity of residue or atom pairs in the molecule can be represented as a square symmetric matrix (Phillips, 1970). Values stored in such a matrix may represent the Euclidean distances between particular atoms and such a matrix is then called a (Euclidean) distance map. If only binary information about residue-residue interactions is included (e.g. qualified by a Euclidean distance below a given threshold), that matrix is called a contact map. The majority of the contact map processing approaches, especially those using contacts in protein 3D structure prediction, use Euclidean metric functions based on the distances between the Cα or Cβ atoms of the residues with the contact threshold set to 8 Å.

A protein 3D structure can be accurately predicted based on information about distances between its individual atoms, encoded as spatial restraints (Sali and Blundell, 1993). It can be also modeled based on its binary contact map, with the reconstructed and original structures similar up to the resolution of the contact map representation (Vendruscolo *et al.*, 1997; Vendruscolo and Domany, 2000). The pioneering method for protein 3D structure reconstruction from 2D contact maps developed by Vendruscolo *et al.* was tested and proven to work with the native (100% accurate) contact maps, as well as with maps into which up to 4% of true contacts were replaced by erroneous ones. A more recent method FT-COMAR enabled 3D structure reconstruction from less accurate contact maps. It was shown that the quality of 3D reconstruction with FT-COMAR is unaffected by deleting up to an average 75% of the real contacts (if the remaining 25% contains no errors) while indeed only a small percentage of randomly generated (wrong) contacts in place of non-contacts are sufficient to hamper 3D reconstruction (Vassura *et al.*, 2008, 2011). A detailed error-rate comparison implies that 8 Å RMSD model quality limit is upheld by Vendruscolo method if the random error rate is about 6% or less, while in FT-COMAR an error rate of 16% or lower can be tolerated.

It must be emphasized that information about chirality in 3D (handedness) is lost upon conversion of a protein structure from 3D to a 2D representation. The determination of the actual biological structure among the two mirror image models that are in agreement with the map is not trivial, but can be achieved. Typically, if the protein contains α-helices, a model with biologically relevant right-handed helices can be selected. Further, the need to instantiate biologically relevant stereoisomeric forms of individual residues may allow the discrimination of models with native-like versus mirrored topologies using scoring functions for protein model quality assessment at the stage of all-atom reconstruction of the 3D model (Kryshtafovych and Fidelis, 2009).

The conservation of 3D structures in globular protein domains imposes strong constraints on amino acid residues. In each protein, favorable interactions are formed between residues with appropriate side chains (e.g. disulfide bridges between cysteine residues, salt bridges between charged residues, hydrophobic interactions between aliphatic residues, etc.). Favorable interactions tend to be preserved in evolution, resulting in correlations among amino acid compositions at different sequence positions when aligned sequences of homologous proteins are considered. Both physico-chemical and evolutionary considerations can be exploited to infer spatial contacts for proteins with unknown structure. Currently, the computational prediction of direct contacts is considered much easier than the prediction of distances between residues. There are many methods that predict contacts between amino acid residues from protein sequence; e.g. NNcon (Tegge *et al.*, 2009), SVMcon (Cheng and Baldi, 2007), Possum (Hamilton *et al.*, 2004), Psicov (Jones *et al.*, 2012), PconsC (Skwark *et al.*, 2013) and MULTICOM (Wang *et al.*, 2010). Typically, they generate a map with predicted contact probabilities for all possible pairs of residues, which can be used as a starting point for 3D structure reconstruction. For a typical structure of a protein of length L, the total number of unique contacts defined as distances between the Cα atoms ≤ 8 Å, is about 4.5–5.0 L (data not shown). A typical predicted map contains a much larger number of tentative contacts; however, a more realistic map (with a native-like number of contacts) may be inferred by taking 4.5 L contacts predicted with highest probability scores.

Over last 20 years, significant effort has been made to benchmark and assess methods for predicting protein structure from amino sequence, in particular in the framework of the biannual Critical Assessment of Protein Structure Prediction (CASP) experiment (Moult *et al.*, 2014). Methods for contact prediction were evaluated in the RR category of CASP. A common evaluation metric for residue-residue contact predictions is the accuracy of only the top L/5, L/10 or 5 predictions scored best by a given method. Accuracy is defined as the number of correctly predicted residue-residue contacts divided by the total number of contact predictions considered (Graña *et al.*, 2005). Unfortunately, the overall contact prediction quality is still at a disappointingly low level as compared with the required levels of 3D structure retrieval protocols. Recently, novel quality measures were introduced to score predictors in CASP10 (Monastyrskyy *et al.*, 2014). MULTICOM was able to predict contacts with best quality levels, defined as a mutual score of sensitivity (recall): 31.4% and precision (accuracy) 6.9% when considering predictions for long-range contacts (i.e. residue pairs separated in the sequence by at least 24 other residues). However, these predictions of contact maps that may be considered state-of-the-art, are significantly worse than the minimal requirements by the state-of-the-art methods for 3D protein structure reconstruction from 2D contacts.

Here, we present a novel approach for predicting 3D structures from 2D maps. It is based on two components: first, a new method for predicting 2D distance maps from 2D contact maps and second, a modeling protocol that involves a combination of coarse-grained and all-atom modeling. Our method was developed to use predicted contact maps that include probabilities of contacts for all residues and can be used in *de novo* 3D structure prediction, based on the output of various existing contact map prediction programs. We implemented this protocol as a publicly available web server GDFuzz3D that takes a contact map as an input and reports an all-atom 3D model of the protein structure, as well as a predicted 2D distance map.

2 Methods

A typical binary contact map lacks sufficient information to be perceived as a comprehensive 2D alternative to a 3D protein model. On the other hand, a Euclidean distance map can be treated as a detailed 2D representation, which can be relatively easily converted into a 3D model. Thus, we developed an algorithm with which to reconstruct a Euclidean distance map from a 2D contact map. Our protocol draws from a concept similar to that described by (Tenenbaum *et al.*, 2000) and applies it to protein structures.

2.1 Graph distance map: definition and features

A contact map can be regarded as an adjacency matrix that represents a residue proximity graph. In this graph, each vertex is a single-point representation of a residue (typically a Cα atom). Edges between vertices (residue pairs) can be deduced or just read out from the contact map (as respective contacts). If any two residues are in contact, there also exists an edge connecting the respective pair of vertices. The graph thus represents the mathematical relation of spatial proximity for all residue pairs in 3D space. In this case, an edge has uniform weight set to 1; however, graph edges can bear different weights and this feature will be used later, for the purpose of graph generalization. Our new distance function is defined as a graph distance function applied to the connectivity graph. Such distance is measured for all possible residue pairs, also for pairs other than those that are in contact. The distance is defined as the shortest path length among all paths for any given residue pair. In the connectivity graph, the path length is calculated by summing the total number of edges connecting the two residues under consideration. This function isn't however a Euclidean distance, and such distances only approximate real distances in 3D. The proposed distance function has natural numbers as values, because the graph distance (edge weight) between a residue pair contacting in 3D is uniformly set to 1. An example graph derived from protein Cα backbone fragments as well as its respective graph distance matrix, are shown in Figure 1.

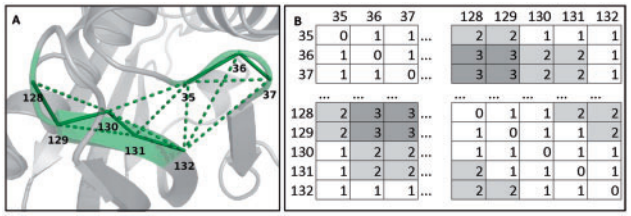

Fig. 1. Visualization of the graph distance map definition for a protein backbone fragment. (**A**) Two backbone fragments are shown as solid lines with the Cα atoms numbered according to the residue position in sequence. Residues in contact (distance ≤8 Å) are connected by vertices shown as dotted lines. (**B**) A graph distance matrix calculated for the residues from the highlighted fragment shown in (A). Distance values greater than 1 are defined as the length of the shortest path connecting respective residue pair in the connectivity graph shown in (A)

Fast numerical routines exists for calculating the graph distance map given a contact map. For example, graph-based algorithms are implemented in the Mathematica package (Wolfram Research, 2015). For a typical protein 200 residues long, the graph distance calculation using this algorithm takes usually below 1 s. An example visualization of a protein contact map, the graph distance map and a Euclidean distance map is shown in Figure 2. The current implementation of our approach assumes that we deal with protein structures that are continuous in space, i.e. are formed from elements that are in physical contact with each other. For such structures, the graph must be connected (as defined in graph theory), such that all residue pairs are connected by a finite path. It is important to note that a gapped graph indicates spatially isolated fragments.

2.2 Multi-dimensional scaling of maps

The example shown in Figure 2 suggests that the graph distance map calculated from the contact map can be considered an approximation of a Euclidean distance map of a given protein. However, the values of the graph distance map are integer numbers, while the real, Euclidean distance map contains real values. We used a Multi-Dimensional Scaling (MDS) algorithm (Kruskal, 1964) in order to transform the graph distance matrix into a corresponding real value distance matrix that should approximate the real Euclidean distance map. The algorithm was set to operate in 3D, which also yields the proposed 3D representation of points corresponding to the final Euclidean distance map.

Directly before running the MDS algorithm, a transition from the unit-less graph distances to the corresponding Euclidean distances (in Ångströms) is made. This is made by using the following formula for a graph distance value of n:

$$d_n = n * d_1$$

where the expected distance value on the right side of the equation equals that for contacts, which in turn is based on statistical measurements on a dataset derived from the PDB (Berman *et al.*, 2000); the procedure is described in detail in Supplementary material (see also equation E1 therein).

The graph distance map values, being discrete natural numbers (0, 1, 2, . . .), are replaced by real values according to the above formula (e.g. all occurring graph distance values of 1 are replaced by

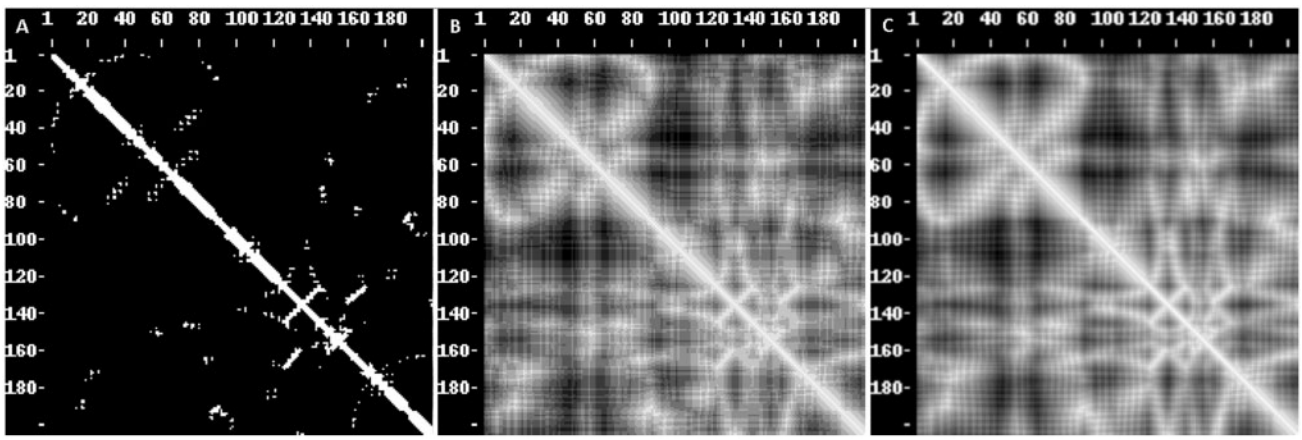

Fig. 2. Visualization of the three types of 2D maps for protein molecules. (**A**) Contact map of manganese(III) superoxide dismutase (PDB id: 1IX9, chain A), calculated for Cα atoms, using an 8 Å threshold. Contacts are shown as white dots. (**B**) Graph distance map with discrete, natural distance values, derived solely based on the contact information from (A). (**C**) Euclidean distance map of the same protein, calculated for Cα atoms. In both (B) and (C) the relative distance is scaled from zero (white) to maximal observed distance (black). All images were generated using PROTMAP2D (Pietal *et al.*, 2007)

5.72 Å). This step is made before MDS in order to make the algorithm transform a crude 2D map into an approximate Euclidean distance map. Thus, a 2D map has the same number of distance values as the graph distance map. However, it contains only several distinct real values. As an example, a Euclidean distance map of a real protein structure of 200 residues contains 19 800 unique real entries. An approximate Euclidean distance map obtained by the MDS procedure is thus a more realistic and detailed representation of a protein structure.

2.3 Graph distance generalization for processing predicted contact maps

The generalization of the graph distance map is achieved by introducing probabilities of contacts as weights of edges in the connectivity graph. The protocol remains the same except for the fact that the graph attains weights as contact probabilities taken from the input 2D map. The map can be a typical output of any contact map prediction method. The only difference is that during calculating paths in the graph, the number of paths between two vertices (i,j) can be fractional, because of processing the probabilities (we deal with expected path lengths). The resulting graph distance map shares the same features with the one derived for binary contact map processing. See Supplementary material for additional details.

An important assumption for the graph distance map calculation algorithm is to have a stop feature: while it may seem trivial in the case of native contact maps, it can be problematic for predicted contact maps that may lack some 'obvious' contacts. If the contact probability for the nearest neighboring (NN) residue $(i,i+1)$, the next NNN $(i,i+2)$ and so on, is less than 0.5, this might cause the algorithm to never stop. It is important that the whole graph is connected, so subsequent residues (NN, NNN, etc.) are assigned a contact probability 1.0 which is true for biological sequences, but not necessarily true for the predicted contact maps.

2.4 3D stage: modeling protocol

The MDS algorithm generates 3D coordinates of a crude Cα model and its mirror image, which both represent the same approximate Euclidean distance map. In either of the crude models, most neighboring Cα distances are non-physical. This is because as the MDS protocol tries to optimize all distance pairs, this often causes the distances between subsequent Cα deviate from the physical value of 3.8 Å. To address this problem, we implemented a simple correction algorithm (see Supplementary Material for details).

Following the initial refinement of a Cα model, we generate an all-atom model using MODELLER (Sali *et al.*, 1995), with restraints on Cα-Cα distances limited to residues with graph distances ≤ 4 and on secondary structure predicted by SSPro4 (Cheng *et al.*, 2005). Because MODELLER accepts distance restraints as normal distribution parameters, we use such parameters (which are: expected distance value and standard deviation) separately for each restraint. Those are the same values as used in transition from graph distance values to Ångströms (see Supplementary Material). The purpose of using MODELLER, is to generate an all-atom model that presents reasonable stereochemistry and packing.

Subsequently, we continue the refinement with the REFINER program (Boniecki *et al.*, 2003), with similar restraints as in the previous step, only with distance restraints further limited to residues with originally predicted graph distance ≤ 3. The purpose of using REFINER is in the use of a knowledge-based force-field that has been developed specifically to refine protein structures to improve the formation of native-like residue-residue interactions, and, for example, to guide the formation of proper β-sheets from neighboring β-strands. We set REFINER to operate in a 'burial mode', which regularizes 3D packing and removes some of other artifacts produced by MODELLER. Both 3D models generated by MDS are processed in parallel, until the refined version of one of them is finally selected according to the REFINER score.

MDS outputs two Cα models, which are mirror images of each other. This is a typical feature of correspondence between the 2D and 3D representation of the same protein. This notion comes from the fact that any distance function is invariant to isometric transformations of a 3D object: translation, rotation and symmetry. Since initially it is not obvious which of the two structures related by mirror symmetry represents a biological structure, we apply MODELLER and REFINER to refine both variants with the assumption of biologically relevant handedness on the local level (e.g. L-amino acids, right-handed α-helices etc.). Thereby, for models of wrong handedness, we introduce a mismatch between global and local handedness, and the refinement often causes models of the wrong global handedness to be poorly folded compared with their counterparts with the correct global handedness. In our experience, the scoring of refined models by REFINER typically allows selection of a model with the proper handedness.

2.5 3D stage: refinement protocol

The conformation selected as potentially best by REFINER is ultimately processed with MMTSB Rebuild (Feig *et al.*, 2004) to generate an all-atom 3D model. In the last step, MODELLER is run with restraints on the secondary structure only to alleviate steric clashes that might arise as a result of conversion from a coarse-grained to all-atom representation.

The modeling protocol is illustrated in Figure 3 and has been implemented as a web server GDFuzz3D that takes as an input a protein sequence and a predicted contact map and outputs an all-atom 3D structural model.

3 Results

3.1 GDFuzz3D comparison with FT-COMAR

We tested our method by comparing it directly against FT-COMAR, the state-of-the-art method for 3D structure reconstruction from 2D maps. For this purpose, we selected maps generated by the MULTICOM predictor for single-domain targets analyzed in the CASP10 experiment (Moult *et al.*, 2014). This dataset comprised 45 targets, with sensitivity (recall) of predicted contact maps ranging from 0 to 0.84 (average sensitivity 0.30), including only 10 maps with sensitivity >0.5. This range of sensitivity values (mostly very bad to some that are reasonable but far from perfect) can be expected for 'real life' situations, i.e. prediction tasks encountered by typical users of protein structure prediction methods.

For this dataset, FT-COMAR was able to generate 10 models with TM-score > 0.5, i.e. with a correct 3D fold (Supplementary Table S1). GDFuzz3D was able to generate correct fold predictions for 16 targets, including all 10 correctly predicted by FT-COMAR. On average, models generated by GDFuzz3D were more accurate, with an average TM-score of 0.41 compared with 0.31 for models returned by FT-COMAR (average RMSD: 11.06 Å for GDFuzz3D versus 14.88 Å for FT-COMAR). GDFuzz3D was able to generate 5 models with RMSD to the reference structure <3.5 Å, while none of the FT-COMAR models met this criterion. If the RMSD threshold is relaxed to 5 Å, GDFuzz3D scores 9 models compared with 5 by FT-COMAR. Supplementary Figures S3 and S4 illustrate examples of successful predictions made by GDFuzz3D.

If only 10 'reasonable' contact maps (sensitivity > 0.5) are taken into account, GDFuzz3D is able to generate 6 out of 10 models with TM-score > 0.5 which indicates correct fold prediction, and eight models generated by GDFuzz3D are significantly better than those of FT-COMAR, and two are of essentially the same accuracy. The average RMSD value for a model generated by GDFuzz3D for these maps is 6.15 Å, while the average RMSD for FT-COMAR is 11.02 Å, and the average TM-score is 0.51 for GDFuzz3D versus 0.39 for FT-COMAR.

3.2 GDFuzz3D comparison with PconsFold

While we prepared our results for publication, another protocol for protein 3D structure prediction based on contact maps was proposed: PconsFold (Michel *et al.*, 2014) uses PconsC as a contact map predictor (Skwark *et al.*, 2013), and ROSETTA (Leaver-Fay *et al.*, 2011) as a method for 3D structure prediction with restraints derived from predicted contact maps. PconsFold was tested on several datasets, of which the largest was the PSICOV dataset (Jones *et al.*, 2012) that consists of 150 single-domain proteins with sequence lengths between 52 and 266 residues. PconsFold was also compared with the EVfold protocol (Morcos *et al.*, 2011; Marks *et al.*, 2012) and was shown to yield more accurate models (Michel *et al.*, 2014). It must be emphasized that the PconsC maps for the PSICOV dataset are of above-average quality, with sensitivity ranging from 0.33 to 0.72, and average sensitivity of 0.55 (i.e. much better than the relatively unbiased set of predictions generated by MULTICOM during CASP10).

Detailed results of comparison between models generated by GDFuzz3D and PconsFold for 150 proteins of the PSICOV dataset and contact maps predicted by PconsFoldC are shown in Supplementary Table S4. Models generated by GDFuzz3D have an average TM-score of 0.49 and average RMSD to the reference structure of 8.2 Å, which is slightly inferior to best models reported by PconsFold (average TM-score 0.55 and average RMSD 7.4 Å), and they are slightly better than models returned by the EVfold protocol (average TM-score 0.47). In this exercise, according to RMSD, 58 models generated by GDFuzz3D were more accurate than models generated by PconsFold, and 92 models generated by PconsFold were better. For example, GDFuzz3D was able to generate a better prediction than PconsFold for the target 1jwq (Michel *et al.*, 2014). There, GDFuzz3D achieved a RMSD of 3.3 Å and TM-score of 0.77 compared with RMSD of 5.1 Å and TM-score of 0.62 for a model generated by PconsFold. Figure 4 illustrates the relative performance of GDFuzz3D, FT-COMAR, and PconsFold, in terms of TM-score of the models depending on the sensitivity of the starting map and the results are summarized in Table 1.

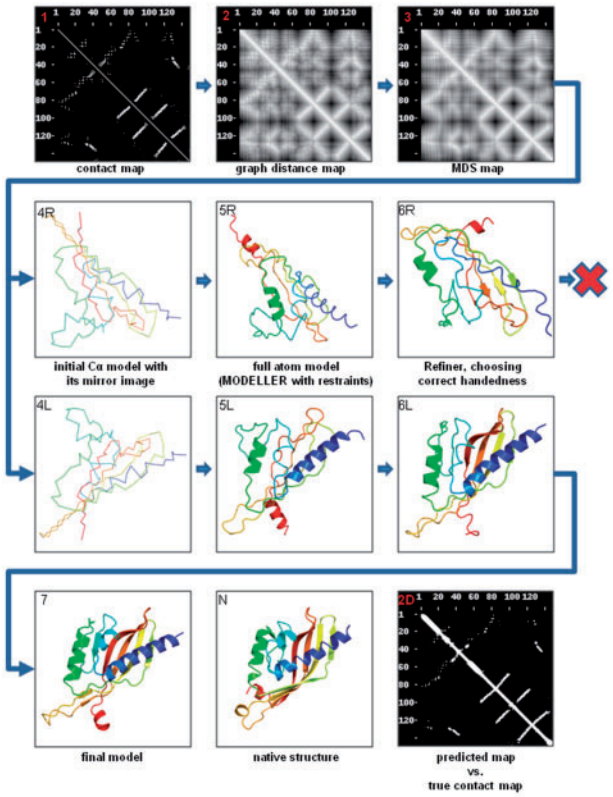

Fig. 3. Schematic data flow of GDFuzz3D. GDFuzz3D workflow with T0752 CASP10 model as an example: (1) contact map prediction with MULTICOM; (2) graph distance map; (3) rescaled, MDS-refined graph distance map; (4 L&4 R) an initial Cα model with fixed Cα neighboring distances and its mirror image; (5 L&5 R) all-atom models generated by MODELLER (with right-handed helices); (6) models optimized by REFINER; (7) top-scoring structure locally optimized with MODELLER; (N) reference structure 4GB5 from PDB. (2D) 2D comparison (Cα, 8 Å definition) between the native structure (lower part of the map) and the model (upper part). RMSD of our model to the reference is 4.8 Å, TM-score is 0.76 and original map sensitivity (recall) is 0.12 (any sequence separation). (See Supplementary Fig. S5 for additional details)

3.3 Improvement of predictions in the course of modeling with GDFuzz3D

We tested to which extent the contact maps improve in the course of modeling by GDFuzz3D. Table 2 and Supplementary Table S5 show that for the PSICOV dataset our method was able to improve the quality of contact maps: the generation of graph distance maps and its subsequent optimization by the MDS procedure increased the average contact map recall from 0.550 to 0.685, and it further increased to 0.785. Accuracy of the maps has also improved at the 3D modeling stage. On the other hand, we were not able to achieve such improvement for the more noisy input maps generated by the MULTICOM method (not shown).

We have also tested to which extent the use of the graph distance function improves 3D structure prediction accuracy compared with

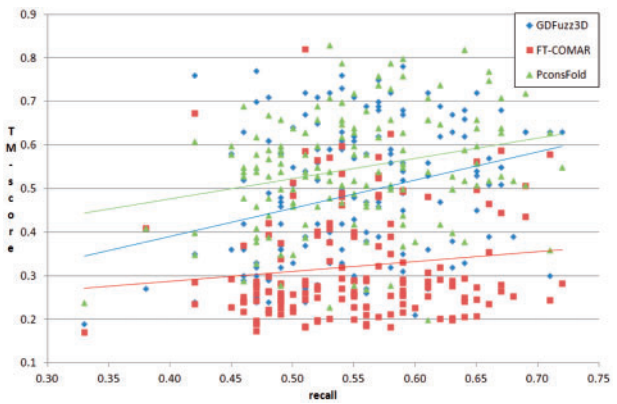

Fig. 4. Final model quality in a form of a TM-score as a function of initial contact map recall, between models generated by GDFuzz3D and those by FT-COMAR and PconsFold tested on the PSICOV dataset (150 proteins). GDFuzz3D performs clearly better than FT-COMAR on a wide range of sensitivity values, and it performs slightly worse than PconsFold on a restricted range of values of the PSICOV dataset. But as the recall of the map improves, GDFuzz3D models catch up with those by PConsFold. For additional comparisons see Supplementary Figure S6

the use of the contacts alone. Table 3 demonstrates that the use of the graph distance improves average model quality in terms of RMSD and TM-score. The improvement is minor for the relatively accurate maps of the PSICOV test set, but it increases for the less accurate maps of the MULTICOM46 test set.

4 Discussion

We developed a method for predicting protein 3D structures from 2D contact maps, which is based on a novel approach for predicting distance maps from predicted 2D contact maps. The main novelty is in the concept of a non-Euclidean 2D distance map, which can be derived directly (and unambiguously) from a protein contact map. The graph distance reflects a series of physical contacts between the residues, and thereby indicates the length of the path of the signal produced by a perturbation of a given residue (due to a substitution or due to any physical interaction with the environment) until it affects other residues in the protein. The graph distance matrix can be therefore treated as a map of perturbation propagation pathways, which describes in detail to which extent a single residue substitution may impact other residues in protein structure.

Most of existing methods for predicting protein 3D structures from 2D contact maps are very sensitive to errors. Currently, the best methods for 2D contact map predictions perform rather poorly on average, and only a small fraction of their predictions have high accuracy. Typical predicted contact maps contain a non-physical number of contacts and they include a large number of contacts that are erroneous. Our intention was to develop a method that overcomes the problem of non-physicality of typical predicted contacts maps, and to handle input maps with probabilities of contacts for all residues. Such maps contain a very high number of predicted contacts, of which many are erroneous, and which often contradict one another. Our algorithm does not tackle cases where only a few highly relevant contacts are known, but this task can be handled by standard methods for protein 3D structure prediction by folding with restraints, such as ROSETTA (Thompson and Baker, 2011) or CABS (Latek *et al.*, 2007).

We tested our 3D structure prediction protocol on contact maps predicted with MULTICOM (the best method for contact map prediction according to CASP10) and found that it generates more accurate 3D models than those produced by FT-COMAR, on maps with a wide range of sensitivity values. We have also tested it on contact maps predicted by PconsC (with a restricted range of sensitivities) and demonstrated that models generated by GDFuzz3D are on the average almost as accurate as those generated by computationally demanding PconsFold protocol.

Our method is modular and elements of our protocol can be incorporated into other protocols for protein 3D structure prediction from contact maps. In particular, the 3D modeling methodology we used (combination of MODELLER and REFINER) can be easily replaced any other method for structure prediction that uses distance restraints. In particular, the use of extensive conformational sampling (as in PconsFold with ROSETTA) and/or the use of alternative methods for model quality assessment (and identification of candidate structures from many decoys) may further improve the accuracy of predicted structures. In particular, we envisage that elements of our protocol could be combined with elements of the PconsFold protocol to generate even better predictions. On the other hand, the choice of a method with which to generate an input contact map can be a significant factor in the procedure. It would be interesting to evaluate various combinations of methods for contact map generation, conversion of distance maps to contact maps, and protein 3D folding with restraints.

We envisage potential applications of graph distance maps beyond protein 3D structure prediction. One possible application is in protein contact map alignment, flexible superposition of protein structures, and structure-based sequence alignment. Algorithms that use contact maps operate on sparse matrices, in which the majority of entries represent non-contacts. The use of graph distance maps, which can be obtained both from experimentally determined protein 3D structures and from predicted contact maps, would allow for the use of more efficient algorithms developed; e.g. for distance matrix comparison, as in DALI (Holm and Sander, 1993). This approach could facilitate comparison of proteins with known structures with those with unknown structures, for which only theoretical predictions (including predicted contact maps) are available.

Another possible application of graph distance maps include RNA 3D structure modeling. Most methods for RNA 3D structure prediction developed to date utilize similar strategies to those developed for protein 3D structure modeling (Rother *et al.*, 2011). RNA 3D structure prediction is usually guided by secondary structure prediction, which identifies canonical Watson-Crick base pairs

Table 1. Models with correct fold (TM score > 0.5)

Method	GDFuzz3D (%)	PconsFold	FT-COMAR (%)
Targets with TM-score > 0.5 (PSICOV test set)	48.7	66.7%	10.7
Targets with TM-score > 0.5 (MULTICOM test set)	34.8	N/A	21.7

Table 3. Comparison of a full GDFuzz3D procedure ('graph distance') with a simplified version ('contacts-only'), where only restraints on contacts are used with MODELLER

Measure	Contacts only	Graph distance	Improvement	Improvement (% of better models)
MULTICOM46				
RMSD	11.39	11.06	0.33	65.2%
TM-score	0.397	0.414	0.017	73.9%
PSICOV				
RMSD	8.36	8.21	0.15	52.7%
TM-score	0.482	0.487	0.005	62.7%

Note: RMSD—the smaller the better, TM-score—the higher the better (details in Supplementary Tables S4 and S5).

Table 2. Improvement of 2D maps in the course of GDFuzz3D modeling, shown for the PSICOV dataset (details in Supplementary Table S3)

Measure	PSI COV	GDF3D, after 2D stage	GDF3D, after 3D stage	2D stage improvement.	2D + 3D improvement
RECALL	0.550	0.685	0.785	0.135	0.235
ACCURACY	0.532	0.576	0.597	0.044	0.064

between ribonucleotide residues. However, there exist methods such as SHEVEK (Pang *et al.*, 2005), MC-Fold (Parisien and Major, 2008) or RNAwolf (zu Siederdissen *et al.*, 2011) that predict other types of contacts in RNA structures, including non-canonical base pairs. Thus, the approach delineated here for protein structure modeling could be used to improve RNA 3D structure modeling by the inference of RNA graph distance maps. Last, but not least, some elements of the methodology described in this article could be adapted to help generation of coarse-grained 3D models of chromatin structure based on long-range contact information from experiments such as Hi-C (Dekker *et al.*, 2013).

5 Conclusions

We developed a novel approach for predicting protein 3D structures, by converting protein residue contact maps into protein residue distance maps, and then into all-atom models. We implemented it as a publicly available web server GDFuzz3D, which accepts a predicted contact map as an input and generates a corresponding 3D structural model. The novel approach to calculating non-Euclidean distance maps from contact maps may find other uses beyond protein 3D structure prediction, such as flexible alignment of protein sequences and structures, modeling of RNA structure, and modeling of chromatin structure.

Acknowledgements

M.J.P. thanks Jacek Koronacki for useful discussions. We thank Grzegorz Chojnowski, Joanna Kasprzak, Tomasz Puton, Michal Boniecki and Wayne Dawson for critical reading of the manuscript and useful comments.

Funding

This project was supported by the European Research Council (grant RNA+P=123D). M.J.P. was supported by the European Social Fund through Subcarpathian Doctoral Stipend Fund and later by the Polish National Science Center (grant 2013/09/B/NZ2/00121). J.M.B. was supported by the 'Ideas for Poland' fellowship from the Foundation for Polish Science. The development of the GDFuzz3D web server was supported by the EU structural funds (grant POIG.02.03.00-00-003/09).

Conflict of Interest: none declared.

References

Anfinsen,C.B. *et al.* (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci USA*, **47**, 1309–1314.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Boniecki,M. *et al.* (2003) Protein fragment reconstruction using various modeling techniques. *J. Comput. Aided Mol. Des.*, **17**, 725–738.

Cheng,J. and Baldi,P. (2007) Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, **8**, 113.

Cheng,J. *et al.* (2005) SCRATCH: a protein structure and structural feature prediction server. *Nucleic Acids Res.*, **33**, W72–W76.

Dekker,J. *et al.* (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.*, **14**, 390–403.

Feig,M. *et al.* (2004) MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graph. Model.*, **22**, 377–395.

Graña,O. *et al.* (2005) CASP6 assessment of contact prediction. *Proteins Struct. Funct. Bioinf.*, **61**, 214–224.

Hamilton,N. *et al.* (2004) Protein contact prediction using patterns of correlation. *Proteins*, **56**, 679–684.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Jones,D.T. *et al.* (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190.

Kruskal,J. (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, **29**, 1–27.

Kryshtafovych,A. and Fidelis,K. (2009) Protein structure prediction and model quality assessment. *Drug Discov. Today*, **14**, 386–393.

Latek,D. *et al.* (2007) Protein structure prediction: combining de novo modeling with sparse experimental data. *J. Comput. Chem.*, **28**, 1668–1676.

Leaver-Fay,A. *et al.* (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.*, **487**, 545–574.

Marks,D.S. *et al.* (2012) Protein structure prediction from sequence variation. *Nat. Biotechnol.*, **30**, 1072–1080.

Michel,M. *et al.* (2014) PconsFold: improved contact predictions improve protein models. *Bioinformatics*, **30**, i482–i488.

Monastyrskyy,B. *et al.* (2014) Evaluation of residue–residue contact prediction in CASP10. *Proteins Struct. Funct. Bioinf.*, **82**, 138–153.

Morcos,F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA*, **108**, E1293–E1301.

Moult,J. *et al.* (2014) Critical assessment of methods of protein structure prediction (CASP)—round X. *Proteins Struct. Funct. Bioinf.*, **82**, 1–6.

Pang,P.S. *et al.* (2005) Prediction of functional tertiary interactions and intermolecular interfaces from primary sequence data. *J. Exp. Zool. B*, **304**, 50–63.

Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.

Phillips,D.C. (1970) The development of crystallographic enzymology. *Biochem. Soc. Symp.*, **30**, 11–28.

Pietal,M.J. *et al.* (2007) PROTMAP2D: visualization, comparison, and analysis of 2D maps of protein structure. *Bioinformatics*, **23**, 1429–1430.

Rother,K. *et al.* (2011) RNA and protein 3D structure modeling: similarities and differences. *J. Mol. Model.*, **17**, 2325–2336.

Sali,A. and Blundell,T. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Sali,A. *et al.* (1995) Evaluation of comparative protein modeling by MODELLER. *Proteins*, **23**, 318–326.

Skwark,M.J. *et al.* (2013) PconsC: combination of direct information methods and alignments improves contact prediction. *Bioinformatics*, **29**, 1815–1816.

Tegge,A.N. *et al.* (2009) NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Res.*, **37**, W515–W518.

Tenenbaum,J.B. *et al.* (2000) A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319–2323.

Thompson,J. and Baker,D. (2011) Incorporation of evolutionary information into Rosetta comparative modeling. *Proteins*, **79**, 2380–2388.

Vassura,M. *et al.* (2008) FT-COMAR: fault tolerant three-dimensional structure reconstruction from protein contact maps. *Bioinformatics*, **24**, 1313–1315.

Vassura,M. *et al.* (2011) Blurring contact maps of thousands of proteins: what we can learn by reconstructing 3D structure. *BioData Min*, **4**, 1.

Vendruscolo,M. and Domany,E. (2000) Protein folding using contact maps. *Vitam Horm*, **58**, 171–212.

Vendruscolo,M. *et al.* (1997) Recovery of protein structure from contact maps. *Fold.Des.*, **2**, 295–306.

Wang,Z. *et al.* (2010) MULTICOM: a multi-level combination approach to protein structure prediction and its assessment in CASP8. *Bioinformatics*, **26**, 882–888.

Wolfram Research, Inc. (2015) Mathematica, Version 10.1, Champaign, IL.

zu Siederdissen,C.H. *et al.* (2011) A folding algorithm for extended RNA secondary structures. *Bioinformatics*, **27**, i129–i136.