Bioinformatics
Structural Bioinformatics Manuscript
Advance Access Submission Date: 5 July 2021
Manuscript Category

Structural Bioinformatics

NMR Data for ab initio protein structure prediction

Christina Kaitatzi [1,*], Merkouris Papamichail [2,*] and Charalambos Tzamos [2,*]

[1] Department of Physics, National and Kapodistrian University of Athens, Athens, 15784, Greece and
[2] Department of Informatics & Telecommunications, National and Kapodistrian University of Athens, Athens, 15784, Greece.

*To whom correspondence should be addressed.

Author names have been set in alphabetical order.

Abstract

Motivation: Three-dimensional protein structure prediction by using NMR input has been a very active field in bioinformatics. There are many methods that approximate solutions for this problem. In this study, we propose an augmentation of an existing numerical method. The approach is based on a distance geometry formulation. We also propose a different representation of the input, that is to split the protein into parts, instead of dealing with it as one.

Results: Our model improves significantly the computational complexity of the existing method by Emiris and Nikitopoulos (2005), while not losing significantly in accuracy. Our observations are based on experimental results.

Availability: `https://github.com/merkouris148/nmr-structure-prediction`

Contact: ckaitatzi@phys.uoa.gr or m.papamichail@di.uoa.gr or ctzamos@di.uoa.gr

1 Introduction

The 3d structure of a protein is a set of points in $\mathbb{R}^3$, such that each point corresponds to an atom of the protein. In nature, not every embeddable 3d structure is valid for a protein. Depending on the amino-acid sequence and its bonds, the protein may fold in various ways in order to minimize its free energy. Observe that, from a geometric point of view, folding means that certain parts of the protein comes closer in $\mathbb{R}^3$, despite being far apart in the amino-acid sequence. Nuclear magnetic resonance, also known as NMR, is an experimental technique that is applied to proteins, in order to extract information about its structure. One of the extracted pieces of information, is the frequency of each atom of the protein. Starting with such frequencies, we can generate the distances between atoms. Yet, these distances does not give us an accurate 3d structure for the protein. That is because of the large error for pairs of atoms with distance greater than 5Å.

It is an important question, whether one can make accurate 3d structure predictions of a protein by using only the output of an NMR experiment. There are several approaches to this. Buchner and Güntert (2015) propose a model which is implemented in the program CYANA, that makes very accurate structure prediction and is based on simulated annealing and torsion angle dynamics algorithm. Automated NOE assignment and the structure calculation are combined in an iterative process of cycles, each of them transfers information about the intermediary 3D structure of the protein. The result of this process is a final structure calculation using only unambiguously assigned distance constraints. CYANA as input receives: the amino-acid sequence, the list of cross peaks, and the list of chemical shifts.

Emiris and Nikitopoulos, 2005, receive the distance intervals between the atoms to predict the 3d structure, while using no other information about the protein, such as the amino-acid sequence. They insert the distance intervals into a square non-symmetric Cayley-Menger matrix D. It is known (Blumenthal, 1970) that, in order for D to be embeddable in $\mathbb{R}^3$, its rank must be equal to 5. But, for the raw NMR constraints, in the general case, the structure that is constucted by these constraints is not embeddable, i.e. $\text{rank}(D) > 5$. By using a Newton Iteration, they achieve to reduce the rank of the matrix enough for the induced structure to be embeddable, and that is the output of their model. Despite this being a geometric calculation, the computed structure is not far from what we encounter in nature.

In this study, we propose an augmentation of Emiris and Nikitopoulos work, by reducing their algorithm's computational complexity. Based on our experimental results, the increase of speed does not decrease notably the quality of the predictions, i.e. the RMSD (root-mean-square-deviation) against the actual protein structure. We also propose a heuristic method, where we divide the protein into parts. The atoms of each part will be close in the amino-acid sequence. Therefore, they will not be "very" far apart in the 3D-structure, i.e. less than 5Å. The latter can be applied to other methods as well, in order to decrease the computational complexity and increase the accuracy of the predictions.

We organize the manuscript as follows: on Section 2 we present the

$$\begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,r} & \cdots & u_{1,n} \\ u_{2,1} & u_{2,2} & \cdots & u_{2,r} & \cdots & u_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{r,1} & u_{r,2} & \cdots & u_{r,r} & \cdots & u_{r,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ u_{n,1} & u_{n,2} & \cdots & u_{n,r} & \cdots & u_{n,n} \end{bmatrix} \cdot \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & \sigma_n \end{bmatrix} \cdot \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,r} & \cdots & v_{1,n} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,r} & \cdots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ v_{r,1} & v_{r,2} & \cdots & v_{r,r} & \cdots & v_{r,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ v_{n,1} & v_{n,2} & \cdots & v_{n,r} & \cdots & v_{n,n} \end{bmatrix}$$

Fig. 1. Illustration of the truncated SVD algorithm. The SVD decomposition into $U \in \mathbb{R}^{n\times n}$, $\Sigma \in \mathbb{R}^{n\times n}$ and $V^T \in \mathbb{R}^{n\times n}$. The blue indices remain after the truncation. The multiplication of the blue sub-matrices in the same order, produces an n-by-n matrix with rank less or equal to 5.

way we generated the dataset and the methods that we propose to predict the protein structure; on Section 3 we show some of our experimental results and on Section 4 we propose some future work.

2 Materials and Methods

2.1 Datasets

The protein structures on which we tested our model have been arbitrarily picked from the Protein Data Bank and used for both generating NMR datasets and as a comparison model. Because of the lack of NMR data that correspond to known 3d structures, we applied a perturbation on the distances between atoms of the corresponding protein structure from PDB, in order to obtain distance intervals in the Cayley-Menger matrix. The error that is applied to each distance, is proportional to the value of the distance.

2.2 Protein structure predictor

Let D be the Cayley-Menger matrix that represents the NMR distance constraints. Moreover, let $\sigma(D)$ denote the number of non-zero singular values of D. Then, $\sigma(D) = \text{rank}(D)$. Emiris and Nikitopoulos, 2005, apply a Newton Iteration to the Cayley-Menger matrix, in order to minimize all its singular values but the largest 5, thus reducing its rank to 5. It is proven that if they do so, the structure at that point would be embeddable in $\mathbb{R}^3$ (Blumenthal, 1970). The iteration has quadratic convergence rate (Wicks and DeCarlo, 1995). This approach minimizes $O(n)$ singular values. In our method we only minimize $\log n$ singular values, that is the $\log n$ smallest singular values of D, thus reducing asymptotically the number of iterations.

However, by minimizing $\log n$ singular values, there is no guarantee that the induced structure is embeddable. To overcome this problem, we apply a simple truncated SVD to the remaining matrix (see Figure 1), so that the resulting matrix should be embeddable in $\mathbb{R}^3$. Truncated SVD, decomposes the matrix into: $U \in \mathbb{R}^{n\times n}, \Sigma \in \mathbb{R}^{n\times n}$, and $V^T \in \mathbb{R}^{n\times n}$, just like SVD does, and then truncates parts of the matrices as follows: $U_r \in \mathbb{R}^{n\times r}$, $\Sigma_r \in \mathbb{R}^{r\times r}$, and $V_r^T \in \mathbb{R}^{r\times n}$, where $r = 5$. Then it produces a new matrix $M' = U_r \cdot \Sigma_r \cdot V_r^T$. It follows that M' has rank equal to 5, hence an embeddable structure. The output of our model is the set of coordinates obtained by M'.

2.3 Split & Glue

NMR distances greater than 5Å are quite inaccurate. Using them could introduce noise to the input and consume unnecessary computational resources. Various methods have been proposed in literature regarding this phenomenon. The most elaborate may be the one used in CYANA (Buchner and Güntert, 2015), where a complex ranking system is utilized.

We introduce a much simpler approach. Consider an input of our model; a Cayley-Menger matrix, where the atoms are in the same order as in the amino-acid sequence. With this ordering on the indices, the atoms of which the indices are close, should also be close in $\mathbb{R}^3$. Based on this observation, we split the protein into parts of consecutive atoms, and make predictions for them, instead of predicting the whole structure at once. We split the parts, so that each pair of consecutive parts, shares 3 atoms, i.e the last three atoms of the former part are the same with the first three of the latter part. This technique helps us glue the part back together, by matching the planes formed by the common triangles. Note that, for the common atoms, their predicted coordinates in different parts may vary. In our implementation, we always translate and rotate the latter part, and then we discard the prediction of the coordinates of the common triangle from the latter. For example, let a_1, a_2, a_3 be 3 consecutive atoms that form a shared triangle. Let x_1, x_2, x_3 be the predicted coordinates from the former part, and y_1, y_2, y_3 be the predicted coordinates from the latter part. We rotate the latter, so that the triangles match, with respect to their indices. Then, translate the latter, in the way that the centres of the triangles match.

Let D be the initial Cayley-Menger matrix. Observe that for each $D(i,j), D(j,i)$, let d_{ij} be the distance between the atoms i, j. We introduce implicitly the constraints,

$$\min\{D(i,j), D(j,i)\} \leq d_{ij} \leq \max\{D(i,j), D(j,i)\}$$

We have $O(n^2)$ constraints as the one above, where n denotes the number of atoms. Splitting the $n \times n$ matrix into parts of k atoms, as discussed previously, results to reducing the number of constraints to $\frac{n}{k}k^2 = nk$. When $k = O(1)$, the number of constraints, or equivalently the number of entries that are taken under consideration, will be linearly depended on the number of atoms, so we choose k arbitrarily. Also, note that our method can take advantage of the parallel computing, since the prediction of a part does not rely on the prediction of other parts.

3 Results and Discussion

In this section we present our results in accuracy against the ground truth, and compare our model's results to the results of the model proposed by Emiris and Nikitopoulos (2005). Based on our dataset, the average RMSD between our predictions and the ground truth obtained by PDB is equal to 6.7Å, while the average RMSD for each part is equal to 1.9Å. Our predictions are formed by gluing the parts together, as discussed previously, to form a protein structure. The average RMSD seems to be high, and that might be because of two factors: the scaling error on each part, and the propagating error that increases after every gluing operation, however, note that our method predicts the coordinates of all the atoms of the protein and not just the carbon atoms on the backbone. Other methods, such as CYANA (Buchner and Güntert, 2015), start with an a priori knowledge about the positions of the atoms for each residue, and they only have to predict the coordinates of the backbone atoms. When we compare our

		SVM		T-SVM	
PDB	no. Parts	Time (s)	avg. RMSD (Å)	Time (s)	avg. RMSD (Å)
1q01	18	2118	**1.588**	**42**	1.642
4ncu	7	795	**2.357**	**8**	2.358
4cvd	12	1412	1.842	**15**	**1.841**
avg. per Part		117	**1.929**	**2**	1.947

Table 1. Results from our experiments on small molecules. SVM corresponds to the Singular value minimization by Emiris and Nikitopoulos, and T-SVM corresponds to our method without the gluing operation. Result in bold indicates the best result for the corresponding experiment.

	PIDs							
	1Q01		4NCU		4CVD		1UAO	
Methods	Time (min)	RMSD (Å)	Time (min)	RMSD (Å)	Time (min)	RMSD (Å)	Time (min)	RMSD (Å)
SVM	74.04	5.52	89.73	8.08	112.48	6.99	40.93	4.6
T-SVM	0.51	**5.52**	0.46	8.09	0.36	6.99	0.35	4.6
TP-SVM	**0.2**	7.42	**0.17**	**5.51**	**0.18**	**6.26**	**0.17**	**4.2**

Table 2. The results on the experiments on first 60 atoms o each protein with PID, 1Q01, 4NCU, 4CVD, 1UAO. With bold we denote the method achieving the best time or accuracy.

method, to the other singular value minimization method (see Table 1), on small molecules without using the gluing operation, we have slightly worse results in terms of accuracy, i.e. the average difference between the predicted models is about 0.01Å, while in terms of speed 58 times faster, and our method runs on each small molecule, for 2 seconds on average. This is a huge upgrade on the running time, for such a minor loss in accuracy. Note that the difference in running time between the two methods increases as the length of the small molecules increases, but the difference in accuracy remains low.

We also run experiments to compare the three methods: Singular Value Minimization (SVM) by Emiris and Nikitopoulos, Truncated-Singular value minimization (T-SVM), and Truncated-Singular Value Minimzation by Parts (TP-SVM) (see Table 2). The test was on the first 60 atoms of each protein. On all three proteins the methods this paper proposes, namely the T-SVM, and TP-SVM, achieve better scores in terms of time and accuracy. Namely, the TP-SVM method achieves better scores in terms of time and accuracy for all the proteins, except the 1Q01, where the T-SVM method gives better results. For the SVM, which is used as subroutine from each method, we used *perturb value* = 0.3, while the *tolerance* was 1e-16. The *perturb value* expresses the impact of the perturbation at each iteration of the algorithm, and the *tolerance* expresses the threshold at which the values that are below, are treated as 0.

For all the experiments we used a Windows 10 machine, with an Intel i3-6006u, 2Ghz CPU. The code was tested on GNU Octave, version 6.2.0.

4 Conclusion and Future Work

In this work, we propose a model for the problem of approximating the 3d structure of a protein by using NMR data, which is a variation of a model proposed by Emiris and Nikitopoulos. Our model achieves lower computational cost, and based on our experiments, we slightly lose in accuracy. We also propose a different way to deal with the 3d protein structure prediction, that is to break the protein into many parts and predict each part individually, and then glue the parts together to create a complete 3d structure.

As for the future work, obtaining the parts in a different way than just in the order of the amino-acid sequence may affect the results. Also, the gluing part plays an important role in the quality of the results. A different gluing function may drastically affect the results. Another idea is to generalize the overlapping atoms. In the presented version we demand only three atoms to overlap between two consecutive parts. One can imagine for k atoms to overlap between the consecutive parts, where k greater than 3. This way we would have more common restraints in each part and thus better accuracy, on times expense. Observe, that in the extreme case, for big values of k, TP-SVM degenerates to T-SVM.

References

Blumenthal, L. (1970). *Theory and Applications of Distance Geometry*. AMS Chelsea Publishing Series. Chelsea Publishing Company.

Buchner, L. and Güntert, P. (2015). Combined automated noe assignment and structure calculation with cyana. *Journal of biomolecular NMR*, **62**.

Emiris, I. and Nikitopoulos, T. (2005). Molecular conformation search by distance matrix perturbations. *Journal of Mathematical Chemistry*, **37**, 233–253.

Wicks, M. A. and DeCarlo, R. A. (1995). Computing most nearly rank-reducing structured matrix perturbations. *SIAM Journal on Matrix Analysis and Applications*, **16**(1), 123–137.