JJBS

Volume 12, Number 1,March 2019
ISSN 1995-6673
Pages 83 - 87

Jordan Journal of Biological Sciences

Mathematical Prediction of Nucleic Acids 3-D Structures Using Inter-Spin Distances and Nonlinear Least Squares Analysis

Samer I. Awad*

The Hashemite University, Faculty of Engineering, Biomedical Engineering Department, Zarqa, Jordan, 13115

Received June 11, 2018; Revised July 27, 2018; Accepted August 4, 2018

Abstract

Nucleic acids consist of several double helical arms (three to eight of them). These arms are connected at a junction point, with or without several unpaired bases in one or more of the different strands. Current structural information on several nucleic acids is still limited. Electron paramagnetic resonance (EPR) can provide distance measurements of site-directed spin labels (SDSL) applied to the double helical arms. These distances can be used to generate several quadratic equations that characterize the three-dimensional (3-D) structure of the nucleic acid. In this work, a nonlinear least squares algorithm is used to solve these equations along with molecular constraints simultaneously. The solution that this algorithm calculates can be used to create the predicted 3-D structure. The algorithm was tested using twenty-five cases for known DNA structures. Root-mean-square deviation (RMSD) was used in this study to evaluate the accuracy of the predicted structures. The calculated RMSD values had an average of 2.56 and a standard deviation of 1.65.

Keywords. DNA; Nucleic Acid; Biological Macromolecule; Site-directed spin labels; Electron paramagnetic resonance; Least-squares; Levenberg-Marquardt.

1. Introduction

The biological function of nucleic acids molecules is highly dependent on their three-dimensional structures (Hvidsten *et al.*, 2009). Most of these structures are poorly understood which limits the understanding of the biological mechanisms of nucleic acids (Dawson and Bujnicki, 2016; Doudna, 2000). Calculating the 3-D structures of nucleic acids and thus having a better understanding of their function play an important role in therapeutic and diagnostic medical applications. Examples of these applications include drug design, ribosome structuring, and nanoengineering (Greer *et al.*, 1994; Maune *et al.*, 2010).

Structural analysis of biological macromolecules is mainly implemented using X-ray crystallography and high-resolution nuclear magnetic resonance (NMR) spectroscopy. Both methods can provide high resolution structural details; however, they have serious limitations (Banaszak, 2000). X-ray crystallography can be difficult to use with macromolecules that have high flexibility in their structures such as membrane proteins (Mchaourab *et al.*, 2011; Jeschke, 2012). Although NMR spectroscopy can provide some information on this flexibility, it has molecular mass limitations that exclude a large number of macromolecules (Mittermaier and Kay, 2009).

Site-directed spin labeling (SDSL) offers an alternative approach for the structure prediction of nucleic acids (Borbat *et al.*, 2002; Jeschke and Polyhach, 2007). In SDSL, nitroxide spin labels are introduced into specific sites of a macromolecule. Distances between pairs of labels can be measured from which valuable structural information can be obtained. Electron paramagnetic resonance (EPR) spectroscopy techniques have been successfully used for providing such information (Schweiger, 2001). Using conventional continuous wave (CW) EPR spectroscopy and SDSL, distances in the range of 8-20 Å can be measured (Hubbell *et al.*, 2000; Steinhoff and Suess, 2003). Pulsed Double Electron−Electron Resonance (DEER or PELDOR) can access longer distances in the range of 20-80 Å to characterize relatively large biological macromolecules (Jeschke, 2004; Sale *et al.*, 2005).

Several studies have employed different techniques for the prediction of 3-D structures of biological macromolecules. Tung *et al.* employed several computations based on deuterium labeling to determine the atomic model of the cAMP-dependent protein kinase (Tung *et al.*, 2002). Pulsed electron paramagnetic resonance (ESR) based on the detection of double quantum coherence (DQC) and nitroxide spin-labels were used by Borbat et al. to establish the structure of T4 Lysozyme (Borbat *et al.*, 2002). Jeschke and Polyhach studied the effects of varying several experimental parameters on the sensitivity of spin labels distance measurements using EPR and DEER coupled with Monte Carlo search algorithm (Jeschke and Polyhach, 2007). Hatmal employed spin label distance measurements using EPR and DEER coupled with the Quasi-Newton method implemented in Fortran and MATLAB to determine the structure of DNA and RNA junctions, and for the central region of endophilin

* Corresponding author. e-mail: samer.awad@gmail.com.

(Hatmal, 2011). Zhang *et al*. also employed Monte Carlo algorithm to detect the global structure of Phi29 packaging RNA based on site-directed spin labeling (Zhang *et al*., 2012). A neural network technique was used by Hamad *et al*. to predict the 3-D protein structure as a function of enzyme family types and amino acid sequences (Hamad *et al*., 2017).

In this work, a least squares algorithm was used to process the distance measurements acquired using SDSL for the prediction of biological macromolecule structure. The acquired distances were used to create a group of quadratic equations (distance equations). These equations were next solved simultaneously using the Levenberg-Marquardt algorithm limited by molecular constraints to provide the 3-D model of the nucleic acid using the internal geometries of the corresponding arms. This algorithm has been developed fully in MATLAB (Mathworks, Inc., Natick, MA, USA), and was successfully tested using twenty-five cases for known DNA structures. Root-mean-square deviation (RMSD) was calculated to evaluate the accuracy of the predicted structures against the true ones.

2. Methods

The following steps were used in the framework of nucleic acid structure prediction in this work (shown in figure 1).

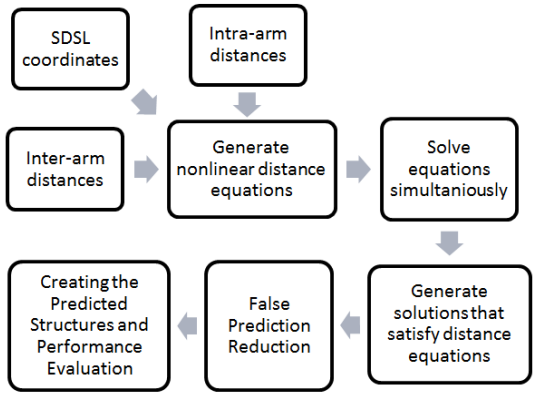


Figure 1. A flowchart of the proposed approach.

2.1. Acquiring Distance Measurements of SDSL Pairs

*Acquiring SDSL Coordinates***:** This process starts with obtaining the structure of each individual arm from a protein databank. In this work, "worldwide protein data bank" (www.wwpdb.org) and "RCSB protein data bank" (www.rcsb.org) were used (Heinz *et al*., 1993). Next, nitroxide spin labels were attached to several positions on each arm using a published algorithm (Beasley *et al*., 2015). The locations of several spin labels were then averaged resulting in one point. This reduces the structural information of each arm into the x, y, and z coordinates of such points. The resulting coordinates identify each arm independent of the nucleic acid and aligned with the positive z-direction.

*Acquiring Intra-Arm Distances***:** The distances between points belonging to the same arm (intra-arm distances) acquired in the previous process were then calculated.

*Acquiring Inter-Arm Distances***:** The distances between points in two different arms (inter-arm distances) can be obtained using the aforementioned EPR spectroscopy techniques. Alternatively, these distances can be obtained computationally for known structures of nucleic acids. The later method was implemented in this work using a published algorithm (Beasley *et al.*, 2015). In this context, the inter-arm distances serve as constraints in the mathematical analysis to reduce the number of degrees of freedom available to the arm with respect to the rest of the macromolecule structure (or with respect to a reference arm). Hence, more estimated distances can reduce the number of possible structures that satisfy the distance equations simultaneously.

2.2. Generating Non-linear Equations Using the Input Information

The inter-spin distances generated were next used to create distance equations of the form:

$$[X_j - X_k]^2 + [Y_j - Y_k]^2 + [Z_j - Z_k]^2 = d^2 \quad (1)$$

where $X_j, Y_j, and\ Z_j$ are the coordinates of the first point and $X_k, Y_k, and\ Z_k$ are the coordinates of the second point, and d is the distance between them. One of the arms in the macromolecule structure was used as a reference. For this arm, the coordinates of the points were kept unchanged with the arm aligned with the positive z-direction, while the coordinates of other arms were changed to account for inter-arm distances. The overall number of distance equations equals the number of the inter-distances plus the number of intra-distances. Three different types of distance equations were created (all having the form of eq. 1):

- Equations between points inside the same arm for arms other than the reference arm. The coordinates of both points are unknowns, since the arm will be reoriented considering the inter-arm distances.
- Equations between a point on the reference arm with known coordinates and a point on a non-reference arm with unknown coordinates.
- Equations between two points on two different non-reference arms where the coordinates of both points are unknowns.

2.3. Solving the Non-Linear Equations

The distance equations generated in the previous step must be solved simultaneously to reorient the arms using a non-linear curve-fitting method. Several least squares methods can be implemented for solving non-linear equations including: Gauss-Newton, trust-region methods, and Levenberg-Marquardt (Björck, 1996). Trust-region methods have the limitation that they can only be used to analyze determined systems of equations in which the number of unknowns equals the number of equations (Byrd *et al*., 1987; Moré and Sorensen, 1983). The Levenberg-Marquardt method can outperform the Gauss-Newton method as it can find a solution even if the initial conditions were far off the final minimum (Pujol, 2007). Since the number of distance equations can be less than the number of unknowns in DNA structure prediction, the Levenberg-Marquardt method was used in this work.

Similar to other minimization algorithms, Levenberg-Marquardt is an iterative procedure. This means that it needs an initial guess to start a minimization. The initial guess was chosen using a series of values through a for-

loop that has a start value, a step value, and an end value. These values were selected taking into consideration the typical dimensions of macromolecules. The final solutions came in the form of new x, y, and z coordinates of the different points on each non-reference arm revealing the structure of the macromolecule.

2.4. False Prediction Reduction

Inter-spin distances acquired using EPR spectroscopy can be difficult to obtain. In many cases, they are fewer than the number required to obtain a unique solution. In this work, an additional step was added to furthermore reduce the number of solutions. The glycosidic N-N distances between arms (the distance at the junctional area between arms) are known to be more than 4 Å and less than approximately 12 + 8n Å, where n is the number of unpaired bases between the ends of arms in the junction area (Banaszak, 2000). Solutions that violated this rule were discarded. Additionally, some solutions can have overlapping arms. Some overlap (or clash) between arms is physically possible and should be accepted if it came up as a structural prediction. Solutions with more than 10 clashes of more than 75% of the sum of the van der Waals radii of any two interacting atoms were also discarded (Banaszak, 2000).

2.5. Creating the Predicted Structures and Performance Evaluation

As previously mentioned, inter-arm distances were obtained computationally for known 3-D structures of DNA (real and hypothetical). This option was chosen to evaluate the accuracy of the calculated solutions generated by the proposed approach against known structures. The evaluation was done using root-mean-square deviation (RMSD):

$$RMSD = \sqrt{\frac{\sum_{n=1}^{N}(x_{1,n}-x_{2,n})^2+(y_{1,n}-y_{2,n})^2+(z_{1,n}-z_{2,n})^2}{N}} \quad (2)$$

where (x_1, y_1, z_1) are the coordinates of a point predicted by the proposed algorithm, (x_2, y_2, z_2) are the coordinates of the same point in the known nucleic acid structure, and N is the total number of points.

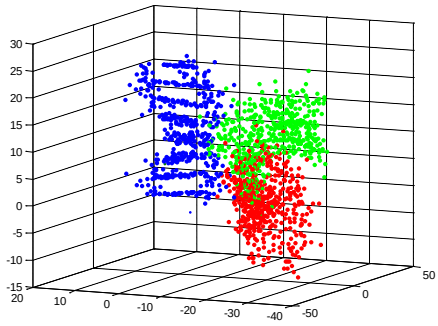
3. Results

To test the proposed algorithm, five different hypothetical DNA structures and one real DNA structure (PDB ID: 1EKW) with different number of inter-arm distances and spin label positions were used (total of 25 cases). For the hypothetical cases, the arms were built using standard B-DNA parameters creating planner and non-planner 3-way junction DNA structures. Table 1 lists the different cases used along with: number of inter-arm distances, number of resulting structures with and without false prediction reduction, and lowest RMSD.

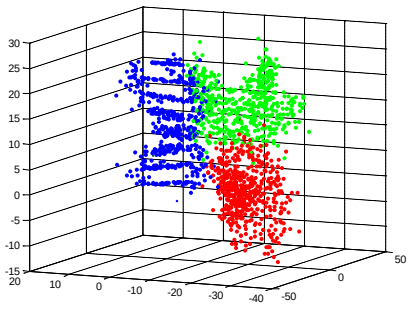
The final output of the proposed algorithm was in the form of reoriented coordinates of points on each non-reference arm along with unchanged coordinates of points on the reference arm. To visualize the DNA structure prediction, all these points were plotted on a 3-D graph. Figure 2 shows the two structure predictions of a sample DNA molecule (the second sample in table 1). Points on the reference arm, second and third arms are colored blue, red, and green, respectively. The difference between the two predictions in figure 2 is the location and orientation of the third arm (colored green).

Table 1. Results of structure predictions for real and hypothetical DNA molecules using different inter-arm distances and label positions.

DNA sample	Number of inter-arm distances	Number of resulting structures without false prediction reduction	Number of resulting structures with false prediction reduction	Lowest RMSD
Planer 3-way DNA junction	9	3	1	0.66
	8	2	1	0.98
	7	4	2	1.68
	6	146	18	1.56
Non-Planer 3-way DNA junction	16	1	1	0.70
	9	4	1	0.96
	9	7	1	0.96
	9	8	1	0.92
	8	4	1	0.97
	7	17	2	1.21
	6	37	5	1.90
Non-Planer 3-way DNA junction (Modified distances)	9	4	1	3.69
	8	6	1	4.78
	7	11	1	4.04
	6	174	1	3.98
1EKW	9	6	1	2.33
	8	4	1	2.33
	7	16	2	2.32
	6	78	4	5.55
	5	428	17	6.52
Non-Planer 3-way distorted DNA junction	9	4	1	2.52
	8	6	1	2.57
	7	10	2	4.91
	6	43	4	3.42
Non-Planer 3-way distorted DNA junction (Modified distances)	9	4	1	2.60



(a)



(b)

Figure 2. Two structure predictions of a sample DNA molecule (the second sample in table 1). Points on the reference arm, second and third arms are colored blue, red, and green, respectively.

4. Discussion and Conclusions

This paper describes structure prediction of biological macromolecules based on distance measurements of spin labels using a least squares algorithm. Inter-arm and intra-arm distances were acquired and used to create non-linear distance equations. These equations were solved simultaneously using the Levenberg-Marquardt algorithm to predict the orientations of macromolecule's arms revealing the 3-D structure. Several predictions were obtained when the number of inter-arm distances was less than the number required to obtain a unique solution (underdetermined cases). In these cases, molecular constraints were applied to reduce the number of predictions. Twenty-five DNA cases were used to test the algorithm. The average of lowest RMSD of the predictions was 2.56 with a standard deviation of 1.65 ranging from 0.66 to 6.52.

A strong correlation can be noticed between a higher number of inter-arm distances (hence more distance equations) and a lower number of predictions and lower RMSD values. This is due to the fact that more equations lead to less degrees of freedom for the non-reference arms orientations and a better chance to reach the true shape of the molecule. Additionally, changing the chosen SDSL pairs (while keeping the same number of distances) could affect the RMSD values (samples 2 and 3 in table 1). The reason for this is that the degrees of freedom of arms could be different which could lead to different predictions and hence different RMSD values.

Future work will include studying the effects of varying the number of distances on the number predictions and RMSD values. Additionally, the effects of varying the chosen spin label pairs will be studied. Future work will also include testing the algorithm using DNA structures with a larger number of arms and using experimental EPR data for other types of biological macromolecules.

References

Banaszak L. 2000. **Foundations of Structural Biology**, 1st Ed. Academic Press, San Diego.

Beasley K, Sutch B, Hatmal M, Langen R, Qin P and Haworth I. 2015. Computer Modeling of Spin Labels: NASNOX, PRONOX, and ALLNOX. *Methods Enzymol*, **563**: 569–593.

Björck A. 1996. **Numerical Methods for Least Squares Problems**, 1st Ed. SIAM, Philadelphia.

Borbat PP, Mchaourab HS and Freed JH. 2002. Protein structure determination using long-distance constraints from double-quantum coherence ESR: Study of T4 lysozyme. *J Am Chem Soc.*, **124**: 5304–5314.

Byrd R, Schnabel R and Shultz G. 1987. A Trust region algorithm for nonlinearly constrained optimization. *SIAM J Numer Anal.*, **24:** 1152–1170.

Dawson WK and Bujnicki JM. 2016. Computational modeling of RNA 3D structures and interactions. *Curr Opin Struct Biol.*, **37**: 22-28.

Doudna J. 2000. Structural genomics of RNA. *Nat Struct Mol Biol.*, **7:** 954 - 956.

Greer J, Erickson J, Baldwin J and Varney M. 1994. Application of the Three-Dimensional Structures of Protein Target Molecules in Structure-Based Drug Design. *J Med Chem.*, **37**: 1035-1054.

Hamad EM, Rawashdeh NA, Khanfar MF, Al-Qasem EN and Al-Gharabli SI. 2017. Neural network based prediction of 3D protein structure as a function of enzyme family type and amino acid sequences. *Jordan J Biol Sci.*, **10**: 73-78.

Hatmal M. 2011. Molecular and computational analysis of spin-labeled nucleic acids and proteins. PhD Thesis, University of Southern California, Los Angeles, USA.

Heinz D, Baase W, Dahlquist F and Matthews B. 1993. How amino-acid insertions are allowed in an alpha-helix of T4 lysozyme. *Nature*, **361**: 561–564.

Hubbell W, Cafiso S and Altenbach C. 2000. Identifying conformational changes with site-directed spin labeling. *Nat Struct Biol.*, **7**: 735-739.

Hvidsten T, Laegreid A, Kryshtafovych A, Andersson G, Fidelis K and Komorowski J. 2009. A comprehensive analysis of the structure-function relationship in proteins based on local structure similarity. *PLOS One*, **4**: e6266.

Jeschke G, Bender A, Paulsen H, Zimmermann H and Godt A. 2004. Sensitivity enhancement in pulse EPR distance measurements. *J Magn Reson.*, **169**: 1-12.

Jeschke G. 2012. DEER Distance measurements on proteins. *Annu Rev Phys Chem.*, **63**: 419–446.

Jeschke G and Polyhach Y. 2007. Distance measurements on spin-labelled biomacromolecules by pulsed electron paramagnetic resonance. *Phys Chem Chem Phys*, **9**: 1895–1910.

Maune H, Han S, Barish R, Bockrath M, Goddard W, Rothemund P and Winfree E. 2010. Self-assembly of carbon nanotubes into two-dimensional geometries using DNA origami templates. *Nat Nanotech.*, **5**: 61 - 66.

Mchaourab H, Steed P and Kazmier K. 2011. Toward the fourth dimension of membrane protein structure: Insight into dynamics from spin-labeling EPR spectroscopy. *Structure*, **19**: 1549–1561.

Mittermaier A and Kay L. 2009. Observing biological dynamics at atomic resolution using NMR. *Trends Biochem. Sci*, **34**: 601–611.

Moré J and Sorensen D. 1983. Computing a trust-region step. *SIAM J Sci Stat Comput.*, **4**: 553–572.

Pujol J. 2007. The solution of nonlinear inverse problems and the Levenberg-Marquardt method. *Geophysics*, **72**: W1-W16.

Sale K, Song LK, Liu YS, Perozo E and Fajer P. 2005. Explicit treatment of spin labels in modeling of distance constraints from dipolar EPR and DEER. *J Am Chem Soc.*, **127**: 9334-9335.

Schweiger A and Jeschke G. 2001. **Principles of Pulse Electron Paramagnetic Resonance**. Oxford University Press, Oxford, UK.

Seeman NC. 1982. Nucleic acid junctions and lattices. *J Theor Biol.*, **99**: 237-247.

Steinhoff H and Suess B. 2003. Molecular mechanisms of gene regulation studied by site-directed spin labeling. *Methods*, **29**: 188-195.

Tung CS, Walsh DA and Trewhella J. 2002. A structural model of the catalytic subunit-regulatory subunit dimeric complex of the cAMP-dependent protein kinase. *J Biol Chem.*, **277**: 12423–12431.

Zhang X, Tung C, Sowa G, Hatmal M, Haworth I and Qin P. 2012. Global structure of a three-way junction in a Phi29 packaging RNA dimer determined using site-directed spin labeling. *J Am Chem Soc.*, **134**: 2644-2652.