Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS

ISSN 2454-6348

Original Research Article **DOI - 10.26479/2017.0302.08**

DISTANCE GEOMETRY WITH IMPROVED SAMPLING OF CONFORMATIONAL SPACE FOR BIOMOLECULAR STRUCTURE DETERMINATION BY NMR SPECTROSCOPY AND PROTEIN DESIGN

Sekhar Talluri*

Department of Biotechnology, GITAM University, Visakhapatnam, India.

ABSTRACT: Protein structure determination from experimental NMR spectroscopic data, Comparative Modeling and Protein Design require a search for low energy conformations with specific spatial relationships between different segments of the same molecule. Generation and characterization of the range of molecular conformations that are consistent with a specified set of restraints is a challenging problem. Distance geometry is an efficient method for generation of molecular conformations that are consistent with a set of specified distance restraints. The effects of Energy directed generation of trial distance matrix for use in Distance Geometry are investigated in this study. Potential Energy directed sampling of distances may direct conformational search towards favorable regions of the conformational space. The use of this method has the potential to improve sampling of conformational space and to avoid the tradeoff between the experimental and the knowledge based information in structure determination and structure refinement.

KEYWORDS: Distance geometry, Protein structure determination, Protein Design, NMR spectroscopy, conformational analysis, structure refinement

***Corresponding Author: Prof. Dr. Sekhar Talluri** Ph.D.

Department of Biotechnology, GITAM University, Visakhapatnam, India

* Email Address: stalluri.gitam@gmail.com

1. INTRODUCTION

Enhancing the sampling efficiency of algorithms for conformational analysis has the potential to lead to substantial improvements in molecular modeling, protein design and protein structure determination [1]. Distance restraints, obtained from protein design specifications or NMR spectroscopy experiments, can be used by Restrained Molecular Dynamics, Target Function Minimization and Distance Geometry based methods for determination of the three dimensional structure of proteins [2-4]. A hybrid Molecular Dynamics-Monte Carlo method (MD-MC) has been demonstrated for sampling enrichment towards target structures using small angle X-ray scattering (SAXS) intensity profiles [5]. A chemical shift driven genetic algorithm for biased Molecular dynamics has been applied for protein structure refinement [6]. Distance geometry can used to determine the coordinates of a set of objects based on distance information between the objects of interest [7, 8]. The input set of distances may consist of exact distances, or they may be specified in the form of distance ranges [9]. In addition, the set of input distances may or may not be sufficient for a unique determination of the coordinates of all the objects of interest [10]. Distance geometry has been used extensively in biomolecular structure determination and drug design [11, 12], and is also useful in other applications such as comparative modeling [13, 14], sensor network localization and graph drawing [15, 16]. In most applications of distance geometry for protein structure determination, a trial distance matrix is generated after Bounds smoothing [17]. The trial distance matrix is then subjected to Embedding [18], Majorization [19, 20] and Structure refinement [21-22]. Alternative, and potentially more efficient methods for obtaining Euclidean coordinates consistent with a specified distance matrix are being investigated [23-24]. Embedding produces a metric matrix from a trial distance matrix. The trial distance matrix has to be generated by choosing a specific set of distances consistent with the set of distance bounds, and these trial distances should, if possible, be consistent with each other. The trial distance matrix can be generated by using a distribution function or by using triangle correlation or Metrization. Metrization may be full or partial, atomwise or pairwise [13, 25]. The choice of an algorithm for generation of the trial distance matrix is generally based on a compromise between the requirements of time, the acceptance ratio and global properties of the conformational ensemble such as RMSD, radius of gyration and Energy. The method of choice of the distribution of distances used for generation of the trial distance matrix is a critical factor that determines the properties of the ensemble of structures generated by application of Distance Geometry [26]. In the earliest attempts at structure determination by distance geometry, trial distances were chosen randomly from the distance bounds. It was observed that such a choice of distances results in preferential sampling of extended conformations that are not typical of globular proteins. To overcome this problem, distribution functions were used to pick distances from the available range, and the choice of the distribution functions was optimized to produce compact conformations [27]. However, a distribution function that is appropriate for a compact globular

protein may not be appropriate for fibrous proteins or for other designed macromolecules with unusual shapes. These abstract mathematical distributions may be supplemented or replaced with force field or knowledge based Energy functions. The results of such changes in the distribution of trial distances are investigated in this study. The potential benefits are improved sampling of favorable regions of conformational space, without reducing the total conformational search space. And the ability to separate the information that directs conformational search from the information used to evaluate the fit to experimental data.

2. MATERIALS AND METHODS

Deca-Alanine: Distance and Torsional restraints were generated based on values expected for a helical conformation [28]. Twenty restraints consisting of 15 torsional and 5 distance restraints were used, corresponding to a mean value of 2 restraints per residue. The distances between 5 pairs of backbone C^{α} atoms separated by four residues, were assigned lower bounds of 1.86A and upper bounds of 6.76A. The bounds for torsional angle phi were specified as -80.0^{o} and -40.0^{o}, for all residues except the first and last residues. The bounds for torsional angle psi were specified as -60.0^{o} and -20.0^{o} for all residues except the first and last two residues. A helical conformation of Deca-Alanine was generated by using PyMOL (Schrodinger Inc.) for use as a reference. The energy minimization was carried out by using the 2008 parameter set of OPLS force field of Jorgensen et al. [29], with implicit solvation based on a Generalized Born – Solvent Accessible Surface Area model (GB-SA) available in the Tinker package (J. Ponder, 2017).

PPM-DG: The Distance Geometry algorithm with Partial Pairwise Metrization, implemented in the Tinker package, was used without any further modification, as a reference (Ponder, 2017). Distance matrices were generated with 5% random Pairwise Metrization.

EDD-PPM-DG (PPM-DG with Energy Directed generation of trial Distances): The implementation of the Distance Geometry for partial pairwise Metrization in the Tinker package was modified as follows. The trial distance used for the pairwise partial metrization was used to evaluate the Lennard-Jones potential energy by using the OPLS force field parameters. The calculated energy value was used to evaluate the Boltzmann factor (exp(-E/RT)) and this value was compared to a random number between 0 and 1. If it exceeded the random number then the trial distance was retained. Otherwise, a new random trial distance was generated within the applicable distance bounds, and a new value of the Boltzmann factor was calculated and compared to a new random number. The process was repeated for a maximum of 100 times. If the calculated value of the Boltzmann factor did not exceed the random number after 100 attempts, the initial value of the trial distance was retained. The trial temperature was set to 298K for all tests. All other protocols and parameters were chosen to be the same for PPM-DG and EDD-PPM-DG.

Evaluation of the conformational ensembles: The Distance Geometry program in the Tinker package uses a simplified forcefield for the target function that is optimized after the Embed step,

and this was not altered in the implementation of the Distance Geometry described in the current study. The mean values of the target function calculated after the final optimization step are listed in the table as "Mean value of DG error function." The radius of gyration values were calculated with the GYRATE subroutine of the Tinker package (Ponder, 2017). The force field based energy and the residual restraint violations in the models generated with the Distance Geometry program were evaluated with the 'Analyze' tool in the Tinker package [30, 31]. The OPLS forcefield available in the Tinker package was used with implicit solvation (GB-SA).

Calculation of RMSD: Pairwise RMSD and ensemble RMSD for backbone CA atoms was evaluated by using the 'fit' and 'align' commands in PyMOL, and custom python scripts. The calculation of ensemble RMSD average and the average of pairwise RMSD from reference conformation was performed by using LibreOffice-Calc version 4.2.8.2.

3. RESULTS AND DISCUSSION

Distance geometry was used to generate 100 conformations for Deca-Alanine using PPM-DG as well as EDD-PPM-DG. The conformational properties of this ensemble of conformations have been assessed by several methods (Table 1). Although the mean value of the target function for the 100 conformations was the same for both methods, the number of conformations with zero residual violation of input restraints was higher in EDD-PPM-DG than in PPM-DG, indicating the potential superiority of the method described in this work. Furthermore, the mean value of the restraint violations as well as the conformational energy evaluated with OPLS force field was lower for EDD-PPM-DG, indicating that the restraints satisfaction improvement was obtained without compromising the structural parameters that determine the potential energy. There was a small increase in the mean value of the radius of gyration for conformations generated with EDD-PPM-DG, and this is likely if the ensemble includes a higher proportion of helical conformers. This conclusion is supported by the observation that the mean value of the RMSD from the reference helical conformation was lower for the ensemble generated with EDD-PPM-DG. There was a small decrease in the mean value of the pairwise RMSD for the ensemble of conformations generated with EDD-PPM-DG compared to the standard method, presumably because more of the conformations generated by this new method satisfy the restraints which restrict the conformational space.

Table 1

Comparison of the conformational ensemble generated by pairwise partial metrization (PPM-DG) and the conformational ensemble generated by energy directed sampling of distances (EDD-PPM-DG) for Deca-Alanine

	PPM-DG	**EDD-PPM-DG**
Mean value of DG error function	0.25 (0.26)	0.25 (0.30)
Number of conformations with zero restraint violations	28	35
Mean value of restraint violation*	0.085 (0.13)	0.059 (0.07)
Mean value of conformational energy	1.02 (169)	-4.15 (184)
Mean value of Radius of Gyration* (A)	5.20 (0.29)	5.29 (0.27)
Mean value of RMSD from a reference helical conformation of Deca-Alanine	1.47 (0.98)	1.34 (0.87)
Mean value of Pairwise RMSD*	2.09 (1.14)	1.90 (1.09)

Note: The calculated values of standard deviation are given in parenthesis. Asterisk indicates that there is a significant difference based on a T-test with P-value less than 0.05.

A 10 residue polypeptide, Deca-Alanine has been used to demonstrate the feasibility of using Energy directed distribution of distances for the generation of the trial distance matrix, in the application of Distance Geometry for structure determination. EDD-PPM-DG can be tuned further by optimizing the temperature used in the Boltzmann factor. In addition, the sampling can be altered by the use of a constant offset to the potential energy used in calculation of the Boltzmann factor. However, in this study, no offset was applied and the behavior of the algorithm was investigated at only one temperature (298 K). Parameter optimization and the use of additional or alternative terms for force field based energy and Potential of Mean Force are currently under investigation and will be described elsewhere. The restraint set used in this study may be considered to be representative of the restraint sets available in the early stages of structure determination by NMR spectroscopy. In the early stages of the NMR structure determination process only a limited number of restraints are available. In some cases, such as large proteins, additional restraints cannot be obtained due to limitations of spectral quality [32]. In such cases, as well as in molecular modeling and protein design applications, the properties of the conformational ensemble are expected to be sensitive to the

force field used for conformational analysis. The torsional conformational preferences of amino acid residues can be used to refine structures generated by NMR spectroscopy by incorporating this information into a potential of mean force [33]. If a potential of mean force is used, a compromise may be necessary between the requirements of satisfaction of experimental restraints and the expected distribution of bond lengths, bond angles, torsional angles and non-bonded interactions. Methods such as Steered Molecular Dynamics [1], incorporation of sampling enrichment schemes into Molecular Dynamics calculations [5], or the use of EDD-PPM-DG have the potential to alleviate such problems.

4. CONCLUSION

The EDD-PPM-DG method, described in this study, can use information regarding the interaction energy from a force field or from a potential of mean force to alter the distribution of distances used for generation of the trial distance matrix, *without changing the distance bounds.* Therefore, conformational search can be directed towards favorable regions of conformational space without explicit restrictions on the conformational space to be searched.

CONFLICT OF INTEREST

The authors have no conflict of interest.

ACKNOWLEDGEMENT

The implementation of Distance Geometry described in this article is a modification of the version included in the Tinker package for molecular modeling by Prof. Jay Ponder, University of Washington. The following free software was also used in this study: Ubuntu 14.04, GNU Fortran 4.8.4, Libre Office 4.2.8.2, Firefox 47.0, Rasmol 2.7.5.2, PyMOL (1.6 and 1.7), PyMOL scripts. The author gratefully acknowledges the following resources: RCSB, Google, Cornell arXiv, NCBI PubMed and PMC, and GITAM University. This research is not the outcome of any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] Maximova T, Moffat R, Ma B, Nussinov R, Shehu A. Principles and overview of sampling methods for modeling macromolecular structure and dynamics. *PLOS Computational Biology.* 2016; 12:e1004619.

[2] Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM. The Xplor-NIH NMR molecular structure determination package. *Journal of Magnetic Resonance.* 2003; 160:65-73.

[3] Guntert P. Automated NMR structure calculation with CYANA. *Methods in Molecular Biology.* 2004; 278:353-378.

[4] Lavor C, Liberti L, Maculan N. Molecular distance geometry problem. Floudas C. and Pardalos P. (Eds.) Encyclopedia of Optimization, Springer, New York. 2009 pp. 2305–2311.

[5] Yang K, Rozycki B, Cui F, Shi C, Chen W, Li Y. Sampling enrichment toward target structures using hybrid molecular dynamics-Monte Carlo simulations. *PLoS One.* 2016;

11:e01566043.

[6] Berjanskii M, Arndt D, Liang Y, Wishart DS. A robust algorithm for optimizing protein structures with NMR chemical shifts. *Journal of Biomolecular NMR.* 2015; 63:255-264.

[7] Blumenthal LM. Theory and applications of distance geometry. Oxford University Press, 1953.

[8] Billinge SJL, Duxbury PM, Gonçalves DS, Lavor C, Mucherino A. *4OR-Q Journal of Operational Research.* 2016; 14:337-376.

[9] Liberti L, Lavor C, Maculan N. A branch-and-prune algorithm for the molecular distance geometry problem. *International Transactions in Operational Research.* 2008; 15:1–17.

[10] Oshiro CM, Thomason J, Kuntz ID. Effects of Limited Input Distance Constraints Upon the Distance Geometry Algorithm. *Biopolymers.* 1991; 31:1049-1064.

[11] Crippen GM, Havel TF. Distance Geometry and Molecular Conformation. John Wiley and Sons. 1988.

[12] Havel T, Wuthrich K. A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of intramolecular H-H proximities in solution. *Bulletin of Mathematical Biology.* 1984; 46:673-698.

[13] Havel TF, Snow ME. A new method for building protein conformations from sequence alignments with homologues of known structure. *Journal of Molecular Biology.* 1991; 217:1-7.

[14] Huang ES, Samudrala R, Ponder JW. Distance geometry generates native-like folds for small helical proteins using the consensus distances of predicted protein structures. *Protein Science.* 1998; 7:1998-2003.

[15] So M, Ye Y. Theory of semidefinite programming for sensor network localization. *Mathematical Programming.* 2007; 109:367–384.

[16] Bachrach J, Taylor C. Localization in sensor networks. Stojmenovi´c I. (Ed.) Handbook of Sensor Networks. Wiley. 2005.

[17] Dress AWM, Havel TF. Shortest-Path Problems and Molecular Conformation. *Discrete Applied Mathematics.* 1988; 19:129-144.

[18] Crippen GM, Havel TF. Stable Calculation of Coordinates from Distance Information. *Acta Cryst. A.* 1978; 34:282-284.

[19] Havel TF. An Evaluation of Computational Strategies for use in the Determination of Protein Structure from Distance Constraints obtained by Nuclear Magnetic Resonance. *Progress in Biophysics and Molecular Biology.* 1991; 56:43-78.

[20] Glunt W, Hayden TL, Raydan M. Molecular Conformations from Distance Matrices. *Journal of Computational Chemistry.* 1993; 14:114-120.

[21] Clore GM, Nilges M, Br¨unger AT, Karplus M, Gronenborn AM. A comparison of the

restrained molecular dynamics and distance geometry methods for determining three-dimensional structures of proteins on the basis of interproton distances. *FEBS Letters.* 1987; 213:269-277.

[22] Kaptein R, Boelens R, Scheek RM, van Gunsteren WF. Protein structures from NMR. *Biochemistry.* 1988; 27:5389-5395.

[23] Dong Q, Wu Z. A geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data. *Journal of Global Optimization.* 2003; 26:321–333.

[24] Grosso A, Locatelli M, Schoen F. Solving molecular distance geometry problems by global optimization algorithms. *Computational Optimization and Applications.* 2009; 43:23–27.

[25] Kuszewski J, Nilges M, Br¨unger AT. Sampling and efficiency of metric matrix distance geometry-a novel partial metrization algorithm. *Journal of Biomolecular NMR.* 1992; 2:33-56.

[26] Havel TF. The Sampling Properties of Some Distance Geometry Algorithms Applied to Unconstrained Polypeptide Chains: A Study of 1830 Independently Computed Conformations. *Biopolymers.* 1990: 29:1565-1585.

[27] Hodsdon ME, Ponder JW, Cistola DP. The NMR Solution Structure of Intestinal Fatty Acid-binding Protein Complexed with Palmitate: Application of a Novel Distance Geometry Algorithm. *Journal of Molecular Biolology.* 1996; 264:585–602.

[28] Talluri, S. Advances in engineering of proteins for thermal stability. *International Journal of Advanced Biotechnology and Research.* 2011; 2:190-200.

[29] Kaminsky GA, Friesner RA, Tirado-Reeves J, Jorgensen W. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *Journal of Physical Chemistry B.* 2001; 105:6474-6487.

[30] Shi Y, Xia Z, Zhang J, Best R, Wu C, Ponder JW, Ren P. The polarizable atomic multipole-based AMOEBA force field for proteins. *Journal Chemical Theory and Computation.* 2013; 9:4046-4063.

[31] Laury ML, Wang LP, Pande VS, Gordon TH, Ponder JW. Revised parameters for the AMOEBA polarizable atomic multipole water model. *Journal of Physical Chemistry B.* 2015; 19:276-291.

[32] Raman S, Lange OF, Rossi P., et al. NMR structure determination for larger proteins using backbone only data. *Science*. 2010; 327:1014-1018.

[33] Kuszewski J, Gronenborn AM, Clore GM. Improvements and extensions in the conformational database potential for the refinement of NMR and X-ray structures of proteins and nucleic acids. *Journal of Magnetic Resonance.* 1997; 125:171-177.