Proteins Structure Determination with Imprecise Distances*

Ivan Sendin,[1] and Siome Klein Goldenstein[2]

[1] *Dept. of Computer Science-CAC, Federal University of Goias, Catalao, Brazil,* sendin@catalao.ufg.br

[2] *Institute of Computing, IC/Unicamp, Campinas, Brazil,* siome@ic.unicamp.br

Abstract The Molecular Distance Geometry Problem is related to protein structure determination using Nuclear Magnetic Resonance information which is imprecise distances of some proteins atoms. Most current methods available to solve this problem work with exact distances. We propose three new methods to propagate uncertainty: using particles, using affine forms and hybrid affine-particles. We use these new methods to propagate uncertainty and determine the protein backbone using NMR like information.

Keywords: proteins structure, uncertainty propagation affine arithmetic, particles

1. Introduction

Proper knowledge of three-dimensional protein structure is a major step in many bioinformatic tasks. On important method to obtain protein structure is the Nuclear Magnetic Resonance(**NMR**) [15] . This process can detect the interaction between pairs of atoms near to each other. So the information given by an NMR experiment is an imprecise distance of some pairs of atoms [6].

The computational problem to determine a protein structure from inter-atomic distances is the **Molecular Distance Geometry Problem (MDGP)** [10]. Usually, we view this problem as a graph problem, where each atom is mapped to one vertex and the edges are the known inter-atomic distances, so this problem is also called **graph embedding problem**. The problem to decide if a graph can be embedded in some k-dimensional space is known to be NP-Complete, even for one dimensional case [12,13]. For a complete graph with exact distances, this is a trivial problem. Dong and Wu [4] presented the Geometric Build-Up(GBU) algorithm that uses a sufficient dense graph with exact distances to iteratively build a solution for the problem in polynomial time.

Most current methods used to address MDGP use exact distances or an optimization process, like Simulated Annealing [9]. In this work, we introduce the use of particles and present a new hybrid method to propagate uncertainty. Applied to GBU, we can reconstruct a protein using a sparse graph with intervalar distances.

*This research was partly supported by CAPES and FAPESP.

1.1 Uncertainty Propagation

Uncertainty representation and propagation is an important field in information theory [8].In this work uncertainty means imprecise information, i.e. the unknown true value lies in an interval. An uncertainty propagation method should represent the uncertainty of each system state and control the uncertainty growth: if the uncertainty grows too much the information can be useless.

We will use two well known methods for uncertainty propagation: Particles and Affine Forms. Also, we will introduce a new hybrid method. All three methods will be applied to the GBU algorithm and tested in protein structure determination.

Particles. Particles is a non-parametric uncertainty representation method [14]. Modelling with particles is straightforward, a set of samples - called particles - is created for each unknown value and computation is applied on these particles.

This approach is interesting because the computational framework is the same as that used on exact values, the selection, filtering and optimization already available can be applied over particles. In this work, we will use two methods to control particles:

Selection To control the amount of uncertainty to be propagated a subset of particles is selected to represent one state. This selection is performed using Mahalanobis Distance [11].

Sample Importance Ressample Using a problem dependent scoring function, the score of each particle is calculated and this score is used to determine the propagation probability of each particle [2].

Affine Forms. A partially known value $\hat{x}$ is defined by its central value and symbolic sum of noise terms

$$\hat{x} = x_0 + \sum_{i=1}^{n} x_i e_i,$$

with $x_i \in \mathcal{R}$ and $e_i \in [-1, 1]$. The unknown terms e_i models the uncertainty of one affine form. To measure the uncertainty of one affine form $\hat{x}$, one can use the *range* function:

$$range(\hat{x}) = \sum_{i=1}^{n} |x_i|.$$

In [3], an Affine Arithmetic (**AA**) is defined, arithmetical operations with real numbers are trivial, other mathematical operations require approximations that create new unknowns values and enlarge the range. One important feature of AA is that noise terms can be cancelled:

$$\hat{x} - \hat{x} = 0.$$

Affine arithmetic ensures that the resulting affine form contains the true value provided that the operands contain the true value. This property, useful for reliable computing, in general is not desirable because it causes the growth of noise range, because unlikely regions are reached by an affine form.

Another drawback of affine representation is its distance to exact representation, that makes the optimization process harder to design and implement.

One can create an exact representation from affine forms sampling values for unknown terms and replacing the sampled values in all affine forms. As the affine correlation is held on unknown sharing, this sampling process can create consistent values for a set of affine forms. Also, it is possible to create a particles representation using this method.

Hybrid Method. Here, we introduce a hybrid method for uncertainty propagation. Like in [7] and in [5], the uncertainty is represented both in parametrical and non-parametrical forms. Our method starts with an affine representation of the problem. Then a exact representation is obtained sampling values for the unknown. The sampling process is repeated and a set of exact instances is created. Now these instances are filtered and optimized (as seen on Section 1.1). We expect that this process will produce narrower limits and use those particles to control the affine forms.

2. Computational Experiments

Three versions of the GBU were created to propagate uncertainty: particles-GBU, affine-GBU and a hybrid-GBU. For particles and hybrid method, at each GBU step we create 60 particles for each state. The uncertainty is controlled as follows: the SIR process uses a quadratic penalty function, and is repeated until the average score is stabilized, and a range that contains 85% of the particles is propagated. After the proteins is determined an interval version of Stochastic Embedding Proximity [1] improves the final structure.

2.1 Dataset and Distances Determination

We obtained all NMR proteins structures available in October 2012 at the PDB bank. As the proposed method uses covalent and C_α distances (see below), we are able to use only proteins whose distances were well defined in PDB bank, making 367 proteins.

The distances used in the tests were determined as follows:

1. **RMN-like distances** With atoms separated up to 5 Å , we use a intervalar distance: 2 to 3 Å , 3 to 4 Å and 4 to 5 Å ,in accordance with the observed real distance;
2. **Molecular Geometry distances** For atoms separated by one or two covalent bonds and for consecutive C_α its exact distance is used;

2.2 Results

The affine GBU method did not work: this method does not control the uncertainty, the range grows too fast and the computation does not work. The results for particles and hybrid methods are summarized on Table 1. The results are grouped by the size of the protein backbone, and we show the percentage of distance restraints satisfied and the RMSD to the original protein. In Figure 3 we show the result for the 2gp8 protein.

Table 1. First the average percentage of distance restraints satisfied by the reconstructed protein and, in parentheses, the average RMSD to the original protein, in Angstroms.

Method/Backbone Size	50	100	150	200	>200
Particles	65,9 (2,8)	68,4 (4,6)	63,0 (7,2)	64,5 (8,4)	63,2 (10,5)
Hybrid	73,2 (3,2)	73,8 (4,6)	62,3 (6,7)	63,7 (8,1)	64,2 (9,9)

3. Conclusions

In this work we presented three methods to build protein structures using imprecise interatomic distances. The pure affine approach did not work. The statistical uncertainty propagation - provided by particles selection and SIR filtering - is efficient to control the uncertainty enabling the particles and the hybrid methods to determine the protein structure.

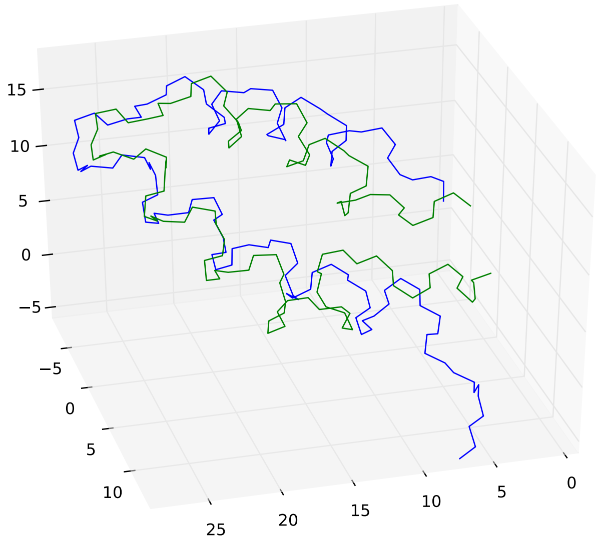

Figure 1. The alignment of 2gp8 protein. In blue, the reconstructed protein using the hybrid method, aligned with the original one, in green.

References

[1] D. Agrafiotis. Stochastic Proximity Embedding. *Journal of Computational Chemistry*, 24(10):1215–1221, 2003.

[2] J. Carpenter, P. Clifford, and P. Fearnhead. An Improved Particle Filter for Non-linear Problems. *IEE Proceedings - Radar, Sonar and Navigation*, 146(1):2–7, 1999.

[3] Luiz H. de Figueiredo and Jorge Stolfi. *Self-Validated Numerical Methods and Applications*. Brazilian Mathematics Colloquium monographs. IMPA/CNPq, Rio de Janeiro, Brazil, 1997.

[4] Qunfeng Dong and Zhijun Wu. A Geometric Build-Up Algorithm for Solving the Molecular Distance Geometry Problem with Sparse Distance Data. *Journal of Global Optimization*, 26:321–333, 2003.

[5] Leyza Baldo Dorini and Siome Klein Goldenstein. Unscented feature tracking. *Computer Vision and Image Understanding*, 115(1):8–15, 2011.

[6] P. Guntert. Structure calculation of biological macromolecules from NMR data. *Quarterly reviews of biophysics*, 31(2):145–237, 1998.

[7] Simon J. J. and J. K. Uhlmann. A new extension of the kalman filter to nonlinear systems. *SPIE*, 1997.

[8] G.J. Klir. *Uncertainty and information: foundations of generalized information theory*. Wiley-IEEE Press, 2006.

[9] C. Lavor, L. Liberti, N. Maculan, and A. Mucherino. Recent advances on the discretizable molecular distance geometry problem. *European Journal of Operational Research*, 219:698–706, 2012.

[10] C. Lavor, L. Liberti, and a. Mucherino. On the solution of molecular distance geometry problems with interval data. *BIBMW*, pages 77–82, 2010.

[11] P.C. Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55, 1936.

[12] J. Saxe. Embeddability of weighted graphs in k-space is strongly NP-hard. *Proceedings of the 17th Allerton Conference on Communication, Control, and Computing*, pages 480–489, 1979.

[13] J. Saxe. Two Papers on Graph Embedding Problems. Technical Report 10-102, Department of Computer Science, Carnegie-Mellon University, 1980.

[14] L. Wasserman. *All of nonparametric statistics*. Springer-Verlag New York Inc, 2006.

[15] D.M. Webster. *Protein Structure Prediction: Methods and Protocols*. Methods in Molecular Biology. Humana Press, 2000.