# Homology Initialization for Protein Structure Determination via Distance Geometry

Abiy Tasissa [1], Rongjie Lai [2] and Chunyu Wang [3]

[1] *Department of Mathematics, Tufts University*
[2] *Department of Mathematics, Rensselaer Polytechnic Institute*
[3] *Deptartment of Biological Sciences, Deptartment of Chemistry and Chemical Biology, Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute*

**Abstract**

The problem of finding the configuration of points given partial information on pairwise inter-point distances, the Euclidean distance geometry problem, appears in multiple applications. In this paper, we propose an approach that integrates homology modeling and a nonconvex distance geometry algorithm for the protein structure determination problem. Preliminary numerical experiments show promising results.

## 1 Introduction

The Euclidean distance geometry problem, abbreviated EDG hereafter, is the problem of finding the configuration of points given partial information on pairwise inter-point distances. The problem is of central importance in many tasks such as protein structure determination [1, 2, 3, 4], sensor localization [5, 6] and dimensionality reduction [7]. Formally, consider a set of $n$ points $\mathbf{Q} = \{\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_n\} \in \mathcal{R}^{n\times r}$ in an $r$ dimensional space and with $\mathbf{D}$ denoting the partially observed squared distance matrix. The exact EDG problem can be formulated as follows

$$\text{Find } \mathbf{Q} \in \mathcal{R}^{n\times r} \text{ such that } ||\mathbf{q}_i - \mathbf{q}_j||_2^2 = \mathbf{D}_{i,j} \quad (i,j) \in \Omega, \tag{1}$$

where $\Omega$ denotes the set of observed indices. The above minimization problem also has the following compact representation

$$\text{Find } \mathbf{X} \in \mathbb{S} \text{ such that } \mathcal{A}_\Omega(\mathbf{X}) = \mathbf{b}, \tag{2}$$

where $\mathbf{X} = \mathbf{Q}\mathbf{Q}^T$ denotes the symmetric positive semi-definite Gram matrix and the linear operator $\mathcal{A}_\Omega(\mathbf{X}) : \mathcal{R}^{n\times n} \to \mathcal{R}^{|\Omega|}$ is defined as $\mathcal{A}_\Omega(\mathbf{X}) = \{\mathbf{X}_{i,i} + \mathbf{X}_{j,j} - 2\mathbf{X}_{i,j}, (i,j) \in \Omega\}$ and $\mathbf{b} = \{\mathbf{D}_{i,j}, (i,j) \in \Omega\}$. The set $\mathbb{S}$ is defined as $\mathbb{S} = \{\mathbf{X} \in \mathcal{R}^{n\times n} \mid \mathbf{X} = \mathbf{X}^T \,\&\, \mathbf{X}\cdot\mathbf{1} = \mathbf{0}\}$. The constraint $\mathbf{X}\cdot\mathbf{1} = \mathbf{0}$ fixes translation ambiguity by centering the points at the origin. Since distance measurements are prone to be corrupted by noise, we relax the equality constraint in the above minimization and consider the following optimization program

$$\min_{\mathbf{X}} \; ||\mathcal{A}_\Omega(\mathbf{X}) - \mathbf{b}||_2^2. \tag{3}$$

Let $\mathbf{q}^*$ denote the optimal point configuration obtained from solving the optimization program. In this paper, we study the molecular conformation problem where the task is to determine the three-dimensional structure of a molecule given some information about inter-point distances. We focus on the protein structure determination problem where the measurements are obtained from nuclear magnetic resonance (NMR) spectroscopy or X-ray diffraction experiments. Due to the limited range of these measurements, the partial distance matrix is specified in terms of a certain Angstrom threshold. To estimate the protein structure, we make use of a homologous protein and build a homology model of the target protein using the homologous protein. Hereafter, the point cloud corresponding to the homology model is denoted by $\mathbf{q}_0$.

## 2 Proposed Approach

Consider the exact EDG problem where the $n$ atoms in consideration are in a 3-dimensional space with $n \gg 3$. It can be verified that the rank of the $\mathbf{X}$ is 3. With this, we can consider a rank regularization on the optimization problem in (2).

$$\min_{\mathbf{X}\in\mathbb{S}} \quad \text{rank}(\mathbf{X}) \;\; \text{subject to} \;\; \mathcal{A}_\Omega(\mathbf{X}) = \mathbf{b} \tag{4}$$

Rank minimization is generally NP-hard [8] and a well known approach is an approximation which replaces the rank function with the convex nuclear norm. The nuclear norm of a matrix is the sum of the singular values of matrix and is denoted by $||\mathbf{X}||_*$. We apply this convex relaxation to (2) to obtain

$$\min_{\mathbf{X}\in\mathbb{S}} \quad ||\mathbf{X}||_* \;\; \text{subject to} \;\; \mathcal{A}_\Omega(\mathbf{X}) = \mathbf{b} \tag{5}$$

The above problem is a convex optimization program which facilitates ease of computation by employing standard convex solvers. The theoretical analysis in [9], under mild assumptions, shows that solving (5) with $O(3n\log n)$ uniformly random distance measurements determines the underlying point configuration with high probability. In practice, the distance measurements are not exact. In addition, the available distance information is not necessarily uniformly at random but rather a deterministic set $\Omega$ provided from measurements. From the optimization perspective, the program in (5) will consist of a step that projects onto the set of rank $r$ matrices which could be expensive for large number of atoms. Based on these considerations, we factorize the Gram matrix $\mathbf{X} = \mathbf{Q}\mathbf{Q}^T$ and consider the following nonconvex optimization program.

$$\min_{\mathbf{Q}\in\mathcal{R}^{n\times r}} \quad \frac{\lambda}{2}||\mathcal{A}_\Omega(\mathbf{Q}\mathbf{Q}^T) - \mathbf{b}||_2^2 + \text{trace}(\mathbf{Q}\mathbf{Q}^T) \tag{6}$$

The trace term in (6) appears since $||\mathbf{X}||_* = \text{trace}(\mathbf{X})$ for a positive semidefinite matrix. The parameter $\lambda$ controls the extent to which the regularization is enforced. Our numerical experiments indicate that, in the exact EDG setting with $O(nr\log n)$ uniformly random measurements, the algorithm corresponding to the above optimization problem typically converges in few iterations to the desired true solution irrespective of the initialization. Compared to its convex counterpart, the algorithm based on the nonconvex formulation is computationally faster. This is due to the employed factorization $\mathbf{x} = \mathbf{Q}\mathbf{Q}^T$ with $\mathbf{Q} \in \mathcal{R}^{n\times r}$ and $r = O(1)$. However, in the general noisy EDG problem, numerical experiments indicate that the initialization of the point cloud is crucial. To this end, the point cloud of the homology model can also be used as a good initialization for the nonconvex EDG problem. To solve the program in (6), we use the algorithm in [9] that employs the augmented Lagrangian method and the Barzilai-Borwein (BB) steepest descent method to update $\mathbf{Q}$. The augmented Lagrangian is given by

$$Ł(\mathbf{Q};\mathbf{\Lambda}) = \text{Trace }(\mathbf{Q}\mathbf{Q}^T) + \frac{\lambda}{2}||\mathbf{A}(\mathbf{Q}\mathbf{Q}^T) - \boldsymbol{b}||_2^2$$

where $\lambda$ denotes the penalty term. The resulting algorithm, Algorithm 1, is detailed in [9] and is summarized below. With this, $\mathbf{q}^*$ denotes the point configuration obtained from the algorithm. We emphasize that the algorithm is initialized with the point cloud of the homologous protein. We posit that this allows us to obtain local convergence for the nonconvex optimization program.

**Algorithm 1** Augmented Lagrangian based scheme to solve (6)

**Initialization:** Set rank estimate $r = 10$, $\mathbf{Q}^{\text{init}} = \mathbf{Q}^0$, $E^0_{\text{Total}} = 0$. Set maxiterations, bbiterations, $r$, $\lambda$, Tol
**for** k = 1:maxiterations **do**
  Barzilai-Borwein(BB) descent for $\mathbf{Q}_k$.
  $E^k_{\text{Total}} = \text{Trace }(\mathbf{Q}^k(\mathbf{Q}^k)^T) + \frac{\lambda}{2}||\mathbf{A}(\mathbf{Q}^k(\mathbf{Q}^k)^T) - \boldsymbol{b}||_2^2$
  **if** $E^k_{\text{Total}} < \text{Tol}$ **then**
    **break**

# 3 Results

To test the proposed algorithm, we consider a synthetic experiment to determine the structure of Ubiquitin. This is identified as 1UBQ_H, a protonated structure based on 1ubq in the protein data bank and the structure has 629 protons. The homologous protein is identified as 2FAZ in the protein data bank. Since pairwise proton distances measured by NMR are not exact, we adopt the following noise model

$$\hat{\mathbf{D}}_{i,j} = \mathbf{D}_{i,j} + KN(0,1)\mathbf{D}_{i,j}\,], \text{ for all observed indices } \quad (i,j) \in \Omega$$

$N(0,1)$ is a Gaussian noise with mean 0 and standard deviation 1 and the parameter $K$ controls the range of the noise as a relative error. For our test problem, we first consider a 6 Å cutoff i.e. distances between protons within 6 Å are observed, a range that is typical observed in protein NMR for structure determination. To resemble a realistic setup, from these available distances, we only consider a a subset of the distances i.e. for each point, among the available distances to this point, we pick a certain percentage of randomly selected distances. We initialize the algorithm with the homology model and use the the root mean square error (RMSE) between the numerically estimated point cloud and the ground truth restricted to H-alpha atoms. Table 1 summarizes our numerical results. For small to moderate noise, we obtain good reconstruction of the underlying protein. Given our noise model, there is a trade-off between percentage of available distances and noise level. Since all the available distances are corrupted by a random noise, the input to the nonconvex distance geometry algorithm is more inaccurate when we have more distance samples and a large noise parameter. This observation aligns with our numerical results where better RMSE is obtained for $r = 30\%$ compared to $r = 40\%$ when $K \geq 0.01$. Fig.1 shows a visualization of the estimated protein obtained from the algorithm. We see a good agreement with the ground truth in the positions of the H-alpha atoms.

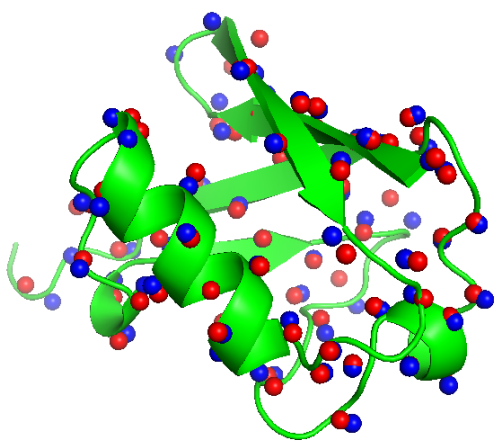

Figure 1: The original structure is shown in green. The red spheres are the H-alpha atoms of 1UBQ_H and the blue spheres are the H-alpha atoms of the estimated structure. The algorithm is run with noise parameter $K = 0.05$ and percentage of available distances, among the distances within the 6 Å cutoff, is $r = 30\%$.

# 4 Conclusion

In this paper, we study the protein structure determination problem based on pairwise distances obtained from solution NMR. Using homology model, we initialize a nonconvex distance geometry algorithm based on a homologous protein. Our preliminary results show that, with limited distance measurements and small to moderate noise, reasonable estimates of the underlying structure could be obtained.

| | r = 80% | r = 60% | r = 40% | r = 30% |
|---|---|---|---|---|
| % of available distances | 4.26% | 3.20% | 2.10% | 1.53% |
| K = 0.01 | 0.0895 | 0.263 | 1.225 | 1.662 |
| K = 0.05 | 0.609 | 1.134 | 1.804 | 1.622 |
| K = 0.1 | 1.406 | 1.904 | 2.334 | 1.936 |
| K = 0.2 | 2.440 | 3.219 | 3.323 | 2.603 |
| K = 0.5 | 7.018 | 7.489 | 5.462 | 4.218 |

Table 1: The root mean square error (RMSE) in the unit of Å, restricted to H-alpha atoms, in estimating the structure of 1UBQ_H given partial distances. The nonconvex algorithm has been initialized with a homologous protein. K denotes the noise parameter and $r$ is the percentage of randomly selected distances among the distances which are within 6 Å for a given atom in consideration. The results reported are average of 20 runs.

# References


[1] Babak Alipanahi, Nathan Krislock, Ali Ghodsi, Henry Wolkowicz, Logan Donaldson, and Ming Li. Determining protein structures from noesy distance constraints by semidefinite programming. *Journal of Computational Biology*, 20(4):296–310, 2013.

[2] Xingyuan Fang and Kim-Chuan Toh. Using a distributed sdp approach to solve simulated protein molecular conformation problems. In *Distance Geometry*, pages 351–376. Springer, 2013.

[3] Ngai-Hang Z Leung and Kim-Chuan Toh. An sdp-based divide-and-conquer algorithm for large-scale noisy anchor-free graph realization. *SIAM Journal on Scientific Computing*, 31(6):4351–4372, 2009.

[4] Michael Souza, Carlile Lavor, Albert Muritiba, and Nelson Maculan. Solving the molecular distance geometry problem with inaccurate distance data. *BMC bioinformatics*, 14(S9):S7, 2013.

[5] Yichuan Ding, Nathan Krislock, Jiawei Qian, and Henry Wolkowicz. Sensor network localization, euclidean distance matrix completions, and graph realization. *Optimization and Engineering*, 11(1):45–66, 2010.

[6] Pratik Biswas, Tzu-Chen Lian, Ta-Chung Wang, and Yinyu Ye. Semidefinite programming based algorithms for sensor network localization. *ACM Transactions on Sensor Networks (TOSN)*, 2(2):188–220, 2006.

[7] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[8] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[9] Abiy Tasissa and Rongjie Lai. Exact reconstruction of euclidean distance geometry problem using low-rank matrix completion. *IEEE Transactions on Information Theory*, 65(5):3124–3144, 2018.