Recreating Protein Structures

Kirill Konovalov (contribution: 3,6), Esther Visser (contribution: 2,5), Sepanta Zeighami (contribution: 1,4)

1 Introduction

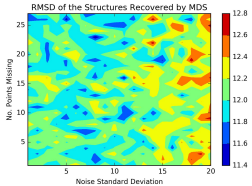
Reconstruction of protein structure by nuclear magnetic resonance (NMR) necessarily uses a step of fitting coordinates to sparse pairwise distance restraints derived from nuclear Overhauser effect (NOE) signals, which as all experimental data is subject to errors.

We have employed several popular methods of restoring coordinates from pairwise distances and tested their performance on incomplete and noisy data, derived from all C_α atoms of the crystal structure of a protein (PDBID 1R9H) This model approximates the experimental data.

2 MDS with missing data

We applied classical MDS[4] on the data. However, if you remove a few distances in your distance matrix, global classical MDS will no longer work. We looked how well MDS performs locally in this case: we split up the dataset into two parts, each part of maximum size such that we have all distances. Then we apply classical MDS on the parts and glue the parts together again.

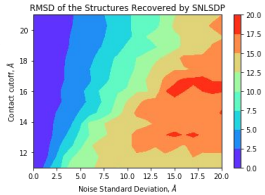
It turned out that the few distances we removed did not really influence the performance if we apply normal noise to the data.



3 SNL

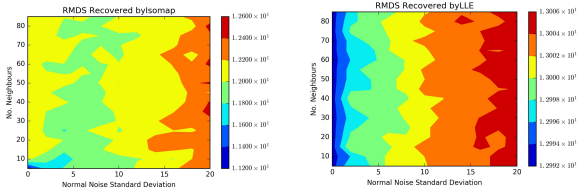
SNL[3] is a semidefinite programming technique that allows to relax the distance constraints and solve the topological embedding problem with sparse data. The obtained solution is refined by a gradient descent. The contact cutoff i.e. the amount of available data favorably affects the reconstructed structure reducing the RMSD, while noise deteriorates the solution.

The problem also has a lower bound of sparsity, which is determined by the relevant SDP problem feasibility. For the presented data, the minimal feasible cutoff was found to be 11Å



4 Manifold Learning

We used Isomap[1] and LLE [2] algorithms to recover the proteins' structure using the geodesic distance between the atoms. Isomap performed better than LLE, best with existence of some low magnitude noise, while LLE recovered the structure with a consistent RMSD even under noise. Moreover, number of neighbours used in the calculation of the geodesic distances only had a marginal effect on the performances.



5 Analysis

The ability to produce structures with low RMSD with increasing noise levels is different for all methods used, the best results for lower noise levels were obtained by SNL, but its performance worsened drastically with the increase of noise.

Secondly, for high noise levels Isomap performed better than the SNL. However, comparing SNL with MDS would not be fair as the SNL works with much sparser data than MDS. The same is true in comparing MDS with Isomap.

Structure reconstructed by SNL: white - original atom positions, red - restored positions at 13Å cutoff after being subjected to 3Å gaussian noise

6 Conclusion & Future Work

Our work has shown that SNL is the best method to recover the structure of a molecule under low noise situations, and when the magnitude of noise is larger, Isomap and MDS outperform the other algorithms. **Future challenges** include testing the performance scaling with the dataset size. Real protein structures can contain on the order of 10^5 atoms so measuring performance in terms of computational time may also be a problem. Employing chemical principles of molecular geometry to impose constraints on the solution may aid in reconstructing high fidelity structures. An interesting hypothesis worth investigating is that various subsets of the given structure may respond differently to reconstruction, thus excluding some of the data may favorably affect the overall quality of the final structure.

[1] "A global geometric framework for nonlinear dimensionality reduction" Tenenbaum, J.B.; De Silva, V.; & Langford, J.C. Science [2] "Nonlinear dimensionality reduction by locally linear embedding" Roweis, S. & Saul, L. Science 290:2323 (2000) [3] P. Biswas, et al., Semidefinite programming approaches for sensor network localization with noisy distance measurements, IEEE Transactions on Automation Science and Engineering, 3 (2006), pp. 360--371 [4] Gower, J. C. (1966) Some distance … analysis. *Biometrika* **53**, 325–328.