John von Neumann Institute for Computing

Dimensionality Reduction Techniques for Protein Folding Trajectories

T. Eitrich, S. Mohanty, X. Xiao, U. H. E. Hansmann

published in

From Computational Biophysics to Systems Biology (CBSB07),
Proceedings of the NIC Workshop 2007,
Ulrich H. E. Hansmann, Jan Meinke, Sandipan Mohanty,
Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **36**, ISBN 978-3-9810843-2-0, pp. 99-102, 2007.

http://www.fz-juelich.de/nic-series/volume36

Dimensionality Reduction Techniques for Protein Folding Trajectories

T. Eitrich, S. Mohanty, X. Xiao, and U. H. E. Hansmann

John von Neumann Institute for Computing,
Research Centre Jülich, 52425 Jülich, Germany
E-mail: {*t.eitrich, s.mohanty, x.xiao, u.hansmann*}*@fz-juelich.de*

In our work we analyze large and high dimensional data from protein folding simulations. The main goals are to extract the underlying dimensionality, to find a small number of features that describe the data with high accuracy and to find interesting clusters in the data: in this work we treat this as a problem of dimensionality reduction. Dimensionality reduction aims to find a mapping of the original space into a space of a few interesting dimensions, which the user then can use for interpretation and analysis. We study modern dimensionality reduction techniques and combine them with promising distance measures, suitable for the description of dissimilarities between the data points generated by the package ProFASi - a Protein Folding and Aggregation Simulator.

1 Introduction

Understanding the folding of proteins is one of the most challenging problems in computational biology. We refer to protein folding as the process by which a protein assumes its native state. Protein folding trajectory data is high dimensional and thus hard to interpret after the simulation. Dimensionality reduction methods, that map the data into a new space of fewer dimensions while preserving as much relevant information as possible, can be used to find meaningful low dimensional structures in the original high dimensional datasets.

Several techniques do exist, which we may divide into

- linear methods, like principal component analysis (PCA)[6] and multi dimensional scaling (MDS)[2], and
- nonlinear methods, like locally linear embedding (LLE)[8], kernel PCA[9] and Isomap[10].

We refer to this methods as unsupervised embedding algorithms. Supervised embedding methods, which take additional properties into account, are discussed in ref 5.

2 Material and Methods

Our dataset for this work consists of a folding trajectory of a 49 residue alpha-beta protein with PDB id 2GJH, generated using the program package ProFASi: a Protein Folding and Aggregation Simulator[4]. Each data point represents the conformation of the molecule at a certain Monte Carlo time. Each successive pair of data points are separated by 1000 Monte Carlo sweeps. For this analysis, we have used 1000 such points, which were selected to span one particular folding event in the simulations.

We have used two dimensionality reduction methods, MDS and Isomap. The following are the main steps in the Isomap[10] method.

- Compute dissimilarities between all points.
- Construct a neighborhood graph according to the number of neighbors k_n
- Based on the neighborhood information, compute shortest paths between all points.
- Embed the data into a new d-dimensional space using an eigenvalue method.

In the case of MDS, only the first and the last step need to be done (no neighborhood parameter k_n), because the dissimilarity measure is assumed to be uniformly good for all pairs of points.

The success of dimensionality reduction methods highly depends on a reasonable choice of the dissimilarity measure, since these methods use distances between objects instead of the objects themselves. Thus, any information we want to preserve must be represented by the distance measure. In our study we have used and compared three different measures, i.e.

- the minimized root mean square deviation (RMSD) between atomic coordinates, as it was used in ref 3,
- the RMSD of the dihedral angels, as introduced in ref 1, and
- a two dimensional structural similarity measure based on dihedral angle distributions and a Fourier transformation process, described in ref 7.

3 Results and Discussion

In our tests we have applied different dimensionality reduction techniques as well as various distance measures for our folding trajectory dataset. We observed, that both, MDS and Isomap (using small neighborhood sizes) lead to interesting embedding dimensions. We observed, that all three distance measures give results which fit to some of the characteristics of the simulation, like energy or RMSD to the native state.

In Fig. 1, we show results using Isomap with 10 neighbors and the atomic RMSD distance measure. Plotted is the RMSD to the native state against the first embedding dimension - a descriptor for how well the protein has folded in the simulation. Please note, that the colors in all pictures express the time steps. In Fig. 2, results using MDS with the dihedral RMSD distance measure are given. Plotted is the total energy value against the first embedding dimension. In Fig. 3, MDS together with the Fourier measure was applied. Given is the helix content against the first embedding dimension.

So far we could not observe huge differences between the linear MDS and the nonlinear Isomap, which leads us to the conclusion, that for the case of protein folding trajectories, we have examined, the choice of a reasonable distance measure also leads to adequate results for linear embedding methods.

4 Summary and Future Work

In our work we studied linear and nonlinear dimensionality reduction techniques. Our application field is protein folding, for which we try to find and analyze new embedding coordinates. Future work will concentrate on the analysis of embedding data. Especially, we will use supervised methods to cluster and classify embedded data.

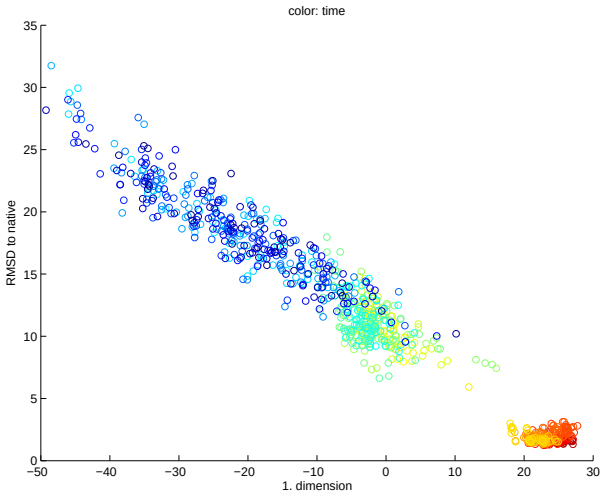

Figure 1. First embedding coordinate is correlated with the RMSD to native value.

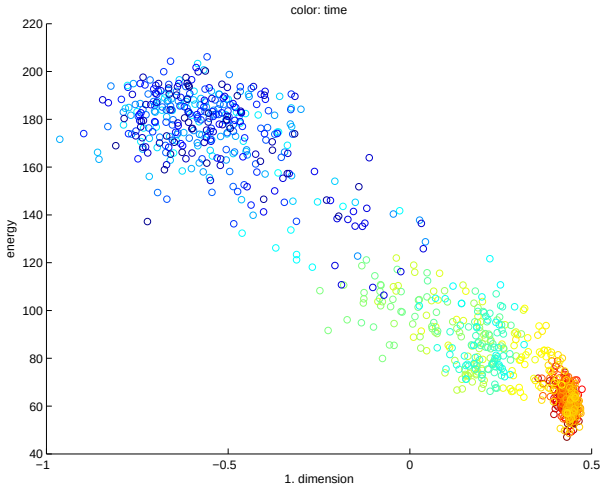

Figure 2. First embedding coordinate is correlated with the energy.

Acknowledgments

Part of this work are supported by research grants of the National Science Foundation, USA (CHE-0313618) and the National Institutes of Health, USA (GM 62838).

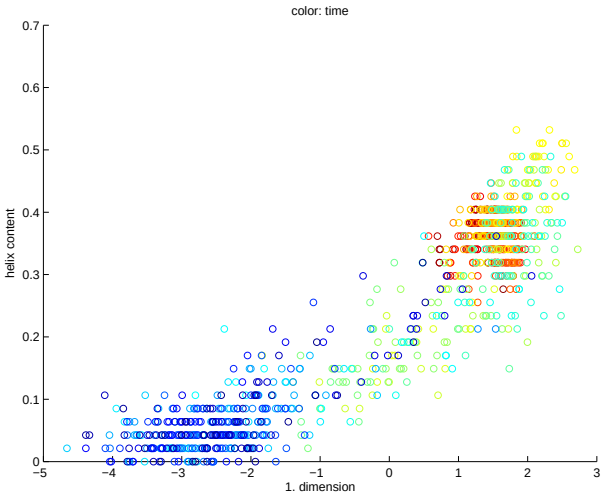

Figure 3. First embedding coordinate is correlated with the helix content.

References

1. A. Collins, R.I. Cooper, and D.J. Watkin *Structure matching: measures of similarity and pseudosymmetry. J. Appl. Cryst.* **39**, *842–849, 2006.*
2. *T.F. Cox and M.A.A. Cox.* Multidimensional scaling. *Chapman & Hall, London, 1994.*
3. *P. Das, M. Moll, H. Stamati, L.E. Kavraki, and C. Clementi* Low-dimensional free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. PNAS **103**(26), 9885–9890, 2006.
4. A. Irbäck and S. Mohanty *ProFASi: a Monte Carlo simulation package for protein folding and aggregation. J. Comput. Chem.* **27**, *1548–1555, 2006.*
5. *T. Iwata, K. Saito, N. Ueda, S. Stromsten, T.L. Griffiths, and J.B. Tenenbaum* Parametric embedding for class visualization. Advances in Neural Information Processing Systems **17** (NIPS2004), 617–624, 2005.
6. I.T. Jolliffe *Discarding variables in a principal component analysis. I: artificial data. Applied Statistics* **21**, *160–173, 1972.*
7. *N. Kandiraju, S. Dua, and S. Conrad* Dihedral angle based dimensionality reduction for protein structural comparison. Proc. Conf. ITCC2005, 14–19, 2005.
8. S.T. Roweis and L.K. Saul. *Nonlinear dimensionality reduction by locally linear embedding. Science* **290***(5500), 2323–2326, 2000.*
9. *B. Schölkopf, A.J. Smola, and K.-R. Müller.* Kernel principal component analysis. In Advances in kernel methods: support vector learning, MIT Press, 327–352, 1999.
10. J.B. Tenenbaum, V. deSilva, and J.C. Langford. *A global geometric framework for nonlinear dimensionality reduction. Science* **290***(5500), 2319–2323, 2000.*