A Principled Comparative Analysis of Dimensionality Reduction Techniques on Protein Structure Decoy Data

Rohan Pandit[1] Amarda Shehu[2,3,4,‡]
[1]Thomas Jefferson High School, Alexandria, VA
[2]Dept. of Computer Science, [3]Dept. of Bioengineering, [4]School of Systems Biology,
George Mason University, Fairfax, VA 22030
[‡]Correspondence: amarda@gmu.edu

Abstract

In this paper we investigate the utility of dimensionality reduction as a tool to analyze and simplify the structure space probed by *de novo* protein structure prediction methods. We conduct a principled comparative analysis in order to identify which techniques are effective and can be further used in decoy selection. The analysis allows drawing several interesting observations. For instance, many of the reportedly state-of-the-art non-linear dimensionality reduction techniques fare poorly and are outperformed by linear techniques that tend to have consistent performance across various protein structure data sets. The analysis in this paper is likely to open the way to new techniques that make use of the reduced dimensions to organize protein structure data so as to automatically detect the elusive native structure of a protein. We show some preliminary results in this direction.

1 Introduction

The three-dimensional (native) structure into which a specific chain of amino acids folds under physiological conditions determines to a great extent the molecular partners onto which a protein will bind and thus the recognition events in which a protein will participate in the cell [3]. In the post-genomic era, there is an increasing need for computational methods to predict such structures in order to keep up with the millions of uncharacterized protein-encoding gene sequences [2]. A popular group of methods that operate in the absence of a template structure, known as *de novo* structure prediction methods, generate thousands or more low-energy structures of a given amino-acid sequence under the umbrella of stochastic optimization [18]. These methods then conduct what is referred to as decoy selection; the computed structures are typically organized into clusters based on structural similarity and offer any of the top k populated clusters as possibly containing the native structure; common values of k range from 1 to 10.

In this paper we investigate the utility of dimensionality reduction as a tool to analyze and simplify protein structure spaces probed by *de novo* structure prediction methods. Dimensionality reduction techniques are often employed to extract the collective variables that explain the dynamics of biomolecules, but their applicability is typically restricted to structures obtained via Molecular Dynamics (MD) or in the wet laboratory [6,8,9,13,16,17]. Moreover, existing studies select a specific dimensionality reduction technique for application, but fail to conduct a principled comparative analysis of different available techniques.

We conduct here a comparative analysis of representative dimensionality reduction techniques on protein structures generated via the Rosetta *de novo* structure prediction protocol [10]. Our analysis builds on established research in the machine learning community on extracting the intrinsic dimension of a dataset and then comparing such-dimensional embeddings obtained from different techniques via informative metrics. Typically, such analyses are conducted on synthetic data, where the intrinsic dimension is a priori known. In contrast, in our setting, the intrinsic dimension of the structure space of a protein is not known. For this reason, we first determine the intrinsic dimension of the structure space of each protein which then facilitates an objective comparative analysis of different dimensionality reduction techniques.

This paper makes several contributions. A principled comparative analysis of dimensionality reduction techniques on protein structure data is likely to be useful in organizing protein structure data for predicting native structures. The findings here are also likely to be of interest to the machine learning community, as they demonstrate the behavior of well-understood and mature techniques on real-world, protein structure data. This data is a discrete representation of an unknown structure space that is possibly sampled at widely-different densities by stochastic optimization algorithms in *de novo* structure prediction. On such data, we show, for instance, that many of the reportedly sophisticated nonlinear dimensionality reduction techniques fare poorly and, instead, linear tech-

niques perform best or have consistent good performance across various datasets and parameter settings.

2 Method

2.1 Maximum Likelihood Estimation of Intrinsic Dimension

We determine the intrinsic dimension of each dataset via the maximum likelihood estimator (MLE) algorithm proposed in [12]. Rather than relying on an analysis of accumulation of variance, the algorithm takes a geometric and statistical approach. Let us regard the natural dataset as a list of points $i \in \{1, \ldots, n\}$, with corresponding coordinates $x_1, \ldots, x_n$ in some high-dimensional space $\mathcal{R}^m$. In our case, n is the number of Rosetta-computed structures for a given protein sequence, and m is the number of Cartesian coordinates of the CA atoms ($3x$ number of amino acids in the protein sequence).

The intrinsic dimension $\hat{m}_k$ of the manifold where the data are assumed to lie can be calculated from the local intrinsic dimension around each point x_i as a function of the number k of nearest neighbors: $\hat{m}_k = 1/n \sum_{i=1}^{n} \hat{m}_k(x_i)$. The local intrinsic dimension around a point x is calculated as: $\hat{m}_k(x) = [1/(k-1) \sum_{j=1}^{k-1} \log (E_k(x)/E_j(x))]^{-1}$, where j varies over the number of selected neighbors, and $E_j(x)$ is the Euclidean distance from a fixed point x to its j-th nearest neighbor.

Rather than relying on a visual interpretation of changes in $\hat{m}_k$ as a function of k, an intrinsic dimension $\hat{m}$ is calculated as the average over a range of small to moderate values $k \in \{k_1, \ldots, k_2$ as in: $\hat{m} = 1/(k_2 - k_1 + 1) \sum_{k=k_1}^{k_2} \hat{m}_k$. We employ the above method for its improved accuracy over correlation and geometric methods on synthetic and natural datasets. Additionally, a publicly-available implementation of this algorithm exists in the Matlab toolbox

2.2 Dimensionality Reduction Techniques

We investigate here linear and nonlinear techniques. Our baseline linear technique is Principal Component Analysis (PCA). Four nonlinear techniques are considered, Locally Linear Embedding (LLE), Isomap, Local Diffusion Map (LDFMap), and kernel PCA (KPCA). A description of each of these techniques is provided in Ref. [19]. Briefly, PCA translates and rotates the dataset so the new axes maximize the variance in the data. In our application of PCA, all structures are first aligned to a reference structure selected arbitrarily to be the first one computed by the Rosetta protocol. The alignment process is the one typically used to compute the least root-mean-squared deviation (lRMSD) between two structures [15]. This is done so as to remove differences due to the trivial rigid-body motions and instead focus on structural differences. The eigendecomposition is conducted on the matrix of atomic displacements from an average model computed over all aligned models. KPCA is a reformulation of PCA in a high-dimensional space resulting from a (polynomial) kernel function.

LLE and Isomap are similar to each-other, both relying on a graph representation of the points, where each point is connected to its k nearest neighbors, with k being a user parameter. In LLE, the data points are linear combinations of their nearest neighbors. In the low-dimensional embedding, LLE seeks to retain the reconstruction weights in the linear combinations. In Isomap, the distance between two points is the shortest path in the graph. The resulting pairwise distance matrix is subjected to eigendecomposition.

While PCA, KPCA, and Isomap are global techniques, LLE and LDFMap are local techniques. LDFMap is a diffusion-based technique proposed in [16] to correct issues in the global diffusion map technique. The latter performs an eigendecomposition of a matrix of so-called diffusion distances and re-interprets the distance between two points as a diffusion distance. LDFMap relies on the notion of a local scale and local dimensionality rather than assuming that all data lie on the same dimensional manifold.

2.3 Performance Metrics

We consider three metrics here, trustworthiness, continuity, and Local Continuity Meta-Criterion (LCMC).

Trustworthiness Trustworthiness and continuity have been proposed in [20]. Trustworthiness measures to what extent the local neighborhood of a point in the lower-dimensional embedding space is correct in the high-dimensional original space. Continuity measures to what extent the local neighborhood in the high-dimensional, original space is preserved in the low-dimensional, embedding space. Specifically, given k nearest neighbors of a point, trustworthiness $T_k(i)$ is measured as $1 - 2/(N \cdot k - 3 \cdot k - 1) \sum_{i=1}^{n} \sum_{j \in U_k(i)} [r(i,j) - k]$. In this equation, $U_k(i)$ is the set of points in the neighborhood of size k of point i in the low-dimensional, embedding space but not in the original, high-dimensional space. $r(i,j)$ is the rank of point j with respect to point i in the original, high-dimensional space; that is, how many points are closer to point i than point j in the original space. This rank must be greater than k, since per the above formula j is not among the k nearest neighbors of point i in the original space. A higher value of trustworthiness means the reduction is more trustworthy in preserving nearest neighbors.

Continuity Continuity is defined in a similar manner. Now let $V_k(i)$ be the set of elements in the neighborhood of point i in the original space but not in the embedding space. Let $\hat{r}(i,j)$ represent the rank in the embedding space. Then, given a number k of nearest neighbors, continuity $C(k)$ is defined as: $1 - 2/(N \cdot k - 3 \cdot k - 1) \sum_{i=1}^{n} \sum_{j \in V_k(i)} [\hat{r}(i,j) - k]$.

Characteristics of Trustworthiness and Continuity Trustworthiness and continuity capture the quality of an embedding, distinguishing two types of errors. Faraway points that become neighbors decrease trustworthiness. Neighbors embedded faraway from each-other decrease continuity. Trustworthiness and continuity on random neighbors are approximately 0.5 [20]. Since both metrics vary with k, it is in principle not clear how to compare two embeddings obtained with two different dimensionality reduction techniques. Since higher values of both these metrics are desired, we summarize the progression of trustworthiness and continuity as a function of k with the area under the curve (AUC). In section 3 we show these curves as a function of k on a selected dataset and then employ AUC to compare the different reduction techniques across many different datasets.

LCMC LCMC belongs to a family of local continuity meta-criteria that is defined as the average size of the overlap of the k nearest neighbors in the high-dimensional data (original space) and the low-dimensional embedding space [4, 5]. The overlap is tracked as the number of nearest neighbors is increased. The embedding space employed here is the one obtained after determining the intrinsic dimension, as described above. The higher the number of nearest neighbors, the higher the size of the expected overlap, eventually reaching 100%. A reduction technique that reaches higher LCMC values with a smaller number of neighbors is considered more useful at providing more faithful embeddings of a high-dimensional space.

Implementation Details Most of the reduction techniques compared here are available via the python scikit.learn library in python. LDFMap has been provided to us by the original authors. The code to calculate LCMC has also been provided to us by the original authors. Determining the intrinsic dimension on each of the protein datasets takes about 10 hours in the Matlab implementation. The running time of the different reduction techniques and the calculation of the various metrics takes anywhere from 10 to 20 hours. The latter were run on the Mason Argo cluster.

3 Results

3.1 Experimental Setup

The 11 proteins on which we conduct our analysis are shown in detail in Table 3.1. Column 1 shows the PDB id of the known native structure in each case (PDB refers to the Protein Data Bank where wet-laboratory experts deposit resolved native structures of proteins [1]). The selected proteins are diverse in fold (column 2); based on PDB *macromolecule annotations* 3 are all or mainly α, 5 are all or mainly β, 2 are $\alpha + \beta$, and 1 is largely void of secondary structures. The selected proteins range in length from 54 to 139 amino acids (column 3). For each protein sequence, more than $50,000$ structures are generated via Rosetta on the Mason Argo cluster; the actual size of the structural ensemble Ω generated for each protein is shown in column 4. For the purpose of analysis, only structures with Rosetta (score12) all-atom energies no higher than the median are retained for each protein. This energetic filtering, which we conduct in order to focus on structures with lower energies, effectively reduces the number of structures subjected to the dimensionality reduction techniques in half.

Table 1: Datasets and Intrinsic Dimension

PDB ID	Fold	Length	$\lvert\Omega\rvert$	$\hat{m}$
1BQ9	β	54	$53,664$	11
1ISUA	`coil`	62	$60,361$	15
1C8CA	β	64	$53,323$	10
1SAP	β	66	$51,210$	9
1WAPA	β	75	$51,842$	9
1AOY	α	78	$52,219$	10
1CC5	α	83	$51,688$	14
1TIG	$\alpha + \beta$	94	$52,100$	10
1HHP	β	99	$52,160$	12
2EZK	α	99	$50,193$	10
1FWP	$\alpha + \beta$	139	$53,134$	10

3.2 Intrinsic Dimension

The intrinsic dimension $\hat{m}$ calculated for each of the 11 dataset with the MLE algorithm is shown in column 5 in Table 3.1. The results show that $\hat{m}$ varies between 9 and 15. No correlation is observed with protein length. It is worth noting that the $\hat{m}$ values are significantly lower than the (original) dimension m fed to the various reduction techniques; based on the lengths shown in column 3 in Table 3.1, m varies from 54 x 3 to 139 x 3 Cartesian coordinates. The determined $\hat{m}$ values represent up to 50 fold reductions over m. Further details are provided by showing $\hat{m}_k$ values over a range $\{k_1, \ldots, k_2\}$ of nearest neighbors.

3.3 Performance Comparison

Figure 2 compares the different dimensionality reduction techniques based on the three performance metrics listed in Section 2 on each of the 11 different datasets. LLE is consistently the worst performer

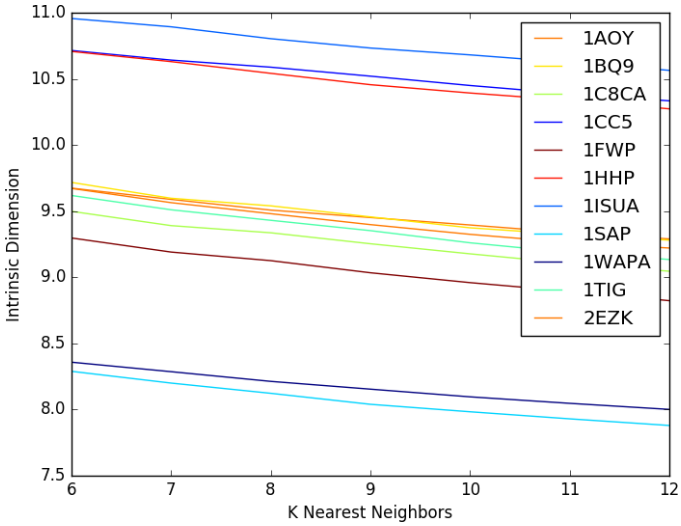

Figure 1: The intrinsic dimension $\hat{m}_k$ is shown as a function of the number k of nearest neighbors for each of the different proteins.

across the different techniques, followed by KPCA. LDFMap also performs unreliably across the different datasets. The top two techniques are Isomap and PCA. PCA is consistently the top performer. Further details are provided on a selected protein (with PDB id of the native structure 1FWP). Figure 3 tracks the metrics as a function of the number of neighbors.

3.4 Utility for Decoy Selection

We demonstrate the utility of dimensionality reduction techniques for discriminating near-native from non-native structures predicted from *de novo* structure prediction software, such as Rosetta. The above results suggest that two top reliable techniques across diverse proteins are Isomap and PCA. So, it is worth investigating these two techniques further on their ability to capture not only the topology of the structure space of a particular protein but also expose characteristics that may be useful in extracting regions of the space near the native structure.

In particular, our analysis (not shown here) is that while Isomap is often outperformed by PCA when metrics are calculated on embedding of intrinsic dimension, it performs very well in two-dimensional embeddings that are often useful for visualization. We illustrate this via a visual demonstration in Figure 4 for a selected protein (with native structure under PDB id 1FWP). The Rosetta-generated structures are compared via CA lRMSD to the native structure. The projections of the computed structures on the top two (based on variance) eigenvectors are color-coded based on their CA lRMSD values from the native structure. We refer to these projections as collective coordinates (CVs), and visualize the two-dimensional maps of the

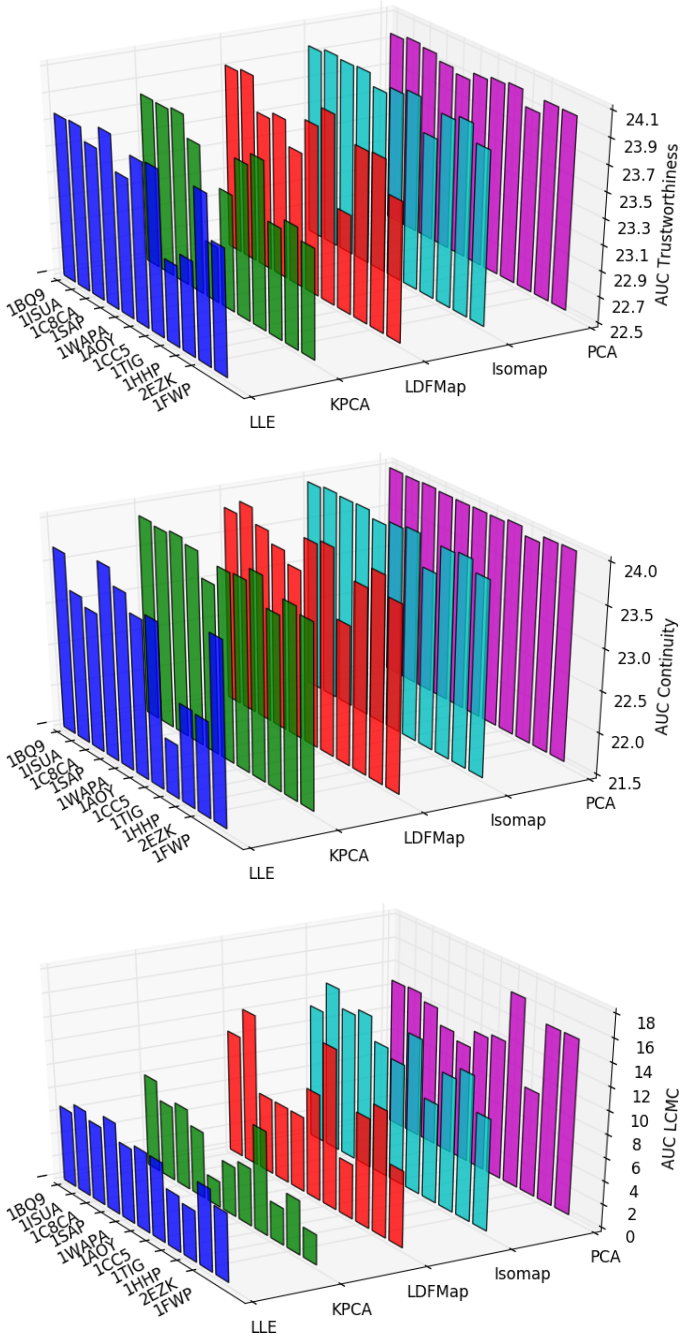

Figure 2: All techniques (in different colors, x axis) are compared across the different proteins (y axis) on three metrics, trustworthiness, continuity, and LCMC (z axis).

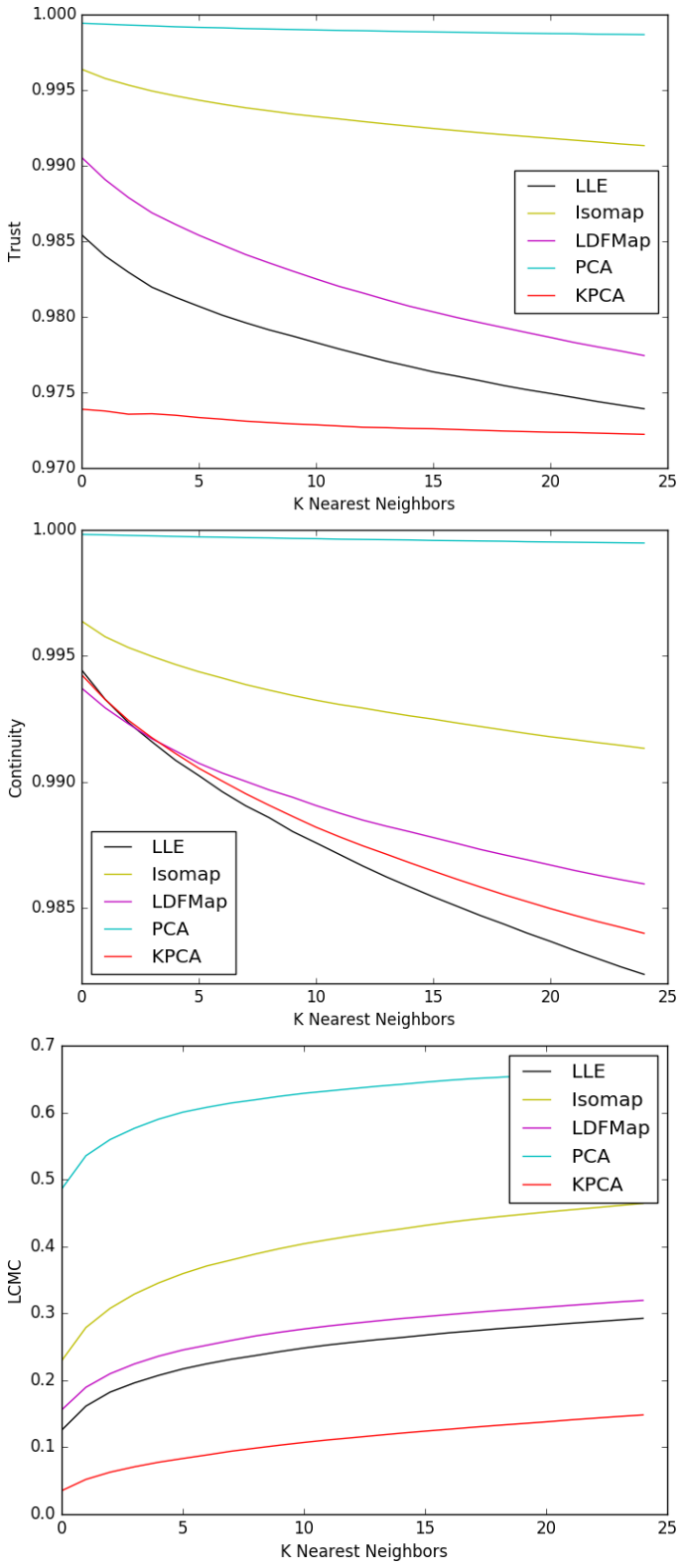

Figure 3: Performance of all techniques (in different colors, legend) is detailed on a selected protein on the three metrics.

structure space probed by Rosetta as obtained from Isomap.

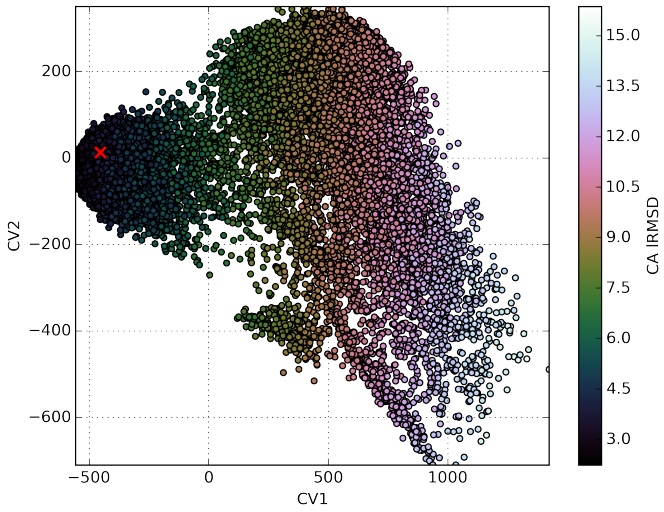

Figure 4: Projections of Rosetta-computed structures on the top two eigenvectors are color-coded based on their CA lRMSD from the known native structure (PDB id 1FWP).

Figure 4 shows strong co-localization; that is, structures with low CA lRMSDs from the native structures are also closer to the native structure (marked with a red X) in the two-dimensional projections. This specific results suggest that proceeding via clustering and other machine learning techniques on the space of collective variables may be useful in decoy selection. In particular, in the context of supervised classification, where the number of features needs to be small and the features need to be informative, the collective coordinates elucidated from dimensionality reduction techniques may prove useful.

4 Conclusion

The results presented in this paper allow drawing several conclusions. First, the intrinsic dimensionality of structure data generated by *de novo* structure prediction methods is likely to be low; at least one order of magnitude less than the number of dihedral angle fragments usually employed for sampling by these methods. This finding is in agreement with growing knowledge of intrinsic co-operative protein dynamics. Second, linear dimensionality reduction techniques are likely to perform reliably well on protein structure data. It is worth noting that previous comparative reviews of dimensionality reduction techniques have reached similar conclusions on different data sets. Work in [19] shows that nonlinear techniques perform well on synthetic datasets but do not outperform PCA on real-world data. One of the con-

tributions of this paper is further supporting this conclusion on protein structure data.

The methodology presented in this paper can also be considered a protocol for determining which dimensionality reduction technique to employ in reducing biomolecular structure data both for analysis of biomolecular dynamics as well as for simulation of the dynamics via stochastic optimization algorithm that operate on reduced variable spaces [7, 14]. Moreover, in the strict context of protein structure data, the reduced dimensions may be useful to organize *de novo* computed structures of a protein sequence for decoy selection. Future research will investigate this direction, as well as broaden the comparative analysis to more datasets, more dimensionality reduction techniques, and more metrics [11].

Acknowledgements

This work is supported in part by NSF-CCF1421001, NSF-ACI1440581, and NSF-IIS1144106. Computations were run on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University.

References

[1] H. M. Berman, K. Henrick, and H. Nakamura. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.*, 10(12):980–980, 2003.

[2] C. E. Blaby-Haas and V. de Crécy-Lagard. Mining high-throughput experimental data to link gene and function. *Trends Biotechnol*, 29(4):174–182, 2013.

[3] D. D. Boehr and P. E. Wright. How do proteins interact? *Science*, 320(5882):1429–1430, 2008.

[4] L. Chen. *Local multidimensional scaling for nonlinear dimensionality reduction, graph layout, and proximity analysis*. PhD thesis, University of Pennsylvania, 2006.

[5] L. Chen and A. Buja. Local multidimensional scaling for nonlinear dimensionality reduction, graph drawing, and proximity analysis. *K Am Stat Assoc*, 104:209–219, 2009.

[6] R. Clausen, B. Ma, R. Nussinov, and A. Shehu. Mapping the conformation space of wildtype and mutant h-ras with a memetic, cellular, and multiscale evolutionary algorithm. *PLoS Comput Biol*, 11(9):e1004470, 2015.

[7] R. Clausen and A. Shehu. A data-driven evolutionary algorithm for mapping multi-basin protein energy landscapes. *J Comp Biol*, 22(9):844–860, 2015.

[8] 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. *Proc. Natl. Acad. Sci. USA*, 103(26):9885–9890, 2006.

[9] B. J. Grant, A. A. Gorfe, and J. A. McCammon. Ras conformational switching: Simulating nucleotide-dependent conformational transitions with accelerated molecular dynamics. *PLoS Comput Biol*, 5(3):e1000325, 2009.

[10] A. Leaver-Fay et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol*, 487:545–574, 2011.

[11] J. A. Lee and M. Verleysen. Quality assessment of nonlinear dimensionality reduction based on k-ary neighborhoods. *J Mach Learn Res*, 4:21–35, 2008.

[12] I. S. Lim, de H. P. Ciechomski, S. Sarni, and D. Thalmann. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, volume 17, pages 50–55, Cambridge, MA, 2004. MIT Press.

[13] G. G. Maisuradze, A. Liwo, and H. A. Scheraga. Principal component analysis for protein folding dynamics. *J Mol Biol*, 385(1):312–329, 2009.

[14] T. Maximova, E. Plaku, and A. Shehu. Computing transition paths in multiple-basin proteins with a probabilistic roadmap algorithm guided by structure data. In *IEEE Intl Conf Bioinf and Biomed (BIBM)*, 2015.

[15] A. D. McLachlan. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallogr. A.*, 26(6):656–657, 1972.

[16] M.A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi. Determination of reaction coordinates via locally scaled diffusion map. *J. Chem. Phys.*, 134(12):124116, 2011.

[17] G. Scarabelli and B. J. Grant. Mapping the structural and dynamical features of kinesin motor domains. *PLoS Comput Biol*, 9(11):e1003329, 2013.

[18] A. Shehu. Probabilistic search and optimization for protein energy landscapes. In S. Aluru and A. Singh, editors, *Handbook of Computational Molecular Biology*. Chapman & Hall/CRC Computer & Information Science Series, 2013.

[19] Laurens JP van der Maaten, Eric O Postma, and H Jaap van den Herik. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(1-41):66–71, 2009.

[20] J. Venna and S. Kaski. Local multidimensional scaling with controlled tradeoff between trustworthiness and continuity. In *5th Workshop on Self-Organizing Maps*, pages 695–702, 2005.