Dimensionality Estimation of Protein Dynamics Using Polymer Models

Joshua L. Phillips
Middle Tennessee State University
Murfreesboro, Tennessee
Joshua.Phillips@mtsu.edu

Michael E. Colvin
University of California, Merced
Merced, California
mcolvin@ucmerced.edu

Shawn Newsam
University of California, Merced
Merced, California
snewsam@ucmerced.edu

ABSTRACT

Molecular dynamics (MD) simulation is a powerful technique for sampling the conformational landscape of natively folded proteins (NFPs) and structurally dynamic intrinsically disordered proteins (IDPs). NFPs and IDPs can be viewed as nonlinear dynamical systems that exercise available degrees of freedom to explore their energetically-accessible conformation landscape. Dimensionality estimators have emerged as useful tools to characterize nonlinear dynamical systems in other domains, but their application to MD simulation has been limited due to thermal noise and a lack of ground-truth data. We develop a series of increasingly complex biopolymer models which exhibit a range of dynamics we seek to characterize in MD simulations (stochastic dynamics, helical structures, partially folded states, and correlated motions) and are of known dimensionality. We utilize the maximum-likelihood dimension (MLD) estimator to investigate the effects of thermal noise and noise-smoothing techniques on the estimates obtained from the polymer models and MD simulations of two NFPs and two IDPs. We find that under certain noise/smoothing conditions, the MLD over/under-estimates the true dimensionality of the models in a predictable manner, allowing us to relate differences between MLD estimates to differences between NFP and IDP motions for classification of biomolecular systems based on their dynamics.

CCS CONCEPTS

• **Applied computing → Molecular structural biology**;

KEYWORDS

Dimensionality Estimation, Molecular Dynamics, Polymer Models

ACM Reference Format:
Joshua L. Phillips, Michael E. Colvin, and Shawn Newsam. 2018. Dimensionality Estimation of Protein Dynamics Using Polymer Models. In *ACM-BCB'18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, August 29-September 1, 2018, Washington, DC, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3233547.3233713

ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA

ACM ISBN 978-1-4503-5794-4/18/08…$15.00
https://doi.org/10.1145/3233547.3233713

1 INTRODUCTION

The "protein structure-function" paradigm, which states that proteins adopt nearly rigid three-dimensional structures that are responsible for their function, is one of the central tenets of molecular biology, yet some protein domains exist as intrinsically disordered forms. The classification of IDPs is challenging because traditional approaches to protein structure classification often rely on static structural features which IDPs often lack. Molecular dynamics (MD) simulation is an approach for sampling the conformation space of IDPs. However, analyzing these simulations remains a challenge due to the overwhelming complexity of the conformation space for even small IDPs. This paper investigates dimensionality estimation as a potential technique for studying the dynamics of MD simulations of biopolymers.

The interest in the dimensionality of protein dynamics stems from the intuitive idea that NFPs, when folded, should exhibit dynamics with a dimensionality of zero because they subsist in one singular point in the high-dimensional conformation space. The more *unfolded* a protein becomes, the higher the dimensionality of the motion will become as well. In contrast, IDPs should never exhibit zero-dimensional dynamics since no native structure exists. Even collapsed IDPs should exhibit higher dimensionality motion than NFPs of equivalent length in terms of the number of residues since the physical forces inducing folding in the NFPs will constrain the conformation space to a lower-dimensional manifold. The dimensionality of protein dynamics could therefore potentially be used to distinguish between different flavors of IDPs or between IDPs and NFPs much like the static structural features of folded proteins have been used to classify NFPs. We introduce a polymer-based framework for studying the dimensionality estimation of unfolded protein dynamics which provides ground truth data for interpreting such results.

Many methods have been developed for calculating the dimensionality of nonlinear processes based on data samples [3, 4, 6, 10, 13, 18]. Perhaps the most commonly employed techniques emerged within the context of dimensionality reduction, where heuristic methods (e.g. gaps in eigenspectra) are used to deduce the intrinsic dimensionality of the data. Howevere, methods based on nearest-neighbor properties and statistics are often preferred in practice.

2 METHODS

2.1 Maximum Likelihood Estimator of Dimensionality

The algorithm of choice for this study is the maximum likelihood dimension (MLD) estimator of Levina and Bickel [13]. We refer the reader to [13] for the details of the MLD algorithm and use

the harmonic mean as suggested by MacKay and Ghahramani [14]. This was shown to produce better estimates for small scales (smaller values of nearest neighbors, k). However we *average over scales before averaging over points*, whereas their implementation averages over all points and then over scales. Our approach thus produces a local (pointwise) estimate of dimensionality rather than a global one which is important since our systems have time-varying dynamics. Global estimates can still be obtained by taking the harmonic mean of the local estimates. Input to the MLD is the *pairwise distances* between points (protein structures) in a dataset. Pairwise distances are computed using common structure comparison measures (e.g. C_α root-mean-squared distance, Euclidean distance between sin-cos pairs of the backbone $\Phi - \Psi$ angles).

2.2 Polymer Models

The models below have known dimensionality and exhibit specific dynamics of interest, allowing us to quantitatively determine the accuracy of estimates. Similar behavior between the dimensionality estimates of the MD simulations and the polymer models should allow us to conclude that similar dynamics are present.

2.2.1 Semirigid Helix. The first model consists of a set of l virtual bond segments all a=3.8Å in length which is the typical distance between subsequent C_α atoms along a protein chain. At the junction between two contiguous links, two angles (θ and ϕ) describe the orientation of the second link relative to the first. The angle θ describes the inclination of a link relative to the prior link, while the angle ϕ describes the azimuthal rotation of the link relative to the prior link. A rigid helix, analogous to the folded protein α-helix, is formed by setting all ϕ angles along the chain to be random values chosen from a Gaussian distribution with mean μ_ϕ=0.83 and standard deviation σ_ϕ=0, and all θ angles chosen from a Gaussian distribution with mean μ_θ=1.54 and standard deviation σ_θ=0. This is considered the fixed, folded conformation. Ensembles model the fluctuations around this conformation and are generated by sampling angles from a zero-mean Gaussian distribution with non-zero standard deviation. The ensemble should have a dimensionality of zero in the absence of noise and achieve the maximum theoretical dimensionality of the polymer when the noise is large.

2.2.2 Half-folded Helix. The half-folded helix model simulates a helical polymer in which one portion may fold/unfold while the other portion remains folded. The half-folded helix is similar to the semirigid helix but is separated into two connected segments of equal size that have different noise properties. The ϕ angles of both segments are sampled from a Gaussian distribution with mean μ_ϕ=0.83 and standard deviation σ_ϕ=0.01, and the θ angles of both segments are sampled from a Gaussian distribution with mean μ_θ=1.54 and standard deviation σ_θ=0.01. However, a varying amount of additional noise is injected into the θ angle of the "unfolded" segment. This noise is sampled from a Gaussian with zero mean and standard deviation $\sigma_{\theta_{unfolded}}$. For large values of this parameter, the polymer will consist of a folded region which remains in a helix conformation while the unfolded region is disordered. Thus, the effects on the dimensionality estimator due to a different number of significant dimensions (or mixed levels of noise) can be investigated using this model.

2.2.3 Correlated Helix. Proteins often exhibit coordinated motions [12], where several parts of the chain move in response to the motions along other parts of the chain. These coordinated motions represent restricted degrees of freedom (DoFs) and thus dynamics with lower dimensionality. This decrease should be detected by the estimator, especially if the amplitudes of the motions are large compared to the noise.

The model is use to generate a polymer trajectory which exhibits coordinated transitions from a helical conformation to an elongated chain, and then the reverse. This is accomplished by starting with all angles set so that the polymer is in a helical conformation (ϕ=0.83 and θ=1.54). The theta angles are then decremented by a small amount, $\epsilon_\theta + \mathcal{N}(0, 0.01)$, and a new conformation is generated. This process in repeated until the θ angles are less than 0.087 radians (nearly a straight rod). The angles are then incremented by $\epsilon_\theta + \mathcal{N}(0, 0.01)$ on each step to bring the polymer back to a helical conformation (until θ>1.54). This model has only one coordinated motion and is thus a simple one-dimensional model.

Models with more than one coordinated motion are crafted by breaking the polymer into distinct segments, s_i, each with a unique increment $\epsilon_\theta^{s_i}$ and a unique amount of noise added to the increment (but with the same standard deviation). Note that even if $\epsilon_\theta^{s_i}$ is set to the same value for all segments, the difference between segments in the small amount of noise added to the increments at each step, $\mathcal{N}(0, \sigma_{\epsilon_\theta})$, would make the dimensionality of the system equal to the number of segments. For simplicity, the total number of independent ϵ_θ^s values used is referred to as the number of correlated dimensions (d_{cor}) in the trajectory. In addition, a small amount of Gaussian noise with zero mean and standard deviation, $\sigma_{\theta,\phi}$, is added to all angles, independent of the coordinated walk in angle space presented above.

2.3 Noise Smoothing

Thermal noise is a challenge and tuning estimators to ignore noise using data prepreprocessing (smoothing) is appealing. We utilize the discrete fourier transform (DFT) to remove noise. We form a temporal vector of complex values for each of the θ or ϕ angles along the chain by transforming these values to $e^{i\theta}$ or $e^{i\phi}$ where $i = \sqrt{-1}$ and then compute the DFTs of these vectors. We set to zero those signal components above a frequency threshold or below an amplitude threshold. After taking the inverse DFT, the smoothed angular values are extracted from the complex vectors. In the experiments below, we compare both the frequency and amplitude cutoff approaches across a range of thresholds.

2.4 Molecular Dynamics Simulations

Molecular dynamics simulations were performed on two NFPs and two IDPs which cover the major structural classes of NFPs and IDPs. The first NFP, GB1 [PDB:1GB1] [7], is a 16 amino acid long fast-folding protein that spontaneously adopts a β-hairpin conformation. The second NFP, Trp-cage [PDB:1L2Y] [15], is a 20 amino acid long fast-folding protein that spontaneously adopts a mainly α-helical conformation. Therefore, these two proteins cover both of the broadest structural classes of NFPs.

The first IDP, Nsp1, is a 25 amino acid subsequence of the full length wildtype FG-nucleoporin NSP1 [GenBank: NP_012494] [5],

and the second, Nup116, is a 25 amino acid long subsequence of the full length wildtype FG-nucleoporin NUP116 [GenBank: NP_013762] [5]. These IDPs have been shown to adopt relaxed coil and collapsed coil structures, respectively, via both theory and experiment [11, 19].

Ten independent replicate simulations were performed for each of the four proteins using the GROMACS version 4.5.5 software package [9], Amber ff03w forcefield [1], and the TIP3P water model neuralized with 150 mMol NaCl. All bonds to hydrogens were constrained using the LINCS algorithm [8], and Bussi's stochastic thermostat [2] was used for temperature control, both with the default parameter settings. Each simulation started from a completely extended conformation that was minimized for 10000 steps of steepest descent. The temperature was set to 300K, but was then linearly increased to 600K for 25ns, then held at 600K for 50ns, then decreased to 300K for 25ns. The simulations then continued for an additional 250ns at 300K. Structures were sampled every 10ps to form a total of 35001 structures per simulation with an aggregate total of 42 microseconds of simulation.

2.5 Available Implementation

MLD estimation tools, polymer models and simulations are available as part of the Molecular Dynamics Spectral Clustering Toolkit (MDSCTK) [16, 17], an open source project which provides a several tools for analyzing MD simulations available at: https://github.com/jlphillipsphd/mdsctk/.

3 RESULTS

In the polymer results below, nearest neighbors between structures were computed using the Euclidean distance between the sin and cos values of the set of θ and ϕ angles along the polymer chain. Pointwise dimensionality estimates were computed for three ranges of scales: small k=[2, 3, 4, 6], medium k=[2, 3, 4, 6, 8, 16, 32, 64], and large k=[2, 3, 4, 6, 8, 16, 32, 64, 128, 256, 512, 1024]. Noise smoothing was also performed.

3.1 Semirigid Helix

The results in Fig. 1A demonstrate that any amount of noise increases the estimated dimensionality above that of a truly rigid helix which has a dimension of zero. A system of this length has $2l$-5=35 DoF and we do see this accurately estimated for large amounts of noise and small values of k (the number of nearest neighbors). We see two decreasing trends away from this value though. First, the estimates decrease as k increases across all noise levels. This indicates that, in this case, the ensemble does not sample the conformation space sufficiently for large values of k. This is confirmed by the results in Supplemental Material S1 [1] for the N=2000 case where the fall-off is even greater. The second trend is that the estimates also decrease with decreasing noise. Fluctuations with smaller magnitudes result in lower dimensionality estimates than those with larger magnitudes even if they occur over all DoFs. We therefore expect that estimates for MD simulations of folded proteins, which have been observed to have a noise level of roughly $\sigma_{\theta,\phi}$=0.1, will be systematically lower than their maximum number of DoF.

[1]Supplemental Material for all calculations are reported at: https://www.cs.mtsu.edu/~jphillips/papers/PhillipsColvinNewsam-ACMBCB-2018-SM.pdf.

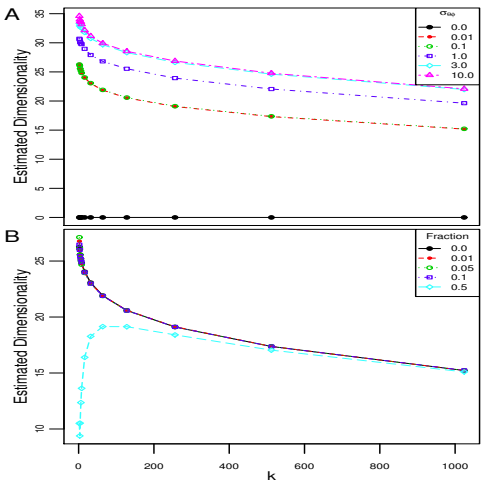

Figure 1: Semirigid helix model results. (A) Estimated dimensionality for a semirigid helical polymer for N=5000 structures of length 20 with various amounts of noise ($\sigma_{\theta,\phi}$) injected into the "folded" ensemble. k indicates the number of nearest neighbors used in the calculation. The expected dimension of a completely rigid model (no noise) is zero. The expected dimension of a model which is exercising all DoFs is 35. (B) Estimated dimensionality for the same polymer under a noise level of $\sigma_{\theta,\phi}$=0.10 but with noise smoothing by removing increasing amounts of the high frequency components.

Fig. 1B shows the results of smoothing the semirigid helix using different frequency cutoffs. This is for a noise level of $\sigma_{\theta,\phi}$=0.10. The smoothing only has an effect when a large fraction of the high frequencies are removed and for small values of k. In this case, the estimated dimension is lower as predicted. The effect of smoothing is reduced as k increases due to insufficient sampling.

The complete results for the semirigid helix are summarized in Supplemental Material S1. These results support the observations made from Fig. 1 above. Smoothing using an amplitude cutoff has less of an effect than a frequency cutoff.

3.2 Half-folded Helix

Fig. 2A shows that having the same amount of noise in both segments ($\sigma_{\theta_{unfolded}}$=0.0) produces the highest estimates. As the noise in the unfolded segment is increased, the estimates decrease. This makes sense because the larger fluctuations in this segment start to dominate the smaller fluctuations in the folded segment; i.e, the smaller fluctuations are truly noise and the algorithm is correcly finding that the system is exercising fewer DoFs. Half-folded polymer systems are expected to have lower dimensionality than either completely unfolded systems or completely folded systems which, due to the noise introduced into the system, appear to be exercising all DoFs. The results from the half-folded helix show that our framework is able to capture this. The results of the semirigid and half-folded helices thus show that the expected transitions of a

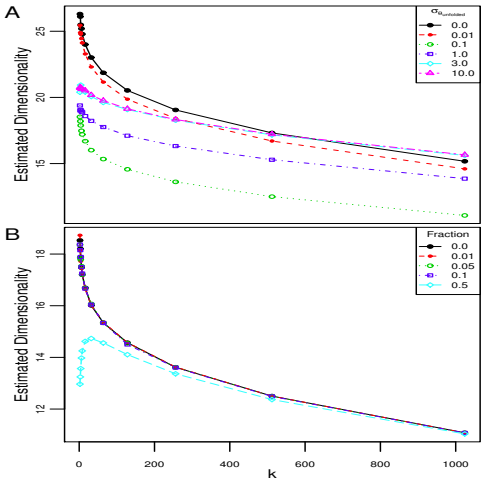

Figure 2: Half-folded helix model results. (A) Estimated dimensionality for a half-folded helical polymer for N=5000 structures of length 20 with various amounts of noise ($\sigma_{\theta_{unfolded}}$) injected into the unfolded segment. (B) Estimated dimensionality for the same polymer with a noise level of $\sigma_{\theta_{unfolded}}$=0.10 but with noise smoothing by removing increasing amounts of the high frequency components.

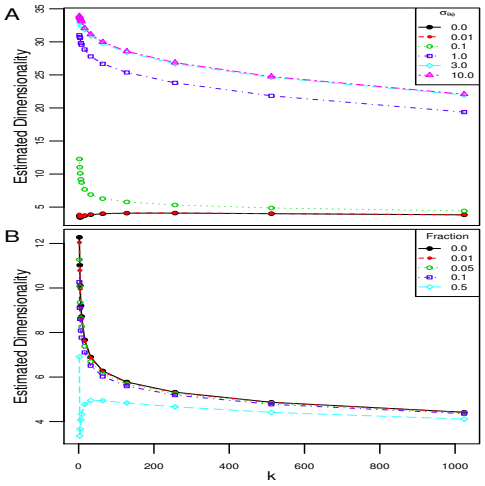

Figure 3: Correlated helix model results. (A) Estimated dimensionality for a correlated helical polymer for N=5000 structures of length 20 and 5 correlated dimensions with various amounts of noise ($\sigma_{\theta,\phi}$). The expected dimensionality of this system without noise is 5. (B) Estimated dimensionality for the same polymer under a noise level of $\sigma_{\theta,\phi}$=0.10 but with noise smoothing by removing increasing amounts of the high frequency components.

folding system are from unfolded states which have high dimensionality, to partially folded states which have lower dimensionality, and finally to the folded states which again have high dimensionality since the only motion is due to noise and involves all DoFs in the polymer. We note that the dimensionality estimates do start to increase again for the half-folded helix for large amounts of noise (high $\sigma_{\theta_{unfolded}}$) in the unfolded segment especially for small values of k.

These results also demonstrate the potential of our framework to distinguish between IDPs and folding proteins that have not yet started to fold. Since disordered proteins are not truly random coils, their dimensionality should be somewhat suppressed and their dynamics should be similar to those of the half-helix model most of the time. In contrast, the yet-to-fold NFPs should should have higher dimensionality.

Fig. 2B shows the results of smoothing the half-folded helix using different frequency cutoffs. Similar to the semirigid helix, the smoothing only has an effect when a large fraction of the high frequencies are removed and for small values of k.

The complete results for the half-folded helix are summarized in Supplemental Material S2. These results support the observations made from Fig. 2 above. Smoothing using an amplitude cutoff again has less of an effect than a frequency cutoff.

3.3 Correlated Helix

The results in Fig. 3A show the estimates are quite accurate for small amounts of noise. The dimensionality is slightly underestimated which is potentially due to the fact that the manifolds are not closed and estimates near the boundaries will therefore be lower. The estimates remain accurate even for modest amounts of noise, including the noise level observed in folded systems, $\sigma_{\theta,\phi}$=0.1 (although it is slightly overestimated at small values of k). For high levels of noise, the estimated dimensionality matches that of a semirigid helix of similar length, as expected.

Overall, the framework accurately estimates the intrinsic dimensionality of the models even at noise levels consistent with thermal motion in molecular systems. It detects the correlated motions that are often present in very extended protein chains as well as in short pieces of folding proteins during the folding process. The results for models with fewer correlated dimensions are similar (see Supplemental Material S3). These results, in combination with those of the half-folded helix above, demonstrate that the dimensionality estimator can distinguish between dynamics that are due to significant (large amplitude) motions and those that are due to noise, and can further detect when these motions are correlated.

Fig. 3B shows the results of smoothing the correlated helix using different frequency cutoffs. Similar to the previous models, the smoothing only has an effect when a large fraction of the high frequencies are removed. In this case, it does result in a more accurate estimate of the dimensionality especially for small to medium values of k. This is significant as it shows that smoothing can further improve the estimates for systems with intrinsic dimensionalities that are relatively small compared to the maximum DoFs but are overestimated due to modest levels of noise.

Complete results for the correlated helix are summarized in Supplemental Material S3. The results support the observations made in Fig. 3 above: smoothing using an amplitude cutoff has less effect than a frequency cutoff. Additionally, smoothing with a large fraction frequency cutoff (0.5) improved estimates. The effect was the most pronounced for small k=(2,3,4,6) (data not shown).

3.4 Molecular Dynamics Simulations

We now apply our dimensionality estimation framework to the protein simulations guided by several key insights from the polymer models: 1) Noise greatly increases the estimated dimensionality especially when its magnitude is of the order of other motions. This is particularly evident in systems that fluctuate around a fixed conformation. The highest estimates correspond to systems whose only motion is due to evenly distributed noise. 2) Dimensionality decreases when parts of the system undergo motions which have greater amplitude than others. Specifically, the larger motions dominate the noise and thus fewer DoFs are truly being exercised. 3) Correlated motion results in significantly reduced dimensionality. This remains true even for modest amounts of noise. 4) Increasing the scale (k) improves the accuracy of the estimates for trajectories (temporally related ensembles) especially when noise is present. 5) Noise smoothing improves the accuracy of the estimates especially for smaller scales.

In all the results below, nearest neighbors were computed within a simulation using the Euclidean distance between the sin and cos values of the Φ and Ψ angles of the protein conformations. Pointwise dimensionality estimates were computed for three ranges of scales: small k=[2, 3, 4, 6], medium k=[2, 3, 4, 6, 8, 16, 32, 64], and large k=[2, 3, 4, 6, 8, 16, 32, 64, 128, 256, 512, 1024]. Noise smoothing was also performed. Comparisons to a reference conformation (e.g. the native folded conformation) were performed for each frame in the trajectories using C_α root-mean-squared distance (RMSD).

3.4.1 Aggregate Estimates. Per-replicate dimensionality estimates were computed separately for the 100ns annealing phase and the 250ns production phase using the harmonic mean. Boxplots of the large scale estimates (k=[2, 3, 4, 6, 8, 16, 32, 64, 128, 256, 512, 1024]) for the ten replicates of each protein simulation are shown in Fig. 4. (Results for the small and medium scales can be found in Supplemental Material S4.) Since the proteins have different lengths, the bottom row shows the normalized results computed by dividing by the total DoFs (the number of Φ and Ψ angles): 30 for GB1; 38 for Trp-cage; and 48 for both Nsp1 and Nup116.

For to the production phase of the simulations (Fig. 4, right), we see that the normalized estimates are lower for the IDPs than the NFPs. This might have seemed contradictory without the insights from the polymer models since the IDPs are expected to be more disordered than the NFPs. However, the polymer studies showed that the estimated dimensionality would be high if a system remained in tightly-packed, frustrated, or possibly even folded structures such as might be the case for the NFPs. The IDPs, on the other hand, likely have partially formed structures and correlated motions.

3.4.2 Individual Replicate Estimates. We selected one replicate from each protein simulation to examine the pointwise dimensionality estimates. The replicates chosen for GB1 and Trp-cage were those that best approached the folded state of these proteins. The replicates chosen for Nsp1 and Nup116 were those that best matched the average conformation over all replicates for these proteins. For comparison, we computed each conformation's RMSD value to the folded structure for GB1 and Trp-cage, and to the average conformation for Nsp1 and Nup116. The dimensionality estimates and RMSD values were averaged over 0.5ns windows.

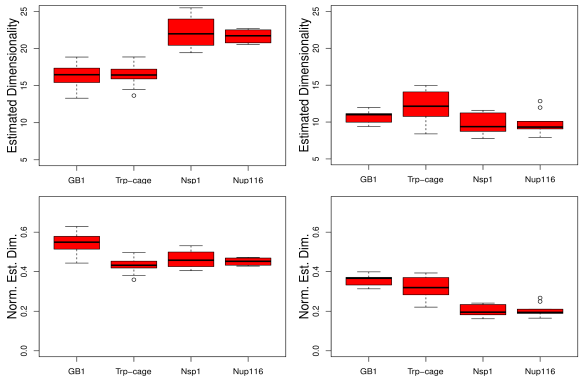

Figure 4: Distribution of dimensionality estimates (large k). Distributions of dimensionality estimation results over 10 replicates for (left) 100ns annealing (300K-600K-300K) and (right) 250ns production (300K) MD simulations of GB1, Trp-cage, Nsp1, and Nup116. Estimates are computed over a large scale (k=[2, 3, 4, 6, 8, 16, 32, 64, 128, 256, 512, 1024]). Results shown for (top) original and (bottom) normalized estimates.

The large scale dimensionality estimates for GB1, Trp-cage, Nsp1, and Nup116 are shown in Fig. 5 along with the RMSD values to the folded (NFPs) or average (IDPs) structure. The most striking feature of these plots is the consistency of the dimensionality estimates during the production phase even though there are clearly large, fast structural transitions that occur according to the RMSD values. GB1, for example, appears to be a very frustrated system that is constantly attempting to fold, but is restricted to suboptimal conformational states. As a result, the thermal fluctuations of the system act as noise, greatly increasing the dimensionality estimates. According to the RMSD values, the Trp-cage simulation comes very close to folding the protein, allowing thermal noise to produce a high estimated dimensionality due to the abundance of sampled structures. According to the RMSD values for Nsp1 and Nup116, these simulations undergoe regular, rather minor structural transitions without becoming frustrated at any particular location. The dimensionality estimates for these IDPs reflect this fact as well and remain relatively suppressed compared to the NFPs.

Results of applying smoothing are shown in part D of Fig. 5 for the proteins. Similar to the polymer model studies, smoothing has limited effect since estimates are computed for large values of k.

4 DISCUSSION

We introduced a polymer framework for examining dimensionality estimation algorithms for studying protein MD simulations. The key contribution is the development of several polymer models which exhibit well-defined dynamics of known dimensionality. Dimensionality estimation results from the polymer models indicated that (1) under/over estimation due to sampling/noise can be reliably predicted for some dynamical transitions, (2) relative ranking of dimensionaility estimates is still accurate in spite of inaccuracies in absolute intrisic dimensionality estimates, and (3) the two previous results can be leveraged to guide interpretation of results obtained from simulations of disordered and folding proteins. We

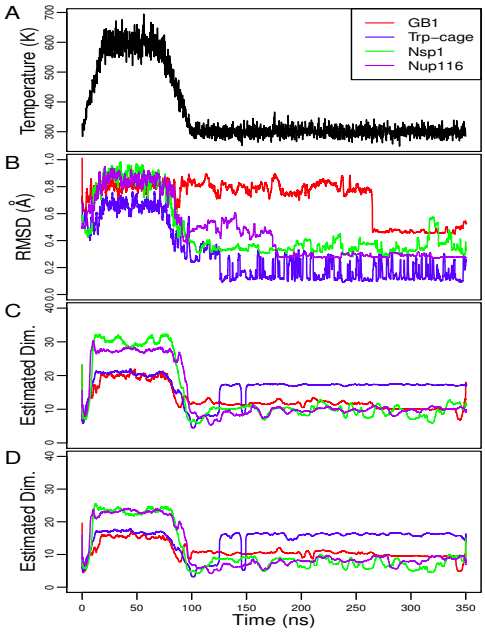

Figure 5: MD simulation dimensionality estimation results. Dimensionality estimation results for a representative GB1, Trp-cage, Nsp1, and Nup116 simulations. (A) Plot of temperature versus time, (B) RMSD from the folded structure versus time (GB1, Trp-cage) or RMSD from the average conformation (Nsp1, Nup116), (C) pointwise dimensionality estimates and (D) pointwise dimensionality estimates after noise smoothing with a fractional frequency cutoff of 0.5. Point estimates were obtained using k=[2, 3, 4, 6, 8, 16, 32, 64, 128, 256, 512, 1024].

conclude therefore that dimensionality estimation is a useful tool for differentiating between protein classes.

The dimensionality estimator was used to compare the dynamics of natively folded and intrinsically disordered proteins. While it was hypothesized that the folded state of proteins was of zero dimensionality, practical limitations (noise) prevent estimates of zero. Instead, the high-dimensionality of the noise allowed the folded or more frustrated states of proteins folding to be predicted to have very high dimensionality relative to less frustrated and more dynamic intrinsically disordered proteins. In particular, even though one of the natively folded proteins used in the study did not fold, the proteins adopted intermediate structures that were also very rigid and of high estimated dimensionality. Future work is needed to ascertain if this result is consistent for other folding proteins and the particular properties driving this effect in the simulations.

ACKNOWLEDGMENTS

This work was supported in part by NSF Award Number 0960480, by NIH Grant GM077520, and by the U.S. Department of Energy, Office of Science, Offices of Advanced Scientific Computing Research, and Biological & Environmental Research through the U.C. Merced Center for Computational Biology. JLP was also supported by an N. C. Metropolis Postdoctoral Fellowship through the Los Alamos National Laboratory Advanced Scientific Computing program and the Center for Nonlinear Studies.

REFERENCES

[1] Robert B Best and Jeetain Mittal. 2010. Protein simulations with an optimized water model: cooperative helix formation and temperature-Induced unfolded state collapse. *The Journal of Physical Chemistry B* 114, 46 (Nov. 2010), 14916–14923. https://doi.org/10.1021/jp108618d

[2] Giovanni Bussi, Davide Donadio, and Michele Parrinello. 2007. Canonical sampling through velocity rescaling. *Journal of Chemical Physics* 126, 1 (Jan. 2007), 014101. https://doi.org/10.1063/1.2408420

[3] F Camastra. 2003. Data dimensionality estimation methods: a survey. *Pattern Recognition* 36, 12 (Dec. 2003), 2945–2954. https://doi.org/10.1016/S0031-3203(03)00176-6

[4] J. A. Costa and A. O. Hero. 2004. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing* 52, 8 (Aug. 2004), 2210–2221. https://doi.org/10.1109/TSP.2004.831130

[5] A Goffeau, B G Barrell, H Bussey, R W Davis, B Dujon, H Feldmann, F Galibert, J D Hoheisel, C Jacq, M Johnston, E J Louis, H W Mewes, Y Murakami, P Philippsen, H Tettelin, and S G Oliver. 1995. Life with 6000 Genes. *Science* 274, November (1995), 546–67. https://doi.org/10.1126/science.274.5287.546

[6] Peter Grassberger and Itamar Procaccia. 1983. Characterization of strange attractors. *Physical Review Letters* 50, 5 (Jan. 1983), 346–349. https://doi.org/10.1103/PhysRevLett.50.346

[7] A M Gronenborn, D R Filpula, N Z Essig, A Achari, M Whitlow, P T Wingfield, and G M Clore. 1991. A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. *Science* 253, August (1991), 657–661. https://doi.org/10.1126/science.1871600

[8] Berk Hess. 2008. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation* 4, 1 (Jan. 2008), 116–122. https://doi.org/10.1021/ct700200b

[9] Berk Hess, Carsten Kutzner, David van der Spoel, and Erik Lindahl. 2008. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation* 4, 3 (March 2008), 435–447. https://doi.org/10.1021/ct700301q

[10] Balazs Kegl. 2003. Intrinsic dimension estimation using packing numbers. In *Advances in Neural Information Processing Systems 15*. MIT Press, Cambridge, MA.

[11] V. V. Krishnan, Edmond Y. Lau, Justin Yamada, Daniel P. Denning, Samir S. Patel, Michael E Colvin, and Michael F Rexach. 2008. Intramolecular cohesion of coils mediated by phenylalanine–glycine motifs in the natively unfolded domain of a nucleoporin. *PLoS Computational Biology* 4, 8 (Jan. 2008), e1000145. https://doi.org/10.1371/journal.pcbi.1000145

[12] Oliver F. Lange and Helmut Grubmüller. 2006. Generalized correlation for biomolecular dynamics. *Proteins: Structure, Function and Genetics* 62, 4 (2006), 1053–1061. https://doi.org/10.1002/prot.20784

[13] Elizaveta Levina and Peter J Bickel. 2005. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou (Eds.). MIT Press, Cambridge, MA.

[14] David J C MacKay and Zoubin Ghahramani. 2005. Comments on "Maximum likelihood estimation of intrinsic dimension" by E . Levina and P . Bickel (2004). , 5 pages. http://www.inference.phy.cam.ac.uk/mackay/dimension/

[15] Jonathan W Neidigh, R Matthew Fesinmeyer, and Niels H Andersen. 2002. Designing a 20-residue protein. *Nature Structural Biology* 9, 6 (June 2002), 425–30. https://doi.org/10.1038/nsb798

[16] Joshua L. Phillips, Michael E. Colvin, Edmond Y. Lau, and Shawn Newsam. 2008. Analyzing dynamical simulations of intrinsically disordered proteins using spectral clustering. In *Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine Workshops*. IEEE, Philadelphia, PA, 17–24. https://doi.org/10.1109/BIBMW.2008.4686204

[17] Joshua L. Phillips, Michael E. Colvin, and Shawn Newsam. 2011. Validating clustering of molecular dynamics simulations using polymer models. *BMC Bioinformatics* 12, 1 (Jan. 2011), 445. https://doi.org/10.1186/1471-2105-12-445

[18] J B Tenenbaum, V de Silva, and J C Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 5500 (Dec. 2000), 2319–2323. https://doi.org/10.1126/science.290.5500.2319

[19] Justin Yamada, Joshua L. Phillips, Samir Patel, Gabriel Goldfien, Alison Calestagne-Morelli, Hans Huang, Ryan Reza, Justin Acheson, Viswanathan V. Krishnan, Shawn Newsam, Ajay Gopinathan, Edmond Y. Lau, Michael E. Colvin, Vladimir N. Uversky, and Michael F Rexach. 2010. A bimodal distribution of two distinct categories of intrinsically disordered structures with separate functions in FG nucleoporins. *Molecular & Cellular Proteomics* 9, 10 (Oct. 2010), 2205–2224. https://doi.org/10.1074/mcp.M000035-MCP201