Higher-order Representations of Protein Structure Space

Kevin Molloy[1], M. Jennifer Van[1], Daniel Barbara[1,*], and Amarda Shehu[1,2,3*]
[1]*Dept. of Computer Science,* [2]*Dept. of Bioengineering,* [3]*School of Systems Biology*
George Mason University, Fairfax, VA, 22030, United States
[*dbarbara, amarda*]*@gmu.edu*
∗*Corresponding Author*

***Abstract*—Fragment-based representations of protein structure have recently been proposed to identify remote homologs with reasonable accuracy. The representations have also been shown through PCA to elucidate low-dimensional maps of protein structure space. In this work we conduct further analysis of these representations, showing that the low-dimensional maps preserve functional co-localization. Moreover, we employ Latent Dirichlet Allocation to investigate a new, topic-based representation. We show through various techniques adapted from text mining that the topics have unique signatures over structural classes and allow a complementary yet informative organization of protein structure space.**

***Keywords*-protein structure space; molecular fragments; Latent Dirichlet Allocation.**

I. Introduction

Functional information is unavailable for millions of protein sequences. Many methods have been proposed to infer this information from homologous proteins of known function. Close homologs can accurately be identified by sequence comparison. Identification of remote homologs, however, where structure is better preserved than sequence, requires comparing protein structures. Structure comparison methods are often computationally expensive and not feasible for employment on large databases of protein structures, such as the Protein Data Bank (PDB), for the identification of remote homologs of a given protein.

$Fragbag$ is a recently proposed structure representation to rapidly identify homologs and embed protein structure space in a few dimensions useful for visualization. In this work we investigate a new representation inspired from research in text mining. We apply the Latent Dirichlet Allocation (LDA) model, which has been used in text mining for identifying topics in documents, to redefine proteins in a topic space. We show here that the topic-based representation is comparable to $fragbag$ in identifying structural neighbors with reasonable accuracy. We also show that the representation offers an alternative organization of structure space, where topics represent signatures that correlate with SCOP classifications.

II. Methods

We first summarize the $fragbag$ representation, followed by a description of LDA and metrics used in this study.

$Fragbag$ is a bag-of-words vector representation for a protein [1]. A library of K molecular fragments, each of length f, defines the "vocabulary" in this bag-of-words model. Each fragment is composed of f C_α atoms. The libraries used in this study are supplied by Kolodny [2]. To transform a protein into the $fragbag$ representation, its structure is first reduced to a trace of its C_α atoms. For each interval $[i, (i + f - 1)]$ of C_α atoms, the best-fitting fragment in terms of lRMSD is selected from the library. Let us assume that this fragment is stored at position $1 < k \leq K$ in the library. The protein structure is thus represented as a vector of K dimensions, where each position k stores the number of times the fragment stored at position k in the library has been selected as the best fit. Various distance measurements can then be defined over these representations to compare protein structures. The cosine distance metric has been shown most effective [1].

The representation we propose and investigate here is a topic-based representation obtained through LDA. LDA is a generative model that locates latent clusters (topics) within a corpus of documents [3]. In this setting, the corpus is an ensemble of proteins represented by $fragbag$ vectors. LDA allows for multi-membership and provides a discrete distribution of topics for each protein. In addition to showing that this representation captures structural neighbors as well as $fragbag$, we conduct various analyses of the information captured in this representation.

We employ a text mining technique from [6] to measure the information content of each topic. The distribution of fragments across the ensemble is used as a baseline. The information gain between this baseline and the distribution of fragments within each LDA topic is then computed using the symmetric Kullback-Leibler (KL) distance function. Topics that are close to the baseline distribution (lower symmetric KL distances) have low information. This technique is used to tune the number of topics. In text mining, the semantic meaning of each topic is determined by analyzing the most frequently occurring words. In this setting, identifying the meaning of a topic by analyzing individual fragments would be difficult. We identify topic "signatures" associated within existing structural classifications (SCOP). This identification is performed by showing the distribution of five SCOP classes (a-e) across the topic-space.

978-1-4673-1321-6/12/$31.00

III. Results

The Steyvers/Griffiths toolkit is employed for the LDA analysis [4], using $\alpha = 50/T$ (T = number of topics) and $\beta = 0.5$. We use a CATH 2.4 dataset of $2,930$ domains and a dataset of $31,155$ domains extracted from SCOP 1.71. Both are transformed into $fragbag$ representations using the Kolodny library of 400 fragments of length 11 [2].

Results presented in [1] are reproduced (data not shown); namely, structural neighbors over the CATH dataset are accurately identified when using the cosine distance function over the $fragbag$ representation. In addition, PCA of the $fragbag$ vectors associated with the SCOP dataset shows that the top two PCs capture 99% of the total variance. Projecting the SCOP dataset on the two top PCs illustrates that the SCOP classes at various levels of the SCOP hierarchy are well separated (results not shown), suggesting that the $fragbag$ representation retains sufficient structural information to elucidate functional co-localization in the protein structure space.

We choose to highlight here two representative results from our application of LDA on the $fragbag$ vectors of SCOP domains. First, we vary the number of topics T obtained through LDA and analyze information gain as described in Methods. Figure 1 (top panel) shows the results for various settings. Lower KL distances indicate "junk" topics, as they show higher similarity to the baseline distribution. Lower values of T have larger KL distances from the baseline, justifying the use of $T = 10$ in the rest of our analysis. The topic-space vectors of dimensionality 10 are then used to identify structural neighbors using the same protocol as in [1]. Results are comparable to those using the $fragbag$ representation [1], [5] (data not shown). Finally, we analyze this topic space of 10 dimensions for semantic meaning. Figure 1 (bottom panel) shows the distribution of α and β SCOP classes across LDA topics. The results show that not all topics are well distributed across the SCOP classes. Instead, most of them have unique signatures correlating with structural classes.

IV. Conclusion

This work shows that text mining techniques can be employed and adapted to investigate novel informative representations of protein structure, laying the foundation for continuing studies to gain further insight into the organization of protein structure space.

Acknowledgments

We thank R. Kolodny for providing us with fragment libraries and datasets for direct comparisons. This work is supported in part by NSF CCF No. 1016995 and NSF IIS CAREER Award No. 1144106 to AS and a Mason OSCAR undergraduate fellowship to JV.

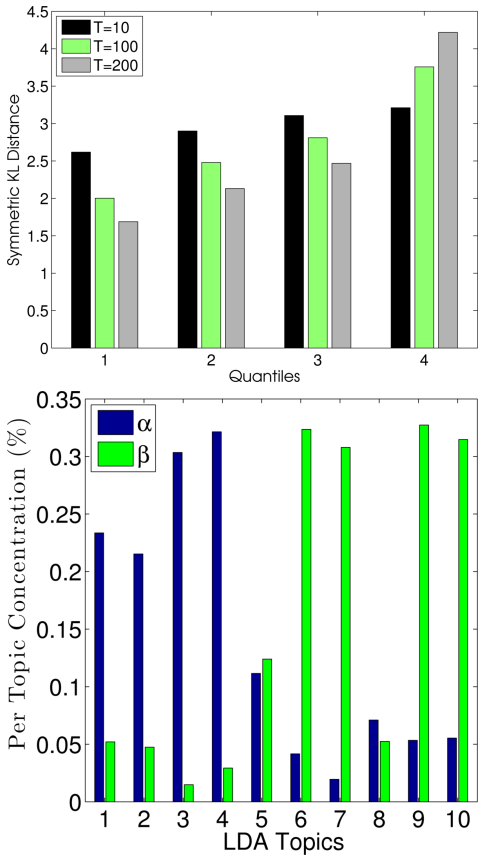

Figure 1: (Top) Information gain is shown while varying the number of topics. (Bottom) Distribution of α and β SCOP classes is shown across topic space.

References

[1] I. Budowski-Tal, , Y. Nov, and R. Kolodny, "Fragbag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately," *Proc. Natl. Acad. Sci. USA*, vol. 107, pp. 3481–3486, 2010.

[2] R. Kolodny, P. Koehl, L. Guibas, and M. Levitt, "Small libraries of protein fragments model native protein structures accurately," *J. Mol. Biol.*, vol. 323, pp. 297–307, 2002.

[3] D. M. Blei, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[4] M. Steyvers and T. Griffiths, "Probabilistic topic models," in *Latent Semantic Analysis: A Road to Meaning.*, T. Landauer, D. Mcnamara, S. Dennis, and . Kintsch, Eds. Laurence Erlbaum, 2006. [Online]. Available: http://cocosci.berkeley.edu/tom/papers/SteyversGriffiths.pdf

[5] S. Shivashankar, S. Srivathsan, B. Ravindran, and A. V. Tendulkar, "Multi-view methods for protein structure comparison using Latent Dirichlet Allocation," *Bioinformatics*, vol. 27, pp. i61–i68, 2011.

[6] L. AlSumait, P. Wang, C. Domeniconi, and D. Barbarà, "Embedding semantics in LDA topic models," in *Text Mining: Application and Theory*, M. W. Berry and J. Kogan, Eds. Fairfax, VA: John & Wiley, 2010, ch. 10.