Learning Protein Embeddings using Variational Autoencoders

Alexandra-Ioana Albu[1], Gabriela Czibula[1]
[1]Department of Computer Science, Babeş-Bolyai University, Cluj-Napoca, Romania
alexandra.albu@stud.ubbcluj.ro, gabis@cs.ubbcluj.ro

Abstract

Modeling protein structures dynamics represent a key aspect in protein structures analysis. We present an approach using Variational Autoencoders for learning low dimensional representations for protein conformations which can capture the structure of the input data. Additionally, we investigate the usefulness of the learned protein embeddings for predicting the evolution of protein trajectories and we propose directions for future improvements.

Keywords: protein dynamics; variational autoencoders; latent representations
Domain: computer science
Section: New (2020) thesis proposals

Motivation

Proteins are molecules that are vital for the biological processes taking place in living organisms. Protein structures are known to be dynamic, as they are able to transition between certain conformations. The ability to understand protein motions is important due to their role in determining the protein's function. In a recent work [AC20], we have introduced an approach using Variational Autoencoders (VAEs) for modeling protein conformational transitions. In this abstract, we extend our approach by presenting preliminary results on the usefulness of the embeddings learned by the VAE for predicting how the proteins' conformations change over time.

Methodology of Research

We represented protein conformations using the structural alphabet (SA) proposed by Pandini et al. [PFK10]. A sequence of SA symbols can be represented numerically either by using a one-hot encoding or by replacing each symbol with the three angles characterizing that conformational state. Two representations have been used in our experiments for the conformations: the representation based on SA letters and the representation combining angles and RSA values, which quantify the degree to which the residue is exposed in the protein structure [TCB+18].
We trained a VAE on the protein conformations and we assessed how well the VAE's latent space models the input data using three measures. The *IntraPS* similarity measure [TCB+18] was used to quantify the average similarity between successive conformations within a protein. In addition, we defined a measure, *Dist*, which assesses the absolute difference between the IntraPS of the original protein and the IntraPS of the encoded protein. To evaluate the capability of VAEs to preserve individual similarities between conformations, we also assessed the correlation between similarities in the input space and between their representations in the latent space. Further, we evaluated the learned embeddings on a downstream prediction task.

Results and Comparison with State-of-the-art

We evaluated our approach on conformational transitions belonging to three proteins. In our experiments, the representation based on SA yielded slightly higher *IntraPS* values than the Angles+RSA representation, but was outperformed by the combined representation when considering the distance between the original protein and its embedding. Moreover, the similarities between successive conformations were better preserved in the Angles+RSA representation, as shown by the

correlation coefficients. We compared our model with a similar approach using a sparse denoising autoencoder [TCB+18]. Table 1 presents a comparison using the Angles+RSA representation for the two proteins used in both studies. The results highlighted that the VAE's latent space preserves more faithfully than autoencoders the similarity with the original proteins, yielding, however, lower IntraPS values.

Representation	Protein	Model	*IntraPS*	*Dist*
Angles+RSA	1JT8	Our VAE	0.9959	**0.0046**
		AE [TCB+18]	**0.9985**	0.0072
	1P1L	Our VAE	0.9863	**0.0290**
		AE [TCB+18]	**0.9962**	0.0389

Table 1. Comparison of our model with the sparse denoising autoencoder (AE) from [TCB+18].

In our preliminary results obtained for the prediction task the Angles+RSA representation yielded smaller error values on the test set compared to the SA representation.

Conclusions

We have conducted a study on using VAEs for modeling protein dynamics, with the end goal of predicting protein conformational transitions. Two representations for the protein conformations were analysed, with comparable performances. As future work, we plan to extend our experimental analysis on a larger dataset and investigate alternative representations for the protein data. Additionally, we aim to compare the performance of our prediction model with similar methods, such as recurrent models in the input space or time-lagged autoencoders [WN18] which map conformations to future time steps.

Acknowledgements

The authors thank lecturer Alessandro Pandini from Brunel University for providing the data sets used in the experiments.

References

[AC20] Alexandra-Ioana Albu and Gabriela Czibula. "Analysing protein dynamics using machine learning based generative models." *2020 IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. IEEE, 2020.

[PFK10] Alessandro Pandini, Arianna Fornili, and Jens Kleinjung. "Structural alphabets derived from attractors in conformational space." In: *BMC bioinformatics* 11.1 (2010): 97.

[TCB+18] Mihai Teletin, Gabriela Czibula, Maria Iuliana Bocicor, Silvana Albert, Alessandro Pandini (2018, October). Deep Autoencoders for Additional Insight into Protein Dynamics. In *International Conference on Artificial Neural Networks* (pp. 79-89). Springer, Cham.

[WN18] Christoph Wehmeyer and Frank Noé. "Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics." *The Journal of chemical physics* 148.24 (2018): 241703.