COMPARING MACHINE LEARNING FRAMEWORKS FOR PROTEIN FOLDING RESEARCH

Bogdan Czejdo, Fayetteville State University, USA
Sambit Bhattacharya, Fayetteville State University, USA

ABSTRACT

In this paper, we describe, compare, and design frameworks for protein folding research with the goal of optimizing an iteration heavy protein scoring efforts. Traditional frameworks use a handful of popular machine learning (ML) algorithms, such as Support Vector Machines (SVM), Random Forest and Neural Networks that are trained on a feature engineered input. New deep learning frameworks have shown to do well even on raw data. The frameworks are generally expandable i.e. they allow researchers to modify the parameters for the included algorithms, or replace the parts of the algorithms with their own. In this paper, we describe a new, multi-level machine learning framework for protein model scoring that is constructed by integrating two frameworks. The deep learning framework is used to automatically discover categories of variable-length amino acid sequences using Recurrent Neural Networks (RNN). The traditional machine learning framework is used to perform model scoring, separately for each cluster of protein models corresponding to a given category. Our multi-level framework incorporates appropriate data manipulation functions to aid researchers with correct preparation of the input.

INTRODUCTION

Due to recent advances in computing power and availability of high amounts of labeled and unlabeled data, applications of machine learning (ML) have emerged as the dominant form of data-driven, applied artificial intelligence (AI). This transformation, which has brought AI out of academic labs, has reshaped it into an area where progress is driven by applications that are of interest to society. One of the greatest challenges of AI has been to develop approaches to analyze sequence data. Recent advances in artificial neural networks with deep architectures, popularly known as deep learning, has shown progressively better solutions for many problems. However the previous approaches of machine learning are still applied, and in many real world problems a combination of deep and previous (or so-called shallow methods) frequently works best. In this paper we concentrate on investigating the use of machine learning in protein folding research. Many protein sequences are known, only a small percentage of those sequences have known 3-D structures that have been experimentally determined. Experimental methods to identify new protein 3-D structures are very expensive and slow. Therefore, including the use of ML techniques seems to be practically unavoidable for progress in this area.

Proteins make up living things and can act as tiny machines for biological processes. They are initially created as a chain of amino acids (protein sequence), which folds into a three dimensional shape called the "native" structure. The various interactions of the amino acids in the chain produce the compact shape (native structure), which determines the protein's biological functionality. The structure is essential in the biological function of the protein, and many diseases have been linked with prions, or misfolded proteins (Alberts et al., 2002).

Many protein sequences are known, however only a small percentage of those sequences have known native structures, which have been experimentally determined. Almost all of known structures have been deposited in the Protein Data Bank, and are available as a learning set for computational experiments (Mirzaei et al., 2016). Since the experimental methods to identify new protein structures are very expensive and slow, an important goal is to accelerate this process by developing computational methods that can quickly predict the native structure of proteins from their sequence of amino acids. The native structure of a protein corresponds to the global minimum of a very complex energy function whose local minima increase exponentially with the number of amino acids in the sequence. There is no known global optimization method that can solve this problem and therefore direct generation of the protein target structure is practically impossible.

Currently the only practical approach is to use statistical and physical/chemical based techniques to generate thousands of approximate protein structures (called protein models), and then use a scoring functions to select the best models among them. The CASP experiments (Mirzaei et al., 2016) have shown that as methods get better and better at sampling, the proper selection is a major factor that limits the success of protein structure prediction (Mirzaei et al., 2016).

Different types of approaches have been used to score protein models. Some initial experiments used clustering based methods (Alberts et al., 2002). These methods were later replaced by energy functions, derived from physical principles (Zhou et al., 2011) or statistical analysis. All these methods do not perform consistently well (Zhou et al., 2011). Yet another, newer approach is to use machine learning ML algorithms such as neural networks (Faraggi et al., 2014), Support Vector Machines (SVM) (Mirzaei et al., 2016), and others that use calculated features to estimate protein model quality. It seems the ML approach is the most rapidly developing and holds the greatest potential for novel and interesting discoveries.

ML is a significant area within AI and is undergoing dramatic changes. These rapid changes are leading to better understand existing approaches (such as neural networks) and allow us to apply them in a more effective way. ML techniques have been used in protein scoring for several years (Mirzaei et al., 2016), but new algorithms and new available hardware should result in significant increase of the efficiency of ML experimentation work. Significantly improved speed of processing on GPUs and High Performance Computers (HPCs) has made possible many new experiments in protein model scoring. In this paper, we describe a new, multi-level machine learning framework for protein model scoring that utilizes both GPUs and HPCs. It is constructed by integrating two frameworks. The deep learning framework is used to automatically discover categories of variable-length amino acid sequences using Recurrent Neural Networks (RNN). The traditional machine learning framework is used to perform model scoring, separately for each cluster of protein models corresponding to a given category. We also incorporate a robust data preparation component for data sets that are commonly present in the protein folding research.

DATA PREPARATION FOR MACHINE LEARNING FRAMEWORKS

Almost all of known protein structures have been deposited in the Protein Data Bank (Mirzaei et al., 2016). The CASP competitions (Khoury G. et al., 2014) produce invaluable data that include hundreds of new protein models for novel experimentally determined protein structures. Another important source of processed data is WeFold (Keasar C et al., 2017), an organization for international collaborative efforts for protein structure prediction.

The datasets can be computed for all protein models based on their various features. Let us discuss the most important structures and processes for extracting information from the protein model datasets. The datasets are typically stored in a hierarchical directory as is shown in Figure 1. The root directory contains subdirectories containing data for each CASP. Each CASP subdirectory contains several directories corresponding to specific target proteins. Inside of each of those directories there are the files that describe around 80-120 best models generated for a specified target protein. More precisely, some target proteins can be logically divided into domains what is indicated by a letter “D” in the model name.

The main challenge related to data preparation is effectively applying a variety of available algorithms (functions) to compute components of the feature vector for the models and integrate these components into alternative feature vectors. We refer to these algorithms (functions) as Feature Computation Functions. Unfortunately Feature Computation Functions have usually incompatible input e.g. requiring a file, group of files or directory as an argument, and incompatible output. Therefore, for an effective feature computations a wrapping protocol needs to be developed and implemented to properly invoke each Feature Computation Function as shown in Figure 2. The first module Processing Tree is relatively simple, and is based on processing of a universal tree (more precisely directed acyclic graph). It is responsible for traversing a tree and selecting the nodes for further processing. The Function Wrapper module is a key module that invokes Feature Computation Functions according to a specially developed wrapping protocol.

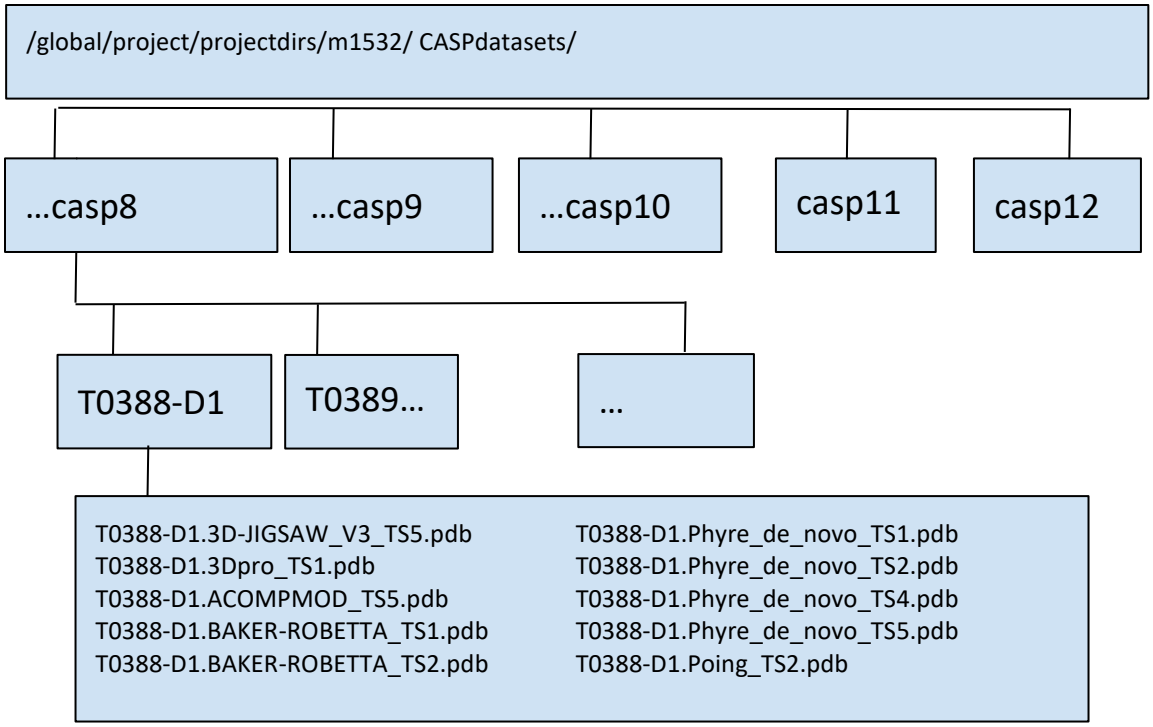

Figure 1. Hierarchical Directory of Protein Models Data.

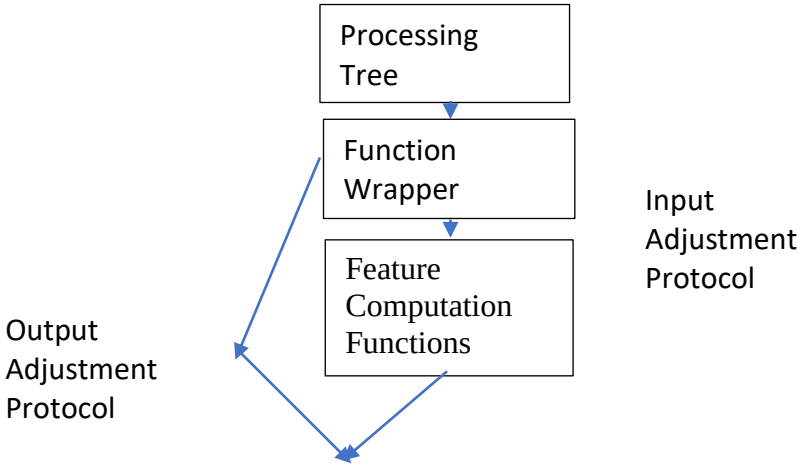

Figure 2. Processing Protein Databank Files.

As a result of such processing we can create four datasets for ML. The first two sets contain data for both models and their native structures:

- Feature Vector set containing algorithmically processed information of each model and each target protein.
- Raw 3-D data set containing direct information about each model and each target protein 3-D structure.

The remaining two sets contain data for the native structures only:

- Sequence Feature Vector set containing algorithmically processed information of amino acid chain treated as a sequence of data e.g. identifying percentage of different secondary structures in the chain.
- Amino acid chain data sets containing direct specification of amino acid chain as a sequence of data.

MACHINE LEARNING FRAMEWORKS FOR PROTEIN FOLDING

In general, the protein folding problem that can be expressed as the main research question "How can we predict the protein three dimensional structures from the amino acid chains?" Practically, it is replaced by two distinct research questions: "How can we generate a set of models from a sequence of amino acids that are representative of low energy portions of the protein folding manifold?", and "How can we select from this set of models the ones which are the closest to the protein native structure using more accurate but higher cost calculations?". Our efforts are concentrated on the latter, i.e. the process of choosing the model closest to the protein native structure. Choosing the best model is often referred to as protein model scoring. For effective protein model scoring we cannot use the same physics and chemistry laws as for the model generation. We can, however, use the data obtained after analyzing protein models created through many CASP years as described in the previous section. The data that we use for analysis include experimentally discovered protein target structures and corresponding protein models that are computationally generated.

As it is shown in Figure 3 and 4, by using machine learning algorithms, we can "learn" from previously generated models by comparing them to the native structures. They can be assigned "scores" based on GDT_TS function that roughly corresponds to geometrical differences between models and the target structure.

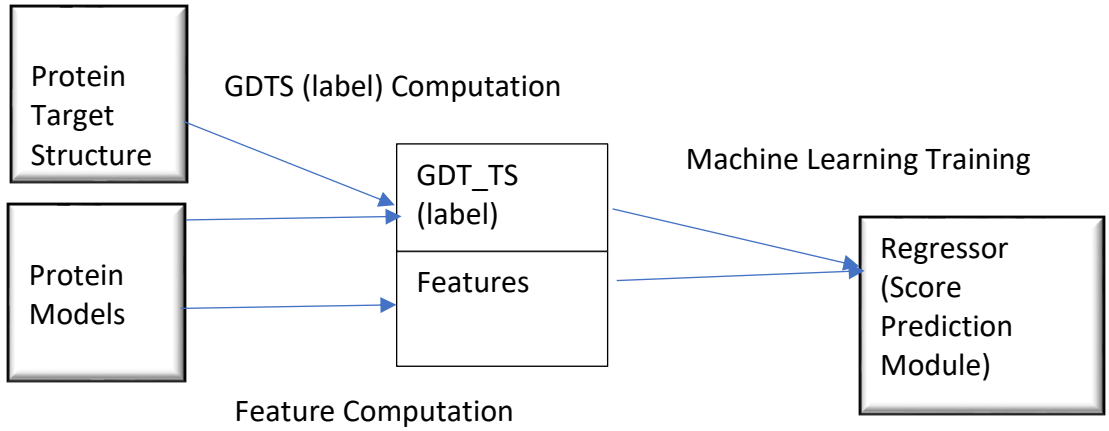

Figure 3. Training phase of machine learning.

At this moment our framework supports workflows for three fundamental ML algorithms: Support Vector Machines (SVM), Random Forest and Neural Networks. This research framework allows researchers to proceed with experiments quicker by avoiding unnecessary impediments like hardware and software configurations.

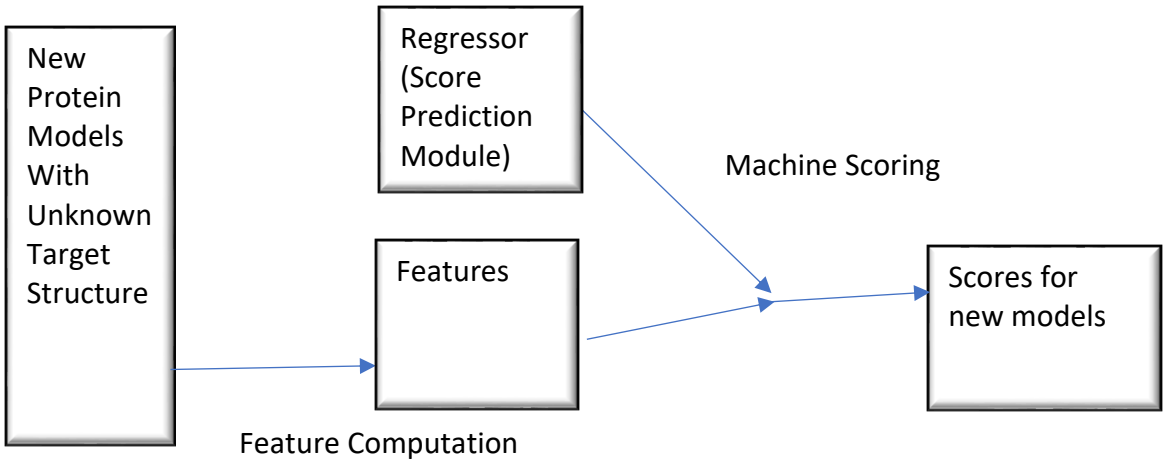

Figure 4. Scoring phase of machine learning.

There are many possible approaches to improve protein model scoring by using deep machine learning algorithms (Pallastri et al., (2002). Most of the approaches cannot provide a complete solution based on raw data but new deep learning frameworks can be used for feature engineering to automatically detect important attributes of the protein 3-D models. Deep learning can also be used to directly analyze amino acid chains. Recurrent Neural Networks (RNN), a form of deep learning, have been recently and successfully used in learning patterns in data which are characterized by values that form a sequence (Zaremba et al., 2014). The sequential structure can ultimately carry advanced knowledge such as information about the folded structure and function of an amino acid chain. It is an open area of research to use RNN for unsupervised learning such as clustering but some promising results are already available (Zaremba et al., 2014). The RNN module consists of one or more sets of "bi--directional" recurrent layers with long--short term memory (LSTM) artificial neurons. These artificial neurons process an input sequence one element at time, where an element in the case of a protein may be a single amino acid or a secondary structure. The sequences are scanned in both directions since doing so utilizes information once going from one end to the other end and vice versa. The network learns patterns by adjusting millions of parameters. These parameters control the degree to which a network remembers patterns that are discovered over longer parts of the sequences versus smaller parts of the sequences.

MULTI-LEVEL MACHINE LEARNING FRAMEWORK WITH DEEP LEARNING

In order to accelerate protein model scoring research we propose an approach based on two-level framework that integrates deep learning with traditional machine learning. On the first level, the deep learning is used to cluster the protein amino acid chains based on their sequences and secondary structures as shown in Figure 5. The motivation for this clustering is that various amino acid chains contain various secondary structure e.g. beta-sheets, that result in significant geometrical differences between models belonging to different clusters.

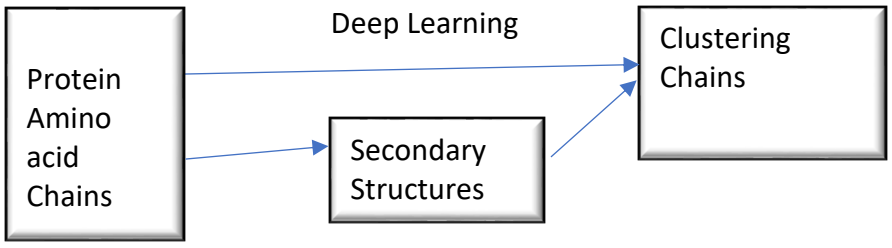

Figure 5. First level of our framework: Protein Amino Acid Chains Clustering

The second level of our framework is responsible for traditional machine learning, but the "learning" is done separately for each cluster. As a result we have a separate Regressor (Score Prediction Module) for each cluster. The score prediction requires placing the given amino acid chain in the proper cluster first, and then using the associated Regressor to compute the final model score.

CONCLUSIONS

In constructing the research framework for protein scoring, we took into consideration the heterogeneity of datasets and newest solutions in machine learning. Our framework allows us to experiment with both deep and shallow (traditional) machine learning. The project can be extended in several directions. One direction is to improve the framework architecture and usability by adding the DBMS system support. Another direction is to improve techniques for parallelization of data preparation and machine learning algorithms.

ACKNOWLEDGEMENTS

We would like to acknowledge support of the summer Workforce Development & Education program at Department of Energy (DOE) Lawrence Berkeley National Laboratory (LBNL) to initiate this project and continuing cooperation with LBNL to improve it. More specifically we would like to acknowledge the selfless efforts of Dr. Silvia Crivelli in guiding us in understanding the complexities of protein model scoring. We would like also to acknowledge contribution to this project by student interns: Casey L. Lorenzen and Catherine L. Spooner, and the support of Mary Ann Leung, director of Sustainable Horizons Institute.

REFERENCES

Alberts B. et al. (2002), The Shape and Structure of Proteins, Molecular Biology of the Cell; Fourth Edition. New York and London: Garland Science. ISBN 0-8153-3218-1.

Faraggi E. and Kloczkowski A. (2014), A global machine learning based scoring function for protein structure prediction, Proteins: Structure, Function, and Bioinformatics, vol. 82, no. 5, pp. 752-759.

Keasar C et al., (2017). "An Assessment of WeFold: A Framework of International Collaborative Pipelines for Protein Structure Prediction." submitted to Scientific Reports.

Khoury G. et al., (2014), WeFold: A Coopetition for Protein Structure Prediction. Proteins: Structure, Function, and Bioinformatics; 82(9): 1850-1868, doi: 10.1002/prot.24538.

Mirzaei S, T. Sidi, C. Keasar, and S.Crivelli (2016), Purely Structural Protein Scoring Functions Using Support Vector Machine and Ensemble Learning, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume: PP Issue: 99. DOI: 10.1109/TCBB.2016.2602269.

Pollastri, Gianluca, et al., (2002), Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles." Proteins: Structure, Function, and Bioinformatics 47.2: 228-235.

Zaremba W. et al. (2014), Recurrent Neural Network Regularization, arXiv preprint arXiv:1409.2329.

Zhou H., and Skolnick J. (2011), GOAP: a Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction, Biophysical Journal, vol. 101, no. 8, pp. 2043-2052.