Using Machine Learning in Accuracy Assessment of Knowledge-Based Energy and Frequency Base Likelihood in Protein Structures

Katerina Serafimova[1], Iliyan Mihaylov[1], Dimitar Vassilev[1(✉)], Irena Avdjieva[1], Piotr Zielenkiewicz[2], and Szymon Kaczanowski[2(✉)]

[1] FMI, Sofia University "St. Kliment Ohridski", 5 James Bourchier Street, 1164 Sofia, Bulgaria
dimitar.vassilev@fmi.uni-sofia.bg
[2] IBB PAN, Warsaw, Poland
szymon@ibb.pan.pl

Abstract. Many aspects of the study of protein folding and dynamics have been affected by the accumulation of data about native protein structures and recent advances in machine learning. Computational methods for predicting protein structures from their sequences are now heavily based on machine learning tools and on approaches that extract knowledge and rules from data using probabilistic models. Many of these methods use scoring functions to determine which structure best fits a native protein sequence. Using computational approaches, we obtained two scoring functions: knowledge-based energy and likelihood of base frequency, and we compared their accuracy in measuring the sequence structure fit. We compared the machine learning models' accuracy of predictions for knowledge-based energy and likelihood values to validate our results, showing that likelihood is a more accurate scoring function than knowledge-based energy.

Keywords: Knowledge-based energy · Statistical potential · Likelihood · Cross-validation · Machine learning · Protein structure prediction

1 Introduction

Proteins are built of one or more linear chains of amino acid residues, which are protein sequences that fold into three-dimensional structures. Correct folding leads to a native structure, and knowledge of the native protein structure is essential for understanding the protein function. A growing amount of structural data in databases such as the Protein Data Bank (PDB) [1] has led to the development of computational approaches for protein structure prediction. However, these approaches are often time-consuming and costly or have low accuracy, so there is a need for effective and accurate computational approaches to protein structure prediction.

Most of the methods for structure prediction use scoring functions to determine which structure best fits a native protein sequence. The native structure generally has a lower free

V. V. Krzhizhanovskaya et al. (Eds.): ICCS 2020, LNCS 12139, pp. 572–584, 2020.
https://doi.org/10.1007/978-3-030-50420-5_43

energy than the other possible structures under the native conditions [2], which means that an accurate free energy function can be applied in the prediction and assessment of protein structures [3], for example, as a scoring function in measuring sequence structure fit. However, calculating the free energy of protein folding (or unfolding) using all-atom coordinates is impractical because it is computationally demanding, so only a small fraction of the available conformational space can be explored in this way. Knowledge-based (KB) approaches that extract knowledge and rules from data are therefore used in the assessment of an ensemble of structural models produced by computational methods to find the correct structure that fits a given sequence.

KB free energy (also known as statistical potential or pseudo-energy potential) is widely used for those purposes. Statistical potential is derived using a mathematical approach, according to which the statistical preferences of interactions between different molecules can be described. However, protein folding is a cooperative process with many driving forces, which means that a residue in a given position has an impact on other residues and, ultimately, on the structure in which the sequence will fold. A way to describe cooperation is by using a likelihood function.

In this study, we used a one-dimensional (1D) representation of the protein structure based on the buried or exposed state of the residues to compare the accuracy of the KB energy (E) and the likelihood of base frequency (L), which are essentially scoring functions and can be used in protein structure prediction.

Machine learning (ML)–based approaches used in the accuracy assessment of sequence-structure fit in proteins provide a large set of models that can contribute to the process by enriching the quantification of different parameters. The ML models can also be efficiently used to validate results related to the accuracy of the functions by assessing the sequence structure fit. The large diversity of models bears some problems related to how and which model is better to choose for a particular case, which can be overcome by different cross-validation approaches.

To validate our results showing that likelihood is a more accurate measure of the sequence-structure fit than KB energy, we used ML models. The application of such ML models to predict likelihood and energy values accordingly provides criteria for assessing the accuracy and predictability of these two approaches.

2 Problem Description

Proteins interact strongly with surrounding solvents, and the exposure of amino acids to solvents is a sensitive parameter that can be used to model energetic features on the protein–solvent boundary [4]. The folding process of soluble proteins also decreases the surface area in contact with the solvent; this is related to the secondary structures of proteins. Accurate knowledge of residue accessibility would thus aid in the prediction of protein structures [5].

The protein residues in a structure are exposed to the solvent to different extents. We applied KB approaches to describe different types of residue preferences for being in a buried or exposed state in the protein structure.

In this work, the model of the protein structure is one-dimensional (1D) and uses only the solvent accessibility of every residue. For simplicity, the solvent accessibility is

categorized into two states: buried (0) and exposed (1). There are 20 types of amino acid residues and each of them can be categorized as buried or exposed in a given position in the protein sequence, so the number of residue classes is 40.

The solvent accessible surface area (SASA) of all the proteins in the two sets is calculated, and the threshold of the SASA per residue is selected to classify a residue as buried or exposed. Then, a buried/exposed pattern is assigned to every sequence to construct an **object** that describes the protein using the amino acid (AA) **sequence** and one structural property – a **pattern** of the buried or exposed status of the residues.

The sequence-structure pair objects based on the solvent accessibility were used for optimization of their KB energy or likelihood, accordingly.

The concept of pseudo-energy was introduced to biology by the seminal paper of Tanaka and Scheraga [6]. They assumed that residues behave like molecules interacting in gas, and they used the observed frequencies of the contacts between different types of residues in known X-ray structures. Using these data, they calculated the "free energies" (ΔG°) of the contact between different types of amino acids using a formula exported from statistical chemistry:

$$\Delta G^{\circ} = -RTX_{ij}X_iX_j \tag{1}$$

where X_{ij} is the frequency of the observed contacts between the residues of type i and the residues of type j, X_iX_j represents a multiplication of these frequencies (statistical expectation of the contact between residue i and residue j), R is the gas constant, and T is temperature.

In this study, using parameters obtained from a set of 200 native protein structures, we calculated the KB energy of proteins, as seen in Eq. 2:

$$kbE(protein) = \sum_{1}^{j} E_{i[0\,or\,1]} \tag{2}$$

where i is the type of residue according to the position of the protein sequence, j is the length of the protein, and $E_{i[0\,or\,1]}$ is the KB energy of the i-th residue, which can be buried [0] or exposed [1].

Likelihood is also widely used in biology, for example, in the case of phylogenetics. In this paper, we applied the following formula:

$$\mathrm{L} = \log\left(\frac{P_1^{n_{1[0]}} \times P_{2[0]}^{n_{2[0]}} \ldots P_{20[0]}^{n_{20[0]}} \times P_{1[1]}^{n_{1[1]}} \times P_{2[1]}^{n_{2[1]}} \ldots P_{20[1]}^{n_{20[1]}} \left((n_{1[0]} + \ldots n_{20[1]})!\right)}{\left(n_{1[0]}!\right)\left(n_{2[0]}!\right)\left(n_{3[0]}!\right)\ldots\left(n_{20[i]}!\right)}\right) \tag{3}$$

where $P_{i[0\,or\,1]}$ is the observed frequency of the residues in the buried/exposed state in the entire database, $n_{i[0\,or\,1]}$ is the number of residues of a given type in a given protein, and $(n_{1[0]} + \ldots n_{20[1]})$ is equal to the length of the protein.

The object design is highly simplified, and the reduction of 3D structural information to 1D lowers the possible accuracy of the scoring. Therefore, it is possible to optimize a native sequence-structure pair, for example, by changing the pattern to better fit the native sequence according to the selected criteria. The resulting pattern will be different from the native one (because the criteria is imperfect), and the accuracy of the criteria can be assessed by calculating the identity of the resulting pattern to the native pattern.

Using the two criteria – KB energy and likelihood, separately, in two optimization experiments, we were able to conclude that the likelihood is less erroneous than the KB energy as a criterion of the sequence-structure fit.

The concept for this study is to test the accuracy of the two criteria for measuring the sequence-structure fit by using a ML approach.

Machine learning can be applied for multiple purposes in protein folding and structure prediction: measuring the sequence-structure fit, designing energy functions, or analyzing protein simulation data.

In this work, we apply ML models to evaluate the accuracy of the two properties, KB energy and likelihood, that can be used as scoring functions. We have already assessed that likelihood is a more accurate measure of the sequence-structure fit. The goal of this study is to validate that using ML models. We want to check whether the model predictions of the likelihood values will be more accurate than those of KB energy. That will show that the likelihood provides the possibility of better use of structural information in prediction than the KB energy.

A common method to estimate the quality of model predictions is to use cross-validation and calculate the average prediction performance across test samples. Here, we use cross-validation in the context of predictive modeling. This is one of the most widely used data resampling methods to assess the generalizability of a predictive model and to prevent overfitting. To build the final model for the prediction of real future cases, the learning function (or learning algorithm) f is usually applied to the entire learning set. The purpose of cross-validation in the model-building phase is to estimate the performance of the final model on new data.

Cross-validation divides the training data into several disjointed cohorts of approximately equal size. Each cohort is used in turn as testing data, while the remaining cohorts are used as training data. The prediction model built on the training data is then applied to predict the class labels of the testing data. This process is repeated until all cohorts have been used as the testing data once, and then the prediction accuracies of all the blinded tests are combined to produce an overall performance estimate.

3 Related Work

Different bioinformatics and statistical approaches can be used to predict the 3D structure of a protein from its amino acid sequence. Many of these approaches can be viewed as sequence-structure fitness problems. In evaluating a hypothetical structure, such as the fitness of a sequence for a structure, one must be able to distinguish between correct and incorrect structures (to identify the structural states that have a high probability of being observed in given environmental conditions). Success or failure depends crucially on the underlying description of structural states and on the evaluation scheme of sequence-structure fitness [7].

Based on the thermodynamic hypothesis [2], computational studies of proteins, including structure prediction, folding simulation, and protein design, depend on the use of a potential function to calculate the effective energy of the molecule. In protein structure prediction, the potential function is used either to guide the conformational search process or to select a structure from a set of possible sampled candidate structures [8].

Two fundamentally different approaches exist to obtain a potential energy function [9]. The first is an inductive approach [4], a mathematical model that describes the system is assumed without previous knowledge about the physical principles. The resulting potential is directly extrapolated to more complex molecules by assuming that a common behavior will exist in both cases [9]. The second approach is deductive (or KB). In order to obtain an accurate description of the potential energy function, experimental data from large macro-molecular-solvent systems should be used [9]. The parameters of the potential functions are extracted from a database of known protein structures [4]. Because of the deductive nature f this approach, which incorporates many physical interactions (electrostatic, van der Walls, cation interactions) and the extracted potentials do not necessarily reflect true energies, it is often referred to as the "knowledge-based", "empirical", or "statistical" effective potential function or scoring function [8].

Current studies are focused on improving knowledge-based potentials used for: protein structure predictions, [10–12] RNA structure predictions [13, 14], and rational drug design [15].

More complex KB approaches use the advances of ML for protein structure prediction and sequence-structure fit assessment. Theoretically, the implementation of ML can be defined as both supervised learning, where the data includes additional attributes that are expected to be predicted, and unsupervised learning, where the training data consists of a set of input vectors without any corresponding target values. The supervised learning set of models consists of two groups: classification and regression. The large background of standard supervised ML methods provides reasonable results, but the advent of methods based on deep residual networks has shown more promising results in some cases.

Different ML methods have been applied as a tool for protein structure prediction based on KB potentials [16, 17]. It is expected that ML forcefields may soon replace forcefields in protein simulations [18].

Some alternative ML methods for structure prediction, such as probabilistic neural networks and deep learning end-to-end differentiable networks, have shown wider applicability [19].

There have also been attempts to apply likelihood functions as a tool for protein structure predictions using the multiple sequence alignment of related proteins as input data [20, 21]. Multiple sequence alignment shows which residues are evolutionarily related. A likelihood function indicates the probability of contact between different residues.

A significant problem in using ML models for sequence-structure fit is how to validate the results. Very often, this process is based on cross-validation of the outcomes of the applied models. Cross-validation is primarily used in applied ML to assess the potential and the accuracy of certain ML models for certain data. This means that it is possible to use a limited sample to estimate how the model will perform in general when used to make predictions on data not used during training. The cross-validation model can be used to estimate any quantitative measure that is appropriate for the data and the model. The use of cross-validation in sequence to structure fit evaluation models is discussed in [22].

4 Data Description

For the purposes of the study, we used two datasets: 1) a set of 200 protein structures for the calculation of the parameters used in KB energy and likelihood determination, and 2) a set of 45 protein structures for the optimization experiments. The first dataset was extracted from a selection of nonhomologous proteins [23]. The second dataset for testing purposes was obtained from the non-redundant PDB chain set of proteins with a sequence-similarity cut-off BLAST p-value of 10e−7, which is the most non-redundant of the given. The testing set contains 45.*pdb* files that meet the following criteria: having 0% unknown, incomplete, or missing residues or residues with incomplete side-chain; having only one chain (subunit) in the PDB entry; and not containing any heterogens (except for water). The models were determined by X-ray crystallography.

The sequence of every one of the 245 selected proteins is extracted from the *.pdb* file using *Biopython* [24, 25].

The solvent accessible surface area (SASA) of the residues is a geometric measure of exposure to the solvent. SASA is typically calculated by methods involving the in- silico rolling of a spherical probe, which approximates a water molecule, around a full-atom protein model [26]. The SASA of the protein molecule is the surface area traced by the center of the probe. A classical approximation commonly used to calculate SASA is the Lee and Richards (L&R) approximation [27], where the surface is approximated by the outline of a set of slices [28].

In this work, the Python module of FreeSASA, an open source C library [28], is used to calculate the solvent-accessible areas. SASA values for every residue in the protein are obtained by a high precision L&R calculation (probe radius: 1.400; slices: 100) using the default on FreeSASA ProtOr radii [29].

The relative solvent accessibility (RSA) of a residue indicates its degree of burial in a structure. The RSA calculation is important because different amino acids are of different sizes, so they also differ in area. To disregard these differences, the relative exposure (RSA) is calculated by normalizing the surface area of the residue in the structure by the surface area of the same type of residue in some reference state (e.g. the residue X in an extended tripeptide, such as Gly-X-Gly). RSA values are calculated by dividing the absolute SASA by the maximum solvent accessibility (maxSASA). Values for maxSASA based on ProtOr radii were extracted from the default reference values used in the FreeSASA classifier.

The calculated RSA was further divided into two states, using an exposure threshold of 0.1 (10%). Namely, a residue is considered buried (marked as 0) when RSA $\leq$ 0.1 and exposed (marked as 1) when RSA $>$ 0.1. Each residue in a chain is then assigned to class 0 if it has an RSA lower than or equal to 0.1 and to class 1 if the RSA is higher.

5 Suggested Methodology

For the purposes of this study, we have developed an ML-driven approach for accuracy assessment of knowledge-based energy (E) and frequency base likelihood (L) for protein structure prediction. Both approaches are based on statistics of the buried/exposed properties of residues.

5.1 Data Preparation for ML

The sequence and pattern of every one of the 245 protein objects are transformed into numerical values that can be used as parameters in ML models.

To every type of amino-acid residue a) in the sequence, a corresponding number from b) is assigned:

a) 'A','R','N','D','C','E','Q','G','H','I','L','K','M','F','P','S','T','W','Y','V'
b) 10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29

The sequence of these numbers is specific for every protein and is used as parameter X1 in the ML models. The numbers representing all the residues in one protein sequence are then added to a value that is later used as parameter (p1) in the ML models.

In the patterns, every buried state (before represented as 0) is assigned the coefficient 0.2, and every exposed state (previously represented as 1) is assigned the coefficient 0.5. The sequence of these coefficients is then summed to obtain the second parameter (p2). The purpose of the coefficients is to represent the structural component as a distance.

The values of the KB energy (E) and the likelihood of base frequency (L) are used in the ML and are calculated for every one of the 245 proteins.

Outliers with values greater than five times the mean distance are removed from the study. After this filtering, a dataset generated from 244 native protein structures is used.

The 244 samples of the parameter values X1 and X2 are then individually normalized using the standard normalizer of the *scikit-learn* Python library [30].

To predict the KB energy and likelihood values, we used three supervised regression ML models. The chosen models are from python *scikit-learn* package: 1) Lasso – *linear*_model (alpha = 0.1), which is a regression analysis method that performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical model it produces; 2) Nearest Neighbors Regression (NNR) – *kNeighborsRegressor* (n_neighbors = 5, algorithm ='kd_tree'), which is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression. We use a regression approach where the output is the property value for the object. This value is the average of the values of k nearest neighbors. 3) Decision tree regression (DTR) – *DecisionTreeRegressor* (max_depth = k). Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

For every one of the models, a k-fold cross-validation is used to split the set into k smaller sets for better estimation.

As input, we used 244 samples of:

- Two normalized parameters, X1 and X2, that were obtained from data about the protein sequence and the protein structure, respectively.
- The actual values for KB energy or likelihood, obtained from formulas (2) and (3).

For k = 3, 5 and 7:

- The original sample is randomly partitioned into k equal sized subsamples.
- Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 subsamples are used as training data.
- The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data.
- The k train score results are then averaged to produce a single estimation (with a standard deviation).
- The predicted values are plotted against the original data.

The purpose of the suggested methodology is to show the difference in accuracy of prediction performance of the applied ML models based on the values of the two scoring functions: KB energy and likelihood.

6 Results and Discussion

The methodology of this study provides results based on the three above-described ML models and produces scores for comparing the accuracy of these models.

After the data set is normalized, we apply three supervised regression ML models: lasso regression, nearest neighbor regression, and decision tree regression. We test the cross-validation splitting strategy of k = 3, k = 5, and k = 7 folds to compare the models in terms of their accuracy of predicting the scores of KB energy and likelihood.

The graphs in Figs. 1, 2 and 3 show the relatedness of the actual to the predicted values of every particular model used for KB energy and for likelihood respectively with cross-validation (cv) k = 5.

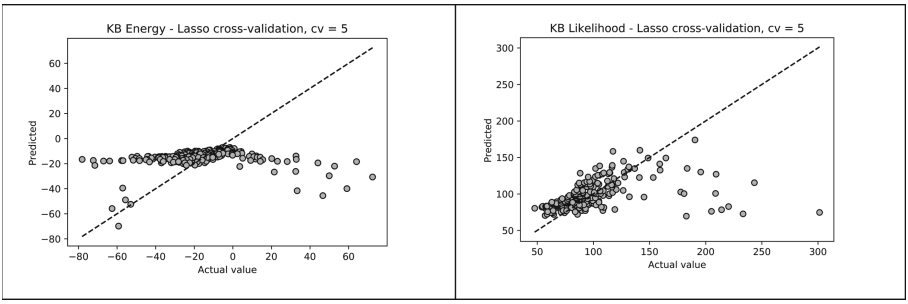

Fig. 1. Lasso cross-validation, k = 5

For the KB energy, lasso has worse predictive results than for likelihood, for which the results are distributed around the fit regression line with very few outliers from the greater actual value of likelihood.

The results of NNR are similar, with the KB energy estimates more dispersed than the values for likelihood.

The DTR produces somewhat similar results for the prediction of KB energy and likelihood values.

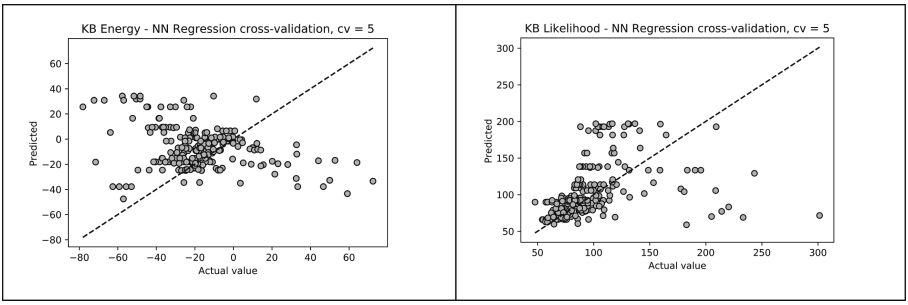

Fig. 2. Nearest neighbor regression cross-validation, k = 5

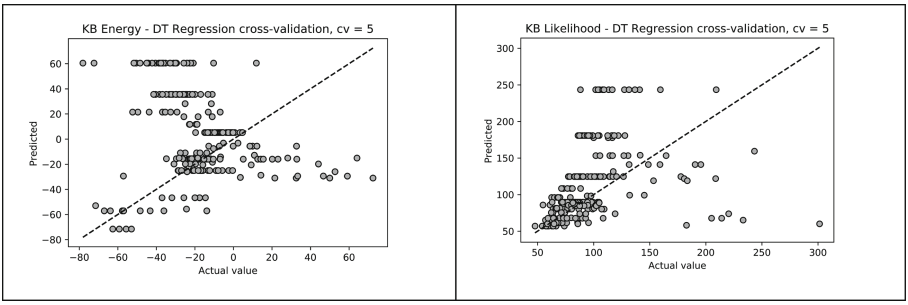

Fig. 3. Decision tree regression cross-validation, k = 5

These results are evidence that the likelihood prediction using is better than the KB energy prediction. These results confirm the analytical computational approach of optimization's finding that likelihood is superior to KB energy as a scoring function.

As a consequence of modelling the relatedness of the predicted to the actual values using the three ML approaches, we can refer to the coefficient of determination resulting from the training scores both for the KB energy and likelihood. The coefficient of determination shows the accuracy of the applied models.

All values given in Figs. 4, 5 and 6 are based on the average values for a particular splitting strategy with different k-fold numbers: 3, 5, and 7.

In Figs. 4, 5 and 6, the greater accuracy of the likelihood prediction ML models over the KB energy prediction ML models is obvious.

In Fig. 4, the NNR (Nearest Neighbor Regression) model with k = 3 produces higher mean values and smaller errors than the other two regression models. Lasso is less accurate model, while DTR (Decision Tree Regression) has an intermediate position.

The accuracy of the applied ML models changed when the splitting training set strategy amounts to five (Fig. 5.) In this case, the NNR and DTR models have very close average mean values and distributions of error values. The lasso regression model is obviously inferior to both NNR and DTR.

The increased accuracy of the DTR model for both KB energy and likelihood predictions is seen in Fig. 6. We can thus infer that, with a higher number of training data

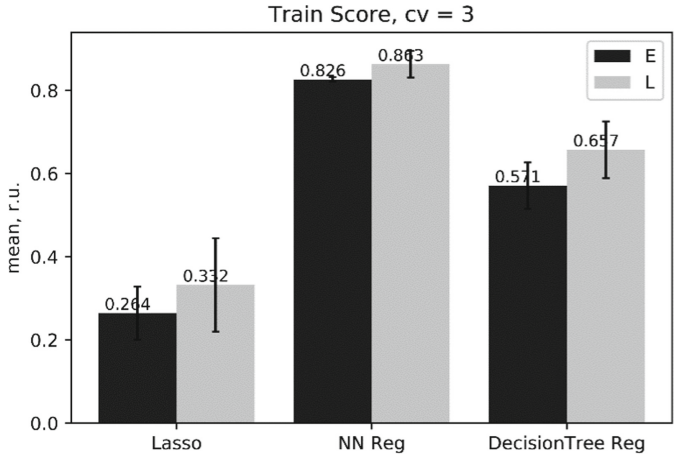

Fig. 4. Mean training scores for the three models with k-fold number cv = 3

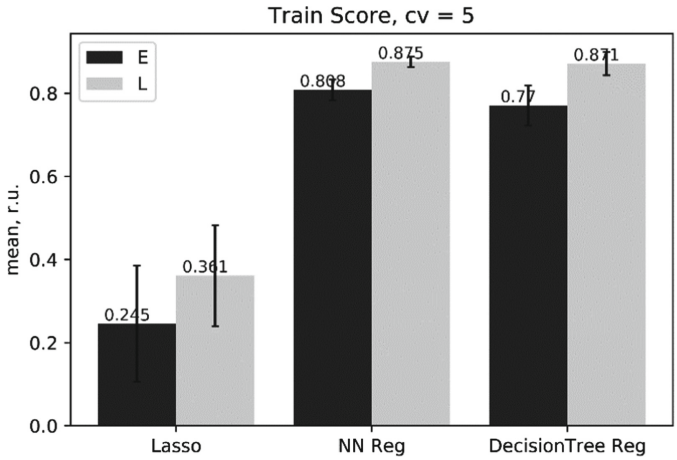

Fig. 5. Mean training scores for the three models with k-fold number cv = 5

sets, we can improve the accuracy of the DTR model. The most important finding is the overall superior accuracy of the likelihood prediction approach.

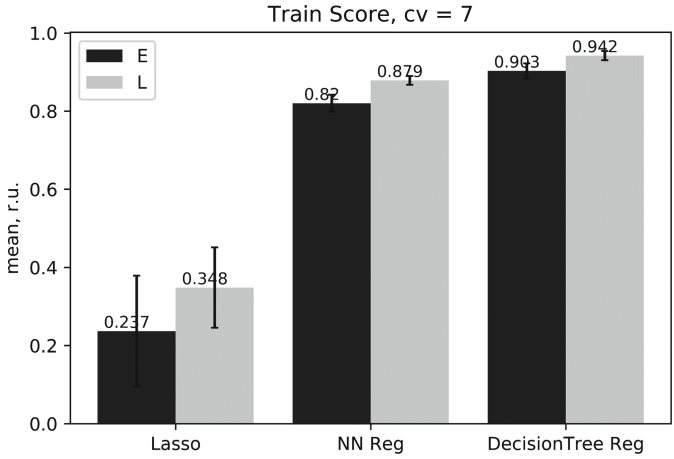

Fig. 6. Mean training scores for the three models with k-fold number cv = 7

7 Conclusions

In bioinformatics, statistical properties can be estimated using likelihood or likelihood function. Recently, ML has been applied as a tool to enhance this classical approach. We showed that ML is more efficient in predicting likelihood parameters than KB energy.

In our study, we developed a ML-driven approach for accuracy assessment of KB energy and frequency base likelihood for protein structure prediction. Both approaches are based on statistics of the buried or exposed properties of residues.

We proposed an approach for model comparison based on cross-validation of the estimated performance.

The ML models were applied to confirm the superiority of the frequency base likelihood approach over the KB based energy approach for assessing sequence-structure fit in proteins.

This study demonstrates the potential of protein structure prediction methods based on ML and indicates that combining ML with frequency base likelihood is more efficient than using KB energy functions.

Acknowledgments. The research presented in this paper was partly supported by the National Scientific Program "Information and Communication Technologies for a Single Digital Market in Science, Education and Security (ICTinSES)", financed by the Bulgarian Ministry of Education and Science. The work planning and conducting the discussed computational experiments was partly supported by the Sofia University SRF within the "A computational approach using knowledge-based energy and entropy for assessment of protein structure prediction" project. The authors are very grateful to the Institute of Biochemistry and Biophysics of the Polish Academy of Sciences for funding the publication of this study.

References

1. Berman, H.M., et al.: The protein data bank. Nucleic Acids Res. **28**(1), 235–242 (2000)

2. Anfinsen, C.B., Haber, E., Sela, M., White, F.H.: The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc. Natl. Acad. Sci. **47**(9), 1309–1314 (1961)
3. Shen, M.-Y., Sali, A.: Statistical potential for assessment and prediction of protein structures. Protein Sci. Publ. Protein Soc. **15**(11), 2507–2524 (2006)
4. Sippl, J.M.: Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. J. Comput. Aid. Mol. Des. **7**(4), 473–501 (1993). https://doi.org/10.1007/BF02337562
5. Lins, L., Thomas, A., Brasseur, R.: Analysis of accessible surface of residues in proteins. Protein Sci. **12**, 1406–1417 (2003)
6. Tanaka, S., Scheraga, H.A.: Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules **9**(6), 945–950 (1976)
7. Ouzounis, C., Sander, C., Scharf, M., Schneider, R.: Prediction of protein structure by evaluation of sequence-structure fitness: aligning sequences to contact profiles derived from three-dimensional structures. J. Mol. Biol. **232**(3), 805–825 (1993)
8. Li, X., Liang, J.: Knowledge-based energy functions for computational studies of proteins. In: Xu, Y., Xu, D., Liang, J. (eds.) Computational Methods for Protein Structure Prediction and Modeling: Volume 1: Basic Characterization, pp. 71–123. Springer, New York (2007). https://doi.org/10.1007/978-0-387-68372-0_3
9. Melo, F., Feytmans, E.: Scoring functions for protein structure prediction. Comput. Struct. Biol. **3**, 61–88 (2008)
10. Ciemny, M.P., Badaczewska-Dawid, A.E., Pikuzinska, M., Kolinski, A., Kmiecik, S.: Modeling of disordered protein structures using monte carlo simulations and knowledge-based statistical force fields. Int. J. Mol. Sci. **20**(3), 606 (2019)
11. López-Blanco, J.R., Chacón, P.: KORP: knowledge-based 6D potential for fast protein and loop modeling. Bioinformatics **35**(17), 3013–3019 (2019)
12. Yu, Z., Yao, Y., Deng, H., Yi, M.: ANDIS: an atomic angle- and distance-dependent statistical potential for protein structure quality assessment. BMC Bioinformatics **20**(1), 299 (2019). https://doi.org/10.1186/s12859-019-2898-y
13. Capriotti, E., Norambuena, T., Marti-Renom, M.A., Melo, F.: All-atom knowledge-based potential for RNA structure prediction and assessment. Bioinformatics **27**(8), 1086–1093 (2011)
14. Zhang, T., Hu, G., Yang, Y., Wang, J., Zhou, Y.: All-atom knowledge-based potential for rna structure discrimination based on the distance-scaled finite ideal-gas reference state. J. Comput. Biol. (2019)
15. Chen, P., et al.: DLIGAND2: an improved knowledge-based energy function for protein–ligand interactions using the distance-scaled, finite, ideal-gas reference state. J. Cheminform. **11**(1), 52 (2019). https://doi.org/10.1186/s13321-019-0373-4
16. Pei, J., Zheng, Z., Merz, K.M.: Random forest refinement of the KECSA2 knowledge-based scoring function for protein decoy detection. J. Chem. Inf. Model. **59**(5), 1919–1929 (2019)
17. Xu, J.: Distance-based protein folding powered by deep learning. Proc. Natl. Acad. Sci. **116**(34), 16856–16865 (2019)
18. Noé, F., De Fabritiis, G., Clementi, C.: Machine learning for protein folding and dynamics. Curr. Opin. Struct. Biol. **60**, 77–84 (2020)
19. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning with Applications in R. STS, vol. 103. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-7138-7
20. Bywater, R.P.: Prediction of protein structural features from sequence data based on Shannon entropy and Kolmogorov complexity. PLoS ONE **10**(4), e0119306 (2015)

21. Aurell, E.: The maximum entropy fallacy redux? PLoS Comput. Biol. **12**(5), e1004777 (2016)
22. Rashid, S., Saraswathi, S., Kloczkowski, A., Sundaram, S., Kolinski, A.: Protein secondary structure prediction using a small training set (compact model) combined with a complex-valued neural network approach. BMC Bioinform. **17**(1), 1471–2105 (2016). https://doi.org/10.1186/s12859-016-1209-0
23. Zhang, Y., Skolnick, J.: TM-align: a protein structure alignment algorithm based on TM-score. Nucleic Acids Res. **33**(7), 2302–2309 (2005)
24. Hamelryck, T., Manderick, B.: PDB parser and structure class implemented in Python. Bioinformatics **19**, 2308–2310 (2003)
25. Cock, P.J.A., et al.: Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics **25**(11), 1422–1423 (2009)
26. Durham, E., Dorr, B., Woetzel, N., Staritzbichler, R., Meiler, J.: Solvent accessible surface area approximations for rapid and accurate protein structure prediction. J. Mol. Model. **15**(9), 1093–1108 (2009). https://doi.org/10.1007/s00894-009-0454-9
27. Lee, B., Richards, F.M.: The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. **55**, 379–400 (1971)
28. Mitternacht, S.: FreeSASA: An open source C library for solvent accessible surface area calculations. F1000Research (2016)
29. Tsai, J., Taylor, R., Chothia, C., Gerstein, M.: The packing density in proteins: standard radii and volumes. J. Mol. Biol. **290**(1), 253–266 (1999)
30. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2012)