International Journal of Information Technology and Management

Vol. VIII, Issue No. XI, February-2015, ISSN 2249-4510

COMPARATIVE STUDY OF MACHINE LEARNING MODELS IN PROTEIN STRUCTURE PREDICTION

AN INTERNATIONALLY INDEXED PEER REVIEWED & REFEREED JOURNAL

www.ignited.in

International Journal of Information Technology and Management
Vol. VIII, Issue No. XI, February-2015, ISSN 2249-4510

Comparative Study of Machine Learning Models in Protein Structure Prediction

Akshay Pandey[1]* Dr. M. K. Sharma[2]

[1]B. Tech. (CSE) R. D. Engineering College, Ghaziabad

[2]Associate Professor, Amrapali Institute, Haldwani

Abstract – Machine learning is a subfield of computer science that incorporates the investigation of frameworks that can gain from information, as opposed to take after just unequivocally modified directions. Probably the most well-known procedures utilized for machine learning are Support Vector Machine, Artificial Neural Networks, K Nearest Neighbor and Decision Tree. Machine learning methods are generally utilized procedures in bioinformatics to take care of various kinds of issues. Protein structure expectation is one of the issues that can be understood utilizing machine learning. The particles which are critical in our cells are Proteins. They are basically associated with all phone capacities. Proteins are arranged on the premise of the event of moderated amino corrosive examples which is the element extraction technique. In the post-genomic time Protein work expectation is a critical issue. Progressions in the trial science have empowered the creation of huge measure of protein-protein communication information. Subsequently, to practically clarify proteins has been widely considered utilizing protein-protein association information. Whenever comment and connection data is deficient in the systems a large portion of the current system based methodologies don't function admirably. In this paper an endeavor has been made to survey diverse papers on proteins capacities and structures that are anticipated utilizing the different machine learning strategies.

Keywords—Protein Structure Prediction, Machine learning, RCGA.

----------♦----------

INTRODUCTION

Knowledge of the three-dimensional (3D) structure of a protein is crucial to understand its function. However, the rapid growth of the number of protein sequences has far outpaced the experimental determination of their structures. Thus, there is a growing need for a computational approach to the problem of protein structure prediction. The prediction of secondary structure, the local structure commonly defined by hydrogen bond patterns and local geometry, is a critical first step towards this end and, therefore, it has attracted a great amount of interest over the past 50 years. With respect to their secondary structure, amino acid residues in protein chains are usually assigned into three main classes, namely helix, extended and coil/loop.

Owing to significant efforts in genome sequencing over nearly three decades (McPherson et al. 2001; Venter et al. 2001), gene sequences from many organisms have been deduced. Over 100 million nucleotide sequences from over 300 thousand different organisms have been deposited in the major DNA databases, DDBJ/ EMBL/GenBank (Benson et al. 2003; Miyazaki et al. 2003; Kulikova et al. 2004), totaling almost 200 billion nucleotide bases (about the number of stars in the Milky Way). Over 5 million of these nucleotide sequences have been translated into amino acid sequences and deposited in the UniProtKB database (Release 12.8) (Bairoch et al. 2005). The protein sequences in UniParc triple this number. However, the protein sequences themselves are usually insufficient for determining protein function as the biological function of proteins is intrinsically linked to three dimensional protein structure (Skolnick et al. 2000). The most accurate structural characterization of proteins is provided by X-ray crystallography and NMR spectroscopy. Owing to the technical difficulties and labor intensiveness of these methods, the number of protein structures solved by experimental methods lags far behind the accumulation of protein sequences. By the end of 2007, there were 44,272 protein structures deposited in the Protein Data Bank (PDB) (www.rcsb.org) (Berman et al. 2000) − accounting for just one percent of sequences in the Uni Prot KB database (http://www.ebi.ac.uk/swissprot). Moreover, the gap between the number of protein sequences and the number of structures has been increasing as indicated.

Proteins represent the most important class of biomolecules in living organisms. They carry out majority of the cellular processes and act as structural constituents, catalysis agents, signaling molecules and molecular machines of every biological system. In all cell functions proteins are virtually involved. Every single protein has specific function within the body. Some of the few proteins are involved in bodily movement, while others are involved in structural support. Proteins differ in functions as well as structures. One of the important goals pursued by bioinformatics and theoretical chemistry is protein structure prediction. It is highly important in biotechnology and medicine.

Proteins are classified according to structural and sequence similarity. The four different levels of protein structure are primary, secondary, tertiary, and quaternary structure. A single protein molecule may contain few of these protein structure types. The structure of protein determines the protein function. The primary structure of a protein is derived from the amino acid sequence of a protein and it is the most fundamental form of information available about the protein. It plays the most critical role in determining various characteristics of the protein such as its sub-cellular localization, structure and function. Because of this, amino acid sequence has tremendous potential to be used extensively for functional annotation of proteins.

Machine learning focuses on prediction, based on known properties learned from the training data. In the field of biology various application extensively uses methods which are based on machine learning algorithms. These methods have been utilized in diverse domains like genomics, proteomics and systems biology. Specifically, supervised machine learning approaches have found immense importance in numerous bioinformatics prediction methods. In this paper we have put different sections were we have explained how machine learning can be applied to protein structure and function predictions.

In this work, we have investigated the machine learning models with physical and synthetic properties to foresee the RMSD (Root Mean Square Deviation) of a displayed protein structure without its actual local state. Physical and Compound properties in particular aggregate exact vitality, optional structure punishment, add up to surface territory, combine number, deposit length and Euclidean separation are utilized. There are add up to 1056 demonstrated imitations structures having 3078 local structures. The demonstrated structures are taken from protein structure expectation focus (CASP-5 to CASP-10 tests), open imitations structures database (Public-Decoy, 2010) and local structure from protein information bank (RCSB).) the highlight significance is dictated by the Real Coded Genetic Calculation (RCGA) .machine learning model shaving the highlights and names Decision Tree, arbitrary woods, Linear model and Neural System for the expectation of RMSD protein structure. By the entirety tests, it is watched that irregular woodland display outflanks the other machine learning approaches in expectation of RMSD. Further, K-overlay cross approval is utilized to quantify the power of the best prescient model. At last, for the benchmarking of model rightness, the execution of best prescient model is thought about with top-performing ProQ2 (Ray et al., 2012).

FEATURES AND METHODS

1. Data set and its features

There are total 1056 modelled structures having 3078 native structures. The modelled structures are fetched from protein structure prediction center (CASP-5 to CASP-10 experiments), public decoys structures database (Public Decoy, 2010) and native structure from protein data bank (RCSB). Table 1 describes the physical and the chemical properties used in this study. A sample of the data set is shown in Table 2. Table 3 shows the correlation between each feature. There is no correlation of energy with euclidean distance, pair number, residue length and area. There is high correlation between

(i). Euclidean distance and pair number,

(ii). Residue length and pair number, and

(iii). Residue length and area

feature	Inform it Lii Hi
A ceil	Total surface arc a.
ED	Euclidean distance.
Energy	Total empirical energy.
55	Secondary structure penalty.
RL	Residue length
PN	Pair number

Table 1. Discription of the features

RMSD	Area	ED	Energy	SS	Rl.	PN
0.00	8243.0	4030.6	-3391.1	86	75.0D	165.00
8.03	791B.2	11984.2	-2273.2	29	153.00	102.00
6.77	0354.8	1 1535.1	-2422.5	66	67.00	186.00
13.26	15664.1	129761.0	-5S20.4	146	104.00	368.00
0.00	8836.1	12108.8	-2926.1	30	6600	101.00
6.76	12620.3	41461.0	-6206.3	146	61.00	116.00

Table 2. Sample dataset

	Energy	SS	El)	PN	RE	Anar
Energy	1.000	a 003	0.001	0.001	0.002	0.002
SS	0.003	1.000	0514	0.572	0.670	0.656
ED	0.001	0514	S.000	0.953	0.838	0.803
PM	0.001	0.572	0.953	1.000	0.913	0.837
Rl.	0.002	0.670	0.838	0.913	1.000	0.942
Area	0.002	0.656	0.803	0.837	0.942	1.000

Table 3. Correlation between each feature.

www.ignited.in

2.1 Feature Measurement

We have explained an overview of the physical and the chemical properties used in this research.

2.2.1 *Root Mean Square Deviation (RMSD)*

The RMSD is calculated using the superposition between matched pairs of Ca in two protein sequences. This superposition is computed using the Kabsch rotation matrix (Betancourt and Skolnick, 2001). The RMSD is calculated as:

RMSD **=** f

where, di is the distance between matched pair i, N is the number of matched pairs. RMSD is calculated using the freely available program at (RMSD, 2011).

2.1.2 Total surface area (Area)

Protein folding is done by various driving forces, which holds minimization of its total surface area. Degree of these external forces depends on the surface of protein exposed to the solvent, which convey the strong dependency of free energy on solvent accessible surface area (SASA) (Durham *et al.,* 2009). SASA has been used as one of the important properties to assess the quality of protein structures. Hydrophobic collapse is considered as a major factor in protein folding and this can be estimated as a loss of SASA of non-polar residues. Each amino acid shows a different affinity to be found on the surface of the protein based on the functional groups present in its side chain (Janin, 1979). Some questions arise with regard to the usage of SASA: (i) should it be the total area or is it the area of the non-polar residues, (ii) what is the standard fixed value of SASA for a native structure and (iii) is the rule of minimum area applicable to non-globular proteins. Here, total SASA have been calculated using Lee & Richards (Janin, 1979) method.

2.2.3 Euclidean distance *(ED)*

Spatial positioning of Ca atoms decides the overall conformation of a protein. Recently, neighborhood profiles of Ca atoms for each pair of residues have been characterized and observed to be invariant in 3618 native proteins suggesting certain geometrical constraints in their positioning (Mittal and Jayaram, 2011). The authors consider four aliphatic non polar residues Alanine (ALA), Valine (VAL), Leucine (LEU) and Isoleucine (ILE); collectively they formed 6 unique pairs among each other. Cumulative inter-atomic distance of their respective Cp atoms were calculated for each residue pair. Euclidean distance is calculated by taking the cumulative difference of Ca and Cp. Euclidean distance between two protein sequences p and q is given as:

$Yj^* - n)^E$

where, n is sequence length.

2.2.4 Total empirical energy (Energy)

The total empirical energy is the absolute sum of electrostatic force, van der Waals force and hydrophobic force (Arora and Jayaram, 1997; Naranget al., 2006). Molecular dynamics simulation package AMBER12 (G¨otz et al., 2012) is used to compute total empirical energy. It is computed as given below:

$$E_{elec}^{ij} = \frac{332 * qi * qj}{r_{ij}}$$

$$E_{vdW}^{ij} = \frac{C_{12}^{ij}}{r_{ij}^{12}} - \frac{C_6^{ij}}{r_{ij}^6}$$

$$E_{hyd}^{ij} = \frac{M_{12}^{ij}}{r_{ij}^{12}} - \frac{M_6^{ij}}{r_{ij}^6}$$

2.2.5 Secondary Structure Penalty (SSP)

Secondary structure prediction has reached to 82% accuracy (Sen et al., 2005) over the last few years. Therefore deviation from ideal predicted secondary structures can be used as a measure to quantify the quality of a structure. Secondary structure penalty is measured from the secondary structure sequence. It is computed as the absolute difference of the STRIDE (Frishman and Argos, 1995) and the PSIPRED (Jones, 1999) scores. STRIDE is used to assign three secondary structure classes, i.e., helix, sheet and coil to each residue in the protein models based on coordinates. PSIPRED is used to predict the probability for the same secondary structure classes.

where, P is the protein sequence ; Sstride(P) and Spsipred(P)are the STRIDE and PSIPRED scores respectively; Shelix(P), Ssheet(P) and Scoil(P) are the STRIDE score for helix, sheet and coil of protein sequence P respectively; F1(P) is the predicted probability from PSIPRED for the secondary structure of the central residue in the sequence window; F2(P) is the correspondence between predicted and actual secondary structure over a 21- residue window; F3(P) is the secondary structure assigned by STRIDE ,binary encoded into three classes over a 5-residue window .

METHODOLOGY

The philosophy is clarified in Fig. 2. In the exceptionally past step, the displayed protein structures are taken from protein structure forecast focus (CASP-5 to CASP-10 tests), open baits database (Public-Decoy, 2010) what's more, local structure from protein information bank (RCSB). They

include estimation, as talked about in segment 2.2, of protein structures is done in second step.in the subsequent stage The evacuation of copies and missing quality passages from dataset were done. There are add up to 1056 fakes structures having 3078 local structures. In the forward stride, the Real Coded Genetic Algorithm (RCGA) is utilized to gauge the significance of each element. Highlight choice makes the forecast of model productive and precise. In the last stride, the four machine learning approaches were prepared and tried on the informational collection with their default parameters. At last, the assessment of the model is done on Root Mean Square Error (RMSE), Coefficient of Determination (R2), Correlation and Precision and K-crease cross approval is utilized to gauge heartiness of the best prescient model.

Real Coded Genetic Algorithms (RCGA)

Real Coded Genetic Algorithms (RCGA) is one of the most popular optimization method among the evolutionary algorithm (EAS). It's a population based stochastic search approach and in general can be regarded as a searching method from multiple positions and directions .It is used for the biological evolution in nature selection and it consists three operations – reproductions, crossover and mutation Multiple good solution are carried out by the reproduction operations. The crossover operation blends genetic information operation between solutions to generate new candidate solution .And the mutation operation convergence to a suboptimum solution. Due to its good results in solving optimization problems, it has been widely applied in science, economics and engineering fields. The crossover operation is considered as important in the evolutionary algorithm as it guides the search by producing new considered solution. In past, the performance of RGCA has been developed by many difference kinds of crossover operators. From technical point of view, the crossover operators developed are mainly on the base of the line segment connection and distribution analysis of parent solutions, e.g., mean-centric and parent centric approaches. As observed in previous studies however, these featured approaches might bring out some problems. We searched firstly, there could be some areas where the crossover operation cannot generate offspring as the size of population so it's relatively small as compared to the whole search space, and/or the distribution of the initial given population does not uniformly scatter over the search space. Secondly, these crossover operators do not work well on the problems when the optimum is located at or near the boundaries of the search space Moreover, due to the inherent nonlinearities, complex constraints and apparent interaction among decision variables, most RCGAs can unavoidably experience the problem of excessive complexity in implementation and the difficulties in locating true global optimal for some practical applications.

2.4.1. Feature Importance using RCGA

The RCGA is used to find the importance of each features. It defines the weight to each feature according to the objective function defined in eq. (3). As consider crossover rate (CR) and mutation rate (MR) are set to be 0.9 and 0.01 respectively. Uniform crossover operator is used for crossover and arithmetic mutation (adding or subtracting a small number) is used as mutation operator. After five different runs, the weight obtained for each feature is described in Table 4. We can see in the above table the average weight of energy is highest and area is lowest that also signifies the importance of each feature in the dataset. As the weight given to each feature is significant so all the features are selected for the experiment where, T is the total number of instances in training data set, R is the RMSD, P is physical and chemical properties, n is the number of properties (6 in this case) and w is the weight given to each feature defined in the range of [0,1].

2.4.2 Machine learning models

In this work, we used four machine learning models (refer, Table 5) for prediction of RMSD of protein structure. The models are available in R open source software. R is licensed under GNU GPL. In precisely the models is presented below:

1. Decision Trees: This model is an extension of C5.0 classification algorithms described by Quinlan.

2. Random forest: It is based on a forest of trees using random inputs.

3. Linear Models: It uses linear models to carry out regression, single stratum analysis of variance and analysis of covariance.

4. Neural Network: Training of neural networks using back propagation, resilient back-propagation with or without weight or the modified globally convergent version.

Runs	Energj[1]	**RL**	**1'N**	S		**i:d**	**Area**
1	0.156	0.184	0.172	0.1	0	0.123	0.115
2	0.250	0.190	**0.169**	0.1	3	0.120	0.118
3	0.253	0.187	**0.172**	0.1	0	0.123	0.115
4	0.249	0.182	**0.174**	0.1	**8**	0.125	0.122
5	0.251	0.184	**0.177**	0.1	*b*	0.117	0.115
Avg.	0.252	0.185	0.173	0.1	1	0.122	0.117
Ranking	**1**	**2**	**3**	**4**		**5**	***b***

Table 4. Importance of each feature using RCGA.

MODEL EVALUATION

We have many ways to measure performance of the prediction, where some are more suitable than the others depending on the application considered. A brief discussion on the performance measures is

explained below. The formula used for all the machine learning models is given by:

RMSD _ Area + ED + Energy + SS + RL + PN

Model	Package	Tuning Pinneter(s) Ref.	
Decision	CSO	winnow,	(Quinlan, 1986)
random forest	Random Foiesl	tutij	(Liaw and Wiener, 2002)
Linear Model	stats	None	(Chambers, 1977)
Neural Network	ncuralnel	layet2, layerl, layer]	(Riedmillcrand Braun, 15)3)

Table 5. Machine learning models used

RESULT

In this area, we watch the forecast aftereffects of all the four machine learning models on the preparation and testing dataset. The machine taking in models may be experience the ill effects of over fitting because of the likelihood of model utilized for preparing the model is not the same as the rule used to judge the viability of a model. Here, to stay away from the over fitting , every one of the four machine learning models are keep running on their default parameters also, the dissemination of information in preparing and testing set are 70% also, 30% separately for every one of the models. Table 6 demonstrates a relative execution of the considerable number of models in the expectation of RMSD on RMSE, Correlation, R2 and Precision. The execution comes about demonstrate that the irregular backwoods display beats the machine learning models in the forecast of RMSD of the protein structure without its actual local state.

The RMSE is utilized to quantify the contrasts between values anticipated by a model and the qualities really watched.

The RMSE is figured utilizing condition 4. The arbitrary woodland have the least RMSE of 0.26 in the preparation dataset and 0.48 in the testing dataset. The connection portrays the factual connection amongst genuine and anticipated esteems and it is ascertained utilizing.

	Tritont dataset			Mnjidital		
Model	m.	Cmlaliim If	Acniracyfc	RMSE If	CofntaUoa	AtonrjOi
IkisiorTrcc	1.20	0.50 0.25	79,55	1,1ft	051 0.26	824ft
RjitdomI'iHl	O,	11,98 UK	9919	0,48	0,900,82	97,02
Linear Model	1.43	0.25 0.0ft	65.51	1.44	0.21 (.05	65.97
ffajial Ndwort	1.39	0.31 0.10	70.19	1.4ft	0.0ft 0.00	67.15

Table 6. Performance comparison of all four models on training and testing data set.

CONCLUSION

In this work, we investigate four machine learning techniques with six physical and compound properties to anticipate the RMSD of protein structure without its actual local state. The correct nature of a model is communicated as far as how the model scoring the normal esteems from a given arrangement of high determination test structures. Here, the strategies machine learning do exclude some other data from different models or option format structures. Every one of the models are assessed on RMSE, relationship, R2 and precision. By the analyses, it is discovered that irregular woodland strategy outflanks the machine learning strategies in the forecast of RMSD. The K-overlay cross approval is utilized to gauge the strength of irregular woods. At long last, for the benchmarking of model rightness, the execution of arbitrary backwoods show is contrasted and top-performing ProQ2 . the benchmark technique is single-display strategy and it is discovered that the irregular woods forecast precision is very great.

REFERENCES

Arora, N. and Jayaram, B. (1997). Strength of hydrogen bonds in a helices. Journal of computational chemistry, 18, pp. 1245–1252.

Baldi, P. and Pollastri, G. (2002). A machine learning strategy for protein analysis. Intelligent Systems, IEEE, 17(2), pp. 28–35.

Betancourt, M. R. and Skolnick, J. (2001). Universal similarity measure for comparing protein structures. Biopolymers, 59(5), pp. 305–309.

Blanco, A., Delgado, M., and Pegalajar, M. (2001). A real-coded genetic algorithm for training recurrent neural networks. Neural networks, 14(1), pp. 93–105.

Bryson, K., Cozzetto, D., and Jones, D. (2007). Computer-assisted protein domain Science, 8(2), pp. 181–188.

Chambers, J. (1977). Computational methods for data analysis. Applied Statistics,(2), pp. 1–10.

Cheng, J., Saigo, H., and Baldi, P. (2005b). Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. Proteins: Structure, Function, and Bioinformatics, 62(3), pp. 617–629.

Cheng, J., Sweredoski, M., and Baldi, P. (2005a). Accurate prediction of protein disordered regions by mining protein structure data.

Data Mining and Knowledge Discovery, 11(3), pp. 213–222.

Durham, E., Dorr, B., Woetzel, N., Staritzbichler, R., and Meiler, J. (2009). Solvent accessible surface area approximations for rapid and accurate protein structure prediction. Journal of molecular modeling, 15(9), pp. 1093–1108.

Fariselli, P., Olmea, O., Valencia, A., and Casadio, R. (2001). Prediction of contact maps with neural networks and correlated mutations. Protein engineering, 14(11), pp. 835–843.

Frishman, D. and Argos, P. (1995). Knowledge based protein secondary structure assignment. Proteins, 23(4), pp. 566–579.

Goldberg, D. E. (1990). Real-coded genetic algorithms, virtual alphabets, and blocking Urbana, 51, pp. 61801.

Herrera, F., Lozano, M., and Verdegay, J. L. (1998). Tackling real coded genetic algorithms: Operators and tools for behavioral analysis. Artificial intelligence review, 12(4), pp. 265–319.

Janin, J. (1979).Surface and inside volumes in globular proteins.

Jones, D. (1999). Protein secondary structure prediction based on position specific scoring matrices. JMB, 292(2), pp. 195–202.

Kim, D., Xu, D., Guo, J., Ellrott, K., and Xu, Y. (2003). PROSPECT II: protein structure prediction program for genome-scale applications. Protein engineering,16(9), pp. 641–650.

Krogh, A., Larsson, B., Von Heijne, G., Sonnhammer, E., et al. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology 305(3), pp. 567–580.

Liaw, A. and Wiener, M. (2002). Classification and Regression by random Forest. R News, 2(3), pp. 18–22.

Mittal, A. and Jayaram, B. (2011). Backbones of folded proteins reveal novel invariant amino acid neighborhoods. Journal of Bio molecular Structure and Dynamics, 28.

Narang, P, Bhushan, K., Bose, S., and Jayaram, B. (2006). Protein structure evaluation using and all-atom energy based empirical scoring function. Journal of Bio molecular Structure and Dynamics, 23(4), pp. 385–406.

Corresponding Author

Akshay Pandey*

B. Tech. (CSE) R. D. Engineering College, Ghaziabad

E-Mail –