Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
Vol. 27 (2024), No. 4, pp. 1129–1138
DOI : 10.47974/JDMSC-1968

Unleashing a cascade of machine learning for turbocharged sequence alignment in discrete mathematical sciences using GPU

Karamjeet Kaur *
School of Engineering and Technology
Sharda University
Greater Noida 201306
Uttar Pradesh
India

Sudeshna Chakraborty †
School of Computing Science & Engineering
Galgotias University
Greater Noida 203201
Uttar Pradesh
India

Anil Kumar Sagar §
School of Engineering and Technology
Sharda University
Greater Noida 201306
Uttar Pradesh
India

Abstract

Prediction and alignment of multiple sequences is the core methodology for information extraction in proteins. The sequence of grouping accuracy predicts the match score of proteins. In recent years, protein sequence alignment has emerged as a critical task in bioinformatics studies. There are two main methods for sequence alignment: pair-wise and multiple sequence alignment. MSA methods have some limitations which were revealed by prior benchmark studies. We have proposed a new benchmark framework for protein grouping based on a machine-learning algorithm. The proposed algorithm is a cascaded vector machine with three levels for data sampling and normalization of protein data. The

* *E-mail:* karam_7378@yahoo.com (Corresponding Author)
† *E-mail:* sudeshna2529@gmail.com
§ *E-mail:* aksagar22@gmail.com

proposed algorithm was simulated in MATLAB tools and analyzed using the standard parameters of prediction. The analysis of the results suggests that proposed algorithm is efficient than CNN and RNN.

Subject Classification: *Primary 93A30, Secondary 49K15.*

Keywords: *MSA, Deep learning, Protein, RNA/DNA, Machine learning, Prediction.*

1. Introduction

The prediction of similar groups of sequence patterns in proteins is a very challenging task. The ability of protein sequence prediction is increased by the development of computational methods which are based on machine learning and deep learning. The prediction of proteins faces various feature-related problems in terms of classification and categorization. Recently, several authors focused on the prediction accuracy of multiple sequences in RNA and DNA [1][2]. The common factor between the approaches of sequence prediction and alignment is distance. The distance factor implied the role of contact prediction in multiple sequence alignment. Modern prediction methods relies on multiple sequence alignment to identify evolutionary relationships between residues [3][4]. The quality of MSA affects accurate contact forecasts. Recent work has generated MSA using a conservative E-value cut-off. Sometime big E-value cut-off give inaccurate MSA and inaccurate co-evolution information and if query protein doesn't have many close homologs then however, big E-value cut-off gives MSA which may provide valuable information [5][6]. The issue is that there should be proper level of sequence homology in protein family. The creation of MSA and its subsequent execution of residue co-evolution analysis serve as the cornerstones of the most recent methods for finding the inter-residue contacts and distances. As a result, residue co-evolutions can be used to estimate contacts. It also helps to find the distance between residues accurately. The fundamental premise is that two residues that are near to one another in space always have a propensity to co-evolve. Typically, residue correlations carried out by MSA are used to infer the co-evolution relationship. However, the indirect correlations between residues are always a barrier to this strategy because they can cause transitivity in the spatial proximity of residues and, as a result, inaccurate estimation of inter-residue contacts [7][8]. This paper proposes an algorithm for the prediction of sequence grouping for proteins using machine learning. The

proposed algorithm improves the performance of contact prediction in a protein sequence.

The following sections comprise the remainder of the paper: section 2 explains related work, section 3 explains proposed methodology, section 4 shows simulation and result analysis and section 5 finally concludes the proposed work.

2. Related work

In the creation of novel medicine and advancement of medicine, the utility of protein sequence alignment and prediction is crucial. Recently, several authors proposed various algorithms based on swarm intelligence and machine learning algorithms. Some significant contributions are described here.

The Felsenstein zone is bias-inducing areas but the CNN-based classification algorithm is very resistant to them, and is extremely accurate in processing data [1]. CNN produced confidence scores and predict appropriate topology according to the authors [2] when compared to traditional techniques used to produce bootstrap and posterior probability scores. To increase the precision of the progressive MSA method, this research developed a novel and simple method. They used the DLPAlign which is a multiple sequence alignment tool to compare this method's performance to that of eleven other top alignment techniques on three empirical alignment benchmarks to assess it. In retrospect, DLPAlign which is based on a decision-making model, performs exceptionally well for protein families with minimal similarity and can align protein families with greater accuracy when compared to top MSA approaches now in use.

Jain, Aashish, and others [3] introduced AttentiveDist, a novel method that increases the evolutionary information available to the model by using many MSA produced with various E-values. To evaluate the relative weights of MSA at the inter-residue stage, attention layer is added to neural network. To develop residue distance/contact prediction tools, this work offers yet another compelling illustration of how protein structure knowledge may be further exploited by utilizing cutting-edge deep learning technology.

Protein interchain interactions [4] are predicted using deep learning techniques that were highly successful at predicting intrachain residue-residue connections. When it is understood that alignment of a monomer which makes homomultimers. It is found that it contains the co-evolutionary signals of residue pairs in contact, at intrachain and

interchain, the predictor's accuracy is significantly higher than random predictor, even though its precision is not particularly high because it was trained primarily for intrachain contact prediction. Sometimes, this approach correctly predicted that there would be more interchain connections than intrachain contacts. The gain is partly attributable to deep learning techniques that forecast distances [5] using multiple sequence alignment and inter-residue interactions.

According to Zhen Chen and others [6], Xuyang Liu and others [7] a sufficient number of effective normalized sequences must be aligned in numerous sequences to estimate the protein contact map. The anticipated contact maps are frequently unsatisfactory when Nf is small. They put forth the MSA dropout innovative data argumentation technique for feature synthesis and utilized a consistency learning network to predict contact maps. They applied MSA's dropout and consistency learning to the prediction architecture that they replicated. They simply selected a range of MSA dropouts that, for proteins with relatively low Nf, result in a considerable improvement for this investigation.

Multiple sequence alignment (MSA) techniques developed by Kasuni Perera et al. [8] are used as the basis for many microbiological investigations to infer homologous regions in biological sequences. While center star technique is an MSA methodology that can handle a sizable dataset, it tends to perform poorly when there are numerous centers present in the collection of sequences. In this technique, they first use the bisecting K-means algorithm to find the groups of sequences within the sequences by using K-means as the clustering attributes. Each subgroup of sequences is subjected to the center star approach separately.

Chao Zhang and others [9] proposed methods for automatic error identification that are more crucial than ever, because it is getting harder and harder to physically check for faults as datasets get bigger. The frequently used alignment masking techniques eliminate all of the aligned sites. Methods that search for unexpected patterns in branch lengths offer workable substitutes for longer mistakes. Other parameters, besides error length, were important. More challenging to detect were mistakes in alignments with few or short sequences, as well as errors in alignments with very divergent alignments.

A multi-objective artificial fish swarm technique (MOAFS) is presented by Ali Dabba et al. [10] to address multiple sequence alignment in this study. The cooperative, decentralized, and parallelism characteristics of the swarm algorithm is used by MOAFS for exploration in search space of sequences. The effectiveness of proposed algorithm has been shown by

comparison to several alignment methods and techniques based on evolutionary algorithms and progressive technique which have either one or several objectives.

Based on the work of Hafiz Asadul Rehman and others [11], this study suggests an Enhanced Bird Swarm Alignment Algorithm. This is recommended to determine best alignment between different sequences. In this study, an EBSAA based on swarm intelligence is presented for MSA problem. In addition to Jeevana Jyothi Pujari [12] in the field of bioinformatics, multiple sequence alignment is a significant ant-search problem. This improved method increases or reduce solution set dynamically in search space for each iteration and investigate greater optimality in MSA. Various techniques have been developed to calculate the ideal sequence alignment, but numerous alignments still pose a difficulty in determining the alignment's ideal accuracy [13].

Hiroyuki Fukuda and coworkers [14] developed neural network model to improve the accuracy of MSA in which weights are given to sequences and their model predict high accuracy. To improve accuracy, they added predictions for secondary structures and solvent-accessible surface areas to model and increased the number of tasks in it. Additionally, they showed that ensembling technique could improve accuracy. They tested their models using previous CASP target protein domains and showed that the final model is better than or on par with the current meta predictors.

The group Geraldo Francisco Donega Zafalon et al. [15] used the progressive method for new hybrid refinement phase which correct the numerous sequence alignment locally and this alignment is generated by genetic algorithm. Their findings show that as compared to GA results, this method improve the quality of the alignment for the Bali Base families. They can therefore conclude that their method can improve MSA results generated by GA-based technologies and deliver alignments with more biological significance. The big data technology is used significantly to cut the time it takes to calculate libraries by utilizing the high level of parallelism offered by the map reduce paradigm [16].

3. Methodology

The cascaded prediction model of sequence prediction is cascading of a support vector machine. The learning classification and detection rate is very high instead of other machine learning algorithms. Here we describe the processing of the algorithm, one part of the algorithm work

with the normalization algorithm for the dimension reduction of features. Then reduced features process through the cascaded models. The cascaded model is a connected feedforwarded neural network model. The first phase of the hybrid algorithm work as training of sample data using vector models which work as a prediction of sequence. The processing of the algorithm describes here.

The sampling of data as

$$SMO_F : F^{Tf} \rightarrow D^{T_f}, \text{ where } F_T \in D^{T_f}$$

The set reduces the dimension features of the protein.

S the transformed vectors $y_j \in D^T$ to map the input of SV

$$y_j = S_j(f_j) + Z, \quad \forall Z \le j \le Z + K$$

The above mention expression represents the data.

Cascaded training of prediction as

$$Cas = yi_0 \rightarrow z(f_0) + Z$$

Define the variance matrix of samples of SV1 and SV2 C_k and R_k

Now the prediction of the cascaded model is

$$\text{Pre(model)} = \left\{ \| D - D_k \| R_k^{-1} + \sum_{j=k}^{k+z} \| ZF_j(f) - y_j \| C_j^{-1} \right\}$$

4. Experimental Analysis

To validate the proposed algorithm for multiple sequence alignment, it is simulated in MATLAB tools. The simulation process was carried out on i7 processor and 16GB RAM with the Linux operating system. The three datasets have been tested, chimpanzee, mouse, and dog protein sequences. The results are shown in Figure 1, 2 and 3. The parameters are chosen as r = 0.1, c = 0.1, q0 = 0.8, b =2. The proposed cascaded machine learning algorithm applies 10 cross-fold ratios. The sampling of prediction as a similar group or not a similar group. The prediction parameters describe as:

$$\text{Accuracy} = \frac{\text{Total No. of Correctly Classified Instances}}{\text{Total No. of Instances}} \times 100$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100$$

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100$$

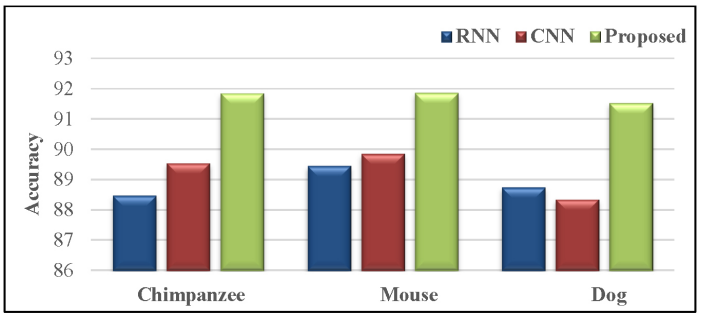

Figure 1

Comparative performance analysis of accuracy using RNN, CNN and proposed method using Chimpanzee, Mouse, and Dog datasets.

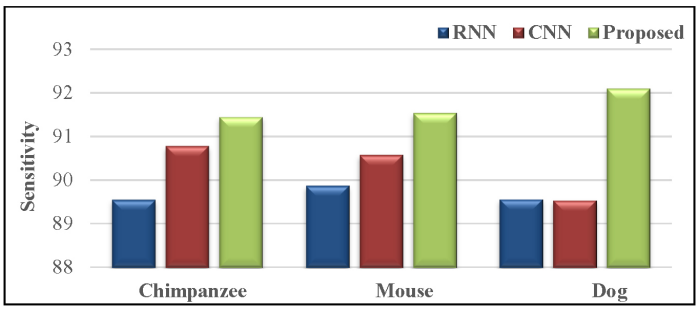

Figure 2

Comparative performance analysis of sensitivity using RNN, CNN and proposed method using Chimpanzee, Mouse, and Dog datasets.

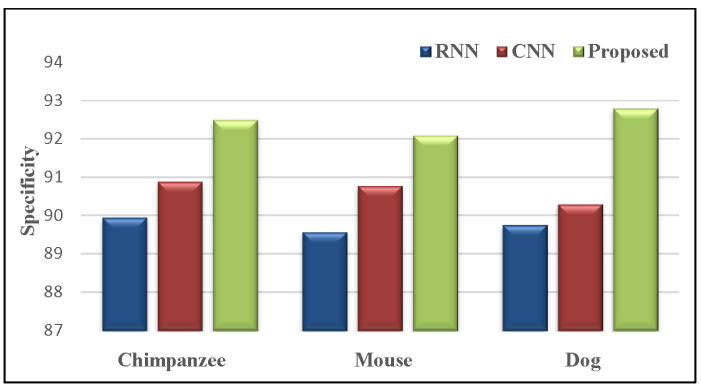

Figure 3

Comparative performance analysis of specificity using RNN, CNN and proposed method using Chimpanzee, Mouse, and Dog datasets

5. Conclusion & Future Scope

The proposed algorithm predicts the sequence of three datasets. The accuracy of the algorithm is improved by adding accurate feature selection method and good results are produced using this. In essence, we found that a prediction model for multiple sequence alignment can be created by choosing appropriate samples from reliable source to find MSA. The dataset must therefore be pre-processed and this is very crucial step in preparing the dataset before applying machine learning algorithm and improving results. For a predictive model to produce reliable results, a suitable algorithm must be utilized. We note that in the majority of models, CNN and RNN have positive benefits for predicting sequence.

References

[1] Suvorov, Anton, Joshua Hochuli, and Daniel R. Schrider, Accurate inference of tree topologies from multiple sequence alignments using deep learning, Systematic Biology 69, no. 2, 221-233, (2020).

[2] Kuang, Mengmeng, Yong Liu, and Lufei Gao, DLPAlign: A Deep Learning based Progressive Alignment Method for Multiple Protein Sequences, In CSBio'20: Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics, pp. 83-92, (2020).

[3] Jain, Aashish, Genki Terashi, Yuki Kagaya, Sai Raghavendra Maddhuri Venkata Subramaniya, Charles Christoffer, and Daisuke Kihara, AttentiveDist: Protein Inter-Residue Distance Prediction Using Deep Learning with Attention on Quadruple Multiple Sequence Alignments, bioRxiv, (2020).

[4] Quadir, Farhan, Raj S. Roy, Randal Halfmann, and Jianlin Cheng, DNCON2_Inter: predicting interchain contacts for homodimeric and homomultimeric protein complexes using multiple sequence alignments of monomers and deep learning, Scientific reports 11, no. 1, 1-10, (2021).

[5] Jain, Aashish, Genki Terashi, Yuki Kagaya, Sai Raghavendra Maddhuri Venkata Subramaniya, Charles Christoffer, and Daisuke Kihara, Analyzing the effect of quadruple multiple sequence alignments

on deep learning based protein inter-residue distance prediction, Scientific Reports 11, no. 1, 1-13, (2021).

[6] Chen, Zhen, Pei Zhao, Fuyi Li, Tatiana T. Marquez-Lago, André Leier, Jerico Revote, Yan Zhu et al., iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA, and protein sequence data, Briefings in Bioinformatics 21, no. 3, 1047-1057, (2020).

[7] Liu, Xuyang, Lei Jin, Shenghua Gao, and Suwen Zhao, Protein contact map prediction using multiple sequence alignment dropout and consistency learning for sequences with fewer homologs, bioRxiv (2021).

[8] Perera, Gamage Kokila Kasuni, and Champi Thusangi Wannige, A hybrid algorithm for identifying partially conserved regions in multiple sequence alignment, *International Journal of Computers and Applications* 43, no. 10, 979-986, (2021).

[9] Zhang, Chao, Yiming Zhao, Edward L. Braun, and Siavash Mirarab, TAPER: Pinpointing errors in multiple sequence alignments despite varying rates of evolution, Methods in Ecology and Evolution 12, no. 11, 2145-2158, (2021).

[10] Dabba, Ali, Abdelkamel Tari, and Djaafar Zouache, Multiobjective artificial fish swarm algorithm for multiple sequence alignment, INFOR: Information Systems and Operational Research 58, no. 1, 38-59, (2020).

[11] Rehman, Hafiz Asadul, Kashif Zafar, Ayesha Khan, and Abdullah Imtiaz, Multiple sequence alignment using enhanced bird swarm align algorithm, *Journal of Intelligent & Fuzzy Systems Preprint*, 1-18, (2021).

[12] Pujari, Jeevana Jyothi, and Kanadam Karteeka Pavan, Multiple Sequence Alignment based on Enhanced Brainstorm Optimization Algorithm with dynamic population size (EBSODP), Annals of the Romanian Society for Cell Biology, 10033-10042, (2021).

[13] Poonia, Ramesh C., and Linesh Raja. "On-demand routing protocols for vehicular cloud computing." In Research Anthology on

Architectures, Frameworks, and Integration Strategies for Distributed and Cloud Computing, pp. 96-122. IGI Global, (2021).

[14] Fukuda, Hiroyuki, and Kentaro Tomii, DeepECA: an end-to-end learning framework for protein contact prediction from a multiple sequence alignment, BMC bioinformatics 21, no. 1, 1-15, (2020).

[15] Zafalon, Geraldo Francisco Donega, Vitoria Zanon Gomes, Anderson Rici Amorim & Carlos Roberto Valencio, A Hybrid Approach using Progressive and Genetic Algorithms for Improvements in Multiple Sequence Alignments, In ICEIS (2), pp. 384-391, (2021).

[16] Sharma, Shilpa, et al. "Hybrid HOG-SVM encrypted face detection and recognition model." *Journal of Discrete Mathematical Sciences and Cryptography* 25.1 (2022): 205-218.