Enrichment of Remote Homology Detection using Cascading Maximum Entropy Markov Model

Manikandan P\textsuperscript{1*}, Ramyachitra D\textsuperscript{2}, Muthu C\textsuperscript{3}, Sajithra N\textsuperscript{4}

\textsuperscript{1}Department of Data Science, Loyola College, Chennai-600 034, India; \textsuperscript{2}Department of Computer Science, Bharathiar University, Coimbatore-641 046, India.

INTRODUCTION

Genome sequence projects provide a huge number of protein sequence data in several databases. The volume of inconsistency among the sum of known protein sequence and the sum of experimentally determined protein structures has been rapidly improved due to the experimental obstacles in analyzing the protein structures and functions. To overcome the limitations, a huge number of computational tools and methodologies have been developed to analyze the protein structure and functions employing the sequence information offered in open-source repositories. If two proteins share evolutionary origins and having more similarities are termed as homolog proteins. Earlier methods in protein homology detection can be clustered into three categories such as pairwise, discriminative and generative methods. Dynamic programming is used in pairwise methods to perform sequence similarity for homology detection.\textsuperscript{1} Dynamic programming method consumes more computation times for relatively long protein sequences. To accelerate the alignment, a heuristic method named Basic Local Alignment Search Tool (BLAST)\textsuperscript{2} is used to find a near-optimal alignment within a rational period. The general hypothesis is that two proteins sequence identity is above 30\% over their entire lengths is
termed as homolog. The disadvantage of pairwise sequence alignment is biological erroneousness of assessment concerning only one identified protein homolog.

Several iterative techniques such as PSI-BLAST, SAM-T98 have been developed to overcome the drawbacks of pairwise methods. However, the generative methods also produce so much false-positive results. Hence, the recent studies on remote homology detection used discriminative methods to make separation among homolog and non-homolog proteins. Remote homology detection of protein sequence plays a crucial role in the field of computational biology because remote homologous proteins contributing similar functions and structures that can be helpful in the protein studies of 3D structure and function. The remaining section of this manuscript is developed as follows: Section 2 designates the background study of existing methods for solving the problem of remote homology detection. The methodology of the proposed C-MEMM algorithm is explained in Section 3. The experimental outcomes of the proposed and the existing algorithms for the Structural Classification of Proteins (SCOP) database is emphasized in Section 4. Lastly, Section 5 spotlights the conclusion and future enhancement.

Background Study

Genome sequence projects generate a huge amount of sequence information and it needs to be annotated. Most of the homology detection methods use the scoring metric to predict the similarity among a query protein with the database sequences. The Smith-Waterman score is frequently used for comparing two protein sequences. In Bioinformatics, the structural and functional annotation of an unknown protein is a crucial task. The annotation of a new protein sequence is based on the similarity among the query protein with the structures and function of the known protein.

If such protein relationship cannot be identified using direct sequence-based approaches, it can be termed as remote homologues and the discovery of remote homologue relationship is a difficult task in computational biology. In several cases, the functional annotation for the new proteins can be done using the protein structures. But the information of protein structures is inadequate and the structure database is also very deliberate. Hence, there is a need to find distant homologues by abusing protein sequence information. Several techniques have been developed to identify the homologous sequences grounded on profile Hidden Markov Models (HMM) and Position-Specific Scoring Matrices (PSSM). The probabilistic model-based profile Hidden Markov Models are further complex in detecting the remote homologue proteins. HHblits and HH search are used to find the remote homologues by the HMM-HMM comparison. The Cascading PSI-BLAST method cover more distant relationships, but the major drawback of this method is computational time, which increases the size and complexity of the databases. Pcons delivers improved automated tools for the prediction of protein structure and analysis of this structure using consensus. The difficult target protein sequence is robotically submitted to open-source servers of fold recognition to find distant structural homologies. Inside the kernel formulation background, the consensus method is used to merge heterogeneous data by semidefinite, second-order cone, semi-infinite linear programming and Bayesian framework. These kernel learning methods are not scalable and guide to very complex frameworks. The Motifs pattern of protein sequence represents highly conserved regions of proteins. The motif content of two protein sequences is used to express a similarity among them. On completion of motifs, significant hints to a protein function can habitually be exposed even if it is not universally similar to any known protein in sequence databases. In consequence, the motif method suits well in remote homology detection and emphasis conserved segments of sequence having more similarity in the vaguely related protein sequence. An enhancement in protein remote homology detection is accomplished by emerging techniques grounded on the profile-sequence comparison. The profile is typically constructed from the query protein sequence or directly from an input multiple sequence alignment.

SAM-T04 approach used Neural Networks (NNs) to evaluate the alignment and map the similarities among pairs of proteins. The main improvements made by the human intervention were from adding constraints to build beta-sheets and from splitting multi-domain proteins into separate domains. Support Vector Machines (SVM) are mainly used to resolve the problem of binary classification. The support vector machine technique builds one-versus-classification models. These models are assessed with esteem to how well each binary classifier can identify the proteins that fit its own modelled class. Long Short - Term Memory (LSTM) is a moderately quicker approach than using alignments and SVM based models. The LSTM is a recurrent NN method that can be used for remote homology detection by determining the discriminating patterns. However, this technique does not attain comparable performance. The Cascade-Hidden Markov Model (C-HMM) involves cascading of HMM to identify the true distant homologues at various levels. The execution time of cascade searches are higher than the other existing methods. Even though several methods have been developed for predicting the remote homologues proteins, but they do not promise to provide better results. Hence, the C-MEMM algorithm has been developed in this research work to overcome the drawbacks of the existing methods.

Methodology

In this research, the existing algorithms, namely PSI-BLAST, C-HMM and the proposed C-MEMM algorithms are compared to predict the remote homology detection.
From the results, it is inferred that the C-MEMM method provides improved results than the existing algorithms.

Diagram 1: Overall architecture diagram for remote homology detection.

Drawing stronger links will be equally valuable for both areas of research is the key motive of Cascading Maximum Entropy Markov Models (C-MEMM) model. Markov models are used for both prediction and estimation, where the states of sequence produce a consistent sequence of annotations. In the maximum entropy method, the annotations are self-determining is also known as logistic regression. Hence, the Maximum Entropy Markov Model can be obtained from logistic regression through an assumption of Markovian between target variables.

In general, the Maximum Entropy Markov Models (MEMM) algorithm is measured by forwarding pass in separation. In HMMs, the backward pass is no longer candid. MEMM algorithm is very efficient, but they grieve from the problem of label bias, where a feeble probability reduces the probability of inactivity across time. To overcome the problem label bias, the linear chain Conditional Random Fields (CRFs) are used. Classifier Chain (CC) is a Cascading MEMM and it can be observed as a T order MEMM. Several existing multi-label methods already split the sequence into subsets.

An associated method is developed in this research work which furthermore links the blocks together, this data is not separated. Most of the existing multi-label techniques are previously transformed the datasets earlier working on them. The same flow of C-HMM is used in C-MEMM, only the Hidden Markov Model (HMM) is replaced with MEMM. All the modules Cluster-MEMM, Cascade MEMM and Custom MEMM are used to classify remote homologues proteins from the sequence databases. The key inspiration of developing the C-MEMM method is to authorize effective detection of remote homology compared to any accessible sequence database, regardless of its dimensions.

C-MEMM algorithm includes cascading of MEMMs by numerous generations to identify true distant remote homologues at various levels such as superfamily and fold. Every true hit of the first group is measured as query sequences to instigate the second group of C-MEMM. Cluster-HMM Module clusters the collected homologues (hits) while penetrating against such enormous databases. Custom-MEMM filter hits gained from the first group are clustered at a specific sequence identity cut-off by Cd-hit. This technique further diminishes the search time without remarkable loss in sequence information in contrast with the two existing modules.

RESULTS AND DISCUSSION

In this research work, the proposed algorithm is tested with the well-known benchmark dataset named SCOP 1.53 for analyzing the performance of the algorithm based on its effectiveness. The proposed C-MEMM algorithm has been assessed by comparing with two existing techniques, namely PSI-BLAST and C-HMM.

Dataset Description

The SCOP 1.53 database has been used as a standard database for remote homology detection. The datasets are retrieved in FASTA format from the ASTRAL compendium for sequence and structure analysis. The SCOP 1.53 database comprises 4352 protein sequences gathered into seventy-four families and fifty-four superfamilies. The dataset was further reduced into 1294 protein sequences which is gathered into forty-six families, one hundred fifty-five superfamilies and two hundred sixty-five folds. The protein domains inside the family are taken as test instances and protein domains inside and outside the superfamily are taken as training positive instances. This produces fifty-four families with ten training and five test positive instances. The negative instances of the protein family are selected via the exterior of the positive sequences fold and were stochastically divided into training tests and test sets in a similar ratio as the positive example.

Performance Measures

The execution of PSI-BLAST, C-HMM and C-MEMM algorithm is assessed by the performance metrics such as coverage and precision values. The coverage is termed as the true positives acknowledged by the method of sequence search via the total amount of true associations extant in the database at various levels namely family, superfamily and fold (Eq.1).

\[
\text{Coverage} = \frac{TP(\text{family} / \text{superfamily} / \text{fold})}{\text{No. of sequences in database}(\text{family} / \text{superfamily} / \text{fold})} \tag{1}
\]

The false-positive can be term as, “any hit that did not belong to the same fold as query sequence”. Precision scores are used to detect the fraction of hits that are related to the sequence search (Eq.2).
Precision = \frac{TP \times 100}{TP + FP} \tag{2}

where TP and FP stand for the number of True Positives and False Positives which are recognized through sequence searches.

### Experimental Results

The Scop 1.53 database is used for testing the performance of the proposed C-MEMM algorithm with the existing algorithms namely PSI-BLAST and C-HMM. The comparison of performance measures such as coverage and precision for the existing and the proposed algorithms are shown in Table 1.

<table>
<thead>
<tr>
<th>S. No</th>
<th>Performance Measures</th>
<th>Organisms</th>
<th>PSI-BLAST</th>
<th>C-HMM</th>
<th>C-MEMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Coverage</td>
<td>Frog</td>
<td>63.73</td>
<td>76.55</td>
<td>84.73</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Bacillus Stearothermophilus</td>
<td>24.2</td>
<td>76.88</td>
<td>90.61</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Escherichia Coli</td>
<td>12.84</td>
<td>65.88</td>
<td>72.52</td>
</tr>
<tr>
<td>4</td>
<td>Precision</td>
<td>Frog</td>
<td>59.54</td>
<td>68.68</td>
<td>78.56</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Bacillus Stearothermophilus</td>
<td>28.6</td>
<td>58.68</td>
<td>82.41</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Escherichia Coli</td>
<td>17.84</td>
<td>52.52</td>
<td>62.52</td>
</tr>
</tbody>
</table>

Fig.1 shows the comparison of Coverage for the proposed and the existing algorithm such as PSI-BLAST and C-HMM. From Fig.1, it is inferred that the C-MEMM algorithm performs better than the other existing techniques in terms of Coverage values. Fig.2 shows the comparison of precision values for the proposed C-MEMM algorithm with the other existing algorithms. From Fig.2, it is inferred that the proposed C-MEMM algorithm provides better results compared to existing PSI-BLAST, C-HMM approaches. For the coverage values in Fig.1, it is inferred that the proposed C-MEMM algorithm performs 24.78 % better than the PSI-BLAST algorithm and 9.65% better than the C-HMM algorithm for the Frog organism.

For the Bacillus stearothermophilus organism, it is inferred that the proposed C-MEMM algorithm performs 82.29 % better than the PSI-BLAST algorithm and 9.16% better than the C-HMM algorithm. For the precision values in Fig.2, it is inferred that the proposed C-MEMM algorithm performs 31.94 % better than the PSI-BLAST algorithm and 12.58% better than the C-HMM algorithm for the Frog organism. For the Bacillus stearothermophilus organism, it is inferred that the proposed C-MEMM algorithm performs 65.29% better than the PSI-BLAST algorithm and 28.79% better than the C-HMM algorithm. For the Escherichia Coli organism, it is inferred that the proposed C-MEMM algorithm performs 71.47% better than the PSI-BLAST algorithm and 51.99% better than the C-HMM algorithm.

### CONCLUSION AND FUTURE ENHANCEMENT

Remote homology detection is one of the crucial tasks for detecting homology in low sequence similarity cases. The discriminative method has classified the proteins using the
sequences that must be transformed into fixed-length feature vectors. In this research work, the C-MEMM model was successfully implemented to detect protein remote homologies. When the C-MEMM method was applied to the feature data set, the sequence returned a prediction score of every instance. The effectiveness of the proposed approach is tested with Coverage and Precision values. Experimental results show that the proposed method effectively improved prediction performance. In future, this research work can be enhanced to discover novel features of protein sequences and applying other language processing methods for the detection of protein remote homology.

ACKNOWLEDGEMENT

The authors acknowledge the immense help received from the scholars whose articles are cited and included in references of this manuscript. The authors are also grateful to authors/editors/publishers of all those articles, journals and books from where the literature for this article has been reviewed and discussed.

Conflict of Interest

The authors declare that they have no conflict of interest.

Source of Funding: NIL

Author’s Contribution

Dr Manikandan P: Manuscript Preparation
Dr Ramyachitra D: Manuscript Editing
Dr Muthu C: Manuscript Reviewing
Ms Sajithra N: Literature Review

REFERENCES