IJRSET August 2015 Volume 2, Issue 6 Pages: 60-65
ISSN 2394-739X
International Journal for Research in Science Engineering and Technology

A Survey on Protein Target Identification using Classification and Prediction

[1] Dr.P. Senthil Vadivu, [2] M. Poornima MCA,
[1] Associate Professor/ Head, [2] M.Phil Research Scholar,
[1] Dept of Computer Application, [2] Dept of Computer Science,
[1&2] Hindustan College of Arts and science, Coimbatore.

Abstract:-

In an era that has been dominated by Structural Biology for the last 30-40 years, a dramatic change of focus towards sequence analysis has spurred the advent of the genome projects and the resultant diverging sequence/structure deficit.

The central challenge of Computational Structural Biology is therefore to rationalize the mass of sequence information into biochemical and biophysical knowledge and to decipher the structural, functional and evolutionary clues encoded in the language of biological sequences.

In investigating the meaning of sequences, two distinct analytical themes have emerged: in the first approach, pattern recognition techniques are used to detect similarity between sequences and hence to infer related structures and functions; in the second prediction methods are used to deduce 3D structure and ultimately to Infer function, directly from the linear sequence. In this article, we Attempt to Provide a Critical assessment of what one may and may not expect from the biological sequences to identify Major issues yet to be resolved.

Keywords: - [Proteins target identification, Classification, Prediction, Data Mining]

1. INTRODUCTION

Over the years a wide range of comparative modeling-based methods have been developed for predicting the structure of a protein (target) from its amino acid sequence. The central idea behind these techniques is to align the sequence of the target protein to one or more template proteins and then construct the target's structure from the structure of the template using the alignment(s) as reference. Recently we developed machine-learning methods 21 that can accurately estimate the root mean squared deviation (RMSD) value of a pair of equal-length protein fragments (i.e., contiguous backbone segments) by considering only sequence and sequence-derived information. Our interest in solving this problem is motivated by the operational characteristics of various dynamic-programming-based

1.1. Prediction

Predictive analysis is an advanced branch of data engineering which generally predicts some occurrence or probability based on data. Predictive analytics uses data-mining techniques in order to make predictions about future events, and make recommendations based on these predictions. The process involves an analysis of historic data and based on that analysis to predict the future occurrences or

events. A model can be created to predict using Predictive Analytics modeling techniques. The form of these predictive models varies depending on the data they are using. Classification & Regression are the two main objectives of predictive analytics. Predictive Analytics is composed of various statistical & analytical techniques used to develop models that will predict future occurrence, events or probabilities. Predictive analytics is able to not only deal with continuous changes, but discontinuous changes as well. Classification, prediction, and to some extent, affinity analysis constitute the analytical methods employed in predictive analytics. Predictive models analyze identify patterns in historical and transactional data to determine various risks and opportunities. Forecasting models capture relationships between many factors to allow assessment of the risks or potential associated with a particular set of conditions, guiding decision making for candidate transactions. Three basic techniques for Predictive analytics are Data profiling and Transformations, Sequential Pattern Analysis and Time Series Tracking. Data profiling and transformations are functions that change the row and column attributes and analyses dependencies, data formats, merge fields, aggregate records, and make rows and columns. Sequential pattern analysis identifies relationships between the rows of data. Sequential pattern analysis involves identifying frequently observed sequential occurrence of items across ordered transactions over time. Time Series Tracking is an ordered sequence of values at variable time intervals at the same distance. Time series analysis gives the fact that the data points taken over time. There are some advanced Predictive analytic techniques like Classification-Regression, Association analysis, Time series forecasting to name a few. Classification uses attributes in data to assign an object to a predefined class or predict the value of a numeric variable of interest. Regression analysis is a statistical tool for the study of relations between variables. Association analysis describes significant associations between data elements. Time series analysis is employed for forecasting the future value of a measure based on past values.

1.2 DATA CLASSIFICATION

The diverse systems that are ordinarily utilized for information classification will be discussed. The most regular systems utilized as a part of data classification are decision trees, rule-based methods, probabilistic methods, SVM methods, instance-based methods, and neural networks.

Feature Selection Methods

The first period of essentially all classification calculations is that of highlight determination. In most information mining situations, a wide mixed bag of components are gathered by people who are regularly not space specialists. Unmistakably, the unessential elements might frequently bring about poor displaying, since they are not all around identified with the class mark. Indeed, such elements will normally decline the classification exactness on account of overfitting, when the preparation information set is little and such components are permitted to be a piece of the preparation model. Case in point, consider a medicinal illustration where the elements from the blood work of distinctive patients are utilized to foresee a specific infection. Unmistakably, a component, for example, the Cholesterol level is prescient of coronary illness, while an element 1, for example, PSA level is not prescient of coronary illness. On the other hand, if a little preparing information set is utilized, the PSA level may have monstrosity relationships with coronary illness as a result of irregular varieties. While the effect of a solitary variable may be little, the combined impact of numerous superfluous components can be significantly. This will bring about a

preparation show that sums up inadequately to concealed test occurrences. There are two broad kinds of feature selection methods:

Filter Models: in this model, a crisp criterion on a single feature, or a subset of features, is used to evaluate their suitability for classification. This method is independent of the specific algorithm being used.

Wrapper Models: in this model, the feature selection process is embedded into a classification Algorithm, in order to make the feature selection process sensitive to the classification algorithm. This approach recognizes the fact that different algorithms may work better with different features.

Probabilistic Methods

Probabilistic techniques are the most central among all information classification strategies. Probabilistic classification calculations use factual surmising to find the best class for a given sample. Not with standing basically appointing the best class like other classification calculation each of the conceivable classes. The back likelihood is defined as the likelihood after watching the specific attributes of the test example. Then again, the former likelihood is basically the division of preparing records having a place with every specific class, with no information of the test occurrence. Subsequent to acquiring the back probabilities, we utilize choice hypothesis to focus class enrollment for each new example. Fundamentally, there are two courses in which we can gauge the back probabilities. In the first case, the back likelihood of a specific class is evaluated by deciding the class-contingent likelihood and the earlier class independently and after that applying Bayes' hypothesis to find the parameters. The most no doubt understood among these is the Bayes classifier, which is known as a generative model.

Decision Trees

Choice trees make a various leveled dividing of the information, which relates the distinctive parcels at the leaf level to the distinctive classes. The various leveled dividing at every level is made with the utilization of a part basis. The part measure might either utilize a condition (or predicate) on a solitary Characteristic or it may contain a condition on different characteristics. The previous is alluded to as a univariate part, while the recent is alluded to as a multivariate part. The general methodology is to attempt to recursively part the preparation information in order to augment the separation among the diverse classes over diverse hubs. The segregation among the distinctive classes is amplified, when the level of skew among the diverse classes in a given hub is augmented. A measure of entropy is utilized as a part of request.

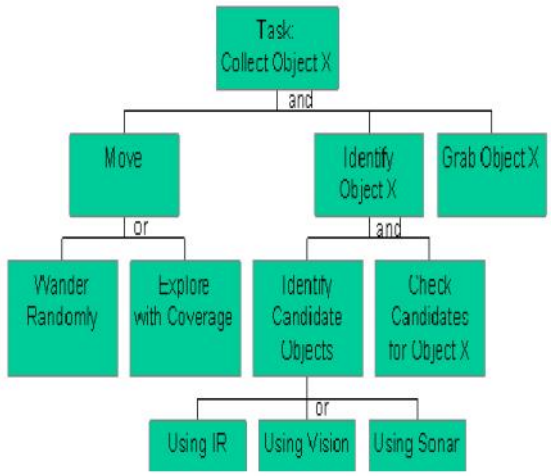

Figure 1: Simple Decision Tree

Rule-Based Methods

Rule-based methods are closely related to decision trees, except that they do not create a strict hierarchical partitioning of the training data. Rather, overlaps are allowed in order to create greater robustness for the training model. Any path in a decision tree may be interpreted as a rule, which assigns a test instance to a particular label. For example, for the case of the decision tree illustrated. It is possible to

create a set of disjoint rules from the different paths in the decision tree. Create related models for both decision tree construction and rule construction. Rule-based classifiers can be viewed as more general models than decision tree models. While decision trees require the induced rule sets to be non-overlapping, this is not the case for rule-based Classifiers. Clearly, second rule overlaps with the previous rule, and is also quite relevant to the prediction of a given test instance. In rule-based methods, a set of rules is mined from the training data in the first phase . During the testing phase, it is determined which rules are relevant to the test instance and the final result is based on a combination of the class values predicted by the different rules. test instance and the final result is based on a combination of the class values predicted by the different rules.

Instance-Based Learning

In instance-based learning, the first phase of constructing the training model is often dispensed with. The test instance is directly related to the training instances in order to create a classification model. Such methods are referred to as lazy learning methods, because they wait for knowledge of the test instance in order to create a locally optimized model, which is specific to the test instance. The advantage of such methods is that they can be directly tailored to the particular test instance, and can avoid the information loss associated with the incompleteness of any training model.

SVM Classifiers

SVM methods use linear conditions in order to separate out the classes from one another. The idea is to use a linear condition that separates the two classes from each other as well as possible. Consider the medical example discussed earlier, where the risk of cardiovascular disease is related to diagnostic features from patients.

2. PATTERN RECOGNITION TECHNIQUES

Pattern recognition methods are built on the assumption that some underlying characteristic of a protein sequence, or of protein structure, can be used to identify similar traits in related proteins. Conserved protein sequence regions are extremely useful for identifying and studying functionally and structurally important regions [9]. Sequence conservation of homologous Sequences is rarely Homogeneous along their length; as sequences diverge, their conservation is localized to specific regions. In order to obtain the general structural features of conserved regions of all proteins, it is necessary to decide the scale of protein clustering, conserved regions and structural features to analyze. Natural choices are generically defined protein families , ungapped protein sequence Motifs (blocks) That separate Proteins into Either conserved or random Signals and The four basic secondary structure elements Namely alpha helices, beta strands, structured turns, and loops. Protein sequence comparison has become one of the most powerful tools for characterizing protein sequences because of the enormous amount of information that is preserved throughout the evolutionary process. One of the early attempts to measure protein sequence comparison was Substitution matrices introduced by Day off .A general approach for functional characterization of unknown proteins is to infer protein functions based on sequence similarity. One of the successful approaches Is to define signatures of known Families of biologically related Proteins (typically at the functional or structural level). Signatures usually identify conserved regions among the family of proteins, revealing the importance for the function of their structural or physicochemical properties. A representative example of this approach is the well-known Prosite database, gathering

Protein sequence patterns and profiles for a large number of families.

3. THE EVALUATION METHODOLOGY

We evaluate the quality of the various alignment schemes by comparing the differences between the generated candidate alignment and the reference alignment generated from structural alignment programs. As a measure of alignment quality, we use the Cline Shift score (CS) to compare the reference alignments with the candidate alignments. The CS score is designed to penalize both under- and over-alignment and crediting the parts of the generated alignment that may be shifted by a few positions relative to the reference alignment. The CS score ranges from a small negative value to 1.0, and is symmetric in nature. We also assessed the performance on the standard Modeler's (precision) and Developer's (recall) score, but found similar trends to the CS score and hence do not report the results here.

3.1 Gap Modeling and Shift Parameters

For all the different scoring schemes, we use a local alignment framework with an affine gap model, and a zero-shift parameter to maintain the necessary requirements for a good optimal alignment. We optimize the gap modeling parameters (gap opening (go), gap extension (ge)),the zero shift value (zs), and weights on the individual scoring matrices for integrating them to obtain the highest quality alignments for each of the schemes. Having optimized the alignment parameters on the care dataset, we keep the alignment parameters unchanged for evaluation on the must ref dataset.

3.2 Optimization Performance

We also performed a sequence of experiments to evaluate the extent to which the two run-time optimization methods which can reduce the number of positions whose fRMSD needs to be estimated while still leading to high-quality alignments. which shows the CS scores obtained by the frmsd scoring scheme on the ceref dataset as a function of the percentage of the residue-pairs whose fRMSD scores were actually estimated. Also, the figure shows the average CS score achieved by the original (not sampled) frmsd scheme. These results show that both the seeded and iterative sampling procedures generate alignments close to the alignment generated from the original complete scheme. The average CS scores of the seeded and iterative sampling alignment by computing just 6% of the original frmsd matrix is 0.822 and 0.715, respectively. The average CS score of the original frmsd scheme is 0.828. Hence, we get competitive scores by our sampling procedures for almost a 20 fold speedup. The seeded based technique shows better performance compared to the iterative sampling technique.

CONCLUSION

In this paper we Surveyed the effectiveness of using estimated fRMSD scores to aid in the alignment of protein sequences. This approach of estimating the fragment-level RMSD is of similar spirit to learning a profile-profile scoring function to differentiate related and unrelated residue pairs using Classification and predictive analysis. Predictive analytics is using business intelligence data for forecasting and modeling. Proper data mining algorithms and predictive modeling can refine search for targeted customers. Predictive Analytics can aid in choosing marketing methods, and marketing more efficiently.

REFERENCES

[1] Bateman, A.; Birney, E.; Cerruti, L.; Durbin, R.; Etwiller, L.; Eddy, S.R.;Griffiths-Jones, S.; Howe, K.L.; Marshall, M.; Sonnhammer, E.L.L. The Pfam protein families database. Nucleic Acids Res.,**2002**, 30, 276-280.

[2] Corpet, F.; Servant, F.; Gouzy, J.; Kahn D. Tools for protein domain analysis And whole genome comparisons, Nucleic Acids Res., **2000**, 28,267-269.

[3] Falquet, L.; Pagni, M.; Bucher, P.; Hulo, N.; Sigrist, C.J.A.; Hofmann, K.; Bairoch, A. The PROSITE database, its status in 2002, Nucleic Acids Res., **2002**, 30, 235-238.

[4] Attwood, T.K.; Blythe, M.J.; Flower, D.R.; Gaulton, A.; Mabey, J.E.; Maudling, N.; McGregor, L.; Mitchell, A.L.; Moulton, G.; Paine, K.; and Scordis. P. PRINTS and PRINTS-S shed light onprotein ancestry., Nucleic Acids Res., **2002**, 30, 239-241.

[5] Lo Conte, L.; Brenner, S.E.; Hubbard, T.J.P.; Chothia, C.; and Murzin, A.G. SCOP database in 2002: refinements accommodate structural genomics, Nucleic Acids Res., **2002**, 30, 264-267.

[6] Pearl, F.M.G.; Martin, N.; Bray, J.E.; Buchan, D.W.A.; Harrison, A.P.; Lee, D.; Reeves, G.A.; Shepherd, A.J.; Sillitoe, I.; Todd, A.E.; Thornton, J.M.; Orengo, C.A. The CATH extended protein family database: providing structural annotations for genome sequences. Nucleic Acids Res., **2001**, 29, 223-227.

[7]. S. F. Altschul, L. T. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Research, 25(17):3389–402, 1997.

[8]. M. Cline, R. Hughey, and K. Karplus. Predicting reliable regions in protein sequence alignments. Bioinformatics, 18:306–314, 2002.

[9]. C. B. Do, S. S. Gross, and S. Batzoglou. Contralign: Discriminative training for protein sequence alignment. In Proceedings of the Tenth Annual International Conference on Computational Molecular Biology (RECOMB), 2006.

[10]. C. B. Do, M. S. P. Mahabashyam, and S. Batzoglou. Probcons: Probabilistic consistency-based multiple sequence alignment. Genome Research, 15:330–340, 2005.

[11]. R. Edgar and K. Sjolander. A comparison of scoring functions for protein sequence profile alignment. BIOINFORMATICS, 20(8):1301–1308, 2004.

[12]. A. Elofsson. A study on protein sequence alignment quality. PROTEINS:Structure, Function and Genetics, 46:330–339, 2002.