The Research of Protein Secondary Structure Prediction System Based on KDTICM

Bingru Yang, Wei Hou, Yonghong Xie, and Lijun Wang

Abstract—**Since 1960s, researchers have proposed several prediction methods, for Protein Secondary Structure Prediction, whereas the accuracies of them are no more than 80%. In this case, there is an urgent need to introduce a novel, high-accuracy prediction method. Based on the theory of KDTICM, we proposed a model, which is composed of four layers by intelligent interfaces, synthesizes several methods, such as SVM, KDD* and so on. From experiments, this model obtained Q3 accuracy 83.06%, 80.49% on RS126 and CB513, and for the proteins containing more alpha/beta structure; the Q3 accuracy obtained is 93.12%.**

Index Terms—**Data mining, KDTICM, Protein secondary structure prediction**

I. Introduction

Although since the mid-1960s, researchers have proposed several prediction methods, the accuracies of them are no more than 80% [1]. In this case, there is an urgent need to introduce a novel, high-accuracy prediction method. As data mining in dealing with massive data has unique advantages, many researchers have focused on the application of data mining in protein structure prediction, and achieved some fruits [2][3].

The existing protein secondary structure prediction methods can be broadly divided into machine learning-based methods, multi-sequence information methods, combine statistical and rules methods. These methods assume that protein secondary structure is determined by the reciprocity of near polypeptide chain, establish a suit of predict rules through analyze and induce the known protein molecules, and according these rules, they can predict other unknown protein secondary structure. Most of methods can obtain satisfied accuracy in a little part of protein, so it is hard to apply generally.

Looking backward of the prediction of protein secondary structure, we can see that, as forecast accuracy increasing, new revolutionary approaches improved. The reasons of why protein secondary structure prediction accuracy improved slowly for a long time are:

1) The lack of solid theoretical foundation for many of current methods. Most of current prediction methods stay in academic stage;

2) Expect for the homology of information, the vast majority of knowledge in the field has not been fully utilized;

3) The hybrid approach in the future development will become the mainstream, but it is still big challenges of how to combine different methods and make them work together effectively.

Aiming at these problems, we propose a new, gradually refining, multi-hierarchical configuration predict model—compound pyramid model. This model integrates methods of SVM, KDD*process model based on KDTICM [4]. Experiments proved that this model will obtain satisfying accuracy for partial to alpha/beta type protein secondary structure by intelligent interfaces and synthesizes several methods.

II. KDTICM (Knowledge Discovery Theory Based on Inner Cognitive Mechanism)

Act as a subject, data mining lack of uniform theory for its basic, this situation restrict the development of data mining for long-term. Jumping out from the mainstream development of KDD in 1997, we originally brought forward a new path to study KD from the perspective of inherent cognitive mechanism. Our idea is regarding the process of KD as a cognitive process and regarding the system of KD as a cognitive system. Our firstly found four principles implied by inherent cognitive mechanism of KD system, derived 8 new process model, 17 new technology methods; and our construct "knowledge discovery theory based on inner cognitive mechanism"(KDTICM) original first in the world in 2002.

A. Double bases cooperation mechanism

Based on "intention creation" and "psychology information restore" in cognitive psychology, we find the relationship of database and knowledge base under specific construction in the process of KDD; demonstrate the conformation mapping theorem; design the heuristic coordinator and the maintaining coordinator; resolve puzzles of "directional searching", "directional mining", independent discovery and real time maintenance.

Manuscript received June 18, 2009. This work was supported in part by the National Natural Science Foundation of China under Grant No. 69835001, No. 60875029, and the National Natural Science Foundation of Beijing under Grant No. 4022008.

Bingru Yang is with the Information Engineering School, University of Science and Technology Beijing, China, Zip 100083, e-mail: bryang_kd@yahoo.com.cn.

Wei Hou is with the Information Engineering School, University of Science and Technology Beijing, China.

Yonghong Xie is with the Information Engineering School, University of Science and Technology Beijing, China.

Lijun Wang is with the Information Engineering School, University of Science and Technology Beijing, China.

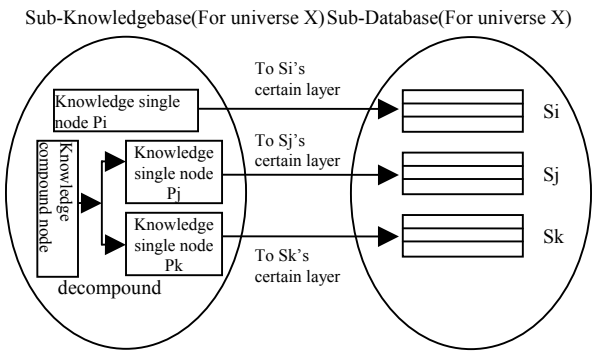

Figure 1 Knowledge single, compound node and their relations in Knowledgebase

(1) Conformation mapping theorem: There is an equivalent relations between the inferential category of universe X, Cr(N) and complete data substructure reachable category $C\propto<\gamma, \Re c(\gamma)>$.

(2) This theorem establishes the one-to-one correspondence between knowledge single node and "data substructure" in database (As shown in figure 1).Double bases cooperation mechanism resolves fundamentally the problem of "directional searching" and "directional mining". Supposed and realized two cooperated algorithms: firstly, real time maintenance for domain knowledge base through maintenance cooperate algorithm and component; secondly, find knowledge shortage independently to produce intention creation through the heuristic coordinate algorithm and component.

(3) Further discussion of double basis cooperation mechanism. For example: (probability estimation theorem of accessible relation): Suppose $p>2\alpha+\alpha 2/(1-\alpha)$; β and B are the parameters defined in definition and $\alpha<\beta<(1-\alpha)p$, $(1-p+p\alpha)/(1-\alpha)<B<1-\alpha$. Each positive rule in primitive knowledge base corresponds to a relation in data sub-class structure base. Therefore, along with the increase of tuples (S(R)) in database $\Re(X)$ of universe of discourse X, the probability of accessibility relation in the data sub-class structure base tends to be one; the probability for inaccessibility relation, which corresponds to each negative rule, also tends to be one.

B. KDD--new process model derive from double cooperation mechanism*

We fuse double basis cooperation mechanism and construction of two cooperators into classical KDD process, form the new process model- KDD* independently, then change the old knowledge discovery process essentially.

(1) Lack of original knowledge discovery process model KDD in technology and function:

1) Domain knowledge cannot step in the process of data mining (knowledge discovery) substantially.

2) System cannot mine the knowledge shortage independently.

3) To make sure the direction of miming and focus only according to user's interests, will bring on a great deal of repeated, redundant rules; it cannot match the knowledge shortage of itself.

4) Cannot maintain knowledge base dynamically on real time.

5) Implement of model is on the base of semantic.

(2) Aiming at shortage above mentioned, KDD* process model given the innovated method and implement technology, details as followed:

1) During the process of mining, domain knowledge step in the mining process directly through two cooperators, the idea is derived from synchronization evolution and cooperation computation.

2) System can produce directional focus through adjacency matrix of directed hyper graph, and mining knowledge shortage independently.

3) Focus problem: direction and process of directional mining will only produce when user's interests match knowledge shortage system find independently. So it will not mine a great deal of repeated, redundant knowledge, decrease rule evaluation greatly. The purpose is to decrease search space, improve the algorithm efficiency, and provide necessary technology support.

4) With the accumulation of knowledge, the knowledge base of knowledge will be more and more, in order to reflect for application quickly, the maintenance coordinator added to the new model, process the repetition, redundancy, confliction, circle and hypostasis effectively, dynamically on real time.

5) The new model is based on the two cognitive features --"intention creation" and "psychology information restore" in cognitive psychology, so the new model has its solid theory base; and the implement of this model is on the base of theory.

C. Associated analysis algorithm

Based on KDD* process model, we supposed a new association analysis algorithm, Maradbcm(for short, we call it as M algorithm)[4].

Process as followed:

Input: Rule strength threshold Min_Intensity, support threshold Min_Sup, confidence threshold Min_Con;
Output: Association rule base KD.
1. Data preprocess;
2. When "shortage of knowledge" is detected
3. Create K_2;//K_m denotes the shortage knowledge whose length is m, namely $K_m=\{r|\ Len(r)=m\}$.
4. m = 2;
5. Create hypothesis of knowledge K_m; // Directional mining the shortage of knowledge r_i in K_m.
6. Repeat
7. For every r_i in K_m
8. If (r_i is conformed with present knowledgebase && the measure of r_i is qualified)
9. move r_i into KD, update reachable matrix;
10. Else
11. delete r_i;
12. Endfor;
13. m = m+1;
14. Until $K_m=\phi$;
15. EndWhen;

Relative to KDD, KDD* fuses the KDD and double bases cooperation mechanism, so that it is a novel knowledge discovery process model, and there are several features of it:

1) KDD* organically integrates and fuses the new knowledge mined by KDD* and the inhere knowledge in basic knowledgebase;

2) In the process of KDD, KDD* relieves the complexity

of data accumulation, meanwhile provides the prior condition for the fusion of new, inhere knowledge;

3) Double bases cooperation mechanism, in itself, is capable of evolving as the structure's changing;

4) KDD* changes and optimizes the process and mechanism of knowledge discovery, realizes "multi-origin" focus, and diminishes the workload of assessment;

5) From the aspect of cognitive science, KDD* enhances and upgrades the intelligence of KDD, and improves its ability of cognitive activeness;

6) Double bases cooperation mechanism, reveals the relation between the sub knowledgebase and data substructure, under certain principle of construction of bases;

7) Double bases cooperation mechanism and KDD* model that is induced from former, derives a novel algorithm, Maradbcm, which is more expansible and effective relatively to prevailing algorithms.

III. Compound pyramid model

For non-trivial problems, such as protein secondary structure prediction, the general single-method models and simple combinations of prediction models, could not obtain satisfied prediction results. The compound pyramid model adopts gradually refining, multi-hierarchical configuration, in which the layers focus on independent functions, so that this model get the higher prediction accuracy comparatively. Its configuration is shown in Fig. 2.

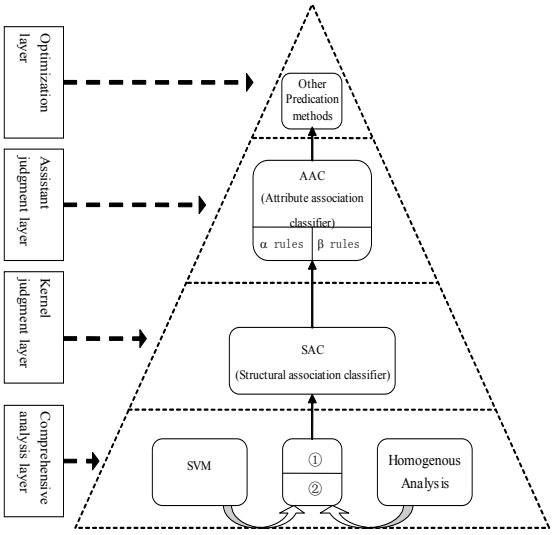

Figure 2 compound pyramid model

The compound pyramid model is composed of 4 layers, whose functions are independent, and the layers work in close coordination.

A. Comprehensive analysis layer

This layer is the basic layer of the whole model, integrate homologous analyze and optimized SVM, and can complete the prediction of feature clearly amino acids more than 50%.

The homologous sequence method is widely used in the field of the protein secondary structure prediction. In our experiment, we use Apssp2 which is a sophisticated method of protein secondary structure prediction technique based on the multi-sequence alignment, and its basic steps are as follows:

1) Using standard neural network technology and multi-sequence alignment result produced by the PSIBLAST algorithm; the input of the standard neural network is a window, and the output is H, E, C (three kinds of secondary structures).

2) Using improved EBL (Example based learning) technical to predict protein secondary structure, and improved EBL is the key technology.

3) Predicting protein's secondary structure according to the former two steps and their combination can be realized through the confidence score.

Among them, to realize EBL, firstly, we must choose those proteins in the PDB database whose resolution degree is more than 2.8A and length is bigger than 50. Then, we use the DSSP database to endow with the secondary structure to the protein to obtain some patterns, the length of each pattern is 17 residues (they are in a window, and the length of window is 17, to indicate residue's secondary structure in the window center). We can obtain more than 6,000,000 patterns, but excluding the redundancy, there are only 1,300,000 non-redundant patterns left. Then, we train EBL method in these 1,300,000 non-redundant patterns. The standard EBL method will come with the speed question when it is carried on the training in these 1,300,000 non-redundant patterns. According to three central residues (central residue, left side, right flank), it will be divided into 8000 sets. There will be 8000 matrices in all by generating a distance matrix for every set, and then use these 8000 matrices to replace 1 matrix like this.

The protein secondary structure has big data dimension and quantity, is partial to alpha/beta type characteristic specifically in experiment, SVM three classification implement accuracy is difficult to be ensured that we have made several improvements in the following aspects.

SVM is designed for small sample space, but the data amount for training and testing without exception is very big. In our experiment we adopt new study tactics. Firstly, we select a small-scale sample set from training set, use this minor sample set to train and get an initial classification implement, use this classification to prune training set in large scale getting a small-scale reduction set. At last, we get the ultimate classifier by training with the reduction set.

We use "rotation" method to resolve the problem of three classifiers. Be constructing six classifiers such as H/ ~ H , E/ ~ E , C/ ~ C , H/E , H/C and E/C, we firstly calculate distance from H/ ~ H , E/ ~ E , C/ ~ C to samples, then elect the maximum one followed by entering the next layer.

The protein secondary structure is unbalanced in data distribution, usually alpha occupies majority, and we have realized the penalty factor so as to eliminate the problem of unbalanced data.

Moreover, we use the dichotomy method to train the most superior parameter, solving the problem of inefficiency of manual test parameter, and carry on a quick seek for the most superior parameter.

SVM method is aimed at amino acid physical and chemical properties, homologous analyze is based on sequence structure, so this layer integrate analyze results of phy-chemical properties and structure sequence. Because homologous analyze using neural network to build predict

model, form aspect of methodology, this layer integrate methods of SVM and neural network.

B. Kernel judgment layer

This layer use association algorithm based on KDD*process model to construct SAC (Structural association classifier) module which takes on the classification of data that is hard to judge in compound layer and after the classification via SVM. The theory bases on the correlative effects among protein secondary structures. In other words, the conformation affects information among secondary structure. The basis of the core theory is KDTICM established by our research institute, and the tool is the KDD* system which can mine the highly precise association rules with good fitness from the training data. The association rules described the effects between secondary structures. SAC module is constructed by the use of these association rules and the improved CBA algorithm.

SAC is based on slide window, we set it to 13; and the basis of prediction is the association rules which training data use amino acids secondary structure in the window to mining rules. According to the proper classify, SAC function will obtain higher prediction accuracy.

C. Optimization judgment layer

The core of this layer is AAC (Attribute association classifier) module. Through the association analysis for the physical and chemical properties of the amino acid, refinement rule base is created to predict the lower layers of data that is not determinable.

The main tool is also the KDD* system based on the original theory KDTICM and the improved CBA algorithm of our institute. Through the alpha, beta rules generated by the KDD* the system, the refined alpha, beta rule bases are obtained after a certain degree of reduction. By using the CBA algorithm, the threshold of support level and confidence level applied to the alpha, beta is received through the repeated experiments. The results are reliable through the experimental verification. It is proved that the alpha, beta rule bases and the threshold of support level and confidence level we obtained can be fixed as knowledge and embedded in the integral pyramid model.

In the use of the improved CBA algorithm, we also set the threshold of support level uncommonly. And we use the accumulation according to the confidence level as the criteria for alpha, beta. We don't use the confidence level as a single measure, but regard the distance between support level and confidence level as a composite measure. It may preferably reflect the action of the two main indicators in protein secondary structure prediction.

According to the characteristics of protein bio-data, we also break the convention, and abandon the use of a common database in the process for generating rules. We use the protein database of which the contents are relatively biased toward alpha, beta-protein respectively at a high starting point. Therefore, the long-standing problems that the support level and confidence level of the generating rules exit in the protein mining association rules are low and the rules are not reliable is solved. There is a relatively increase of the accuracy compared to the former common protein database proved by experiment.

D. Optimization layer

This layer mainly designs 3 methods of tendency factor, potential function and reasonable inference. The first two kinds methods are attribute to bioinformatics methods, these methods predict the structure using bioinformatics background, reasonable inference method is on the base of the physical and chemical properties rules which every kinds of secondary structure behaves, these three kinds of method optimize the results of three layer, then it can improve the whole predict accuracy.

Noteworthiness, compound pyramid model use Causal Cellular Automata[5] based on knowledge discovery finally choose the hydrogen bond, hydrophobia and electricity as considered properties to reduce the complexity of this model, by analyzing the associational relations among the phy-chemical properties such as hydrogen bond, carbon circle, hydrophobia, electricity, residue size, fat and electric quantity and so on.

IV. Experiment

We select ILP (a dataset containing more alpha/beta structure), RS126, and CB513 datasets to experiment. At the same time, we use standard of Q3 to value it, Q3 is the ratio of correct amino acids and total amino acids which predicted. Experimental results are shown as table 1-3.

While the module in the Compound Pyramid Model are independent in function, the model is gradually enhanced, multi-layer systematic. Based on Comprehensive analysis layer, the model can predict most unpredicted amino acid with obvious feature, and the accuracy is high up to 90%, which assure the accuracy of the whole model. The second layer and the third one are most important. Comprehensive analysis layer only can predict part of amino acid effectively, moreover, the feature in structure are most apparent such that the amino acid in the Kernel judgment layer and the assistant one scaly own obvious structure feature. As a result, more sophisticated methods are required. The Kernel judgment layer is most important and it mainly enhances the unsure result in the Optimization layer to ensure the unpredicted amino acid with the lest Coil to sent to Assistant judgment layer. Although the number of amino acid is less in this layer, its structure feature is not apparent and it adopts sequence alignment and homologous analysis, which render it difficult to get a better result. The Compound Pyramid Model builds another road. It uses KDD* to mine α/β base. Because α/β base adopt high purified method and KDD* can mine unexpected rules, the association rules set obtained is essential, i.e., it is different from the common result. This rules set with high support and reliability build the solid foundation for the association rules classification. Assistant judgment layer will solve most of unpredicted amino acid through double verification of AAC module and improved CBA method. Optimization layer is the assistant function on the compound pyramid model, while these three methods is traditional, and predict accuracy is low when use it single, but we use tendency factor, potential function and reasonable inference methods, we find we can obtain more higher

accuracy when we use these methods modify this predict result, this is because of cohesion knowledge is hard to obtain by other methods, so it can complement each other to other three layers.

Table 1 predict result of ILP

module	Accuracy	percent
Comprehensive Analysis Layer	344/352=97.73%	352/523= 67.3 %
Kernel judgment layer	142/171=83.04%	171/523= 32.7%
Assistant judgment layer	1/2=50%	2/523=0.38%
Total	487/523= 93.12%	

Table 2 predict result of RS126

module	Accuracy	percent
Comprehensive Analysis Layer	12765/13981=91.3%	13981/24806 = 56.36%
Kernel judgment layer	7786/10552=73.79%	10552/24806 = 42.54%
Assistant judgment layer	54/273=19.78%	54/24806=0.22%
Total	20605/24806 = 83.06%	

Table 3 predict result of CB513

module	Accuracy	percent
Comprehensive Analysis Layer	73145/80530 = 90.83%	80530/146233 = 55.07%
Kernel judgment layer	44240/64534=68.55%	64534/146233 = 44.13%
Assistant judgment layer	325/1169=27.8%	1169/146233=0.8%
Total	117710/146233 = 80.49%	

Predict accuracy compare between the predict system based on compound pyramid model and typical research literatures related international.

We conclude that the predict system based on Compound pyramid model has innovations as followed:

1) There is no comparability between our research achievements and others' reported in literature ahead. Because our achievements are independent, the models and algorithms involved in predict system are all original or improved, we doesn't make use of any output result provided by outside server, but other research fruits put forward by literature make use of the output by outside server (some are entirely, some are partly) , and they only integrate and optimize them. Still, our predict precision exceed others' reported in formal publication of international (use Q3 method).

2) This research is belong to prediction on secondary structure of protein, which is not only researching one predicting method, it contains many core components such as system model, system method and system optimisation. Up to now, we have obtained breakthrough in science idea, technological method and methodology.

3) Form aspect of system model, there isn't any other researcher using composing pyramid model, and no researcher combining physical attribute determinant and structure sequence determinant to form method, we firstly use KDTICM theory to construct associate classifier.

4) Form aspect of methodology, we form system methods, which are formed by many relevant predicting methods. Among those methods, some are original, such as KDD* model and M-algorithm based on KDTICM theory; some are improved methods, such as SVM apperception analysis. Especially for causal cellular automata theories, this can optimize physical attributes and improve the predicting accuracy.

5) Form aspect of system optimization, in the deduction of system model, granularity space of every layer is becoming thinner, which shows perforation of field knowledge and background knowledge, and can ensure the prediction accuracy.

According to the result mentioned above, we conclude that the "Composed Pyramid model" predict system adding the SAC module and the main thread of innovative DM technology can not only apply to protein secondary structure prediction in RS126 and CB513 data sets, but get very good classification effect. Predict results: we achieved 93.12% Q3 accuracy on ILP dataset(maximal 81% we have known in the world); 83.06% Q3 accuracy on RS126 dataset(maximal 81.65% we have known in the world); 80.49% Q3 accuracy on CB513 dataset(maximal 78.44% we have known in the world). Our research result is the highest in the world. We can improve the prediction accuracy in optimization layer until this achievement.

In bioinformatics, it is a fellow felling, that is when the accuracy of protein secondary structure prediction surpasses 80%, its tertiary structure can be predicted generally. As a result, we will get breakthrough in protein tertiary structure prediction based on the predict system we used.

V. Conclusion

A new, gradually refining, multi-hierarchical configuration prediction model for protein secondary structure, Compound Pyramid Model, is proposed in this paper. This model obtained better results in several classic protein sets. The protein 3D structure prediction based on this model is one of the most hopeful works in the future.

References

[1] X. Wu, L. Jain, J. Wang et al., *Data Mining in Bioinformatics*. Berlin: Springer, 2005, ch. 1.

[2] A. Haoudi, H. Bensmail, "Bioinformatics and data mining in proteomics(Book style)," *Expert Review of Proteomics*, vol.3, 2006, pp. 333-343.

[3] J. Y. Li, L. S. Wong, Q. Yang, "Data mining in Bioinformatics (Periodical style)," *IEEE Intelligent Systems*, vol. 20, 2005, pp. 16-18.

[4] B. Yang, *Knowledge discovery based on theory of inner cognition mechanism and application*, Beijing: Electronic Industry Press, 2004.

[5] B.Yang, X. Li, W. Song, "Generalized Causal Inductive Reasoning Model Based on Generalized Causal Cellular Automata (Conference style)," In: Proc. of *ICNN&B'05*, 2005, pp. 375-378.