Rule Generation Using NN and GA for SARS-CoV Cleavage Site Prediction

Yeon-Jin Cho and Hyeoncheol Kim*

Department of Computer Science Education,
Korea University, Seoul, 136-701, Korea
{jx,hkim}@comedu.korea.ac.kr

Abstract. Cleavage site prediction is an important issue in molecular biology. We present a new method that generates prediction rules for SARS-CoV protease cleavage sites. Our method includes rule extraction from a trained neural network and then enhancing the extracted rules by genetic evolution to improve its quality. Experimental results show that the method could generate new rules for cleavage site prediction, which are more general and accurate than consensus patterns.

1 Introduction

Strong interest in automated identification and prediction of cleavage sites have been evoked not only by the huge amount of unprocessed data available but also by the commercial need. The identification and prediction problem is domain-specific, and machine learning methods such as neural networks have therefore been widely used and been successful. In this paper, we present new approaches to rule generation for the cleavage site prediction, and the rule is represented in an explicit form such as "if L@p2 and R@p3, then cleavage". Prediction of SARS-CoV protease cleavage site was selected as a subject of study.

The first cases of severe acute respiratory syndrome (SARS) were identified in China in November, 2002 and have spread many countries around the world [17]. By late June 2003, the World Health Organization (WHO) has recorded more than 8000 cases of SARS and more than 750 SARS related deaths, and a global alert for the illness was issued due to the severity of the disease [20]. Corona virus (CoV) family consists of four groups as illustrated in figure 1. SARS-CoV is a mutant virus of CoV and belongs to the Group 4. One of the ways to make SARS-CoV incapable is to obstruct the increase of the virus by constraining the activity of proteinase (3CLpro), which is one of the core viruses of SARS-CoV. Therefore, the analysis of CoV cleavage site in the other three groups makes it possible to predict the cleavage site of SARS-CoV and tackle the deseases caused by other upcoming CoV mutant viruses.

SARS-CoV can be incapacitated by restraining the activity of protease (3CL^{pro}), which is one of the main proteins consisting SARS-CoV [11]. Several researches indicate that virus proliferation can be arrested using specific proteinase inhibitors.

Consequently, the protease inhibitors that restrains the virus proliferate can be found by analysing the cleavage site which cleaved by protease, a main protein, in Corona virus. This fact motivated the computational research on the cleavage site

* This work was supported by a Korea University Grant K0400911

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 785–791, 2005.

prediction and analysis. It reduces time and expense of processes of the pathology experiments. The cleavage site analysis enables us to recognize the cleavage candidates and to design the inhibitors of protease [4], [10], [11]. It will assist the cure for SARS and for other diseases caused by Corona viruses.

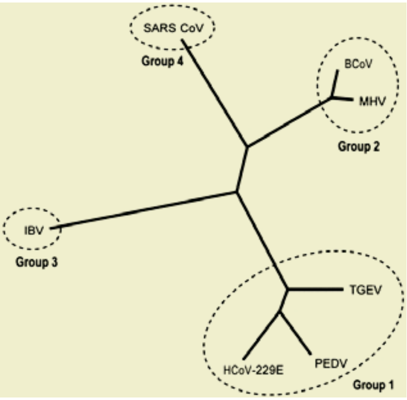

Fig. 1. SARS Corona virus and six other Corona viruses[1]

Machine learning approaches including neural networks have been applied to cleavage site analysis successfully [1], [3], [4], [8], [15]. They however focused on identification or classification of cleavage sites, but not on the explanation of the classification. Prediction knowledge or rule in an explicit format will help us to understand how the sites are classified and provide us with better insights about the specific domain. The idea of prediction rule is not new since consensus patterns and decision trees were used among biologists. However, the consensus patterns are not complete and not accurate enough. Our goal in this paper is to present accurate and robust methods to generate prediction rules of good quality. We used the methods of rule extraction from neural networks and knowledge-based genetic algorithms in this paper.

2 Materials and Methods

In the search for potential inhibitors, important issue is to predict which peptides can be cleaved by SARS-CoV protease. Even limited in the range of an octapeptide, experimental test would be very expensive because the number of possible octapeptides formed from 20 amino acids runs into $20^8 = 2.56\times10^{10}$. Thus, computational methods for cleavage site prediction would be very useful.

2.1 Data Set Preparation

Twenty-four genomic sequences of coronavirus and the annotation information were downloaded from the GenBank database [2], of which 12 are SARS-CoV and 12 are

[1] Marra Ma et al.: The Genome Sequence of the SARS-Associated Coronavirus. SCIENCE VOL 300 (2003)

other groups of coronaviruses. Each sequence of coronavirus genome includes 11 cleavage sites and thus total 264 (= 24×11) sites are available. We eliminated duplicated ones out of the total 264 results and identified final 70 cleavage sites. Each cleavage site of octapeptides includes 8 regions (i.e., 8 positions of P4, P3, P2, P1, P1', P2', P3', P4'). The position p1 is just before the cleavage site; p4 through p1 is N-terminal to the cleavage site and p1' through p4' is C-terminal to the cleavage site. Each region represents one of the 20 amino acids.

For a classification problem such as cleavage site classification, both positive and negative examples are needed. It searches or induces rules that cover positive examples as much as possible and negative examples as little as possible. Negative examples (presumed non-cleavage sites) are created by defining all other Glutamines (Q) in the viral polyproteins as non-cleavable sites [11]. Therefore we obtained 70 positive (i.e., cleavage) and 1267 negative (i.e., non-cleavage) examples in our experimental dataset. Since every site in our dataset has a glutamine (Q) in position P1 (the position just before the cleavage site), the position p1 does not play any role in our classification experiments and thus the symbol "Q@p1" (i.e., amino acid 'Q' at position p1) is ignored in the rules generated in our experiments.

For neural network training, each region value that is one of 20 amino acids is converted into 20 binary digits. For example, Alanine(A) among 20 amino acids is represented by 20 bits of 10000000000000000000. Thus, each cleavage sites (i.e., octapeptides) composed of 8 regions is encoded into 160 bits. Class is encoded into either 1 (i.e., cleavage) or 0 (i.e., non-cleavage).

2.2 Methods

Analysis of the Cleavage Sites: Sequence Logo and Decision Tree

Amino acid conservation in multiple sequence alignments may be visualized using sequence logo. Sequence logo is useful for a quick examination of the range in which a sequence signal is present. From the sequence logo in Figure 2, a very strong consensus is evident around the cleavage site. Three consensus patterns from the sequence logo are 'LQ', 'LQ[S/A]' and '[T/S/A]X[L/F]Q[S/A/G]' [11].

Decision tree is one of the best-known classification techniques in symbolic machine learning. We used C5.0 algorithm and generated the following: "if L@p2 ^ [A/C/G/N/S]@p1', then cleavage".

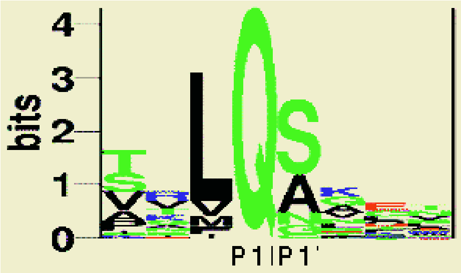

Fig. 2. The sequence logo of SARS-CoV cleavage sites. P1 = N-terminal to cleavage site, P1'= C-terminal to cleavage site

Feed-Forward Neural Network and Rule Extraction

Kiemer, *et al.* used feed-forward neural networks for SARS-CoV cleavage site analysis [11]. They showed that the neural network outperforms three consensus patterns in terms of classification performance. However, they used the neural network for just cleavage site prediction, but not for expressing the sites in explicit knowledge.

There have been many studies for efficient extraction of valid and general rules from a trained neural network [1], [6], [7], [8], [9], [12], [16], [18], [19]. In this paper, we used the OAS (Ordered-Attribute Search) algorithm to extract *if-then* rules from the neural networks [12].

Genetic Algorithm and Knowledge-Based Genetic Algorithm

GA (Genetic algorithm) can be used for searching generalized rules [5], [13], [14]. Individual chromosome in a GA population is a sequence of 8 symbols in which each symbol represents an amino acid or `*' (i.e., don't_care symbol). Then we can say that, for example, the chromosome [**LQS***] represents the rule "If L@p2 and S@p1', then cleavage". We ignore the Q at p1 as mentioned before. Therefore the size of rule space is as huge as 21^8. The GA-based model searches for the best fitted set of chromosomes (i.e., rules) among the 21^8 candidates. One-point crossover is used and crossing point is selected randomly. Mutation occurs on each symbol by 1% and changes its symbol to one of other 20 symbols. Fitness function for a chromosome *n* is defined as follows.

$$f(n) = \frac{nt}{nt + nf + 1} \times 100 + d$$

where *nt (or nf)* is the number of positive (or negative) instances matched by the chromosome rule and *d* is the number of *s (i.e., don't_care symbols) in the chromosome.

The GA-based model generates rules, but the performance is not good enough. The performance was very sensitive to the initial population of chromosomes which was generated by random. Knowledge-based approach to the GA-model is used to restrict the random search space. Domain-knowledge was used as an initial population and GA-model refines and explores from the initial rules. The knowledge-based approach also reduces the GA learning time significantly because it restricts GA search space.

3 Experimental Results

Our experiment includes the following steps; (1) Rule extraction from a trained neural network and comparison of the rules with consensus patterns and decision trees; (2) Rule generation by GA-based model with initial knowledge from consensus patterns, decision tree and neural network rules; (3) Comparison of the rules from neural networks, decision tree and knowledge-based GA. A rule is in the form of "IF condition, THEN class" where class is either of cleavage or non-cleavage. Performance of a rule is evaluated by its coverage and accuracy defined as follows:

$$coverage = \frac{\text{\# of examples matched by the condition part}}{\text{Total \# of examples}}$$

$$accuracy = \frac{\text{\# of true positive examples}}{\text{\# of examples matched by the condition part}}$$

A feed-forward neural network was configured with 160 input nodes, 2 hidden nodes and 1 output node. The neural network was trained and tested by 3-fold cross-validation. Generalization of the neural network is as high as 97.9% while training accuracy is 99.6%. Then we extracted if-then rules from the trained neural network by Kim's OAS algorithm [12], and compared them with consensus rules and decision tree rules.

Next, GA-based model was used with domain knowledge incorporated initially. Domain knowledge was obtained by extracting rules from consensus patterns, decision tree and neural networks. Our experiment shows that the GA rules from neural network knowledge outperform others in terms of the number of quality rules.

Finally, we compare the rule performances between decision tree, neural network and knowledge-based genetic algorithm (KBGA). Table 1 lists the rules with coverage greater than 17% and accuracy greater than 60%. Decision Tree generates rules no better than consensus patterns, while neural network generates four other useful rules in addition to the decision tree rules. The neural network rules were enhanced by GA evolution which was initialized with the neural network rules. The new rule, [S@p4 ^ S@p1'], discovered only by the KBGA is of good quality with accuracy 86.6% and the rule was not discovered by any other methods.

Table 1. The performance of the rules extracted from each classifier algorithm (standard coverage: +17%, standard accuracy: +60%)

	Positive Rules	Coverage(%)	Accuracy(%)
Consensus rules	L@p2 ^ S@p1'	36.35	75.76
	L@p2 ^ A@p1'	26.11	78.26
DT(C5.0) rule	L@p2 ^ S@p1'	36.35	75.76
	L@p2 ^ A@p1'	26.11	78.26
NN rules	L@p2 ^ S@p1'	36.35	75.76
	L@p2 ^ A@p1'	26.11	78.26
	L@p2 ^ E@p3'	21.9	71.43
	V@p4 ^ L@p2	20.71	60.87
	T@p4 ^ L@p2	20.32	77.78
	R@p3 ^ L@p2	17.3	85.71
KBGA rules initialized by NN rules	L@p2 ^ S@p1'	36.35	75.76
	L@p2 ^ A@p1'	26.11	78.26
	L@p2 ^ E@p3'	21.9	71.43
	V@p4 ^ L@p2	20.71	60.87
	T@p4 ^ L@p2	20.32	77.78
	R@p3 ^ L@p2	17.3	85.71
	S@p4 ^ S@p1'	18.73	86.67

4 Conclusion

Prediction or classification rules provide us with explanation about the classification and thus better insights about a domain. We presented a new method that generates rules and improves quality of the rules with the subject of SARS-CoV protease cleavage site prediction. Rules were extracted from a well-trained neural network and then enhanced by genetic evolution. Our experiment presents the rules generated by four different types of approaches:

- Consensus patterns
- Decision Tree
- Neural networks
- Genetic Algorithm initialized by neural network rules

Neural network could generate the rules of high quality that were not discovered by decision trees or consensus patterns. Knowledge-Based Genetic Algorithm (KBGA) model in which the neural network rules were incorporated initially could discover new rules in addition to the neural network rules. The KBGA can be considered as a hybrid model of neural networks and genetic algorithm since knowledge learned by a neural network is enhanced and expanded by GA evolution. The experimental result demonstrates that the hybrid model improves quality of rule generation.

References

1. Andrews, Robert, Diederich, Joachim, Tickle, Alam B.: Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-Based Systems 8(6) (1995) 373-389
2. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL.: GenBank: update. Nucleic Acids Res, 32 Database issue: (2004)D23-26
3. Blom N, Hansen J, Blaas D, Brunak S.: Cleavage site analysis in picornaviral polyproteins: discovering cellular targets by neural networks. Protein Sci (1996) 5:2203-2216
4. Chen LL, Ou HY, Zhang R, Zhang CT.: ZCURVE-CoV: a new system to recognize protein coding genes in coronavirus genomes, and its applications in analyzing SARS-CoV genomes. SCIENCE DIRECT, BBRC (2003) 382-388
5. De Jong, K.A. and Spears, W.M.: Learning Concept Classification Rules Using Genetic Algorithms. Proceedings of the I Zth. international Conference on Artificial Intelligence (1991) 651-656
6. Fu, LiMin.: Neural Networks in Computer Intelligence. McGraw Hill, Inc (1994)
7. Fu, LiMin.: Rule generation from neural networks. IEEE Transactions on Systems, Man, and Cybernetics 24(8) (1994) 1114-1124
8. Fu, LiMin.: Introduction to knowledge-based neural networks. Knowledge-Based Systems 8(6) (1995) 299-300
9. Fu, LiMin and Kim, Hyeoncheol.: Abstraction and Representation of Hidden Knowledge in an Adapted Neural Network. unpublished, CISE, University of Florida (1994)
10. Gaoa F, Oua HY, Chena LL, Zhenga WX, Zhanga CT.: Prediction of proteinase cleavage sites in polyproteins of coronaviruses and its applications in analyzing SARS-CoV genomes. FEBS Letters 553 (2003) 451-456
11. Kiemer L, Lund O, Brunak S, Blom N.: Coronavirus 3CL-pro proteinase cleavage sites: Possible relevance to SARS virus pathology. BMC Bioinformatics (2004)
12. Kim, Hyeoncheol.: Computationally Efficient Heuristics for If-Then Rule Extraction from Feed-Forward Neural Networks. Lecture Notes in Artificial Intelligence, Vol. 1967 (2000) 170-182
13. Michalewicz, Z.: Genetic Algorithms + Data Structures = Evolution Programs. 3rd edn. Springer-Verlag, Berlin Heidelberg New York (1996)
14. Mitchell, M.: An Introduction to Genetic Algorithms. MIT Press (1996)

15. Narayanan, A., Wu, X., Yang, Z.R.: Mining viral protease data to extract cleavage knowledge. bioinformatics, 18(1) (2002) s5-s13.
16. Setino, Rudy, Liu, Huan: Understanding neural networks via rule extraction. Proceedings of the 14th International Conference on Neural Networks. (1) Montreal, Canada (1995) 480-485
17. Stadler K, Masignani V, Eickmann M, Becker S, Abrignani S, Klenk HD, Rappuoli R.: SARS - BEGINNING TO UNDERSTAND A NEW VIRUS. NATURE REVIEWS, MICROBIOLOGY VOLUME 1 (2003) 209-218
18. Taha, Ismali A. and Ghosh, Joydeep: Symbolic interpretation of artificial neural networks. IEEE Transactions on Knowledge and Data Engineering 11(3) (1999) 443-463
19. Towell, Geoffrey G. and Shavlik, Jude W.: Extracting refined rules from knowledgebased neural networks. Machine Learning 13(1) (1993)
20. Yap YL, Zhang XW, Danchin A.: Relationship of SARS-CoV to other pathogenic RNA viruses explored by tetranucleotide usage profiling. BMC Bioinformatics (2003)