Protein Families Classification using Support Vector Machine

Joo Chuan Tong[1], Khar Heng Choo[2], Teck Kwong Lee[3], Lesheng Kong[4], Soon Heng Tan[5], Tin Wee Tan[6], Shoba Ranganathan[7]

Keywords: Support Vector Machines, Transporter Protein Families, Protein Structure

1 Introduction.

Proteins play an important role in biological processes and detailed knowledge about protein functions is fundamental to understand the complex biological pathways that occur in living organisms. A large number of sequence information has been experimentally determined and deposited in numerous databases. However, only a small fraction of protein sequences have been experimentally characterized. In this context, theoretical prediction of protein functions is becoming critically important in furthering our understanding of biological processes. In recent years, several groups have adopted the use of SVM as a prediction tool for protein functional family classification. This study investigates the 11 amino acid attributes (surface tension, hydrophobicity, normalized Van der Waals volume, relative mutability, polarity, polarizability, charge, bulkiness, solvent accessibility, predicted secondary structure and predicted trans-membrane region) to the accuracy of transporter protein family classification using the leave-one-property-out procedure.

2 Methods

Our dataset currently covers 81 protein functional families from the class 2A super family of TCDB [6]. These include 774 Swiss-Prot protein records from TCDB, 1065 Swiss-Prot records collected from Pfam, and 9764 NCBI's non-redundant protein database records. The program SVM^{light} [3] was adopted in this study for SVM calculations. Secondary structure assignment was performed using PSIPRED [4] and trans-membrane assignment was achieved using MEMSAT [5].

3 Results

[1] Department of Biochemistry, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260. Email: victor@bic.nus.edu.sg
[2] Department of Biochemistry, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260. Email: justin@bic.nus.edu.sg
[3] Department of Biochemistry, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260. Email: bernett@bic.nus.edu.sg
[4] Department of Biochemistry, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260. Email: lesheng@bic.nus.edu.sg
[5] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613. Email: soonheng@i2r.a-star.edu.sg
[6] Department of Biochemistry, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260. Email: tinwee@bic.nus.edu.sg
[7] Research Institute for Biotechnology, Macquarie University, NSW 2109, Australia. Email: shoba@bic.nus.edu.sg

A three-fold cross validation was adopted to investigate the contribution of each amino acid property on the accuracy of the classifier. From the amino acid properties used in our current study, 11 iterations were performed using the leave-one-property-out validation with the exclusion of a specific property during each train-test phase in order to determine the significance of each property on the accuracy of prediction. Our study reveals that radial basis function presents the best performing kernel function and the exclusion of polarity (#5) and polarizability (#6) from the protein-chain descriptors resulted in the optimal prediction accuracy of 84.00% to 99.95% for 81 functional families from the 2A family of transporter in Transporter Commission Database (TC-DB). Refined versions of our SVM can serve as a useful tool for transporter protein classification of the entire TC-DB. In this way, the effort to create a universal classification system for all currently identified and yet-to-be recognized transport proteins could be greatly facilitated.

4 Figures and tables.

Property	All	Exclude Property										
		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11
Ave. Acc. (%)	99.31	99.10	99.14	99.29	99.28	99.43	99.40	99.27	99.29	99.28	99.21	99.26
Ave. Acc. of top 5 biggest subclasses (%)	97.04	95.76	96.15	96.95	96.93	96.34	96.21	96.94	96.99	96.93	96.41	96.94
Prec/Recall Score Freq	0.543	0.193	0.218	0.523	0.519	0.860	0.852	0.482	0.535	0.519	0.420	0.457

Table 1: SVM prediction accuracy using the leave-one-property-out validation

5 References

[1] Cai CZ, Han LY, Ji ZL, Chen X, and Chen YZ. 2003. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res*. 31, 3692-3697.

[2] Dubchak I, Muchnik I, Holbrook SR, and Kim SH. 1995. Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. U. S. A. 92*, 8700-8704.

[3] Joachims T. 2001. Learning To Classify Text Using Support Vector Machines - Methods, Theory, *Algorithms*. Kluwer Academic Publishers

[4] Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol*. 292, 195-202.

[5] Jones DT, Taylor WR, and Thornton JM. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry (Mosc)*. 33, 3038-3049.

[6] Saier MH Jr. 2000. A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol. Mol. Biol. Rev*. 64, 354-411.