REFERENCES

[1] Pauling L, Corey RB. Configurations of polypeptide chains with favoured orientations around single bonds: Two new pleated sheets. *Proc Natl Acad Sci U.S.A.* 1951;37:729–740.

[2] Pauling L, Corey RB, Branson HR. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U.S.A.* 1951;37:205–234.

[3] Branden C, Tooze J. *Introduction to Protein Structure.* Garland Publishing 1999.

[4] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982;157:105-132.

[5] Kabsch W, Sander C. Dictionary of protein secondary structure, pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983; 22(12):2577-637.

[6] Minor DL Jr, Kim PS. Context-dependent secondary structure formation of a designed protein sequence. *Nature* 1996;380:730-734.

[7] Garnier J, Osguthorpe DJ, Robson B. Analysis and implication of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* 1978;120:97-120.

[8] Qian N, Sejnowski TJ. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol* 1988;202:865-884.

[9] Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* 1993;232:584-599.

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

[11] V. Eyrich DS, Friesner R. Prediction of protein tertiary structure to low resolution: performance for a large and structurally diverse test set. *J Mol Biol* 1999;288:725–742.

[12] C. Chen JS, Altman R. Using imperfect secondary structure predictions to improve molecular structure computations. *Bioinformatics* 1999;15:53–65.

[13] Samudrala Y, Xia EH, Levitt M. *Ab initio* protein structure prediction using a combined hierarchical approach. *Proteins* 1999;Suppl:194–198.

[14] M. Young K, Kirshenbaum KD, Highsmith S. Predicting conformational switches in proteins. *Prot. Sci.* 1999;8:1752–1764.

[15] Davies G, Martin I, Sturrock S, Cronshaw A. On the structure and operation of type i dna restriction enzymes. *J Mol Biol* 1999;290:565–579.

[16] Szent-Gyrgyi, AG, Cohen C. Role of proline in polypeptide chain configuration of proteins. *Science* 1957;126:697.

[17] Blout E, de Loz C, Bloom S, Fasman G. Dependence of the conformation of synthetic polypeptides on amino acid composition. *J. Am. Chem. Soc.* 1960;82: 3787–3789.

[18] Rost B. Rising accuracy of protein secondary structure prediction. 2002

[19] Chou PY, Fasman GD. Conformational parameters for amino acids in helical, sheet and random coil regions calculated from proteins. *Biochemistry* 1974;13:211-222.

[20] Bishop C. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford. 1995.

[21] Haykin S. *Neural Networks - A comprehensive foundation.* Prentice Hall. 1999.

[22] Baldi P, Brunak S, Frasconi P, Soda G, Pollastri G. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics* 1999; 15:937-946.

[23] Holley H, Karplus M. Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Science of the United States of America,* 1989;86:152–156.

[24] Kneller D, Cohen F. Improvements in secondary structure prediction by an enhanced neural network. *J Mol Biol* 1990;214(1):171–182.

[25] Stolortz P, Lapedes A, Xia Y. Predicting protein secondary structure using neural net and statistical methods. *J Mol Biol* 1990;214(1):171–182.

[26] Riis S, Krogh A. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J. Comput. Biol.* 1996;3:163–183.

[27] Huyen M. Exploring phenotype space through neutral evolution. *J. Mol. Evol.* 1996; 43:165–169.

[28] Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. *J Mol Biol* 1994;235:13-26.

[29] Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of sov, a segment-based measure for protein secondary prediction assessment. *Proteins: Struct Funct Genet* 1999;34:220-223.

[30] Vapnik V. *Statistical Learning Theory*. New York: John Wiley and Sons;1998.

[31] Schökopf B, Burges CJC, Smola AJ. *Advances in Kernel Methods –Support Vector Learning*. MIT Press;1999.

[32] Jaakkola T, Diekhans M, Haussler D. A discrimitive framework for detecting remote protein homologies. *J. Comput. Biol.* 2000;7, 95-114.

[33] Zien A, Ratsch G, Mika S, Schökopf B, Lengauer T, Muller KR. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics* 2000;16(9):799-807.

[34] Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. *J Mol Biol* 2001;308:397-407.

[35] Guo J, Chen H, Sun Z, Lin Y. A Novel Method for Protein Secondary Structure Prediction Using Dual-Layer SVM and Profiles. *Proteins: Struct Funct Bioinf* 2004;54:738-743.

[36] Durbin R, Eddy S, Krogh A Mitchison G. *Biological Sequence Analysis – Probabilistic models of proteins and nucleic acids.* Cambridge University Press;1998.

[37] Salzberg SL. A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Comput Appl Biosci* 1997;13(4):365-367.

[38] Audic S, Claverie J. Detection of eukaryotic promoters using Markov transition matrices. *Computers Chem* 1997;21:223-227.

[39] Borodovsky M, McIninch J. GENMARK: parallel gene recognition for both DNA strands. *Computers Chem* 1993;17:123-133.

[40] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25(17): 3389-3402.

[41] Osuna E, Freund R, Girosi F. An improved training algorithm for support vector machines. *Proceeding of IEEE Workshop on Neural Networks for Signal Processing VII*. New York; 1997. p 276-285.

[42] Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* 1999;34:508-519.

[43] Hobohm U, Scharf M, Schneider R, Sander C. Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Protein Science* 1992;1:409-417.

[44] Frishman D, Argos P. Knowledge-based secondary structure assignment. *Proteins* 1995;32:566–579.

[45] Richards, F. and Kundrot, C. Identification of structural motifs from protein coordinate data: secondary structure and first level super-secondary structure. *Proteins*, 1998;3:71–84.

[46] Matthews BW. Comparison of the predicted and observed secondary structure of t4 phase lysozyme. *Biochim. Biophys. Acta* 1975;450, 442-451.

[47] Joachims T. *Making large-scale SVM learning practical. In Advances in Kernel Methods – Support Vector Learning* (Schökopf B, Burges C, Smola A) MIT-Press; 1999. p 42-56.

[48] Garnier J, Gibrat JF, Robson B. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol* 1996;266: 540-553.

[49] Frishman D, Argos P. 75% accuracy in protein secondary structure prediction. *Proteins* 1997;27:329-335.

[50] Salamov AA, Solovyev VV. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignment. *J Mol Biol* 1995;247:11-15.

[51] Kasemsant Kuphanumat, Chidchanok Lursinsap. Prediction of protein secondary structure by combination of support vector machine and Markov models. *2nd International Conference on Bioinformatics*, 2003;100.

APPENDIX

Appendix I

Publications

[I] Kasemsant Kuphanumat and Chidchanok Lursinsap, "Prediction of protein secondary structure by combination of support vector machine and Markov models", *2nd International Conference on Bioinformatics 2003 (InCoB2003)*, September 2003.

Prediction of Protein Secondary Structure by Combination of Support Vector Machine and Markov models

Kasemsant Kuphanumat[1,2] and Chidchanok Lursinsap[1]

[1]Advanced Visual and Intelligent Computing Research Center (AVIC),
Department of Mathematics, Faculty of Science, Chulalongkorn University,
Bangkok 10330, Thailand.
[2]National Center for Genetic Engineering and Biotechnology (BIOTEC),
National Science and Technology Development Agency (NSTDA),
Pathumthani 12120, Thailand.

Support Vector Machine (SVM) currently is a novel approach for protein secondary structure prediction. Based on statistical learning theory and its generalization, SVM was reported to have out-performed results for many applications on Bioinformatics. Even through the method of protein secondary structure prediction based on SVM achieved a good performance, it did not produce the remarkably high results. The primary obstacle that vastly inhibits the power of the predicting model is an inappropriately encoding scheme of protein sequence data. Thus, finding an efficiently representative set of input features for protein sequences is, extremely, an important process to achieve the goal of protein secondary structure prediction. In our research, we introduce a new method based on Makov process to encode the protein sequences. With this simple method, input vectors that contain the essential features of protein sequence can be extracted and efficiently used to train SVM classifiers. Our method achieved the highest results that out-perform other advanced methods at present. The SVM together with Markov transition matrix encoding scheme produces the performance of three-state overall per-residue accuracy measure $Q_3 = 81.39$ and segment overlap accuracy measure SOV = 78.64% through a seven-folded cross validation on the data set of 513 non-homologous protein chains (CB513). That is the next improving step closing to the theoretical limitation.

Availability: The programs' source code and data sets are available upon request.

Contact: kasemsant@biotec.or.th; lchidcha@chula.ac.th

Keywords: Protein secondary structure prediction, Support Vector Machine (SVM), Markov transition matrix, Machine learning, Patterns classification.

Biography

Mr Kasemsant Kuphanumat

PERSONAL DETAILS:

Data of Birth:	July 19, 1970
Place of Birth:	Nakornsawan, Thailand

EDUCATION:

Jun'01 – Mar'06	Ph.D. Program in Computer Science, Department of Mathematics, Chulalongkorn University, Thailand.
Jun'93 – Mar'97	M.Sc. Program in Computer Science, Department of Mathematics and Computer Science, Kingmongkut Institute of Technology Lardkrabang, Thailand.
June'89 – Mar'93	B.Sc. Program in Biochemistry, Department of Biochemistry, Chulalongkorn University, Thailand.

SCHOLARSHIPS:

June'01 – Mar'06	Scholarship of the Royal Golden Jubilee (RGJ) Ph.D. Program.

PUBLICATIONS:

Sep'03	Kasemsant Kuphanumat and Chidchanok Lursinsap,"Prediction of protein secondary structure by combination of support vector machine and Markov models", 2nd *International Conference on Bioinformatics 2003 (InCoB2003)*, September 2003.