Journal of Computational Intelligence in Bioinformatics
ISSN: 0973-385X Volume 4 Number 2 (2011) pp. 173-181

http://www.ripublication.com/jcib.htm

A Survey of Support Vector Machines and an Analysis of DNA Methylation Information Processing using an SVM based Online Methylator Software.

D.N.T. Kumar[*, 1, 2], Qufu Wei[1], Joseph Kost[2], Fenglin Huang[1]
Tao Dan[1], Li Jing[1] and Lu Bing[1]

[1]Key Laboratory of Eco-textiles, Ministry of Education,
Jiangnan University, Wuxi, 214122, P.R. China.
[2]Laboratory of Controlled Release Systems and Gene Therapy,
Department of Chemical Engineering,
Ben-Gurion University of the Negev, Beer Sheva, 84105, Israel.
[*]Corresponding Author E-mail:tejdnk@gmail.com/qfwei@jiangnan.edu.cn

Abstract

Support vector machines (SVMs), are a set of related supervised learning methods, that analyze data and recognize patterns, used for classification and regression analysis. SVMs have been widely applied to many areas of bioinformatics, including protein function prediction, protease functional site recognition, transcription initiation site prediction and gene expression data classification. In this survey paper, we discussed and highlighted, the principles of SVMs and their applications, to perform the analysis, of mainly DNA Methylation concept. Tutorial approach is adopted in our current paper.

Keywords: Machine Learning, SVM, Algorithms, Classification, Regression Analysis, Biological data, DNA, DNA-Methylation.

Introduction

Background theory and Historical Perspectives of SVMs

The original SVM algorithm, was invented by Vladimir Vapnik and the current standard incarnation(soft margin), was proposed by Corinna Cortes and Vladimir Vapnik. In bioinformatics, a demanding and tough task, is the classification and prediction of biological data. With the rapid increase, in size of the biological data banks, it is essential to use computer programs based on machine learning techniques, to automate the classification process."At present, the computer programs that give

the best prediction performance, are support vector machines (SVMs).This is because SVMs, are designed to maximize, the margin to separate two classes, so that, the trained model generalises, well on unseen data. Most other computer programs, implement a classifier, through the minimization of error, occurred in training, which leads to poorer generalization".[1-2]

"Classifying is a common task, in machine learning. Suppose some given data points, each belong to one of two classes, and the goal is to decide, which class a new data point will be in. In the case of support vector machines, a data point is viewed as a p-dimensional vector (a list of p numbers) and we want to know whether, we can separate such points with a, $p - 1$-dimensional hyperplane. This is called a linear classifier. There are, many hyperplanes that might classify, the data. One reasonable choice as the best hyperplane, is the one that represents the largest separation, or margin, between the two classes. So we choose the hyperplane, so that the distance from it, to the nearest data point on each side is maximized. If such a hyperplane exists, it is known, as the maximum-margin hyperplane and the linear classifier it defines, is known as a maximum margin classifier".[1-6]

SVMs belong to a family, of generalized linear classifiers."They can also be, considered a special case of Tikhonov regularization. A special property, is that they simultaneously minimize, the empirical classification error and maximize the geometric margin, hence, they are also known as, maximum margin classifiers. A comparison, of the SVM to other classifiers, has been made by Meyer, Leisch and Hornik".[7-10]

In searching for the best hyper-plane, "SVMs find a set of data points, that are the most difficult training points to classify. These data points are referred to as support vectors. In constructing an SVM classifier, the support vectors are closest to the hyper-plane and are located on the boundaries of the margin between two classes. The advantage of using SVMs is that the hyper-plane is searched through maximising this margin, because of this, the SVM classifier is the most robust and hence has the best generalisation ability".[11-16]

Background theory and applications of DNA methylation concept

"After every cycle of DNA replication, several modifications occur in the DNA.DNA methylation is one such post-synthesis modification. DNA methylation has been proven by research to be manifested in a number of biological processes such as regulation of imprinted genes, X chromosome inactivation and tumor suppressor gene silencing in cancerous cells. It also acts as a protection mechanism adopted by the pathogen DNA (mainly bacterial against the endonuclease activity that destroys any foreign DNA. In eukaryotes, methylation plays a vital role in gene expression regulation. Furthermore, it has been observed that abnormal methylation of the DNA, can lead to coordinate changes in expression of genes, resulting in cancer growth and metastasis".[17-25]

"The most common type of DNA modification, consists of the methylation of cytosine in the CpG dinucleotide. Methylation in non-CpG sequences, is less frequent represting up to 15-20% of total 5'-methylcytosine.CpGs are present at an average, of one per 80-dinucleotides throughout most part of the genome. However, there are

regions within the genome, where CpGs are around five times greater than the average. These regions are known as CpG islands and comprise 1-2% of the genome. CpG islands, have a high G+C content (greater than 50%) and a size ranging from, 200bp to several thousand base pairs".[26-29]

Main function of DNA methylation is the repression of gene expression. Thus, "methylation of CpGs may disrupt the binding, of certain transcription factors."Also, DNA methylation, promotes the association to the DNA of specific 5-methylctosine binding proteins and other structural proteins, which results in the packing of the DNA into a structure that is inaccessible to transcription factors. Not surprisingly, it has been suggested that patterns of methylation may compartmentalize the genome into transcriptionally active (non-methylated) and non-active regions (methylated). DNA methylation, can alter the flow, of the genetic information and reprogram the genome function and therefore it is recognized, as the major epigenetic modification. Genomic methylation patterns, in non-dividing somatic differentiated cells are, generally stable and heritable. However, there are instances where methylation patterns, undergo significant changes to alter the phenotype".[30-32]

Description of SVMs based Software Tools and their Bio-Applications

The relationship, between levels of DNA methylation and gene activity has been known for some time."Many of the early procedures developed gave only somewhat limited information about methylation patterns, for example, the total level of 5-methyl cytosine in the genome or the frequency of methylation of cytosines, within certain restriction sites". However, in the last few years, there has been an explosion of interest in DNA methylation and with it, many new and powerful techniques, have been developed to facilitate its study.[16-19]

Ever since methylcytosine, was found in genomic DNA, this epigenetic alteration has become a center of scientific attraction, especially, because of its relation to gene silencing in disease."There is currently, a wide range of methods designed, to yield quantitative and qualitative information on genomic DNA methylation. The earliest approaches were concentrated on the study of overall levels of methylcytosine, but more recent efforts have focused on the study of the methylation status of specific DNA sequences".[18-23]

"Particularly, optimization of the methods based on bisulfite modification of DNA permits the analysis of limited CpGs in restriction enzyme sites(e.g., combined bisulfite restriction analyses and methylation-sensitive single nucleotide primer extension) and the overall characterization based on differential methylation states", (e.g., methylation-specific PCR, MethyLight and methylation-sensitive single-stranded conformational polymorphism) and allows very specific patterns of methylation to be revealed (bisulfite DNA sequencing)."In addition, novel methods designed to search for new methylcytosine hot spots have yielded further data without requiring prior knowledge of the DNA sequence". Quantitative assessment of DNA methylation, has an important potential in applications for, disease diagnosis, classification and prognosis in clinical settings.[30-35]

Genetic information is not merely contained in the arrangement, of the four nucleotide bases, but also in the "covalent addition of methyl groups to cytosine within CpG dinucleotides. Methylation and related chromatin changes are important processes in the regulation of gene expression. The relevance of DNA methylation has been demonstrated in mammalian development, imprinting and X-chromosome inactivation, suppression of parasitic DNA and various cancer types. The ability to detect and to quantify methylation is particularly important to the field of cancer diagnostics. Changes in the methylation status of DNA have the potential to serve as an early detection marker for malignancies".[28-35]

Free/Open Source and Online Software Tools/Reports used in our Paper

FASTA Software and DNA Methylator Software Usage.

FASTA-Compares a protein sequence to another protein sequence or to a protein database, or a DNA sequence to another DNA sequence or a DNA library.

"DNA Methylator-Users can enter a nucleotide sequence in one of the standard formats such as GenBank, EMBL, GCG or plain format. The method provides the option of pasting the sequence in the text area or uploading the direct sequence file. All non-standard characters except the four nucleotides bases adenine, guanine, cytosine and thymine will be ignored from the sequence. The method only allows the prediction for single sequence in one run of prediction and there is limit size of 10, 000 nucleotides per query".[22]

Additional Information that could be useful for an in-depth study

i. http://bio.dfci.harvard.edu/Methylator/index.html
ii. http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml
iii. http://www.premierbiosoft.com/molecular_beacons/dna-methylation/index.html
iv. http://www.ncbi.nlm.nih.gov/pubmed/14501188
v. http://www.ncbi.nlm.nih.gov/pubmed/12238773
vi. http://en.wikipedia.org/wiki/Support_vector_machine

Analysis and Discussion

We have shown here a simple procedure using the concepts, described above, to demonstrate our Viewpoint.

FASTA Software based Sequencing File of HomoSapiens(TTN), transcript variantN2-B, mRNA

We reproduce here only "part of the sequence file", as the length of the Sequence is too long.

>gi|20143913|ref|NM_003319.2| Homo sapiens titin (TTN), transcript variant N2-B, mRNA

AGCAGTCGTGCATTCCCAGCCTCGCCTCGGGTGTAGGGATTGCATAGAAA
AGCAAAACTACACAGTCTTGACTGTGTAGTTTTGTTTTTAGGATTAGAGGC
TCACCGATTCATGTCGGAGATGGTCAGAAAAACCAACTCTCCATAGGACG
TCGTTTCAGAAGCAACCTTGGGCTTAGTCCCACCCTTTTTAGGCACTCTTG
AGAAATCAAGTGCCTAGAAAGATGACAACTCAAGCACCGACGTTTACGCA
GCCGTTACAAAGCGTTGTGGTACTGGAGGGTAGTACCGCAACCTTTGAGG
CTCACATTAGTGGTTTTCCAGTTCCTGAGGTGAGCTGGTTTAGGGATGGCC
AGGTGATTTCCACTTCCACTCTGCCCGGCGTGCAGATCTCCTTTAGCGATG
GCCGCGCTAAACTGACGATCCCCGCCGTGACTAAAGCCAACAGTGGACGA
TATTCCCTGAAAGCCACCAATGGATCTGGACAAGCGACTAGTACTGCTGA
GCTTCTCGTGAAAGCTGAGACAGCACCACCCAACTTCGTTCAACGACTGC
AGAGCATGACCGTGAGACAAGGAAGCCAAGTGAGACTCCAAGTGAGAGT
GACTGGAATCCCTACACCTGTGGTGAA

DATA File 1: Sequence File used in our Simulation.
[This Example sequence file is taken from the FASTA Sequencing Software Examples directory]

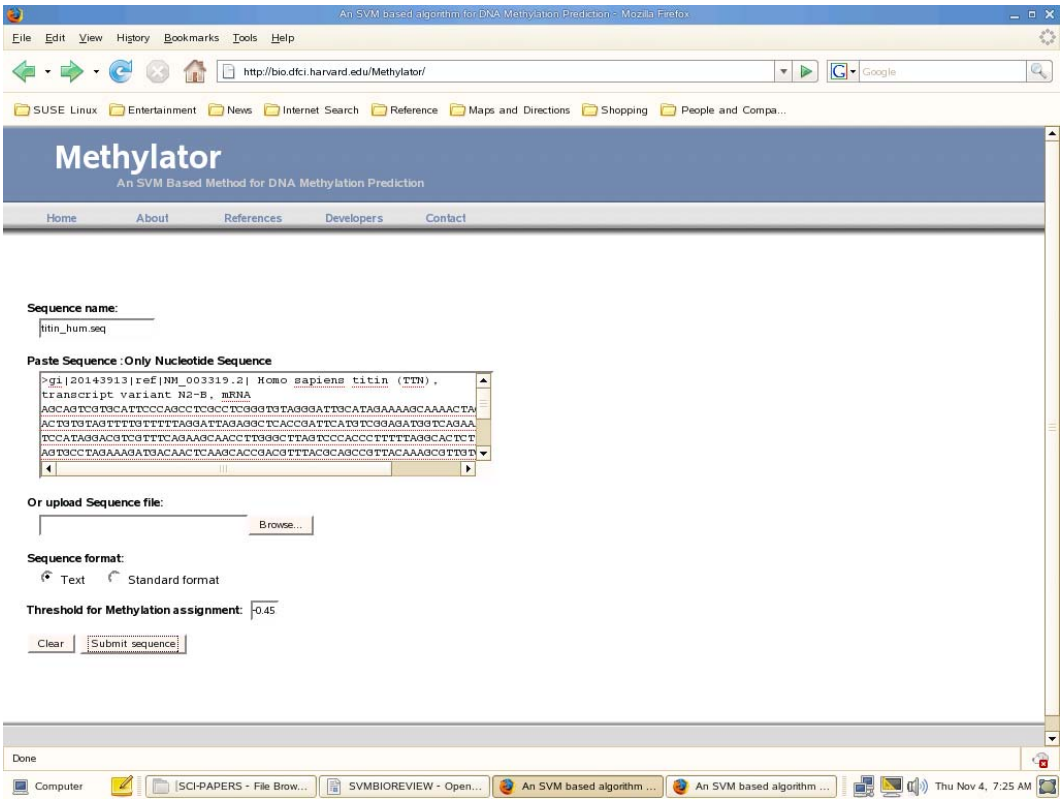

Figure 2: Methylator Screen Shot to initiate the DNA Methylation prediction Simulation. [http://bio.dfci.harvard.edu/Methylator/].

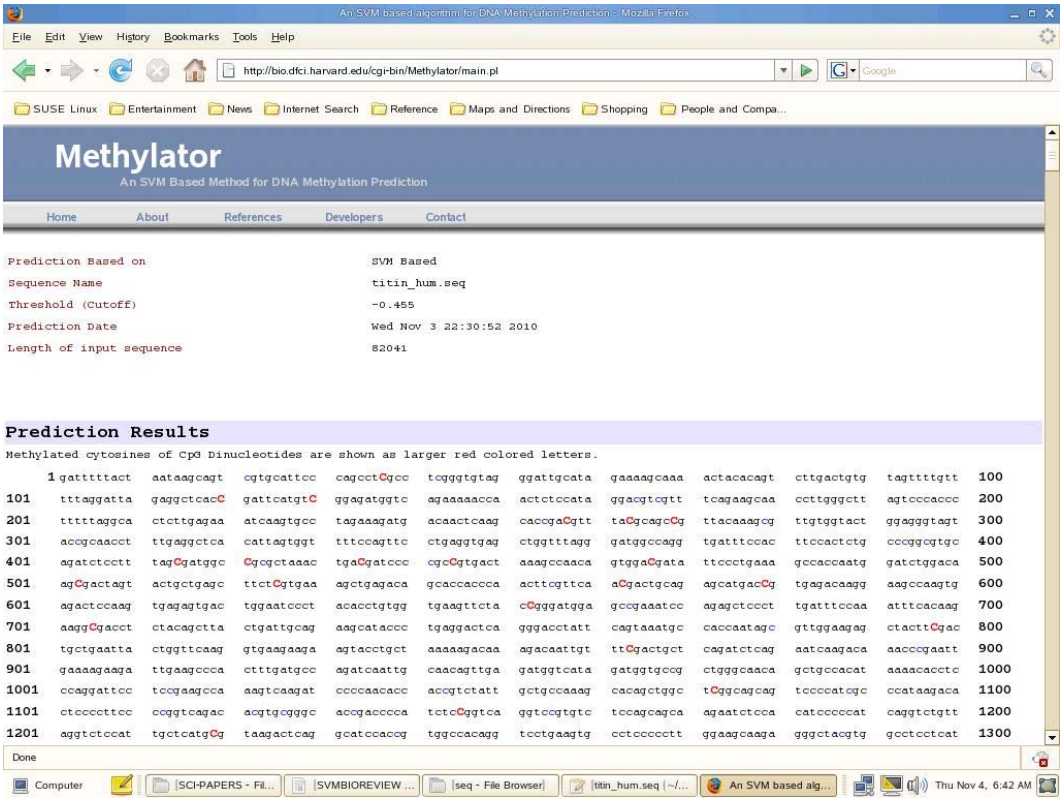

Figure 3: Methylator Screen Shot from our prediction simulation. Methylated cytosines of CpG Dinucleotides are shown as larger red colored letters by the Software.[http://bio.dfci.harvard.edu/cgi-bin/Methylator/main.pl]

We performed the DNA Methylation prediction, using the online SVM based Methylator Software, from Harvard University, Cambridge, Massachusetts, USA.

Simulation File Details

Prediction Based on	SVM Based
Sequence Name	titin_hum.seq
Threshold (Cutoff)	-0.455
Prediction Date	Wed Nov 3 22:30:52 2010
Length of input sequence	82041

Conclusion and Future perspectives

DNA methylation contributes, to the control of gene expression and plays an essential role, in cellular physiology. Well-defined patterns of DNA methylation, are established and fixed during embryonic development and changes, in these patterns may be a contributing factor in developmental disorders, cancer and aging. Not least,

the possibility of using DNA methylation, as a marker for disease, has created a strong need for techniques, to detect and measure DNA methylation. Different techniques provide information on DNA methylation at different levels, spanning from genome-wide methylation content to methylation of single residues in specific genes. The limitations, of individual techniques strongly affect, interpretation of data. In this survey, we discussed some general themes about DNA methylation analysis and outline, the basic principles of current key techniques.

We discussed, the advantages of using SVM based computational techniques, suggested some overall guidelines, that may be instructive for a rational choice of methodology. As we have seen, from the published literature, the SVM boasts a strong theoretical underpinning, coupled with remarkable empirical results across a growing spectrum of applications. Thus, SVMs will probably continue to yield valuable insights into the growing quantity and variety of molecular biology data.

Acknowledgements

We sincerely thank, all the members involved, in making this paper a possibility. Further, we thank Jiangnan University, Wuxi, P.R. China and Ben-Gurion University of the Negev, Beersheva, Israel for providing us, with the research support and the necessary conducive environment. The Authors declare, no competing financial interests. Further, we extend our sincere thanks, to some of the authors' for giving us permissions, to reproduce some of their published works, in our survey and analysis sections.

References

[1] Noble W S., What is a support vector machine?., Nat Biotech, Vol (24):12, 2006.

[2] Yang Z.R., Biological applications of support vector machines, Brief Bioinform (2004) 5 (4):328-338., [doi: 10.1093/bib/5.4.328.]

[3] Li, E., "Chromatin modification and epigenetic reprogramming in mammalian development." Nat Rev Genet, 2002. 3(9): p.662-73.

[4] Walsh, C.P., J.R. Chaillet and T.H. Bestor., "Transcription of IAP endogenous retroviruses is constrained by cytosine methylation."Nat Genet, 1998.20(2): p.116-7.

[5] Costello, J.F., et al., "Aberrant CpG-island methylation has non-random and tumour-type specific patterns."Nat Genet., 2000.24(2): p.132-8.

[6] Jones, P.A. and S.B. Baylin, "The fundamental role of epigenetic events in cancer." Nat Rev Genet, 2002.3(6): p.415-28.

[7] Costello, J.F. and C. Plass., "Methylation matters." J Med Genet, 2001. 38(5): p.285-303.

[8] Feinberg, A.P., "Cancer epigenetics takes center stage." Proc Natl Acad Sci U S A, 2001. 98(2): p.392-4.

[9] Stanssens, P., et al, "High-Throughput MALDI-TOF Discovery of Genomic Sequence Polymorphisms., Genome Research, 2004. 14(1): p.126-133.
[10] Grunau, C., Renault, E., Rosenthal, A., Roizes, G., Worm, J., Guldberg, P., Brusic, V., vanEndert, P., Zeleznikow, J., Daniel, S. et al., (2001) Nucleic Acids Res, 29, 270-274.
[11] Amoreira, C.Hindermann, W.Grunau, C. Issa, J.P. Adorjan, P.Distler, J.Lipscher, E. Model, F.Muller, J.Pelet, C et al., (2003)Nucleic Acids Res, 31, 75-77.
[12] Issa, J.P., (2004) Nat Rev Cancer, 4, 988-993.
[13] Takai, D. and Jones, P.A., (2002) Proc Natl Acad Sci U S A, 99, 3740-3745.Epub 2002, Mar 3712.
[14] Costello, J.F., Plass, C., Singal, R., Ginder, G.D., Takai, D., Jones, P.A., Feltus, F.A., Lee, E.K., Vertino, P.M., Amoreira, C.et al., (2001) J Med Genet, 38, 285-303.
[15] Feltus, F.A., Lee, E.K., Costello, J.F., Plass, C., Vertino, P.M., Amoreira, C., Hindermann, W., Grunau, C., Issa, J.P., Adorjan, P. et al. (2003) Proc Natl Acad Sci U S A, 100, 12253-12258.Epub 12003, Sep 12230.
[16] Singal, R. and Ginder, G.D., (1999) Blood, 93, 4059-4070.
[17] Szyf, M., Pakneshan, P. and Rabbani, S.A., (2004) Biochem Pharmacol, 68, 1187-1197.
[18] Adorjan, P., Distler, J., Lipscher, E., Model, F., Muller, J., Pelet, C., Braun, A., Florl, A.R., Gutig, D., Grabs, G. et al., (2002) Nucleic Acids Res, 30, e21.
[19] Worm, J., Guldberg, P., Brusic, V., van Endert, P., Zeleznikow, J., Daniel, S., Hammer, J., Petrovsky, N., Szyf, M., Pakneshan, P. et al., (2002) J Oral Pathol Med, 31, 443-449.
[20] Burges, C.J.C., (1998) A tutorial for support vector Machines for pattern recognition.
[21] Bhasin, M. and Raghava, G.P., (2004) J Biol Chem, 279, 23262-23266. Epub 22004, Mar 23223.
[22] Bhasin, M. and Raghava, G.P., (2004) Nucleic Acids Res, 32, W414-419.
[23] Cai, Y.D., Zhou, G.P. and Chou, K.C., (2003) Biophys J, 84, 3257-3263.
[24] Yang, Z.R. and Chou, K.C., (2004) Bioinformatics, 20, 735-741.Epub 2004 Jan 2029.
[25] Joachims, T., (1999) In Smola, B. S. a. C. B.a. A. (ed.), Advances in Kernel methods of support vector learning. MIIT Press, Cambridge massachusetts, London, England.
[26] Bhasin, M and Raghava, G.P., (2004) Vaccine, 22, 3195-3204.
[27] Swets, J.A., (1986) Psychol Bull, 99, 181-198.
[28] Brusic, V., van Endert, P., Zeleznikow, J., Daniel, S., Hammer, J., Petrovsky, N., Szyf, M., Pakneshan, P. and Rabbani, S.A. (2004) Biochem Pharmacol, 68, 109-121.
[29] Campbell, M.K., (1995) Biochemistry. Saunders College:Philadelphia, pgs.615-16, 181.
[30] Maclean, N., S.P. Gregory and R.A. Flavell (1993) Eukaryotic Genes.Butterworth and Co., London, pgs.53-67.

[31] Xu, DG., HZ He, GG Zhang, B.Ganswendt and H. Peter, (1993) DNA methylation of mono halogenated methanes of F344 rats, J. Tongi Med.Univ., 13(2)

[32] Freiefelder, D., (1987).Molecular Biology.Jones and Bartlett: Boston, pgs 550-559, DNA Methylation

[33] "SNP Discovery Using the MassARRAY System."Application note.Sequenom web site: http://www.sequenom.com/Assets/pdfs/appnotes/SNP_Discovery_Application_Note.pdf.

[34] Ehrich, M et al, "Multiplexed discovery of sequence polymorphisms using base-specific cleavage and MALDI-TOF MS." Nucleic Acids Research, 2005. 33(4): e38.

[35] Vu, TH et al, "Symmetric and assymetric DNA methylation in the human IGF2-H19 imprinted region.", Genomics, 2000. 64(2): p.132-143.