EXCLI Journal 2005;4:114-129 – ISSN 1611-2156
Received: 5. October 2005, accepted: 11. October 2005, published 14. October 2005

Original article:

Recognition of DNA Splice Junction via Machine Learning Approaches

Chanin Nantasenamat[1], Thanakorn Naenna[2],
Chartchalerm Isarankura-Na-Ayudhya[1], Virapong Prachayasittikul[1*]

[1]Department of Clinical Microbiology, Faculty of Medical Technology, Mahidol University, Bangkok 10700, [2]Department of Industrial Engineering, Faculty of Engineering, Mahidol University, Nakhon Pathom 73170, Thailand, Telephone: 662-849-6318, Fax: 662-849-6330, e-mail: mtvpr@mucc.mahidol.ac.th ([*]corresponding author)

ABSTRACT

Successful recognition of splice junction sites of human DNA sequences was achieved via three machine learning approaches. Both unsupervised (Kohonen's Self-Organizing Map, KSOM) and supervised (Back-propagation Neural Network, BNN; and Support Vector Machine, SVM) machine learning techniques were used for the classification of sequences from the testing set into one of three categories: transition from exon to intron, transition from intron to exon, and no transition. The dataset used in this study is comprised of 1,424 DNA sequences obtained from the National Center for Bioinformatics Information (NCBI). Performance of the machine learning approaches were assessed by the construction of learning models from 1,000 sequences of the training set and evaluated on the 424 sequences of the testing set that is unknown to the learning model. Each sequence is a window of 32 nucleotides long with regions comprising -15 to +15 nucleotides from the dinucleotide splice site. Since the nucleotides (A, C, G, and T) are represented by four digit binary code (e.g. 0001, 0010, 0100, and 1000) the number of descriptors increased from 32 to 128. The performance of machine learning techniques in order of increasing accuracy are as follows SVM > BNN > KSOM, suggesting that SVM is a robust method in the identification of unknown splice site. Although KSOM gave lower prediction accuracy than the two supervised methods, it is fascinating that it was able to make such prediction based only on knowledge of the input whereas the supervised method requires that the output be known during training. It is expected that the Support Vector Machine method can provide a powerful computational tool for predicting the splice junction sites of uncharacterized DNA.

Keywords: DNA splice junction, machine learning, Kohonen's Self-Organizing Map, Back-propagation Neural Network, Support Vector Machine

INTRODUCTION

The deoxyribonucleic acid (DNA) of humans comprises of over three billion nucleotides and an estimated 30,000 genes (Venter et al., 2001; Collins et al., 2003; International Human Genome Sequencing Consortium 2004). Gene expression is the multi-step processes by which DNA expresses the gene product that it encodes. First, certain region of the DNA is transcribed into RNA in the form of pre-mRNA. Next, the introns of the pre-mRNA are excised, leaving only exons intact to become the mature mRNA. The ribosome then translates the mRNA into a polypeptide chain of amino acids that eventually becomes a protein (Cooper et al., 2004).

DNA splice junction sites (Figure 1) are boundaries where splicing occurs and are found between the regions of DNA that code for gene products (exon) and those that do not (intron) (Hastings et al., 2001). The presence of introns in eukaryotic organisms are believed to be involved in exon shuffling (or alternative splicing) that is responsible for the higher diversity of gene products found in eukaryotic organisms than that of prokaryotic organisms (Fedorova et al., 2003; Long et al., 2003; Roy, 2003). A typical example of exon shuffling is the generation of antibodies against foreign antigens that may invade the host system. The dinucleotide AG are splice sites that borders the transition from intron to exon (Intron/Exon border) going from 5’ to 3’, while GT are associated with the transition from exon to intron (Exon/Intron border). The GT dinucleotide is usually referred to as “donor” whereas the AG dinucleotide is known as “acceptor” (Snyder et al., 1995).

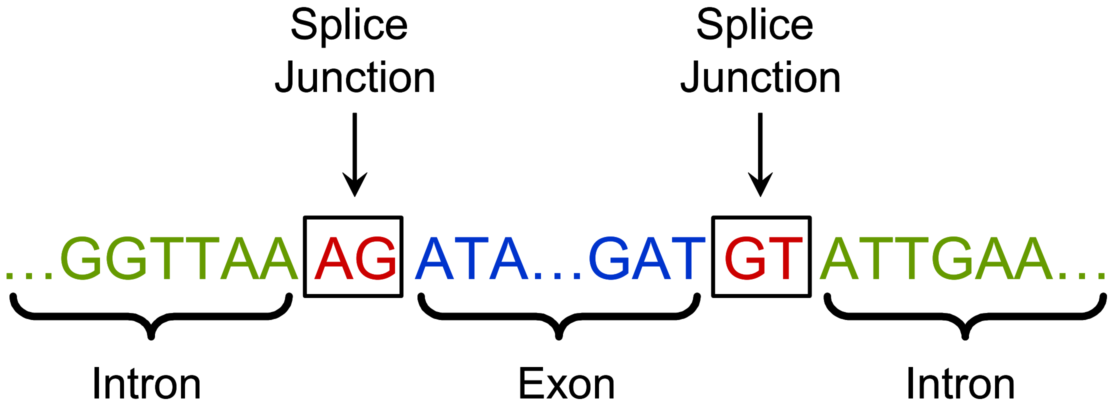

Figure 1: Schematic representation of the splice junction site.

The Human Genome Project as well as other genome projects aimed toward the unraveling of the genetic makeup of various organisms has generated a large volume of uncharacterized DNA sequences (Celniker, 2000; Johnston, 2000; Rubin, 2001; Venter et al., 2001). Much effort has been geared toward the prediction of gene products based on the DNA sequence alone by identifying regions of the DNA that serve as signals essential in gene expression (Burge et al., 1997; Burge et al., 1998; Brent et al., 2004). These signals include transcription initiation (Hannenhalli et al., 2001) and termination (Lesnik et al., 2001) sites, translation initiation (Zhu et al., 2004) and termination (Williams et al., 2004) sites as well as splice junction sites (Reese et al., 1997). Apart from constitutive splicing sites, much interest has been focused on the identification of alternative splicing sites, in which a gene is capable of producing several distinct mRNAs that code for different proteins (Rätsch et al., 2005; Zheng et al., 2003; Dror et al., 2005). In order to determine the protein that is to be produced it is essential to precisely identify regions of the DNA that are to be translated. Therefore, the ability to predict the location of splice sites in DNA sequences has great implications for the identification of potential gene product (Pertea et al., 2001; Zhang et al., 2003; Chen et al., 2005).

Machine learnings are methods that are capable of making predictions by learning from a set of training data and extrapolating the newfound knowledge on a set of testing data (Witten et al., 2000; Han et al., 2001). Different learning techniques use different algorithms to extract useful knowledge from the dataset (Han et al., 2001). The machine learning approaches used in this study comprise of two types of learning method, namely the supervised learning methods, back-propagation neural network and support vector machine, and the unsupervised learning method, Kohonen's self-organizing map. Supervised learning approaches require the knowledge of the output to be known while training, whereas unsupervised learning techniques can be trained without knowing about the actual output values (Han et al., 2001). In other words, supervised learning involves learning by example while the unsupervised approach learns by finding similarities among the instances of the data to assign them class labels.

Machine learning holds great promise in predictive medicine, preventive medicine, and personalized medicine (Weston et al., 2004). Genetic screening together with machine learning would make it possible to predict the probable future health history of patients based on the genes found in their blood sampling. In the case that a genetic abnormality is discovered treatments and precautions could be instigated to prevent the disease from occurring through the use of "drugs, embryonic stem cell therapy, engineered proteins, genetically-engineered cells, and many others" (Bensmail et al., 2003; Hood et al., 2004; Weston et al., 2004; Pennisi 2005; Singer 2005; Institute for Systems Biology, 2005). Personalized treatment offers great benefits to patients since each individual are unique and may require slightly different treatment than off-the-counter drugs that may cause side-effects. It is also anticipated that prediction, preventive, and personalized medicine may extend the lifespan by 10-30 years (Weston et al., 2004).

In this study, we aim to predict the location of splice sites in DNA sequences by classifying the instances of the testing set into one of three categories of splice sites: Exon/Intron, Intron/Exon, and unknown splice site. This was put forth by analyzing a dataset in which the DNA sequence contains one of two possible dinucleotide splice site that is flanked by neighboring nucleotides. The machine learning approaches will learn how to classify an unknown DNA sequence into one of three categories of splice site based on knowledge gained from the training dataset. Different learning approaches were able to correctly predict the type of splice site at varying degree of accuracy with support vector machine outperforming the rest.

MATERIALS AND METHODS

Data Collection

The DNA dataset used in this study was obtained from the website of the National Center for Bioinformatics Information (NCBI). The data used was taken from four different genes to take into consideration possible variability that may exist among the genes. This dataset comprises of 1,424 sequences of Human DNA that is split into two portions: 1) a training set of 1,000 sequences, and 2) a testing set of 424 sequences. Each sequence contains as input a total of 32 nucleotides with 15 nucleotides flanking upstream and downstream of the dinucleotide splice junction site; and an output that contains three possible values, which is associated with the following types of splice junction going from left to right of the dinucleotide splice site: Intron-AG-Exon, Exon-GT-Intron, and unknown-AG or GT-unknown.

Data pre-processing

The entries of the dataset were processed to contain information describing the regions surrounding the splice site by leaving nucleotides flanking the dinucleotide splice site from -15 to +15 to obtain a total of 32 nucleotides. The DNA nucleotides were converted into four digit binary code as to facilitate ease of processing by machine learning software. The nucleotides adenine, cytosine, guanine, and thymine are represented as 0001, 0010, 0100, and 1000, respectively. Therefore, each entry of the dataset comprises of 32*4 or 128 descriptors where each descriptor is a binary number (0 or 1).

Overview of machine learning approaches

Kohonen's self-organizing map (KSOM) is an unsupervised learning neural network developed by Kohonen (Kohonen, 2001). KSOM transforms input data from a high-dimensional space into a lower-dimensional space in such a way that the topology of the input data and the relative distance between input data are preserved (Kohonen, 1998; Kohonen, 2001). Since KSOM is an unsupervised learning method it does not require an output to be known when training instead it converts the high-dimensional DNA sequence onto a two-dimensional map known as the U-Matrix. Input data points that are located close to each other in the input space are mapped to nearby neurons on the output map (Kohonen, 2001). Thus, KSOM is widely used for the visualization of high-dimensional data (Kaski et al., 1999). In KSOM training, output neurons compete with each other where only the winning neuron and to a lesser degree its neighboring neurons are adjusted. After the training is complete the U-Matrix, a map that visualizes the cluster structure of the input data, is generated. Similarities among the input data found clustered together thereby yielding neurons that have low distances from one another. The distances are represented by a color spectrum and the map is shaded to indicate the clustering tendency. Areas having low distance values in the U-Matrix form clusters, while those with high distance values on the U-Matrix indicate a cluster border.

Back-propagation neural network (BNN) is a supervised learning method that is capable of adaptive learning, in which weights that connect the neurons are adjusted accordingly with respect to the error (Zupan et al., 1999). In this study, the back-propagation neural network comprises of three layers, namely the input, hidden, and output layer. The input layer receives input data and relays it to the hidden layer for further processing and transformation of the input data and the output layer transmit the final results (Zupan et al., 1999). Each layer contains processing units called neurons (nodes). Each node of the hidden and output layer contains two components, namely the summation function and the transfer function. The summation function is computed from the weighted sum of all nodes that are sent to each node of the successive layer in a feed-forward manner. The sum is then processed by the transfer function based on certain pre-defined threshold to output values. For example, the summed value that is less than 0.5 are sent out as 0 while those that are greater than 0.5 are sent out as 1. A neural network is trained by adjusting the weights until they are optimal in which the predicted output value is as similar as possible to the actual output value (Zupan et al., 1999). Since each run starts with a random seeding of the weight value, multiple runs must be carried out in order to ensure reproducibility.

Support vector machines (SVM) are learning techniques, developed by Vapnik, based on the Statistical Learning Theory (Cristianini et al., 2004). SVM is usually used for binary classification where the output can have two possible values (e.g. 0 or 1, -1 or +1). Multi-class SVM (Hsu et al., 2002) is also possible and multi-class implementation of WEKA was used in this study. SVM learning comprises of two essential steps. The first involves the use of kernel functions to linearly or non-linearly transform input data from a low-dimensional space to a high-dimensional space (Cristianini et al., 2004). Next, generate numerous hyperplanes that segregates the data objects into two distinctive regions based on the output binary value. The sequential minimal optimization algorithm was used for the construction of hyperplanes (Platt 1998). The hyperplane that maximizes the distance between the data objects has been selected, and this hyperplane is referred to as the maximal hyperplane (Cristianini et al., 2004).

Machine learning calculations

All machine learning calculations were performed on a personal computer running Windows XP with Intel Pentium M 1.73 GHz CPU and 512 MB of RAM. Kohonen's self-organizing map was trained with the K2D program and visualized with the SOM TOOLBOX (Alhoniemi et al., 2005) running under MATLAB (The MathWorks, USA). Back-propagation neural network was computed with WEKA (Witten et al., 2000). Support vector machine calculations were made using the sequential minimal optimization (SMO) algorithm under WEKA.

Kohonen's self-organizing map calculations were carried out by allowing the learning rate parameter and neighborhood function to decrease during the training process. Training is divided into two phases. In the first phase or ordering phase, the parameters are set to the following values: the initial learning parameter, *$\alpha(t)$*, which controls the step size of the update, is 0.8, and is gradually linearly decreased to 0.01; the initial variance of the Gaussian function used for controlling the neighborhood size, *$\sigma(t)$*, is 5, and is gradually decreased with an exponential decay function during the training phase to 1. The first phase training is terminated after 20,000 iterations. The initial weight vectors are selected from a uniform random distribution in [-0.1, 0.1]. After the ordering phase has been trained, the final weight vectors of each neuron are used for the initial weight vectors of the second phase. The parameters of the second phase or fine tuning phase are set to the following values: the initial learning parameter is 0.05 and the initial variance of the neighborhood neurons is 2. The training is terminated after 200,000 steps. After training, the U-Matrix is calculated and all 1,000 DNA sequences of the training set are mapped and labeled on to the U-Matrix (Figure 2). Next, the 424 DNA sequences of the testing set are mapped and labeled onto their corresponding winning neurons and those that mapped in an incorrect cluster are defined as misclassified (Figure 3).

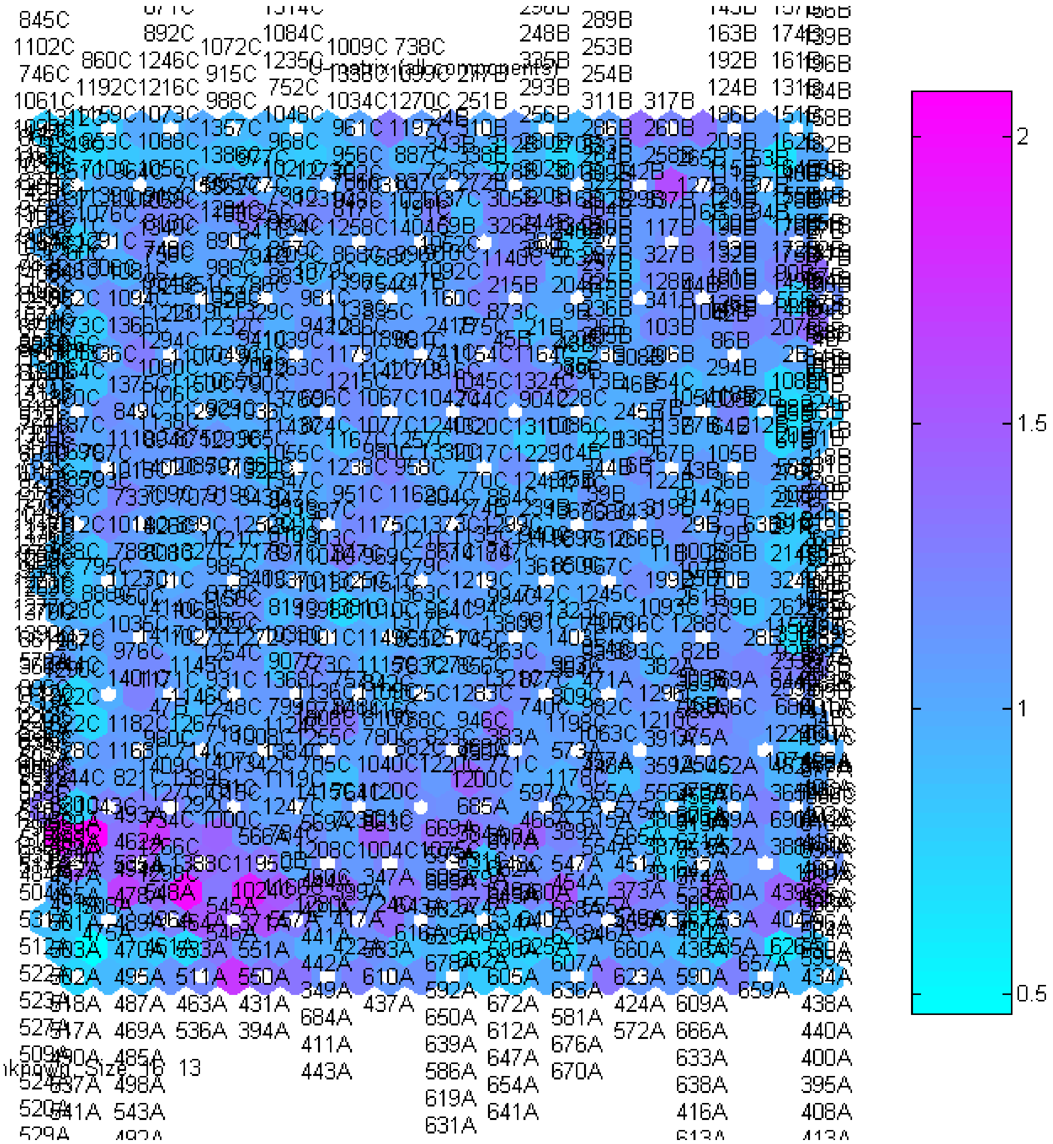

Figure 2: The U-matrix map of the training set.

Back-propagation calculations were carried out by first seeking the optimal network parameters through trial-and-error calculations. First, the optimal number of nodes to use for the hidden layer was varied from 2 to 32 and the number of nodes that gave the lowest root mean square error (RMS). Next, the optimal learning epoch was selected from the plot of the RMS as a function of the number of learning epochs that ranged from 1 to 100. Finally, the pair of learning rate (η) and momentum (μ) value that gave the lowest RMS was selected from the contour plot of the RMS as a function of the learning rate and momentum constants that ranged from 0 to 1. The optimal network parameters were used for the actual calculations and the average of five runs were used.

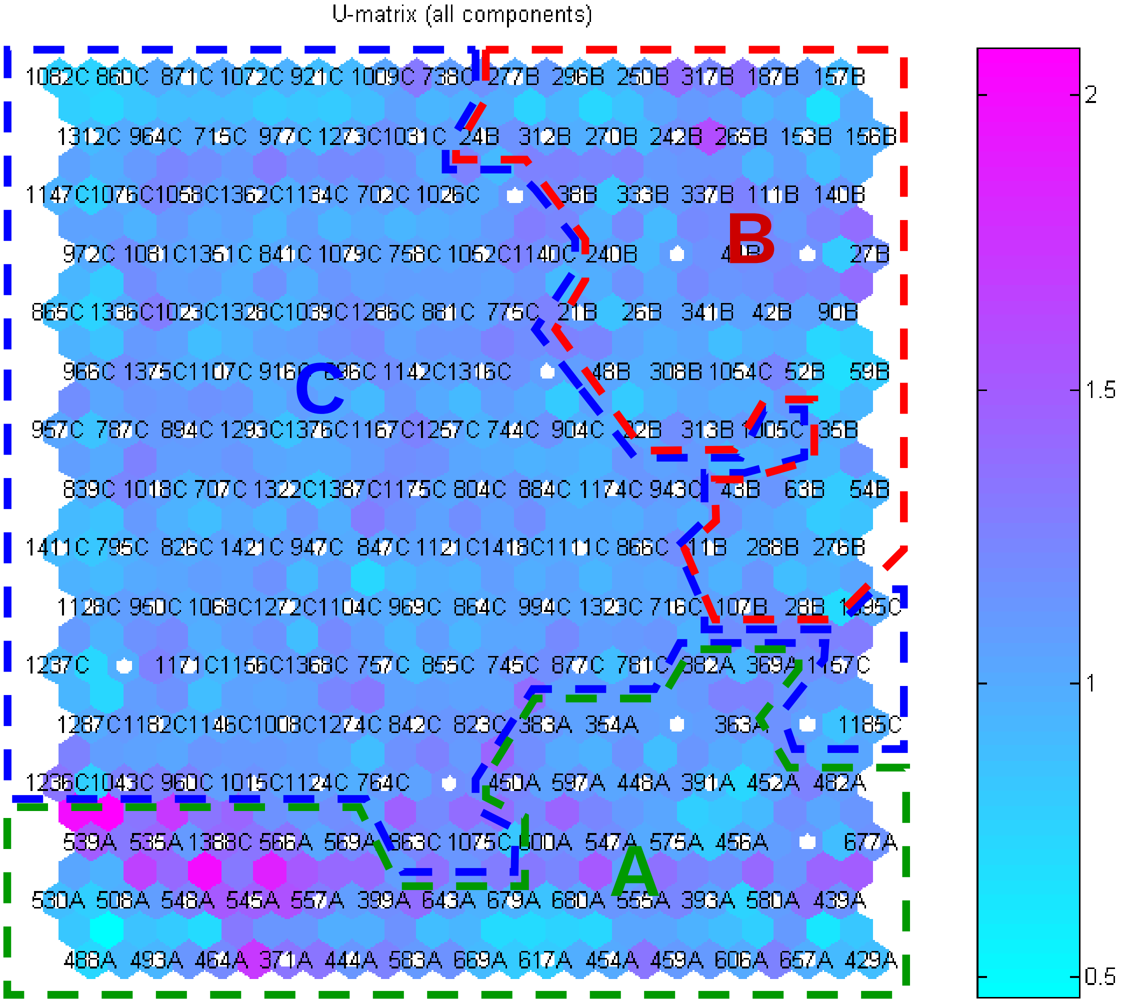

Figure 3: Re-labeled neurons of the DNA testing set on the U-matrix map of the training set.

Radial basis function (RBF) was used as the kernel and the sequential minimal optimization as the algorithm in our support vector machine calculations. Support vector machine based on RBF require the search for optimal values of C and γ. Two sequential searches were performed where the first performs a loose grid search and the second is a local grid search. The region of the plot of loose grid search that gave high prediction accuracy is used for successive local grid search where a more detailed search is carried out. Once the optimal C and γ values are found, actual runs are then performed.

RESULTS AND DISCUSSION

The dataset comprises of 1,424 DNA sequences and each belongs to one of three categories of splice junction site. The first category is marked by the transition from intron to exon and is labeled “A” or “0”; the second category refers to the transition from exon to intron and are labeled “B” or “1”; and the third category represents those that have no transition and are labeled “C” or “0.5” (see Table 1). The dataset was randomly split into a training set of 1,000 sequences and a testing set of 424 sequences. Both the training set and testing set contain approximately the same proportion of sequences belonging to the three output categories, therefore knowledge gained from the training set can be extrapolated on the testing set in the prediction of the type of splice site. Table 1 also summarizes the data distribution by class of the training set and testing set.

Table 1 Summary of the DNA dataset.

Categories	Training set	Test set	Total
Non-gene to gene (Class A)	238	110	348
Gene to non-gene (Class B)	246	99	345
No transition (Class C)	516	215	731

Kohonen's self-organizing map calculation

The calculated U-Matrix of the DNA training set (Figure 2) comprises of three regions corresponding to the three categories of splice junction sites. Class A dominates the lower region of the U-Matrix, while Class B takes up the upper right region, and Class C occupying much of the left region. DNA sequences that have similar patterns are located closely to each other on the map. The output neurons on the U-Matrix map are labeled by the highest frequency categories of the DNA sequences having the particular output neuron as their winning neuron. The map was re-labeled in Figure 3 showing three clusters marked by the dotted lines. The 424 DNA sequences of the testing set are mapped and labeled on their corresponding winning neurons where sequences that are mapped in an incorrect cluster are defined as misclassified.

Back-propagation neural network parameter optimization and calculation

The optimization of the neural network architecture was carried out by determining the parameters by trial-and-error and using the RMS as a measure of prediction performance. Parameters that yielded low RMS were chosen as the optimal value. The averages of five runs were used for each parameter calculations as the seeding of the weight value of the neural network were randomized at the beginning of each run. The optimal parameters that were determined empirically include: the number of nodes in the hidden layer, the number of learning epochs, and the learning rate and momentum. To obtain the optimal parameters training was carried out using 5-fold cross-validation on the training set.

The optimal number of hidden nodes was determined by making a plot (Figure 4) of RMS as a function of the number of nodes in the hidden layer, which was varied from 2 to 32. The optimal number of hidden nodes was found to be 23 since it exhibited the lowest RMS.

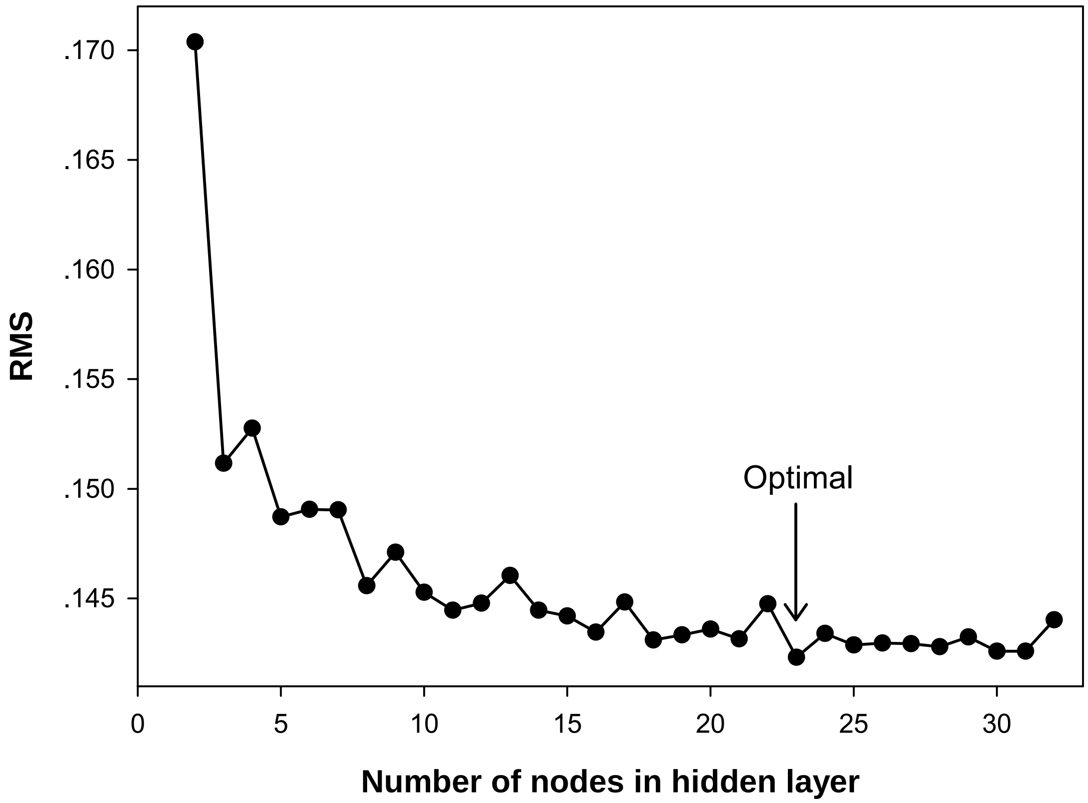

Figure 4: RMS as a function of the number of nodes in the hidden layer for the testing set.

The most favorable learning epoch was determined by plotting RMS as a function of the learning epoch size (Figure 5). RMS was recorded after the first learning epoch and subsequent RMS were calculated after every 10 epochs. The best learning epoch was determined to be 20 epochs.

The optimal learning rate and momentum were chosen from the contour plot of RMS as a function of the learning rate and momentum constants (Figure 6). The best pair of learning rate and momentum is found in the lower left region of the plot to be 0.2 and 0.3, respectively.

Once the optimal parameters were obtained, the assessment of the prediction performance was evaluated on the testing set.

Support vector machine parameter optimization and calculation

The search for the optimal parameters was performed using 5-fold cross-validation using the training set. The optimal value of C and γ was obtained by performing a loose grid search followed by a local grid search as described by Chih-Jen Lin and colleagues (Hsu et al., 2003). Briefly, the region that gave good prediction performance on the loose grid search (Figure 7) was examined in more detail by conducting an exhaustive local grid search (Figure 8). The value of C and γ that gave the best performance was determined to be $2^{0.75}$ and 2^{-5}, respectively.

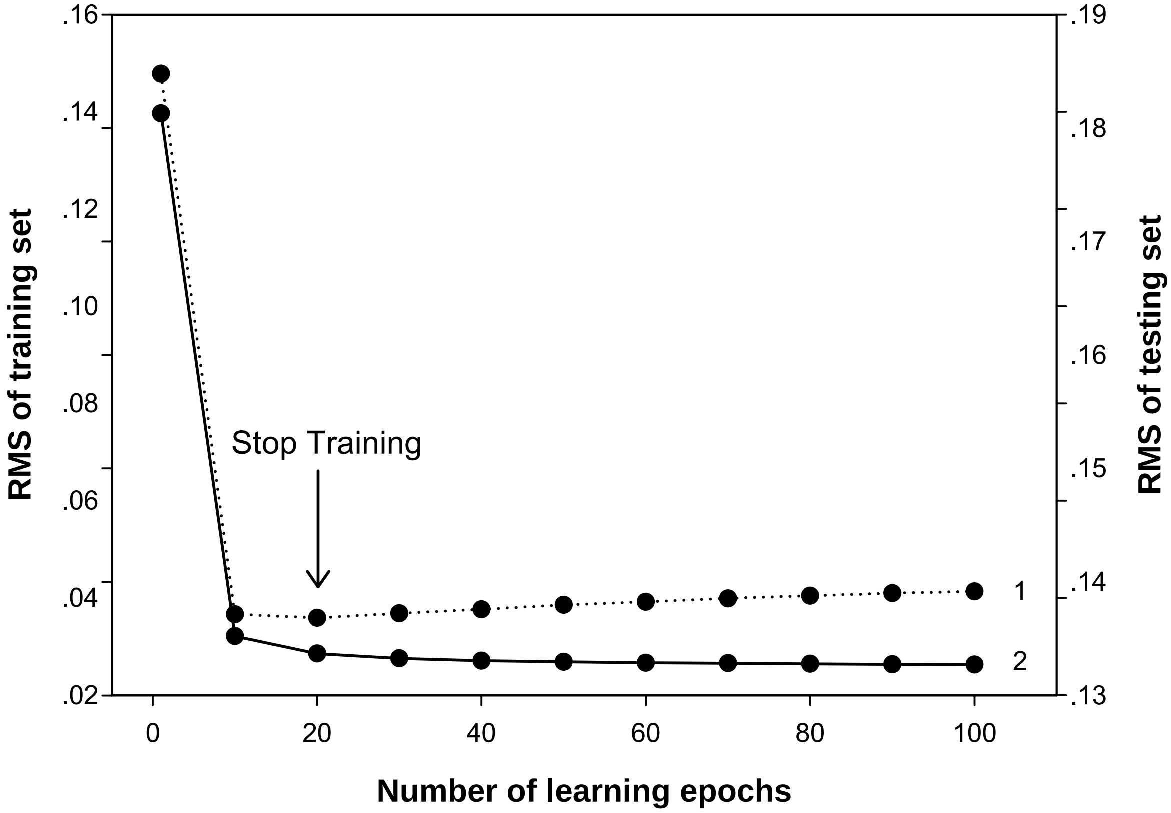

Figure 5: RMS as a function of the number of learning epochs for the testing set. Curves 1 and 2 represent the testing set and training set, respectively.

The empirically determined parameters C and γ that gave the best performance were then evaluated on the testing set.

Prediction of splice junction sites in Human DNA sequences

Each of the machine learning approaches is trained with the training set upon which extrapolation is evaluated on the testing set. It is observed that the supervised learning approach (BNN and SVM) gave better results than the unsupervised learning method (KSOM). It can be seen that SVM outperforms both BNN and KSOM in all three output categories (see Table 2). Out of the total of 424 DNA sequences of the testing set, KSOM made 27 misclassifications, BNN made 13 misclassifications, and SVM made 9 misclassifications that correspond to prediction error of 6.368, 3.066, and 2.123 percent, respectively (see Table 3). For all three learning approaches the prediction performance follows a general trend in which predictions of Class A performed better than Class B. Furthermore, the prediction performance of Class C of KSOM performed poorer than those of BNN and SVM, both of which achieved the same prediction performance. In KSOM, the performance of Class C was better than both Class A and B; for BNN, the performance of Class C was slightly less than Class A but better than Class B; and for SVM, the performance of Class C was lower than those of Class A and B.

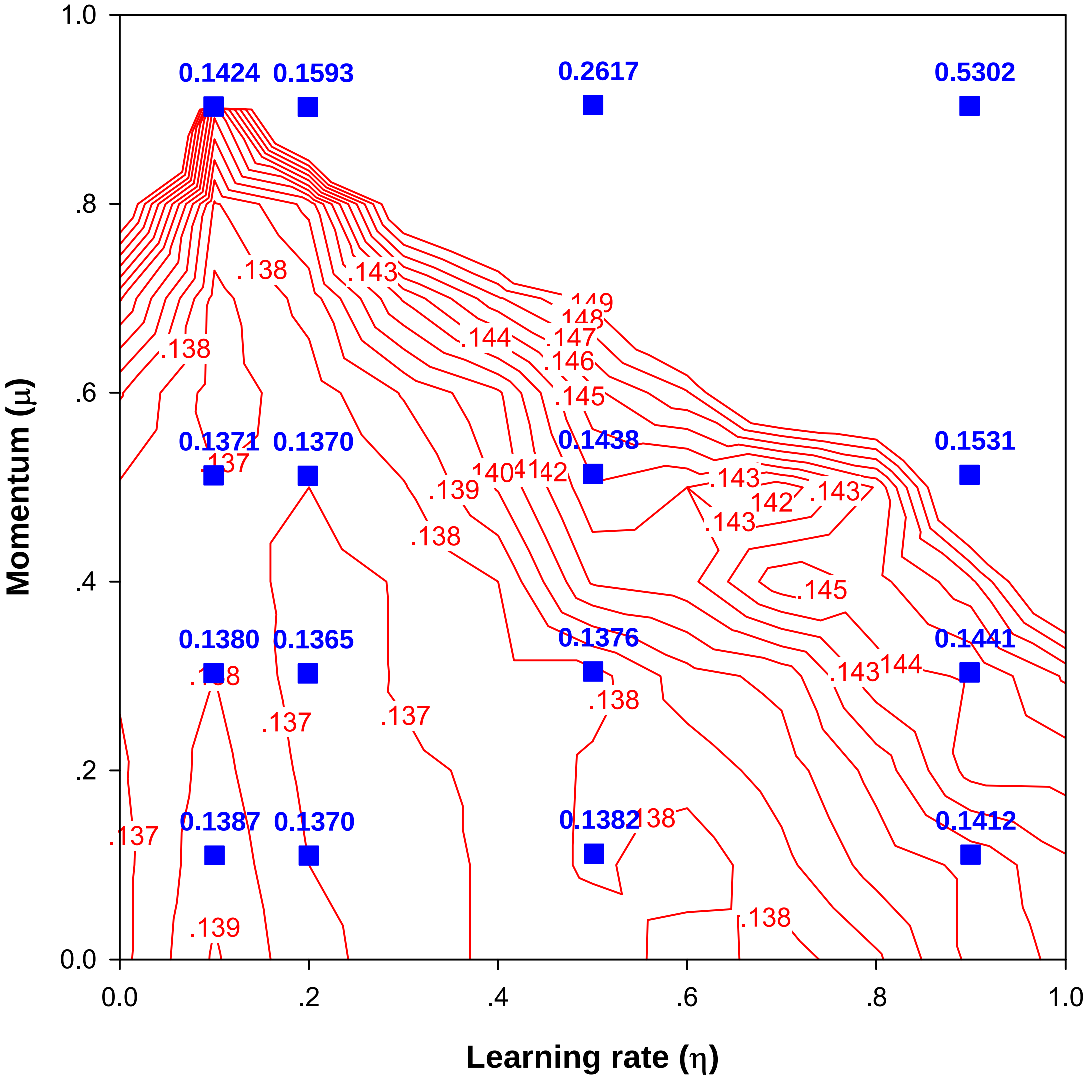

Figure 6: Contour plot of RMS in testing set versus learning rate (*η*) and momentum (*μ*) for the testing set. Each line represents constant value of the RMS, while shaded boxes represent RMS values obtained from the training procedure and fitted onto the same surface model of the contour plot.

Table 2 Prediction accuracy by class of DNA splice junction sites.

Machine learning methods	Prediction accuracy (%)		
	Class A	Class B	Class C
KSOM	92.727	90.909	95.349
BNN	97.273	95.960	97.209
SVM	99.091	97.980	97.209

CONCLUSION

In this study, both supervised and unsupervised machine learning approaches were employed for the recognition of splice junction sites in Human DNA sequences. It was demonstrated that the supervised approaches yielded better prediction than that of the unsupervised approach. Furthermore, all three machine learning approaches demonstrated the same trend in which IE splice sites provided higher prediction accuracy than those of EI splice sites. It was also demonstrated that the Support Vector Machine method holds great potential for the prediction of splice sites in uncharacterized DNA for the elucidation of possible gene products. The methods used in this study could be applied to solve other relevant biological problems in light of the heap of information derived from the Human Genome Project.

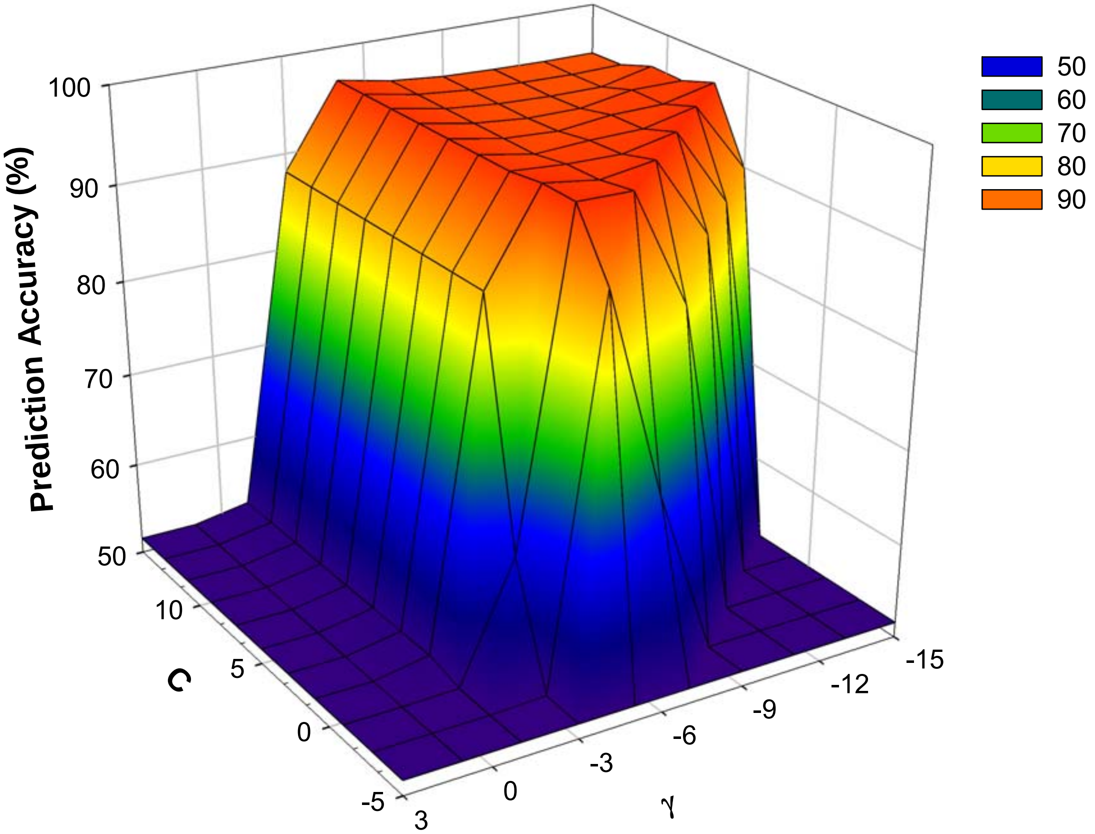

Figure 7: Three-dimensional mesh plot of the loose grid search depicting prediction accuracy as a function of parameters C and γ.

Table 3 Confusion matrix of the three machine learning methods.

Kohonen's self-organizing map (KSOM)

Predicted Class			Actual Class
Class A	**Class B**	**Class C**	
102	2	6	Class A
0	90	9	Class B
4	6	205	Class C

Back-propagation Neural Network (BNN)

Predicted Class			Actual Class
Class A	**Class B**	**Class C**	
107	1	2	Class A
0	95	4	Class B
5	1	209	Class C

Support vector machine (SVM)

Predicted Class			Actual Class
Class A	**Class B**	**Class C**	
109	1	0	Class A
0	97	2	Class B
5	1	209	Class C

Acknowledgements

The authors would like to thank Robert A. Bress for the preparation of the dataset used in this study. C.N. under V.P. supervision is thankful for the Royal Golden Jubilee Ph.D. Scholarship from The Thailand Research Fund. This project was also supported in part by the Thailand Toray Science Foundation (TTSF) and a grant from the annual budget of Mahidol University (B.E.2546-2550).

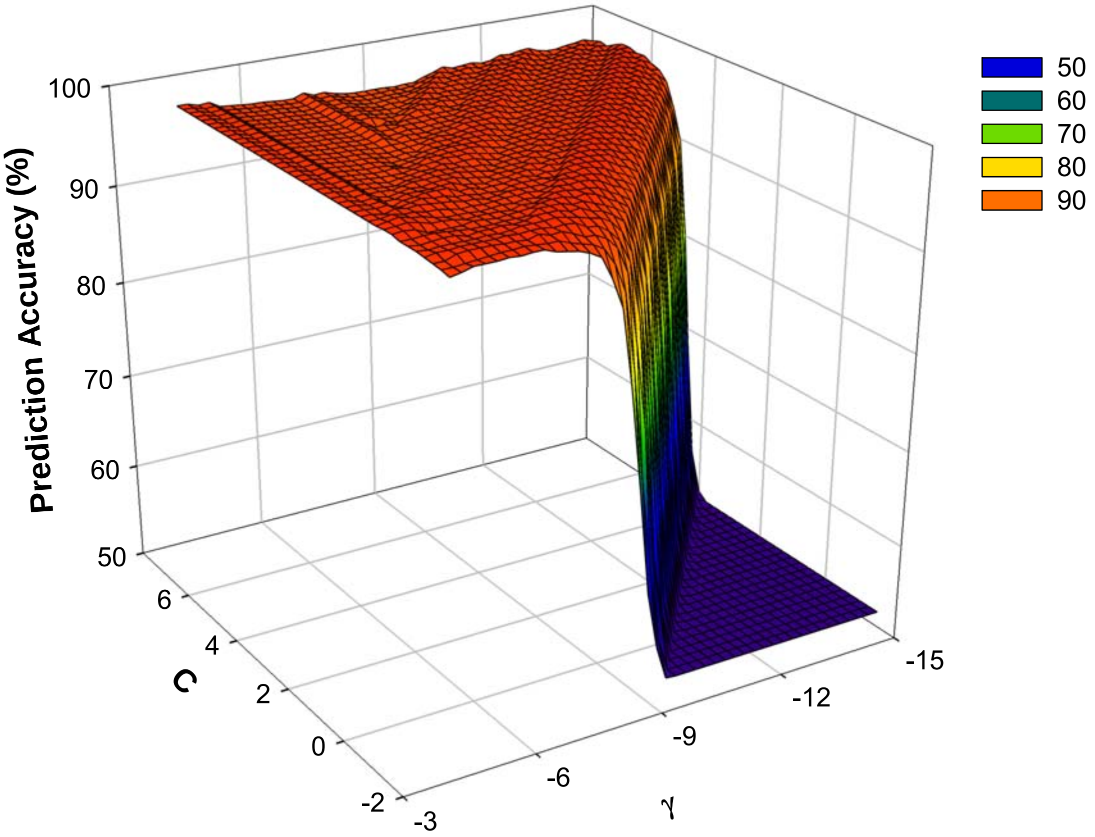

Figure 8: Three-dimensional mesh plot of the local grid search depicting prediction accuracy as a function of parameters C and γ.

REFERENCES

Alhoniemi E, Himberg J, Parhankangas J, Vesanto J. SOM Toolbox, version 2.0 beta, 2005

Bensmail H, Haoudi A. Postgenomics: Proteomics and Bioinformatics in Cancer Research. *J Biomed Biotechnol* 2003;4:217-30

Brent MR, Guigo R. Recent advances in gene structure prediction. *Curr Opin Struct Biol* 2004;14:264-72

Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 1997;268:78-94

Burge C, Karlin S. Finding the genes in genomic DNA. *Curr Opin Struct Biol* 1998;8:346-54

Celniker SE. The Drosophila genome. *Curr Opin Genet Dev* 2000;10:612-6

Chen TM, Lu CC, Li WH. Prediction of splice sites with dependency graphs and their expanded bayesian networks. *Bioinformatics* 2005;21:471-82

Collins FS, Green ED, Guttmacher AE, Guyer MS; US National Human Genome Research Institute. A vision for the future of genomics research. *Nature* 2003;422:835-47

Cooper GM, Hausman RE. The Cell: A Molecular Approach, 2004, Sinauer Associates, Inc, Washington, D.C.

Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, 2004, Cambridge University Press, Cambridge

Dror G, Sorek R, Shamir R. Accurate identification of alternatively spliced exons using support vector machine. *Bioinformatics* 2005;21:897-901

Fedorova L, Fedorov A. Introns in gene evolution. *Genetica* 2003;118:123-31

Han J, Kamber M. Data Mining: Concepts and Techniques, 2001, Morgan Kaufmann, San Francisco

Hannenhalli S, Levy S. Promoter prediction in the human genome. *Bioinformatics* 2001;17:S90-6

Hastings ML, Krainer AR. Pre-mRNA splicing in the new millennium. *Curr Opin Cell Biol* 2001;13:302-9

Hood L, Heath JR, Phelps ME, Lin B. Systems biology and new technologies enable predictive and preventative medicine. *Science* 2004;306:640-3

Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, 2003

Hsu CW, Lin CJ. A comparison of methods for multi-class support vector machines. *IEEE Trans Neural Netw* 2002;13:415–25

Institute for Systems Biology. Health Care in the 21st Century: Predictive, Preventive and Personalized. http://www.systemsbiology.org, 2005

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* 2004;431:931-45

Johnston M. The yeast genome: on the road to the Golden Age. *Curr Opin Genet Dev* 2000;10:617-23

Kaski S, Venna J, Kohonen T. Coloring that Reveals High-Dimensional Structures in Data. *International Conference on Neural Information Processing* 1999;2:729-34

Kohonen T. The Self-Organizing Map. *Neurocomputing* 1998;21:1-6

Kohonen T. Self-Organizing Maps, 2001, Springer, New York

Lesnik EA, Sampath R, Levene HB, Henderson TJ, McNeil JA, Ecker DJ. Prediction of rho-independent transcriptional terminators in *Escherichia coli*. *Nucleic Acids Res* 2001;29:3583-94

Long M, Betran E, Thornton K, Wang W. The origin of new genes: glimpses from the young and old. *Nat Rev Genet* 2003;4:865-75

Pennisi E. Genomic medicine. Gene sequence study takes a stab at personalized medicine. *Science* 2005;308:1102

Pertea M, Lin X, Salzberg SL. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res* 2001;29:1185-90

Platt JC. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, 1998, Microsoft Research MST-TR-98-14

Rätsch G, Sonnenburg S, Schölkopf B. RASE: recognition of alternatively spliced exons in *C. elegans*. *Bioinformatics* 2005;21:i369-77

Reese MG, Eeckman FH, Kulp D, Haussler D. Improved splice site detection in Genie. *J Comput Biol* 1997;4:311-23

Roy SW. Recent evidence for the exon theory of genes. *Genetica* 2003;118:251-66

Rubin GM. The draft sequences. Comparing species. *Nature* 2001;409:820-1

Singer E. Personalized medicine prompts push to redesign clinical trials. *Nat Med* 2005;11:462

Snyder EE, Stormo GD. Identification of protein coding regions in genomic DNA. *J Mol Biol* 1995;248:1-18

Venter JC, Adams MD, et al. The sequence of the human genome. *Science* 2001;291:1304-51

Weston AD, Hood L. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *J Proteome Res* 2004;3:179-96

Williams I, Richardson J, Starkey A, Stansfield I. Genome-wide prediction of stop codon readthrough during translation in the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res* 2004;32:6605-16

Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations, 2000, Morgan Kaufmann, San Francisco

Zhang L, Luo L. Splice site prediction with quadratic discriminant analysis using diversity measure. *Nucleic Acids Res* 2003;31:6214-20

Zheng CL, de Sa VR, Gribskov M, Nair TM. On selecting features from splice junctions: an analysis using information theoretic and machine learning approaches. *Genome Inform Ser Workshop Genome Inform* 2003;14:73-83

Zhu HQ Hu GQ, Ouyang ZQ, Wang J, She ZS. Accuracy improvement for identifying translation initiation sites in microbial genomes. *Bioinformatics* 2004;20:3308-17

Zupan J, Gasteiger J. Neural Networks in Chemistry and Drug Design, 1999, Wiley-VCH, Weinheim