Predicting single nucleotide polymorphisms (SNP) from DNA sequence by support vector machine

Waiming Kong, and Keng Wah Choo

Bioinformatics Group, Nanyang Polytechnic,180 Ang Mo Kio Ave 8, S(569830), Singapore

TABLE OF CONTENTS

1. ABSTRACT

Recently, SNP has gained substantial attention as genetic markers and is recognized as a key element in the development of personalized medicine. Computational prediction of SNP can be used as a guide for SNP discovery to reduce the cost and time needed for the development of personalized medicine. We have developed a method for SNP prediction based on support vector machines (SVMs) using different features extracted from the SNP data. Prediction rates of 60.9 % was achieved by sequence feature, 59.1% by free-energy feature, 58.1% by GC content feature, 58.0% by melting temperature feature, 56.2% by enthalpy feature, 55.1% by entropy feature and 54.3% by the gene, exon and intron feature. We introduced a new feature, the SNP distribution score that achieved a prediction rate of 77.3%. Thus, the proposed SNP prediction algorithm can be used to in SNP discovery.

2. INTRODUCTION

A single nucleotide polymorphism (SNP) is defined as a location in the human genome that is different from one individual to another. Most of the SNPs occur in the human genome where they do not affect any gene or protein. However, SNPs found in any gene region may have some effects on the function of the gene. In the case of the sickle-cell anemia disease, red-blood cells are found to have sickle shape that caused them to be removed by the body. This resulted in less oxygen supply for the body.

SNPs are useful as genetic markers because of their frequency and stability in the human genome. Their frequent occurrence provides a large source of genetic markers that are more likely to be located close to target genes of interest. SNPs that are located close to genes tend to be inherited together over many generations. SNPs that frequently differ in individuals with a disease compared with individuals without the disease act as beacons to warn scientists that a disease susceptibility gene may be nearby.

Recently, SNPs have been receiving a lot of attention as it is discovered that the SNP patterns of patients will result in different responses to medicine. Researchers are trying to associate SNPs with diseases to predict which group of people are more susceptible to certain diseases and to work on prevention. By associating SNPs with medicine, researchers are exploring personalized drugs for individual based on the genetic composition.

Due to the importance of SNPs, many genotyping techniques are developed, including DNA arrays (3,10), mass spectrometry (6), DNA melting analysis (9), denaturing gradient gel electrophoresis and denaturing HPLC. Although, the cost of SNP genotyping is decreasing, SNP prediction algorithm can be used as a initial test to prioritize the regions where SNPs are likely to be found and this will result in even lower cost for SNP genotyping.

In this paper, Support Vector Machine (SVM) is used for the SNP prediction. SVM, an effective method for general purpose supervised pattern recognition (13,14), has been applied successfully to many biological data recently, including the identification of unknown genes using the gene expression data from DNA microarray hybridization experiments by Brown et al.(1), classification of ovarian cancer tissue (4) and prediction of protein structural test (3). In addition, SVM was also used for searching translation initiation sites (16) and for splice site recognition (11). We trained the SVM with positive and negative SNP data using the sequence feature, melting temperature feature , free-energy feature, GC content feature, entropy feature, enthalpy feature and gene, exon and intron feature and obtained prediction results of 54.3% - 60.9%. Finally, we introduce the SNP distribution score feature that is able to give prediction rate of 77.3%.

3. MATERIALS AND METHODS

3.1. Data Generation

The positive SNPs are obtained from the Japan SNP database (5) as the data are wet-lab verified and well organized. The data come with a standard flanking sequence length of 60bp upstream and 60bp downstream. This resulted in a total sequence length of 121bp. 10000 SNPs are randomly selected from the Japan SNP database as positive SNPs.

A negative SNP is defined as a continuous stretch of nucleotide sequence where no SNP has been reported. There is a possibility that SNP might occur in the sequence but not detected, however the probability is quite low. In this paper, we have extracted randomly 10000 nucleotide sequences of 121 nucleotides length to be negative SNPs. The negative SNPs are randomly chosen over the 24 chromosomes with the condition that there must not be any SNP within 60 positions on either side of the chosen nucleotide.

A positive or negative SNP data refers to the chosen nucleotide and the immediate left and right neighbors. Most of the features are generated with lengths of 3, 7 and 11 nucleotides. A length of 3 nucleotides refers to the SNP and the immediate left and right neighbor.

3.2. Features used in SVM Training

The features used in this paper are free energy, entropy, enthalpy, GC content, melting temperature, SNP distribution score, sequence, gene, intron and exon features. Many features can be extracted from the DNA sequences, such as the GC content and sequence information. Other features required additional transformations applied to the sequence to create information used in the SNP prediction, such as the melting temperature, free energy, entropy and enthalpy. These features can be used together or separately. The features are calculated in the following sections.

3.2.1. Free Energy

Free energy (ΔG) of a reaction is a measure of its spontaneity, and of how much energy is released or required to drive the reaction. In this paper, we used lengths of 3, 7 and 11 nucleotides to generate 3 sets of values for the free energy. In the example of 3 nucleotides, free energy of AAT = free energy of AA + free energy of AT+ energy(Initiation) = -1.02 -0.73+2.8 = 1.05 kcal/mol

3.2.2. Entropy

Entropy (ΔS) is a measure of the randomness or disorder in either the system or the surroundings. The calculation of entropy for 3, 7 and 11 nucleotides are similar to the free energy using the parameters in Table 1 for entropy.

3.2.3 Enthalpy

Enthalpy (ΔH) is the heat/energy released or consumed during a chemical reaction due to differences between the chemical bond energies of the products and reactants of the reaction. The calculation of enthalpy for 3, 7 and 11 nucleotides are similar to the free energy using the parameters in Table 1 for enthalpy.

3.2.4. Melting Temperature

Melting temperature is defined as the temperature at which 50% of the oligonucleotide and its perfect complement are in duplex. The melting temperature is calculated using Wallace rule (15). The melting temperature is calculated for 3 , 7 and 11 nucleotides with the equation below.

Melting Temperature = 4 x (#C + #G) + 2 x (#A + #T) °C
For example, if the sequence is ATG, #C=0, #G=1, #A=1, #T=1, Melting temperature = 4x(0+1)+2x(1+1)=8 °C.

3.2.5. GC Content

The GC content of genomic DNA is defined as the mean percentage of guanine (G) and cytosine (C). For example, a sequence of 10 bases, ATGTACCCCG, will have GC content of 60%. GC rich regions melt at higher temperatures than regions that are AT rich. The GC content is calculated for 3,7 and 11 nucleotides.

Table 1. Thermodynamic Parameters for DNA Helix Initiation and Propagation

DNA sequence	Free Energy (ΔG)	Entropy (ΔS)	Enthalpy (ΔH)
AA/TT	-1.02	-23.6	-8.4
AT/TA	-0.73	-18.8	-6.5
TA	-0.6	-18.5	-6.3
CA	-1.38	-19.3	-7.4
GT	-1.43	-23	-8.6
CT	-1.16	-16.1	-6.1
GA	-1.46	-20.3	-7.7
CG	-2.09	-25.5	-10.1
GC	-2.28	-28.4	-11.1
GG	-1.77	-15.6	-6.7
Initiation at GC[b]	+1.82	-5.9	0
Initiation at AT[c]	+2.8	-9	0

(The data below is extracted from (17)). [b]Initiation parameter for duplexes that contain at least one GC base pair. [c]Initiation parameter for duplexes that contain only AT base pairs.

Table 2. Numeric code for sequence feature

A	0 0 0 1
C	0 0 1 0
T	0 1 0 0
G	1 0 0 0

Table 3. Calculation of Sensitivity and Specificity

Test Results	**Disease**	
	+	-
+	Number of true positive	Number of false positive
-	Number of false negative	Number of true negative

Sensitivity = True Positive / (True Positive + False Negative), Specificity = True Negative / (True Negative + False Positive)

GC content for ATG = (No of G & C)/(Number of nucleotides)

= (0+1)/(3)

=0.33

3.2.6 SNP Distribution Score

The human genome is divided into equal segments of fixed predetermined length. The SNP distribution score measures the number of SNP found in these fixed segments of the DNA where the chosen nucleotide position is situated. The SNP distribution score is calculated for segment lengths of 10000, 50000 and 100000 nucleotides. The SNP distribution score parameter is calculated as follows:

I. The human genome is divided into smaller segments of predetermined length (10000,50000 and 100000 nucleotides lengths are used in this paper).

Ii. The number of SNPs located in each region is recorded for all the segments in the entire human genome.

III. The position of the SNP data is used to locate the segment that it belongs to and the number of SNPs found in the segment is recorded as the SNP distribution score parameter.

3.2.7. Sequence

Sequence feature contains the actual nucleotide sequence of 7 nucleotides length. The nucleotides are converted to numeric codes based on the table below. For example, ACTACTA is represented by 0001001001000001001001000001.

3.2.8. Gene, Intro & Exon

The gene parameter is given a value of 1 if the chosen nucleotide position is found in a gene and 0 otherwise. The intron and the exon parameters are calculated in the same manner.

3.3. Accuracy of Diagnostic Tests

Accuracy of a diagnostic test can be expressed through sensitivity and specificity. Sensitivity refers to the ability of a certain diagnostic test to detect a particular disease. It is expressed as the probability of testing positive if the particular disease is truly present, i.e., the probability of having both a positive test and a positive diagnosis. Hence a test with 98% sensitivity means that 98% of those with the disease will test positive. Specificity, on the other hand, refers to the probability of testing negative if the disease is truly absent. In other words, 98% specificity means that 98% of those who are truly negative for the disease or problem will have a negative test while 2% of them will have a false positive test. See Table 3 for calculation.

3.4. Support Vector Machines

SVM is a learning algorithm that was developed by Vapnik (13,14). The basic idea behind SVM is the formation of a linear classifier to separate the positive and negative training data. This linear classifier is then used on the test data. Depending on the position of the point falling on the positive or negative side of the linear classifier, the prediction result is calculated. In real world situation, a linear classifier may not be able to separate the positive and negative data sets. To overcome this limitation, kernels are used to transform the data into a feature space where the linear classifier can be used for the separation. Choosing the right kernel to use is therefore essential for the separation in the feature space.

3.5. Design and Implementation

SVMLight (7,8) was used to implement the SVM. The training sets consist of 2500 SNPs, each for the positive and negative sets. The SVM is trained with the training sets using linear and radial basis function (RBF) kernels with different C values (trade-off between training error and margin). The result is repeated by reversing the training and test sets. The experiment is further repeated with another group of training and test sets containing 2500 SNPs each. Table 4 shows 4 sets of results generated by reversing the test and training data. The sensitivity and the specificity are calculated from the prediction results. Table 4 shows the best results obtained with different features using different kernels and parameters.

4. RESULTS

The results from the SNP prediction in Table 4 can be broadly divided into 3 main groups. The worst performance comes from enthalpy, entropy and gene, exon and intron features with an average percentage of 54-56%.

Table 4. Prediction results for different features and with selected parameters

Features	Kernel	gamma	C	Sensitivity	Specificity	Average
Free Energy 1	RBF	0.5	0.05	0.5908	0.6156	0.591
Free Energy 2	RBF	0.5	0.5	0.568	0.6156	
Reverse Free Energy 1	RBF	0.5	0.5	0.5636	0.6024	
Reverse Free Energy 2	RBF	0.5	0.2	0.574	0.5956	
GC Content 1	RBF	0.0001	0.05	0.5032	0.6752	0.581
GC Content 2	RBF	0.0001	0.05	0.4928	0.6708	
Reverse GC Content 1	RBF	0.0	0.1	0.5716	0.5832	
Reverse GC Content 2	Linear	0	0.1	0.5584	0.596	
Gene, Exon & Intron1	RBF	0.0001	0.05	0.4108	0.6816	0.543
Gene, Exon & Intron2	RBF	0.0001	0.05	0.396	0.6836	
Reverse Gene, Exon & Intron1	RBF	0.0001	0.05	0.4068	0.6752	
Reverse Gene, Exon & Intron2	RBF	0.0001	0.05	0.41	0.6836	
Enthalpy 1	RBF	0.001	0.2	0.2792	0.8444	0.562
Enthalpy 2	RBF	0.01	0.5	0.2588	0.8776	
Reverse Enthalpy 1	RBF	0.1	0.05	0.2768	0.8428	
Reverse Enthalpy 2	RBF	0.1	0.5	0.2328	0.8868	
Melting Temperature1	Linear	0	0.1	0.634	0.5548	0.58
Melting Temperature2	Linear	0	0.05	0.6088	0.542	
Reverse Melting Temperature1	Linear	0	0.05	0.524	0.6228	
Reverse Melting Temperature2	RBF	0.0001	0.2	0.512	0.6384	
Entropy 1	RBF	0.5	0.2	0.5804	0.5424	0.551
Entropy 2	RBF	0.5	0.5	0.586	0.514	
Reverse Entropy 1	RBF	0.5	0.5	0.6052	0.518	
Reverse Entropy 2	RBF	0.5	0.5	0.552	0.5292	
Sequence 1	RBF	0.5	0.2	0.5716	0.644	0.609
Sequence 2	RBF	0.5	0.1	0.5968	0.6308	
Reverse Sequence1	RBF	0.5	0.05	0.608	0.604	
Reverse Sequence2	RBF	0.1	0.2	0.5652	0.6516	
SNP Distribution Score 1	RBF	1	0.5	0.8392	0.7148	0.773
SNP Distribution Score 2	RBF	0.1	0.1	0.858	0.668	
Reverse SNP Distribution Score 1	RBF	1	0.5	0.8192	0.726	
Reverse SNP Distribution Score 2	RBF	0.1	0.5	0.818	0.7404	
Combined Energy1	RBF	0.01	0.5	0.542	0.6756	0.599
Combined Energy 2	RBF	0.0001	0.2	0.4688	0.7196	
Reverse Combined Energy 1	RBF	0.01	0.5	0.5124	0.6904	
Reverse Combined Energy 2	RBF	0.001	0.5	0.502	0.6796	
All1	RBF	0.01	0.5	0.782	0.7208	0.759
All2	RBF	0.01	0.5	0.8136	0.7184	
Reverse All1	RBF	0.001	0.5	0.7668	0.7516	

Abbreviations: RBF: Radial Basis Function.

The next higher group consists of free energy, GC content and melting temperature with average prediction of 58-59%. Better performance comes from the sequence feature with prediction rate of 60.9 % while the SNP distribution score gives prediction rate of 77.3 %. The high prediction rate indicates that large number of SNPs nearby is a good indicator for the presence of SNP. Of the 3 energy features, free energy features gives slightly higher prediction rate than enthalpy and entropy. However the prediction rates of the these 3 energy features are low when compared to other features showing that the relationship between energies and SNP is not strong. Surprisingly, gene, intron and exon feature do not have a close relationship to SNP as seen by the low rate of prediction. Sequence feature gives an prediction rate of 60.9 % which suggests that SNP is more closely related to the sequence than to the energy and gene features. Finally, the best result is achieved by the SNP distribution score which shows that the number of SNPs in the surrounding region does indicate whether a SNP is likely to be present. The combined energy feature combined the free-energy, entropy and enthalpy features together and the prediction rate improved to 59.9 %. By combining all the available features together, the SVM is able to give a prediction rate of 75.9 % .

5. CONCLUSION

SNP will be playing a important role in improving the health care of tomorrow. Through the SVM prediction, we are able to show the relationship between SNP and the different features. It is demonstrated in this paper that SNP can be predicted with a reasonable high level of accuracy using the SNP distribution score feature. The prediction algorithm can be used as early indicator of the presence of SNP. This will allow the researchers to set priority on which regions to search for SNP and will result in lower cost and shorter time for SNP discovery.

6. REFERENCES

1. Brown MPS, W.N. Grundy, D. Lin, N. Cristianini , C.W. Sugnet, T.S. Furey , M. Jr. Ares, D. Haussler: Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Nat Acad Sci USA* 97: 262-267 (2000)

2. Cai Yu-Dong, Xiao-Jun Liu, Xue-biao Xu and Guo-Ping Zhou: Support Vector Machines for predicting protein structural class. *BMC Bioinformatics* 2:3 (2001)

3. Fan J.B., X. Chen, M.K. Halushka, A. Berno, X. Huang, T. Ryder: Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Res*, 10:853–60 (2000)

4 Furey T. S., N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler: Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906-914 (2000)

5. Hirakawa M., T. Tanaka, Y. Hashimoto, M. Kuroda, T. Takagi, and Y. Nakamura: JSNP: a database of common gene variations in the Japanese population. *Nucleic Acid Res*, 30:158-162 (2002)

6. Jackson P.E., P.F. Scholl, J.D. Groopman : Mass spectrometry for genotyping: an emerging tool for molecular medicine. *Mol Med Today*, 6:271–6 (2000)

7. Joachims T.: Making large-scale SVM learning practical. In: scholkppf, B.,Burges, C.,Smola, A. (Eds.), Advances in Kernel Methods-Support Vector Learning. MIT Press, Cambridge, MA (1999)

8. Joachims T.: *Proceedings of the International Conference on Machine Learning* (1999)

9. Lipsky R.H., C.M. Mazzanti, J.G. Rudolph, K. Xu, G. Vyas, D. Bozak: DNA melting analysis for detection of single nucleotide polymorphisms. *Clin Chem*, 47:635–44 (2001)

10. Pastinen T., M. Raitio, K. Lindroos, P. Tainola, L. Peltonen, A.C. Syvanen: A system for specific, high-throughput genotyping by allelespecific primer extension on microarrays. *Genome Res*, 10:1031–42 (2000)

11. Sonnenburg S., G. Rätsch, A. Jagota, K.R. Müller: New Methods for Splice Site Recognition. *Proceedings of the International Conference on Artifical Neural Networks* (2002)

12. Hua Sujun and Zhirong Sun: Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721-728 (2001)

13. Vapnik V.: The Nature of Statistical Learning Theory. Springer, Berlin (1995)

14. Vapnik V.: Statistical Learning Theory. Wiley-Interscience, New York (1998)

15. Wallace R.B., J. Shaffer, R.F. Murphy, J. Bonner, T. Hirose, K. Itakura: *Nucleic Acids Res*, 6, 3543 (1979)

16. Zien A., G. Ratsch, S. Mika, B. Scholkopf, T. Lengauer, K.R. Muller: Engineering support vector machine kernels that recognize translation initiation sites. *BioInformatics*, 16(9):799-807 (2000)

17. Santa-Lucia J. Jr., H.T. Allawi, P.A. Seneviratne: Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, 35, 3555-62 (1996)

Key Words: SVM, SNP, SNP prediction

Send correspondence to: Dr Waiming Kong, Bioinformatics Group, Nanyang Polytechnic,180 Ang Mo Kio Ave 8, S(569 830), Singapore, Tel: 65-6550-0441, Fax: 65-6452-0400, E-mail: KONG_Wai_Ming@nyp.gov.sg

http://www.bioscience.org/current/vol12.htm