정규상호정보와 지지벡터기계를 이용한 천식 관련 단일염기다형성 유전형 자료 분석
(Analysis of Asthma Related SNP Genotype Data Using Normalized Mutual Information and Support Vector Machines)

이중섭 [†] (Jungseob Lee)   김승현 [††] (Seung-Hyun Kim)   신기섭 [†††] (Ki-Seob Shin)   임규철 [†††] (Kyucheol Lim)

요 약

서론

천식에는 아스피린 과민증 (aspirin hypersensitivity)에 따라 아스피린 불내성 (aspirin intolerant asthma, AIA)과 내성 천식 (aspirin tolerant asthma, ATA) 두 가지 유형이 있다. 천식과 관련된 유전적 위험 요인들은 집중적으로 또한 광범위하게 연구되고 있다. 그러나 단일염기다형성들의 조합의 효과에 대해서는 거의 평가되지 않았다. 본 논문에서는 두 유형의 천식 진단에 유용한 단일염기다형성의 최상의 조합을 찾는다.

방법

본 논문에서는 246명의 천식환자들을 조사하였다. (94명은 아스피린 불내성 천식을 152명은 아스피린 내성 천식을 가지고 있다) 그리고 천식과 관련된 것으로 추측되는 25개의 단일염기다형성들을 분석하였다. 단일염기다형성의 조합의 정규상호정보 값을 계산하여 높은 정규상호정보 값을 갖는 단일염기다형성들의 조합을 선택하고 선택된 조합들의 예측 정확도를 지지벡터기계를 사용하여 계산하였다.

결과

최상의 조합은 4개짜리이고 ALOX5_p1_1708, B2ADR_q1_46, CCR3_p1_520, CysLTR1_p1_634로 구성된 모델이다. 이것은 0.053의 정규상호정보 값과 71.14%의 ATA 질병에 대한 예측 정확도를 갖는다.

키워드 : 천식, 단일염기다형성, 상호정보, 지지벡터기계

Abstract

Introduction

There are two types of asthma according to aspirin hypersensitivity: aspirin intolerant asthma (AIA) and aspirin tolerant asthma (ATA). The genetic risk factors that are related with asthma have been investigated intensively and extensively. However the combinatory effects of single nucleotide polymorphisms (SNPs) have hardly been evaluated. In this paper we searched the best set of SNPs that are useful to diagnose the two types of asthma.

Methods

We examined 246 asthmatic patients (94 having aspirin intolerant asthma and 152 having aspirin tolerant asthma) and analyzed 25 SNPs typed in them, which are suspected to be associated with asthma. Normalized mutual information values of combinations of typed SNPs are calculated, and those with high normalized mutual information values are selected. We use support vector machines to evaluate the prediction accuracy of the selected combinations.

· This study was supported by grants from the Basic Research Program of the Korea Science and Engineering Foundation(R01-2006-000-10775-0).

[†] 정 회 원 : 아주대학교 자연과학부 교수
jslee@ajou.ac.kr
[††] 비 회 원 : 아주대학교 의과대학 교수
kimsh@ajou.ac.kr
[†††] 비 회 원 : 아주대학교 수학과
k1430@ajou.ac.kr
kyucheol@ajou.ac.kr

논문접수 : 2008년 10월 2일
심사완료 : 2009년 8월 3일

Results

The best combination model turns out four-locus and consists of ALOX5_p1_1708, B2ADR_q1_46, CCR3_p1_520, CysLTR1_p1_634. Its normalized mutual information value is 0.053 and the accuracy in predicting ATA disease risk among asthmatic patients is 71.14%.

Key words : asthma, single nucleotide polymorphism, mutual information, support vector machine

1. Introduction

It has been thought that finding a relationship between DNA sequence variations and susceptibility to a disease can make significant improvement in diagnosing and treating it. The collection of single nucleotide polymorphisms (SNPs) variants that an individual possesses in a number of key genes are assumed to play an important role in conferring variability to drug response[1] and susceptibility to human disease. Therefore, association studies using SNPs have been advocated in common complex diseases and traits. In [2], Kim et al used the multifactor dimensionality reduction (MDR) method to identify the best model predicting AIA disease risk among asthmatic patients. In this paper, we exploit normalized mutual information and support vector machines to predict ATA disease risk among them. First using normalized mutual information values, we reveal sets of SNPs that separate phe-onotypically distinct classes of the sample according to their genotype signatures. Then the method of support vector machine is used to evaluate the prediction accuracies of the selected sets of SNPs.

The main contribution of the current work is to provide an efficient method for selecting a set of SNPs that is associated with disease phenotype.

2. Data

Twenty-five SNPs were selected for analysis and the genotype data of 246 asthmatics (94 with AIA and 152 with ATA) are provided by Ajou University Medical Center. The mean age and sex proportion between AIA and ATA groups are not different. Table 1 shows the names of SNPs analyzed in this paper. For their detailed informations, see [2].

Table 1 Candidate SNPs associated with asthma

ID	SNP symbol
SNP1	ALOX5_p1_1708_G>A
SNP2	B2ADR_q1_46_A>G
SNP3	B2ADR_q2_79_C>G
SNP4	CCR3_p1_520_T>G
SNP5	CCR3_p2_174_C>T
SNP6	CysLTR1_p1_634 C>T
SNP7	CysLTR2_q1_2079_C>T
SNP8	CysLTR2_q2_2534_A>G
SNP9	FCER1B_p1_109_T>C
SNP10	FCER1B_q1_237_A>G
SNP11	IL10_p1_1082_A>G
SNP12	IL10_p2_819_T>C
SNP13	IL13_p1_1510_A>C
SNP14	IL13_p2_1055_C>T
SNP15	IL18_p3_137_G>C
SNP16	LTC4S_p2_444_A>C
SNP17	NAT2_q5_197_G>A
SNP18	NAT2_q7_286_G>A
SNP19	TBXA2R_q1_795_T>C
SNP20	TGFB_p1_509_C>T
SNP21	TNFA_p1_1031_T>C
SNP22	TNFA_p2_863_C>A
SNP23	TNFA_p3_857_C>T
SNP24	TNFA_p5_308_G>A
SNP25	TNFA_p6_238_G>A

3. Methods

3.1 Normalized mutual information

Let S be a whole data set with $|S|$ elements. Consider a partition $X=\{X_1, X_2, \ldots, X_n\}$ of S such that $X_i \cap X_j = \varnothing$, $|X_1| = d_1$, $|X_2| = d_2, \ldots,$ $|X_n| = d_n$ and $|S| = \sum_{i=1}^{n} |X_i|$. For our study, the whole data set S consists of asthma patients and X is divided into two groups by the asthma types. We want to score each SNP according to its relevance to the partition X for S. The Shannon entropy of the partition $X=\{X_1, X_2, \cdots, X_n\}$ is defined as

$$H(X) = -\sum_{i=1}^{n} \frac{d_i}{|S|} \log \frac{d_i}{|S|}.$$

Let Y be another partition for S which is induced by the genotype. For example, if Y is induced by the SNP ALOX5_p1_1708_G>A then Y consists of three members according to the geno-types GG, AG and AA. We are interested in the interaction between X and Y. This measure is defined by mutual information (MI) [3,4].

$$\begin{aligned} MI(X;Y) &= H(X)+H(Y)-H(X\vee Y) \\ &= H(X)-H(X|Y) \\ &= H(Y)-H(Y|X) \end{aligned}$$

where $H(X|Y)$, the conditional entropy, is the remaining uncertainty about X with the knowledge of Y.

This score measures the amount of information that the genotype at the locus gives about membership in each of the sets $X_1, X_2, \cdots, X_n$.

However, we find out conventional mutual information has bias when the number of samples in some partition members of X or Y is relatively large compared with the others. In fact, if the partition X and/or Y are not relatively equally distributed, $H(X)$ and/or $H(Y)$ may be small, and then $MI(X;Y)$ can be quite small even if X and Y are highly correlated. Thus, the effect of original genotype partition is not ignorable when it comes to ordering and selecting good attributes. To resolve this problem, we normalized the conventional mutual information by dividing it by the sum of the entropies of two partitions. Equation (1), named normalized mutual information (NMI) [5,6], is the new scoring method used in our analysis.

$$NMI(X;Y)=\frac{MI(X;Y)}{H(X)+H(Y)}. \qquad (1)$$

We notice that the distribution of the NMI values is more normal then that of the MI values as shown Figure 1.

The mutual Information and normalized mutual information can be defined for any number of partitions. Let Y_1, Y_2,..., Y_k be partitions for S. We define

$$\begin{aligned} MI(X;\vee_{i=1}^{k} Y_i) = H(X) + H(\vee_{i=1}^{k} Y_i) \\ - H(X \vee (\vee_{i=1}^{k} Y_i)) \end{aligned}$$

and

$$NMI(X;\vee_{i=1}^{k} Y_i)=\frac{MI(X;\vee_{i=1}^{k} Y_i)}{H(X)+H(\vee_{i=1}^{k} Y_i)}. \qquad (2)$$

The amount of information that k combinations of SNPs give about membership in each component of the partition X is expressed by Equation (2). By examining this combinatorial result, it is possible to observe some synergistic effect of SNPs on types of asthma.

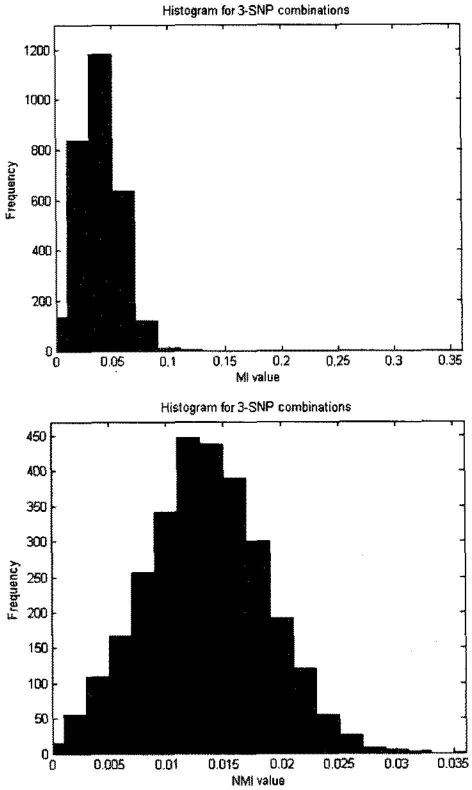


Figure 1 Comparison of histograms for MI and NMI

3.2 Support Vector Machines

The support vector machine (SVM) is a supervised learning method used for pattern classification and regression. A property of SVM is that it simultaneously minimizes the empirical classification error and maximizes the geometric margin. Hence, it is known as maximum margin classifier.

Suppose that the data set X consists of a series of objects $x_1, x_2, ..., x_N$ together with a series of labels associated with the objects

$$y_1, y_2, ..., y_N \in \{-1, 1\}.$$

In linear separable case, we find the optimal hyperplane by minimizing the cost function

$$\phi(w)=\frac{1}{2}w^T w=\frac{1}{2}\parallel w \parallel^2$$

such that

$$y_i(w^T x_i + b) \ge 1 \text{ for } i=1,2,...,N.$$

In non-linear separable case, we take the Lagrangian in the usual manner:

$$L(w,\zeta,b,a)=\frac{1}{2}w^T w + C\sum_{i=1}^{N}\zeta_i + \sum_{i=1}^{N}\alpha_i[1-\zeta_i - y_i(x_i^T w+b)].$$

Since the Lagrangian function is linear in α, we can't set the gradient with respect to α to zero. We obtain the following dual optimization problem:

$$Max : Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

such that

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \text{ and } 0 \le \alpha_i \le C \text{ for } i = 1, 2, \ldots, N$$

where $K(x_i, x_j)$ is a kernel function.

To find out hyperplanes about training data using the support vector machine, we transformed the original data in the following way. The AIA class labeled by 1 and ATA by -1. Each of SNPs is assigned to 1 or -1 according to their genotype signatures. For example, if SNP data is ALOX5_p1_1708_G>A then, the major allele is G and the minor allele is A. We exchanged $(1, 1)$ for GG, $(-1, 1)$ for AG and $(-1, -1)$ for AA.

4. Results

Our results were obtained by a two-step procedure. In the first step, the NMI was used to analyze the association of combinations of 3 SNPs with asthma and determine which combinations of SNPs were best in classifying the types of asthma (AIA or ATA). We calculated NMI values of all 3-SNP combinations using Equation (2), and selected 20 combinations with highest NMI values. In the second step, we applied support vector machine method, with the Gaussian radial basis function kernel, to evaluate the prediction accuracy of each of top 20 SNP combinations. Here we used SVM-light program[7], and the accuracy was obtained by the leave-one-out validation method. This result is summarized in Table 2, and it shows that the highest accuracy is achieved by the combination of 1st, 7th and 22nd SNPs.

The same procedure is applied to 4-SNP combinations and to 5-SNP combinations; the results are summarized in Table 3 and Table 4 respectively. Each of the highest prediction accuracy combinations is indicated in bold. Finally, the best model among all 3-SNP, 4-SNP and 5-SNP combinations is obtained by comparing the 3 highest accuracy combinations (Figure 2). It turns out to be the four SNP combination of ALOX5_p1_1708, B2ADR_q1_46, CCR3_p1_520 and CysLTR1_p1_634. Its NMI value is 0.053 and the prediction accuracy of the ATA disease risk among asthmatic patients is 71.14%.

Table 2 Top 20 combinations of 3 SNPs

3-SNP combinations	NMI value	MI value	Prediction accuracy(%)
1, 7, 21	0.03515	0.11590	65.85
2, 4, 6	0.03187	0,11970	66.67
1, 4, 6	0.03110	0.11105	65.85
3, 14, 23	0.03066	0.07861	65.04
1, 5, 6	0.03062	0.10767	60.16
1, 7, 22	0.03044	0.09816	**68.29**
2, 14, 23	0.02998	0.09027	64.63
14, 16, 23	0.02951	0.07878	64.63
2, 14, 19	0.02835	0.09739	59.35
3, 13, 23	0.02823	0.07739	66.67
1, 5, 13	0.02805	0.09353	61.79
10, 14, 23	0.02800	0.07312	63.41
1, 4, 23	0.02794	0.08726	63.41
12, 14, 23	0.02753	0.07944	66.67
12, 13, 23	0.02733	0.08353	67.07
4, 6, 9	0.02731	0.09860	59.35
1, 4, 13	0.02691	0.09104	57.72
12, 14, 16	0.02679	0.07903	60.57
1, 12, 13	0.02674	0.08703	65.04
2, 6, 23	0.02665	0.08945	56.91

Table 3 Top 20 combinations of 4 SNPs

4-SNP combinations	NMI value	MI value	Prediction accuracy(%)
2, 4, 6, 9	0.05873	0.26633	67.48
2, 4, 6, 19	0.05351	0.25177	68.29
1, 2, 4, 6	0.05330	0.23934	**71.14**
1, 2, 5, 6	0.05139	0.22767	64.23
4, 6, 19, 20	0.05138	0.24202	66.26
4, 6, 9, 21	0.05120	0.22620	62.20
4, 6, 19, 22	0.05115	0.23014	62.20
2, 4, 6, 23	0.05000	0.21471	68.70
1, 4, 6, 20	0.04974	0.22347	51.63
2, 4, 6, 21	0.04958	0.22397	65.85
2, 4, 6, 8	0.04930	0.22746	63.41
1, 4, 6, 22	0.04829	0.20514	59.76
4, 6, 19, 21	0.04799	0.21999	60.98
2, 4, 6, 20	0.04783	0.22253	60.16
2, 5, 6, 9	0.04780	0.21503	61.38
4, 6, 9, 23	0.04758	0.19917	67.48
2, 4, 6, 12	0.04756	0.21662	65.45
4, 6, 9, 22	0.04754	0.20591	61.38
2, 4, 6, 22	0.04715	0.20902	63.82
2, 5, 6, 8	0.04687	0.21614	65.45

The NMI code was programmed in MATLAB 2007b and SVM code was programmed in C++.

Table 4 Top 20 combinations of 5 SNPs

5-SNP combinations	NMI value	MI value	Prediction accuracy(%)
2, 4, 6, 19, 22	0.07992	0.41904	65.45
2, 4, 6, 19, 20	0.07992	0.42691	60.16
2, 4, 6, 19, 21	0.07777	0.41231	61.38
1, 2, 4, 6, 9	0.07773	0.39892	63.01
4, 6, 9, 19, 21	0.07698	0.40227	63.01
4, 6, 9, 19, 22	0.07636	0.39479	63.01
1, 2, 4, 6, 13	0.07610	0.39224	61.38
2, 4, 6, 9, 21	0.07507	0.38711	63.41
2, 4, 6, 9, 23	0.07498	0.37194	64.23
2, 4, 6, 9, 19	0.07498	0.39811	59.76
2, 4, 6, 9, 13	0.07469	0.38747	64.63
1, 2, 4, 6, 8	0.07446	0.38848	59.76
2, 4, 6, 8, 9	0.07430	0.38812	60.57
4, 6, 19, 20, 22	0.07398	0.38880	64.63
2, 5, 6, 19, 22	0.07363	0.38328	63.01
4, 6, 7, 9, 21	0.07344	0.37767	62.60
4, 6, 19, 20, 21	0.07337	0.39018	61.79
1, 2, 4, 6, 22	0.07334	0.36830	**68.70**
1, 2, 4, 6, 20	0.07310	0.37997	63.82
4, 6, 8, 19, 20	0.07303	0.37909	59.76

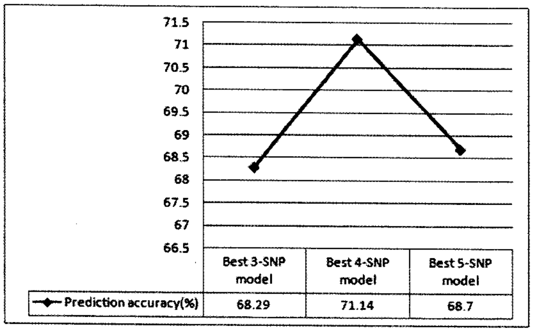


Figure 2 Comparison of prediction accuracy of the best 3-SNP, 4-SNP and 5-SNP models

5. Discussion

We described how to select SNPs which discriminate between AIA and ATA. By applying normalized mutual information, more than single SNP can be related so that we can explore the synergistic effect of three, four and five SNPs on asthma. Finally, we discussed how to predict asthma type from SNPs on the basis of SVM method. The results of this study can also be used as basic information between SNPs and disease or among SNPs for further implementation such as cluster analysis or gene-gene interaction. In conclusion, our results suggest that effect of multi-locus genetic interaction exists in the susceptibility to aspirin tolerance in asthmatic patients, which will help to better understand the complex genetic basis of ATA. The best 4-SNP combination model found in the previous section can be applied as a potential gene marker for diagnosing ATA.

Applying the MDR method to the same SNP data, we found that the best 4-SNP combination predicting ATA disease risk among asthmatic patients is B2ADR_q1_46, CCR3_p2_17, CysLTR1_p1_634, TBXA2R_q1_795 and the prediction accuaracy is 59.75%. The best SNP combination found in this paper provide us higher prediction accuracy than the one obtained by the MDR method.

References

[1] Tsalenko, A., Ben-Dor, A., Cox, N., et al., "Methods for analysis and visualization of SNP genotype data for complex diseases," *Pac. Symp. Biocomput.*, vol.8, pp.548-561, 2003.

[2] S. H. Kim H. H. Jeong, B. Y. Cho, et al, "Association of four-locus gene interaction with aspirin-intolerant asthma in Korean asthmatics," *J. Clin. Immunol.*, vol.4, no.4, pp.336-342, 2008.

[3] Cover, T. M. and Thomas, J. A., *Elements of Information Theory*, 2nd Ed., Wiley, 2006.

[4] Furey, T. S., Cristianini, N., Duffy, N., et al, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol.16, no.10, pp. 906-914, 2006.

[5] Zhou, X., Wang, X., Dougherty, E. R., et al. "Gene clustering based on clusterwide mutual information," *J. Comput. Biol.*, vol.11, no.1, pp.147-161, 2004.

[6] http://en.wikipedia.org/wiki/Mutual_information.

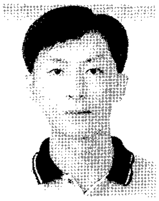
[7] http://svmlight.joachims.org/.

이 중 섭

1980년 서울대학교 수학과 졸업. 1982년 서울대학교 수학과 석사. 1989년 미시간 대학교 수학과 박사. 1992년~현재 아주대학교 자연과학부 교수. 관심분야는 Bioinformatics, Symbolic Dynamics

김 승 현

1990년 숙명여자대학교 화학과 졸업(학사). 1993년 서울대학교 화학과(이학석사). 1996년 서울대학교 화학과(이학박사). 2003년~2006년 아주대의대 연구전임강사. 2007년~2008년 아주대의대 연구 조교수. 2009년~현재 아주대 의대 지역임상시험센터 조교수. 관심분야는 생명정보학



신 기 섭

2000년 청주대학교 사범대학 수학교육학과 졸업(학사). 2002년 인하대학교 수학과(이학석사). 2003년~현재 아주대학교 수학과 박사과정. 2008년~현재 아주대학교 강사. 2009년~현재 강남대학교, 강릉대학교 강사. 관심분야는 Cluster Analysis, Bioinformatics

임 규 철

2006년 아주대학교 수학과 졸업(학사). 2008년 아주대학교 수학과 졸업(이학석사). 2008년~현재 군복무중. 관심분야는 패턴인식, 데이타마이닝