BMB
reports

Identification of epistasis in ischemic stroke using multifactor dimensionality reduction and entropy decomposition

*Jungdae Park, Younyoung Kim & Chaeyoung Lee**
Department of Bioinformatics and Life Science, Soongsil University, Seoul 156-743, Korea

We investigated the genetic associations of ischemic stroke by identifying epistasis of its heterogeneous subtypes such as small vessel occlusion (SVO) and large artery atherosclerosis (LAA). Epistasis was analyzed with 24 genes in 207 controls and 271 patients (SVO = 110, LAA = 95) using multifactor dimensionality reduction and entropy decomposition. The multifactor dimensionality reduction analysis with any of 1- to 4-locus models showed no significant association with LAA ($P > 0.05$). The analysis of SVO, however, revealed a significant association in the best 3-locus model with P10L of TGF-β1, C1013T of SPP1, and R485K of F5 (testing balanced accuracy = 63.17%, $P < 0.05$). Subsequent entropy analysis also revealed that such heterogeneity was present and quite a large entropy was estimated among the 3 loci for SVO (5.43%), but only a relatively small entropy was estimated for LAA (1.81%). This suggests that the synergistic epistasis model might contribute specifically to the pathogenetsis of SVO, which implies a different etiopathogenesis of the ischemic stroke subtypes. [BMB reports 2009; 42(9): 617-622]

INTRODUCTION

A limited number of ischemic stroke patients have shown a Mendelian inheritance pattern caused by a single gene, whereas many patients have shown complex patterns caused by multiple genes under various environmental exposures (1). Nevertheless, studies that have examined the genetic dissection of the complex ischemic stroke have been quite limited (2). Recently, research efforts have been devoted to identify associations of ischemic stroke with individual candidate genes. For example, a DNA sequential association (P10L) with ischemic stroke was identified in transforming growth factor-β1 (TGF-β1), an important cytokine involved in the process of inflammation that could cause plaque rupture, fatty streak, thrombosis and atherosclerosis (3-5). Another significant missense polymorphism, V66M, was identified in the gene of brain-derived neurotrophic factor (BDNF), which plays important roles in the survival, growth, and differentiation of neurons (2). Significant variants in promoter regions were identified in front of klotho (6) and thrombomodulin (7). Some intronal sequence variants were also significantly associated with the genes of secreted phosphoprotein 1 (8, 9) and neuropeptide Y (5), although their functions are still unclear.

By examining only individual genetic associations, it has been difficult to understand the genetic architecture of ischemic stroke. These questions can only be fully addressed using simultaneous analysis with multiple genes, since these methods can be used to accurately assess the genetic effects of such complex traits. The objective of this study was to conduct epistasis analyses using multifactor dimensionality reduction (MDR) and entropy decomposition (ED) to better understand the genetic associations of ischemic stroke, with a particular focus on its subtypes. A nonparametric approach was employed for epistatic analysis because potentially low power or non-estimable statistics might be caused by the large number of parameters used in the parametric analytical models. This is the first study that examines ischemic stroke using MDR or ED.

RESULTS

Single gene analysis

Statistically significant associations between ischemic stroke and its subtypes were found in the following genes: BDNF, LIF, NPY, SPP1, and TGF-β1 (Supplementary Table 1) (http://clee11.cafe24.com/mdred) ($P < 0.05$). The LIF, NPY, and SPP1 showed haplotypic association whereas BDNF and TGF-β1 showed only single locus association. These genes showed subtype-specific effects.

Multifactor dimensionality reduction analysis

MDR analysis with the combined data showed no significant interaction effect in any of the one- to four-locus models (Table 1) ($P > 0.05$). The best model was the single locus model with the TGF-β1 P10L, showing an average cross-validation consistency (CVC) of 7.57 and an average testing balanced accuracy (TBA) of 51.39%. Subsequent analysis with the subtypes revealed no significant association with large artery athero-

*Corresponding author. Tel: 82-2-820-0455; Fax: 82-2-824-4383; E-mail: clee@ssu.ac.kr

Received 21 January 2009, Accepted 10 May 2009

Keywords: Entropy, Epistasis, Genetic association, Genetic factor, Stroke

Table 1. Best candidate model selected for ischemic stroke using multifactor dimensionality reduction

No. of loci	Best candidate model	Avg. CVC[a]	Avg. TBA (%)[b]
	Combined ischemic stroke		
1	*TGFB1* P10L	7.57	51.39
2	*LIF* T4524G / *LIF* C3640A	5.63	50.84
3	*LIF* T4524G / *LIF* C3640A / *SPP1* C5891T	3.95	51.08
4	*TGFB1* P10L / *SPP1* C1013T / *MTHFR* C677T / *F5* R485K	4.90	50.92
	Large artery atherosclerosis		
1	SPP1 C1013T	5.04	49.06
2	LIF T4524G MTHFR C677T	3.22	46.03
3	LIF T4524G MTHFR C677T IL6R D358A	7.21	56.35
4	LIF T4524G SPP1 C1013T MTHFR C677T IL6R D358A	8.04	57.34
	Small vessel occlusion		
1	TGF-β1 P10L	9.43	53.32
2	TGF-β1 P10L SPP1 C2140T	5.35	54.47
3	TGF-β1 P10L SPP1 C1013T F5 R485K	9.67	63.17*
4	TGF-β1 P10L LIF T4524G SPP1 C1013T F5 R485K	4.87	52.79

[a]Stands for the average of 100 replicates for cross-validation consistency. [b]Stands for the average of 100 replicates for testing balanced accuracy. This was determined by testing the model built with a training set. *P < 0.05

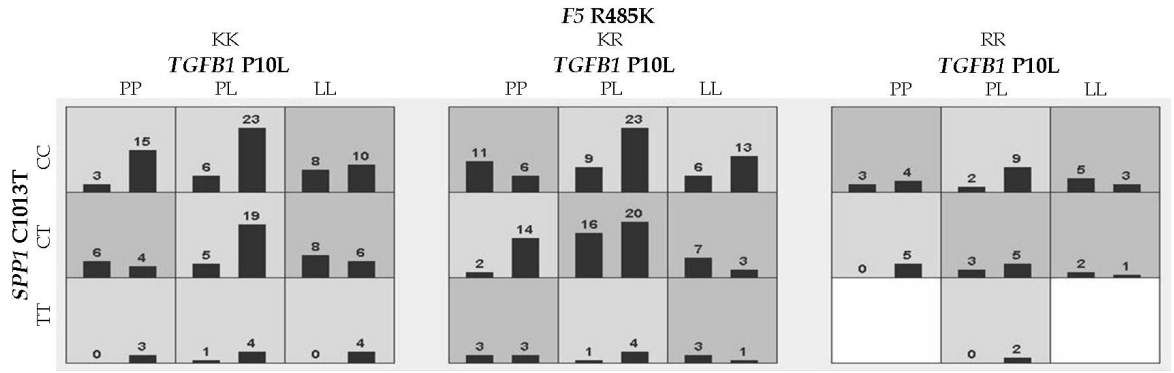

Fig. 1. The best multi-locus model for small vessel occlusion showing the genotypes combined with 3 loci and their observed numbers of patients and controls. The dark gray cell indicates the high risk genotype and the gray cell indicates the low risk genotype, along with the corresponding distribution of cases (left bar) and of controls (right bar) for each combination. The white cell shows the genotype without observation.

sclerosis (LAA) (Table 1) (P > 0.05). On the other hand, significant associations were observed in the analysis with small vessel occlusion (SVO) when the 3-locus model was used, which included P10L of TGF-β1, C1013T of SPP1, and R485K of F5 (Table 1). The permutation test showed a significant average TBA (63.17%) for the model, and its P values ranged from 0.028 to 0.029. The statistical significance of the 3-locus model for SVO was confirmed using a logistic regression analysis (Supplementary Table 2) (http://clee11.cafe24.com/mdred). For example, the odds ratio estimate for the combined genotype of KR (F5 R485K) CT (SPP1 C1013T) LL (TGF-β1 P10L) was 4.55, and its corresponding 95% confidence interval (CI) ranged between 1.15-17.98 (P < 0.05). The odds ratio estimate for all the risk genotype groups was 4.48 with a 95% CI ranging between 2.72-7.36 ($P = 1.08 \times 10^{-9}$).

Genotypic combinations with the three loci selected in the MDR analysis were further analyzed to determine whether each combination belonged to the risk or protective genotype in regards to the susceptibility to SVO. Assignment to the risk or protective genotype was determined by the ratio of the case to control number obtained in the analysis (Fig. 1). For example, the group with the genotype RR (F5 R485K) CT (SPP1 C1013T) PP (TGF-β1 P10L) was most likely predisposed to SVO and the KR CT LL genotype was least susceptible to SVO.

Entropy decomposition analysis

An entropy-based interaction graph was established using the most significant variants in the MDR analysis and in a preliminary ED analysis (Fig. 2). The polymorphism of T235M in AGT was added because of its considerable epistatic contribution

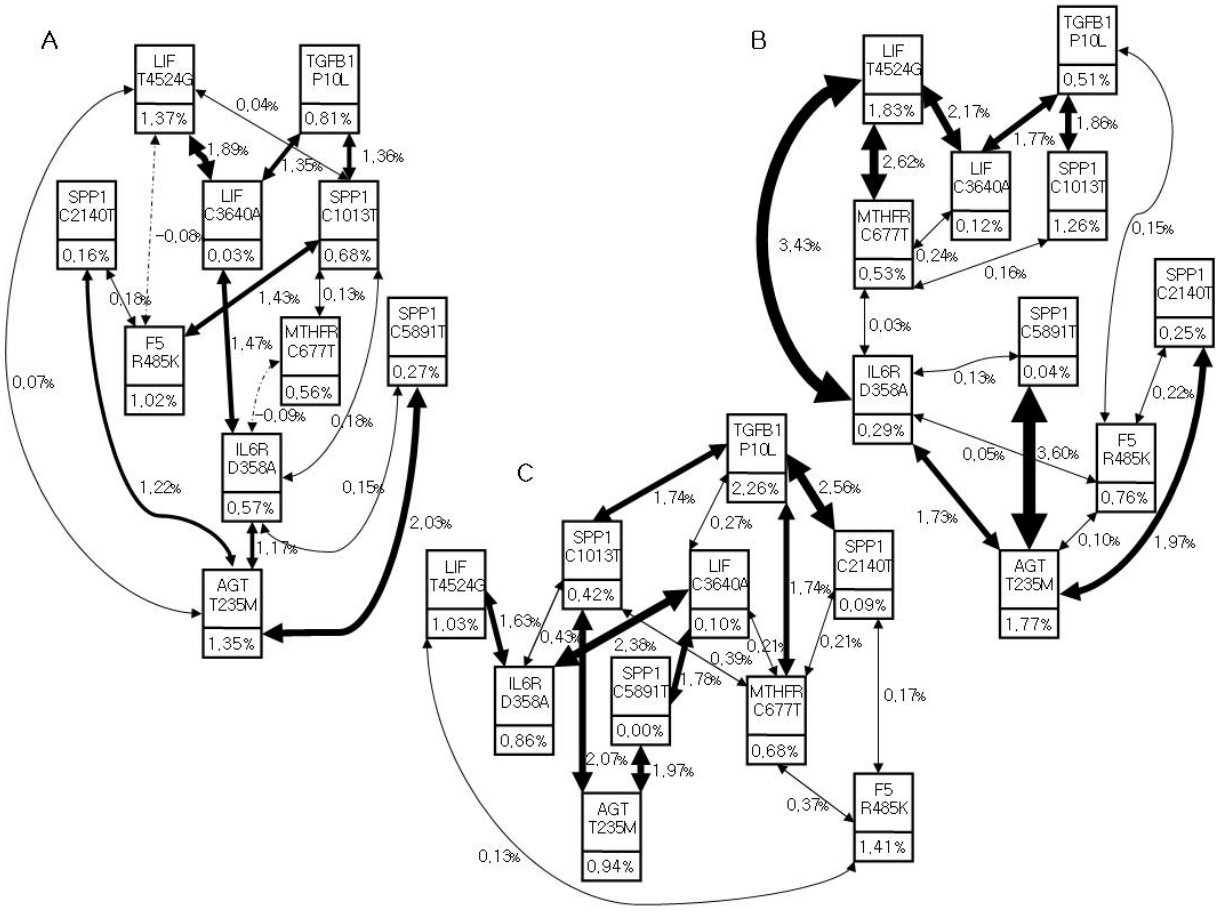

Fig. 2. Orange canvas interaction graph for combined ischemic stroke (A), large artery atherosclerosis (B), and small vessel occlusion (C). The hierarchical interaction graph shows the percentage of entropy removed in the case-control by the main individual locus effect (node) and by their pairwise interaction effect (arrow). Solid (dotted) arrows indicate each positive (negative) interaction.

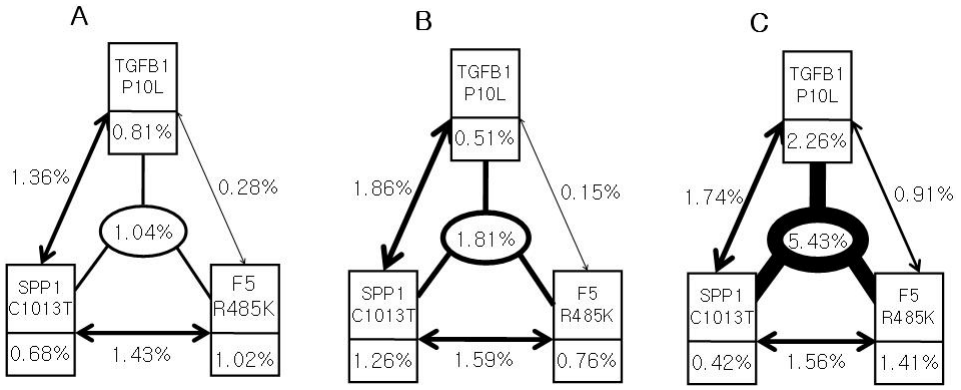

Fig. 3. Entropy decomposition with the three loci, which were selected as the best multi-locus model for small vessel occlusion. (A) combined ischemic stroke, (B) large artery atherosclerosis, and (C) small vessel occlusion. This hierarchical interaction graph shows the percentage of entropy removed in the case-control by the main individual locus effect (node), by their pairwise interaction effect (arrow), and by the 3-locus interaction effect (circle). Gray arrow and circle indicate positive interactions.

with other genes included in the graph. For example, individual entropy estimates for AGT T235M and SPP1 C5891T were small or negligible for LAA (1.77% and 0.04%, respectively), but their interaction estimate was considerably larger (3.60%).

The entropy information in the case-control status also indicated that susceptibility genes by stroke subtypes were heterogeneous. For example, LIF T4524G had the largest entropy for LAA whereas TGF-β1 P10L had the largest entropy for SVO. The heterogeneity was also estimated in the interactions. For example, the interaction between SPP1 C5891T and AGT 235 M had the largest entropy for LAA, and the interaction between TGF-β1 P10L and SPP1 C2140T had the largest entropy for SVO (Fig. 2).

Further investigation of the 3-locus interaction models also revealed differences between LAA and SVO in the interaction estimates. For example, the entropy estimate from the 3 loci identified as the most significant factors for SVO in the MDR analysis (TGF-β1 P10L, SPP1 C1013T, and F5 R485K) was quite large (5.43%) relative to the entropy estimate obtained for LAA (1.81%) (Fig. 3).

DISCUSSION

A clear understanding of the genetic dissection of ischemic stroke has been quite limited because of its variety in clinical endpoints as well as its genetic complexity. In the current study, we conducted genetic analysis simultaneously with multiple genes, focusing on their epistasis, in an attempt to explain the genetic architecture of this disease. Furthermore, we analyzed the main subtypes of the ischemic stroke, LAA and SVO. As a result, we found that these subtypes display genetic heterogeneity. For example, in the MDR analysis, the 3-locus model with TGF-β1 P10L, SPP1 C1013T, and F5 R485K was determined to be the best in explaining the susceptibility to SVO ($P < 0.05$). On the other hand, statistical signficance was not observed for the other subtype (LAA) of ischemic stroke for any multi-locus model ($P > 0.05$).

Further epistatic analysis by ED showed detailed relationships among multiple loci. This entropy-based analysis revealed that the significance detected for the 3-locus model from the MDR analysis might be largely explained by the 3-locus synergistic effect (5.43%). This interaction estimate was even larger than the sum (4.09%) of the individual effects (2.26%, 0.42%, and 1.41%) and also larger than the sum (4.21%) of the pair-wise interaction effects (1.74%, 1.56%, and 0.91%). The 3-locus interaction effect (1.81%) for LAA was comparable to some pair-wise interaction effects (1.86% and 1.59%), which also suggests the presence of heterogeneity in the subtypes (Fig. 3).

The considerable significance found in the 3-locus interaction could never be predicted using only the pair-wise entropy information in Fig. 2. Also, the largest pair-wise interaction (3.60% between SPP1 C5891T and AGT T235M for LAA, and 2.56% between TGF-β1 P10L and SPP1 C2140T for SVO) could not be predicted by the entropy information obtained from an individual loci. Even some individual effects among the variants were negligible, 0.04% for SPP1 C5891T in LAA and 0.09% for SPP1 C2140T in SVO. This implied that we were hardly able to predict higher order interaction effects with a lower order interaction model.

In conclusion, the current genetic analysis of ischemic stroke provided the first evidence that an epistatic model including TGF-β1 P10L, SPP1 C1013T, and F5 R485K is associated with the susceptibility to SVO as assessed by MDR and ED. However, it is worth noting that the false negative results obtained for LAA may be attributable to the small sample size. Thus, our findings should be replicated using larger subgroups of ischemic stroke patients for practical applications. Network analyses (10) and functional studies will also be necessary to better understand the underlying mechanism of the epistasis.

MATERIALS AND METHODS

Subjects

Ischemic stroke patients were recruited from Hallym University Hospital. A positive diagnosis was determined by performing computed tomography or magnetic resonance imaging scans from acute stroke patients within 7 days of onset (4). We included a total of 271 patients with ischemic stroke diagnosed from 2002 to 2005 and further categorized them into its subtypes such as SVO (n = 110), LAA (n = 95), cardioembolism (CE, n = 20), and the other strokes with rare or undetermined etiology, using the TOAST classification system (11). We utilized the classified data sets as well as the combined data set in the current study, and the subtype analysis was limited to LAA and SVO because of the small sample sizes of the other subtypes. Two hundred and seven subjects who served as the control group did not have any history of cerebral ischemic events and were randomly selected among healthy people from routine health checkups including chest X-ray, gastroscopy, basic health checkup (blood test, urinalysis, liver function test, heart function test, and etc), optional cancer examinations, and a routine survey prior to consultation. A detailed description and summary of the data has been presented in a previous study (4). Written informed consent was obtained from all subjects, and the study protocol was approved by an Ethical Committee.

Genotyping

We used genotypic data on 36 sequence variants (35 SNPs and 1 Ins/Del polymorphism) from our previous association studies (2, 4, 6), and additional genotyping was conducted for sequence polymorphisms in coagulation factor V (F5), interleukin 6 receptor (IL6R), and 5,10-methylenetetrahydrofolate reductase (MTHFR) genes using the TaqMan polymerase chain reaction (PCR) assay (Applied Biosystems, Foster City, CA, USA). The three candidate genes selected in the current study were first examined for ischemic stroke. Reactions were carried out following the manufacturer's protocol, and the prod-

ucts were analyzed by ABI PRISM 7900HT (Applied Biosystems, Foster City, CA, USA). Genotyping was conducted with laboratory personnel blind to the case-control status of the samples.

Single gene analysis

Associations of each individual locus with ischemic stroke or their subtypes were tested by odds ratio (OR) statistics. The ORs and their 95% confidence intervals were estimated using SAS Release 9.1 (SAS Institute, Cary, NC, USA).

Multifactor dimensionality reduction analysis

Joint analyses with multiple loci were conducted using MDR, which is a data reduction approach for identifying combinations of multi-locus genotypes that were associated with a susceptibility to a specific disease (12). This method allowed high-dimensional genetic data to be collapsed into a single dimension and thus made it possible to infer epistasis in a relatively small sample size by grouping genotypes. A cross-validation strategy was incorporated with the MDR to estimate the classification and prediction error of multifactor models. The current MDR under the case-control design was conducted using 10-fold cross validation, and the data were randomly and equally partitioned into 10 pieces. We utilized 9 pieces as a training data set and 1 piece as a testing data set for each of the 10 possible partitions. The training set was used to build a genetic model for predicting susceptibility to ischemic stroke, and the testing set was used to test the model built by using the training set.

In order to identify the best n-locus model, a contingency table for the combined genotypes produced with every possible n-locus was first created to display the case vs. control status in n-dimensional space. The risk level (i.e. a high risk or a low risk) of each cell was determined by comparing the case-control ratio estimate to the corresponding total ratio. The total ratios were 1.31, 2.18, and 1.88 for the combined ischemic stroke, SVO, and LAA, respectively. The possible combinations of n loci were evaluated based on minimum classification error. The final step was to estimate the TBA of the selected model. This procedure was repeated 10 times by 10-fold cross-validation.

All of the above procedures were replicated 100 times by shuffling data sets, and the average estimates of CVC and TBA for the replicates were then calculated. The final model among the best models with 1-, 2-, ⋯, n-loci was determined with the maximum estimates of CVC and TBA. The statistical significances of the best candidate models were determined by a permutation test. We excluded covariates such as gender, age, BMI, hyperlipidemia, smoking, and hypertension in the analysis because our preliminary analysis revealed that the best models when these factors were incorporated in the generalized MDR (13) did not differ from those without incorporation (data not shown). The MDR analysis was conducted using the MDR software package available for free at http://www.multifactordimensionalityreduction.org.

Entropy decomposition analysis

The interaction among multiple loci associated with a susceptibility to ischemic stroke was further interpreted by displaying a graph with entropy-based pair-wise interaction estimates suggested by Jakulin and Bratko (14). This complementary method provides distinguishable additive and non additive genetic effects, which can not be separated in the MDR analysis, and thus we were able to detect the epistatic effects and their directions (synergistic and redundant effects). This graph does not provide information on genotypes whereas the MDR analysis does. The graph is comprised of a node for each variant and line connections between them. The estimate in the node is the portion of entropy removed by each variant, and the estimate by the connection is the portion of entropy removed for each pair-wise interaction information of the variants. The entropy-based interaction information among multiple variants was further extended as follows:

$$\begin{aligned} I(L_i;L_j;L_k;C) = {} & H(L_i) + H(L_j) + H(L_k) + H(C) - H(L_i,L_j) - \\ & H(L_i,L_k) - H(L_j,L_k) - H(L_i,C) - H(L_j,C) - \\ & H(L_k,C) + H(L_i,L_j,L_k) + H(L_i,L_j,C) + H(L_i,L_k,C) \\ & + H(L_j,L_k,C) - H(L_i,L_j,L_k,C) \end{aligned}$$

where $I(L_i;L_j;L_k;C)$ is the size of interaction information for 3 sequence variants(L_i, L_j, and L_k) and one class variable (C), and $H(\bullet)$,$H(\bullet,\bullet)$,$H(\bullet,\bullet,\bullet)$, and $H(\bullet,\bullet,\bullet,\bullet)$ are measures of unpredictability as the single entropy of 1 attribute and the joint entropy of 2, 3, and 4 attributes, respectively. For example, two-way interaction analysis reduces the uncertainty of either of the two attributes with the knowledge of the other attribute, and the joint entropy is calculated as follows:

$$H(L_1,L_2) = -\sum\sum p(l_1,l_2)\log p(l_1,l_2)$$

Finally, the entropy removed in the case-control by main individual locus effect was estimated as $I(L_i;C)/H(C)$ and the entropy removed by their pairwise interaction effect was estimated as $I(L_i;L_j;C)/H(C)$.

The direction of the interaction effect was determined by its positive (synergistic effect) or negative (redundant effect) value. This provided a complementary inference to the MDR study, which only determined if an interaction effect existed (12, 14, 15). We selected 10 variants for their display by the Orange Canvas in the current study, and the variant selection was based on the best candidate model estimated by MDR analyses. The entropy decomposition analysis was conducted using the freely available Orange machine learning software at http://www.ailab.si/orange.

Acknowledgements

This study was supported by a grant of the Korea Healthcare Technology R&D Project, Ministry of Health & Welfare, Republic of Korea (A080042).

REFERENCES

1. Hassan, A. and Markus, H. S. (2000) Genetics and ischaemic stroke. *Brain* **123**, 1784-1812.
2. Lee, C. and Kong, M. (2007) An interactive association of common sequence variants in the neuropeptide Y gene with susceptibility to ischemic stroke. *Stroke* **38**, 2663-2669.
3. Cipollone, F., Fazia, M., Mincione, G., Iezzi, A., Pini, B., Cuccurullo, C., Ucchino, S., Spigonardo, F., Di Nisio, M., Cuccurullo, F., Mezzetti, A. and Porreca, E. (2004) Increased expression of transforming growth factor-β1 as a stabilizing factor in human atherosclerotic plaques. *Stroke* **35**, 2253-2257.
4. Kim, Y. and Lee, C. (2006) The gene encoding transforming growth factor (1 confers risk of ischemic stroke and vascular dementia. *Stroke* **37**, 2843-2845.
5. Sie, M. P., Uitterlinden, A. G., Bos, M. J., Arp, P. P., Breteler, M. M., Koudstaal, P. J., Pols, H. A.,Hofman, A., van Duijn, C. M. and Witteman, J. C. (2006) TGF-β1 polymorphisms and risk of myocardial infarction and stroke: the Rotterdam study. *Stroke* **37**, 2667-2671.
6. Kim, Y., Kim, J. H., Nam, Y. J., Kong, M., Kim, Y. J., Yu, K. H., Lee, B. C. and Lee, C. (2006) Klotho is a genetic risk factor for ischemic stroke caused by cardioembolism in Korean females. *Neurosci. Lett.* **407**, 189-194.
7. Li, Y. H., Chen, J. H., Wu, H. L., Shi, G. Y., Huang, H. C., Chao, T. H., Tsai, W. C., Tsai, L. M., Guo, H. R., Wu, W. S. and Chen, Z. C. (2000) G-33A mutation in the promoter region of thrombomodulin gene and its association with coronary artery disease and plasma soluble thrombomodulin levels. *Am. J. Cardiol.* **85**, 8-12.
8. Brenner, D., Labreuche, J., Touboul, P. J., Schmidt-Petersen, K., Poirier, O., Perret, C., Schonfelder, J., Combadiere, C., Lathrop, M., Cambien, F., Brand-Herrmann, S. M. and Amarenco, P. (2006) Cytokine polymorphisms associated with carotid intima-media thickness in stroke patients. *Stroke* **37**, 1691-1696.
9. Kim, Y. and Lee, C. (2008) Haplotype analysis revealed a genetic influence of osteopontin on large artery atherosclerosis. *J. Biomed. Sci.* **15**, 529-533.
10. Rho, S., You, S., Kim, Y. and Hwang, D. (2008) From proteomics toward systems biology: integration of different types of proteomics data into network models. *BMB Rep.* **41**, 184-193.
11. Adams, H. P. Jr, Bendixen, B. H., Kappelle, L. J., Biller, J., Love, B. B., Gordon, D. L. and Marsh, E. E. 3rd. (1993) Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. *Stroke* **24**, 35-41.
12. Moore, J. H., Gilbert, J. C., Tsai, C. T., Chiang, F. T., Holden, T., Barney, N. and White, B. C. (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J. Theor. Biol.* **241**, 252-261.
13. Lou, X. Y., Chen, G. B., Yan, L., Ma, J. Z., Zhu, J., Elston, R. C. and Li, M. D. (2007) A generalized combinatorial approach for detecting gene by gene and gene by environment interactions with application to nicotine dependence. *Am. J. Hum. Genet.* **80**, 1125-1137
14. Jakulin, A. and Bratko, I. (2003) Analyzing attribute interactions. *Lect. Notes Artiff. Intell.* **2838**, 229-240.
15. McGill, W. J. (1954) Multivariate information transmission. *Psychometrica* **19**, 97-116.