Biochemical and Biophysical Research Communications 419 (2012) 99–103

Contents lists available at SciVerse ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

Partition dataset according to amino acid type improves the prediction of deleterious non-synonymous SNPs

Jing Yang [a,b], Yuan-Yuan Li [a,b], Yi-Xue Li [a,b,*], Zhi-Qiang Ye [c,d,*]

[a] School of Biotechnology, East China University of Science and Technology, Shanghai 200237, China
[b] Shanghai Center for Bioinformation Technology, Shanghai 200235, China
[c] Laboratory of Chemical Genomics, School of Chemical Biology and Biotechnology, Peking University Shenzhen Graduate School, Shenzhen 518055, China
[d] Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

ARTICLE INFO

Article history:
Received 18 January 2012
Available online 4 February 2012

Keywords:
Non-synonymous SNP
Disease-association
Dataset partition
Machine learning

ABSTRACT

Many non-synonymous SNPs (nsSNPs) are associated with diseases, and numerous machine learning methods have been applied to train classifiers for sorting disease-associated nsSNPs from neutral ones. The continuously accumulated nsSNP data allows us to further explore better prediction approaches. In this work, we partitioned the training data into 20 subsets according to either original or substituted amino acid type at the nsSNP site. Using support vector machine (SVM), training classification models on each subset resulted in an overall accuracy of 76.3% or 74.9% depending on the two different partition criteria, while training on the whole dataset obtained an accuracy of only 72.6%. Moreover, the dataset was also randomly divided into 20 subsets, but the corresponding accuracy was only 73.2%. Our results demonstrated that partitioning the whole training dataset into subsets properly, *i.e.*, according to the residue type at the nsSNP site, will improve the performance of the trained classifiers significantly, which should be valuable in developing better tools for predicting the disease-association of nsSNPs.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

Single nucleotide polymorphisms (SNPs), representing the genomic differences of a single nucleotide between individuals, are widely distributed on the human genome [1]. The total amount of human SNPs is now about 40 millions according to the statistics from NCBI SNP database (dbSNP build 135). Non-synonymous single nucleotide polymorphism (nsSNP) is a class of SNP which occurs in the coding region and leads to amino acid substitution in the protein product, *i.e.*, an original residue is changed to a substituted residue, and is thus called single amino acid polymorphism (SAP) as well. Many nsSNPs play important roles in the pathogenesis of inherited diseases, cancers, diabetes, and cardiovascular diseases [2–4], hence sorting out disease-associated nsSNPs from neutral ones becomes necessary when we obtain a large set of nsSNPs in a specific study. Since wet lab experiments for confirming nsSNPs' disease-association one by one are expensive, slow, and laborious, developing computational tools to make predictions before experiments is desirable.

The last decade has witnessed the publishing of various prediction methods for this purpose, and most of them adopted machine learning methods to train prediction classifiers [5–12]. To achieve good performance and to improve the statistical significance, they utilized all the training data as a single dataset. The recent accumulation of nsSNPs with known disease-association has provided a much larger training dataset than in previous studies, which has enabled us to partition the overall heterogeneous training set into homogeneous subsets while keeping substantial data samples left in each one. It is supposed to be optimal to train individual classifiers on each subset separately. In this study, we partitioned the nsSNP dataset into 20 subsets based on the amino acid residue type at the nsSNP position, either the original or the substituted residue type. After a set of well-known attributes were extracted, we trained support vector machine (SVM) classifiers for each subset. We found that training classification models based on each subset partitioned according to the original (substituted) residue type at the nsSNP site resulted in an accuracy of 76.3% (74.9%), while training on 20 randomly generated subsets or the whole dataset obtained an accuracy of only 73.2% or 72.6%, demonstrating that appropriately partitioning a heterogeneous training set into homogeneous subsets will improve the performance of the trained classifiers significantly.

* Corresponding authors. Address: School of Biotechnology, East China University of Science and Technology, Shanghai 200237, China. Fax: +86 21 54065058 (Y.-X. Li); Laboratory of Chemical Genomics, School of Chemical Biology and Biotechnology, Peking University Shenzhen Graduate School, Shenzhen 518055, China. Fax: +86 755 26033196 (Z.-Q. Ye).
E-mail addresses: yxli@sibs.ac.cn (Y.-X. Li), yezq@pkusz.edu.cn (Z.-Q. Ye).

0006-291X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2012.01.138

2. Materials and methods

2.1. Data sources

The Swiss-Prot knowledgebase (version 57, included in Uniprot database version 15.0) which contains 53,846 amino acid substitutions of 11,086 human proteins served as the major data source for implementing our methods [13,14]. To extract proper features, we also used the data from AAIndex (http://www.genome.ad.jp/aaindex/) and IMGT/HLA (http://www.ebi.ac.uk/imgt/hla/, released in April 2011) [15,16].

2.2. Feature construction

It is assumed that the changes in protein sequence, structure, and physicochemical characteristics caused by disease-associated nsSNPs and neutral nsSNPs are different, so the features reflecting the sequence, structure, and physicochemical characteristics will contain the information for differentiating the disease-associated nsSNPs and neutral ones. Previously published studies have already established a set of sequence and structural features with good predictive power. In order to insure the scalability, we adopted a set of 55 features which could be calculated or derived from sequence information [17–19].

2.2.1. Physicochemical properties

AAIndex is a database containing various physicochemical and biochemical properties of amino acids or residue pairs [15]. For each type of amino acid residues, seven physicochemical properties including molecular weight, isoelectric point (pI) [20], hydrophobicity index [21], bulkiness [20], residue volume [22], average volume of buried residue [23], and side chain volume [24] were extracted from AAIndex [15]. In addition, the residue's relative frequencies in three secondary structural states (helix, sheet, and turn) [25] were also retrieved from AAIndex [15]. Since one nsSNP contains both original and substituted residue, there will be 20 features (10 for the original residue and 10 for the substituted one) of physicochemical properties describing it in total.

2.2.2. BLOSUM and GRANTHAM scores

BLOSUM62 matrix provides scores representing the residue substitution probabilities in the evolutionary history of protein families [26], and GRANTHAM scores depict the physicochemical distances between residue pairs [27]. Since each nsSNP represents a substitution from the original residue to the substituted one, which corresponds to a score in BLOSUM62 and GRANTHAM matrix respectively, we obtained two features here.

2.2.3. PSSM and conservation score

Using a protein sequence as the query, a PSI-BLAST run against a protein sequence database iteratively can generate a position-specific scoring matrix (PSSM), which contains frequencies of 20 types of amino acids at each sequence position [28]. The frequencies represent the occurring probabilities of each type of residue at a specific sequence position of the protein family in the evolutionary history. In addition to these 20 frequencies at the nsSNP position, we also calculated the conservation score according to

$$\text{Conservation} = -\sum_{i=1}^{20} p_i \log_2 p_i, \tag{1}$$

where p_i represents the occurring frequency of residue type i at the nsSNP position in the PSSM [29]. If a position has a high conservation score, we consider that this position is a functionally important position and its variation is prone to be disease-associated. In this study, we ran PSI-BLAST against Swiss-Prot version 57 by setting the e-value with 1e-5, the h-value with 0.02, and iterative rounds with 2. By parsing the checkpoint file, we obtained each numeric values in the PSSM. Taken together, 21 PSSM-related features were obtained for each nsSNP.

2.2.4. NsSNPs' microenvironment

Previous studies demonstrated the usefulness of the microenvironment of the nsSNP in this kind of prediction [12,17]. We defined the 4 upstream and 4 downstream neighboring residues as the nsSNP's microenvironment, whose characteristics were also represented by the 10 physicochemical properties mentioned above. Specifically, we calculated the average of each property regarding the residues in the microenvironment, and we thus got 10 features describing it. If the nsSNP is located in the 4 N-terminal (C-terminal) amino acids of the protein, we calculated the average of all the upstream (downstream) residues and 4 downstream (upstream) residues. In addition, we also calculated the conservation at each residue position at the microenvironment, and then the average was taken as a feature.

2.2.5. HLA family

As indicated in previous studies, we defined a feature to indicate whether the protein where the nsSNP is located is a member of HLA (Histocompatibility leukocyte antigen) family or not, since HLA family may hold some special characteristics to carry mainly neutral polymorphisms. Specifically, we ran BLAST against the IMGT/HLA database [16,28], and we assigned a protein to belong to HLA family if it hit a sequence in IMGT/HLA with the e-value less than 0.01 and sequence identity greater than 70%.

2.3. Dataset partition

The training dataset contains a large number of heterogeneous nsSNPs, whose propensities of disease-association is due to different underlying reasons. Hence it should be optimal to partition the dataset into more homogeneous subsets, and then to train individual classifiers for each subset. In detail, our strategy contained two approaches, one of which was to partition the dataset into 20 subsets according to the original residue type at the nsSNP position, and the other of which was to partition according to the substituted residue type. For comparison, we also partitioned the dataset into 20 subsets randomly.

2.4. SVM Classifiers

Support Vector Machine (SVM) is a useful and popular tool in classification problems. It has also been widely used in building classifiers for predicting the disease-association of nsSNPs. We adopted the LIBSVM package (http://www.csie.ntu.edu.tw/~cjlin/libsvm, version 2.91) to train the classifiers using the RBF (radial basis function) kernel [30]. The parameters of C and γ, which should be set before training, were optimized through the grid-search tool implemented in the LIBSVM package. The cross-validation results with the optimized C and γ, were used to calculate the accuracy (ACC) and Matthew's correlation coefficient (MCC) to evaluate the trained classifiers [31]. The formulae of these two measurements are listed in the following:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \tag{2}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{3}$$

where TP is the number of correctly predicted positive samples, TN is the number of correctly predicted negative samples, FP is the

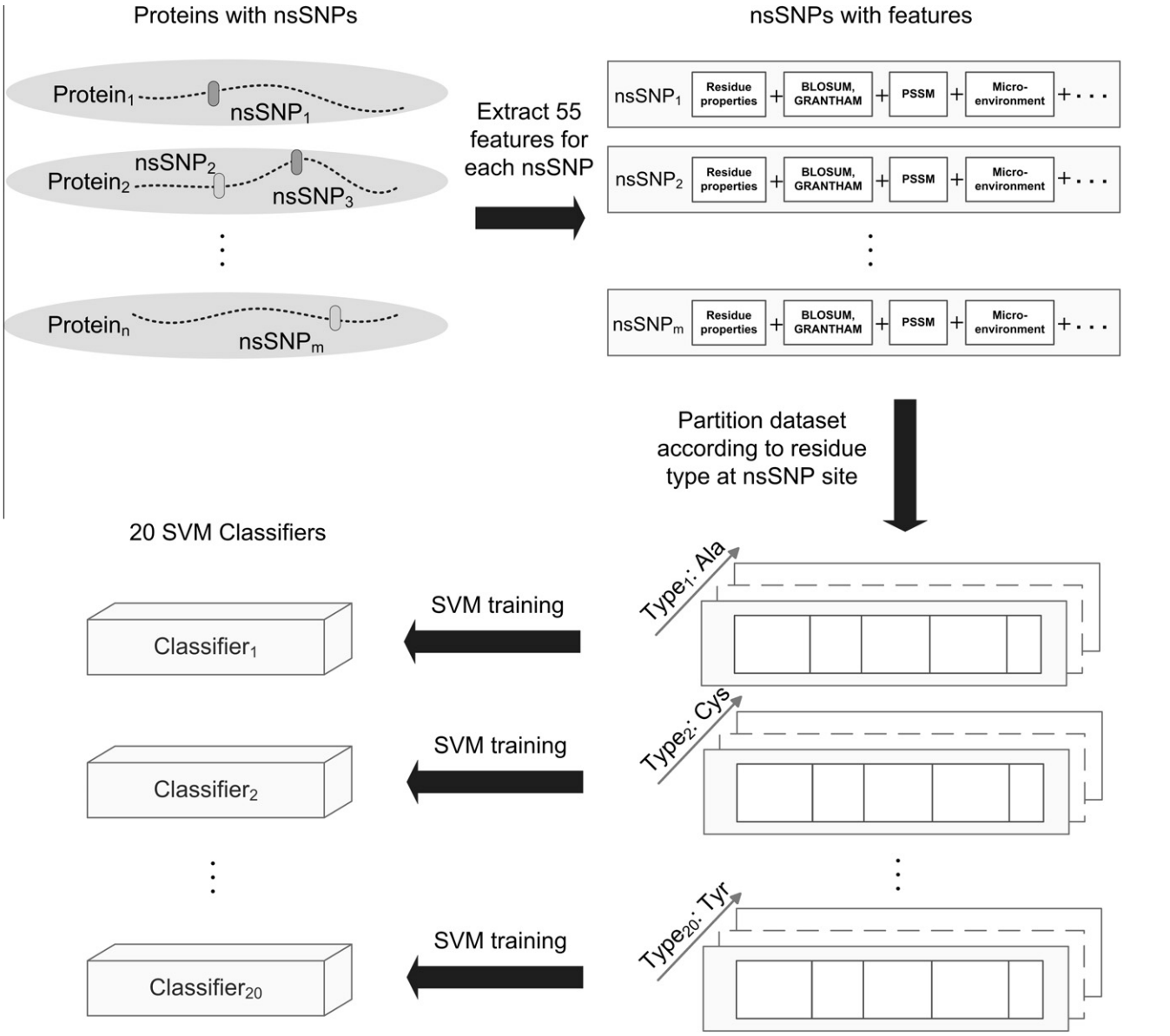

Fig. 1. The overall working pipeline. The dataset partition was performed in two approaches: according to the original residue type and according to the substituted residue type.

number of negative samples which are incorrectly predicted as positive, and *FN* is the number of positive samples which are incorrectly predicted as negative samples. The overall pipeline of our method is illustrated in Fig. 1.

3. Results

3.1. Dataset preparation

In the dataset derived from Swiss-Prot, each substitution (nsSNP or SAP) was annotated as "Disease", "Polymorphism", or "Unclassified", and we used the data annotated as "Disease" (similar to "deleterious" and means that the nsSNP is proved experimentally to affect the protein's function or be associated with disease) and "Polymorphism" (similar to "neutral" or "tolerated" and means that the nsSNP is not reported to be associated with disease) as positive and negative samples, respectively. Finally we collected 20915 disease-associated and 28326 neutral nsSNPs from 10476 proteins. In order to extract features of each nsSNP, the amino acid sequences of these proteins were also retrieved from the Swiss-Prot knowledgebase. For each nsSNP, we extracted 55 features including residue physicochemical properties, conservation score both at the nsSNP site and its microenvironment, PSSM profiles at the nsSNP site, substitution scores between original and substituted residue, etc. The details were described in the Section 2.

3.2. Performance of classifiers with dataset partition

As described, we partitioned the dataset into 20 subsets according to the original residue type at the nsSNP site. For each subset we utilized the grid-search tool of LIBSVM package to find optimal parameters of C and γ, which generated highest 5-fold cross-validation accuracy among all the combinations of C and γ that were tried by default. At the same time, we obtained the *TP*, *TN*, *FP*, and *FN* counts for each subset's classifier with its optimal C and γ. After summing up the corresponding counts of *TP*, *TN*, *FP*, and *FN* from the 20 classifiers, we calculated the overall *ACC* and

Table 1
Performance of different dataset partition approaches.

Partition methods[a]	*ACC* (%)	*MCC*
Original	76.3	0.509
Substituted	74.9	0.479
Random	73.2	0.444
Whole dataset	72.6	0.431

[a] Original and substituted mean that partitioning the dataset according to the original and substituted residue type respectively; random means that partitioning the dataset randomly; whole dataset means no partition.

MCC. The same procedure was applied to the 20 subsets into which the dataset was divided according to the type of the substituted residue at the nsSNP site, and we thus obtained another set of *ACC* and *MCC*. As a control, we also applied this procedure to the 20 randomly partitioned subsets. Moreover, training on the whole dataset without partitioning was adopted for additional comparison.

The cross-validation accuracies based on different approaches are listed in Table 1 and detailed results are provided in the online Supplementary file. We will only discuss *ACC* measurements here, since the similar trend occurred when *MCC* measurements were considered. First, the *ACC* difference between training on the randomly partitioned subsets and training on the whole dataset is only 0.6 percent (73.2% *vs.* 72.6%), which means that randomly partitioning of the dataset will result in accuracy similar to that of training on the whole dataset. This is reasonable since these subsets can be considered to be randomly sampled from the population, and thus represent the population in many aspects. However, the *ACC*s of training on subsets partitioned according to the original and substituted residue type at the nsSNP site are 76.3% and 74.9% respectively, which are 3.7 and 2.3 percents higher than that of training on the whole dataset separately, *i.e.*, both of these two partition methods outperformed the approach without partition apparently. These outcomes demonstrated that training on subsets partitioned appropriately but not randomly would lead to higher accuracy than that on the whole dataset. This ought to be due to that these two partition approaches generated more homogeneous subsets than taking all the data as a whole set, and hence improved the classifiers' performance. In addition, the *ACC* of training on subsets based on original residue type at the nsSNP site is 1.4 percents higher than that of training on subsets based on substituted residue type, indicating that the former is even more optimal than the later.

Taken together, the two dataset partition approaches presented here resulted in classifiers better than training on the whole dataset. This was because they were able to generate more homogeneous subsets than random partition. After partitioning properly, we actually grouped homogeneous data samples into the same group, which was favored in the training of SVM classifiers since there was less noise or random errors within the subsets than in the whole dataset. Moreover, since classifiers based on dataset partition according to original residue type gave rise to better performance than partition according to substituted residue type, it seemed that the former partition strategy was able to generate more homogeneous subsets than the latter, although it was hard to interpret its biological basis currently. Our findings here demonstrated that partitioning the nsSNP training dataset into homogeneous subsets according to the amino acid type at the nsSNP can improve the prediction of disease-associated nsSNPs.

4. Discussion

The continuously accumulated SNP data generated in the next generation sequencing projects require prediction and proper interpretation for their functional impacts and disease-association. Previous studies mainly focused on searching informative prediction features to build classifiers. In this work, we attempted a new strategy for improving current machine learning-based approaches, *i.e.*, properly partitioning the training dataset into subsets that were supposed to be more homogeneous than the whole dataset.

In order to train a good and stable classifier, two important aspects of the training set should be concerned: homogeneous samples for insuring high signal-to-noise ratio and adequate samples for avoiding over-fitting. Many previous studies didn't partition the training set into subsets probably due to that the number of training samples at that time were not enough for partition while keeping each subset with considerable sample size, so they had to treat all the data as a whole to guarantee adequate samples. The recent accumulation of nsSNP data with known functional annotation made the partition strategy feasible. The criteria adopted here partitioned the overall heterogeneous dataset into homogeneous subsets with considerable number of data samples. The improved classification accuracies demonstrated its superiority.

Jiang and colleagues' work has already described the method of data partition according to the original amino acid type at the nsSNP site actually, but hasn't evaluated its advantage and rationality [17]. In this work, we systematically compared this partition strategy with both random partition and the whole dataset without partition, and found that it did improve the classifiers' performance. We also tried the partition according to the substituted residue type at the nsSNP site, and found that it outperformed the random partition as well. Our results demonstrated that partition the training set appropriately, *e.g.*, the criteria adopted here, could improve the performance of the classifiers, and thus ascertained the advantage and rationality of dataset partition.

We can expect that further accumulation of the data in the future will enable us to partition the inflated training data into 380 subsets according to all the substitution types. There may exist more rational partition strategies which are worth exploring further. Moreover, this strategy adopted here can also be extended to other machine learning-based prediction models.

Acknowledgments

This work was supported by grants from the National "973" Basic Research Program of China (2012CB316501), the National '863' Hi-Tech Research and Development Program of China (2009AA022710), the National Natural Science Foundation of China (30800641, 31171268, 31000380, 30900834). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank Hui Yu, Rudong Li, Fudong Yu, Yunqin Chen, and Baohong Liu for helpful discussions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2012.01.138.

References

[1] F.S. Collins, A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, L. Walters, New goals for the U.S. Human Genome Project: 1998–2003, Science 282 (1998) 682–689.
[2] A.M. Glazier, J.H. Nadeau, T.J. Aitman, Finding genes that underlie complex traits, Science 298 (2002) 2345–2349.
[3] P.D. Stenson, E.V. Ball, M. Mort, A.D. Phillips, J.A. Shiel, N.S. Thomas, S. Abeysinghe, M. Krawczak, D.N. Cooper, Human.gene.mutation database (HGMD) update, Hum. Mutat. 21 (2003) (2003) 577–581.
[4] A. Ruepp, O.N. Doudieu, J. van den Oever, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, C. Skornia, S. Wanka, T. Rattei, P. Pagel, L. Riley,

D. Frishman, D. Surmeli, I.V. Tetko, M. Oesterheld, V. Stumpflen, H.W. Mewes, The mouse functional genome database (MfunGD): functional annotation of proteins in the light of their cellular context, Nucleic Acids Res. 34 (2006) D568–571.
[5] L. Bao, Y. Cui, Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information, Bioinformatics 21 (2005) 2185–2190.
[6] Y. Bromberg, B. Rost, SNAP: predict effect of non-synonymous polymorphisms on function, Nucleic Acids Res. 35 (2007) 3823–3835.
[7] E. Capriotti, R. Calabrese, R. Casadio, Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information, Bioinformatics 22 (2006) 2729–2734.
[8] J. Hu, C. Yan, Identification of deleterious non-synonymous single nucleotide polymorphisms using sequence-derived information, BMC Bioinformatics 9 (2008) 297.
[9] V.G. Krishnan, D.R. Westhead, A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function, Bioinformatics 19 (2003) 2199–2209.
[10] C.T. Saunders, D. Baker, Evaluation of structural and evolutionary contributions to deleterious mutation prediction, J. Mol. Biol. 322 (2002) 891–901.
[11] P. Yue, Z. Li, J. Moult, Loss of protein structure stability as a major causative factor in monogenic disease, J. Mol. Biol. 353 (2005) 459–473.
[12] Z.Q. Ye, S.Q. Zhao, G. Gao, X.Q. Liu, R.E. Langlois, H. Lu, L. Wei, Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP), Bioinformatics 23 (2007) 1444–1450.
[13] Y.L. Yip, H. Scheib, A.V. Diemand, A. Gattiker, L.M. Famiglietti, E. Gasteiger, A. Bairoch, The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants, Hum. Mutat. 23 (2004) 464–470.
[14] The_UniProt_Consortium, Ongoing and future developments at the Universal Protein Resource, Nucleic Acids Res. 39 (2011) D214–219.
[15] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, AAindex: amino acid index database, progress report, Nucleic Acids Res. 36 (2008) (2008) D202–205.
[16] J. Robinson, K. Mistry, H. McWilliam, R. Lopez, P. Parham, S.G. Marsh, The IMGT/HLA database, Nucleic Acids Res. 39 (2011) D1171–1176.
[17] R. Jiang, H. Yang, L. Zhou, CC. Kuo, F. Sun, T. Chen, Sequence-based prioritization of nonsynonymous single-nucleotide polymorphisms for the study of disease mutations, Am. J. Hum. Genet. 81 (2007) 346–360.
[18] C.Y. Li, Q. Yu, Z.Q. Ye, Y. Sun, Q. He, X.M. Li, W. Zhang, J. Luo, X. Gu, X. Zheng, L. Wei, A nonsynonymous SNP in human cytosolic sialidase in a small Asian population results in reduced enzyme activity: potential link with severe adverse reactions to oseltamivir, Cell Res. 17 (2007) 357–362.
[19] T. Huang, P. Wang, Z.Q. Ye, H. Xu, Z. He, K.Y. Feng, L. Hu, W. Cui, K. Wang, X. Dong, L. Xie, X. Kong, Y.D. Cai, Y.X. Li, Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties, PLoS One 5 (2010) e11900.
[20] J.M. Zimmerman, N. Eliezer, R. Simha, The characterization of amino acid sequences in proteins by statistical methods, J. Theor. Biol. 21 (1968) 170–201.
[21] P. Argos, J.K. Rao, P.A. Hargrave, Structural prediction of membrane-bound proteins, Eur. J. Biochem. 128 (1982) 565–575.
[22] D.E. Goldsack, R.C. Chalifoux, Contribution of the free energy of mixing of hydrophobic side chains to the stability of the tertiary structure of proteins, J. Theor. Biol. 39 (1973) 645–651.
[23] C. Chothia, Structural invariants in protein folding, Nature 254 (1975) 304–308.
[24] W.R. Krigbaum, A. Komoriya, Local interactions as a structure determinant for protein molecules: II, Biochim. Biophys. Acta 576 (1979) 204–248.
[25] M. Levitt, Conformational preferences of amino acids in globular proteins, Biochemistry 17 (1978) 4277–4285.
[26] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. USA 89 (1992) 10915–10919.
[27] R. Grantham, Amino acid difference formula to help explain protein evolution, Science 185 (1974) 862–864.
[28] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[29] T.D. Schneider, G.D. Stormo, L. Gold, A. Ehrenfeucht, Information content of binding sites on nucleotide sequences, J. Mol. Biol. 188 (1986) 415–431.
[30] C.-C.a.L. Chang, C.-J, LIBSVM: a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/≃cjlin/libsvm (2001).
[31] B.W. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta 405 (1975) 442–451.