遗传 HEREDITAS (Beijing) 2008 年 9 月, 30(9): 1157－1162
ISSN 0253-9772 www.chinagene.cn

DOI: 10.3724/SP.J.1005.2008.01157

离子通道亚型与其基因共表达的关联研究

杨德武[1], 李霞[1,2], 肖雪[1], 杨月莹[1], 王靖[1]

1. 首都医科大学生物医学工程学院, 北京 100069;
2. 哈尔滨医科大学生物信息系, 哈尔滨 150081

摘要: 离子通道亚型与其基因共表达的关联对研究离子通道功能有重要意义。文章采用主成分分析和模糊 C-均值聚类算法对数据进行分析, 将方法应用到人类和小鼠两套表达谱数据, 结果发现离子通道亚型中钾离子通道、钙离子通道、氯离子通道和受体激活型离子通道的表达谱聚类结果与生物学分类有较好的一致性, 体现了离子通道亚型在 mRNA 水平上的共表达, 并证实了通过离子通道表达谱能很好的对离子通道的功能亚型进行分类。

关键词: 离子通道; 基因表达谱; 离子通道亚型; 共表达; 关联

Association between ion channel subtype and its gene co-expression

YANG De-Wu[1], LI Xia[1,2], XIAO Xue[1], YANG Yue-Ying[1], WANG Jing[1]

1. *Department of Biological Engineering*, *Capital Medical University*, *Beijing* 100069, *China*;
2. *Department of Bioinformatics*, *Harbin Medical University*, *Harbin* 150081, *China*

Abstract: Association between ion channel functional subtype and its genes expression is important for exploring function of ion channel, annotating function of an unknown subtype and probing into molecular mechanism of ion channel diseases. In this study, we began with noise reduction by standardizing original micro-array data, which consisted of human and mouse gene expression profiles, and then we employed principle component analysis (PCA) together with fuzzy C-mean clustering algorithm to analyze the pre-processed gene expression profiles. PCA is applied to rebuild the feature space of human gene in 21 dimensions as well as the feature space of mouse gene in 26 dimensions. Using this method we largely reduced computational complexity without losing much information involved in the original data. Subsequently, fuzzy C-mean clustering was used to classify the ion channel genes of human and mouse in their reduced feature space. In the end, four ion channel functional subtypes, such as potassium ion channels, calcium ion channel, chloride ion channel, and receptor-mediated ion channel were clustered in both human and mouse gene feature space. We applied two statistic ways to conduct significance test of the findings. In one way, we randomly sampled the data for each functional subtype of the ion channel genes and recorded the true positive rate. As a result, in both human and mouse gene feature spaces, genes that

收稿日期: 2008–04–12; 修回日期: 2008–06–05
基金项目: 国家自然科学基金(编号：30571034, 30370798), 国家高技术研究发展计划项目(863 项目) (编号：2007AA02Z329), 2006 年北京市新世纪百千万人才工程, 北京市教育委员会科技发展计划项目(编号：KM200610025011)资助[Supported by the National Natural Science Foundation of China (No. 30571034 and 30370798), National High-Tech Research and Development Plan (No. 2007AA02Z329), New Century Hundred-Thousand-Ten Thousand Talents Project of Beijing City, Scientific Research Common Program of Beijing Municipal Commission of Education (No. KM200610025011)]
作者简介: 杨德武(1981–), 男, 硕士, 专业方向：生物信息学。Tel: 013241777617; E-mail: yt2610@163.com
通讯作者: 李霞(1957–), 女, 博士, 教授, 研究方向：生物信息学。Tel: 0451-86615922; E-mail: lixia@hrbmu.edu.cn

belong to one functional subtype were more likely to be clustered together than expected by chance. In the other way, we performed Kappa test and used the functional subtypes as gold standard. The result showed that consistency between the ion channel gene clusters and the ion channel gene subtypes was significantly high for both human and mouse. These results indicate that ion channel genes within the same functional subtype tend to be co-expressed at least at the mRNA-level.

Keywords: ion channel; gene expression profile; ion channel subtype; co-expression; association

离子通道(Ion channel)是细胞膜上的一类特殊亲水性蛋白质微通道，根据生物学功能特性可对离子通道进行分类：第一类是电压门控的离子通道，包括Na^+、K^+、Ca^{2+} 通道等；第二类是受体激活的离子通道，包括神经递质、激素等外源性化学物质以及机械和渗透压力刺激所激活的离子通道；第三类是第二信使激活的通道，包括由细胞内Ca^{2+}、IP_3、G蛋白以及蛋白激酶激活的离子通道。许多离子通道病，如某些先天性或获得性疾病就是离子通道基因缺陷与功能改变的结果[1~3]。

表达谱数据是通过基因芯片技术获取的，基因芯片又称DNA芯片。利用基因芯片可以同时检测数以千计的基因表达数据[4],而后利用生物信息学技术提取出这些高通量信息背后的生物学意义[5]。

有证据表明许多功能相关基因是共表达的，共表达基因可以揭示很多调控机制，而且基因在不同的细胞类型和细胞状态下具有不同的表达水平，基于以上原因我们可以通过对基因表达谱的分析实现对离子通道功能的生物学分类方面的研究[6]。在基因表达谱分析中,首先要进行预处理，另一个很重要的任务就是应用聚类分析技术对样本分类。如果聚类结果与离子通道的功能分类之间存在较好的一致性，我们就可以依此推断若某未知功能类型的离子通道与某已知功能类型的离子通道在基因表达水平上是相近的，那么它们就具有相同的生物学功能[7]。本文在数据预处理方面采用主成分分析方法来降低数据处理的复杂性，通过模糊 C-均值聚类对离子通道在 mRNA 表达水平进行分类，同时应用 Kappa 一致性检验来评价离子通道表达谱分类的显著性意义，结果显示有四类离子通道具有显著性意义。

1 材料和方法

1.1 材料

我们主要针对离子通道表达谱数据进行分析，数据来源于圣地亚哥化学与细胞生物学组织[8]，此套基因芯片数据为不同离子通道在不同组织中的mRNA表达水平，分为人类和小鼠两大类数据的 99 670 个条目，从中提取含 89 项指标的 120 个离子通道通蛋白表达谱数据(小鼠)和含 101 项指标的 141 个离子通道蛋白表达谱数据(人类)，其数据形式可以用x_{ij}表示，x_{ij}代表第i离子通道蛋白对应第j项条件下的mRNA表达水平。

1.2 方法

1.2.1 除噪标准化[8]

本文所研究的是离子通道表达谱数据，该数据中含有小值甚至是负值的存在，一般视为噪声来处理，在数据值小于 20 的时候将该值重置为 20[9]。同时表达谱数据的原始观测值由于各种变量的量纲和数量大小上是不一致的，需采用标准化方法给每种变量以统一度量[10]，这里使用的是标准差标准化方法。

1.2.2 主成分分析

芯片表达谱数据量比较大，我们需要考虑在这些基因中是否存在某些指标并未提供有显著意义的信息，精简数据提高分析结果的精确性，这一任务可由主成分分析(PCA)[11]完成:

$$S=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})^T \quad (1)$$

$$\lambda\omega=S\omega \quad (2)$$

其中，x_j代表第 i 个离子道蛋白所对应的向量，$\bar{x}$ 代表所有样本的平均向量，N 表示样本的总个数，把特征值按降序排列$\lambda_i\geq\lambda_{i+1}$，选择对应前 m 个非零特征值的特征向量作为主元。

在主成分分析中，我们用类间距水平来评价提取主成分的数目不同对分类的影响。类间距水平代表各聚类样本的平均距离的远近，D 值越大说明聚类效果越好，其公式如下:

$$D(m)=tr(S_w+S_b) \quad (3)$$

其中，S_b 为类间离散度矩阵，S_w 为类内离散度矩阵，

m 代表所提取的主成分的数目。

1.2.3　表达谱聚类算法

PCA技术在与那些需要事先确定类的个数的聚类分析方法合并应用时，算法意义更加突出，本文采用模糊C-均值聚类[12]对这些数据进行分析，依据此方法挖掘有显著性共表达的离子通道基因聚成相应的类，研究该分类与离子通道的功能分类之间的一致性。

模糊C-均值聚类[13]就是求使聚类目标函数J最小的模糊划分矩阵$U=\lfloor u_{ij} \rfloor_{C\times N}$，以及类别中心$V$。

$$J=\sum_{j=1}^{N}\sum_{i=1}^{C}(u_{ij})^m \parallel x_j - v_i \parallel^2$$

其中: v_i表示第i个聚类中心, i=1, 2,…, C; j=1, 2, ..., N; u是隶属度，代表第i类的加权向量，$m\in(1,\infty)$是加权指数,目标函数表示了各类数据到相应聚类中心的加权距离平方和。具体算法如下:

第一步: 确定聚类数目 C,初始化 m 及聚类中心 v_i;

第二步: 对第 t 次迭代，根据式

$$u_{ij}=\left(\sum_{t=1}^{C}\left[\frac{d_{ij}}{d_{tj}}\right]^{\frac{2}{m-1}}\right)^{-1} \quad 和 \quad v_i=\sum_{j=1}^{N}(u_{ij})^m x_j \Big/ \sum_{j=1}^{N}(u_{ij})^m \ ,$$

计算新的隶属度函数 u 和 C 个聚类中心 v_i;

第三步: 若$|J^{(t)}-J^{(t-1)}|\leqslant\varepsilon$，则停止，否则返回第二步继续迭代。

1.2.4　一致性检验

本文采用Kappa值作为评价一致性程度的指标，Kappa 值是 1960 年 Cohen 提出的，是用来描述一致性的较好指标。此外对 mRNA 表达水平的聚类结果与随机状态进行比较，进一步检验离子通道功能分类与表达一致性。

2　结　果

我们首先对基因表达谱数据进行除噪、标准化、特征提取等一系列的处理，在进行特征提取时，主成分分析应用于小鼠和人类两套数据。图 1 所示为表达谱数据进行特征提取的显著性分析结果，按照特征值从大到小的排序，横坐标代表提取的前 N 个特征值对应的主成分，纵坐标代表模糊 C-均在所提取的主成分数目不同时相应的类间距水平，从图中可以看出并没有明显的波峰对应显著性的主成分提取的个数，但是可以看出当提取的指标达到 26(小鼠)、21(人类)的时候其类间距的水平已不再有很明显的变化。其中对于小鼠数据，我们利用主成分分析方法提取了 26 个主成分后进行了分类，此时 $D_{26}\geq(D_{89}-D_1)\times0.95$; 对于人类的数据，我们提取了其中的 21 个主成分，同时也满足，$D_{21}\geq(D_{101}-D_1)\times0.95$ 这表明处理后的数据基本包含了原数据的全部信息，其中D_m代表提取的前m个主成分进行分类的类间距水平。这种处理也可以看作是为了精简数据降低复杂性，而将小鼠和人类两套数据分别降维至 26、21 进行聚类分析。

另一方面，根据离子通道的生物学功能特性可将离子通道分成若干功能类，其中有些类包含的数目太少，在考虑了各类包含数目差异之后我们选取 8 为阈值，将符合条件的功能类进行下一步的研究。

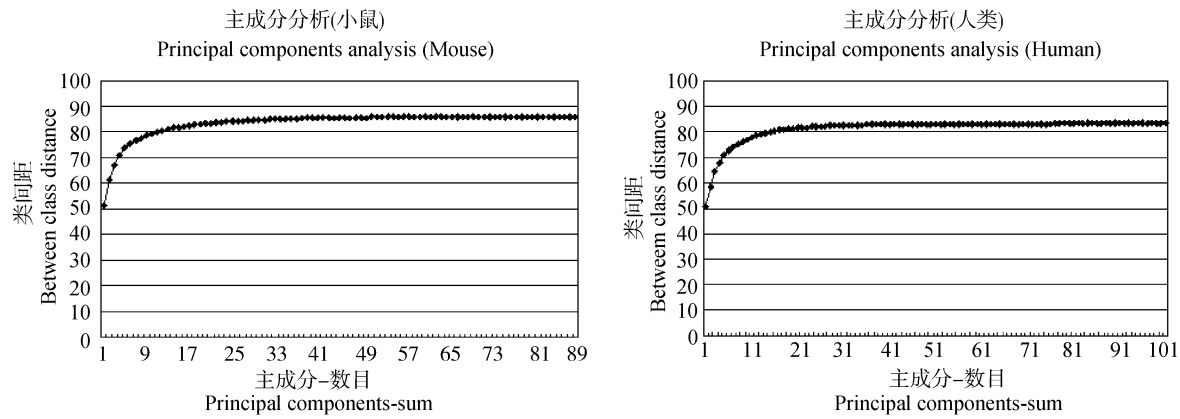

图 1　主成分分析显著性效果示意图

Fig. 1　The diagram of significance of principal components analysis

在设定阈值的条件下，共采用了 4 类进行研究，将表达谱的分类结果与离子通道生物学功能分类作Kappa一致性检验[14]，这里以生物学功能分类作为金标准，此外对生物学分类做了随机化处理，检验分类结果在随机状态下的*P*值，从两方面分析表达谱的分类与生物学分类之间一致性的显著性意义，其结果如表 1 和表 2 所示。其中，离子通道亚型所包含的离子通道数目，其与表达谱分类一致的离子通道数目，以及随机情况下最大可能的两种分类一致的离子通道数目均列于表中，两种检验方法计算的结果显示，Kappa一致性检验和随机状态对比检验的*P*值都明显小于 0.05，即基于离子通道表达谱数据的分类与离子通道功能分类之间存在较好的一致性。

表 1 离子通道表达谱亚型分类与生物学分类一致性(小鼠)
Table 1 The consistency between clustering result of ion-channel-gene expression profile and functional subtypes (mouse)

离子通道类型 Ion channel subtype	功能类数目 No. of ion channel	Kappa 一致性检验 Kappa test		随机对比检验 Random testing	
		共表达数 No. of co-expression	*P* 值 *P*-value	随机最大数 Maximum number in random	*P* 值 *P*-value
钾通道 Potasium channel subtype	49	26	0.002	20	0.010
钙通道 Calcium ion channel	9	4	0.008	1	0.011
氯通道 Chloride ion channel	11	5	0.001	1	0.001
受体型通道 Receptor-mediated ion channel	11	5	0.011	1	0.002

表 2 离子通道表达谱亚型分类与生物学分类一致性(人类)
Table 2 The consistency between clustering result of ion-channel-gene expression profile and functional subtypes (human)

离子通道类型 Ion channel subtype	功能类数目 No. of ion channel	Kappa 一致性检验 Kappa test		随机对比检验 Random testing	
		共表达数 No. of co-expression	*P* 值 *P*-value	随机最大数 Maximum number in random	*P* 值 *P*-value
钾通道 Potassium ion channels	57	33	0.001	23	0.013
钙通道 Calcium ion channel	26	18	0.001	4	0.001
氯通道 Chloride ion channel	14	7	0.001	2	0.001
受体型通道 Receptor-mediated ion channel	11	6	0.001	2	0.001

图 2 和图 3 分别为小鼠和人的显著性效果检验图，对小鼠和人类两套数据分别进行显著性描述。其中，灰色柱条代表随机状态下抽取一定数量的样本后包含该类对应横坐标数目的概率，黑色柱条代表表达谱数据进行一系列分析之后得到最终的聚类中含有该类对应数目的真阳性率。例如小鼠的离子通道生物学功能分类中钾离子通道数为 49 个，在 120 个样本中随机抽取 1 000 次 49 个样本，随机情况下最大可能出的钾离子通道数目为 20 个(灰色柱条横坐标值)，出现概率为 0.133(灰色柱条纵坐标值)，而按照离子通道表达谱的分类结果看，该类型离子通道数目为 26 个(黑色柱条的横坐标值)，其随机出现的概率仅为 0.002(黑色柱条的纵坐标值)，其分布如图 2A 所示，其他如受体型离子通道一致性分析如图 2D 所示，其 *P* 值为 0.011。图 3 人类的 4 种离子通道亚型也具有显著性意义，由以上结果可以看出表达谱数据分类结果比随机状态下分类更有显著性意义。

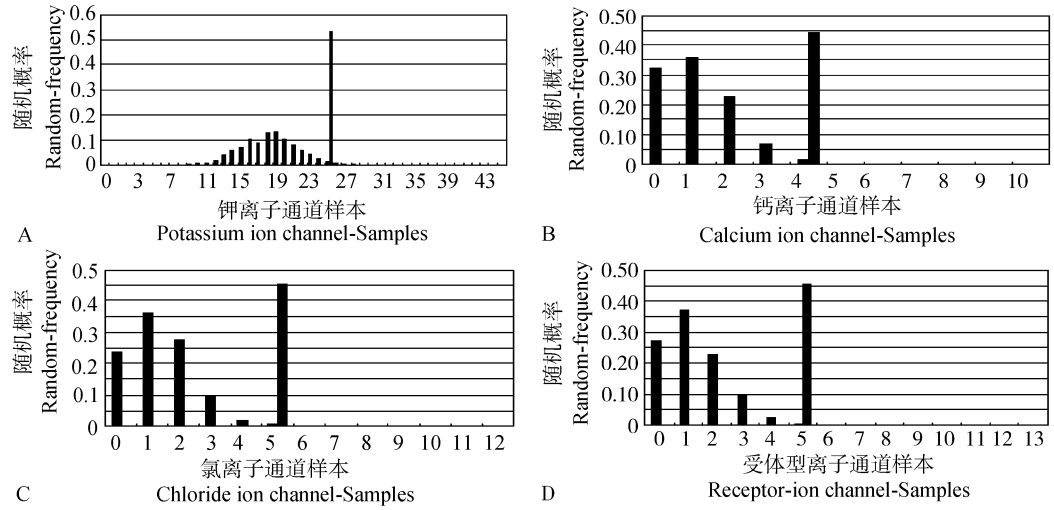

图 2　小鼠显著性效果检验图
Fig. 2　Significance test in mouse

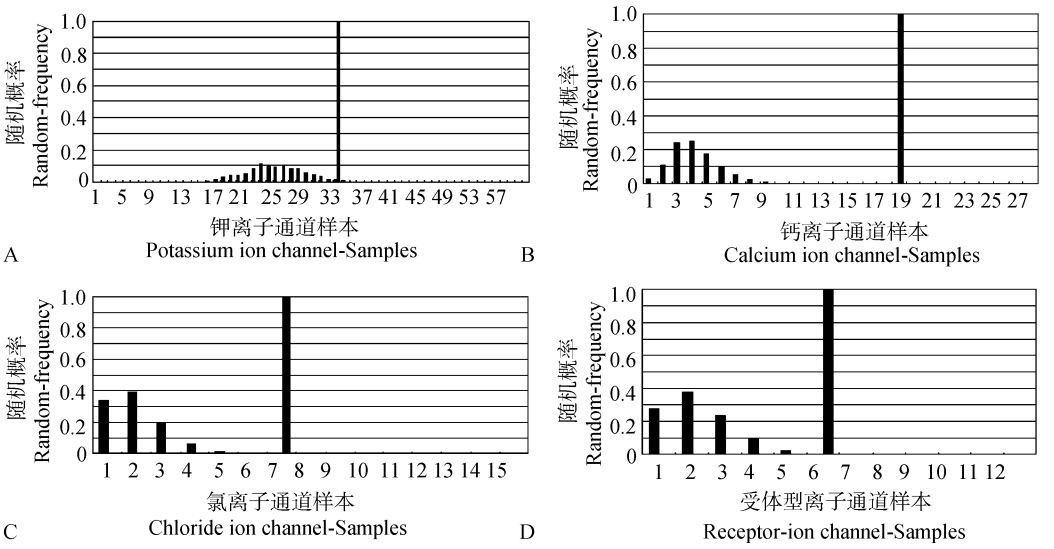

图 3　人类显著性效果检验图
Fig. 3　Significance test in human

3　讨　论

随着基因芯片技术的不断发展，大量的实验数据提供了相关研究的基础，但是如何准确挖掘数据背后的生物学意义仍然是科学研究的一大难题。同时，离子通道作为所有真核生物细胞保持正常生理功能必需的一大类跨膜蛋白[15]，许多复杂疾病的发生都与离子通道的功能异常有关[16]。

国内外已有很多关于离子通道的重要性的研究，也有很多研究围绕着基因表达差异与疾病的发生发展的联系[17]，但从离子通道mRNA表达水平层面研究离子通道功能的工作还不是很多，本文采用人类和小鼠的两套表达谱数据研究了离子通道功能分类与其基因共表达之间关系，并对离子通道亚型：钾离子通道、钙离子通道、氯离子通道和受体激活型离子通道分别进行了生物学亚型分类与表达谱分类之间一致性的检验。在进行除噪、标准化和主成分分析等一系列的处理之后，较之于随机状态下，离子通道表达谱数据的模糊 C-均值聚类结果与离子通道生物学功能分类之间存在显著的一致性，结果也同时证实了离子通道亚型在 mRNA 水平上是倾向于

共表达的。不过，由于离子通道蛋白本身数量较少，因此本研究所采用的数据量比较有限，在以后的工作中我们将进一步扩大样本含量。

从我们的工作可以看到，表达谱数据能较好的反映离子通道生物学功能，两者之间的相关性提示我们，离子通道mRNA水平的研究有助于离子通道疾病的早期诊断，并能够帮助我们判定一些新出现的未知类型离子通道疾病的性质，同时也为深入探索离子通道疾病发生及发展的分子机制提供了重要依据。我们的下一步工作将利用GO及KEGG数据库所提供的功能及通路信息进一步验证离子通道表达谱数据分类结果的生物学意义，并将结论应用到离子通道相关疾病的研究中去[18]。

参考文献(References):

[1] Brugada J, Brugada R, Brugada P, Channelopathies: a new category of diseases causing sudden death. *Herz*, 2007, 32(3): 185–191.[DOI]

[2] Heron SE, Scheffer IE, Berkovic SF, Dibbens LM, Mulley JC. Channelopathies in idiopathic epilepsy. *Neurotherapeutics*, 2007, 4(2): 295–304.[DOI]

[3] Ryan A, Matthews E, Hanna M. Skeletal-muscle channelopathies: periodic paralysis and nondystrophic myotonias. *Curr Opin Neurol*, 2007, 20(5): 558–563.

[4] YANG Chang, FANG Fu-De. Data analysis in microarray experiment. *Chinese Bull Life Sci*, 2004, (1): 41–48
杨畅，方福德. 基因芯片数据分析. 生命科学, 2004, (1): 41–48.

[5] Montero-Conde C, Martín-Campos JM, Lerma E, Gimenez G, Martínez-Guitarte JL, Combalía N, Montaner D, Matías-Guiu X, Dopazo J, de Leiva A, Robledo M, Mauricio D. Molecular profiling related to poor prognosis in thyroid carcinoma. Combining gene expression data and biological information. *Oncogene*, 2008, 27(11): 1554–1561.[DOI]

[6] Clements M, van Someren EP, Knijnenburg TA, Reinders MJ. Integration of known transcription factor binding site information and gene expression data to advance from co-expression to co-regulation. *Genomics Proteomics Bioinformatics*, 2007, 5(2): 86–101.[DOI]

[7] Huttenhower C, Hibbs M, Myers C. Troyanskaya OG. A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*, 2006, 22(23): 2890–2897.[DOI]

[8] Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 2003, 19(2): 185–193.[DOI]

[9] Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG, Hogenesch JB. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci USA*, 2002, 99(7): 4465–4470.[DOI]

[10] Lim WK, Wang K, Lefebvre C, Califano A. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, 2007, 23(13): 282–288.[DOI]

[11] Mamedov TG, Padhye NV, Viljoen H, Subramanian A. Rational de novo gene synthesis by rapid polymerase chain assembly (PCA) and expression of endothelial protein-C and thrombin receptor genes. *J Biotechnol*, 2007, 131(4): 379–387.[DOI]

[12] Asyali M, Alci M. Reliability analysis of microarray data using fuzzy c-means and normal mixture modeling based classification methods. *Bioinformatics*, 2005, 21(5): 644–649.[DOI]

[13] GONG Gai-Yun, MAO Yong-Cai, GAO Xin-Bo, LIU San-Yang. Fuzzy c-mean clustering method for analyzing microarray gene expression data. *J Xi'an Univ*, 2004, (3): 291–295.
宫改云，毛用才，高新波，刘三阳. 基于模糊 c-均值聚类的微阵列基因表达数据分析. 西安电子科技大学学报, 2004, (3): 291–295.

[14] Cohen J. A coefficient of agreement for nominal scales. *Educ Psych Meas*, 1960, 20(1): 37–46.[DOI]

[15] Viviani B, Gardoni F, Marinovich M. Cytokines and neuronal ion channels in health and disease. *Int Rev Neurobiol*, 2007, 82: 247–263.[DOI]

[16] Varga Z, Hajdu P, Panyi G, Gáspár R, Krasznai Z. Involvement of membrane channels in autoimmune disorders. *Curr Pharm Des*, 2007, 13(24): 2456–2468.[DOI]

[17] Ma X, Lee H, Wang L, Sun F. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics*, 2007, 23(2): 215–221.[DOI]

[18] Yi G, Sze SH, Thon MR. Identifying clusters of functionally related genes in genomes. *Bioinformatics*, 2007, 23(9): 1053–1060.[DOI]