❖医学物理与工程学

Application of imaging genetics method based on group sparse canonical correlation analysis in schizophrenia

WEI Fengxian[1], WU Jie[1*], YANG Ye[2], XIE Zhongxiang[1]
(1. School of Medical Instrument and Food Engineering, University of Shanghai For Science and Technology, Shanghai 200093, China; 2. EST Yishitong [Shanghai] Medical Equipment Co. Ltd, Shanghai 200093, China)

[**Abstract**] **Objective** To explore the correlation between imaging data and genetic data of schizophrenia patients using imaging genetics method. **Methods** A group sparse canonical analysis method was proposed, group sparse constraints $\lambda_1 \| u \|_G$ and $\lambda_2 \| v \|_G$ were added to sparse canonical correlation analysis model to select features groups. Then, features within one group were selected by sparse constraints $\tau_1 \| u \|_1$ and $\tau_2 \| v \|_1$. The imaging genetics method based on group sparse canonical correlation analysis method was used to analyze the correlation between brain regions and genes of schizophrenia, and the stability and ability of this method to select biomarkers were also verified. **Results** Several pairs canonical brain regions and genes were identified. The left insula and gene AKT1 demonstrated the most significant correlation (r=0.653 8), and r value between right rectus and gene DAOA, MAGI2 were larger than 0.6. The correlation coefficients of selected features were 0.626 9±0.016 1 with group sparse canonical correlation analysis and 0.625 5±0.018 1 with sparse canonical correlation analysis. After 10 selections, the proportion of 75 genes related to schizophrenia was higher than that of non-related genes randomly selected in the most related 20 genes selected by group sparse canonical correlation analysis. **Conclusion** Several pairs canonical brain regions and genes can be identified by the group sparse canonical analysis method, which may provide a new way for the study of schizophrenia and other complex mental disorders.

[**Keywords**] schizophrenia; sparse representation; canonical correlation analysis; single nucleotide polymorphism
DOI:10.13929/j.1003-3289.201807106

基于组稀疏典型相关分析方法的影像遗传学方法在精神分裂症中的应用

魏凤仙[1],武 杰[1*],杨 叶[2],谢忠翔[1]
[1.上海理工大学医疗器械与食品学院,上海 200093;2.伊士通(上海)医疗器械有限公司,上海 200093]

[**摘 要**] **目的** 采用影像遗传学研究方法探索精神分裂症患者的影像学数据与遗传学数据间的相关性。**方法** 提出一种组稀疏典型相关分析方法,在稀疏典型相关分析模型的基础上增加组稀疏惩罚项 $\lambda_1 \| u \|_G$ 和 $\lambda_2 \| v \|_G$ 进行变量组选择;对于选择的特征组,再利用组内惩罚项 $\tau_1 \| u \|_1$ 和 $\tau_2 \| v \|_1$ 进行组内的变量选择。采用基于组稀疏典型相关分析方法的影像遗传学方法分析精神分裂症患者脑区与相关基因位点的相关性,并验证其稳定性和筛选生物标记物的能力。**结果** 采用组稀疏典型相关分析方法获得了多组精神分裂症相关脑区和基因,其中左侧脑岛与基因 AKT1 的相关性最

[**第一作者**] 魏凤仙(1994—),女,河南周口人,在读硕士。研究方向:医学图像处理与分析。E-mail:octaviawfx@163.com
[**通信作者**] 武杰,上海理工大学医疗器械与食品学院,200093。E-mail:wujie3773@sina.com
[**收稿日期**] 2018-07-13 [**修回日期**] 2018-10-22

大,相关系数为 0.653 8;右侧直回与基因 DAOA 和 MAGI2 的相关系数均大于 0.6。组稀疏典型相关分析筛选出的特征的相关系数为 0.626 9±0.016 1,稀疏典型相关系数为 0.625 5±0.018 1。经过 10 次实验,在采用组稀疏典型相关分析方法筛选出的最相关的前 20 组特征中,属于已知的精神分裂症相关 75 个基因的比例大于随机选出的非相关基因的比例。**结论** 通过组稀疏典型相关分析方法能够筛选出多组精神分裂症的相关基因和脑区,为今后对精神分裂症等复杂精神类疾病的研究提供了新的思路。

[**关键词**] 精神分裂症;稀疏表示;典型相关分析;单核苷酸多态性

[**中图分类号**] R749.3;R445.2 [**文献标识码**] A [**文章编号**] 1003-3289(2019)02-0277-05

精神分裂症表现为情感、思维和行为的分裂,是一类复杂的精神疾病[1-2],研究其发病机制对于诊断及治疗均有重大意义[3]。精神分裂症的遗传度高达 81%[4]。在医学影像学方面,fMRI 已被广泛用于精神分裂症研究[5]。目前国内外学者多单独从影像学或遗传学层面进行研究,未能将二者结合。本研究尝试综合影像遗传学和信息,探讨精神分裂症患者影像学与遗传学数据的相关性。

人脑的每个脑区结构和功能各异,来自于同一个脑区的 fMRI 图像特征之间具有相关性,且同一个基因的单核苷酸多态性(single nucleotide polymorphisms, SNP)位点功能也具有相似性,即这两类特征均具有明显的组结构。特征的组结构属于先验信息,在研究中的价值不可忽视。既往研究[6-7]表明,针对生物学数据,分组处理具有有效性。考虑到稀疏典型相关模型局限于变量水平上的稀疏性,只有变量选择能力而无变量组选择能力,笔者提出组稀疏典型相关分析方法,针对 208 个样本,从每个样本中提取 41 236 组 fMRI 数据和 722 177 组 SNP 数据,以综合分析和 fMRI 影像数据遗传学 SNP 数据,探讨精神分裂症患者异常脑区与基因变异之间的关系。

1 资料与方法

1.1 一般资料 采用 MCIC 联盟(Mind Clinical Imaging Consortium)公开数据库中的 208 个样本,包括 92 例精神分裂症患者,男 70 例,女 22 例,年龄23~45 岁,平均(34.6±5.9)岁,均符合《精神障碍诊断与统计手册》第 4 版(DSM-Ⅳ)中关于精神分裂症的诊断标准[8];116 名健康志愿者,男 72 名,女 44 名,年龄21~43岁,平均(33.6±4.8)岁。本研究共使用来自 208 个样本的 41 236 组 fMRI 特征数据和722 177组 SNP 特征数据[9]。

1.2 方法

1.2.1 典型相关分析 给定来自于同一模式的 2 个样本集 X 和 Y,均含 n 个样本,假设 X 样本集的特征维数为 p,Y 样本集的特征维数为 q,表示如下:

$$X=\begin{matrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{matrix} \quad Y=\begin{matrix} y_{11} & \cdots & y_{1q} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nq} \end{matrix} \tag{1}$$

典型相关分析的目的是找到 2 组投影 α 和 β,以最大化 X 和 Y 的线性组合关系,公式如下:

$$(u,v)=\arg\max_{u,v}\left|corr(\alpha^T X,\beta^T Y)\right|$$
$$\text{s.t. } \alpha^T S_{XX}\alpha=1,\beta^T S_{YY}\beta=1 \tag{2}$$

其中 S_{XX}、S_{YY} 分别表示 X 和 Y 的方差矩阵和协方差矩阵,u、v 为典型向量,投影后得到的一维向量分别为 $X'=\alpha^T X$,$Y'=\beta^T Y$,即典型变量[10]。

此过程的求解一般有 2 种方法,即奇异值分解法(singular value decomposition, SVD)和特征值分解法,后者较为繁琐,对矩阵 K 进行 SVD 应用更为广泛,本研究采用 SVD 法首先求出初始的 u_0 及 v_0。矩阵 K 的定义为:

$$K=d_1u_1v_1^T+d_2u_2v_2^T+\cdots+d_ku_kv_k^T \tag{3}$$

k 是矩阵 K 的秩,d_i 是矩阵 K^TK 或 KK^T 的第 i 个特征值,u_i 及 v_i 是对应的典型向量,投影向量为 $\alpha=S_{XX}^{-1/2}u$ 和 $\beta=S_{YY}^{-1/2}v$。

1.2.2 稀疏典型相关分析 为观察精神分裂症患者影像学特征和遗传学特征之间的关系,针对 208 个样本,从每个样本中提取 41 236 组 fMRI 数据和 722 177 组 SNP 数据。由于特征数远大于样本数,式(2)的求解将导致过拟合问题。为此,本研究引入稀疏表示的思想,通过添加约束项使部分变量的系数收敛为 0,将式(2)改写为:

$$\min_{u,v}\|K-duv^T\|_F^2+\lambda\Psi(u)+\tau\Phi(v)$$
$$\text{s.t. } \|u\|_2^2=1,\|v\|_2^2=1 \tag{4}$$

$\Psi(u)$ 和 $\Phi(v)$ 分别为 u 和 v 的约束项。由于式(4)是非凸的,将条件放松至 $\|u\|_2^2\leqslant1$,$\|v\|_2^2\leqslant1$[11],但求最优解时,要求满足 $\|u\|_2^2=1$,$\|v\|_2^2=1$,如此,模型可以表示为:

$$\min_{u,v}\|K-duv^T\|_F^2+\lambda\Psi(u)+\tau\Phi(v)$$
$$\text{s.t. } \|u\|_2^2\leqslant1,\|v\|_2^2\leqslant1 \tag{5}$$

稀疏典型相关分析的约束项常采用 L_1 范数或 L_1 和 L_2 范数的组合,即弹性网络。

1.2.3 组稀疏典型相关分析 考虑到来自同一个脑区的体素和 SNP 在生理机制上有明显相关性,本研究对两类特征进行分组处理,提出组稀疏典型相关分析模型。此模型从特征组和个体特征两个层面施加约束项,将稀疏典型相关分析模型的 $\Psi(u)$ 和 $\Phi(v)$ 表示为如下形式:

$$\Psi_G u+\Psi(u)\rightarrow\Psi(u)$$
$$\Phi_G v+\Phi(v)\rightarrow\Phi(v) \quad (6)$$

Ψ_G 和 Φ_G 为组约束项。由此,结合稀疏表示的思想,将模型(5)表示为式(7)的最优化问题。

$$\min_{u,v}\|K-duv^T\|_F^2+\lambda_1\|u\|_G+\tau_1\|u\|_1+\lambda_2\|v\|_G+\tau_2\|v\|_1 \quad (7)$$

从式(7)可以看出,当 $\lambda_1=\lambda_2=0$ 时,模型则变成不存在组约束的模型。但模型(7)依然是非凸的,通过分解问题将其转化为式(8)和式(9)的双凸优化模型:

$$\min_u\|K-duv^T\|_F^2+\lambda_1\|u\|_G+\tau_1\|u\|_1+\Delta(\|u\|_2^2-1) \quad (8)$$

$$\min_v\|K-duv^T\|_F^2+\lambda_2\|v\|_G+\tau_2\|v\|_1+\Delta(\|v\|_2^2-1) \quad (9)$$

基于块坐标下降算法求解模型(8)和模型(9),使用软阈值函数来设置阈值。步骤如下:

输入:矩阵 K、初始的 u_0、v_0,稀疏系数 λ_1、λ_2、τ_1、τ_2 及精度 ε。

输出:典型向量 u、v。

① $\|soft_l(Kv)\|=S[(Kv_l),\tau_1]$,$l=1,2,3,\cdots L$,S 为软阈值函数;

②当 $\|soft_l(Kv)\|\leqslant\lambda_1$,$u_l^{j+1}=0$;否则 $u_l^{j+1}=\frac{Sg_l(Kv)}{\|Sg_l(Kv)\|}$,其中 $Sg_l(Kv)=\frac{1}{2}\left[soft(Kv)-\lambda_1\omega_l\frac{soft(Kv)}{\|soft(Kv)\|}\right]$;

③重复上述步骤直到 $\|u_l-u_{l-1}\|\leqslant\varepsilon$;

④以上求解 u_l 的方法同样适用于 v_l。

2 结果

2.1 影像学数据和基因数据之间的相关性分析 采用基于组稀疏的典型相关分析方法对 fMRI 数据以及 SNP 数据进行综合分析,使用交叉验证的方法确定模型的稀疏系数 λ_1、λ_2、τ_1 及 τ_2,目的是确定精神分裂症患者的影像学特征数据和遗传学特征数据之间的相关性。

依据精神分裂症基因库(https://bioinfo.uth.edu/SZGR1/index.jsp)提供的精神分裂症相关的 75 个基因[12-13],本研究从 208 个样本的 722 177 组位点中提取出上述 75 个相关基因的 6 151 组 SNP 位点,根据所在基因将这些位点分为 75 组;根据自动标记模板(anatomical automatic labeling, AAL),将 fMRI 数据根据其所在脑区分为 116 组。采用组稀疏典型相关分析方法,确定每个基因位点与 116 个脑区的 fMRI 特征数据之间的相关性。对选择到的特征按照相关性大小进行排序,相关系数>0.6 的基因和脑区见表 1。

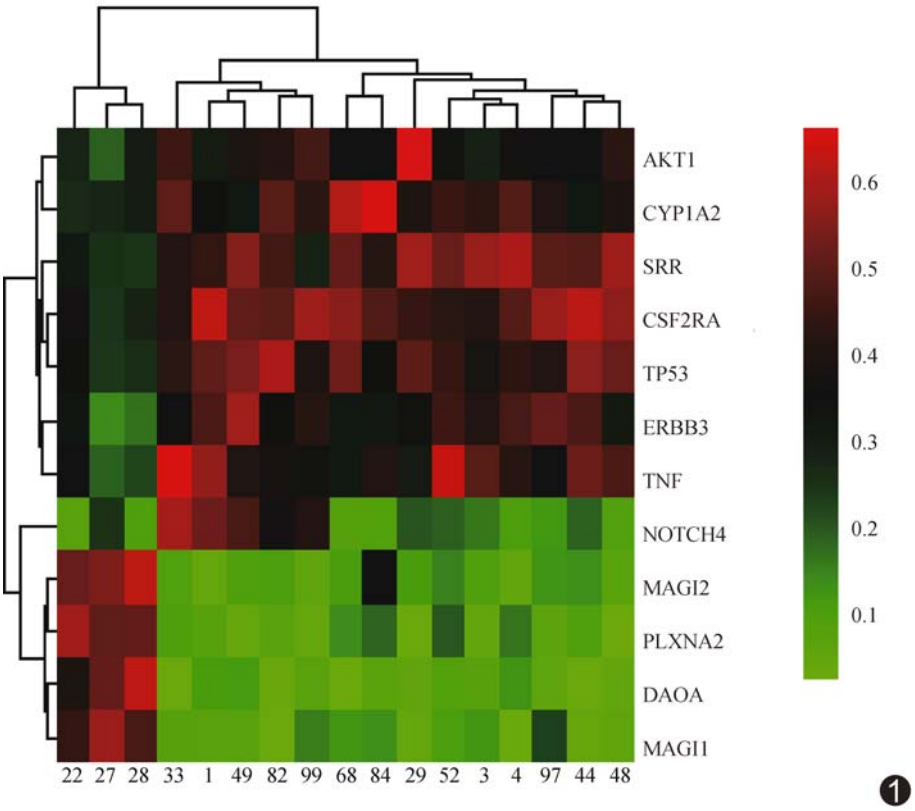
表 1 精神分裂症影像学数据与基因数据间相关系数>0.6 的基因和脑区

脑区	AAL 分区	基因	相关系数
左侧脑岛	29	AKT1	0.653 8
右侧颞上回	84	CYP1A2	0.642 7
左侧内侧和旁扣带回	33	TNF	0.642 3
右侧枕中回	52	TNF	0.629 6
左侧中央前回	1	CSF2RA	0.618 6
右侧直回	28	DAOA	0.617 4
右侧直回	28	MAGI2	0.617 1
右侧距状裂周围皮层	44	CSF2RA	0.615 3
右侧楔前叶	68	CYP1A2	0.605 4

2.2 相关性最大的 20 组数据特征 将全部 116 个脑区和 75 个基因的相关性按照大小排序,取相关性最大的前 20 组特征数据,并以热图的形式表示(图 1)。在图 1 中,列数据为按照 AAL 模板划分的脑区的编号,行数据代表各个基因,其中 AAL 分区为 28 的脑区右侧直回在前 20 对特征中出现 3 次,AAL 分区为 29、33 的左侧脑岛、左侧内侧和旁扣带脑回脑区分别出现 2 次,提示以上 3 个脑区与精神分裂症的遗传特性有比较显著的关系。

为对比本研究提出的组稀疏典型相关分析模型与稀疏典型相关分析模型进行相关性分析的能力和稳定性,计算了 2 个模型筛选出的特征的相关系数的均值和标准差,组稀疏典型相关系数为 0.626 9±0.016 1,稀疏典型相关系数为 0.625 5±0.018 1。

2.3 精神分裂症生物标记物的筛选 为观察本方法是否具有筛选精神分裂症生物标记物的能力,确定模型的参数后,将随机选出的非相关基因的位点与 75 个相关基因的位点置于同一个数据集中,使用组稀疏典型相关分析方法进行特征筛选;共进行 10 次筛选,每次实验均取相关基因与非相关基因各 20 组。根据选出的特征的相关性大小进行排序,结果显示,在采用组稀疏典型相关分析方法选出的最相关的前 20 组特征中,属于已知的精神分裂症相关 75 个基因的比例大于随机选出的非相关基因的比例(图 2)。



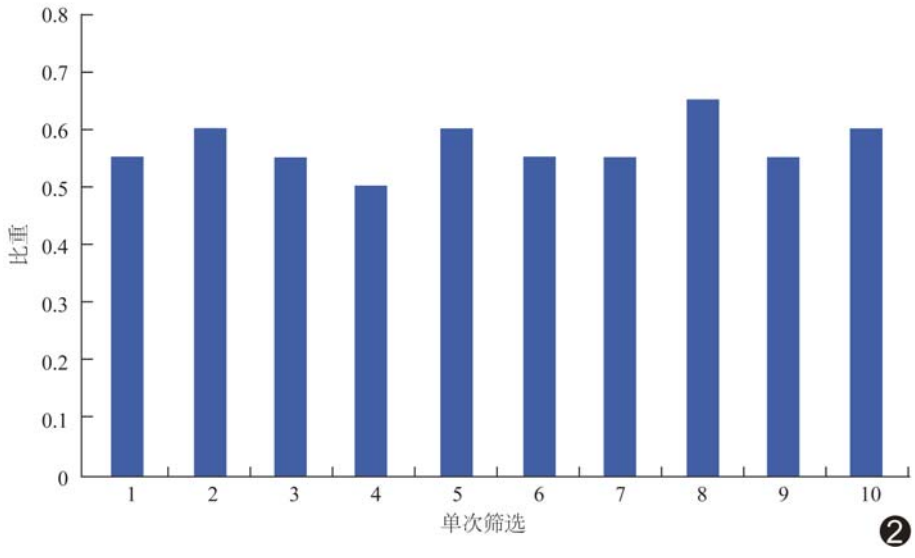


图 1　相关性最大的 20 组数据特征　　**图 2**　筛选能力分析

3　讨论

为分析精神分裂症的脑部病变与其易感基因之间的相关性，本研究提出基于组稀疏的典型相关分析方法，以整合分析两类影像遗传学数据——fMRI 数据和 SNP 数据，最终从 41 236 组 fMRI 特征数据和 722 177组 SNP 特征数据中找出相关系数大于 0.6 的 8 个脑区和 6 个易感基因，而对模型稳定性以及筛选能力的评估结果也提示此方法具有一定可靠性。

与稀疏典型相关模型相比，本研究提出的基于组稀疏的典型相关分析模型引入了组惩罚参数 λ_1 及 λ_2，令 $\lambda_1=\lambda_2=0$，则组稀疏典型相关分析模型转化为稀疏典型相关分析模型。本研究中，组稀疏典型相关分析模型所得相关系数的均数略大于稀疏典型相关分析，而标准差略小于稀疏典型相关分析，提示组稀疏典型相关分析模型可以提升稀疏典型相关分析模型的稳定性。

组稀疏典型相关分析模型在稀疏典型相关分析模型的基础上增加组惩罚项 $\lambda_1\|u\|_G$ 和 $\lambda_2\|v\|_G$ 来进行变量组选择，对于选择到的特征组再利用惩罚项 $\tau_1\|u\|_1$ 和 $\tau_2\|v\|_1$ 进行组内的变量选择。因此，组稀疏典型相关分析模型不仅可以在 fMRI 特征和 SNP 特征水平上进行筛选，还具有在脑区和基因水平上进行特征选择的能力。组稀疏典型相关模型实现了变量组水平上的稀疏性控制，具有变量组选择能力。由于在一次运行中确定 4 个参数的时间成本较高，本研究采用交叉验证方法确定模型的参数，首先固定 τ_1、τ_2 确定组惩罚参数 λ_1、λ_2，再确定组内惩罚参数 τ_1、τ_2。

本研究结果显示精神分裂症患者右侧直回与基因 DAOA 和 MAGI2 的相关系数均大于 0.6，提示此脑区与精神分裂症的基因生物标记物相关性显著[14-15]；此外，左侧脑岛与基因 AKT1 的相关系数最大[16]，为 0.653 8，后续研究可据此开展进一步实验。在相关性最大的前 20 组影像遗传学特征结果中，左侧脑岛、左侧内侧和旁扣带脑回与多个基因的相关性均较显著，故在热图中，行数据与列数据分别为 17 个脑区和 12 个基因，而非 20 个脑区和 20 个基因。

精神分裂症的遗传学研究结果错综复杂,每项研究只能从局部揭示其发病机制,对其发病机制和遗传机制尚需深入开展全面系统的研究。本研究基于稀疏表示利用典型相关分析方法研究脑区与精神分裂症的基因生物标记物之间的相关性,并根据 fMRI 数据和 SNP 数据自身的结构进行分组,结果表明采用组稀疏典型相关分析方法可以提取出与学界公认的精神分裂症的 75 个易感基因相关性较大的脑区,推测也可将其用于确定抑郁症等复杂精神类疾病的易感基因。

[参考文献]

[1] Sawa A, Snyder SH. Schizophrenia: Diverse approaches to a complex disease. Science, 2016,296(5568):692-695.

[2] 刘晋,林敬晖,洪楠.精神分裂症影像学研究.中国介入影像与治疗学,2014,11(2):121-124.

[3] Yuan Q, Yang F, Xiao Y, et al. Regulation of brain-derived neurotropic factor exocytosis and gamma-aminobutyric acidergic interneuron synapse by the schizophrenia susceptibility gene Dysbindin-1. Biol Psychiatry, 2016,80(4):312-322.

[4] Sullivan PF, Kendler KS, Neale MC. Schizophrenia as a complex trait: Evidence from a Meta-analysis of twin studies. Arch Gen Psychiatry, 2003,60(12):1187-1192.

[5] Du Y, Pearlson GD, Lin D, et al. Identifying dynamic functional connectivity biomarkers using GIG-ICA: Application to schizophrenia, schizoaffective disorder, and psychotic bipolar disorder. Hum Brain Mapp, 2017,38(5):2683-2708.

[6] Chen X, Liu H. Anefficient optimization algorithm for structured sparse CCA, with applications to eQTL mapping. Statistics in Biosciences, 2012,4(1):3-26.

[7] Chen J, Bushman FD, Lewis JD, et al. Structure-constrained sparse canonicalcorrelation analysis with an application to microbiome data analysis. Biostatistics, 2014,(2):244-258.

[8] Mendelson G. Diagnostic andstatistical manual of mental disorders, fourth edition (DSM-Ⅳ). Washington: American Psychiatric Association, 2015:4189.

[9] Gollub RL, Shoemaker JM, King MD, et al. The MCIC collection: A shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. Neuroinformatics, 2013,11(3):367-388.

[10] Chu D, Liao LZ, Ng MK, et al. Sparse canonical correlation analysis: New formulation and algorithm. IEEE Trans Pattern Anal Mach Intell, 2013,35(12):3050-3065.

[11] Gossmann A, Zille P, Calhoun V, et al. FDR-corrected sparse canonical correlation analysis with applications to imaging genomics. IEEE Trans Med Imaging, 2018,37(8):1761-1774.

[12] Zhao Z, Webb BT, Jia P, et al. Association study of 167 candidate genes for schizophrenia selected by a multi-domain evidence-based prioritization algorithm and neurodevelopmental hypothesis. PLoS One, 2013,8(7):e67776.

[13] Sun J, Han L, Zhao Z. Gene- and evidence-based candidate gene selection for schizophrenia and gene feature analysis. Artif Intell Med, 2010,48(2-3):99-106.

[14] Koide T, Banno M, Aleksic B, et al. Common variants in MAGI2 gene are associated with increased risk for cognitive impairment in schizophrenic patients. PLoS One, 2012, 7(5):e36836.

[15] Ma J, Sun J, Zhang H, et al. Evidence for transmission disequilibrium at the DAOA gene locus in a schizophrenia family sample. Neurosci Lett, 2009,462(2):105-108.

[16] Emamian ES, Hall D, Birnbaum MJ, et al. Convergent evidence for impaired AKT1-GSK3beta signaling in schizophrenia. Nat Genet, 2004,36(2):131-137.

关键词

关键词又称主题词,是位于摘要之后,在论文中起关键作用的、最能说明问题的、代表论文特征的名词或词组。它通常来自于题目,也可以从论文中挑选。一般每篇论文要求 2～5 个关键词。每个关键词都可以作为检索论文的信息,若选择不当,会影响他人的检索效果。医学上现在主要使用美国《医学索引》(Index Medicus)的医学主题词表(Medical Subject Headings, MeSH)最新版作为规范,亦可参考中国医学科学院情报研究所翻译地英汉对照《医学主题词注释字顺表》。非主题词表的关键词为自由词,只有必要时,才可排列于最后。有些新词也可选用几个直接相关的主题词进行搭配。