生物信息学在脑疾病研究中的应用

刘冰
中国科学院自动化研究所中法信息、自动化与应用数学联合实验室(LIAMA)计算医学研究中心，北京 100190

摘　要　随着脑疾病相关的各种大规模生物学数据的产生，生物信息学的研究方法与策略正开始深入到脑疾病研究的各个层面，并取得了很多振奋人心的科研成果；同时，不断涌现的各种脑疾病全基因组数据也从生物信息学计算理论与方法层面上对信息科学以及系统科学提出了巨大的挑战。本文分别从基因组、转录组及蛋白质组、表型组以及交互作用组等不同层次总结阐述生物信息学在脑疾病研究中的具体应用。

关键词： 生物信息学；脑疾病；生物学标记；系统生物学

Applications of Bioinformatics in Brain Disease Study

Liu Bing
Research Center of Computational Medicine, Sino-French Laboratory for Computer Science, Automation and Applied Mathematics (LIAMA), and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract: Based on various brain disease related large-scale biological datasets, bioinformatics methods are being applied on each level of brain disease study and many successful research results have been achieved. Meanwhile, it is also a great challenge for informatics science and systems science to deal with various novel genome-wide biological data for brain disease. In this paper, we summarized specific applications of bioinformatics in brain disease study respectively from genome, transcriptome, proteome, phenotome, and interactome levels.

Keywords: Bioinformatics; Brain Disease; Biomarker; Systems Biology

1. 引言

随着社会老龄化以及现代化进程的加快，各种神经精神疾病等脑疾病，尤其精神分裂症(Schizophrenia)、阿尔茨海默氏病(Alzheimers' Disease, AD)等常见疾病，已经成为21世纪威胁人类健康最严重的疾病类型之一，给患者、家庭以及社会造成沉重的负担，日益成为世界重大的公共卫生问题。然而，到目前为止，这些疾病的病因和发病机制仍未完全阐明，基础研究的滞后极大地影响了对疾病的早期诊断和有效治疗。

越来越多的研究提示，各种常见脑疾病属于复杂的多基因遗传疾病，是由多基因、多因素、遗传和环境共同作用的结果[1]，对于此类复杂疾病分子机制的理解是21世纪医学科学中的巨大挑战。然而，到目前为止，绝大多数的遗传学研究仍然采用传统的单基因疾病的研究方法。对于神经精神疾病等复杂疾病的研究，传统的遗传学方法尽管取得了巨大的研究进展，但总体来说并没有获得本质上的突破，遇到了诸多难以逾越的困难。例如，越来越多的研究证实：经典的连锁研究结合突变分析，对于复杂疾病的研究不是很有效，主要由于这种情况下单个基因的作用太小；而候选基因的病例对照关联研究，虽然简单和应用广泛，但其结果却很难被重复验证，而且在疾病的病理生理尚不清楚前，候选基因方法不能完全地解释疾病的遗传基础。近年来全基因组的关联研究（Genome-Wide Association Study）正成为一种重要的发展方向，但这种方法的应用还有很多障碍需要克服，包括成本、样本量、交互作用的分析、假阳性率等一系列问题[2] [3]。

与此同时，与征服宇宙的计划相媲美的人类基因组计划（Human Genome Project，HGP），给21世纪的社会带来了很大的冲击，给人类全面认识自我以及彻底攻克人类重大疾病带来了极大的鼓舞和希望。人类基因组计划完成之后，我们进入了后基因组时代，也即功能基因组学研究这一崭新的生物学世纪。基因组学研究基因组序列变异和基因表达产物在机体发育、分化及疾病中的作用，其中包括基因组（Genome）研究、转录组（Transcriptome）研究、蛋白质组（Proteome）研究、交互作用组（Interactome）研究等。而且，基于上述“组学”发展了各种高通量的生物技术，带动了各种大规模生物学数据（包括DNA/蛋白质序列数据、基因芯片数据以及各种基因-蛋白质交互作用数据等）的迅猛增长，生物学进入了“淘宝”时代。我们以DNA序列数据为例，从来自GenBank数据中心自1982年到2008年间的数据统计图（图1），我们可以大致看出目前大规模生物学数据呈现何种

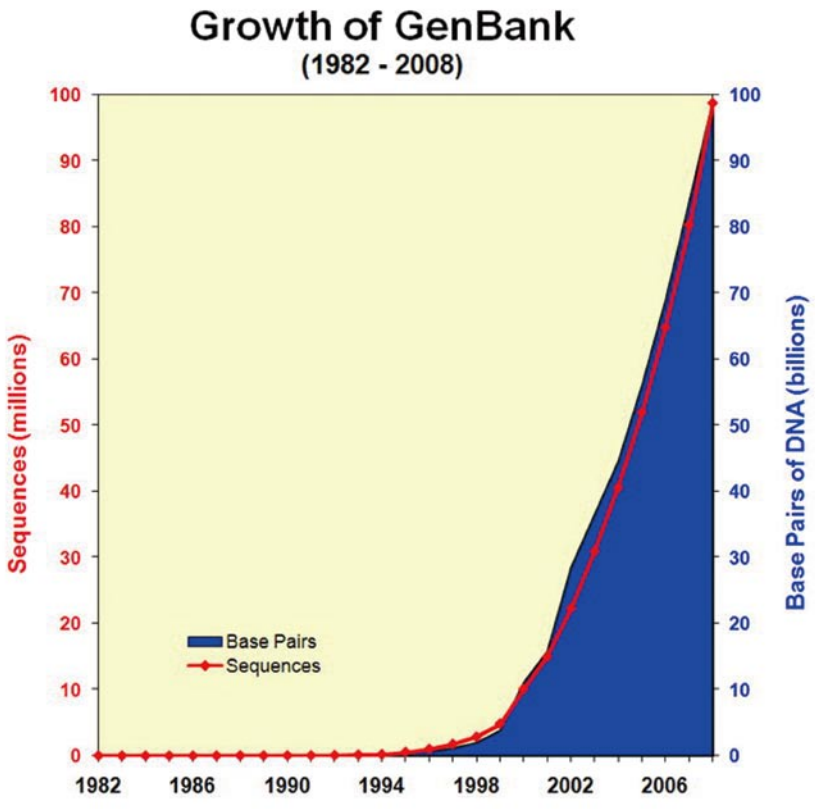

图1 DNA序列数据的增长趋势

（本图来自http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html）

程度的增长速度。

然而，如何充分利用这些海量数据，从中发现对人类健康有用的信息，从而真正实现最初“造福全人类”的口号，离不开信息科学的介入。在这种形势下，一个名为生物信息学(Bioinformatics)的新研究领域出现了，它正成为自然科学中多学科交叉的、有活力的、有影响的新领域。生物信息学，是20世纪随着计算机技术的迅猛发展、以及生物学数据的海量增长而迅速发展起来的。实际上，生物信息学到现在也没有一个统一的定义。在此，本文采用1995年美国国立卫生研究院给出的定义：生物信息学是一门包括生物学数据的获取、处理、存储、发布、分析和解释的科学学科，它综合了数学、计算机科学以及生物学的工具和技术，其目标是理解海量数据的生物学意义[4]。基于各种高通量的生物学数据，借助生物信息学技术与手段对脑疾病进行研究，有望突破单基因病分析方法在神经精神疾病研究中的局限，能够促进神经精神疾病的研究向功能系统的方向发展，并在疾病的诊断、治疗、药物开发等方面提供有价值的理论指导和分析。

近年来，生物信息学的研究方法与策略正开始深入到脑疾病研究的各个层面，并取得了一些振奋人心的科研成果；同时，不断涌现的神经精神疾病全基因组数据也从生物信息学计算理论与方法层面上对信息科学以及系统科学提出了巨大的挑战。概括来讲，我们下面分别从基因组、转录组及蛋白质组、表型组以及交互作用组等层次分别阐述生物信息学在脑疾病研究中的具体应用。

2. 在基因组层次研究中的应用

随着HGP的实现、分子技术的巨大进步、以及后基因组学时代生物信息学研究的不断深入，神经精神疾病等脑疾病的遗传学研究亦得到迅速发展。首先，在后基因组时代，人们基于对基因组的较全面认识，对各种经典遗传学模型进行了修改与完善，并发展了更加准确和复杂的模型。人们不仅对数量遗传学（Quantitive Genetics）和群体遗传学（Population Genetics）有了新的诠释，也发展了表观遗传学(Epigenetics)[5]等新的研究领域。表观遗传学实际上是与传统遗传学相对应的概念。遗传学，是指基于基因序列改变所致基因表达水平变化，如基因突变、基因杂合丢失和微卫星不稳定等；而表观遗传学，则是指基于非基因序列改变所致基因表达水平变化，如DNA甲基化、组蛋白修饰以及非编码RNA调控等。表观遗传学补充了“中心法则”（图2）忽略的两个问题，即哪些因素决定了基因的正常转录和翻译，以及核酸并不是存储遗传信息的唯一载体。近几年，表观遗传学已经深入到各种脑疾病机制的研究中[6-9]。我们通过对表观遗传中各种因子的突变，以及所导致疾病的研究，将有助于我们了解表观遗传机制，进而指导复杂疾病的治疗和新药的研制。其次，人类

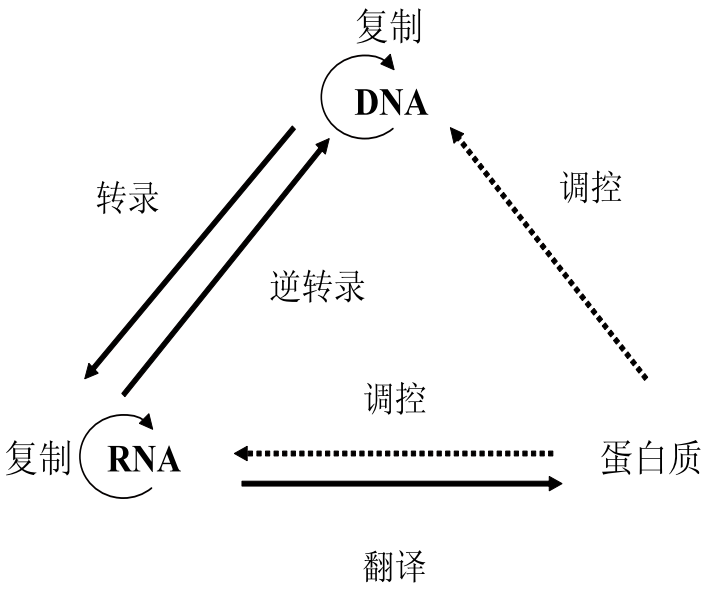

图2 中心法则

基因组计划以及其后的人类单体型计划（HapMap Project）的实施与完成，对后基因组时代神经精神疾病的遗传学研究也有着巨大的影响。以此为基础，多项神经精神疾病的全基因组关联研究已经陆续发布，包括精神分裂症、双相情感障碍等[2][3]。然而，如何充分利用这些人类共有资源，引入基因位点之间的相互作用等复杂信息，对如此海量的脑疾病全基因组关联研究数据进行有效的分析，从而发现疾病的有效生物标记，是信息科学界面临的巨大挑战。

此外，随着计算技术以及文献挖掘方法的发展，科学家们希望能够充分利用已有的研究成果，用来探索脑疾病的分子机制。首先，可以通过文献信息的挖掘，预测复杂疾病相关的基因。例如，Hristovski等人开发了BITOLA系统，可以通过挖掘PubMed中的文献信息，预测复杂疾病相关的基因[10]。Tiffin等人通过结合文献挖掘，以及复杂疾病的基因表达数据分析，发展了有效的复杂疾病基因预测算法[11]。其次，通过对已有研究的meta分析，可以客观定量地评价已有遗传学研究的结果。在脑疾病的遗传学研究领域，meta分析得到了广泛应用。例如，Levinson等人发展了有效的针对全基因组连锁研究的meta分析方法，并应用于精神分裂症和双向情感障碍（Bipolar Disorder）全基因组扫描研究的分析，最终确定了显著连锁的一些区域[12-14]。Bertram等人完成了有关AD所有关联研究的系统meta分析，就是综合分析了已有有关AD的关联研究，并建立了可动态更新的数据库，为进一步的遗传学研究提供了很有价值的工具与线索[15]。

3. 在转录组及蛋白质组层次研究中的应用

近年来基因以及蛋白芯片技术迅速发展，由于它可以同时监测成千上万个基因或蛋白在某一状态下的表达情况，为科学家们系统阐明生命以及疾病的本质提供了可能性。然而，大规模基因表达谱数据的合理利用，同样离不开高通量的计算技术。因此，基因芯片数据的分析与利用，是脑疾病生物信息学研究的一个重要方面。概括起来，基因芯片技术及其相应的生物信息学方法在脑疾病研究中的应用，大致可以概括为以下三个方面的研究：发现疾病差异表达基因、构建疾病基因调控网络以及疾病的分类与诊断。首先，发现疾病差异表达基因，是通过搜寻在不同疾病状态、以及疾病与正常人之间表达水平有显著差异的基因，以此确定在疾病发生发展过程中起重要作用的基因，从而为复杂疾病的理解以及治疗提供理论依据。其次，基于大规模基因表达谱数据，亦可以探索基因与基因之间的相互作用关系，即构建复杂的基因调控网络，从而能够从系统层次上认识脑疾病的发生发展机制。最后，由于基因芯片可以发现某特定疾病特定的基因表达模式，因此可以用于脑疾病的分类与诊断。基于基因芯片数据，人们已经发展了各种分类方法，可以对多种脑疾病进行有效准确的分类与诊断。我们基于基因组和蛋白质组的海量芯片数据建立了新的生物信息学算法模型，发展了一种基于多特征融合的集成神经网络方法，通过融合多种特征，可充分利用已有信息，取得比目前已有算法更高的准确率[16]。

事实上，神经精神疾病的基因芯片研究，相对于肿瘤等疾病的研究要困难的多，主要在于实验样本的难以收集，尤其对于国内的研究人员来说，收集病例的尸检脑组织几乎是不可能的。于是，有研究者试图通过检测血液中全基因组的表达情况，鉴别出血液中有效的神经精神疾病生物学标记[17-18]。尽管这种研究有一定的局限性，并不一定能完全真实地反映脑组织的病变情况，却为神经精神疾病的基因组研究开辟了新的曙光，另外，血液生物学标记也能为疾病的早期诊断提供方便可行的指标[19]。因此，从生物信息学的角度，系统分析具体脑疾病的脑组织与血液的基因以及蛋白芯片数据，寻找系统层次上的共同生物标记，以及进行相应分类研究是脑疾病研究中的一个重要内容。

4. 在表型组层次研究中的应用

由于各种神经精神疾病有着相似的临床症状，例如精神分裂症与双相情感障碍的相似程度很高，阿尔茨海默氏病也有一些类似抑郁症的临床症状，通常假设这些疾病可能存在一些共同的分子机制。近期的两项研究通过从不同角度分析不同疾病间的相似性，建立了大规模的疾病表型－表型之间的关联关系——疾病表型网络[20-21]。我们曾通过对疾病表型网络进行自动地网络模块化分析以及重要特性分析，发现疾病表型网络存在着很好的模块化性质，相似的疾病趋向于在同一个网络模块中（图3），各种神经精神疾病基本上聚集在同一网络模块中，为神经精神疾病可能存在一些共同的分子机制这一假说提供了一定的证据支持[22]。因此，考虑不同神经精神疾病表型以及表型之间的相互关系，结合大规模基因芯片数据，全基因组关联研究数据等，进行疾病表型组层次的分析，以试图发现神经精神疾病特有表型对应的生物学标记是生物信息学在表型组层次研究中的重要应用。

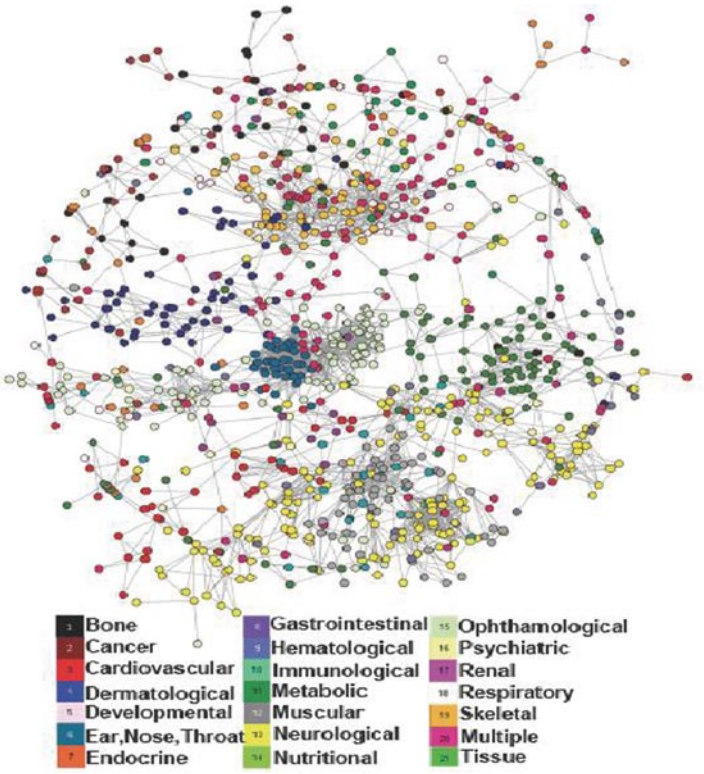

图3 疾病表型网络模块化与疾病分类的图示

5. 在交互作用组层次研究中的应用

既往脑疾病的研究方式多是基于“还原论”的思路，无法完整地理解系统水平的行为规律，因此对脑疾病机理的认识很难取得本质上的突破。而事实上，各种生物学机制都不可避免地要涉及分子之间的相互作用，以及反应的时空依赖性。因此，基于各种脑疾病相关的大规模生物医学数据，如何从“相互作用—网络－功能”的模式出发，有效地组织这些数据，最大限度的挖掘有用信息，最真实地反映生物体内的网络拓扑结构以及动态特性，并利用基因调控网络、蛋白质相互作用网络以及疾病表型网络的丰富信息，对神经精神疾病的分子机制进行研究，是亟待研究的热点问题。我们已发展了一种基于复杂脑基因网络批量预测阿尔茨海默氏病基因的计算方法，初步提出了一种全新的大批量预测复杂脑疾病基因的计算方法，基于复杂脑基因网络方便快捷地从系统层次上预测复杂脑疾病基因。本方法使用计算系统生物学的研究策略，基于贝叶斯统一框架，结合基因组学、蛋白质组学以及遗传学等信息，试图建立复杂脑疾病的致病模型，并将这一新方法应用于阿尔茨海默氏病基因的预测，取得了较好的预测效果，可以为相关遗传学研究提供非常有价值的参考信息[23]。这一研究方法，为深入理解神经精神疾病的病理机制开辟了一个新的研究思路。随着各种新的大规模生物学数据的出现，这一方法有待于进一步改进和提高，并应用于其它神经精神疾病的研究，预测的实验结果也有待深入

的分子实验验证。

此外，交互作用组不仅应该研究不同组学（基因组、转录组、蛋白质组以及表型组）内部之间的交互作用关系，同时应该关注不同层次之间的交互作用，构建从基因组-〉转录组-〉蛋白质组-〉表型组之间不同层次生物网络的桥梁，从而真正从整体上对神经精神疾病进行系统性研究。2007年发表在Nature Genetics上的一项研究通过对193例正常脑组织同时进行全基因组的SNP芯片和基因表达芯片分析，初步建立了脑组织中基因组到转录组之间的对应关系[24]。这一研究领域，也可以称做遗传基因组学（Genetical Genomics）[25]，也是一个非常有前景的发展方向。如图4所示，传统的遗传学研究，通过发现复杂性状所含有基因序列上的变异，寻找表型与基因型的对应关系，以解释复杂性状的遗传特点；而定位于转录组层次的基因芯片技术，则通过发现某些条件下基因表达水平的差异，从而解释复杂性状或者疾病；而遗传基因组学则是通过结合遗传学与基因组学的研究来寻找复杂疾病基因。

近年来，陆续有多项研究，通过集成全基因组的连锁分析、以及全基因组的基因表达谱来寻找复杂疾病的基因[26-28]，揭开了复杂疾病研究的新篇章。其研究思路大致可以分为两类：一种是分别进行遗传学的连锁研究以及复杂疾病的基因表达研究，然后寻找在两个水平上均有显著变化的基因，作为所研究疾病的候选基因；另外一种研究思路则是，通过把全基因组的基因表达谱作为复杂疾病的数量遗传性状（Expression Quantitive Trait Loci, eQTL），然后结合复杂疾病的表型性状（Physiological Quantitive Trait Loci, pQTL），进行序列标记的连锁研究[27-28]。总之，遗传基因组学为脑疾病的研究开辟了新的思路。其目标是基于现有的各种公共“组学”数据，通过建立有效的数学模型，以发现不同组学层次之间的内在联系，从整体上对脑疾病进行系统性研究。

6. 总结与展望

综上，基于多种公共“组学”数据，通过发展相应生物信息学的计算理论与方法，建立有效的数学模型，对神经精神疾病的分子机制进行系统性研究，深入理解疾病发病机制，并发现其有效生物学标记和建立早期诊断模型是国际前沿研究热点之一。我们相信，以系统观为特点的生物信息学研究策略与方法，有望突破单基因病分析方法在脑疾病

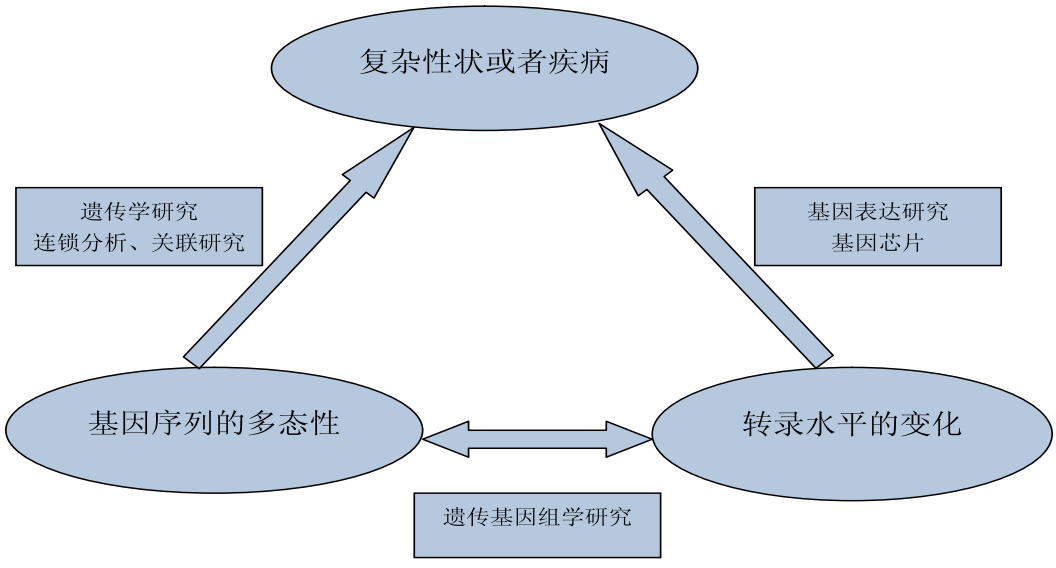

图4 遗传基因组学研究示意图

研究中的局限，能够在生物学和临床医学的诊断、治疗、药物开发等方面提供有价值的理论指导和分析。生物信息学，作为当今生命科学研究最重要的平台技术，不仅能够分析脑疾病相关的多种生物分子数据，同时更适于综合多种生物分子及其相互作用的知识来了解系统的功能，由此促进脑疾病的研究向功能系统的方向发展。

参考文献:

[1] Tsuang M. Schizophrenia: genes and environment. Biol. Psychiatry , 2000, 47: 210-220.

[2] Lin S, Chakravarti A and Cutler DJ. Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. Nat Genet , 2004, 36: 1181-1188.

[3] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature , 2007, 447(7145):661-678.

[4] Understanding Our Genetic Inheritance. The US Human Genome Project: The First Five Years , 1991-1995. NIH Publibcation No. 901590, April, 1995.

[5] Singh SM, Murphy B, and O'Reilly R. Epigenetic contributors to the discordance of monozygotic twins. Clin Genet , 2002, 62: 97-103.

[6] Jones PA, and Baylin SB. The epigenomics of cancer. Cell, 2007, 128: 683-692.

[7] Ducasse M, and Brown MA. Epigenetic aberrations and cancer. Mol Cancer, 2006, 5: 60.

[8] Abdolmaleky HM, Thiagalingam S, and Wilcox M. Genetics and epigenetics in major psychiatric disorders: dilemmas, achievements, applications, and future scope. Am J Pharmacogenomics, 2005, 5: 149-160.

[9] Bjornsson HT, Fallin MD, and Feinberg AP. An integrated epigenetic and genetic approach to common human disease. Trends Genet , 2004, 20: 350-358.

[10] Hristovski D, Peterlin B, Mitchell JA, et al. Using literature-based discovery to identify disease candidate genes. Int J Med Inform, 2005, 74: 289-298.

[11] Tiffin N, Kelso JF, Powell AR, et al. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res , 2005, 33:1544-1552.

[12] Levinson DF, Levinson MD, Segurado R, et al. Genome scan meta-analysis of schizophrenia and bipolar disorder, part I: Methods and power analysis. Am J Hum Genet , 2003, 73: 17-33.

[13] Lewis CM, Levinson DF, Wise LH, et al. Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. Am J Hum Genet , 2003, 73: 34-48.

[14] Segurado R, Detera-Wadleigh SD, Levinson DF, et al. Genome scan meta-analysis of schizophrenia and bipolar disorder, part III: Bipolar disorder. Am J Hum Genet , 2003, 73: 49-62.

[15] Bertram L, McQueen MB, Mullin K, et al. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet , 2007, 39: 17-23.
[16] Liu B, Cui Q, Jiang T, Ma S. A combinational feature selection and ensemble neural network method for classification of gene expression data. BMC Bioinformatics , 2004, 5:136.
[17] Cui DH, Jiang KD, Jiang SD, et al. The tumor suppressor adenomatous polyposis coli gene is associated with susceptibility to schizophrenia. Mol Psychiatry , 2005, 10(7):669-677.
[18] Maes OC, Xu S, Yu B, et al. Transcriptional profiling of Alzheimer blood mononuclear cells by microarray. Neurobiol Aging, 2007, 28(12):1795-1809.
[19] Ray S, Britschgi M, Herbert C, et al. Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins. Nat Med, 2007, 13(11):1359-1362.
[20] Goh KI, Cusick ME, Valle D, et al. The human disease network. Proc Natl Acad Sci USA , 2007, 104(21): 8685-8690.
[21] Lage K, Karlberg EO, Storling ZM, et al., A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol , 2007, 25(3): 309-316.
[22] Jiang X, Liu B, Jiang J, et al. Modularity in the genetic disease-phenotype network, FEBS Letters, 2008, 582(17): 2549-2554.
[23] Liu B, Jiang T, Ma S, et al. Exploring candidate genes for human brain diseases from a brain-specific gene network. Biochem Biophys Res Commun , 2006, 349(4):1308-1314.
[24] Myers AJ, Gibbs JR, Webster JA, et al. A survery of genetic human cortical gene expression. Nat Genet , 2007, 39(12): 1494-1499.
[25] Li J, and Burmeister M. Genetical genomics: combining genetics with gene expression analysis. Hum Mol Genet, 2005, 14: R163-R169.
[26] Mootha VK, Lepage P, Miller K, et al. Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics. Proc Natl Acad Sci USA , 2003, 100: 605-610.
[27] Bystrykh L, Weersing E, Dontje B, et al. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. Nat Genet , 2005, 37: 225-232.
[28] Hubner N, Wallace CA, Zimdahl H, et al. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat Genet , 2005, 37: 243-253.

收稿时间：2009年5月7日

作者信息

刘冰

博士，中国科学院自动化研究所模式识别国家重点实验室助理研究员。
研究方向为生物信息学、计算系统生物学和影像基因组学。