遗传 HEREDITAS (Beijing) 2011 年 8 月, 33(8): 809－819
ISSN 0253-9772 www.chinagene.cn

综 述

DOI: 10.3724/SP.J.1005.2011.00809

肿瘤相关生物学通路的发现和建模

郭昊, 朱云平, 李栋, 贺福初

军事医学科学院放射与辐射医学研究所, 蛋白质组学国家重点实验室, 北京蛋白质组研究中心, 北京 100850

摘要: 肿瘤是一种严重影响人类健康和生命的复杂疾病。某些生物学通路在肿瘤的发生、发展和转移的过程中发挥了关键作用, 如何发现和研究肿瘤相关通路是人们面临的一大挑战。随着以基因芯片数据为代表的海量实验数据的产出, 很多研究小组提出了一系列算法和模型通过整合和分析实验数据, 鉴定和模拟肿瘤相关的生物学通路, 发现了很多重要的生物学结论。文章对这些研究工作进行了综述, 给出了一些常用的算法、软件和数据库资源, 并讨论了该领域存在的问题和以后的发展方向。

关键词: 肿瘤; 基因芯片; 基因富集分析; 生物通路建模

Identification, modeling and simulation of key pathways underlying certain cancers

GUO Hao , ZHU Yun-Ping, LI Dong, HE Fu-Chu

State Key Laboratory of Proteomics, *Beijing Proteome Research Center*, *Beijing Institute of Radiation Medicine*, *Beijing* 100850, *China*

Abstract: Cancer is a complex disease which greatly affects the human health. It has been widely reported that certain biologcial pathways play important roles in the process of tumorigenesis, tumor progression and metastasis. Identification and simulation of these pathways can help to understand the underlying mechanisms. With the deposition of huge amount of microarray data, many groups have developed a series of algorithms and models to analyze the microarray datasets for the identification and simulation of tumor related pathways. In this paper, firstly we review the recent development of these algorithms and models; then list the related software and data sources; and finally discuss the existence problems and perspectives in this field.

Keywords: cancer; microarray; gene set enrichment analysis; biological pathway modeling

肿瘤是一类严重影响人类健康和生命的复杂疾病, 通常由细胞内多种生物学通路的扰乱所导致。很多生物学通路同肿瘤的发生发展过程有紧密的联系, 例如几乎所有肿瘤细胞的关键增殖和存活通路都会存在异常, 某些特定肿瘤细胞还拥有特异的异常通路。因此, 寻找和模拟特异影响肿瘤发生发展的生物通路和重要分子, 深入理解关键通路机制的异同, 对肿瘤研究具有重要的理论指导意义。

多年来, 肿瘤的遗传学研究对疾病的临床诊治起的作用很小。但近年来, 随着基因组学的发展, 生物信息学开始在肿瘤研究中发挥越来越重要的作用, 例如各种组学数据的整合、比对分析、模式抽取、

收稿日期: 2011–04–15; 修回日期: 2011–05–08
基金项目: 国家重点基础研究发展计划(973)(编号：2011CB910202)和国家自然科学基金项目(编号：30800200)资助
作者简介: 郭昊, 博士研究生, 研究方向：疾病相关生物学通路的鉴定和建模。Tel: 010-80727777-1125; E-mail: haosmail@gmail.com
通讯作者: 贺福初, 研究员, 中国科学院院士, 研究方向：蛋白质组学。Tel: 010-66931246; E-mail: hefc@nic.bmi.ac.cn
李栋, 副研究员, 研究方向：生物学网络研究。 Tel: 010-80705999; E-mail:lidong.bprc@foxmail.com

建模仿真等。这些生物信息学的分析方法能够通过鉴定肿瘤定量数据鉴定和预测肿瘤相关基因，寻找和发现肿瘤相关生物学网络，提供潜在的肿瘤诊断分子和通路。从而为人们研究肿瘤的发生、发展和转移提供新的理论依据和研究方向。

上述领域中的很多方面对于肿瘤临床研究者来说都不是很熟悉，特别是利用组学数据鉴定肿瘤相关生物学通路以及相关信号通路的建模仿真。为了更好地研究肿瘤发生发展机制，并提供相关参考，本文从聚类分析、数据降维分析、功能富集分析和网络发现 4 个方面介绍了目前主流的肿瘤相关通路鉴定方法，概括介绍了目前几种肿瘤相关生物学通路的建模和仿真研究，列举了一些常用的肿瘤相关基因芯片数据库和文献挖掘方法，并对这一领域存在的问题和前景做了简单的讨论。

1 鉴定肿瘤相关的生物学通路

当前，基于组学数据尤其是基因芯片数据研究肿瘤的疾病机制和预后已成为系统生物学领域一个重要的研究内容。人们发展了很多利用芯片数据信息鉴定肿瘤基因和通路的算法。在这里我们将简要介绍这些算法的基本原理和应用情况。

1.1 聚类分析法

聚类[1]就是将具有不同特征的数据划分到不同组或簇的过程。同组数据之间具有较高的相似性。对基因表达数据进行聚类分析的目的就是通过调整基因芯片数据矩阵的行、列，从而优化矩阵排列并揭示一定的生物学意义。从系统生物学的度量角度来讲，这种调整是通过分组归类实现的，例如将具有相同行为的信号蛋白归为一类。聚类方法最终抽取信息的能力取决于使用者对实验数据组织结构的先验生物学理解。

通过使用聚类方法，可以鉴定出肿瘤细胞中某些基因或者蛋白组成的特殊功能模块，提示潜在的新的肿瘤相关基因或者生物通路。

聚类的第一步就是定义相似性(Similarity)。而数据间的相似性取决于数据是如何被组合到一起的。通常情况下，距离的定义远比聚类使用的算法更为重要。对于基因芯片数据来说，通过基于向量分析方法比较容易解释聚类方法的基本原理(如图 1 所示)。在这种方法中，两组数据的相似性被定义为两个向量间的距离(Distance)[2]。通常情况下，对行向量(样本)聚类主要利用欧几里德距离(Euclidean distance)，而对列向量(基因)进行聚类主要使用皮尔森距离(Pearson distance)。

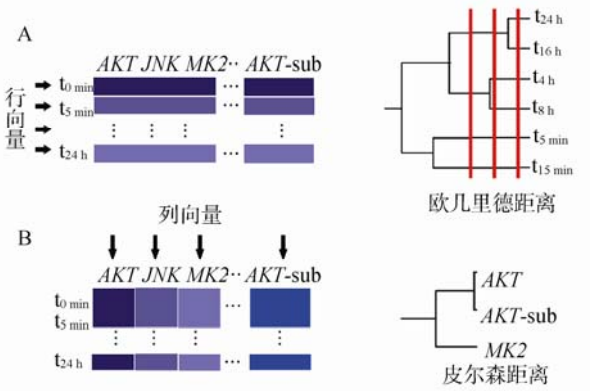

图 1 利用不同的距离尺度来进行行向量和列向量的聚类[3]

欧几里德距离又称几何距离。对于样本进行聚类的最简单方法，就是将几何距离上离得最近的两个样本行向量分到同一个簇里面。n 维行向量 A 和 B 之间的欧几里德距离由(1)式定义:

$$Euclidean\ distance = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \tag{1}$$

而对列向量，即基因表达向量，进行聚类的时候，由于必须考虑基因之间的协同变化，因此就不能再使用欧几里德距离了。对列向量聚类必须考虑到两个列变量之间的协方差(Covariance, cov)。通常用皮尔森距离来描述列向量之间的相似性。列向量 A 和 B 之间的皮尔森距离定义如(2)式所示:

$$Pearson\ distance = 1 - Pearson\ coefficient = \frac{\mathrm{cov}(A,B)}{\sqrt{\mathrm{cov}(A,A)\cdot\mathrm{cov}(B,B)}} \tag{2}$$

其中皮尔森相关系数的范围是–1 到+1，–1 代表负相关，+1 代表正相关。通常使用皮尔森相关系数可以衡量两基因表达水平的相关程度。而皮尔森距离就定义为 1 减去皮尔森相关系数。

两种比较常见的聚类方法是分裂聚类和凝结聚类[4]。分裂聚类通常是自上而下，首先将整个数据集分成两个簇，然后不停地将两个簇细分成更多的子簇，直到每一个子簇仅包含一个向量。该聚类计算量通常比较大，也很少用。凝结聚类通常是自下而

上，从每个向量出发，将相似的向量分配到同一个簇中，直到每一个向量都被包含在簇中，从而形成一个大的分层树状结构。凝结聚类比较常用，其优势在于算法简单，且分类结果比较容易可视化。通过使用凝结聚类可以比较容易的发现基因的一致表达模式。

对于常用的凝结聚类算法—— 分级聚类[5]来讲，首先将两个具有最小距离的向量归到一个簇里面，然后寻找剩余向量中和该簇距离最近的那个向量，将其和之前的那个簇重新归入一个新簇里面，以此不断往下类推。可以看出在图 1 A 中越往左移动，簇的分支越少，越往右移动，簇的分支就越多。如果对簇的个数 k 有一定要求的话，那么分级聚类就变成了 K 均值聚类[6]。其他聚类方法还包括自组织映射神经网络(Self-organization mapping net, SOM)[7]、模糊 C 均值聚类[8]、谱聚类[9]、双向聚类[10]等等。大量的多种多样的聚类方法在基因芯片的处理中的应用相对还是比较广泛的。研究者根据不同的生物学数据和研究对象，使用不同的聚类方法。

聚类分析在肿瘤研究方面应用较广，Segal 等[11]分析了 22 种肿瘤细胞的 1 975 套已发表的基因芯片数据。通过自下而上的分级聚类分析，作者提取了在上述基因表达数据中的一些能够一致执行某种功能的模块，并用这些模块的激活和未激活来区分基因表达谱。作者发现其中有一些模块是特殊的肿瘤类型所特有的，这些模块可能成为肿瘤细胞无限制增殖的基础。例如在急性淋巴细胞白血病中，生长抑制模块就是被特异性地抑制的。其他的一些模块在一些不同的临床条件下是不同的肿瘤细胞所共同拥有的。例如，成骨细胞模块在很多类型肿瘤细胞中普遍存在，并且包含生长因子及其受体，这就暗示了基础肿瘤增殖和骨转移之间存在一种特殊的关联机制。Koumakpayi 等[12]分析了 63 个前列腺癌症患者的组织基因芯片数据，利用多种统计学方法来进行数据分析，最后按照复发风险，利用非监督分级聚类分析方法来对肿瘤患者进行分组。聚类结果显示，前列腺癌患者按照生化复发性(Biochemical recurrence)可以被分成 5 组。数据分析和聚类的结果也从一方面验证了 ErbB 在 NF-κB 通路激活中的作用，提供了证据证明 ErbB/PI3K/Akt/NF-κB 信号通路在前列腺肿瘤发生发展中扮演着重要角色。

1.2　数据降维方法

从数学方法上来讲，聚类的方法是一种特殊的数据降维方法。我们在该部分介绍的数据降维方法，是一种不同于聚类方法，主要基于矩阵压缩的数学模型，包括主成分分析(Principle component analysis, PCA)[13]、独立成分分析(Independent component analysis, ICA)[14]、奇异值分解(Singular value decomposition, SVD)[15]和偏最小二乘法(Partial least square, PLS)[16]等算法。这种降维方法可以从多维空间中解析出主要的影响因素，简化复杂的数据结构。通常我们可以从基因芯片的数据矩阵中抽取一系列主成分或者元基因(Metagene)，将高维数据矩阵投影到较低维度的空间里。这些主成分或者元基因都是原有数据向量的线性组合，能够近似地反映原始矩阵的特征，起到了压缩原始矩阵的作用。实验数据之间的独立性会影响到数据降维算法分析基因表达数据结果的准确性。使用降维分析方法分析肿瘤芯片数据，可以降低数据分析的复杂度，抽取原始芯片数据的主要特征，发现潜在的新的肿瘤相关信号通路，并且可以用得到的相关通路区分不同的肿瘤表型，为肿瘤诊断提供候选基因和通路。

Bild 等[17]利用人类基础乳腺上皮细胞培养基来培养一系列通路特征。用重组腺病毒的方法在其他休眠细胞中来传递不同的致瘤行为，使得后续发生的独立事件都可以被定义为某条通路的激活或者下调。通过使用奇异值分解算法确定一系列跟细胞表型高度相关的元基因。并用这些挑选出来的元基因来描述每个通路特征的表达模式，然后作者利用回归模型得到了肿瘤细胞采样中通路下调的相对概率。这种基于通路特征的对几条通路的预测可以鉴定可以用来区分特殊肿瘤细胞及其表型的通路下调的模式。缺点就是作者挑选的通路数目有限，仅仅采用了五条通路信息。因此在用来寻找与表型相关的新的通路时，该方法就显得无能为力了。

Janes 等[16]使用偏最小二乘法(Partial least squares, PLS)模型预测了在 19 种时间依赖信号谱中测量到的 12 种凋亡响应。这种特殊的模型可以通过计算信息量最大的信号组合来预测细胞功能，并且可以有效地捕捉细胞因子诱导的凋亡响应。而这种响应在不同类型的细胞中是具有差别的。这暗示了不同细

胞类型可能通过一个公共的网络将信号转化为细胞内响应。上述两种方法采取的都是线性的降维方法, 而 Ivakhno 等[18]引用了非监督非线性降维方法、等距映射(Isomap), 来寻找在相同刺激条件下两种细胞信号网的输出响应。Isomap 可以找到与不同细胞因子处理相对应的簇。并且和 PCA、PLS、PLS-DA 作对比后发现, 这些方法都不能找到有生物学意义的簇, 而 Isomap 能鉴定更多生物功能连贯的簇, 并且在前两个主成分能够抓住更多信息。在系统的层面上理解细胞决策过程, 通过对细胞内信号分子空间的非线性降维, 找到不同刺激条件相对应的簇。该方法可以被应用于凋亡强度预测模型的监督学习中。扩展的 Isomap 方法可以用于信号网络的可视化, 预测和描述性建模。但是当更多的信号网络需要分析的时候, Isomap 的准确性就降低了, 这表示当信号网络变大时, 信号和响应之间的映射函数就会变的非线性。

1.3 功能富集分析的统计检验方法

高通量的基因组学实验往往产生了很多令人感兴趣的基因, 比如表达水平显著改变的基因。解释这些基因背后蕴含的生物学意义成为生物信息学的主要任务。很多研究小组基于各种生物知识数据库如基因本体数据库(Gene Ontology, GO)、KEGG (Kyoto Encyclopedia of Genes and Genomes)通路数据库等, 利用不同的统计分析策略, 系统分析了与这些基因所富集的生物过程及信号通路。

常用的基因富集分析可以分为 3 种[19]: 单基因富集分析(Singular enrichment analysis, SEA)、基因集富集分析(Gene set enrichment analysis, GSEA)和模块富集方法(Modular enrichment analysis)。

单基因富集分析是最常用的传统富集分析策略。研究人员通过单基因统计分析, 首先得到一系列实验组与对照组相比表达水平具有显著差异的基因列表, 然后逐一检验功能注释条目在这些基因中的富集程度, 并给出显著富集的 p 值。有很多统计分析方法可以用来检验功能富集的显著性, 例如卡方分析[20]、Fisher 精确检验[21]、二项概率分布以及超几何分布[22]等。单基因分析在抽取海量芯片数据背后的生物学意义方面显得非常有效。例如 Zeeberg 等[23]开发的软件 GoMiner, 可以方便的对基因芯片数据中差异表达的基因进行 GO 功能分析。GoMiner 首先将一组基因功能注释映射到基因本体树(GO tree)上, 然后在 GO tree 上标记基因芯片中上调和下调的基因, 通过统计检验来对这些基因进行功能富集分析。其中, GO tree 是一种通过分级控制的基因功能词汇表, 各功能注释条目来源于 GO 数据库。其他类似的分析工具还包括 Onto-Express[24]、DAVID[25]、GeneXPress[11]等。这种富集分析的缺点是找到的功能注释条目数目庞大, 不便于进行生物功能和通路分析。

基因集富集分析吸取了单基因富集分析的优点, 但是采用了不同的富集显著性分析策略。该分析不用预先挑选差异表达的基因, 而是使用全部的基因表达信息。应用比较广泛的是 Subramanian 等[26]提出的成套基因集通路鉴定方法—— 基因组富集分析(Gene set enrichment analysis, GSEA)。该方法首先利用先验的生物学知识, 例如一些已发表的生物通路信息或者是 GO 功能条目, 确定一套基因组(Gene set)。然后通过统计计算赋予每一个基因组一个富集打分(Enrichment score, ES), 进而检测不同分组(例如肿瘤和正常)的基因组 ES 的差异显著性水平, 最后调整多重假设检验估计的显著性水平, 同时通过控制假阳性率来得到差异表达的通路列表。通过这一系列步骤最终确定基因组中的成员基因的表达水平是随机分布还是有一定的分布趋势的。具体应用到肿瘤数据分析的时候, 通常的单基因分析方法在两个独立的肿瘤基因表达数据集中分别找到的通路很少有重复的, 而 GSEA 则在两个数据集中鉴定出很多共有的生物通路。通过利用该方法对糖尿病数据集分析发现氧化磷酸化通路的成员基因在糖尿病患者的肌肉中一致下调[27]。该分析方法常用到的统计分析策略包括 Kolmogorov-Smirnov-like 统计分析方法[26]、t 检验和 Z 分值[20]等。这种功能富集分析策略有两种优点: 减少差异基因挑选过程对富集分析的影响; 使用了芯片实验的全部信息。但该方法也有一定的限制, 它忽略了基因表达水平之间的相关性, 会过高估计显著性水平, 进而导致假阳性。但尽管如此, 这种方法仍然不失为提出假设的一种有利工具。除了 GSEA 分析之外, 相似的分析策略还包括 ErmineJ[28]、ADGO[29]等。

模块富集分析不仅继承了单基因富集分析的优

点, 还集成了一些基于功能注释条目间关系的网络发现算法。一些最近发布的分析软件, 如 Ontologizer [30]、GENECODIS[31]等, 通过在富集计算过程中考虑 GO 条目之间的相互关系, 提高了功能富集的敏感性和特异性。该分析策略的主要优点在于, 研究人员可以考虑功能注释之间的关系, 揭示那些彼此交叉的功能注释条目背后蕴藏的独特生物学含义。该分析的限制主要在于孤立基因或者注释条目(与邻居注释条目或者基因)可能会被该分析忽略。另外单基因的缺点它也具备, 即挑选差异表达基因的过程能影响最终的分析结果。

常用的基因表达数据分析软件包及平台的功能和下载地址如表 1 所示。

1.4　基于生物学网络的方法

近年来, 组学数据的爆发式增长以及数据可信度的提高, 大大促进了生物学网络的相关研究。大量基于组学实验得到的基因表达数据和蛋白质相互数据, 以及代谢组数据构成了相互关联、相互影响的复杂的生物学网络。基于这些复杂的生物学网络, 目前有很多研究小组提出了一系列算法和模型, 鉴定肿瘤关联通路和致病基因以及蛋白。

Bernardo 等[40]提出了一种基于基因芯片数据的网络建模方法, 网络鉴定模式响应模型(Mode-of-action

表 1　近年来用于基因芯片数据分析鉴定肿瘤相关通路的软件包和分析平台

软件	功能	网址
ADGO[29]	检测来源于不同 GO 功能条目的基因集之间的基因差异表达程度	http://array.kobic.re.kr/ADGO
ArrayTrack[32]	储存多种格式的芯片数据, 通过统计检验发现表达模式, 提供数据标准化的各种方法, 并提供基因, 蛋白和通路的功能信息	http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/index.htm
caGEDA[33]	提供了很多芯片数据分析的统计学工具, 包括预处理, 特征提取, 病人预测模型发展等	http://bioinformatics.upmc.edu/GE2/GEDA.html
DAVID[25]	基因列表的功能注释, 可视化和整合发现	http://david.abcc.ncifcrf.gov/
ErmineJ[28]	芯片表达水平的基因集分析, 鉴定差异基因所在的生物通路	http://bioinformatics.ubc.ca/ermineJ/
Genecluster[34]	采用 Java 语言, 整合多种聚类方法和统计学检验方法	http://www.broad.mit.edu/cancer/software/genecluster2/gc2.html
GENECODIS[31]	基于网格计算, 鉴定与基因集特意相关的功能注释	http://genecodis.dacya.ucm.es/
GeneXPress[11]	鉴定基因集中显著表达的序列和这些显著序列富集的功能注释	http://GeneXPress.stanford.edu/.
GenMAPP[35]	浏览和分析生物通路的表达情况, 能够通过输入芯片数据, 显示通路上各个基因的激活情况	http://www.genmapp.org
GoMiner[23]	GO 功能聚类分析, 可以分析多重芯片数据集, 方便基因表达的时序研究	http://discover.nci.nih.gov/gominer/
GoSurfer[36]	GO 基因功能可视化; 通过统计检验搜索 GO 注释条目	http://www.biostat.harvard.edu/complab/gosurfer/
GSEA-P[37]	采用基因组富集分析算法, 检测基因表达模式, 拥有跨平台图形用户接口, 用 R 和 JAVA 编写方便整合	http://www.broad.mit.edu/GSEA
NUDGE[38]	采用 R 语言, 使用简单单变量正态统一混合模型结合新标准化方法来寻找 cDNA 芯片中的不同表达的基因	http://www.bioconductor.org/
Onto-Express[24]	鉴定差异基因富集的 GO 功能注释	http://vortex.cs.wayne.edu/projects.htm
Ontologizer [30]	利用 Fisher 精确检验和 GO 功能的拓扑网络算法鉴定差异表达的基因集	http://compbio.charite.de/index.php/ontologizer2.html
PLAGE[39]	PLAGE 利用奇异值分解的方法鉴定预选通路基因的不同表达水平的显著性.	http://dulci.biostat.duke.edu/pathways/

by network identification, MNI), 该方法是反转工程学建模方法的一种。首先将对实验得到的某种感兴趣的表型组织的芯片数据进行预处理, 然后将全基因表达谱数据作为训练集, 利用他们设计的模型来分析药物处理过的细胞表达谱数据, 从而鉴定与给药响应相关的基因和通路。该模型是基于有向图模型的, 网络中的边表示一个基因的产物会影响到另一个基因的表达水平。大量基因相互之间的复杂影响关系就形成了一个有权重的表达网络, 其中权重表示影响的强度。作者利用该模型分析了经过药物或者基因敲除等处理过的 515 个酵母全基因表达谱, 成功鉴定出了已知的药物靶标和关联通路。Ergun 等[41]利用 Bernardo 等开发的这种 MNI 的网络建模方法分析了没有复发和发生转移的两组前列腺癌基因芯片数据, 成功鉴定出了雄性激素受体(Androgen receptor, AR)基因以及雄性激素受体通路是前列腺癌转移的生物标志物。

除了基于基因芯片数据的基因网络分析之外, 近年来基于大规模蛋白相互作用数据的研究也方兴未艾。除了直接利用蛋白之间的相互作用来直接预测疾病相关基因之外[42], 还可以通过研究蛋白在整个蛋白相互作用网络中的位置和拓扑性质从而发现疾病相关基因[43]。Sam 等[44]还发展了一种算法, 鉴定不同疾病下的蛋白相互作用网络, 进而用来比较不同疾病下的蛋白相互作用子网的重叠部分, 从而提示了不同疾病在分子层次水平的相关性。最近, Raj 等[45]整合了基因芯片数据和蛋白相互作用数据以及通路信息, 建立了一个混合的生物网络模型, 鉴定出了一系列肿瘤相关基因, 以及潜在的跟肿瘤表型相关的信号子网。作者利用该方法成功鉴定了肿瘤发展过程中整联蛋白 $\alpha 6\beta 4$—— 肿瘤抗原的转录靶标。

2 信号通路的建模

在发现了与肿瘤相关的生物学通路后, 为了进一步研究它们的行为和特征, 人们发展了一系列计算模型, 利用各种组学数据来描述特定生物细胞中特定通路的反应动力学[46]。

最常用的反应动力学建模方法是微分方程[47, 48]。这种方法可以利用基因或者蛋白质表达数据模拟基因调控和信号传导的过程。通过使用反应动力学方程, 对基因调控过程或者信号的传导过程进行建模, 鉴定目标基因的上游调控基因或者模拟信号蛋白的合成和降解。以基因调控过程为例, 微分方程的基本过程是使用一组混合调控子来估计目标基因的调控模式, 进而拟和一个线性模型。其一般形式如(3)式所示:

$$X_i(t) = G_i(t) - \lambda_i X_i(t) + \xi_i(t) \quad (3)$$

其中, $X_i(t)$代表实验中第 i 个基因在时刻 t 的表达水平; $G_i(t)$和λ_i 表示每个基因 i 的转录速率和自身降解速率; 最后一项$\xi_i(t)$表示数据的未知噪声和模型的残差。动力学方程(1)式定义了 mRNA 水平改变的速率为转率速率 $G_i(t)$控制合成的速率和 mRNA 分子自然降解速率$\lambda_i X_i(t)$的差。如果用该模型来描述信号传导过程中的信号蛋白的合成和降解过程, 则 $X_i(t)$代表了信号响应的大小, 如信号蛋白的浓度; $G_i(t)$代表了信号强度, 如 mRNA 浓度。

利用微分方程模型可以简单的估算基因间的调控关系, 推测出基因的调控能力和调控激活的延迟时间, 推测潜在的基因调控通路; 也可以定量的描述信号响应过程中信号蛋白浓度变化情况, 推测信号传递的速度和最终的响应输出。目前已有很多研究小组通过使用各种动力学模型对一些特定的通路进行了分析和建模。

2.1 JAK-STAT 通路

JAK-STAT 通路在调控人细胞突变响应和动态平衡方面起着重要的作用。它是由自分泌蛋白(细胞因子)激活的。该通路主要包括 3 个组分: 酪氨酸激酶 JAK 受体(4 种)、酪氨酸激酶 JAK 和转录因子 STAT(7 种)。Swamaye 等[49]提出了一种模拟 JAK-STAT 通路的数学模型, 描述了每一步信号传递步骤的反应过程。由于 STAT5 分子必须呆在细胞核内一段时间, 微分方程中必须用一个固定的延迟来模拟。该模型揭示了 STAT5 分子在细胞内的循环。而该循环是 JAK-STAT 通路活化循环的重要环节, 并且对解释实验数据很重要。STAT3 和 STAT5 可以刺激细胞增殖和阻止细胞凋亡, 从而促进肿瘤的产生。针对 JAK-STAT 通路中转录因子 STAT 家族的研究可以提供潜在的治疗干预策略。

2.2 EGF 受体信号转导通路

表皮生长因子(Epidermal growth factor, EGF)受

体属于酪氨酸激酶受体家族, 它调控细胞生长、存活、增值和分化。对 EGF 通路建模可以量化通路的瞬时信号, 并预测不同外界刺激对 EGF 通路的影响和信号变化情况。Schoeberl 等[46]发展了一个 EGF 受体激活 MAPK 级联通路的全面模型。该模型包含了 94 个组分, 可以研究在 EGF 受体内移的情况下, 通路信号的特征以及 ERK 的磷酸化过程。该模型可以从定量、动力学、拓扑等多角度描述 EGF 受体信号通路, 显示 EGF 及其受体在细胞表面绑定, 并引发下游蛋白在信号级联放大过程中的激活等一系列过程。该模型得到了 EGF 受体信号通路的一个重要特征: 当配体浓度在 100 倍变化范围内时, EGF 诱导的响应能够保持稳定; 决定信号传递效率的关键参数是受体激活的初始速度。同时该模型揭示了 EGF 对 ERK1/2 的磷酸化和调控 *c-fos* 基因的表达, 并且得到和实验结果一致的结论。模型发现受体内移在强刺激过程中贡献比较小, 但是在弱刺激全局响应中通过保留信号强度而产生较强的影响。

2.3　Wnt 通路

Wnt 蛋白是许多细胞内信号转导的中介物, 因此 Wnt 家族蛋白的信号通路研究也至关重要。Lee 等[50]通过构建微分方程来研究 Wnt/β-catenin 信号通路, 每一步反应的动力学机制都尽可能地通过恒定速率或者质量反应动力学来描述, 同时考虑了 DVL、TCF、GSK3β 以及 APC 的保守关系。模型揭示了两种支架蛋白 axin 和 APC 促进降解复合体形成的不同方式, 并且解释了 axin 降解在放大和增强 Wnt 信号中的重要性。模型同时揭示了阻止 β-catenin 在低 APC 浓度时聚化的最关键过程是 APC 可以调控 axin 的降解。该模型的分析结果可以定量地解释 Wnt 信号通路中一些潜在的抑癌效应(APC、GSK3β 和 axin)和致癌效应(PP2A、TCF、Dsh 和 β-catenin)。

除了以上通路, 针对 TGF-β/Smads 信号通路[51]、干扰素信号传导通路[52]、细胞毒素相关通路[53]、NF-κB [54]等信号通路也发展出了一系列模型来描述通路响应和预测新的信号蛋白。

这些信号通路的建模, 能够将实验得到的知识整合成连贯的图片, 帮助提出、测试、支持或者推翻潜在的生物机制假说。由于目前设计的揭示复杂生物信号网络机制的实验技术越来越定量化和多元化, 利用各种各样的计算模型可以尽可能多地整合已知定量生物学知识和抽取基因组、蛋白质组数据中蕴含的信息。

当然, 目前基于反应动力学进行通路建模也存在一定的问题。一是可靠的定量模型难以得到; 二是初始模型是基于预先搜集好的数据的, 如浓度、动力学参数、流量测量、显微图像等, 这些数据可能来自不同的来源(文献数据库、自己测量); 三是需要用新的实验不断地改进原始模型, 模型发展和实验设计就需要协同发展。模型和对象的研究可能会被反复定义, 这样就存在一些问题, 比如说模型输出结果和生物学观察到的结果有可能存在差异或者丢失一些相关信息。解决办法就是通过产生新的实验数据, 然后不断的改进原始的模型, 提出更多的参数或者提出更准确的模型和新的实验。或者将多种通路整合到一起形成通路网络来分析, 以获得更精确的模型。整合网络拥有单一通路所没有的特性, 例如延长的信号滞留时间, 反馈回路的激活, 生物学效应的阈值或者多信号的输出。或者可以提出一些通路建模的标准, 用来规范现有的通路模型以及开发新的模型。

3　肿瘤通路建模研究常用的数据库和文献挖掘工具

3.1　常用的数据资源

目前许多组织和机构在网上公布了基因芯片数据库, 并且提供下载和相关服务。这些数据库资源对于研究各种条件下的采样数据, 鉴定不同条件下鲁棒性比较强的模式或者区分不同的肿瘤类型, 或者通过研究这些基因表达数据中蕴含的模式来将一部分基因映射到肿瘤相关通路中来, 都有着相当大的推动作用。目前网上大多数工具和文章分析方法用到的数据库如表 2 所示。

3.2　文本挖掘工具

通常情况下, 我们研究肿瘤相关的通路信息需要查阅很多科学文献。为此, 研究人员开发出了一系列文献挖掘工具。利用这些工具可以通过对已发表文献的摘要或者全文进行自动文献分析, 提取一些跟我们感兴趣的生物学过程相关联的已发表的蛋白、基因或者通路信息。

表 2 常用的基因芯片数据库资源

数据库	来源	网址
GEO	NCBI, NIH	http://www.ncbi.nlm.nih.gov/projects/ geo/
ArrayExpress	EMBL-EBI	http://www.ebi.ac.uk/arrayexpress/
Oncomine Cancer Microarray Database	Univ. of Michigan	http://www.oncomine.org
Standard Microarray Database (SMD)	Stanford Univ.	http://smd.stanford.edu/
The Gene Expression Database (GXD)	The Jackson Lab.	http://www.informatics.ja.org/mgihome/GXD/aboutGXD.shtml
Center for Genomic Research	Whitehead Institute	http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi
CIBEX	Nat. Inst. Genetics, Japan	http://cibex.nig.ac.jp
The Tumor Gene Family Databases	Baylor College of Medicine	http://condor.bcm.tmc.edu/ermb/tgdb/tgdf.html
Cancer Gene Data Curation Project	US National Cancer Institute	http://ncicb.nci.nih.gov/NCICB/projects/cgdcp
Cancer Gene Expression Database (CGED)	Osaka University School of Medicine	http://cged.hgc.jp/cgi-bin/input.cgi
Cancer Cell Map	Memorial Sloan-Kettering Cancer Center	http://cancer.cellmap.org/cellmap

在利用蛋白或者基因名称检索相关文献方面，Tanabe 等[55]开发的 AbGene 能够比较精确地通过名称条目和一系列规则设计，找到相关程度很高的文献(精确度 85.7%)。Chang 等[56]开发的 GAPSCORE 系统基于不同的算法通过对文献语句中词条的打分来判断是否与感兴趣的信息相关，也得到了比较好的结果(精确度 74%)。Chen 和 Friedman[57]提出了 MEDLEE 系统来识别文献中和表型信息相关的短语。该系统使用了自然语言技术来鉴定文献摘要中的表型语句，并且识别这些短语，包括在段落中分开的关键词。

上述几个都是从词条出发进行文献挖掘的工具，而近年来一些综合的文献挖掘工具不断的涌现出来，可以满足用户不同的需要。例如 Hoffmann 等[58]开发出来一种基于 web 的开放易用的文献挖掘系统 iHOP (Information Hyperlinked over Proteins)，该系统可以建立和保留一些文献库中的生物数据库的信息及其原始资源的链接。用户可以通过基因、蛋白名称或者蛋白相互作用的类型，来找到与这些基因、蛋白或者信号通路相关的已发表的文献。Novichkova 等[59]开发的 MedScan 系统将句法和语义词典模版整合到了一个通用的文本挖掘系统中来提取生物学和医学词条之间的关系。Becker 等[60]开发出来 PubMatrix 基于 PubMed 多重查询的结果来显示基因名称和对功能条目的对比。其他相类似的文献挖掘平台还有 TXTGATE[61]、BioRAT[62]、Textpresso[63]等，除了一些基本的信息挖掘功能之外，有的系统还提供了一些功能可以方便蛋白相互作用网络的重建，数据库和文字信息的整合，并且支持用户用公式来提出一些假说用户可以根据自己的需要以及感兴趣的生物学词条来选择不同的数据库。

4 结论与展望

由于基因表达数据和定量蛋白质组数据的海量产出，使得人们有了大量可以利用的数据来源。利用生物信息学的方法，分析整合湿实验环境下得到的基因或蛋白表达数据，发现新的基因表达模式和肿瘤相关的蛋白和通路成为生物信息学一个重要的研究分支。通过对一些肿瘤相关的基因芯片数据和定量蛋白质组数据进行分析，可以发现大量肿瘤相关的基因和相关通路信息。通过这方面的研究可以使人们对一些肿瘤细胞生理过程有重要影响的信号通路和分子有更加深刻的理解和认识。

本文介绍了一些当前应用到肿瘤相关通路研究中的一些主要的数学模型、芯片数据库以及几年来发展的一系列软件、工具和 web 平台。虽然这个领域目前的已经有了很大的发展，但是还是存在一些亟待解决的问题，例如：(1) 芯片数据质量方面，如何降低数据背景噪声的影响，如何利用有限的表达信息准确的推测整个生物过程，如何评估其准确性，如何解决不同数据标准化方法的影响，目前还有待进一步的研究。(2) 利用各种算法得到的蕴藏在基因芯片数据中的基因表达模式的生物学意义如何评估。不可否认，很多统计学方法包括聚类分析，降维方法等得到了很多具有统计学意义的模块或者是表

达模式，但是这些模块的生物学意义如何评估，反映真实生物过程的程度如何评估，还没有一个统一的方法。(3) 如何考虑基因之间的相关性，并在此基础上合理的利用芯片数据；如何将各种通路数据库以及各种实验已知的共表达信息和芯片数据结合在一起挑选疾病相关的基因或者通路信息，目前都还没有成熟的策略。(4) 如何整合现有的一些分析工具和方法。目前针对各种数据源、大量不同平台、不同编程语言的工具包被开发出来针对性的研究一些问题。但是这些工具包研究的问题都比较具体化，没有一个成型的通用性跨平台的综合软件，各种软件之间相互独立，重复调用性不强。面对诸如此类的问题，亟待系统生物学家们引入开发新的更完备的模型来进一步研究肿瘤相关通路。

未来该领域的发展，有很多可行性很强的热点方向，比如数学模型和统计学验证更好的结合，芯片设计和数学模型相结合，芯片数据和蛋白网络拓扑结构的结合。这样得到的结果将更精确，更具有生物学合理性，也能够更好的描述一些生物学现象或者反应一定的生物过程和表征。其次就是整合组学层次上的数据，将基因组数据和蛋白质组数据结合起来，挖掘出更多疾病相关信息。该方向目前风头正劲，例如 Rhodes 等[64]通过使用一种基于朴素贝叶斯分类器概率分析的整合模型分析了相互作用组数据，蛋白结构域数据，基因组层次上的基因表达数据和功能注释数据，鉴定出了肿瘤细胞中激活的几个相互作用子网。最后就是利用现有的调控通路或者信号通路拓扑结构信息，挖掘基因之间的相关性信息，发展出更符合生物学意义的模型和工具。在这些模型鉴定出相关基因和通路以后，可以开发出一些利用肿瘤相关通路和基因作为特征进行肿瘤分型和给药设计的生物信息学工具，进一步改进人类对肿瘤的发现、分类、监控以及治疗的方法。

参考文献**(References)**:

[1] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, 1998, 95(25): 14863–14868.

[2] D'Haeseleer P. How does gene expression clustering work? *Nat Biotechnol*, 2005, 23(12): 1499–1501.

[3] Janes KA, Yaffe MB. Data-driven modelling of signal-transduction networks. *Nat Rev Mol Cell Biol*, 2006, 7(11): 820–828.

[4] Sherlock G. Analysis of large-scale gene expression data. *Curr Opin Immunol*, 2000, 12(2): 201–205.

[5] Tjaden B. An approach for clustering gene expression data with error information. *BMC Bioinformatics*, 2006, 7: 17.

[6] Wilkin GA, Huang XZ. A practical comparison of two K-Means clustering algorithms. *BMC Bioinformatics*, 2008, 9(Suppl. 6): S19.

[7] Blackhall FH, Wigle DA, Jurisica I, Pintilie M, Liu N, Darling G, Johnston MR, Keshavjee S, Waddell T, Winton T, Shepherd FA, Tsao MS. Validating the prognostic value of marker genes derived from a non-small cell lung cancer microarray study. *Lung Cancer*, 2004, 46(2): 197–204.

[8] Dembélé D, Kastner P. Fuzzy C-means method for clustering microarray data. *Bioinformatics*, 2003, 19(8): 973–980.

[9] Kluger Y, Basri R, Chang JT, Gerstein M. Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research*, 2003, 13(4): 703–716.

[10] Lapointe J, Li CD, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U, Ekman P, DeMarzo AM, Tibshirani R, Botstein D, Brown PO, Brooks JD, Pollack JR. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci USA*, 2004, 101(3): 811–816.

[11] Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 2003, 34(2): 166–176.

[12] Koumakpayi IH, Le Page C, Mes-Masson AM, Saad F. Hierarchical clustering of immunohistochemical analysis of the activated ErbB/PI3K/Akt/NF-κB signalling pathway and prognostic significance in prostate cancer. *Br J Cancer*, 2010, 102(7): 1163–1173.

[13] Ma SG, Kosorok MR. Identification of differential gene pathways with principal component analysis. *Bioinformatics*, 2009, 25(7): 882–889.

[14] Huang DS, Zheng CH. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics*, 2006, 22(15): 1855–1862.

[15] Sandberg R, Ernberg I. Assessment of tumor characteristic gene expression in cell lines using a tissue similarity index (TSI). *Proc Natl Acad Sci USA*, 2005, 102(6): 2052–2057.

[16] Janes KA, Albeck JG, Gaudet S, Sorger PK, Lauffenburger

DA, Yaffe MB. A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science*, 2005, 310(5754): 1646–1653.

[17] Bild AH, Yao G, Chang JT, Wang QL, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, Olson JA Jr, Marks JR, Dressman HK, West M, Nevins JR. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, 2006, 439(7074): 353–357.

[18] Ivakhno S, Armstrong JD. Non-linear dimensionality reduction of signaling networks. *BMC Syst Biol*, 2007, 1: 27.

[19] Huang D W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucl Acids Res*, 2009, 37(1): 1–13.

[20] Curtis RK, Orešič M, Vidal-Puig A. Pathways to the analysis of microarray data. *Trends Biotechnol*, 2005, 23(8): 429–435.

[21] Drăghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA. Global functional profiling of gene expression. *Genomics*, 2003, 81(2): 98–104.

[22] Rivals I, Personnaz L, Taing L, Potier MC. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, 2007, 23(4): 401–407.

[23] Zeeberg BR, Feng WM, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 2003, 4(4): R28.

[24] Draghici S, Khatri P, Bhavsar P, Shah A, Krawetz SA, Tainsky MA. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res*, 2003, 31(13): 3775–3781.

[25] Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*, 2003, 4(5): P3.

[26] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, 2005, 102(43): 15545–15550.

[27] Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. PGC-1α- responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 2003, 34(3): 267–273.

[28] Lee HK, Braynen W, Keshav K, Pavlidis P. ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, 2005, 6: 269.

[29] Nam D, Kim SB, Kim SK, Yang SJ, Kim SY, Chu IS. ADGO: analysis of differentially expressed gene sets using composite GO annotation. *Bioinformatics*, 2006, 22(18): 2249–2253.

[30] Bauer S, Grossmann S, Vingron M, Robinson PN. Ontologizer 2.0-a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, 2008, 24(14): 1650–1651.

[31] Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol*, 2007, 8(1): R3.

[32] Tong W, Harris S, Cao X, Fang H, Shi L, Sun H, Fuscoe J, Harris A, Hong H, Xie Q, Perkins R, Casciano D. Development of public toxicogenomics software for microarray data management and analysis. *Mutat Res*, 2004, 549(1–2): 241–253.

[33] Patel S, Lyons-Weiler J. caGEDA: a web application for the integrated analysis of global gene expression patterns in cancer. *Appl Bioinformatics*, 2004, 3(1): 49–62.

[34] Reich M, Ohm K, Angelo M, Tamayo P, Mesirov JP. GeneCluster 2.0: an advanced toolset for bioarray analysis. *Bioinformatics*, 2004, 20(11): 1797–1798.

[35] Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet*, 2002, 31(1): 19–20.

[36] Zhong S, Storch KF, Lipan O, Kao MC, Weitz CJ, Wong WH. GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology(TM) space. *Appl Bioinformatics*, 2004, 3(4): 261–264.

[37] Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. *GSEA-P*: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics*, 2007, 23(23): 3251–3253.

[38] Dean N, Raftery AE. Normal uniform mixture differential gene expression detection for cDNA microarrays. *BMC Bioinformatics*, 2005, 6: 173.

[39] Tomfohr J, Lu J, Kepler TB. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 2005, 6: 225.

[40] di Bernardo D, Thompson MJ, Gardner TS, Chobot SE, Eastwood EL, Wojtovich AP, Elliott SJ, Schaus SE, Collins JJ. Chemogenomic profiling on a genome-wide scale using

reverse-engineered gene networks. *Nat Biotechnol*, 2005, 23(3): 377–383.

[41] Ergün A, Lawrence CA, Kohanski MA, Brennan TA, Collins JJ. A network biology approach to prostate cancer. *Mol Syst Biol*, 2007, 3: 82.

[42] Oti M, Snel B, Huynen MA, Brunner HG. Predicting disease genes using protein-protein interactions. *J Med Genet*, 2006, 43(8): 691–698.

[43] Xu JZ, Li YJ. Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics*, 2006, 22(22): 2800–2805.

[44] Sam L, Liu Y, Li J, Friedman C, Lussier YA. Discovery of protein interaction networks shared by diseases. *Pac Symp Biocomput*, 2007: 76–87.

[45] Loganantharaj R, Chung J. Integrating diverse information to gain more insight into microarray analysis. *J Biomed Biotechnol*, 2009, 2009: 648987.

[46] Schoeberl B, Eichler-Jonsson C, Gilles ED, Müller G. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat Biotechnol*, 2002, 20(4): 370–375.

[47] Chen HC, Lee HC, Lin TY, Li WH, Chen BS. Quantitative characterization of the transcriptional regulatory network in the yeast cell cycle. *Bioinformatics*, 2004, 20(12): 1914–1927.

[48] Chen KC, Wang TY, Tseng HH, Huang CYF, Kao CY. A stochastic differential equation model for quantifying transcriptional regulatory network in *Saccharomyces cerevisiae*. *Bioinformatics*, 2005, 21(12): 2883–2890.

[49] Swameye I, Müller TG, Timmer J, Sandra O, Klingmuller U. Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. *Proc Natl Acad Sci USA*, 2003, 100(3): 1028–1033.

[50] Lee E, Salic A, Krüger R, Heinrich R, Kirschner MW. The roles of APC and Axin derived from experimental and theoretical analysis of the Wnt pathway. *PLoS Biol*, 2003, 1(1): E10.

[51] 张云艳, 李雪, 隋丽华, 王琦, 李璞, 傅松滨. 卵巢癌中 TGF-β/Smads 传导通路的功能研究. 遗传学报, 2004, 31(8): 759–765.

[52] 崔建军, 田庚善, 田地, 曾争. 干扰素信号传导通路与其基因组多态性网络模型的建立. 遗传, 2008, 30(6): 788–794.

[53] Nagasaki M, Doi A, Matsuno H, Miyano S. A versatile petri net based architecture for modeling and simulation of complex biological processes. *Genome Inform*, 2004, 15(1): 180–197.

[54] Ihekwaba A, Broomhead D, Grimley R, Benson N, Kell D. Sensitivity analysis of parameters controlling oscillatory signalling in the NF-κB pathway: the roles of IKK and IkBa. *Syst Biol*, 2004, 1(1): 93.

[55] Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. *Bioinformatics*, 2002, 18(8): 1124–1132.

[56] Chang JT, Schütze H, Altman RB. GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics*, 2004, 20(2): 216–225.

[57] Chen L, Friedman C. Extracting phenotypic information from the literature via natural language processing. *Stud Health Technol Inform*, 2004, 107(Pt 2): 758–762.

[58] Hoffmann R, Krallinger M, Andres E, Tamames J, Blaschke C, Valencia A. Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci STKE*, 2005, 2005(283): pe21.

[59] Novichkova S, Egorov S, Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics*, 2003, 19(13): 1699–1706.

[60] Becker KG, Hosack DA, Dennis G Jr, Lempicki RA, Bright TJ, Cheadle C, Engel J. PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics*, 2003, 4: 61.

[61] Glenisson P, Coessens B, van Vooren S, Mathys J, Moreau Y, De Moor B. TXTGate: profiling gene groups with text-based information. *Genome Biol*, 2004, 5(6): R43.

[62] Corney DPA, Buxton BF, Langdon WB, Jones DT. BioRAT: extracting biological information from full-length papers. *Bioinformatics*, 2004, 20(17): 3206–3213.

[63] Müller HM, Rangarajan A, Teal TK, Sternberg PW. Textpresso for neuroscience: searching the full text of thousands of neuroscience research papers. *Neuroinformatics*, 2008, 6(3): 195–204.

[64] Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM. Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, 2005, 23(8): 951–959.