

· 生物信息学 ·

生物信息学跨学科研究

武妍¹ 胡德华^{2△}

(1 中南大学湘雅三医院 湖南 长沙 410013 2 中南大学湘雅医学院 湖南 长沙 410013)

摘要 目的 研究生物信息学起源、发展趋势,与其他学科相互交叉渗透关系的发展及强度。方法 利用美国《科学引文索引》(SCI)数据库 web of science,运用文献计量学方法对 8 种权威生物信息学期刊 2001 年至 2010 年于 2011 年 1 月 15 日之前上传至 web of science 的全部文献进行统计及分析。通过研究生物信息学相关论文的主题分类,被引情况及施引文献的分类,寻找其跨学科的趋势及相关研究领域的进展情况、主要内容。结果 生物信息学的相关文献数在 2001-2010 间逐年增加,在 2009-2010 年达到高峰。跨学科领域广泛,并以生物化学、分子生物学、计算生物学、微生物学、数学、统计学等学科为主要交叉学科。各交叉学科与生物信息学之间跨学科研究的文献数也呈逐年递增趋势。结论 生物信息学的跨学科范围广泛,发展迅速。

关键词 web of science 跨学科研究 文献计量学

中图分类号 Q-31,R-058 **文献标识码** A **文章编号** 1673-6273 (2012) 01-137-05

The Research on Bioinformatics Interdisciplinary

WU Yan¹, HU De-hua^{2△}

(1 The third Xiangya hospital of Central South University, Changsha 410013, China;

2 Xiangya medical school of Central South University, Changsha 410013, China)

ABSTRACT Objective: To study the origin, developing trends of bioinformatics, relationship with other disciplines, the intensity and development of interdisciplinary situation. **Methods:** In this paper, the United States "Science Citation Index" (SCI) database, web of science, and the method of bibliometrics are used to do a research about all literature from 8 authoritative bioinformatics journals from the year 2001 to 2010 with metrology statistics. Through researches about the classification related to the theme of the papers of the literature, and the classification of cited literature's discipline, the trends and the research areas of bioinformatics' inter-disciplinary development can be known. **Results:** Bioinformatics literature in the decade from 2001 to 2010 is increased year by year and reached a peak in 2009 and 2010 year. The interdisciplinary fields of bioinformatics become wider. Among all the interdisciplinary fields, biochemistry, molecular biology, computational biology, microbiology, mathematics, statistics and other disciplines are the main cross-disciplinary disciplines. **Conclusion:** The cross-disciplinary research is wide and develops quickly.

Key word: Web of science; Interdisciplinary research; Bibliometrics

Chinese Library Classification(CLC): Q-31, R-058 **Document code:** A

Article ID: 1673-6273(2012)01-137-05

引言

生物信息学 (Bioinformatics) 是基于分子生物学与多种学科交叉,以计算机为工具对生物相关信息进行储存、检索和分析的科学,是当今生命科学和自然科学的重大前沿领域之一。其发展依赖于生物学,计算科学等相关学科不断突破的同时,又为这些学科提供信息、材料及研究方法。随着科学研究逐步细化,开拓出更多新的领域并形成更多新兴学科,生物信息学在越来越多的学科中得到应用,因此衍生的跨学科研究也更加广泛。研究相关跨学科的学科性质以及与生物信息学间的交叉强度,能追溯生物信息学发展的路线,并为其发展趋势做出预

测。同时,在了解已有跨学科领域的基础上,挖掘新兴跨学科领域的发展动向,有利于启发相关研究,并能促进生物信息学与这些学科的发展。

1 材料和方法

本文利用国际权威的科技引文索引数据库 web of science, 统计分析了 2001-2010 年生物信息领域 8 种重要期刊 Bioinformatics, BMC Bioinformatics, Briefings IN Bioinformatics, Current Bioinformatics, International Journal of Data Mining and Bioinformatics, IEEE-ACM Transactions on Computational Biology and Bioinformatics, Journal OF Computational Biology, PLoS Computational Biology 于 2011 年 1 月 15 日之前上传至 web of science 的全部文献,共 13253 篇。从来源文献及被引文献的学科分类角度,揭示生物信息学跨学科的分布规律。

2 结果及分析

2.1 生物信息学研究的趋势分析

作者简介 武妍 (1988), 女, 大学本科, 电话: 13548622264,

E-mail: wy4192221@yahoo.com.cn

△通讯作者 胡德华 (1972), 男, 博士, 硕士生导师, 副教授, 主要从事生物信息学研究, 发表论文 50 多篇。

Email: hudehua2000@163.com

(收稿日期 2011-06-18 接受日期 2011-07-13)

将 13253 篇文献按其发表的年代进行分析,其结果见图 1。

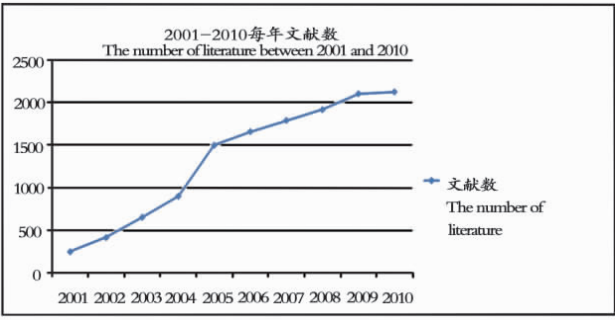


图 1 2001 年至 2010 年共 13253 篇文献的年限分布
Fig.1 The chronological distribution of the total 13253 literature between 2001 and 2010

从图 1 中可看出,2001 至 2010 年生物信息学的研究论文在数量上不断增加,尤其在 2004 及 2005 两年增长速度最快,并于 2010 年达到高峰。09 及 10 年发展速度逐渐趋于平稳,预示着生物信息学发展平台期即将到来。因此,寻找新突破对于生物信息学及相关跨学科的发展是至关重要的。

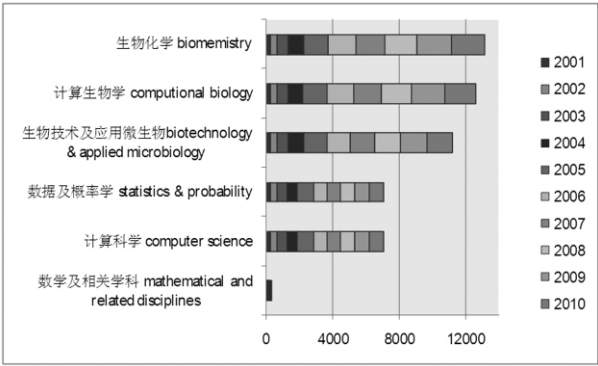


图 2 2001 年至 2010 年共 13253 篇文献的主题分布
Fig.2 The subjective distribution of the total 13253 literature between 2001 and 2010

2.2 来源文献的跨学科分布

将 13253 篇文献按主题分类进行分析,结果见图 2。由图 2 可看出 13253 篇文献的主题分类为生物化学、计算生物学、生物技术及应用微生物、数据及概率学、计算科学、数学及相关学科等 6 个主要学科。在这十年里,各跨学科的研究均有快速的发展。这一同步性表示生物信息学与这些交叉学科的发展关系密切。其中,计算科学、数据及概率学的文献数在 2005 年达到高峰,之后一直保持持平。数学及其相关交叉学科与生物信息学的跨学科文献数较其它学科少,但其跨学科的发展从年文献增加数据上看并不亚于其他学科(读者可查看图表源数据)。

2.3 被引文献跨学科分析

利用 web of science 对 13253 篇文献的施引文献进行分析。通过这个角度的分析,可以了解生物信息学在其他学科中

所具有的影响力。

由于 web of science 的引文分析不能对超过 10000 篇的文献同时进行分析,故本研究中将 2001-2010 年间共 13253 篇文献分为 2001-2005 及 2006-2010 两部分进行分析。

2001-2005 年论文数共 3714 篇,被引频次共 151703 次,每篇论文平均被引 40.86 次。H 因子为 137。去除自引的引文共 91485 篇(以上数据均来源于 web of science)。对施引文献按跨学科强度排列依次为生物化学与分子生物学、基因及遗传学、生物技术及应用为微生物、生物化学及研究方法、计算生物学、进化生物学、微生物学、计算科学、植物科学、数据及概率学等 10 门学科(图 3)。

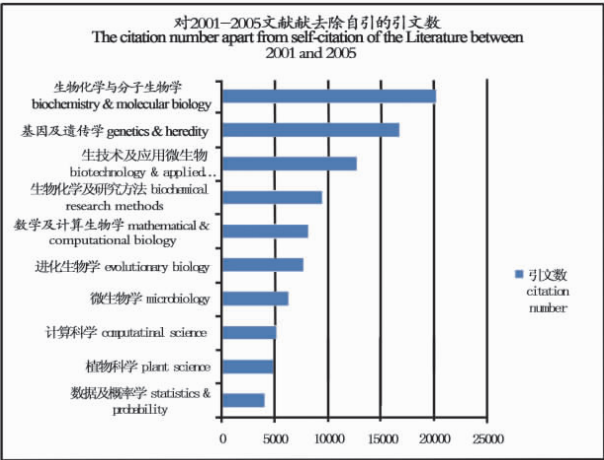


图 3 2001 年至检索时对 2001-2005 年间发表的生物信息学文献引文的学科类别分布
Fig.3 Subject category distribution of citation about the literature published between 2001 and 2005 from the year 2001 to the searching point

2006-2010 年论文数共 9539 篇,被引频次共 7013 次,每篇论文平均被引 7.35 次。H 因子为 74。去除自引的引文共 35816 篇(以上数据均来源于 web of science)。对施引文献按跨学科强度排列依次为生物化学与分子生物学、基因及遗传学、生物技术及应用为微生物、交叉学科、生物学、生物化学及研究方法、细胞生物学、进化生物学、生物物理学、微生物学、计算生物学等 10 门学科(图 4)。

对比 2001-2005 年间的论文被引用的情况,2006-2010 年间论文的被引频次较少,影响因子较小。一方面,2005 年后的论文发表时间较短。另一方面,2005 年后论文被引频率与其论文数的增加不成比例,表示 2005 年后论文数增加较快的同时,质量却相较 2005 前有所下降。这提示在学科之间跨学科研究的中应注重研究的实际意义以及论文质量。

3 讨论

3.1 生物信息学科的发展

研究发现生物信息学正处发展阶段,相关的论文数不断增加,并在 2009-2010 年趋于平台期。其发展与生物化学等学科的发展息息相关。因此,加强学科间研究对于生物信息学的快速发展至关重要。

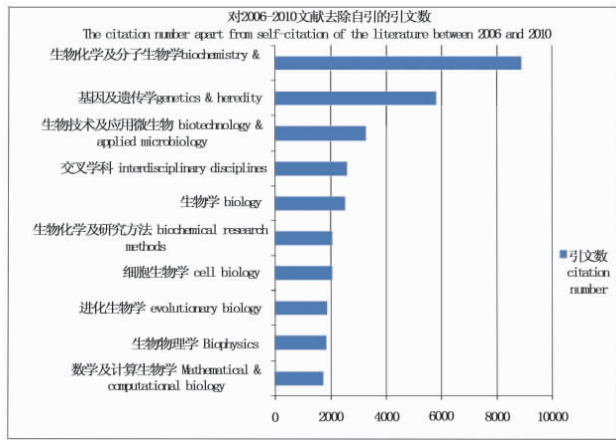


图4 2006年至检索时对2006-2010年间发表的生物信息学文献引文的学科类别分布

Fig.4 Subject category distribution of citation about the literature published between 2006 and 2010 from the year 2006 to the searching point

3.2 生物信息学科交叉与融合

从来源文献的学科分类来看,生物化学、计算生物学、生物技术及应用微生物、数据及概率学、计算科学、数学及相关学科等6个与生物信息学直接的跨学科研究较多。从引文的学科分类来看,2001-2005年间的文献跨学科类别主要是生物化学与分子生物学、基因及遗传学、生物技术及应用为微生物、生物化学及研究方法、计算生物学、进化生物学、微生物学、计算科学、植物科学、数据及概率学等10门学科。2006-2011年间的文献跨学科类别主要是生物化学与分子生物学、基因及遗传学、生物技术及应用为微生物、交叉学科、生物学、生物化学及研究方法、细胞生物学、进化生物学、生物物理学、微生物学、计算生物学等10门学科。其大致分类与主题的学科分类相吻合。因此下文主要针对主题的学科分类中6个学科进行跨学科内容的分析。

3.3 生物信息跨学科特征

3.3.1 生物信息学与生化、分子生物学研究 本研究中的13253篇文献可归类为生化及分子生物学的文献共12466篇。其中被引频次最高的可达7351次,被引频次超过500的文献共有25篇。主要内容主要包括核酸及蛋白质相关研究,如利用DNA及蛋白质序列信息对分子遗传的进化过程进行的研究^[1-3]、DNA多态性数据分析软件包^[6]、消除高密度寡核苷酸阵列中非生物起源阵列变异的标准化软件^[8]、增强计算效率的DNA树^[9]、同源蛋白质簇的数据库^[10,11]、核酸与蛋白质的多序列对比^[15]、蛋白质排序简化方法^[16]、基因显著性^[19]、蛋白质模型的最优化处理^[20]、同源蛋白基因组分类^[21]等研究。以及基因相关研究,如分子进化遗传分析软件包^[14]、疾病基因的连锁不平衡^[5]、多序列对比程序^[7]、利用上千个混合模型进行的基于最大似然比的物种分析方式^[12]、基因谱及变异的研究^[13,14]、微阵列基因表达数据集^[17]及用于分析的Bayesian体系^[18]等。

从文章的被引频次可看出,这些技术被广泛应用并衍生出更多新技术。这类研究大多是遗传调控网络、数学建模、仿真学、生物信息学、数学、生物学等学科交叉研究的产物^[23]。例如,在同类技术的不断革新中,对复杂生理过程的计算方法得到不

断简化,使得人们更容易获得细胞信号通路、物质代谢通路、基因调节等方面的信息^[21,23]。这类研究的共同特征是在分子生物学基础上,建立复杂的系统模型。在此基础上,Hucka M et al.^[21]统一了分子生物学各模型间的语言,从而解决了各生物模型间相互交融因语言不同而产生的障碍。这些资源都有由研究者创建的网站或数据库,便于提供免费应用。如Clamp M用于核酸及蛋白质多序列对比的网页^[15],Tusnady GE et al.^[22]创立的HMMTOP跨膜拓扑结构预测服务器。这不但推动了同类技术的进步,同时也使得文献得到更高的引用频次。这些研究成果一方面可以转化为生物信息学资料,同时也可以生物信息学领域的研究者根据已有信息进行更深更广的研究。

3.3.2 生物信息学与计算生物学 本研究中的13253篇文献可归类为计算生物学的文献共11951篇。其中被引频次最高的可达6479次,被引频次超过500的文献共有24篇。跨学科部分的文献具有明显的数学及计算学科特征,如单纯用于生物信息学分析的算法^[7],基于方差及偏倚对高密度核苷酸序列进行标准化的数学方法^[8],利用最大似然比或对置信度进化树进行分析^[19,24],基因直接相关联系的计算^[13,25],对cDNA微阵列表达的数据进行显著性分析^[19]等。这些研究结果有助于提高同类研究中预测精度及实验结果的准确性,如Troyanskaya O et al.^[17]提出的可正确评估DNA微阵列实验中丢失值的方法,Hudson RR^[14]创新的人群取样方法使研究中产生的偏倚减少了。这类文献的特征是与生化、分子生物学、生物信息学有着明显的交叉。说明生物信息学在与计算生物学跨学科领域的研究在较大程度上依赖于生物化学及分子生物学的发展。

3.3.3 生物信息学与生物技术及应用微生物学 本研究中的13253篇文献可归类为生物技术及应用微生物学的文献共10540篇。其中被引频次最高的可达7351次,被引频次超过500的文献共有25篇,与以上分析过的生化、分子生物学的25篇文献完全相同,因此也包含了计算生物学的24篇文献。如Barrett JC^[5]所研发的HaploView软件包。它不仅提供了人口统计数据中主要基因型的单体模式及连锁不平衡计算,并且让使用者拥有视觉吸引及互动界面。这类文献同时也包括了专业用于DNA多态性分析的软件包^[6]。这表示生物信息学与生物技术及应用微生物学之间跨学科的研究有与生物化学、分子生物学,计算生物学的发展是密不可分的。

3.3.4 生物信息学与计算科学及其交叉学科,统计学及概率论 本研究中的13253篇文献可归类为计算科学及其交叉学科,统计学及概率论的文献各6442篇。被引频次最高的可达6479次,被引频次超过500的文献共有22篇。这提示着计算科学作为一种工具被广泛应用于生物相关学科,并且也提示计算科学是生物信息学发展过程中一门不可或缺的基础学科。生物信息学正是在计算科学及计算机的普及基础上建立的。这一方面可验证生物信息学作为一门新兴学科的起源,同时也作为联系不同跨学科的锁链,使得不同领域的研究得以相连。在这五个学科中均出现的22篇文献有可能正是因为其跨学科性质较强,故被引频次较高。这也提示五种学科其他的文献引用频次相对较低,影响力相对较弱。加强学科内及学科间的研究有望加强相互间跨学科的研究。

3.3.5 生物信息学与数学及相关交叉学科 本研究中的13253

篇文献可归类为数学及相关交叉学科文献共 10540 篇。其中被引频次最高的为 306 次,被引频次超过 10 次的文献共有 20 篇。

这一学科的研究侧重数学方法,包括生物学中的数学分析,如微阵列基因的聚类分析^[26,34]及显著性分析^[27],基因相似性的数学分析^[28],不同物种基因组之间的比较^[29],利用基因对疾病(如癌症)进行分类^[30,31]等。以及相关研究中提出的数学算法,如快速寻找 motif 的算法^[32,33],用于基因定位、寻找隐性复杂性状遗传疾病标记链的新算法^[42],解决系统进化树研究中分割不兼容问题的 SplitsTree4 算法^[5]。可为生物网络中结构域的有效探测减少偏倚的算法^[35],基因组组合计算法则^[41]等。这些改进的算法以提出有意义的研究结果或对陈旧的法则进行改进等为目的^[37,38]减少了计算精度对结果的影响,降低了搜索成本^[39]。其他研究还包括多药治疗 HIV 方案选择的评价法则^[40],哺乳动物大脑自动映射的表达谱数据库及相关神经信息的基因表达通路^[36]等。

这类研究为生物学、生化、分子生物学等学科提供不断完善的数学工具,如公开的图、算法、法则、网站等,推动着相关交叉学科及生物信息学的发展。这一类引用频次较多的文献均侧重于应用,因此可发挥数学及其相关交叉学科的实际价值。该学科文献数相较之前的生物相关学科则少,最高引用频次也较小。同时,因为与生化、计算机、生物技术等学科存在交叉,故可能包含在以上几类学科的文献中。然而从文献增加的趋势来看,还处于持续发展的阶段,仍拥有着巨大的发展潜力。若加强其跨学科的研究则有可能获得更多的研究成果。

4 结论

生物信息学是一门热点前沿学科,它与多种学科交叉渗透。本研究揭示了生物信息学近十年的发展趋势,以及它与生物化学、计算生物学、生物技术及应用微生物、数据及概率学、计算科学、数学及相关学科等自然学科相交叉的情况。生物信息学的进步依赖于相关学科不断突破的同时,其发展又为这些学科的进步提供信息、材料及研究方法。

生物信息学及其跨学科的研究仍处于不断发展的阶段。其在 2010 年里的发展速度相对减缓,意味着过去 10 年的研究思路仍需进一步突破。加强数学、生物物理学、植物科学等学科与生物信息学之间的跨学科研究,并扩大跨学科的范围至多学科,对这些学科的研究能起到推动作用。同时,研究者应扩宽眼界,不应仅着眼于论文的发表数量,而应注重研究的实际应用。这样不仅能受益更多的科研工作者,同时也能让自己的论文获得更多的引用频次。另外,全球的研究者应将自己的研究成果最大化的共享,正如 2001 至 2005 年间很多的研究者那样把开发的程序、数据库或改进的算法和分析思路放在开放的网络平台上,开拓出更多新的领域能更大程度地推动科技发展。

参考文献 (References)

- [1] Kumar S, Tamura K, Nei M. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment [J]. Brief Bioinform, 2004, 5(2):150-163
- [2] Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees[J]. Bioinformatics, 2001, 17(8):754-755
- [3] Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models [J]. Bioinformatics, 2003, 19(12): 1572-1574
- [4] Kumar S, Tamura K, Jakobsen IB, et al. MEGA2: molecular evolutionary genetics analysis software [J]. Bioinformatics, 2001 Dec; 17(12): 1244-1245
- [5] Barrett JC, Fry B, Maller J, et al. Haploview: analysis and visualization of LD and haplotype maps[J]. Bioinformatics, 2004, 21(2): 263-265
- [6] Rozas J, et al. DnaSP, DNA polymorphism analyses by the coalescent and other methods[J]. Bioinformatics, 2003, 19(18): 2496-2497
- [7] Larkin MA, et al. Clustal W and clustal X version 2.0[J]. Bioinformatics, 2007, 23(21): 2947-2948
- [8] Bolstad BM. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias [J]. Bioinformatics, 2003, 19(2): 185-193
- [9] Schmidt HA, et al. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing [J]. Bioinformatics, 2002, 18(3): 502-504
- [10] Tatusov RL, Fedorova ND, Jackson JD, et al. The COG database: an updated version includes eukaryotes [J]. BMC Bioinformatics, 2003, 11; 4:41
- [11] Arnold K, Bordoli L, et al. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling[J]. Bioinformatics, 2006, 22(2): 195-201
- [12] Stamatakis A. RAXML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models [J]. Bioinformatics, 2006, 22(21): 2688-2690
- [13] Purcell S et al. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits [J]. Bioinformatics, 2003, 19(1): 149-150
- [14] Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation[J]. Bioinformatics, 2002, 18(2): 337-338
- [15] Clamp M, et al. The Jalview Java alignment editor[J]. Bioinformatics, 2004, 20(3): 426-427
- [16] Edgar RC, et al. MUSCLE: a multiple sequence alignment method with reduced time and space complexity [J]. BMC Bioinformatics, 2004, 5:1-19
- [17] Troyanskaya O, et al. Missing value estimation methods for DNA microarrays[J]. Bioinformatics, 2001, 17(6): 520-525
- [18] Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes[J]. Bioinformatics, 2001, 17(6): 509-519
- [19] Wolfinger RD. Assessing gene significance from cDNA microarray expression data via mixed models[J]. Journal Of Computational Biology, 2001, 8(6): 625-637
- [20] Federico Abascal, et al. ProtTest: selection of best-fit models of protein evolution[J]. Bioinformatics, 2005, 21(9): 2104-2105
- [21] M. Hucka, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models[J]. Bioinformatics, 2003, 19(4): 524-531
- [22] Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server[J]. Bioinformatics, 2001, 17(9): 849-850
- [23] De Jong H. Modeling and simulation of genetic regulatory systems: A

- literature review [J]. Journal of Computational Biology, 2002, 9(1): 67-103
- [24] Shimodaira H, Hasegawa M. CONSEL: for assessing the confidence of phylogenetic tree selection [J]. Bioinformatics, 2001, 17 (12): 1246-1247
- [25] Sanderson MJ. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock[J]. Bioinformatics, 2003, 19 (2): 301-302
- [26] Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: A survey[J]. IEEE-ACM Transactions on Computational Biology AND Bioinformatics, 2004, 1(1): 24-45
- [27] Zhang CL. Significance of gene ranking for classification of microarray samples [J]. IEEE-ACM Transactions on Computational Biology AND Bioinformatics, 2006, 3(3): 312-320
- [28] Popescu M, et al. Fuzzy measures on the gene ontology for gene product similarity [J]. IEEE-ACM Transactions on Computational Biology AND Bioinformatics, 2006, 3(3): 263-274
- [29] Zheng CF, et al. Removing noise and ambiguities from comparative maps in rearrangement analysis[J]. IEEE-ACM Transactions on Computational Biology AND Bioinformatics, 2007, 4(4): 515-522
- [30] Wang LP, et al. Accurate cancer classification using expressions of very few genes [J]. IEEE-ACM Transactions on Computational Biology AND Bioinformatics, 2007, 4(1): 40-53
- [31] Tang YC, et al. Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis [J]. IEEE-ACM Transactions on Computational Biology AND Bioinformatics, 2007, 4(3): 365-381
- [32] Handl J. Multiobjective optimization in bioinformatics and computational biology [J]. IEEE-ACM Transactions on Computational Biology AND Bioinformatics, 2007, 4(2): 279-292
- [33] Brown DG, Harrower IM. Integer programming approaches to haplotype inference by pure parsimony [J]. IEEE-ACM Transactions on Computational Biology AND Bioinformatics, 2006, 3(2): 141-154
- [34] Au WH, et al. Attribute clustering for grouping, selection, and classification of gene expression data[J]. IEEE-ACM Transactions on Computational Biology AND Bioinformatics, 2005, 2(2): 83-101
- [35] Wernicke S. Efficient detection of network motifs [J]. Ieee-acm Transactions on Computational Biology and Bioinformatics, 2006, 3(4): 347-359
- [36] Ng L, et al. Neuroinformatics for genome-wide 3D gene expression mapping in the mouse brain [J]. IEEE-ACM Transactions on Computational Biology AND Bioinformatics, 2007, 4(3): 382-393
- [37] Kucherov G, et al. Multiseed lossless filtration[J]. IEEE-ACM Transactions on Computational Biology AND Bioinformatics, 2005, 2 (1): 51-61
- [38] Elias I, Hartman T. A 1.375-approximation algorithm for sorting by transpositions[J]. IEEE-ACM Transactions on Computational Biology AND Bioinformatics, 2006, 3(4): 369-379
- [39] Cameron M, et al. Improved gapped alignment in BLAST [J]. Ieee-acm transactions on computational biology and bioinformatics, 2004, 1(3): 116-129
- [40] Neri F, et al. An adaptive multimeme algorithm for designing HIV multidrug therapies [J]. IEEE-ACM Transactions on Computational Biology AND Bioinformatics, 2007, 4(2): 264-278
- [41] Berard S, et al. Perfect sorting by reversals is not always difficult[J]. IEEE-ACM Transactions on Computational Biology AND Bioinformatics, 2007, 4(1): 4-16
- [42] Dawy Z, et al. Gene mapping and marker clustering using Shannon's mutual information [J]. IEEE-ACM Transactions on Computational Biology AND Bioinformatics, 2006, 3(1): 47-56