文章编号：2095-2163(2019)01-0272-05 中图分类号：TM614 文献标志码：A

基于内存计算的基因疾病搜索系统

杨 勤，臧天仪
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要： 随着现代测序技术的发展，产生海量生物数据，快速发展的生物信息学也在不断剖析这些数据的隐藏生物信息。通过生物网络研究基因型与疾病表型的关联关系从而实现致病基因的预测和寻找基因导致的疾病。基于疾病基因模块性特征，提出整合蛋白质相互作用网络、疾病表型相似性网络、疾病-基因对应网络，构建异构生物网络，改进网页排序算法 TrustRank，对候选基因与疾病进行优先级排序，实现预测功能。本文还将通过 Spark 平台开发基因疾病搜索系统，数据存储在 HBase 中，形成大数据存储、处理、分析的解决方案，对临床诊断和疾病治疗提供新思路。

关键词： 基因型；疾病表型；大数据；TrustRank

Genetic disease search system based on memory computing

YANG Qin, ZANG Tianyi
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 With the development of modern sequencing technology, resulting in massive biological data, the rapid development of Bioinformatics is also constantly analyzing the hidden information of these data. Through the biological network to study the relationship between genotype and disease phenotype, the prediction of pathogenic genes could be achieved and diseases caused by the genes found. Based on the modular nature of the disease gene, the paper proposes to integrate the protein interaction network, disease phenotype similarity network, disease-gene correspondence network, construct heterogeneous biological network, improve the web page sorting algorithm TrustRank, prioritize candidate genes and realize diseases forecasting function. This paper also develops the genetic disease search system through the Spark platform. The data are stored in HBase, which form a large data storage, processing and analysis solution, and provide new ideas for clinical diagnosis and disease treatment.

【Key words】 genotype; disease phenotype; big data; TrustRank

0 引 言

近年来随着人类基因组计划(Human Genome Project, HGP)的顺利交付，生物技术不断进步，从而诞生了生命科学和计算机科学结合起来的新科学—生物信息学。生物信息学是结合应用数学、统计学和计算机科学的方法研究生物问题，是建立在分析生物学的基础上，其研究重点主要体现在基因组学(Genomics)和蛋白质组学(Proteomics)2 方面。具体就是从核酸和蛋白质序列出发，分析序列中表达的结构功能的生物信息。由于下一代高通量测序(NGS)技术的迅猛发展，生物实验方法和检测手段正日趋丰富多样，由此也就生成了海量生物数据。如何有效利用这些数据来研究疾病的发生机理，寻找致病基因即已成为生物信息学的重要研究分支。

利用现有的数据进行致病基因的预测，提高致病基因的检验效率是目前生物信息学的研究热点。作为生物重要特征，基因型(Genotype)指的是一个生物体内的 DNA 所包含的基因，表型是指受基因、环境等影响而在生物体上表现出来的特征。整合生物学数据库是为了更好地研究基因与疾病之间关系，而目前各种分析策略在整体上可划定为基于文本的研究、基于网络的研究和基于本体的研究三大类。生物大数据一直都是业界瞩目的焦点，况且生物数据量庞大，数据格式未能统一，传统方法分析颇显复杂繁冗，许多生物项目都迁移到大数据平台亟待处理，对其进行大数据分析、整合与挖掘则已尤其显得紧急与迫切。

本次研究主要分 2 个部分。一部分是整合相关的基因型、表型和疾病数据，充分利用蛋白质相互作用网络，疾病相似性网络。疾病-基因二分网络构建异构综合网络，将已知的致病基因作为种子节点，改进网页排序 TrustRank 算法，设计提出 YSearch 算法模型预测致病基因；另一部分则是系统实现，将整合后的数据存储在 NoSql 数据库 HBase，继而通过 Spark 大数据框架构建搜索引擎，编码实现预测算

作者简介：杨 勤(1992-)，男，硕士研究生，主要研究方向：生物信息学、大数据分析；臧天仪(1968-)，男，博士，教授，博士生导师，主要研究方向：服务计算与服务网络、生物医学大数据计算、数据密集型计算等。

收稿日期：2017-06-23

法，输出致病基因的排序结果，从而为疾病治疗提供参考。研究内容可论述如下。

1　相关理论与方法

1.1　相关生物网络

疾病网络与基因网络是根据 OMIM 数据构建，OMIM 是持续更新的、关于人类基因和遗传紊乱的数据库。对于疾病相似性网络 D，其中节点 $d_1, d_2, \cdots, d_n \in V_D$ 在网络中代表一种疾病，2 个节点间的连线 E_D 表示 2 种疾病是相似的，边的权重 $\boldsymbol{A}_{\boldsymbol{D}_{i,j}}$ 表示相似程度。表型描述的更标准化方法包括每个特征的频率估计，将大大增加基因型-表型相关性分析的产量。邻接矩阵归一化 $\boldsymbol{D}$ 满足：

$$\boldsymbol{D}_{i,j} = p(d_j \mid d_i) = \boldsymbol{A}_{\boldsymbol{D}_{i,j}} / \sum_j \boldsymbol{A}_{\boldsymbol{D}_{i,j}} \tag{1}$$

蛋白质相互作用网络 G，其中节点 $g_1, g_2, \cdots, g_m \in V_G$ 在网络中代表一个蛋白质，如果蛋白质 g_i 与蛋白质 g_j 存在相互作用关系，则邻接表 $\boldsymbol{A}_{\boldsymbol{G}_{i,j}} = 1$，否则为零。PPI 网络数据来自于 HPRD 数据库，通过预处理将数据存储在邻接表中。PPI 邻接矩阵归一化 $\boldsymbol{G}$ 满足：

$$\boldsymbol{G}_{i,j} = p(g_j \mid g_i) = \boldsymbol{A}_{\boldsymbol{G}_{i,j}} / \sum_j \boldsymbol{A}_{\boldsymbol{G}_{i,j}} \tag{2}$$

基因-疾病对应网络 DG 与 GD，其中从疾病到基因的转移矩阵 $\boldsymbol{G}_{DG}$，如果基因 g_i 是疾病 d_j 的致病基因，则邻接矩阵 $\boldsymbol{A}_{i,j} = 1$，否则为零。OMIM 数据文件中每条记录表示每个疾病表型描述对应的致病基因条目，处理后的转移概率满足：

$$\boldsymbol{M}_{DG} = p(g_i \mid d_j) = \boldsymbol{A}_{i,j} / \sum_j \boldsymbol{A}_{i,j} \tag{3}$$

同理，从基因到疾病的转移矩阵 $\boldsymbol{M}_{GD}$ 满足：

$$\boldsymbol{M}_{GD} = p(d_j \mid g_i) = \boldsymbol{A}_{j,i} / \sum_i \boldsymbol{A}_{j,i} \tag{4}$$

首先构建生物信息的异构网络表明来自多个公共资源的先验信息，表示成 $G = (V, E)$，其中 V 表示节点集合，E 表示边集合，蛋白质相互作用网络是无向无权图；疾病-基因对应网络是有向无权图；疾病表型相似性网络是无向有权图。在异构网络中存在着 4 种状态转移，抽象出来即如图 1 所示。

1.2　改进 TrustRank 算法

本文设计疾病表型-基因关联算法 YSearch 是基于网页排序算法 TrustRank 的设计改进，算法包括 2 种形式：查询疾病的致病基因以及查询基因导致的疾病。分别是在疾病表型相似性网络与蛋白质相互作用网络随机游走，还有疾病-基因二分网络的迭代处理。算法的设计代码描述如下。

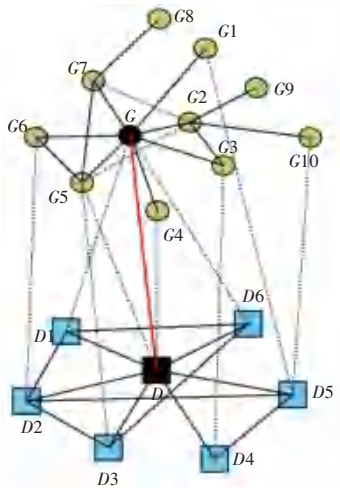

图 1　整合后的异构网络

Fig. 1　Integrated heterogeneous network

算法：YSearch

输入：$\boldsymbol{G}$ 为蛋白质相互作用转移矩阵；$\boldsymbol{D}$ 为疾病表型相似性转移矩阵；$\boldsymbol{M}$ 为疾病 - 基因对应转移矩阵；α 为调整参数；n 为算法迭代次数

输出：TR 为分数

```
Begin
    s = SelectSeed( )// 种子集
    TR^0 = s
    for i = 1 to n do
        TR^(k+1) = α · M · TR^k + (1 − α) · s
     Return TR
end
```

算法的工作原理可概括为：先人工识别高质量节点（即种子集），种子集指向的节点质量也可能高，即 TR 值高，与种子集节点连接越远，节点的 TR 值越低。综上可知，TR 算法也就是一个利用网络的拓扑性质在全局网络中进行排序的设计过程。

2　系统框架

2.1　数据

考虑到迄今尚未见到专门的基因相互作用网络数据，且假设蛋白质与基因相对应，因此研究利用了蛋白质相互作用网络（PPI）。数据来源是 HPRD 数据库（Human Protein Reference Database）。这是与人类蛋白质有关的蛋白质组学信息数据库。本文的 HPRD 数据是通过网址 http://hprd.org/download 下载给定的 txt 文件。文件存储着蛋白质相互作用数据，文件格式详见表 1。

表 1 HPRD 数据格式

Tab. 1 HPRD data format

Gene symbol	HPRD id	RefSeq id	Gene symbol	HPRD id	RefSeq id	Experiment type	Pubmed id
ITGA7	2761	NP_001138468.1	CHRNA1	7	NP_001034612.1	in vivo	10910772

疾病表型相似性的数据来源是 MimMiner 网站，van 等人使用 MeSH 解析了 OMIM 数据库，对其中 5 000多种人类表型进行文本挖掘，生成疾病表型相似性网络。通过网址 http://www.cmbi.ru.nl/MimMiner/suppl.html 下载数据文件。文件的每一行开头都是一个蛋白质 MIM 编号，其后就依序排布着一系列与其有相互作用关系的蛋白质 MIM 编号和相似度，一共是 5 080 * 5 080 的对称矩阵。

基因－疾病对应网络的数据来源是 OMIM (Online Mendelian Inheritance in Man)数据库，通过网址 https://omim.org/downloads/，下载 morbidmap.txt 文件。文件格式详情可参见表 2。

表 2 morbidmap 文件数据格式

Tab. 2 morbidmap file data format

疾病表型描述	基因标志	MM 编号	染色体位置
17-alpha-hydroxylase/17,20-lyase deficiency, 202110(3)	CYP17A1, CYP17, P450C17	609300	10q24.32

2.2 Spark 平台

本文采用的是 Spark on yarn 平台，Apache Spark 是一个以速度、易用性和复杂分析为特点构建的大数据处理框架。Spark 在数据处理过程中使用成本更低的洗牌(Shuffle)方式，提升 MapReduce 性能，由于内存数据存储和实时的处理能力，Spark 比其它的大数据处理技术的性能要更加出色。还支持大数据查询的延迟计算，可以优化大数据处理流程。关于优化，仍需补充的一点就是，当需要多次处理同一数据集时，将中间结果保存在内存中而不是将其写入磁盘的 Spark 的设计初衷就是研发既可以在内存中、又可以在磁盘上工作的执行引擎。当内存中的数据过期时，Spark 操作符就会执行外部操作，可以将某个数据集的一部分送入内存而剩余部分置于磁盘中。Spark 的性能优势得益于这种内存中的数据存储。基于此，可得 Spark 生态系统的架构设计如图 2 所示。

2.3 HBase

大量生物网络数据都存储在 NoSQL 数据库 HBase 中，通过 Spark 平台操作数据。HBase 是一个面向列、可靠性高的分布式存储系统，一个开源的非关系型分布式数据库(NoSQL)。在技术上，改进了谷歌的 BigTable 方法，利用 HBase 技术可在廉价 PC 机群上搭建起大规模存储集群，HBase 为了扩展海量数据，可采用增加节点实现线性扩展，从而可以在集群上管理大量非结构化或半结构化的稀疏数据。HBase 仅能通过主键或主键的 range 检索数据来支持单行事务操作。需要注意的是，HBase 的数据存储形式与其它数据库不一样，其中包含了：行键(Row Key)、时间戳(Timestamp)、列族(Column Family)、表和区域(Table&Region)和单元格(Cell)。HBase 系统内部框架如图 3 所示。

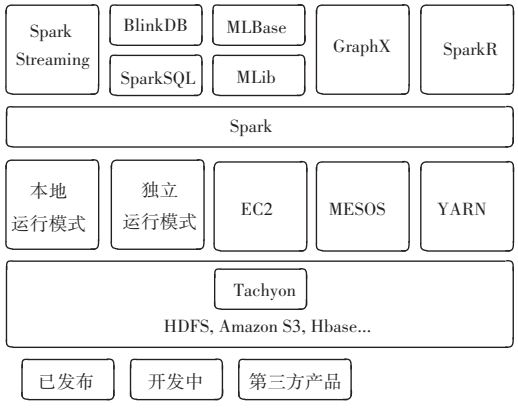

图 2 Spark 生态系统

Fig. 2 Spark system

3 系统设计实现

基因型是控制生物性状的基因座上特定等位基因的组合，在不考虑具体基因座时可以泛指生物个体的全部遗传组成。与基因型相对的是表型，而表型是指生物个体表现的性状。基因型是表型的遗传基础。不同基因型表现相同表型、以及相同基因型表现不同表型的现象广泛存在，这使得从基因型到表型的遗传调控这一科学难题颇显研究难度。而研究中通过运用计算机大数据技术，汇聚整合这些数据库的问题就归结集中在数据格式不统一上。输入文件的数据来源主要有注释变异文件、VCF 文件、xml 文件等，整体

大数据框架是 Spark on yarn 平台，主要选用技术包括 MongoDB 数据库存储和 Spark SQL 查询，输出的是与输入相关的基因型或者表型。设计课题提供一个可扩展且高性能的存储、处理、分析基因大数据的解决方案，设计构建框架如图 4 所示。

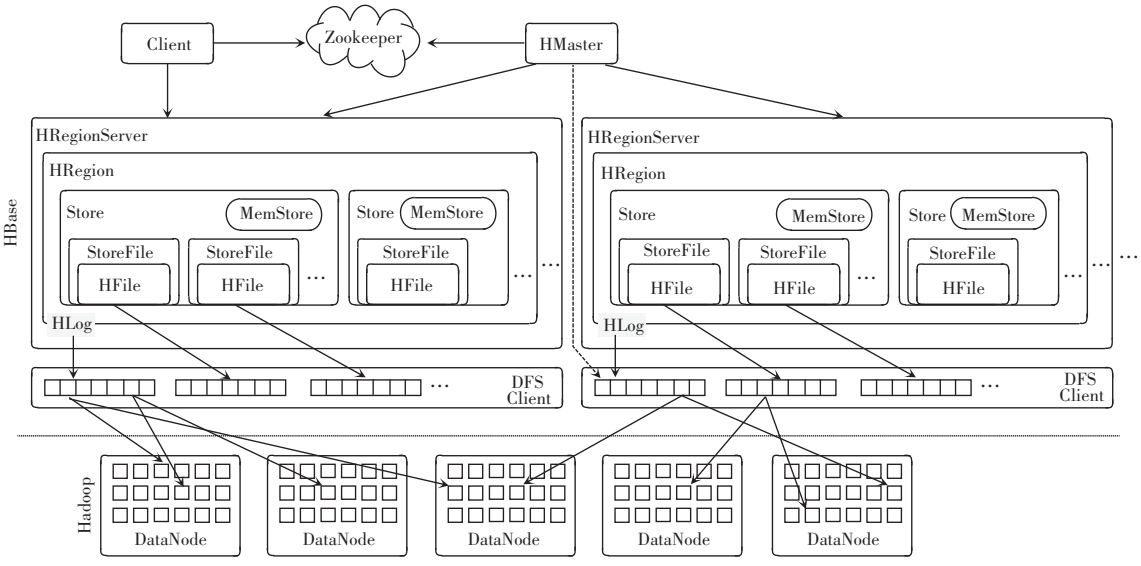

图 3　HBase 系统架构

Fig. 3　HBase system structure

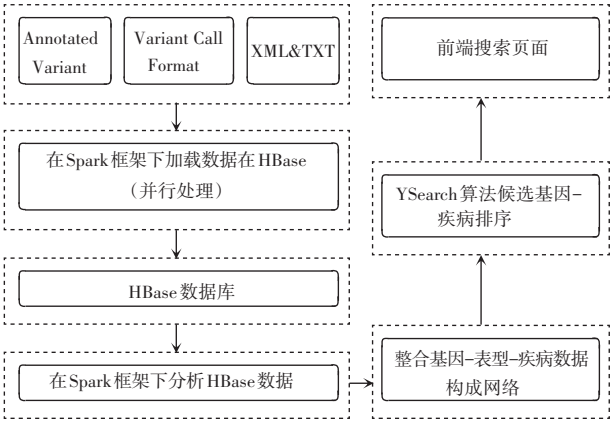

图 4　YSearch 系统框架

Fig. 4　YSearch system framework

系统首页是搜索页面，如图 5 所示。可以手动输入疾病或基因名称，也可以点击示例输入，再点击搜索，生成如图 6 所示的结果页面。比如搜索疾病寻找致病基因，通过输入疾病表型名称，在结果页面中展示了相关联的蛋白质名称与其相关度，然后可以点击蛋白质名称跳转到 HPRD 数据库中相关蛋白质的详尽解读内容。链接 HRRD 效果界面则如图 7 所示。同样查询基因导致的疾病也是类似操作。

图 5　系统首页

Fig. 5　System home page

图 6　搜索结果页面

Fig. 6　Search results page

图 7 链接 HPRD 页面

Fig. 7 Link the HPRD page

4 结束语

使用计算方法预测候选基因-疾病相关性既可以研究发病机理,而且也有助于疾病诊断、治疗,以及预防。近年来随着精准医疗的提出与广受关注,个人医疗数据正日趋丰富,通过大数据技术来展开处理研究已成为未来的热点领域。同时个人健康也正成为时下的热、焦点话题,随之优化疾病-基因关联算法,完善搜索系统则更加显现出兼具着不容忽视的社会和经济双重效益。在此基础上,该研究对生物医药产业的发展也必将发挥不可低估的重要推动作用。

参考文献

[1] LI Yongjin, LI Jinyan. Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data [J]. BMC Genomics, 2012, 13(S7):S27.

[2] ZHAO Z Q, HAN G S, YU Z G, et al. Laplacian normalization and random walk on heterogeneous networks for disease - gene prioritization[J]. Computational Biology & Chemistry, 2015, 57(C):21-28.

[3] KÖHLER S, BAUER S, HORN D, et al. Walking the interactome for prioritization of candidate disease genes[J]. American Journal of Human Genetics, 2008, 82(4):949-958.

[4] LAGE K, KARLBERG E O, STøRLING Z M, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders[J]. Nature biotechnology, 2007, 25(3): 309-316.

[5] MASINO A J, DECHENE E T, DULIK M C, et al. Clinical phenotype - based gene prioritization: An initial study using semantic similarity and the human phenotype ontology[J]. BMC bioinformatics, 2014, 15(1): 248.

[6] 魏春水. 基于随机游走的疾病-基因关联算法[D]. 西安:西安电子科技大学, 2012.

[7] LE D H, DANG V T. Ontology-based disease similarity network for disease gene prediction[J]. Vietnam Journal of Computer Science, 2016, 3(3): 197-205.

[8] RAMOS E M, HOFFMAN D, JUNKINS H A, et al. Phenotype-genotype integrator (PheGenI): Synthesizing genome - wide association study (GWAS) data with existing genomic resources[J]. European Journal of Human Genetics, 2014, 22(1): 144-147.

[9] VAN DRIEL M A, BRUGGEMAN J, VRIEND G, et al. A text-mining analysis of the human phenome[J]. European Journal of human genetics, 2006, 14(5): 535-542.

[10] SCOTT A F, AMBERGER J, BRYLAWSKI B, et al. OMIM: Online mendelian inheritance in man [M]// ETOVSKY S. Bioinformatics: Databases and Systems. Boston, MA: Springer, 2002:57-61.

[11] GYÖNGYI Z, GARCIA-MOLINA H, PEDERSEN J. Combating web spam with trustrank[C]//Proceedings of the 30th international conference on Very large data bases - Volume 30. Toronto, Canada:Morgan Kaufmann, 2004: 576-587.

[12] GUO Xingli, GAO Lin, WEI Chunshui, et al. A computational method based on the integration of heterogeneous networks for predicting disease-gene associations[J]. PloS one, 2011, 6(9): e24171.

(上接第 271 页)

参考文献:

[1] 江蜀华,王薇. 电工电子技术基础[M]. 西安:西安电子科技大学出版社,2009.

[2]杨宇,赵玉霞,徐刚. 555 定时器功能及应用的探讨[J]. 科技广场, 2007(5): 234- 235.

[3]刘琳霞. 单片机在温度控制系统设计中的应用研究[J]. 内燃机与配件 ,2017(21):107.

[4]李鑫宇,肖雪. 基于单片机的温度控制系统设计研究[J]. 通信电源技术,2018,35(2):15-16.

[5]张洪润. 实用自动控制[M]. 成都:四川科技出版社, 1993.

[6] 李新刚,于巍巍. 智能电加热温控系统的研制[J]. 机械工程师,2005(7):68-69.

[7] 余瑾 姚燕. 基于 DS18B20 测温的单片机温度控制系统[J]. 微计算机信息,2009, 25(3-2):105-106,112.