基于蚁群算法的基因路径预测①

张伟娟, 张红梅, 陈 峰

(河南工业大学 信息科学与工程学院, 郑州 450000)

摘 要: 随着基因测序速度的快速增长, 人们开始着手于研究具体生物过程的基因路径, 但生物实验的限制性使之进展缓慢. 文中提出了基于蚁群算法(Ant colony optimization ACO)的基因路径预测技术, 以 wnt 信号通路为例, 使用对应的基因集合内部的联系设置算法参数, 对其进行基因路径预测. 实验结果显示, 使用该方案预测的路径具有很大的可参考性, 能够为基因路径绘制实验指引研究方向, 提高研究的速度.

关键词: 基因路径; 蚁群算法; wnt 信号通路; 自适应

Prediction of Gene Pathway Based on Ant Colony Optimization

ZHANG Wei-Juan, ZHANG Hong-Mei, CHEN Feng

(College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450000, China)

Abstract: With the rapid growth of gene sequencing, people began to study gene pathway of specific biological processes, but biological experiments restrictive make it slowly. In this paper, it proposed a gene pathway prediction technical which based on ant colony optimization, used the corresponding gene in wnt signal pathway as data and its internal relations set algorithm parameters to predict gene pathway. The results show that this program has a great reference value, it can guide the path of research and increase the speed of research.

Key words: gene pathway; ant colony optimization; wnt signal pathway; adaptive

1 引言

随着人类基因组(测序)计划的逐步实施以及分子生物学相关学科的迅猛发展, 基因序列数据正在以前所未有的速度迅速增长, 人们致力于研究基因的各种功能, 希望揭开生命的秘密.

其中, 基因路径是指在同一生物过程中发生作用的基因集合, 在路径上的映射[1]. 其基本特点要求对路径的研究不仅需要基因功能的确定, 还需要对其在生物过程中的的关系, 如时间关系、功能关系等十分了解. 因此工作量十分庞大, 需要耗费大量的时间、人力及物力, 但也未必能得到正确的结果.

在已知具有疾病调控作用的抗病基因的情况下, 如何确定基因路径, 目前的基因路径多数使用标记基因等技术测得, 然后人工绘制, 如 KEGG 数据库中的路径[2,3], 这样得到的路径速度缓慢, 耗资巨大. 而新兴的生物信息学能够对基因功能进行注释, 分析基因、基因产物之间的相互作用关系, 绘制基因调控网络图, 达到预测基因路径的目的, 越来越广泛的应用到功能基因组学、药物基因组学和人类复杂疾病的研究中[4].

因此, 文中提出的方案, 正是基于生物信息学的方法, 利用蚁群算法的正反馈并行自催化机制, 在 wnt[5,6]生物过程所涉及的基因中, 利用已知的各个基因间的相互关系设置路径搜索参数, 探索可能的基因路径, 结果发现搜索结果对于路径的研究具有较强的参考性.

2 方法

要使用生物信息学的方法求取基因路径, 必须忠实于基因本身的特点. 通过目前的研究, 可知基因序

① 基金项目:国家自然科学基金(61203265);河南省重点项目(122102110106);创新人才项目(171144)
收稿时间:2014-09-16;收到修改稿时间:2014-10-16

列, 基因所属的基因组, 以及部分基因的功能等基因属性. 其中基因序列, 即 DNA 序列是多数生物信息研究的对象. 要在如此混乱无序的数据信息中寻找基因路径, 对路径求取算法的要求是十分严格的.

当前使用最为普遍的路径求取算法有 Dijkstra 算法、Floyd 算法、Kruskal 算法、A*算法等, 它们广泛的应用在各种求取最短路径的领域中, 但用在这里却不适合. 首先, 因为基因的各种属性数据多且无序, 需要算法具有众多的参数以便求取时对路径的寻找进行控制, 而这些算法参数少而且求取方法简单, 不易控制; 其次, 基因功能在研究中不断的更新, 如果使用这些算法, 就必须对初识的路径信息进行全部更新, 工作量巨大. 因此, 文中提出了使用对众多蚂蚁动态寻找最短路径的仿真算法——蚁群算法.

蚁群算法最初是由意大利学者 Dorigo M 于 1991 年首次提出[7-10], 是一种模拟昆虫王国中蚂蚁群体智能行为的仿生学优化算法, 蚂蚁根据根据当前的信息素等因子使用公式(1):

$$p_{ij}^{k}(t)=\begin{cases}\dfrac{[\tau_{ij}(t)]^{\alpha}[\eta_{ik}(t)]^{\beta}}{\sum_{s\subset allowed_k}[\tau_{is}(t)]^{\alpha}[\eta_{is}(t)]^{\beta}} & 若 j\subset allowed_k \\ 0 & 否则\end{cases} \quad (1)$$

进行概率选择策略, 找寻最短路径, 并通过时时刷新信息素保证正反馈机制. 其中, $allowed_k$ 表示蚂蚁 k 下一步能选择的点. 具体的参数意义可以参考文献 9, 在此不再赘述.

使用蚁群算法, 可以处理基因的序列信息作为路径信息, 将基因的基因组信息、功能信息等标记到蚂蚁的信息素矩阵中, 指导蚂蚁进行路径寻找, 并使用各种因子对路径寻找进行控制. 这样一方面避免了对基因特性的忽视, 能够按照基因的基本属性寻找路径, 另一方面, 避免了功能更新时的大量的工作, 只需对信息素中相关数据进行更新即可. 所以, 使用蚁群算法对基因路径进行预测.

基于相似数据离散的特点以及预测基因路径的目的, 算法采用 Ant-Cycle 模型. 此外, 为了避免出现停滞现象, 文中采用自适应蚁群算法, 使用确定选择和随机选择相结合的选择策略, 可以在搜索中动态的调整状态选择概率, 下文提到的蚁群算法均是指自适应蚁群算法.

这次实验以 wnt 信号通路中的各个基因为实验对象, 首先使用目前较为成熟的序列比对技术 BLAST[11], 得到基本的基因相似性参数, 作为蚁群算法最初进行路径选择的路径信息参数矩阵, 之后根据对各基因现有的研究成果更新信息素, 最后使用蚁群算法预测路径.

通过上述方法, 就可以对已知的基因进行基因路径的探索, 完成基因路径预测的目的.

3 实验

3.1 实验过程

3.1.1 流程简介

实验流程如下所示:

① 采集数据 实验的初始数据数据来自 WNT 路径中的基因. 序列数据在 NCBI, KEGG 数据可中可搜索到.

② 处理数据 使用BLAST进行序列比对, 得到序列之间的相似性数据, 消除离群点.

③ 定义蚁群算法 此步是基因路径预测与蚁群算法的接口, 也是实验中最关键的部分. 重点是将基因的已知信息融入蚁群算法的路径搜索中去: 第一, 将流程 2 得到的序列相似数据作为路劲搜索的依据—路径信息矩阵 D. 第二, 综合各类基因研究成果对信息素矩阵进行定义与更新, 具体操作如下一小节所示.

④ 进行预测 依次对每个基因与剩余基因的路径进行预测.

3.1.2 参数分析与设置

使用蚁群算法前, 由于基因数据的特殊性, 需要对信息素矩阵谨慎定义. 需要在了解基因集合中各个基因之间现有的关系研究的情况下, 对其进行差别定义, 即对那些已经证明具有相同、相似、相继等作用的基因间的信息素要差别定义. 如在实验中的 wnt 路径中, Homo sapiens: 27121[12] 和 Homo sapiens: 27123[13,14]都属于 dkk 基因组, 它们的信息素在本次实验中设置为一般信息素的 2 倍.

此外, 公式 1 中的参数 α 和 β 对蚁群路径的寻找也是至关重要的, 其中, α 表示信息素的相对重要性, α 越大路径越偏向于信息素的选择, 随机性减小; 参数 β 表示期望启发因子的相对重要性, β 越小, 算法局部最优, 收敛越快. 如图 1 所示, 设置了多个 α 和 β 的值进行结果比较, 可以看出, 当 α=4、β=3 时结果最好.

选定合适的 α 和 β 后, 对所有基因使用蚁群算法进行遍历, 搜索所有可能的路径. 得到最终的路径预测结果.

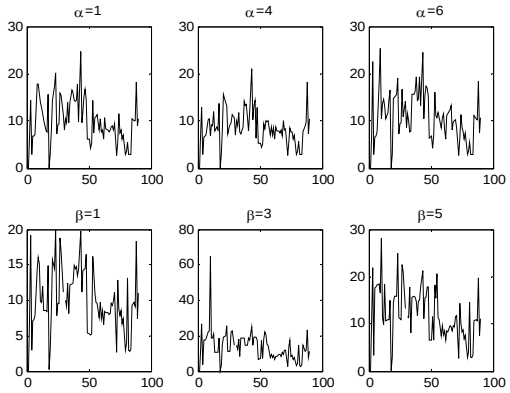

图 1 参数比较

3.2 实验结果

在实验中，选取各种基因点作为起始点和终止点，预测可能基因路径，图 2 是其中部分有代表性的基因路径.

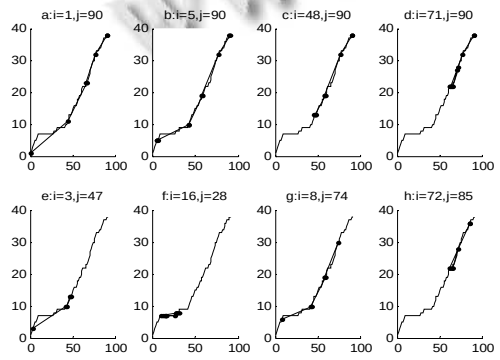

图 2 基因路径

图 2 中每个图中皆有两条路径曲线，固定不变的是正确的路径，变化的是求取到的路径，图中显示的较大的黑点是路径中所涉及到的基因.

分析实验结果可知，使用蚁群算法得到的基因路径具有以下特点:

① 路径方向的正确性 从图 6 中的 a 到 h 不难发现，所有求得的路径与正确路径具有具有高趋向性，能够正确的指示信号通路中的基因路径. 对生物实验验证具有较高的指导意义.

② 短路径的高准确性 比较所有图可以发现: d、f、h 所代表的小路径求解得到的路径更详细，更趋于原始路径，具有较高的准确率.

③ 不稳定性 由图可知，图 a 中的路径对图 b、c、d 中路径应具有覆盖作用，但是得到路径中的基因并不是相同的，因此路径具有不稳定性.

此外，较于其他算法，蚁群算法的结果易受蚁群规模及迭代次数的影响，从而产生多样性结果，即特点 3 的路径中包含基因的不稳定性. 如 a 中路径所包含的 b、c、d 中的路径为路径提供了更多可选的不同的正确基因，对基因的作用及关系的预测具有较高的参考价值.

综上所述可知，使用蚁群算法对已知的具有同一生物过程的作用基因寻找基因路径的方法是有效的，能够一定程度的预测基因路径，对各种生物实验求解基因路径具有较高的指导意义.

4 结语

文中基于蚁群算法预测关联基因的路径，能够充分利用已有的基因情况的研究，定义预测路径的方向，最终得到较符合正确路径的预测结果. 未来的研究应该在不断提高准确率及稳定性的前提下，增加路径各基因的功能注释，并将基于序列的算法扩展到基于序列的二维乃至三维结构上.

参考文献:

1 Huang R, Wallqvist A, Thanki N, Covell DG. Linking pathway gene expressions to the growth inhibition response from the National Cancer Institute’s anticancer screen and drug mechanism of action. The Pharmacogenomics Journal, 2005, 5: 381–399.

2 周婷婷,容健锋,王正华,董蕴源,王勇献,朱云平.一种基于 KEGG 数据库的重构代谢网络的新方法.计算机工程与科学,2010,32(8):104–111.

3 李向真,刘子朋,李娟,方慧生.KEGG 数据库的进展及其在生物信息学中的应用.药物生物技术,2012,19(6):535–539.

4 李霞,李亦学,廖飞.生物信息学.北京:人民卫生出版社,2010

5 尹定子,宋海云.Wnt 信号通路:调控机理和生物学意义.中国细胞生物学学报,2011,33(2):103–111.

6 王震凯,朱人敏.Wnt 信号转导通路在肿瘤中的研究进展.医药研究生报,2007,20(12):1294–1301.

7 倪庆剑,邢汉承,张志政,王蓁蓁.蚁群算法及其应用研究进展.计算机应用与技术,2008,25(8):12–16.

8 蒋玲艳,张军,钟树鸿.蚁群算法的参数分析.计算机工程与应用,2007,43(20):31–36.

9 Dorigo M, Gambardella LM. Ant colony system: A cooperative

learning approach to the traveling salesman problem. IEEE Trans. on Evolutionary Computation, 1997, 1(1): 53–66.

10 詹士昌,徐婕,吴俊.蚁群算法中有关算法参数的最优选择. 科技通报,2003,19(5):381–386.

11 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol, 1990, 215: 403–412.

12 Gregory SG, Barlow KF, Mclay KE, et.al. The DNA sequence and biological annotation of human chromosome. Nature, 2006, 441(7091): 315–321.

13 Hillier LW, Graves TA, Fulton RS, et al. Generation and annotation of the DNA sequences of human chromosomes 2 and 4. Nature, 2005, 434(7034): 724–731.

14 Muzny DW, Scherer SE, Kaul R, et al. The DNA sequence, annotation and analysis of human chromosome 3. Nature, 2006, 440(7088): 1194–1198