P.O.Box 8718, Beijing 100080, China	Journal of Software, Feb 2006,17(2):232-241
E-mail: jos@iscas.ac.cn	ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
http://www.jos.org.cn	Copyright © 2006 by *Journal of Software*

基于二分频率变换的序列相似性查询处理技术

王国仁, 葛　健, 徐恒宇, 郑若石

Full-Text PDF　Submission　Back

王国仁, 葛　健, 徐恒宇, 郑若石
(东北大学　信息科学与工程学院,辽宁　沈阳　110004)
作者简介: 王国仁(1966－),男,湖北崇阳人,博士,教授,博士生导师,CCF高级会员,主要研究领域为XML数据管理,生物信息学,P2P数据管理,多媒体索引技术.葛健(1978－),男,硕士,主要研究领域为生物信息学.徐恒宇(1978－),男,硕士,主要研究领域为生物信息学.郑若石(1980－),男,硕士,主要研究领域为生物信息学.
联系人: 王国仁　Phn: +86-24-83681250, Fax: +86-24-23893138, E-mail: wanggr@mail.neu.edu.cn, http://mitt.neu.edu.cn

Received 2004-06-20; Accepted 2005-03-11

Abstract

As a main method for predicting the functionality of genes, the sequence similarity querying technique is becoming one of the research hotspots in bioinformatics. The similarity of gene sequence and structure usually determines the similarity of gene functionality, and the function of an unknown gene can be predicted by sequence similarity querying. After analyzing the advantages and shortcomings of related work such as frequency transformation and wavelet transformation used in MRS, a new sequence similarity query processing technique based on the two-Partitioning Frequency Transformation 2-PFT is proposed. Firstly, the Two-partitioning frequency transformation and the corresponding distance function are designed. They have a higher filtering ability than frequency transformation and wavelet transformation, and the system performance is thus improved significantly. Secondly, the problem of processing the queries with any length is solved. Theoretical proof and experimental results show that the 2-PFT system outperforms the MRS system greatly.

Wang GR, Ge J, Xu HY, Zheng RS. A sequence similarity query processing technique based on two-partitioning frequency transformation. *Journal of Software*, 2006,17(2):232-241.
DOI: 10.1360/jos170232
http://www.jos.org.cn/1000-9825/17/232.htm

摘要

作为基因功能预测的主要手段,序列相似性查询技术是生物信息学领域的研究热点.基因序列和结构的相似性往往决定了基因功能的相似性,因此可以通过基因序列的相似性查找来预测新基因的功能.分析了MRS索引中频率变化和小波变换等相关技术,讨论了它们的缺点和不足,提出了一种基于二分频率变换2-PFT的序列相似性查询处理技术.首先,设计了二分频率变换和相应的距离函数,使得系统较之频率变换和小波变换具有更高的过滤能力,极大地提高了系统的性能;其次,解决了处理任意长度查询的问题.理论证明和实验结果均表明,2-PFT系统的性能远远优于MRS系统.

基金项目：Supported by the National Natural Science Foundation of China under Grant No.60273079 (国家自然科学基金)

References:

[1] The human genome project (HGP). 2006. http://www.nhgri.nih.gov/

[2] National Center for Biotechnology Information. Genbank database. 2005. http://www.ncbi.nlm.nih.gov/

[3] Benson DA, Karsh-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL. Genbank. Nucleic Acids Research, 2000,28(1): 15-18.

[4] Gusfield D. Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology. Cambridge: Cambridge University Press, 1997.

[5] Myers E. An O(ND) difference algorithm and its variations. Algorithmica, 1986,1(2):251-266.

[6] Myers E. A sublinear algorithm for approximate keyword matching. Algorithmica, 1994,12(4-5):345-374.

[7] Baeza-Yates RA, Navarro G. Faster approximate string matching. Algorithmica, 1999,23(2):127-158.

[8] Kahveci T, Singh AK. An efficient index structure for string databases. In: Apers P, Atzeni P, Ceri S, Paraboschi S, Ramamohanarao K, Snodgrass R, eds. Proc. of the 27th Int'l Conf. on Very Large Data Bases (VLDB 2001). Roma: Morgan Kaufmann Publishers, 2001. 351-360.

[9] Sun H, Ozturk O, Ferhatosmanoglu H. CoMRI: A compressed muti-resolution index structure for sequence similarity queries. In: Peter M, Xu Y, ed. Proc. of the 2nd IEEE Computer Society Bioinformatics Conf. (CSB 2003). Califonia: IEEE Computer Society, 2003. 553-559.

[10] Ferhatosmanoglu H, Ozturk O. Effective indexing and filtering for similarity search in large biosequence databases. In: Jamil H, Megalooikonomou V, ed. Proc. of the 3rd IEEE Int'l Symp. on BioInformatics and BioEngineering (BIBE 2003). Washington DC: IEEE Computer Society, 2003. 359-366.