比较基因组学平台的设计与实现

刘 娜[1],郭云波[1],孙显赫[1],马 骊[2],邓亲恺[1](1南方医科大学基础医学院生物信息学研究室,2生物技术学院分子免疫学研究所,广东 广州 510515)

摘要:目的 开发一个基于 Web 的本地服务器支持的功能全面的基因组比较和可视化平台,以加快对基因组的分析。**方法** 构建以 Apache HTTP 服务器为平台的 WEB 服务器,采用 Perl 语言编程,优选整合了 MUMmer、LAGAN、Mauve 等多个基因组学研究的软件和算法,针对生物学家不同的应用目的,可直接在网页上提交基因组序列数据和参数选项,经平台处理的结果通过网页以图形化的方式返回用户。**结果** 该平台可以处理多种序列数据输入格式,实现了完整的基因组序列和草图基因组序列比对,近远源物种的双基因组或多基因组比对,基因组间的同线性区域寻找,以及定位大范围的基因组重组(基因插入/缺失、重复、重排和水平转移)和小的核苷酸突变等功能;并将结果以图形化的方式显示。应用本平台,对 10 种新型甲型流感病毒株作基因组同源性分析,表明 PB1 基因可能来自于人 H3N2,PB2、PA 基因可能来自于禽类 H3N2,而 HA、NS 基因可能来自于猪 H1N1。在本平台上还对结核分枝杆菌(H37Rv、CDC1551)和牛分支杆菌(AF2122/97)基因组的研究分析,发现插入/缺失和重复序列是导致三个菌株基因组差异的主要来源。**结论** 该平台功能资源整合较全面,应用界面友好,结果显示直观,故可作为比较基因组学常规研究的有效平台。

关键词:比较基因组学;基因组比对可视化;同线性;基因组重组;生物信息学;甲型流感病毒;结核分枝杆菌
中图分类号:R34 **文献标识码:**A **文章编号:**1673-4254(2010)02-0219-05

Design and construction of the platform for comparative genomics

LIU Na[1], GUO Yun-bo[1], SUN Xian-he[1], MA Li[2], DENG Qin-kai[1]
[1]Institute of Bioinformatics, School of Basic Medical Science, [2]Research Institute of Molecular Immunology, College of Biotechnology, Southern Medical University, Guangzhou 510515, China

Abstract: Objective To design a versatile genome comparison and visualization platform based on browser/server mode supported by a local server. **Methods** The server of the platform was Apache HTTP server. Perl was used to integrate such genome alignment package and algorithms as MUMmer, LAGAN, and Mauve for different comparison purposes, and the users could submit data and parameters to the platform via the web page. The results of analysis were also returned via the web page. **Results** The platform could handle multiple data input formats, compare complete and draft genome sequence, alignment pair-wise or multi genome of more divergent species, identify regions of high similarity, locate local nucleotide mutations and large-scale recombination, and display the results by visualization interface. Analysis of the homology of 10 new strains of influenza A virus indicated that PB1 gene might evolve from human H3N2 viruses, PB2 and PA genes from avian H3N2 viruses, and HA and NS genes from swine H1N1 viruses. Alignment of *Mycobacterium tuberculosis* (H37Rv, CDC1551) and *Mycobacterium bovis* (AF2122/97) genomes revealed that sequence insertion/deletion and duplication were the major source of genomic differences. **Conclusion** The platform integrate comprehensive resources with a user-friendly interface and intuitive result visualization to facilitate conventional study of comparative genomics.

Key words: comparative genomics; genomes alignment visualization; synteny; genome recombinant; bioinformatics; influenza A virus; *mycobacterium tuberculosis*

近年来,随着高速测序技术的迅猛发展和众多物种的全基因组测序计划的实施,基因组数据大量产出,呈海量增长趋势。大规模的全基因组数据的功能分析需要新的算法、软件和强大的计算平台的支持。

全基因组比对往往是进行基因组分析的第一步,通过不同亲缘关系物种的基因序列比对,能够鉴定出编码序列、非编码调控序列以及给定物种的独有序列,可以了解不同物种的核苷酸组成和基因顺序方面的异同,进而可以揭示基因潜在的功能,阐明物种进化关系及基因组的内在结构。比较基因组学已成为一门崭新的学科,在基因组分析,药物发现,以及人类疾病和物种进化研究中发挥着越来越重要的作用[1-3]。

目前研究人员开发了许多进行比较基因组学分析的算法和软件包。以识别保守区域和重组事件的同线性比对工具有:MUMmer [4-6]、LAGAN 系列[7-8]、BLASTZ [9]、AVID/MAVID [10-11]、MGA [12]、Mauve [13]、CGAT[14]、CoCoNUT[15]等。以开发基因组比对和可视

收稿日期:2009-08-13
基金项目:传染病防治国家重大科技专项(2008ZX10003013-02)
作者简介:刘 娜(1984-),女,E-mail: liuna520@fimmu.com;郭云波(1971-),男,在读博士研究生,刘娜和郭云波为共同第一作者
通讯作者:邓亲恺,男,教授,E-mail: dqk001@fimmu.com

化的比较基因组工具有：ACT [16]、COMPAM [17]、GenAlyzer[18]、GenomeComp[19]等。此外也有许多比较基因组学分析的网络服务资源：VISTA、ECRbrowser、ZPicture、PLATCOM、MBGD、CoGe、CFGP、CCG、Compagen、IMG 等。

但是上述软件均存在着一定的局限性：软件的功能比较单一，每个软件都有各自特殊的数据输入输出格式；不同的软件采用不同的算法，侧重点不同；有的只能在特定的操作系统下运行。此外有些软件的设置参数较多，一般生物学家往往不易熟悉，很难选择，结果同样的序列用不同的软件得到的结果也不同；而特别值得指出的是，一个全基因组数据往往非常庞大，尤其是进行多重基因组比对时，需要耗费大量的计算时间和存储空间，个人计算机往往不能满足要求[20-21]。

因此有必要开发一个具有高性能计算能力，操作界面友好，功能较全面的比较基因组学服务平台，以方便生物学家在远程个人计算机上进行有效的操控，实施全基因组分析。本平台采用 Perl 语言编程，整合了一批功能较好的基因组比较分析软件，在基于 Linux 操作系统和浏览器 / 服务器(B/S)模式的高性能计算机(HPC)环境下工作。

1 平台的构建与设计

1.1 基本构架

本平台采用浏览器 / 服务器(Browser/Server，B/S）网络结构，用户可以在个人计算机上通过 web 浏览器，将基因组数据提交到相应的 Web 服务器，同时选择参数，服务器进行分析和处理后，将结果返回到用户浏览器。平台采用了全球使用最为广泛的 Apache HTTP 服务器，在基于 Linux 操作系统的高性能计算机环境下运行，保证了平台的数据安全和运算速度。平台的数据管理用 MySQL 数据库。

1.2 实现方法

平台设计选用 Perl 语言，一种源代码开放且功能强大的编程语言。由于基因组数据的输入文件大多是文本文件，而 Perl 语言具有强大的正则表达式，非常适合操作文本文件；Perl 还擅长 Web 编程、数据库处理、XML 处理、图像处理等。此外还应用了 PHP 语言编写部分动态网页。

在数据分析方法上，本平台整合了多个基因组分析比对工具，包括 MUMmer、LAGAN 系列软件和 Mauve 来进行基因组比较分析。

本平台旨在为研究者提供合适的基因组比对、分析和结果显示工具，通过对不同物种的基因组进行比较分析，揭示彼此的相似性和差异性，以了解不同物种间进化上的差异。平台的数据流处理与主要功能见图 1。

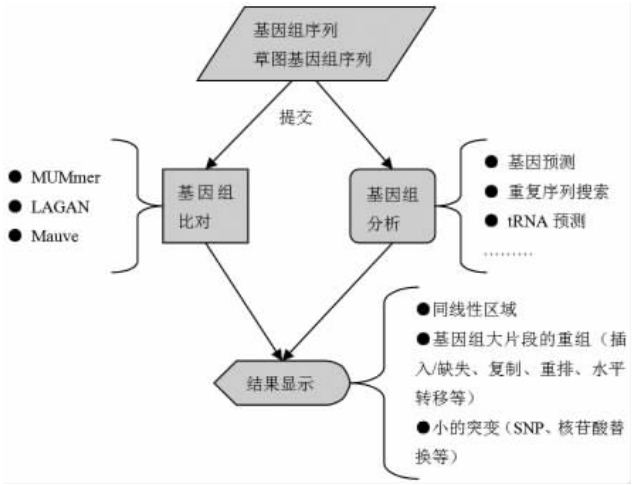


图 1 平台数据流处理及主要功能

2 结果

2.1 平台的功能

目前平台的主要功能已经实现，可以处理 fasta、multi-fasta、Genbank/GB、EMBL 等文件格式的数据和用户直接提交的序列。平台整合了多种算法，涉及到基因组比对的各种情况。分析结果将以网页的方式动态显示。

2.1.1 全基因组序列的比对与分析 基因组比对的方法很多，根据要比对的基因组个数，基因组比对可以划分为双基因组比对和多重基因组比对。根据比对物种的亲缘关系远近，可以分为：近源或近距离物种间比对，中等距离间物种比对和远源或远距离物种间比对。不同的基因组个数、不同亲缘关系的物种比较分析的侧重点不同[22-23]。为了尽可能多地满足研究者各种研究的需求，提高数据分析的质量，本平台整合了 MUMmer、LAGAN 和 Mauve 等软件。MUMmer 是一个可以快速比对两个基因组序列的软件，它采用后缀树算法快速地寻找两序列间的最大精确匹配片段，可以精确的识别出基因组间的插入 / 删除、重复、串联重复、小的突变区域和 SNPs（单核苷酸多态性）。LAGAN 软件包包括 LAGAN、S-LAGAN 和 MLAGAN 三种算法。LAGAN 算法是专门针对相对远源物种比对开发的一个有效可靠的全局比对算法，它采用 CHAOS 局部比对程序查找局部同源序列，允许错配，有利于比对非编码区域。S-LAGAN 是 LAGAN 算法的改进，在比对中考虑了序列的方向，可以检测到基因组间的重组事件。MLAGA 算法是多重基因组比对算法。Mauve 软件包包括 MauveAligner 和 ProgressiveMauve 两个算法。MauveAligner 可以有效的识别多重基因组中的保守基因组区域，重新排列区域，保守区域中的插入序列和精确的断点等。此外还可以识别核苷酸替换，小片段的插入和缺失。但是

MauveAligner 算法只能识别在所有比对基因组中都存在的保守序列,存在于部分物种中的保守序列不能被识别,所以适合比对近源物种。与 mauveAligner 相比 ProgressiveMauve 算法可以检测到只在部分基因组中存在的同源序列,可以比对分化距离较远的序列。研究者可以根据不同的比对目的选择适合的算法。

除了基因组比对，平台还提供一些扩展功能,如序列格式转换,基因组数据库的链接,基因组分析如基因组重复序列搜索,基因预测,tRNA 预测等。

2.1.2 基因组比对可视化　基因组比对的结果通过网页以图形化的方式在用户端显示(图 2)。由于平台运用了多个基因组比较分析算法,每个算法都有各自的输出文件格式,为此本平台把每次比对的结果都转换成直观的图形,通过网页显示出来。显示的内容包括同线性区域，基因组大片段的重组，如片段的插入 / 删除,重复和翻转,基因组中一些小的突变,如 SNP,核苷酸替换等。在结果显示网页用户还可以根据自己的需求对图像进行放大、缩小、平移。

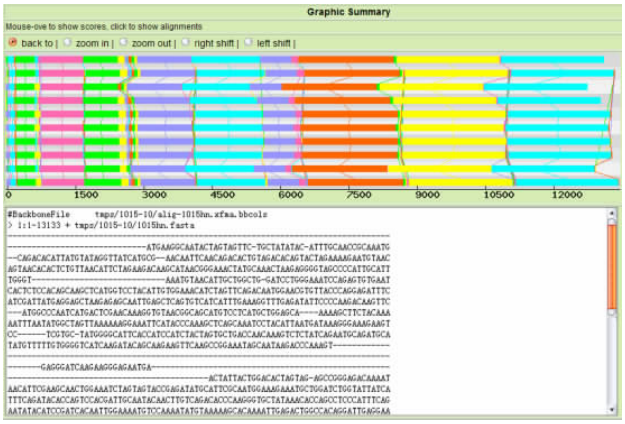


图 2 基因组比对结果显示界面

2.2 案例研究

为了对所构建的比较基因组学平台的可操作性、数据分析的有效性进行验证,选取了两组基因组数据在本平台上进行分析，即：针对新型甲型流感病毒 A/H1N1 的同源性分析，以及结核分枝杆菌(H37Rv、CDC1551)和牛分支杆菌(AF2122/97)基因组比较分析。

2.2.1 新型甲型流感病毒 A/H1N1 同源性分析　以 2009 年流行的新型甲型流感病毒 A/H1N1 病毒株为参考序列,分别与不同地区、不同时期、不同亚型和不同宿主的甲型流感病毒 9 种全基因组参考序列进行比对。首先从 NCBI 上下载病毒的全基因组参考序列(表 1),每个株病毒的 8 个基因按 HA、MP、NA、NP、NS、PA、PB1、PB2 顺序以 mulit-fasta 格式保存。采用本平台中的 nucmer 算法，通过图 3 所示的界面提交数据。经分析(表 2)发现:PB1 基因来自于人 H3N2,同源性大于 91%;NP 基因与人类波士顿 H3N2 同源性最高为 92.32%;PB2、PA 与禽类 H3N2 同源性较高,分别为 89.91%、90.38%;HA、NS 与猪 H1N1 同源性较高，分别为 89.42%、91.89%;MP 与人 H5N1、猪 H3N2、禽类 H5N1、禽类 H3N2 同源性均很高分别为 90.84%、95.08%、91.27%、91.27%;NA 与其它病毒株同源性都不高。

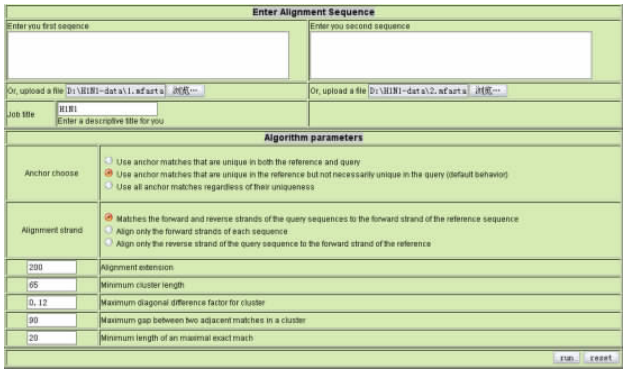


图 3 新型甲型流感病毒 A/H1N1 同源性分析数据提交和参数设置界面

表 1 用于比对分析的 10 种甲型流感病毒株的全基因组序列列表

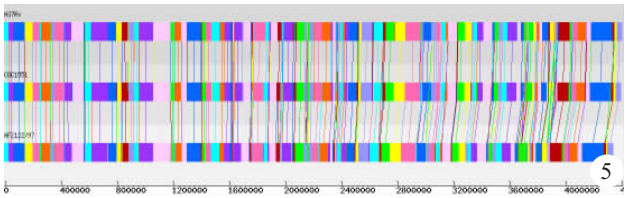
序列名称	GenBank accession number
A/Sichuan/1/2009(H1N1)	GQ 166223-GQ166230
A/Kentucky/UR06-0007/2006(H1N1)	CY028195-CY028202
A/Guangdong/03/2005(H3N2)	EU604298-EU604305
A/Jiangsu/1/2007(H5N1)	EU434686-EU434693
A/Memphis/104/1976(H3N2)	CY022309-CY022316
A/Boston/5/2008(H3N2)	CY044461-CY044468
A/swine/Tennessee/2/1978(H1N1)	CY027307-CY027314
A/swine/Spain/33601/2001(H3N2)	CY009372-CY009379
A/Cygnus olor/Germany/R1372/2007(H5N1)	FM165519-FM165526
A/green-winged teal/Minnesota/Sg-00131/2007(H3N2)	CY041884- CY041891

2.2.2 结核分枝杆菌杆菌(H37Rv、CDC1551)和牛分支杆菌 (AF2122/97) 基因组比较　结核分枝杆菌(*Mycobacterium tuberculosis*)H37Rv (NC_000962)和 CDC1551(NC_002755)是人类肺结核治病菌的典型菌株。而牛分支杆菌 (Mycobacterium bovis) AF2122/97(NC_002755)是多种动物的有效致病菌。三个菌株(H37Rv、CDC1551、AF2122/97)的基因组比对采用本平台中的 progressiveMauve 算法,重复序列搜索采用本平台中的 repseek 软件。按图 4 所示提交基因组序列并设置参数。经比对发现:三个基因组极其相似(>99.9%),基因组之间存在整体共线性,几乎没有出现重组(图 5)。但在 H37Rv 基因组中存在大量的插入元件,AF2122/97 基因组有部分基因缺失。与 H37Rv 和 AF2122/97 基因组相比,CDC1551 基因组的基因插入 / 缺失较少(图 6)。此外,三个基因组都

表 2 新型甲型流感病毒 A/H1N1 全基因组序列与流感病毒参考序列的同源性比较结果

病毒株	与 A/Sichuan/1/2009(H1N1)同源性(%)							
	PB2	PB1	PA	HA	NP	NA	MP	NS
Kentucky/UR06-0007/2006(H1N1)	83.50	-	-	-	84.58	-	87.28	-
Guangdong/03/2005(H3N2)	84.25	93.62	83.25	-	83.33	-	86.82	-
Jiangsu/1/2007(H5N1)	84.74	84.91	89.12	-	83.12	87.13	90.84	-
Memphis/104/1976(H3N2)	84.42	91.20	83.85	-	84.55	-	88.39	-
Boston/5/2008(H3N2)	84.03	93.27	83.15	-	92.32	-	86.56	-
swine/Tennessee/2/1978(H1N1)	83.98		82.44	89.42	84.19	-	88.59	91.89
swine/Spain/33601/2001(H3N2)	83.64	85.29	86.16	-	84.12	-	95.08	80.48
Cygnus olor/Germany/R1372/2007(H5N1)	84.69	85.49	88.36	-	82.80	86.34	91.27	85.05
green-winged teal/Minnesota/Sg-00131/2007(H3N2)	89.91	86.98	90.38	-	83.32	-	91.27	84.73

存在不同程度的重复序列。H37Rv 的重复序列最多，并且有将近一半的重复序列长度大于 1000 bp。其它两个基因组重复序列的长度主要集中在 50~500 bp 之间(表 3)。这些重复序列的差异使相应的蛋白质之间差异极大。由以上分析可以推断插入 / 缺失和重复序列是导致这些基因组特异性的主要来源。对这些片段进行分析有可能发现重要的基因或结构，从而为结核杆菌抗药性的研究和疫苗的开发提供线索。



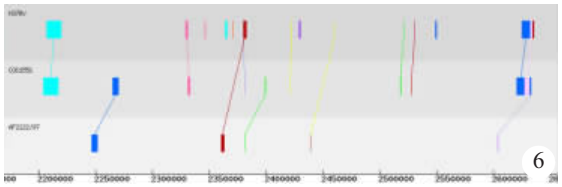


图 4 结核分枝杆菌 H37Rv、CDC1551 及牛分支杆菌 AF2122/97 基因组序列提交和参数设置界面

图 5 H37Rv、CDC1551、AF2122/97 基因组同线性分析结果

图 6 H37Rv、CDC1551、AF2122/97 基因组的插入 / 缺失片段分析

表 3 H37Rv、CDC1551、AF2122/97 基因组重复序列搜寻结果

序列	>50 bp		>100 bp		>500 bp		>1000 bp	
	Direct	Inverted	Direct	Inverted	Direct	Inverted	Direct	Inverted
H37Rv	182	92	128	83	77	75	65	66
CDC1551	126	38	69	27	23	18	9	9
AF2122/97	123	31	69	21	20	16	8	8

3 讨论

为适应大规模的全基因组数据的功能分析需求，构建了基于 Linux 操作系统和 Apache HTTP 服务器的高性能计算机平台，平台有效整合了 MUMmer、LAGAN 和 Mauve 等基因组比对分析工具和算法，能处理多种输入数据格式，比对完整的基因组和 draft 基因组，由此可寻找基因组同线性区域，定位大范围的基因组重组(基因插入 / 缺失、重复、重排和水平转移)和小的核苷酸突变等。利用本平台对新型甲型流感病毒 A/H1N1 同源性分析，推断 PB1 基因来自于人 H3N2，PB2、PA 基因来自于禽类 H3N2，HA、NS 基因来自于猪 H1N1，这些结论与相关文献中的报道相符[24-25]。利用本平台对三个分支杆菌的基因组比对分析，表明三个基因组具有很高的相似性，每个基因组都存在不同程度的插入 / 缺失和重复序列，这些差异是种间和种内多样性的主要来源。平台友好的 B/S 界面、动态化图形的结果显示和强大的后台计算能力给生命科学工作者从事比较基因组学研究提供了较大的方便。目前，本平台正在局域网内测试，运行状态良好。

参考文献：

[1] Shi WJ, Zhang XL, Wang HH. Comparative genomics and anti-tuberculosis drug target discovery[J]. Chin J Tuberculosis Respire Dis, 2008, 31(1): 54-7.

[2] Moreno C, Lazar J, Jacob HJ, et al. Comparative genomics for detecting human disease genes[J]. Adv Genet, 2008, 60: 655-97.

[3] Nobrega MA, Pennacchio LA. Comparative genomic analysis as a tool for biological discovery[J]. Physiol Soc, 2004, 554(Pt1): 31-9.

[4] Kurtz SF, Phillippy A, Delcher AL, et al. Versatile and open software for comparing large genomes[J]. Genome Biol, 2004, 5: R12.

[5] Delcher AL, Phillippy A, Carlton J, et al. Fast algorithms for large-scale genome alignment and comparison[J]. Nucl Acids Res, 2002, 30(11): 2478-83.
[6] Delcher AL, Kasif S, Fleischmann RD, et al. Alignment of whole genomes [J]. Nucl Acids Res, 1999, 27(11): 2369-76.
[7] Brudno M, Do CB, Cooper GM, et al. LAGAN and Multi-LAGAN : efficient tools for large-scale multiple alignment of genomic DNA [J]. Genome Res, 2003, 13(4):721-31.
[8] Brudno M, Malde SK, Poilakov AL, et al. Global alignment: finding rearrangements during alignment [J]. Bioinformatics, 2003, 19: 54i-62i.
[9] Schwartz S, Kent WJ, Smit A, et al. Human-mouse alignments with BLASTZ[J]. Genome Res, 2003, 13: 103-7.
[10] Bray N, Dubchak I, Pachter L. AVID: A global alignment program [J]. Genome Res, 2003, 13(1): 97-102.
[11] Bray N, Pachter L. MAVID: Constrained ancestral alignment of multiple sequences[J]. Genome Res, 2004, 14: 693-9.
[12] Höhl M, Kurtz S, Ohlebusch E. Efficient multiple genome alignment [J]. Bioinformatics, 2002, 18: S312-20.
[13] Darling AC, Mau B, Blattner FR, et al. Mauve: multiple alignment of conserved genome research with rearrangements genomes [J]. Genome Res, 2004, 14(7): 1394-403.
[14] Uchiyama I, Higuchi T, Kobayashi I. CGAT: a comparative genome analysis tool for visualizing alignments in the analysis of complex evolutionary changes between closely related genomes [J]. BMC Bioinformatics, 2006, 7: 472.
[15] Abouelhoda MI, Kurtz S, Ohlebusch E. CoCoNUT: an efficient system for the comparison and analysis of genomes [J]. BMC Bioinformatics, 2008, 9: 476.
[16] Carver TJ, Rutherford KM, Berriman KM, et al. ACT: the Artemis Comparison Tool[J]. Bioinformatics, 2005, 21: 3422-3.
[17] Lee D, Choi JH, Dalkilic MM, et al. COMPAM: visualization of combining pairwise alignments for multiple genomes [J]. Bioinformatics, 2006, 22(2): 242-4.
[18] Choudhuri JV, Schleiermacher C, Kurtz S, et al. Genalyzer: interactive visualization of sequence similarities between entire genomes[J]. Bioinformatics, 2004, 20: 1964-5.
[19] Yang J, Wang JH, Yao ZJ, et al. GenomeComp: a visualization tool for microbial genome comparison[J]. J Microbiol Methods, 2003, 54(3): 423-6.
[20] Batzoglou S. The many faces of sequence alignment[J]. Brief Bioinformatics, 2005, 6(1): 6-22.
[21] Frazer KA, Elnitski L, Church DM, et al. Cross-species sequence comparisons: a review of methods and available resources [J]. Genome Res, 2003, 13(1): 1-12.
[22] Margulies EH, Chen CW, Green ED. Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons [J]. Sciences, 2006, 4(22): 187-93.
[23] Hardison RC. Comparative genomics[J]. Plos Biol, 2004, 2(1): 156-60.
[24] Michaelis M, Doerr HW, Cinatl Jr J. Novel swine-origin influenza A virus in humans: another pandemic knocking at the door[J]. Med Microbiol Immunol, 2009, 198(3): 175-83.
[25] Smith JD, Vijaykrishna D, Bath J, et al. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic[J]. Nature, 2009, 459: 1122-5.

(上接 218 页)

the potential role of immune regulatory cells[J]. Expert Opin Biol Ther, 2004, 4(9): 1387-93.
[14] Daugherty A, Rateri DL. T lymphocytes in atherosclerosis: the yin-yang of th1 and th2 influence on lesion formation[J]. Circ Res, 2002, 90(10): 1039-40.
[15] Fontenot JD, Gavin MA, Rudensky AY. Foxp3 programs the development and function of CD4+CD25+ regulatory T cells [J]. Nat Immunol, 2003, 4(4): 330-6.
[16] Maloy KJ, Salaun L, Cahill R, et al. CD4 +CD25 + T(R) cells suppress innate immune pathology through cytokine-dependent mechanisms[J]. J Exp Med, 2003, 197(1): 111-9.
[17] 眭维国, 孙燕燕, 黄 河, 等. 慢性肾功能不全患者外周血调节性 T 细胞的表达[J].中国免疫学杂志, 2008, 24(7): 652-4.
[18] Meier P, Meier R, Blanc E. Influence of CD4+/CD25+ regulatory T cells on atherogenesis in patients with end-stage kidney disease [J]. Expert Rev Cardiovasc Ther, 2008, 6(7): 987-97.
[19] Meier P, Golshayan D, Blanc E, et al. Oxidized LDL modulates apoptosis of regulatory T cells in patients with ESRD[J]. J Am Soc Nephrol, 2009, 20(6): 1368-84.