

应用系统发育分析方法鉴别近缘病原菌

曹清毅, 侯晓丽, 吴昊, 陈智

(浙江大学医学院附属第一医院传染病研究所, 浙江 杭州 310003)

[摘要] 目的:以 *gyrB* 基因序列为基础,采用系统发育分析方法构建进化树,对种系关系密切的病原菌进行区分鉴别。方法:测序获得本地区19株大肠埃希菌、13株志贺菌、2株豚鼠气单胞菌、2株嗜水气单胞菌、1株维隆气单胞菌的 *gyrB* 基因序列,并将其同公共数据库已有序列合并,构建进化树。分析各序列在树中的聚类关系,以此对菌株进行种水平分类鉴别,并同BLAST查询结果相比较。结果:除鲍氏和痢疾志贺菌外,对检测的所有序列,在进化树上与之最邻近的5条序列均为该种菌株。而对一部分菌株,其BLAST查询结果中含有其它菌种的或分类关系不明确的序列。结论:以系统发育方法对种系发生关系密切的病原菌序列进行鉴别,具有较好的分辨力和准确度。

[关键词] 细菌; 基因; 鉴别; 系统发育树; *gyrB* 基因; BLAST; 序列分析

[中图分类号] R 372 **[文献标识码]** A **[文章编号]** 1008-9292(2007)06-0531-06

Identification of closely related bacteria via phylogenetic methods

CAO Qing-yi, HOU Xiao-li, WU Hao, et al (Institute of Infectious Diseases, The First Affiliated Hospital, College of Medicine, Zhejiang University, Hangzhou 310003, China)

[Abstract] **Objective:** To differentiate closely related pathogenic bacteria via phylogenetic method on the basis of *gyrB* gene sequences. **Methods:** *CyrB* sequences of 19 strains of *E. coli*, 13 *Shigella* spp., 2 *Aeromonas caviae*, 2 *Aeromonas hydrophilia*, 1 *Aeromonas veronii* were determined and combined with sequences retrieved from public databases to construct phylogenetic trees. For each sequence tested, the identification deduced from the clustering relation of sequences was compared with the phenetic identification. **Results:** All the tested sequences, except those of *Shigella boydii* and *Shigella dysenteriae*, were corresponded with the 5 closest sequences on the tree at the species level. While the BLAST queries returned some other bacteria organisms or undetermined entries. **Conclusion:** Phylogenetics displays good discriminative power in bacterial sequences differentiation.

[Key words] Bacteria; Genes; Identification; Phylogeny; *gyrB*; BLAST; Sequence analysis

[J Zhejiang Univ (Medical Sci), 2007, 36(6):531-536.]

收稿日期: 2007-07-16 修回日期: 2007-08-23

基金项目: 浙江省科学技术厅重点科研项目(2003C13015, 021103128)。

作者简介: 曹清毅(1982—), 男, 博士研究生。

通讯作者: 陈智(1956—), 男, 教授, 从事病毒性肝炎研究; E-mail: chenzhi@zju.edu.cn

临床常见的感染性疾病涉及的病原菌范围广泛、种类繁多,快速准确地鉴定细菌对临床诊断与治疗有重要的指导作用。常规的病原学诊断方法包括细菌培养、分离及生化反应鉴定,耗时久,一般需24~72 h才能获得较确切的结果,难以及时指导疾病的治疗。另外,对于难以培养的,或者种系发生关系密切的近缘菌种(其生化反应差异模糊),很难依靠传统方法鉴定。

近年来,随着分子生物学和生物信息学技术的发展,病原菌分子诊断技术逐渐成熟。其具有灵敏度高、检测快捷、分辨率好、结果明确的特点。典型的分子诊断方法包括限制性片段长度多态性(Restriction fragment length polymorphisms, RFLP),定量PCR,基因芯片(MicroArray),DNA序列测序分析^[1]等。不同物种间,核苷酸碱基排列顺序的差异是这些分析方法的本质依据,因此,精确的序列测定是分子诊断方法的金标准,其结果明确而又唯一,具有最大的信息量。

取得待测菌株DNA序列后,如何做分类学鉴定?原核生物的系统发育关系极其庞杂;传统分类学中定义的科、属、种、亚种等分类单位,更多地反映了细菌外在的生化表型特征和临床显著性,很难与其内在的基因型得到一一对应。对此,除了选择能尽可能准确地反映绝大多数菌种系统发育关系的基因作为检测靶分子外,也需要根据问题的性质选择合适的生物信息学工具,对目的序列和参考序列进行比对分析。

一般为人们所惯用的方法是将测序获得的序列输入NCBI网站的BLAST服务器,根据返回序列中的相似性最显著者推断查询序列所属物种。这种方法较费时费力,此外,受程序原理所限,BLAST对一些近缘物种(同一属内各种间及种以下水平)序列的分辨力不足。在本研究中,我们以细菌鉴定中常用的*gyrB*基因为靶分子,将公共数据库序列与待检序列合并构建系统发育树^[2],根据待检序列在所构树上的位置及其与近缘种系距离,分别对大肠埃希菌—志贺菌属复合体,气单胞菌属的共37株临床株进行了鉴别。

1 材料与方法

1.1 细菌菌株来源

5个标准菌株:鲍氏志贺

菌(NICPBP51315, 51318与51519),痢疾志贺菌(NICPBP51252与512555),为浙江大学传染病研究所保存菌株。临床分离株:所有的临床菌株于2004年4月至2005年8月从浙江大学附属第一医院收集获得。所有的菌株事先都经临床微生物实验室用生化反应或Vitek-AMS系统(BioMerieux, France)确认。

1.2 基因组DNA提取 细菌接种于3 ml液体培养基中,37℃摇床(300 r/min)培养过夜。取100 μl菌体培养物于一灭菌Ep管中,15 000×g离心2 min,弃上清液,收集菌体。加入100 μl超纯灭菌水,15 000×g离心5 min,弃上清(此步骤是为了去除培养基中可能存在的PCR扩增抑制成分),再加入100 μl超纯灭菌水重悬混匀,100℃煮沸10 min,15 000×g离心5 min,收集上清液作为PCR扩增细菌DNA模板,置4℃保存备用。

1.3 *gyrB*基因扩增与测序 PCR反应总体积为50 μl,内含2 μl(1:10)稀释的菌株DNA模板,5 μl 10×PCR缓冲液,200 μmol/L dNTP,1.25 U Pyrobest Taq酶,12 pmol引物。扩增条件:35个循环(96℃变性1 min,62℃退火1 min,72℃延伸2 min),然后72℃终末延伸7 min。PCR产物约1.2 kb,经割胶纯化后,由上海英骏生物技术有限公司测序。对含有该菌株特异性突变位点的序列,首先检查测序波峰图,仍难以确定者行二次测序以资确定。

1.4 网络数据库序列获取 从Genbank及ICB数据库^[3],采集截止于2006年7月的所有大肠埃希菌、志贺菌属、气单胞菌属的*gyrB*基因序列,用于构建系统发育树。

1.5 序列获取与处理中的质量控制及多序列联配 对于网络数据库采集序列,首先去除长度过短的,以及没有明确“种”水平分类信息的条目;再通过序列初步比对,剔除由插入/删除造成读码框偏移,以及含有明显低测序质量区的序列。使用Lasergene软件包中的Megalign组件,应用Clustal W^[4]算法,对待检临床株序列合并网络数据库所采集的序列进行多序列对齐排列。根据联配结果行进一步校正:为了减少计算量,对多条完全相同的序列只保留一条作为代表;最后在头尾将联配结果裁减对齐,并使起

始位置对准该基因编码区读码框。剪切后的序列长为1125个联配位点(position),起始于大肠埃希菌K12菌株(ATCC8735,GenBank 序列号U00096)gyrB基因编码区的第346位核苷酸。考虑到该段基因序列是连续编码区,为使联配结果具有较好的生物学意义,裁减后的序列先按照标准遗传密码子表,翻译为相应的氨基酸序列,同样以ClustalW算法联配,获得氨基酸序列,算法相关参数中插入空位罚分值(gap opening penalty)设为35,空位延伸罚分(gap extension penalty)设为0.75^[6],再将氨基酸联配结果原位转回核苷酸序列用于构建系统发育树。如此获得的核苷酸联配结果中的所有空位(gap),都以三联体形式存在于密码子之间而非插入密码子内^[5]。这样就得到了将用于系统发育分析的最终核苷酸序列联配结果。

1.6 系统发育树构建与处理 使用PHYLP软件3.65版^[6],以距离法(JC69模型构建距离矩阵,Neighbor-joining法聚类)构建系统发育树。距离法中所需设置的具体参数(如Ts/Tv,alpha等)通过PAML软件包3.14版获得。解析树中各序列间距离的程序通过PERL程序语言

自行编写。以系统发育树上与待检序列相距最近的5条序列作为鉴定依据。

1.7 公共数据库BLAST查询 采用NCBI网站提供的BLAST服务(<http://www.ncbi.nlm.nih.gov/BLAST/>)^[7],对Genbank数据库进行查询,采用BLASTN方法,Database设为Others。其余参数为默认值。以返回的前5条结果作为鉴定依据。

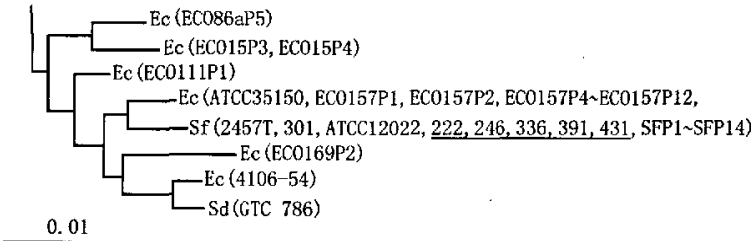
2 结 果

2.1 大肠埃希菌-志贺菌复合体系统分析与BLAST查询结果比较 如表1所示,所测的19株大肠埃希菌临床株中共有9种不同的gyrB序列,在系统发育树上与之最临近的5条序列都是大肠埃希菌的。在BLAST结果中,除79号外,对各株序列查询返回的前5条序列也均为大肠埃希菌的。对福氏、宋氏志贺菌,两种方法返回结果均为同种菌株序列。鲍氏、痢疾志贺菌在系统发育树上部分混杂在一起;BLAST对两者返回的5条最近缘序列中均有一大肠埃希菌条目。作为示例,将福氏志贺菌在gyrB基因系统发育树中的聚类情况列出,如图1所示。

表1 比较大肠埃希菌与志贺菌分离株序列BLAST查询结果与系统发育分析结果
Table 1 Comparison of phylogenetic analysis and BLAST query results for *E.coli* and *Shigella* spp. isolates

菌 株	系统发育树中最近缘的5条序列	BLAST返回的前5条序列
Ec(147,163,261,346,350,352,356,410)	Ec(5/5)	Ec(5/5)
Ec(178)	Ec(5/5)	Ec(5/5)
Ec(179)	Ec(5/5)	Ec(5/5)
Ec(152,182,183)	Ec(5/5)	Ec(5/5)
Ec(167)	Ec(5/5)	Ec(5/5)
Ec(49)	Ec(5/5)	Ec(5/5)
Ec(58,488)	Ec(5/5)	Ec(5/5)
Ec(67)	Ec(5/5)	Ec(5/5)
Ec(79)	Ec(5/5)	Ec(4/5),Sb(1/5)
Sb(51315,51318,51519)	Sb(4/5),Sd(1/5)	Sb(4/5),Ss(1/5)
Sd(51252)	Sb(4/5),Sd(1/5)	Sb(4/5),Ss(1/5)
Sd(51255)	Sb(4/5),Sd(1/5)	Sb(4/5),Ss(1/5)
Sf(222,246,336,391,431)	Sf(5/5)	Sf(5/5)
Ss(404,447,463)	Ss(5/5)	Ss(5/5)

Ec表示大肠埃希菌;Sf表示福氏志贺菌;Ss表示宋氏志贺菌;Sb表示鲍氏志贺菌;Sd表示痢疾志贺菌



Ec 表示大肠埃希菌;Sf 表示福氏志贺菌;Sd 表示痢疾志贺菌。本研究中测序的临床分离株用下划线标识

图 1 大肠埃希菌和志贺菌的系统发育树局部

Fig. 1 Partial phylogenetic tree for *E. coli* and *Shigella* spp.

2.2 气单胞菌属系统分析与 BLAST 查询结果比较 如表 2 所示,系统发育树中,对于 5 株豚鼠、维隆及嗜水气单胞菌分离株,最近邻的 5 株序列均为其所属种。而在 BLAST 结果中,返回的与 2 株豚鼠气单胞菌最近缘的序列均为嗜水气单胞菌。对嗜水、维隆气单胞菌返回结果中均有所属种不确定的单胞菌属菌株(*Aeromonas* spp.)。所涉及的嗜水气单胞菌分离株在 *gyrB* B 基因系统发育树中的聚类情况如图 2 所示。

表 2 比较嗜水、豚鼠、维隆气单胞菌分离株序列 BLAST 查询结果与系统发育分析结果

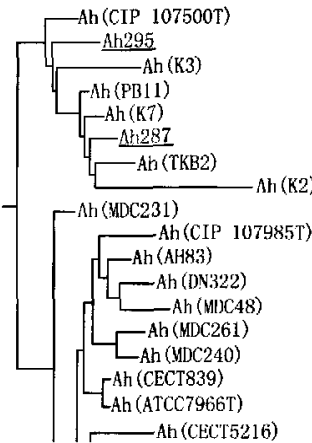
Table 2 Comparison of phylogenetic analysis and BLAST query results for *Aeromonas* spp. isolates

菌 株	系统发育树中最 近缘的 5 条序列	BLAST 返回的 前 5 条序列
Ac(8)	Ac(5/5)	Ah(1/5), N/A(2/5), Ac(2/5)
Ac(425)	Ac(5/5)	Ah(1/5), Ac(4/5)
Ah(287)	Ah(5/5)	Ah(4/5), N/A(1/5)
Ah(295)	Ah(5/5)	Ah(4/5), N/A(1/5)
Av(453)	Av(5/5)	Av(3/5), N/A(2/5)

Ac 表示豚鼠气单胞菌, Ah 表示嗜水气单胞菌, Av 表示维隆气单胞菌, N/A 表示仅确认为气单胞菌属菌株

3 讨 论

3.1 鉴定结果的生物学意义 大肠埃希菌和志贺菌被归为两个不同的菌属,很大程度上是



Ah 表示嗜水气单胞菌。本研究中测序的临床分离株用下划线标识

图 2 气单胞菌属的系统发育树局部

Fig. 2 Partial phylogenetic tree for *Aeromonas* spp.

出于临床意义的考虑。但从分子生物学与基因组学的角度来看,它们的相似性极高。当前普遍认为志贺菌是一群特殊的大肠埃希菌克隆^[8]。而气单胞菌属所含菌种较多,属内各种间及种内的序列差异大^[9];对某些种,其基因型难以很好地与其分类相关。鉴于以上两者在基因分型方面的复杂性,我们就以此作为本方法应用的实例。

对大肠-志贺菌群,使用两种方法都得到比较理想的结果。这同大肠埃希菌作为模式生物相关研究较多,相应参考序列较多有关。对大肠埃希菌 79 号菌株, BLAST 查询返回了 1 条

志贺菌的结果;而在系统发育树中,虽然该菌株位于大肠埃希菌中一个比较边缘化的类群,仍有多株高度近缘的大肠埃希菌序列和它聚类成一支。由此看来,对于序列与该种标准株相差较远的菌株,系统发育树较BLAST查询具有更高的可靠性。这主要是因为系统发育树是在同时综合地考虑所有序列相互关系的情形下建立的,而不是简单的“一对多”式比较。

对鲍氏和痢疾志贺菌,两种方法都返回了模糊的结果。这主要是由于两种菌的分类尚不能与其gyrB基因型很好关联所致。

对于单胞菌属,通过系统发育树获得的鉴别结果要明显好于BLAST查询。所测试的2株豚鼠气单胞菌,BLAST给出的最近缘序列均为嗜水气单胞菌AN-3。而在系统发育树中,2菌株序列均明确无误地聚类在豚鼠气单胞菌聚集的枝条中,该分支中虽然也混杂了一株嗜水气单胞菌但并未对两菌株的鉴定产生影响。这说明系统发育树中的位置关系不易受到个别异常序列的影响。

需要指出的是,对于287及295两株嗜水气单胞菌临床株,两种方法都给出了正确的鉴定。但系统发育树在它们与其它菌株的关系上提供了更多信息。从图2可见,树中的嗜水气单胞菌株明显分为两簇。同我们的两临床株聚类的序列基本都来自ICB数据库,即本地区分离株。说明该分支可能有一定的地区特异性。以上特征在BLAST返回结果中自然无法得到体现。

3.2 BLAST在此类应用中的局限 BLAST是一个基于局部序列比对的数据库搜索工具,随着十几年来各种生物序列数据的大量积累,现有数据库的规模堪称“海量”,并仍然在以指数级增长。在这样的前提下,以传统的序列比对启发式算法(如Needleman-Wunsch和Smith-Waterman)无法进行有效的数据库搜索,因为所消耗的时间将是一个天文数字。在BLAST的实现中,对待查询序列先按一定的字段长度分解并赋予匹配得分值,再对得分高于某个阈值的字段行数据库扫描,并对可能的匹配向两端延伸。由于以得分高于某阈值的定长字段的完全匹配作为序列比对单位,使得BLAST算法的效率大大提高,实现了对现有数据库的有

效搜索,但也是以牺牲了灵敏度为代价的。虽然目前NCBI网站的BLAST默认以Megablast方式运行,它对高相似度序列进行了优化。但总的来说,它仍是一种启发式算法,理论上无法保证得到最优匹配。对细菌种属水平分类鉴定这一特殊应用,它的敏感度仍不足,因为很多菌种/血清型间的看家基因序列差异仅仅是几个碱基位点的变异。

本研究所用的方法在鉴别结果上相对于常规BLAST方法的优势,很大程度上得益于本方法所用的数据集仅包含经过手工筛选的特定靶基因的序列。用于分类鉴定的序列数据库,其包含的序列数目并非越多越好,但质量一定要可靠。现今的GenBank数据太过于庞杂,存在数据质量良莠不齐,非冗余数据库难以做到真正意义上的“非冗余”以及序列注释标准不一等问题。采用经过人工采集并甄选的数据集,一定程度上避免了这些干扰;再加上多序列联配及剪切对齐等前期工作,也是对序列数据的一个标准化和归一化的过程。基于以上优点,专用靶分子数据库已经在细菌的基因分型中得到越来越多的采用,如rRNA数据库RDP^[10],gryB数据库ICB等,专门的细菌序列鉴别网站BIBI^[11],提供了十几种细菌分类靶分子的数据库供查询。直接对这类专用靶分子数据库进行BLAST查询,一般也能得到比较理想的结果。但系统发育方法的优势在于,能根据待检菌株与参考菌株间关系的亲疏远近给出全面而详尽的图形化阐释。

3.3 未来的可能趋势 当然,从网络数据库手工采集选择特定靶基因的序列需要一定的前期工作量,但这基本上是一次性的工作。对于细菌鉴别临床实验室而言,在一定时空内所检测菌种的大概范围以及所用靶分子是相对固定的。只要在最初搜集一定数量的相应序列就能方便地构建系统发育树,从而对临床株进行检测。同时,还可以将每次检测结果得到证实的菌株序列加入数据集,在以后的鉴定中参与构树;这样可以解决参考序列与本地菌株序列存在潜在的地区性差异的问题,使其呈现一种检测数量越多,效果越好的良性循环。此外,随着生物信息学工具应用的进一步深入,序列采集、序列质量

评估及选择,以及进化树构建等各个环节,在一个统一的平台上自动化或半自动化地进行成为可能,工作强度也将随之减轻。

目前,核苷酸序列测序虽已在分子生物学领域是较成熟的技术,但直到不久前,直接测序在分子诊断中的应用还不多,主要是大多数临床诊断实验室需要将样本外送测序,造成时间的延误和成本的上升。2006年Delmas等^[12]报道了用Beckman Coulter染料终止法测序*gyrB*基因片断,进行菌种分型,并以BIBI系统鉴别所测序列,全过程可在8 h内完成。随着新的高通量快速测序技术的进一步发展,如焦磷酸测序(pyrosequencing)^[13],毛细管电泳微流控(CAE)^[14],使得在临床实验室内完成测序工作,不仅在技术上,而且在时间和成本上均成为可能。而与之相伴的是,包括系统发育方法在内的各种序列分析与聚类的生物信息学方法,也必将在病原体鉴定中得到更为广泛的应用。

References:

- [1] PERRI C A. Detection and identification of microorganisms by gene amplification and sequencing [J]. *Clin Infect Dis*, 2007, 44: 1108-1114.
- [2] OWEN R J. Bacterial taxonomics, finding the wood through the phylogenetic trees [J]. *Methods Mol Biol*, 2004, 266: 353-383.
- [3] KASAI H, Watanabe K, Gasteiger E, et al. Construction of the *gyrB* Database for the Identification and Classification of Bacteria [J]. *Genome Inform Ser Workshop Genome Inform*, 1998, 9: 13-21.
- [4] THOMPSON J D, Higgins D G, Gibson T J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice [J]. *Nucleic Acids Res*, 1994, 22: 4673-4680.
- [5] HALL B G. *Phylogenetic trees made easy: a how-to manual* [M]. 2 ed, Sunderland, Massachusetts, Sinauer Associates, Inc, 2004. 50, 153-155.
- [6] FELSENSTEIN J. *Phylogeny Inference Package version 3. 6* [CP/OL]. Department of Genome Sciences, University of Washington. 2004. <http://evolution.genetics.washington.edu/phylip.html>.
- [7] MCGINNIS S, Madden T L. BLAST: at the core of a powerful and diverse set of sequence analysis tools [J]. *Nucleic Acids Res*, 2004, 32: W20-25.
- [8] ESCOBAR-PARAMO P, Giudicelli C, Parsot C, et al. The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised [J]. *J Mol Evol*, 2003, 57: 140-148.
- [9] SOLER L, Yanez M A, Chacon M R, et al. Phylogenetic analysis of the genus *Aeromonas* based on two housekeeping genes [J]. *Int J Syst Evol Microbiol*, 2004, 54: 1511-1519.
- [10] MAIDAK B L, Cole J R, Parker C T, et al. A new version of the RDP (Ribosomal Database Project) [J]. *Nucleic Acids Res*, 1999, 27: 171-173.
- [11] DEVULDER G, Perriere G, Baty F, et al. BIBI, a bioinformatics bacterial identification tool [J]. *J Clin Microbiol*, 2003, 41: 1785-1787.
- [12] DELMAS J, Breyse F, Devulder G, et al. Rapid identification of Enterobacteriaceae by sequencing DNA gyrase subunit B encoding gene [J]. *Diagn Microbiol Infect Dis*, 2006, 55: 263-268.
- [13] HAANPERA M, Jalava J, Huovinen P, et al. Identification of alpha-hemolytic streptococci by pyrosequencing the 16S rRNA gene and by use of VITEK 2 [J]. *J Clin Microbiol*, 2007, 45: 762-770.
- [14] KREMSER L, Blaas D, Kenndler E. Capillary electrophoresis of biological particles: viruses, bacteria, and eukaryotic cells [J]. *Electrophoresis*, 2004, 25: 2282-2291.

[责任编辑 张荣连]