文章编号：1000-0615(2021)11-1921-13　　水产学报, 2021, 45(11): 1921−1933

水产学报
JOURNAL OF FISHERIES OF CHINA
DOI: 10.11964/jfc.20210612935
Science Press

水环境细菌病原数据库的构建及应用

董鹏生[1,2]，郭海朋[1,2]，王艳婷[1,2]，程皇位[1,2]，王　凯[1,2]，洪　慢[1,2]，侯丹迪[1,2]，吴宇华[3]，张德民[1,2*]

(1. 宁波大学，农产品质量安全危害因子与风险防控国家重点实验室，浙江 宁波　315211；
2. 宁波大学海洋学院，浙江海洋高效健康养殖协同创新中心，浙江 宁波　315211；
3. 国家海洋局东海预报中心，上海　200136)

摘要：水环境病原菌对人类和水生动物的健康以及水产品生物安全带来了重大威胁，是公共卫生、水产养殖、食品安全等行业的重点监测对象。然而水环境病原菌数据库建设相对滞后，相关数据库分散在临床医学和水产动物病害等领域，且缺乏信息交流与融合，完整性仅限于各自独立的学科，不能满足区域尺度或生态学视角下，大规模水源性病原鉴定及生物安全评价等高通量监测的需求。因此，本研究通过整理人类介水传染病、水生动物、哺乳动物、植物和跨宿主疾病等 7 大类细菌病原信息，构建多线程可调度通讯模型和全局序列匹配算法，开发了水环境细菌病原数据库 (DPiWE，dayuz.com)。DPiWE 收集了 14 门、27 纲、54 目、116 科、221 属、1 097 种、9 070 株细菌病原的物种分类、16S rRNA 基因、宿主 (195 种) 和感染类型 (21 种) 信息。并在 Web 端实现信息检索、序列比对和注释结果可视化等功能。案例分析显示，DPiWE 构建的系统发育网络，清晰地将养殖环境菌株 DS10−D19 划分为鳗发光杆菌；用 DPiWE 对海水混养系统细菌高通量测序结果进行注释，揭示 3 种养殖动物病原分布具有明显差异，患病组水体有传播人体和鱼类共患病病原的风险。DPiWE 及配套分析流程可为水环境生物安全高通量评价、渔业生态健康维护和水产动物病害个性化防治提供新的思路和数据基础。

关键词: 细菌病原；16S rRNA；多线程调度数据库；鉴定；注释；水环境

中图分类号: S 917.1　　**文献标志码:** A

水环境生态系统与人类的生产生活密切相关，同时也是病原重要的“物种池”之一[1-2]。受人类活动影响，地表水污染日益严重[2]，病原微生物通过水环境传播的风险逐渐增加[3]，不仅直接威胁人类健康[1]，同时也增加了水生动物疾病发生的频率[4]，对食品安全和渔业可持续发展造成不利影响[5]。因此，Sagova-Mareckova 等[3] 建议将水体病原菌等微生物纳入水环境生态系统中的生物常规监测，用于环境生态评价、病原菌溯源追踪以及水生动物病害防治等工作[6-8]。

水环境病原微生物鉴定依赖病原菌数据库的完整性和准确度[7-8]。目前鉴定技术主要分为三大类：①基于多相分类 (polyphasic taxonomy) 策略，综合形态学和生理生化特征观察[9]、自动化微生物鉴定系统[10] 和基于 16S rRNA 基因 Sanger 测序[9, 11] 等传统分类学方法，需要使用数据库进

收稿日期：2021-06-30　　**修回日期：**2021-08-15
资助项目：国家重点研发计划 (2016YFC1402205)；国家自然科学基金 (31672658)；宁波市农业重大专项 (2017C110001)
第一作者：董鹏生 (照片)，从事养殖微生物生态研究，E-mail：dpsh@foxmail.com
通信作者：张德民，E-mail：zhangdemin@nbu.edu.cn

行系统进化分析以准确确定病原菌的分类地位；②基于基因芯片[12]、抗原抗体反应[13]、荧光定量PCR[14]、数字PCR[15]和微流控定量[16]等PCR反应检测，以及其他特异性病原微生物快速检测系统[13]等，均需依赖病原菌基因序列信息才能完成特异性的引物或抗体设计；③基于高通量测序和生物信息学技术检测环境样本中微生物的方法[7, 14]，虽然通量高、信息量大，但需要依赖完善的微生物分类特征和序列信息等先验知识，才能实现对环境中尽可能多的病原菌进行物种注释。不完整的数据库可能导致水环境中的DNA信息得不到有效捕获或准确注释，易造成物种漏检或鉴定错误。因此，建立和完善水环境病原菌信息，提高参考数据库的完整性和质量，不仅可以支撑传统分类学的水环境病原菌菌种鉴定工作，而且可以提升水环境病原监测的可靠性和准确度，对水环境病原菌的鉴别以及水产动物疾病预警和监测具有重要意义。

病原菌数据库的建设得到研究人员的充分关注，但是仅局限于各自的领域，缺乏对水环境病原菌信息的系统梳理和整合[17-20]。目前，国际上公开的病原菌数据库主要有：① 病原体-宿主互作数据库 (pathogen-host interactions, PHI-base, http://www.phi-base.org/)，共收集276种动植物病原 (植物病原菌为主) 的8 070个功能基因[17]；② 美国国立卫生研究院 (national institutes of health, NIH) 病原数据库 (http://www.ncbi.nlm.nih.gov/pathogens)，共收录34种细菌性病原菌的92 466株菌株信息；③ 京都基因与基因组百科全书 (Kyoto encyclopedia of genes and genomes, KEGG) 病原耐药数据库，共收录552种 (包括204种病毒、151种细菌、61种真菌和136种其他真核病原) 1 340株人类和人畜共患病病原菌信息[18]；④全球医生联盟也提供了包含64种致病细菌的参考信息 (https://globalrph.com/bacteria/)。国内人类感染性疾病的病原微生物信息主要被中国医学科学院病原生物学研究所的细菌毒力因子数据库[19] (a reference database for bacterial virulence factors, VFDB, http://www.mgc.ac.cn/VFs/)、中科院微生物所的微生物与病毒主题数据库 (http://www.micro.csdb.cn/) 和华大基因的国家基因库子库病原菌变异数据库 (Pathogen variation database，PVD，http://db.cngb.org/pvd/) 收录。水产动物疾病相关的数据库主要有：① AquaPathogen X病原模板数据库 (aquatic animal pathogens template database, https://pubs.usgs.gov/fs/2012/3015/)，提供了基于商业软件FileMaker® Pro的病原菌信息收集标准化模板，方便用户制作自己的本地数据库，但未提供病原菌信息；② 鱼类病原基因组数据库 (fish pathogen genome database, https://pubmlst.org/projects/fish-pathogens) 具有Web端界面，收录了气单胞菌属 (*Aeromonas*)、弧菌属 (*Vibrio*)、短螺旋体属 (*Brachyspira*) 等鱼类病原的203个基因标志物[20]；③ 水生病原数据库清单 (list of aquatic pathogens database，http://aahl.res.in/repositories/list-of-aquatic-pathogens-database/) 提供了20种水生动物病毒和5种细菌病原名称。遗憾的是，与人类感染性疾病等病原数据库相比，水环境病原菌数据库的建设工作存在明显的滞后性，收录信息及使用方法有待进一步完善。因此，有学者呼吁加强水环境病原菌的数据库建设和数据分析等工作，帮助推动水产养殖可持续发展和全球水环境保护等工作[5, 8]。本研究通过收集整理人类、鱼类、无脊椎动物、植物疾病及跨宿主共患病的细菌病原分类学、感染类别和宿主的信息，并基于多线程可调度数据库系统和多任务模式的全局序列匹配算法，开发了DPiWE数据库及配套生物信息学分析方法，以期为水环境病原菌的鉴定、溯源和水产动物病害防治等工作提供基础数据保障和分析流程参考，也为渔业生态健康评价及渔业资源保护提供科学支撑。

1 材料与方法

1.1 病原菌信息收集及数据库构建

DPiWE数据库的构建流程如图1-a所示，包括病原菌物种信息汇总和16S rRNA基因序列收集两部分工作。首先，对目前水环境病原菌物种信息进行人工整理和数据库构建。对于水产动物的病原细菌学信息，参考Austin等[21]和房海等[22]的研究，收集病原菌物种或菌株名及其感染的宿主信息。其他种类的细菌病原信息收集采用以下策略：首先根据Cabral等[7]的报道，收集了250株志贺氏菌属 (*Shigella*)、沙门氏菌属 (*Salmonella*) 等人类介水传染病的病原菌株信息；然后考虑到水环境较为复杂，容易受工农业等人类活动干扰[2-3]，尤其是医院附近的水体[7]，病原菌多样性高、组成复杂，为了使数据库适用性更广，数据库同时收录了KEGG病原耐药数

据库[18]和 VFDB 细菌毒力因子数据库[19]中所有的细菌病原信息。为防止物种分类信息被错误记录，利用原核生物名称列表数据库 (list of prokaryotic names with standing in nomenclature, LPSN)[23-24]，对收集到的病原菌分类学信息进行人工核对和校准，共获得1 097 个病原菌物种。然后利用 KEGG 数据库对病原菌的感染类型进行注释，共得到 21 种感染类型。最后，使用 NCBI (national center for biotechnology information) 核酸数据库获取病原菌菌株对应的 16S rRNA 基因序列信息。若没有该菌株的序列，则用该物种模式株的 16S rRNA 基因序列；若病原菌只有属的分类信息，则下载该属下各物种的模式菌株 16S rRNA 基因序列。

1.2　基于多线程可调度通讯模型的 DPiWE 网络端实现方法

DPiWE 的网络端系统搭建流程如图 1-b 所示，包含应用展示层、业务逻辑层、数据仓储层、数据实体层和基础设施层 5 个子应用层。其中，应用展示层使用 HTML5 技术，渲染用户使用界面网页，响应用户上传数据、匹配计算等用户交互操作；其他应用层采用 rabbitmq 消息队列和 ASP.Net-Core 框架进行开发，业务逻辑层主要负责组织整个应用的流程，定义了软件需要完成的任务，是调用其他模块的通道；数据仓储层负责对数据库表的封装，提供 CRUD [增加 (create)、检索 (retrieve)、更新 (update) 和删除 (delete)] 功能；数据实体层用于实现对用户数据的封装、定义业务状态信息和业务规则；基础设施层为其他各层提供层间信息传递、数据持久化机制等通用技术[25-27]。管理系统效率和比对软件的运算速率是影响数据库匹配和注释序列时效的核心因素。为了解决数据并发 (如序列比对运算) 和系统内通信速率对服务器系统资源的分配问题，在 DPiWE 网络端中采用陈小辉等[25]和左朝树[26]的方法，构建满足高并发的线程调度模型，通过将用户任务写入消息队列，以异步方式处理非重点业务，从而加快响应速度。当并发量大时，按照数据库的承载量，从消息队列中拉取并处理消息，对内部通讯进行并发接收和发送数据。同时根据运算节点的系统资源占有率，动态调度通讯和运算节点，合理协调数据库内部的通信性能和服务器系统对序列的处

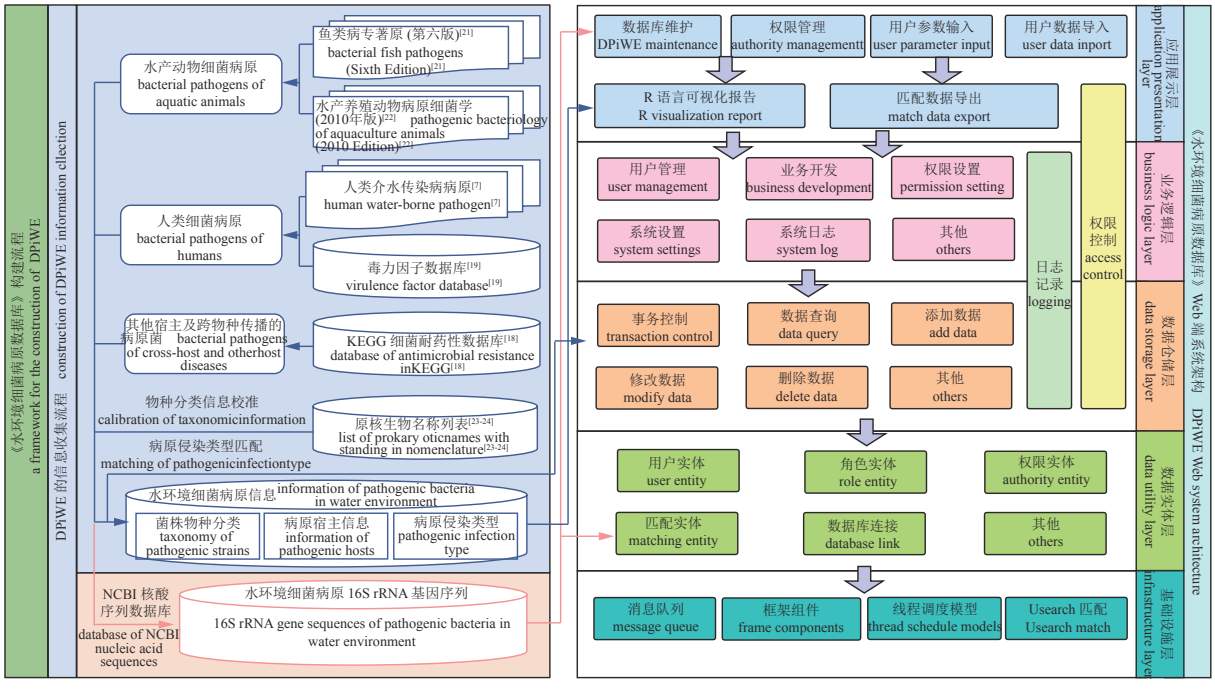


图 1　水环境细菌病原数据库 (DPiWE) 的构建流程

(a) DPiWE 的信息收集流程；(b) 网络端系统构架

Fig. 1　A framework for the construction of database of pathogenic bacteria in water environment (DPiWE)

(a) construction of DPiWE information collection; (b) system architecture in Web site

理能力，最大限度地提高用户提交任务的运行效率和服务器资源的利用率。数据库管理采用NoSQL的方法[27]，实现对病原菌物种信息等数据库数据的维护、访问以及用户序列比对和结果注释等功能。最后，使用R语言服务器程序Rserve对序列比对和注释结果进行统计汇总，并提供基于比对结果的可视化报告(图1-b)。

1.3 不同序列比对软件用于匹配DPiWE的性能评价方法

首先利用DPiWE数据库序列，分别构建用于模拟全长16S rRNA基因测序序列(如Sanger, PacBio RS等一代或三代测序平台的数据)和16S rRNA基因V4可变区测序序列(如Illumina Hiseq 2 500或MiSeq测序平台等二代测序数据)的测试数据集，参考McDonald等[28]的方法，并进行改进：① 对于全长16S rRNA基因序列，利用Seqtk-1.3软件(https://github.com/lh3/seqtk，版本号r106)分别随机抽取1、1 000、2 000、3 000、4 000和5 000条DPiWE数据库序列，每次抽取重复3次；② 对于V4可变区序列，先使用Usearch软件[29](版本号11.0.667)的电子PCR模块(search_pcr2)提取DPiWE数据库序列对应的V4区，再进行序列抽取，抽取方法与上述全长序列一致，每次抽取均为独立事件。在考虑到服务器性能与负荷的情况下，本研究选用当前微生物序列比对分析中常用的软件进行测试。QIIME 1作为经典的微生物高通量数据分析软件，在水产动物微生物组和其他水生环境微生物群落研究中得到了非常广泛的应用[6, 14, 30]。因此使用QIIME 1软件的注释结果作为选择比对软件的最低标准，即选用的比对软件注释正确率不应低于QIIME 1的结果。由于数据库面临多用户及多线程使用压力，基于机器学习分类器的方法对运算资源要求较高，运算时间过长，因此并未使用QIIME 2等采用此类方法的软件进行测试[31]。综合以上因素，本研究选用BLAST+软件[32](https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/，版本号2.11.0)中的Blastn模块，Qiime软件(https://github.com/biocore/qiime, 版本号1.9.1)中的“*pick_closed_reference_otus.py*”脚本[30]，Sortmerna软件[33](https://github.com/biocore/sortmerna，版本号4.3.2)和Usearch软件的全局搜索模块[29]等4个常用的序列比对算法进行测试。依次运行4种软件，将测试数据集序列对比到DPiWE数据库中，其中比对相似度(percent identity或similarity)和最低共识度(minimum consensus fraction)等相关运算参数分别设置为0.98和0.51，其他均采用软件默认参数。以上4种软件的序列比对运算均在相同测试环境(DELL™ T630服务器，处理器：Intel Xeon E5-2620V3×2，内存64 G，SAS硬盘300G×3，操作系统为Ubuntu 16.04.7 LTS Linux平台)下进行，并记录软件比对不同测试数据集的运行时间和计算结果。以测试序列是否匹配到原始数据的种分类水平为依据，采用交叉验证的方法[31, 34]对4种软件的匹配性能进行分析。使用R语言的“verification”(版本号1.42)[35]和“ROCR”(版本号1.0-11)[36]软件包计算真阳性率(true positive rate)和假阳性率(false positive rate)，绘制接受者操作特性(ROC)曲线，并计算ROC曲线下的面积(AUC)。AUC值越接近于1，说明软件匹配可靠性越高，比对性能越好[36]。

1.4 案例研究及其数据分析方法

使用DPiWE数据库分析了一株未鉴定细菌16S rRNA基因全长序列和3种海水养殖动物及养殖环境细菌群落高通量测序数据[37]。

一株来源于海水养殖区的细菌16S rRNA基因序列分析 菌株DS10−D19来自宁波大学微生物与水域生态健康研究团队菌种保藏中心，样本采集地为宁波象山港黄避岙某养殖场(29.60°N, 121.80°E)。提取菌株基因组DNA后，使用16S rRNA基因的引物对(27F: 5′-AGAGTTTGATCCTGGCTCAG-3′，1492R: 5′-TACCTTGTTACGACTT-3′)[38]进行PCR扩增，PCR反应体系和反应条件参考陈慧等[9]的方法。PCR产物纯化后使用上海生工生物工程有限公司的Sanger平台进行16S rRNA基因测序。将该序列比对到DPiWE数据库，得到物种分类信息。并从DPiWE数据库中下载该菌株所在属的所有病原菌16S rRNA基因序列，用PyNAST软件[39]参考Greengenes 13.5数据库(http://greengenes.secondgenome.com)中的16S rRNA基因预聚类序列进行多重比较，使用R语言“phangorn”(版本号2.5.5)[40]，“ggtree”(版本号2.2.1)[41]和“ggnetworx”(版本号0.99.0)[42]软件包进行系统发育网络(phylogenetic network)分析。

三种海水养殖动物肠道及其养殖环境的细菌群落高通量测序数据分析 Sun等[37]于2017年8月采集了中国莆田登封海水养殖区的

养殖水体、底泥，以及凡纳滨对虾 (*Litopenaeus vannamei*)、三疣梭子蟹 (*Portunus trituberculatus*) 和硬壳蛤 (*Mercenaria mercenaria*) 的肠道微生物样本，其中有 3 个池塘暴发疾病，其他均为健康组。其高通量测序序列和实验设计储存在 NCBI 的 SRA (Sequence Read Archive) 数据库中 (项目编号 PRJNA542997)。使用 Usearch 软件[29] 对原始数据进行质控和去噪[43]，得到最小测序深度为 11 197 序列数的 zOTU (零半径操作分类单元) 数据集，使用 DPiWE 数据库注释其代表序列，统计各样本的病原菌组成及丰度信息，使用 R 语言 “ggalluvial” (版本号 0.11.3，http://corybrunson.github.io/ggalluvial/) 软件包进行可视化和溯源分析。

2　结果

2.1　DPiWE 数据库细菌病原信息收录及网站功能结构

根据数据库病原信息构建流程 (图 1-a)，最终收集整理了 9 070 个病原菌菌株信息，共分为鱼类/两栖动物 (7 146 株)、人类介水传染病 (250 株)、人类其他疾病 (595 株)、哺乳动物 (358 株)、植物 (118 株)、无脊椎动物 (23 株) 以及跨宿主 (580 株) 病原等 7 大类数据，其中包括了 14 门、27 纲、54 目、116 科、221 属、1 097 种的病原菌分类信息 (图 2)。属于 γ-变形菌纲 (Gammaproteobacteria) 的病原在鱼类、人类和哺乳动物病原中的比例最高，其次是厚壁菌门 (Firmicutes)，放线菌门 (Actinobacteria) 和拟杆菌门 (Bacteroidetes) 的病原也超过 500 株。数据库还收录了鲑属 (*Oncorhynchus*)、罗非鱼属 (*Oreochromis*) 等 195 种被病原感染的宿主信息，以及“弧菌感染 (*Vibrio* infection)”、“气单胞菌感染 (*Aeromonas* infection)”、“食源性产气荚膜梭菌中毒 (foodborne *Clostridium perfringens* intoxication)”等 21 种细菌感染类型。Web 端数据库按照图 1-b 的系统架构设计，实现了基于多线程可调度的分布式通信模型、全局匹配算法和 R 语言可视化方案。同时，开发了

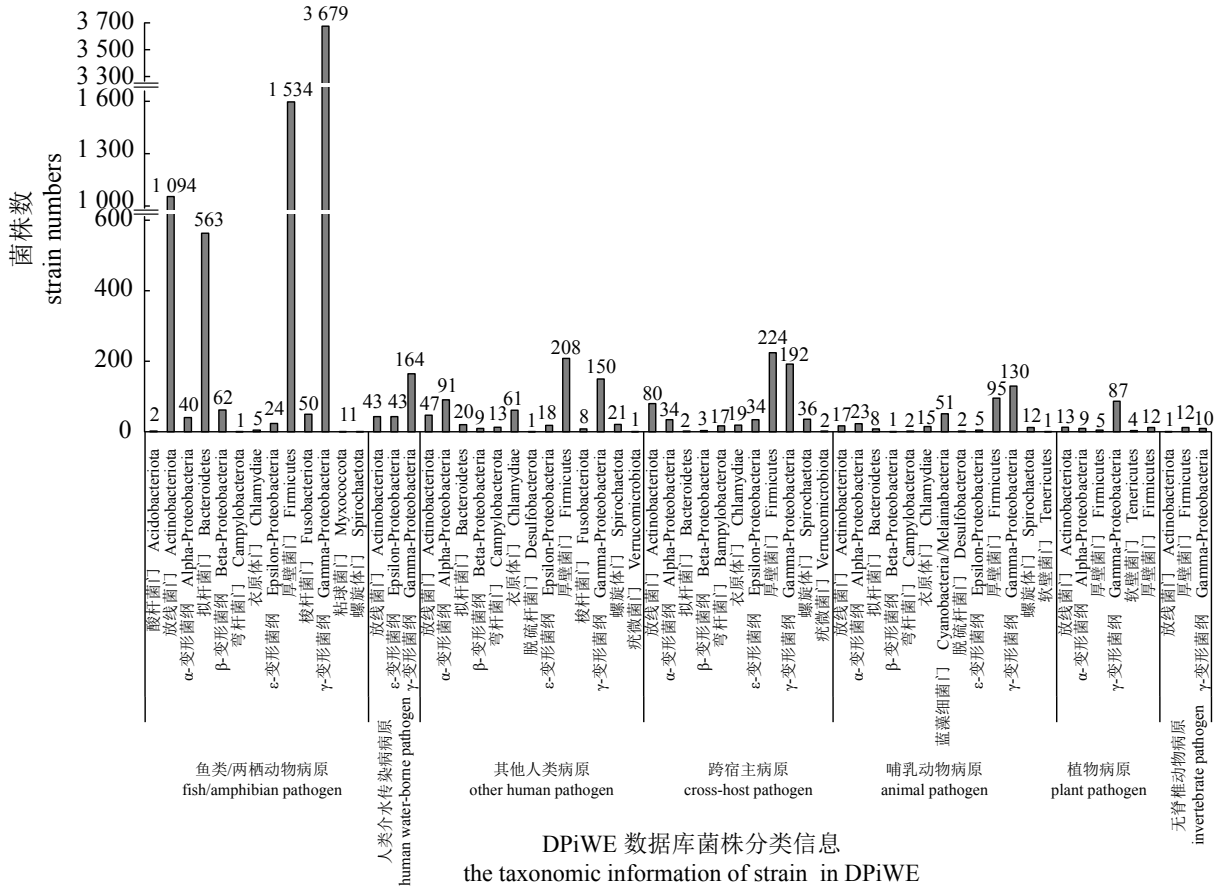
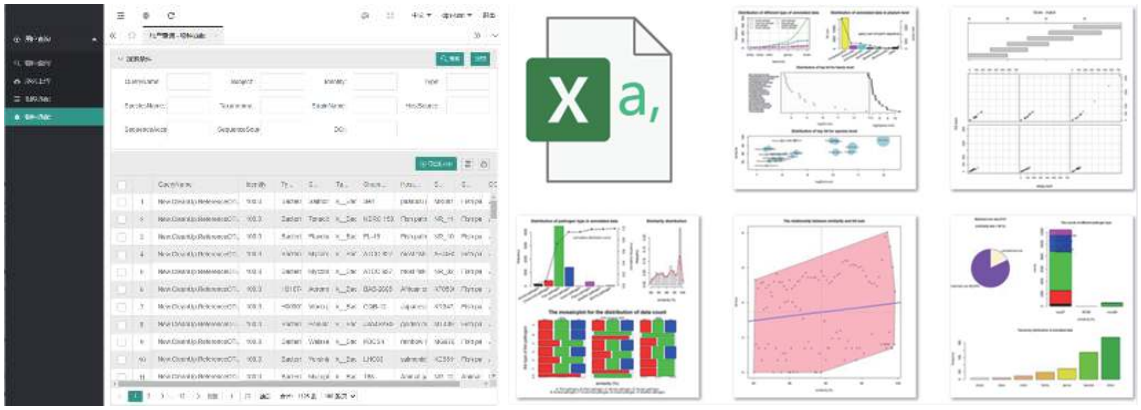


图 2　DPiWE 数据库中细菌病原类群和宿主信息统计

Fig. 2　Summary of pathogenic bacteria taxonomy and host information in DPiWE

信息发布 、用户管理、病原菌数据库的信息检索、序列上传与存储、序列比对 (图 3-a) 和匹配结果可视化等功能 (图 3-b)。



(a) (b)

图 3 DPiWE 数据库 Web 端功能

(a) 信息发布、用户管理、信息检索、序列上传与存储和序列比对功能；(b) 匹配结果可视化报告

Fig. 3 Web function for DPiWE

(a) web function of information release, user management, database retrieval, upload and storage and alignment for sequences; (b) visual report of matching results

2.2 四种序列比对软件与 DPiWE 数据库的匹配性

从软件运行时间和匹配效果的角度，评价 Usearch、Sortmerna、Qiime 和 Blast 这 4 种常用序列比对软件与 DPiWE 数据库的匹配效能 (图 4)。对于 16S rRNA 基因全长和 V4 可变区序列的测试数据集，4 种软件在比对运行时间上有着相似的规律 (图 4-a)，开启多线程的任务比一般任务执行的时间更短。Usearch 软件在多线程模式下运行的时间最短，其次是 Sortmerna 软件，但 Blast 软件在多线程和一般模式下均比其他软件的运行时间长。序列匹配性能结果表明，Usearch 软件拥有最高的序列比对精确度，AUC 值接近 1，Blast 和 Qiime 软件的精确度最低，AUC 均为 0.72 (图 4-b)。这可能与不同软件的序列比对算法有关，Blast 采用的是局部搜索算法，由于 16S 基因的保守性较高，容易造成假阳性比对结果，Usearch 可以调用全局搜索算法，在保证较高比对覆盖度的情况下，得到的结果更加可靠。因此，在使用 DPiWE 时，推荐利用 Usearch 软件进行序列比对，可以在提高运算速率的情况下，得到正确率较高的比对结果。

2.3 基于 DPiWE 数据库比对结果的菌株 DS10-D19 系统发育网络构建

把菌株 DS10-D19 的 16S rRNA 基因序列比对到 DPiWE 数据库中，结果显示其与数据库中的鳗发光杆菌 (*Photobacterium leiognathid*，NCBI 登录号为 NR_029253.1，DPiWE 数据库菌株编号为 7a78d052deced33f69703a3f81761205) 的相似度达到 99.1%。然后，把 DPiWE 中所有的发光杆菌属 (*Photobacterium*) 细菌 [鳗发光杆菌、纤细发光杆菌 (*P. angustum*)，*kishitanii* 发光杆菌 (*P. kishitanii*)，美人鱼发光杆菌亚种 (*P. damselae* subsp. *piscicida* 和 *P. damselae* subsp. *damselae*) 和美人鱼发光杆菌两个菌株 (ATCC33539 和 MTO71398.1)] 的 16S rRNA 基因序列与菌株 DS10-D19 一起进行多重比对，并构建菌株 DS10-D19 的系统发育网络 (图 5)。结果显示，菌株 DS10-D19 与鳗发光杆菌遗传距离最接近，结合 16S rRNA 基因相似度信息，初步判断菌株 DS10-D19 属于鳗发光杆菌。此外，通过两次分支，菌株 DS10-D19 与纤细发光杆菌和鳗发光杆菌相互连接并位于网络的同一侧，因此，该菌株分支 (δ 和 β) 与鳗发光杆菌分支 (γ) 和纤细发光杆菌分支 (δ) 在网络中均比较接近。

2.4 基于高通量测序数据的 DPiWE 分析揭示水产动物和养殖环境病原群落组成及溯源结果

使用 DPiWE 数据库，对 3 种海水养殖动物

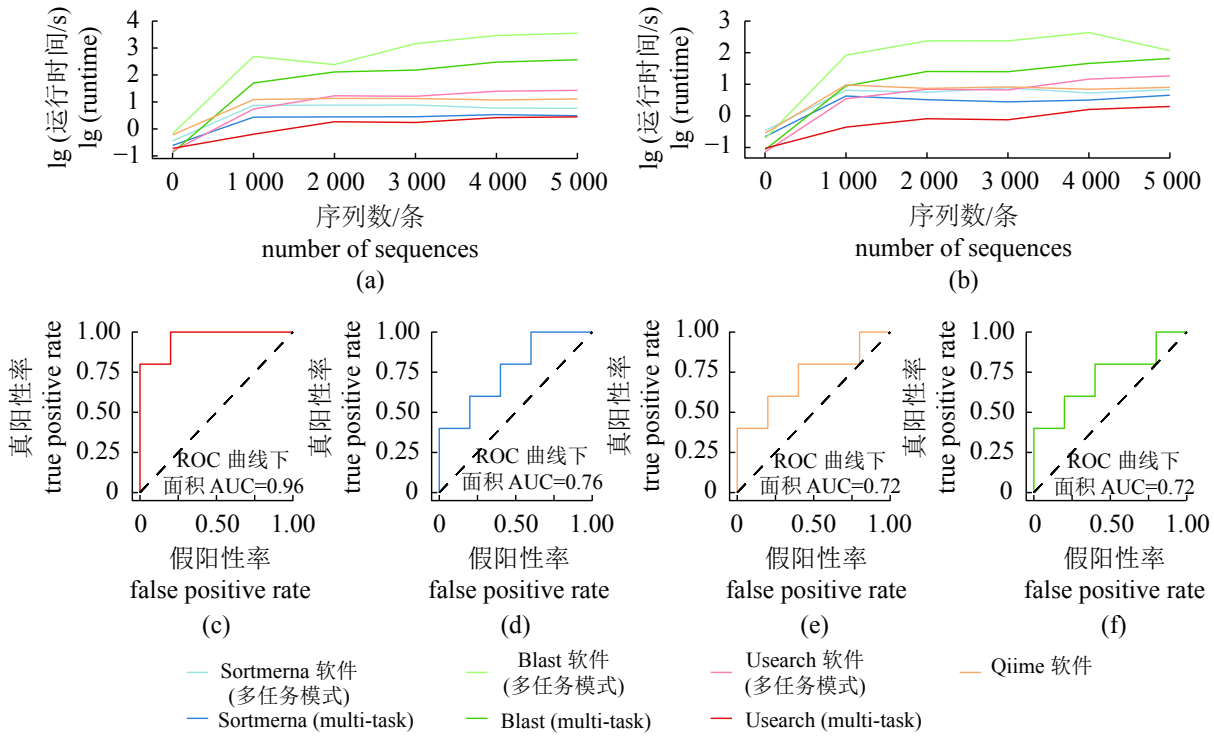


图 4　四种序列比对软件与 DPiWE 数据库的匹配性

(a)~(b) 4 种软件比对 16S rRNA 基因全长 (a) 和 V4 可变区 (b) 序列的运行时间比较，多任务模式下 4 种软件使用相同的 CPU 线程数；(c)~(f) Usearch (c)、Sortmerna (d)、Qiime (e) 和 Blast (f) 4 种软件基于交叉检验的比配精确度，AUC 为曲线 (接受者操作特性曲线) 下面积，AUC 越接近于 1，说明准确性测试的结果越好

Fig. 4　Performance of four software for sequence alignment matching the sequences in DPiWE

(a)-(b) runtime of the four software for full length (a), and V4 region (b) of 16S rRNA gene, the four-software using the same number of CPU threads in multitasking mode; (c)-(f) The performance of Usearch (c), Sortmerna (d), Qiime (e), and Blast (f) on the cross-validated sequence datasets for DPiWE. The value of AUC [Area Under ROC (receiver operating characteristic) Curve] the closer to 1 means the better the accuracy in test results

肠道及养殖环境细菌高通量测序数据进行序列比对和病原菌物种注释 (图 6)。该养殖环境样本中的病原菌主要为弧菌属 (7.39%)、发光杆菌属 (5.64%)、乳球菌属 (*Lactococcus*, 1.40%)、假单胞菌属 (*Pseudomonas*, 0.26%) 和假交替单胞菌属 (*Pseudoalteromonas*, 0.10%) 等。健康和疾病组水产动物的病原菌也主要来自弧菌属和发光杆菌属，但凡纳滨对虾和三疣梭子蟹肠道中病原菌的丰度和物种组成模式均不同。患病对虾肠道中的主要病原菌是发光杆菌属 (45.03%) 和弧菌属 (24.91%)，其丰度均高于健康组 (8.70%、8.02%)；而患病梭子蟹肠道中弧菌属 (47.17%) 和发光杆菌属 (10.57%) 的丰度均明显高于健康组 (0.48%、0.20%)。值得注意的是，在患病梭子蟹和对虾肠道中属于弧菌属的病原，超过 45% 来自鱼类和人类共患病病原；水体中病原菌相对丰度虽然较低，但是仍有超过 50% 的病原菌属于跨宿主共患病病原。

3　讨论

水体病原菌是影响人类、动物和环境健康的重要因素[2-3]，也是全健康 (One Health) 全球使命重点关注的领域之一[44]。描述与水体病原菌相关的分类学和宿主信息，对更好地了解病原菌群落多样性和对水圈环境的影响至关重要[3]。本研究开发的 DPiWE 数据库，可用于水环境细菌病原的分类学和群落生态学等领域的研究。目前，KEGG 细菌耐药性数据库[18] 和 NIH 病原数据库等综合性病原数据库，仅包含 48–201 株鱼类相关的病原 (2020 年 12 月 22 日数据)。常用的水产动物病原菌数据库，如鱼类病原基因组数据库和水生病原数据库清单等，也存在收录信息较少，病原菌种类覆盖不足，不支持高通量数据注释等问题[19]。针对以上情况，本数据库提供了更多的人类介水传染病、水产动物疾病及跨宿主共患病等病害的细菌病原的分类学和宿

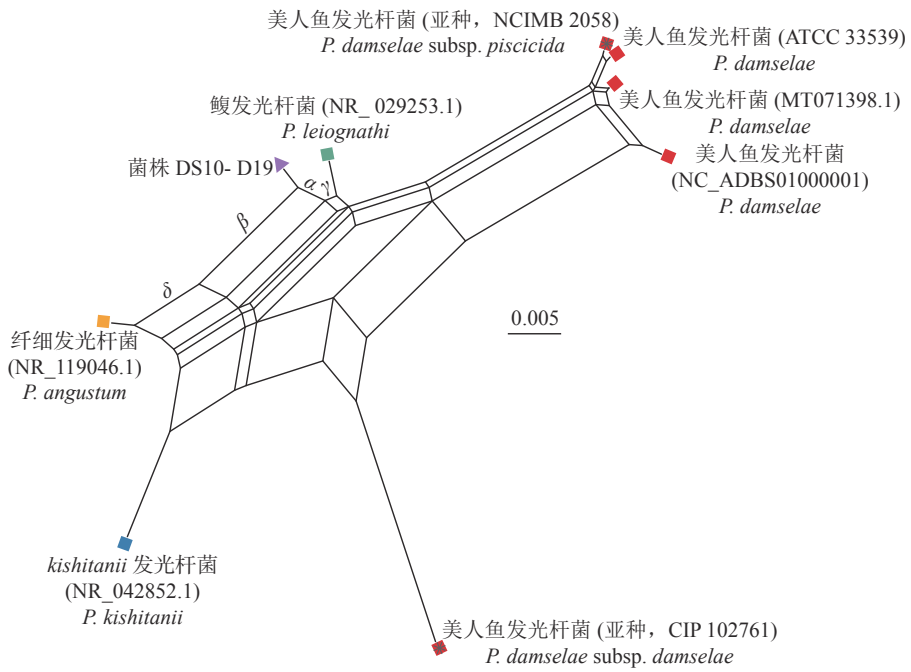


图 5　基于 DPiWE 构建菌株 DS10-D19 在发光杆菌属中的系统发育网络

■. DPiWE 中的参考菌株；▲. 菌株 DS10−D19；＊. 模式菌株；α 和 β 分支.菌株 DS10−D19 发出的分支；γ 分支.鳗发光杆菌发出的分支；δ 分支.纤细发光杆菌发出的分支；括号内为菌株名或 GenBank 登录号

Fig. 5　Phylogenetic network of strain DS10-D19 in *Photobacterium* based on DPiWE

■. strains in DPiWE; ▲. strain DS10−D19; * type Strains; α, β branch. split from strain DS10−D19; γ branch. split from *P. leiognathid*; δ branch. split from *P. angustum*; the strain names or GenBank access numbers are in brackets

主信息。此外，DPiWE 的注释信息和其配套的生物信息学数据分析流程，可以在不同的水环境群落中进行快速、可靠地高通量病原菌鉴定，病原菌群落特征分析以及基于大数据角度对水环境病原溯源追踪等工作。DPiWE 是已知第 1 个基于高通量测序数据，分析水体环境病原菌群落特征和溯源的综合数据库。

准确度在菌株鉴定以及高通量测序数据分析中至关重要，这在很大程度上取决于参考数据库[45]和用于序列比对的算法[34]。参考数据库的完整性和质量是影响病原菌鉴定，以及高通量测序数据生物信息学分析可靠性和再现性的重要因素[46]。本研究在收集整理病原分类信息时，统一不同菌株信息源的物种分类标准，使用原核生物名称列表数据库 (LPSN)[24] 对病原菌分类系统进行人工校准，确保分类信息符合当前国际主流分类标准。不同的序列匹配算法对数据的处理方法不同[34]，本研究通过评估 4 种常见序列比对软件的匹配性能，选择性能最佳的软件作为 DPiWE 数据库的比对推荐算法。Usearch 软件在运算时间和匹配精确度上均优于 Blast 软件。这可能是因为 16S rRNA 基因具有较强的保守性，尤其是 DPiWE 的数据主要集中在厚壁菌门和 γ-变形菌纲中，Blast 使用的局部搜索算法，对于全长 16S 序列的保守区不敏感，容易造成假阳性结果[34]。而 Usearch 软件通过调用全局匹配算法[29]，在保证序列比对高覆盖率的情况下，使用 K-mers 数估计整体比对序列的同一性，在保持较高灵敏度的同时，显著提高了运算速率。另外，PHI-base 数据库[17] 和 VFDB 数据库[19]，在 Web 端均提供了 Blast 比对模块，可对功能基因进行相似性比较；而 SCycDB 硫循环数据库[46] 将 Usearch、Blast 和 Diamond 这 3 种工具结合在一起，用于硫循环功能基因丰度估计和物种分类鉴定等分析。由于不同数据库所提供的参考序列种类并不相同，针对不同的数据库序列类型，需要选择合适的方法来实现数据库的比对搜索等功能。根据本研究结果，推荐使用 Usearch 多全局搜索算法的多任务模式对 DPiWE 数据库进行序列匹配。

基于 16S rRNA 基因序列的物种鉴定是细菌分类学上最常用的标准之一。目前已建立了多

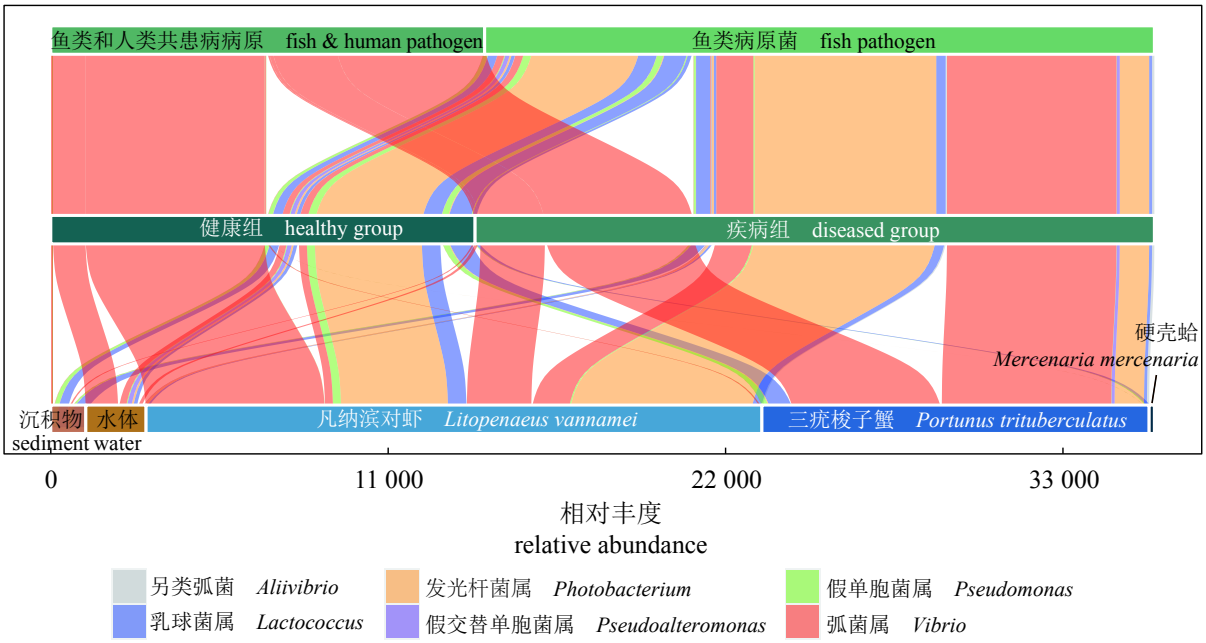


图 6　三种海水养殖动物肠道及养殖环境中病原菌组成及其溯源

Fig. 6　Composition and source-tracking of pathogens in intestine and culture environment of three mariculture animals

种基于 16S rRNA 序列的病原菌分子生物学分类鉴定方法[47-48]，但是由于部分细菌的 16S rRNA 基因具有较高的多态性，种间 16S rRNA 基因的同源性较高[49]，尤其是在弧菌中[38] (部分菌株的 16S rRNA 基因同源性达到 99%)，目前该标准的权威性受到一定的挑战[16, 50]。尽管如此，在病原鉴定工作中，依然可以在高质量的测序基础上，通过生物信息学方法解析出同源性之外的其他进化信息，提高数据的利用效率和解释度。例如张晓华等[11] 利用 159 种弧菌的 16S rRNA 基因序列构建系统进化树，得到了较为清晰的弧菌科内部物种进化关系。因此，16S rRNA 基因序列仍是细菌新种鉴定[51]、病原菌检测[38, 48] 等工作的重要组成部分。本研究使用 DPiWE 数据库对菌株 DS10-D19 进行初步鉴定，虽然菌株 DS10-D19 与纤细发光杆菌的 16S rRNA 基因序列相似度不足 99%，但是系统发育网络提示，该菌株与鳗发光杆菌和纤细发光杆菌具有相似的网络结构，它们可能在进化过程中经历了相似的事件[52]，在今后的研究中可结合基因组信息开展进一步研究。

基于高通量测序的病原检测和疾病诊断作为一项新兴技术，具有免培养、耗时短、通量高等特点[53]，在医学研究和临床诊断等领域得到了广泛重视[54]，但在水产动物疾病诊断和水环境病原菌检测方面，不仅应用较少，而且缺乏相关的规范化标准[14]。本研究构建的 DPiWE 数据库，可以对基于高通量测序的环境和水生动物微生物群落数据，进行快速、准确的病原菌注释，解析水环境病原菌在不同样本中的组成和分布规律，并可通过溯源分析，推断病原微生物与宿主的可能来源和传播途径。对 3 种海水养殖动物肠道及养殖环境微生物高通量测序数据的案例分析结果表明，尽管位于同一区域，基于高通量测序数据的 DPiWE 注释信息仍然可以区别出不同患病动物主导病原在物种组成和分布上的差异，这也与原始案例中描述的患病动物临床特征一致[42]，体现了 DPiWE 在分析病原群落生态学中的优势。养殖环境微生物高通量测序与 DPiWE 数据库的联合使用，不仅解析了不同水产动物肠道病原菌群落的结构和组成差异，而且检测到疾病组养殖水体具有传播人体和鱼类共患病病原的风险，为今后水环境生物安全高通量测序分析和水产动物病害个性化防治提供新的思路和数据基础。

参考文献 (References):

[1] Pedley S, Bartram J, Rees G, *et al*. Pathogenic Mycobacteria in water: a guide to public health consequences, monitoring and management[M]. London, UK: IWA Publishing, 2004.

[2] Wright R J, Erni-Cassola G, Zadjelovic V, *et al*. Marine plastic debris: a new surface for microbial colonization[J]. Environmental Science & Technology, 2020, 54(19): 11657-11672.

[3] Sagova-Mareckova M, Boenigk J, Bouchez A, *et al*. Expanding ecological assessment by integrating microorganisms into routine freshwater biomonitoring[J]. Water Research, 2021, 191: 116767.

[4] Bondad-Reantaso M G, Fejzic N, MacKinnon B, *et al*. A 12-point checklist for surveillance of diseases of aquatic organisms: a novel approach to assist multidisciplinary teams in developing countries[J]. Reviews in Aquaculture, 2021, 13(3): 1469-1487.

[5] Naylor R L, Hardy R W, Buschmann A H, *et al*. A 20-year retrospective review of global aquaculture[J]. Nature, 2021, 591(7851): 551-563.

[6] Mansour I, Heppell C M, Ryo M, *et al*. Application of the microbial community coalescence concept to riverine networks[J]. Biological Reviews, 2018, 93(4): 1832-1845.

[7] Cabral J P S. Water microbiology. Bacterial pathogens and water[J]. International Journal of Environmental Research and Public Health, 2010, 7(10): 3657-3703.

[8] Chen Q L, An X L, Li H, *et al*. Long-term field application of sewage sludge increases the abundance of antibiotic resistance genes in soil[J]. Environment International, 2016, 92-93: 1-10.

[9] 陈慧, 张德民, 王龙刚, 等. 一株反硝化光合细菌的生物学特性及系统发育分析[J]. 微生物学报, 2011, 51(2): 249-255.
Chen H, Zhang D M, Wang L G, *et al*. Biological characteristics and phylogenetic analysis of a denitrifying photosynthetic bacterium[J]. Acta Microbiologica Sinica, 2011, 51(2): 249-255 (in Chinese).

[10] 童桂香, 韦信贤, 黎小正, 等. BIOLOG自动微生物鉴定系统在水产动物病原菌检测中的应用[J]. 安徽农业科学, 2011, 39(13): 7846-7848.
Tong G X, Wei X X, Li X Z, *et al*. Application of BIOLOG automatic microbiological assay system in the detection of aquatic animals pathogen[J]. Journal of Anhui Agricultural Sciences, 2011, 39(13): 7846-7848 (in Chinese).

[11] 张晓华, 林禾雨, 孙浩. 弧菌科分类学研究进展[J]. 中国海洋大学学报, 2018, 48(8): 43-56.
Zhang X H, Lin H Y, Sun H. Taxonomy of Vibrionaceae: a review[J]. Periodical of Ocean University of China, 2018, 48(8): 43-56 (in Chinese).

[12] 徐晓丽, 邵蓬, 丁子元, 等. 生物芯片技术及其在水产动物病原检测中的应用[J]. 河北渔业, 2013(4): 59-63.
Xu X L, Shao P, Ding Z Y, *et al*. Description and application of biochip in pathogen detection of aquatic animals[J]. Hebei Fisheries, 2013(4): 59-63 (in Chinese).

[13] 贺电, 吴后波. 分子生物学技术在水产养殖动物病原快速检测中的应用[J]. 海洋科学, 2007, 31(3): 76-81.
He D, Wu H B. The application of molecular methods to rapid detection of aquatic pathogens[J]. Marine Sciences, 2007, 31(3): 76-81 (in Chinese).

[14] Dong P S, Guo H P, Wang Y T, *et al*. Gastrointestinal microbiota imbalance is triggered by the enrichment of *Vibrio* in subadult *Litopenaeus vannamei* with acute hepatopancreatic necrosis disease[J]. Aquaculture, 2021, 533: 736199.

[15] 马新冉, 肖雨晴, 雷春, 等. 数字PCR技术在水产病原菌检测中的应用[J]. 鲁东大学学报(自然科学版), 2020, 36(1): 48-54.
Ma X R, Xiao Y Q, Lei C, *et al*. Application of digital PCR in detection of aquatic pathogens[J]. Journal of Ludong University (Natural Science Edition), 2020, 36(1): 48-54 (in Chinese).

[16] 陈京, 于永翔, 张正, 等. 基于*toxR*基因的轮虫弧菌荧光定量微流控快速检测技术的建立[J]. 水产学报, 2020, 44(12): 2066-2075.
Chen J, Yu Y X, Zhang Z, *et al*. Establishment of a fluorescence quantitative microfluidic rapid detection technique for *Vibrio rotiferianus* based on *toxR* gene[J]. Journal of Fisheries of China, 2020, 44(12): 2066-2075 (in Chinese).

[17] Urban M, Cuzick A, Rutherford K, *et al*. PHI-base: a new interface and further additions for the multi-species pathogen-host interactions database[J]. Nucleic Acids Research, 2017, 45(D1): D604-D610.

[18] Kanehisa M. Inferring antimicrobial resistance from pathogen genomes in KEGG[M]//Mamitsuka H. Data mining for systems biology. Methods in bolecular biology. New York: Humana Press, 2018, 1807: 225-239.

[19] Liu B, Zheng D D, Jin Q, *et al*. VFDB 2019: a comparat-

ive pathogenomic platform with an interactive web interface[J]. Nucleic Acids Research, 2019, 47(D1): D687-D692.

[20] Bayliss S C, Verner-Jeffreys D W, Ryder D, *et al*. Genomic epidemiology of the commercially important pathogen *Renibacterium salmoninarum* within the Chilean salmon industry[J]. Microbial Genomics, 2018, 4(9): e000201.

[21] Austin B, Austin D A. Bacterial fish pathogens: disease of farmed and wild fish[M]. 6th ed. Chichester: Springer International Publishing, 2016.

[22] 房海, 陈翠珍, 张晓君. 水产养殖动物病原细菌学 [M]. 北京: 中国农业出版社, 2010: 171-693.
Fang H, Chen C Z, Zhang X J. Aquacultural animal pathogenic bacteriology[M]. Beijing: China Agricultural Press, 2010: 171-693 (in Chinese).

[23] Parte A C, Carbasse J S, Meier-Kolthoff J P, *et al*. List of prokaryotic names with standing in nomenclature (LPSN) moves to the DSMZ[J]. International Journal of Systematic and Evolutionary Microbiology, 2020, 70(11): 5607-5612.

[24] Parte A C. LPSN-list of prokaryotic names with standing in nomenclature (bacterio. net), 20 years on[J]. International Journal of Systematic and Evolutionary Microbiology, 2018, 68(6): 1825-1829.

[25] 陈小辉, 刘心松, 左朝树, 等. 分布式并行数据库中基于调度的多线程通信模型之研究[J]. 小型微型计算机系统, 2005, 26(4): 604-608.
Chen X H, Liu X S, Zuo C S, *et al*. Study on multithreaded-communication-model based on scheduling in distributed and parallel database system[J]. Mini-Micro Systems, 2005, 26(4): 604-608 (in Chinese).

[26] 左朝树. 基于寄生式故障检测的分布式并行服务器系统容错技术 [D]. 成都: 电子科技大学, 2005.
Zuo C S. Fault tolerant technology based on autoecious fault detection in distributed parallel server system[D]. Chengdu: University of Electronic Science and Technology of China, 2005 (in Chinese).

[27] 许鑫, 时雷, 何龙, 等. 基于NoSQL数据库的农田物联网云存储系统设计与实现[J]. 农业工程学报, 2019, 35(1): 172-179.
Xu X, Shi L, He L, *et al*. Design and implementation of cloud storage system for farmland internet of things based on NoSQL database[J]. Transactions of the Chinese Society of Agricultural Engineering, 2019, 35(1): 172-179 (in Chinese).

[28] McDonald D, Price M N, Goodrich J, *et al*. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea[J]. The ISME Journal, 2012, 6(3): 610-618.

[29] Edgar R C. Search and clustering orders of magnitude faster than BLAST[J]. Bioinformatics, 2010, 26(19): 2460-2461.

[30] Caporaso J G, Kuczynski J, Stombaugh J, *et al*. QIIME allows analysis of high-throughput community sequencing data[J]. Nature Methods, 2010, 7(5): 335-336.

[31] Bokulich N A, Kaehler B D, Rideout J R, *et al*. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin[J]. Microbiome, 2018, 6: 90.

[32] Camacho C, Coulouris G, Avagyan V, *et al*. BLAST+: architecture and applications[J]. BMC Bioinformatics, 2009, 10: 421.

[33] Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data[J]. Bioinformatics, 2012, 28(24): 3211-3217.

[34] Ji D J, Ye W, Chen H F. Revealing the binding mode between respiratory syncytial virus fusion protein and benzimidazole-based inhibitors[J]. Molecular BioSystems, 2015, 11(7): 1857-1866.

[35] Jolliffe I T, Stephenson D B. Forecast verification: a practitioner's guide in atmospheric science[M]. 2nd ed. Chichester: John Wiley & Sons, 2012.

[36] Sing T, Sander O, Beerenwinkel N, *et al*. ROCR: Visualizing classifier performance in R[J]. Bioinformatics, 2005, 21(20): 3940-3941.

[37] Sun F L, Wang C Z, Chen L J, *et al*. The intestinal bacterial community of healthy and diseased animals and its association with the aquaculture environment[J]. Applied Microbiology and Biotechnology, 2020, 104(2): 775-783.

[38] 郑嘉来, 阎永伟, 唐姝, 等. 哈维弧菌16S rRNA基因拷贝数的种内变异[J]. 生物学杂志, 2015, 32(6): 6-11.
Zheng J L, Yan Y W, Tang S, *et al*. Intra-species variation of 16S rRNA gene copy number of *Vibrio harveyi*[J]. Journal of Biology, 2015, 32(6): 6-11 (in

Chinese).

[39] Caporaso J G, Bittinger K, Bushman F D, *et al*. PyNAST: a flexible tool for aligning sequences to a template alignment[J]. Bioinformatics, 2010, 26(2): 266-267.

[40] Schliep K P. phangorn: phylogenetic analysis in R[J]. Bioinformatics, 2011, 27(4): 592-593.

[41] Yu G C. Using ggtree to visualize data on tree-like structures[J]. Current Protocols in Bioinformatics, 2020, 69(1): e96.

[42] Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R[J]. Bioinformatics, 2019, 35(3): 526-528.

[43] Edgar R C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing[J]. BioRxiv, 2016.

[44] Franklin A M, Brinkman N E, Jahne M A, *et al*. Twenty-first century molecular methods for analyzing antimicrobial resistance in surface waters to support One Health assessments[J]. Journal of Microbiological Methods, 2021, 184: 106174.

[45] Edgar R. Taxonomy annotation and guide tree errors in 16S rRNA databases[J]. PeerJ, 2018, 6: e5030.

[46] Yu X L, Zhou J Y, Song W, *et al*. SCycDB: a curated functional gene database for metagenomic profiling of sulphur cycling pathways[J]. Molecular Ecology Resources, 2021, 21(3): 924-940.

[47] 高晓建, 姚东瑞, 孙晶晶, 等. 4株长牡蛎(*Crassostrea gigas*)致病性哈维氏弧菌(*Vibrio harveyi*)鉴定及其毒力基因检测[J]. 海洋湖沼通报, 2015(3): 87-96.
Gao X J, Yao D R, Sun J J, *et al*. Identification of 4 pathogenic *Vibrio harveyi* strains isolated from diseased Oyster (*Crassostrea gigas*) and detection of their virulence genes[J]. Transactions of Oceanology and Limnology, 2015(3): 87-96 (in Chinese).

[48] 徐鹏昊, 罗武松, 何恩明, 等. 16S rDNA特异性引物设计优化及其在淞江鲈体表微生物鉴定中的应用[J]. 复旦学报 (自然科学版), 2018, 57(1): 59-67, 78.
Xu P H, Luo W S, He E M, *et al*. Improvement of specific primer design on 16S rDNA and its application on skin microbes detection of *Trachidermus fasciatus*[J]. Journal of Fudan University (Natural Science), 2018, 57(1): 59-67, 78 (in Chinese).

[49] Brown M V, Ostrowski M, Grzymski J J, *et al*. A trait based perspective on the biogeography of common and abundant marine bacterioplankton clades[J]. Marine Genomics, 2014, 15(5): 17-28.

[50] Kim Y B, Okuda J, Matsumoto C, *et al*. Identification of *Vibrio parahaemolyticus* strains at the species level by PCR targeted to the *toxR* gene[J]. Journal of Clinical Microbiology, 1999, 37(4): 1173-1177.

[51] 刘巍 郭海朋 董鹏生, 等. 别样玫瑰变色杆菌(*Aliiroseovarius* sp.)Z3基因组测序及比较基因组分析[J]. 热带海洋学报, 2021.
Liu W, Guo H P, Dong P S, *et al*. Draft genome sequence and comparative genome analysis of *Aliiroseovarius* sp. Z3[J]. Journal of Tropical Oceanography, 2021 (in Chinese).

[52] Jinam T, Kawai Y, Kamatani Y, *et al*. Genome-wide SNP data of Izumo and Makurazaki populations support inner-dual structure model for origin of Yamato people[J]. Journal of Human Genetics, 2021, 66(7): 681-687.

[53] Gu W, Deng X D, Lee M, *et al*. Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids[J]. Nature Medicine, 2021, 27(1): 115-124.

[54] 中华医学会检验医学分会临床微生物学组, 中华医学会微生物学与免疫学分会临床微生物学组, 中国医疗保健国际交流促进会临床微生物与感染分会. 宏基因组高通量测序技术应用于感染性疾病病原检测中国专家共识[J]. 中华检验医学杂志, 2021, 44(2): 107-120.
Clinical Microbiology Group of Chinese Society of Laboratory Medicine, Clinical Microbiology Group of Chinese Society of Microbiology and Immunology, Society of Clinical Microbiology and Infection of China International Exchange and Promotion Association for Medical and Healthcare. Chinese expert consensus on metagenomics next-generation sequencing application on pathogen detection of infectious diseases[J]. Chinese Journal of Laboratory Medicine, 2021, 44(2): 107-120 (in Chinese).

DPiWE: a curated database for pathogenic bacteria involved in water environment

DONG Pengsheng [1,2], GUO Haipeng [1,2], WANG Yanting [1,2], CHENG Huangwei [1,2], WANG Kai [1,2], HONG Man [1,2], HOU Dandi [1,2], WU Yuhua [3], ZHANG Demin [1,2*]

(1. *State Key Laboratory For Managing Biotic and Chemical Threats to the Quality and Safety of Agro-products*, *Ningbo University*, *Ningbo* 315211, *China*;
2. *Collaborative Innovation Center for Zhejiang Marine High-efficiency and Healthy Aquaculture*, *School of Marine Sciences*, *Ningbo University*, *Ningbo* 315211, *China*;
3. *East China Sea Forecast Center of State Oceanic Administration*, *Shanghai* 200136, *China*)

Abstract: Pathogenic bacteria in the water environment are mainly monitored in the public health, food safety, aquaculture and other industries due to their major threats to the health of humans and aquatic animals, and the biosafety of aquatic products. However, pathogenic database involved in water environment pathogen is mainly constructed according to independent disciplines, and scattered in the fields of clinical medicine and aquatic animal diseases, which can no longer meet the high-throughput identification and biosafety evaluation of pathogenic bacteria involved in the water environment in the regional scale or ecological perspective. In this study, a database of pathogenic bacteria involved in water environment (DPiWE) was constructed by collecting the taxonomic information of pathogenic bacteria from humans, aquatic animals, mammals, plants, and cross-host comorbidities. A multi-threaded schedulable communication model and a multi-task mode global sequence matching algorithm were developed to construct DPiWE. The database collected 9 070 pathogenic bacteria strains, which belong to 14 phyla, 27 classes, 54 orders, 116 families, 221 genera and 1 097 species. The corresponding 16S rRNA gene sequences, host information and infection types of these strains were also collected in DPiWE. This database was deployed at a website (http://dayuz.com/) with the functions including web user management, pathogenic information retrieval, sequence upload, storage and alignment, and visualization of annotation result. Two examples were used to test the functions of DPiWE. The first example showed that, DPiWE can accurately construct a phylogenetic network of an unidentified bacterium (strain DS10−D19) isolated from cultural seawaters, according to its 16S rRNA gene sequence, and identified it as *Photobacterium leiognathid*. The result of network also showed that the network structure of strain DS10−D19 was similar to *P. leiognathid* and *P. angustum*. The second example showed that the compositions of pathogens in the intestines of three mariculture animals were significantly different through annotating the high-throughput sequencing data using DPiWE, and the rearing water in diseased groups had potential risk of spreading the comorbid pathogenic bacteria of human and fish. The DPiWE and its supporting data analysis process can provide new ideas and data foundations for high-throughput detection of the biosafety of water environment, protecting health of fishery ecology, and controlling diseases of aquatic animals, in the future.

Key words: pathogenic bacteria; 16S rRNA gene; multi-thread scheduling database; identify; annotation; water environment

Corresponding author: ZHANG Demin. E-mail: zhangdemin@nbu.edu.cn

Funding projects: National Key Research and Development Program of China (2016YFC1402205); National Natural Science Foundation of China (31672658); Major Agricultural Projects in Ningbo (2017C110001)