基于“天河二号”的水产病原生物信息分析平台构建及其在水产病原分析中的应用

方翔，李宁求，付小哲，李凯彬，林强，刘礼辉，石存斌，吴淑勤

中国水产科学研究院珠江水产研究所，农业部渔用药物创制重点实验室，广东省水产动物免疫技术重点实验室，广州 510380

摘要: 作为生命科学的关键组成，生物信息学已被广泛地应用于基因组学、转录组学和蛋白质组学中。然而，生物信息分析平台的构建需要高性能计算机而非普通的个人电脑，从而极大地限制了生物信息学在水产科学中的应用。本研究基于“天河二号”超级计算机，构建了水产病原生物信息分析平台。该平台由基因组与转录组测序数据分析、蛋白质结构预测和分子动力学模拟 3 个功能模块组成。为了验证该平台的实用性，以水生动物病原微生物为例进行了生物信息学分析。通过 Blast 检索、GO 和 InterPro 注释，鉴定了约氏黄杆菌(*Flavobacterium johnsoniae*)M168 株的功能基因并对其进行了注释。通过同源模建，构建了草鱼呼肠孤病毒(Grass carp reovirus，GCRV)HZ-08 的 5 个小节段的蛋白结构模型。对嗜水气单胞菌(*Aeromonas hydrophila*)外膜蛋白 A 进行了分子动力学模拟，并观察了平衡过程中系统温度、总能量、均方根偏差和环区构象的变化。以上结果均显示本研究成功建立了在“天河二号”超级计算机上运行的水产病原生物信息分析平台。此项研究将为其他学科生物信息分析平台的构建提供思路和线索。

关键词: 生物信息学；天河二号；水产病原

Construction and application of bioinformatic analysis platform for aquatic pathogen based on the MilkyWay-2 supercomputer

Xiang Fang, Ningqiu Li, Xiaozhe Fu, Kaibin Li, Qiang Lin, Lihui Liu, Cunbin Shi, Shuqin Wu

Key Laboratory of Aquatic Animal Immune Technology of Guangdong Province, Key Laboratory of Fishery Drug Development of Ministry of Agriculture, Pearl River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou 510380, China

Abstract: As a key component of life science, bioinformatics has been widely applied in genomics, transcriptomics, and proteomics. However, the requirement of high-performance computers rather than common personal computers for constructing a bioinformatics platform significantly limited the application of bioinformatics in aquatic science.

收稿日期: 2015−01−16; **修回日期:** 2015−04−01
基金项目: 广州市珠江科技新星专项(编号: 2012J2200078)和中国水产科学研究院院级中央级公益性科研院所基本科研业务费(编号: 2013A0609)资助
作者简介: 方翔，助理研究员，研究方向：水产生物信息学。E-mail: wayj86@gmail.com
通讯作者: 李宁求，副研究员，研究方向：水产病害。E-mail: liningq@126.com；
吴淑勤，研究员，研究方向：水产病害。E-mail: wushuqing001@21cn.com
DOI: 10.16288/j.yczz.15-038
网络出版时间: 2015-5-20 16:33:19
URL: http://www.cnki.net/kcms/detail/11.1913.R.20150520.1633.001.html

In this study, we constructed a bioinformatic analysis platform for aquatic pathogen based on the MilkyWay-2 supercomputer. The platform consisted of three functional modules, including genomic and transcriptomic sequencing data analysis, protein structure prediction, and molecular dynamics simulations. To validate the practicability of the platform, we performed bioinformatic analysis on aquatic pathogenic organisms. For example, genes of *Flavobacterium johnsoniae* M168 were identified and annotated via Blast searches, GO and InterPro annotations. Protein structural models for five small segments of grass carp reovirus HZ-08 were constructed by homology modeling. Molecular dynamics simulations were performed on out membrane protein A of *Aeromonas hydrophila*, and the changes of system temperature, total energy, root mean square deviation and conformation of the loops during equilibration were also observed. These results showed that the bioinformatic analysis platform for aquatic pathogen has been successfully built on the MilkyWay-2 supercomputer. This study will provide insights into the construction of bioinformatic analysis platform for other subjects.

Keywords: bioinformatics; MilkyWay-2; aquatic pathogen

生物信息学是计算机科学与生命科学相融合形成的交叉学科，是当今生命科学的核心领域和发展前沿[1,2]。基因组学、转录组学、蛋白质组学和结构生物学等新兴学科产生的海量数据都需要通过生物信息学来进行采集、处理、分析、阐释和存储。生物信息学已经和这些组学技术形成了密不可分的整体，相关的分析方法也随着组学技术的进步而不断发展和改进[3,4]。

近年来，基因组学、转录组学和蛋白质组学等组学技术以及蛋白质结构分析技术开始逐步应用到水产病害科学上, 取得了许多突破性的研究成果。2010 年，Cheng 等[5]通过冷冻电子显微镜(Cryo-Electron Microscopy)和生物信息学结合的方法，成功构建了水生呼肠孤病毒(Aquareovirus)粒子的骨架模型(Backbone model)。与此同时，Verma 等[6]通过同源模建、分子对接和分子动力学模拟等方法，研究斑节对虾(*Penaeus monodon*)和白斑综合征病毒(White spot syndrome virus, wssv)之间的相互作用，揭示了白斑综合征致病分子机理。2013 年，为了研究大黄鱼(*Pseudosciaena crocea*)的抗病毒分子机制，Mu 等[7]对大黄鱼的脾脏转录组进行了测定，并通过 KEGG 注释分析鉴别出了 5389 个有注释信息的非冗余基因(Unigene)以及与免疫相关的 Pathway 信息。2014 年，Wiens 等[8]测定了嗜冷黄杆菌(*Flavobacterium psychrophilum*)CSF259-93 株的全基因组序列，为抗细菌性冷水病(Bacterial cold water disease)的虹鳟的筛选提供了遗传背景。最近，Khemiri 等[9]通过蛋白质组测序技术研究了水生病害菌嗜肺军团病杆菌(*Legionella pneumophila*)在生物膜形成过程中的蛋白调控，发现毒力因子的水平和对氧化压力的响应分别出现了降低和升高。在这些研究中，生物信息学都起到了极为关键的作用，表明生物信息学已逐渐成为解决水产科学问题的重要工具。

然而由于分析过程涉及大规模运算，需要消耗较多的 CPU 与内存资源(运行分子动力学模拟甚至需要进行并行运算)，搭建生物信息分析平台往往需要高性能计算机，常规的个人电脑几乎无法进行。目前，对国内许多科研单位而言，生物信息分析大都依赖相关的生物技术公司，或者自行购买大型服务器，极大地限制了生物信息学的实际应用。另一方面，随着国家超级计算天津中心、深圳中心、长沙中心、济南中心和广州中心的建立，超级计算服务在我国蓬勃发展，已有越来越多的科研单位向超级计算中心申请了计算资源进行高性能运算和数据分析。因此，如果能借助超级计算平台的强大运算能力，建立具有学科特色的生物信息学分析平台，既能实现高效的分析运算和安全的数据存储，又能极大地降低运算成本。

“天河二号”是由国防科技大学研制的超级计算机平台，是当今世界上速度最快的超级计算机，坐落于国家超级计算广州中心[10,11]。“天河二号”共有 16 000 个运算节点，共 312 万个计算核心，峰值计算速度达到每秒 5.49 亿亿次，计算能力连续三年位居世界第一，同时拥有强大的网络传输和数据存储能力[12,13]。经过为期半年多的运算调试，“天河二号”目前已面向社会和广大科研单位开放申请使用，能够提供高性能计算所需的计算资源。本研究将基于“天河二号”构建水产病原生物信息分析平台，其中

包含核酸测序数据分析、蛋白结构研究及分子动力学模拟三大模块，能够独立完成生物信息学的标准化分析。同时，通过对几种水生动物病原进行分析来验证平台的实用性。本平台的成功构建将为其他学科建立具有学科特色的生物信息学分析平台提供思路和参考。

1 材料和方法

1.1 配置和完善平台基础架构

“天河二号”超级计算机运行 64 位的 Ubuntu 操作系统，每个运算节点含有 2 个多核中央处理器和 3 个众核加速器共计 16 个 CPU 核心, 内存为 88 GB。本平台运算时需要 4 个节点，所需硬盘空间约为 10 TB。平台中已整合了生物信息分析常用的脚本编程语言 Perl 和 Python。同时, 为了流畅运行 Blast2go[14] 和 InterProscan[15]基因功能注释软件，家目录下还安装了 JAVA 语言开发工具包 JDK(java development kit)和 MYSQL 数据库管理程序。此外，系统环境变量也专门进行过设置和优化，以确保各个生物信息学软件的正常运行。

1.2 核酸测序数据分析模块

核酸测序数据分析模块包含基因组和转录组高通量测序数据分析两个子模块。由于这两个组学的结果都是核酸数据，而且很多分析都用到相似的软件和数据库，因此将二者整合在一个大的分析模块中。基因组测序数据分析首先采用软件 Velvet[16]或 SOAPdenovo[17]对高通量测序产生的 reads 进行组装，拼接成长的 contig 以及更长的 scaffold。然后用 GeneMarks[18](针对原核生物)或 Augustus[19](针对真核生物)软件进行基因预测，获得基因组中编码基因的 DNA 序列及其编码的氨基酸序列。核酸和蛋白序列比对分析则通过 Blast 程序进行，相对应的数据库包括常用的 Nt(核苷酸数据库)、Nr(非冗余蛋白序列数据库)、PDB(Protein data bank，蛋白质晶体结构数据库)、Swissprot(经过校验的蛋白数据库)，GO(基因本体数据库)和 KEGG(京都基因与基因组百科全书数据库)。此外，还包含具有特色的毒力基因数据库 VFDB[20](Virulence factors of bacterial pathogens)和耐药菌株抗性基因数据库 ARDB[21](Antibiotic resistance genes database)等。以上数据库均已实现本地化。基因功能注释包含 GO(Gene ontology)注释和 InterPro 注释, 分别通过软件 Blast2go[14]和 InterProscan[15]进行。转录组数据分析则首先通过软件 Trinity[22](有参考基因组)或 Soap2[23](无参考基因组)进行序列从头组装或比对；基因表达差异分析通过软件 Cufflinks[24]进行，表达模式聚类分析通过软件 Cluster 进行。

1.3 蛋白结构研究模块

蛋白结构研究模块包含结构预测和结构分析两个子模块。结构预测子模块首先通过软件 ClustalW[25]将模板序列和目标序列进行多重序列比对，为结构预测做准备；然后采用蛋白模建程序 Moderller[26]，通过同源模建方法构建目标蛋白的三维结构模型。结构分析子模块包括二级结构分析、结构域分析、复合物结合模式分析、极性相互作用分析和抗原位点预测等，主要通过可视化操作软件 VMD[27]和 Pymol[28]进行；氨基酸点突变预测通过采用软件 SwissPDBviewer[29]进行。

1.4 分子动力学模拟模块

分子动力学模拟模块包含分子预处理和模拟运算两个子模块。其中分子预处理模块主要通过软件 VMD 对需要进行运算的分子进行旋转平移，加水框和加抗衡离子等预处理, 并对模拟后的结果进行分析和计算。在本研究中主要分析了模拟过程中平台的温度、压力和整个分子的均方根偏差 (Root mean square deviation RMSD)；模拟运算子模块主要通过软件 NAMD[30]进行，对能量最小化以后的分子进行平衡，自由动力学模拟和拉伸分子动力学模拟，研究生物大分子(DNA，蛋白质乃至生物膜)在不同温度和压强下的动态变化。

1.5 实验材料

约氏黄杆菌(*Flavobacterium johnsoniae*)M168 株和鲫鱼诺卡氏菌*(Nocardia seriolea)*分别从患烂鳃病的草鱼腮部和患诺卡氏病的大口黑鲈肾脏分离纯化得到；提取总 DNA 后进行全基因组测序, 获得基因组序列。草鱼呼肠孤病毒(Grass carp reovirus，GCRV) HZ-08 株从患有出血病的草鱼的肝肾组织中分离纯化得到；提取病毒基因组并进行测序，获取小节段(S7~S11)DNA 序列后将其翻译为相应的氨基酸序列。嗜水气单胞菌(*Aeromonas hydrophila*)GYK1 株从患出血病的鳜鱼肾脏中分离纯化得到；提取其总 DNA 后，对外膜蛋白 A(Out membrance protein A, *OMPA*)基因进行扩增及

测序，并将测序后的 DNA 序列翻译为氨基酸序列。

2　结果与分析

2.1　平台构建情况

经过为期 4 个月的架设与试运算，核酸测序数据分析、蛋白结构研究和分子动力学模拟三大模块均已配置完成，表明水产病原生物信息分析平台已成功构建在“天河二号”超级计算机上。本平台没有采用图形界面，通过虚拟专用网络(Virtual private network，VPN)以安全套接层(Secure sockets layer，SSL)方式连接到“天河二号”提交运算任务。平台的工作流程图如图 1 所示。截至目前，该平台的运算量已超过 4 万 CPU 小时，已对包括约氏黄杆菌、鰤鱼诺卡氏菌、水生呼肠孤病毒和嗜水气单胞菌等在内的多种水生动物病原进行了生物信息学分析，获得了初步的研究成果，并证实了平台的实用性。

2.2　约氏黄杆菌 M168 和鰤鱼诺卡氏菌的全基因组测序数据分析

约氏黄杆菌 M168 是本研究团队新近分离并进行基因组测序的一株病原菌，它可导致草鱼烂鳃病[31]。为了验证核酸测序数据分析模块的功能，本文对该菌的全基因组测序数据(大小为 5.39 Mb)进行了分析。GeneMarks 基因预测结果表明，该基因组共包含 4797 个编码基因, 平均长度为 992 bp。为了识别编码基因，对 M168 基因的核酸序列和蛋白质序列分别进行了 Blast 比对，选取的数据库分别为 Nt 和 Nr。每次 Blast 运算只采用一个节点中的一个 CPU 核心, 运行时间约为 72 小时。Blast 完成后，以 1E-5 为期望值对结果进行筛选。结果共标记出 2964 个基因的核酸序列，以及 4499 个蛋白质的氨基酸序列。此外，通过对 VFDB 数据库的检索，在基因组中共发现了 647 个毒力基因。在 Blast 比对结果的基础上, 采用 Blast2go 程序对约氏黄杆菌 M168 的全部基因进行了注释。Blast2go 运算同样只采用单节点单核心进行, 运行时间约为 48 小时。GO 注释包含细胞组分(Cellular component)、分子功能(Molecular function)和生物过程(Biological process)3 个层面[32]。在细胞组分方面，基因主要集中在细胞膜和胞质上(图 2A)；在分子功能方面，大部分基因与 ATP 结合和 DNA 结合相关；在生物过程方面，参与氧化还原和转录调控的基因数目最多。InterPro 注释同样在 Blast 比对的基础上采用单节点单核心运算，通过 InterProscan 软件包内置的多个数据库和软件，历经约 40 小时，预测并注释了 M168 蛋白所属的蛋白家族及其内在的关键结构域。

鰤鱼诺卡氏菌能够感染虹鳟、乌鳢、大黄鱼和大口黑鲈等重要水生经济动物，对水产养殖业造成了较大的威胁[33]。采用同样的方法和流程，本文分析了鰤鱼诺卡氏菌的全基因组序列(7.52 Mb), 并对其 8221 个编码基因进行了 GO 注释(图 2B)和 InterPro 注释，为进一步研究相关蛋白的结构和功能奠定了坚实的基础。

2.3　呼肠孤病毒 HZ-08 小节段蛋白的结构预测和分析

草鱼呼肠孤病毒 HZ-08 是 Wang 等[34]近年来分离得到的一个新毒株，它可导致草鱼出血病。该病毒基

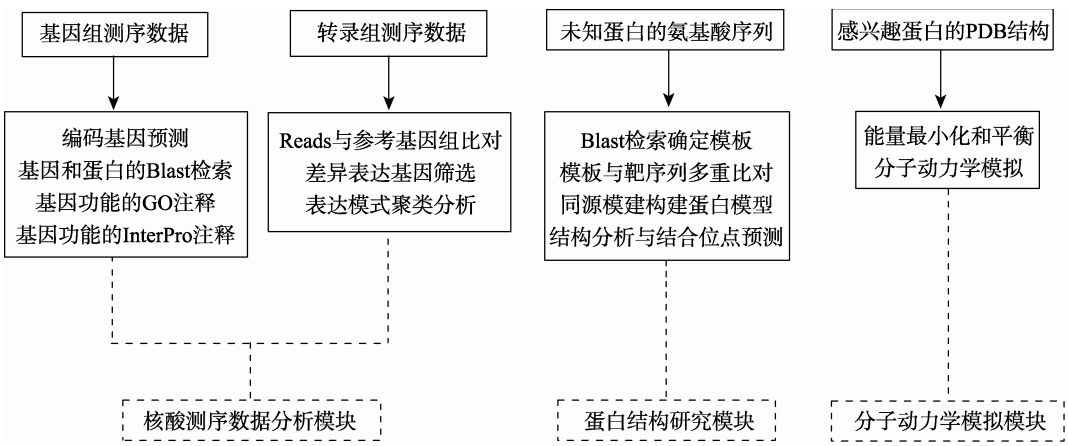

图 1　生物信息分析平台工作流程图

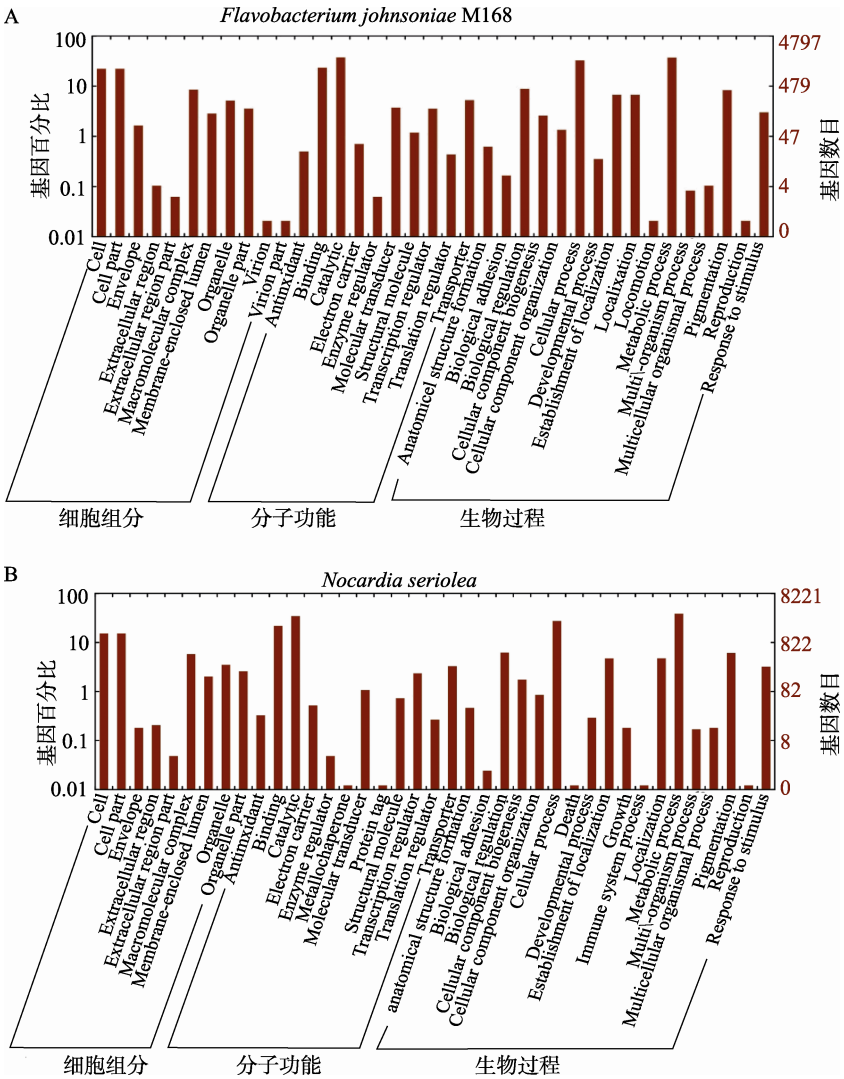

图 2　基因功能的 GO 注释结果

A：约氏黄杆菌 M168 基因的 GO 注释；B：鳞鱼诺卡氏菌基因的 GO 注释。

因组共分为 11 个节段，其中 3 个(S1~S3)为大节段，3 个(S4~S6)为中节段，剩下的 5 个(S7~S11)为小节段。与大节段和中节段相比，小节段的保守性最低，目前还没有对其结构和功能进行研究的报道。为了验证蛋白结构研究模块的功能，本文对 HZ-08 的 5 个小节段结构域进行了结构预测。S7-S11 节段的长度分别为 512、361、418、345 和 310 个氨基酸残基。首先，对 S7~S11 的氨基酸序列分别进行了 BlastP 检索，从 PDB 数据库中选取了比对最优的百日咳杆菌粘附素毒力因子(PDB 编号为 1DAB)、α-1,2-甘露糖苷酶(PDB 编号为 1DL2)、呼肠孤病毒核心结构(PDB 编号为 1EJ6)、人源赖氨酸转甲基酶复合物(PDB 标号为 3RIB)和外甘露糖苷酶(PDB 编号为 1UUQ) 5 个已知的晶体结构，分别作为 S7、S8、S9、S10 和 S11 节段的模板。然后，分别将目标序列与模板序列进行了多重比对，在比对基础上通过 Modeller 进行了同源模建。每个蛋白分别构建 8 个备选模型；在单节点单核心的条件下，只需 30 min 即可完成 8 个备选模型的构建及后续的结构优化。5 次运算共获得 40 个备选模型，选取每个蛋白得分最优的备选模型作为最终的结构模型(图 3)。其中，S7 节段模型的正面主要由平行的 β 片层组成了一个略为弯曲的弧形，背面是无规则卷曲。这一特殊的分子构象与人源血小板糖蛋白 GPIbα 的 N 末端结构域相类

似[35]，表明这些 β 片层有可能是参与受体配体结合的关键部位。相比之下，S8~S11 节段则主要由 α 螺旋及无规则卷曲组成。其中，S8 的结构也形成了一个弯曲的拱形, 向内凹陷的部分有可能包含结合位点。S9~S11 的构象稍显复杂，是由多个 α 螺旋组成的球蛋白结构，其结合位点可能位于表面的 α 螺旋上。

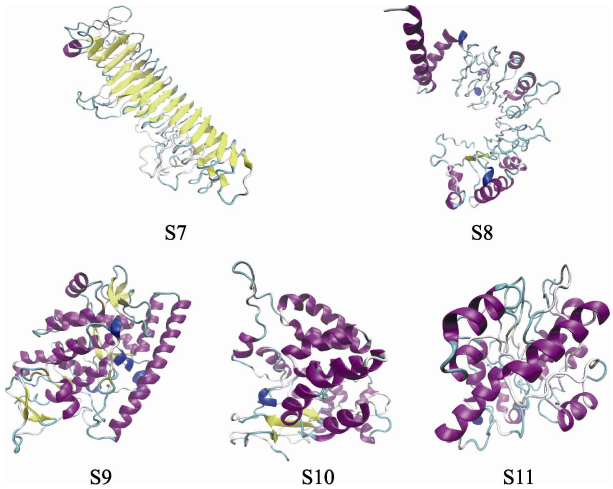

图 3　HZ-08 小节段(S7~S11)的蛋白结构模型

2.4　嗜水气单胞菌外膜蛋白 A 的分子动力学模拟

嗜水气单胞菌是一种典型的人–兽–鱼共患病病原菌, 它可引起多种水产动物的败血症和人类腹泻[36]。其外膜蛋白 A(OMPA)是一种主效的毒力蛋白，具有许多重要的生物学功能[37]。为了验证分子动力学模拟模块的功能，本文对嗜水气单胞菌 OMPA 的三维结构模型进行了模拟。首先，通过 VMD 软件的 Autopsf 插件来为 OMPA 添加缺失的氢原子，然后将其浸没在一个尺寸为 11.5 nm×7.3 nm×6.1 nm 长方形 TIP3 水盒子(Water box)中，并加入 53 个钠离子和 52 个氯离子来中和平台的电荷, 使离子的浓度与生理盐水的浓度一致。系统总原子数为 67 554。接下来，分别在固定蛋白质的全部原子和放开原子约束的情况下，通过 NAMD 软件分别进行了各 5000 步能量最小化，每次能量最小化采用单节点单核心，运算时间需要约半个小时。此后，在 20 ps(皮秒)内将温度从 0 K 升至 310 K；在此基础上，以 2 fs(飞秒)为积分步长，在周期性边界条件(Periodic boundary condition)下对 OMPA 进行了时长为 5 ns(纳秒)的平衡模拟。在平衡过程中, 温度用 Langevin 动力学控制在 310 K, 压力用 Langevin 活塞方法稳定在 1 atm。平衡需采用并行运算，在 4 个节点 64 个 CPU 核心的规模下，所需运算时间约为 40 min。平衡过程中系统温度、总能量和 RMSD 的变化过程如图 4A 所示，从中可见随着模拟的进行，系统温度和总能量分别保持在 310 K 和−270000 kcal/mol，而重原子的 RMSD 在开始阶段呈逐步上升趋势，经过约 4.6 ns 后逐步变得稳定。OMPA 蛋白的构象变化过程如图 4B 所示。OMPA 上位于分子左侧的 4 个环区的构象逐步趋于稳定，与 RMSD 曲线的变化相一致，表明这些环区已逐步接近其真实生理条件下的结构，能够用来进一步研究其与细胞表面受体之间的相互作用。

3　讨　论

生物信息学是当今生命科学中方兴未艾的重要学科，随着生命科学的不断发展，其应用也日趋广泛。然而，与之形成鲜明对比的是，目前国内关于生物信息平台的研究寥寥无几。2007 年，赵友杰等[38]构建了基于多层结构模型的新城疫病毒生物信息分析平台，发现其模型能够较好地解决生物数据更新、数据集成、应用集成等问题；但该平台只能实现序列的检索、聚类和 BLAST 分析，其功能较为简单。2009 年，马相如等[39]建立了过基于局域网的生物信息学开发与应用平台，并集成了核酸和蛋白的一系列序列分析工具。然而，该平台主要注重生物信息的教学、软件应用和开发，并非为分析高通量测序的海量数据而设立；同时，该平台构建在 PC 上，运算能力不够强大。此外，截至目前国内尚无水产病原相关生物信息平台的报道。为了填补这一空白，本研究以“天河二号”超级计算机为基础构建了水产病原生物信息分析平台, 并通过对约氏黄杆菌 M168、鲈鱼诺卡氏菌、草鱼呼肠孤病毒 HZ-08 和嗜水气单胞菌分别进行了基因组测序数据分析、蛋白结构预测分析和分子动力学模拟，获得了初步的研究成果并证实了平台的实用性。与现有的生物信息平台相比, 本平台具有 3 个优势。第一是分析运算速度快。本平台采用单节点单核心运算时，比当今 6 万元级别的服务器(CPU 为英特尔 Xeon E5-2680v2)快 10%左右；而采用多节点并行运算时，速度则可达到后者的 1.5~150 倍(运算体系越大，所用节点数越多，差距越明显)，从而显著缩短了运算时间，提高了分析效率。第二是数据处理能力强。利用“天河

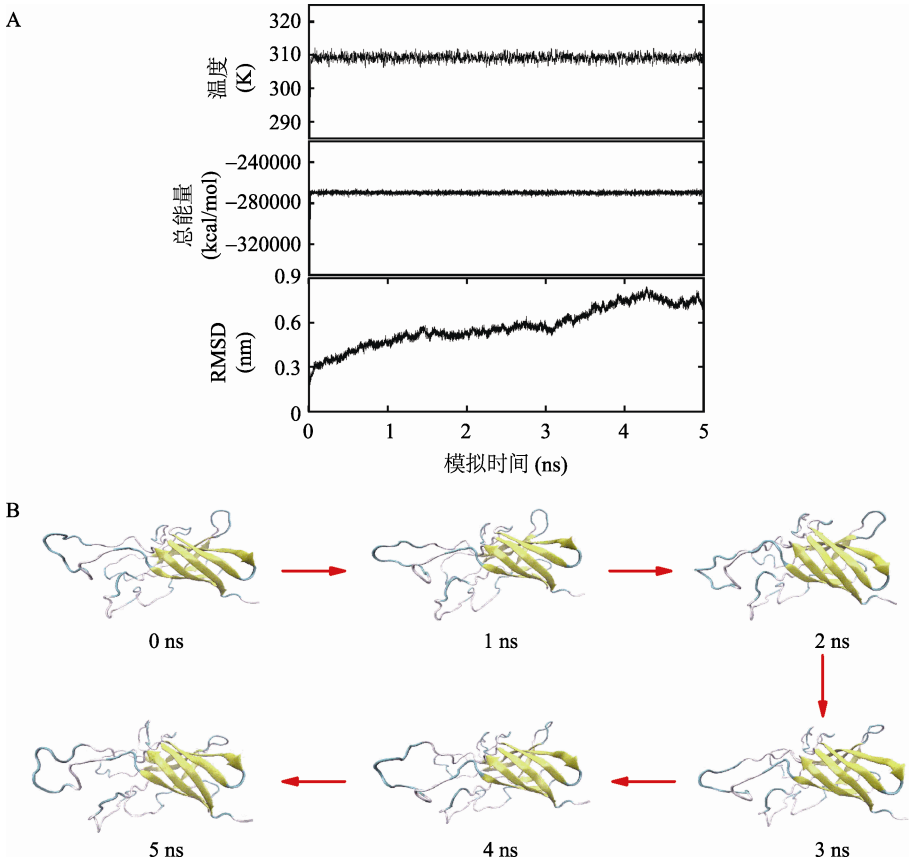

图 4　嗜水气单胞菌外膜蛋白 A 的分子动力学模拟。

A：OMPA 蛋白在平衡过程中系统温度、总能量和 RMSD 的变化；B：OMPA 蛋白在平衡过程中的构象变化。

二号”的超大内存和多个运算节点，本平台能够分析大型的基因组或转录组测序数据，以及进行超大分子体系(原子数目 1 千万以上)的分子动力学模拟。第三是分析工具全面，融合了基因组和转录组数据分析、蛋白结构建模以及分子动力学模拟的相关技术，能够沿着“基因组和转录组测序结果→差异表达基因→基因编码蛋白的结构→蛋白的动力学性质→蛋白的关键残基和分子间相互作用”这一线索进行持续探索，有利于深入研究水产病害背后的分子机制。

此外，依托“天河二号”的强大运算和存储能力，本平台还有较大的可扩展空间。首先，当前的信息学分析越来越注重数据的个性化分析，因为只有个性化分析才能深入地挖掘数据中蕴含的生物学机理。目前的分析平台中只包含生物信息学数据的标准化分析流程，个性化分析尚未能实现。因此，下一阶段我们将在现有的各个分析模块中加入更多的数据处理软件，绘图软件和运算脚本(主要是 perl 和 python 脚本)，以实现对数据的个性化分析。其次，现代组学的发展已经超出了基因组学，转录组学和蛋白组学的范畴，还包括代谢组学和宏基因组学等等。尤其是宏基因组学，作为研究某一特定环境中的全部微生物的新兴学科[40]，能够极大地加深人们对与环境相关的生命现象的理解。为此，我们将在当前的分析平台中加入宏基因组的分析模块，为宏基因组测序数据提供分析工具。第三，目前平台运行的是单机版的 Blast 和 InterProscan，只能在单节点单核心下运行，影响了计算效率。因此，下一阶段我们将安装能够实现并行运算的 mpiBlast 和 InterProscan，以及其他软件的并行版，利用“天河二号”的强大并行运算能力提高运算和分析效率。最后，目前本平台分析过的数据都只和水产病原相关，但病害学科仅仅是水产科学的一个领域；水产科学还包括遗传育种、资源生态、环境保护，水产生物技术和观赏渔业等众多研究领域，几乎每个领域都能与生物信息学关联起来。因此，未来我们将致力于对现有的功能进行扩展，添加例如真核生物基因组

和转录组分析软件，以及分子对接软件等生物信息分析工具，使该平台能够涵盖病害以外的其他水产学科，成为为整个水产科学服务的综合性生物信息分析平台，同时也可为相关科学问题的解决提供思路和线索。

参考文献

[1] Dolled-Filhart MP, Lee M, Jr., Ouyang CW, Haraksingh RR, Lin JC. Computational and bioinformatics frameworks for next-generation whole exome and genome sequencing. *ScientificWorldJournal*, 2013, 2013: Article ID 730210. [DOI]

[2] Bishop ÖT, Adebiyi EF, Alzohairy AM, Everett D, Ghedira K, Ghouila A, Kumuthini J, Mulder NJ, Panji S, Patterton HG. Bioinformatics education-perspectives and challenges out of Africa. *Brief Bioinform*, 2015, 16(2): 355–364. [DOI]

[3] Good BM, Su AI. Crowdsourcing for bioinformatics. *Bioinformatics*, 2013, 29(16): 1925–1933. [DOI]

[4] Hamada M. Fighting against uncertainty: an essential issue in bioinformatics. *Brief Bioinform*, 2014: 15(5): 748–767. [DOI]

[5] Cheng LP, Zhu J, Hui WH, Zhang XK, Honig B, Fang Q, Zhou ZH. Backbone model of an aquareovirus virion by cryo-electron microscopy and bioinformatics. *J Mol Biol*, 2010, 397(3): 852–863. [DOI]

[6] Verma AK, Gupta S, Verma S, Mishra A, Nagpure NS, Singh SP, Pathak AK, Sarkar UK, Singh M, Seth PK. Interaction between shrimp and white spot syndrome virus through PmRab7-VP28 complex: an insight using simulation and docking studies. *J Mol Model*, 2013, 19(3): 1285–1294. [DOI]

[7] Mu YN, Li MY, Ding F, Ding Y, Ao JQ, Hu SN, Chen XH. De novo characterization of the spleen transcriptome of the large yellow croaker (*Pseudosciaena crocea*) and analysis of the immune relevant genes and pathways involved in the antiviral response. *PLoS One*, 2014, 9(5): e97471. [DOI]

[8] Wiens GD, LaPatra SE, Welch TJ, Rexroad C, III, Call DR, Cain KD, LaFrentz BR, Vaisvil B, Schmitt DP, Kapatral V. Complete genome sequence of *Flavobacterium psychrophilum* strain CSF259–93, used to select rainbow trout for increased genetic resistance against bacterial cold water disease. *Genome Announc*, 2014, 2(5): e00889–14. [DOI]

[9] Khemiri A, Lecheheb SA, Chi Song PC, Jouenne T, Cosette P. Proteomic regulation during Legionella pneumophila biofilm development: decrease of virulence factors and enhancement of response to oxidative stress. *J Water Health*, 2014, 12(2): 242–253. [DOI]

[10] Liao XK. MilkyWay-2: back to the world Top 1. *Front Comput Sci*, 2014, 8(3): 343–344. [DOI]

[11] Liao XK, Xiao LQ, Yang CQ, Lu YT. MilkyWay-2 supercomputer: system and application. *Front Comput Sci*, 2014, 8(3): 345–356. [DOI]

[12] Pang ZB, Xie M, Zhang J, Zheng Y, Wang GB, Dong DZ, Suo G. The TH Express high performance interconnect networks. *Front Comput Sci*, 2014, 8(3): 357–366. [DOI]

[13] Xu WX, Lu YT, Li Q, Zhou EQ, Song ZL, Dong Y, Zhang W, Wei DP, Zhang XM, Chen HT, Xing JY, Yuan Y. Hybrid hierarchy storage system in MilkyWay-2 supercomputer. *Front Comput Sci*, 2014, 8(3): 367–377. [DOI]

[14] Conesa A, Gotz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics*, 2008, 2008: 619832. [DOI]

[15] Jones P, Binns D, Chang HY, Fraser M, Li WZ, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 2014, 30(9): 1236–1240. [DOI]

[16] Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, 2008, 18(5): 821–829. [DOI]

[17] Luo RB, Liu BH, Xie YL, Li ZY, Huang WH, Yuan JY, He GZ, Chen YX, Pan Q, Liu YJ, Tang JB, Wu GX, Zhang H, Shi YJ, Liu Y, Yu C, Wang B, Lu Y, Han CL, Cheung DW, Yiu SM, Peng SL, Zhu XQ, Liu GM, Liao XK, Li YR, Yang HM, Wang J, Lam TW, Wang J. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience*, 2012, 1: 18. [DOI]

[18] Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res*, 2001, 29(12): 2607–2618. [DOI]

[19] Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res*, 2005, 33(Suppl.2): W465–W467. [DOI]

[20] Chen LH, Xiong ZH, Sun LL, Yang J, Jin Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res*, 2012, 40(D1): D641–D645. [DOI]

[21] Liu B, Pop M. ARDB-Antibiotic resistance genes database. *Nucleic Acids Res*, 2009, 37(Suppl .1): D443–D447. [DOI]

[22] Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*, 2013, 8(8): 1494–1512. [DOI]

[23] Li RQ, Yu C, Li YR, Lam TW, Yiu SM, Kristiansen K, Wang J. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 2009, 25(15): 1966–1967. [DOI]

[24] Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, 2012, 7(3): 562–578. [DOI]

[25] Thompson JD, Gibson TJ, Higgins DG. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics*, 2002, Chapter 2: Unit 2.3. [DOI]

[26] Joo K, Lee J, Seo JH, Lee K, Kim BG. All-atom chain-building by optimizing MODELLER energy function using conformational space annealing. *Proteins*, 2009, 75(4): 1010–1023. [DOI]

[27] Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph*, 1996, 14(1): 33–38. [DOI]

[28] Lill MA, Danielson ML. Computer-aided drug design platform using PyMOL. *J Comput Aided Mol Des*, 2011, 25(1): 13–19. [DOI]

[29] Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, 1997, 18(15): 2714–2723. [DOI]

[30] Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. Scalable molecular dynamics with NAMD. *J Comput Chem*, 2005, 26(16): 1781–1802. [DOI]

[31] Zhou XX, Ding YT, Wang YB. Proteomics: present and future in fish, shellfish and seafood. *Rev Aquacult*, 2012, 4(1): 11–20. [DOI]

[32] Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 2004, 32(Suppl.1): D258–D261. [DOI]

[33] 孙东雨, 沈理, 金珊, 缪燕萍, 赵青松, 陈寅儿. 鲫鱼诺卡氏菌对乌鳢血液指标的影响. 基因组学与应用生物学, 2012, 31(3): 289–294. [DOI]

[34] Wang Q, Zeng WW, Liu C, Zhang C, Wang YY, Shi CB, Wu SQ. Complete genome sequence of a reovirus isolated from grass carp, indicating different genotypes of GCRV in China. *J Virol*, 2012, 86(22): 12466. [DOI]

[35] Dumas JJ, Kumar R, McDonagh T, Sullivan F, Stahl ML, Somers WS, Mosyak L. Crystal structure of the wild-type von Willebrand factor A1-glycoprotein Ibalpha complex reveals conformation differences with a complex bearing von Willebrand disease mutations. *J Biol Chem*, 2004, 279(22): 23327–23334. [DOI]

[36] 潘厚军, 吴淑勤, 董传甫, 石存斌, 叶美茜, 林天龙, 黄志斌. 鳜致病性嗜水气单胞菌 GYK1 株的鉴定、毒力及溶血性. 上海水产大学学报, 2004, 13(1): 23–29. [DOI]

[37] Confer AW, Ayalew S. The OmpA family of proteins: roles in bacterial pathogenesis and immunity. *Vet Microbiol*, 2013, 163(3–4): 207–222. [DOI]

[38] 赵友杰, 张剑峰, 曹永忠, 陆王红. 基于多层结构模型的生物信息分析平台研究. 计算机应用研究, 2007, 24(11): 55–56, 59. [DOI]

[39] 马相如, 王红梅, 顾延生, 葛继稳. 基于局域网的生物信息学应用与开发平台的建立. 计算机应用, 2009, 29(Suppl.1): 387–389, 392. [DOI]

[40] Sharon I, Banfield JF. Microbiology. Genomes from metagenomics. *Science*, 2013, 342(6162): 1057–1058. [DOI]

(责任编委: 赵方庆)