基于网格的分布式数据计算预测蛋白质相互作用关系

谢 江[1,2] 张 武[1,2] 梅 健[1]

(上海大学计算机工程与科学学院 上海 200072)[1] (上海大学系统生物技术研究所 上海 200444)[2]

摘 要 随着不同物种蛋白质相互作用关系数据的大量积累,蛋白质相互作用特别是大规模蛋白质相互作用的研究成为生命科学领域的又一个研究热点。目前互联网上存在大量分布式的、异构的蛋白质相互作用关系数据源,用户难以高效地整合这些数据源中的信息。本文针对多个大容量的蛋白质相互作用关系数据库,在数据网格 BD-Grid 环境下,提出了预测蛋白质相互作用关系通用结构框架,通过有效的元数据管理服务和合理的数据分类服务,屏蔽异构数据库间的差异,实现了用户对数据的透明存取,并在该数据平台上,实现了预测蛋白质相互作用关系的应用,使相关领域的研究人员能以生物信息学的方法研究蛋白质的相互作用关系。初步的实验结果说明,该 BD-Grid 具有较好的性能。

关键词 数据网格,分布式异构数据,蛋白质相互作用关系,生物信息学

Predicting Protein-protein Interaction through Computing Based on Data Grid

XIE Jiang[1,2] ZHANG Wu[1,2] MEI Jian[1]

(School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China)[1]

(Institute of Systems Biology, Shanghai University, Shanghai 200444, China)[2]

Abstract With the emergence of the massive data collections for protein-protein interactions in various species, research on protein-protein interactions, especially large scale protein-protein interactions, becomes a new hot area. There are so many distributed heterogeneous data sources for protein-protein interactions on Internet that it is difficult for users to integrate information within the data sources. In order to utilize high-capacity protein-protein interaction databases, this paper implements transparent access data and produces a common framework of predicting the protein-protein interaction on BD-Grid, which gets rid of the differences between heterogeneous databases through effective metadata management service and reasonable data classification service. The predication for the protein-protein interactions has been fulfilled on this data grid. As a result, the biological researchers can study protein-protein interaction with bioinformatics method efficiently and effectively. Test results show that the BD-Grid has good performance.

Keywords Data grid, Distributed heterogeneous data, Protein-protein interactions, Bioinformatics

1 引言

蛋白质是生命的物质基础,是人体内三大组成部分(蛋白质、脂肪、碳水化合物)之一,研究蛋白质-蛋白质相互作用关系是从分子水平揭示蛋白质功能的重要途径。随着后基因时代的来临,人们正试图通过研究蛋白质相互作用关系网络来控制细胞和组织的活动。

目前,在互联网上存在着各种关于蛋白质相互作用关系的数据,这些数据来自于不同的研究机构,其中的许多数据都来自于研究人员的生物学实验,甚至于有的数据的获得要花费大量的时间和成本且不可重复。著名的关于蛋白质相互作用关系的数据库有 HPRD[1], MINT[2], BIND[3], DIP[4], MIPS[5]等,整合和分析这些数据将有效地实现信息共享,节省大量的时间和经济成本,加快蛋白质研究的进展。

另一方面,由于实验条件的局限性以及生命科学本身的复杂性,任何研究者都不可能完成不同生命阶段的所有生物学实验,往往只能选择特定的阶段做少量可能具有代表性的实验。所有这些已有的数据都是不完整的,任何一个数据源都不可能囊括其它所有数据源的信息。因此,作为对传统的生物学实验的补充,人们需要通过新的实验方法来获得更多的关于蛋白质相互作用关系的数据,而计算预测就是新的方法之一。由于互联网上现有的蛋白质相互作用关系数据源都以分布式的、异构的、大容量的形式存在,用户难以高效地整合这些数据源中的信息;各种计算预测算法也没有统一的规范和接口,限制了用户对它们的使用。数据网格的兴起则能有效地解决这些问题,它通过虚拟企业中的管理域使共享异构、分布式存贮资源和基于本地或全局策略的数字实体成为可能[6]。目前已经有许多关于数据网格应用的研究,如体系结构、数据复制管理、任务执行路径等。

在生物信息学领域,网格是支持个性化生物信息学实验所必需的新一代基础架构。其中数据网格作为网格研究的一个重要方面,起着非常重要的作用。特别是现在主要的生物信息网格,如亚太生物信息网格(the Asia Pacific BioGrid Initiative, http://www.apbionet.org/apbiogrid/)、美国北卡罗来纳州的生物信息网格(the North Carolina BioGrid, http://www.ncbiogrid.org/)、欧洲经济共同体共同出资研制的 EUROGRID (http://www.eurogrid.org/)以及生物医学信息研究网络 (the Biomedical Informatics Research Network, http://www.nbirn.net/)等。这些信息网格建立了大型数据库,收集并定期更新有关数据,供各国研究者通过 Web 界面

谢 江,梅 健 博士研究生;张 武 教授,博导。

免费获取数据及分析算法(主要是数据挖掘程序,这些程序的运行需耗费大量CPU机时和I/O流量),这在一定程度上提高了数据的共享性和处理效率。但是,这种共享方式还存在耗费大量CPU机时和I/O流量等许多局限性,成为阻碍生物信息学进一步快速发展的瓶颈。

首先,相对于这领域中的大批研究者而言,可通过Web访问的计算资源是有限的,资源之间的组合利用就更加困难了。其次,数据处理常常需要对不同应用程序的运算结果做流水线处理,而从基于Web的应用程序检索其运算结果是很容易出错的,所以很多研究者将所需数据从主要研究机构的数据库中下载到自己的本地计算机上处理。这样下载的数据缺乏集中管理,不可避免地存在数据的冗余和数据差异,这就降低了实验数据的质量和处理结果的可信度。虽然一些生物方面的数据网格在一定程度上解决了数据共享的问题,但是它们并没有一个数据平台来定义数据源,并对数据源进行组合和过滤,也没有对数据进行分类和整理。

本文提出在生物信息学数据网格BD-Grid环境下,有效利用各种研究机构提供的分布式的、异构的生物学实验数据和计算算法,计算预测蛋白质相互作用关系。我们的主要工作是:

①设计并实现了一个统一、标准的三层数据平台,用户可以通过UDDI发布、获取可用的数据源数据。

②我们对元数据(Metadata)进行类别的规划,让用户在共享数据资源的时候,定位到相应的目录;用户在使用资源的时候,可以准确及时地定位到数据源,并为以后的数据留下扩展接口。

③提供一个数据操作的平台,让用户定义数据源,对数据进行操作。

④在BD-Grid环境下,计算预测蛋白质相互作用关系。通过实验证明,BD-Grid为在生物信息学方面研究蛋白质相互作用关系提供了一条有效途径。

下面将从BD-Grid体系结构、计算实验方法和实验结果几个方面分别进行阐述。

2 BD-Grid体系结构

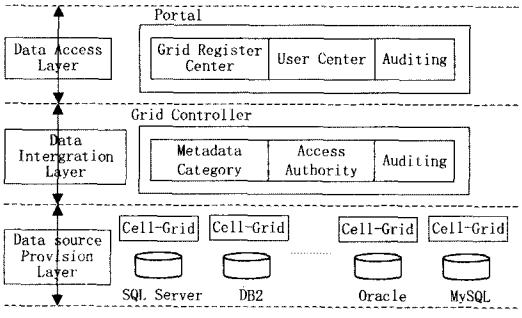

图1 BD-Grid体系结构

BD-Grid(Biologic Data Grid)以一种数据联盟的形式提供数据的呈现和整合,把异地的分布数据提供给用户和应用。其间,数据网格平台抽取来自多个不同类型数据源的数据,把它们整理成为用户需要的形式,使不同的研究机构可以安全地共享数据,用户和应用通过标准的界面访问数据而不需要知道数据存储的物理位置。BD-Grid体系结构如图1所示,分成三层:数据供应层DPL、数据集成层DIL以及数据访问层DAL。

网格的应用采用网格节点控制器和子网格节点两种数据服务。考虑到容错的需要,我们把网格节点控制器分为两种:一种是主网格节点控制器,另一种是副网格节点控制器,它们的功能是完全一致的。如果主网格节点控制器宕机,则副网格节点控制器可以代替它继续工作。

2.1 数据供应层

数据供应层将数据以有序和安全的方式呈现给各种用户。

我们把提供数据的主机作为子网格节点,这些节点按照生物学数据的分类分成不同的域。例如常用的蛋白质数据库有SWISS-PROT(http://www.ebi.ac.uk/swissprot/)、HPRD[1]等,核酸数据库有EMBL[7]和GenBank[8]等,基因组数据库有GDB(http://www.gdb.org/)和FlyBase[9]等。这些数据库中的特定的数据集都可以作为子网格节点的相关的域。网格子节点从多个不同类型的数据源抽取数据,将它们转换成为用户需要的形式,并安全地被各种研究机构使用。网格子节点本身也具备网格的特征:发布数据服务、响应服务请求以及反映实时状态。我们前期开发的用于与监控模块的通信的中间件[10]应用于BD-Grid的网格子节点上,用户通过它可以获取每个服务节点的实时信息,通过对原始数据的链接避免了数据复制,在随时保持同步的同时节省大量存储成本。

2.2 数据集成层

合并到BD-Grid中的数据往往来自不同类型的数据库。在计算预测蛋白质相互作用关系的时候,用户需要从多个数据库中挖掘蛋白质数据,或者从不同系统整合数据以得到某个蛋白质的信息,这样他们就需要花费大量时间去查找相关数据,编写代码访问和转换数据。为了解决这个问题,我们设计了数据集成层来集成数据呈现层所呈现的数据。通过数据集成层的标准接口,用户可以方便、安全地从不同类型的多个数据源抽取数据而不用知道数据的物理存储位置。如果数据发生变化,用户只需要变换数据链接而不用修改他们的应用程序。

数据集成层包括三个模块:元数据的分类、数据服务状态监控以及访问授权。

元数据分类模块用于建立一套命名机制来组织元数据,每个元数据类都有个对应的分类目录。目前BD-Grid中有蛋白质相互作用关系、蛋白质域、蛋白质复合体等三个元数据类,它们分别抽取于MINT[2],BioGrid[11],HPRD[1],Pfam[12]以及PDB[13]。元数据分类目录和普通目录一样分层,不同的是分类目录的物理位置是相互独立的。用户不需要了解数据的物理存储位置,只需要从分类目录中抽取有用的数据。在分类目录中新建条目的时候,也不需要进行源数据的复制,而是在分类目录和源数据之间建立链接,同时把链接参数如数据库类型、数据库名称、数据库驱动、用户名、密码、证书标示等提交给访问授权模块。

我们将数据服务状态监控模块独立出来,实时地反映注册的数据资源的服务状态。用户在查询相应数据资源时,可以通过这个模块,有效地提高目标的命中率和执行效率。

访问授权模块处理登录访问BD-Grid的授权。我们将用户分组并授权给组。一个包括多个用户的组可以加入到访问列表中并和一个普通用户一样有被允许或拒绝的权限。

2.3 数据访问层

数据访问层的目的在于使BD-Grid的用户方便地找到所

需要的数据，其中包括 Web 界面、网格注册中心、用户管理模块和审核模块。

用户在 Web 门户完成通用的任务，包括设置任务，如代理服务配置和目录服务集成；数据管理任务，如使数据有效和设置访问权限；日常任务，如监控、日志和备份。

网格注册中心类似于统一描述、发现和集成协议 UDDI。在这里采用了域的概念，用户因此可以在 BD-Grid 上方便地找到需要的数据或把元数据发布到合适的分类目录下。通过 BD-Grid 的域，可以清晰地列出元数据，降低基础架构的复杂度，因此在方便对元数据的管理的同时，降低了管理成本。

用户管理模块用来管理用户的访问权限。根据用户指定的时间，审核模块不断地从数据集成层的审核模块获取信息，包括登录响应时间、出错、响应时间等。用户可以通过这个模块跟踪一些工作流，获得实时的执行信息。

3 计算实验方法

在 BD-Grid 环境下，我们集成蛋白质相互作用关系和蛋白质域两种元数据类，通过 ODBC，JDBC，SOAP 等一系列协议和接口使授权用户易于访问和整合各种数据库中的异构、分布式蛋白质数据，并以蛋白质相互作用关系计算算法库中的 APM[14] 算法为例，实现了以 α-synuclein 为目标蛋白的蛋白质相互作用关系网络的计算预测，如图 2 所示。

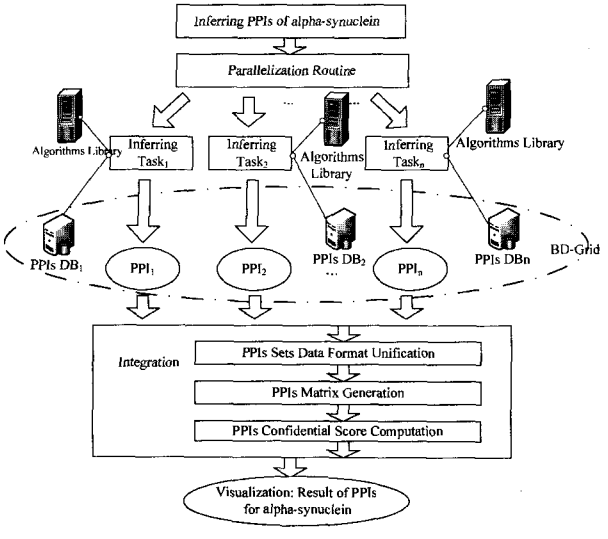

图 2 计算预测蛋白质相互作用关系的方法实现

计算预测的实现可以如下完成。

(1)任务划分：根据用户在计算过程中需要用到的数据集的个数，将计算划分为几个任务在子网格节点上并行执行。每一个任务负责返回一个在给定数据集上计算得到的结果。

(2)服务调用：这些数据集是分布式、异构且各不相同的，计算方法库中又有各种不同的算法。根据用户的请求，系统通过 BD-Grid 中统一的目录为每个任务调用合适的算法服务和相关的数据服务。这样，用户可以专注于蛋白质相互作用关系的计算过程，而避开如何访问各种原始数据的问题。

(3)中间结果输出：每个任务输出从相应的元数据计算得到的蛋白质相互作用关系结果。

(4)数据整合：生成蛋白质相互作用关系矩阵，矩阵元素 C_{ij} 代表在第 i 个数据集中蛋白质 j 和目标蛋白 α-synuclein 是否存在相互作用关系。

(5)根据用户要求计算的网络规模，重复(2)-(4)步骤，或者停止计算，可视化输出。

结束语 运用上述方法，在 BD-Grid 上实现了目标蛋白 α-synuclein 相互作用关系的计算，结果如表 1 所示(所有数字均表示是人类蛋白质相互作用关系数据)。

表 1 各数据库中蛋白质相互作用关系实验数据及计算结果

数据库	总数(条)	用于计算的数目(条)	数据库中 α-synuclein 关系(条)	计算结果 α-synuclein 关系(条)
HPRD	37,581	23,187	29	769
MINT	18,980	12,264	4	364
BioGrid	38,225	14,031	20	553

本文提出在数据网格 BD-Grid 环境下计算预测蛋白质相互作用关系。通过清晰合理的数据网格系统架构、有效的元数据管理服务和合理的数据分类服务将异构、分布的数据集成在 BD-Grid 平台，并在该数据平台上，实现了预测蛋白质相互作用关系的应用。通过实验证明，我们的 BD-Grid 是一个高效的、具有借鉴意义的应用型数据网格，为生物数据的整合和研究提供一个有效的数据平台。

参 考 文 献

[1] Peri S, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Research, 2003, 13(10): 2363-2371

[2] Zanzoni A, et al. MINT: a Molecular INTeraction database. FEBS Letters, 2002, 513: 135-140

[3] BIND Biomolecular Interaction Network Database. http://bond.unleashedinformatics.com

[4] DIP Database of Interacting Proteins. http://dip.doe-mbi.ucla.edu

[5] Pagel P, Kovac S, Oesterheld M, et al. The MIPS mammalian protein-protein interaction database. Bioinformatics, 2005, 21(6): 832-834

[6] Moore R W, Rajasekar A, Wan M. Data Grids, Digital Libraries, and Persistent Archives: An Integrated Approach to Sharing, Publishing, and Archiving Data. Proceedings of the IEEE, 2005, 3(93): 578-588

[7] Kulikova T, Akhtar R, Aldebert P, et al. EMBL Nucleotide Sequence Database in 2006. Nucleic Acids Research, 2006, 00(Database issue): D1-D5

[8] Benson D A, Karsch-Mizrachi I, Lipman D J, et al. GenBank. Nucleic Acids Research, 2006, 34(Database issue): D16-D20

[9] Grumbling G, Strelets V. The FlyBase Consortium. FlyBase: anatomical data, images and queries. Nucleic Acids Research, 2006, 34: D484-D488

[10] Mei J, Zhang W, Xie J. A General Data Grid: Framework and Implementation // Alexandrov V N, et al, eds.: ICCS 2006, Part IV, LNCS 3994. 2006: 669-676

[11] Stark C, Breitkreutz B J, Reguly T, et al. BioGRID: A General Repository for Interaction Datasets. Nucleic Acids Res, 2006, 34(Database issue): D535-D539

[12] Finn R D, Mistry J, Schuster-Bockler B, et al. Pfam: Clans, Web Tools and Services. Nucleic Acids Research, 2006, 34(Database issue): D247-D251

[13] Berman H M, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acids Research, 2000, 28(1): 235-242

[14] Chen LN, Wu LY, Wang Y, et al. Inferring Protein Interactions from Experimental Data by Association Probabilistic Method. PROTEINS: Structure, Function, and Bioinformatics, 2006, 62: 833-837