基于GPU的非标记定量软件QuantWiz并行化实现

费 辉[1,2,3] 张云泉[1,2] 王 靖[1,2]
(中国科学院软件研究所并行软件与计算科学实验室 北京 100190)[1]
(中国科学院软件研究所计算机科学国家重点实验室 北京 100190)[2]
(中国科学院研究生院 北京 100190)[3]

摘 要 QuantWiz是一款基于质谱的非标记定量软件,可很好地应用于定量蛋白质组学。实验数据的日益增大,使定量的计算量巨大,耗费时间长。GPU以几百GFlops甚至上TFlops的运算能力,为定量蛋白质组学这样的计算密集型应用提供了良好的加速方案。对QuantWiz软件做了深入的研究与分析,找到了软件性能的热点模块所在,提出了该软件在GPU上的加速方案——GPU-QuantWiz,并进行了实现。性能测试显示,在Tesla C1060上,该方案的平均加速比达到9.66倍,得到了良好的加速效果。同时,该方案还可以扩展到两块及以上的GPU上,具有良好的可扩展性。

关键词 非标记定量,蛋白质组,QuantWiz,GPU,并行计算
中图法分类号 TP312 **文献标识码** A

Parallelization of Label-free Protein Quantification Software QuantWiz Based on GPU

FEI Hui[1,2,3] ZHANG Yun-quan[1,2] WANG Jing[1,2]
(Laboratory of Parallel Software and Mathematic Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)[1]
(State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)[2]
(Graduate University, Chinese Academy of Sciences, Beijing 100190, China)[3]

Abstract QuantWiz is a label-free quantitative software based on mass spectrometry, well used in quantitative proteomics. The increasing experimental data causes the enormous workload. Having hundreds of GFlops or even TFlops performance, GPU can speed up such compute-intensive quantitative proteomics applications. This article analyzed the software QuantWiz, to find the hotspot module of this software. Then we presented an accelerated program on GPU for this software called GPU-QuantWiz and implemented it under CUDA Framework. Statistical performance results show that the accelerated program can achieve good performance, with 9.66 speedup. Moreover, our algorithm can be extended on two or more GPUs, with a good scalability.

Keywords Label-free quantification, Proteome, QuantWiz, GPU, Parallel computing

1 引言

蛋白质组学是在细胞的整体蛋白质水平上进行研究、从蛋白质整体活动的角度来认识生命活动规律的一门新学科[1]。研究内容包括对各种蛋白质的识别和定量化,确定它们在细胞内外的定位、修饰、相互反应、活性,最终确定它们的功能。随着生命科学研究已经进入了后基因时代,蛋白质组学开始蓬勃发展。蛋白质组在时间和空间上存在多样性,因此精确描述细胞内全部蛋白质的形态和数量是比较困难的,目前主要依赖于定量蛋白质组学[2]来解决这个问题。定量蛋白质组学,就是把一个基因组表达的全部蛋白质或一个复杂的混合体系中所有的蛋白质进行精确的定量和鉴定,它在蛋白质组学研究中起着重要作用。

目前定量蛋白质组研究策略主要有:基于双向凝胶电泳的定量蛋白质组研究策略和基于生物质谱的定量蛋白质组研究策略等[3]。生物质谱的定量蛋白质组研究策略又包括基于质谱的稳定同位素标记定量方法和基于质谱的非标记定量方法。近几年来,由于对大规模的蛋白质进行定量研究,基于质谱的非标记定量方法备受关注。随着蛋白质分离技术和质谱技术的发展,基于质谱的非标记定量方法省去了同位素标记等繁杂的化学处理过程,只需要分析处理质谱原始数据。因此,基于质谱的非标记定量成为当前国内外研究的热点。随着质谱技术和生物实验手段的飞速发展,定量实验所产生的质谱数据量日益增大,定量性能的问题已经显现出来,已成为

到稿日期:2011-08-12 返修日期:2011-10-15 本文受中科院知识创新工程重大项目(KGCX1-YW-13)资助。
费 辉(1985—),男,硕士生,主要研究方向为并行算法与并行软件;张云泉(1973—),博士,研究员,博士生导师,主要研究方向为高性能计算及并行数值软件、并行计算模型、并行数据库、海量数据并行处理;王 靖(1982—),硕士,主要研究方向为生物信息学和并行软件,E-mail:wangjing0625@163.com(通信作者)。

制约定量蛋白质组学向前发展的瓶颈。对于超大规模的定量实验数据，传统定量软件分析处理需要长达数天甚至数月之久，已经不能满足大规模质谱数据分析处理的要求，因此迫切需要有效的加速技术手段来解决这样的问题。

本文针对非标记定量软件 QuantWiz 进行了深入研究与分析，首先分析其运行的热点模块，然后对热点模块在 GPU 上进行并行化。测试结果显示，经过 GPU 加速的 QuantWiz 比 CPU 版本的性能有了明显的提高，平均加速比达到 9.66 倍。

本文第 2 节介绍相关工作；第 3 节介绍非标记定量软件 QuantWiz 的研究与分析；第 4 节介绍 QuantWiz 的 GPU 并行加速算法设计及实现；第 5 节介绍实验结果与分析；最后对全文做了总结。

2 相关工作

MapQuant[4]、SuperHirn[5]、OpenMS[6] 和 Census[7] 等是较为广泛使用的非标记定量软件。文献[7]对当前主流的非标记定量软件进行了较为详细的分析比较。QuantWiz 是一款我们自主研发的、基于质谱的非标记定量软件，可很好地应用于定量蛋白质组学中。QuantWiz 目前已有 MPI 版本，在 32 核上并行效率达到 63%[8]。

随着 GPU 的快速发展，到现在它已有几百 GFlops 甚至上 TFlops 的运算能力，其强大的计算能力使得它在很多方面得到了应用。天河一号 A 采用了 CPU+GPU 的混合架构，系统效率有了很大提升，它的 Linpack 峰值性能高达 2.57 petaflop/s，在 Top500 的第 36 版排行榜中排名第一[9]。GPU 强大的计算能力使得它已经越来越多地应用到生命科学以及其他一些领域当中。文献[10]利用 GPU 实现的分子动力学(MD)模拟，加速比达到 20～60 倍，通过方腔流及颗粒-气泡接触等实例初步展示了 GPU 加速方式从微观上模拟介观行为的能力。文献[11]为了精确高效地进行生物序列比对，提出一种 GPU 加速的 Smith—Waterman 算法，该算法使用菱形数据布局以更充分地利用 GPU 对的并行处理能力，与 CPU 上的串行算法相比，可以获得 120 倍以上的性能提升。文献[12]通过实验分析了 BLAST 软件包中的典型程序 BLASTN 的运行热点，并选定关键热点模块，利用 GPU 进行并行化改造，可明显缩短该模块的运行时间，最终可获得超过 35 倍的加速比。

利用 GPU 来加速一些计算密集型应用已成为当今的研究热点之一。由于 GPU 出色的计算能力以及相对较低的价格，基于 GPU 的通用计算已经对许多应用领域产生了极大的推动作用。可见，利用 GPU 来加速非标记定量软件 QuantWiz 是一个可行的方案。

3 非标记定量软件 QuantWiz 的研究与分析

一般来说，基于质谱的非标记定量的数据处理流程可以分为质谱对齐、质谱预处理、色谱峰平滑、峰识别、峰下面积计算等。图 1 为基于质谱的非标记定量流程[8]。

非标记定量实际上就是根据鉴定结果，对多个质谱数据进行分析处理，然后将定量结果进行输出。非标记定量的步骤如下：第一步是质谱对齐，也就是对多个文件的质谱数据进行对齐，确定鉴定结果在多个质谱文件中鉴定位置的关系，根据某一质谱文件中鉴定到的肽段 P，能够预测肽段 P 在其它质谱文件中的鉴定位置，这样就能较为精确地找到该肽段在多个质谱中的色谱峰。第二步是质谱预处理，这个过程包含质谱数据的读取和数据的降噪处理两部分，由于实验数据存在系统误差，因此需要对实验数据进行降噪处理，以减小系统误差。第三步是色谱峰平滑、峰识别、峰下面积计算，其中，色谱峰平滑是用特定的数学函数对降噪处理后的数据进行拟合平滑，以便很好地识别色谱峰；峰识别是根据平滑后的色谱峰图形，利用鉴定结果找出相应的色谱峰的波峰和波谷；峰下面积计算，就是对色谱峰进行求积分，得到的结果可以衡量肽段的量。第四步是定量输出，根据计算的峰下面积求出肽段在多个质谱文件中的比值，将定量结果输出。

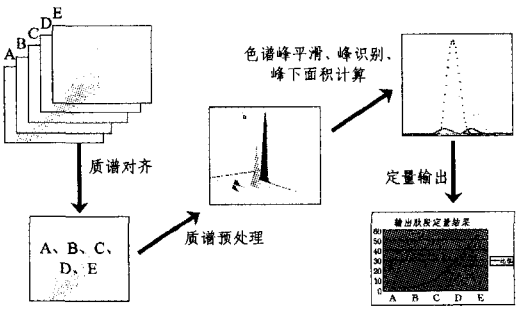

图 1 基于质谱的非标记定量数据处理流程

QuantWiz 软件的具体处理流程如图 2 所示。

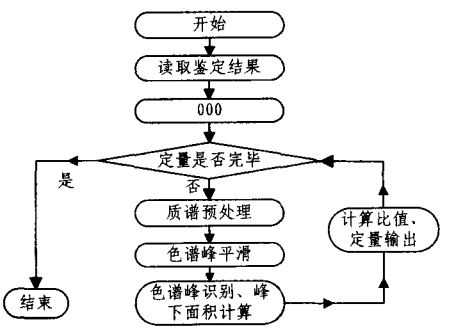

图 2 QuantWiz 软件流程图

从 QuantWiz 的软件流程图中可以看到，QuantWiz 的核心处理部分是一个循环，在循环内部针对每一个鉴定结果对质谱进行预处理、色谱峰平滑、色谱峰识别、峰下面积计算。QuantWiz 的算法具体描述如算法 1 所示。

算法 1(QuantWiz 串行算法)

1)从鉴定文件中读取全部鉴定结果；

2)对质谱文件进行对齐；

3)选择一个鉴定结果，读取对应的质谱数据；

4)对数据进行降噪处理；

5)对数据进行平滑处理；

6)色谱峰识别、峰面积计算；

7)计算该鉴定结果在不同质谱文件中量的比值；

8)跳转到 3)，直到所有的鉴定结果定量完毕；

9)输出定量结果。

对 QuantWiz 进行了实验分析，实验数据为上海生命科学研究院系统生物学重点实验室提供的 1:1:2:5:10 的标准非标记定量测试数据，质谱文件共 5 个(每个质谱文件大小约 200MB，共 1G 左右)，鉴定文件 1 个(鉴定结果 836 个)。软件的编译及运行环境如下：操作系统为 Ubuntu 4.3.3-5；编译器为 gcc 4.3.3；CPU 为 Intel(R) Xeon(R) CPU X5472@3.00 GHz。软件各个模块的运行时间分布如表 1 所列。

表 1 QuantWiz 中各个模块的运行时间分布

模块名称	质朴数据读取	数据降噪	其它	总计
时间(s)	9.95888	216.87	0.27212	227.101
所占百分比(%)	4.39	95.49	0.12	100

各个模块的运行时间所占百分比如图 3 所示。从图 3 中可以看出，数据降噪这个模块占用的时间是最长的，占了运行时间的 95.49%，质谱数据读取模块的运行时间占了4.39%，其他部分的运行时间占了 0.12%。可见数据降噪这个模块就是热点模块，因此选择对这个模块进行并行化。

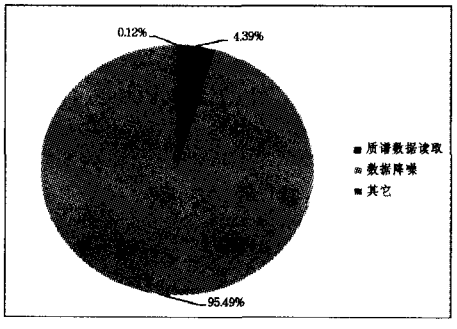

图 3 QuantWiz 中各个模块运行时间分布

4 QuantWiz 的 GPU 加速算法设计及实现

数据降噪这部分为 QuantWiz 的热点模块，而且这个模块是计算密集型的，因此分析出这个热点模块的并行性，然后将这个模块放到 GPU 上进行计算，便可以得到良好的加速效果。

通过观察分析 QuantWiz 串行算法，可以发现各个鉴定结果的计算在一个循环中依次执行，它们之间是相互独立的，没有依赖关系。这种情况下，每个鉴定结果的数据降噪处理部分也是相互独立的，所有鉴定结果的这一部分是可以并行处理的，因此降噪处理这部分可以放在 GPU 上进行并行化计算。

这样，得到 QuantWiz 在 GPU 上的并行加速方案 GPU-QuantWiz，具体描述如算法 2 所示。

算法 2(GPU-QuantWiz 算法)

1)从鉴定文件中读取鉴定结果；

2)对质谱文件进行对齐；

3)读取所有鉴定结果的质谱数据；

4)将数据从主机内存拷贝到 GPU 显存；

5)调用 RemoveBaseLineKernel 对数据进行降噪处理；

6)将数据从 GPU 显存拷贝到主机内存；

7)选择一个鉴定结果的数据进行平滑处理；

8)色谱峰识别、峰面积计算；

9)计算该鉴定结果在不同质谱文件中量的比值；

10)跳转到 8)，直到所有的鉴定结果定量完毕；

11)输出定量结果。

在算法 2 的描述中，该算法主要包含 5 部分：第一步是读取鉴定结果，对质谱文件作对齐处理；第二步是读取鉴定结果所需要的数据；第三步是将数据从 CPU 内存拷贝到 GPU 显存中，调用 GPU 上的 RemoveBaseLineKernel 算法(见算法 3)，处理数据降噪部分，然后将计算结果从 GPU 显存拷贝到 CPU 内存中；第四步是对所有鉴定结果的数据进行平滑处理、峰识别、峰面积计算，得到鉴定结果在不同质谱文件中量的比值；第五步是输出定量结果。

算法 3(RemoveBaseLineKernel 算法)

1. get_global_id，得到线程 id 为 x；

2. 对第 x 个鉴定结果的质谱数据进行降噪处理。

在上述算法的实现中，CPU 部分的代码采用 C++语言实现，GPU 部分的代码采用 CUDA C 实现。

5 实验与分析

本节主要对 QuantWiz 和 GPU-QuantWiz 的定量性能进行实验验证和分析。实验的编译及运行环境如表 2 所列。

表 2 实验的编译及运行环境

操作系统	Ubuntu 4.3.3-5
编译器	gcc4.3.3
	nvcc3.0
CPU	Intel(R) Xeon(R) CPU X5472 @ 3.00GHz
GPU	Tesla C1060
	Quadro FX 4800

5.1 GPU-QuantWiz 实验结果及分析

实验数据由上海生命科学研究院系统生物学重点实验室提供，数据规模如表 3 所列。

表 3 生物实验数据信息

实验组别	质朴文件大小	质朴文件数目	鉴定结果数目
50fv100f	107MB	2	1920
50fvs1p	133MB	2	3489
50fvs200f	124MB	2	2645
50fvs3p	160MB	2	4587
50fvs500f	130MB	2	2925
3pv3p	216MB	2	7680

我们对上述的生物实验数据分别进行了测试，QuantWiz 的串行版本在 CPU 上的执行时间和 QuantWiz 的 GPU 并行版本 GPU-QuantWiz 在 CPU+GPU 上的执行时间如表 4 所列，平均加速比为 9.66 倍。加速比的计算方法如式(1)所示。

$$Speedup=\frac{T_{CPU}}{T_{CPU+GPU}} \tag{1}$$

表 4 实验结果数据信息

实验组别	CPU(s)	CPU+GPU(s)	加速比
50fv100f	22.8743	2.3660	9.6678
50fvs1p	43.4752	4.4243	9.8264
50fvs200f	31.8807	3.4703	9.1867
50fvs3p	56.5672	6.6526	8.5030
50fvs500f	35.2676	3.8071	9.2636
3pv3p	93.9497	8.1691	11.5006

GPU-QuantWiz 在 CPU+GPU 上的实现相对于 Quant-Wiz 的串行版本在 CPU 上的加速比趋势如图 4 所示。从图 4 可以看出，相对于 QuantWiz，GPU-QuantWiz 的性能比较稳定，加速比在平均值 9.66 的上下波动，且波动范围很小。因此对于一般的定量应用，与 QuantWiz 相比，GPU-QuantWiz 可以使性能提高到 9.66 倍左右。

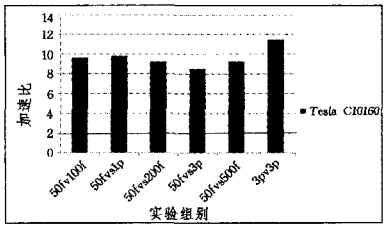

图 4　GPU-QuantWiz 加速效果

为了测试 GPU-QuantWiz 处理大规模实验数据的能力，我们对上海生命科学研究院系统生物学重点实验室提供的两组大规模实验数据进行了测试。实验一的数据中，质谱文件 5 个，共 1GB；鉴定文件 1 个，鉴定结果 836 个。在实验二的数据中，质谱文件 7 个，共 563 MB；鉴定文件 1 个，鉴定结果 16190 个。两组实验数据利用 GPU-QuantWiz 进行处理的加速比如图 5 所示。

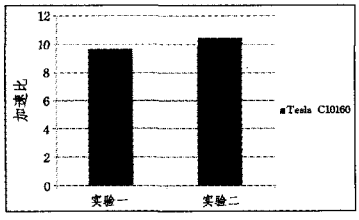

图 5　GPU-QuantWiz 处理两组大规模数据

从图 5 中可以看出，在进行大规模数据处理时，GPU-QuantWiz 依然有良好的性能提升，实验一对应的加速比为 9.64，实验二对应的加速比为 10.44，平均加速比为 10.04。当鉴定结果数目较多、质谱文件较小时，读取质谱数据部分所占比重降低，热点模块的计算所占比例提高，使得并行部分占用比重稍大，加速比略微高些。

5.2　GPU-QuantWiz 可扩展性分析

GPU-QuantWiz 算法的另一个特性是扩展性较强。若有两个以上的 GPU 参与计算，则可以按照计算任务平均划分。我们将该算法扩展到 2 个 GPU 上进行了实现，2 块 GPU 分别为 Tesla C1060 和 Quadro FX 4800，性能统计的结果如图 6 所示。Quadro FX 4800 的性能较 Tesla C1060 差，根据两块 GPU 的峰值性能：Tesla C1060 的单精度浮点峰值性能为 933 GFlops，Quadro FX 4800 的单精度浮点峰值性能为 462GFlops。由式(2)可以算出，利用 Tesla C1060+Quadro FX 4800 达到的性能理论上应该为 TeslaC1060 的 1.5 倍。

$$P=\frac{Peak_{tesla}+Peak_{\text{Quadro FX}}}{Peak_{tesla}} \qquad (2)$$

在实际的定量计算中，利用 Tesla C1060 和 Quadro FX 4800 两块 GPU 加速之后，实验一的加速比为 12.83，实验二的加速比为 14.33，平均加速比为 13.58，Tesla C1060+Quadro FX 4800 的性能约为 Tesla C1060 的 1.35 倍。由于两块 GPU 的性能不一样，需要等待两块 GPU 都完成计算之后再进行在 CPU 段的计算，因此实际值略低于理论值。

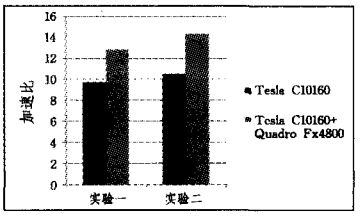

图 6　GPU-QuantWiz 的扩展性

结束语　本文对非标记定量软件 QuantWiz 做了深入的研究，找到了其中的热点模块。针对热点模块进行了更进一步分析，发现其中的并行性，提出了基于 GPU 的并行化算法 GPU-QuantWiz，并将 GPU-QuantWiz 在 CUDA 框架下进行了实现。测试结果显示，在 Tesla C1060 显卡上，GPU-Quant-Wiz 相对于串行 QuantWiz，平均加速比可以达到 9.66 倍左右；利用 GPU-QuantWiz 来处理两组大规模数据，相对于串行 QuantWiz，其性能加速比也能达到 9.6 倍左右。同时，GPU-QuantWiz 还可以扩展到两块及以上的 GPU 上，具有良好的可扩展性。

参 考 文 献

[1]　甄朱. 蛋白质组学进展[J]. 生物工程学报，2001，17(5)：491-493

[2]　Blackstock W P，Weir M P. Proteomics：quantitative and physical mapping of cellular proteins[J]. Trends Biotechnology，1999，17(3)：121-127

[3]　胡维新. 医学分子生物学[M]. 北京：科学出版社，2007

[4]　Leptos K C，Sarracino D A，Jaffe J D，et al. MapQuant：open-source software for large-scale protein quantitation[J]. Proteomics，2006，6(6)：1770-1782

[5]　Mueller L N，Rinner O，Schmidt A，et al. SuperHirn-a novel tool for high resolution LC-MS based peptide/protein profiling[J]. Proteomics，2007，7(19)：3470-3480

[6]　Sturm M，Bertsch A，Groepl C，et al. OpenMS-An open-source software framework for mass spectrometry[J]. BMC Bioinformatics，2008，9(1)：163-173

[7]　Park S K，Venable J D，Xu T，et al. A quantitative analysis software tool for mass spectrometry-based proteomics[J]. Nat Methods，2008，5：319-322

[8]　胡泽林，张云泉，王靖，等. P-QuantWiz：一种基于质谱的并行非标记定量软件[J]. 计算机工程与科学，2009，31(11)：124-127

[9]　Top500 supercomputer sites[EB/OL]. http://www.top500.org

[10]　陈飞国，葛蔚，李静海. 复杂多相流动分子动力学模拟在 GPU 上的实现[J]. 中国科学 B 辑：化学，2008，38(12)：1120-1128

[11]　林江，唐敏，童若锋. GPU 加速的生物序列比对[J]. 计算机辅助设计与图形学学报，2010，22(3)：420-427

[12]　万宁，谢海波，张清，等. 使用 GPU 加速 BLAST 算法初探[J]. 计算机工程与科学，2009，31(11)：98-101