文章编号:1671-4598(2018)12-0234-04 DOI:10.16526/j.cnki.11—4762/tp.2018.12.052 中图分类号:TP274 文献标识码:A

差分隐私 DNA 模体识别安全共享平台的设计与实现

魏裕阳，马 凯，吴 响，毛亚青
（徐州医科大学 医学信息学院，江苏 徐州 221006）

摘要：在数据共享平台进行 DNA 模体识别的过程中，如何防止个体信息泄露已成为该领域发展的研究热点；对此，设计并实现了基于差分隐私的 DNA 模体识别安全共享平台，内置多种满足差分隐私保护模型的 DNA 模体识别算法，实现数据源选择、算法选择、隐私预算设置、结果评估、图形化结果展示等功能；同时，除具备对内置 DNA 数据的模体识别外，还允许科研人员自主上传、共享 DNA 数据库，并对共享的 DNA 数据进行差分隐私模体识别，为科研工作人员提供安全可靠的基因序列分析研究平台和数据共享平台；通过平台测试证明，安全共享平台能够实现 DNA 数据的有效识别和安全共享，设计方案有效可行。

关键词：DNA 模体识别；差分隐私；安全共享

Design and Implementation of Differential Private DNA motif finding Secure Sharing Platform

Wei Yuyang，Ma Kai，Wu Xiang，Mao Yaqing
（School of Medical Informatics，Xuzhou Medical University，Xuzhou 221006，China）

Abstract：In the process of DNA motif finding in data sharing platform，preventing individual information leakage become a research focuses in this field. Aim at this problem，this paper designs and implements a differential privacy—based DNA motif finding security sharing platform，adopts sever DNA motif finding algorithms that satisfy the differential privacy protection model，and realizes data source selection，algorithm selection，privacy budget setting，result evaluation，graphical result display and etc. Moreover，in addition to finding motifs of the built—in DNA datasets，the platform also allows researchers to upload and share DNA datasets and find motifs of the sharing DNA datasets，providing a safe and reliable genetic sequence analysis research and data sharing platform for researchers. The tests prove that the secure sharing platform can effectively find motifs and safely share DNA data，and the design is feasible.

Keywords：DNA motif finding；differential privacy；secure sharing

0 引言

DNA 模体识别（motif finding）作为生物序列分析的基础研究方法之一，对研究基因的表达调控机制、发现 DNA 功能位点有着重要意义[1-2]。但是，DNA 数据蕴含丰富的隐私信息，这些隐私信息的泄露问题成为了 DNA 序列分析发展的瓶颈之一[3-5]。与此同时，Homer 等人也通过实验证实：基因序列分析研究中确实存在极高的隐私泄露风险[6]。该结论导致多个知名生物数据平台暂停 DNA 数据共享服务，严重阻碍 DNA 序列分析研究的发展，隐私泄露已经成为了 DNA 序列分析技术发展中亟待解决的关键性问题。

目前，国外学者对 DNA 序列分析的隐私保护研究主要集中在差分隐私保护技术上，并取得了一些成果[7-11]。差分隐私技术设定了一个严格的攻击模型，能够对隐私泄露风险进行严谨、定量化的推导与证明。而差分隐私模型的特性是能够在攻击者已掌握除某一条 DNA 序列之外的所有数据信息时，仍然保证该 DNA 序列隐私信息的安全性。但是，由于 DNA 数据的高度敏感性，往往容易造成差分隐私对 DNA 序列分析结果的过度加噪，从而导致分析结果失去应有价值。因此，在进行差分隐私 DNA 序列分析研究时，分析方法既要保证结果安全性又要保证结果的高可用性。

对此，Uhler 等人[7]将差分隐私加噪融入到 DNA 序列分析过程中，并提出差分隐私 MAFs（Minor allele frequencies）、差分隐私卡方检验、差分隐私 p—values 等数据发布方法，且从理论和实验两个方面证明了这些方法的可行性。其后，Simmons 等人[11]对已有研究成果进行改进，并针对人口分层因素影响差分隐私 DNA 序列分析方法精确度的问题，提出了 PrivSTRAT 算法和 PrivLMM 算法，该研究成果引起国内外学术界广泛关注。

而在模体识别领域，Chen 等人[12]指出利用差分隐私可

收稿日期:2018-04-17； **修回日期**:2018-05-15。
基金项目:江苏省高等学校自然科学研究面上项目(18KJB520049)；徐州市科技计划项目(XM13B021)；国家安全生产重大事故防治关键技术科技项目(Jiangsu—0006—2016AQ)。
作者简介:魏裕阳(1993-)，男，江苏徐州人，硕士研究生，主要从事隐私保护和医学数据分析方向的研究。
通讯作者:马 凯(1972-)，男，江苏徐州人，博士，教授，硕士研究生导师，主要从事医学信息学和嵌入式系统方向的研究。

以有效地解决 DNA 模体识别的隐私泄露问题，并成功提出了一种基于 n－gram 的差分隐私保护方法（以下简称 N－gram 算法），该方法一种单纯追求效率的识别方法，在处理较大数据集时需要消耗较多隐私预算，无法保证识别结果的精确度。对此，作者在文献［13］提出一种高精度的方法 DP－CFMF（differential privacy－closed frequent motif finding），该方法在利用闭频繁模式的概念对识别模体中的冗余度进行约减，并减少了隐私预算分配过程，从而在保证 DNA 隐私安全的同时提高了模体识别的精确度。但是，国内外尚未有数据共享平台支撑 DNA 模体的安全识别和研究工作。因而，建立一个 DNA 模体识别安全共享平台成为了模体识别研究领域中亟待解决的问题。

基于以上研究，本文设计并实现了一种差分隐私 DNA 模体识别安全共享平台。该平台通过客户端实现数据源选择、算法选择、隐私预算设置、结果评估及图形化结果等功能，并利用多种差分隐私模体识别方法实现不同需求的 DNA 模体安全识别任务。此外，该平台允许用户自主上传、共享 DNA 数据集，并对上传的数据集进行差分隐私模体识别，在实现 DNA 数据安全共享的同时，为 DNA 模体识别领域研究人员的科研工作提供了有力支撑。

1　平台总体设计

差分隐私模体识别平台主要由平台运行端、DNA 数据库服务器端及客户端三部分组成（图 1 所示为平台总体结构图）。用户通过客户端对模体识别过程中的 DNA 数据库连接、隐私预算配置、算法参数配置及结果显示方式等相关信息进行配置，信息配置包含任务开启、结果显示、DNA 数据导入导出和 DNA 数据上传及共享等指令，并通过多元网络将指令传输给平台运行端；平台运行端在收到任务执行指令后，读取隐私预算配置信息、数据源选择信息、数据规约信息，并执行 DNA 模体识别操作；最后，平台运行端将处理完成后的结果通过多元网络呈现给客户端，并提供结果集展示、本地存储、结果质量评估及图形化展示等功能。

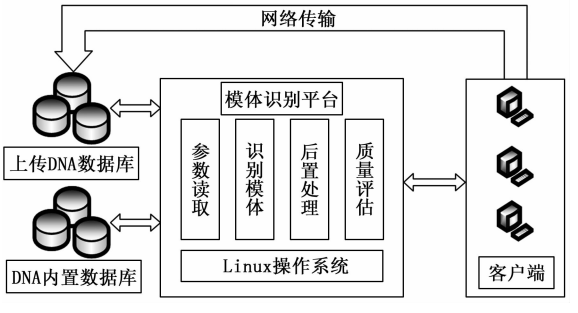

图 1　平台总体结构图

2　平台软件设计

差分隐私模体识别平台主要由平台运行端、DNA 数据库服务器端及客户端三部分组成（图 1 所示为平台总体结构图）。用户通过客户端对模体识别过程中的 DNA 数据库连接、隐私预算配置、算法参数配置及结果显示方式等相关信息进行配置，信息配置包含任务开启、结果显示、DNA 数据导入导出和 DNA 数据上传及共享等指令，并通过多元网络将指令传输给平台运行端；平台运行端在收到任务执行指令后，读取隐私预算配置信息、数据源选择信息、数据规约信息，并执行 DNA 模体识别操作；最后，平台运行端将处理完成后的结果通过多元网络呈现给客户端，并提供结果集展示、本地存储、结果质量评估及图形化展示等功能。平台各子程序具备的功能见表 1。

表 1　各程序具备功能

程序名称	具备功能
主程序	系统初始化、硬件设施配置（如 CPU 数量、内存等）、软件设施配置（平台环境、.Net 框架等）、用户认证
数据库配置	连接功能、数据传输功能、断点续传功能、数据库选择功能
DNA 数据共享	数据集描述、数据存储、数据共享协议
差分隐私模体识别	N－gram[12]、Simple[12]、DP－CFMF[13] 识别算法，一致性约束后置处理方法[13]，数据规约方法
结果反馈	结果存储格式、存储路径、结果质量评估、图形化表示、结果展示

主程序进行平台初始化和各子程序的调用，多元网络通信子程序负责客户端的配置信息及数据库的上传。而平台端在收到客户端的任务开始指令后，将调用服务器内置 DNA 数据库或者用户上传的数据库，并对其进行差分隐私模体识别，最后将识别结果和数据可用性评估通过客户端图形化界面显示给用户。平台软件流程图如图 2 所示。

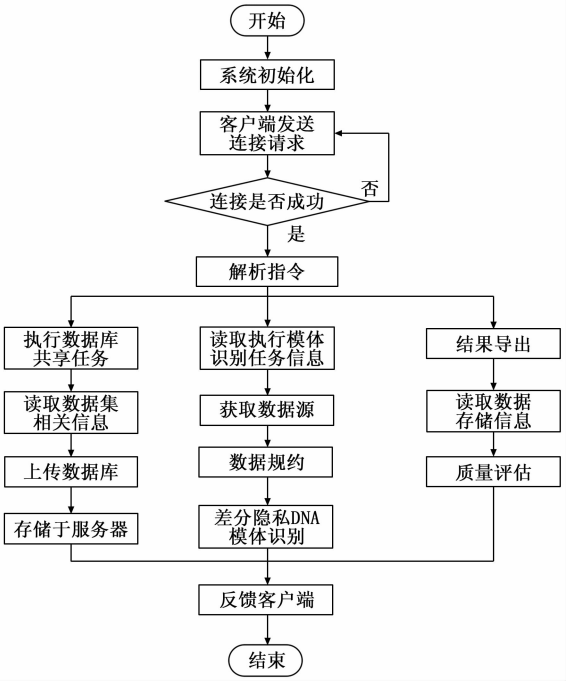

图 2　平台软件流程图

3 平台 DNA 模体识别算法设计

3.1 差分隐私基本概念

差分隐私是一种基于数据失真的隐私保护模型，该模型通过向查询结果中添加适当噪音实现数据分析与共享的隐私保护。差分隐私模型建立在严格的数学推导之上，能够在攻击者拥有最大背景知识情况下保护数据中的个人隐私信息。该模型的原理为：在任一数据集中添加或删除一条数据，这一操作不会影响数据分析的结果。差分隐私模型的具体定义如下：

定义 1：给定两个数据集 D 和 D'，这两个数据集之间最多相差一条数据，即兄弟数据集。同时，给定一个具有隐私保护的算法 A，$range(A)$ 是算法 A 分析结果的取值范围，若算法 A 在给定的两个数据集 D 和 D' 上的任一分析结果 O（其中 $O \in range(A)$）满足下列不等式，则算法 A 满足 ε －差分隐私。

$$|\Pr[A(D)=O]| \leqslant e^{\varepsilon} \times |\Pr[A(D)=O]|$$

上述不等式中，查询结果的概率 $\Pr[\cdot]$ 取决于算法 A 的随机性，也代表着数据集中个人隐私泄露的风险。而隐私预算参数 ε 表示对数据集的隐私保护程度。一般来说，ε 越小，数据集的隐私保护程度越高。

为实现差分隐私模型，一般方法是向算法分析的结果中添加噪声，噪声添加技术主要分为拉普拉斯机制和指数机制，而基于不同噪声机制且满足差分隐私的数据分析算法所需噪音大小与算法的全局敏感性密切相关。

定义 2：对于任意函数 $f: D \rightarrow R^d$，该函数 f 的全局敏感性 Δf 可以表示为：

$$\Delta f = \max_{D,D'} \| f(D) - f(D') \|_p$$

由定义 1 可知，两个数据集 D 和 D' 为兄弟数据集，即两个数据集最多相差一条数据。R 表示通过函数 f，数据集 D 能够映射的实数空间，d 表示映射结果的维度，p 表示全局敏感度 Δf 是利用 L_p 进行度量距离，而本文涉及到的算法均使用 L_1 度量距离。

为使 DNA 模体识别方法满足差分隐私模型，本文使用的噪音机制均为拉普拉斯机制，该机制主要通过拉普拉斯分布产生的随机算子扰动真实 DNA 模体识别频率来实现差分隐私保护。

定义 3：对于任一函数 $f: D \rightarrow \mathbb{R}^d$，如果算法 A 的分析结果满足以下等式，则可以认为算法 A 满足 ε －差分隐私。

$$A(D) = f(D) + < Lap_1(\Delta f/\varepsilon), \ldots, Lap_d(\Delta f/\varepsilon) >$$

在定义 3 中，任一拉普拉斯变量 $Lap_i(\Delta f/\varepsilon)(1 \leqslant i \leqslant d)$ 相互独立。由等式可知，拉普拉斯机制添加的噪音量与 Δf 成正比，与 ε 成反比。换而言之，算法 A 全局敏感性越大，需要添加的噪音量越大。

3.2 差分隐私 DNA 模体识别算法

在平台运行端内置多种差分隐私模体识别方法，除了经典的 N－gram 算法、Simple 算法外，还包括自主设计的基于差分隐私保护模型的 DNA 闭频繁模体识别算法——DP－CFMF，其原理通过构建闭频繁扰动探索树，利用闭频繁模体模型对扰动探索树进行剪枝，该步骤能够减少模体结果集冗余的同时，减少隐私预算的消耗；而且，利用探索树结构能够提高内存使用和模体搜索的效率，并能够快速有效地分配隐私预算；此外，该方法采用最优线性无偏估计对加噪支持度计数进行一致性约束处理，提高数据的可用性。该方法主要包括模式分解单元、构建闭频繁扰动树单元、识别模体单元和一致性约束后置处理单元，其具体流程如下：

1）模式分解单元：利用 $n_{\max}$ 参数对 DNA 原始数据集进行模式分解，获得数据集中长度为 $n_{\max}-1$ 和 $n_{\max}$ 模体及其支持度计数；

2）构建闭频繁扰动树单元：利用长度为 $n_{\max}-1$ 和 $n_{\max}$ 模体构建探索树，利用闭频繁模体等价关系进行剪枝，然后对每一个模体的支持度计数添加相应的拉普拉斯噪声，获得由剪枝后 $n_{\max}-1$ 模体和 $n_{\max}$ 模体组成的闭频繁扰动探索树；

3）一致性约束后置处理单元：利用最优线性无偏估计方法对扰动探索树的每一个节点的支持度计数进行一致性约束后置处理，获得满足树的一致性约束的支持度计数；

4）识别模体单元：在 N－gram 模型的基础上利用马尔可夫假设方法进行预测所有 $n_{\max}+1$ 模体的支持度计数，不断迭代获取长度在 $[n_{\max}, L_u]$ 之间的模体，求解每个模体的联合支持度计数，获得长度在 $[n_{\max}, L_u]$ 之间的频繁模体。

相比于 N－gram 方法来说，DP－CFMF 具有较高的精确度，且其需要使用到的隐私预算较少，可以满足多数情况下的隐私保护；而 N－gram 算法具有较高的效率，但其处理较大数据集时需要消耗大量的隐私预算，甚至可能超出隐私预算上限，从而导致识别过程异常，因此 N－gram 适用于较小 DNA 数据集的安全识别。在使用该平台时，用户可以根据自己不同的情况做出相应的选择。

4 平台测试与分析

4.1 差分隐私模体识别算法测试

本文将真实数据集 Upstream 数据作为内置数据源对平台算法性能进行测试，该数据集包含 487760 条 DNA 序列。测试时，在客户端配置差分隐私保护预算、模体识别参数、图像化显示等信息。实验所使用的软硬件环境为：4G 内存，平台端运行环境为 Linux，算法开发语言为 Python，客户端运行环境为 Window10，客户端开发语言为 C＃，数据库为 SQL sever 2008。图 3 是在不同隐私预算下对 Upstream 数据集执行平台算法测试，其他参数默认值见文献[13]。由图可知，两种方法均可以完成在 Upstream 数据集上的差分隐私模体识别，且具有良好的精确度。此外 DP－CFMF 精确度要高于 N－gram 方法，更适合于高精度要求的任务，而 N－gram 方法相对来说精确度略低，比较适合

处理效率要求较高的任务。

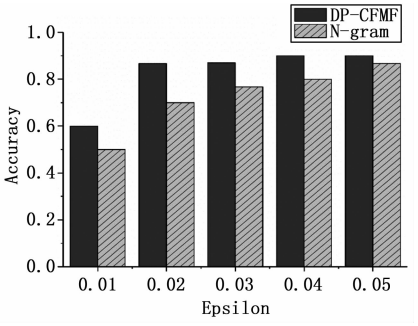

图 3　Upstream 数据集在不同 epsilon 下的精确度对比

为测试研究人员在共享 DNA 数据库场景下的算法运行效果，本文在客户端中将真实数据集 Washington 数据设置为待共享数据集，该数据集共包含 14126 条数据。实验中，客户端通过互联网将 Washington 数据集传输到服务器端。数据共享到服务器端后，本文对 Washington 集进行了不同隐私预算的模体识别测试，测试结果如图 4 所示，DP－CFMF 和 N－gram 算法的精确度均可达到 70％以上。由此可知，通过该平台可以较好地实验 DNA 数据的安全共享。

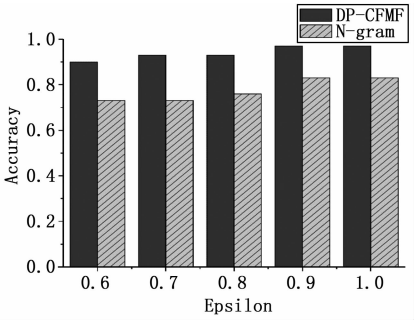

图 4　Washington 数据集在不同 epsilon 下的精确度对比

4.2　客户端总体功能测试

在客户端总体功能测试中，本文主要对安全共享平台进行了参数设置、数据共享、模体识别质量评估等功能的测试。通过测试可知，客户端能够实现内置 DNA 数据进行选择、规约数据大小、描述共享数据集、设置差分隐私模体识别参数、选择结果反馈方式等操作，并将相关指令发送给平台端。平台端对于客户端的请求均做出了响应，并进行了相应操作后将结果反馈给客户端。测试结果表明：平台端和客户端各子程序模块均能成功运行，能满足设计需求。

5　结论

本文描述了差分隐私 DNA 模体识别安全共享平台设计与实现，该平台利用 C/S 架构，允许用户在客户端进行隐私预算及算法参数配置、选择 DNA 数据库、上传及共享 DNA 数据集、结果保存方式等操作，并通过多元网络将指令传入平台端。平台端接收到客户端端指令后，读取、导入用户所选择的数据源，利用差分隐私 DNA 模体识别方法对 DNA 数据进行识别，然后将结果通过客户端的客户端图形化展示给用户。测试结果证明，该平台提供的差分隐私模体识别方法能够有效实现 DNA 数据的安全识别，并能满足用户多种需求。同时，平台提供的自主上传数据和隐私预算配置等功能帮助生物学研究人员开展定制化研究工作，为生物序列的安全共享与研究提供有力支撑。

参考文献：

[1] Mai S M, Abdelhalim M B, Elewa E S. A developed system based on nature－inspired algorithms for DNA motif finding process [J]. Neural Computing & Applications, 2018 (7): 1－11.

[2] Kuttippurathu L, Hsing M, Liu Y, et al. CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments [J]. Bioinformatics, 2011, 27 (5): 715－7.

[3] Check H E. Cloud cover protects gene data [J]. Nature, 2015, 519 (7544): 400.

[4] Marcia McNutt. Genetic privacy [J]. Nature, 2013, 493: 451.

[5] Malin B A. Protecting genomic sequence anonymity with generalization lattices [J]. Methods of Information in Medicine, 2005, 44 (5): 687.

[6] Homer N, Szelinger S, Redman M, et al. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High－Density SNP Genotyping Microarrays [J]. Plos Genetics, 2008, 4 (8): e1000167.

[7] Uhlerop C, Slavkovi? A, Fienberg S E. Privacy－Preserving Data Sharing for Genome－Wide Association Studies [J]. Journal of Privacy & Confidentiality, 2012, 5 (1): 137.

[8] Yu F, Fienberg S E, Slavkovi? A B, et al. Scalable privacy－preserving data sharing methodology for genome－wide association studies [J]. Journal of Biomedical Informatics, 2014, 50 (8): 133.

[9] Yu F, Rybar M, Uhler C, et al. Differentially－Private Logistic Regression for Detecting Multiple－SNP Association in GWAS Databases [M]. Privacy in Statistical Databases. Springer International Publishing, 2014: 170－184.

[10] Zhao Y, Wang X, Jiang X, et al. Choosing blindly but wisely: differentially private solicitation of DNA datasets for disease marker discovery [J]. Journal of the American Medical Informatics Association, 2015, 22 (1): 100－8.

[11] Simmons S, Berger B. Realizing privacy preserving genome－wide association studies [J]. Bioinformatics, 2016, 32 (9): 1293－1300.

[12] Chen R, Peng Y, Choi B, et al. A private DNA motif finding algorithm [J]. Journal of Biomedical Informatics, 2014, 50 (8): 122－132.

[13] Wu X, Wei Y, Mao Y, et al. A differential privacy DNA motif finding method based on closed frequent patterns [J]. Cluster Computing, 2018 (21): 1－13.