摘要：

DNA甲基化是在高等生物中的基因组调节的重要的表观遗传修饰，其在疾病的起始和进展中起关键作用。通过甲基化特异性PCR和全基因组分析技术整合和挖掘DNA甲基化数据可以大大有助于发现新的候选疾病生物标志物。然而，如果没有人类疾病的全面DNA甲基化储存库，这是困难的。因此，我们开发了DiseaseMeth，一种人类疾病甲基化数据库。其重点是对来自各种疾病的DNA甲基化数据集的有效存储和统计分析。已收集来自多个来源的超过14000个条目和175个高通量数据集的实验信息，并将其纳入DiseaseMeth。最新版本包括来自各种技术和平台的72种人类疾病的基因为中心的甲基化数据。为了方便数据提取，DiseaseMeth支持多种搜索选项，如基因ID和疾病名称。DiseaseMeth基于疾病和正常样品的交叉数据集分析提供综合基因甲基化数据。这些可以用于深入鉴定差异甲基化基因和基因疾病关系的调查。

引言

DNA甲基化是哺乳动物基因组中的酶修饰之一（1,2）。甲基化的胞嘧啶几乎仅在CpG二核苷酸序列中。**DNA甲基转移酶（DNMT）是催化CpG甲基化的主要酶。DNA甲基转移酶1（DNMT1）负责预先存在的CpG甲基化模式后复制，而DNMT3A和DNMT3B负责从头DNA甲基化。**以前的研究表明，DNA甲基化异常是人类疾病中最常见的表观遗传事件之一，疾病组织中的DNA甲基化模式与正常对照组不同。异常低甲基化可能导致基因组不稳定，致癌基因的转录激活，印迹的丢失，而局部区域的高甲基化可能与癌细胞的选择性优势相关（6-8）。由于启动子甲基化在功能调节中的重要作用，DNA甲基化已经在诸如神经发育障碍，神经变性和神经疾病，自身免疫疾病和癌症的疾病中广泛研究。已经报道了一些重要病例。基因座特异性方法，如甲基化特异性PCR，焦磷酸测序和双酶切测序，被广泛应用于实验室。最近，基于阵列和下一代测序在基因组范围分析的高通量方法已被青睐（14）。由这些技术产生的疾病甲基化数据的集合将是有用的，并且可以用于探查来自人类疾病状态中的全甲基化的潜在的甲基化标记物/表型。

总体来说，甲基化数据库有两种：实验证据数据库和大型数据库。

MethDB保存在从多生物体CpG甲基化分析获得的DNA测定信息中甲基化胞嘧啶的发生的信息。MethPrimerDB储存了关于CpG甲基化分析的四种PCR方法的合成引物和实验信息。 MethCancerDB提供关于各种癌症中DNA甲基化的现有信息的文档，包括研究大小，癌症类型和使用的方法。 PubMeth和MeInfoText基于Medline / PubMed摘要的文本挖掘，以提取癌症中甲基化的信息。

与这些数据库相反，有两个大型甲基化数据库MethyCancer（20）和NGSmethDB（21）。 MethyCancer承载大规模甲基化参考数据，癌症相关基因和癌症信息从公共数据源。 NGSmethDB宿主基于几个基于序列的参考甲基化数据集，可用于获得基因特异性和差异甲基化信息。

然而，需要一种可以集成分散数据并提供用于深入数据挖掘的方便方式的数据库。为此，我们开发了DiseaseMeth将来自位点特异性技术的实验甲基化信息与来自甲基化构建技术的推断的基因中心甲基化状态相结合。

应该可以通过整合同一疾病的数据集来鉴定疾病中差异甲基化的基因。对特定疾病的跨数据集分析是有用的，因为对于实验生物学家来说，发现潜在的新的疾病基因/区域是困难和昂贵的。此外，结果往往是矛盾的，需要进一步确认。为了解决这些挑战，DiseaseMeth开发了通过用户友好的提取界面有效地存储和挖掘数据。目前发布的DiseaseMeth整合了72种疾病类型。此外，DiseaseMeth存储许多参考甲基化数据集，衍生自正常组织/细胞，可用于鉴定异常甲基化的基因，以及基因组数据如CpG岛，组蛋白修饰和注释基因。此外，DiseaseMeth提供：（i）可用于统计学识别基因中心甲基化差异的搜索选项，提取差异甲基化基因在疾病中与正常组织相比的详细信息，并计算特异性甲基化基因的显着性; （ii）用于计算DNA甲基化与基因 - 基因，基因疾病和疾病疾病的成对关系的相关性的工具，其可以帮助发现疾病特异性和疾病一致性基因/标记;和（iii）基因组甲基化浏览器和显示基因中心疾病甲基化信息结合基因组信息的定制视图基因组规模。简而言之，DiseaseMeth提供全面的疾病甲基化数据，并提供工具来探索疾病和DNA甲基化之间的关系。

概述和数据处理

DiseaseMeth包括基于文献的实验信息和大规模甲基化数据。超过14000条实验信息通过文本挖掘从超过25000的PubMed的DNA甲基化研究的已发表论文收集。 DiseaseMeth还拥有175个大规模甲基化数据集的50种疾病，主要从各种网站和研究所收集，如ArrayExpress，基因表达综合（GEO），UCSC。下载页面列出了有关数据集的详细信息，包括名称/ ID，疾病，数据分析，发布链接，实验平台，样本大小和下载链接

对于原始高通量数据，仅保留关于RefSeq基因的启动子区内的甲基化的信息用于分析。我们将启动子区域定义为从参考基因的转录起始位点（TSS）上游1.5kb到TSS的500bp下游的区域。对于来自UCSC 2006年3月人类参考序列（hg18，NCBI build 36.1）以外的组件的数据，我们使用UCSC（26）的LiftOver工具将其他组件的坐标转换为hg18。直接使用归一化和标准化数据（0-100％）。根据常用程序转换非规范化数据。所有归一化的甲基化数据随后在最终存储之前通过反映相关甲基化水平的百分位数归一化转变为一致的区间[0,100]。所有可下载的数据存储在GFF（通用特征格式）文件中。 DiseaseMeth数据库中的基本操作是搜索，查看，下载和分析（图1）。提供了基于MySQL后端的灵活的搜索引擎，以允许用户友好的数据挖掘和下载。可以使用基于Perl Bio :: Graphics软件包的可视化模块查看DiseaseMeth中的甲基化信息以及一些其他注释。

数据库使用和访问

使用搜索工具检索启动子的甲基化状态当提供起始和结束染色体坐标时，可以检索给定染色体区域的所有甲基化状态。 还可以检索所选疾病，组织，细胞系（cell line），技术和基因ID的数据。 提供的查询参数越多，检索的条目的范围就越窄。 有效的RefSeq ID（NM \*）或官能基因符号来获得基因为中心的甲基化信息。 默认情况下，输出将显示为概述基因的甲基化状态并提供疾病和其他信息的概述表。该表包含基于指定搜索参数生成基于疾病和正常样品的基因中心甲基化信息面板的链接。为了便于查看基因和其他相关信息，在概述面板和基因中心信息表中显示与其他资源（如GeneCard数据库）的基因特异性连接。 可以从概览面板中的链接下载查询的完整结果。

分析工具来探索基因和疾病之间的关系

然而，使用在DiseaseMeth中存储的数据，使用已经开发的定量工具很容易确定基因和疾病之间的关系的初步概况：（i）疾病 - 基因关系分析工具; （ii）疾病 - 疾病关系分析工具; 和（iii）基因 - 基因关系分析工具。 这些工具的优点之一是它们是高度可定制的，用于分析给定的区域，疾病等; 从而有助于特定分析侧重于连续区域，如印记聚类（31）。 可以下载三种分析工具。

系统设计和实现

DiseaseMeth包含三个主要软件组件：Apache HTTP服务器，MySQL数据库和使用Bio :: Perl，Bio :: Graphics和DBI软件包的Perl安装。 后端数据分析程序写在Perl并部署为CGI程序。 分析工具的Perl程序可在网站上获得。

我们将开发新的数据处理算法来处理DNA甲基化测序数据的大规模性质。由于整合分析的重要性，我们将定期收集来自新来源的数据，以增强DiseaseMeth的分析深度。我们还将鼓励将新数据直接提交给DiseaseMeth，以便更新DiseaseMeth并使其全面。基因组甲基化观察者将被改进以显示更多的（epi）基因组资源，并且将被扩展以包括更多可配置的功能。最后，将开发新的分析工具，以提供更好的集成和增强数据挖掘功能。作为研究DNA甲基化的潜在调节功能的资源，DiseaseMeth可以扩展到包括更多的数据集和工具，可以开发用于使用整合的差异甲基化鉴定算法鉴定候选基因的疾病相关的DNA甲基化标记（32） 。我们期望不断努力使用和改善DiseaseMeth将有助于我们理解DNA甲基化驱动的人类疾病。

人类疾病甲基化数据库。其重点是对来自各种疾病的DNA甲基化数据集的有效存储和统计分析。

基于疾病和正常样品的交叉数据集分析提供综合基因甲基化数据。整合同一疾病的数据集来鉴定疾病中差异甲基化的基因。集成分散数据并提供用于深入数据挖掘的方便方式的数据库。为此DiseaseMeth开发了通过用户友好的提取界面有效地存储和挖掘数据。目前发布的DiseaseMeth整合了72种疾病类型。

DiseaseMeth存储许多参考甲基化数据集，衍生自正常组织/细胞，可用于鉴定异常甲基化的基因，以及基因组数据如CpG岛，组蛋白修饰和注释基因。

DiseaseMeth提供：（i）可用于统计学识别基因中心甲基化差异的搜索选项，提取差异甲基化基因在疾病中与正常组织相比的详细信息，并计算特异性甲基化基因的显着性; （ii）开发了定量的工具用于计算DNA甲基化与基因 - 基因，基因疾病和疾病疾病的成对关系的相关性的工具，其可以帮助发现疾病特异性和疾病一致性基因/标记;（iii）基因组甲基化浏览器和显示基因中心疾病甲基化信息结合基因组信息的定制视图基因组规模。

简而言之，DiseaseMeth提供全面的疾病甲基化数据，并提供工具来探索疾病和DNA甲基化之间的关系。

DiseaseMeth数据库中的基本操作是搜索，查看，下载和分析为了方便数据提取，DiseaseMeth支持多种搜索选项，如基因ID和疾病名称。

可以使用基于Perl Bio :: Graphics软件包的可视化模块查看DiseaseMeth中的甲基化信息以及一些其他注释。

用在DiseaseMeth中存储的数据，使用已经开发的定量工具很容易确定基因和疾病之间的关系的初步概况：（i）疾病 - 基因关系分析工具; （ii）疾病 - 疾病关系分析工具; 和（iii）基因 - 基因关系分析工具。 这些工具的优点之一是它们是高度可定制的，用于分析给定的区域，疾病等;

我们开发了DiseaseMeth将来自位点特异性技术的实验甲基化信息与来自甲基化构建技术的推断的基因中心甲基化状态相结合。

DiseaseMeth开发了通过用户友好的提取界面有效地存储和挖掘数据。目前发布的DiseaseMeth整合了72种疾病类型。

整合同一疾病的数据集来鉴定疾病中差异甲基化的基因。

DiseaseMeth开发了通过用户友好的提取界面有效地存储和挖掘数据。目前发布的DiseaseMeth整合了72种疾病类型。此外，DiseaseMeth存储许多参考甲基化数据集，衍生自正常组织/细胞，可用于鉴定异常甲基化的基因，以及基因组数据如CpG岛，组蛋白修饰和注释基因。此外，DiseaseMeth提供：（i）可用于统计学识别基因中心甲基化差异的搜索选项，提取差异甲基化基因在疾病中与正常组织相比的详细信息，并计算特异性甲基化基因的显着性; （ii）用于计算DNA甲基化与基因 - 基因，基因疾病和疾病疾病的成对关系的相关性的工具，其可以帮助发现疾病特异性和疾病一致性基因/标记;和（iii）基因组甲基化浏览器和显示基因中心疾病甲基化信息结合基因组信息的定制视图基因组规模。简而言之，DiseaseMeth提供全面的疾病甲基化数据，并提供工具来探索疾病和DNA甲基化之间的关系。

DiseaseMeth数据库中的基本操作是搜索，查看，下载和分析

可以使用基于Perl Bio :: Graphics软件包的可视化模块查看DiseaseMeth中的甲基化信息以及一些其他注释。

用在DiseaseMeth中存储的数据，使用已经开发的定量工具很容易确定基因和疾病之间的关系的初步概况：（i）疾病 - 基因关系分析工具; （ii）疾病 - 疾病关系分析工具; 和（iii）基因 - 基因关系分析工具。 这些工具的优点之一是它们是高度可定制的，用于分析给定的区域，疾病等;