3.1 研究方法及方案

本项目拟采用分子动力学及量子化学计算的方法，从分子层面研究我们在BS-seq基因测序大数据分析中观察到的CpG分布及其甲基化状态对称性及对称性破缺的现象。通过计算模拟获取不同特征DNA序列的能量图谱，我们希望能从体系能量变化的角度解释CpG位点分布及其甲基化状态对称性破缺发生的内在驱动力和进化方向。在此基础之上，我们将利用这些微观理论进行模拟、分析和解释肿瘤样品中CpG位点甲基化状态相关现象发生的分子机理。本项目将采用从宏观现象到微观机理，再促进宏观现象研究的研究方案进行。具体来说，我们将使用分子动力学方法结合量子化学计算修正来从微观分子层面解释宏观测序大数据分析的结果，并将微观机理应用于更精准的测序大数据分析与挖掘之中，如图 1所示。

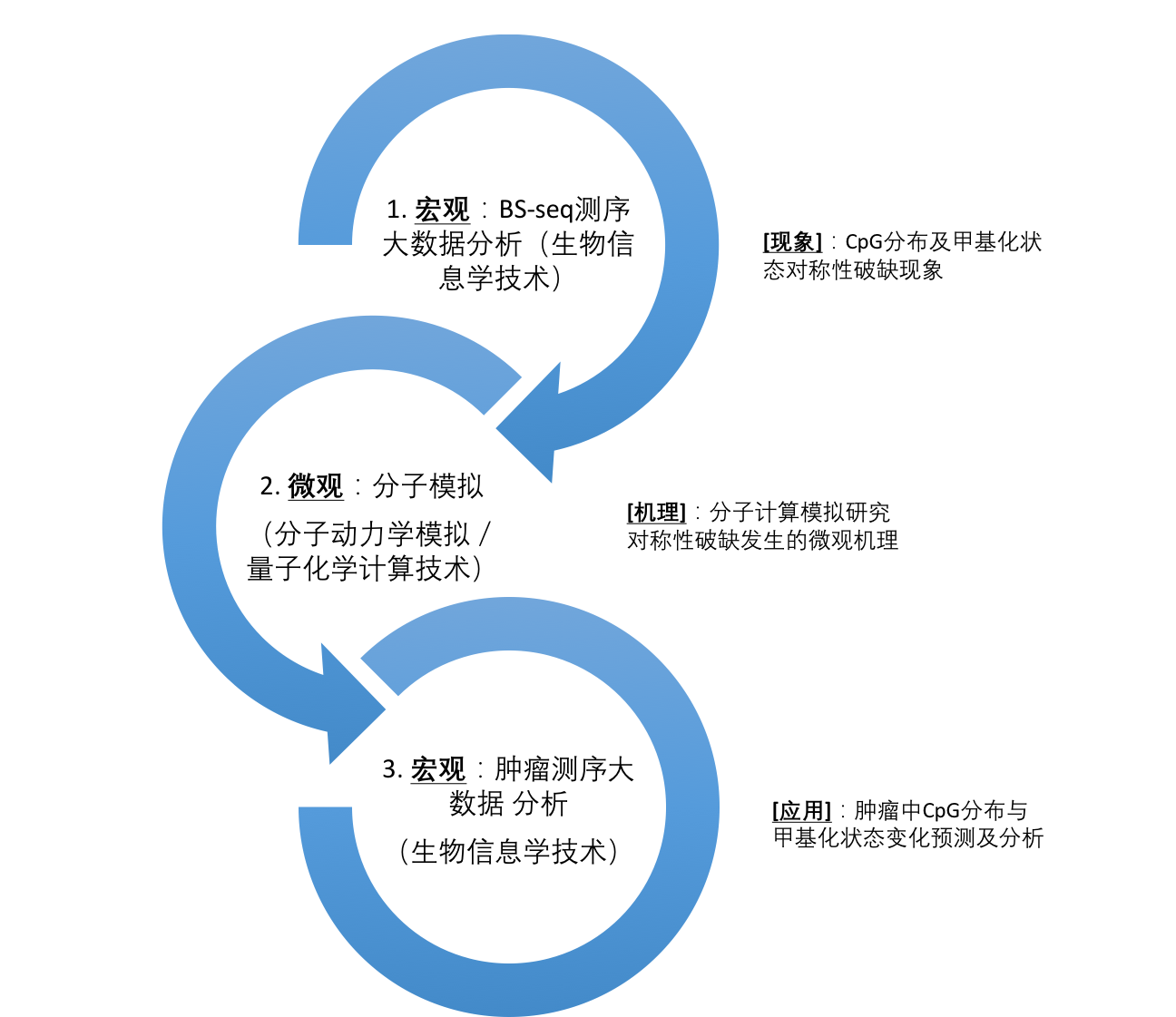


图 1本项目的研究方案

下面我们将对我们拟采用的技术进行详细的介绍：

1) 获取大量的基因测序分析数据

从低等生物到高等生物收集不同物种以及不同组织的BS-Seq测序数据。项目申请人已经建立了完善的甲基化数据处理Pipeline（https://github.com/dlmeduLi/mtbr-pipline）以及等位基因甲基化差异分析Pipeline（https://github.com/dlmeduLi/asm-pipeline）以及相关的测序数据分析与处理辅助程序。我们使用Segemhl (*Bioinformatics* 2012,28:1698-1704) 程序进行BS-Seq测序数据的比对，随后通过脚本程序来获取每个胞嘧啶位置的甲基化状态信息。经过数据质量控制（根据每个位点的覆盖次数、甲基化Read的数量，甲基化水平等等综合确定）筛选出比较可信的数据。这些高质量的数据将被用来进行后续的深入分析（例如：通过使用Fisher检验、χ方检验等等统计方法进行甲基化差异性水平分析 ，基于Tag模型建立tagmeth indexing做等位基因甲基化差异分析等等）。测序数据的分析是我们后续工作的基石，目前该阶段的研究工作已经基本完成，后续会根据项目的需求进行更加细化的 算法开发及程序编写工作。

2) 建立对称性及对称性破缺相关的数学模型

如何建立科学系统的数据描述模型是生物信息学数据挖掘的一个非常重要的方面。我们往往可以通过数学模型观察到大量复杂数据背后隐藏的规律，进而加深我们对跟测序数据相关的各种宏观现象的理解。在本项目中我们打算使用对称性语言来系统化地建模与分析DNA中CpG位点分布及甲基化状态相关的现象。除了常见的各种统计分布模型之外（例如β分布、负二项分布等等），在该项目中我们将尝试引入群论方法来系统化地描述我们在测序数据分布中观察到的各种对称性及对称性破缺的现象。与直观的空间对称性相比较，序列数据的对称性不是那么显然。我们将从分子组装的角度定义序列单元（例如CpG等二核苷酸序列）的对称操作元素。从序列数据产生的角度建立跟测序数据分析吻合的数学模型。

3) 获取特征DNA序列数据

由于计算资源与计算能力的限制，目前计算机能模拟的原子规模是非常有限的。对于普通的小规模集群计算机来说，经典分子动力学大约能模拟十万原子的系统，而量化动力学只能模拟数百到数千的原子。虽然理论上越长的序列我们越能得到更精确的结果，实际只能模拟有限长度的序列。这一步的研究分为两个方面：

A）产生有限长度的具有特征对称性特性的DNA序列

特征序列的产生将使用项目研究的第二步中确定的对称性模型。本项目拟使用隐马尔可夫模型（HMM）作为序列数据生成器。大量的满足我们在测序数据中观测到的分布状态序列将被产生，每个被产生的序列将被做对称性检测，如果不符合对称性分布要求的将被抛弃，满足条件的将作为下一个序列启动序列。直到产生指定长度的序列。

B）确定特征DNA序列长度的边界值

原则上我们将选取符合对称性要求的最短长度的特征序列。这里我们将引入序列特征能量的概念。当特征序列的长度增加时，该特征能量将有明显的收敛趋势。亦即当序列长度大到一定值之后，长度的增加只会影响体系能量的细微变化。这些细微的变化不会影响项目研究的序列对象，因此可以当成微扰忽略。我们将不停增加序列长度，同时计算序列的特征能量，特征能量的变化满足我们的判决阀值时，我们将此时的长度作为序列长度的边界值。

综合起来说，如果我们定义序列字母集，定义HMM产生的符合我们观测的序列为，其中，定义DNA序列的特征能量为，那么特征DNA序列的产生产生算法如图 2所示。

初始序列

HMM产生序列



满足对称性要求？

计算序列特征能量





收敛标准？

最终序列

是

否

是

否

图 2特征DNA序列生成算法

4) 甲基化三维空间结构的量子化学修正

使用的DNA结构模型越准确，获取的能量图谱将越精确。量子化学的计算精度和准确度远远高于经典分子动力学。由于计算规模的限制，完全使用量子化学计算模拟的方式将消耗大量的计算资源，甚至由于计算规模过大而无法获取到计算结果。这里我们采用QM/MM结合的方法，对于甲基化CpG等影响项目计算精度与结果的关键位点使用第一性原理计算的方法进行模型局部结构的精细调整。本项目拟采用基于密度泛函理论(DFT)的第一性原则原理计算程序VASP（J. Chem. Phys. 1984, 81, 511）计算DNA甲基化反应路径，从而确定精确的甲基化修正模型。CpG甲基化反应的反应路径及反应过渡态确定将使用Nudged Elastic Band （NEB）方法（J. Chem. Phys. 2012, 136, 074103）来确定。

5) 分子动力学计算模拟获取能量图谱及对称性破缺机理分析

本项目将使用分子动力计算的方法获取特征DNA序列的最合理的三维空间结构和能量信息。项目拟采用分子动力学程序NAMD（J. Comput. Chem. 2005, 26:1781-1802）计算上几个阶段产生的大量特征序列结构模型的最合理状态的空间构型和体系能量。特征 DNA分子将浸泡在经过KCl中和过的水溶液模型中，计算过程将在NPT系宗下进行，特征DNA分子将在300K温度下进行充分的弛豫反应。项目将使用VMD程序结合相应的TCL/TK脚本程序进行结果分析。项目将建立对称性及对称性破缺特征DNA序列的详细能量图谱图。

本项目拟通过获取完整的对称性及能量变化关系来研究并分析DNA序列中的对称性及对称性破缺现象。通过分子动力学计算得到的能量图谱及其他相关数据将用来解释CpG分布及其甲基化状态对称性及对称性破缺的分子机理。

6) 肿瘤测序数据分析

本项目拟通过对原子及分子组装层次的机理研究来解释我们观察到的DNA序列中对称性及对称性破缺现象。在我们的分析中，肿瘤基因组中这些对称性及对称性破缺现象更加多样化及复杂化。项目的最终步骤是进行肿瘤基因组数据的分析，一方面肿瘤数据的分析验证我们得到的结论，另一方面对肿瘤基因组中CpG及其甲基化相关的对称性破缺现象进行预测性解释。 这一阶段工作包括 肿瘤组织样品的收集、建库、测序和数据的分析工作。

3.2 技术路线

本项目的技术路线：

甲基化状态分析MTBR-Pipeline

https://github.com/dlmeduLi/mtbr-pipline

CpG分布及密度分析

https://github.com/dlmeduLi/methyutils

等位基因甲基化差异分析ASM-Pipeline

https://github.com/dlmeduLi/asm-pipeline

组织及DNA特征区域甲基化差异分析

Bs-seq测序数据

公共数据库

自主建库测序

CpG分布及其甲基化状态的对称性及对称性破缺的数学模型

满足对称性模型的特征DNA序列

使用NAB语言进行不同 CpG状态DNA分子结构模型建立

使用NAMD进行分子动力学模拟与计算，获取不同特征序列的能量完整能量图谱，并分析计算结果

使用NAB语言进行不同甲基化状态DNA分子结构模型建立

量子化学计算进行甲基化反应局域结构优化

生物信息学数据分析技术

对称性建模

分子模拟技术

图 3. DNA甲基化

图 4 本项目拟采用的技术路线图

3.2 可行性分析

本研究项目涉及到新一代测序大数据分析和分子动力学计算两个大的领域，其中涉及到的数学模型建立、计算及程序编写以及分子模型建立等等各方面，申请人都有非常丰富的研究经历并且在该项目的研究方向上已经做好了扎实的前期基础工作。

1) 测序大数据分析

本项目涉及到Bs-seq测序大数据的分析，在这个研究方面申请人所在的实验室已经有了丰富的积累。由申请人主导已经完成了两条Bs-seq测序数据分析的pipeline（CpG分布及甲基化状态分析MTBR-Pipeline <http://www.github.com/dlmeduli/mtbr-pipline/> 以及等位基因甲基差异表达区域分析ASM-Pipeline <http://www.github.com/dlmeduli/asm-pipline/>），同时也积累了大量相关的测序数据分析的算法和程序。申请人在科学计算程序算法设计及开发方面的积累，保证了在项目需要时，能发展新的方法来推进项目进度。

2) 对称性及对称性破缺数学建模

本项目拟从对称性方面解释我们在测序数据分析中观察到的现象，并以此为基础发展新的数据分析方法和手段。申请人在晶体结构学方向有开发新算法的研究经历，对其背后的物理数学知识包括空间群、空间变换、群论等等都有深刻的理解。这些数学物理方面的基础是该项目顺利能完成的基石。

3) 分子动力学模拟与量子化学计算

申请人长期以来一只从事原子尺度的科学计算模拟研究工作。在分子动力学研究方向上，申请人自主开发过分子力场，并使用新的力场做了大量分子模拟方向上的研究工作。这表明申请人不仅有丰富的实践经验，对分子动力学底层机理及机制有深刻的理解，这些保证了项目在分子动力学方向能顺利进行。申请人在大分子与固体表面化学物理吸附反应的第一性原理计算研究经历是本项目中提出的使用量子化学做局域结构修正的的保证。

5) 计算资源

申请人所在的实验室拥有完全自主的计算资源。目前实验室拥有12台高性能（24计算核心、128G内存、24T硬盘）完全能满足项目所设计的测序数据分析以及分子动力学计算模拟等各方面计算资源的需求。

4) 研究团队

申请人所在的团队是一个积聚了生物、医学、计算机科学等等生物信息学相关的各个领域优秀人才的交叉学科的团队。强大的智囊与各种技术支持保证了该项目的顺利实施。

4. 本项目的特色与创新之处

1) 本项目提出从对称性及对称性破缺的角度对DNA中CpG分布及其甲基化状态进行系统描述。这些新的数学物理视角可以推广并应用到其他测序数据（例如癌症测序数据等等）分析上。

2) 本项目尝试从分子结构与能量等原子水平的角度入手，解释宏观测序数据观测到的结果。宏观观察到的结果一般都有其微观方向的影响因素，本项目拟基于测序数据分析的结果，从更底层解释其发生机理，而不单纯发掘其在更高尺度上的应用。

5. 年度研究计划

5.1 年度研究计划

1) 2017.1 – 2017.12 搜集更多的Bs-seq测序数据，编写该项目需求的更详细化的数据分析pipeline。完成更多物种、更多组织的数据分析，建立起系统化的CpG及其甲基化相关的数据仓库。

2) 2018.1 – 2018.12 建立描述CpG及其甲基化对称性与对称性破缺的数学模型，开发算法进行模拟数据的验证。同时，基于该数学模型完善特征DNA序列生成算法，完成相应的程序编写工作，并根据这些特征序列进行自动化分子结构模型建模。在该阶段开始探索性分子动力学计算。

3) 2019.1 – 2019.12 使用VASP进行第一性原理计算，进行DNA甲基化反应结构调整。完善之前的分子结构模型，并开始大规模进行分子动力学的模拟与计算。获取计算结果对计算结果进行分析总结。

5.2 预期研究成果

1) 建立完善的CpG及其甲基化状态相关的对称性及对称性破缺的数学模型，并基于此开发出容易使用的测序数据分析pipeline。

2) 从分子结构层面上阐述CpG及其甲基化状态相关的对称性及对称性破缺的机理，并发表SCI学术论文1-2篇。

3) 培养1-2名硕士研究生

5.3 学术交流

1) 邀请1-2 位统计生物信息学领域著名的专家教授交流1-2 次。

2) 参加生物信息学或分子计算与模拟相关领域的国际会议1-2 次。