（一） 立项依据与研究内容（4000-8000 字）：

1. 项目的立项依据（研究意义、国内外研究现状及发展动态分析，需结合科学研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录）：

1.1 对称性及对称性破缺是生命进化的一个重要驱动力，对 DNA 中对称性及对称性破缺现象发生的机理进行深入研究是解密 DNA 中蕴含的生命信息的重要理论基础

生命的进化是一个从混沌到有序，从低序到高序的过程，也是一个从完全对称性到对称性破缺以及非对称性的过程。P.W. Anderson 在他物理诺贝尔得奖演讲中指出多体系统中对称性破缺水平的高低与系统复杂度和功能专门化密切相关[1]。在生命科学中，各个尺度上系统功能的分化都与相应的对称性破缺密切相关，从分子组装到亚细胞结构，从多种多样的细胞类型的出现到机体组织的分化，以至于到胚胎发育等等都是如此[2]。对称性破缺也被认为是生物新突变形式的来源与生物进化的内在驱动力[3]。

此外，生物在大尺度上的对称性破缺往往根源于其小尺度上对称性破缺的发生。亚细胞结构层次的对称性破缺可能会导致持续的极性生长从而产生不同的细胞形状来满足细胞分化、细胞融合、神经细胞轴突等等各种需求[4–7]。例如，P. Alexis 等在研究拟南芥胚轴生长对称性破缺的原因时发现拟南芥胚轴生长对称性破缺源于壁力学上的细胞非对称性的发生，而细胞壁受力非对称性的发生又是由双极果胶的甲酯化来触发[8]。因此，我们观察到的生物系统在宏观大尺度上的对称性破缺现象往往可以通过生物系统的微观小尺度上的对称性破缺机理来解释。生命的多样性与复杂的功能性都是某种形式的分子组装水平的对称性破缺造成的。

申请人在前期研究工作中发现，随着物种的进化 DNA 中 CpG 的含量发生了非常明显的向不均匀方向变化的趋势，与此同时 CpG 的甲基化水平也有明显的关联性变化。一方面从低等生物到高等生物的进化中 CpG 密度比其他二核苷酸的密度逐渐降低；另外一方面，CpG 的分布逐渐由对称性分布向非对称性分布逐渐转变（研究基础图 7）。同时我们也发现 DNA 中 CpG 的密度与其甲基化

水平存在互补关系（研究基础图 8），然而在一些特定的 DNA 功能区这些对称性关系又发生了破缺（研究基础图 9）。DNA 中 CpG 二核苷酸的对称性及对称性破缺现象必然有着其内在分子水平上的作用机制。本项目拟建立描述这些变化趋势的对称性模型。目前，基于群论等数学工具的各种对称性模型（例如 Z-曲线）已经开始应用于核酸研究之中[9]。

对称性破缺往往意味在外界扰动作用下，体系能态发生变化，体系从稳定状态转变为的亚稳态，此时系统的对称性不发生变化。位于亚稳态状态下的系统极其不稳定，任何微扰都会使系统从亚稳态到新的稳态的转变，系统从对称性转变为非对称性，对称性发生破缺，如图 1 所示。**体系能态的变化是研究对称性破缺发生的关键，CpG 的对称性破缺必定会反应出相应的能态变化**。分子动力学是计算和获取 DNA 分子构型及能量的非常有效的手段。因此，本项目拟通过分子动力学方法从体系能态变化的角度入手研究我们观察到的 DNA 中 CpG 相关的对称性与对称性破缺现象，深入理解对称性破缺发生的诱导因素和发生机理。

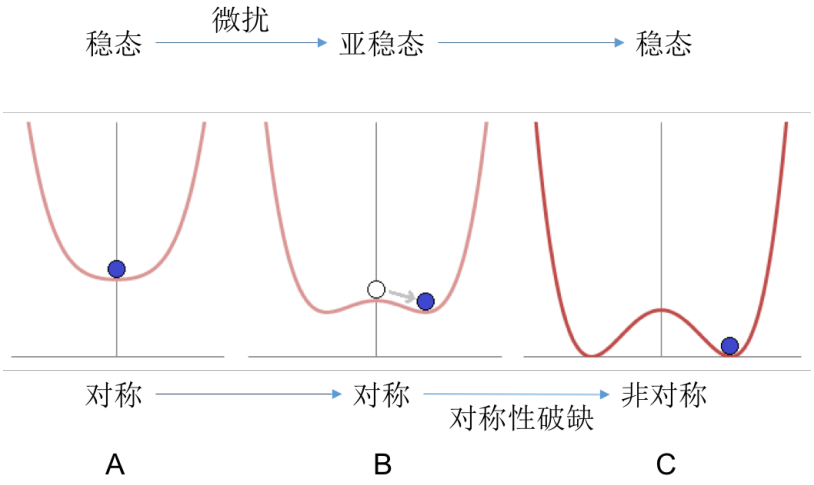

图 1 系统发生对称性破缺示意图[10]。A) 初始状态下系统处于稳态，具有对称性；B）系统在外界扰动作用下，能态发生变化，转变为极其不稳定的亚稳态，此时系统对称性不变；C）系统从亚稳态向稳定状态发生转变，系统获得新的稳定，对称性发生破缺。

1.2 DNA 甲基化是哺乳动物发育以及各种致病发生的重要影响因素，其主要发生在 CpG 位点上

DNA 甲基化（如图 2 所示）是基因组中最关键的表观遗传修饰之一[11,12]，它与许多的细胞活动都密切相关，包括胚胎发育[11,13]、基因转录[14]、染色质结构[15]、X 染色体失活[16,17]、基因印记[18–20]以及染色体稳定性[21,22]等等。正是因为 D

NA 甲基化的这些重要作用，越来越多的人类疾病被发现与 DNA 甲基化异常相关。

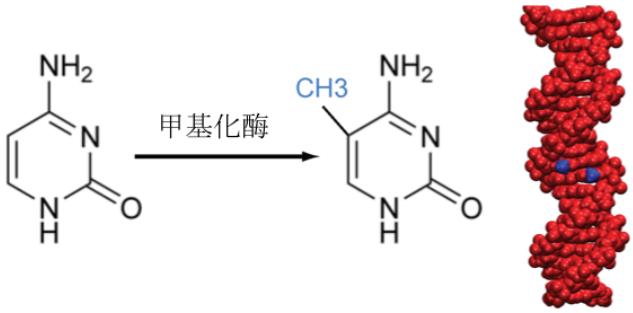

图 2. DNA 甲基化修饰反应示意图。

DNA 中的甲基化修饰大部分都发生在 CpG 位点上。人类基因组中被甲基化的腺嘌呤中大约有 55-80%位于 CpG 上，而相比之下只有 1.3%左右的甲基化发生在非 CpG 位点上[23,24]。因此，基因组的甲基化异常主要反应在 CpG 位点的甲基化异常上。理解 CpG 分布及甲基化对称性与对称性破缺的发生机理将有利于我们理解基因组甲基化修饰异常发生的原因，进而给我们理解相关疾病的发生提供重要的理论指导。

1.3 分子动力学已经成为核酸结构功能研究的一个不可或缺的重要手段

一直以来我们知道 DNA 的结构具有多态性而且极其不稳定。DNA 结构的实验数据主要来自 X-射线衍射晶体结构测定的结果，然而跟 DNA 巨大的构型空间比较起来，X-射线衍射的方法能获取到的 DNA 分子结构构型只能占其中非常小的一部分。除此之外，由于 DNA 的不稳定性，与 DNA 结构测定的相关实验实验都有很多的局限性。与此相比分子动力学模拟（MD）具有很大的优势，该方法不仅消除了诸多实验上的限制，而且能直观观察到分子的反应变化过程，**分子动力学模拟（MD）方法已经成为研究核酸的一个不可或缺的重要手段**[25]。

如图 3 所示，分子动力学已经成功应用在 DNA 研究的各个方面，包括物理性能[22]（例如 DNA 的拉伸力学特性[26,27]），DNA 的构型转变（例如从$A \leftrightarrow B$以及从$B \leftrightarrow Z$的构型转变[28,29]），DNA 与蛋白质大分子的相互作用[30]等等。因此，分子动力学方法是研究 DNA 的一个非常有效的方法。

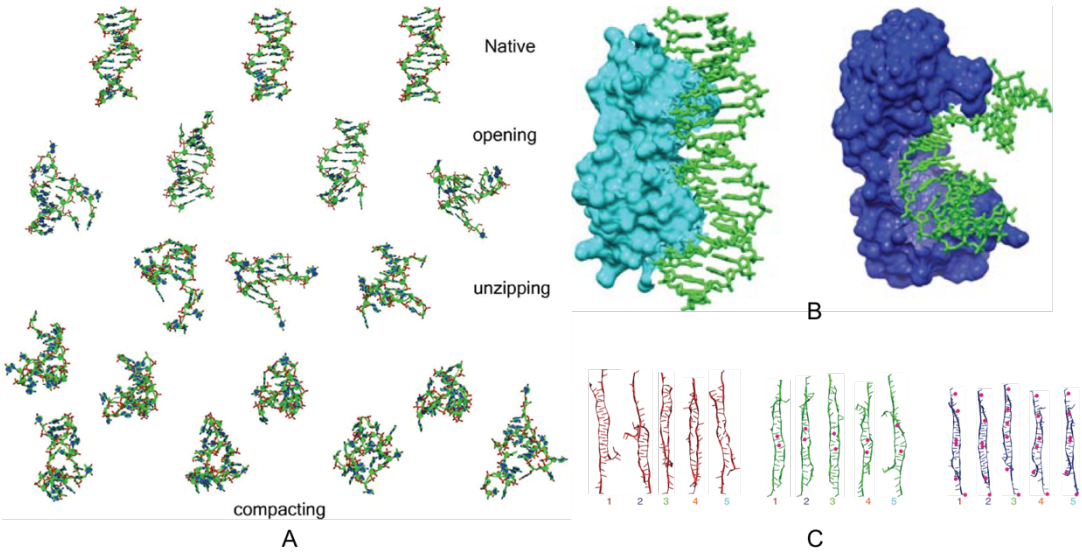

图 3 分子动力学被广泛应用于 DNA 研究中。A）分子动力学方法用来获取 DNA 分子的各种理论构型 [25]；B）分子动力学用于研究 DNA 分子与蛋白质的相互作用 [31]；C）分子动力学用来研究 DNA 分子的力学拉伸 [27]。

申请人在分子动力学方向上以及原子水平的计算模拟方向上有非常丰富的研究经验，包括使用量子化学方法计算大分子的表面吸附反应（Appl. Surf. Sci.，2014）以及自主构建分子立场并进行大量分子动力学模拟（J. Appl. Phys.，2014）等等。基于这些研究基础，本项目拟采用分子动力学方法，对不同 CpG 对称性与非对称性状态下的 DNA 分子构型和能态进行系统的研究，获取对称性与能量图谱和分子构型之间的关系，进而深入理解对称性破缺发生的内在机理。

1.4 小结

近些年来越来越多的疾病被发现与 CpG 的甲基化异常相关，因此深入理解 DNA 中 CpG 分布及其甲基化的状态发生变化的机理是至关重要的。本项目拟从对称性及对称性破缺的角度来描述 DNA 中 CpG 及及甲基化状态的变化并采用分子动力学计算的方法，从分子结构和能量图谱的变化上深入研究这些变化发生机理。该项目结合了宏观的生物信息学分析技术以及微观的分子动力学模拟技术，是研究思路上的创新，是理解 DNA 中甲基化修饰相关的表观遗传信息的关键，对揭示由甲基化异常所引起的疾病的分子机理有着重要的意义。

【参考文献】

1. Anderson, P. W. More is different. *Science* **177,** 393–396 (1972).
2. Li, R. & Bowerman, B. Symmetry Breaking in Biology. *Cold Spring Harb.*

Perspect. Biol. **2,** (2010).
3. Palmer, A. R. Symmetry Breaking and the Evolution of Development. *Science* **306,** 828–833 (2004).
4. Dworkin, J. Cellular Polarity in Prokaryotic Organisms. *Cold Spring Harb. Perspect. Biol.* **1,** (2009).
5. Chang, F. & Martin, S. G. Shaping Fission Yeast with Microtubules. *Cold Spring Harb. Perspect. Biol.* **1,** (2009).
6. Slaughter, B. D., Smith, S. E. & Li, R. Symmetry Breaking in the Life Cycle of the Budding Yeast. *Cold Spring Harb. Perspect. Biol.* **1,** (2009).
7. Tahirovic, S. & Bradke, F. Neuronal Polarity. *Cold Spring Harb. Perspect. Biol.* **1,** (2009).
8. Peaucelle, A., Wightman, R. & Höfte, H. The Control of Growth Symmetry Breaking in the Arabidopsis Hypocotyl. *Curr. Biol.* **25,** 1746–1752 (2015).
9. Zhang, R. & Zhang, C.-T. A Brief Review: The Z-curve Theory and its Application in Genome Analysis. *Curr. Genomics* **15,** 78–94 (2014).
10. https://en.wikipedia.org/wiki/Spontaneous_symmetry_breaking.
11. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat Rev Genet* **14,** 204–220 (2013).
12. Robertson, K. D. DNA methylation and human disease. *Nat Rev Genet* **6,** 597–610 (2005).
13. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell* **99,** 247–257 (1999).
14. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–1022 (2011).
15. Thomson, J. P. *et al.* CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464,** 1082–1086 (2010).
16. Augui, S., Nora, E. P. & Heard, E. Regulation of X-chromosome inactivation by the X-inactivation centre. *Nat. Rev. Genet.* **12,** 429–442 (2011).
17. Gendrel, A.-V. *et al.* Smchd1-Dependent and -Independent Pathways Determine Developmental Dynamics of CpG Island Methylation on the Inactive X Chromosome. *Dev. Cell* **23,** 265–279 (2012).
18. Falls, J. G., Pulford, D. J., Wylie, A. A. & Jirtle, R. L. Genomic Imprinting: Implications for Human Disease. *Am. J. Pathol.* **154,** 635–647 (1999).
19. Reik, W. & Walter, J. Genomic imprinting: parental influence on the genome. *Nat Rev Genet* **2,** 21–32 (2001).
20. Feinberg, A. P., Cui, H. M. & Ohlsson, R. DNA methylation and genomic imprinting: insights from cancer into epigenetic mechanisms. *Semin. Cancer Biol.* **12,** 389–398 (2002).
21. Derreumaux, S., Chaoui, M., Tevanian, G. & Fermandjian, S. Impact of CpG methylation on structure, dynamics and solvation of cAMP DNA responsive element. *Nucleic Acids Res.* **29,** 2314–2326 (2001).
22. Pérez, A. *et al.* Impact of Methylation on the Physical Properties of DNA. *Biophys. J.* **102,** 2140–2148 (2012).

23. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462,** 315–322 (2009).
24. Laurent, L. *et al.* Dynamic changes in the human methylome during differentiation. *Genome Res.* **20,** 320–331 (2010).
25. Pérez, A., Luque, F. J. & Orozco, M. Frontiers in Molecular Dynamics Simulations of DNA. *Acc. Chem. Res.* **45,** 196–205 (2012).
26. Severin, P. M. D., Zou, X., Gaub, H. E. & Schulten, K. Cytosine methylation alters DNA mechanical properties. *Nucleic Acids Res.* gkr578 (2011). doi:10.1093/nar/gkr578
27. Severin, P. M. D., Zou, X., Schulten, K. & Gaub, H. E. Effects of cytosine hydroxymethylation on DNA strand separation. *Biophys. J.* **104,** 208–215 (2013).
28. Kastenholz, M. A., Schwartz, T. U. & Hünenberger, P. H. The transition between the B and Z conformations of DNA investigated by targeted molecular dynamics simulations with explicit solvation. *Biophys. J.* **91,** 2976–2990 (2006).
29. Noy, A., Pérez, A., Laughton, C. A. & Orozco, M. Theoretical study of large conformational transitions in DNA: the B<-->A conformational change in water and ethanol/water. *Nucleic Acids Res.* **35,** 3330–3338 (2007).
30. Mackerell, A. D. & Nilsson, L. Molecular dynamics simulations of nucleic acid-protein complexes. *Curr. Opin. Struct. Biol.* **18,** 194–199 (2008).
31. Laughton, C. A. & Harris, S. A. The atomistic simulation of DNA. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1,** 590–600 (2011).

2. 项目的研究内容、研究目标，以及拟解决的关键科学问题（此部分为重点阐述内容）；

2.1 研究内容

本项目围绕揭示进化中 CpG 分布及甲基化状态对称性与对称性破缺的发生机理展开研究。我们拟使用分子动力学方法（Molecular Dynamics）结合基于第一性原理的量子化学计算来建立与分析不同对称性状态与DNA分子构型和能态变化之间的关系，进而阐述对称性破缺的发生机理。随后，我们将通过对肿瘤 Bs-seq 测序数据的分析从实验的角度来验证我们的理论。下面，我们将逐步对拟开展的研究内容进行阐述。

1） 运用生物信息学分析方法获取不同物种的CpG分布及其甲基化分布状态，建立完善的对称性描述数学模型（**已部分完成**）

获取大量的试验数据将是我们进行后续研究的工作基础，我们将在这一阶段通过生物信息学方法，编写数据分析 pipeline，获取 CpG 分布及其甲基化状态数据，并据此建立系统的数学方法用来描述 CpG 及其甲基化状态的对称性及其对称性破缺现象。

2） 分析测序实验数据，根据对称性数学模型，设计算法产生符合对称性特征的有限长度的DNA序列（**已部分完成**）

由于计算资源的限制，实际能进行模拟的序列长度必定将受到限制。理论上越长的序列越容易得到更符合实验条件下的模拟结果，但是由于目前计算机的计算能力以及计算资源限制，我们只能使用有限长度的 DNA 序列在精度允许的范围内进行模拟。如何在实验数据的结果之上建立数学模型，开发算法，产生有限长度的、符合对称性特征的 DNA 序列将是我们研究内容的一个非常重要的方面。

3） 根据特征DNA序列的建立初始DNA分子三维结构模型并使用量子化学计算方法对CpG甲基化进行局域结构修正（**已部分完成**）

DNA 三维空间构型将对计算模拟的结果有重要的影响。合理的初始结构模型不仅能让计算模拟快速收敛，而且能得到更精确的结果。此外，CpG 位点的甲基化结构是影响 DNA 构型和能态的一个重要方面。甲基化修饰能会影响 DNA 的一些物理特性，如稳定性、亲和性等等。因此，为了获取更精确的计算结果，

我们将使用量子化学计算的方法对 DNA 甲基化的反应机理进行细致的研究，优化甲基化局域结构模型，以期获得更合理更准确的模拟结果。

4） 使用优化之后的初始DNA分子结构模型进行分子动力学模拟，获取不同的CpG分布及其甲基化对称性状态下的能态图谱，分析计算结果并阐述对称性破缺发生的机理

该项目将使用分子动力学方法获取不同对称性特征序列的能量变化图谱，并根据对称性与能态变化之间的关系来分析 DNA 中 CpG 分布及其甲基化状态发生对称性破缺的内在机理。

5） 使用生物信息学方法分析肿瘤Bs-seq测序数据，验证我们根据分析结果得出的结论

分析得到的对称性破缺理论需要经过实验数据的进一步验证。在理论模型的基础之上，有针对性的对肿瘤测序数据进行细化分析，一方面可以验证我们得出的理论模型，另一方面也尝试可以发展新的生物信息学分析思路与方法。

2.2 研究目标

本研究项目的目标是希望通过分子动力学及量子化学计算方法，建立对称性与能态变化之间的关系图谱，进而研究和解释进化过程中 CpG 分布及其甲基化发生对称性破缺的内在机理。这对深入理解甲基化修饰相关的疾病的发生有重要的理论指导意义。此外，该项目拟采用的方法比较新颖，可以为生物信息学数据分析提供新的思路。总结起来，该项目有以下几个的目标：

1） 为CpG分布及其甲基化状态建立完善的对称性数学模型。

2） 通过量子化学计算揭示 CpG位点甲基化反应与非CpG位点甲基化反应的微观反应过程及二者之间的区别。

3） 揭示CpG分布及其甲基化对称性破缺发生的机理，包括激发条件、变化趋势和驱动力等。

4） 发展基于对称性破缺机理的新的测序数据分析手段和方法，并为其建立可公开使用的pipeline。

2.3 拟解决的关键科学问题

本项目将围绕以下几个科学问题进行：

1） 建立CpG分布及其甲基化状态对称性数学模型。

尽管目前统计学及数据挖掘等方法已经大量用于测序数据分析，但是这些都是非确定性数学模型（nondeterministic model）。我们将使用传统的统计学方法结合群论等数学工具建立 CpG 及其甲基化的确定性数学模型（deterministic model）。

2） 产生符合对称性条件的、有限长度的特征DNA序列。

由于实际模拟计算中，计算机能力和计算资源的限制。我们只能模拟有限长度的 DNA 序列。我们将采用隐马尔可夫模型（HMM）结合特征能量条件判据，迭代产生特征符合要求的特征 DNA 序列。

3） CpG分布及其甲基化状态发生对称性破缺的机理。

本项目将采用分子动力学模拟的方法，结合量子化学计算，获取不同特征序列的最优化空间构型和能量图谱。随后我们建立新的分析算法对分子动力学计算结果进行分子构型和能量变化等各方面详尽的分析，建立对称性与能量变化的关系，进而揭示对称性破缺的内在机理。

3. 拟采取的研究方案及可行性分析（包括研究方法、技术路线、实验手段、关键技术等说明）；

3.1 研究方法及方案

本项目将采用分子动力学及量子化学计算的方法，从原子尺度研究我们在 BS-seq 基因测序大数据分析中观察到的进化中 CpG 分布及其甲基化状态发生变化的对称性破缺的机理。通过计算模拟获取不同特征 DNA 序列的能量图谱，我们将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，验证我们的理论并尝试发展新的生物信息学分析方法。本项目将采用从宏观现象到微观机理，再反过来改进宏观现象分析的螺旋式研究方案进行。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 4 所示。

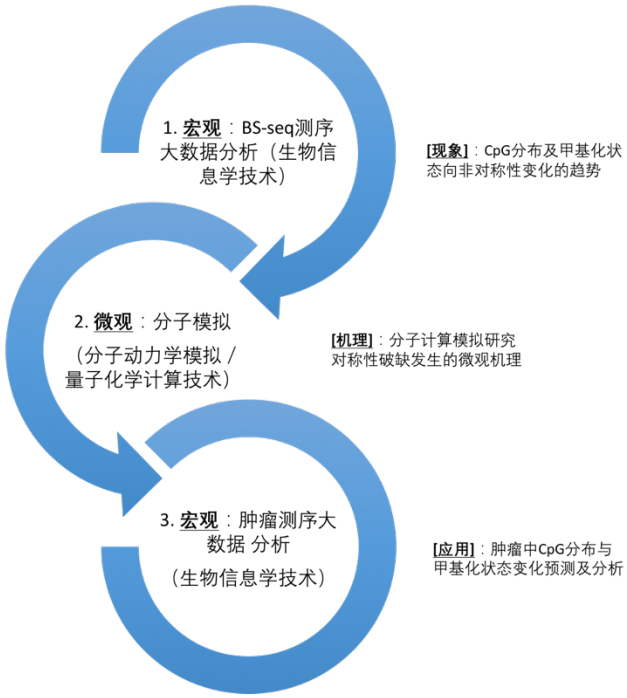

图 4 本项目的研究方案

下面我们将对我们拟采用的研究方案进行详细的介绍：

1） 获取大量的基因测序分析数据

收集从低等生物到高等生物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）、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 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、卡方检验等等统计方法进行甲基化差异性水平分析或者基于 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，进而加深我们对测序数据相关的各种宏观现象的理解。在本项目中我们拟从对称性角度来系统化地建模与分析 DNA 中 CpG 位点分布及甲基化状态发生的各种变化。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。与具有显式变量的空间对称性不同，序列数据的对称性变量是隐式的。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对称性数学模型。

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，而基于第一性原理的量化动力学只能模拟数百左右的原子。虽然理论上越长的序列我们越能得到更精确的结果，但是考虑到实际情况，我们只能通过模拟

有限长度的序列在误差允许的范围内进行近似。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和计算允许的精度的最长特征序列。这里我们将使用序列特征能量来进行判据。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，当特征能量的变化满足我们的项目允许的计算精度时，我们将此时的长度作为序列长度的边界值。

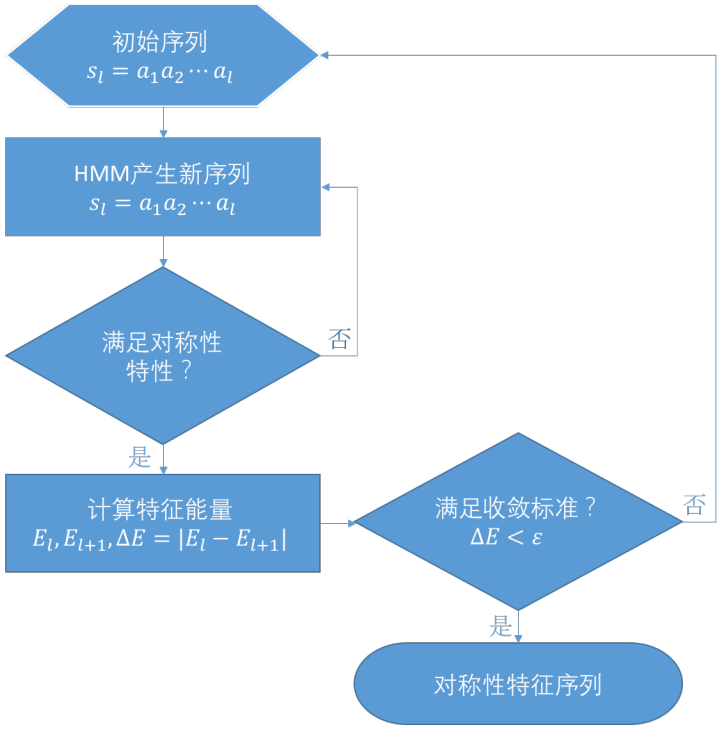

图 5 特征 DNA 序列生成算法

将这两步结合起来，特征 DNA 序列的产生产生算法如图 5 所示。其中，序列字母集$\Omega = \{T, A, C, G, CpG\}$，HMM 产生的序列为$s_l = a_0 a_1 \cdots a_l$，$a \in \Omega$，$E_l$为 DNA 序列$s_l$的特征能量。

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 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 NEB（Nudged Elastic Band）方法来确定。

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。项目将使用分子动力学程序 NAMD 计算前面几个步骤中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 分子将浸泡在使用 KCl 离子中和过的水溶液模型中，整个模拟将在 NPT 系宗条件下，在指定的温度范围内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并建立对称性及对称性破缺特征 DNA 序列的详细能量图谱图。

本项目将通过获取完成的对称性及能量对应关系来研究和分析 DNA 序列中的对称性及对称性破缺现象。使用分子动力学计算得到的能量图谱及其他相关的数据将用来解释 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） 从肿瘤测序数据分析验证对我们提出的称性破缺机理

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进化中 CpG 及其甲基化状态向对称性变化的趋势。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，另一方面针对肿瘤基因组中 CpG 及其甲基化相关的对称性破缺现象发展新的数据分析方法。这一阶段工作包括肿瘤组织样品的收集、建库、测序和生物信息学分析工作。

3.2 技术路线

本项目拟采用的的技术路线如图 6 所示：

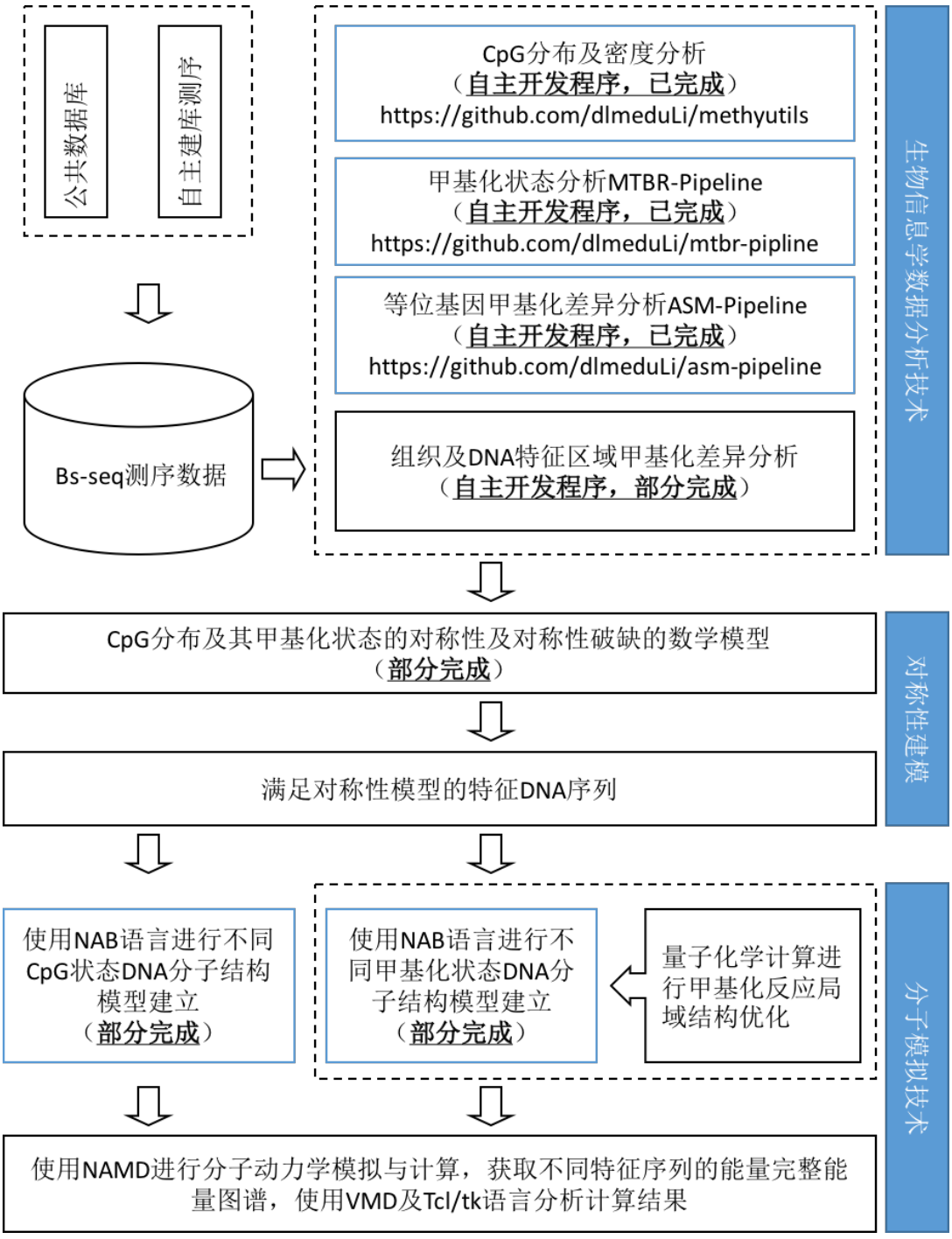

图 6 本项目拟采用的技术路线

3.3 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 的建立包括 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 CpG 及其甲基化状态向非对称性变化趋势的现象，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

申请人长期以来一直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可行性保证。

4） 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。

5） 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。强大的智囊团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；

A. 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数学建模。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，并发展出新的分析方法。

B. 本项目尝试结合宏观的生物信息学分析技术和微观的分子动力学模拟技术，从分子结构与能量等微观原子水平入手，解释宏观测序数据观测到的结果，该项目可以为测序数据分析提供新的研究和分析思路。

5. 年度研究计划及预期研究结果（包括拟组织的重要学术交流活动、国际合作与交流计划等）。

5.1 年度研究计划

A. 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尽的数据分析pipeline。完成更全面的物种、更丰富的组织样品的数据分析，建立起系统化的CpG及其甲基化相关的数据仓库。

B. 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。在该阶段开始尝试性分子动力学计算。

C. 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。获取计算结果并对计算结果进行分析总结。

5.2 预期研究成果

A. 建立完善的CpG及其甲基化状态对称性与对称性破缺的数学模型，并基于此开发出面向公众的测序数据分析pipeline。

B. 从分子结构层面上阐述CpG及其甲基化状态相关的对称性与对称性破缺的机理，并发表SCI学术论文2-3篇。

C. 培养1-2名硕士研究生

5.3 学术交流

A. 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。

B. 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。

（二） 研究基础与工作条件

1. 研究基础（与本项目相关的研究工作积累和已取得的研究工作成绩）；

申请人具有多年原子水平的计算模拟研究经验以及科学计算算法与程序开发经验。自 2009 年来，申请人在这些相关领域共发表国际 SCI 期刊 6 篇（其中通讯作者及第一作者文章 4 篇，单篇最高影响因子 IF=5.152）。这些研究成果中有 3 篇是晶体结构与对称性方向的研究，3 篇是分子动力学模拟以及第一性原理计算方向的研究。除此之外，申请人还自主开发一套分子力场以及发展了一套晶体结构相关的算法并开发了对应的的科学计算程序 EPCRYST（http://www.epcryst.com）。这些研究成果都表明申请人在数学物理方面尤其是对称性相关的研究方面具有非常扎实的基础，在科学计算算法设计与程序编写方面也有丰富的积累和经验。申请人的研究经历和研究基础都保证了该项目能顺利开展与完成。

该项目已经在 BS-Seq 测序数据分析方向上做出了大量的基础工作，建立了两条完整的甲基化状态分析 Pipeline（甲基化信息分析：https://github.com/dlmeduLi/mtbr-pipline 和等位基因甲基化差异区域分析：https://github.com/dlmeduLi/asm-pipeline）以及相关的测序数据分析辅助工具（https://github.com/dlmeduLi）。

使用这些工具，申请人已经做了大量关于 DNA 中 CpG 位点分布状态及其甲基化状态的数据分析工作。我们在数据分析的结果中观察到了一些非常明显的规律性现象和趋势，该项目的根本出发点就是为了从分子组装层面更好、更深入地对这些现象和规律性趋势进行阐述，并据此对我们后续的分析提供更好的理论基础和指导方向。在该项目拟采用的分子动力学方面，申请人已经完成了基于 NAB 的自动化 DNA 结构模型建立及甲基化修饰的程序开发工作。

下面我们将介绍一下该项目已经完成的一些工作和发现：

1） 进化中DNA中CpG位点的含量与分布向不均匀性方向变化的趋势。

DNA 中的 CpG 位点分布呈现与其它二核苷酸序列不同的趋势。首先，CpG 在 DNA 中的含量明显偏低，如图 7A 所示。按照随机理论，DNA 中不同二核苷酸序列的含量应该呈现出均匀分布的趋势，但是在实际生物 DNA 序列中 CpG 的含量明显偏低。而且随着生物从低等到高等的进化，DNA 含量的这种变化趋势越来越显著。对于有着固定字母集合的语言，其字母的使用频率有着非常明显的不对称现象，这种不对称现象代表着有序性和信息量。同样，DNA 中的这种 CpG 碱基序列含量对称性破缺现象及趋势也代表着 DNA 序列信息量的变化。不对称性的程度越高，信息商越大，信息含量也越大。其次，CpG 在 DNA 序列中呈明显的团簇化效应，如图 7B 所示。DNA 序列中大部分的二核苷酸的间距分布都呈现典型的正态分布状态，但是 CpG 的间距却呈现出非常明显的团簇化现象（例如趋向于形成 CpG 岛等等），其间距的分布也表现为非对称性分布状态，而且随着生物的进化，其分布非对称性程度越来越明显。

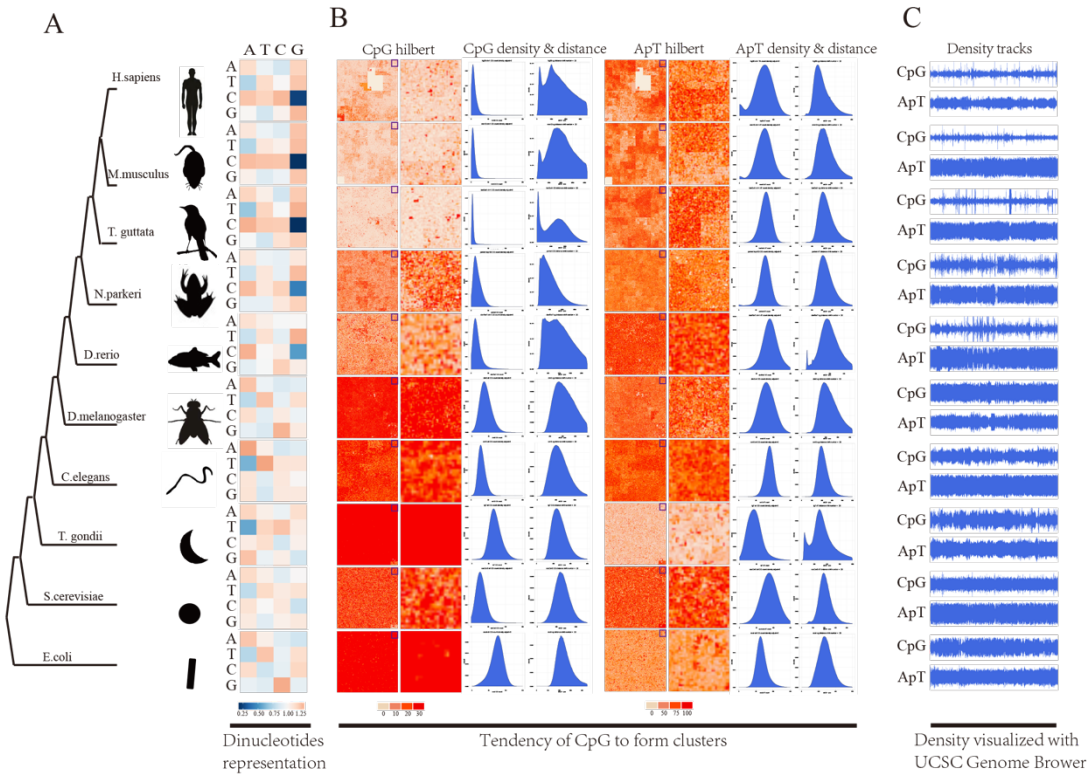

图 7 CpG 含量及分布的对称性破缺现象。A) DNA 中二核苷酸含量的热图，可以看到 CpG 的含量明显比其他二核苷酸低，而且随着物种的进化程度这种含量变低的趋势越来越明显；B) CpG 与 ApT 含量分布 Hilbert 图以及密度和间距分布图。与 ApT 相比，CpG 的 Hilbert 图呈现出明显的团簇化现象。同时 CpG 的分布也比 ApT 表现出更明显的非对称性。随着物种的进化程度 CpG 分布的非对称性越来越明显；C) UCSC Genome Broswer 上 CpG 与 ApT 密度对比图，可以看到 ApT 呈现出更加均匀的分布状态，而 CpG 的密度偏低而且更加不均匀，这种不均匀性随着物种的进化程度而逐渐显著。

2） DNA中CpG的密度与其甲基化状态呈现出明显的负相关。

从信息论的角度来定义，信息对称性是指随机变量X中所包含的关于随机变量Y的互信息量$I(X{:}Y)$与随机变量Y中所包含的随机变量X的互信息量有相等或者相近似的关系（实际中非完美对称性关系），即$I(X{:}Y) = I(Y{:}X)$或者$I(x{:}y) \approx I(y{:}x)$，其中$I(X{:}Y) = H(X) - H(Y|X)I(X{:}Y) = H(X) - H(Y|X)$，$H$为香农熵。总体上看来 DNA 中 CpG 位点的密度与该点的甲基化水平有着非常明显的负相关性，如图 8 所示，这表明 CpG 位点的密度与其甲基化状态是具有内在的信息对称性。

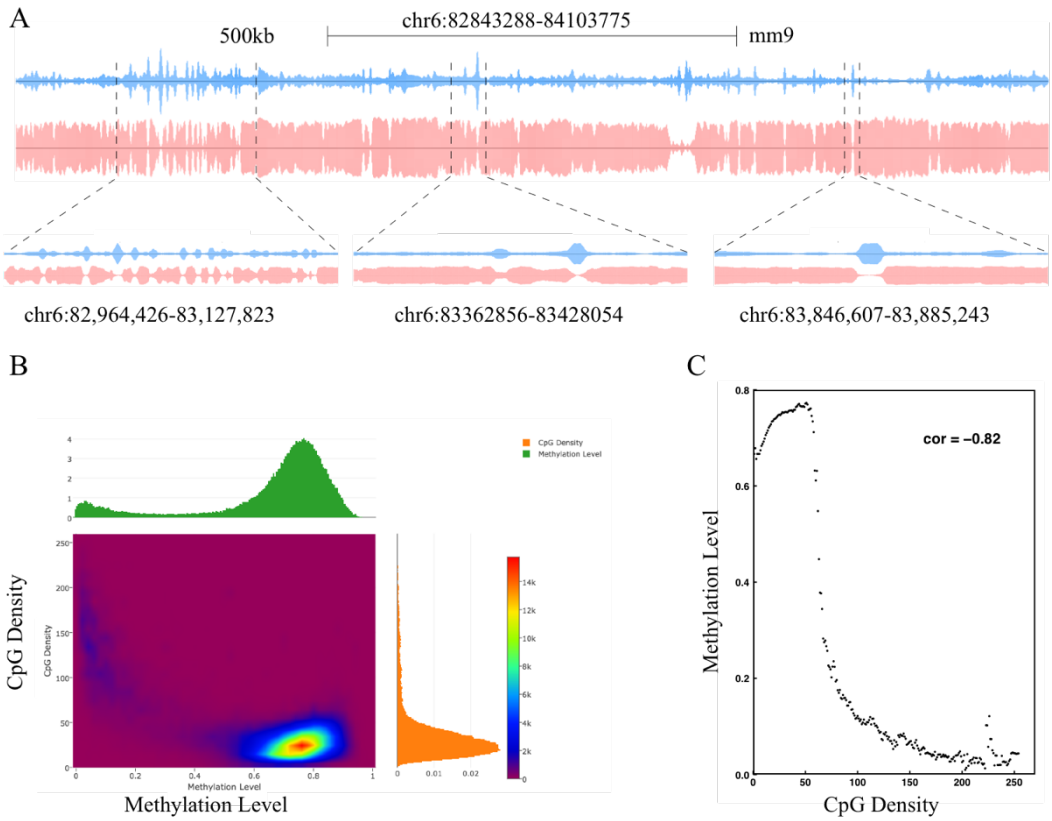

图 8 CpG 密度与其甲基化水平的负相关性及信息对称性（互补特性）。A) UCSC Genome Browswer 截图显示 CpG 密度与其甲基化水平具有非常明显的信息对称性（互补特性），CpG 密度高的区域甲基化水平偏低而，CpG 密度低的区域甲基化水平偏高；B)小鼠（19 号染色体）DNA 中 CpG 密度与其甲基化水平的总体分布图。 我们可以很明显地看出，CpG 密度与其甲基化有显著的负相关性，而且大部分数据都集中分布在甲基化水平偏高而 CpG 密度偏低的区域。C) CpG 密度与其甲基化水平的负相关性曲线，二者负相关性高达-0.82。

然而，这种对称性关系也在一些重要的功能区域存在破缺现象，如图 9 所示。

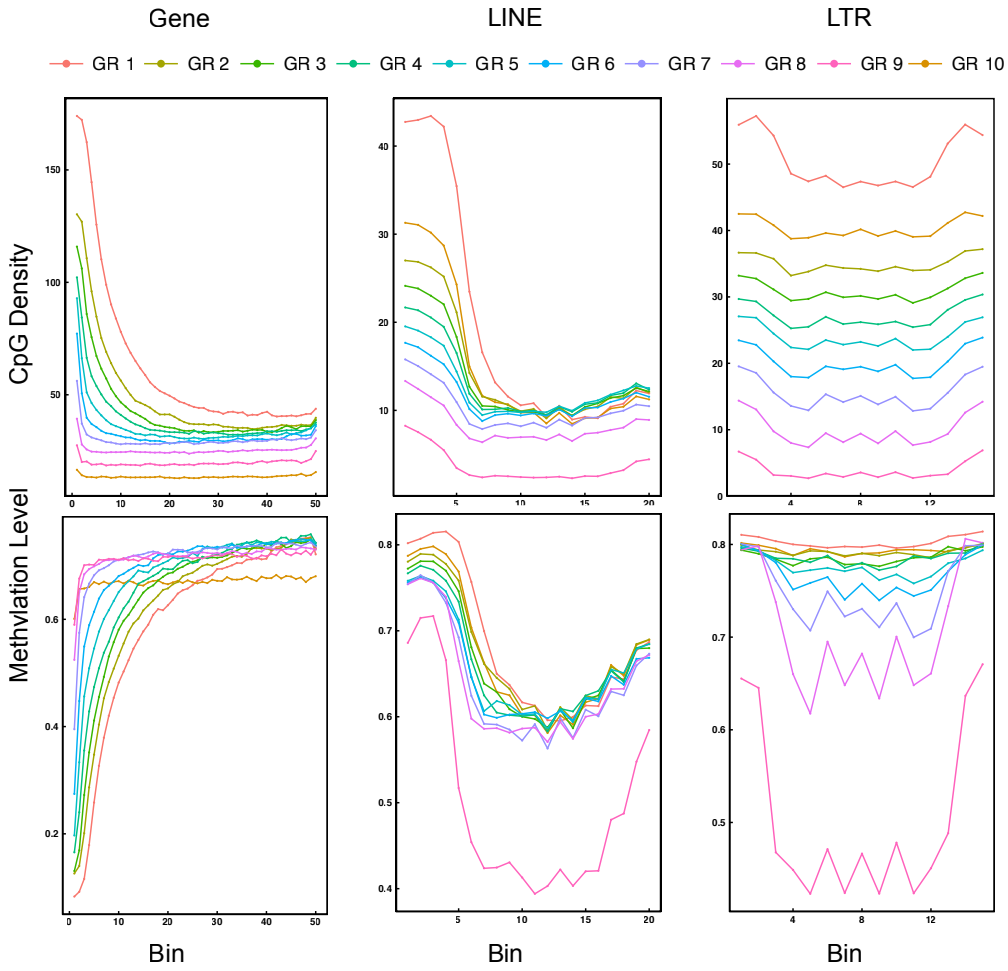

图 9 DNA 中特征区域 CpG 密度与甲基化水平的对称性（互补性）以及非对称性现象。DNA 中 CpG 密度与其甲基化水平有着明显的对称性（互补性），在一些特征区域，这些互补性或者得到增强例如基因（Gene）区域，或者发生对称性破缺现象例如重复序列区域 LINE（Long interspersed nuclear elements）和 LTR（Long terminal repeat elements），这表明 CpG 密度与其甲基化水平之间的对称性或者对称性破缺与 DNA 功能区域密切相关。

我们将 CpG 与其甲基化之间的对应关系分为 Overlap 区域（OR）、Gap 区域（GR）以及互补性区域或对称性区域（CR），其中CR占据了整个体系绝大部分，如图 10 所示。我们在研究中发现，这些不对称性关系可能与组织基因的差异性表达有非常密切的关系。

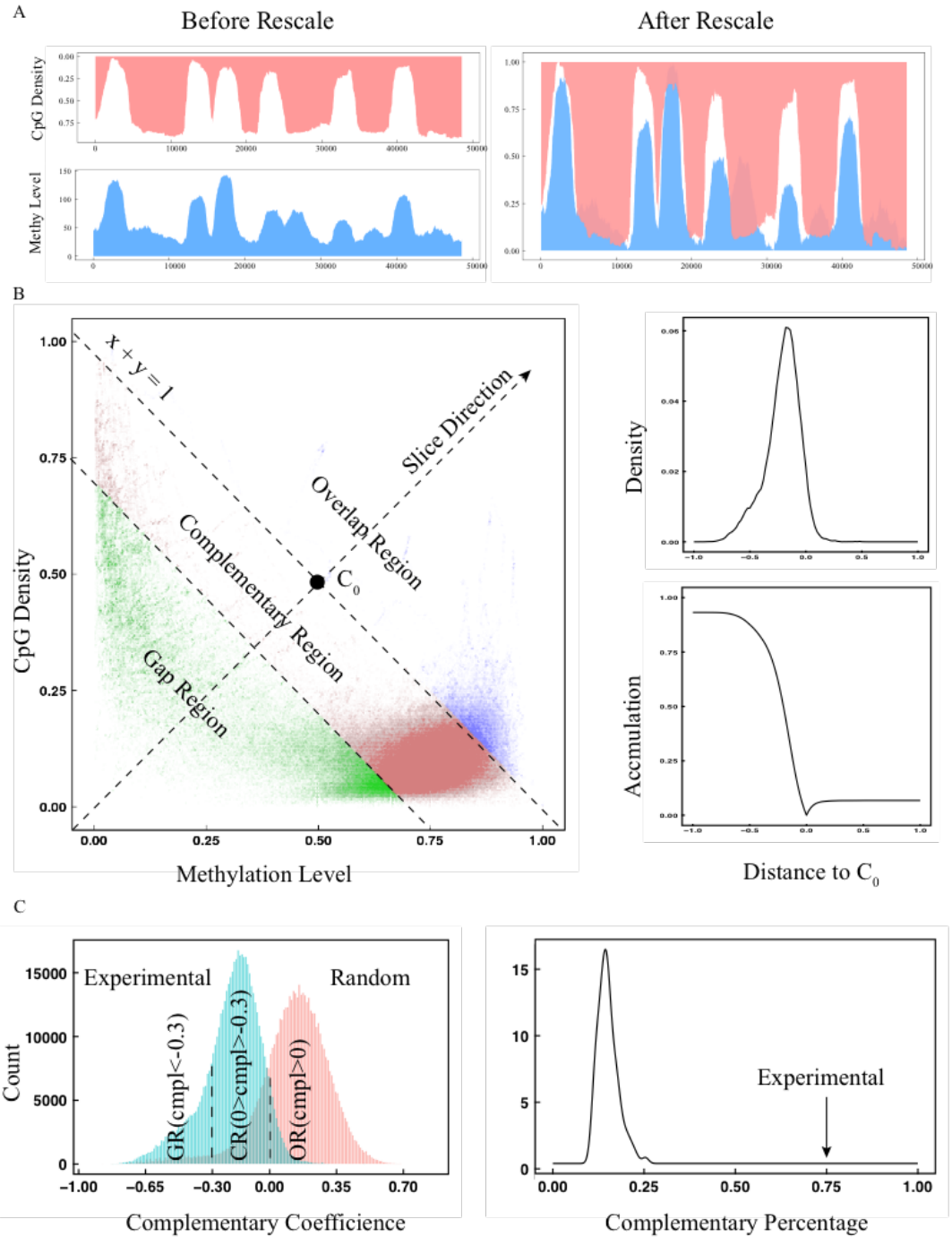

图 10 CpG 密度与其甲基化水平对称性数学描述。A) 根据项目提出的算法，对 CpG 及其甲基化水平的信息对称性进行数学化描述。变化前二者分别使用不同的量度，变换后二者使用同一量度；B) 对非完美对称性进行容差性分析，根据非对称性程度将其分为三大类型 Gap（GR）、互补（对称性 CR）以及 Overlap（OR），右图为其非对称性程度的分布曲线；C) 使用随机数据（项目提出的根据置换操作所生成的模拟 DNA 序列数据）来规范三大类型区域的定义范围，右图为 100 组随机数据完全对称性比例的组织分布图。这表明实际测序数据（75%完全对称性比例）具有很高的非随机性可信度。

3） 特征DNA序列分子动力学建模

申请人已经完成了分子动力学计算模型的建立。项目采用 AmberTools 的 NAB 程序建立 DNA 分子结构模型。NAB（Nucleic acid builder）提供了一种建立核酸模型的高级语言，它具有与 C 语言类似的语法，并基于大的分子和分子片段来建立核酸模型。使用 NAB 提供的语言可以通过编程完成几乎所有非常规的核酸结构模型的建立。目前，申请人已经完成了特征 DNA 序列及其不同甲基化状态的自动结构模型生成程序，图 11 给出了一个特征 DNA 序列的普通结构模型和甲基化之后的模型。

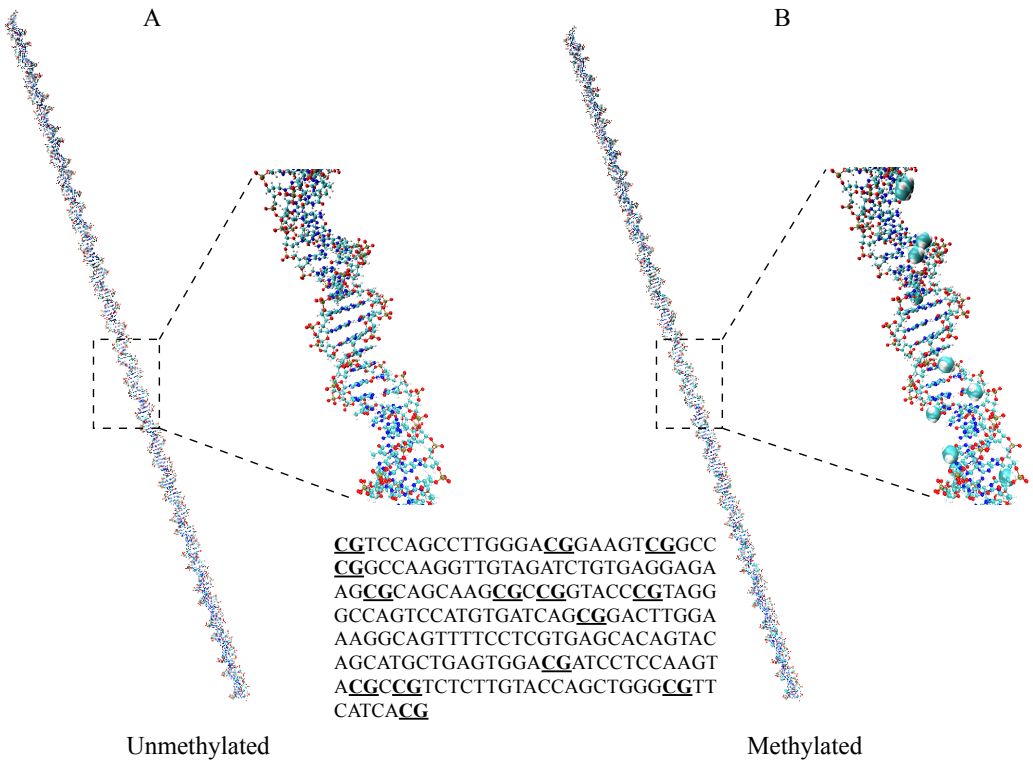

图 11 项目计算模拟拟采用的 DNA 结构模型图。该项目已经开发完成了根据特征序列进行结构模型建立的程序，图为特征序列为 d(CGTCCAGCCTTGGGACGGAAGTCGGCCCGGCCAAGGTTGTAGATCTGTGAGGAGAAGCGCAGCAAGCGCCGGTACCCGTAGGGCCAGTCCATGTGATCAGCGGACTTGGAAAGGCAGTTTTCCTCGTGAGCACAGTACAGCATGCTGAGTGGACGATCCTCCAAGTACGCCGTCTCTTGTACCAGCTGGGCGTTCATCACG)的结构模型图。A) CpG 未甲基化的结构模型；B) CpG 甲基化之后的结构模型。

2. **工作条件**（包括已具备的实验条件，尚缺少的实验条件和拟解决的途径，包括利用国家实验室、国家重点实验室和部门重点实验室等研究基地的计划与落实情况）；

申请人所在的实验室隶属于大连医科大学肿瘤干细胞研究院下的基因组中心。针对国家对于精准医疗的战略发展需求，基因组中心建设了计算集群中心。目前基因组中心拥有超过500核的计算能力，拥有300Tb以上的数据存储能力，完全可以满足对二代测序数据的大数据挖掘与分析需求。在研究院发展计划中，超过千核的大型计算集群也正在筹备当中。基因组中心同时还配备了各种建库及测序仪器以及相应的工作人员，能按照项目需求自行设计实验获取测序数据。申请人所在实验室以及基因组中心有多名生物信息学以及计算机相关的科研人员。这些工作人员有丰富的测序数据分析以及程序编写经验。这些人员及技术储备，是本项目顺利实施的重要保证。

3. **正在承担的与本项目相关的科研项目情况**（申请人和项目组主要参与者正在承担的与本项目相关的科研项目情况，包括国家自然科学基金的项目和国家其他科技计划项目，要注明项目的名称和编号、经费来源、起止年月、与本项目的关系及负责的内容等）；

无

4. **完成国家自然科学基金项目情况**（对申请人负责的前一个已结题科学基金项目（项目名称及批准号）完成情况、后续研究进展及与本申请项目的关系加以详细说明。另附该已结题项目研究工作总结摘要（限500字）和相关成果的详细目录）。

无

（三） 其他需要说明的问题

1. 申请人同年申请不同类型的国家自然科学基金项目情况（列明同年申请的其他项目的项目类型、项目名称信息，并说明与本项目之间的区别与联系）。

无

2. 具有高级专业技术职务（职称）的申请人或者主要参与者是否存在同年申请或者参与申请国家自然科学基金项目的单位不一致的情况；如存在上述情况，列明所涉及人员的姓名，申请或参与申

请的其他项目的项目类型、项目名称、单位名称、上述人员在该项目中是申请人还是参与者，并说明单位不一致原因。

无

3. 具有高级专业技术职务（职称）的申请人或者主要参与者是否存在与正在承担的国家自然科学基金项目的单位不一致的情况；如存在上述情况，列明所涉及人员的姓名，正在承担项目的批准号、项目类型、项目名称、单位名称、起止年月，并说明单位不一致原因。

无

4. 其他。

无