研究内容

本项目拟通过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研究不同CpG分布以及对应的甲基化状态时DNA体系能量的变化。我们希望通过建立与分析系统的能量图谱来解释CpG及其甲基化状态的对称性以及对称性破缺现象发生的机理，同时我们也将通过对基因测序数据的分析从测序实验的角度来验证我们的理论。下面，我们将逐步对拟展开的研究内容逐步进行阐述。

1) 运用生物信息学方法获取不同物种中CpG分布及其甲基化分布状态，建立完善的数学模型

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) 在试验数据的基础之上，运用群论、统计等方法产生特征DNA序列

由于计算资源的限制，我们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精确的结果，但是考虑到实际的计算资源，我们只能计算模拟有限长度的DNA序列。这就需要我们根据相应的对称性模型获取用于模拟的非常特征化的DNA序列 。如何在实验数据的结果之上建立模型，产生有限长度的、能反映我们计算模拟结构特性的DNA序列将是我们研究内容的一个非常重要的方面。

3) CpG甲基化反应机理及DNA三级结构模型的量子化学计算修正

DNA三维空间结构将对计算模拟的结果有重要的影响。CpG位点的甲基化结构是该研究项目中影响DNA空间结构的一个重要方面。已经有相关报道表明CpG甲基化会影响DNA的一些物理特性，如亲和性等等(Pérez et al. 2012; Derreumaux et al. 2001)。因此，为了获取更精确的计算结果，我们将结合该项目的研究目标，从量子化学计算的层面上对DNA甲基化的反应机理进行细致的研究，获取最合理的DNA空间结构模型。

4) 分子动力学模拟获取不同的CpG分布及其甲基化状态对应的能量变化图谱，并据此提出相应的理论解释

该项目将使用分子动力学方法获取不同特征序列的能量变化图谱。我们希望能从能量图谱的变化上来解释对应的特征DNA序列所反应的CpG分布及其甲基化状态的对称性与对称性破缺现象。

5) 从生物信息学的角度重新细化数据分析内容来验证我们根据分子动力学能量变化图谱得到的理论解释

从能量图谱得到的理论模型需要经过实验数据的进一步验证。与之前的数据分析不同，在有了理论模型支持的基础之上，我们希望对数据的分析更加细化，更加具有针对性。这一步除了验证之前的理论模型之外，我们也希望能通过这一整套的方法从大量的测序数据中分析并挖掘出更加有价值的结果。

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的方法从能量图谱变化的角度上来研究和解释DNA中CpG分布及其甲基化对称性及对称性破缺的现象。并希望通过这一手段能对目前生物信息学研究中大量测序数据的挖掘在表观遗传方面提供指导性的方向。总结为以下几个具体化的目标：

1) 为CpG分布及其与甲基化的关系建立详细的数据描述模型。包括分布统计模型以及对其对称性和对称性破缺关系的数学描述模型。

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描述方法。引入确定性的描述方法，例如通过群论等方法建立描述DNA对称性及对称性破缺的模型，将是我们研究的重要目标之一。

2) 通过量子化学计算研究甲基化反应机理及探索不同甲基化基团三维空间构型对DNA特性的影响，并据此确定合理的甲基化的DNA三维空间结构模型。

该项目需要使用量子化学计算的方法从分子水平上确定不同序列中不同CpG位点最合理的甲基化模型，以此来保证获取得更加精确的计算结果。

3) 使用分子动力学模拟方法获取不同CpG分布及其甲基化状态时的能量图谱，并据此来分析和解释其对应的对称性及对称性破缺发生的机理包括发生条件、进化趋势和进化驱动力等。

从分子水平上了解DNA中CpG分布及其甲基化状态对称性及对称性破缺的机理是本项目研究的最重要的目标。同时我们也希望通过对这些机理的深入理解，我们对生物信息数据挖掘有更明确的方向，获取更多新的、有价值的结果。

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） 如何应用精准的数学语言来描述DNA中CpG的分布及其甲基化状态的对称性及对称性破缺关系。

2）对称性及对称性破缺是DNA中CpG位点特殊性变化的内在驱动力吗？如果是，我们能否从能量的角度，在分子水平上来了解对称性破缺是如何发生以及其发展趋势如何？DNA中的这种对称性破缺又会对基因表达等层面的机制产生怎样的影响？

3）在了解对称性破缺的机理情况下，我们能否从头开始地对DNA的序列的变化做出模拟、分析以及预测？

我们将使用分子动力学以及量子化学计算，并结合生物信息学分析手段逐一回答以上几个科学问题。这些科学问题的回答将对我们生物信息的挖掘和理解有重要的理论意义和实践指导意义。