# An overview of the Chinese version [10]

## Sequence motif

序列模体（Sequence motif）是指 DNA、RNA 或蛋白质序列中的一小段具有特定的序列模式。这些模式通常是重复出现的单元或者具有一定生物学功能的结构域。在生物学研究中，寻找并分析序列模体可以帮助我们理解生物分子的结构和功能。通过识别和分析序列模体，我们可以推断蛋白质的结构和功能、DNA 的修饰和转录调控以及 RNA 的折叠和相互作用等生物学过程。

在计算生物学中，序列模体通常通过模式匹配算法来识别。常见的模式匹配算法包括正则表达式、PWM（Position Weight Matrix）和 HMM（Hidden Markov Model）等。正则表达式是一种简单而强大的模式匹配工具，可以识别具有固定序列的模体。PWM 和 HMM 则是更为复杂的算法，可以识别具有变异的序列模体，这些变异可能是由于基因突变或种属演化等原因导致的。

序列模体在生物学研究中具有广泛的应用。例如，DNA 中的转录因子结合位点（TFBS）就是一种常见的序列模体，可以用来预测基因的转录调控。另外，许多疾病和病原体也可以通过识别其特定的序列模体来进行诊断和治疗。总之，序列模体的研究为我们深入理解生物分子的结构和功能提供了重要的工具和方法。

## MEME

MEME 算法（Multiple EM for Motif Elicitation，多重 EM 算法）是一种在生物信息学领域中用于识别 DNA 或蛋白质序列中“motif”（基序）的算法。该算法是基于 EM（Expectation-Maximization）算法的一种变体。

在 MEME 算法中，基序被定义为一种短序列模式，该模式在序列中出现的频率比随机期望高得多。该算法旨在从一组相关序列中识别这些基序，并确定它们在序列中的位置和出现频率。

MEME 算法主要分为两个步骤：第一步是使用 EM 算法识别候选基序；第二步是使用一些统计方法对候选基序进行筛选和排序。该算法的优点是能够在不需要先验信息的情况下识别基序，并且能够发现具有变异或杂合性的基序。

使用 EM 算法发现基序的步骤如下：

1. 收集序列数据集：首先需要准备一个序列数据集，该数据集由多个序列组成，每个序列都包含一个要发现基序的区域。
2. 设定基序的长度：在使用 EM 算法之前，需要确定基序的长度。基序的长度通常由研究人员根据数据集的特征和先验知识来确定。
3. 初始化基序矩阵：将基序矩阵的初始值设为随机值。基序矩阵是一个大小为$L \times 4$的矩阵，其中$L$是基序的长度，4 代表碱基（A、C、G、T）的数量。
4. E 步：对于每个序列，计算基序的每个位置出现的概率，并根据这些概率来计算每个序列中基序的期望出现次数。
5. M 步：根据每个序列中基序的期望出现次数来更新基序矩阵。
6. 重复步骤 4 和步骤 5，直到基序矩阵收敛为止。
7. 选取最优基序：使用一些统计方法对基序进行筛选和排序，选取具有较高出现频率和保守性的基序。

EM 算法能够通过反复迭代来优化基序矩阵，从而提高基序的准确性和可靠性。基序的准确性和可靠性对于进一步的生物信息学研究具有重要意义。

## Motif based prediction & Supervised machine learning modeling

Motif based prediction 和 Supervised machine learning modeling 是两种常用的从 DNA 序列预测表观遗传标记的方法。

Motif based prediction 是指基于 DNA 序列中出现频率较高的基序(motif)来预测表观遗传标记的位置。基序是指在 DNA 序列中出现频率较高的一段 DNA 序列，通常是长度为 6-20 个碱基对的 DNA 序列。基于这种方法，可以通过比对 DNA 序列与已知基序数据库，来预测表观遗传标记的位置。其中比对算法包括 MEME、FIMO、HOMER 等。

Supervised machine learning modeling 是指利用已知表观遗传标记的数据来训练机器学习模型，从而预测未知序列中的表观遗传标记位置。这种方法需要先准备一个标记好的训练集，然后通过对这个训练集进行特征提取、特征选择和分类器训练等步骤，得到一个能够预测表观遗传标记位置的模型。其中特征提取和选择的算法包括 k-mer、PWM 等，常用的分类器包括支持向量机、随机森林、神经网络等。

这两种方法各有优劣，Motif based prediction 速度较快，但需要已知基序数据库，适用于已知的表观遗传标记；而 Supervised machine learning modeling 适用于未知的表观遗传标记，但需要准备标记好的训练集。因此，在具体应用中，可以根据实际情况选择合适的方法。

## CpG island

CpG 岛是指在 DNA 序列中 GC 含量高、且连续分布的区域。其中 CpG 是指在 DNA 序列中的 Cytosine 和 Guanine 碱基之间存在一个磷酸二酯键连接，是 DNA 甲基化的重要部位。

CpG 岛通常指的是长度大于 200bp、CpG 含量大于 50%、CpG 比例/期望值大于 0.6 的 DNA 区域。它们通常出现在启动子区域、基因组重要区域、调控元件等地方。CpG 岛在基因组的结构和功能中发挥着重要的作用，比如在基因表达、转录调控、染色质重塑等方面都扮演着重要角色。

CpG 岛的检测可以通过多种方法实现，包括计算 GC 含量、期望 CpG 比例和观察 CpG 岛长度等方法。常用的算法包括 CpG 岛预测软件 EMBOSS CpGPlot、CpG 岛检测软件 CpGIE、CpG 岛分析工具 MethylCoder 和 CpG 岛识别器 CpGcluster 等。

总之，CpG 岛是基因组中的一个重要特征，对于理解基因组的结构和功能以及人类疾病的研究具有重要意义。

## Hidden Markov model

隐藏马尔可夫模型（Hidden Markov Model，HMM）是一种常用于生物信息学中 CpG 岛识别的方法。HMM 是一种基于概率的统计模型，可用于描述一个系统的状态序列，以及各个状态与可观测变量之间的概率分布。在 CpG 岛识别中，HMM 被用来对 DNA 序列进行建模，以便找到 CpG 岛的位置和边界。

HMM 通常由三个部分组成：状态集合、转移概率矩阵和发射概率矩阵。在 CpG 岛的识别中，状态集合通常包括三个状态：CpG 岛内部状态、CpG 岛外部状态和起始/终止状态。转移概率矩阵描述了各个状态之间的转移概率，即 CpG 岛的起始和结束位置，以及岛内外状态之间的转移概率。发射概率矩阵描述了各个状态下不同碱基对的出现概率，用于计算当前状态下观测序列的概率。

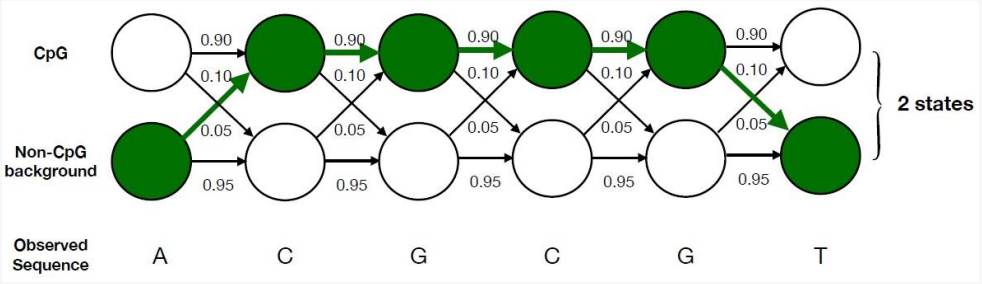
通过对训练数据进行模型训练，即给定已知的 CpG 岛和非 CpG 岛序列，可以得到最优的模型参数。在对新的 DNA 序列进行 CpG 岛识别时，HMM 会根据序列的发射概率和转移概率计算每个位置上的状态概率，然后通过 Viterbi 或者 Forward backward 算法等方法找到最可能的 CpG 岛状态序列。

HMM 是一种基于统计学习的方法，已经在 CpG 岛识别、蛋白质结构预测、信号序列识别等方面取得了广泛的应用。

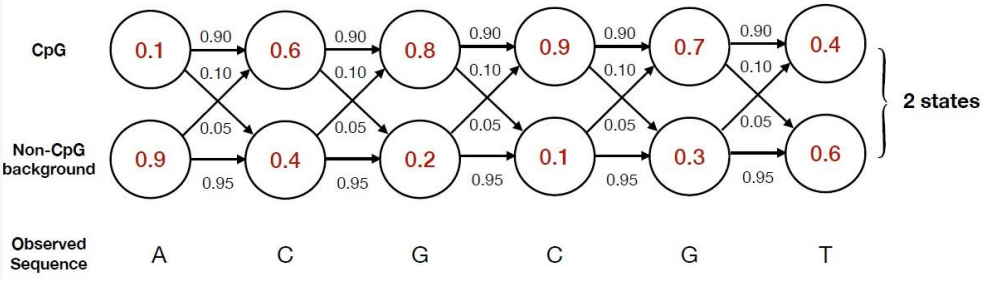
## Viterbi algorithm & Forward backward algorithm

Viterbi 算法和 Forward-Backward 算法是两种用于隐马尔可夫模型（Hidden Markov Model，HMM）中状态序列推断的常用算法。

Viterbi 算法用于找到最可能的状态序列，通常用于分类、标注和预测等任务。算法的核心是动态规划，即在给定观测序列的情况下，从左到右递推计算每个时刻各个状态的最大概率值和对应的最优路径，最终得到最可能的状态序列。具体实现中，通过定义递推公式和 Viterbi 表格，可实现高效的计算。




Forward-Backward 算法用于计算给定观测序列的条件概率，即 P(O|λ)，其中 O 为观测序列，λ 为 HMM 的参数。算法的核心是在给定观测序列的情况下，利用 Forward 算法计算从开始到当前位置的所有状态序列出现的概率，并利用 Backward 算法计算从当前位置到结束的所有状态序列出现的概率。然后，通过将两个概率相乘并除以总概率即可得到条件概率。




Viterbi 算法和 Forward-Backward 算法是 HMM 中常用的两种算法，分别用于状态序列推断和条件概率计算。两个算法在 HMM 的应用中互相补充，可共同用于 CpG 岛识别、蛋白质结构预测、信号序列识别等生物信息学领域的任务中。

## AUROC

AUROC 是指 Area Under the Receiver Operating Characteristic Curve 的缩写，即 ROC 曲线下面积。

ROC 曲线是衡量二元分类器性能的一种常用方法，通过将分类器在不同阈值下的真阳性率（True Positive Rate）和假阳性率（False Positive Rate）作为坐标轴，绘制出一条曲线。而 AUROC 则是用来衡量 ROC 曲线的性能，其取值范围在 0.5 至 1 之间，值越大表示分类器性能越好。

AUROC 的计算方法通常是先计算出不同阈值下的真阳性率和假阳性率，然后将这些点连成一条曲线，并计算曲线下方的面积。如果分类器的性能完全随机，其 AUROC 值应该接近 0.5，而完美的分类器的 AUROC 值为 1.0。

AUROC 在医学、生物信息学、机器学习等领域广泛应用，例如在肿瘤诊断、基因预测、药物筛选等方面的评估中都有使用。

## Workflow of sequence based supervised learning (ANN&HMM)

基于序列的监督学习（Sequence-based Supervised Learning）的工作流程通常包括以下步骤：

1. 数据预处理（Data Preprocessing）：将原始的序列数据进行预处理，如去除噪声、标准化、特征提取等。
2. 特征选择（Feature Selection）：从预处理后的序列中提取出有代表性的特征，用于建立模型。常用的特征包括序列长度、碱基组成、氨基酸序列等。
3. 建立模型（Model Building）：选择合适的监督学习算法，如人工神经网络（Artificial Neural Network，ANN）和隐马尔可夫模型（Hidden Markov Model，HMM），并使用特征选择得到的特征训练模型。
4. 模型评估（Model Evaluation）：使用测试集数据评估模型的性能，常用的评价指标包括准确率、精确度、召回率、F1 值等。
5. 模型优化（Model Optimization）：根据评估结果，对模型进行优化，如调整模型参数、增加训练数据量等。
6. 模型应用（Model Application）：使用优化后的模型对新的序列数据进行分类或预测，如基因分类、药物筛选等应用。

人工神经网络和隐马尔可夫模型是常用的序列分类和预测算法，其中人工神经网络通过学习输入序列和对应的标签之间的关系来进行分类或预测，而隐马尔可夫模型则通过学习隐含状态和可观测序列之间的关系来进行序列分类或预测。在建立模型时，通常需要考虑选择合适的网络结构或模型参数，并使用交叉验证等方法来评估模型性能。

## ANN&HMM details

人工神经网络（Artificial Neural Network, ANN）和隐马尔可夫模型（Hidden Markov Model, HMM）都是机器学习中的重要模型。

人工神经网络是受到生物神经元启发的计算模型，通过输入层、隐藏层和输出层等多个层次的神经元节点之间的连接来模拟神经元的传递过程，从而实现对输入数据的分类、预测等任务。通俗地讲，可以把它看作是由多个小决策器组成的大决策器，通过对每个小决策器的输出进行综合，得出最终的预测结果。

隐马尔可夫模型则是一种基于概率的统计模型，它由多个状态以及状态之间的转移概率和观测概率构成。在 HMM 中，状态是不可见的，而只能根据观测到的数据推断出所处的状态。通俗地讲，可以将其看作是一个“黑盒子”，输入一些观测数据，然后得到一系列潜在的状态。

举个例子来说，假设我们想要预测一个人是否会购买某个产品，我们可以使用人工神经网络来训练一个分类器，输入这个人的性别、年龄、收入等信息，输出预测结果。而在这个预测过程中，我们可能需要使用隐马尔可夫模型来对用户的行为进行建模，以更好地理解他们的购买行为。