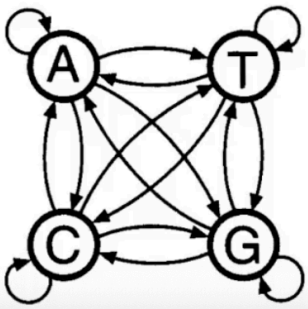
隐马尔可夫模型

马尔可夫模型（马尔可夫链）

*** 马尔可夫模型：描述了一类重要的随机过程（随时间而随机变化的过程），某一时间的状态只由前一个时间的状态决定，状态是已知的，状态之间变化的概率称为 转化概率 。

*** 生物中应用：CpG岛[1]预测问题：一个位置的碱基状态由前一个位置的碱基状态决定。



$$a_{st} = P(x_i = t | x_{i-1} = s).$$

$$\begin{aligned} P(x) &= P(x_L, x_{L-1}, \ldots, x_1) \\ &= P(x_L | x_{L-1}, \ldots, x_1) P(x_{L-1} | x_{L-2}, \ldots, x_1) \cdots P(x_1) \\ P(x) &= P(x_L | x_{L-1}) P(x_{L-1} | x_{L-2}) \cdots P(x_2 | x_1) P(x_1) \\ &= P(x_1) \prod_{i=2}^{L} a_{x_{i-1} x_i}. \end{aligned}$$

定义

马尔可夫链是一组具有马尔可夫性质的离散随机变量的集合。具体地，对概率空间 $(\Omega, \mathcal{F}, \mathbb{P})$ 内以一维可数集为指数集（index set）的随机变量集合 $\boldsymbol{X} = \{X_n : n > 0\}$，若随机变量的取值都在可数集内：$X = s_i$，$s_i \in \boldsymbol{s}$，且随机变量的条件概率满足如下关系[2]：

$$p(X_{t+1} | X_t, \ldots, X_1) = p(X_{t+1} | X_t)$$

则 $\boldsymbol{X}$ 被称为马尔可夫链，可数集 $\boldsymbol{s} \in \mathbb{Z}$ 被称为状态空间（state space），马尔可夫链在状态空间内的取值称为状态[2]。这里定义的马尔可夫链是离散时间马尔可夫链（Discrete-Time MC, DTMC），其具有连续指数集的情形虽然被称为连续时间马尔可夫链（Continuous-Time MC, CTMC），但在本质上是马尔可夫过程（Markov process）[19]。常见地，马尔可夫链的指数集被称为“步”或“时间步（time-step）”[2]。

上式在定义马尔可夫链的同时定义了马尔可夫性质，该性质也被称为“无记忆性（memorylessness）”，即t+1步的随机变量在给定第t步随机变量后与其余的随机变量条件独立（conditionally independent）：$X_{t+1} \perp (X_{t-1}, X_0) | X_t$[2]。在此基础上，马尔可夫链具有强马尔可夫性（strong Markov property），即对任意的停时（stopping time），马尔可夫链在停时前后的状态相互独立[1]。

解释性例子

马尔科夫链的一个常见例子是简化的股票涨跌模型：若一天中某股票上涨，则明天该股票有概率p开始下跌，1-p继续上涨；若一天中该股票下跌，则明天该股票有概率q开始上涨，1-q继续下跌。该股票的涨跌情况是一个马尔可夫链，且定义中各个概念在例子中有如下对应：

- 随机变量：第t天该股票的状态；状态空间：“上涨”和“下跌”；指数集：天数。
- 条件概率关系：按定义，即便已知该股票的所有历史状态，其在某天的涨跌也仅与前一天的状态有关。
- 无记忆性：该股票当天的表现仅与前一天有关，与其他历史状态无关（定义条件概率关系的同时定义了无记忆

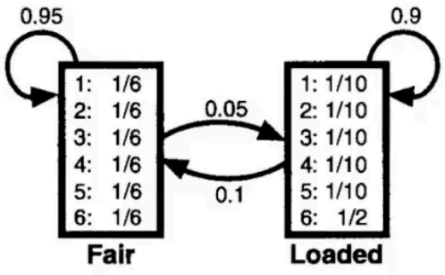
性）。

- 停时前后状态相互独立：取出该股票的涨跌记录，然后从中截取一段，我们无法知道截取的是哪一段，因为截取点，即停时t前后的记录（t-1和t+1）没有依赖关系。

隐马尔可夫模型（Hidden Markov Model, HMM）

***隐马尔可夫模型：描述了一类重要的随机过程（随时间而随机变化的过程），其一时间的状态只由前一个时间的状态决定， 状态是未知的 ,状态之间发生的概率称为 转化概率 ，在某个未知的状态下观测到某个信息的概率称为发射概率。

***赌场骰子预测问题（解码过程）：赌场会交替使用正常和作弊骰子，每次骰子的状态未知，但每次骰子呈现的数字已知。



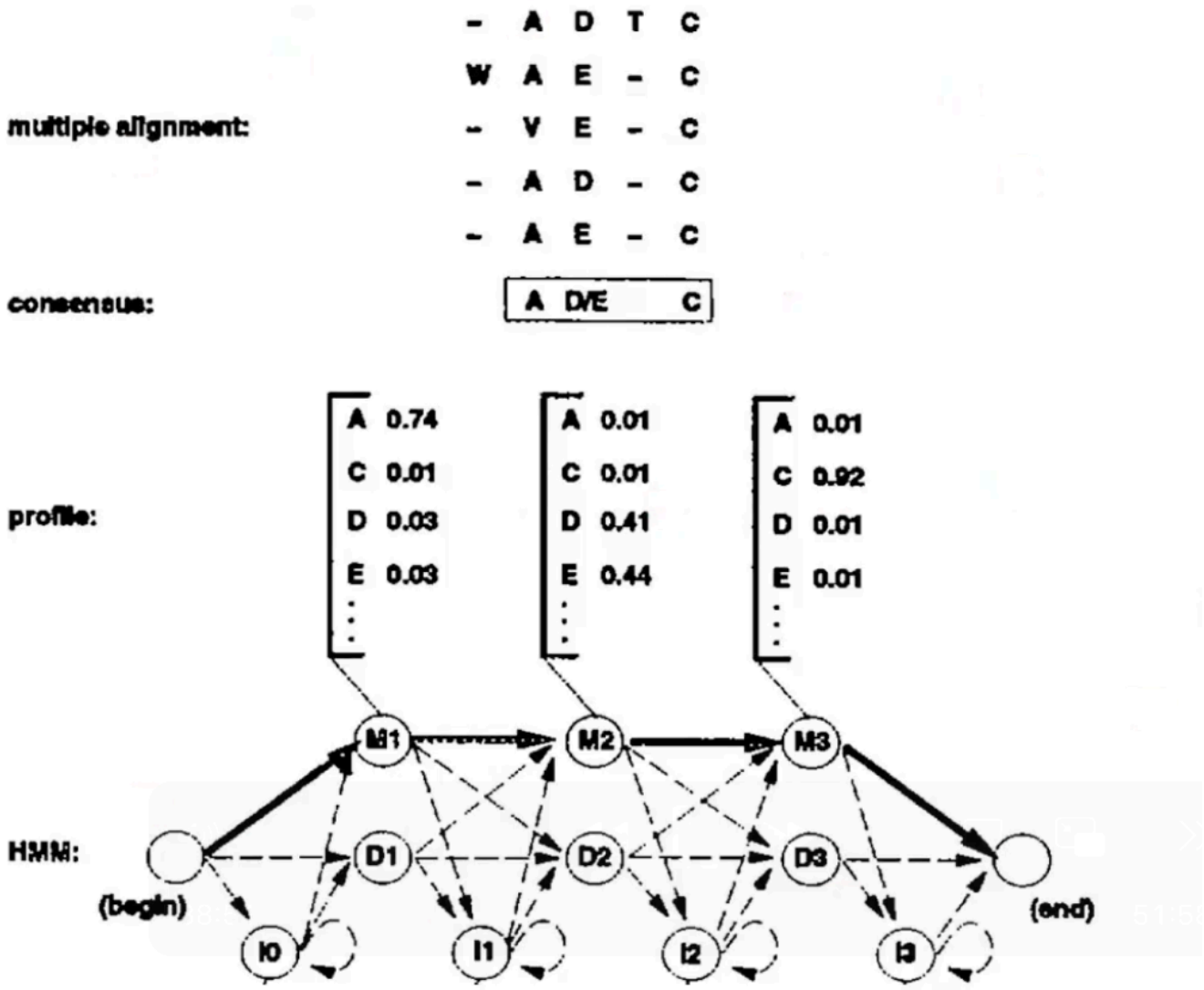
$$P(x,\pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}}$$

蛋白结构域相关的隐马尔可夫模型构建（建模）

*隐马尔可夫模型的构建基础是蛋白的结构域的多序列比对

*多序列比对的状态分为匹配（match，M），插入（insert，I）和删除（delete，D）

*构建的隐马尔可夫模型要保证观测到该多序列比对的概率最大



Pfam (http://pfam.xfam.org)

The Pfam database is a large collection of protein families [2], each represented by *multiple sequence alignments* [3] and *hidden Markov models (HMMs)* [4].

Proteins are generally composed of one or more functional regions, commonly termed *domains*. Different combinations of domains give rise to the diverse range of proteins found in nature. The identification of domains that occur within proteins can therefore provide insights into their function.

Pfam also generates higher-level groupings of related entries, known as *clans*. A clan is a collection of Pfam entries which are related by similarity of sequence, structure or profile-HMM.

The data presented for each entry is based on the UniProt Reference Proteomes but information on individual UniProtKB sequences can still be found by entering the protein accession. Pfam *full* alignments are available from searching a variety of databases, either to provide different accessions (e.g. all UniProt and NCBI GI) or different levels of redundancy.

在pfam数据库将多序列比对分为两个层级

seed alignments（非常重要的！！

*形成过程：繁复的人工审阅，比对的事否合适，序列事否存在保守型

*稳定

full alignments

*会更新的

*有助于提高敏感性

---

1. CpG 作为作为表观调控的重要组成部分，在基因调控方面起着重要作用，及是观察其C事否被甲基化修饰。CpG 岛主要位于基因的启动子（promotor）和第一外显子区域，约有 60% 以上基因的启动子含有 CpG 岛。CpG 岛的 GC 含量大于 50%，长度超过 200bp。（CpG岛预测方法：https://www.biomart.cn/experiment/792/2713475.htm）。但是如果在基因组中CG连在一起，C很容易被甲基化修饰，且修饰后是不稳定，大多会转化为T，所以在基因组中很少看到CG连在一起，但在CpG岛却可以经常看见。 ↩

2. ~~mistaken 并非利用蛋白全长（从n端到c端）.~~而是利用的domn（结构域，也是蛋白发挥功能的基本单位） ↩

3. 多序列比对，利用的UniProt Reference Proteomes 序列进行多序列比对，并在其中存在一定的先验知识：对应的几条序列均具有大致相同的功能，某一些区段经过实验的验证。在pfam数据库将多序列比对分为两个层级 ↩

4. 隐马尔可夫模型 ↩