中国科学 E 辑: 信息科学
2007 年 第 37 卷 第 10 期: 1285~1303
http://www.scichina.com

《中国科学》杂志社
SCIENCE IN CHINA PRESS

RNA 二级结构预测的模糊模型

宋丹丹 邓志东*

(清华大学计算机科学与技术系, 信息科学与技术国家实验室, 北京 100084)

摘要 基于模糊集合理论, 提出了RNA二级结构预测的模糊模型. 该模型通过状态空间的模糊分割以及模糊目标的引入等, 可有效利用模糊动态规划算法给出该模糊模型的最优决策序列, 并进而获得待预测RNA最优与次优的二级结构. 基于模糊模型的方法具有许多优点, 如计算复杂性的降低, 最优与多个次优二级结构的一并获得, 以及定性先验知识的有效融入等. 完整地给出了RNA二级结构的模糊模型及其计算方法, 并进行了具体的实现. 将一个具体的 BJK 模糊模型结构实际应用于 tRNA 及 tmRNA 的数据集中, 并与基于最小自由能的 mfold 工具以及基于 SCFG 的 BJK 文法模型进行了比较研究. 实验结果表明该模型的有效性, 相应的预测精度得到了进一步的提高.

关键词 **RNA 二级结构预测** **模糊模型** **模糊动态规划**

0 引言

近年来有关分子生物学的研究发现, RNA 分子不仅仅是 DNA 与蛋白质间信息的传递中介, 而且在细胞的新陈代谢与基因调控等方面也起着关键作用. RNA 分子的种类多样、结构复杂、功能繁多, 而由于"结构决定功能", 这些特点促使人们大量集中于 RNA 具体结构的研究. 随着大规模高通量 DNA 测序技术的突破性进展, 各种核酸数据库(如 GenBank)中积累了海量的 RNA 一级序列数据, 并且其数量正随着时间的推移而呈指数增长. 相比之下, 目前对于包括RNA在内的生物大分子三维空间结构的测定, 主要是通过X-晶体衍射及核磁共振等实验方法得到的. 尽管这些实验方法具有很高的精度, 但由于实验过程过于复杂, 测定一个RNA分子将花费很长的时间和较高的实验费用, 从而导致RNA三维结构数据相对于其线性序列数据之保有量, 其差距正不断扩大.

基于计算方法预测RNA的二级结构, 则能较好地解决这个意义重大却十分困难的问题. 利用已知的各种计算模型, 计算机可根据输入的 RNA 序列数据, 直接预测出相应的二级结构, 从而为 RNA 分子的功能分析提供必要的依据. 因此目前有关 RNA 二级结构预测方法的研究,

收稿日期: 2006-12-19; 接受日期: 2007-01-09
国家自然科学基金(批准号: 60621062)和国家教育部高等学校优秀青年教师教学科研奖励计划资助项目
* 联系人, E-mail: michael@tsinghua.edu.cn

正持续成为计算生物学或生物信息学的重大研究热点之一.

目前已有许多预测RNA二级结构的典型计算方法, 如基于最大碱基配对数的Nussinov算法[1], 基于最小自由能的Zuker算法[2]及其各种变形等, 其中Zuker算法经过不断地改进与提高, 已经发展出几个广泛使用的折叠工具, 如mfold[3], RNAfold(vienna RNA package)[4]等. 但由于RNA种类的多样性与广泛性, 最小自由能方法中先验给定的热力学参数往往存在一定的局限, 不可能对各种情形都完全准确有效, 对某些问题, 其预测精度有可能急剧下降, 且会出现对参数的过度敏感等问题. 目前Zuker算法中通常采用先后预测出最优与多个次优二级结构的方式来缓解这个问题, 但相应地会带来计算复杂性的增加. 从理论上说, 上述算法本质上都属于确定性的动态规划类算法.

一些随机模型也逐渐应用到RNA二级结构的预测问题中, 如基于随机上下文无关文法(stochastic context-free grammar, 简称SCFG)[5~7]、Bayes估计[8], 以及分割函数[9]等的预测方法. 在这些随机的动态规划类算法中, SCFG方法由于对RNA二级结构的描述简单直观, 而受到人们的普遍重视, 具有相当的代表性. 此外, 也还有一些基于遗传算法[10]、神经网络[11]等的启发式动态规划类方法. 但在此类算法中, 由于预测精度大多较低, 目前并没有公认的最优预测算法.

上述大部分预测方法, 由于存在较大的计算复杂度, 因此一般不能处理太长的 RNA 序列, 尤其是对于长度大于 1000 个碱基的 RNA 序列, 通常存在很大的困难, 甚至不能实现. 对部分数据集(如 ncRNA), 预测精度仍待大幅提高.

另一方面, 自Zadeh于 1965 年开创了模糊集合理论以来[12], 由于模糊语言模型能更好地描述并应用基于人类自然语言表达的定性知识与经验, 并且具有近似推理的能力, 因此能更有效地解决随机数学所不能解决的某些不确定性问题. 模糊推理系统事实上已成为计算智能中最为成熟的方法之一, 已成功地应用于诸多领域. 目前在计算生物学领域, 也有一些模糊集合理论的应用实例[13,14].

基于模糊集合理论的上述优点, 受已有确定性动态规划类算法及随机动态规划类算法的启发, 本文提出了一种RNA二级结构预测的模糊模型. 该模型对状态空间进行模糊分割, 并引入模糊目标. 利用模糊动态规划算法[15], 通过计算出该模糊模型的最优模糊策略, 进而利用所谓脊线计算, 可同时获得待预测RNA的最优与次优二级结构. 基于模糊模型的方法具有许多明显的优点, 如计算复杂性的降低, 最优与多个次优二级结构的一并获得, 以及定性先验知识的有效融入等, 本文完整地提出了基于模糊模型的RNA二级结构预测方法, 并进行了具体的实现. 我们将一般模型的一个具体的BJK模糊模型结构, 实际应用于tRNA及tmRNA的数据集中, 并同最具代表性的mfold折叠工具与基于SCFG的BJK文法模型进行了比较研究. 实验结果不仅表明所提模型的有效性, 而且可使相应的预测精度得到进一步的提高.

1 方法

1.1 模糊模型的基本思想

在各种确定性及随机性RNA二级结构预测模型中, 通常以碱基为研究对象. 在这些模型定义的状态空间中, 行与列的每个元素对应于待预测序列的各个碱基, 所有碱基两两组合构成一个多层二维上三角动态规划矩阵. 每层二维上三角矩阵中的元素分别代表该行与该列两个碱基处于不同结构状态时, 两者之间子序列结构的最优值. 因此通过设定RNA二级结构的最优化准则, 可将RNA二级结构预测问题转化为最优决策问题, 其中每条决策对应于不同碱

基状态之间的局部转移关系. 由决策(子)序列组成的一个(子)策略对应着各碱基状态的一条转移路径, 从而对应着 RNA 当前(子)序列的一个二级结构. 利用确定性或随机性动态规划算法, 对上述上三角动态规划矩阵进行迭代填充与回溯(traceback), 即可得到最优策略与最优路径, 相应可得到设定准则下的 RNA 最优二级结构.

RNA二级结构除了可以用序列中每个碱基的配对或不配对状态简单表示之外, 还可以将其按基本结构单元进行细化分解, 进而可通过其所包含的结构单元, 以及每个结构单元所包含的碱基来表示该二级结构. 这些基本结构单元包括茎(stem)、发夹环(hairpin)、内环(internal loop)、凸环(bulge loop)、多分枝环(multi-branched loop), 以及单链(single stranded)等[16], 且每个基本结构单元通常包含多个相邻碱基. 因此, 如果能够以多个相邻碱基作为研究单位, 在正确预测出多个碱基所属的基本结构单元之后, 再对其中每个碱基的结构进行细化, 也可以得到RNA的二级结构; 而且在生物学意义上, 与氨基酸残基类似, RNA中的单个碱基并不能维持其结构的稳定, 其周围的碱基结构情形必定会对结构单元的稳定产生影响. 因此, 通过将状态空间进行模糊分割, 利用模糊子集代表相邻多个碱基的信息, 从而减小计算复杂度, 这是我们研究模糊模型的基本出发点.

另一方面, 在 RNA 结构中, 特别是对于同源的 RNA 结构, 各个基本结构单元的分布情况具有一定的规律. 若能利用这些规律, 则对提高RNA结构的预测精度具有重要的意义, 而有效融入先验知识正是使用模糊语言的最大优势所在.

本文提出RNA二级结构预测的模糊模型, 其基本思想就是以模糊子集为研究对象, 通过定义模糊目标引入先验知识, 在降低计算复杂性的基础上提高预测的精度. 具体包括: 通过将以碱基为单位的状态空间模糊分割为彼此重叠的模糊子集, 使用模糊子集来表示相邻多个碱基的整体性质, 以减小计算复杂性; 同时通过引入模糊目标, 选取RNA保守结构信息作为该模糊子集的特征, 从而可将先验知识引入预测过程中, 使预测精度能够进一步提高.

模糊模型中定义的模糊决策对应了模糊子集间的状态转移关系. 通过模糊动态规划算法中对模糊动态规划矩阵的迭代填充过程与寻求最优决策的回溯过程, 可计算出最大隶属度准则下的最优模糊策略, 从而得到由结构单元构成的最可能的模糊化的二级结构, 进而通过清晰化过程对每个碱基进行细化, 以便最终得到反模糊化后的 RNA 最优及次优二级结构.

本文首先给出了预测 RNA 二级结构的模糊模型之基本概念, 进而提出了阶段数隐含的一般模糊模型之结构与参数的确定方法. 其次, 通过使用阶段数隐含的模糊动态规划算法, 可以计算出对应于待预测 RNA 序列多个二级结构的最优模糊策略. 最后, 待预测 RNA 的最优二级结构, 可以通过对最优模糊策略“脊线”的清晰化计算得到.

1.2　模糊建模

1.2.1　模糊模型的基本概念

一般地, 本文提出的用于预测 RNA 二级结构的模糊模型主要涉及模糊状态空间、允许模糊决策集、最优模糊策略、模糊状态转移规则, 以及终止模糊状态集等基本概念.

- 模糊状态空间.

在RNA二级结构的各种预测模型中, 状态空间通常使用如下的 n 个(层)不同的二维动态规划矩阵表示:

$$\mathbf{H}=\{H_1,H_2,\cdots,H_n\},$$

其中, 每个状态子空间 $H_r(r=1,2,\cdots,n)$ 是以序列的碱基作为行与列构成的二维矩阵, 因此该矩阵中的元素是该序列中所有可能的两两碱基组合. 每个元素 (i,j) 的数值表示在当前结构状态下(如碱基配对时的茎), 第 i 个至第 j 个碱基间的子序列 $[i,i+1,\cdots,j]$ (当 $i\leqslant j$ 时)的结构最优值, 这里不同的状态子空间即代表不同的结构状态. 假设 RNA 序列的长度为 l, 从序列的5′端到 3′端依次给各碱基标号为 $1,2,\cdots,l$, 对应各碱基 $B(i)$ $(i=1,\cdots,l)$, 则每个状态子空间为 $l\times l$ 大小的矩阵. 又由于 (i,j) 与 (j,i) 意义相同, 因此我们只需对 (i,j) 且 $i\leqslant j$ 的情形进行计算. 显然在该结构状态下, RNA 序列中所有可能的两两碱基组合就构成了一个上三角矩阵, 如图 1 所示. 所有的 n 个状态子空间便构成 n 层的 $l\times l$ 上三角动态规划矩阵.

而在模糊模型中, 我们用彼此重叠的模糊子集来描述序列中相邻多个碱基的整体性质. 具体地, 对每层状态子空间 $H_r(r=1,2,\cdots,n)$, 将图 1 所示上三角动态规划矩阵的行与列, 即 i 与 j 维方向上的 RNA 碱基序列, 分别利用如下所示的 N_a^r 及 N_b^r 个模糊子集进行模糊分割, 即

$$i:(A_1^r,A_2^r,\cdots,A_{N_a^r}^r),$$

$$j:(B_1^r,B_2^r,\cdots,B_{N_b^r}^r),$$

其中 $r=1,2,\cdots,n$. 进而在二维空间上, 通过(1)式中对状态子空间 H_r 中 i 维与 j 维模糊子集的直积(cartesian product)运算, 可得到如图 2 所示用二维模糊子集表示的上三角模糊动态规划矩阵, 从而将以碱基为单位的清晰化状态子空间 H_r 转化为以模糊子集为单位的模糊状态子空间 $\tilde{H}_r(r=1,2,\cdots,n)$.

$$C_{p,q}^r=(A_p^r,B_q^r), \tag{1}$$

这里 $r=1,2,\cdots,n$, $p\leqslant q$, 且 $p(q)=1,2,\cdots,N_a^r(N_b^r)$.

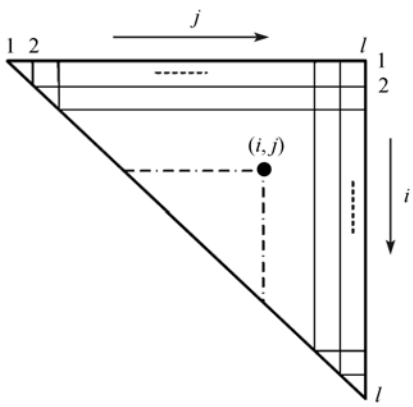

图 1 第 r 层状态子空间 H_r 的示意图 $(r=1,2,\cdots,n)$

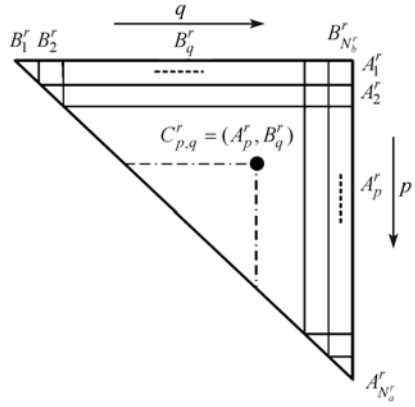

图 2 以 $N_a^r=N_b^r$ 为例, 用模糊子集表示的模糊状态子空间 $\tilde{H}_r$ 示意图

相应二维模糊子集的隶属函数则由每一维隶属函数的直积确定, 如(2)式所示. 这个二维模糊子集隶属函数的定义, 可用于以后的清晰化计算.

$$\mu_{A\times B}(C^r_{p,q}) = \min\{\mu_A(A^r_p), \mu_B(B^r_q)\}, \tag{2}$$

其中, μ 表示元素的模糊隶属度. 我们称每个二维模糊子集中具有最大隶属度的点为“脊点”(ridge point),

$$\arg\min_{p,q}(\mu_{A\times B}(C^r_{p,q})).$$

因此模糊状态的取值范围是 n 层的模糊状态空间 $\tilde{\mathbf{H}}$, 即由 n 个模糊状态子空间中的全体模糊子集所组成的集合

$$X_t \in \mathbf{X} = \tilde{\mathbf{H}} = \{C^r_{p,q} \mid r = 1,2,\cdots,n; p(q) = 1,2,\cdots,N^r_a(N^r_b)\}. \tag{3}$$

- 允许模糊决策集.

允许决策集中的决策与状态之间的转移关系一一对应. 在已有的各种 RNA 二级结构预测模型中, 允许决策集中的各个决策分别对应于碱基状态间的各种转移关系, 相应决策的值代表执行该决策时状态转移的得分或概率. 而在以模糊子集为对象的模糊模型中, 允许模糊决策集中的各模糊决策对应了代表多个碱基性质的模糊子集间的状态转移关系, 其隶属度为执行该决策的可能性大小. 因此, 允许模糊决策集由所有可能的模糊决策组成, 即

$$U_t \in \mathbf{U} = \{U_1, U_2, \cdots, U_m\},$$

其中每个模糊决策对应于一条模糊状态转移规则, 即根据各模糊状态子空间的实际定义, 特别是各模糊子集之间关系的实际含义, 对应于从模糊子集到模糊子集的一种模糊状态转移关系,

$$\text{If} \quad X_t \text{ is } C^r_{p,q} \quad \text{and} \quad U_t \text{ is } U_d, \quad \text{then} \quad X_{t+1} \text{ is } C^{r'}_{p',q'}, \tag{4}$$

其中 $d = 1,2,\cdots,m,\ r(r') = 1,2,\cdots,n,\ p \leqslant q, p(q) = 1,2,\cdots,N^r_a(N^r_b),\ p' \leqslant q', p'(q') = 1,2,\cdots,N^{r'}_a(N^{r'}_b)$. 对于不同的模糊状态子空间 $\tilde{H}_r$ 中的模糊状态, 对应的模糊决策集是允许模糊决策集 $\mathbf{U}$ 的子集, 即 $\mathbf{U}_r \subseteq \mathbf{U}$.

- 最优模糊策略准则.

在不同的 RNA 二级结构预测算法中, 设定的最优化准则全然不同. 例如, Nussinov 算法以碱基配对数最大作为二级结构的最优化准则; Zuker 算法的最优化准则为自由能最小; SCFG 算法的最优化准则为似然度最大. 在本文的模糊模型中, 我们定义的最优化准则为隶属度最大.

对于模糊推理系统, 当前 k–1 阶段的模糊策略是由当前及所有过往阶段模糊决策所组成的子序列 $(U_0, U_1, \cdots, U_{k-1})$, 而其中隶属度最大的为当前阶段的最优模糊策略, 即

$$\mu(U^*_0, U^*_1, \cdots, U^*_{k-1}) = \arg\max_{U_0, U_1, \cdots, U_{k-1}} \mu(U_0, U_1, \cdots, U_{k-1}), \tag{5}$$

其中 k–1 为模糊阶段数. 在 RNA 二级结构预测的模糊模型中, 同样需满足局部最优化原理, 即最优策略的子策略应是对应子问题的最优策略. 因此, 在对模糊动态规划矩阵的迭代填充过程中, k 表示从初始值 1 不断递推增大的 RNA 子序列中平均模糊子集的个数.

- 终止模糊状态集.

在 RNA 二级结构预测的模糊模型中, 对于回溯的转移推导过程, 从模糊动态规划矩阵右上角的初始模糊状态出发, 当前模糊状态按模糊决策进行状态转移, 而当其进入终止模糊状

态集, 即到达模糊动态规划矩阵的对角线方向时, 即满足停止条件, 转移推导过程就此完成. 显然这里并不需要终止阶段数的定义, 而且对于不同的模糊策略, 相应的终止阶段数不同, 这称为阶段数隐含的模糊模型. 此时, 终止模糊状态集 W 应满足

$$\mathbf{W}=\{\hat{X}_1,\hat{X}_2,\cdots,\hat{X}_w\}\subset\mathbf{X}, \tag{6}$$

显然, 终止模糊状态集中不包含初始模糊状态, 即应满足: $X_0\in\mathbf{X}\setminus\mathbf{W}$, 这里 w 为终止模糊状态的个数.

1.2.2 BJK 模糊模型结构的确定

针对包括tRNA与rRNA在内的结构RNA的二级结构预测问题, 本文给出了一种性能较优的BJK模糊模型结构. 所谓BJK模糊模型结构实际是前述一般模糊模型结构的一个实现, 其中模糊状态空间、允许模糊决策集、模糊状态转移规则等的选择, 参考了随机上下文无关文法(SCFG)模型中BJK文法[6]的有关定义. 但与后者相比, 由于这里的模糊模型结构使用了模糊分割、模糊目标与模糊动态规划算法等, 因此在计算复杂性的降低、更多定性知识的融入, 以及预测精度的提高等方面都具有明显的优势.

- 模糊状态空间的分割.

在 BJK 模糊模型结构中, 我们首先定义 3 层状态子空间, 分别记为 L, S, F, 即状态空间 $\mathbf{H}=\{L,S,F\}$, 状态子空间数 $n=3$, 其中, S 为主要产生环结构的结构状态, F 为主要产生茎结构(即连续的碱基配对)的结构状态, L 的结构状态则决定了对应位置上是单个不配对碱基或是一个新的茎结构的起始. 在将这3层状态子空间模糊分割成以模糊子集为单位的模糊状态子空间 $\tilde{\mathbf{H}}=\{\tilde{L},\tilde{S},\tilde{F}\}$ 时, 可采用不同的隶属函数, 且对于不同的隶属函数, 都有相应的参数化表达. 这些参数一般可先验给定初值, 然后利用自组织聚类等方法进行细化调整.

为了简便起见, 这里对于二维模糊状态子空间 H_r 的 i 维及 j 维方向, 分别采用如图 3 和 4

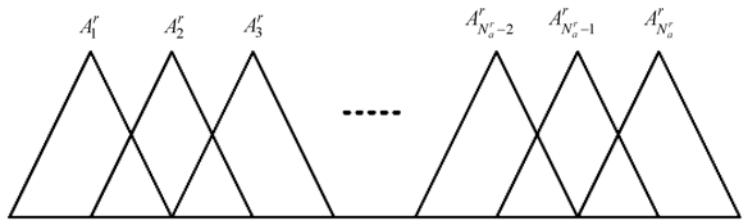

图 3 模糊状态子空间 H_r 中定义在 i 维上的三角形隶属函数

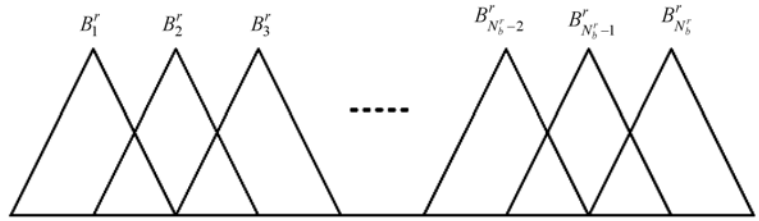

图 4 模糊状态子空间 H_r 中定义在 j 维上的三角形隶属函数

所示的对称三角形隶属函数进行模糊分割, 相应得到各层模糊状态子空间 $\tilde{H}_r$. 每个三角形隶属函数的中心及宽度相应为 $M^r_{p(q)}$ 及 $\delta^r_{p(q)}$, 其中 $p(q)=1,2,\cdots,N^r_a(N^r_b)$.

因此按照(2)式给出的二维隶属函数的直积计算方法, 每个二维模糊子集的隶属函数为如图 5 所示的四棱锥. 此时, 每个模糊子集的"脊点"即为锥的顶点.

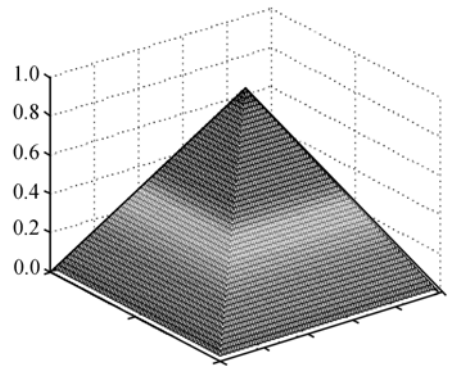

图 5　模糊状态子空间 $\tilde{H}_r$ 中定义二维模糊子集 $C^r_{p,q}$ 的四棱锥形隶属函数

● 允许模糊决策集.

对上述模糊状态子空间, 允许模糊决策集定义为 $\mathbf{U}=\{U_1,U_2,U_3,U_4,U_5,U_6\}$, 其中各模糊决策 $U_d(d=1,2,\cdots,6)$ 对应于一条模糊状态转移规则, 且给出为

$$U_1:\text{If}\quad X_t \text{ is } C^{\tilde{L}}_{p,q}\quad\text{and}\quad U_t \text{ is } U_1,\quad \text{then}\quad X_{t+1} \text{ is } C^{\tilde{F}}_{p+1,q-1} \text{ and } \mathit{Pair}(C^{\tilde{L}}_{p,q}); \tag{7}$$

$$U_2:\text{If}\quad X_t \text{ is } C^{\tilde{L}}_{p,p}\quad\text{and}\quad U_t \text{ is } U_2,\quad \text{then}\quad X_{t+1} \text{ is } \mathit{Single}(C^{\tilde{L}}_{p,p}); \tag{8}$$

$$U_3:\text{If}\quad X_t \text{ is } C^{\tilde{S}}_{p,q}\quad\text{and}\quad U_t \text{ is } U_3,\quad \text{then}\quad X_{t+1} \text{ is } C^{\tilde{L}}_{p,v} \text{ and } C^{\tilde{S}}_{v+1,q}\quad (p\leqslant v<q); \tag{9}$$

$$U_4:\text{If}\quad X_t \text{ is } C^{\tilde{S}}_{p,q}\quad\text{and}\quad U_t \text{ is } U_4,\quad \text{then}\quad X_{t+1} \text{ is } C^{\tilde{L}}_{p,q}; \tag{10}$$

$$U_5:\text{If}\quad X_t \text{ is } C^{\tilde{F}}_{p,q}\quad\text{and}\quad U_t \text{ is } U_5,\quad \text{then}\quad X_{t+1} \text{ is } C^{\tilde{L}}_{p,v} \text{ and } C^{\tilde{S}}_{v+1,q}\quad (p\leqslant v<q); \tag{11}$$

$$U_6:\text{If}\quad X_t \text{ is } C^{\tilde{F}}_{p,q}\quad\text{and}\quad U_t \text{ is } U_6,\quad \text{then}\quad X_{t+1} \text{ is } C^{\tilde{F}}_{p+1,q-1} \text{ and } \mathit{Pair}(C^{\tilde{F}}_{p,q}), \tag{12}$$

其中 *Pair* 和 *Single* 为下面将要介绍的观测字符模糊集. 显然, 这里的模糊状态转移规则并不限于一对一的映射, 而是包含了一对多的映射.

进一步地, 分别对应于 $\tilde{L}$, $\tilde{S}$, $\tilde{F}$ 3个模糊状态子空间, 相应的允许模糊决策子集为 $\mathbf{U}$ 的子集, 分别确定如下:

$$\begin{aligned}\mathbf{U}_{\tilde{L}}&=\{U_1,U_2\}\subset\mathbf{U},\\ \mathbf{U}_{\tilde{S}}&=\{U_3,U_4\}\subset\mathbf{U},\\ \mathbf{U}_{\tilde{F}}&=\{U_5,U_6\}\subset\mathbf{U}.\end{aligned} \tag{13}$$

● 观测字符模糊集.

观测字符集为序列中的碱基, 因此包括 $\{A,C,G,U\}$ 4 种字符, 而观测字符集的生成包括碱

基配对出现及碱基不配对出现两种情形, 分别对应 RNA 二级结构中碱基的两种不同状态. 记 *Pair* 和 *Single* 为观测字符模糊集合, 其中 *Pair* 的元素是广义上所有可能的碱基配对, 因此可以用 4×4 的矩阵表示; *Single* 的元素是所有单个碱基(不配对情形), 因此具有 A,C,G,U 4 个元素, 可以用一维向量表示. 此时, 这两个模糊集合中元素的隶属度分别表示产生这些碱基对或单个碱基的可能性大小, 可以表示为

$$\mu_{Pair}=\begin{array}{c} \\ A\\U\\G\\C\end{array}\begin{array}{c}\begin{array}{cccc}A & C & G & U\end{array}\\\begin{bmatrix}\mu_{AA} & \mu_{AC} & \mu_{AG} & \mu_{AU}\\ \mu_{CA} & \mu_{CC} & \mu_{CG} & \mu_{CU}\\ \mu_{GA} & \mu_{GC} & \mu_{GG} & \mu_{GU}\\ \mu_{UA} & \mu_{UC} & \mu_{UG} & \mu_{UU}\end{bmatrix}\end{array}, \tag{14}$$

$$\mu_{Single}=\begin{bmatrix}\mu_A & \mu_U & \mu_G & \mu_C\end{bmatrix}, \tag{15}$$

其中各隶属度的确定将在后续模型参数确定部分中介绍. 因此, 上述模糊状态转移规则中的 $Pair(C_{p,q}^r)$, 是指以模糊子集 $C_{p,q}^r$ 脊点位置对应的碱基对为变量, 观测字符集 *Pair* 中的对应元素. 而 $Single(C_{p,p}^r)$ 则是指模糊子集 $C_{p,p}^r$ 脊点位置对应的一对碱基中, 在 *Single* 集合中隶属度较大的那个碱基. 目前在BJK模糊模型中, 我们采用模糊子集脊点的碱基代表该模糊子集中各碱基的整体性质来进行预测(这里 $r\in\{\tilde{S},\tilde{L},\tilde{F}\}$).

- 终止状态与终止模糊状态集.

根据上述 BJK 模糊模型中具体定义的模糊状态子集与模糊决策子集, 在转移推导过程中, 当所有当前状态都转化为观测字符模糊集 *Single* 时, 模糊状态转移过程终止. 由于 *Single* 是从模糊状态 $C_{p,p}^{\tilde{L}}$ 使用决策 U_2 转移得到的, 因此我们可在状态子空间 H_L 之上三角矩阵的对角线方向, 增加一条虚拟的下次对角线 $L(i,i-1)(i=1,2,\cdots,l)$ 作为转移终止状态. 这些次对角线上的点, 实际对应了每个可能的不配对碱基 $B(i)(i=1,2,\cdots,l)$, 它们是对状态子空间 H_L 的扩展, 与状态子空间 H_L 一起进行模糊分割, 从而得到用模糊子集表示的扩展的模糊状态子空间 $\tilde{H}_L$.

因此终止模糊状态集 **W** 中的元素为那些使某个下次对角线点 $L(i,i-1)(i=1,2,\cdots,l)$ 隶属度大于 0 的模糊子集, 即为模糊状态子空间 $\tilde{H}_L$ 中满足下式的模糊子集:

$$\exists i\in\{1,\cdots,l\}\quad \mu_{C_{p,q}^{\tilde{L}}}(L(i,i-1))>0. \tag{16}$$

- 模糊目标集合.

对于阶段数隐含的模糊动态规划, 模糊目标集合一般为终止状态的模糊集合, 因此这里为不配对碱基模糊集. 该模糊集合的元素——模糊目标, 即为状态子空间 H_L 下次对角线上的点 $L(i,i-1)(i=1,\cdots,l)$. 事实上, 如此选取的模糊目标对应每个可能的不配对碱基, 而每个模糊目标在模糊目标集合中的隶属度即为该碱基在二级结构中属于不配对碱基集合的隶属度, 它表示了该碱基不配对的可能性大小. 由此可见, 可以通过训练过程中对模糊目标的训练, 将同源RNA的结构特性引入到预测过程中, 从而可通过先验知识的引入, 进行后续的预测过程. 目前引入的特征仅为不配对碱基在序列中的分布情形.

显然, 通过选择不同的结构参数, 就可以得到不同的模糊模型结构. BJK 模糊模型结构只

是其中参考 BJK 文法模型的一个先验选择. 更为一般地, 我们可以通过自组织聚类的方法, 发展 RNA 二级结构预测模糊模型的结构学习算法.

1.2.3　模糊模型的参数估计

在模糊模型结构确定之后, 训练样本数据集可用来估计模糊模型的参数, 这里主要采用了模糊统计的方法进行参数估计. 在文献[17]中模糊统计也称为"投票"方法. 与递推的学习方法比较, 该法可以大大地降低计算复杂度, 并减少计算时间. 尽管在计算精度上有所损失, 但本文的研究表明, 目前的计算可以基本满足精度的要求.

在计算模糊模型的隶属度参数时, 为避免过度学习出现隶属度为零的情形, 通常使用如下Laplace先验(加一)计算公式[18]:

$$\mu_i = \frac{n_i + 1}{\sum_i n_i + I},\quad i = 1,2,\cdots,I, \tag{17}$$

其中, n_i 为第 i 种情形出现的次数, 且 I 表示总的情形数.

- 模糊决策的隶属度.

每个模糊决策的隶属度表示执行该决策的可能性大小. 由于采用阶段数隐含的模糊模型, 阶段数预先无法确定, 因此模糊决策的隶属度不能随阶段数变化, 但可随当前模糊状态的不同而不同. 为简单起见, 本例中模糊决策对所有的模糊状态都取同样的隶属度, 即模糊决策的隶属度与状态独立:

$$\mu_C(U_t \mid X_{t1}) = \mu_C(U_t \mid X_{t2}) = \mu_C(U_t). \tag{18}$$

每个模糊决策隶属度的计算如下:

$$\mu_C(U_i^r) = \frac{n_i^r + 1}{\sum_j n_j^r + n^r}, \tag{19}$$

其中, n_i^r 为模糊决策 U_i^r 在训练集中出现的次数, 而 n^r 为模糊状态子空间 $\tilde{H}_r$ 对应的允许模糊决策子集 $\mathbf{U}_r$ 中元素的个数. 在本文的 BJK 模糊模型结构中, $n^r = 2$, 这里 $r \in \{\tilde{L}, \tilde{S}, \tilde{F}\}$.

- 观测字符模糊集元素的隶属度.

观测字符模糊集 *Pair* 与 *Single* 中元素的隶属度分别表示在该模糊集中产生这些碱基对或单个碱基的可能性, 同样采用模糊统计的方法确定. 为此, 可首先统计出已知的RNA二级结构中不同配对碱基出现的个数, 然后依据(17)式, 确定 *Pair* 中各元素的隶属度. 此时, 有

$$\mu_{Pair}(X,Y) = \frac{n_{X,Y} + 1}{\sum_{X,Y} n_{X,Y} + 16},\quad X(Y) \in \{A,C,G,U\}. \tag{20}$$

类似地, 可计算出已知RNA二级结构中不同单碱基的出现个数, 以确定 *Single* 中各元素的隶属度. 我们有

$$\mu_{Single}(X) = \frac{n_X + 1}{\sum_X n_X + 4},\quad X \in \{A,C,G,U\}. \tag{21}$$

- 模糊目标的统计.

目前选取的模糊目标及其特征, 为序列中的碱基及其不配对的可能性大小. 由于训练样本集中不同的RNA序列长度略有差别, 因此对于训练样本集中的每个已知二级结构数据, 均

使用相对位置的概念来统计模糊目标. 在训练阶段, 只需对所有不配对碱基的相对位置($rp_{l'}^{i}=i/l'$)进行统计保存(l'为序列长度, i 为该碱基的位置标号). 模糊目标隶属度的计算, 将在模糊动态规划算法的初始化阶段对待预测 RNA 序列进行.

1.3 模糊推断: 模糊动态规划算法

在本文提出的 RNA 二级结构预测的模糊模型中, 使用模糊动态规划算法来确定最优模糊策略/最优模糊路径. 与 Zuker 算法和 SCFG 算法中分别使用的确定性与随机动态规划算法类似, 模糊动态规划算法首先通过从终止状态到初始状态方向对模糊动态规划矩阵的迭代填充过程, 得到最优策略的隶属度值, 然后再通过回溯过程得到相应从初始状态到终止状态的最优模糊策略/最优模糊路径. 而与这两种算法不同的是, 这里要解决的问题是模糊系统的多阶段最优决策问题. 最优模糊策略定义为具有最大隶属度的策略, 即

$$\mu(U_0^*,\cdots,U_{K-1}^*\mid X_0)=\max_{U_0,\cdots,U_{K-1}}(\mu_C(U_0)\circ\cdots\circ\mu_C(U_{K-1})\circ(\mu_{X_K}(x_K)\circ\mu_{G^K}(x_K))), \tag{22}$$

其中 K 为总的模糊阶段数. 对于阶段数隐含的模糊动态规划算法, K 不能先验给定, 且随路径的不同而变化, 因此上式中第 K 阶段仅指终止阶段, 而 $K-1$ 则是终止前一个阶段, 依此类推. $\mu_{X_K}(x_K)$ 表示终止状态在终止状态集中的隶属度, 此时为序列碱基 x_K 在终止状态集 X_K 中的隶属度; 而 $\mu_{G^K}(x_K)$ 则表示其在模糊目标集中的隶属度, 代表了该碱基不配对的可能性. 此时, 模糊状态转移规则满足

$$\text{If}\quad X_t \text{ is } X_k \quad\text{and}\quad U_t \text{ is } U_k,\quad \text{then}\quad X_{t+1} \text{ is } X_{k+1}, \tag{23}$$

其中, $k=0,1,\cdots,K-1$.

在(22)式中, 模糊算子$\circ$定义为代数积, 即 $x\circ y=xy$. 特别地, 为了减少截断误差, 本文将这里的隶属度相乘运算, 通过如下的取对数并扩大 1000 倍取整, 从而将其转化为整数相加运算:

$$Lg(x)=[1000*\log_2\mu(x)]. \tag{24}$$

1.3.1 迭代填充计算

基于模糊目标与模糊约束的迭代填充计算过程, 采用迭代的步骤, 从对应子序列长度最短的模糊子集出发(即模糊动态规划矩阵对角线上的模糊子集)开始, 首先计算子序列决策中具有最大隶属度的最优决策, 继而选取使子序列长度逐渐增大的模糊子集进行迭代计算, 直至达到包含整个序列长度的初始状态.

(ⅰ) 初始化. 这里只需考虑模糊目标集中各模糊目标隶属度的初始化, 即计算待预测 RNA 序列中每个碱基位置在训练样本集中不配对的频度. 具体计算步骤如下:

在训练模型参数时, 已经统计了每个训练样本中各个模糊目标的相对位置. 此时, 对于待预测RNA序列, 用相对位置乘以该序列的长度 l, 就可得到各模糊目标对于该序列的绝对位置, 即

$$rp_{l'}^{i}\times l, \tag{25}$$

其中, $rp_{l'}^{i}$ 为序列长度为 l' 的训练样本中第 i 个位置上的模糊目标的相对位置.

将每个绝对位置就近取整, 从而将训练中统计的各个模糊目标离散化映射到待预测RNA序列的每个碱基位置上.

使用(17)式, 计算待预测序列$1,2,\cdots,l$位置上各个模糊目标的隶属度, 即

$$\mu(L(i,i-1))=\frac{n_i+1}{\sum_{j=1,\cdots,l} n_j+l}\quad (i=1,2,\cdots,l), \tag{26}$$

其中n_i与n_j分别表示在第i个及第j个位置上模糊目标出现的次数.

进一步将上述模糊目标隶属度归一化,

$$\mu(L(i,i-1))=\frac{1}{\max_{j=1,\cdots,l}(\mu(L(j,j-1)))}*\mu(L(i,i-1)). \tag{27}$$

由于模糊目标隶属度都是取[0,1]之间的值, 因此对于具有多个模糊目标的策略, 隶属度相乘运算后, 模糊目标越多的策略隶属度将越小, 从而导致最优策略偏向于只有一个模糊目标的情形. 为了避免出现这种计算偏向, 我们在对模糊目标隶属度进行如(24)式所示的取对数计算后, 再进行如下式所示的按均值平移坐标的处理, 即

$$Lg(L(i,i-1))=Lg(L(i,i-1))-\frac{\sum_{j=1,\cdots,l} Lg(L(j,j-1))}{l}, \tag{28}$$

此时对于待预测 RNA, 以每个碱基的相对位置作为自变量, 便可计算出用对数表示的每个碱基的模糊目标隶属度. 在下面的填充及回溯计算过程中, 实际均采用对隶属度取对数后相加的方法, 但为描述得直观, 仍用隶属度进行介绍.

(ii) 模糊动态规划矩阵的迭代填充过程. 对于本文提出的 RNA 二级结构预测的模糊模型, 所谓迭代填充过程就是从图2所示的上三角模糊动态规划矩阵的对角线出发, 向右上方迭代, 相应计算该模糊动态规划矩阵中每个模糊子集的隶属度, 直到当前模糊状态进入使$S(1,l)$隶属度大于 0 的模糊子集组成的集合, 即初始模糊状态为止.

与传统动态规划算法思想相似, 模糊动态规划算法满足局部最优原理, 填充过程中计算得到的每个模糊子集的隶属度, 即为从对角线到该模糊子集的局部决策问题的最优模糊策略之隶属度; 填充过程结束时初始模糊状态的隶属度便为整个决策问题的最优模糊策略之隶属度. 最优模糊策略及相应的最优模糊路径, 则由回溯过程确定.

对阶段数隐含的情形, (22)式所示之最优策略的迭代公式, 可给出如下:

$$\begin{aligned}\mu_{G_{K-1}}(X_{K-1})&=\max_{U_{K-1}}\max_{x_K}(\mu_C(U_{K-1})\circ\mu_{X_K}(x_K)\circ\mu_{G^K}(x_K))\\&=\max_{U_{K-1}}(\mu_C(U_{K-1})\max_{x_K}\circ(\mu_{X_K}(x_K)\circ\mu_{G^K}(x_K))),\end{aligned} \tag{29}$$

$$\mu_{G_{K-v}}(X_{K-v})=\max_{U_{K-v}}(\mu_C(U_{K-v})\circ\mu_{G_{K-v+1}}(X_{K-v+1})), \tag{30}$$

其中, 与前面所述相同, $\mu_{X_K}(x_K)$表示终止状态在终止状态集中的隶属度, 而$\mu_{G^K}(x_K)$则表示模糊目标的隶属度, $v=2,\cdots,K$. 对于 BJK 模糊模型结构, 相应的迭代填充公式具体给出如下:

1) 对于模糊状态子空间$\tilde{L}$中的模糊子集$C_{p,q}^{\tilde{L}}$, 当$p=q$时,

$$\mu(C_{p,p}^{\tilde{L}})=\mu_{Single}(C_{p,p}^{\tilde{L}})\circ\mu_U(U_2)\circ\max_{\mu_{C_{p,p}^{\tilde{L}}}(L(i,i-1))>0}(\mu_G(L(i,i-1))\circ\mu_{C_{p,p}^{\tilde{L}}}(L(i,i-1))). \tag{31}$$

2) 对于模糊状态子空间$\tilde{S}$中的模糊子集$C_{p,q}^{\tilde{S}}$, 当$p\leqslant q$时,

$$\mu(C_{p,q}^{\tilde{S}})=\max\begin{cases}\max\limits_{p\leqslant v<q}\mu(C_{p,v}^{\tilde{L}})\circ\mu(C_{v+1,q}^{\tilde{S}})\circ\mu_U(U_3),\\ \mu(C_{p,q}^{\tilde{L}})\circ\mu_U(U_4).\end{cases}\tag{32}$$

3) 对于模糊状态子空间 $\tilde{F}$ 中的模糊子集 $C_{p,q}^{\tilde{F}}$，当 $p\leqslant q$ 时，

$$\mu(C_{p,q}^{\tilde{F}})=\max\begin{cases}\max\limits_{p\leqslant v<q}\mu(C_{p,v}^{\tilde{L}})\circ\mu(C_{v+1,q}^{\tilde{S}})\circ\mu_U(U_5),\\ \mu(C_{p+1,q-1}^{\tilde{F}})\circ\mu_{Pair}(C_{p,q}^{\tilde{L}})\circ\mu_U(U_6).\end{cases}\tag{33}$$

4) 对于模糊状态子空间 $\tilde{L}$ 中的模糊子集 $C_{p,q}^{\tilde{L}}$，当 $p<q$ 时

$$\mu(C_{p,q}^{\tilde{L}})=\mu(C_{p+1,q-1}^{\tilde{F}})\circ\mu_{Pair}(C_{p,q}^{\tilde{L}})\circ\mu_U(U_1).\tag{34}$$

注意这里的迭代并非对所有两两碱基组合进行计算，而是对将序列碱基进行模糊分割后数量相对大大减少的模糊子集进行计算，因此可以有效地降低计算复杂度.

1.3.2 回溯

在迭代填充过程结束后，使用回溯过程来寻找最优模糊策略/最优模糊路径. 回溯过程从初始模糊状态出发，直至到达终止模糊状态处为止. 回溯过程一般可使用堆栈结构来完成.

首先将模糊子集 $C_{1,N_b^{\tilde{S}}}^{\tilde{S}}$ 入栈，最优模糊决策序列初始设为空集，进而每次循环时使一个模糊子集出栈，根据模糊子集及隶属度不同，其计算过程如下:

(ⅰ) 若出栈状态 $C_{p,q}^{\tilde{S}}$ 满足下式:

$$\mu(C_{p,q}^{\tilde{S}})==\max_{p\leqslant v<q}\mu(C_{p,v}^{\tilde{L}})\circ\mu(C_{v+1,q}^{\tilde{S}})\circ\mu_U(U_3),\tag{35}$$

则将 $C_{v+1,q}^{\tilde{S}}$ 及 $C_{p,v}^{\tilde{L}}$ 入栈，并将决策 U_3 加入最优决策序列.

否则若出栈状态 $C_{p,q}^{\tilde{S}}$ 满足下式:

$$\mu(C_{p,q}^{\tilde{S}})==\mu(C_{p,q}^{\tilde{L}})\circ\mu_U(U_4),\tag{36}$$

则将 $C_{p,q}^{\tilde{L}}$ 入栈，并将决策 U_4 加入最优决策序列.

(ⅱ) 若出栈状态 $C_{p,q}^{\tilde{F}}$ 满足下式:

$$\mu(C_{p,q}^{\tilde{F}})==\max_{p\leqslant v<q}\mu(C_{p,v}^{\tilde{L}})\circ\mu(C_{v+1,q}^{\tilde{S}})\circ\mu_U(U_5),\tag{37}$$

则将 $C_{v+1,q}^{\tilde{S}}$ 及 $C_{p,v}^{\tilde{L}}$ 入栈，并将决策 U_5 加入最优决策序列.

否则若出栈状态 $C_{p,q}^{\tilde{F}}$ 满足下式:

$$\mu(C_{p,q}^{\tilde{F}})==\mu(C_{p+1,q-1}^{\tilde{F}})\circ\mu_{Pair}(C_{p,q}^{\tilde{L}})\circ\mu_U(U_6),\tag{38}$$

则将 $C_{p+1,q-1}^{\tilde{F}}$ 入栈，并将决策 U_6 加入最优决策序列.

(iii) 若出栈状态 $C_{p,q}^{\tilde{L}}(p>q)$ 满足下式:

$$\mu(C_{p,q}^{\tilde{L}}) == \mu(C_{p+1,q-1}^{\tilde{F}}) \circ \mu_{Pair}(C_{p,q}^{\tilde{L}}) \circ \mu_U(U_1), \tag{39}$$

则将 $C_{p+1,q-1}^{\tilde{F}}$ 入栈, 并将决策 U_1 加入最优决策序列.

若出栈状态为 $C_{p,q}^{\tilde{L}}(p=q)$, 则将决策 U_2 加入最优决策序列.

1.4　清晰化过程

由于使用了模糊分割及模糊子集, 模糊动态规划过程计算出的最优模糊策略/最优模糊路径通常对应于最优的模糊化 RNA 结构, 其中包含多个 RNA 二级结构. 在将最优模糊路径清晰化为二级结构的过程中, 可根据模糊子集的隶属函数来确定相应的最优与次优二级结构. 对于RNA的最优二级结构预测, 可将最优模糊路径中各个模糊子集的"脊点"进行连接, 而构成所谓的"脊线" (ridge line), 此时对应于该清晰化"脊线"的二级结构即为最优二级结构. 进一步地, 次大隶属度路径对应的二级结构即是次优二级结构. 显然, 次优二级结构不止一个. 不失一般性, 我们这里假设对状态空间进行模糊分割的三角隶属函数对称, 且宽度均为 2δ.

对于 BJK 模糊模型结构, 根据当前模糊状态 $C_{p,q}^{r}$ 所在模糊状态子空间 $r \in \{\tilde{L}, \tilde{S}, \tilde{F}\}$ 的不同, 由"脊线"确定 RNA 最优二级结构预测的具体算法可给出如下:

- 若当前模糊状态子空间 $r = \tilde{L}$.

如果应用于当前模糊子集 $C_{p,q}^{\tilde{L}}$ 的规则是 U_1, 继而转移到模糊子集 $C_{p+1,q-1}^{\tilde{F}}$, 则根据二者之间跳过的碱基

$$B(R(p)+1), B(R(p)+2), \cdots, B(R(p)+\delta-1) = B(R(p+1)-1)$$

及

$$B(R(q)-\delta+1) = B(R(q-1)+1), \cdots, B(R(q)-2), B(R(q)-1)$$

是否可以彼此配对, 即

$$(R(p)+1, R(q)-1), (R(p)+2, R(q)-2), \cdots, (R(p)+\delta-1, R(q)-\delta+1)$$

是否可以形成Watson-Crick碱基对, 来决定最优二级结构中的配对结果. 其中 $R(v)$ 为模糊状态子空间一维上第 v 个模糊子集脊点的碱基位置标号.

当 $p=q$ 时, 应用的规则为 U_2 并产生 $Single(C_{p,p}^{\tilde{L}})$, 则没有碱基被跳过.

- 若当前模糊状态子空间 $r = \tilde{S}$.

如果应用于当前模糊子集 $C_{p,q}^{\tilde{S}}$ 的规则是 U_3, 继而转移到模糊子集 $C_{p,v}^{\tilde{L}}$ 与 $C_{v+1,q}^{\tilde{S}}$, 则这两个模糊子集脊点之间的碱基

$$B(R(v)+1), B(R(v)+2), \cdots, B(R(v)+\delta-1) = B(R(v+1)-1)$$

应为不配对的.

如果应用的规则为 U_4 且转移到 $C_{p,q}^{\tilde{L}}$, 则没有碱基被跳过, 因此不需要作特别的处理.

- 若当前模糊状态子空间 $r = \tilde{F}$.

如果应用于当前模糊子集 $C_{p,q}^{\tilde{F}}$ 的规则是 U_5 且转移到模糊子集 $C_{p,v}^{\tilde{L}}$ 与 $C_{v+1,q}^{\tilde{S}}$, 则清晰化的方法与规则 U_3 的清晰化方法相同.

如果规则为U_6，则应转移到$C^{\tilde{F}}_{p+1,q-1}$，此时与规则U_1的清晰化方法相同.

图 6 给出了 BJK 模糊模型结构下 RNA 最优二级结构的一个推导示例. 图 7 是与该推导相对应的最优二级结构.

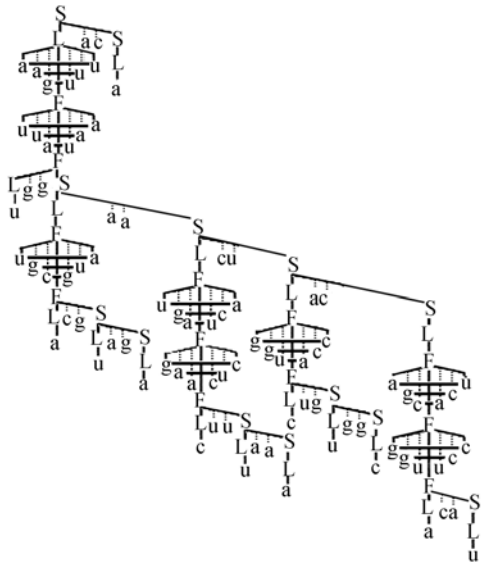

图 6 以 tRNA 及 $\delta^r_{P(q)} \equiv \delta = 3$ 为例的模糊状态转移过程

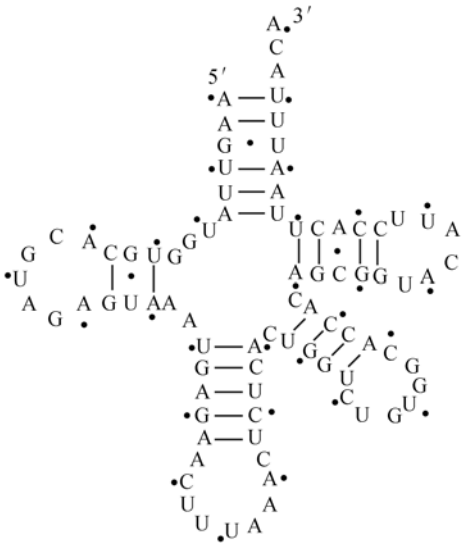

图 7 与图 6 对应的 tRNA 最优二级结构

旁边以点标记的碱基是最优模糊路径中隶属度最大的脊点

2 实验结果与分析

本文实现了上述BJK模糊模型结构. 基于包括tRNA和rRNA在内的多个结构RNA数据集, 我们将本文提出的模糊模型方法与基于最小自由能的mfold折叠工具[3]及在文献[7]中实现的基于SCFG的BJK文法模型进行了比较研究. 前已指出, mfold是Zuker等人开发的目前最常用的RNA二级结构预测程序, 而本文的BJK模糊模型借用了BJK文法模型的产生式规则与参数训练方法, 因此两者具有可比之处. 实验结果表明, 在大多数情形下本文的方法都具有更高的预测精度.

2.1 构造训练与测试样本集

实验中我们使用了tRNA数据集及一部分rRNA数据集. 具有已知二级结构的 843 条tRNA序列, 取自 EMBL 数据库[19], 包括了 virus, archaea, eubacteria, cyanellae, cytoplasm, 以及mitochondria在内的多种tRNA序列及其二级结构数据. 对tRNA, 我们采用类似于文献[20]的方式构造了两个训练样本集. 第 1 个训练样本集MT100, 从mitochondria类中随机地抽取了 100个tRNA序列. 第 2 个训练样本集MT10CY10, 分别从cytoplasm类及mitochondria类中随机地抽取了各 10 个tRNA作为训练集. 测试样本集则由各类tRNA中的其他数据组成.

对rRNA, 我们使用了tmRNA数据集[21], 该数据集中共有 95 条具有已知二级结构的rRNA序列. 我们随机地抽取了其中的40条序列作为训练集, 其余的作为测试集. 每次实验中使用的是通过不同的随机抽取得到的训练及测试样本集.

由于目前采用的数据集中序列长度普遍较短, 即 tRNA 为 51~93 个碱基, tmRNA 为 189~425 个碱基, 并不存在严重的计算复杂性问题, 因此实际实现时, 我们的模型使用了 δ 为零的特例. 换句话说, 在进行状态空间的模糊分割时, 上三角矩阵中的每个点均对应一个单点模糊子集, 因此相应模糊模型中的最优策略/最优路径也就唯一地对应一个最优的二级结构.

2.2　实验结果及比较研究

为分析并比较 RNA 二级结构的预测精度, 我们采用了二级结构预测中评价预测精度常用的灵敏度(Sensitivity)与特异度(Specificity)指标, 即

$$\begin{aligned}\text{Sensitivity} &= \frac{TP}{TP+FN},\\ \text{Specificity} &= \frac{TP}{TP+FP},\end{aligned} \tag{40}$$

其中, TP(true positive, 真阳性)表示在所有预测得到的碱基配对中, 真实配对的碱基对个数, 而 FN (false negative, 假阴性)及 FP(false positive, 假阳性)分别表示预测不配对但真实配对的碱基对个数, 以及预测配对但真实不配对的碱基对个数. 因此灵敏度指标表示的是所有真实配对碱基中被预测正确的百分比, 特异度指标则表示了所有预测配对碱基中被预测正确的百分比.

在本文的比较实验中, 使用了 mfold 的最新版本 RNA3.0(包括其能量规则及默认参数), 分别分析了利用该工具预测出的最小自由能结构的预测精度, 以及 5%最小自由能范围内最优与多个次优结构的最优预测精度(灵敏度与特异度之和最大), 并与我们的方法进行了比较研究.

文献[7]对目前 9 种有关RNA二级结构预测的SCFG算法进行了系统的比较研究, 指出其中的BJK文法模型是目前综合性能最优的SCFG方法. 通过使用相同的训练及测试样本集, 将本文提出的BJK模糊模型与之进行了比较, 给出了相应的预测精度评估. 需要指出的是, 与文献[7]中 BJK 文法的实现程序略有不同, 为了统一起见, 我们在实验中将发夹环的最小长度(HLEN)取为 3.

表 1~3 分别给出了利用 tRNA 数据集时 mfold 折叠工具($mfold$)的预测精度, 以及 SCFG 的BJK 文法(BJK_G)和本文方法(BJK_F)在不同训练集时的预测精度比较结果. 表中的列表示不同的预测方法及精度指标, 而行表示用来测试预测性能的测试样本数据集. 由于 mfold 工具不需要训练集, 因此表 1 中 $mfold_{opt}$ 列中为该工具预测得到的最小自由能结构的预测精度, $mfold_{5\%subopt}$ 列中为 5%最小自由能范围内得到的多个二级结构的最优预测精度. 而基于 SCFG 模型的 BJK 文法及本文 BJK 模糊模型方法是训练集相关的, 以表 2 中的第 1 行第 1 列的 83.83 为例, 它表示使用 MT10CY10 数据集来训练参数, 继而利用 ARCHAE 数据集进行测试时, BJK_G 文法模型所得到的预测精度. 容易看出, 对于 tRNA 数据集, 尽管结构相对保守, 但 mfold 的预测精度很低, 即使是最小能量值5%范围内多个预测结果的最优值也不甚理想, 因此无论是与 mfold 折叠工具, 还是与 SCFG 的 BJK 文法相比较, 本文方法的预测精度均有提高.

表 4 和 5 中列出的是使用 tmRNA 数据集时, mfold 折叠工具, 以及基于 SCFG 的 BJK 文法和本文的模糊模型方法的预测精度比较结果. 实验证明, 对于 tmRNA 数据集, 相较于由 mfold 预测得到的最小自由能结构与由本文方法预测的结果, 两者的灵敏度指标基本一致, 但后者的

表 1 对 tRNA 数据集和 mfold 方法的预测结果

数据集	$mfold_{opt}$					$mfold_{5\%subopt}$				
	Sens.	Spec.	TP	FP	FN	Sens.	Spec.	TP	FP	FN
ARCHAE	65.10	58.53	2854	2022	1530	78.16	70.73	3392	1404	948
CY	59.70	55.56	5812	4648	3924	73.82	70.36	7156	3014	2538
CYANELCHLORO	72.19	65.99	5616	2894	2164	80.64	75.18	6240	2060	1498
EUBACT	68.88	61.80	5892	3642	2662	79.87	72.98	6800	2518	1714
VIRUS	62.96	58.51	612	434	360	74.89	72.11	698	270	234
MT	60.43	60.20	10480	6930	6862	69.56	69.29	12032	5340	5266
PARTIII	65.75	67.67	1390	664	724	73.48	74.63	1524	518	550
合计	64.18	60.60	32656	21234	18226	74.80	71.45	37842	15124	12748

表 2 对 tRNA 数据集, BJK_G 与 BJK_F 方法在使用训练集 MT10CY10 时的预测精度指标结果

数据集	BJK_G					BJK_F				
	Sens.	Spec.	TP	FP	FN	Sens.	Spec.	TP	FP	FN
ARCHAE	83.83	75.51	3675	1192	709	90.47	88.88	3966	496	418
CY	81.89	77.89	7973	2263	1763	93.59	91.93	9112	800	624
CYANELCHLORO	84.56	79.78	6579	1667	1201	89.25	88.73	6944	882	836
EUBACT	90.15	83.94	7711	1475	843	90.37	88.48	7730	1006	824
VIRUS	82.00	77.91	797	226	175	89.71	88.80	872	110	100
MT	78.14	76.46	13551	4172	3791	88.51	90.21	15350	1666	1992
PARTIII	77.11	75.74	1630	522	484	84.48	87.12	1786	264	328
合计	82.38	78.45	41916	11517	8966	89.93	89.75	45760	5224	5122

表 3 对 tRNA 数据集, BJK_G 与 BJK_F 方法在使用训练集 MT100 时的预测精度指标结果

数据集	BJK_G					BJK_F				
	Sens.	Spec.	TP	FP	FN	Sens.	Spec.	TP	FP	FN
ARCHAE	81.73	74.09	3583	1253	801	89.42	88.25	3920	522	464
CY	81.59	77.00	7944	2373	1792	92.48	92.07	9004	776	732
CYANELCHLORO	83.89	78.61	6527	1776	1253	89.59	90.71	6970	714	810
EUBACT	87.58	81.54	7492	1696	1062	89.20	89.09	7630	934	924
VIRUS	79.63	74.21	774	269	198	84.16	88.72	818	104	154
MT	78.82	76.76	13669	4138	3673	87.46	90.45	15168	1602	2174
PARTIII	77.39	76.77	1636	495	478	79.56	84.02	1682	320	432
合计	81.81	77.62	41625	12000	9257	88.82	90.09	45192	4972	5690

表 4 对 tmRNA 数据集, 7 次独立实验中 mfold 方法预测精度结果

数据集	$mfold_{opt}$					$mfold_{5\%subopt}$				
	Sens.	Spec.	TP	FP	FN	Sens.	Spec.	TP	FP	FN
1	44.99	30.96	3648	8136	4460	58.83	41.78	4678	6518	3274
2	43.76	29.62	3382	8036	4346	57.12	40.01	4326	6486	3248
3	42.51	28.70	3410	8470	4612	56.94	39.45	4486	6886	3392
4	41.00	27.50	3192	8414	4594	54.62	37.74	4164	6868	3460
5	43.80	29.99	3464	8086	4444	59.73	42.42	4642	6302	3130
6	45.69	30.92	3488	7792	4146	58.35	40.68	4362	6362	3114
7	44.20	30.85	3496	7836	4414	59.15	42.67	4588	6164	3168
合计	43.71	29.78	24080	56770	31016	57.83	40.67	31246	45586	22786

表 5　对 tmRNA 数据集, 7 次独立实验中 BJK_G 及 BJK_F 方法预测精度结果

数据集	BJK_G					BJK_F				
	Sens.	Spec.	TP	FP	FN	Sens.	Spec.	TP	FP	FN
1	39.44	37.79	3198	5265	4910	43.54	51.55	3530	3318	4578
2	37.44	34.35	2893	5528	4835	43.19	49.42	3338	3416	4390
3	39.78	37.52	3191	5313	4831	43.38	48.63	3480	3676	4542
4	37.48	35.69	2918	5258	4868	40.77	46.66	3174	3628	4612
5	38.76	36.14	3065	5415	4843	45.83	52.16	3624	3324	4284
6	38.66	36.33	2951	5172	4683	40.53	48.65	3094	3266	4540
7	39.63	38.14	3135	5085	4775	43.82	51.81	3466	3224	4444
合计	38.75	36.57	21351	37036	33745	43.03	49.85	23706	23852	31390

特异度指标则提高较多. 即使与由 mfold 方法在最小自由能 5%范围内预测得到的最好结果相比, 尽管由本文方法给出的灵敏度指标略低, 但特异度指标仍较高. 而与 SCFG 的 BJK 文法模型相比, 模糊模型方法的预测精度已有了很大程度的提高.

此外, 本文对不同类型的 RNA 结构预测进行了实验, 即用两种不同类型的 RNA 分别进行训练和测试. 这里选用上述实验中的 tmRNA 数据集作为训练数据集, tRNA 数据集作为测试数据集. 表 6 中列出的是在这样的数据集下, 基于 SCFG 模型的 BJK 文法及本文的 BJK 模糊模型方法的预测精度. 与表 2 及表 3 中使用 tRNA 数据集作训练集时的预测精度相比, 由于基于 SCFG 模型的 BJK 文法及本文的 BJK 模糊模型方法均是基于机器学习的思想, 因此当使用不同来源的 RNA 数据时, 二者的预测精度都有所降低. 同时由于本文的方法更多地融入了保守信息, 因此在部分数据集上, 本文方法的灵敏度指标略低. 但在比较难预测的 PartIII tRNA 数据集上要比 BJK 文法预测精度高. 在表 7 中, 我们总结了对于 tRNA 数据集, mfold 方法的平均特异度指标, 及使用 tmRNA 数据集训练参数时, 基于 SCFG 模型的 BJK 文法和本文的 BJK 模糊模型方法的平均特异度指标. 实验结果表明: mfold 模型中最优自由能结构的平均特异性指标为 60.60%, 5%次优自由能范围内得到的多个二级结构的平均特异性指标为 71.45%, BJKG 模型为 71.42%, 而本文的 BJKF 模型, 其平均特异性指标为 74.14%, 明显高于其他方法.

表 6　当利用tmRNA数据集作为训练数据集和以tRNA数据集作为测试数据集时, 本文的模糊模型方法(BJK_F)与文献[7]中实现的基于SCFG模型的BJK文法(BJK_G)的预测精度比较

数据集	BJK_G					BJK_F				
	Sens.	Spec.	TP	FP	FN	Sens.	Spec.	TP	FP	FN
ARCHAE	72.03	74.83	3158	1062	1226	64.19	75.00	2814	938	1570
CY	61.12	68.90	5951	2686	3785	51.95	69.52	5058	2218	4678
CYANELCHLORO	67.33	78.78	5238	1411	2542	59.18	79.85	4604	1162	3176
EUBACT	64.38	73.46	5507	1990	3047	47.51	66.06	4064	2088	4490
VIRUS	59.67	77.54	580	168	392	39.92	61.20	388	246	584
MT	35.55	65.18	6165	3293	11177	41.76	79.22	7242	1900	10100
PARTIII	28.67	68.47	606	279	1508	37.75	83.47	798	158	1316
合计	53.47	71.42	27205	10889	23677	49.07	74.14	24968	8710	25914

表 7　当利用tmRNA数据集作为训练数据集和以tRNA数据集作为测试数据集时, 本文的模糊模型方法(BJK_F), 与文献[7]中实现的基于SCFG模型的BJK文法(BJK_G)及mfold方法(opt与 5%$subopt$)的平均特异度指标比较

$mfold_{opt}$	$mfold_{5\%subopt}$	BJK_G	BJK_F
60.60	71.45	71.42	74.14

3 讨论

与基于最小自由能的 Zuker 算法以及 SCFG 算法类似, 本文提出的模糊模型中, 利用模糊动态规划算法, 同样通过迭代填充计算及回溯过程确定相应的最优策略. 但不同的是, 我们的RNA二级结构预测模型是一个模糊系统, 其中的模糊目标及模糊决策等都是使用隶属度表示的, 因此我们的方法本质上是基于模糊集合理论的, 与 Zuker 算法的确定性系统以及 SCFG 的随机系统具有本质的不同.

特别地, 文中给出的具体的BJK模糊模型结构与SCFG的BJK文法类似. 在本文的模糊模型中, 采用了与BJK文法中产生式规则类似的模糊状态转移规则, 但由于利用了模糊子集的优势, RNA二级结构的保守信息可非常自然而容易地引入到预测过程中, 从而得到了比BJK文法更佳的性能. 文献[7]指出, Knudsen/Hein的BJK文法模型的性能在该文比较的所有 9 种SCFG模型中, 仅次于对其进行了一阶Markov链扩展的 $G6^S$ 文法模型. 但 $G6^S$ 模型增加了stack约束, 大大增加了文法的复杂性, 因此BJK文法模型本身可以说是综合性能最优的SCFG方法.

由于本文的模糊模型与 SCFG 的文法模型使用的都是基于样本集训练的机器学习思想, 因此在使用不同类别的 RNA 数据集时, 二者的预测精度有所下降. 在后续工作中, 可通过对不同类别的 RNA 分别建立模型, 并添加根据 RNA 长度等特征对待预测序列进行简单分类的预处理, 则可望解决此一局限.

需要着重指出的是, 类似于最小自由能方法可以采用各种具有实际含义的自由能经验计算公式, 以及 SCFG 方法可以使用各种文法规则一样, 本文给出的一般模糊模型并不局限于特定的 BJK 模糊模型结构预测方法. 文中给出的模糊状态分割与模糊状态转移规则的具体定义, 可以根据不同的问题来进行不同的设计与发展.

相较于其他确定性及随机模型, 模糊推理系统的优点在于能够有效地引入专家的主观经验与定性知识. 本文发展的模糊模型方法, 通过将 RNA 二级结构预测过程视为模糊推断过程, 并利用模糊动态规划算法计算相应的最优模糊策略, 从而能够将训练集中 RNA 二级结构的更多保守特征引入模糊模型的建立中, 进而可有效地提高预测精度. 例如, 在本文的预测例子中, 我们使用了不配对碱基的相对位置作为模糊目标, 通过计算各个碱基的模糊目标隶属度, 使得相应的预测精度有一定程度的提高. 显然, 模糊模型的提出, 相当方便各种专家定性知识与经验的有效利用, 这是这种方法相对于确定性模型与随机模型的最大优势.

4 结论

本文提出了一种基于模糊模型的 RNA 二级结构预测的新方法. 文中系统地给出了该模型的基本概念、模型结构与参数的确定, 以及基于模糊动态规划的迭代填充与相应的回溯过程, 并进一步给出了通过已计算出的最优模糊策略, 计算 RNA 最优二级结构的具体过程. 本文的研究表明, 通过状态空间的模糊分割, 能够有效地降低计算复杂性, 并可同时预测出最优与多个次优二级结构. 本文提出的模糊模型的另一个明显优点是, 通过增加模糊目标集隶属函数, 可将定性的先验知识引入预测问题中. 我们将一个具体的 BJK 模糊模型结构实际应用于 tRNA 与 tmRNA 的数据集中, 并进行了相应的比较研究. 实验结果表明了所提模型的可行性, 相应的预测精度得到了进一步的提高. 由于我们的模型具有很强的可扩展性, 未来的工作包括对模糊模型进行深入的发展, 以便融入更多的启发式生物学知识, 进一步提高二级结构预测的

精度, 并对长度超过 1000 碱基的较长 RNA 序列, 以及对保守性较差的 ncRNA 序列, 乃至对 RNA 的高级结构(如伪结)提供可行的解决方案.

参　考　文　献

1 Nussinov R, Pieczenik G, Griggs J R, et al. Algorithms for loop matchings. SIAM J Appl Math, 1978, 35: 68—82

2 Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res, 1981, 9(1): 133—148[DOI]

3 Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res, 2003, 31(13): 3406—3415[DOI]

4 Hofacker I L, Fontana W, Stadler P F, et al. Fast folding and comparison of RNA secondary structures. Mon Chem, 1994, 125: 167—188[DOI]

5 Eddy S R, Durbin R. RNA sequence analysis using covariance models. Nucleic Acids Res, 1994, 22(11): 2079—2088[DOI]

6 Kundsen B, Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics, 1999, 15(6): 446—454[DOI]

7 Dowell R D, Eddy S R. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. BMC Bioinformatics, 2004, 5: 71—99[DOI]

8 Ding Y. Statistical and Bayesian approaches to RNA secondary structure prediction. RNA, 2006, 12: 323—331[DOI]

9 McCaskill J S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers, 1990, 29(6-7): 1105—1119[DOI]

10 Hu Y J. GPRM: A genetic programming approach to finding common RNA secondary structure elements. Nucleic Acids Res, 2003, 31: 3446—3449[DOI]

11 Steeg E W. Neural network algorithms for RNA secondary structure prediction. Technical Report CRG-TR-90-4, University of Toronto Computer Science Dept. 1990

12 Zadeh L A. Fuzzy sets. Inform Contr, 1965, 8: 338—353

13 Blankenbecler R, Ohlsson M, Peterson C, et al. Matching protein structures with fuzzy alignments. Proc Natl Acad Sci USA, 2003, 100(21): 11936—11940[DOI]

14 Jacob E, Sasikumar R, Nair K N R. A fuzzy guided genetic algorithm for operon prediction. Bioinformatics, 2005, 21: 1403—1407[DOI]

15 Kacprzyk J, Esogbue A O. Fuzzy dynamic programming: Main developments and applications. Fuzzy Set Syst, 1996, 81: 31—45[DOI]

16 Rivas E, Eddy S R. A dynamic programming algorithm for RNA structure prediction including pseudoknots. J Mol Biol, 1999, 285(5): 2053—2068[DOI]

17 Bilgic T, Türksen I B. Measurement of membership functions: Theoretical and empirical work, Chapter 3. In: Dubois D, Prade H, eds. Handbook of Fuzzy Sets and Systems Vol 1, Fundamentals of Fuzzy Sets. Norwell, MA: Kluwer Academic, 1999. 195—232

18 Durbin R, Eddy S R, Krogh A, et al. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. London: Cambridge University Press, 1998. 299—325

19 Steinberg S, Misch A, Sprinzl M. Compilation of tRNA sequences and sequences of tRNA genes. Proc Natl Acad Sci USA, 1993, 21(13): 3011—3015

20 Sakakibara Y, Brown M, Hughey R, et al. Stochastic context-free grammars for tRNA modeling. Nucleic Acids Res, 1994, 22(23): 5112—5120[DOI]

21 Zwieb C, Gorodkin J, Knudsen B, et al. tmRDB (tmRNA Database). Nucleic Acids Res, 2003, 31(1): 446—447[DOI]