Институт прикладной математики
имени М.В. Келдыша
Российской Академии Наук

Е.Е. Козырева, Е.И. Кугушев, А.В. Майков

Методика математического моделирования процесса образования вторичной структуры РНК.

Москва 2001

Аннотация. Описывается универсальная методика ускорения вычислений в задачах математического моделирования процесса структурообразования биологических макромолекул, синтезируемых в

живой клетке (нуклеиновые кислоты, белки). Развитие таких процессов определяется соотношением скорости удлинения растущей молекулярной цепи и скорости распада и образования структурных связей в ней. Вычислительная сложность моделирования здесь оценивается как N^5, где N – полная длина молекулярной цепи. Задача сводится к многократному построению локально-минимального пути на ориентированном динамически меняющемся графе, описывающем возможные межструктурные переходы. Множество таких переходов для каждой вершины графа монотонно растет вместе с ростом молекулярной цепи. Это свойство позволяет снизить вычислительные затраты до N^4. Проводится доказательство корректности методики, анализируются ее возможные варианты. Методика использовалась при математическом моделировании и изучении прицесса образования вторичных структур рибонуклеиновых кмслот (РНК) различных типов.

Ключевые слова: математическое моделирование, вычислительная сложность, вторичная структура РНК.

Abstract. Universal method is described of calculations acceleration in tasks of mathematical simulation of biological macromolecules structures formation process in living cell (nucleic acids, proteins). The way of such processes running is determined by the ratio of velocity of molecular chain lengthening and the velocity of structure bonds breaking and arising. Computational complication of simulation in this task is estimated as N^5, where N – is full length of molecular chain. The task is reduced to multiple calculations of locally optimal path on an oriented and dynamically changed graph, which describes possible inter-structural transitions. The set of such transitions in every graph vertex monotonously growth according to the growth of molecular chain. This property allows us to decrease computational expenditures up to N^4. The proof of method correctness is given. The method possible variants are analyzed. This method has used in mathematical simulation and investigation of process of various ribonucleic acids (RNA) secondary structures formation.

Key words: mathematical simulation, computational complication, secondary RNA structure.

Содержание

1. Введение

Определение геометрической структуры таких биологических макромолекул как белки и нуклеиновые кислоты (их геометрической формы, топологии межатомных связей) является фундаментальной проблемой молекулярной биологии. Эти молекулы представляют собой длиныые цепочки однотипных элементов – нуклеотидов для нуклеиновых кислот или аминокислот для белков. Молекулярная цепь образуется и растет путем постепенного добавления к ней этих однотипных элементов в ходе соответствующего процесса, происходящего в живой клетке – трансляции (для белков), транскрипции или репликации (для нуклеиновых кислот). Функциональные свойства биологической макромолекулы в значительной степени определяются той формой (или структуой), в которую сворачивается ее молекулярная цепь.

В связи с большим количеством (тысячи) известных к настоящему времени различных биологических макромолекул и сложностью прямых физических и биохимических методов распознавания их структур особый интерес представляет разработка математических моделей структурообразования биологических макромолекул и определение структур реальных макромолекул методом математического моделирования.

Биологическая макромолекула, сворачиваясь, приобретает структуру, которой соответствует локальный минимум ее свободной энергии. Линейные (продольные) связи вдоль молекулярной цепи достаточно сильны, поэтому в ходе структурообразования молекулярная цепь не разрывается, молекула принимает энергетически локально - минимальную форму за счет изгибов молекулярной цепи и образования поперечных, более слабых, связей.

Гибкость молекулярной цепи и способность образовывать поперечные связи определяется ее составом или, как принято говорить, ее первичной структурой. Однотипные элементы, составяющие молекулярную цепь могут отличаться друг от друга. Молекулярная цепь рибонуклеиновой кислоты (РНК) составляется из нуклеотидов четырех видов, то же имеет место и для дезоксироибонуклеиновых кислот (ДНК). Белковая молекулярная цепь составляется из аминокислот двадцати различных видов. Первичная структура молекулы определяется тем, какие и в какой последовательности однотипные элементы составляют молекулярную цепь.

Целью математического моделирования процесса структурообразования биологических макромолекул является определение пространственной структуры в которую сворачивается молекулярная цепь по ее первичной структуре. В данной работе рассматриваются методы такого математического моделирования для процесса структурообразования РНК. Однако, в силу общности основных черт механизмов роста цепи биологических макромолекул и их структур, эти методы могут найти применение применимы и для математического моделирования процессов структурообразования биологических макромолекул других типов.

Число потенциально возможных устойчивых структур для биологической макромолекулы велико. Сворачивание макромолекулы в ту или иную структуру зависит не только от ее первичной структуры, но и от условий, в которых образуется молекулярная цепь. Математическая модель процесса структурообразования должна учитывать эти условия. Характерной особенностью синтеза нуклеиновых кислот и белков в живой клетке является то, что макромолекула образуется постепенно; молекулярная цепь ее растет с некоторой конечной скоростью (скорость элонгации). При этом на уже синтезированном участке молекулы постоянно идут процессы структурообразованияи и, как показали наши исследования для РНК [1 - 3], развитие структуры определяется соотношением скорости элонгации и скорости структурообразования (т.е. образования новых и распада старых поперечных структурных связей в молекуле). Численную меру этого соотношения мы назвали периодом стуктуризации T. В нашей модели величина T определяется, как число нуклеотидов, на которые удлиняется молекулярная цепь за время структуризации уже образовавшегося участка молекулярной цепи. В связи с тем, что данная молекула РНК может порождаться как часть более длинной молекулярной цепи (т.н. первичного транскрипта), процесс ее структурообразования может начинаться не сразу (с появления первого нуклеотида), а тогда, когда молекулярная цепь уже имеет в своем составе некоторое число (L_0) нуклеотидов.

При неизвестных заранее значениях T и L_0 вычислительная сложность математического моделирования процесса структурообразования оценивается как N^5, где N – полная длина молекулярной цепи (число нуклеотидов в ней). В данной работе описывается универсальная методика ускорения вычислений. Задача сводится к многократному построению локально-минимального пути на ориентированном динамически меняющемся графе, описывающем возможные межструктурные переходы. Множество таких переходов для каждой вершины графа монотонно растет вместе с ростом молекулярной цепи. Это свойство позволяет снизить вычислительные затраты до N^4.

Проводится доказательство корректности методики, анализируются ее возможные варианты. Методика использовалась при математическом моделировании и изучении прицесса образования вторичных структур рибонуклеиновых кмслот (РНК) различных типов.

2. Вторичные структуры РНК и моделирование их образования.

Как уже говорилось, молекулярная цепь рибонуклеиновой кислоты (РНК) составляется из нуклеотидов четырёх видов. Это адениновая, гуаниловая, цитидиловая и уридиловая кислоты. Основания этих кислот обозначаются символами: A – аденин, G – гуанин, C – цитозин, U – урацил. Первичная структура РНК описывает последовательность нуклеотидов в молекулярной цепи; её можно представить как строку текста составленного из указанных четырёх букв. Среди поперечных связей, которые образуются при сворачивании молекулярной цепи в пространственную структуру, наиболее сильными являются водородные Уотсон-Криковские связи, возникающие между основаниями некоторых нуклеотидов. Такие связи могут возникать между A и U, между C и G и между G и U. Пары оснований A–G и C–G называются классическими, комплементарными; пара G–U – неклассической. Пары оснований, между которыми возникли Уотсон-Криковские связи, называются спаренными. Связь C–G – самая сильная, связь G–U – самая слабая и может возникать только при некоторых условиях.

В естественной молекуле РНК часть оснований находится в спаренном, а часть в неспаренном, свободном состоянии. Это определяет т.н. вторичную структуру РНК и вторичные (Уотсон-Криковские) взаимодействия в молекуле. При сворачивании молекулы РНК в пространстве могут возникать добавочные (более слабые) связи между её атомами. Это определяет т.н. третичную структуру РНК и третичные взаимодействия в молекуле. Первичная структура у молекулы РНК одна, а возможных вторичных (и третичных) структур много.

Формально, вторичная структура РНК – это описание всех спаренных и свободных оснований в молекулярной цепи. Пронумеруем последовательно нуклеотиды в молекулярной цепи так, что нуклеотид, присоединяющийся к растущей цепи позже, имеет больший номер. Отрезок $[n_1, n_2]$ свободных оснований в молекулярной цепи называется однонитевым. Отрезки $[n_1, n_2]$ и $[n_3, n_4]$ спаренных оснований так, что n_1 спарен с n_4, $n_1 + 1$ с $n_4 - 1$,…, n_2 с n_3 образуют двухнитевый или двуспиральный участок во вторичной структуре РНК. Однонитевые участки также называются петлями, а двухнитевые – стеблями. Стебель можно представить себе, как участок винтовой лестницы, где ступеньки

– это поперечные, Уотсон-Криковские связи. Длиной стебля называется чсло пар оснований в нём: $n_2 - n_1 + 1 = n_4 - n_3 + 1$.

Таким образом , вторичная структура РНК – это совокупность стеблей и петель. Других элементарных составляющих во вторичных структурах реальных РНК к настоящему времени не обнаружено.

Общее число возможных стеблей для данной молекулы РНК оценивается как N^2, где N – число нуклеотидов в цепи РНК. Получим эту оценку для случая, когда разрешено только классическое спаривание C–G и A–U (случай с G–U рассматривается аналогично). Вероятность того, что наугад выбранная пара оснований окажется классической коплементарной равна ¼ (четыре пары A–U, U–A, C–G, G–C из 16 возможных комбинаций). Число различных пар оснований в цепи РНК равно $\frac{N(N-1)}{2}$. Поэтому число стеблей длины 1 оценивается как $Q_1 = \frac{N(N-1)}{8}$. Вероятность того, что наугад выбранные k пар оснований все окажутся комплементарными равна $(¼)^k$. Число четверок (n_1, n_2, n_3, n_4) таких, что $n_1 \le n_2 < n_3 \le n_4$ и $n_2 - n_1 + 1 = n_4 - n_3 + 1 = k$, не превосходит $\frac{N(N-1)}{2}$. Поэтому число стеблей длины k в среднем не превосходит $Q_k = \frac{N(N-1)}{8}\left(\frac{1}{4}\right)^k$. Поскольку $\frac{1}{4} + \frac{1}{16} + \ldots + \frac{1}{4^k} + \ldots = \frac{1}{3}$, то среднее число стеблей m можно теперь оценить так:

$$\frac{1}{4}\frac{N(N-1)}{2} \le m \le \frac{1}{3}\frac{N(N-1)}{2},$$

что и требовалось показать.

Отметим, что вероятности появления нуклеотидов любого типа (или их цепочек) здесь предполагались равными. Обратный случай рассматривается аналогично.

В работах [1 – 4] моделируетсяя процесс образования вторичной структуры РНК, развивающийся по пути локальной минимизации свободной энергии молекулы. Процесс является пошаговым и отражает последовательный рост молекулярной цепи РНК. На первом шаге длина молекулярной цепи (т.е. число нуклеотидов в ней) полагается равной некоторой величине L_0, называемой начальной длиной. На каждом шаге сначала моделируется структуризация, т.е. процесс образования вторичной структуры РНК в пределах уже выросшего участка молекулярной цепи. ”Выращивание” структуры производится повторяющимся добавлением к ней новых стеблей так, чтобы обеспечивался минимальный отрицательный прирост свободной

энергии молекулы. По окончании структуризации молекулярная цепь удлиняется на некоторое постоянное число нуклеотидов T и происходит переход к следующему шагу. Параметр T называется периодом структуризации. Он равен величине удлинения молекулярной цепи за время её структуризации и, таким образом, характеризует отношение скоростей двух процессов – элонгации (роста молекулярной цепи) и структуризации (роста вторичной структуры).

Моделирование проводилось для четырёх опубликованных моделей свободной энергии вторичной структуры РНК: G_1 – [5], G_2 – [6], G_3 – [7], G_4 – [8]. Модель свободной энергии – это семейство термодинамических параметров, позволяющих вычислить приращение свободной энергии ΔG молекулы по её нуклеотидной последовательности. Во всех этих моделях свободная энергия вторичной структуры РНК складывается из суммы свободных энергий её элементов – свободной энергии стеблей и свободной энергии образования (инициации) петель. Параметры в этих моделях были определены экспериментально (в простейшей модели G_1 этого не требовалось).

Исследования проводились в полной области целочисленных значений параметров T, L_0. Для каждой молекулы РНК моделировалось $\frac{N(N+1)}{2}$ вариантов процесса структурообразования (N – полная длина молекул, значение L_0 лежит в пределах $0 \le L_0 \le T - 1$). Вычислительную сложность моделирования одного варианта можно оценить как N^4. Она складывается из следующих величин: N – число шагов процесса, N^2 – число стеблей, перебираемых на каждом шаге процесса, N – сложность вычисления приращения свободной энергии ΔG при добавлении к структуре нового стебляя. Таким образом, общая сложность моделирования процесса образования вторичной структуры РНК достигала N^6. В следующих разделах данной работы описывается методика вычислений для данной задачи, которая позволила понизить сложность до N^4.

3. Общая модель процесса образования вторичной структуры РНК.

В данном разделе дается математическая модель процесса образования вторичной структуры РНК в несколько более общем виде, чем в [1 – 5]. Введём некоторые обозначения. Нуклеотиды в молекулярной цепи РНК пронумеруем начиная с единицы последовательно, в порядке её роста. Пусть L – длина РНК, т.е. число нуклеотидов в полной цепи. S(n) обозначает вторичную структуру на отрезке молекулярной цепи состоящем из нуклеотидов 1, 2, 3,…, n. S(n) – это множество пар (i, j),

означающих, что нуклеотид i образовал Уотсон-Криковскую связь (спарен) с нуклеотидом j. Там, где это ясно, или не важно, мы будем обозначать вторичную структуру просто S, не указывая значения n. Совпадение вторичных структур понимается в теоретико-множественном смысле. Формула $S_1 = S_2$ означает, что множества пар (i, j) спаренных нуклеотидов для обоих структур совпадают.

Обозначим $n^0(S)$ – максимальный номер спаренного нуклеотида:

$$n^0(S) = \max_{(i,j) \in S} (\max\{i, j\})$$

Данная структура S может реализоваться только тогда, когда число нуклеотидов в молекулярной цепи станет не меньшим, чем $n^0(S)$. Для всех n: $n^0(S) \le n \le L$, имеем $S(n^0) = S(n) = S(L)$.

Стебель C – это четвёрка целых чисел $C = (n_1, n_2, n_3, n_4)$, $n_1 \le n_2 \le n_3 \le n_4$, таких, что нуклеотиды n_1, $n_1 + 1$, $n_1 + 2$,…, n_2 могут образовывать Уотсон-Криковские связи (спариваться) соответственно с нуклеотидами n_4, $n_4 - 1$, $n_4 - 2$,…, n_3. Длина стебля l(C) – это число пар оснований в нём: $l(C) = n_2 - n_1 + 1$. Любую Уотсон-Криковскую пару (n_1, n_2) можно рассматривать как стебель длины 1: (n_1, n_1, n_2, n_2). В дольнейшем мы не будем различать пару (n_1, n_2) и соответствующий ей стебель.

Стебель можно определить и как вторичную структуру, состоящую из Уотсон-Криковских пар вида (n_1, n_4), $(n_1 + 1, n_4 - 1)$, $(n_1 + 2, n_4 - 2)$,…, (n_2, n_3). В этом смысле выражение $C \subseteq S$ означает, что все пары оснований из C спариваются в структуре S. Выражение $n^0(C)$ означает максимальный номер нуклеотида спаренного в стебле, т.е. $n^0(C) = n_4$. Два стебля C^1 и C^2 называются совместимыми, если они могут одновременно участвовать в какой-либо вторичной структуре S: $C^1 \subseteq S$ и $C^2 \subseteq S$. Не любые два стебля C^1 и C^2 совместимы. Например, если $(n, p) \in C^1$, $(n, q) \in C^2$ и $p \ne q$, то стебли C^1 и C^2 несовместимы, поскольку в данной вторичной структуре основание n может быть спарено только с одним основанием.

Часто используется стерическое условие совместимости [9]. Пары (p_1, p_2) и (q_1, q_2) стерически совместимы, если их связи не “перекрещиваются”, т.е. выполнено одно из условий:

либо $[p_1, p_2] \subseteq [q_1, q_2]$,
либо $[q_1, q_2] \subseteq [p_1, p_2]$,
либо $[p_1, p_2] \cap [q_1, q_2] = \varnothing$,

где [n, m] – это целочисленный отрезок от n до m. Два стебля стерически совместимы, если совместимы любые две составляющие их Уотсон-Криковские пары.

Условия совместимости могут быть разнообразными. Мы не будем их конкретизировать здесь, поскольку это не влияет на излагаемый здесь материал.

В ходе роста вторичной структуры в ней исчезают (разрываются) некоторые старые Уотсон-Криковские связи и возникают новые. В описываемой математической модели это достигается добавлением к структуре тех или иных стеблей. Множество допустимых стеблей, которые можно добавлять к структуре S(n) обозначается D(S, n). Операцию добавления будем условно обозначать знаком +. При добавлении к структуре S(n) стебля $C \in D(S, n)$ возникает некая новая структура $S_1(n)$:

$$S + c = S_1$$

Операция добавления может происходить по-разному. В основном случае из структуры S сначала удаляются пары (i, j) несовместимые с C (разрыв Уотсон-Криковских связей), а затем добавляются пары $(i, j) \in C$ (возникновение новых Уотсон-Криковских связей). В других вариантах, например, возникновение связей или их разрушение может быть опущено. Трактовка операции добавления регулируется признаком, который приписывается стеблю при формировании множества допустимых стеблей D(S, n).

Формирование множества допустимых стеблей D(S, n) происходит по правилам модели процесса структурообразования. Допускаются различные модели. В работах [1 – 5] рассматривалась простейшая модель, в которой не допускались перестройки структур с разрушением Уотсон-Криковских связей. В этом случае D(S, n) состоит из всех стеблей C совместимых с S и таких, что $n^0(C) \le n$ (если $n^0(C) > n$, то не все Уотсон-Криковские пары в стебле могут возникнуть). В другом (предельном) случае также рассмотренном в этих работах допускаются любые перестройки и в D(S, n) включаются любые стебли C, у которых $n^0(C) \le n$. Можно также, например, запретить все стебли, содержащие неклассические пары G–U и т.п.

Выше уже говорилось, что рост вторичной структуры РНК направляется локальной минимизацией её свободной энергии. Пусть молекулярная цепь имеет длину n нуклеотидов. Будем обозначать $\Delta G(S, n)$ – приращение свободной энергии при переходе молекулярной цепи из свободного однонитевого состояния, в котором нет спаренных оснований, в состояние со структурой S(n). Отметим, что хотя $S(n_1) =$

$S(n_2)$ (для $n_1, n_2 \geq n^0(S)$), но в некоторых моделях свободной энергии может быть $\Delta G(S, n_1) \neq \Delta G(S, n_2)$ при $n_1 \neq n_2$.

Введём функцию локального приращения свободной энергии F. Пусть S(n) – некая структура, к которой добавляется стебель C, тогда

$$F(S, C, n) = \Delta G(S + C, n) - \Delta G(S , n)$$

Отметим свойство использующееся в следующем разделе. Для всех моделей энергии G_1, G_2, G_3 значение F не зависит от длины молекулярной цепи:

$$F(S, C, n) = F(S, C, m), \quad \forall\ m \geq n$$

Для модели G_4 можно полагать

$$F(S, C, n) \leq F(S, C, m), \quad \forall\ m \geq n$$

В этой модели свободная энергия зависит от галичия или отсутствия одного свободного нуклеотида в конце цепи (справа). Поэтому

$$F(S, C, n^0(C)) \neq F(S, C, n^0(C) + 1),$$

Но

$$F(S, C, n^0(C)) = F(S, C, m), \quad \forall\ n, m \geq n^0(C) + 1.$$

В этом случае каждый стебель C берётся в двух экземплярах C и C^*, при этом экземпляр C считается допустимым при $n \geq n^0(C) + 1$ и величина F для него берётся как обычно; экземпляр C^* считается доаустимым при $n \geq n^0(C)$, но

$$F(S, C^*, n)\begin{cases} F(S,C,n) & при\ n = n^0(C) \\ +\infty & при\ n > n^0(C) \end{cases}$$

Обозначим

$$F_{\min} = \min_{C \in D(S,n)} F(S,C,n)$$

F_{min} – минимальное приращение свободной энергии при добавлении к структуре S стебля из множества допустимых стеблей. Множество стеблей , на которых $F = F_{min}$ обозначим D_{min}:

$$D_{min}(S, n) = \{C \in D(S, n): F(S, C, n) = F_{min}\}.$$

Если $F_{min} > E_{act}$, т.е. больше энергии активации, то $D_{min} = \varnothing$.

Если $D_{min}(S, n) = \varnothing$, то это означает, что структурные перестройки возможны только при поступлении дополнительной энергии, т.е. что структура S энергетически устойчива.

Введём также

$$D^{\sim}(S, n) = \{C \in D(S, n): F(S, C, n) < E_{act}\},$$

Тогда $D_{min} \subseteq D^{\sim}$.

Введём граф межструктурных переходов. Вершины графа – это вторичные структуры, рёбра – переходы от структуры к структуре при добавлении какого-либо стебля. Этот граф меняется с ходом времени. В каждый момент времени t молекулярная цепь имеет длину n(t). При этом в графе могут существовать только те вершины S, у которых $n^0(S) \leq n(t)$. У каждой существующей вершины S графа имеется множество допустимых переходов: $\{S \rightarrow S + c, \forall C \in D(S, n(t)\}$.

Моделирование процесса структурообразования начинается в момент t = 0 из пустой вторичной структуры S^0 при исходной длине молекулярной цепи равной L_0 нуклеотидов. Строится путь на графе межструктурных переходов. Это отражает рост молекулярной цепи и вторичной структуры на ней. Если в момент t молекулярная цепь имеет длину n(t) и мы находимся в вершине S графа (т.е. имеем вторичную структуру S), то выбирается стебель C, обеспечивающий минимальный прирост свободной энергии $C \in D_{min}(S, n(t))$ и делается переход $S \rightarrow S^* = S + C$. Если $D_{min} = \varnothing$ (пустое), то делается переход $S \rightarrow S^* = S$, причём считается, что на это затрачивается время Δt, в течение которого молекулярная цепь удлиняется на Δn нуклеотидов.

Модели элонгации (роста молекулярной цепи) могут быть разнообразными. В простейшем случае, как в [1 – 5]

$$\Delta n = \begin{cases} 0, & если\ D_{\min} \neq \varnothing \\ T, & если\ D_{\min} = \varnothing \end{cases}$$

т.е. структура полностью формируется в пределах выросшего участка молекулярной цепи, а когда этот процесс кончается, молекулярная цепь удлиняется на T нуклеотидов. В другой модели

$$\Delta n = \begin{cases} \tau, & если\ D_{\min} \neq \varnothing \\ T, & если\ D_{\min} = \varnothing \end{cases}$$

т.е. за время добавления любого стебля молекулярная структура вырастает на τ нуклеотидов. Если же структура стабилизировалась, то молекулярная цепь растёт с постоянной скоростью T нуклеотидов за единицу времени. Интересны модели, в которых время переформирования структуры (и, следовательно, Δn) растёт по мере сложности происходящих переформирований.

Процесс структурообразования кончается, когда молекулярная цепь дорастает до конца (n = L) и структура стабилизируется ($D_{min} = \varnothing$).

4. Методика ускорения вычислений.

В данном разделе представляется методика ускорения вычислений при математическом моделировании процесса образования вторичной структуры РНК. Модель процесса описана в предыдущем разделе.

Основное время при моделировании уходит на вычисление множества энергетически минимальных допустимых стеблей $D_{min}(S, n)$ для данной вторичной структуры S при длине молекулярной цепи n нуклеотидов. Поскольку процесс моделируется при разных значениях начальной длины L_0 молекулярной цепи и периода структуризации T , то мы можем попадать в данную структуру S много раз (порядка L^2 в ходе моделирования). Основное время при вычислении D_{min} уходит на определение приращений свободной энергии F(S, C, n) = ΔG(S + C, n) – ΔG(S , n) для различных стеблей C. В использовавшихся моделях свободной энергии величина F либо вовсе не зависит от n (при $n \geq n^0(C)$), либо не зависит от n при $n \geq n^0(C) + 1$ (модель G4 [9]). Поэтому очень эффективным оказался следующий метод.

Метод 1. При попадании в структуру S вычисляется и заполняется

$$D^0_{\min}(S) = \bigcup_{n \geq n^0(S)} D_{\min}(S, n)$$

Кроме того заполняются значения F(S, C, n) для всех $n \geq n^0(C)$ и $C \in D^0_{min}(S)$. При дальнейших попаданиях в структуру S множество $D_{min}(S, n)$ отыскивается как подмножество $D^0_{min}(S)$. Заметим, что число стеблей,

реально попадавших в $D^0_{min}(S)$ оказалось достаточно малым (порядка N/10).

Метод 1 универсален. Он не использует специальных черт моделируемого процесса, однако требует определённой памяти. Обратимся теперь к методам, использующим специфику и за счёт этого ускоряющих вычисления или уменьшающих размер требуемой памяти.

Свойство 1. Монотонность локального приращения свободной энергии при росте молекулярной цепи.

$$F(S, C, n) \le F(S, C, m), \quad \forall\ m > n.$$

Это свойство означает, что если молекулярная цепь увеличивается, а набор Уотсон-Криковских связей остаётся неизменным, то свободная энергия неубывает, т.е. вторичная структура не становится стабильнее. Это свойство имеет место для всех исследовавшихся нами математических моделей структурообразования (см. п. 2). Оно даёт возможность использовать следующий метод ускорения вычислений.

Метод 2. Обозначим

$$D^-(S,n) = \bigcup_{k \ge n} D_{\min}(S,k)$$

Нетрудно видеть, что $\forall$ m > n имеем

$$D^-(S,m) \subseteq D^-(S,n) \bigcup (D(S,m) \setminus D(S,n)) \qquad (4.1)$$

В самом деле, пусть $C \in D^-(S,m)$. Если $C \notin D(S, n)$, то $C \in D(S, m) \setminus D(S, n)$, если $C \in D(S, n)$, то

$$F(S, C, n) \le F(S, C, m) < E_{act}$$

и, значит, $C \in D^-(S,n)$, что и доказывает (4.1).

Поскольку $D_{\min}(S,m) \subseteq D^-(S,m)$, то поиск D_{min} можно вести на стеблях из множества, стоящего в правой части (4.1), что значительно сокращает поиск. Это особенно выигрышно, если структура S стабильна при длине молекулярной цепи n. Тогда

$$D_{\min}(S,n) = D^-(S,n) = \varnothing,$$

и

$$D_{\min}(s,m) \subseteq D(s,m) \setminus D(S,n)$$

При малых скоростях роста молекулярной цепи может оказаться, что m = n + Δn близко к n и, что D(S, m) = D(S, n). Тогда $D_{min}(S, m) \subseteq \varnothing$, и перебор полностью отсутствует.

Свойство 2. Отметим важное свойство множеств допустимых стеблей, которое нам понадобится в дальнейшем. Это свойство монотонности по отношению к длине молекулярной цепи:

$$\text{если} \quad n < m, \quad \text{то} \quad D(S, n) \subseteq D(S, m) \qquad (4.2)$$

Оно означает, что с ростом молекулярной цепи множество структур, в которые можно перейти из структуры S, растёт. Именно на этом свойстве основаны описанные ниже методы ускорения вычислений при математическом моделировании. В исследовавшихся нами математических моделях процесса структуризации это свойство имело место.

В связи с этим свойством можно ввести числовую функцию d(C) – предел допустимости для стебля C. Это минимальное n, при котором $C \in D(S, n)$. Величина d(C) зависит от структуры S, но мы не будем её указывать, т.к. это не приводит к двусмысленности. Ясно, что $d(C) \geq n^0(C)$, таким образом, стебель C допустим в диапазоне $d(C) \leq m \leq L$.

Свойство 3. Имеет место для моделей свободной энергии G_1, G_2, G_3 :

$$F(S, C, n) = F(S, c, m), \ \forall \ m \geq n \qquad (4.3)$$

(см. п.2).

Метод 3. Этот метод применим для моделей свободной энергии G_1, G_2, G_3, для которых выполнено (4.3) и (4.2).

Имеем $\forall \ m \geq n$

$$D_{\min}(S,m) \subseteq D_{\min}(S,n) \bigcup (D(S,m) \setminus D(S,n)) \qquad (4.4)$$

В самом деле, пусть $C \in D_{min}(S, m)$ и $C \in D(S, n)$. Из свойства 2 следует, что $D_{min}(S, n) \subseteq D(S, m)$. Тогда $F(S, C, m) \leq F(S, C^*, m)$ и, из (4.3) имеем $F(S, C, n) \leq F(S, C^*, n)$ и, значит $C \in D_{min}(S, n)$, что и требовалось доказать.

Использование (4.4) вместо (4.1) ещё более сокращает перебор.

В дальнейшем будем использовать следующее обозначение. Если D – некое множество стеблей, то $D \mid n = \{C \in D : d(C) \le n\}$ – подмножество стеблей из D, которые допустимы на молекулярной цепи длиной n и выше. Пусть также

$$P(D) = \min_{C \in D} d(C)$$

Утверждение 1. Пусть $P_n = P(D_{min}(S, n))$, тогда

$$D_{min}(S, m) = D_{min}(S, n) \mid m, \quad \forall\, m: \max\{P_n, n^0(S)\} \le m \le n$$

Доказательство. Докажем, что $D_{min}(S, n) \mid m \subseteq D_{min}(S, m)$. Пусть $C \in D_{min}(S, n) \mid m$, тогда $d_0(C) \le m$ и $F(S, C) \le F(S, C^*)$, $\forall\, C^* \in D(S, n)$. Но $D(S, m) \subseteq D(S, n)$, значит $F(S, C) \le F(S, C^*)$, $\forall\, C^* \in D(S, m)$. Поскольку $d_0(C) \le m$, то $C \in D_{min}(S, m)$.

Докажем, что $D_{min}(S, m) \in D_{min}(S, n) \mid m$. Возьмём $C^* \in D_{min}(S, n) \mid m \in D_{min}(S, m)$, тогда $\forall\, C \in D_{min}(S, n)$ имеем $F(S, C) = F(S, C^*)$. Значит

$$F(S, C^*) \le F(S, C'),\ \forall\, C' \in D(S, n)$$

Кроме того $d(C^*) \le m \le n$, поэтому $C^* \in D(S, m) \subseteq D(S, n)$, значит $C^* \in D_{min}(S, n)$, что и требовалось доказать.

Нетрудно видеть, что верно следующее следствие.

Следствие 1. Пусть $n^0(S) \le m \le n$. Если $D_{min}(S, n) \mid m = \varnothing$, то $D_{min}(S, m) = D_{min}(S, n) \mid m$.

Сформулируем теперь два метода ускорения вычислений, опирающихся на это утверждение.

Метод 4. Для каждой структуры S, которая возникла при моделировании вычисляется и сохраняется множество стеблей $D = \bigcup_k D_k$ и значения локального приращения свободной энергии F(S, C). Множества D_k определяются рекуррентными соотношениями:

$$D_1 = D_{min}(S, L), \qquad n_1 = P(D_1)$$

$$D_{k+1} = D_{min}(S, n_k - 1), \quad n_{k+1} = P(D_{k+1})$$

Нетрудно видеть, что $\forall$ n

$$D_{min}(S, n) \subseteq \{C \in D: d(C) \le n\}$$

В отличие от метода 1 здесь не надо вычислять $D_{min}(S, n)$ для всех значений n, а только для $n = n_k - 1$, k = 1, 2,… Это существенно ускоряет вычисления.

Прежде, чем описывать следующий метод ускорения вычислений, определим множество N(S). В ходе математического моделирования (для различных значений L_0, T) мы можем попадать в данную структуру S неоднократно. Длина молекулярной цепи n при разных попаданиях может быть различной. Обозначим её n_i для i-го попадания в структуру S. Обозначим $N_k(S)$ – множество содержащее величины n_i для i = 1, 2, …, k: $N_k(S) = \{n_1, n_2,\dots ,n_k\}$.

Метод 5. Для каждой вторичной структуры S, в которую мы попали в ходе математического моделирования в k-й раз, хранится множество D_k :

$$D_k(S) = \bigcup_{n \in N_k(S)} D_{\min}(S,n)$$

Помимо этого множества для любого стебля C $\in$ $D_k(S)$ хранится величина $n^{max}(C)$ – максимальная длина молекулярной цепи из N_k, при которой стебель C давал минимальный локальный прирост свободной энергии (и, следовательно, мог участвовать в переформировании структуры):

$$\forall C \in D_k(S), \qquad n^{\max}(C) = \max_{\substack{n \in N_k(s) \\ C \in D_{\min}(S,n)}} \{n\}$$

Вся эта информация накапливается рекуррентным образом:

$$D_1 = D_{\min}(S, n), \quad D_m(S) = D_{m-1}(S) \bigcup D_{\min}(S, n_m),$$
$$n_1^{\max}(C) = n_1, \qquad n_m^{\max}(C) = \max\{n_{m-1}^{\max}(C), n_m\}$$

(если C $\in$ $D_{min}(S, n_m)$).

Пусть теперь мы попали в структуру S в k + 1-й раз, при длине молекулярной цепи РНК равной n_{k+1}. Определим множество Q стеблей из $D_k(S)$ допустимых для данной длины цепи РНК:

$$Q = \{C \in D_k(S): d(C) \le n_{k+1}\} = D_k(S) \mid n_{k+1}.$$

Если Q не пусто ($Q \neq \varnothing$), то определим

$$n^* = \max_{C \in Q} n_k^{\max}(C), \quad D^* = \{C \in Q : n_k^{\max}(C) = n^*\}$$

Если $Q = \varnothing$, то положим $D^* = \varnothing$, $n^* = -\infty$. Если $n^* \ge n_{k+1}$, то это означает, что на предыдущих шагах моделирования мы попадали в структуру S при длине молекулярной цепи $n^* \ge n_{k+1}$ и, значит, $D_{min}(S, n^*) \subseteq D_k$. При этом

$$D_{min}(S, n^*) \mid n_{k+1} \neq \varnothing.$$

Значит (см. Следствие 1 из Утверждения 1)

$$D_{min}(S, n_{k+1}) = D_{min}(S, n^*) \mid n_{k+1}.$$

Поэтому

$$D_{min}(S, n_{k+1}) = D^* \qquad (4.5)$$

Если $n^* < n_{k+1}$ или $Q = \varnothing$, то

$$D_{\min}(S, n_{k+1}) \subseteq D^* \bigcup \{C \in D(S, n_{k+1}) : d(c) > n^*\} \qquad (4.6)$$

В силу определения $D^* \subseteq D(S, n_{k+1})$ и, кроме того, $D^* \subseteq D(S, n^*)$, причём $F(S, C^*) \le F(S, C)$, $\forall C^* \in D^*$ и $\forall C \in D(S, n^*)$ и

$$F(S, C^*) \le F(S, C), \quad \forall C^* \in D^* \text{ и } \forall C \in D(S, n^*) \setminus D^* \qquad (4.7)$$

Если $C \in D(S, n_{k+1})$ и $d(C) \le n^*$, то $C \in D(S, n^*)$. Если $C \notin D^*$, то из (4.7) $C \notin D_{min}(S, n_{k+1})$.

Таким образом, $D_{min}(S, n_{k+1})$ определяется либо из (4.5), либо из (4.6). И то, и другое значительно сокращает перебор.

5. Анализ эффективности.

Эффективность рассматриваемой методики во многом зависит от конкретного вида моделируемого процесса: от того, какие перестройки структуры допустимы, какие временные задержки они вызывают и т.п. Поскольку именно этим определяется сложность или простота, разбросанность или компактность семейства путей на графе межструктурных переходов, по которым будет развиваться процесс для различных значений (T, L_0). Поэтому здесь мы сможем дать только довольно грубую аналитическую оценку, и дополнить её некоторыми числовыми данными, полученными при математическом моделировании.

Напомним обозначения:

L – полная длина молекулы РНК,

T – период структуризации.

Если считать, что, грубо, процесс развивается так: сначала структура наращивается, затем молекулярная цепь удлиняется, то число шагов в развитии процесса составляет $\frac{2L}{T}$. При этом возникают последовательно $\frac{L}{T}$ структур. На каждом шаге процесса, при текущей длине РНК – n нуклеотидов, имеем порядка $\sim 0.3n^2$ стеблей, которые можно добавить к структуре. Для выбора энергетически минимального стебля надо затратить порядка $\sim 0.3n^2\tau$ времени, где τ – время расчёта свободной энергии для одного стебля. Модели свободной энергии таковы, что $\tau \sim n$. Таким образом, общее время счёта на одном шаге процесса составляет $\sim n^3$. Просуммировав это по $\sim \frac{L}{T}$ шагам процесса, получим, что моделирование одного варианта процесса структурообразования требует времени $\sim \frac{L^4}{T}$. Поскольку для данного T мы моделируем процесс на различных значениях начальной длины молекулы $L_0 = 0, 1, 2, \ldots, T - 1$, то для данного T (на всех L_0) моделирование требует времени $\sim L^4$. Поскольку T также меняется от 1 до L (с шагом 1), то общее время моделирования для одной молекулы РНК без применения методки ускорения оценивается как $\sim L^5$.

Методы ускорения вычисления, изложенные в п.4 можно разбить на две группы. Это метод 2, не требующий памяти, и остальные (1, 3 – 5), использующие дополнительную память ЭВМ. Рассмотрим сначала

метод 2. Сокращение времени здесь происходит за счёт того, что после стабилизации структуры, при удлинении молекулярной цепи РНК, перебор происходит для стеблей, возникших на участке цепи от n до n + T нуклеотидов. Число их можно оценить как ~$0.3((n + T)^2 – n^2) \sim nT$, а число таких шагов ~$\frac{L}{T}$. Таким образом, время вычислений для этих шагов (при фиксированных T и L_0) можно оценить как ~L^3. А при всех (T, L_0), как ~L^5. Значит, в этом методе время вычислений, хотя и сокращается, но всё равно имеет порядок L^5. Реальные эксперименты с этим методом показали, что, для молекул РНК длиной около 100 нуклеотидов, этот метод уменьшает время вычислений в полтора – два раза.

Обратимся теперь к группе методов с памятью. Здесь основное время вычислений приходится на анализ новых (ещё не возникавших) структур. Общее число возможных структур на отрезке молекулярной цепи длиной L можно оценить как ~L^3 (экспериментальная оценка). Оценка каждой структуры требует время ~L, а в целом получается ~L^4. После того, как структуры оценены, остаётся расчет ~L^2 вариантов (для всех T и L^0). Это время добавляется к предыдущему и в целом получается ~L^4.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект 99-01-00029).

ЛИТЕРАТУРА

1. Кугушев Е.И., Козлов Н.Н. Компьютерное моделирование структурообразующих свойств плюс- и минус-цепей ДНК, кодирующих 5s р РНК и тРНК. (1993), Доклады Академии Наук, 333 (1), стр. 107 – 111.
2. Kozlov N.N., Kugushev E.I. (1993), Computer simulation of tRNA secondary structure folding. CABIOS, v. 9, p. 253 - 258.
3. Кугушев Е. И., Козлов Н.Н. (1995), Моделирование влияния структурных перестроек на процесс образования вторичной структуры РНК. Доклады Академии Наук, 340 (2), стр. 263 – 267.
4. Козлов Н. Н., Кугушев Е.И., Энеев Т. М. (1998), Структурообразующие характеристики транскрипционного процесса. Математическое моделирование, 10(6), стр. 3 – 19.
5. Туманян В.Г., Сотникова Л.Е., Холопов А.Е. (1966), Об определении вторичной структуры РНК по последовательности нуклеотидов. Докл. АН СССР, т. 166, N 6, с. 1465 - 1468.
6. Salser,W. (1977), Globin mRNA sequences: Analysis of base pairing

and evolutionary implications. Cold Spring Harbor Symp. Quant. Biol., v. 42, p. 985 - 1002.
7. Jacobson A.B. Good L., Simonetti J., Zuker M. (1984), Some simple computational methods to improve the folding of large RNAs. Nucleic Acids Res., v. 12, N 1, p 45 - 52.
8. Freier,S.M., Kierzek,R., Jalger,J.A., Sugimoto,N., Caruthers,M.H., Neilson,T., and Turner,D.H. (1986), Improved free-energy parameters for predictions of RNA duplex stability. Proc. Natl. Acad. Sci. USA, **83**, 9373–9377.
9. Ninio,J., (1979), Biochimic, 61, 1133-1150.