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

Д. И. Сабитов

Об идентификации параметров математической модели пространственной структуры РНК

Москва 2004

Аннотация. В работе описываются наиболее важные параметры, относящиеся к математическим моделям вторичных и третичных структур молекул рибонуклеиновых кислот (РНК). Дается строгая количественная оценка разнообразия возможных вторичных структур молекул РНК. Ставится задача идентификации параметров модели третичной структуры, основывающаяся на данных о третичных структурах реальных молекул транспортных РНК, полученных методом рентгеноструктурного анализа. Первый этап этой задачи (определение геометрических параметров двуспиральных участков третичной структуры молекулы РНК) решается на основе рентгеноструктурных данных о форме дрожжевой фенилаланиновой транспортной РНК.

Ключевые слова: вторичная, третичная структуры РНК, генетический алгоритм, идентификация параметров

Abstract. The most important parameters of mathematical models of secondary and tertiary structures of molecules of ribonucleic acids (RNA) are described in this work. A severe quantitative estimation of a variety of possible secondary structures of RNA molecules is given. A task of identification of the tertiary structure model's parameters based on the structures of real transport RNAs received by a method of the X-Ray analysis is put. The first stage of this task (determination of geometrical parameters of two-spiral sites of tertiary structure of an RNA molecule) is solved on the basis of the X-Ray data on the form of a yeast phenilalanine transport RNA.

Key words: secondary, tertiary structure of a RNA, genetic algorithm, identification of parameters

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проекты 04-01-00358, 02-01-00352 и 02-07-90027) и Программы фундаментальных исследований Президиума РАН № 17.

Содержание

Введение

Математическое моделирование пространственной структуры биологических макромолекул, таких как РНК и белки, является в настоящее время одной из наиболее бурно развивающейся ветвей молекулярной биологии. Фундаментальность этой проблемы определяется тем, что основные процессы жизнедеятельности клетки зависят в первую очередь от пространственной структуры упомянутых макромолекул.

В последнее время в фармакологической промышленности интенсивно развивается метод создания лекарственных препаратов на основе РНК-содержащих соединений (SELEX, Systematic Evolution of Ligands by Exponential enrichment [11, 12]). Он основан на том, что даже сравнительно короткие молекулы РНК (около 100 нуклеотидов) обладают гигантским количеством пространственных форм. Такое разнообразие в принципе позволяет подобрать подходящую по форме молекулу РНК, закрывающую активные центры патогенного соединения (белка или фермента) и препятствующую его разрушительной работе (docking-метод [8]). Основная проблема здесь – это поиск достаточно дешевого метода определения пространственной структуры короткой молекулы РНК по ее нуклеотидному составу (первичной структуре). Наиболее перспективными для решения этой задачи мы считаем методы математического моделирования.

Для определения пространственной структуры молекулы чаще всего используется рентгеноструктурный анализ. Однако, этот физический метод является дорогостоящим и длительным. Помимо чисто практических нужд фармакологии, необходимость в дешевых методах определения третичной структуры молекул РНК вызывается и тем, что поток информации о вновь открытых молекулах и расшифровке их нуклеотидного (для нуклеиновых кислот) и аминокислотного (для белков) состава год от года возрастает.

Предлагаемая нами математическая модель пространственной структуры РНК [1] является довольно простой, и вряд ли она будет давать точную информацию о пространственной структуре длинных молекул РНК. Тем не менее, при определении пространственной структуры другими методами нашу модель можно использовать для построения начального приближения структуры. Затем ее можно последовательно уточнять.

Молекулярная цепь рассматривается как тонкий упругий стержень, имеющий в свободном состоянии форму винтовой линии. С шагом, равным длине одного нуклеотида, в стержень вделаны жесткие перемычки, равные по длине половине Уотсон-Криковской связи. В простейшей модели параметры винтовой линии и ориентация перемычек выбираются так, что два свободных стержня одинаковой длины, будучи правильно расположены, образуют двойную спираль в А-форме. Пространственная структура молекулы собирается из стеблей и петель в соответствии с заданной вторичной структурой. Каждая петля состоит из семейства тонких упругих стержней со взаимно согласованными краевыми условиями равновесия. В соответствии с краевыми условиями концы стержней ориентируются так же, как концы нитей в стебле А-формы

РНК.

Качество предсказания пространственной структуры молекулы РНК напрямую зависит от правильного определения параметров математической модели. В данной работе рассматриваются вопросы, связанные с решением этой задачи.

Параметры модели разбиваются на несколько групп. Это геометрические параметры стержня, моделирующего молекулярную цепь, его упругие характеристики, геометрические параметры жестких перемычек, моделирующих Уотсон-Криковскую связь, и геометрические параметры двуспиральных участков.

Задача идентификации параметров модели естественным образом разделяется на два этапа. Первый этап – это идентификация геометрических параметров двуспиральных участков и жестких перемычек, моделирующих Уотсон-Криковскую связь. Второй этап – это идентификация геометрических и упругих параметров стержня, моделирующего молекулярную цепь, с использованием результатов, полученных на первом этапе.

В работе основное внимание уделяется первому этапу задачи идентификации. Основная идея идентификации параметров состоит в использовании опубликованной информации о третичных структурах некоторых реальных молекул транспортных РНК, полученной методом рентгеноструктурного анализа [10]. Эти данные включают, в частности, координаты атомов рибозофосфатных остовов обеих нитей, составляющих двуспиральные участки стеблей тРНК. Идентифицируемые параметры подбираются таким образом, чтобы нити модельного двуспирального участка как можно лучше приближались к нитям рибозофосфатных остовов двуспирального участка реальной молекулы.

Изложим теперь кратко содержание препринта. Для большей ясности, помимо материала, посвященного собственно задаче идентификации параметров, в работу включен раздел, в котором кратко описываются основные понятия, задачи и параметры, относящиеся к общей проблеме изучения пространственных структур молекул РНК. В этом же разделе дается строгая количественная оценка разнообразия структур молекул РНК, которая показывает перспективность использования РНК-содержащих препаратов в фармакологии.

Материал, относящийся к задаче идентификации параметров, разбит на три раздела. Первый раздел посвящен оценке разнообразия потенциально возможных структур молекул РНК. Во втором разделе дается строгая формулировка задачи идентификации, рассматриваются проблемы, возникающие при ее решении. Также описываются методы и алгоритмы разрешения этих проблем. Третий раздел посвящен описанию результатов первых экспериментов по определению геометрических параметров двуспиральных участков третичной структуры молекулы РНК на основе данных рентгеноструктурного анализа формы дрожжевой фенилаланиновой транспортной РНК [10]. Опишем содержание работы более подробно.

В первом разделе дается краткое описание иерархии структур нуклеиновых кислот и, в частности, структур молекул РНК. Принято выделять четыре структурных уровня – *первичную, вторичную, третичную и четвертичную*

структуры молекулы РНК.

В рассматриваемой модели пространственная (третичная) структура строится на основе известной вторичной. Краткий обзор по вторичным структурам можно найти, например, в [6, 7]. Задача об определении вторичной структуры в данной работе не рассматривается. Однако, здесь дается оценка числа потенциально возможного количества вторичных структур. Оказывается, что количество вторичных структур, имеющих n связей, пропорционально $k(n+1)$. Символом $k()$ мы обозначаем так называемые числа Каталана, изучаемые в комбинаторике [3].

Во втором разделе дается краткое описание используемой математической модели и ее параметров, подлежащих идентификации. Рассказывается о постановке задачи идентификации, дается описание данных рентгеноструктурного анализа молекул тРНК, называются основные шаги процесса идентификации. Далее описывается генетический алгоритм поиска экстремумов и способ построения хромосомы для задачи идентификации. Также рассматривается задача сравнения двух ломаных с различным числом звеньев. Предлагаются три методики ее решения.

В третьем разделе приводятся результаты некоторых вычислительных экспериментов по оценке геометрических параметров модели двуспиральных участков РНК на основе рентгеноструктурных данных. В задаче сопоставления геометрических форм пространственных кривых функционал сравнения может иметь не одну точку экстремума. Следует ожидать, что их может быть довольно много. Поэтому в строгом математическом смысле задача не является корректной (т. е. ее решение не единственно). Однако, наличие минимума у оценочного функционала можно гарантировать, т. к. он ограничен снизу на компактном пространстве параметров. Поэтому осмысленной является задача изменения геометрических параметров модели таким образом, чтобы оценочный функционал стал меньше при разумной форме и взаимном расположении сопоставляемых кривых в точке экстремума. В силу сказанного результаты процесса идентификации контролировались нами для того, чтобы отбрасывать заведомо неподходящие варианты. В разделе приводятся результаты наиболее удачных процессов идентификации параметров.

1. Разнообразие структур молекул РНК

В данном разделе рассматриваются вопросы, связанные с оценкой разнообразия потенциально возможных структур молекул РНК.

1.1 Иерархия структур нуклеиновых кислот

Химически молекула РНК представляет собой полинуклеотид – цепочку соединенных в линейной последовательности мономеров-нуклеотидов. В РНК нуклеотиды бывают 4х типов: аденин (A), гуанин (G), цитозин (C) и урацил (U). Последовательность расположения нуклеотидов в молекулярной цепочке несет информационную нагрузку: это могут быть копии генов, необходимых клетке для процесса жизнедеятельности (матричная РНК). В иных случаях по-

следовательность нуклеотидов задает форму молекулы, определяющую функцию РНК (транспортные РНК, рибосомальные РНК). Нуклеотидная последовательность молекулы называется ее *первичной структурой*.

Некоторые нуклеотиды в молекулярной цепочке связаны попарно т.н. Уотсон-Криковскими связями. Структура этих связей называется *вторичной структурой* РНК. В силу конечности множества возможных Уотсон-Криковских связей число вторичных структур, которые может принимать данная молекула РНК, конечно, но весьма значительно. Определение реальной вторичной структуры РНК по ее известной первичной структуре также является важной фундаментальной проблемой молекулярной биологии.

Под *третичной структурой* молекулы РНК (ДНК) понимается пространственная форма, которую принимает ее молекулярная цепочка в пространстве под воздействием Уотсон-Криковских и других более слабых потенциалов. *Четвертичной структурой* называется форма молекулы, которую она приобретает, связываясь в комплекс с другими биомолекулами. Первичная структура у молекулы РНК одна, а возможных вторичных (третичных, четвертичных) структур много.

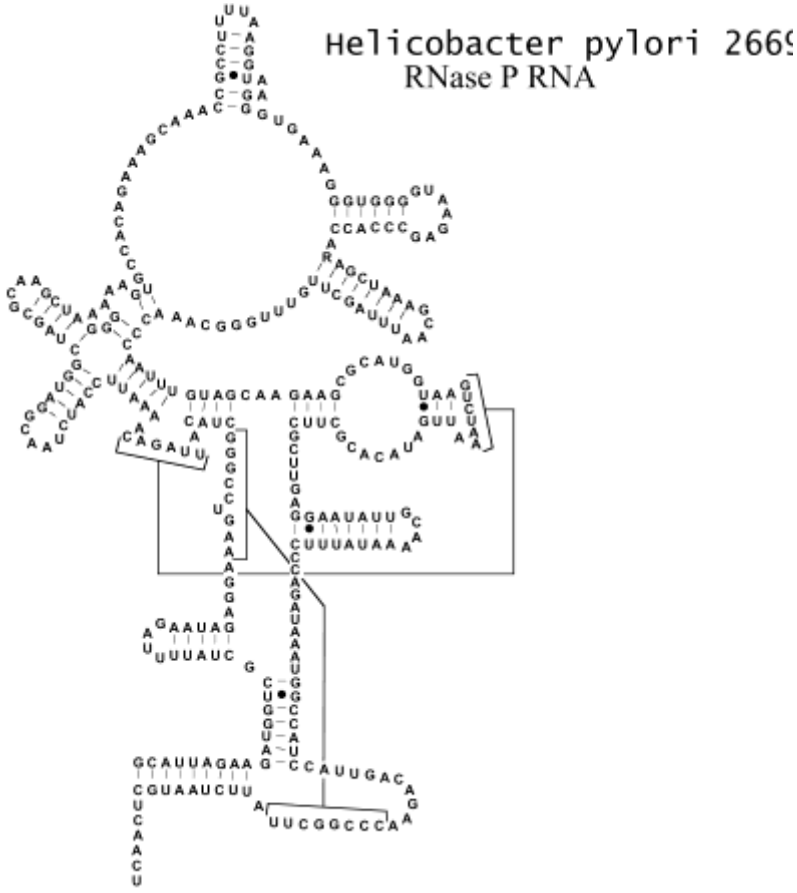


Рис. 1.1 Первичная и вторичная структуры молекулы Helicobaster pyori RNase P RNA [9].

1.2 Элементы пространственных структур РНК

С точки зрения вторичной структуры все нуклеотиды в молекуле можно разбить на два класса – спаренные (т.е. образующие Уотсон-Криковскую связь с каким-либо другим нуклеотидом) и свободные (неспаренные). Формально,

вторичная структура РНК – это описание всех спаренных и свободных оснований в молекулярной цепи.

Отрезок молекулярной цепи, состоящий из неспаренных оснований, 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 (и семейства таких участков) называются *петлями*, а двухнитевые – *стеблями*. Стебель можно представить себе, как участок винтовой лестницы, где ступеньки – это поперечные, Уотсон-Криковские связи. Длиной стебля называется число пар оснований в нём: $n_2 - n_1 + 1 = n_4 - n_3 + 1$. Таким образом, вторичная структура РНК – это совокупность стеблей и петель.

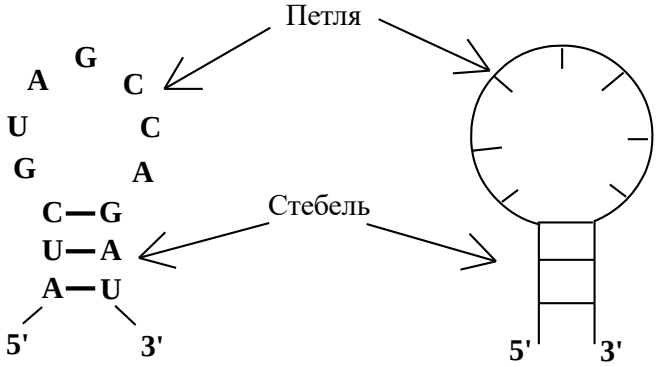
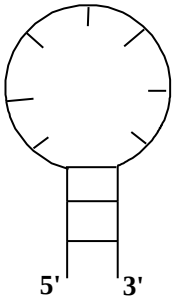


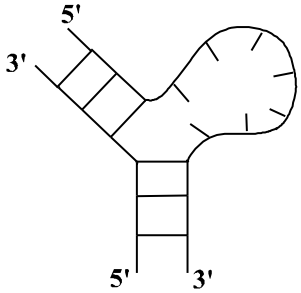
Рис. 1.2 Однонитевый (петля) и двухнитевый (стебель) элементы вторичной структуры РНК.

1.3 Типы петель во вторичной структуре

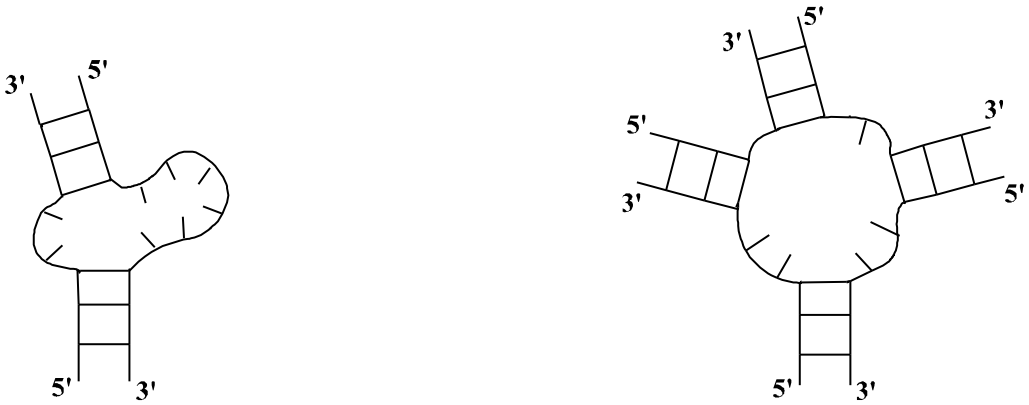
Петлей называется замкнутая последовательность однонитевых участков РНК, концы которых соединены Уотсон-Криковскими вторичными связями. При этом начало каждого следующего участка соединено с концом предыдущего, а конец последнего участка соединен с началом первого. Однонитевые участки, входящие в состав петли, называются ее *ветвями* или *звеньями*. *Длиной* петли называется число свободных нуклеотидов, входящих в ее состав. Выделяют следующие типы петель.



a) Шпилечная петля



b) Боковая петля



c) Внутренняя петля d) Многозвенная петля

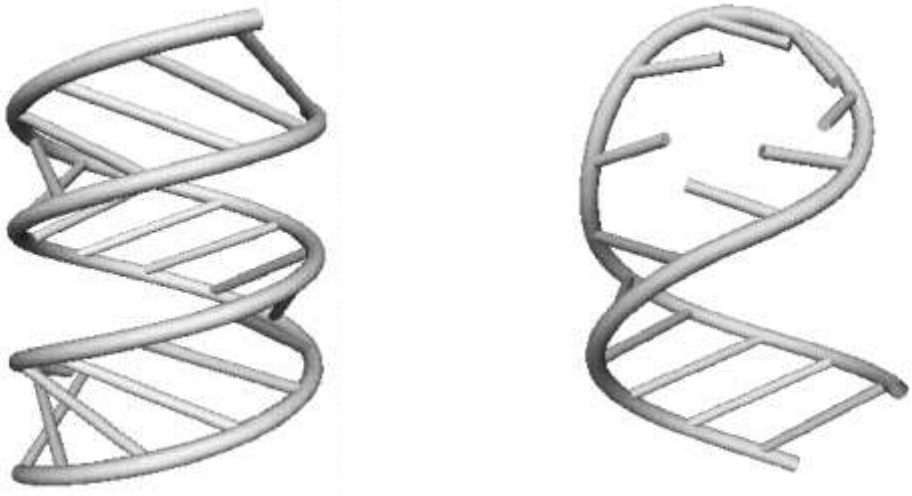
Рис 1.3 Типы петель (однонитевых участков вторичной структуры).

Шпилечная петля соединяет первую и вторую нить в одном стебле. Это простейшая однозвенная петля (она состоит из одного однонитевого участка). Считается, что шпилечная петля всегда содержит не менее трех нуклеотидов.

Боковая петля содержит два однонитевых участка, один из которых вырожден – имеет нулевую длину (не содержит ни одного несвязанного нуклеотида). Длина же второго участка называется длиной боковой петли.

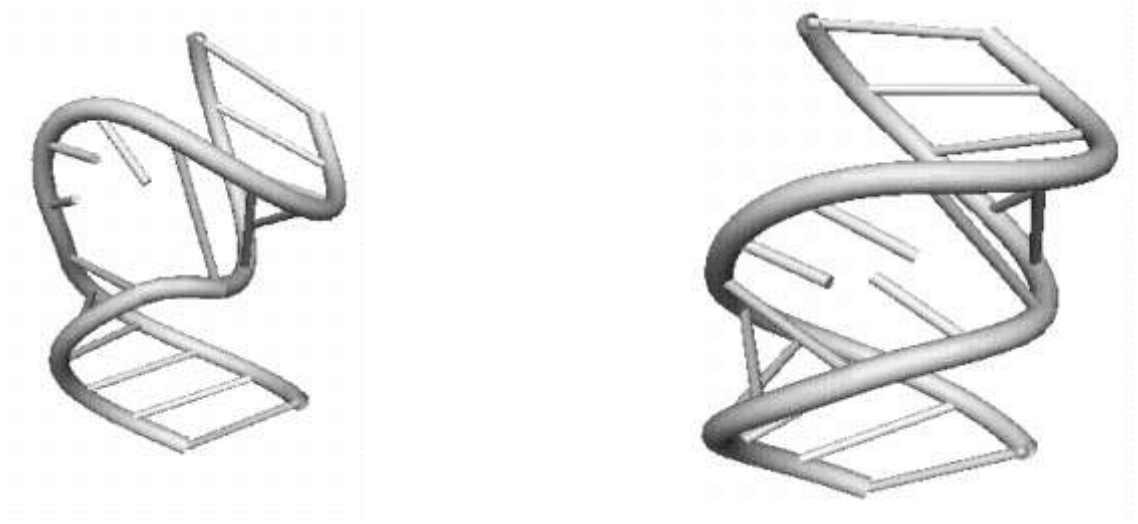
Внутренняя петля содержит два однонитевых участка. Длины этих участков являются параметрами, определяющими петлю.

Многозвенная петля содержит несколько однонитевых участков. Число этих участков и их длины являются параметрами, определяющими петлю.



a) b)

Рис 1.4 Элементы третичной структуры РНК: a) стебель; b) шпилечная петля длиной в семь нуклеотидов.



a) b)

Рис 1.5 Элементы третичной структуры РНК: a) боковая петля, состоящая из трех нуклеотидов; b) внутренняя петля, звенья которой состоят из одного и двух нуклеотидов.

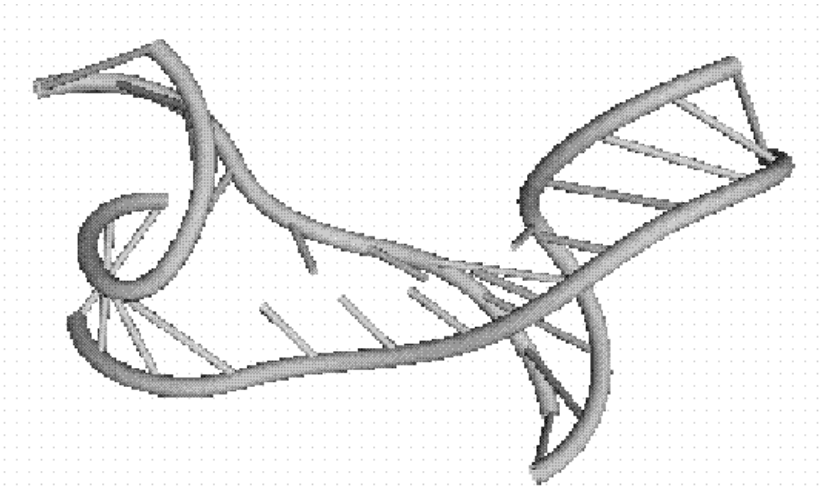


Рис 1.6 Элементы третичной структуры РНК: пример многозвенной петли, состоящей из четырех звеньев.

1.4 Способы описания вторичной структуры.

Существует несколько способов описания вторичных структур РНК. Мы опишем два основных.

Геометрический способ – молекулярная цепь располагается на плоскости так, чтобы стебли образовывали прямоугольные “лесенки”, а петли – замкнутые контуры (см. Рис 1.2, 1.3). Такой способ дает некоторое представление о реальной геометрии структуры молекулы.

Представление на окружности – молекулярная цепь располагается на плоскости по окружности, так что движению от *5’* конца к *3’* концу соответствует движение против часовой стрелки. Уотсон-Криковские связи при этом изображаются хордами окружности. Если никакие две хорды не пересекаются, то говорят, что выполняется *стерическое условие.*

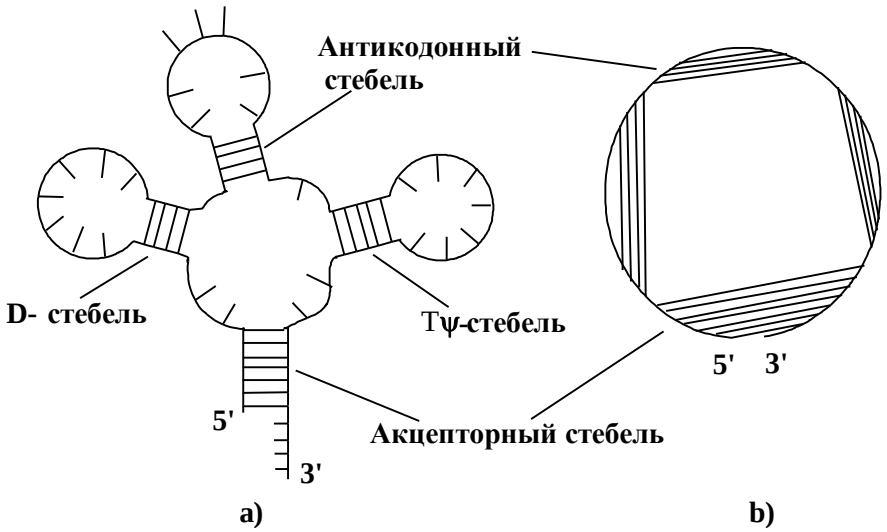


Рис 1.7 Представление вторичной структуры тРНК геометрически (a) и на окружности (b). Для такой структуры стерическое условие выполнено.

1.5 Стерическое условие и псевдоузлы.

Если в представлении вторичной структуры на окружности никакие две хорды не пересекаются, то говорят, что выполняется *стерическое условие*. Перенумеруем все нуклеотиды в молекулярной цепи, начиная с *1*. Две пары нуклеотидов (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. Если стерическое условие не выполнено, то вторичная структура содержит *псевдоузлы.* В описываемой модели предполагается, что псевдоузлы во вторичной структуре отсутствуют.

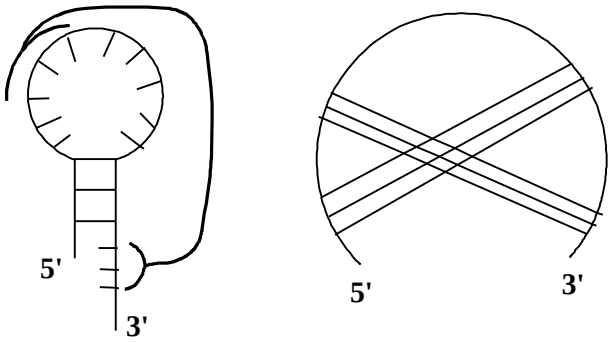


Рис. 1.8 Псевдоузел. Связи, входящие в него, пересекаются на круговой диаграмме, т. е. нарушается стерическое условие.

1.6 Оценка количества вторичных структур РНК

В подавляющем большинстве случаев различные вторичные структуры соответствуют различным третичным. Поэтому оценка количества вторичных структур дает представление о количестве третичных структур. Попробуем

оценить количество вторичных структур, которые могут возникать в молекулярной цепи, первичная структура которой (т.е. нуклеотидный состав) случайна.

Будем называть *длиной структуры* количество Уотсон-Криковских связей в ней. Напомним, что мы рассматриваем только структуры, удовлетворяющие стерическому условию. Обозначим за $V(N,n)$ среднее число структур длины n при случайном выборе молекулярной цепи, состоящей из N нуклеотидов. Количество структур длины *1* равно числу стеблей, содержащих только одну Уотсон-Криковскую связь:

$$V(N,1) = U(N,1) = \frac{N(N-1)}{8} \qquad (1.1)$$

Рассмотрим теперь структуру длины n. В ней какие-то $2n$ нуклеотидов попарно связаны Уотсон-Криковскими связями. Всего возможно C_N^{2n} различных выборов $2n$ нуклеотидов. После того, как нуклеотиды выбраны, вторичная структура определяется системой связей между ними. Перенумеруем выбранные нуклеотиды последовательно от *1* до $2n$ и расположим их в вершинах правильного $2n$-угольника. Соединим связанные нуклеотиды диагоналями. Каждой вторичной структуре взаимно однозначно соответствует набор из n непересекающихся диагоналей.

Обозначим за $s(n)$ число различных расположений n непересекающихся диагоналей в правильном $2n$-угольнике, что отвечает рассмотрению вторичной структуры в виде хорд на окружности с выполнением стерического условия. Тогда (с учетом того, что вероятность комплементарности выбранных нуклеотидов равна $\frac{1}{4^n}$) имеем

$$V(N,n) = \frac{1}{4^n} C_N^{2n} s(n) = \frac{1}{4^n} \frac{N!}{(2n)!(N-2n)!} s(n) \qquad (1.2)$$

Вывод формулы для *s(n).* Выведем формулу для функции $s(n)$, описывающую количество способов проведения n непересекающихся диагоналей в выпуклом *2n*-угольнике, *включая в понятие "диагонали" также и стороны фигуры*. Действуя по индукции, предположим $s(i)$, $1 \le i \le n$ известными и примем $s(0) = 1$. Зафиксируем в многоугольнике вершину и проведем через нее какую-нибудь диагональ (см. Рис. 1.9).

Пусть диагональ является стороной *2n*-угольника. Удалим из его ломаной границы эту диагональ, пару вершин и стороны, соединяющие эти вершины с соседними, и замкнем новую ломаную, получив многоугольник с *2(n-1)* вершинами. Таким образом, зафиксировав эту диагональ, получим $s(0)s(n-1)$ вариантов проведения непересекающихся диагоналей. Через одну вершину многоугольника, в том числе и через зафиксированную нами, проходят *2* стороны, и зафиксировав вторую из них, мы также получим $s(0)s(n-1)$ вариантов.

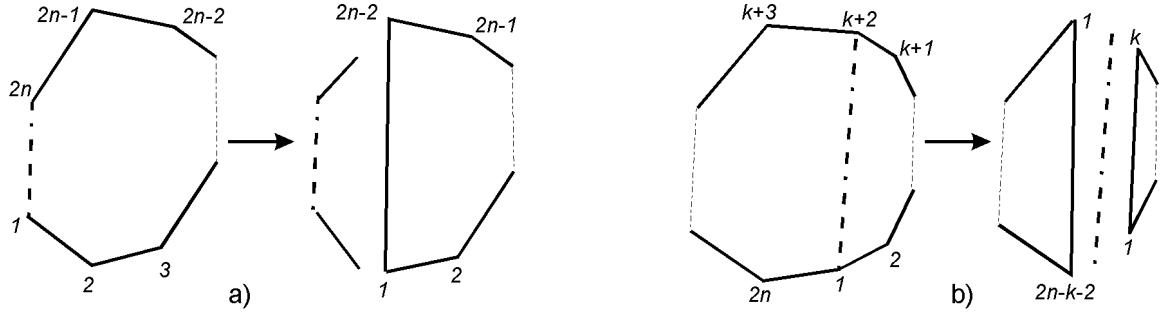


Рис. 1.9. При разбиении многоугольника диагональю, стороны включаются в число диагоналей. а) диагональ является стороной многоугольника, в результате разбиения получаются *0*-угольник и *(2n-2)*-угольник). b) диагональ не является стороной многоугольника, в результате получаются *k*-угольник и *(2n-k-2)*-угольник).

Пусть диагональ не является стороной. Пара вершин, соединенных этой диагональю, разобьет многоугольник на *2* ломаные, лежащие в разных полуплоскостях относительно нее. К каждой из них применим ту же операцию удаления-замыкания, что и в предыдущем случае, и получим *2* новых многоугольника с вершинами, количество которых будет одинаковой четности: *k* и *2n - k - 2*. Понятно, что в многоугольнике с нечетным количеством вершин g можно провести не более $(g-1)/2$ непересекающихся диагоналей: каждая диагональ соединяет *2* вершины, и одна вершина многоугольника останется свободной. Таким образом, если *k* нечетное, то в сумме в двух полученных многоугольниках мы сможем провести не более $(k-1)/2+(2n-k-3)/2=n-2$ непересекающихся диагоналей, что не соответствует нашей цели: проведению *n* диагоналей. Следовательно, *k* должно быть четным *(k = 2p)*. Тогда нужное нам число вариантов будет $s(p)s(n-1-p)$. Проводя в исходном многоугольнике различные диагонали из фиксированной точки и суммируя получающиеся количества вариантов, получим формулу:

$$s(n)=\sum_{i=1}^{n} s(i-1)s(n-i) \qquad (1.3)$$

Вычислим по этой формуле $s(1)=1,\ s(2)=2,\ s(3)=5$. Эти значения доказывают базу индукции.

Оценим функцию снизу при *n > 2*. В такой сумме есть *2* слагаемых вида $s(0)s(n-1)$. Так как все слагаемые положительны, то

$$s(n)>2s(n-1)>2^2 s(n-2)>\ldots>2^{n-1}s(1)=2^{n-1} \qquad (1.4)$$

При *n = 2* достигается равенство.

Проведем оценку сверху. Слагаемых в сумме *n*, и каждое меньше, чем $s^2(n-1)$:

$$s(n)<ns^2(n-1)<n(n-1)^2 s^4(n-2)<\ldots<\prod_{k=0}^{n-2}(n-k)^{2^k}<n^{\sum_{k=0}^{n-2}2^k}=n^{2^{n-1}-1} \qquad (1.5)$$

При $n = 2$ также достигается равенство.

Полученное рекуррентное соотношение представляет собой вариант известных *чисел Каталана* $k(n) = s(n-1)$, о которых можно прочитать в [3]. Для чисел Каталана известно и прямое выражение:

$$k(n) = \frac{1}{2n-1} C_{2n-1}^{n-1} \qquad (1.6)$$

1.7 Разнообразие вторичных структур для тРНК

Поскольку $s(n) > 2^{n-1}$, то

$$V(N,n) = \frac{1}{4^n} C_N^{2n} s(n) > \frac{1}{2^{n+1}} \frac{N!}{(N-2n)!} \qquad (1.7)$$

причём степень полинома от N, стоящего в правой части, равна $2n$. Отсюда можно сделать два вывода:

С ростом молекулярной цепи количество возможных вторичных структур, состоящих из n Уотсон-Криковских связей, растет как N^{2n}.

С ростом молекулярной цепи количество всех возможных вторичных структур растет быстрее, чем любой полином (иначе говоря, скорость роста количества вторичных структур больше полиномиальной).

Для транспортных РНК средние значения длины молекулы $N=75$ и количество Уотсон-Криковских связей составляет $n=20$. Подставляя эти значения в формулу (1.2), получим:

$$V(75,20) = 1.75679 \cdot 10^{19}$$

Для малых ядерных РНК, длина которых ≈ 30 нуклеотидов, а n около 5, число возможных структур составляет $1.23232 \cdot 10^{6}$.

2. О задаче идентификации параметров

В данном разделе дается строгая формулировка задачи идентификации и описываются проблемы, возникающие при ее решении.

2.1 Математическая модель пространственной структуры РНК

Отчетливо выделяются по крайней мере 3 типа моделей пространственной структуры нуклеиновых кислот:

- Атомарная модель. Относится к классу задач, решаемых с использованием методов молекулярной динамики. Молекула рассматривается как совокупность атомов (материальных точек), которые перемещаются друг относительно друга под воздействием межатомных сил.
- Нуклеотидная модель. Молекула представляется в виде цепочки элементов-нуклеотидов, связанных между собой потенциалами фосфодиэфирных и Уотсон-Криковских связей.
- Непрерывная модель. В данной модели молекулярная цепь рассматривается как тонкий непрерывный стержень с поперечными стяжками, соответствующими Уотсон-Криковским связям.

В работе [1] описана упрощенная модель в непрерывной постановке, где поперечные стяжки представляют собой абсолютно твердые палочки, а стержень в свободном состоянии имеет форму винтовой линии. При правильном выборе параметров модели она будет приближенно описывать пространственную форму молекулы нуклеиновой кислоты. Модели такого типа уже стали классическим средством для изучения пространственной формы молекул ДНК, для которой экспериментальным путем удалось определить упругие параметры ее нити. Для РНК такие параметры неизвестны.

Пространственная структура молекулы собирается из элементов (стеблей, петель), каждый из которых представляет собой замкнутый контур, состоящий из семейства упругих стержней, связанных жесткими перемычками.

2.2 Система уравнений модели

Выпишем систему уравнений исследуемой модели, описывающих равновесную форму упругого стержня.

$$\begin{cases} \frac{d}{ds}\left(B_1(\omega_1-\omega_1^0)\right)+B_3\omega_2(\omega_3-\omega_3^0)-B_2\omega_3(\omega_2-\omega_2^0)=0 \\ \frac{d}{ds}\left(B_2(\omega_2-\omega_2^0)\right)+B_1\omega_3(\omega_1-\omega_1^0)-B_3\omega_1(\omega_3-\omega_3^0)-F_3=0 \\ \frac{d}{ds}\left(B_3(\omega_3-\omega_3^0)\right)+B_2\omega_1(\omega_2-\omega_2^0)-B_1\omega_2(\omega_1-\omega_1^0)+F_2=0 \\ \frac{d}{ds}F_1+\omega_2F_3-\omega_3F_2=0 \\ \frac{d}{ds}F_2+\omega_3F_1-\omega_1F_3=0 \\ \frac{d}{ds}F_3+\omega_1F_2-\omega_2F_1=0 \\ \frac{d\vec{r}}{ds}=\vec{e}_1 \\ \frac{d\vec{e}_1}{ds}=\omega_3\vec{e}_3\times\vec{e}_1-\omega_2\vec{e}_3 \\ \frac{d\vec{e}_3}{ds}=\omega_1\vec{e}_1\times\vec{e}_3+\omega_2\vec{e}_1 \end{cases} \qquad (2.1)$$

Здесь $\vec{e}_1$ - касательный вектор к осевой линии стержня, $\vec{e}_3 \perp \vec{e}_1$, $\vec{e}_2 = \vec{e}_3 \times \vec{e}_1$ - образуют главный сопутствующий репер осевой линии стержня, совпадающий с репером Френе при повороте на угол θ вокруг $\vec{e}_1$; $\vec{r}$ - радиус-вектор осевой линии стержня; B_1 - крутильная жесткость, $B_{2,3}$ - изгибные жесткости стержня; F_i - компоненты силы, действующей на левое поперечное сечение стержня в точке s; ω_i - вектор Дарбу осевой линии стержня в изогну-

том состоянии; ω_i^0 - вектор Дарбу осевой линии стержня в свободном состоянии.

Для определения пространственной формы двуспиральных участков, где стержень в терминах модели находится в свободном состоянии, достаточно поставить задачу Коши со следующими начальными условиями:

$$\vec{r}(0)=\vec{r}_0,\quad \vec{e}_1(0)=\vec{e}_{10},\quad \vec{e}_3(0)=\vec{e}_{30},\quad \omega_i(0)=\omega_i^0, i=1,2,3,\quad F(0)=0$$

Здесь первое условие задает начальную точку стержня, второе и третье – ориентацию главного сопутствующего репера, четвертое и пятое фактически определяют момент и силу, приложенные к его левому концу.

Проинтегрировав уравнения модели с заданными начальными условиями, мы получим одну нить из двуспирального участка, представляющую собой винтовую линию. Вторая нить получается из первой поворотом вокруг оси винтовой линии на угол ϕ и сдвигом по этой же оси на расстояние ΔZ.

2.3 Параметры модели

Параметры данной математической модели естественным образом разбиваются на 4 группы:

- Упругие параметры стержня: жесткости B_i и угол θ;
- Геометрические параметры стержня для однонитевых участков, определяющие его форму в свободном состоянии: вектор Дарбу $(\omega_1^0, \omega_2^0, \omega_3^0)$;
- 3 координаты вектора Уотсон-Криковской связи в трехграннике Френе;
- Геометрические параметры стержня для двуспиральных участков: вектор Дарбу $(\omega_1^d, \omega_2^d, \omega_3^d)$.

Параметры ϕ и ΔZ, упомянутые в предыдущем пункте, можно вычислить, используя вектор Уотсон-Криковской связи $\vec{u}$, вектор направления оси двуспирального участка $\vec{e}$, радиус винтовой линии двуспирального участка r и вектор направления его оси $\vec{e}$. ΔZ есть проекция вектора $\vec{u}$ на вектор $\vec{e}$. Для определения угла ϕ найдем проекцию вектора $\vec{u}$ на плоскость с нормалью $\vec{e}$:

$$l=\sqrt{|\vec{u}|^2-\Delta Z^2} \qquad (2.2)$$

и воспользуемся теоремой косинусов:

$$\cos\phi=\frac{2r^2-l^2}{2r^2}. \qquad (2.3)$$

2.4 Постановка задачи идентификации

В данной работе мы опишем процесс идентификации 2-й группы параметров. Процесс идентификации происходит следующим образом.

В файле, описывающем молекулу tRNA, содержатся координаты атомов, входящих в состав молекулы, определенные путем рентгеноструктурного анализа. Из этого множества атомов нам необходимы только координаты атомов фосфора, содержащихся в рибозофосфатном остове молекулы, по одному на

каждый нуклеотид.

Для определения геометрических параметров стержня проще всего взять атомы фосфора из достаточно длинного стебля, так как в терминах модели силы и моменты на концах стебля равны нулю, упругие параметры поэтому не влияют на форму стержня и нет необходимости решать краевую задачу. Таким образом, для нашей цели необходимо взять атомы фосфора, лежащие на одной нити длинного стебля молекулы. После этой операции у нас появится набор точек (или ломаная линия) в пространстве, координаты которых есть координаты атомов фосфора (*физические* точки). Численное интегрирование уравнений модели даст также набор точек (*модельных* точек), описывающий форму стержня. Количество модельных точек, или звеньев модельной ломаной, будет отличаться от количества физических. Вариацией геометрических параметров модели необходимо добиться того, чтобы сгенерированная модельная ломаная проходила около каждой точки из набора по возможности ближе.

В качестве алгоритма вариации параметров был взят генетический алгоритм поиска экстремумов.

2.5 Генетический алгоритм поиска экстремумов

Генетические алгоритмы (ГА) - сравнительно новый класс вычислительных алгоритмов, предназначенных для поиска экстремумов функций. Для поиска экстремальных значений используется эволюционный метод поиска в пространстве решений, позаимствованный из живой природы. Рассмотрим работу типичного ГА.

Для работы ГА необходимы две вещи. Во-первых, необходимо установить взаимно однозначное соответствие между параметрами оптимизируемой задачи (*решениями*) и битовыми строками фиксированной длины (данный этап называют "конструированием хромосомы"). Во-вторых, требуется т.н. *fitness-функция*, или функционал оценки, на входе получающий хромосому и возвращающий некоторое число - тем больше, чем "лучше" в терминах задачи данное решение. Следует сразу же отметить, что для алгоритма важна лишь длина хромосомы. Что в ней находится реально, известно лишь функционалу. Имея сконструированную хромосому и функционал оценки, можно приступать к решению задачи.

Итак, пусть задана длина хромосомы n и fitness-функция. На начальном этапе работы генерируются k случайных чисел, имеющих разрядность n бит. Они оцениваются, и запускается классический оператор отбора – «рулетка с весами» (Рис. 2.1). Ее отличие от обычной рулетки заключается в том, что площадь ее k секторов пропорциональна приспособленностям наших k решений. Вероятность для попадания шарика в какой-либо сектор будет тем выше, чем больше его площадь. Запустив рулетку k раз, получим первое поколение, приспособленность которого в среднем (по вероятности) будет выше, чем у предыдущего. Но новых решений при этом не получается. Для генерации новых решений используются два оператора: кроссинговер и мутация.

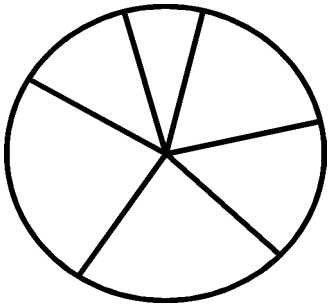


Рис. 2.1. «Рулетка с весами». Количество секторов равно количеству особей в популяции, их площади пропорциональны приспособленности особей.

Кроссинговер работает следующим образом. На вход подаются две хромосомы. Случайным образом выбирается позиция (номер бита от *1* до *n*), и начиная с этой позиции хромосомы меняются битами. На выходе имеются две новые хромосомы.

Оператор мутации работает с одной хромосомой. Случайно выбирается позиция, и бит в этой позиции инвертируется.

После кроссинговера и мутации происходит оценка получившихся решений, затем снова запускается рулетка. Критерием автоматической остановки является либо схождение алгоритма к одному решению (т.е. вся популяция решений становится одинаковой), либо достижение заданного количества итераций.

Построение хромосомы. В случае, когда у задачи имеются *n* параметров, для которых известны континуумные ограниченные множества возможных значений, хромосома строится следующим образом.

Прежде всего каждое множество значений дискретизируется. Задаются натуральные числа $d_i, i = 1...n$, обычно равные 2^k. Из *i*-го множества выбираются d_i точек, обычно же равномерно распределенные внутри множества. Затем точкам присваиваются порядковые номера.

Длина хромосомы задается как $\sum_{i=1}^{n}]\log_2 d_i[$, где $]d[$ - взятие целой части от d с избытком. Внутри хромосомы имеются *n* областей длины $]\log_2 d_i[$, содержащие двоичные числа, которые являются номерами точки из соответствующего множества значений параметра.

2.6 Задача сравнения двух ломаных

Как было сказано выше, для работы генетическому алгоритму необходим функционал оценки, в котором заложена информация о решаемой задаче. Генетический алгоритм в процессе генерации поколений будет искать его минимум. Данный функционал, получая на входе хромосому, должен декодировать генетическую информацию, заложенную в ней, получить численные значения параметров модели и решить задачу Коши с полученными значениями параметров. Далее необходимо получить численную оценку сходства пространственных форм полученной модельной ломаной и ломаной, вершинами кото-

рой являются атомы фосфора участка реальной молекулы. Задача оценки сходства двух ломаных осложняется тем, что, во-первых, они имеют различное количество вершин и различные длины, во-вторых, их взаимное расположение в пространстве с введенной системой координат далеко от наилучшего. Таким образом, видятся два различных подхода к получению данной оценки. Первый подход заключается в приведении двух ломаных в наилучшее взаимное положение, при котором, например, суммарное расстояние от вершин физической ломаной до модельной ломаной наименьшее. Второй подход состоит в построении функционала, который не использует координатное представление ломаных и оперирует с инвариантными относительно движений величинами. Ниже мы рассмотрим некоторые способы построения оценочных функционалов, использующие эти подходы.

Квазимеханический способ сопоставления кривых. Для приведения двух ломаных в наилучшее относительное расположение решается система неявных разностных уравнений:

$$\begin{cases} \vec{x}^{i+1} = \vec{x}^{i} + \tau \cdot \vec{N}(x^{i+1}, \vec{\phi}^{i+1}) \\ \vec{\phi}^{i+1} = \vec{\phi}^{i} + \tau \cdot \vec{M}(\vec{x}^{i+1}, \vec{\phi}^{i+1}) \end{cases}, \qquad (2.2)$$

где $\vec{N}(\vec{x},\vec{\phi}) = k\sum_{i=1}^{n} \frac{\vec{r}_i(\vec{x},\vec{\phi})}{|\vec{r}_i(\vec{x},\vec{\phi})|}$, $\vec{M}(\vec{x},\vec{\phi}) = l\sum_{i=1}^{n} \vec{e}_i(\vec{x},\vec{\phi}) \times \vec{r}_i(\vec{x},\vec{\phi})$, $\vec{x} = (cx, cy, cz)$ - координаты центра масс кривой, $\vec{\phi} = (\phi, \varphi, \theta)$ - углы поворота кривой вокруг осей Кенига, k, l – безразмерные коэффициенты, влияющие на скорость и характер сходимости процесса.

На начальном этапе работы алгоритма совмещаются центры масс кривой и набора. Затем у кривой вычисляются собственные векторы центрального тензора инерции, и кривая поворачивается до совпадения их с соответствующими векторами набора. Так как изначально неясно, какому вектору набора какой вектор кривой соответствует (возможны три случая), то рассматривают их все и выбирают тот, норма которого наименьшая.

Далее начинается итерационный процесс.

Вычисляются векторы $\vec{e}_i$, соединяющие центр масс кривой и ближайшие точки, и векторы $\vec{r}_i$, соединяющие ближайшие точки и соответствующие им точки набора (см. Рис. 2.2). Составляются суммы $\vec{N}(\vec{x},\vec{\phi})$, $\vec{M}(\vec{x},\vec{\phi})$, их значения подставляются в систему неявных разностных уравнений (2.2).

Система решается методом Эйлера. Для найденных значений координат центра масс и углов поворота вновь вычисляются суммы $\vec{N}$ и $\vec{M}$, и процесс повторяется. Критерием остановки служит одновременная малость скалярных произведений $(\vec{N},\vec{N})$ и $(\vec{M},\vec{M})$.

Процесс минимизирует потенциал $k\sum_{i=1}^{n} d_i$, где d_i - расстояние от i-й ближайшей точки до соответствующей ей точки набора.

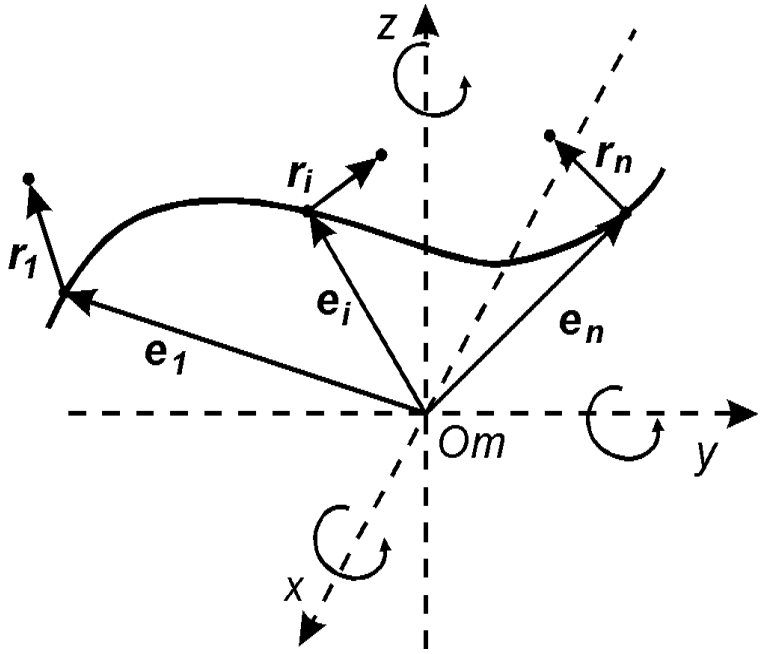


Рис. 2.2 Предварительное совмещение кривых производится квазимеханическим методом, при котором точки $\vec{e}_i$ модельной кривой как бы притягиваются к точкам $\vec{e}_i + \vec{r}_i$ физической кривой. Такое «притяжение» инициирует повороты вокруг координатных осей и смещение центра масс в направлении, уменьшающем рассогласование.

Данный способ вычисления меры близости пространственных форм двух ломаных имеет то преимущество, что получаемые в конце его работы ломаные линии близки в смысле координатного их представления, и можно, построив по координатам эти объекты, оценить визуально, насколько хорошо форма модельной ломаной описывает форму физической. Однако, отсюда же происходит и недостаток этого метода: непозволительно большое время расчета. Минимизируемый потенциал имеет так называемый "овражистый рельеф". Попадая в овраги, итерационный процесс надолго застревает в них и зачастую не доходит до минимума.

Дискретные аналоги кривизны и кручения. Рассмотрим пространственную ломаную линию $P_0, P_1, \ldots, P_n$, составленную из отрезков $a_{i+1} = P_i P_{i+1}$, $i = 0,1,\ldots,n-1$. Определим для этой ломаной две последовательности углов: ϕ_i, $i = 1,\ldots,n-1$, и θ_i, $i = 1,\ldots,n-2$.

Будем считать $\vec{a}_i$ векторами с началом в точке P_i и концом в точке P_{i+1}. Также примем, не ограничивая общности, что любые три последовательно идущие вершины ломаной P_{i-1}, P_i, P_{i+1} не лежат на одной прямой. Тогда два вектора, $\vec{a}_i$ и $\vec{a}_{i+1}$, определяют плоскость и ориентацию в ней. Определим угол ϕ_i как угол между векторами $\vec{a}_i$ и $\vec{a}_{i+1}$, отсчитываемый в положительном направлении согласно ориентации плоскости. Для данного угла будут выполняться неравенства $0 < \phi_i < \pi$.

Ориентированная плоскость, построенная выше, определяет единичную

нормаль $\vec{\beta}_i$. Построим еще одну плоскость, теперь уже на векторах $\vec{\beta}_i$ и $\vec{\beta}_{i+1}$, и зададим на ней ориентацию при помощи ортогонального ей вектора $\vec{a}_{i+1}$. Угол θ_i между векторами $\vec{\beta}_i$ и $\vec{\beta}_{i+1}$ будем отсчитывать в положительном направлении согласно этой ориентации. Тогда для этого угла будет выполнено $0 \le \theta_i \le 2\pi$. Назовем углы ϕ_i углами кривизны, а θ_i - углами кручения.

Имеет место следующая **теорема** (доказательство см. в [2]):

Пусть заданы три последовательности чисел: $s_1,\ldots,s_n$, $\bar{\phi}_1,\ldots,\bar{\phi}_{n-1}$, $\bar{\theta}_1,\ldots,\bar{\theta}_{n-2}$, такие, что $s_i > 0$, $0 < \bar{\phi}_i < \pi$, $0 \le \bar{\theta}_i \le 2\pi$. Пусть в пространстве заданы точка P_0, проходящая через нее ориентированная плоскость α и направление отрезка a_1. Тогда в пространстве существует единственная ломаная с отрезками a_i длины s_i, углами кривизны $\phi_i = \bar{\phi}_i$ и углами кручения $\theta_i = \bar{\theta}_i$.

Имея в виду данную теорему, можно построить для двух ломаных p^1 и p^2 с одинаковым количеством вершин n функционал с тремя весовыми коэффициентами a,b,c:

$$A(p^1, p^2) = a\sum_{i=0}^{n-1}(s_i^1 - s_i^2)^2 + b\sum_{i=1}^{n-1}(\phi_i^1 - \phi_i^2)^2 + c\sum_{i=1}^{n-2}(\theta_i^1 - \theta_i^2)^2, \qquad (2.3)$$

равный *0*, когда p^1 совпадает с p^2, и больший *0* в противном случае. Но как быть, когда количество вершин у ломаных различное?

Для решения этой проблемы можно применить алгоритм редуцирования количества вершин ломаной линии.

Итак, у нас имеются две ломаные линии: физическая, порожденная атомами фосфора из молекулы, и модельная, полученная путем численного интегрирования уравнений равновесия. Количество вершин модельной ломаной m намного превышает количество вершин n физической, к тому же эти ломаные имеют различные длины $l^{физ}$ и $l^{мод}$.Поделим вначале длины звеньев физической ломаной $s_i^{физ}$ на $l^{физ}$, получив относительные длины звеньев $r_i^{физ} = \dfrac{s_i^{физ}}{l^{физ}}$. Попробуем задать лежащие на модельной ломаной n новых вершин $P_i^{нов}$ так, чтобы выполнялись равенства $r_i^{физ} = r_i^{нов}$, где $r_i^{нов}$ вычисляются для ломаной с новыми вершинами $P_i^{нов}$ по тому же алгоритму, что и $r_i^{физ}$. Потребуем также, чтобы новые начальная и конечная вершины совпадали со старыми.

Зафиксировав некоторое значение параметра $\lambda > 0$, построим сферу S_1 радиуса $\lambda s_1^{физ}$ с центром в начальной точке модельной ломаной $P_0^{мод} = P_0^{нов}$. Пусть сфера пересечет модельную ломаную, возможно, в нескольких точках. Выберем из них ту, путь к которой от центра сферы по ломаной меньше. Обозначим эту точку за $P_1^{нов}(\lambda)$, отбросим мысленно часть ломаной, ограничен-

ной точками $P_0^{нов}$ и $P_1^{нов}(\lambda)$, и построим сферу S_2 радиуса $\lambda s_2^{физ}$ с центром в точке $P_1^{нов}$. Пусть, опять же, сфера пересечет модельную ломаную в нескольких точках (без учета отброшенной части). Снова отберем среди них одну по принципу, указанному выше, обозначим ее за $P_2^{нов}(\lambda)\ldots$ и будем продолжать так до тех пор, пока либо не построим все точки $P_i^{нов}(\lambda)$ количеством n, либо не обнаружим на шаге k, что остаток модельной ломаной целиком лежит внутри сферы S_k. Тогда мы проведем из точки $P_{k-1}^{нов}$ луч, направленный по вектору $P_{k-2}^{нов}P_{k-1}^{нов}$, и продолжим процесс, двигаясь далее по лучу вместо ломаной, пока не построим все n точек.

После проведения данного построения можно написать функцию

$$\xi(\lambda) = (P_{n-1}^{нов}P_n^{нов}, P_m^{мод}P_n^{нов}),\ 0 < \lambda < \infty, \qquad (2.4)$$

где $(\cdot)$ - скалярное произведение векторов, со следующими свойствами:

- $\xi(\lambda) > 0$, если точка $P_n^{нов}(\lambda)$ лежит вне модельной ломаной;
- $\xi(\lambda) < 0$, если точка $P_n^{нов}(\lambda)$ лежит внутри модельной ломаной;
- $\xi(\lambda) = 0$, если точка $P_n^{нов}(\lambda)$ совпадает с концом модельной ломаной.

Выберем отрезок $[\lambda_1;\lambda_2]$ такой, что $\xi(\lambda_1) < 0$ и $\xi(\lambda_2) > 0$. Если для любого λ из данного отрезка при построении точек $P_i^{нов}(\lambda)$ не встречаются ситуации, подобные показанной на Рис. 2.3, то можно утверждать, что:

- функция $\xi(\lambda)$ непрерывна на отрезке $[\lambda_1;\lambda_2]$;
- ее единственный нуль находится внутри отрезка.

Тогда нуль этой функции можно найти, например, методом деления отрезка пополам.

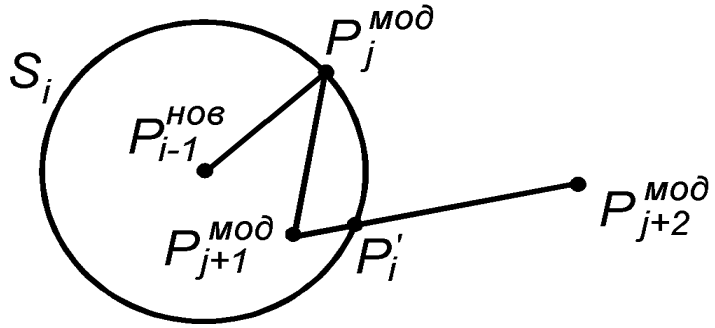


Рис. 2.3 Случай скачка функции $\xi(\lambda)$: $\left|P_{i-1}^{нов}P_j^{мод}\right| = \lambda s_i^{физ}$. При увеличении λ на $\varepsilon > 0$ произойдет скачок положения точки $P_i^{нов}$ на ломаной из положения $P_j^{мод}$ в точку $P_i^{'}$ пересечения отрезка $P_{j+1}^{мод}P_{j+2}^{мод}$ и сферы S_i.

Таким образом, из ломаной $P_0^{мод},\ldots,P_m^{мод}$ мы получили ломаную

$P_0^{нов},\ldots,P_n^{нов}$, с числом вершин, равным количеству вершин n физической ломаной, и дополнительным полезным свойством $r_i^{физ} = r_i^{нов}$. Теперь можно подсчитать значение функционала (2.3).

Построенный функционал замечателен тем, что он не зависит от взаимной ориентации ломаных и значений координат их вершин. Он оперирует лишь с их пространственными формами. В то же время функционал имеет и существенный недостаток: при вычислении его значения складываются величины различных размерностей.

Построение тетраэдров. Ломаная, полученная путем интегрирования уравнений модели, редуцируется до физической ломаной с меньшим числом звеньев по алгоритму, описанному в предыдущем пункте. Далее в обеих ломаных проводятся следующие построения:

- Измеряются расстояния между соседними точками, включая первую и последнюю (т.е. ломаная достраивается до замкнутого многозвенника).
- В получившемся многозвеннике проводятся все хорды, соединяющие точки, между которыми лежит одна вершина. Те же действия производятся для точек, между которыми лежат две вершины.
- Зафиксировав длины хорд и сторон в каждом многозвеннике, мы получим неизгибаемую конструкцию, состоящую из набора тетраэдров, построенных на вершинах многозвенника.

Вычислив сумму квадратов разностей между соответствующими хордами и сторонами, получим число, количественно выражающее различие между пространственными формами двух ломаных.

Функционал данного типа, сохраняя достоинство предыдущего - независимость от координатного представления ломаных, свободен от его недостатка: теперь все действия совершаются над величинами одинаковой размерности.

2.7 Определение параметров Уотсон-Криковской связи

В качестве параметров двухнитевого участка молекулы РНК, помимо вектора Дарбу $(\omega_1^d, \omega_2^d, \omega_3^d)$, можно указать также угол ϕ и смещение ΔZ, упомянутые в разделе 2.2. Их значения можно получить, используя знание вектора Уотсон-Криковской связи и вектора направления оси двуспирального участка, как было указано в разделе 2.3.

При построении моделей однонитевых участков молекулы РНК, каковыми являются петли различных типов, возникает необходимость в решении краевых задач для системы уравнений (2.1). Знание параметров Уотсон-Криковской связи необходимо для вычисления граничных условий этих краевых задач. Однако, на практике в этих двух случаях удобнее использовать непосредственно параметры ΔZ и ϕ. С помощью несложных геометрических построений можно получить из них и координаты вектора Уотсон-Криковской связи в главном сопутствующем репере стержня.

В данном разделе описывается алгоритм, позволяющий определить параметры ΔZ и ϕ и найти вектор оси двуспирального участка молекулы РНК.

Необходимо отметить, что Уотсон-Криковской связью мы будем называть отрезок, соединяющий два атома фосфора, которые принадлежат спаренным нуклеотидам молекулы.

Из рентгеноструктурных данных о третичной структуре молекулы РНК нам известны, в частности, координаты атомов фосфора, находящихся в двуспиральном участке молекулы. Так как модель двуспирального участка представляет собой две винтовые линии с общей осью, попробуем найти направление этой оси для данного участка. Для этого построим цилиндр бесконечной длины, сумма квадратов расстояний от атомов фосфора до поверхности которого минимальна. Его осевую линию будем считать осевой линией двуспирального участка.

Теперь можно найти параметры ΔZ и ϕ. Подсчитав среднее арифметическое проекций векторов Уотсон-Криковских связей в двуспиральном участке на осевую линию участка, мы получим смещение ΔZ. Для вычисления угла ϕ нам потребуется среднее арифметическое длин Уотсон-Криковских связей в стебле $\bar{s}$. Воспользуемся формулой (2.2), подставив в нее в качестве $|\vec{u}|$ среднюю длину $\bar{s}$, и получим среднюю длину проекции связи на плоскость, перпендикулярную оси цилиндра. Зная радиус цилиндра, вычислим значение ϕ, используя формулу (2.3).

Построение наилучшего цилиндра. Итак, у нас имеется множество точек $M = \{(x_i, y_i, z_i)\}$, $i = 1, \ldots, n$, для которого нужно построить бесконечный цилиндр таким образом, чтобы сумма квадратов расстояний от точек множества до цилиндра была минимальной.

Бесконечный цилиндр можно однозначно задать следующим набором параметров:

- вектор $\vec{e} = (A, B, C)$, $A^2 + B^2 + C^2 \neq 0$, определяющий направление оси цилиндра;
- точка (X, Y, Z), принадлежащая оси цилиндра;
- радиус цилиндра r.

Найдем выражение для суммы квадратов расстояний от точек из множества M до произвольного цилиндра. Вначале проведем плоскость π_i, проходящую через точку (x_i, y_i, z_i) и перпендикулярную оси цилиндра:

$$A(x - x_i) + B(y - y_i) + C(z - z_i) = 0 \quad (2.5)$$

Найдем координаты точки пересечения плоскости π_i с осью цилиндра. Для этого запишем уравнение для оси цилиндра в параметрической форме и подставим его в (2.5):

$$A^2 t + AX + B^2 t + BY + C^2 t + CZ - (Ax_i + By_i + Cz_i) = 0$$

Отсюда найдем нужные нам координаты (x^0, y^0, z^0):

$$x^0 = -A\frac{AX + BY + CZ - (Ax_i + By_i + Cz_i)}{A^2 + B^2 + C^2} + X;$$

$$y^0 = -B\frac{AX + BY + CZ - (Ax_i + By_i + Cz_i)}{A^2 + B^2 + C^2} + Y;$$

$$z^0 = -C\frac{AX + BY + CZ - (Ax_i + By_i + Cz_i)}{A^2 + B^2 + C^2} + Z.$$

Длина вектора $\vec{\eta}_i = (x_i - x^0; y_i - y^0; z_i - z^0)$ будет равна расстоянию от i-й точки множества M до оси цилиндра. Тогда сумма квадратов расстояний от точек множества до цилиндра будет выглядеть следующим образом:

$$S(A,B,C,X,Y,Z,r) = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (r - |\vec{\eta}_i|)^2$$

Аналитически из условия $\frac{\partial S}{\partial r} = 0$ можно найти оптимальное значение радиуса цилиндра:

$$\frac{\partial S}{\partial r} = \sum_{i=1}^{n} (2r - 2|\vec{\eta}_i|) = 0,$$

откуда $r_{опт} = \frac{\sum_{i=1}^{n} |\vec{\eta}_i|}{n}$, т.е. оптимальный радиус равен среднему арифметическому расстояний от точек из множества M до оси цилиндра. Точное нахождение оптимальных значений других параметров цилиндра представляет уже серьезную проблему. Задачу поиска минимума функции S можно решить численно. Для этого желательно еще больше уменьшить количество параметров.

Не ограничивая общности, можно наложить на вектор оси цилиндра $\vec{e}$ условие $A^2 + B^2 + C^2 = 1$. При этом вектор $\vec{e}$ будет определяться уже двумя параметрами. Попробуем найти эту зависимость, предполагая областью допустимых значений двух параметров *a* и *b* всю числовую прямую.

Введем в трехмерном пространстве декартову систему координат. Вокруг точки *O* построим сферу радиуса *1*. Через две точки, $(a,b,0)$ и $(0,0,1)$, проведем прямую, которая будет иметь со сферой две точки пересечения: $(0,0,1)$ и (A,B,C). Нужное свойство, $A^2 + B^2 + C^2 = 1$, следует из способа построения точки. Выпишем выражения для *A, B* и *C* через значения параметров *a* и *b*:

$$A = \frac{2a}{a^2 + b^2 + 1},\ B = \frac{2b}{a^2 + b^2 + 1},\ C = \frac{a^2 + b^2 - 1}{a^2 + b^2 + 1}.$$

При данном способе параметризации можно получить все возможные векторы $k\vec{e}$, $|k| = 1$, с точностью до знака коэффициента *k*, знание которого не нужно для задания направления оси цилиндра. Выражения для *A, B* и *C* являются непрерывными функциями от *a* и *b*.

Для задания точки, лежащей на оси цилиндра, также достаточно двух параметров *x* и *y*. Эти параметры представляют собой координаты на плоскости, перпендикулярной вектору $\vec{e}$. Зависимость трехмерных координат (X,Y,Z) этой точки от *x* и *y* можно найти следующим образом.

Если вектор $\vec{e}$ коллинеарен базисному вектору $\vec{e}_z$, то значения трехмерных координат получаются автоматически: $X = x$, $Y = y$, $Z = 0$. В противном случае найдем вектор $\vec{\xi} = \vec{e} \times \vec{e}_z$ и угол между единичными векторами $\vec{e}$ и $\vec{e}_z$:

$$\cos\phi = (\vec{e}, \vec{e}_z),\ \sin\phi = \left|\vec{\xi}\right|,$$

и запишем выражение:

$$(X,Y,Z)^T = Rot(\vec{\xi},\phi)\cdot(x,y,0)^T$$

Здесь $Rot(\vec{\xi},\phi)$ - матрица 3*x*3, реализующая поворот вокруг вектора $\vec{\xi}$ с началом в точке *O* на угол ϕ ([4]).

Таким образом, мы свели число параметров задачи до четырех, что значительно ускорило нахождение минимума функции S методом градиентного спуска.

Заметим, что функция расстояния непрерывно зависит от параметров цилиндра (ориентации и положения оси, радиуса). Пространство ориентации оси цилиндра компактно (RP^2). При неограниченном сдвиге оси, а также увеличения радиуса, функция расстояния неограниченно возрастает. Отсюда следует существование минимума у функции расстояния, т.е. минимальный цилиндр всегда существует. Однако единственность минимального цилиндра не гарантируется.

Вычисление граничных условий на примере шпилечной петли. Для нахождения формы шпилечной петли необходимо вычислить значения следующих величин:

$$r(0) = r_b,\quad r(L) = r_e,\quad e_1(0) = e_{b1},\quad e_1(L) = e_{e1},\quad e_3(0) = e_{b3},\quad e_3(L) = e_{e3}.$$

Здесь $\vec{r}(s)$-радиус-вектор, $\vec{e}_i(s)$ - главный сопутствующий репер осевой линии стержня, совпадающий с репером Френе при повороте на угол θ вокруг $\vec{e}_1$.

На Рис. 1.4 b) видно, что шпилечная петля находится на конце двуспирального участка молекулы. Следовательно, относительное расположение точек $\vec{r}_b$ и $\vec{r}_e$ должно совпадать с расположением концов двуспирального участка, а $\vec{e}_{b1}$, $\vec{e}_{b3}$, $\vec{e}_{e1}$ и $\vec{e}_{e3}$ задаются положениями главных сопутствующих реперов на концах стебля ([1]).

Пусть осевая линия двухспирального участка совпадает с осью *Oz*. Зададим координаты точки $\vec{r}_b = (-d/2,0,0)$. Главный сопутствующий репер в этой точке (вектора $e_1(0) = e_{b1}$ и $e_3(0) = e_{b3}$) нам известен из решения задачи Коши для одной нити двуспирального участка. Необходимо найти репер и положение конца второй нити. Для этого применим к точке $\vec{r}_b$ и векторам $\vec{e}_i(0)$ операцию поворота вокруг оси *Oz* на угол ϕ, а затем операцию параллельного переноса на вектор $(0,0,\Delta Z)$. В результате этих действий мы получим значения коорди-

нат точки $\vec{r}_e$ и векторов $\vec{e}_i(L)$.

Решение краевой задачи необходимо при проведении второго этапа идентификации – определения упругих параметров стержня.

3. Результаты вычислительных экспериментов

В данном разделе приводятся результаты некоторых вычислительных экспериментов по оценке геометрических параметров двуспиральных участков РНК на основе рентгеноструктурных данных.

При использовании генетических алгоритмов поиска экстремумов очень важным является правильный выбор параметров этого алгоритма. От этого зависит быстродействие процесса, скорость его сходимости, точность окончательного результата. Нами были опробованы различные вектора параметров. Наиболее сбалансированным с точки зрения указанных требований оказался следующий вектор параметров:

Число идентифицируемых параметров: 2 ($\omega_2^0 = Q$, $\omega_3^0 = R$)

Число особей в популяции: 70

Количество бит на параметр: 8

Вероятность кроссинговера: 1.0

Вероятность мутации: 0.08

Оператор кроссинговера: двухточечный

Оператор мутации: замена каждого бита на случайный с вероятностью мутации

Оператор селекции: турнирный.

В качестве алгоритма сравнения двух ломаных был взят квазимеханический алгоритм. В процессе работы генетического алгоритма минимизировалось максимальное расстояние между кривой и набором атомов фосфора.

Следует сказать несколько слов о корректности поиска экстремумов в данной задаче. Конечно, в задаче сопоставления геометрических форм пространственных кривых функционал сравнения может иметь не одну точку зкстремума. Следует ожидать, что их может быть достаточно много. Поэтому в строгом математическом смысле задача не является корректной (т. е. ее решение не единственно). Однако, наличие минимума у оценочного функционала можно гарантировать, т. к. он ограничен снизу на компактном пространстве параметров. Поэтому осмысленной является задача изменения геометрических параметров модели таким образом, чтобы оценочный функционал стал меньше при разумной форме и взаимном расположении сопоставляемых кривых в точке зкстремума. В силу сказанного результаты процесса идентификации контролировались нами для того, чтобы отбрасывать заведомо неподходящие варианты.

Ниже приводятся результаты наиболее удачных процессов идентификации параметров.

Для идентификации параметров выбран участок нити из двухнитевого стебля молекулы tRNA-Phe SACCHAROMYCES CEREVISIAE ([14]), состоя-

щая из пяти нуклеотидов. Идентифицировались параметры Q и R, изменяющиеся в пределах $[0.0;0.1]$. Для достижения минимума функционала генетический алгоритм сгенерировал порядка 300 поколений за время 2 часа (на машине класса Pentium166-MMX).

В результате идентификации для параметров были найдены следующие значения:

Q = 0.08981 R = 0.05098 func = 0.11727

Визуальный результат идентификации представлен на Рис. 3.1.

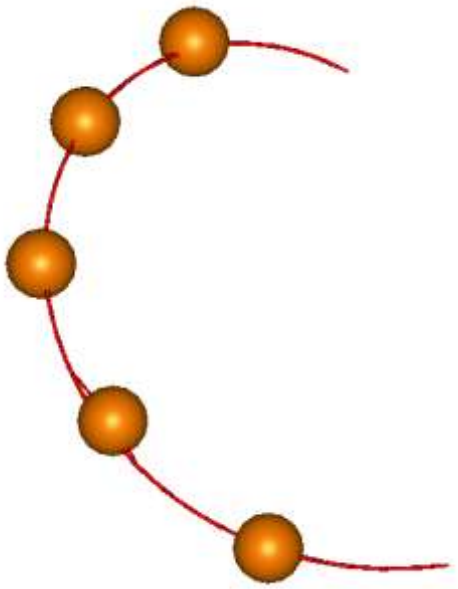


Рис. 3.1. Упругий стержень, аппроксимирующий участок из 5 нуклеотидов. Шарики соответствуют атомам фосфора, положение которых определено методом рентгеноструктурного анализа. Непрерывная линия соответствует нити двухспирального участка, полученной в результате процесса идентификации.

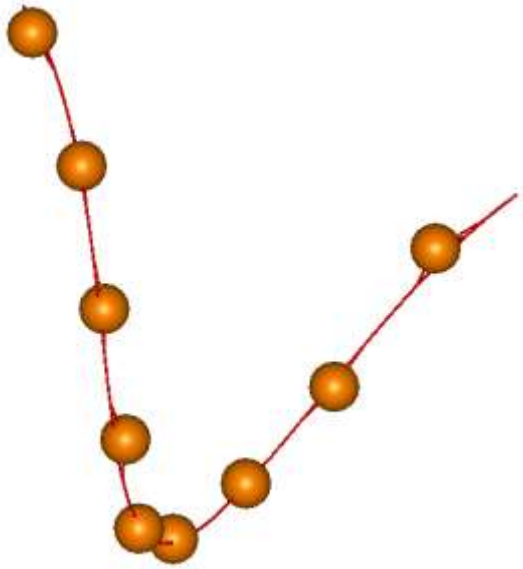


Рис. 3.2. Упругий стержень, аппроксимирующий участок из 9 нуклеотидов. Смысл изображения такой же, как на Рис. 3.1.

Из двухнитевого стебля той же молекулы был выбран участок, состоящий из девяти нуклеотидов. Для параметров Q и R и области их изменения $[0.0;0.1]$ минимум был найден спустя 90 поколений, сгенерированных за 4 часа. Значения параметров получились такие:

Q = 0.0845 R = 0.0375 func = 0.8072744

Представление о том, как выглядит форма стержня, можно получить из Рис. 3.2.

Литература

1. Козлов Н. Н, Кугушев Е. И., Сабитов Д. И., Энеев Т. М. Компьютерный анализ процессов структурообразования нуклеиновых кислот. Препринт ИПМ им М. В. Келдыша РАН, 2002, N 42.
2. Аминов Ю. А. Дифференциальная геометрия и топология кривых. Наука, 1987.
3. Ширшов А. Об одной комбинаторной задаче. «Квант» № 9, 1979, с. 19
4. Бронштейн И. Н., Семендяев К. А. Справочник по математике для инженеров и учащихся втузов. - 13-е изд., исправленное. Наука, 1986.
5. Батищев Д. И. Генетические алгоритмы решения экстремальных задач / Под ред. академика АЕН Львовича Я. Е.: Учеб. пособие. Воронеж. гос. техн. ун-т; Нижегородский гос. ун-т. Воронеж, 1995.
6. Энеев Т.М., Козлов Н.Н., Кугушев Е.И. Процессы структуризации биомолекул. Результаты математического моделирования. Препринт ИПМ им. М.В. Келдыша РАН, N 69, 1995, с. 22.
7. Козлов Н.Н., Кугушев Е.И., Энеев Т.М. Структурообразующие характеристики транскрипционного процесса. Математическое моделирование т.10, N 6, с.3-19, 1998.
8. www.moldyn.ru/library/md/default.htm
9. Tomb et al. The complete genome sequence of the gastric pathogen Helicobacter pyori. Nature 1997, Aug 7; 388 (6642).
10. Brown R. S., Dewan J. C., Klug A. Crystallographic and biochemical investigation of the lead-catalyzed hydrolysis of yeast phenilalanine tRNA. Biochemistry. v. 24, 1985.
11. http://www.tulane.edu/~biochem/nolan/lectures/rna/rnapro02.htm
12. http://wwwmgs.bionet.nsc.ru/mgs/gnw/selex
13. Попов Е.П. (1948) Нелинейные задачи статики тонких стержней. Л.М. ОГИЗ, с.170.
14. Benham C.J. (1983) Geometry and mechanics of DNA superhelicity. Biopolymers, **22**, 11, 2477 - 2495.