Правительство Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего профессионального образования
«Национальный исследовательский университет
«Высшая школа экономики»

Факультет Бизнес-информатики
Отделение Прикладной математики и информатики
Базовая кафедра Яндекс

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА
на тему

Классификация элементов вторичной структуры РНК

Выполнил студент группы 471
Баулин Евгений Федорович

Научный руководитель:
Зав. Кафедрой, д.ф.-м.н.
Ройтберг Михаил Абрамович

Москва 2013

Оглавление

Введение

Изучение вторичной структуры РНК, в частности, предсказание вторичной структуры по последовательности путем минимизации энергии, - одна из классических задач биоинформатики [1-2]. В рамках общепринятой в настоящее время модели Цукера-Мэтьюза-Тернера (Nearest Neighbour Model, NNM, см. [3-4]) структура РНК разбивается на петли (loops). Каждая петля относится к определенному классу, для каждого класса дается формула вычисления энергии петли, зависящая от длин соответствующих участков РНК, нуклеотидов и т.п.

К сожалению, эта модель неприменима при изучении структур, которые содержат псевдоузлы. Проблема предсказания структур РНК, которые содержат псевдоузлы, изучена значительно хуже, чем проблема предсказания «классических» структур, не содержащих псевдоузлов (pseudoknot-free structures) [5]. При этом в настоящее время не существует общепринятой классификации псевдоузлов. Описаны простейшие виды псевдоузлов (kissing hairpins, H-structures и др., см. [6]), псевдоузлы, играющие важную роль в клеточных процессах, см. например, [7]; в [8] дан анализ классов псевдоузлов, введенных в различных работах по предсказанию вторичной структуры РНК по ее последовательности. В базе данных [9, 10] собраны примеры псевдоузлов в экспериментально определенных структурах РНК. Однако единого способа описания элементов вторичной структуры при наличии псевдоузлов, подобного тому, который был предложен в [3-4] для описания структур, которые не содержат псевдоузлов, в настоящее время нет.

Цель настоящей работы – восполнить этот пробел. Мы предлагаем способ описания элементов вторичной структуры, пригодный как для описания классических петель, так и для описания псевдоузлов. Этот способ основан на разложении плоских графов в элементарные циклы, ср. с работой

[11]. Основной принцип предложенной классификации состоит в том, что каждой спирали ставится в соответствие отдельная петля. Таким образом, мы избегаем проблемы идентификации класса петли в связи с наличием пересечений между спиралями. Однако при таком подходе нарушается однозначность принадлежности однонитевых участков, то есть отдельно взятая нить может относиться к более чем одной петле, в отличие от модели Мэтьюза-Тернера.

Дополнительный интерес данной работы состоит в анализе взаимосвязи между вторичной структурой РНК и РНК-Белковыми взаимодействиями. На основе предложенного способа классификации элементов вторичной структуры РНК была разработана схема базы данных, содержащая всю информацию, необходимую для анализа структур РНК. В качестве исходных данных были использованы документы банка данных белковых структур (Protein Data Bank, PDB, version 3.3), содержащие РНК. Для разметки вторичной структуры цепей РНК, представленных в выбранных документах, была использована функция find_pair из программного пакета X3DNA (www.x3dna.org, version 1.5) [12].

В рамках работы по созданию описанной базы данных на языке программирования Python (version 3.2.1) нами была разработана программа, выполняющая полный цикл переработки исходных данных в формат, соответствующий предложенной классификации и пригодный для дальнейших исследований.

Такая впервые созданная база найдет широкое применение в исследованиях (в частности, изучение достоверности и полноты предсказаний; анализ закономерностей РНК-Белковых взаимодействий).

В настоящее время для тестирования базы данных разрабатывается веб-интерфейс, который поможет выявить основные требования к формату

представленной информации для наиболее эффективного её применения. Бета-версия веб-интерфейса доступна онлайн любому пользователю и располагается по адресу http://server2.lpm.org.ru/~baulin/home.html.

В процессе создания описанной базы данных в рамках предварительного анализа нами был поставлен вопрос о существовании такой вторичной структуры РНК, которая не может быть представлена в виде планарного графа. Проведённые исследования показали, что такие структуры имеют место в экспериментально определенных структурах. Все такие структуры относятся к фрагменту 23S (в единичных случаях 25S и 26S) рибосомальных РНК различных организмов.

Работа имеет следующую структуру. Глава 1 описывает предлагаемую нами классификацию элементов вторичной структуры РНК. В разделе 1.1 «Основные определения» вводится вся необходимая терминология. В разделе 1.2 «Спирали и петли» вводится основное для предлагаемого подхода понятие петли, обобщающее понятие петли по Мэтьюзу-Тернеру. Далее, в разделе 1.3 «Структура петель» доказываются утверждения, позволяющие установить общий вид петель и ввести их классификацию, которая согласована с классификацией Мэтьюза-Тернера. Глава 2 «База данных» описывает предлагаемую схему базы данных и объясняет ключевые моменты её создания. В главе 3 «Предварительные результаты» представлены выводы предварительного анализа изучаемых структур. В «Заключении» обсуждаются перспективы практического применения предложенных определений.

Глава 1. Теоретические основы предлагаемой модели описания вторичной структуры РНК

1.1 Основные определения

Молекулу РНК мы будем представлять, как последовательность нуклеотидов, иначе говоря, как символьную последовательность в алфавите {A, C, G, U}. Каждый нуклеотид в молекуле имеет свой номер от 1 до L, где L – длина последовательности.

Связь – это пара нуклеотидов (i, j), где i < j, которая образует водородную связь. При этом допускаются не только связи между комплементарными нуклеотидами (Watson-Crick pairs) и G-U связи (Wobble pairs), но и неканонические связи, см. [12]. Причем комплементарными нуклеотидами являются пары A-U и G-C.

Спираль – это последовательность пар нуклеотидов вида (i, j), (i+1, j-1),…, (i+k, j-k) такая, что

1) i < j, i+k < j-k, k > 1;
2) все пары вида (i+x, j-x), где x = 0, …, k, образуют связи, причём связи (i, j) и (i+k. j-k) – Уотсон-Криковские связи (УК-связи), т.е. связи между комплементарными нуклеотидами, или G-U связи.

Участок цепи [i, i+k] будем называть *левым крылом* спирали, соответственно участок [j-k, j] будем называть *правым крылом* спирали.

Пару (i, j) будем называть *внешней парой* спирали или *торцом* спирали, пару (i+k, j-k) будем называть *внутренней парой* спирали.

Замечание. В популярной программе 3DNA [12] используется другое определение спирали (основанное исключительно на геометрических параметрах цепи). Далее в случае элементов согласно 3DNA мы будем использовать термины *Д-спираль* и *изолят* для описания спирали и

одиночного спаривания соответственно. Описанный ниже подход применим к любому определению спирали. Говоря неформально, мы считаем, что на последовательности РНК каким-то образом уже размечены левые и правые крылья спиралей и установлено соответствие между крыльями одной спирали.

Нить – это такой участок цепи [i, j], где i < j, что

1) не существует такой связи (k, t), что i <= k <= j или i <= t <= j.

2) существуют пары, в которые входят нуклеотиды i-1 и j+1.

Замечание. Допускаются нити «нулевой длины», для их обозначения используется запись [i+1,i], где i – номер последнего нуклеотида предшествующего крыла.

Вторичная структура РНК – это такое множество связей, что

1) каждый нуклеотид входит не более чем в одну связь;

2) каждая пара входит в некоторую спираль.

Отметим, что в экспериментально определенных пространственных структурах РНК есть значительное число водородных связей, не входящих в спирали [13], роль таких связей в настоящее время изучена слабо. Мы исходим из предположения, что полезно отдельно рассматривать «базовую» вторичную структуру, образованную спиралями, и (над этой структурой) – одиночные водородные связи, называемые *линками*.

Линк – одиночное спаривание (i, j), не являющееся частью спирали.

Будем говорить, что две спирали (спираль и линк, два линка) находятся *в конфликте,* если между крыльями одной спирали (линка) находится одно, и только одно крыло другой спирали (линка).

Псевдоузел – участок вторичной структуры, содержащий хотя бы одну пару спиралей, находящихся в конфликте друг с другом.

По наличию конфликтов линки делятся на три типа. Линк называется *внутренним*, если он не конфликтует с другими спиралями (линками); *связанным*, если он конфликтует с линками, но не со спиралями; *свободным*, если он конфликтует со спиралями.

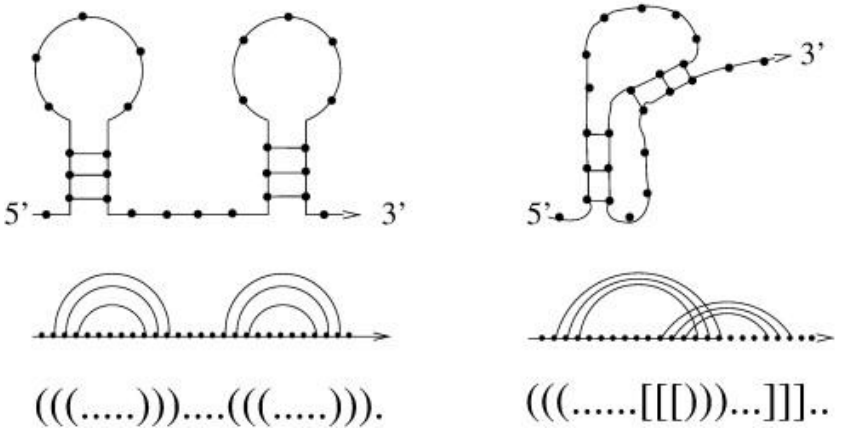

Рисунок 1.1 Пример структуры без псевдоузлов (слева) и с псевдоузлами (справа)

1.2 Спирали и петли

Здесь и далее будем считать фиксированной цепь РНК с заданной на ней вторичной структурой. Эту цепь можно рассматривать как чередующуюся последовательность нитей и крыльев. Для удобства изложения мы будем считать, что перед первым и после последнего нуклеотида цепи добавлены крылья «внешней спирали».

С каждой спиралью связан *внутренний* по отношению к ней участок цепи – участок между концом левого крыла и началом правого крыла, иначе говоря, - между нуклеотидами, образующими внутреннюю пару спирали. Для фиктивной внешней спирали внутренним участком является вся исходная последовательность РНК.

Пусть H – спираль и (i, j) – ее внутренняя пара.

Определение 1. Позиция цепи *t – внутренняя* для спирали H (синоним: *лежит внутри* H) , если *i<t<j*. Фрагмент цепи - *внутренний* для спирали H (синоним: *лежит внутри* H) , если все его позиции – внутренние для спирали H. Спираль H1 *лежит внутри* спирали H (является *внутренней* для H), если все позиции ее крыльев – внутренние для H.

Определение 2. Позиция цепи *t принадлежит* спирали H, если она внутренняя для H и не существует спирали H1, лежащей внутри H, такой, что *x<t< y*, где *(x, y)* – внешняя пара (торец) H.

Определение 3. *Петля* спирали H – это множество всех позиций, которые принадлежат спирали H.

Очевидно, каждая позиция, не входящая в связь, принадлежит хотя бы одной петле – обычной или внешней. При этом если какая-то позиция нити (крыла) принадлежит некоторой петле, то и вся нить (все крыло) принадлежит этой петле.

Если в структуре нет псевдоузлов, то каждая петля в смысле определения 3 является петлей по Мэтьюзу-Тернеру и наоборот. При этом каждая нить принадлежит ровно одной петле (возможно, внешней), а ни одно крыло не принадлежит какой-либо петле. Для структур с псевдоузлами оба эти свойства нарушаются.

1.3 Структура петель

Определение 4. Пусть H – спираль и (u, v) – ее внутренняя пара. Участок, [i, j] называется *замкнутым* относительно H**,** если

1) [i, j] лежит внутри H;

2) не существует таких связей (k, t), что *(i <= k <= j < t < v*) или *(u <k < i <= t <= j);*

3) существуют связи (i, .k) и (t, j), где *k <= j; i<= t.*

4) не существует отличного от [i, j] участка [i’, j’] такого, что i<= i’ < j’<=j и участок [i’, j’] удовлетворяет условиям 1) - 3).

Пара нуклеотидов (i, j) называется *торцом* замкнутого относительно H участка.

Утверждение 1. Пусть *Z = [f, g]* – участок, замкнутый относительно спирали H; *(u, v)* – внутренняя пара спирали H. Тогда:

1. Участок *Z* целиком лежит внутри спирали H.
2. Крыло либо целиком лежит в *Z*, либо целиком лежит вне *Z*.
3. Замкнутый относительно H участок начинается левым крылом некоторой спирали H1, лежащей внутри H, и заканчивается правым крылом некоторой спирали H2, лежащей внутри H.
4. Если H1=H2 – это одна и та же спираль, то торец *(s, t)* участка *Z* – это торец спирали *H*. В противном случае *s* – это начало левого крыла спирали H1, *t* – это конец правого крыла спирали H2.

   Доказательство – следует из определения 4 и того, что крылья не пересекаются.

Определение 5. Пусть *Z* – это участок, замкнутый относительно спирали H. Участок *Z* называется *простым*, если его торец – это торец некоторой спирали и *сложным* в противном случае. Сложные участки для краткости будем называть *блоками*.

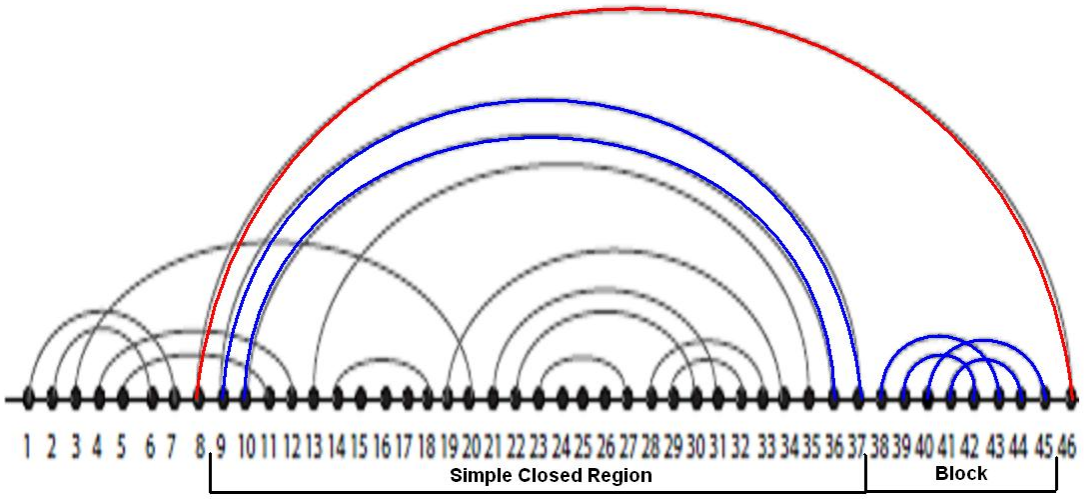

Рисунок 1.2 Примеры закрытых участков

Утверждение 2. Пусть H – спираль; *(u, v)* – ее внутренняя пара. Тогда

1. Никакие два участка, замкнутых относительно H, не пересекаются.
2. Пусть позиция *t* лежит внутри спирали H. Позиция *t* НЕ принадлежит спирали H тогда и только тогда, когда *t* лежит внутри некоторого участка *Z*, замкнутого относительно H (т.е. лежит в *Z*, но не входит в его торец).

Доказательство – следует из определений 1, 2 и 4.

Определение 6. Пусть H – спираль и *(u, v)* – ее внутренняя пара. Пусть $(s_1, t_1), \ldots, (s_n, t_n)$ – торцы всех участков, замкнутых относительно H; $s_1 < t_1 < \ldots < s_n < t_n$. Для удобства пусть $t_0 = u$; $s_{n+1} = v$. Пусть k – целое; $1 <= k <= n+1$. Тогда k-я *грань* петли H – это фрагмент $[t_{k-1}+1, s_k-1]$.

Замечание. Если $s_k = t_{k-1}+1$, то k-я грань петли H – пустой отрезок.

Утверждение 3. Пусть H – спираль и *(u, v)* – ее внутренняя пара. Пусть $(s_1, t_1), \ldots, (s_n, t_n)$ – торцы всех участков, замкнутых относительно H; $s_1 < t_1 < \ldots < s_n < t_n$. Для удобства пусть $t_0 = u$; $s_{n+1} = v$. Тогда петля спирали H – это

объединение торцов всех участков, замкнутых относительно H, и расположенных между ними граней.

Доказательство – следует из утверждения 2.

Утверждение 4. Пусть H – спираль и *(u, v)* – ее внутренняя пара и позиция *x* принадлежит грани *(t, s)* петли спирали H. Тогда

1. Позиция *x* либо не участвует в связи, либо принадлежит крылу спирали H', другое крыло которой лежит вне спирали H.
2. Если *x* принадлежит нити (крылу спирали), то все позиции этой нити (этого крыла) принадлежат той же грани петли спирали H.

Доказательство – следует из определения граней и того, что крылья не пересекаются.

Утверждения 3 и 4 описывают возможные структуры петель. Отметим, что в случае структур, которые не содержат псевдоузлов, все замкнутые участки – простые и каждая грань состоит из единственного однонитевого участка. Поэтому можно дать такое определение.

Определение 7. Петля называется *классической*, если она не содержит крыльев и торцов блоков. Петля называется *изолированной*, если она не содержит крыльев. и *узловой***,** если она содержит крылья.

Спираль называется *узловой*, если ее петля – узловая.

Применим классификацию петель по Мэтьюзу-Тернеру к введенному нами обобщению, основываясь на количестве торцов, входящих в петлю. Отметим, что в нашем случае торцы могут быть как торцами спиралей (иными словами – простых замкнутых участков), так и концами блоков (сложных замкнутых участков).

Определение 8. Петля называется *тупиковой* (hairpin), если она не содержит торцов и, соответственно имеет одну грань. Петля называется *внутренней* (internal), если она содержит ровно один торец, и, соответственно, имеет две грани. Петля называется *ветвящейся* (multiple), если она содержит более одного торца, и, соответственно, более двух граней.

Замечание 1. Будем называть выпуклостью (Buldge) такую внутреннюю петлю, одна из граней которой является нитью нулевой длины.

Замечание 2. Данная классификация распространяется как на обычные, так и на внешние петли (принадлежащие «внешним» спиралям).

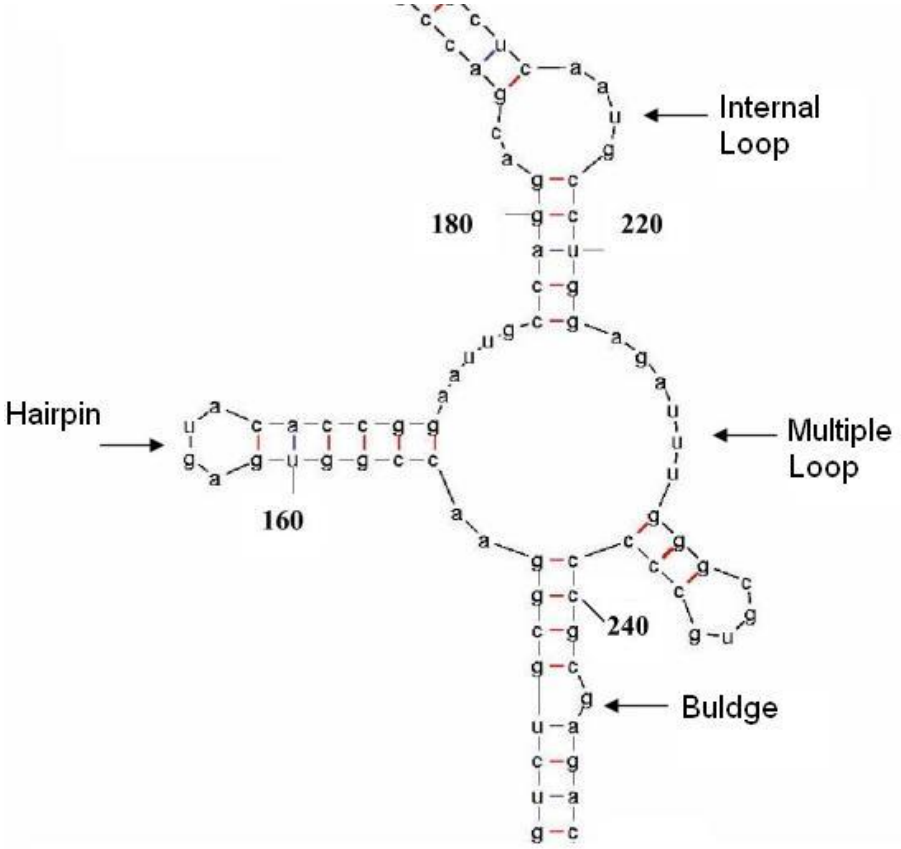

Рисунок 1.3. Различные типы петель

1.4 Выводы и результаты по главе

В данной главе даётся описание предлагаемой нами классификации элементов вторичной структуры РНК, а именно: вводится вся необходимая терминология, включая основное для предлагаемого подхода понятие петли, обобщающее понятие петли по Мэтьюзу-Тернеру, а также доказываются утверждения, позволяющие установить общий вид петель и согласовать новую классификацию с моделью NNM.

В главе вводятся принципиально новые понятия, такие как *линк, грань, блок* и др., которые позволяют обобщить модель Мэтьюза-Тернера на случай псевдоузловых структур. Также, наглядно доказано, что при отсутствии псевдоузлов представленные классификации полностью совпадают.

Глава 2. База данных

2.1 Исходные данные

В качестве исходных данных были выбраны документы из базы пространственных структур PDB (Protein Data Bank, www.rcsb.org, version 3.3). Среди всех документов были выбраны два класса структур: содержащие только РНК (907 документов) и содержащие РНК-Белковые комплексы (1239 документов). Поскольку некоторые структуры были представлены в одном и том же документе в нескольких вариациях, все документы были разделены по принципу один файл – одна вариация структуры (т.н. модель). Таким образом, в первоначальной выборке содержалось 7610 моделей из 2146 документов.

```
HEADER    RNA                                     10-MAY-97   1AJU
TITLE     HIV-2 TAR-ARGININAMIDE COMPLEX, NMR, 20 STRUCTURES
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: TAR RNA;
COMPND   3 CHAIN: A;
COMPND   4 SYNONYM: TAR;
COMPND   5 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 SYNTHETIC: YES
KEYWDS    COMPLEX (RIBONUCLEIC ACID/LIGAND), NMR, TRANSCRIPTIONAL
KEYWDS   2 ACTIVATION, PROTEIN-RNA INTERACTIONS
EXPDTA    SOLUTION NMR
NUMMDL    20
AUTHOR    A.S.BRODSKY,J.R.WILLIAMSON
REVDAT   2   24-FEB-09 1AJU    1       VERSN
REVDAT   1   17-DEC-97 1AJU    0
JRNL        AUTH    A.S.BRODSKY,J.R.WILLIAMSON
JRNL        TITL    SOLUTION STRUCTURE OF THE HIV-2 TAR-ARGININAMIDE
JRNL        TITL 2 COMPLEX.
JRNL        REF     J.MOL.BIOL.                   V. 267   624 1997
JRNL        REFN                  ISSN 0022-2836
JRNL        PMID    9126842
JRNL        DOI     10.1006/JMBI.1996.0879
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH    A.GELBIN,B.SCHNEIDER,L.CLOWNY,S.-H.HSIEH,W.K.OLSEN,
REMARK   1  AUTH 2 H.M.BERMAN
REMARK   1  TITL    GEOMETRIC PARAMETERS IN NUCLEIC ACIDS: SUGAR AND
REMARK   1  TITL 2 PHOSPHATE CONSTITUENTS
REMARK   1  REF     J.AM.CHEM.SOC.                V. 118   519 1996
REMARK   1  REFN                  ISSN 0002-7863
REMARK   1 REFERENCE 2
REMARK   1  AUTH    F.ABOUL-ELA,J.KARN,G.VARANI
REMARK   1  TITL    THE STRUCTURE OF THE HUMAN IMMUNODEFICIENCY VIRUS
REMARK   1  TITL 2 TYPE-1 TAR RNA REVEALS PRINCIPLES OF RNA
REMARK   1  TITL 3 RECOGNITION BY TAT PROTEIN
ATOM      5  C3'    G A  16       1.702  24.452  -0.456  1.00  0.00           C
ATOM      6  O3'    G A  16       3.108  24.240  -0.350  1.00  0.00           O
ATOM      7  C2'    G A  16       0.997  24.283   0.884  1.00  0.00           C
ATOM      8  O2'    G A  16       1.736  24.844   1.953  1.00  0.00           O
ATOM      9  C1'    G A  16      -0.269  25.109   0.648  1.00  0.00           C
ATOM     10  N9     G A  16      -1.379  24.330   0.102  1.00  0.00           N
ATOM     11  C8     G A  16      -1.778  24.256  -1.211  1.00  0.00           C
ATOM     12  N7     G A  16      -2.803  23.471  -1.397  1.00  0.00           N
ATOM     13  C5     G A  16      -3.102  22.995  -0.128  1.00  0.00           C
ATOM     14  C6     G A  16      -4.120  22.099   0.305  1.00  0.00           C
ATOM     15  O6     G A  16      -4.983  21.533  -0.377  1.00  0.00           O
ATOM     16  N1     G A  16      -4.075  21.879   1.685  1.00  0.00           N
ATOM     17  C2     G A  16      -3.154  22.457   2.535  1.00  0.00           C
ATOM     18  N2     G A  16      -3.260  22.130   3.831  1.00  0.00           N
ATOM     19  N3     G A  16      -2.201  23.293   2.142  1.00  0.00           N
ATOM     20  C4     G A  16      -2.233  23.516   0.809  1.00  0.00           C
ATOM     21  H5'    G A  16       2.385  26.089  -2.694  1.00  0.00           H
ATOM     22 H5''    G A  16       0.724  25.484  -2.799  1.00  0.00           H
ATOM     23  H4'    G A  16       2.164  26.512  -0.364  1.00  0.00           H
ATOM     24  H3'    G A  16       1.343  23.755  -1.213  1.00  0.00           H
ATOM     25  H2'    G A  16       0.758  23.233   1.055  1.00  0.00           H
ATOM     26 HO2'    G A  16       2.452  24.241   2.152  1.00  0.00           H
ATOM     27  H1'    G A  16      -0.616  25.599   1.558  1.00  0.00           H
ATOM     28  H8     G A  16      -1.291  24.794  -2.010  1.00  0.00           H
ATOM     29  H1     G A  16      -4.760  21.255   2.086  1.00  0.00           H
ATOM     30  H21    G A  16      -3.982  21.492   4.133  1.00  0.00           H
ATOM     31  H22    G A  16      -2.615  22.519   4.504  1.00  0.00           H
ATOM     32 HO5'    G A  16       0.014  27.588  -2.215  1.00  0.00           H
ATOM     33  P      G A  17       3.692  22.744  -0.259  1.00  0.00           P
```

Рисунок 2.1. Фрагмент pdb-файла.

В процессе работы часть структур была отброшена в силу различных причин, таких как отсутствие вторичной структуры РНК, конфликт с используемым нами сторонним ПО и др. Также часть документов была отложена в связи с возникшими трудностями их обработки (например, наличие конфликтующих спиралей между различными цепями РНК). В дальнейшем планируется активное сокращение числа отложенных документов. Таким образом, на данный момент проанализировано 6716 моделей структур из 1674 документов. Всего проанализировано 3169 цепей РНК (без учета представления одной цепи в нескольких моделях).

Далее была поставлена задача описать вторичную структуру РНК. Во-первых, была проведена проверка нуклеотидов на наличие более чем одного спаривания. Структур с подобными нуклеотидами не оказалось. Во-вторых, все оставшиеся документы были разделены по наличию различных особенностей вторичной структуры на три типа:

1) содержащие пары цепей РНК, имеющие общую вторичную структуру;
2) содержащие цепи РНК, во вторичной структуре которых присутствуют конфликтующие друг с другом спирали (т.н. псевдоузлы);
3) документы без особенностей.

Таблица 2.1 Категории представленных документов.

Категория документов (основная характеристика)	Количество моделей		Статус
	РНК	РНК+Белок	
Отсутствует вторичная структура	38	574	Отброшены
Конфликт со сторонним ПО	36	41	Отложены
Конфликт парсинга	47	94	Отложены
Присутствуют псевдоузлы	310	329	Обработаны
Межцепочечные псевдоузлы	5	1	Отложены
Пары цепей с общей вторичной структурой	1063	174	Обработаны
Общая вторичная структура у трех и более цепей	5	53	Отложены
Без осложнений	3760	1080	Обработаны
Всего обработано	5133	1583	
	6716		
ВСЕГО	5264	2346	
	7610		

Для разметки водородных связей, образующих вторичную структуру РНК, была использована функция find_pair из пакета инструментов X3DNA (www.x3dna.org, version 1.5) [12]. С её помощью были получены out-файлы, содержащие информацию о спариваниях между нуклеотидами и о спиралях, образованных данными спариваниями.

```
S:\PDB\R\models\1A9L.pdb2
S:\PDB\R\models\1A9L.out
    2          # duplex
   14          # number of base-pairs
    1    1     # explicit bp numbering/hetero atoms
    1   38  0 #    1 | A:...1_:[..G]G-----C[..C]:..38_:A  0.86  0.09  4.80  8.98 -0.47
    2   37  0 #    2 | A:...2_:[..G]G-----C[..C]:..37_:A  1.10  0.11 12.06  8.81 -0.18
    3   36  0 #    3 | A:...3_:[..G]G-----C[..C]:..36_:A  0.95  0.20 12.17  8.91 -0.15
    4   35  0 #    4 | A:...4_:[..U]U-----A[..A]:..35_:A  0.53  0.29 11.35  8.85 -0.38
    5   34  0 #    5 | A:...5_:[..G]G-----C[..C]:..34_:A  0.24  0.22 23.56  9.07 -0.81
    6   33  9 #    6 x A:...6_:[..A]A-----U[..U]:..33_:A  0.81  0.06 22.23  8.74 -0.57
    7   29  0 #    7 | A:...7_:[..C]C-----G[..G]:..29_:A  0.56  0.27 16.77  9.02 -0.39
    8   28  0 #    8 | A:...8_:[..U]U-----A[..A]:..28_:A  0.29  0.12 11.47  9.09 -0.97
    9   27  0 #    9 | A:...9_:[..C]C-----G[..G]:..27_:A  1.22  0.22 12.26  8.74  0.16
   10   26  9 #   10 x A:..10_:[..C]C-----G[..G]:..26_:A  0.37  0.32 11.19  9.15 -0.49
   14   25  0 #   11 | A:..14_:[..G]G-----C[..C]:..25_:A  0.74  0.72 13.31  9.10  0.68
   15   24  0 #   12 | A:..15_:[..G]G-----C[..C]:..24_:A  1.05  0.17 11.99  8.88 -0.10
   16   23  0 #   13 | A:..16_:[..U]U-----A[..A]:..23_:A  0.63  0.59  2.65  8.86  0.31
   17   22  0 #   14 | A:..17_:[..C]C-----G[..G]:..22_:A  0.78  0.40 13.33  8.95  0.08
##### Base-pair criteria used:   4.00 15.00  2.50 65.00  4.50  7.50
##### 0 non-Watson-Crick base-pairs, and 3 helices (0 isolated bps)
##### Helix #1 (6): 1 - 6
##### Helix #2 (4): 7 - 10
##### Helix #3 (4): 11 - 14
```

Рисунок 2.2 Пример out-файла

2.2 Обработка информации

Для обработки исходных данных, представленных файлами PDB и out-файлами, была написана программа на языке Python (version 3.2.1). Программа выполнена в виде библиотеки функций и содержит 17 модулей. Общий объем кода составляет 2963 строк и занимает 110 килобайт.

Описанная библиотека содержит все необходимые функции для парсинга исходных данных, их обработки, построения базы данных и анализа, как полученных элементов вторичной структуры, так и взаимодействий РНК с другими молекулами. По выполняемым действиям программа делится на две части:

1) Обработчик исходных данных (разделение документов на модели, прогонка моделей через 3DNA, проверка документов на наличие различных особенностей вторичной структуры и разбиение документов на соответствующие категории).

Общее время работы на всей выборке: 17 часов 20 минут (значительную долю занимает время работы функции find_pair из пакета 3DNA).

2) Конструктор текстовых файлов для наполнения БД (парсинг моделей и out-файлов, разметка элементов вторичной структуры, разметка РНК-Белковых и других взаимодействий)

Общее время работы на всей выборке: 6 часов 30 минут

Помимо описанной функциональности каждый из модулей может использоваться автономно, например, для промежуточного анализа результатов.

Стоит отметить, что на данный момент разметка межатомных взаимодействий осуществляется без использования стороннего ПО. Эту функцию выполняет один из модулей разработанной библиотеки, распознающий взаимодействия атомов нуклеотидов и аминокислот, учитывая расстояние между их центрами, их химические элементы и отсутствие третьего атома между ними. Контакты между атомами нуклеотидов и аминокислот классифицируются, как:

наложение (любая пара атомов на расстоянии меньше 2.5 ангстрем);

водородная связь (пара {O,N}-{O,N} на расстоянии от 2.5 до 3.5 ангстрем);

связь через воду (водородная связь при наличии молекулы воды между атомами);

гидрофобный контакт (пара {C,S}-{C,S} на расстоянии от 2.5 до 5.0 ангстрем);

атипичный первого типа (пара {O,N}-{C,S} на расстоянии от 2.5 до 5.0 ангстрем);

атипичный второго типа (пара, хотя бы один атом которой не из {O,N,C,S}, на расстоянии от 2.5 до 5.0 ангстрем).

Также, помимо контактов класса РНК-Белок мы размечаем взаимодействия Белок-Лиганд и РНК-Лиганд. В дальнейшем планируется усовершенствовать используемый метод разметки межатомных взаимодействий, что позволит избежать существенной доли «ложных» контактов.

2.3 Схема, данные, тестирование.

Для анализа экспериментально полученных структур РНК была разработана схема базы данных, основанная на новом способе описания петель. Данная разработка направлена на углубленное изучение вторичной структуры РНК и сбор статистики для последующего применения в рамках предсказания реальных последовательностей РНК, учитывая наличие псевдоузлов.

Предлагаемая схема состоит из 27 таблиц и содержит исчерпывающий набор данных, необходимых для дальнейших исследований, включая таблицы спиралей, нитей, петель, граней и др. Подробное описание таблиц см. в приложении А.

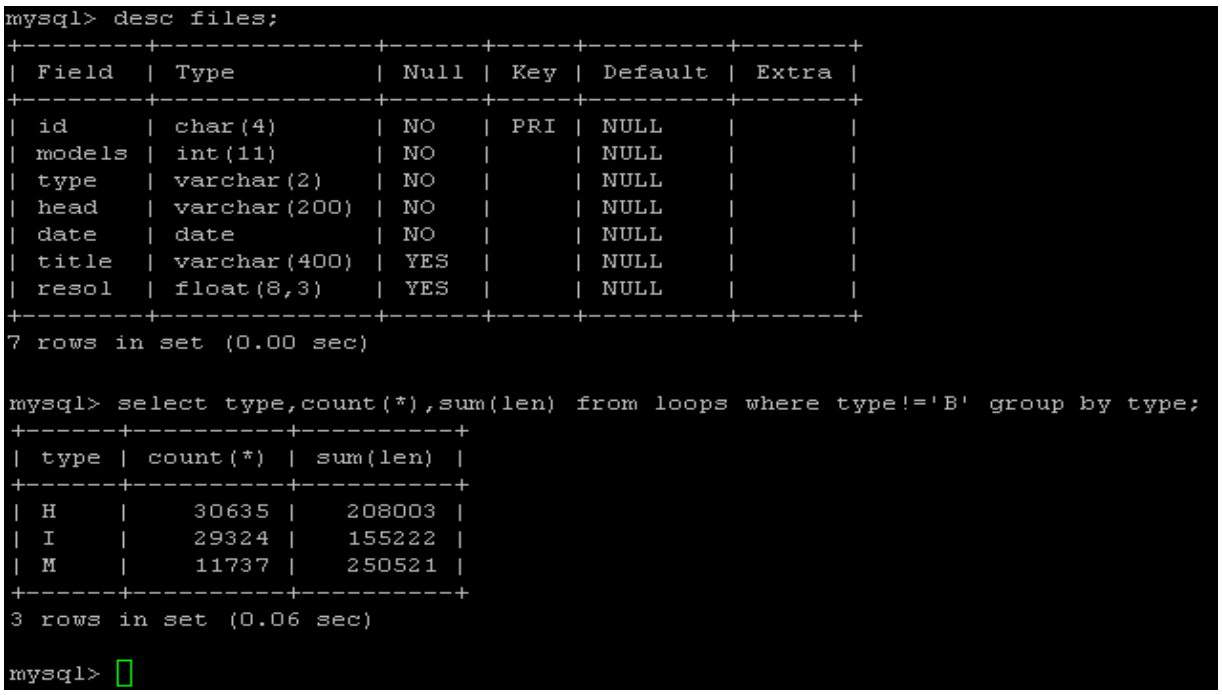

Рисунок 2.3. Пример взаимодействия с базой данных

Полный объём текстовых данных, полученных с помощью написанной нами программы, составляет 2.6 ГБ. Для поднятия базы данных на сервере был написан скрипт на языке MySQL (общее время работы скрипта: 13 минут). Для загрузки текстовых файлов и для взаимодействия с сервером нами использовались сторонние программы, такие как Putty и FileZilla.

Для тестирования базы данных в данный момент разрабатывается веб-интерфейс, позволяющий взаимодействовать с данными онлайн. Интерфейс реализован в виде html-страницы с вставками CSS и JavaScript и CGI-скриптов, написанных на Python 2.7. Сайт доступен любому пользователю и располагается по адресу http://server2.lpm.org.ru/~baulin/home.html.

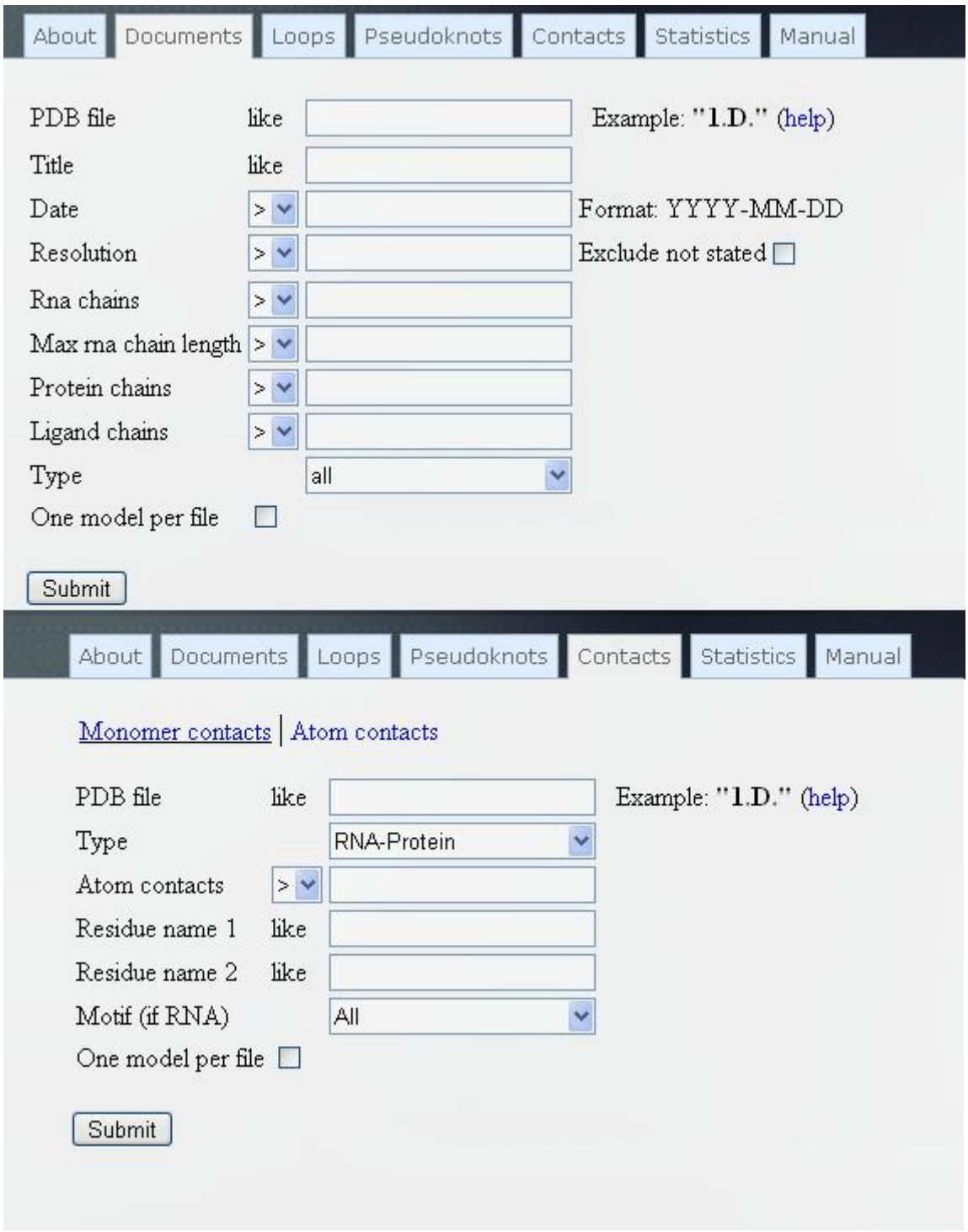

Рисунок 2.4. Веб-интерфейс: поиск документов (сверху) и поиск контактов (снизу).

В настоящее время реализован поиск по трём объектам: документы, атомарные контакты и мономерные контакты. В ближайшее время планируется расширить данный список до 7-10 объектов, после чего начать

непосредственное тестирование работоспособности базы данных и оптимальности её структуры.

2.4 Выводы и результаты по главе

Данная глава описывает предлагаемую схему базы данных и объясняет ключевые моменты её создания. Даётся подробное описание отбора исходных данных и их обработки. Также представлено описание процесса наполнения базы данными, с помощью разработанной библиотеки функций. В главе также рассматриваются существующие проблемы на различных этапах работы, в том числе детали формата исходных данных и тестирование существующей схемы с помощью созданного веб-интерфейса.

В качестве вывода можно отметить, что нами был проделан большой объем работы, но в то же время предстоит ещё очень много сделать.

Глава 3. Предварительные результаты

3.1 Анализ тройных узлов

На данный момент создана предварительная версия базы данных (бета-версия), которая находится на этапе открытого тестирования. В дальнейшем планируется приступить к анализу информации, представленной в ней. В ближайших планах – сбор различных случаев псевдоузловых структур, их сортировка, систематизация и анализ, а также исследование роли вторичной структуры РНК в образовании РНК-Белковых комплексов. Кроме того стоит задача изучения роли линков в образовании вторичной структуры РНК. Также планируется разработать полноценный веб-интерфейс для работы с базой данных, что даст возможность взаимодействия с ней пользователям, незнакомым с языком SQL.

Кроме того, во время работы над новой разметкой петель, был поставлен вопрос, имеются ли вторичные структуры с псевдоузлами, которые нельзя представить в виде плоского графа (графа, который можно разместить на

плоскости без самопересечений). Было выяснено, что неплоский граф структуры может иметь место, если в ней присутствует тройной узел.

Будем называть *тройным узлом* структуру РНК, в которой попарно конфликтуют три спирали. Более точно, пусть A, A’; B, B’ и C, C’ – комплементарные участки (крылья) спиралей A, B и C. Мы говорим, что спирали A, B и C образуют тройной узел, если их крылья расположены на цепи РНК в следующем порядке: A B C A’ B’ C’. Интерес к тройным узлам определяется различными причинами. С одной стороны, такая структура должна быть весьма компактной, и было интересно, реализуется ли она в действительности. С другой стороны, вопрос о наличии в РНК тройных узлов представляет и теоретический интерес.

Был проведён поиск тройных узлов и выявлено 233 структуры, которые содержат тройной узел.

При этом:

- все такие узлы находятся в гомологичных друг другу участках 23S РНК;
- в двух из трех спиралей крылья комплементарны, а в одной – нет (во всех случаях, кроме одного, «неправильная» спираль - это спираль A);
- все спирали, как правило, короткие (2-3 звена), в небольшом количестве случаев спираль C содержит 4 звена.

В связи с этим было проведено детальное исследование найденного участка 23S РНК. Были отобраны все рибосомальные РНК, представленные в PDB. После этого были оставлены только те документы, в которых цепь РНК содержит не менее 2000 нуклеотидов (таких оказалось 282, см. таблицу 1).

Почти все подобные структуры содержат цепи 23S РНК (малыми количествами представлены 25S РНК и 26S РНК).

Узел BCBC был найден во всех структурах. С помощью программы JMOL (jmol.sourceforge.net; version 12.2), в качестве примера, была рассмотрена структура 2D3O. Оказалось, что спирали имеют неправильную структуру (несмотря на комплементарные нуклеотиды). В 233 случаях find_pair для спирали A указывает две пары нуклеотидов. В 32 случаях – указывает одну пару (линк). В 7 случаях find_pair спираль не распознает.

Из 7 структур, в которых спираль A не найдена, только 3 имеют разрешение меньше 4 ангстрем (эти структуры: 2O44, 1P9X, 2VHN). Кроме того, 10 структур были отброшены в связи с некорректным форматом.

В ближайших планах – детальное изучение спариваний нуклеотидов, сбор статистики по геометрическим параметрам данных связей и дальнейший анализ выявленных тройных узлов.

Таблица 3.1. Организмы, чьи рибосомальные РНК содержат тройные узлы.

Организм	**К-во**	**Царство**	**Домен**
CANIS FAMILIARIS	1	Животные	Эукариоты
DEINOCOCCUS RADIODURANS	32	Бактерии	Прокариоты
DEINOCOCCUS RADIODURANS R1	3	Бактерии	Прокариоты
ESCHERICHIA COLI	78	Бактерии	Прокариоты
ESCHERICHIA COLI DH1	1	Бактерии	Прокариоты
HALOARCULA MARISMORTUI	62	Бактерии	Прокариоты
HALOARCULA MARISMORTUI ATCC 43049	1	Бактерии	Прокариоты
METHANOTHERMOBACTER THERMAUTOTROPHICUS	1	Эуархеи	Археи
SACCHAROMYCES CEREVISIAE	5	Грибы	Эукариоты
SACCHAROMYCES CEREVISIAE S288C	1	Грибы	Эукариоты

Таблица 3.1. Организмы, чьи рибосомальные РНК содержат тройные узлы (продолжение).

SPINACEA OLERACEA	1	Растения	Эукариоты
TETRAHYMENA THERMOPHILA	4	Протозоа	Эукариоты
THERMOMYCES LANUGINOSUS	1	Грибы	Эукариоты
THERMUS THERMOPHILUS	84	Бактерии	Прокариоты
THERMUS THERMOPHILUS HB8	6	Бактерии	Прокариоты
TRITICUM AESTIVUM	1	Растения	Эукариоты
ВСЕГО	282		

3.2 Выводы и результаты по главе

В данной главе даётся описание проведённого анализа нестандартных структур, которые не могут быть представлены в виде плоского графа. Было выяснено, что такие структуры имеют место, если они содержат т.н. тройные узлы, т.е. псевдоузлы, в которых три спирали попарно находятся в конфликте. Приводятся количественные результаты исследований подобных структур. Также в главе даётся объяснение значимости тройных узлов в экспериментально полученных структурах РНК.

В ближайшее время планируется провести подробный анализ геометрических параметров выявленных структур и выяснить, какую они играют роль в функционировании рибосомальных РНК.

Заключение

Нами предложено новое определение петли, которое, с одной стороны, является обобщением понятия петли по Мэтьюзу-Тернеру, а с другой позволяет разбить на петли произвольную вторичную структуру, а не только структуры, не содержащие псевдоузлов. Возможность такого описания важна при создании баз данных экспериментально определенных структур РНК

Определение основывается на наблюдении, что модель Мэтьюза-Тернера устанавливает взаимно-однозначное соответствие между спиралями и петлями, причем каждая петля по Мэтьюзу-Тернеру состоит из однонитевых участков, которые разделены торцами спиралей. Основные отличия наших петель от петель Мэтьюза-Тернера состоят в следующем.

1) В петлях для вторичных структур общего вида торцы могут быть торцами, как спиралей, так и сложных замкнутых участков (в петлях Мэтьюза-Тернера – только торцы спиралей).

2) Грань петли – это чередующаяся последовательность петель и крыльев спиралей (в петлях Мэтьюза-Тернера грани не содержат крыльев спиралей).

3) Однонитевой участок РНК может принадлежать нескольким петлям (в петлях Мэтьюза-Тернера – только одной петле).

На основе новых определений предложена классификация петель, являющаяся обобщением классификации Мэтьюза-Тернера, а также разработана основанная на ней база данных. Представляется интересным выяснить, какие из возможных видов петель реализуются в экспериментально определенных структурах РНК.

В ближайших планах – совершенствование используемых методов и программ для более точного анализа и обработки исходных данных,

разработка полноценного веб-интерфейса, который позволит эффективно использовать предложенную нами базу данных. Также планируется дальнейшее изучение и сбор подробной статистики различных случаев псевдоузловых структур, в частности структур содержащих тройные узлы.

Библиографический список

1. Waterman M.S.(ed.) *Mathematical methods for DNA sequences*. CRC Press, Boca Raton, FL. 1989. 293 p.
2. Nussinov R., Jacobson, A.B. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl Acad. Sci. USA.* 1980. Vol. 77. P. 6309–6313.
3. Xia T., SantaLucia J. Jr., Burkard M.E., Kierzek R., Schroeder S.J., Jiao X., Cox C., Turner D.H. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson–Crick base pairs. *Biochemistry.* 1998. Vol.37 4735.
4. Zuker M., Mathews D.H., Turner D.H., et al. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In: *RNA Biochemistry and Biotechnology*, J. Barciszewski & B.F.C. Clark, eds., NATO ASI Series, Kluwer Academic Publishers, 1999. P. 42-57.
5. Mathews D.H., Turner D.H. Prediction of RNA secondary structure by free energy minimization. *Current Opinion in Structural Biology 2006,* Vol. 16. P. 270–278.
6. Pleij, C.W.A. RNA Pseudoknots. In: *The RNA world.* Gesteland, R. F. and Atkins, J. F., eds. Cold Spring Harbor Laboratory Press, 1993. P. 609-613.
7. Gultyaev AP, Olsthoorn RC. A family of non-classical pseudoknots in influenza A and B viruses. *RNA Biol.* 2010. Vol. 7. P. 125-129.
8. Condon A., Davy B., Rastegari B., Zhao S., Tarrant F. Classifying RNA pseudoknotted structures. *Theoretical Computer Science* 2004. Vol. 320. P. 35 – 50.
9. Batenburg F.H.D. van, Gultyaev A.P., Pleij C.W.A., Ng J., Oliehoek J. Pseudobase: a database with RNA pseudoknots. *Nucl. Acids Res.* 2000. Vol. 28. P. 201-204.
10. Taufer M., Licon A., Araiza R., Mireles D., Batenburg F.H.D. van, Gultyaev A.P., Leung M.-Y. PseudoBase++: an extension of PseudoBase for easy searching, formatting and visualization of Pseudoknots. *Nucl. Acids Res*. 2009. Vol. 37. P. D127–D136

11. Haslinger C., Stadler P.F. RNA Structures with Pseudo-knots: Graph-theoretical, Combinatorial, and Statistical Properties. *Bulletin of Mathematical Biology*.(1999. Vol. 61. P. 437–467
12. Zheng G., Lu X.-J., Olson W.K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures', Nucleic Acids Res. 2003. Vol. 31. P. 5108-5121.
13. Baulin E., Ivankov D., Roytberg M. Statistics of RNA structures. In: *Proceedings of Moscow Conference on Computational Molecular Biology, July 21-24, 2011*. M. 2011. P. 325-326.
14. Баулин Е. Ф., Ройтберг М. А., Астахова Т. В. Классификация элементов вторичной структуры РНК // Математическая биология и биоинформатика. 2012. Т. 7. № 2. С. 567-571.

Приложение А. Схема базы данных

Список таблиц.

Номер таблицы	Название	Сущность	Объем данных (строк)
1	files	Таблица файлов	1674
2	models	Таблица моделей	6716
3	molecules	Таблица молекул	12297
4	rnachains	Таблица цепей РНК	9141
5	protchains	Таблица цепей белка	11018
6	ligchains	Таблица цепей лигандов	119
7	wings	Таблица крыльев	156034
8	threads	Таблица нитей	149637
9	helices	Таблица спиралей	78017
10	loops	Таблица петель	78017
11	sides	Таблица фрагментов граней	182579
12	faces	Таблица торцов	71198
13	links	Таблица линков	119181
14	nucleotides	Таблица нуклеотидов	1162169
15	aminoacids	Таблица аминокислот	1435821
16	ligands	Таблица лигандов	68689
17	pairs	Таблица спариваний	457357
18	ratoms	Таблица атомов РНК	24869144
19	patoms	Таблица атомов белка	11010798
20	latoms	Таблица атомов лигандов	141537
21	watoms	Таблица атомов воды	506645
22	rpatompairs	Таблица межатомных пар РНК-Белок	6458500

Список таблиц (продолжение).

23	rlatompairs	Таблица межатомных пар РНК-Лиганд	545725
24	platompairs	Таблица межатомных пар Белок-Лиганд	103325
25	rpmonopairs	Таблица межмономерных пар РНК-Белок	1009363
26	rlmonopairs	Таблица межмономерных пар РНК-Лиганд	194819
27	plmonopairs	Таблица межмономерных пар Белок-Лиганд	33461

Таблица 1. Файлы.

Номер	Поле	Значение
1	name	Уникальное имя документа (4 символа)
2	models	К-во моделей данного кристалла
3	type	Тип структуры (R - только РНК; RP - РНК с Белком)
4	head	Строка HEADER из pdb-файла
5	date	Дата создания
6	title	Заголовок файла
7	resol	Разрешение файла (в ангстремах)

Таблица 2. Модели.

Номер	Поле	Значение
1	id	Уникальный номер id
2	number	Номер модели (по порядку внутри pdb-файла)
3	file	имя pdb-файла, к которому принадлежит модель
4	main	1 - "рабочая" модель; 0 - нет
5	type	('CH' - есть связанные цепи; 'PS' - есть псевдоузлы; 'NM' - без осложнений)
6	rnachains	К-во цепей РНК
7	maxrnalen	Длина максимальной цепи РНК
8	rnalen	Общая длина цепей РНК
9	protchains	К-во цепей белка
10	maxprotlen	Длина максимальной цепи белка
11	protlen	Общая длина цепей белка
12	ligchains	К-во цепей лигандов
13	maxliglen	Длина максимальной цепи лигандов
14	liglen	Общая длина цепей лигандов
15	ligands	К-во лигандов
16	ligandtypes	К-во типов лигандов

Таблица 3. Молекулы.

Номер	Поле	Значение
1	id	Уникальный номер id
2	file	Имя файла
3	mol_id	Номер молекулы внутри файла
4	mol	Имя макромолекулы
5	orgsc	Имя организма
6	frag	Фрагмент молекулы
7	syn	Список синонимов имени молекулы
8	ec	Комиссионное число ферментов молекулы
9	eng	Способ получения
10	mut	Присутствует ли мутация
11	details	Прочие детали

Таблица 4. Цепи РНК.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели, которой принадлежит цепь
3	molecule	id молекулы, которой принадлежит цепь (Null, если не указана)
4	letter	Идентификатор цепи из pdb-файла
5	len	Реальная длина цепи
6	glen	Длина цепи, учитывая разрывы
7	bound	К-во спаренных нуклеотидов в 3DNA
8	share	Доля связанных нуклеотидов
9	chain2	id цепи РНК, с которой связана данная цепь (Null если нет)
10	dhelices	К-во собственных Д-Спиралей
11	dhelices2	К-во общих Д-Спиралей
12	isolates	К-во собственных изолятов
13	isolates2	К-во общих изолятов
14	helices	К-во собственных стандартных спиралей
15	helices2	К-во общих стандартных спиралей
16	towers	К-во собственных башен
17	towers2	К-во общих башен
18	all_links	К-во собственных линков
19	all_links2	К-во общих линков
20	ilinks	К-во собственных внутренних линков
21	ilinks2	К-во общих внутренних линков
22	blinks	К-во собственных связанных линков
23	blinks2	К-во общих связанных линков
24	flinks	К-во собственных свободных линков
25	flinks2	К-во общих свободных линков
26	multi_loops	К-во собственных Мульти-петель
27	multi_loops2	К-во общих Мульти-петель
28	int_loops	К-во собственных внутренних петель
29	int_loops2	К-во общих внутренних петель

Таблица 4. Цепи РНК (продолжение).

30	hairpins	К-во собственных Hairpin'ов
31	buldges	К-во собственных выпячиваний
32	buldges2	К-во общих выпячиваний
33	rpatompairs	К-во водородных связей между атомами нуклеотидов и аминокислот
34	rlatompairs	К-во водородных связей между атомами нуклеотидов и лигандов
35	rpcontacts	Количество пар (водородных связей) вида нуклеотид-аминокислота
36	rlcontacts	Количество пар (водородных связей) вида нуклеотид-лиганд
37	pnucleos	К-во нуклеотидов, имеющих водородные связи с аминокислотами
38	lnucleos	К-во нуклеотидов, имеющих водородные связи с лигандами
39	in_helices	К-во нуклеотидов в спиралях, имеющих водородные связи с аминокислотами

Таблица 5. Цепи белка.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели, которой принадлежит цепь
3	mol	id молекулы, которой принадлежит цепь
4	letter	Идентификатор цепи из 3db-файла
5	len	Реальная длина цепи
6	glen	Длина цепи, учитывая разрывы
7	rpatompairs	К-во водородных связей между атомами аминокислот и нуклеотидов
8	platompairs	К-во водородных связей между атомами аминокислот и лигандов
9	rpcontacts	Количество пар (водородных связей) вида нуклеотид-аминокислота
10	plcontacts	Количество пар (водородных связей) вида лиганд-аминокислота
11	raminos	К-во аминокислот, имеющих водородные связи с нуклеотидами
12	laminos	К-во аминокислот, имеющих водородные связи с лигандами

Таблица 6. Цепи лигандов.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели, которой принадлежит цепь
3	molecule	id молекулы, которой принадлежит цепь (Null, если не указана)
4	letter	Идентификатор цепи из 3db-файла
5	len	Длина цепи
6	rlatompairs	К-во водородных связей между атомами аминокислот и нуклеотидов
7	platompairs	К-во водородных связей между атомами аминокислот и лигандов
8	rlcontacts	Количество пар (водородных связей) вида нуклеотид-лиганд
9	plcontacts	Количество пар (водородных связей) вида лиганд-аминокислота
10	rligands	К-во лигандов, имеющих водородные связи с нуклеотидами
11	pligands	К-во лигандов, имеющих водородные связи с аминокислотами

Таблица 7. Крылья.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели, которой принадлежит крыло
3	chain	id цепи РНК
4	helix	id спирали, которой принадлежит крыло
5	another	id спаренного крыла
6	next	id следующего крыла (Null, если крыло последнее)
7	type	Тип крыла (L – левое (раньше по цепи); R – правое (позже по цепи))
8	start	id первого нуклеотида (с наименьшим номером)
9	end	id последнего нуклеотида (с наибольшим номером)
10	len	длина крыла
11	fake	1 - фиктивное крыло; 0 - иначе

Таблица 8. Нити.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели, которой принадлежит нить
3	chain	id цепи РНК
4	leftwing	id крыла, примыкающего слева (Null, если нить на краю цепи)

Таблица 8. Нити (продолжение).

5	rightwing	id крыла, примыкающего справа (Null, если нить на краю цепи)
6	start	id первого нуклеотида (если нить нулевой длины, то id начала примыкающей справа спирали) общее правило: нуклеотид после последнего нуклеотида спирали слева
7	end	id последнего нуклеотида (если нить нулевой длины, то id конца примыкающей слева спирали) общее правило: нуклеотид перед первым нуклеотидом спирали справа
8	len	Реальная длина нити
9	glen	Длина нити, учитывая разрывы
10	gap	1 - в нити есть разрыв; 0 - нет
11	links	К-во концов линков, принадлежащих нити
12	ext	1 - нить на краю цепи; 0 - иначе

Таблица 9. Спирали.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели, которой принадлежит спираль
3	chain1	id первой цепи РНК
4	chain2	id второй цепи РНК (если спираль не общая, то совпадает с chain1)
5	leftwing	id левого крыла
6	rightwing	id правого крыла
7	len	длина спирали
8	tower	номер башни (по порядку внутри модели)
9	fake	1 - фиктивная спираль; 0 - иначе

Таблица 10. Петли.

Номер	Поле	Значение
1	helixid	Уникальный номер id (совпадает с id закрывающей спирали)
2	model	id модели, которой принадлежит петля
3	chain1	id первой цепи РНК
4	chain2	id второй цепи РНК (совпадает с первой, если петля для неё является собственной)
5	threads	К-во нитей в петле
6	wings	К-во крыльев в петле
7	sides	К-во граней в петле
8	hfaces	К-во торцов спиралей в петле
9	bfaces	К-во торцов блоков в петле
10	type	I - Internal Loop, H - Hairpin, M - Multi-Loop, B - Buldge
11	ptype	C - классическая; I - изолированная; P - узловая
12	gap	1 - если петля содержит разрыв(ы); 0 - иначе

Таблица 10. Петли (продолжение).

13	len	Реальная длина петли (суммарная реальная длина нитей и длина крыльев)
14	glen	Длина петли, учитывая разрывы
15	links	К-во концов линков, принадлежащих петле
16	ext	Является ли петля внешней (external) = True/False

Таблица 11. Фрагменты граней.

Номер	Поле	Значение
1	loopid	id петли, которой принадлежит грань
2	model	id модели, которой принадлежит петля
3	chain	id цепи РНК
4	snum	номер грани по порядку внутри петли
5	len	Реальная длина грани
6	glen	Длина грани, учитывая разрывы
7	num	Номер фрагмента по порядку внутри грани
8	type	RW - правое крыло; LW - левое крыло; TH - нить
9	partid	id крыла или нити
10	links	количество концов линков в грани

Таблица 12. Торцы.

Номер	Поле	Значение
1	loopid	id петли, которой принадлежит торец
2	model	id модели, которой принадлежит петля
3	chain1	id первой цепи РНК
4	chain2	id второй цепи РНК
5	num	номер торца по порядку внутри петли
6	type	B - торец блока; H - торец спирали
7	helix1	id открывающей спирали
8	helix2	id закрывающей спирали (если тип H, то совпадает с helix1)

Таблица 13. Линки.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели, которой принадлежит линк
3	chain1	id первой цепи РНК
4	chain2	id второй цепи РНК (если линк не общий, то совпадает с chain1)
5	left_thread	id нити, которой принадлежит первый нуклеотид линка
6	right_thread	id нити, которой принадлежит второй нуклеотид линка
7	type	I - Internal link, B - Bound link, F - Free link

Таблица 13. Линки (продолжение).

8	nucl1	id первого нуклеотида
9	nucl2	id второго нуклеотида

Таблица 14. Нуклеотиды.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели, которой принадлежит нуклеотид
3	chain	id цепи, которой принадлежит нуклеотид
4	name	название нуклеотида (см. список возможных названий ниже)
5	place	место нуклеотида (W - крыло; T - нить)
6	motif	id крыла или нити
7	link	1 - нуклеотид составляет линк (только если place==T); 0 - иначе
8	linkid	id линка (Null, если link == 0)
9	number	номер нуклеотида (из 3DNA)
10	pdbnumber	номер нуклеотида (из PDB)

Таблица 15. Аминокислоты.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели, которой принадлежит аминокислота
3	chain	id цепи, которой принадлежит аминокислота
4	name	название аминокислоты
5	pdbnumber	номер аминокислоты (из pdb)

Таблица 16. Лиганды.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели, которой принадлежит лиганд
3	chtype	тип цепи, которой принадлежит лиганд (R - РНК, P - Белок, L - Лиганды)
4	chain	id цепи, к которой принадлежит лиганд
5	name	название лиганда
6	pdbnumber	номер лиганда (из pdb)

Таблица 17. Спаривания.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели, которой принадлежит линк
3	chain1	id первой цепи РНК
4	chain2	id второй цепи РНК (если спаривание не общее, то совпадает с chain1)
5	type	H - Helix I - Internal link, B - Bound link, F - Free link
6	bond	связь (пример G-----C)
7	part	id нити или крыла
8	motif	id линка или спирали
9	stack	0 - изолят; 2 - середина д-спирали; 1 - край д-спирали
10	dnum	номер д-спирали (изолята) по порядку в модели
11	tower	номер башни (по порядку внутри модели), Null для линков
12	nucl1	id первого нуклеотида
13	nucl2	id второго нуклеотида

Таблица 18. Атомы РНК.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели, которой принадлежит атом
3	chain	id цепи РНК
4	nucl	id нуклеотида
5	name	имя атома в нуклеотиде
6	number	номер атома в файле
7	elem	наименование химического элемента
8	x	координата x
9	y	координата y
10	z	координата z

Таблица 19. Атомы белка.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели, которой принадлежит атом
3	chain	id цепи белка
4	amino	id аминокислоты
5	name	имя атома в аминокислоте
6	number	номер атома в файле
7	elem	наименование химического элемента
8	x	координата x
9	y	координата y
10	z	координата z

Таблица 20. Атомы лигандов.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели, которой принадлежит атом
3	chtype	R - рнк; P - белок; L - лиганды
4	chain	id цепи
5	ligand	id лиганда
6	name	имя атома в лиганде
7	number	номер атома в файле
8	elem	наименование химического элемента
9	x	координата x
10	y	координата y
11	z	координата z

Таблица 21. Атомы воды.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели, которой принадлежит атом
3	chtype	R - рнк; P - белок; L - лиганды
4	chain	id цепи
5	number	номер атома в файле
6	elem	наименование химического элемента
7	x	координата x
8	y	координата y
9	z	координата z

Таблица 22. Межатомные пары РНК-Белок.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели
3	atom1	id атома РНК
4	atom2	id атома белка
5	dist	длина связи
6	type	тип связи ('OL' - наложение, 'WT' - через воду, 'HP' - гидрофобная, 'HG' - водородная, 'AT' – атипичная(1), 'SW' - атипичная(2))
7	watom	id атома воды (если WT, иначе Null)

Таблица 23. Межатомные пары РНК-Лиганд.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели
3	atom1	id атома РНК
4	atom2	id атома лиганда
5	dist	длина связи
6	type	тип связи ('OL' - наложение, 'WT' - через воду, 'HP' - гидрофобная, 'HG' - водородная, 'AT' - атипичная, 'SW' - атипичная(2))
7	watom	id атома воды (если WT, иначе Null)

Таблица 24. Межатомные пары Белок-Лиганд.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели
3	atom1	id атома белка
4	atom2	id атома лиганда
5	dist	длина связи
6	type	тип связи ('OL' - наложение, 'WT' - через воду, 'HP' - гидрофобная, 'HG' - водородная, 'AT' - атипичная, 'SW' - атипичная(2))
7	watom	id атома воды (если WT, иначе Null)

Таблица 25. Межмономерные пары РНК-Белок.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели
3	nucl	id нуклеотида
4	amino	id аминокислоты
5	atoms	К-во связей в паре

Таблица 26. Межмономерные пары РНК-Лиганд.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели
3	nucl	id нуклеотида
4	lig	id лиганда
5	atoms	К-во связей в паре

Таблица 27. Межмономерные пары Белок-Лиганд.

Номер	Поле	Значение
1	id	Уникальный номер id
2	model	id модели
3	amino	id аминокислоты
4	lig	id лиганда
5	atoms	К-во связей в паре