

**Федеральное агентство по образованию**  
Федеральное государственное бюджетное образовательное учреждение высшего  
профессионального образования  
**Пушкинский государственный естественно-научный институт**  
**Учебный центр математической биологии**

На правах рукописи

**Баулин Евгений Федорович**

**МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ**

*База данных пространственных структур РНК*

**Направление подготовки магистра**  
*(010 400 Прикладная математика и информатика)*

**Магистерская образовательная программа**  
**(Математическое моделирование)**

«Работа допущена к защите»:

Научный руководитель  
д.ф.-м.н., проф. \_\_\_\_\_ М.А. Ройтберг

Руководитель магистерской  
образовательной программы,  
д.ф.-м.н., проф. \_\_\_\_\_ М.Н. Устинин

**Пушино – 2015**

## СОДЕРЖАНИЕ

	Стр.
ВВЕДЕНИЕ.....	3
ГЛАВА 1. Теоретические основы предлагаемой модели описания структуры РНК.....	5
1.1 Основные определения.....	5
1.2 Спирали и петли.....	7
1.3 Структура петель.....	8
1.4 Псевдоузлы.....	12
1.5 Стемовые мотивы.....	18
ГЛАВА 2. База данных.....	21
2.1 Исходные данные.....	21
2.2 Обработка структур.....	23
2.3 Схема и данные.....	25
ГЛАВА 3. Веб-интерфейс.....	26
3.1 Детали реализации.....	26
3.2 Общая структура и функционал.....	27
3.3 Обзор существующих аналогов.....	31
ГЛАВА 4. Результаты.....	33
4.1 Анализ псевдоузлов.....	33
ЗАКЛЮЧЕНИЕ.....	35
БИБЛИОГРАФИЧЕСКИЙ СПИСОК.....	36
ПРИЛОЖЕНИЕ А. Структура программного обеспечения.....	40
ПРИЛОЖЕНИЕ В. Схема базы данных.....	42
ПРИЛОЖЕНИЕ С. Список рассмотренных баз данных.....	60

## ВВЕДЕНИЕ

Магистерская диссертация выполнена на базе Института математических проблем биологии РАН, в лаборатории Прикладной математики, под руководством д.ф.-м.н. Ройтберга М.А.

Методика реализации баз данных и веб-интерфейсов была освоена с помощью:

- Яковлев Виктор Вадимович, к.ф.-м.н., научный сотрудник Лаборатории прикладной математики Института математических проблем биологии РАН.
- Хачко Денис Викторович, техник Лаборатории прикладной математики Института математических проблем биологии РАН.

Консультантом в области молекулярной биологии был:

- Спириин Сергей Александрович, к.ф.-м.н., старший научный сотрудник НИИ физико-химической биологии имени А.Н. Белозерского МГУ имени М.В. Ломоносова.

Исследование РНК и ее пространственной структуры – одно из основных направлений современной молекулярной биологии, как с точки зрения теории (понимание внутриклеточных процессов, теория эволюции), так и с точки зрения практики (создание лекарств) [1-3]. Развитие исследований сдерживается тем, что не существует универсальной аннотированной базы пространственных структур РНК. Одной из причин этого является отсутствие общепринятой классификации элементов вторичной структуры РНК, которая позволила бы описывать как классические вторичные структуры, так и структуры, содержащие псевдоузлы.

В рамках общепринятой в настоящее время модели Цукера-Мэтьюза-Тернера (Nearest Neighbour Model, NNM, [4,5]) вторичная 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) [6]. При этом в настоящее время не существует общепринятой классификации псевдоузлов. Описаны простейшие виды псевдоузлов (kissing hairpins, H-structures и др., см. [7]), псевдоузлы, играющие важную роль в клеточных процессах, см. например, [8]; в [9] дан анализ классов псевдоузлов, введенных в различных работах по предсказанию вторичной структуры РНК по ее последовательности. В базе данных [10,11] собраны примеры псевдоузлов в экспериментально определенных структурах РНК. Однако единого способа описания элементов вторичной структуры при наличии псевдоузлов, подобного тому, который был предложен в [4,5] для описания структур, которые не содержат псевдоузлов, в настоящее время нет.

В рамках данной работы впервые создана общедоступная аннотированная база данных экспериментально определенных пространственных структур РНК (The «Universe of RNA Structures» Database, URSDB), основанная на едином языке описания произвольной вторичной структуры и включающая в себя все доступные РНК-содержащие структуры и их элементы. Основной принцип предложенной классификации элементов вторичной структуры 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 NNM.

# ГЛАВА 1. Теоретические основы предлагаемой модели описания 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 (Спаривание)* – это пара нуклеотидов  $(i, j)$ , где  $i < j$ , которая образует водородные связи. При этом допускаются не только связи между комплементарными нуклеотидами (A-U и G-C Watson-Crick pairs) и G-U связи (Wobble pairs), но и неканонические связи, см. [12,13].

*Спираль (Стем)* – это последовательность пар нуклеотидов вида  $(i, j)$ ,  $(i+1, j-1), \dots, (i+k, j-k)$  такая, что

1)  $i < j, i+k < j-k, k > 1$ ;

2) все пары вида  $(i+x, j-x)$ , где  $x = 0, \dots, k$ , образуют связи, причём связи  $(i, j)$  и  $(i+k, j-k)$  – Уотсон-Криковские связи (WC-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 [13] используется другое определение спирали (основанное исключительно на геометрических параметрах цепи). Далее в случае элементов согласно 3DNA мы будем использовать термины *D-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 WC-связи или G-U связи  $(k, t)$ , что  $i \leq k \leq j$  или  $i \leq t \leq j$ ;

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 WC и G-U 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 спираль [14], роль таких связей в настоящее время изучена слабо. Мы исходим из предположения, что полезно отдельно рассматривать «базовую» вторичную структуру, образованную спиралью, 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 (см. Рис. 1.1).

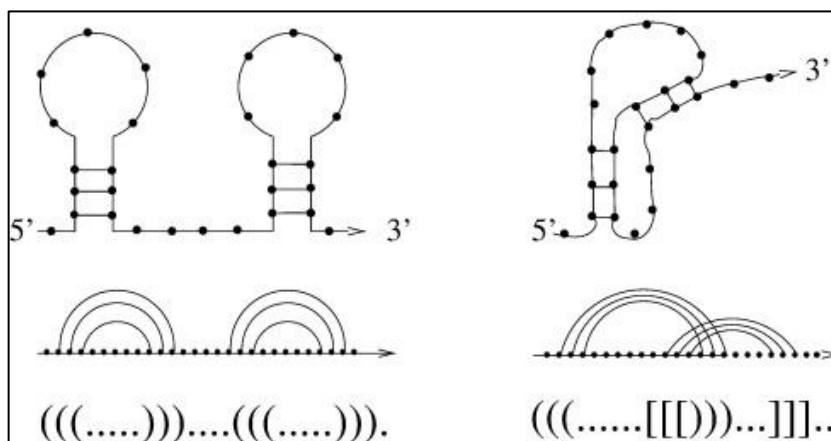


Рисунок 1.1. Пример структуры без псевдоузлов (слева) и с псевдоузлами (справа)

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.

## 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» (ср. с [15]).

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 элементарным замкнутым участком относительно *H* (в общем случае элементарный замкнутый относительно стема участок, *stem-related elementary closed region, S-ECR*), если

- 1)  $[i, j]$  лежит внутри *H*;
- 2) не существует таких связей  $(k, t)$ , что  $(i \leq k \leq j < t < v)$  или  $(u < k < i \leq t \leq j)$ ;
- 3) существуют связи  $(i, k)$  и  $(t, j)$ , где  $k \leq j$ ;  $i \leq t$ .



4) не существует отличного от  $[i, j]$  участка  $[i', j']$  такого, что  $i' \leq i < j \leq j'$  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 данной спирали. 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* (см. Рис. 1.2).

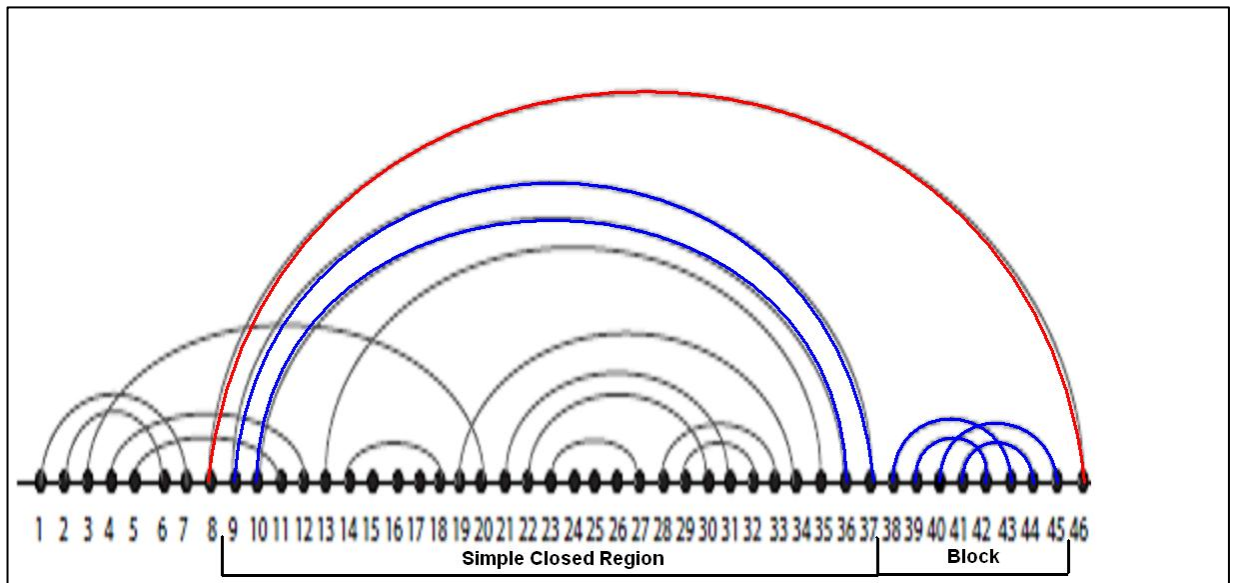


Рисунок 1.2. Примеры закрытых участков

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  $(s_1, t_1), \dots, (s_n, t_n)$  – торцы всех участков, замкнутых относительно  $H$ ;  $s_1 < t_1 < \dots < s_n < t_n$ . Для удобства пусть  $t_0 = u$ ;  $s_{n+1} = v$ . Пусть  $k$  – целое;  $1 \leq k \leq n + 1$ . 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  $(s_1, t_1), \dots, (s_n, t_n)$  – торцы всех участков, замкнутых относительно  $H$ ;  $s_1 < t_1 < \dots < s_n < t_n$ . 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 одностороннего участка. Поэтому можно дать такое определение.

Определение 7. Петля называется *классической* (classical), если она не содержит крыльев и торцов блоков. Петля называется *изолированной* (isolated), если она не содержит крыльев. и *узловой* (pseudoknotted), 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 (см. Рис. 1.3).

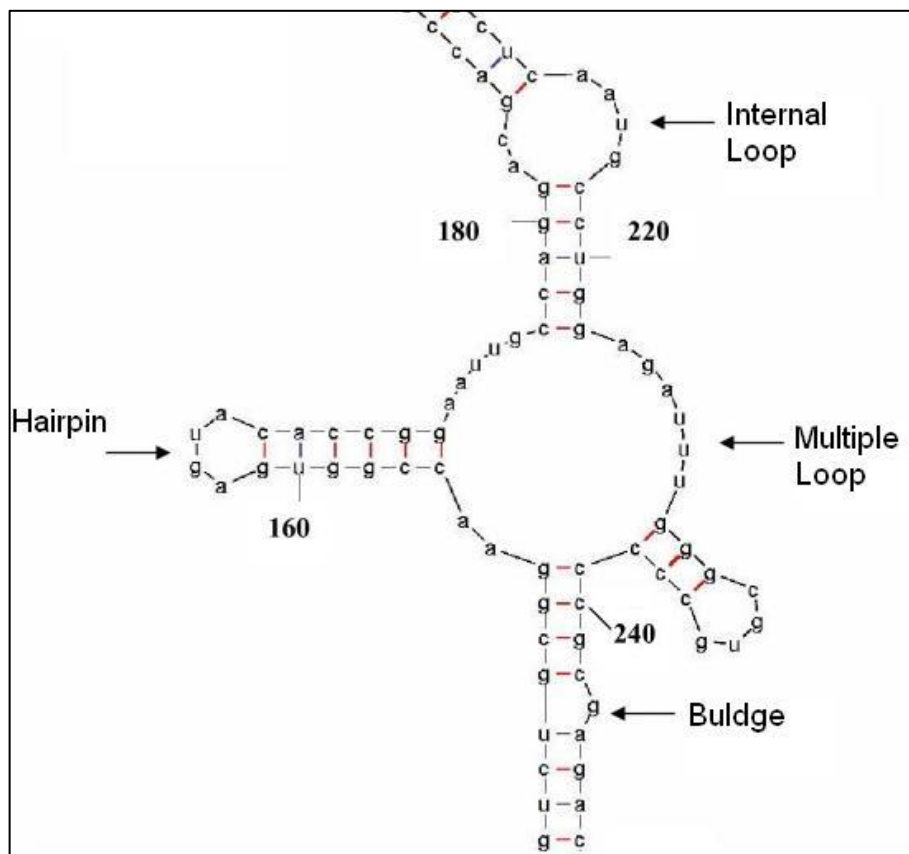


Рисунок 1.3. Различные типы петель

Замечание 1. Будем называть выпячиванием (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» спиральям).

#### 1.4 Псевдоузлы

Элементарный замкнутый участок (ЭЗУ, *elementary closed region, ECR*) – минимальный участок  $[i, j]$ , такой что:

1) Не существует связи  $(k, l)$ , такой что  $(i \leq k \leq j; l > j)$  или  $(k < i; i \leq l \leq j)$ ;

2) Не существует позиции  $l$ , такой что  $i < l < j$  и оба участка  $[i, \dots, l]$  и  $[l + 1, \dots, j]$  удовлетворяют условию 1);

3) Существуют связи  $(i, k)$  and  $(l, j)$ ; допускаются равенства  $k = j$  и  $i = l$ .

Пара  $(i, j)$  называется *торцом* ЭЗУ  $[i, j]$ . Отметим, что если пара  $(i, j)$  является связью и принадлежит некоторой спирали, то торец ЭЗУ совпадает с торцом данной спирали.

ЭЗУ  $[k, l]$  является *суб-ЭЗУ* (sub-ECR) относительно ЭЗУ  $[i, j]$ , если  $i < k < l < j$  и не существует такого ЭЗУ  $[m, n]$ , что  $i < m < k < l < n < j$ .

ЭЗУ называется *псевдоузлом* (синоним: *псевдоузловой ЭЗУ*) если принадлежащие ему спирали находятся в конфликте. В противном случае ЭЗУ называется свободным от псевдоузлов или *классическим*.

Классификация псевдоузлов, реализованная в разработанной базе данных URSDb, основана на понятии *сигнатуры*. Классификация схожа с топологической классификацией псевдоузлов, предложенной в работе [16]. Основным отличием нашей классификации является исключение из рассмотрения одиночных спариваний (линков).

Рассмотрим все спирали ЭЗУ и обозначим их буквами латинского алфавита в соответствии с позициями их крыльев от 5`- к 3`-концу. Левое крыло будем обозначать строчными буквами, например, **a**, а правое крыло – заглавными буквами, например, **A**. Таким образом, каждая спираль будет обозначена двумя буквами, например, **aA**.

*Полной сигнатурой* ЭЗУ называется последовательность его крыльев в соответствии с их позициями от 5`- к 3`-концу.

Пример 1. Пусть ЭЗУ  $[10, 70]$  содержит три спирали,  $([10, 15]; [65, 70])$ ,  $([20, 25]; [45, 50])$ ,  $([30, 35]; [55, 60])$ , здесь  $[10, 15]$  и  $[65, 70]$  – крылья спирали  $([10, 15]; [65, 70])$ , и т.д. Тогда спираль  $([10, 15]; [65, 70])$  обозначим как **aA**, спираль  $([20, 25]; [45, 50])$  – **bB**, а спираль  $([30, 35]; [55, 60])$  – **cC**. Полной сигнатурой данного ЭЗУ будет последовательность **abcBCA**. Участок  $[20, 60]$  является суб-ЭЗУ относительно исходного ЭЗУ (см. Рис. 1.4, 1.5).

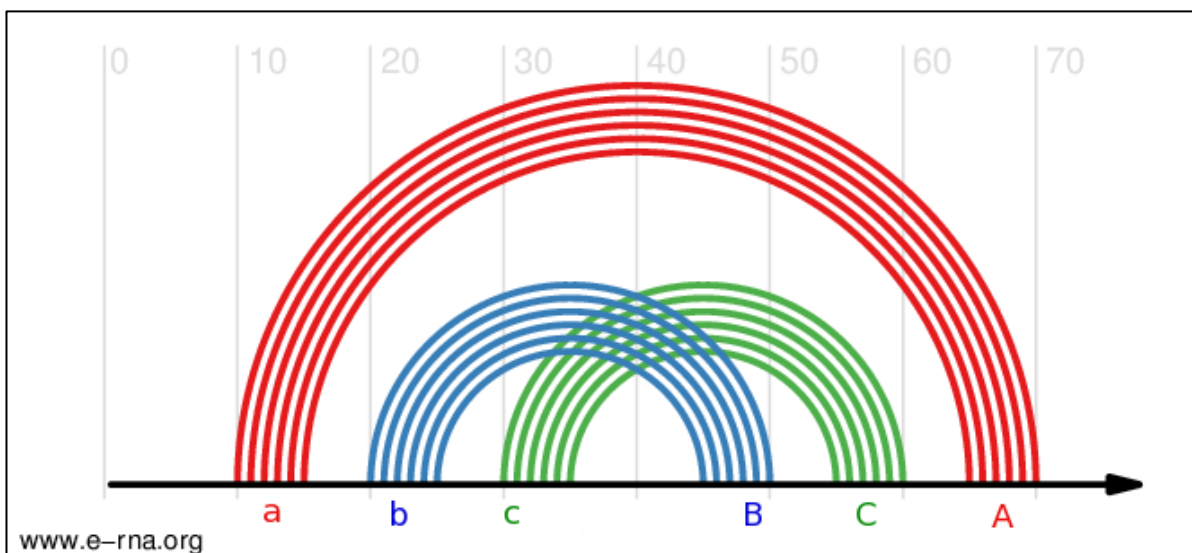


Рисунок 1.4. Дуговая диаграмма крыльев ЭЗУ из примера 1.

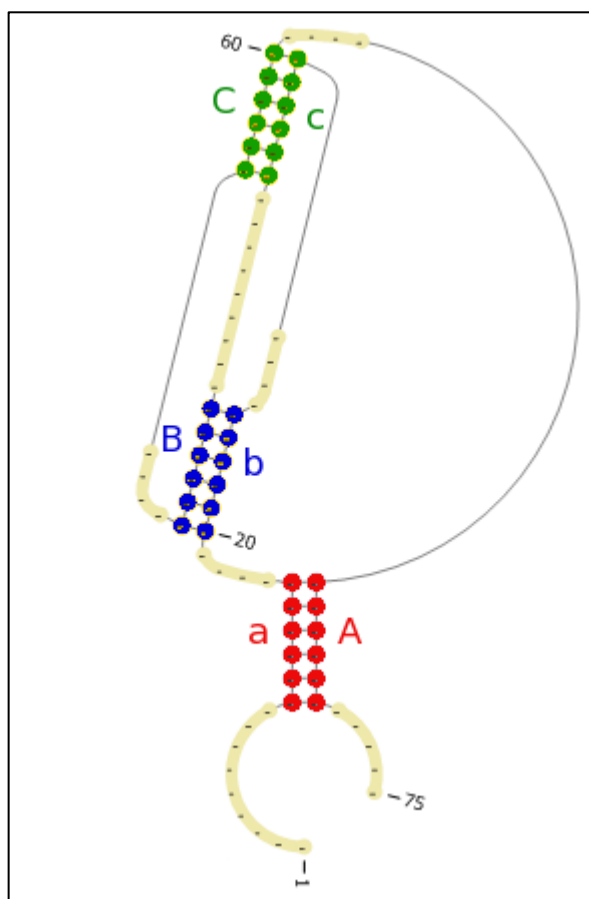


Рисунок 1.5. Схема вторичной структуры ЭЗУ из примера 1.

Пример 2. Пусть ЭЗУ содержит четыре спирали,  $([10, 15]; [70, 75])$ ,  $([20, 25]; [50, 55])$ ,  $([30, 35]; [40, 45])$ ,  $([60, 65]; [80, 85])$ , здесь  $[10, 15]$  и  $[70, 75]$  – крылья спирали  $([10, 15]; [70, 75])$ , и т.д. Тогда спираль  $([10, 15]; [70, 75])$

обозначим как **aA**, спираль  $([20, 25]; [50, 55])$  – **bB**, спираль  $([30, 35]; [40, 45])$  – **cC**, а спираль  $([60, 65]; [80, 85])$  – **dD**. Полной сигнатурой данного ЭЗУ будет последовательность **abcCBdAD**. Участок  $[20, 55]$  является суб-ЭЗУ относительно исходного ЭЗУ, а участок  $[30, 45]$  является суб-ЭЗУ относительно участка  $[20, 55]$  (см. Рис. 1.6, 1.7).

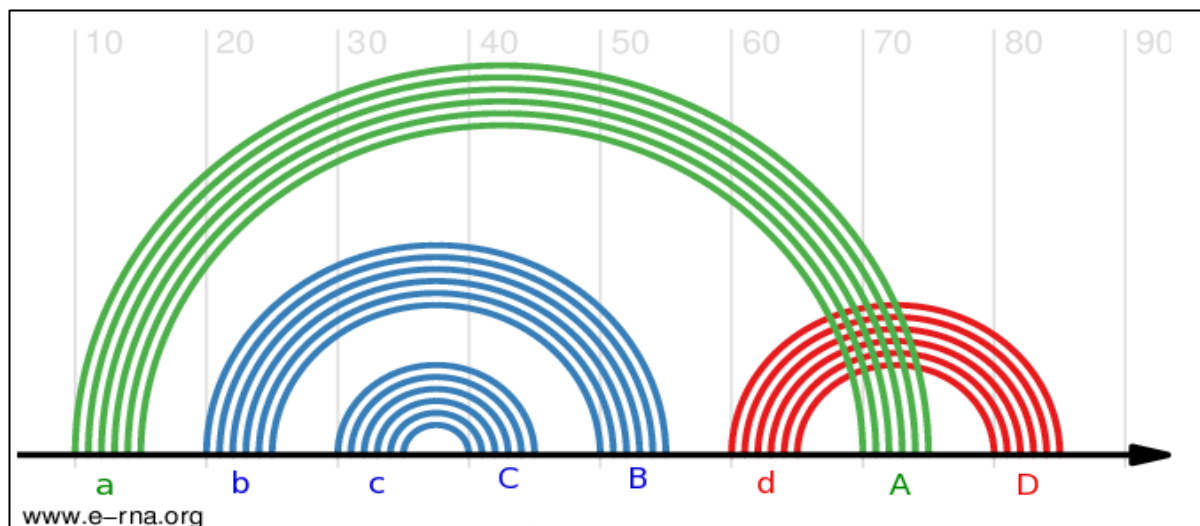


Рисунок 1.6. Дуговая диаграмма крыльев ЭЗУ из примера 2.

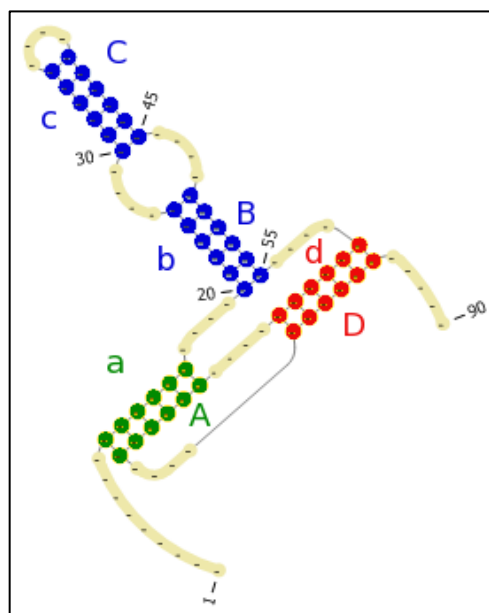


Рисунок 1.7. Схема вторичной структуры ЭЗУ из примера 2.

Пример 3. Пусть ЭЗУ содержит шесть спиралей,  $([2, 7]; [90, 95])$ ,  $([10, 15]; [80, 85])$ ,  $([20, 25]; [50, 55])$ ,  $([30, 35]; [40, 45])$ ,  $([60, 65]; [120, 125])$ ,  $([70, 75]; [110, 115])$ , здесь  $[2, 7]$  и  $[90, 95]$  – крылья спирали  $([2, 7]; [90, 95])$ , и т.д.

Тогда спираль  $([2, 7]; [90, 95])$  обозначим как **aA**, спираль  $([10, 15]; [80, 85])$  – **bB**, спираль  $([20, 25]; [50, 55])$  – **cC**, спираль  $([30, 35]; [40, 45])$  – **dD**, спираль  $([60, 65]; [120, 125])$  – **eE**, а спираль  $([70, 75]; [110, 115])$  – **fF**. Полной сигнатурой данного ЭЗУ будет последовательность **abcdDCefBAFE**. Участок  $[20, 55]$  является суб-ЭЗУ относительно исходного ЭЗУ, а участок  $[30, 45]$  является суб-ЭЗУ относительно участка  $[20, 55]$  (см. Рис. 1.8, 1.9).

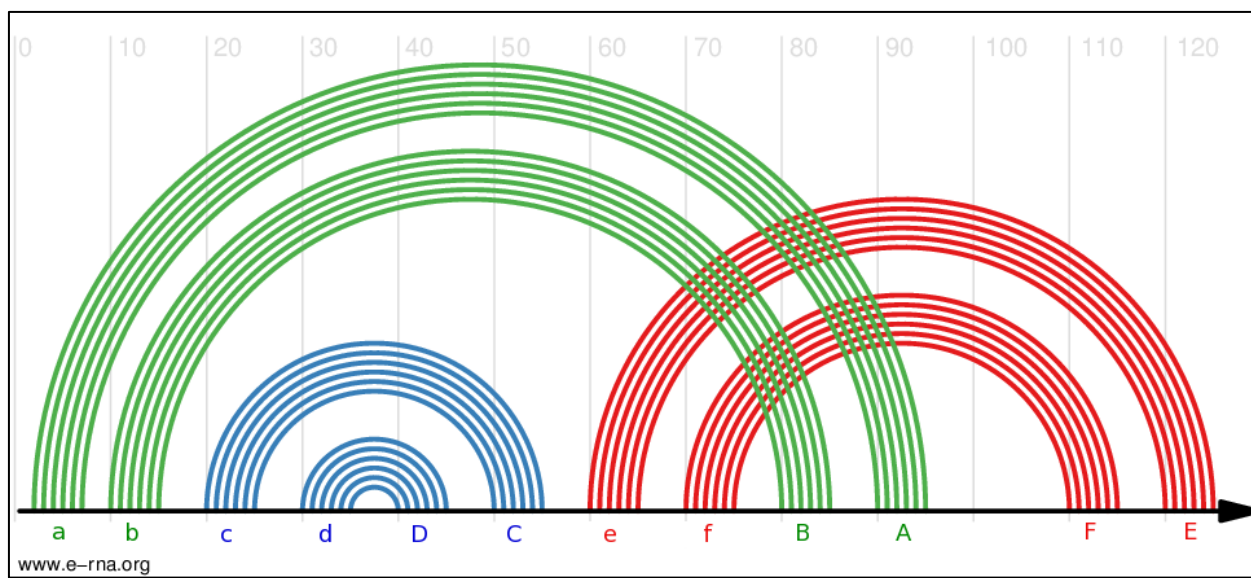


Рисунок 1.8. Дуговая диаграмма крыльев ЭЗУ из примера 3.

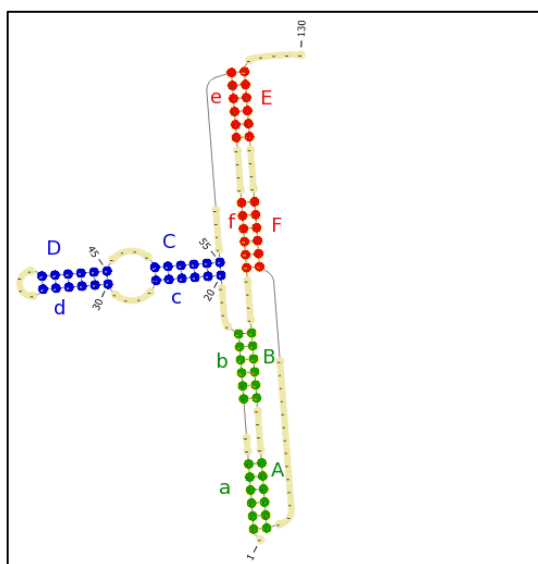


Рисунок 1.9. Схема вторичной структуры ЭЗУ из примера 3.

Верхняя сигнатура ЭЗУ получается из полной сигнатуры в результате:

- 1) Удаление крыльев, соответствующих всем суб-ЭЗУ;



2) Переименования спиралей в порядке использования последовательных букв латинского алфавита (сохраняя порядок крыльев).

Верхней сигнатурой ЭЗУ из примера 1 будет последовательность **aA**; участок **bcBC**, отвечающий суб-ЭЗУ  $[20, 60]$ , был удален из полной сигнатуры **abcBCA**.

Верхней сигнатурой ЭЗУ из примера 2 будет последовательность **abAB**. Сначала, участок **bcCB**, отвечающий суб-ЭЗУ  $[20, 55]$ , был удален из полной сигнатуры **abcCBdAD**. Затем мы, заменяя **d** и **D** на **b** и **B**, получем **abAB**.

Аналогично, верхней сигнатурой ЭЗУ из примера 3 будет последовательность **abcdBADC**.

Спирали **xX**, **yY**, **zZ**, ... называются связанными в верхней сигнатуре, если оба слова **xuz...** и **...ZYX** являются подстроками верхней сигнатуры.

*Сигнатура ЭЗУ (усечённая сигнатура ЭЗУ)* – это последовательность, полученная из верхней сигнатуры в результате:

1) удаления всех букв, кроме **x** и **X** (первая буква левой части и последняя буква правой части), соответствующих цепочкам связанных спиралей.

2) Переименования спиралей в порядке использования последовательных букв латинского алфавита (сохраняя порядок крыльев).

Сигнатуры ЭЗУ из примеров 1 и 2 совпадают со своими верхними сигнатурами. Сигнатурой ЭЗУ из примера 3 будет последовательность **abAB**, что совпадает с сигнатурой ЭЗУ из примера 2.

Примеры типичных сигнатур:

a) Н-узел (H-knot): **abAB**;

b) «Целующиеся петли» (Kissing Loops): **abAcBC**;

c) Тройной узел (Triple knot): **abcABC**.

## 1.5 Стемовые мотивы

*Обобщенный стем* (free stem) – это такая последовательность пар  $(i, j)$ ,  $(i + 1, j - 1)$ , ...,  $(i + k, j - k)$ , что:

1)  $k \geq 1$ ;

2)  $i + k < j - k$ ;

3) Все пары  $(i + x, j - x)$ , где  $x = 0, \dots, k$ , образуют спаривания любого типа.

*Обратный стем* (reversed stem) – это такая последовательность пар  $(i, j)$ ,  $(i + 1, j + 1)$ , ...,  $(i + k, j + k)$ , что:

1)  $k \geq 1$ ;

2)  $i + k < j$ ;

3) Все пары  $(i + x, j + x)$ , где  $x = 0, \dots, k$ , образуют спаривания любого типа.

Здесь и далее под словом *линк* мы будем понимать одиночное спаривание, которое не является частью никакого обобщенного или обратного стема.

*Стемовый мотив* – это такой связный ориентированный граф  $(V, E)$ , что:

1)  $V = V_F + V_R + V_N$ , где:

a)  $V_F$  – это множество крыльев обобщенных стемов;

b)  $V_R$  – это множество крыльев обратных стемов;

c)  $V_N$  – это множество нуклеотидов;

2)  $E = E_F + E_R + E_L + E_E + E_I + E_S$ , где для вершин  $V_1, V_2$ :

a)  $(V_1, V_2)$  принадлежит  $E_F$ , если  $V_1$  и  $V_2$  соответственно левое и правое крыло одного обобщенного стема;

b)  $(V_1, V_2)$  принадлежит  $E_R$ , если  $V_1$  и  $V_2$  соответственно левое и правое крыло одного обратного стема;

c)  $(V_1, V_2)$  принадлежит  $E_L$ , если существует такой линк, что его левый конец принадлежит  $V_1$ , а правый конец принадлежит  $V_2$ ;

d)  $(V1, V2)$  принадлежит  $EE$ , если  $V1$  вкладывается (является частью) в  $V2$ ;

e)  $(V1, V2)$  принадлежит  $EI$ , если  $V1$  и  $V2$  пересекаются, причем  $V1$  располагается ближе к 5'-концу цепи РНК;

f)  $(V1, V2)$  принадлежит  $ES$ , если  $V2$  примыкает к  $V1$  справа (вершины располагаются одна за другой в цепи РНК).

#### Алгоритм разметки стемовых мотивов:

1) Вход: множества  $VF$  и  $VR$ , множества  $EF$  и  $ER$ , множество линков;

2) Из множеств  $VF$  и  $VR$  извлекаем множества  $EE$ ,  $EI$  и  $ES$  (попарное сравнение последовательностей крыльев);

3) Из множества линков получаем множества  $EL$  и  $VN$ , причем  $VN$  состоит только из тех концов линков, которые не являются частью никакого крыла (если нуклеотид не принадлежит никакому крылу и является концом более, чем одного линка, то для него создается только одна вершина из  $VN$ );

4) Дополняем множество  $ES$  ребрами между крыльями и нуклеотидами;

5) Компоненты связности такого графа являются стемовыми мотивами (пример см. Рис. 1.10).

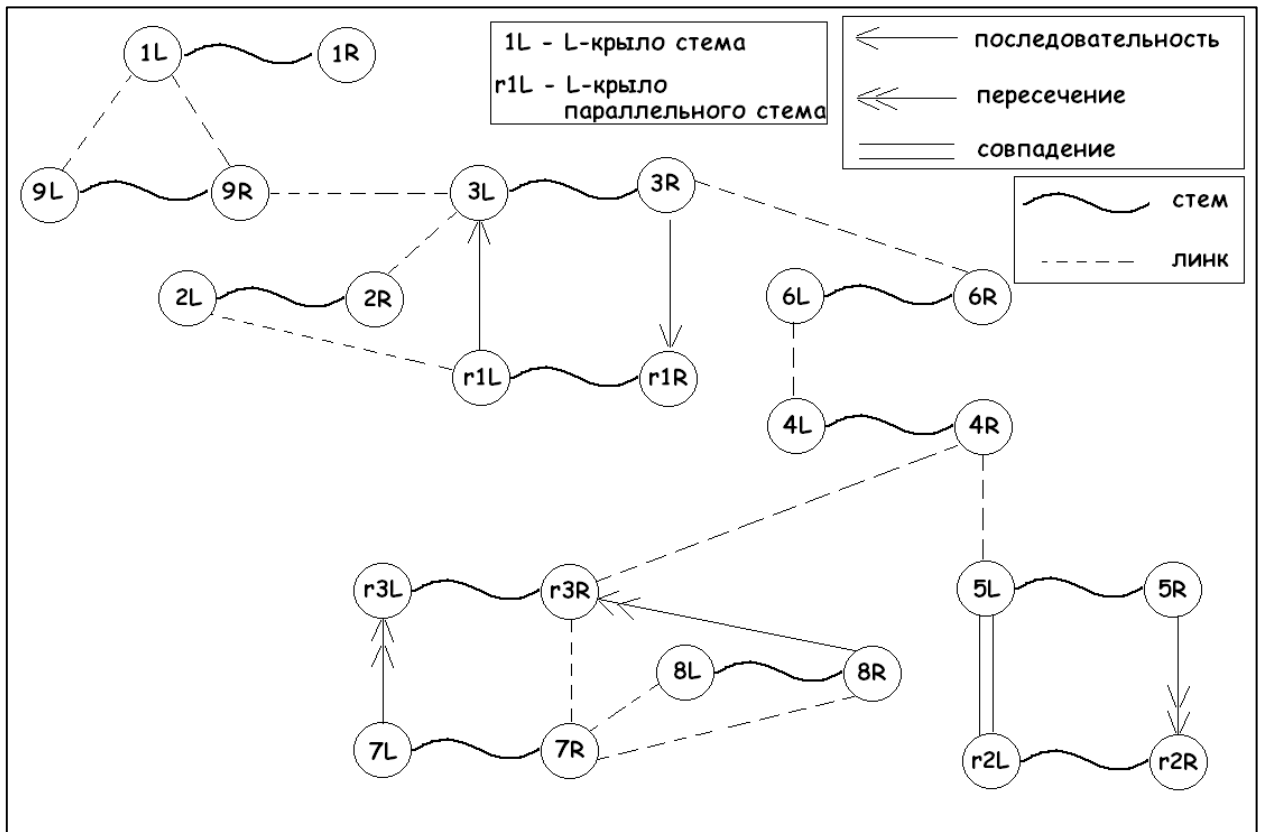


Рисунок 1.10. Пример сложного стемового мотива.

Замечание 1. Если существует линк между нуклеотидами, принадлежащими некоторому крылу  $V1$ , то в EL добавляем ребро  $(V1, V1)$ .

Замечание 2. Простейшими видами стемовых мотивов являются отдельные линки и отдельные стемы (обобщенные или обратные).

## ГЛАВА 2. База данных

### 2.1 Исходные данные

В качестве исходных данных были выбраны все РНК-содержащие экспериментально определенные пространственные структуры из банка данных PDB (Protein Data Bank, [www.rcsb.org](http://www.rcsb.org), version 3.3, [17]). По состоянию на декабрь 2014 года база данных включала структуры из более чем 3300 документов банка PDB в .pdb формате (см. Рис. 2.1). В связи с тем, что с начала 2015 года основным форматом PDB стал формат mmCIF [18], разработанная нами база данных также перешла на использование документов в данном формате. Преимуществом формата mmCIF (см. Рис. 2.2) является отсутствие ограничений на размер документа, что позволило объединить в один файл многие разделенные из-за ограничений формата .pdb структуры.

LINK	MG	MG 02926		OP1	A 02483	1555	1555	1.69
LINK	MG	MG 02926		OP1	C 02534	1555	1555	1.98
SITE	1 AC1	1 CYS X 22						
SITE	1 AC2	3 G 02102	G 02482		C 02536			
SITE	1 AC3	7 A 02483	A 02532		C 02533		C 02534	
SITE	2 AC3	7 HON 02927	HON 02928		HON 02929			
SITE	1 AC4	7 G 0 627	G 02482		A 02483		C 02534	
SITE	2 AC4	7 HON 02930	HON 02931		HON 02932			
CRYST1	211.660	299.670	573.770	90.00	90.00	90.00	C 2 2 21	8
ORIGX1	1.000000	0.000000	0.000000		0.000000			
ORIGX2	0.000000	1.000000	0.000000		0.000000			
ORIGX3	0.000000	0.000000	1.000000		0.000000			
SCALE1	0.004725	0.000000	0.000000		0.000000			
SCALE2	0.000000	0.003337	0.000000		0.000000			
SCALE3	0.000000	0.000000	0.001743		0.000000			
ATOM	1 O5'	U 0 10	17.684	150.891	105.893	1.00	57.93	O
ATOM	2 C5'	U 0 10	17.611	149.829	104.909	1.00	59.38	C
ATOM	3 C4'	U 0 10	18.984	149.234	104.658	1.00	59.35	C
ATOM	4 O4'	U 0 10	19.482	148.632	105.882	1.00	60.32	O
ATOM	5 C3'	U 0 10	19.095	148.129	103.612	1.00	58.65	C
ATOM	6 O3'	U 0 10	19.139	148.716	102.291	1.00	55.97	O

Рисунок 2.1. Фрагмент документа 1FFK из банка PDB в формате .pdb

```

_struct_ref_seq.pdbx_auth_seq_align_beg
_struct_ref_seq.pdbx_auth_seq_align_end
1 1 2KMJ A 1 ? 28 ? 2KMJ 16 43 16 46
2 2 2KMJ B 1 ? 4 ? 2KMJ 1 4 1 4
3 2 2KMJ C 1 ? 4 ? 2KMJ 1 4 1 4
#
loop_
  _chem_comp.id
  _chem_comp.type
  _chem_comp.mon_nstd_flag
  _chem_comp.name
  _chem_comp.pdbx_synonyms
  _chem_comp.formula
  _chem_comp.formula_weight
G 'RNA linking' y "GUANOSINE-5'-MONOPHOSPHATE" ? 'C10 H14 N5 O8 P' 363.223
C 'RNA linking' y "CYTIDINE-5'-MONOPHOSPHATE" ? 'C9 H14 N3 O8 P' 323.199
A 'RNA linking' y "ADENOSINE-5'-MONOPHOSPHATE" ? 'C10 H14 N5 O7 P' 347.224
U 'RNA linking' y "URIDINE-5'-MONOPHOSPHATE" ? 'C9 H13 N2 O9 P' 324.183
DAR 'D-peptide linking' . D-ARGININE ? 'C6 H15 N4 O2 1' 175.210
ZUK 'D-peptide linking' . 5-PYRIMIDIN-2-YL-D-NORVALINE ? 'C9 H13 N3 O2' 195.221
NH2 NON-POLYMER . 'AMINO GROUP' ? 'H2 N' 16.022
#

```

Рисунок 2.2. Фрагмент документа 2KMJ из банка PDB в формате mmCIF.

По состоянию на май 2015 года база URSDB включает 2812 документов банка PDB, из которых 1538 документов содержат структуры РНК-белковых комплексов, 1087 документов содержат структуры одиночных РНК, 143 документа содержат структуры РНК в комплексе с ДНК и белками и 44 документа содержат структуры РНК в комплексе с ДНК. Стоит отметить, что документ банка PDB может содержать несколько моделей одной и той же структуры. Это связано с тем, что в ходе экспериментального определения структуры не всегда удается однозначно восстановить координаты всех её атомов. Описанные 2812 документов в формате mmCIF содержат в общей сложности 9443 модели структур. Всего представлено 7302 цепи РНК (без учета представления одной цепи РНК в нескольких моделях).

Для разметки водородных связей, образующих вторичную структуру РНК, была использована программа DSSR [19]. Данная программа была выбрана среди более распространенных аналогов [20-22] как самая свежая и обладающая наиболее богатым функционалом. Более того, мы принимали активное участие в её тестировании. Выходные данные программы DSSR

содержат 7 файлов с аннотациями спариваний, элементов вторичной структуры и некоторых третичных мотивов. Основной выходной файл имеет формат .out (см. Рис. 2.3) и содержит подробное описание спариваний между основаниями нуклеотидов, а также других водородных связей и стекинг-взаимодействий.

```

List of 20 base pairs
  nt1          nt2          bp name          Saenger    LW DSSR
1  .A.GB.1.     .A.CC.18.         g-c WC        19-XIX     cWW cW-W
   [-175.5(anti) C2'-exo lambda=53.1] [-158.3(anti) C3'-endo lambda=49.9]
   d(C1'-C1')=10.87 d(N1-N9)=9.03 d(C6-C8)=9.78 tor(C1'-N1-N9-C1')=-8.7
   H-bonds[3]: "O6(carbonyl)-N4(amino)[2.69],N1(imino)-N3[2.94],N2(amino)-O2(carbonyl)[3.13]"
   bp-pars: [0.08  -0.07  -0.35  -21.13  -17.16  -6.56]
2  .A.G.2.     .A.A.16.         G-A Imino     08-VIII    cWW cW-W
   [-168.8(anti) C3'-endo lambda=33.4] [-156.8(anti) C3'-endo lambda=31.3]
   d(C1'-C1')=13.31 d(N1-N9)=10.82 d(C6-C8)=10.44 tor(C1'-N1-N9-C1')=9.3
   H-bonds[1]: "O6(carbonyl)-N6(amino)[2.47]"
   bp-pars: [0.34  1.25  -2.48  -9.11  -27.21  -44.75]
3  .A.G.2.     .A.C.17.         G-C WC        19-XIX     cWW cW-W
   [-168.8(anti) C3'-endo lambda=52.2] [-154.9(anti) C3'-endo lambda=47.4]
   d(C1'-C1')=10.96 d(N1-N9)=9.06 d(C6-C8)=9.79 tor(C1'-N1-N9-C1')=-25.3
   H-bonds[3]: "O6(carbonyl)-N4(amino)[2.80],N1(imino)-N3[2.92],N2(amino)-O2(carbonyl)[3.19]"
   bp-pars: [0.03  -0.12  0.09  -16.39  -27.61  -8.11]
4  .A.U.3.     .A.A.16.         U-A WC        20-XX     cWW cW-W
   [-157.4(anti) C3'-endo lambda=49.8] [-156.8(anti) C3'-endo lambda=50.9]
   d(C1'-C1')=10.97 d(N1-N9)=9.09 d(C6-C8)=9.87 tor(C1'-N1-N9-C1')=-24.4
   H-bonds[2]: "N3(imino)-N1[2.94],O4(carbonyl)-N6(amino)[2.88]"
   bp-pars: [-0.11  -0.07  0.16  -3.57  -27.58  -8.11]

```

Рисунок 2.3. Фрагмент основного выходного файла программы DSSR в формате .out.

Для каждой структуры из банка PDB в качестве исходных данных использовались координаты атомов и описание цепей биополимеров из документа в формате mmCIF и разметка спариваний оснований из выходного файла программы DSSR в формате .out.

## 2.2 Обработка структур

Для обработки исходных данных был разработан программный комплекс, реализованный на языке программирования Python 3. Программа выполнена в виде библиотеки, состоящей из 27 модулей. Такая реализация позволяет добавлять новые модули, изменять или удалять существующие модули независимо от остального кода. Общий объём кода составляет 9391 строку и занимает 324 килобайта.

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 три части:

1) Подготовка исходных данных (разделение документов на модели, прогонка моделей через DSSR);

Общее время работы на всей выборке: 20 часов 46 минут (99% времени занимает работа программы DSSR).

2) Конструктор текстовых файлов для наполнения базы данных (парсинг моделей и out-файлов, разметка элементов вторичной структуры, разметка РНК-белковых и других взаимодействий, разметка сигнатур псевдоузлов, разметка стемовых мотивов).

Общее время работы на всей выборке: 3 часа 10 минут.

3) Конструктор скрипта в формате SQL для создания базы данных на SQL сервере и её наполнения.

Общее время работы скрипта MySQL: 24 минуты 12 секунд.

Алгоритм разметки водородных связей между РНК и белком был взят из [23] с согласия авторов. Данный алгоритм использует численный инвариант, который отражает достоверность водородных связей – от 0.1 (недостоверная связь) до 1.0 (достоверная связь).

Начиная с марта 2015 года для работы с разработанным программным комплексом используется система контроля версий Git [24]. Также, с помощью веб-ресурса GitLab (аналога GitHub), запущенного в нашей лаборатории, создан открытый репозиторий с исходным кодом, доступный по адресу <http://git.lpm.org.ru/baulin/ursdb>.

Приложение А содержит подробное описание модулей разработанного комплекса.



## 2.3 Схема и данные

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

Предлагаемая схема состоит из 51 таблицы и содержит исчерпывающий набор данных, необходимых для дальнейших исследований, включая таблицы спиралей, нитей, петель, граней, псевдоузлов и др. Подробное описание таблиц см. в приложении В.

Таблицы базы данных URSDB можно условно разделить на 4 блока данных:

- 1) Блок данных, полученных в результате парсинга mmCIF документов (таблицы моделей, молекул, цепей биополимеров, мономеров, атомов);
- 2) Блок данных, полученных в результате переработки данных о вторичной структуре РНК (таблицы спариваний, спиралей, петель, крыльев, линков и пр.);
- 3) Блок данных, полученных непосредственно из выходного файла программы DSSR, предназначенный для верификации разметки (таблицы некоторых третичных мотивов, мультиплетов, и пр.)
- 4) Блок данных, полученных в процессе исследования структуры РНК (таблицы псевдоузлов, стемовых мотивов, РНК-белковых контактов и пр.)

Полный объем текстовых данных составляет 8,5 ГБ. По состоянию на май 2015 года база данных содержит описания 7302 цепей РНК, около 1,2 млн. спариваний и 4764 псевдоузлов. Данные регулярно обновляются (не реже, чем раз в месяц).

## ГЛАВА 3. Веб-интерфейс

### 3.1 Детали реализации

Для тестирования базы данных разработан веб-интерфейс, позволяющий взаимодействовать с базой в режиме онлайн. На данный момент веб-интерфейс находится на стадии бета-версии и доступен по адресу <http://server3.lpm.org.ru/urs/>.

Веб-интерфейс реализован в виде набора CGI скриптов, выполненных на языке программирования Python 2. В процессе работы использовалась система контроля версий Git. Также, в процессе работы были освоены языки разметки HTML и CSS, а также скриптовый язык программирования JavaScript.

В рамках реализации веб-интерфейса использовались следующие технологии:

1) Local Storage – возможность HTML5, позволяющая сохранять данные на стороне клиента. Была использована для сохранения выдачи результатов поиска в течение сессии пользователя.

2) jQuery – сторонняя библиотека языка JavaScript. Была задействована для выполнения асинхронных запросов к серверу (Ajax) в процессе выдачи информации об отдельной структуре.

3) Ajax – технология выполнения асинхронных запросов к серверу. Использовалась для выдачи информации об отдельной структуре и её визуализации.

4) Jmol – Java-апплет визуализации пространственной структуры. Был интегрирован в окно отдельной структуры для её визуализации.

5) JSmol – HTML5-версия апплета Jmol. Был интегрирован в окно отдельной структуры для её визуализации.

б) ExtJS – сторонняя библиотека CSS. Использовалась для оформления окна отдельной структуры.

Общий объем кода HTML, CSS и JavaScript составляет 4455 строк (без учета сторонних библиотек). Общий объем CGI скриптов составляет 5267 строк.

### **3.2 Общая структура и функционал**

Разработанный интерфейс позволяет пользователю формировать выборку интересующих его РНК-содержащих структур, после чего собирать статистику по имеющимся в них структурным элементам. Кроме того пользователю доступны индивидуальные структуры и элементы с подробными данными для детального анализа.

Формирование выборки структур происходит согласно запросу к базе данных; запрос состоит из конъюнкции нескольких дизъюнкций элементарных условий. Элементарные условия, доступные пользователю на данный момент, состоят из 4 групп: (1) общая информация о PDB документе, (2) информация о содержащихся макромолекулах, (3) шаблоны структурных элементов, а также (4) информация о содержащихся спариваниях между основаниями в РНК и между атомами РНК и белка. Результатом запроса является список структур, удовлетворяющих введенным параметрам (см. Рис. 3.1).

Universe of RNA Structures

Structures    Statistics    Help    About Us    Links

Query:  New Search ▾  
First Model Only ▾

Help ⓘ

- General Information Help ⓘ
- Contained Molecules Help ⓘ
- Contained RNA Structure Patterns Help ⓘ
- Contained Interactions Help ⓘ

Time used: 2.82 sec

Query History Help ⓘ

(Method:X-RAY\_DIFFRACTION AND Resolution<3.0A)

Select Fields Help ⓘ

N    PDB ID    Header    Date    Method    Resolution

**Sort by:**    PDB ID    1 to N (A to Z)   

Structures List         Help ⓘ

Check All    Structures found: 1271

N	PDB ID (# Models)	Header	Date	Method	Resolution
<input checked="" type="checkbox"/> 1	<a href="#">157D</a> (1)	RNA	1994-02-01	X-RAY DIFFRACTION	1.8
<input checked="" type="checkbox"/> 2	<a href="#">165D</a> (1)	DNA-RNA HYBRID	1994-03-21	X-RAY DIFFRACTION	1.55
<input checked="" type="checkbox"/> 3	<a href="#">1A34</a> (1)	Virus/RNA	1998-01-28	X-RAY DIFFRACTION	1.81

Рисунок 3.1. Страница поиска структур.

Для индивидуального анализа интересующей пользователя структуры имеется окно структуры, доступное по клику на ID структуры в списке результатов поиска структур. Данное окно содержит детальную информацию об имеющихся в структуре цепях, спариваниях, петлях, спиралях и псевдоузлах. В будущем этот список будет пополняться. Также в окно встроен 3D-визуализатор молекул JSmol [25] (см. Рис. 3.2).

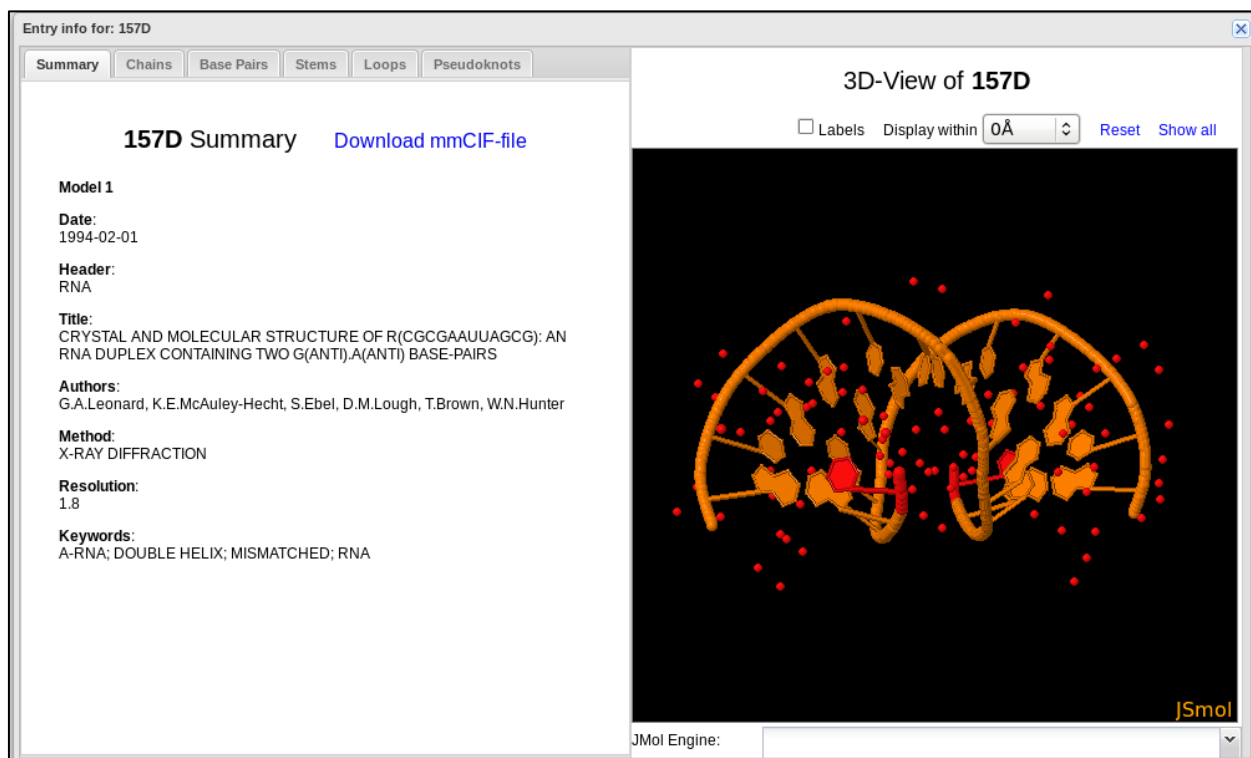


Рисунок 3.2. Окно отдельной структуры

После формирования выборки документов пользователь имеет возможность собирать статистику структурных элементов, содержащихся в выбранных структурах. На данный момент в качестве таких элементов доступны цепи, спаривания, петли, спирали, псевдоузлы, мультиплеты, а также РНК-белковые взаимодействия. По каждому виду элемента доступны фильтры, позволяющие ограничивать элементы необходимыми параметрами, а также полные списки найденных элементов с возможностью вывода детальной информации об отдельном элементе по клику (см. Рис. 3.3,3.4).

**Filters**

ECR Pattern (as stem description)

Rank1 from  to  Rank2 from  to  Depth from  to

Signature:

Show a list of pseudoknots Sort by:

---

**Statistics**

**Pseudoknots:**  
Count: 1137

Depth (number of parent ECFs): max = 9 min = 0 mean = 2.15 std = 2.19  
 Rank1 (#different brackets): max = 4 min = 2 mean = 2.12 std = 0.38  
 Rank2 (max #loops per thread): max = 3 min = 2 mean = 2.03 std = 0.16

Signature	Rank 1	Rank 2	Count
abAB	2	2	786
abAcBC	2	2	206
abAcdeDfgFEChBHiGI	3	2	68
abcdCefgFhiHGEjkAKIDLjmIMB	3	2	15
abcdCADB	2	2	13
abcBdefEghGFDijkJlCLImHMAK	4	3	12
abAcdeDfgFEChijlKbKHIGLJ	4	3	8
abAcBdCD	2	2	4
abcdBCAD	3	3	4
abAcdeBEfDFC	2	2	4
abAcdefEghGFDiBijHjC	3	2	4
abcBdefECDAF	3	3	3
abAcBdefEghiHjkJlGlmCMnFNLoKOD	3	2	3
abAcBDeCE	2	2	2
abcdCefAFDEB	3	2	1
abcBdeAEfgDGChFH	2	2	1
abcBdefEghGFDijAjkCKlIHL	3	3	1
abcBdefEghGFDijkJlCLImHMANoNKO	4	3	1
abAcdeDBefghGijkJlmLkInoCOpHPNqMQF	3	2	1

Рисунок 3.3. Результат запроса статистики псевдоузлов

Element info for: 1FFK, pseudoknot

**1FFK, pseudoknot**

Scheme: 1 2 -1 -2  
 Signature: abAB  
 Brackets: (())

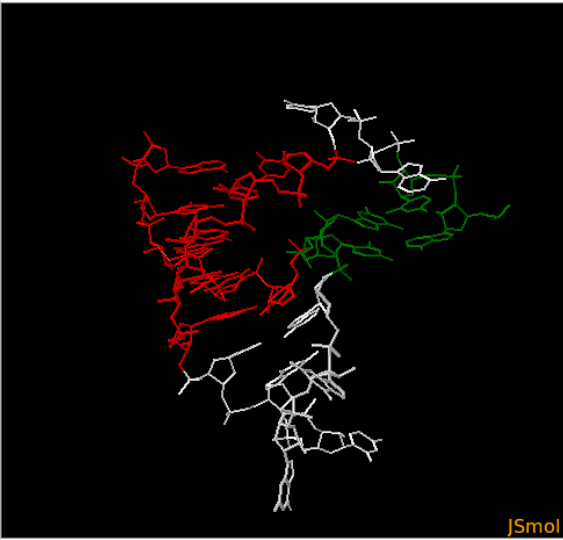
Rank 1 = 2  
 Rank 2 = 2  
 Depth = 1

**Wings:**

No	Length	Start	End	Sequence
1	2	0.A.2060.	0.C.2061.	AC
2	5	0.U.2064.	0.G.2068.	UCCAG
-1	2	0.G.2075.	0.U.2076.	GU
-2	5	0.C.2077.	0.A.2081.	CUGGA

**3D-View of 1FFK, pseudoknot**

Labels Display within



JMol Engine:

Рисунок 3.4. Окно отдельного псевдоузла

### 3.3 Обзор существующих аналогов

В настоящее время существует около 15 баз данных, которые содержат сведения о пространственных структурах РНК. Отметим, что кроме баз структур большое значение при изучении РНК имеют базы, основанные на множественном выравнивании семейств РНК, например, Rfam [26], однако в настоящем обзоре они не рассматриваются.

Все базы пространственных структур РНК основаны на данных PDB (в случае RNAStrand [27] – и на других источниках) и различаются возможностями поиска (например, поиск только по структурам в целом или по отдельным элементам), полнотой охвата известных структур (например, включены ли в базу сведения о РНК-белковых комплексах), полнотой описания отдельных структур (некоторые базы содержат сведения только о петлях, в большинстве баз нет сведений о псевдоузлах). К сожалению, ни одна из них не обеспечивает полный набор возможностей. Полный список рассмотренных баз данных см. в приложении С.

Все существующие базы данных можно условно разделить на три типа: вспомогательные базы для структурного моделирования, базы, посвященные отдельным элементам вторичной структуры РНК и универсальные базы пространственных структур РНК. Остановимся подробнее на последнем классе баз, к которому принадлежит и разрабатываемая нами база. Наиболее интересными с точки зрения полноты имеющихся данных и функциональности веб-интерфейса представляются базы RnaFRABASE [28] и RNAStrand. Можно отметить такие специфические достоинства указанных баз, как наличие гибкого поиска структурных элементов (RnaFRABASE) и обширность и полноту представленных данных (RNAStrand). Однако данные базы имеют и важные недостатки, так, например, в RnaFRABASE исключены псевдоузловые структуры, а в RNAStrand отсутствует возможность поиска отдельных

структурных элементов. Более того, как было выяснено, в RNAStrand представлены не все содержащие РНК структуры из PDB.

К универсальным можно отнести и базу RNA3DHub [29]. Однако данная база не рассматривает такой важный элемент вторичной структуры РНК, как спирали, а, значит, не вполне соответствует требованиям универсальности. Тем не менее, с точки зрения выделения мотивов, которые встречаются в различных пространственных структурах, база RNA3DHub, как и другие разработки группы Н.Леонтиса и К. Зирбея, представляют большой интерес. С точки зрения удобства пользования и полноты охвата известных структур интерес представляет и база NPIDB [23], однако эта база специализирована для анализа нуклеиново-белковых контактов и не содержит разметки вторичной структуры РНК.

Следует отметить, что, в отличие от многих биоинформатических баз данных, в частности, баз данных по первичным структурам РНК и белков, базы данных пространственных структур РНК используются, в основном, для гомологического моделирования, подбора knowledge-based параметров и тестирования программ предсказания и сравнения структур РНК [30-36] (в дополнение к упоминавшимся базам отметим специализированную базу BRASERO [37]). По нашему мнению, их потенциал для собственно биологических исследований используется недостаточно. Разработанная нами база данных URSDB призвана заполнить этот пробел.



## ГЛАВА 4. Результаты

### 4.1 Анализ псевдоузлов

Был проведен анализ имеющихся в базе псевдоузлов. Все псевдоузлы были разбиты на классы согласно предложенному методу. Каждому классу псевдоузлов соответствует определенная сигнатура крыльев. Количественный анализ показал, что наиболее распространены H-узлы (сигнатура abAB, 3503 случаев) и kissing hairpins (сигнатура abAcBC, 669 случаев). Всего в базе имеется 47 различных классов псевдоузлов. В таблице 4.1 указаны сигнатуры классов и количество реализаций этих сигнатур в базе данных.

Таблица 4.1. Распределение сигнатур псевдоузлов

Сигнатура	Количество
abAB	3503
abAcBC	669
abAcdeDfgFEChBHIgI	213
abAcdeBEfDFC	54
abAcdefEghGFDiBIjHJC	46
abAcBdefEghiHjkJlGlmCMnFNLoKOD	35
abcdCefgFhiHGEjkAKIDLmIMB	30
abAcBdeDfghGijlHFkIClEMKnJN	30
abcBdefEghGFDijAJkCKIIHL	15
abcdCADB	14
abAcdeBEfDF	14
abcdBCAD	14
abcBdefEghGFDijkJlClmHMAK	14
abAcBdCD	13
abAcBDeCE	13
abcBdefEghGFDijkJlClmHMANoNKO	12
abAcdeDfgFEChijkBKHIGLJ	11
abcBdefECDAF	7
abcABC	5
abAcdeDBEfgHijkJlmLKInoCOpHPNqMQF	5
abAcdeDBfghGijkJlmLKInoEOpHPNqMQF	5
abcBdACefgFhijklKJHmndNoGOMpLPE	4
abcdCefAFDEB	3
abcBdefEghGFDijkJlClmHMANoNpqPKQO	3
abAcdeDfghiHFECjkGKIBLmIM	3
abcBdeDCAfEF	2
abAcdeDfBFgEGC	2
abAcdefEghijlGFDklHlMbmKnJNC	2
abAcBdefEghCHIFIGD	2
abcdeDfgEGhFBCAH	2
abcdCefAFgDGEB	2
abcBdefEghGFDijkCKIIHLAJ	2
abcdCABeDE	1
abcdeDBCAfghiEIGHF	1
abcBdefEghiHjCJgFKDAI	1
abcdCefghGijlHFklAlmDMKnJNEB	1

abcBdeAEfgDGChFH	1
abcdCeBEAfDF	1
abcdCeAEfghGDFBH	1
abcBdefEghGFDijkJCLImHMAnopOKNP	1
abcdBDeAEfCF	1
abcBdefEghGFDijkJCLImHMAnopOqrQstSKNPTR	1
abcBdeDfgFhACGEijkJlmnMopONLqRHRsKSQtPTI	1
abcBdefgFhiHGEjCkIKIADL	1
abcBdefEghGFDijkJCLImHMAnopOKqrQNRp	1
abAcdeDfgFEChijkBKHIGLmnMJN	1
abcACdBD	1

Стоит отметить, что сложные сигнатуры (длина которых больше 10) зачастую описывают псевдоузлы, образованные несколькими цепями РНК, т.е. не являющиеся псевдоузлами в строгом смысле. Так, например, сигнатура **abcBdefEghGFDijkJCLImHMAnoNKO** описывает псевдоузел, в котором учатсвуют три цепи. При этом, если удалить спирали между разными цепями, то данная сигнатура разобьется на две сигнатуры: **bcBdefEghGFDiCIIHL** и **abAB**.

## ЗАКЛЮЧЕНИЕ

В рамках работы была спроектирована и реализована база данных экспериментально определенных пространственных структур РНК. Все цепи РНК и их структурные элементы проаннотированы. В частности, размечены все элементы вторичной структуры, в том числе, связанные с псевдоузлами, и нестандартные мотивы, образованные с участием неканонических водородных связей. Подобная база разработана впервые. Она существенно расширит возможности исследования пространственных структур РНК, в том числе, – их неканонических элементов.

Для взаимодействия конечных пользователей с базой разработан веб-интерфейс, обеспечивающий поиск нужных структур в базе, сбор статистики об этих структурах, визуализацию и хранение результатов поиска. Веб-интерфес позволяет работать не только с целыми структурами РНК, но и с различными типами их структурных элементов, таких как, спаривания, петли, псевдоузлы, РНК-белковые контакты и др. Также, веб-интерфейс содержит подробную страницу помощи.

На предварительном этапе исследований были проанализированы такие виды неканонических элементов структур РНК, как псевдоузлы и стемовые мотивы. Был составлен список всех псевдоузловых сигнатур, встречающихся в пространственных структурах РНК. Как и ожидалось, значительное большинство реальных псевдоузлов являются Н-узлами. Сложные сигнатуры, в свою очередь, представлены межцепочечными взаимодействиями и в строгом смысле псевдоузлами не являются.

Также, было разработано описание стемовых мотивов РНК, основанное на графах. Целью данного метода является единая классификация третичных мотивов РНК, которая на данный момент отсутствует. Данный подход является развитием нашей более ранней работы (см. [38]).

## 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. Gorodkin J., Ruzzo W. L. (ed.). *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*. – Humana Press, 2014.
4. 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.
5. 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.
6. 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.
7. 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.
8. Gulyaev AP, Olsthoorn RC. A family of non-classical pseudoknots in influenza A and B viruses. *RNA Biol*. 2010. Vol. 7. P. 125-129.
9. Condon A., Davy B., Rastegari B., Zhao S., Tarrant F. Classifying RNA pseudoknotted structures. *Theoretical Computer Science* 2004. Vol. 320. P. 35 – 50.
10. Batenburg F.H.D. van, Gulyaev 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.

11. Taufer M., Licon A., Araiza R., Mireles D., Batenburg F.H.D. van, Gulyaev 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
12. Leontis N. B., Westhof E. Geometric nomenclature and classification of RNA base pairs // *Rna*. – 2001. – Т. 7. – №. 04. – С. 499-512.
13. 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.
14. 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.
15. Exterior loop in RNA secondary structure: [Электронный ресурс]. URL: <http://x3dna.org/articles/exterior-loop-in-rna-secondary-structure>. (Дата обращения: 18.01.2015)
16. Andersen J. E. et al. Topological classification and enumeration of RNA structures by genus // *Journal of mathematical biology*. – 2013. – Т. 67. – №. 5. – С. 1261-1278.
17. Berman H. M. et al. The protein data bank // *Nucleic acids research*. – 2000. – Т. 28. – №. 1. – С. 235-242.
18. Bourne P. E. et al. The macromolecular crystallographic information file (mmCIF) // *Meth. Enzymol.* – 1997. – Т. 277. – С. 571-590.
19. Lu X. J., Olson W. K., Bussemaker H. J. The RNA backbone plays a crucial role in mediating the intrinsic stability of the GpU dinucleotide platform and the GpUpA/GpA miniduplex // *Nucleic acids research*. – 2010. – Т. 38. – №. 14. – С. 4868-4876.
20. Sarver M. et al. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures // *Journal of mathematical biology*. – 2008. – Т. 56. – №. 1-2. – С. 215-252.

21. Yang H. et al. Tools for the automatic identification and classification of RNA base pairs //Nucleic acids research. – 2003. – T. 31. – №. 13. – C. 3450-3460.
22. Gendron P., Lemieux S., Major F. Quantitative analysis of nucleic acid three-dimensional structures //Journal of molecular biology. – 2001. – T. 308. – №. 5. – C. 919-936.
23. Kirsanov D. D. et al. NPIDB: nucleic acid—protein interaction database //Nucleic acids research. – 2012. – C. gks1199.
24. Loeliger J., McCullough M. Version Control with Git: Powerful tools and techniques for collaborative software development. – " O'Reilly Media, Inc.", 2012. - -
25. Hanson R. M. et al. JSmol and the Next Generation Web Based Representation of 3D Molecular Structure as Applied to Proteopedia //Israel Journal of Chemistry. – 2013. – T. 53. – №. 3 4. – C. 207-216.
26. Griffiths-Jones S. et al. Rfam: annotating non-coding RNAs in complete genomes //Nucleic acids research. – 2005. – T. 33. – №. suppl 1. – C. D121-D124.
27. Andronescu M. et al. RNA STRAND: the RNA secondary structure and statistical analysis database //BMC bioinformatics. – 2008. – T. 9. – №. 1. – C. 340.
28. Popena M. et al. RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures //BMC bioinformatics. – 2010. – T. 11. – №. 1. – C. 231.
29. Petrov A. I., Zirbel C. L., Leontis N. B. Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas //RNA. – 2013. – T. 19. – №. 10. – C. 1327-1340.
30. Kirillova S., Tosatto S. C. E., Carugo O. FRASS: the web-server for RNA structural comparison //BMC bioinformatics. – 2010. – T. 11. – №. 1. – C. 327
31. Wang C. W., Chen K. T., Lu C. L. iPARTS: an improved tool of pairwise alignment of RNA tertiary structures //Nucleic acids research. – 2010. – T. 38. – №. suppl 2. – C. W340-W347.

32. Muppirala U. K., Honavar V. G., Dobbs D. Predicting RNA-protein interactions using only sequence information //BMC bioinformatics. – 2011. – Т. 12. – №. 1. – С. 489.
33. Nebel M. E., Weinberg F. Algebraic and combinatorial properties of common RNA pseudoknot classes with applications //Journal of Computational Biology. – 2012. – Т. 19. – №. 10. – С. 1134-1150.
34. Rother M. et al. ModeRNA: a tool for comparative modeling of RNA 3D structure //Nucleic acids research. – 2011. – Т. 39. – №. 10. – С. 4007-4022.
35. Zhao Y. et al. Automated and fast building of three-dimensional RNA structures //Scientific reports. – 2012. – Т. 2.
36. Puton T. et al. CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction //Nucleic acids research. – 2013. – Т. 41. – №. 7. – С. 4307-4323.
37. Allali J. et al. BRASERO: A resource for benchmarking RNA secondary structure comparison algorithms //Advances in bioinformatics. – 2012. – Т. 2012.
38. Баулин Е. Ф., Ройтберг М. А. Стемовые мультиплеты: новый подход к описанию третичных мотивов РНК // Математическая биология и биоинформатика. 2015. Т. 10. № 1. С. 54-59. doi: 10.17537/2015.10.54

## ПРИЛОЖЕНИЕ А. Структура программного обеспечения

- **Build.py** – Главный скрипт, конструктор текстовых файлов для наполнения базы данных.
- **DB.py** – Скрипт вызова конструкторов таблиц базы данных; объединяет Tables.py, LuTables.py и ExtraTables.py.
- **db\_script.py** – Конструктор SQL скрипта для инициализации и наполнения базы данных.
- **DSSR.py** – Парсер выходного файла программы DSSR.
- **ExtraTables.py** – Конструктор дополнительных и тестовых таблиц базы данных.
- **LuTables.py** – Конструктор таблиц, описывающих аннотации элементов структур РНК, полученные из выходного файла программы DSSR.
- **Merge.py** – Скрипт, объединяющий классы mmCIF\_Model и DSSR\_Model в один класс.
- **mmCIF.py** – Парсер mmCIF моделей.
- **PDB.py** – Парсер PDB моделей.
- **RSS.py** – Скрипт вызова разметок отдельных элементов структур РНК; объединяет модули библиотеки SecStruct.
- **SplitmmCIF.py** – Скрипт разделения mmCIF файлов на модели.
- **SplitPDB.py** – Скрипт разделения PDB файлов на модели.
- **Tables.py** – Конструктор основных таблиц базы данных.
- **Tools.py** – Библиотека вспомогательных функций и классов; вызывается из парсеров входных файлов.



- **SecStruct library**
  - **ChainOrder.py** – Модуль сортировки цепей РНК при наличии межцепочечных взаимодействий.
  - **Graph.py** – Модуль для работы с графами; вызывается из Multiplets.py.
  - **Interactions.py** – Модуль разметки РНК-белковых взаимодействий.
  - **Links.py** – Модуль разметки и классификации линков.
  - **Loops.py** – Модуль разметки и классификации петель.
  - **Maps.py** – Набор карт атомов, которые могут образовывать водородные связи; вызывается из Interactions.py.
  - **Mask.py** – Модуль предварительной разметки стемов; вызывается из Wings.py.
  - **Multiplets.py** – Модуль разметки стемовых мотивов.
  - **Sign.py** – Модуль разметки и классификации псевдоузлов.
  - **Stems.py** – Модуль разметки стемов.
  - **Threads.py** – Модуль разметки нитей.
  - **Towers.py** – Модуль разметки башен, т.е. наборов последовательных стемов, разделенных внутренними петлями и выпячиваниями.
  - **Wings.py** – Модуль разметки крыльев.

## ПРИЛОЖЕНИЕ В. Схема базы данных

Таблица А.1. Список таблиц URSDb.

Номер	Таблица	Комментарий
1	files	Table of Files
2	models	Table of Models
3	molecules	Table of Molecules
4	chains	Table of chains
5	nucls	Table of Nucleotides
6	aminos	Table of Aminoacids
7	ligands	Table of Ligands (excluding metals and waters)
8	atoms	Table of all atoms
9	BasePairs	Table of Base-Pairs
10	towers	Table of Towers
11	links	Table of Links
12	stems	Table of DSSR Stems
13	StemsOld	Table of Old Stems
14	StemsFull	Table of Full Stems
15	wings	Table of Wings
16	WingsOld	Table of Old Wings
17	WingsFull	Table of Full Wings
18	threads	Table of Threads
19	hairpins	Table of Hairpins
20	bulges	Table of Bulges
21	internals	Table of Internal loops
22	junctions	Table of Junctions
23	threadloop	Table of Thread-Loop relations
24	wingloop	Table of Wing-Loop relations
25	faceloop	Table of Face-Loop relations
26	ChainsConcat	Table of relations between chains
27	LuHelices	Table of Helices
28	LuStems	Table of LuStems
29	LuStemHelix	Table of LuStem-LuHelix relations
30	LuMultiplets	Table of Multiplets
31	LuNphbs	Table of Non-Pairing interactions
32	LuNonLoops	Table of Non-Loop segments
33	LuKissing	Table of kissing loops
34	LuAminors	Table of A-minors
35	LuUturns	Table of U-turns
36	LuZippers	Table of Ribose Zippers
37	LuKturns	Table of Kink-turns
38	LuHairpins	Table of Hairpins
39	LuBulges	Table of Bulges
40	LuInternals	Table of Internal loops
41	LuJunctions	Table of Junctions
42	LuJuncThreads	Table of Junction threads
43	RevStems	Table of Reversed Stems
44	NuclMults	Table of nucleotide multiplets
45	StemMults	Table of stem-multiplets
46	StemNPairs	Table of stem-pairs (edge=nucl)
47	StemLPairs	Table of stem-pairs (edge=link)
48	diagrams	Table of bracket diagrams
49	ecfs	Table of elementary closed fragments
50	atompairs	Table of H-bonds between RNA and protein atoms
51	monopairs	Table of H-bonds between RNA and protein monomers

Таблица А.2. Файлы.

Номер	Поле	Комментарий	Тип	Пример
1	name	Unique PDB document identifier	char(4)	157D
2	models	Number of models	int	1
3	type	R - only RNA; RP - RNA+Protein; RD - RNA+DNA; RPD - RNA+DNA+Protein	varchar(3)	R
4	head	Header of document	text	RNA
5	date	Date of deposition	date	1994-02-01
6	title	Title of document	text	CRYSTAL AND MOLECULAR STRUCTURE OF R(CGCGAAUUAGCG): AN RNA DUPLEX CONTAINING TWO G(ANTI).A(ANTI) BASE-PAIRS
7	keywds	Key words	text	A-RNA;DOUBLE HELIX;MISMATCHED
8	mdltyp	Model details	text	CA ATOMS ONLY, CHAIN A, B ; P ATOMS ONLY, CHAIN C
9	author	Authors of experiment	text	G.A.LEONARD,K.E.MCAULEY-HECHT,S.EBEL,D.M.LOUGH,T.BROWN,W.N.HUNTER
10	expdata	Experiment data	text	X-RAY DIFFRACTION
11	resol	Resolution	float(8,3)	1.800

Таблица А.3. Модели.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	1
2	number	Number of model in PDB document	int	1
3	file	File name	char(4)	124D
4	main	1 - main model; 0 - otherwise	int	1
5	rnachains	Number of RNA chains	int	1
6	maxrnalen	Max RNA chain length	int	8
7	rnalen	Total RNA length	int	8
8	dnachains	Number of DNA chains	int	1
9	maxdnalen	Max DNA chain length	int	8
10	dnalen	Total DNA length	int	8
11	protchains	Number of Protein chains	int	0
12	maxprotlen	Max Protein chain length	int	0
13	protlen	Total Protein length	int	0
14	ligchains	Number of Ligand chains	int	0
15	maxliglen	Max Ligand chain length	int	0
16	liglen	Total Ligand length	int	0
17	atoms	Number of atoms	int	328
18	nucls	Number of nucleotides	int	16
19	aminos	Number of aminoacids	int	0
20	ligands	Number of ligands	int	0
21	liglist	List of Ligands	text	,PO4,
22	metlist	List of Metals	text	,MG,NA,K,

Таблица А.4. Молекулы биополимеров.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	10373
2	file	File name	char(4)	2IX1
3	mol_id	Molecule id in PDB document	int	1
4	mol	Molecule name	text	EXORIBONUCLEASE 2

5	orgsc	Organism name	text	ESCHERICHIA COLI
6	frag	Molecule fragment name	text	RESIDUES 6-644
7	syn	Molecule name synonyms	text	RNASE II D209N MUTANT, EXORIBONUCLEASE II, RIBONUCLEASE II, RNASE II
8	ec	Enzyme Commission number	text	3.1.13.1
9	eng	Engineered or not	text	YES
10	mut	Mutation	text	YES
11	details	Other details	text	NULL

Таблица А.5. Цепи.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	2507
2	model	Model ID	int	1366
3	mol	Molecule ID	int	2188
4	letter	Letter identifier in PDB document	varchar(5)	I
5	len	Length	int	30
6	liglen	Number of ligands	int	0
7	type	RNA/DNA/Protein/Unknown	varchar(7)	RNA
8	seq	Sequence	text	A,G,A,U,U,C,C,G,U,U,U,U,C,A,G,U, C,G,G,G,A,A,A,A,C,U,G,A,A
9	ligseq	Sequence of ligands	text	
10	lubrackets	DSSR Bracket Diagram	text	...[.(((.[[[[[[[]]]].)]...]]]]]
11	brackets	Bracket Diagram (stems only)	text	....(((.[[[[[[[]]]].)]...]]]]]
12	slbrackets	Bracket Diagram (including wc/wb links)	text	...(((((.[[[[[[[]]]].)]...]]]]]
13	start	First residue dssr-id	varchar(20)	I.A.4.
14	end	Last residue dssr-id	varchar(20)	I.A.33.
15	bound	Number of bound nucleotides	int	23
16	share	Share of bound nucleotides	float(8,3)	0.767
17	ligbound	Number of bound ligands	int	0
18	ligshare	Share of bound ligands	float(8,3)	0.000
19	fake	0 - chain has sequence; 1 - chain has only ligands	int	0
20	diagram	Diagram ID	int	1362
21	seqbrack	seq + slbrackets	text	A.G.A.U(U.C(C(G(U.U.U[U[C[A[G[U(C)G)G)G.A)A.A.A]C]U]G]A]A]
22	sharewcb	Share of nucleotides in WC/WB bps	float(8,3)	0.543

Таблица А.6. Нуклеотиды.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	205
2	dssr	dssr-id	varchar(20)	A.A.6.
3	model	Model ID	int	15
4	chain	chain ID	int	28
5	letter	chain letter-id	varchar(5)	A
6	thread	thread ID	int	NULL

7	wing	wing ID	int	37
8	oldwing	OldWing ID	int	31
9	fstems	Number of StemsFull	int	1
10	place	0 - Sequence; 1 - Ligands	int	0
11	num	Number in Sequence or Ligands	int	6
12	zip	zipper ID	int	NULL
13	mult	NuclMult ID	int	1
14	lumult	LuMult ID	int	NULL
15	bps	Number of BasePairs	int	1
16	miss	missed - 1; else - 0	int	0
17	chnumindiag	Chain Number in Diagram	int	1
18	diagram	Diagram ID	int	57
19	res	Residue	varchar(3)	A
20	bracket	Char from dot-bracket notation (no links)	char(1)	(
21	slbracket	Char from dot-bracket notation (with links)	char(1)	[
22	cifid	Number in Sequence from mmCIF	int	6

Таблица А.7. Аминокислоты.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	50
2	dssr	dssr-id	varchar(20)	A.THR.50.
3	model	Model ID	int	26
4	chain	chain ID	int	39
5	letter	chain letter-id	varchar(5)	A
6	place	0 - Sequence; 1 - Ligands	int	0
7	num	Number in Sequence or Ligands	int	50
8	miss	missed - 1; else - 0	int	0
9	res	Amino Acid	varchar(3)	CYS
10	cifid	Number in Sequence from mmCIF	int	50

Таблица А.8. Лиганды.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	50
2	dssr	dssr-id	varchar(20)	A.GE3.30.
3	model	Model ID	int	392
4	chain	chain ID	int	571
5	letter	chain letter-id	varchar(5)	A
6	place	0 - Sequence; 1 - Ligands	int	1
7	num	Number in Sequence or Ligands	int	3
8	miss	missed - 1; else - 0	int	0
9	cifid	Number in Sequence from mmCIF	int	3

Таблица А.9. Атомы.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	808
2	model	model ID	int	2
3	chain	chain ID	int	4
4	res	dssr-id of residue	varchar(20)	B.A.21.
5	restype	Type of residue (N/A/L/M/W)	char(1)	R

6	name	Atom name (inside the residue)	varchar(5)	C5'
7	num	Atom No (inside the model)	int	428
8	elem	Chemical element	varchar(3)	C
9	occup	Occupancy	float(8,3)	1.000
10	altloc	Alternative Location	char(1)	
11	x	x-coordinate	float(8,3)	0.097
12	y	y-coordinate	float(8,3)	9.887
13	z	z-coordinate	float(8,3)	3.398
14	rescifid	CIFID of residue	int	25

Таблица А.10. Спаривания.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	50
2	model	Model ID	int	7
3	num	Number in DSSR	int	4
4	chain1	First chain ID	int	13
5	chain2	Second chain ID	int	14
6	nucl1	First nucleotide dssr-id	varchar(20)	A.CPN.4.
7	bond	bond description (example: G-C)	char(3)	c-G
8	nucl2	Second nucleotide dssr-id	varchar(20)	B.G.9.
9	type	Type of pair	char(2)	WC
10	class1	Saenger class	varchar(10)	19-XIX
11	class2	Leontis class	varchar(10)	.WW
12	class3	DSSR class	varchar(10)	.W-W
13	stem	stem ID (stems)	int	10
14	oldstem	stem ID (StemsOld)	int	7
15	fullstem	stem ID (StemsFull)	int	7
16	revstem	stem ID (RevStems)	int	NULL
17	lustem	stem ID (LuStems)	int	28
18	link	link ID	int	NULL
19	helix	helix ID (LuHelices)	int	7
20	nuclmult	NuclMult ID	int	NULL
21	info1	first nucleotide details	text	n/a(n/a) n/a lambda=n/a
22	info2	second nucleotide details	text	-160.6(anti) C3'-endo lambda=n/a
23	dist1	d(C1'-C1')	float(8,3)	NULL
24	dist2	d(N1-N9)	float(8,3)	9.190
25	dist3	d(C6-C8)	float(8,3)	10.100
26	tor	tor(N1-C1'-C1'-N9)	float(8,3)	NULL
27	hbonds	list of H-Bonds	text	N3-N1(imino)[3.01]; N3- N2(amino)[3.08]; O2(carbonyl)- N2(amino)[3.03]; N4(amino)- O6(carbonyl)[3.06]
28	hbnum	Number of hbonds	int	4
29	shear	Shear parameter of bp	float(8,3)	-0.80
30	stretch	Stretch parameter of bp	float(8,3)	0.02
31	stagger	Stagger parameter of bp	float(8,3)	-0.30
32	buckle	Buckle parameter of bp	float(8,3)	-23.61
33	propeller	Propeller parameter of bp	float(8,3)	7.20
34	opening	Opening parameter of bp	float(8,3)	1.35

35	shift	Shift parameter of dinucl step with next bp	float(8,3)	0.10
36	slide	Slide parameter of dinucl step with next bp	float(8,3)	-1.54
37	rise	Rise parameter of dinucl step with next bp	float(8,3)	3.14
38	tilt	Tilt parameter of dinucl step with next bp	float(8,3)	2.72
39	roll	Roll parameter of dinucl step with next bp	float(8,3)	2.03
40	twist	Twist parameter of dinucl step with next bp	float(8,3)	34.39

Таблица А.11. Башни.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	50
2	model	model ID	int	50
3	stems	Number of stems	int	1
4	chain1	First chain ID	int	88
5	chain2	Second chain ID	int	88
6	pseudo	3 - tower has pseudoknotted loops inside and pseudoknotted edge; 2 - tower has pseudoknotted loops inside; 1 - tower has pseudoknotted edge ; 0 - else	int	0
7	type	Type of end (0 - hairpin; 1 - junction; 2 - nonloop)	int	0

Таблица А.12. Линки.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	50
2	model	Model ID	int	56
3	chain1	First RNA chain ID	int	100
4	chain2	Second RNA chain ID	int	101
5	bp	Base pair ID	int	466
6	bptype	Type of BasePair	char(2)	WC
7	lthread	First thread ID	int	223
8	lwing	First wing ID	int	NULL
9	rthread	Second thread ID	int	226
10	rwing	Second wing ID	int	NULL
11	type	0 - no intersections; 1 - intersections with bad pairs; 2 - intersections with WC/WB links; 3 - intersections with stems	int	1
12	depth	0 - only for stem; 1 - for stem and oldstem; 2 - for all stems	int	0
13	nucl1	First nucleotide dssr-id	varchar(20)	A.C.6.
14	nucl2	Second nucleotide dssr-id	varchar(20)	B.G.22.
15	class1	Saenger type	varchar(10)	19-XIX
16	class2	Leontis type	varchar(10)	.WW
17	class3	DSSR type	varchar(10)	.W-W
18	dist	Distance between nucls (on the chain)	int	10

Таблица А.13. Стемы.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	39
2	model	Model ID	int	25
3	num	Number in model	int	1

4	chain1	First chain ID	int	38
5	chain2	Second chain ID	int	38
6	lseq	Left wing sequence	text	G,G,C,G,U,A
7	rseq	Right wing sequence	text	U,A,U,G,C,C
8	len	Length	int	6
9	leftw	Left wing ID (wings)	int	77
10	rightw	Right wing ID (wings)	int	80
11	pairseq	Sequence of pairs	text	WC,WC,WC,WB,WC,WC
12	tower	tower ID	int	25
13	loopclass	Class of Loop (H/B/I/J)	char(1)	B
14	looptype	Type of Loop (P/I/C)	char(1)	C
15	oldstem	OldStem ID	int	36
16	fullstem	FullStem ID	int	36
17	mult	StemMult ID	int	NULL
18	diagram	Diagram ID	int	25
19	numindiag	Number in diagram's scheme	int	1
20	ecf	ECF ID	int	39
21	numinecf	Number in ecf's scheme	int	1
22	jmol	Jmol code for stem	text	18-20:A,30-32:A
23	loopjmol	Jmol code for loop	text	18-20:A,30-32:A;21-24:A;;;25-29:A

Таблица А.14. Закрытые стемы.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	3
2	model	Model ID	int	3
3	num	Number in model	int	1
4	chain1	First chain ID	int	5
5	chain2	Second chain ID	int	6
6	lseq	Left wing sequence	text	G,C,U,U,C,G,G,C
7	rseq	Right wing sequence	text	G,C,U,U,C,G,G,C
8	leftw	Left wing ID (WingsOld)	int	5
9	rightw	Right wing ID (WingsOld)	int	6
10	pairseq	Sequence of pairs	text	WC,WC,WB,NA,NA,WB,WC,WC
11	fullstem	FullStem ID	int	3
12	stems	Number of stems	int	2
13	len	Length	int	8
14	mult	StemMult ID	int	NULL
15	jmol	Jmol code for stem	text	18-20:A,30-32:A

Таблица А.15. Обобщенные стемы.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	100
2	model	Model ID	int	70
3	num	Number in model	int	1
4	chain1	First chain ID	int	128
5	chain2	Second chain ID	int	129
6	lseq	Left wing sequence	text	G,G,C,G,U,C,A
7	rseq	Right wing sequence	text	A,G,U,C,G,C,C
8	len	Length	int	7



9	leftw	Left wing ID (WingsFull)	int	199
10	rightw	Right wing ID (WingsFull)	int	202
11	stems	Number of stems	int	1
12	oldstems	Number of StemsOld	int	1
13	pairseq	Sequence of pairs	text	WC,WC,WC,WC,NA,WC,NA
14	mult	StemMult ID	int	NULL
15	jmol	Jmol code for stem	text	18-20:A,30-32:A

Таблица А.16. Крылья.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	100
2	model	Model ID	int	35
3	chain	chain ID	int	58
4	stem	stem ID	int	50
5	another	Complementary wing ID	int	99
6	prev	Previous Thread ID	int	147
7	next	Next Thread ID	int	148
8	prevw	Previous wing ID	int	99
9	nextw	Next wing ID	int	NULL
10	type	L - left; R - right	char(1)	R
11	start	First nucleotide dssr-id	varchar(20)	B.G.213.
12	end	Last nucleotide dssr-id	varchar(20)	B.C.220.
13	len	Length	int	8
14	seq	Sequence	text	G,C,U,A,G,U,C,C
15	ecf	ECF ID	int	5332
16	numinecf	Number in ecf's scheme	int	-2
17	jmol	Jmol code for wing	text	213-220:A

Таблица А.17. Крылья закрытых стемов.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	37
2	model	Model ID	int	16
3	chain	chain ID	int	29
4	stem	OldStem ID	int	19
5	another	Complementary wing ID	int	36
6	prevw	Previous wing ID	int	36
7	nextw	Next wing ID	int	38
8	type	L - left; R - right	char(1)	R
9	start	First nucleotide dssr-id	varchar(20)	A.C.14.
10	end	Last nucleotide dssr-id	varchar(20)	A.C.15.
11	len	Length	int	2
12	seq	Sequence	text	C,C

Таблица А.18. Крылья обобщенных стемов.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	483
2	model	Model ID	int	137
3	chain	chain ID	int	218
4	stem	FullStem ID	int	241

5	another	Complementary wing ID	int	480
6	prevw	Previous wing ID	int	482
7	nextw	Next wing ID	int	484
8	type	L - left; R - right	char(1)	R
9	start	First nucleotide dssr-id	varchar(20)	A.G.26.
10	end	Last nucleotide dssr-id	varchar(20)	A.G.29.
11	len	Length	int	4
12	seq	Sequence	text	G,G,A,G

Таблица А.19. Нити.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	205
2	model	Model ID	int	53
3	chain	chain ID	int	94
4	prev	Previous wing ID (wings)	int	137
5	next	Next wing ID (wings)	int	138
6	start	dssr-id of nucleotide after the last nucleotide of previous wing	varchar(20)	A.U.5.
7	end	dssr-id of nucleotide before the first nucleotide of next wing	varchar(20)	A.A.7.
8	len	Length	int	3
9	links	Number of link ends	int	2
10	seq	Sequence	text	U,C,A
11	place	0 - Chain Sequence; 1 - Chain Ligands	int	0
12	full	1 - thread = chain; 0 - else	int	0
13	ext	1 - External; 0 - else	int	0
14	jmol	Jmol code for thread	text	5-7:A

Таблица А.20. Шпильки.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	698
2	stemid	stem ID (stems)	int	1196
3	model	Model ID	int	555
4	chain	chain ID	int	858
5	wings	Number of wings	int	2
6	type	C - classical; P - pseudoknotted	char(1)	P
7	form	Description of loop	text	5.w3.0.w7.0
8	len	Length	int	15
9	links	Number of link ends	int	0
10	breaks	Number of breaks in chain	int	0
11	sjmol	Jmol code for closing stem	text	18-20:A,30-32:A
12	tjmol	Jmol code for threads	text	21-24:A,25-27:A
13	sfjmol	Jmol code for stem faces	text	21-24:A,25-27:A
14	bfjmol	Jmol code for block faces	text	21-24:A,25-27:A
15	wjmol	Jmol code for wings	text	21-24:A,25-27:A

Таблица А.21. Выпячивания.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	657
2	stemid	stem ID (stems)	int	1193
3	model	Model ID	int	555

4	chain1	First chain ID	int	858
5	chain2	Second chain ID	int	858
6	wings	Number of wings	int	1
7	side	Left or Right (L/R)	char(1)	L
8	form	Description of loop	text	3.w6.0- S -0
9	type	C - classical; I - isolated; P - pseudoknotted	char(1)	P
10	len	Length	int	9
11	links	Number of link ends	int	0
12	breaks	Number of breaks in chain	int	0
13	sjmol	Jmol code for closing stem	text	18-20:A,30-32:A
14	tjmol	Jmol code for threads	text	21-24:A,25-27:A
15	sfjmol	Jmol code for stem faces	text	21-24:A,25-27:A
16	bfjmol	Jmol code for block faces	text	21-24:A,25-27:A
17	wjmol	Jmol code for wings	text	21-24:A,25-27:A

Таблица А.22. Внутренние петли.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	656
2	stemid	stem ID (stems)	int	2158
3	model	Model ID	int	966
4	chain1	First chain ID	int	1545
5	chain2	Second chain ID	int	1545
6	wings	Number of wings	int	2
7	form	Description of loop	text	5.w6.5.w2.11- S -10
8	type	C - classical; I - isolated; P - pseudoknotted	char(1)	P
9	len	Length	int	39
10	sym	1 - symmetric; 0 - asymmetric	int	0
11	links	Number of link ends	int	28
12	breaks	Number of breaks in chain	int	0
13	sjmol	Jmol code for closing stem	text	18-20:A,30-32:A
14	tjmol	Jmol code for threads	text	21-24:A,25-27:A
15	sfjmol	Jmol code for stem faces	text	21-24:A,25-27:A
16	bfjmol	Jmol code for block faces	text	21-24:A,25-27:A
17	wjmol	Jmol code for wings	text	21-24:A,25-27:A

Таблица А.23. Мульти-петли.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	656
2	stemid	stem ID (stems)	int	6046
3	model	Model ID	int	1808
4	chain1	chain ID (first thread)	int	3347
5	chain2	chain ID (last thread)	int	3347
6	threads	Number of threads	int	6
7	sides	Number of sides	int	3
8	wings	Number of wings	int	3
9	sfaces	Number of stem faces	int	2
10	bfaces	Number of block faces	int	0
11	form	Description of loop	text	0.w2.4.w7.3- S -1- S -1.w2.0
12	type	C - classical; I - isolated; P - pseudoknotted	char(1)	P

13	len	Length	int	20
14	links	Number of link ends	int	8
15	breaks	Number of breaks in chain	int	0
16	sjmol	Jmol code for closing stem	text	18-20:A,30-32:A
17	tjmol	Jmol code for threads	text	21-24:A,25-27:A
18	sfjmol	Jmol code for stem faces	text	21-24:A,25-27:A
19	bfjmol	Jmol code for block faces	text	21-24:A,25-27:A
20	wjmol	Jmol code for wings	text	21-24:A,25-27:A

Таблица А.24. Зависимости нить-петля.

Номер	Поле	Комментарий	Тип	Пример
1	model	model ID	int	21
2	loopid	Loop ID	int	21
3	looptype	Type of loop (H/B/I/J)	char(1)	H
4	thread	Thread ID	int	96
5	num	Number in loop	int	1
6	side	Number of side	int	1
7	links	Number of link ends	int	2
8	seq	Sequence	text	A,U,U,A
9	len	Length	int	4

Таблица А.25. Зависимости крыло-петля.

Номер	Поле	Комментарий	Тип	Пример
1	model	model ID	int	463
2	loopid	Loop ID	int	38
3	looptype	Type of loop (H/B/I/J)	char(1)	J
4	wing	Wing ID (wings)	int	1697
5	num	Number in loop	int	3
6	side	Number of side	int	1
7	seq	Sequence	text	G,G,C,U
8	len	Length	int	4

Таблица А.26. Зависимости торец-петля.

Номер	Поле	Комментарий	Тип	Пример
1	model	model ID	int	68
2	loopid	Loop ID	int	73
3	looptype	Type of loop (N/H/B/I/J)	char(1)	B
4	stem1	First Stem ID (stems)	int	99
5	stem2	Second Stem ID (stems)	int	99
6	type	S - Stem; B - Block	char(1)	S
7	num	Number in loop	int	1

Таблица А.27. Связанные цепи.

Номер	Поле	Комментарий	Тип	Пример
1	model	model ID	int	548
2	chain1	chain1 ID	int	843
3	chain2	chain2 ID	int	844

Таблица А.28. Спирали DSSR.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	88
2	model	model ID	int	72
3	len	Length	int	19
4	seq1	Left Sequence	text	5'- GGCCGAUGGUAGUGUGG GG-3'
5	seq2	Right Sequence	text	3'- CCGGAUGAGAGCGUACC CC-5'
6	form	Helix-form	text	AAA.....A..
7	stems	Number of stems	int	3
8	lones	Number of lone basepairs	int	0

Таблица А.29. Стемы DSSR.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	87
2	model	Model ID	int	24
3	dssr	ID in DSSR	int	1
4	chain1	Chain1 ID	int	37
5	chain2	CHain2 ID	int	37
6	len	Length	int	6
7	lseq	Left Sequence	text	5'-GGCGUA-3'
8	rseq	Right Sequence	text	3'-CCGUAU-5'
9	form	Helix-form	text	AA..A

Таблица А.30. Зависимости стем-спираль DSSR.

Номер	Поле	Комментарий	Тип	Пример
1	model	model ID	int	24
2	stem	stem ID	int	87
3	helix	helix ID (LuHelices)	int	30

Таблица А.31. Мультиплеты DSSR.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	87
2	model	model ID	int	332
3	len	Length	int	3
4	seq	Sequence	text	UAU

Таблица А.32. Контакты кроме спариваний DSSR.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	87
2	model	model ID	int	16
3	nucl1	Last nucleotide dssr-id	varchar(20)	A.U.12.
4	nucl2	First nucleotide dssr-id	varchar(20)	A.A.13.
5	hbnum	Number of hbonds	int	2
6	hbonds	list of H-Bonds	text	O2'(hydroxyl)-OP1[2.97],O2'(hydroxyl)-OP2[2.94]
7	stack	Stacking details	text	2.4(2.3)--pm(>>)

Таблица A.33. Концевые участки (вне петель) DSSR.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	120
2	model	model ID	int	261
3	chain	chain ID	int	379
4	len	Length	int	4
5	seq	Sequence	text	GCCA
6	break	Break (1 - yes; 0 - no)	int	0
7	start	First nucleotide dssr-id	varchar(20)	R.G.673.
8	end	Last nucleotide dssr-id	varchar(20)	R.A.676.

Таблица A.34. Kissing loops DSSR.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	120
2	model	model ID	int	2567
3	loop1	First LuHairpin ID	int	5106
4	loop2	Second LuHairpin ID	int	5108
5	kiss	LuStem ID	int	15844

Таблица A.35. A-minors DSSR.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	120
2	model	model ID	int	966
3	nucl	Nucleotide dssr-id	varchar(20)	0.A.2883.
4	pair	Basepair ID	int	13671
5	bpdesc	Basepair description	varchar(40)	0.C.1786.-0.G.1806.
6	hbonds1	Left H-bonds	int	1
7	hbonds2	Right H-bonds	int	2
8	type	Type	varchar(5)	I
9	class	Class	varchar(5)	A C-G

Таблица A.36. U-turns DSSR.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	120
2	model	model ID	int	589
3	chain	chain ID	int	941
4	nucl1	First nucleotide dssr-id	varchar(20)	A.G.11.
5	nucl2	Second nucleotide dssr-id	varchar(20)	A.A.14.
6	hbonds	Number of H-Bonds	int	2

Таблица A.37. Ribose zippers DSSR.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	120
2	model	model ID	int	1356
3	len	Length	int	4
4	seq	Sequence	text	GGAG

Таблица A.38. Kink turns DSSR.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	120
2	model	model ID	int	1877
3	bp	BasePair ID	int	51024
4	helix	helix ID (LuHelices)	int	4870
5	stem1	First Stem ID (LuStems)	int	8889
6	stem2	Second Stem ID (LuStems)	int	8888
7	iloop	Internal Loop ID (LuInternals)	int	2796
8	type	0 - Normal; 1 - Reverse; 2 - Undecided; 3 - [w]	int	0

Таблица A.39. Шпильки DSSR.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	120
2	model	model ID	int	119
3	chain	chain ID	int	200
4	len	Length	int	6
5	seq	Sequence	text	GGGUCA
6	stem	LuStem ID	int	280

Таблица A.40. Выпячивания DSSR.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	120
2	model	model ID	int	360
3	chain	chain ID	int	539
4	prev	Previous LuStem ID	int	812
5	next	Next LuStem ID	int	813
6	form	Description of loop	text	1,0
7	seq	Sequence	text	C
8	len	Length	int	1

Таблица A.41. Внутренние петли DSSR.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	117
2	model	model ID	int	235
3	chain1	First chain ID	int	351
4	chain2	Second chain ID	int	351
5	prev	Previous LuStem ID	int	530
6	next	Next LuStem ID	int	531
7	form	Description of loop	text	4,3
8	sym	1 - symmetric; 0 - asymmetric	int	0
9	lseq	Left Sequence	text	UGCU
10	rseq	Right Sequence	text	ACG
11	len	Length	int	7

Таблица A.42. Мульти-петли DSSR.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	117

2	model	model ID	int	1112
3	threads	Number of threads	int	3
4	form	Description of loop	text	0,7,1
5	len	Length	int	8

Таблица А.43. Нити мульти-петель DSSR.

Номер	Поле	Комментарий	Тип	Пример
1	junc	junction ID (LuJunctions)	int	45
2	model	model ID	int	953
3	chain	chain ID	int	1526
4	num	Number in Junction	int	3
5	start	First nucleotide dssr-id	varchar(20)	D.A.44.
6	end	Last nucleotide dssr-id	varchar(20)	D.C.48.
7	len	Length	int	5
8	seq	Sequence	text	AGgUC
9	prev	Previous LuStem ID	int	2412

Таблица А.44. Обратные (параллельные) стемы.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	112
2	model	model ID	int	969
3	chain1	chain1 ID	int	1578
4	chain2	chain2 ID	int	1578
5	len	Length	int	3
6	lseq	Left Sequence	text	U,U,U
7	rseq	Right Sequence	text	A,A,G
8	pairseq	Sequence of BasePairs	text	NA,HG,NA

Таблица А.45. Мультиплеты.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	110
2	model	model ID	int	221
3	nts	number of Nucleotides	int	3
4	bps	Number of BasePairs	int	2
5	wcwb	Number of WC/WB BasePairs	int	1
6	type	Type of multiplet	text	1(2(1)),1(2(1)),2(1.1)

Таблица А.46. Стемовые мотивы.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	110
2	model	model ID	int	765
3	fstems	Number of StemsFull	int	3
4	nts	Number of shared nucleotides	int	3
5	links	Number of links	int	2
6	nedges	Number of nucl-edges	int	2
7	ledges	Number of link-edges	int	1
8	edges	Total Number of edges	int	2
9	npairs	Total Number of related StemNPairs	int	3
10	lpairs	Total Number of related StemLPairs	int	2



11	type	Type of multiplet	text	1(2(1)),1(2(1)),2(1.1)
----	------	-------------------	------	------------------------

Таблица А.47. Общие нуклеотиды (в стемовых мотивах).

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	110
2	model	model ID	int	721
3	fstem1	First FullStem ID	int	1482
4	fstem2	Second FullStem ID	int	1483
5	nucl	Nucleotide dssr-id	varchar(20)	A.A.5.
6	mult	StemMult ID	int	79

Таблица А.48. Линки (в стемовых мотивах).

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	110
2	model	model ID	int	771
3	link	Link ID	int	2767
4	fstem1	First FullStem ID	int	1641
5	ostem1	First OldStem ID	int	1567
6	stem1	First Stem ID	int	1650
7	fstem2	Second FullStem ID	int	1640
8	ostem2	Second OldStem ID	int	NULL
9	stem2	Second Stem ID	int	NULL
10	mult	StemMult ID	int	116
11	nucl1	First Nucleotide dssr-id	varchar(20)	A.G.25.
12	nucl2	Second Nucleotide dssr-id	varchar(20)	A.A.27.

Таблица А.49. Вторичные структуры.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	51
2	model	model ID	int	51
3	seq	Sequence of chain letters	varchar(20)	BC
4	brackets	Bracket diagram (stems only)	text	..((((((...))))))...
5	slbrackets	Bracket diagram (including wc/wb links)	text	(.((((((...)))))).)
6	sldibrank	Rank1 (number of different brackets for slbrackets)	int	1
7	sllptrank	Rank2 (max number of loops per thread for slbrackets)	int	1
8	scheme	Scheme of stems	text	1,-1
9	stembrack	Bracket diagram of scheme	text	()
10	dibrank	Rank1 (number of different brackets)	int	1
11	lptrank	Rank2 (max number of loops per thread)	int	1
12	depth	Max depth of ecf in the diagram	int	0
13	nuclseq	Total Sequence of chains	text	A,G,A,U,U,C,C,G,U,U, U,U,C,A,G,U,C,G,G,G, A,A,A,A,A,C,U,G,A,A
14	seqbrack	Sequence + dot-bracket (with links)	text	A(G.C(U.G)U.U)

Таблица А.50. Элементарные закрытые участки.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	157
2	model	model ID	int	102
3	diagram	Diagram ID	int	102
4	fullscheme	Scheme of ecf (including children, as in diagram)	text	2,3,-2,-3
5	scheme	Scheme of ecf	text	1,2,-1,-2
6	signat	Signature of ecf	text	abAB
7	brackets	Bracket diagram of signature	text	([])
8	dibrank	Rank1 (number of different brackets)	int	2
9	lptrank	Rank2 (max number of loops per thread)	int	2
10	depth	Depth of ecf (Number of parents)	int	0
11	parent	Parent ecf ID	int	NULL
12	wingseq	Sequence of wings	text	1:3:C,C,C.2:5:U,C,C,G, A.-1:3:G,G,G.- 2:5:U,C,G,G,A
13	jmol	Jmol code	text	(18-20:A,30-32:A;21- 24:A;;;25-29:A.[25- 29:A,36-40:A;33- 35:A;;;30-32:A

Таблица А.51. Водородные связи между РНК и белком.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	1
2	model	model ID	int	1
3	rchain	RNA chain ID	int	1
4	pchain	Protein chain ID	int	3
5	nucl	Nucleotide name	varchar(3)	G
6	amino	Amino acid name	varchar(3)	LYS
7	rdssr	Nucleotide dssr code	varchar(20)	Q.G.4.
8	pdssr	Amino acid dssr code	varchar(20)	B.LYS.20
9	ratomid	RNA atom ID	int	102
10	patomid	Protein atom ID	int	489
11	rname	RNA atom name	varchar(5)	O6
12	pname	Protein atom name	varchar(5)	NZ
13	relem	RNA atom element	char(1)	O
14	pelem	Protein atom element	char(1)	N
15	rjmol	RNA atom jmol code	varchar(20)	4:Q.O6
16	pjmol	Protein atom jmol code	varchar(20)	20:B.NZ
17	power	Power parameter value	float(8,3)	0.23
18	dist	Distance (angstrom)	float(8,3)	φев.58
19	rcos	Cosine of theta angle of RNA atom	float(8,3)	0.69
20	pcos	Cosine of theta angle of Protein atom	float(8,3)	0.79
21	monopair	Monomer pair ID	int	1

Таблица А.52. РНК-белковые взаимодействия.

Номер	Поле	Комментарий	Тип	Пример
1	id	Unique ID number	int	1
2	model	model ID	int	1
3	rchain	RNA chain ID	int	1
4	pchain	Protein chain ID	int	3
5	nucl	Nucleotide name	varchar(3)	G
6	amino	Amino acid name	varchar(3)	LYS
7	rdssr	Nucleotide dssr code	varchar(20)	Q.G.4.
8	pdssr	Amino acid dssr code	varchar(20)	B.LYS.20
15	rjmol	Amino acid jmol code	varchar(15)	4:Q
16	pjmol	Nucleotide jmol code	varchar(15)	20:B
21	atompairs	Number of atomic pairs	int	2

## ПРИЛОЖЕНИЕ С. Список рассмотренных баз данных

1. **CoSSMos**. <http://cossmos.slu.edu/>.

*Инструмент анализа pdb-файла* – MC-Annotate.

*Класс* – база, посвященная отдельным структурным элементам (петлям).

*Поиск* – по петлям.

*Охват* – РНК.

*Достоинства*: высокий качественный отбор петель для хранения в базе.

*Недостатки*: ограниченный набор данных (только петли ограниченного размера, без мульти-петель).

2. **DARTS**. <http://bioinfo3d.cs.tau.ac.il/DARTS/>.

*Инструмент анализа pdb-файла* – ARTS.

*Класс* – универсальная база структур РНК.

*Поиск* – по структурам в целом.

*Охват* – РНК.

*Достоинства*: имеются данные сравнения структур на трех уровнях (1D, 2D, 3D).

*Недостатки*: рассматриваются только структуры в целом.

3. **MINAS**. <http://www.minas.uzh.ch/home/>.

*Инструмент анализа pdb-файла* – собственный.

*Класс* – база, посвященная отдельным структурным элементам (ионам металла).

*Поиск* – по структурам в целом.

*Охват* – РНК и РНК-белковые комплексы.

*Достоинства*: простая реализация громоздких запросов.

*Недостатки*: нет дополнительных данных, только возможность поиска по pdb.

4. **NPIDB**. <http://npidb.belozersky.msu.ru/>.

*Инструмент анализа pdb-файла* – PDBeFold.

*Класс* – база, посвященная отдельным структурным элементам (РНК-белковым контактам).

*Поиск* – по структурам в целом, по доменам РНК-белковых комплексов.

*Охват* – РНК-белковые комплексы.

*Достоинства:* хороший веб-интерфейс, интеграция со многими существующими базами.

*Недостатки:* отсутствие вторичной структуры РНК.

5. **PRIDB**. <http://pridb.gdcb.iastate.edu/>.

*Инструмент анализа pdb-файла* – FR3D.

*Класс* – база, посвященная отдельным структурным элементам (РНК-белковым контактам).

*Поиск* – по структурам в целом.

*Охват* – РНК-белковые комплексы.

*Достоинства:* возможность совместного поиска заданных подструктур РНК и белка.

*Недостатки:* отсутствие вторичной структуры РНК.

6. **PseudoBase++**. <http://pseudobaseplusplus.utep.edu/>.

*Инструмент анализа pdb-файла* – собственный.

*Класс* – база, посвященная отдельным структурным элементам (псевдоузлам).

*Поиск* – по псевдоузлам.

*Охват* – РНК.

*Достоинства:* возможность добавлять собственные псевдоузлы.

*Недостатки:* эмпирический характер описания имеющихся элементов.

7. **RLooM**. <http://rloom.mpimp-golm.mpg.de/cgi-bin/index.py>.

*Инструмент анализа pdb-файла* – MC-Annotate.

*Класс* – вспомогательная база для структурного моделирования.

*Поиск* – по петлям.

*Охват* – РНК.

*Достоинства:* возможность поиска по шаблонам структур на языке MC-Search.

*Недостатки:* присутствуют только петли классических вторичных структур.

8. **RNA3DHub**. <http://rna.bgsu.edu/rna3dhub>.

*Инструмент анализа pdb-файла* – FR3D.

*Класс* – универсальная база структур РНК.

*Поиск* – по структурам в целом, по петлям.

*Охват* – РНК.

*Достоинства:* использование геометрических параметров для кластеризации элементов.

*Недостатки:* нет единого языка описания структуры, приводится только список содержащихся спариваний и петель; долгое ожидание загрузки объемных страниц.

9. **RnaBricks**. <http://iimcb.genesilico.pl/rnabricks/>.

*Инструмент анализа pdb-файла* – MC-Annotate.

*Класс* – вспомогательная база для структурного моделирования.

*Поиск* – по петлям, спиральям и однонитевым сегментам.

*Охват* – РНК и РНК-белковые комплексы.

*Достоинства:* учитывается любое окружение элементов (лиганды, металлы, белки, взаимодействия между ячейками кристалла).

*Недостатки:* нет описания структур в целом.

10. **RnaFRABASE**. <http://rnafrabase.ibch.poznan.pl/index.php>.

*Инструмент анализа pdb-файла* – RNAView.

*Класс* – универсальная база структур РНК.

*Поиск* – по структурам в целом, по нуклеотидам и их спариваниям, по динуклеотидным шагам и спиральям, по петлям.

*Охват* – РНК.

*Достоинства:* гибкий поиск по всему что есть.

*Недостатки:* исключены псевдоузлы.

11. **RNAJunction**. <http://rnajunction.abcc.ncifcrf.gov/>.

*Инструмент анализа pdb-файла* – RNAView.

*Класс* – база, посвященная отдельным структурным элементам (петлям).

*Поиск* – по петлям.

*Охват* – РНК.

*Достоинства:* используются геометрические параметры спиралей для кластеризации мульти-петель.

*Недостатки:* нет ничего кроме петель и kissing loops.

12. **RNAStrand**. <http://www.rnasoft.ca/sstrand/>.

*Инструмент анализа pdb-файла* – RNAView.

*Класс* – универсальная база структур РНК.

*Поиск* – по структурам в целом.

*Охват* – РНК.

*Достоинства:* большое количество имеющихся атрибутов структуры, большой охват исходных данных, описание псевдоузлов.

*Недостатки:* нет поиска по отдельным элементам, нет белков, неполноценный язык описания псевдоузлов.

13. **SCOR**. <http://scor.lbl.gov>.

*Инструмент анализа pdb-файла* – собственный.

*Класс* – универсальная база структур РНК.

*Поиск* – неизвестно, так как база не поддерживается в настоящее время.

*Охват* – РНК.

*Достоинства:* имеется классификация третичных мотивов по их функциям.

*Недостатки:* не поддерживается в настоящее время.