

Рис. 2. Структуры белка

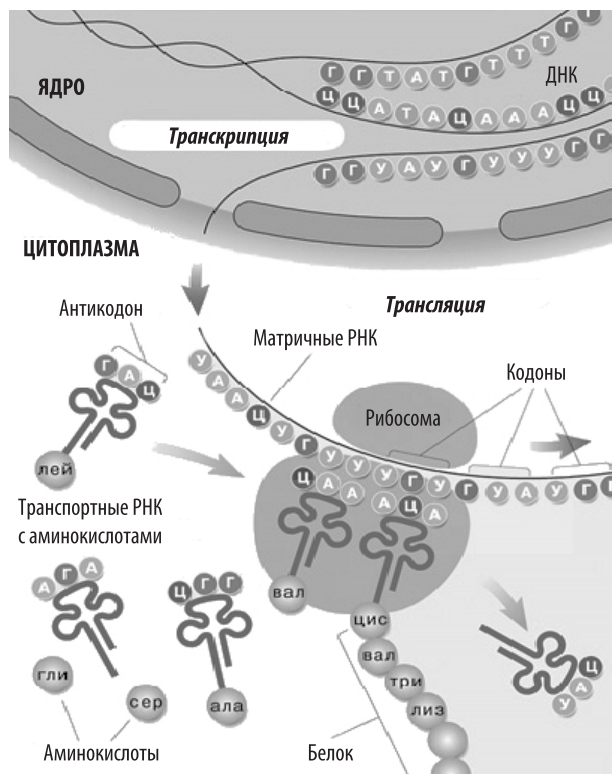


Рис. 3. Синтез белка

Е. Л. Калишенко, К. В. Крикин

донам информационной РНК. Соединенные аминокислоты взаимодействуют между собой, образуя полипептидную цепь, специфичную для данного белка, т. е. его первичную структуру. В дальнейшем она подвергается спирализации и определенной «упаковке» в пространстве, в результате чего формируются вторичная и третичная структуры данного белка.

Задача моделирования топологической структуры белка

Понимание механизмов функционирования живых систем, а значит, и возможность влиять на них, например, с помощью лекарственных средств, требует знания структуры белковых молекул и глубокого понимания их функций.

Знание пространственной организации белковых молекул является ключом не только к пониманию их функций и механизма работы, но и основой для разработки эффективных и безопасных лекарственных средств. В то же время определить структуру белков в прямом эксперименте не всегда возможно или целесообразно — из-за сложности, дороговизны и ограниченности возможностей экспериментальных методик. Однако иногда удается преодолеть эти сложности, подойдя к проблеме «с другого конца»: структуру биомакромолекул можно «предсказать», используя теоретические подходы — основанные на физических или эмпирических приближениях.

С термодинамической точки зрения самосворачивание белка является переходом белковой молекулы в наиболее статистически вероятную конформацию (что практически можно приравнять к конформации с наименьшей потенциальной энергией). Ограниченность понимания механизмов фолдинга (сворачивания белков) связана с тем, что его сложно наблюдать экспериментально: это достаточно быстрый динамический процесс, «разглядывать» который нужно на уровне отдельных молекул [4].

Целью разрабатываемой системы является построение трехмерной структуры белка, максимально приближенной к той, которая воссоздается в живой клетке на основе синтезированной первичной структуры белка.

В настоящее время существуют две основных системы предсказания структуры белков.

1. Rosetta. Чтобы предсказать форму, которую специфический белок принимает в природе, выполняется поиск сворачивания с самой низкой энергией. Проект использует кластер из узлов сети Интернет. Каждый может установить на свою домашнюю машину легкий интернет-клиент и в фоновом режиме проводить часть вычислений. Кратко алгоритм работы Rosetta выглядит так:

- начать с полностью развернутой цепочки аминокислот;
- переместить часть цепочки, чтобы создать новую форму;
- вычислить энергию новой формы;
- принять или отклонить движение в зависимости от изменения энергии;
- повторять со 2 по 4 шаг, пока каждая часть цепочки не будет перемещена много раз [5].

2. Tasser. Короткие структурные фрагменты «собираются» в специализированном силовом поле, а результат (модель, предположительно близкая к нативной) выбирается из ансамбля предсказаний с помощью идентификации наиболее плотного структурного кластера, являющегося, по мнению исследователей, «гнездом» физически реалистичных моделей [4].

В рамках разрабатываемой системы был выбран менее «физичный» метод предсказания структуры молекул — использование абстракции МВМ (Молекулярной Векторной Машины), суть которого описана ниже.

Молекулярная векторная машина

МВМ — абстракция процесса последовательного синтеза белка из 20 различных аминокислот, представляющая формирование пространственной структуры белка через действие физических операторов. Физический оператор (ФО) — абстракция аминокислоты, представляющая ее боковую цепь как инструмент воздействия на сформированную ранее структуру белка через усиление/ослабление возникающих в процессе синтеза водородных связей. Область действия физических операторов — водородная связь между крайними

звеньями четырехзвенного цикла в цепи полимера (рис. 4). Длина структурного ребра соответствует константе k_s и для цепного полимера является постоянной величиной (ее можно принять за 1), а ребра связности — константе k_c , которая может варьироваться в пределах $0 - k_c$. Может существовать несколько констант k_c . В зависимости от свойств боковых цепей аминокислот выделяются:

- операторы связности — обеспечивают дополнительную фиксацию четырехзвенного фрагмента полимера; боковая цепь имеет на конце группы, способные к образованию укрепляющих водородных связей;
- операторы антисвязности — препятствуют формированию замкнутого четырехзвенного цикла, не допускают образования водородной связи.

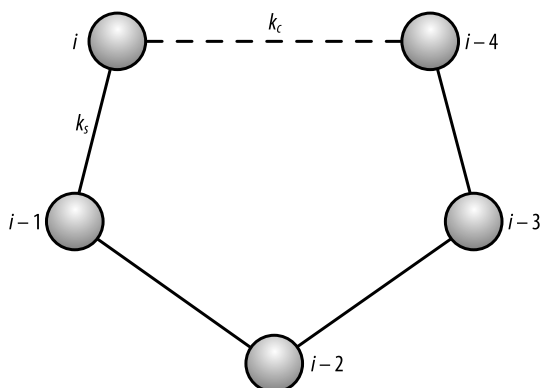


Рис. 4. Четырехзвенный цикл аминокислот

Геометрической интерпретацией МВМ является додекаэдр (рис. 5). Рассмотренный на рис. 4 четырехзвенный фрагмент цепного полимера на рис. 5 преобразован так, чтобы в область связи между $(i-1)$ и $(i-4)$ элементами стало удобно помещать МВМ. С помощью додекаэдра, имеющего 20 вершин, заданы 20 направлений векторов. Размеры додекаэдра определились исходя из параметров четырехзвенного фрагмента полимера. Для того чтобы задать вектор, необходимо знать положение двух точек — начальной, из которой исходит вектор, и конечной, куда он направлен. В нашем случае за начальную точку всех векторов можно принять центр додекаэдра, а конечными точками будут вершины додекаэдра [1].

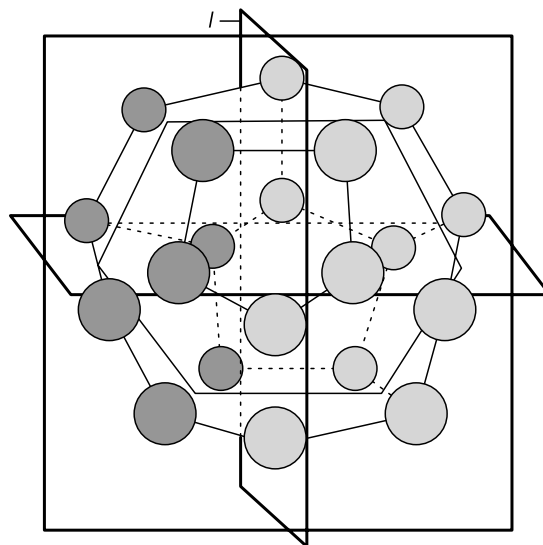


Рис. 5. Схема МВМ

Таким образом, действие физического оператора определяется действием МВМ на связь посредством вектора, исходящего из центра додекаэдра. Рассмотрим алгоритм работы МВМ при построении такой структуры как «альфа-спираль» подробно (рис. 6).

1. Считывание триплета — три нуклеотида (триплет) ДНК однозначно кодируют аминокислоту. Одну аминокислоту могут кодировать несколько триплетов, однако обратное неверно.
2. Определение аминокислоты — простое табличное сопоставление «триплет — аминокислота». Если на входе аминокислота Пролин, цепь прерывается, так как Пролин является оператором антисвязности и не может сформировать четырехзвенный цикл с участием водородной связи.
3. Продолжение спирали — в случае прихода Глицина, спираль продолжается без изменений, так как Глицин в модели МВМ — нейтральный элемент.
4. Поворот структуры — определяется действием ФО, т. е. образование водородной связи или ее разрыв неизбежно приводит к изменению положения атомов последнего цикла.
5. Добавление аминокислоты — «полноценная» установка пришедшей аминокислоты на свое место.

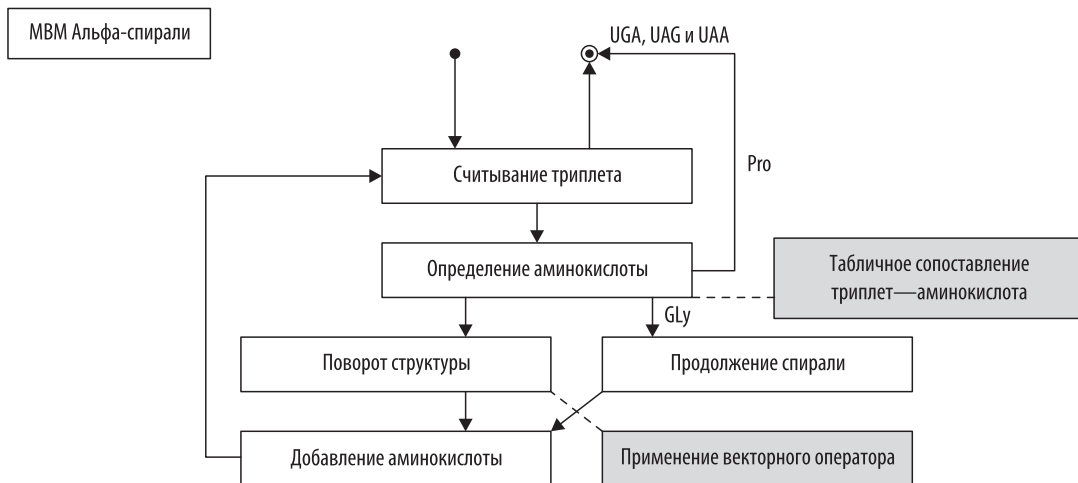


Рис. 6. Алгоритм МВМ

МВМ как компилятор ДНК

Выше была рассмотрена работа МВМ в рамках одного вида структуры белка — альфа-спирали. Теперь пришло время распространить модель МВМ на всю входную последовательность аминокислот с входной цепочкой компилятора, МВМ — с обработчиком токенов, а белка — с выходной цепочкой позволяет представить процесс моделирования структуры белка как исполнение машины состояний, схема которой представлена на рис. 7.

Свободное состояние этой машины служит для накопления достаточного количества ами-

нокислот, чтобы четко определить, в рамках какого вида структуры белка должна работать МВМ. На начальных этапах работы системы обработка свободного состояния может заключаться в использовании заранее известной разметки структур белка. А для более точного моделирования в рамках конкретных видов структур свободное состояние позволяет пропускать виды структур, с которыми система работать еще не научилась.

Коэффициенты модели МВМ для разных видов структур будут отличаться, поэтому машина состояний предусматривает работу МВМ в рамках какой-либо одной известной структуры.

Система топологического моделирования структуры белковых молекул

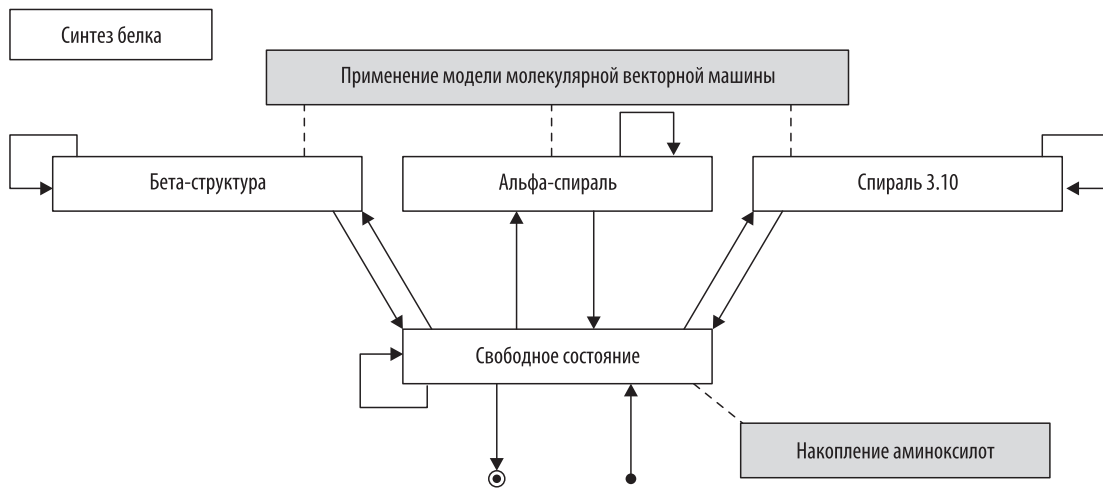


Рис. 7. Компилятор ДНК

По состоянию на 25 марта 2008 года число структур в Бруксхэйвенском банке белковых структур (Protein Data Bank, PDB) составляло около 10 000, что соответствует приблизительно 1–2% от общего числа практически важных белков. Все эти экспериментальные данные находятся в свободном доступе и являются основными входными данными разрабатываемой системы. Кроме того, существует дополнительная разметка белков по структурам, которая может понадобиться на начальных этапах работы системы для упрощения ее обучения. Таким образом, на вход системы поступают:

- набор экспериментально определенных пространственных структур макромолекул или комплексов молекул; одна запись в банке данных (один файл) соответствует одному эксперименту; в файле содержатся координаты атомов в некоторой произвольной системе координат, аннотации, первичная и вторичная структура белков и т. д.
- разметка структуры белка — метаинформация, позволяющая определить, какая структура получается при добавлении к цепи очередной аминокислоты (альфа-спираль, бета-структура и т. д.).

Архитектура системы

Архитектура системы представлена на рис. 8.

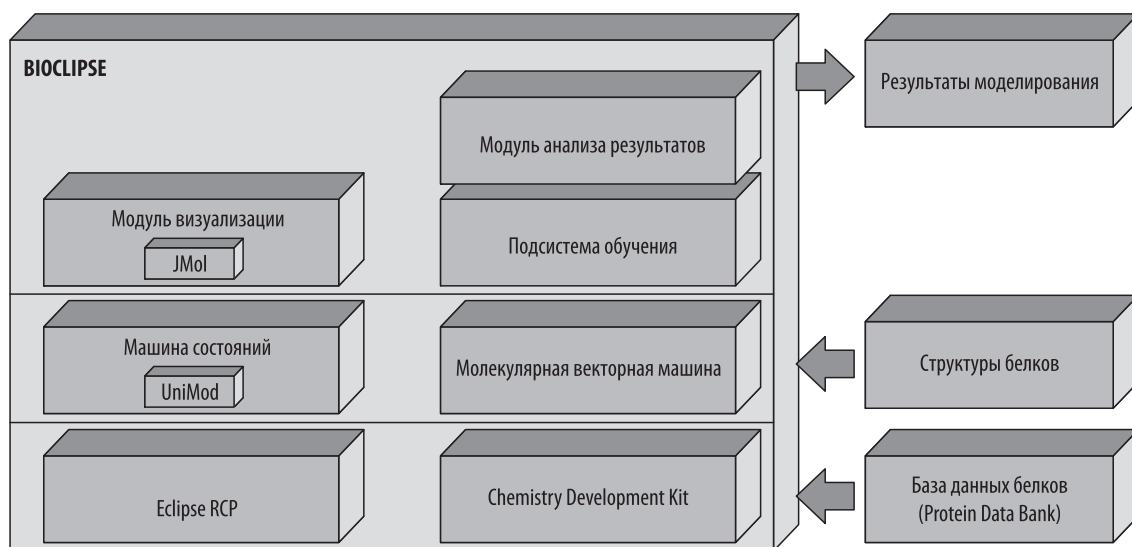


Рис. 8. Архитектура системы

Рассмотрим используемые сторонние разработки.

Bioclipse — основанная на Eclipse RCP (от англ. Rich Client Platform — среда разработки приложений) платформа для исследований в области биоинформатики, визуализации химической и биологической информации. Имеет следующие возможности:

- 2D-редактирование и 3D-визуализация молекул;
- расчет различных химических коэффициентов;
- преобразование форматов данных;
- работа со спектрами (NMR, MS);
- встроенный скриптовый язык, основанный на Mozilla Rhino;
- встраивание плагинов с использованием стандартного механизма расширений и точек расширений Eclipse RCP.

Проект призван обеспечивать исследователей единой программной средой с открытым исходным кодом и, как следствие, возможностью модифицировать/дополнять заложенные в систему алгоритмы.

Jmol — 3D-визуализатор молекулярных структур (рис. 9). Поддерживает множество форматов (более 20), возможно использование скриптов, использование в качестве Java-апплета. Среди базовых функций можно выделить:

- вращение молекулы;

- изменение масштаба изображения;
- различные способы изображения и раскраски молекул.

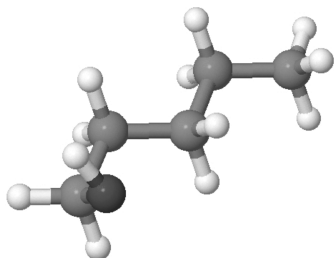


Рис. 9. Молекула

Chemistry development kit — библиотека Java-классов для разработки химического ПО. Содержит основные абстракции предметной области (молекулы, валентности, заряд и т. д.) и алгоритмы работы с этими объектами. CDK содержит методы загрузки данных из принятых международных форматов, таких как PDB.

UniMod — программный пакет для разработки объектно-ориентированных приложений на основе автоматного подхода. Содержит набор инструментов, позволяющих визуальное проектировать и реализовывать программы. При этом первоначально строится схема связей, состоящая из источников событий, системы управления и объектов управления, в которых реализованы вызываемые из автоматов выходные воздействия и опрашиваемые автоматами входные переменные. Частью UniMod является Java Finite State Machine Framework — система построения и исполнения конечных автоматов. В разработке для создания машины состояний используется плагин к Eclipse, позволяющий визуализировать процесс создания и отладки машин состояний.

Рассмотрим назначение и принципы работы каждого из модулей системы.

Машина состояний

Применение формализма машины состояний к моделированию пространственной структуры белков позволяет сопоставить входной цепочке последовательность аминокислот, а операциям в различных состояниях — действие физических операторов на оконча-

ние уже построенной белковой структуры. С помощью программного средства UniMod была построена машина состояний с правилами переходов между состояниями и операциями внутри состояний.

Описание правил перехода и вызова соответствующих методов можно увидеть на изображении графа состояний. Java Finite State Machine Framework обеспечивает вызов одного из методов контролируемого объекта (рис. 10).

Таким образом, машина состояний контролирует процесс моделирования структуры белка применением нужной модели к различным белковым структурам. Если рассмотреть систему моделирования с позиции шаблона проектирования MVC (Model-View-Controller), то машину состояний можно считать контроллером (Controller), хранилища данных и внутренние структуры — моделью (Model), визуализацию процесса моделирования — отображением (View).

Молекулярная векторная машина

Модуль обеспечивает:

- возможность визуализации модели МВМ;
- анализ корректности перемещения оператором окончания цепи, в зависимости от пришедшей аминокислоты и, как следствие, наложенных ограничений на создание/разрушение водородных связей;
- хранение коэффициентов текущей модели для каждой аминокислоты (для определения воздействия достаточно хранить два угла поворота и изменение расстояния до центра додекаэдра);
- сохранение последовательности операций пользователя для возможности их воспроизведения;
- реализацию алгоритмов корректировки коэффициентов модели.

Модуль визуализации

Процесс моделирования требует наличия возможности видеть как моделируемую, так и экспериментальную структуры белка. Это достигается благодаря параллельному размещению двух областей визуализации. К обеим

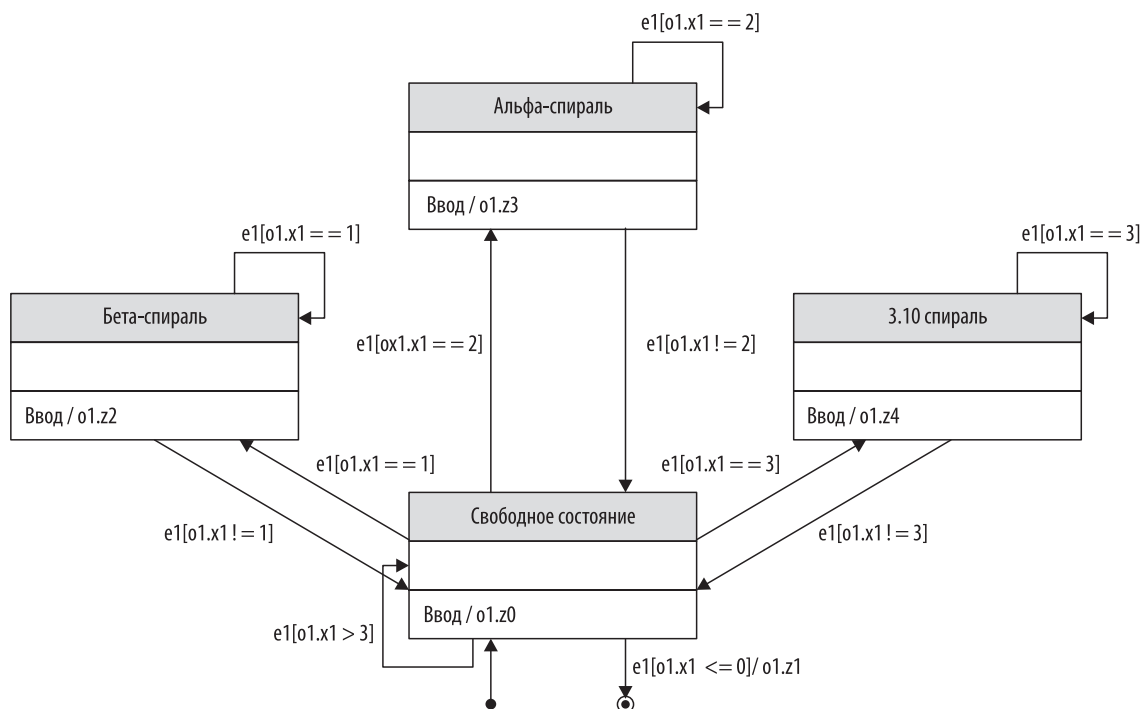


Рис. 10. Машина состояний:

x1 — предсказать структуру белка по принятой аминокислоте; z0 — подготовить структуры данных при первом запуске машины; z1 — завершить работу; z2 — применить МВМ к бета-спирали; z3 — применить МВМ к альфа-спирали; z4 — применить МВМ к 3.10-спирали

областям предъявляются следующие требования:

- выбор между отображением текущей структуры белка и всем построенным на этот момент полимером;
- возможность выделения и идентификации аминокислот;
- измерение расстояний между атомами;
- масштабирование модели;
- синхронное вращение областей визуализации;
- выбор отображения только альфа-углеродных атомов.

Визуализация МВМ накладывает дополнительные требования:

- отображение модели додекаэдра;
- возможность физического изменения оператором положения последних пяти аминокислот;
- визуализация корректности применения МВМ (например, в случае увода добавляемой аминокислоты в недопустимую область).

Модуль анализа результатов

Система, рассчитанная на моделирование чего бы то ни было, должна предоставлять средства оценки корректности своих результатов. Входными данными модуля являются две сложных пространственных структуры белка — экспериментальная и смоделированная — представленные в виде координат атомов в произвольной трехмерной системе координат. Сложность сравнения таких структур более обусловлена необходимостью их качественного, а не количественного анализа: простого учета соответствующего положения атомов в моделях здесь недостаточно.

Представим себе, что смоделированная структура повернута относительно экспериментальной на неизвестный угол. Уже в этом случае простое сравнение координат атомов перестает работать. Еще хуже дело обстоит в ситуации, которая неизбежно возникает в процессе моделирования. Относительное

положение атомов полученной структуры может отличаться от экспериментального, хотя качественно структуры могут быть полностью идентичны.

Таким образом, процедура анализа результатов должна быть направлена на глубокий анализ структурной похожести белковых молекул. Сравнение смоделированной и экспериментальной белковых структур производится с применением трех подходов.

1. Визуальное сравнение. Исследователь имеет возможность наглядно сравнить результат моделирования и экспериментальные данные, проводить операции вращения и масштабирования для визуальной оценки корректности параметров примененной модели.

2. Оценка СКО. СКО в молекулярном моделировании используется в качестве меры пространственной близости двух моделей: низкое СКО обозначает близость двух структур. Позволяет количественно ценить похожесть относительного расположения атомов в имеющихся структурах.

3. Сравнение структур как узлов. Проекция трехмерной структуры белка на плоскость представляет собой сложный узел (рис. 11). Теория узлов позволяет представить запутанные структуры в виде полинома, причем качественно одинаковые структуры описываются одинаковыми полиномами. При таком подходе происходит абстрагирование от меры изогнутости в пользу анализа структурной похожести объектов.

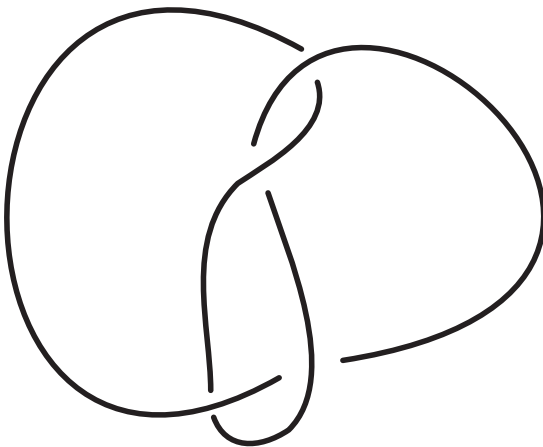


Рис. 11. Узел

Познакомимся немного ближе с теорией узлов. Существует ли алгоритм, с помощью которого по любой паре диаграмм можно узнать, эквивалентны они или нет? Теорема звучит следующим образом: «Два узла эквивалентны тогда и только тогда, когда от диаграммы одного узла к диаграмме другого можно перейти с помощью четырех операций». Теорема Рейдемейстера сводит трудную пространственную задачу определения эквивалентности двух узлов к более простой «плоской» задаче о превращении одной диаграммы узла в другую с помощью трех известных операций [5].

Английский математик Дж. Конвей постулировал, что каждой диаграмме узла или зацепления поставлен в соответствие полином (многочлен) от переменной x с целыми коэффициентами [3].

Таким образом, используя полином Конвея, возможно представить пространственную структуру белка (ее проекцию) в виде полинома. При этом незначительное смещение координат атомов смоделированной структуры от экспериментальной не приведет к ошибочному результату. Как сказано выше, одинаковые по структуре узлы имеют одинаковые полиномы Конвея. Более того, при таком подходе отпадает необходимость приведения структур к одному углу поворота и наклона — построение полинома Конвея позволит сравнить структуру узлов на качественном уровне.

Модуль обучения

Наличие тысяч белков, на которых может проводиться обучение системы и постепенное приближение коэффициентов модели МВМ к реальности, требует введения в систему механизма самообучения. На ранних этапах работы ручное обучение системы необходимо для целей как отладки модели, так и определения ее начальных коэффициентов. Однако приближение модели к реальности не может обойтись без тестирования системы на большом наборе входных данных.

Принцип коррекции коэффициентов модели после каждого предъявления нового белка хорошо отражается в алгоритмах, основанных

на обучении нейронных сетей. Для улучшения обучаемости системы можно применить:

- генетические алгоритмы — как инструмент обучения нейронных сетей;
- сеть Хопфилда — так как в процессе работы динамика таких сетей сходится к одному из положений равновесия, что можно считать минимальной энергией белковой структуры. Кроме того, поскольку сети такого типа можно интерпретировать как ассоциативную память, логично ожидать улучшения работы нейронной сети при обработке участков структур, встречавшихся ранее.

Процесс моделирования

Система предусматривает два варианта моделирования.

1. Ручное: оператор при добавлении каждой аминокислоты в интерактивном режиме корректирует положение атомов окончания белковой цепи, тем самым корректируя коэффициенты модели MBM, что при следующем добавлении такой же аминокислоты позволит расположить ее более корректно.

2. Автоматическое: после нескольких циклов ручного моделирования возможно проводить автоматическое моделирование белковых структур с применением самообучения системы. Начальные коэффициенты модели MBM, полученные в процессе ручного моделирования, могут быть существенно изменены предоставлением большого числа известных структур белков для обучения системы.

Потоки данных

Система моделирования получает данные из двух типов источников: внешних и внутренних. Внешними по отношению к системе являются следующие источники.

Аминокислоты белка — хранилище экспериментально полученных структур белков в формате PDB. Основные данные, поставляемые этим источником и используемые в системе — последовательность аминокислот (используется при последовательном моделировании пространственной структуры) и ко-

ординаты атомов (используются для оценки результатов моделирования и корректировки параметров модели).

Структура белка — хранилище метаинформации о белковых структурах. Отсюда система извлекает данные о разметке белка — указание вида белковой структуры (вида спирали), в рамках которого сейчас происходит моделирование. В зависимости от типа структуры параметры модели изменяются.

Внутренний источник данных системы — **Параметры модели** — предназначен для:

- получения начальных параметров модели, основанных на теоретических расчетах и скорректированных за время предыдущих сеансов моделирования;
- сохранения результатов корректировки модели, получаемых после моделирования структуры очередного белка;
- сохранения последовательности преобразований, выполненных системой при моделировании структуры конкретного белка, с целью иметь возможность воспроизвести процесс моделирования в автоматическом режиме.

В системе можно выделить два основных потока данных — поток экспериментальных данных и поток параметров модели (рис. 12).

Поток экспериментальных данных обеспечивает:

- сбор статистики — определение числа аминокислот каждого типа и их общего количества;
- отображение экспериментальной структуры — визуализация экспериментальной структуры белка с использованием трехмерных координат атомов;
- определение типа структуры — выделение текущей аминокислоты и выделение из хранилища структур белка вида моделируемой спирали;
- применение модели MBM — экспериментальные координаты атомов используются для начального расположения молекулы аминокислоты;
- сравнение и отображение результатов — результат работы системы моделирования сравнивается с экспериментально определенной пространственной структурой белка.

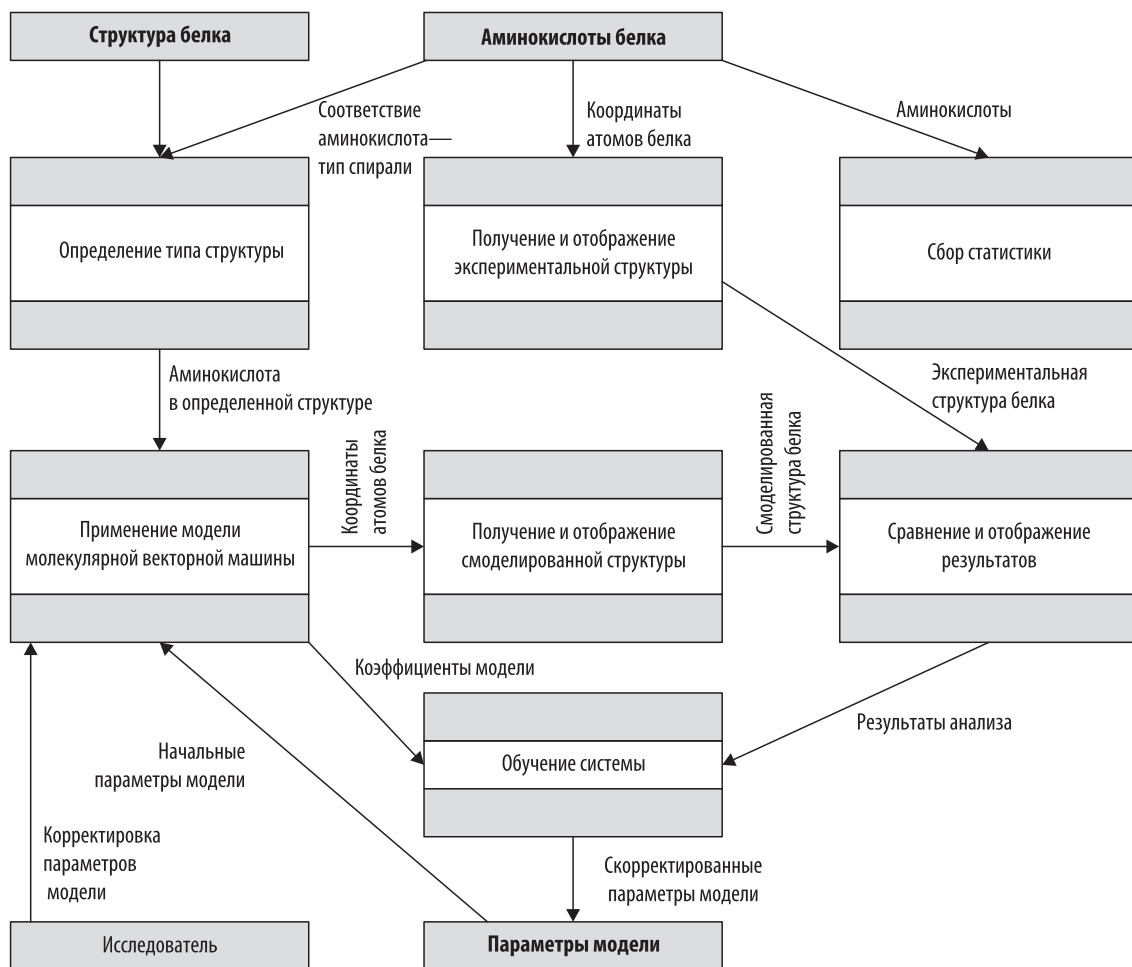


Рис. 12. Диаграмма потоков данных

Поток параметров модели обеспечивает:

- применение модели МВМ — начальные параметры модели, взятые из хранилища «Параметры модели», используются для корректировки начального положения текущей аминокислоты и окончания построенной белковой цепи.
- обучение системы — на основе ручной корректировки параметров модели и на основе анализа результатов происходит изменение параметров модели и их сохранение.

Результаты

В итоге модель молекулярной векторной машины реализована как исполнение конечного автомата. Реализован прототип системы, позволяющий проводить ручное моделирова-

ние пространственной структуры белка. Информация о проекте и исходные коды доступны по адресу <http://osll.spb.ru/wiki/genecode>.

СПИСОК ЛИТЕРАТУРЫ

1. Карасев В. А., Лучинин В. В. Введение в конструирование бионических наносистем. М: Физматлит, 2009.
2. Карасев В. А. Генетический код: новые горизонты. СПб.: Тесса, 2003.
3. Сосинский А. Б. Узлы и косы. М.: МЦНМО, 2001.
4. <http://biomolecula.ru>
5. <http://www.ams.org/featurecolumn/archive/knots-dna.html>
6. <http://boinc.bakerlab.org/rosetta>