на правах рукописи

УДК 533.9

ВЛАСОВ ПЁТР КОНСТАНТИНОВИЧ

Анализ конформаций аминокислотных остатков в глобулярных белках. Предсказание левой спирали типа полипролин II.

03.00.02. – биофизика

автореферат диссертации на соискание ученой степени
кандидата физико-математических наук

Москва 2003

Работа выполнена в Институте молекулярной биологии им. В.А. Энгельгардта РАН
на кафедре молекулярной биофизики Московского физико-технического института

Научные руководители:
доктор физико-математических наук, профессор
Владимир Гайевич Туманян
кандидат физико-математических наук
Михаил Абрамович Ройтберг

Официальные оппоненты:
доктор физико-математических наук
Владимир Абрамович Намиот
доктор физико-математических наук, профессор
Валентин Иванович Лобышев

Ведущая организация:
НИИ Биомедицинской химии им. В.Н. Ореховича РАМН

Защита состоится 19.06 2003 г. в 10 час. 00 мин. на заседании диссертационного совета К 212.156.03 при Московском физико-техническом институте (141700, Московская обл., г. Долгопрудный, Институтский пер. 9, МФТИ).

С диссертацией можно ознакомиться в библиотеке МФТИ.

Автореферат разослан «14» мая 2003 г.

Ученый секретарь
диссертационного совета
кандидат физико-математических наук В.Е. Брагин

Общая характеристика работы

Введение. Актуальность проблемы. В настоящее время установлено, что существуют три основные регулярные конформации полипептидной цепи в белках, как глобулярных, так и фибриллярных. Это правая α-спираль и левая спираль типа полипролин II, а также плоская структура типа параллельного или антипараллельного β-слоя, которая может рассматриваться как вырожденный случай спирали. Эти регулярные конформации являются одновременно вторичными структурами в том смысле, что, входя в иерархию структурных уровней белка, формируются на основе первичной структуры, - аминокислотной последовательности, - и образуют третичную, - пространственную, - структуру белковой цепи.

В то время как расшифровка первичной структуры белка, - получение его аминокислотной последовательности, - является ныне относительно простой процедурой, экспериментальные исследования каждого отдельного белка на предмет выяснения его вторичной (а уж тем более третичной) структуры пока что попросту невозможны, - слишком дорогостоящими таковые исследования являются, а потому носят совсем немассовый характер. В связи с этим предсказание вторичной структуры по аминокислотной последовательности белковой цепи является одной из важнейших задач прикладной биоинформатики. Но, несмотря на полвека исследований вторичной структуры, проблема далека от своего решения как с точки зрения чистой, так и прикладной науки. Так и нет однозначной физической теории, которая объясняла бы формирование вторичной структуры с позиций фундаментальной науки.

Для α-спирали и β-структуры уже существуют эффективные методы предсказания, использующие для такового аминокислотные последовательности аннотируемых белковых цепей. Но самое существенное, что до сих пор нет ни общего метода, пригодного для предсказания всех трех типов вторичной структуры (α, β, и полипролин II), - а вместе с тем до сих пор не создано метода, отдельно предсказывающего места локализации участков типа левой спирали PPII.

Цель и задачи исследования. Главной целью проведённой работы было изучение конформационных свойств аминокислотных остатков, составляющих последовательности

белковых цепей, и попытка использовать выявленные свойства в решении задачи предсказания вторичной структуры белковых цепей.

Дополнительный интерес вызвали конформационные свойства коротких последовательностей аминокислотных остатков, - олигопептидов, - как составных частей белковых цепей. Понимание, при каких длинах набор аминокислотных остатков уже способен проявлять конформационную стабильность, важно для выявления физических механизмов, ответственных за формирование и поддержание структуры белка.

Полученные результаты использовались в формализации и реализации нового метода предсказания вторичной структуры цепи по её аминокислотной последовательности, позволяющего в числе прочих предсказывать и структуру левой спирали типа полипролин II, составляющей значительную долю регулярных конформаций белковых цепей. Проверка высокого содержания структуры такого типа также значилась среди целей работы.

Методы исследования. Исследований конформационных особенностей олигопептидных фрагментов белковых цепей требует наличия большой выборки белков. На март 2002 года в международном банке данных PDB, суммирующем все публично доступные данные по пространственным структурам природных полимеров, содержалось более 17000 файлов со структурами белков, нуклеиновых кислот, комплексов белок-лиганд и т.п. При работе с полным банком и его подвыборками предварительно приходится избавляться от файлов, содержащих небелковые структуры. Мы подвергли такой обработке банк структур и выделили из него только цепи глобулярных белков и их комплексы. Вообще, в дальнейшем изложении термин «белковая цепь» будет подразумевать именно аминокислотную последовательность одной самостоятельной цепи некоторого глобулярного белка.

Увы, но банк PDB и его подвыборки до сих пор представляют собой наборы простых текстовых файлов, в каждом из которых содержится разнообразная информация о представленной структуре или комплексе. Сожаление в предыдущем предложении вызвано тем, что работа с данными подобного формата (удобного, конечно, для «просмотра глазами») не позволяет производить более-менее сложные исследования без создания громоздкого аппарата чтения данных, прикладных вычислений и вывода результатов в удобном виде, - который и был нами самостоятельно реализован.

Для проведения исследований был создан набор процедур на языке Perl, которые производят поиск интересующих отдельных остатков или олигопептидных фрагментов и аннотацию тех или иных конформационных свойств. Выбор языка Perl объясняется

удобством его использования в задачах чтения, записи и анализа больших массивов текстовых данных (которые и представляют собой файлы формата PDB). Для более громоздких вычислительных процедур использовался инструментарий, реализованный на языке C++, - наиболее подходящем для создания программ, выполняющих трудоёмкие вычисления.

Научная новизна. Несмотря на обширные исследования конформаций белковых цепей и составляющих их остатков, ранее не предпринимались попытки провести подобные разработки на всём многообразии представленных в публичном доступе структур белков и их комплексов. Потому ранее полученные данные о конформационных свойствах аминокислотных остатков, олигопептидов и целых белковых цепей зачастую отражали свойства небольшой группы белков, служивших объектом исследования. Мы впервые собираем обобщающую статистику конформаций для всех остатков, составляющих все белковые цепи из банка данных PDB, в котором воедино собрана вся информация о структурах, полученная мировым академическим сообществом за долгие года исследования биополимеров.

Отсутствие методов, успешно предсказывающих участки структуры левой спирали типа полипролин II, несло не только заметное неудобство в задачах аннотации белковых цепей, но и служило темой для спекуляций о возможной неоправданности выделения таковой конформации в отдельный класс вторичных структур. Предъявляемая нами статистика встречаемости конформаций различных типов в белках, а также метод предсказания левой спирали, дающий хорошие результаты не только для неё самой, но и для других вторичных структур, позволяют со всей ответственностью заявить о левой спирали как о самостоятельном типе конформаций белковой цепи.

Практическое значение работы. Основными сферами практического применения полученных нами результатов видятся три области:

- выявленные конформационные свойства аминокислотных остатков и олигопептидных фрагментов белковых цепей позволяют оценить вклад локальных взаимодействий в формирование структуры белка, сделать выводы о том, какие именно из таких взаимодействий играют главную роль в поддержании структуры белковой цепи,

- разработанный алгоритм предсказания вторичной структуры уже используется на практике, как для общей аннотации белковых цепей, так и, главным образом, для разметки участков левой спирали типа полипролин II,
- наборы олигопептидных фрагментов, обладающие особенно ярко выраженной конформационной стабильностью и хорошо предсказуемыми свойствами можно использовать в задачах практического синтеза искусственных пептидов с заданными структурными и/или функциональными характеристиками.

Апробация работы. Результаты работы были представлены на международной конференции BGRS'2002.

Публикации. За время работы над диссертацией опубликованы две статьи в реферированных журналах.

Объём и структура диссертации. Работа изложена на 95 страницах, иллюстрирована 19-ю рисунками и содержит 12 таблиц. Диссертация состоит из Введения и шести глав, включая литературный обзор. Список цитированной литературы содержит 82 наименования.

Содержание работы.

Первая глава

Содержит критический обзор литературы, посвящённый нескольким темам:

- исследования конформационных свойств аминокислотных последовательностей,
- методы предсказания вторичной структуры белковой цепи по её аминокислотной последовательности,
- ранее полученные результаты по левоспиральной конформации в белках.

Конформационные свойства аминокислотных остатков давно привлекали внимание исследователей. В приведённом обзоре литературы продемонстрировано, что большая часть ранее полученных результатов отражало свойства небольшой группы белков, выбранных исследователями в конкретной работе. Особенно мы подчёркиваем, что ранее само состояние банка структур белков (PDB, Protein Data Bank) не позволяло производить существенно обширные исследования конформационных свойств аминокислотных остатков и их

последовательностей. Лишь сейчас в распоряжении исследователей имеются не единицы и десятки, а тысячи белковых структур, а точность их разрешения существенно повысилась за последние годы.

Исследованию вторичной структуры посвящено большое число работ. Их целью является понимание физических основ реализации того или иного типа структуры, или, по крайней мере, феноменологическое описание экспериментально наблюдаемой картины вторичных структур. Такого рода знания и/или результаты статистических исследований структурных банков данных позволяют разрабатывать методы предсказания вторичной структуры по первичной структуре. Существует большое число методов, позволяющих с достаточно высокой точностью (около 80% верных предсказаний) предсказывать участки α-спирали и β-структуры. Описание таких методов, необходимое для их сопоставления с предлагаемым нами алгоритмом, приведено в этой работе.

Левая спираль всегда являлась «бедным родственником» среди вторичных структур. Экспериментальные данные подтверждали её заметное содержание в структурах белков, - но по сравнению с α-спиралью и β-структурой её доля всё равно признавалась несущественной. Мы уделяем особое внимание в обзоре тем редким работам о левоспиральной конформации, в которых авторы уже пытались аннотировать её как самостоятельный класс вторичных структур.

Вторая глава

Содержит описание и обсуждение результатов той части диссертационной работы, которая касается изучения конформационных свойств аминокислотных остатков на основе анализа структур белков банка PDB.

Конформационные характеристики аминокислотных остатков и их последовательностей. Главными конформационными характеристиками отдельного аминокислотного остатка выступают значения двугранных углов φ и ψ, являющиеся углами вращения вокруг связей N-Cα и C-C(O). В принципе, для конкретного остатка возможны любые сочетания этих двугранных углов, но комплекс стерических ограничений приводит к тому, что многие пары двугранных углов энергетически невыгодны остатку. Отражению возможных сочетаний углов φ и ψ служит карта Рамачандрана, - в координатах которой по горизонтали откладывается угол φ, в диапазоне от –180 до +180 градусов, а по вертикали, -

угол ψ, в тех же пределах. Точками на карте отмечаются реализуемые пары двугранных углов.

Дальнейшим исследованиям, результаты которых и составляют предмет данной работы, подвергались массивы двугранных углов аминокислотных остатков, составляющих белковые цепи. Типичный (и простейший) алгоритм статистического теста выглядел следующим образом:

- по очереди анализируются все цепи, составляющие исследуемую выборку,
- программа «бежит» по аминокислотной последовательности белка, рассматривая все взятые в рассмотрение остатки или олигопептидные фрагменты определённой длины,
- для каждого из остатка, определяется конформационая характеристика, - соответствующая точка на карте Рамачандрана,
- после обработки всех цепей, для отдельных остатков и различных позиций олигопептидов и последовательностей определяется частота появления различных значений конформационных характеристик.

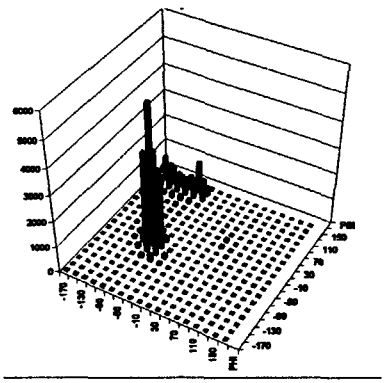
Статистика конформаций отдельных аминокислотных остатков. Приведённые в работе гистограммы для всех 20-ти основных белковых аминокислотных остатков демонстрируют распределение каждого из них по карте Рамачандрана. Высота столбцов, пропорциональна количеству встреч остатка с соответствующим набором двугранных углов.

Не видя смысла отдельно останавливаться на каждом аминокислотном остатке (результат по каждому из них в диссертации представлен отдельной иллюстрацией), можно сделать некоторые общие наблюдения:

- все аминокислотные остатки способны образовывать α-спиральную конформацию – более того, для всех них большее число появлений наличествует именно в таковой конформации,
- серин, пролин практически не встречаются в β-структурах,
- глицин, естественно, демонстрирует самую разнообразную конформационную подвижность.

В качестве примера мы приводим гистограмму встречаемости сочетаний двугранных углов для остатка аланина. В основании гистограммы находится карта Рамачандрана. Высота столбиков соответствует встречаемости соответствующей пары углов φ и ψ.

Рисунок 1.



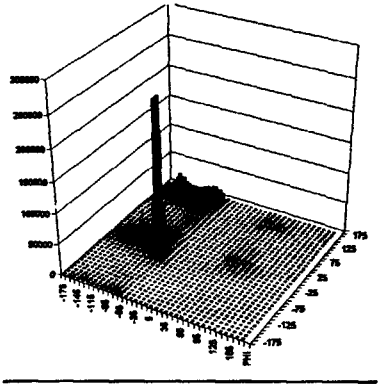
Максимум распределения приходится на значения φ = -65^0 и ψ = -45^0, что характеризует аланин как предпочитающий α-спиральную конформацию.

Третья глава

Представленные в главе результаты относятся к изучению свойств регулярных (в пространстве) структур, образовываемых в белковых цепях наборами аминокислотных остатков.

Встречаемость различных типов регулярной конформации в белках. Левая спираль как элемент регулярной конформации и вторичной структуры. До сих пор к регулярным вторичным структурам причислялись участки цепи, образующие систематические сетки межпептидных водородных связей. Третья конформация полипептидной цепи, причём формирующаяся в наиболее распространённом белке, коллагене, - а именно, левая спираль типа полипролин-II – рассматривалась как часто встречающаяся, но саттелитная конформация. В данной работе представляются данные о встречаемости левой спирали типа PP-II в банке белков. На гистограмме внизу приведены встречаемости различных сочетаний двугранных углов всех остатков всех белковых цепей, входящих в банк PDB.

Рисунок 2



В основании рисунка находится карта Рамачандрана.

Как видно, наибольшую встречаемость имеет α-спираль ($\varphi \sim -65^0$ и $\psi \sim -45^0$). Нас же больше интересует область левого верхнего угла карты, - кроме β-структур, занимающих область около $\varphi \sim -120^0$ и $\psi \sim +110^0$, также плотно заполнена и область, соответствующая левой спирали, $\varphi \sim -60^0$, $\psi \sim +150^0$. Более подробные цифры и их интерпретация приведены в ссобственно самой диссертационной работе, в соответствующем разделе, - а резюме выражается в том, что содержание левой спирали как конформации аминокислотных остатков, составляющих белковые цепи, лишь немного уступает содержанию β-структур!

Примерное процентное содержание трёх основных типов конформации:

- α-спираль составляет ~30% конформаций,
- β-структуры: ~20%,
- левая спираль ~15%.

Аминокислотный состав участков вторичной структуры.

Исследование аминокислотного состава участков вторичной структуры может дать богатую пищу для размышлений о физических механизмах формирования и стабилизации различных регулярных конформаций белковой цепи. Проведённые нами исследования

суммированы и представлены в таблицах. Данные собраны для полного банка PDB и негомологичной выборки белков.

Все таблицы имеют единообразный вид:

- заголовками столбцов выступают длины участков одной из трёх основных регулярных структур,
- заголовки строк, - аминокислотные остатки,
- в ячейках указана процентная доля соответствующего остатка в структуре соответствующей длины.

В качестве примера приведены таблицы по аминокислотному составу α-спиралей и левых спиралей типа полипролин II.

Таблица 3

α-спирали, полный PDB-банк.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	12	13	13	12	12	13	12	12	13	13	11	11	15	7	5
R	5	5	5	5	5	5	5	5	5	8	5	5	7	7	13
N	2	3	3	3	3	2	3	3	3	3	5	2	0	7	5
D	5	5	5	5	5	5	5	4	6	3	4	3	5	0	3
C	1	0	1	1	1	1	1	0	1	2	1	0	0	0	0
Q	3	4	4	4	5	5	5	5	4	4	5	9	6	7	5
E	8	9	8	8	9	8	7	8	7	8	8	10	8	17	3
G	3	3	4	3	3	3	3	3	2	2	3	2	2	1	6
H	1	1	2	1	1	1	2	2	2	1	2	3	2	5	0
I	4	4	4	5	5	5	5	5	6	6	4	2	1	0	1
L	9	9	11	12	12	12	12	13	11	9	12	16	13	1	23
K	7	7	6	6	6	7	6	6	7	7	7	8	6	19	6
M	2	2	2	3	3	3	3	3	3	5	1	3	3	5	0
F	2	3	3	3	3	3	3	4	2	6	1	1	0	5	0
P	8	5	2	1	1	1	0	1	1	0	1	0	1	0	0
S	5	5	5	4	3	4	4	4	4	2	6	2	1	0	0
T	4	3	4	3	3	3	3	3	4	2	4	2	6	4	16
W	1	1	1	1	1	1	1	0	0	1	1	0	0	1	5
Y	2	2	2	3	2	2	3	2	2	1	2	1	1	5	0
V	5	5	5	6	6	6	7	6	5	6	7	10	9	2	5

Таблица 4

Левые спирали, полный PDB-банк.

	1	2	3	4	5	6	7
A	8	8	8	9	11	6	14
R	3	3	4	3	4	1	4
N	3	2	2	2	1	1	1
D	5	4	4	5	4	4	3
C	1	1	1	1	1	1	2
Q	2	2	3	4	3	2	0
E	4	5	6	6	4	8	9
G	4	3	3	3	2	1	0
H	2	1	1	1	1	0	0
I	3	2	2	2	1	2	1
L	7	8	8	7	9	8	9
K	4	6	5	5	5	4	4
M	1	1	1	1	1	0	2
F	3	2	2	1	3	1	1
P	16	23	25	26	28	33	25
S	9	7	5	4	6	8	8
T	6	5	4	5	4	8	5
W	1	0	0	0	0	0	3
Y	3	2	1	1	1	1	0
V	4	3	4	4	2	1	3

Наиболее заметными и интересными характеристиками аминокислотного состава различных регулярных конформаций можно считать следующие:

I. α-спирали

- аминокислотный состав незначительно меняется в зависимости от длины участков,
- единственным существенным скачком является наличие немалого числа пролина в участках длиной в 1-2 остатка и практически полное его отсутствие в участках большей длины, - что, скорее, характеризует сам пролин как конформационно способного соответствовать альфа-спирали, но не встраивающегося в неё ввиду известных ограничений.

2. β-структуры
   - доля лейцина растёт с длиной структурного участка,
   - доля аспарагина падает с длиной структуры.
3. Левая спираль
   - содержание пролина увеличивается (что естественно, - его конформационно наиболее выгодные состояния соответствуют левому повороту цепи), особенно заметен скачок при переходе от единичных «левоповёрнутых» остатков к фрагментам большей длины,
   - доля аланина растёт с длиной спирали, – впрочем, малое количество левых спиралей длины более 5-6 остатков не позволяет достоверно судить о значимости этого результата.

В таблице порой наличествуют заметные колебания (резкие увеличения и уменьшения) долей тех или иных аминокислот в составе участков вторичной структур больших длин (около 10 и более остатков). Но следует учитывать, что таковые участки могут содержаться в гомологичных белках (и как показывают выборочные проверки, так и обстоит дело), и наблюдаемые особенности аминокислотного состава могут отражать лишь особенности конкретного белкового семейства, чьи представители преобладают в банке PDB. Такое уточнение делает очевидным необходимость осторожности выводов относительно наблюдаемых резких подвижек аминокислотного состава, если закономерности обнаружены лишь для протяжённых участков регулярных конформаций, серьёзно превышающих среднюю длину для таковой структуры.

Четвёртая глава

Посвящена изучению конформационных свойств наборов, - небольших последовательностей, - аминокислотных остатков. Такие наборы, называемые также олигопептидами, представляют из себя отдельные блоки целых белковых цепей и зачастую обладают самостоятельными конформационными свойствами, не зависящими от контекста последовательности, в которую они «встроены».

Встречаемостей олигопептидов различной длины в банке PDB и его подвыборках. Число комбинаторно возможных, различных олигопептидов длины L определяется степенной функцией и равняется 20 в степени L. Таким образом, для длины в

три остатка количество разнообразных «олигов» составляет 8000, для длины в четыре остатка – 160000, и т.д.

При работе даже с большим банком аминокислотных последовательностей (каковым является PDB) на предмет составления статистики олигопептидов необходимо представлять, в коей мере в банке будет представлено разнообразие изучаемых «олигов». Для ответа на этот вопрос было проведено предварительное сканирование базы последовательностей PDB и её упоминавшейся негомологичной выборки PDB-Select на предмет подсчёта встречаемостей три-, тетра- и пентапептидов. Выяснилось следующее:

- олигопептиды длиною до 4 остатков представлены в цепях PDB полностью, а встречаемости каждого из них «зашкаливают» за тысячи случаев,
- тетрапептиды представлены также хорошо, 90% из них имеют более сотни появления в белковых цепях, что позволяет действительно говорить о «статистически значимых» закономерностях их конформаций в случае выявления таковых,
- пентапептиды уже не обладают полнотой встречаемости в PDB, более того, - большая часть (из всех комбинаторно возможных) даже не встречается в банке. Соответственно, и олигопептиды большей длины не представляли для нас дальнейшего интереса, т.к. обладали очевидно низкой встречаемостью, не позволяющей делать серьёзных обобщений статистического характера относительно их конформационных свойств.

Таким образом, в дальнейшем в исследованиях основное внимание уделялось тетрапептидам.

Конформационная статистика тетрапептидов. В процессе работы возникла возможность собрать данные о тенденциях остатков конкретных олигопептидов находится в областях конформационной карты, соответствующих трём основным типам вторичной структуры. Т.е., собрать статистику соответствия пары двугранных углов φ и ψ каждого из четырёх остатков тетрапептида одной из трёх областей карты Рамачандрана, обозначающих регулярные конформации - вторичные структуры. Естественно, каждый из остатков тетрапептида может в доле случаев и не находиться в регулярной конформации.

Фрагмент конформационной статистической таблицы приведён ниже:

Olig	Count	A1	B1	P1	A2	B2	P2	A3	B3	P3	A4	B4	P4
AAAA	533	193	48	10	200	63	7	211	47	38	213	47	22
AAAR	228	102	15	6	117	16	10	98	14	2	84	15	19

...

Содержимое столбцов (в порядке следования слева направо):

- последовательность тетрапептида,
- общее число наблюдений тетрапептида.

...далее идут характеристики пар двугранных углов φ и ψ каждого из 4 остатков:

- количество соответствий пары двугранных углов 1-го остатка области α-спирали (A) на карте Рамачандрана,
- количество соответствий пары двугранных углов 1-го остатка области β-слоя (B),
- количество соответствий пары двугранных углов 1-го остатка области левой спирали (P),
- количество соответствий пары двугранных углов 2-го остатка области α-спирали,

и т.д....

Из приведённого фрагмента, в частности, следует, что все остатки тетрапептида AAAR имеют заметную тенденцию располагаться в конформации α-спирали, и особенно это свойственно второму остатку – 117 случаев, 228 появлений.

Разместить в данной работе таблицу для всех 160000 тетрапептидов не представляется возможным, - следует лишь упомянуть, что в действительности более 150000 тетрапептидов встретились в PDB более одного раза! Т.е. в нашем распоряжении имеется поистине огромный массив информации, отражающий конформационные предпочтения абсолютного большинства тетрапептидов и составляющих их остатков.

Тетрапептиды с высоким предпочтением определённой конформации. В ходе работы выяснилось, что существуют тетрапептиды, обладающие очень избирательной конформационной предпочтительностью! Их доля составляет около 10% от всех комбинаторно возможных 160000 тетрапептидов, а хорошо продемонстрировать их конформационную стабильность способны гистограммы встречаемости двугранных углов составляющих тетрапептиды аминокислотных остатков. А именно, для каждого случая появления конкретного терапептида в белковой цепи из PDB можно отметить пары двугранных углов аминокислотных остатков этого терапептида на карте Рамачандрана, причём высота столбиков пропорциональна частоте реализации соответствующего сочетания двугранных углов. Естественно, если все четыре составляющих тетрапептид остатка предпочитают (находясь в составе рассматриваемого тетрапептида, а не «вообще») принимать одну и ту же конформацию, то и тетрапептид в целом мы вправе считать «предпочитающим» таковую конформацию.

Ниже на рисунке приведены характерные тетрапептиды для трёх основных типов вторичной структуры, и один тетрапептид «смешанного типа».

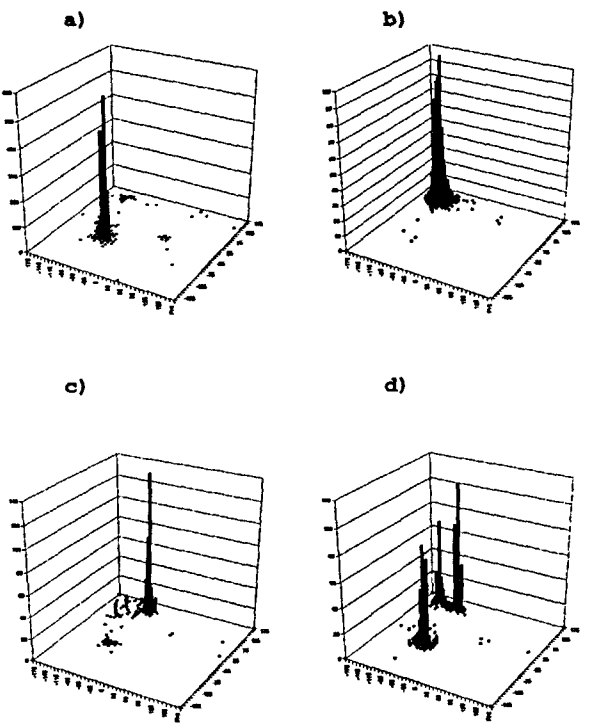
Рисунок 3

a) α-спиральный терапептид LFQN; встречается в белковых цепях PDB 447 раз; максимум на гистограмме распределения двугранных углов: $\varphi = -55$, $\psi = -45$,

b) β-структурный тетрапептид YTCE; встречаемость = 236; максимум гистограммы: $\varphi = -115$, $\psi = -125$,

c) терапептид ETPS левой спирали типа полипролин II; встречаемость = 123; максимум гисограммы: $\varphi = -65$, $\psi = -145$,

d) тетрапептид LKSA; встречаемость = 232; примерно одинаково часто представлены все три основных типа вторичной структуры.



Тетрапептиды с предпочтительной конформацией «смешанного» типа. На данный момент уже имеется немало информации о коротких аминокислотных последовательностях (пептидах), предпочтительно формирующих определённую непрерывную регулярную структуру из числа «канонических», - α- и β-структуры. Как указано выше, в представляемой работе выделены и олигопептидные фрагменты, предпочитающие образовывать участки левой спирали. В связи с упомянутыми данными возникает вопрос: существуют ли аминокислотные последовательности, в большой мере предпочитающие образовывать в белках участки стабильной, но «неканонической», «смешаной» конформации, - т.е.:

- для каждой позиции «олига» свойственна та или иная конформация,
- предпочтительная конформация не однозначна, а меняется от позиции к позиции.

Собранная в ходе работы статистика конформаций тетрапептидов позволила выделить те из них, которые удовлетворяют упомянутым свойствам.

Результаты по 10-ти наиболее встречающимся тетрапептидам приведены в таблице.

Таблица 5

Олиг	Встречаемость в PDB	Конформация	поз №1	поз №2	поз №3	поз №4
TKDE	634	PAAA	88	84	78	78
FPPS	400	BPPP	84	84	85	86
YYTI	395	PPBB	81	87	85	92
SRWY	391	BAAA	84	87	90	85
RWDE	389	BAAA	86	87	85	85
WVAW	387	BAAA	84	85	78	85
KSRW	381	BBAA	85	85	92	94
TFPA	355	BPPP	71	79	87	85
PPSS	327	PPPA	83	82	82	72
HTFP	321	BBPP	76	77	84	92

Для упомянутых тетрапептидов приведена доля предпочтительной конформации, причём на каждой из четырёх позиций. Цифры в данном случае говорят сами за себя, - конформационное предпочтение явное, а варьируется оно от позиции к позиции.

Выясняется, что существуют даже тетрапептиды со стабильной конформаций, в состав которых входят все три основных типа вторичной структуры! Т.е. в таких «олигах» наличествует и α-, и β-, и левоспиральная конформация, - и каждая из них стабильно присуща некоторой определённой позиции данного тетрапептида. В приведённой ниже таблице упомянуты 10-ть наиболее часто встречающиеся такие последовательности.

Таблица 6

Олиг	Встречаемость в PDB	Конформация	поз №1	поз №2	поз №3	поз №4
ISLA	272	BPAA	59	70	90	72
IVLK	215	BBPA	69	71	68	66
LAQV	233	PAPB	52	59	52	55
LKSA	232	PABP	66	71	65	66
KPVN	215	PABB	71	54	77	75
KSAA	245	ABPP	63	64	62	73
FTPA	274	BPAA	66	60	67	55
YSLG	211	BPAA	82	82	63	56
YTKV	257	BPBA	79	86	86	65
VFER	281	BPPA	75	65	72	71

Довольно высокая частота встречаемости тетрапептидов со «смешанной» конформацией отвергает возможность объяснения наличия у них предпочтительных конформаций лишь фактором гомологичности цепей белков, в которых эти тетрапептиды встречаются. Приведённые терапептиды встречаются в белках с существенно различными аминокислотными последовательностями и различными типами организации третичной структуры. Это наводит на мысль о большей роли локальных взаимодействий в образовании таких стабильных конформаций на уровне малых длин последовательностей.

Пятая глава

Представляет собой описание разработанного и реализованного нами алгоритма предсказания вторичной структуры белка по его аминокислотной последовательности. Приведены результаты тестирования нашего алгоритма, позволяющие оценить степень достоверности даваемых им результатов и убедиться в высоком уровне предсказания именно

левой спирали типа полипролин II, до того не предсказываемой с высоким качеством ранее созданными алгоритмами.

Алгоритм предсказания вторичной структуры аминокислотной последовательности, основанный на конформационной статистике олигопептидов. Наличие конформационной статистики олигопептидов позволяет создать простой алгоритм предсказания конформаций для аминокислотных последовательностей-запросов. Работа такого алгоритма основана на рассмотрении каждого аминокислотного остатка последовательности в контексте включающих его (в данной последовательности) олигопептидов.

В качестве «опорных» олигопептидов, чья статистика и будет использоваться в предказательном алгоритме, предлагается использовать тетрапептиды. Такой выбор определяется двумя причинами:

- чем больший размер взят для олигопептидов, тем больше причин считать, что взаимодействия именно входящих в него остатков ответственны за формируемую локальную конформацию и/или участок вторичной структуры,
- довольно полная (по крайней мере, с точки зрения количества встреч и разнообразий контекста появлений) статистика по всему банку пространственных структур PDB имеется для терапептидов, но не может быть (пока...) собрана для олигопептидов больших длин.

Например, остаток аланина, находящийся в третьей позиции подпоследовательности ISPALKW, входит в состав четырёх тетрапептидов: ISPA, SPAL, PALK и ALKW. В предположении наличия достоверной конформационной статистики данных тетрапептидов, а именно информации об их тенденциях образовывать определённые конформации, можно предсказать и конформацию данного участка последовательности и конкретного остатка, аланина. Качество такого предсказания, предположительно, будет зависеть от того, в коей мере стабильные конформации характерны для рассматриваемых олигопетидов, и сколь полна имеющаяся статистика (т.е. отражает ли она всё многообразие возможных конформаций).

Перейдём к строгому описанию работы алгоритма предсказания конформации.

1. Рассмотрим фрагмент некоторой последовательности, для остатков которой пытаемся предсказать конформацию, как последовательность перекрывающихся олигопептидов длины L, соответствующих движению «рамки» размера L с шагом в 1 остаток.

2. В дальнейшем под «олигопептидами» будем подразумевать именно тетрапептиды (L=4). Например, подпоследовательность NLPMAWRT представляется как набор тетрапептидов: NLPT, LPTA, PTAW, MAWR, AWRT.
3. Пусть N(Olig) – общее число наблюдений олигопептида Olig в банке данных.
4. Обозначим как C(Olig, S, I) число нахождений остатка с позиции I олигопептида Olig в области карты двугранных углов, соответствующей структуре S (S – это (α-спираль (A), β-слой (B), левая спираль (P) или область нерегулярных структур (-)).
5. Возвращаясь к рассматриваемой подпоследовательности, при предсказании конформации каждого из остатков обратимся к его конформационной статистике в контексте каждого из содержащих его (в данной последовательности) терапептидов. Например, в данном случае, остаток аланина на пятой позиции последовательности содержится в тетрапептидах LAAA, AAAA, AAAR и AART. Обозначим эти олиги индексом J, 1<=J<=4 (Olig[1]=LAAA, Olig[2]=AAAA,...). Присвоим рассматриваемому остатку (аланина) вероятность быть в конформации S по формуле:

$$P(S) = ¼*(C(Olig[1],S,4)/(N(Olig[1]) + C(Olig[2],S,3)/N(Olig[2]) + C(Olig[3],S,2)/N(Olig[3]) + C(Olig[4],S,1)/N(Olig[4])),$$

$$S = A, B, P, - .$$

т.е. вероятность просто усредняет имеющиеся четыре значения для разных содержащих рассматриваемый остаток тетрапептидов.

Такая математическая модель имеет вполне понятное, интуитивное объяснение: комбинируя информацию статистического характера о конформации рассматриваемого остатка в контексте всех включающих его олигопептидов, мы повышаем «информационный вес» данных наблюдений по часто наблюдаемым (имеющим общирную статистику) олигопептидам и понижаем таковой «вес» для данных по редко наблюдаемым олигопептидам (чья конформационная статистика менее достоверна).

Тестирование алгоритма предсказания вторичной структуры. Для каждого остатка каждой цепи тестовой выборки предсказывался тип структуры (конформации) по приведённому алгоритму. Результат предсказания сравнивался с действительной конформацией, известной из хранимой в PDB структуры цепи. Далее для трёх типов конформации подсчитывались две характеристики:

- доля остатков, действительно находящихся в некоторой конформации и верно размеченных алгоритмом, - качество работы метода предсказания.
- доля остатков, размеченных алгоритмом как находящиеся в некоторой конформации и действительно (по структурным данным) в неё входящих, - точность метода.

Результаты предсказания выражены в процентах.

Нетрудно понять, что можно завысить качество за счёт перепредсказания, - например, размечая всё остатки как относящиеся к α-спирали, мы достигнем 100% качества анотации α-спирали, - но ведь при этом упадёт точность метода. Потому, как и для большинства аналогичных предсказательных методов, для представляемого алгоритма оптимальным видится режим работы, при котором качество и точность предсказания примерно равны.

Для каждого белка подсчитывались упомянутые качество и точность предсказания. Тестирование метода предсказания велось на основе двух обучающих выборок, - массивов *конформационной статистики*:

- статистика, построенная по полному банку пептидных цепей PDB, - за исключением лишь самих цепей тестовой выборки,
- статистика по банку с исключёнными цепями всех близких гомологов цепей тестовой выборки.

И результирующих таблиц представлено две, - для каждой из обучающих выборок.

Одна строка соответствует одной белковой цепи, для каждого из трёх основных типов вторичной структуры приведены, через «слэш» («/»), значения качества и точности предсказания. В конце таблиц указаны средние значения, - причём усреднение делалось по сумме всех позиций в белках тестовой выборки.

Таблица 7А

1a6m	98/91	75/100	100/57
1aho	50/62	70/58	50/53
1b0y	100/56	58/87	54/86
1brf	40/22	50/80	83/76
1bxo	61/50	66/77	72/70
...	...	...	...
2pvb	72/67	33/42	0/0
3al1	100/90	0/0	0/0
3lzt	96/83	84/78	89/89
3pyp	37/43	59/72	45/62
7a3h	83/77	83/87	76/68
Среднее:	72/73	63/64	61/61

Таблица 7Б

1a6m	96/92	75/75	100/57
1aho	40/57	70/58	50/53
1b0y	71/45	58/87	54/72
1brf	40/22	50/44	75/69
1bxo	57/45	67/78	76/72
...	...	...	...
2pvb	66/66	22/33	0/0
3al1	100/90	0/0	0/0
3lzt	93/83	84/78	89/89
3pyp	29/33	48/65	27/50
7a3h	82/75	77/88	76/70
Среднее:	68/70	59/61	56/57

Естественно, качество работы метода предсказания меняется при использовании «сильно обеднённого» статистического массива. Как видно из результирующих таблиц, падение качества и точности предсказания составляет порядка 5-10%, не более, при переходе от «полной» к «негомологичной» статистике. Это, по сути незначительное изменение работы метода, уже само по себе демонстрирует, что именно короткие олигопептидные фрагменты предоставляют достаточное количество информации для неплохого предсказания конформации отдельных составляющих их остатков. Следует также отметить тот факт, что при исключении гомологичных последовательностей мы существенно уменьшаем объём данных, использованных для составления конформационной статистики олигопептидных фрагментов, - т.е. элементарно «усекаем» обучающую выборку.

Обобщая итоги тестов, можно оценить точность и качество работы алгоритма предсказания следующими цифрами:

- для α-спирали ~70%,
- для β-структуры ~ 65%,
- для левой спирали ~ 60%.

Выводы

1. На основе анализа статистики встречаемостей различных типов конформаций аминокислотных остатков цепей глобулярных белков на карте двугранных углов определены области, характерные для различных типов вторичной структуры. Определено содержание и подтверждена высокая встречаемость двух типов вторичной структуры: α-спирали ($\varphi \sim -65^0$ $\psi \sim -45^0$) и β-структуры ($\varphi \sim -120^0$ $\psi \sim +110^0$). Показано наличие «плотно заполненной» области левой спирали ($\varphi \sim -60^0$ $\psi \sim +150^0$).
2. Обнаружено, что встречаемости остатков в данной области и в β-структуре примерно равны: ~15-20%. Это опровергает звучащие утверждения о несущественном наличии левоспиральной конформации в белках и демонстрирует обоснованность введения классификации конформаций по трём основным типам: α-спирали, β-структуры, левые спирали.

3. Для всех возможных 160000 тетрапептидов впервые составлены таблицы конформаций составляющих их остатков. Получены частоты появления каждого остатка (с 1-го по 4-тый) тетрапептидов в каждой из трёх основных областей конформационной карты: α-спирали, β-структуры и левые спирали типа полипролин II. Для каждой из трёх основных вторичных структур впервые сформированы полные списки «предпочитающих» (встречаемость в структуре > 80%) их тетрапептидов и тетрапептидов, таковые структуры «избегающих» (встречаемость в структуре < 5%).
4. Создан алгоритм предсказания конформации белковой цепи по её аминокислотной последовательности, впервые дающий предсказание участков левой спирали. Тестирование работы этого алгоритма на белковых цепях, составляющих банк PDB-Select дало следующие результаты:
   - для α-спиралей верно предсказываются ~ 70% позиций,
   - для β-структур точность ~ 65%,
   - для левых спиралей точность ~ 60%.

Список публикаций

1. V.E. Ramenskii, P.K. Vlasov, Sh.R. Syunyaev, V.G.Tumanyan
How do point amino acid substitutions affect the protein structure?
Biophysics, 2000, Vol. 45(2), 220-227.

2. P.K. Vlasov, G.T. Kilosanidze, D.L. Ukrainskii, A.V.Kuzmin, V.G. Tumanyan, N.G.Esipova
Left-handed conformation of poly-L-proline II type in globular proteins. Statistics of incidence and a role of sequence.
Biophysics, 2001, Vol. 46(3), 573-576.

3. P.K. Vlasov, G.T.Kilosanidze, D.L. Ukrainskii, V.G. Tumanyan, N.G. Esipova
Predominant conformations of oligopeptide fragments of globular proteins.
Proceedings of the third international conference on bioinformatics of genome regulation and structure, BGRS'2002, Vol.3, 136-138.

Власов Пётр Константинович

АНАЛИЗ КОНФОРМАЦИЙ АМИНОКИСЛОТНЫХ ОСТАТКОВ В ГЛОБУЛЯРНЫХ БЕЛКАХ. ПРЕДСКАЗАНИЕ ЛЕВОЙ СПИРАЛИ ТИПА ПОЛИПРОЛИН II

Подписано в печать 23.05.2003
Формат 60 x 84 1/16. Бумага офсетная. Печать офсетная. Усл. печ. л. 0,5.
Тираж 70 экз. Заказ № ф-27

Московский физико-технический институт
(государственный университет)
Отдел автоматизированных издательских систем
«ФИЗТЕХ-ПОЛИГРАФ»
141700б Московская обл., г. Долгопрудный, Институтский пер., 9