ГЕОМЕТРИЯ АМИНОКИСЛОТ И ПОЛИПЕПТИДОВ

А.О. Иванов, А.С. Мищенко, А.А. Тужилин

Московский государственный университет им. М.В. Ломоносова,
Московский государственный технический университет им. Н.Э. Баумана

aoiva@mech.math.msu.su, asmish@mech.math.msu.su, tuz@mech.math.msu.su

Поступила 15.04.2014

В работе описан ряд геометрико-статистических методов, которые позволяют анализировать информацию, содержащуюся в Protein Data Bank и других подобных базах даных, на на внутреннюю непротиворечивость и согласованность с общей теорией. Также демонстрируется ряд примеров, показывающих, что данные PDB требуют тщательного отбора и анализа.

УДК 514.8+57.087+51-76

Введение

Задача изучения конформации полипептидов хорошо известна и является чрезвычайно важной. Ее огромное значение связано, а частности, с тем, что изменение конформации входящих в клетку белков, скажем, в следствии мутации, может приводить к серьезным заболеваниям организма. Отметим, что экспериментальное определение последствий той или иной мутации является длительным и дорогим процессом. Поэтому возникает естественное желание — подобные эксперименты смоделировать

на компьютере. Однако задача оказывается крайне сложной, и несмотря на усилия огромного научного сообщества, в настоящее время эта задача все еще далека от решения.

Отметим, что обсуждаемая проблема находится на стыке многих наук, а именно, биологии, химии, физики, геометрии, вариационного исчисления, теории вероятностей, вычислительной математики. Возникает насущная необходимость привлечения внимания к этой проблеме ученых самых разных специальностей и создание возможности для заинтересовавшихся быстро войти в курс дела и начать экспериментировать с накопленными данными.

Центральное место в сборе информации о пространственной структуре полипептидов играет Protein Data Bank (Банк Данных Белков). Мы опишем, как можно достаточно быстро выбрать нужную для исследования геометрическую информацию, а затем расскажем о целом ряде досадных препятствий, с которыми мы столкнулись при работе с базой.

Когда мы только собирались писать настоящую статью, мы предполагали продемонстрировать, как некоторые геометрические аналогии могут оказаться полезными при изучении конформаций. Но в какой-то момент мы решили проверить данные PDB на “типичность”, чтобы иметь более точное представление о том, с чем мы имеем дело. Результаты нашего исследования изложены в следующих параграфах. После ряда “чисток”, мы наконец получили часть базы, которая, хочется верить, описывает “типичное” поведение полипептидов. Мы надеемся, что описанная нами методика тестирования геометрической базы данных статистико-геометрическими методами окажется полезной для создания более надежных баз такого типа.

1. Protein Data Bank

> Зона — это ... очень сложная система ... ловушек, что ли ... и все они смертельны!
> ...
> Может даже показаться, что она капризна, но в каждый момент она такова, какой мы ее сами сделали ... своим состоянием.
>
> из фильма А.А.Тарковского “Сталкер”

В данном разделе мы хотим поделиться своими впечатлениями от использования Protein Data Bank (http://www.rcsb.org/pdb/home/home.do) для изучения геометрии аминокислот в составе полипептидов. Собственно говоря, мы хотели извлечь из этой базы информацию о трехмерной структуре полипептидов. Для этого мы решили воспользоваться файлами с расширением pdb: в этих файлах содержатся строки вида

```
ATOM 1 N ASN A 32     65.950  -13.181  61.696     1.00  0.00 N
```

в которых для атома, описанного в 3-ей позиции, указаны его координаты в позициях 7–9. Позиция 4 говорит о том, в какой аминокислоте этот атом содержится, а позиция 2 — номер этой кислоты в рассматриваемом полипептиде. Детали, описывающие структуру формата pdb, можно найти в двухсотстраничном файле

[1] ftp://ftp.wwpdb.org/pub/pdb/doc/format_descriptions/Format_v33_Letter.pdf

Для работы с базой данных мы решили написать свои программы в пакете Mathematica (Wolfram Research Group). Те, кто обладаете быстрым интернетом, могут не скачивать заранее pdb-файлы. Например, чтобы воспользоваться файлом, описывающим полипептид с коротким именем 2LTR, достаточно выполнить команду

pdbFile=Import[“http://www.rcsb.org/pdb/files/2LTR.pdb”, “List”];

после этого переменная pdbFile будет представлять собой список строк, содержащихся в файле 2LTR.pdb, и эти строки можно дальше обрабатывать с помощью команд пакета Mathematica, извлекая интересующую информацию. Для работы с другим полипептидом, нужно в приведенной выше команде просто заменить 2LTR на короткое имя интересующего полипептида. Осталось выяснить, где же взять список коротких имен полипептидов?

1.1. Как извлечь нужную нам информацию с PDB?

Вот наш алгоритм:

(1) заходим на сайт http://www.rcsb.org/pdb/home/home.do

(2) в левом верхнем углу нажимаем на кнопку Advanced, расположенную сразу под кнопкой Search

(3) в появившемся разделе Advanced Search Interface

   (a) в выпадающем меню Choose a Query Type выбираем All/Experimantal Type/Molecule Type;
   внизу появляются два поля:
   Experimental Method и Molecule Type

   (b) в Molecule Type выбираем Protein

   (c) справа нажимаем на кнопку Result Count; под этой кнопкой появляется количество выбранных объектов: в нашем случае было написано 89790 PDB Entries (Structures)

   (d) кликаем в эту надпись — внизу получаем список всех выбранных полипептидов, вместе с их короткими именами и многочисленной другой информацией; по умолчанию в каждом квадратике перед коротким именем полипептида стоит галочка — это означает, что данный полипептид выбран.

Конечно, удобней сохранить где-нибудь список всех коротких имен. Для этого кликаем в поле Filter (в строке меню сразу над списком полипептидов) и в выпадающем меню выбираем Download Checked. Открывается страничка Structure Downloaded, в котором есть поле Enter PDB IDs:, где перечисляются все короткие названия выбранных полипептидов (они совпадают с именами соответствующих pdb-файлов). Кликаем в это поле, метим все элементы этого списка (Ctrl-A в Windows), копируем в буфер обмена (Ctrl-C), открываем какой-нибудь текстовый файл (или создаем новый) и копируем содержимое буфера в этот файл (Ctrl-V). Сохраняем файл: теперь в нем содержатся короткие имена всех полипептидов, имеющихся на данный момент в Protein Dada Bank.

Замечание 1.1. Обратите внимание на то, имеется ли пробел после имени в последней строчке полученного файла: если да — уберите его (наличие пробела в нашем случае приводило к ошибкам при выполнении некоторых функций Математики).

Кстати, этим же алгоритмом можно скачать все pdb-файлы с описаниями полипептидов на свой компьютер. Для этого после появления Structure Downloaded, вместо сохранения списка имен, можно

(1) в столбце Download Type: убрать галочку с поля mmCIF Format и поставить галочку в PDB Format;

(2) в столбце Compression Type: переместить точку в поле uncompressed;

(3) внизу нажать Next

(4) теперь Вам нужно, чтобы у Вас был включен JAVA-плагин: это плагин вызовет кнопку Browse, которая позволит выбрать путь для сохранения файлов; выбрав его, нажмите кнопку Start Download. Через некоторое время вы станете обладателем pdb-файлов, описывающих все выбранные полипептиды.

Давайте начнем с визуализации выбранных полипептидов или отдельных их фрагментов.

1.2. Первые шаги самостоятельной визуализации полипептидов

Несложно выделить строки вида “ATOM ...” (см. выше); сгруппировать эти строки по последовательным аминокислотным остатками (в дальнейшем, для краткости, — *аминокислотам*), используя для этого 6-ое поле, где расположен номер аминокислоты в рассматриваемой полипептидной цепи; извлечь из каждой такой строки название атома, его координаты, название аминокислоты, в которую атом входит. Однако, чтобы изобразить структуру связей, эти связи нужно откуда-то взять. В самом pdb-файле связи не представлены. Конечно, структура аминокислот хорошо известна, но в одну и ту же аминокислоту может входить несколько атомов одного и того же типа. Как установить соответствие между последовательными строками “ATOM...” описания одной аминокислоты из pdb-файла и атомами в структурной формуле аминокислоты? В файле [1] такой информации мы не обнаружили. Зато нашли старую версию, а именно,

[2] http://www.wwpdb.org/documentation/PDB_format_1992.pdf

где имеются соответствующие картинки (стр. 27). Впрочем, здесь ничего не сказано про водороды. Но их расположение можно сообразить по обозначениям. Например, если водород записывается как HD, то нужно найти в данной аминокислоте атом или углерода, или азота, или кислорода, или серы, который записывается соответственно через CD, или ND, или OD, или SD. В результате мы построили следующую таблицу связей:

N→CA, CA→C, C→O, CA→CB, C→OXT,
CB→CG, CB→CG1, CB→CG2, CB→OG, CB→OG1, CB→SG,
CG→CD, CG→CD1, CG→CD2, CG→ND1, CG→ND2, CG→OD1, CG→OD2, CG→SD,
CG1→CD1,
CD→CE, CD→NE, CD→OE1, CD→OE2,
CD1→CE1, CD1→NE1,
CD2→CE2, CD2→CE3, CD2→NE2,
CE→NZ, CE→SD,
CE1→CZ, CE1→ND1, CE1→NE2,

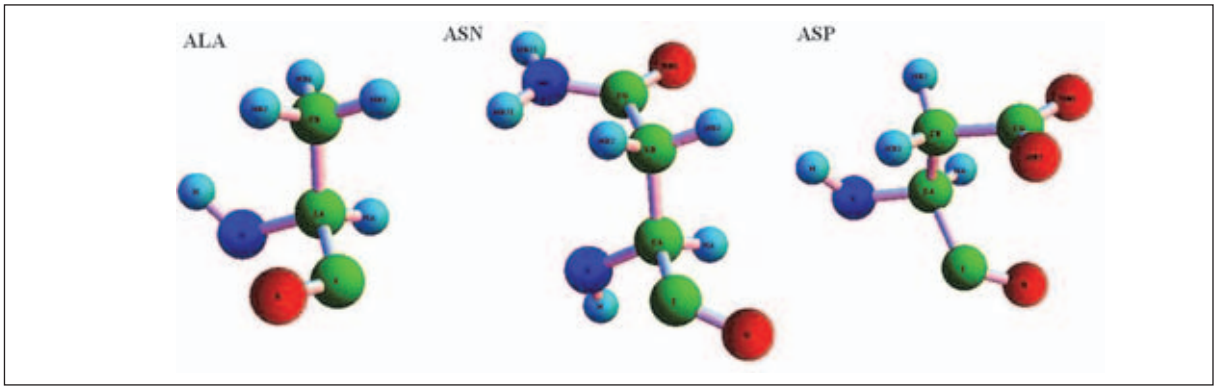

Рис. 1. Пример изображений аминокислот

CE2→CZ, CE2→CZ2, CE2→NE1,
CE3→CZ3,
CH2→CZ2, CH2→CZ3,
CZ→NE, CZ→NH1, CZ→NH2, CZ→OH,
CA→HA, CA→HA2, CA→HA3,
CB→HB, CB→HB1, CB→HB2, CB→HB3,
CG→HG, CG→HG2, CG→HG3,
CG1→HG11, CG1→HG12, CG1→HG13,
CG2→HG21, CG2→HG22, CG2→HG23,
CD→HD2, CD→HD3, **CD→N**, CD→NE2,
CD1→HD1, CD1→HD11, CD1→HD12, CD1→HD13,
CD2→HD2, CD2→HD21, CD2→HD22, CD2→HD23,
CE→HE1, CE→HE2, CE→HE3,
CE1→HE1,
CE2→HE2,
CE3→HE3,
CH2→HH2,
CZ→HZ,
CZ2→HZ2,
CZ3→HZ3,
N→H, N→H1, N→H2, N→H3,
ND1→HD1,
ND2→HD21, ND2→HD22,
NE→HE,
NE1→HE1,
NE2→HE21, NE2→HE22,
NH1→HH11, NH1→HH12,
NH2→HH21, NH2→HH22,
NZ→HZ1, NZ→HZ2, NZ→HZ3,
OG→HG,
OG1→HG1,
OH→HH,
SG→HG

Используя возможности Mathematica, связанные с рисованием графов, мы научились изображать как отдельные аминокислоты полипептида, так и разные его фрагменты. На рис. 1 приведено несколько примеров изображений аминокислот.

Однако некоторые аминокислоты оказались изображенными неправильно, см. рис. 2.

В чем же дело? Обратите внимание на выделенную связь CD→N в приведенной выше таблице. Оказывается, это она все портит. Мы выяснили, что наличие этой связи — единственная проблема в списке связей. А именно, если все аминокислоты, кроме пролина (рис. 3), изображать с помощью всех оставшихся связей, то ошибок не возникает. Для изображения пролина приходится дополнительно добавлять связь CD→N.

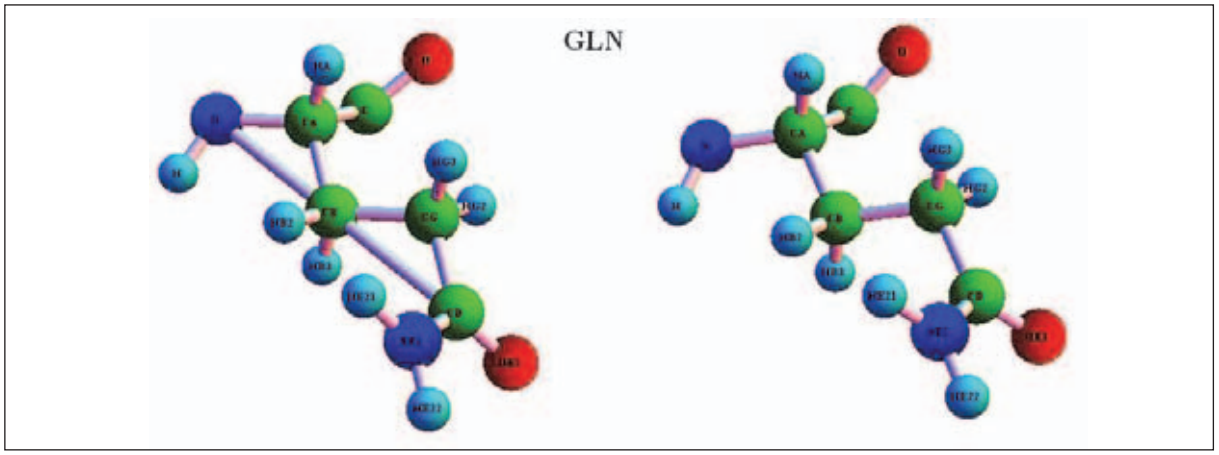

Рис. 2. Слева изображена неправильная структурная формула глутамина. Правильная изображена справа

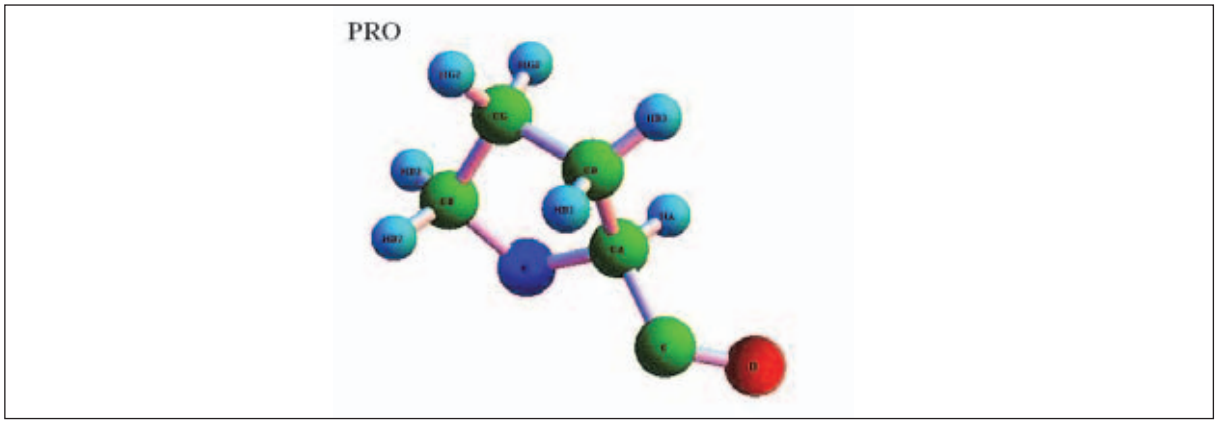

Рис. 3. Пролин

Можно ли решить эту проблему с помощью модификации обозначений атомов аминокислот? Да, это очень просто: достаточно в пролине переименовать CD-атом, скажем, в CF или CW (т.е. дать такое имя, которое не присутствует в названиях атомов других аминокислот). Видимо, разработчикам обозначений это было не нужно.

Разобравшись со связями, мы решили заняться сравнением геометрии одноименных аминокислот, находящихся в разных местах одного полипептида и, более общо, в разных полипептидах. Но тут мы столкнулись с рядом проблем.

1.3. Ряд трудностей, возникающих при работе с pdb-файлами

(1) Не во всех pdb-файлах присутствуют атомы водорода. Оказалось, что в ряде методов получения трехмерной структуры полипептидов атомы водорода просто не видны, как, скажем, при использовании рентгена. Мы выяснили, что наиболее подходящим для нас методом является ядерно-магнитный резонанс (NMR). Чтобы выбрать файлы, полученные именно таким методом, мы вернулись на сайт PDB и, в описанном выше алгоритме получения имен всех нужных нам файлов, в поле Experimental Method, которое раньше мы не модифицировали, выбрали NMR. Теперь в Result Count оказалось почти в 10 раз меньше файлов (около 9000).

(2) В следующих файлах не присутствуют строки “ATOM...”:

1R9V.pdb, 1S1O.pdb, 1S4A.pdb, 2KVJ.pdb

(в них имеются только строки “HETATM...”). Эти файлы мы исключили из списка.

(3) Во многих файлах содержится не один вариант полипептида, а несколько. Эти варианты называются *моделями* (models) и наличие нескольких моделей можно отловить по ключевому слову MODEL, с которого начинается строка, стоящая непосредственно перед строками “ATOM...”. Каждая модель начинается с MODEL N, где N — номер модели, и заканчивается строкой “ENDMODEL”. Мы решили выбирать в каждом файле первую модель. Отметим, что непосредственно перед строкой “ENDMODEL...” стоит заключительная строка последовательности строк “ATOM...”, а именно, *терминальная строка* “TER...” (в ней уже нет координат). Тем самым, при обработке pdb-файла мы выкидываем все начало, до первого вхождения строки “ATOM...”, и собираем в список последующие строки до тех пор, пока не появится строка “TER...”.

(4) Описанный только что простой алгоритм выделения первого связного фрагмента натолкнулся на следующее препятствие: оказывается, в некоторых файлах строки между первым вхождением “ATOM...” и первым появлением строки “TER...” могут перемешиваться со строками “HETATM...” (соответствующими *гетерогенным группам*, отличающимся от аминокислотных остатков стандартных 20-и аминокислот). Вот пример из файла 1A13.pdb

```
.......................
ATOM    235 HD22 LEU A  14      -7.905  -4.241   0.457  1.00  0.00          H  ,
ATOM    236 HD23 LEU A  14      -9.150  -4.650   1.637  1.00  0.00          H  ,
HETATM  237  N   NH2 A  15      -4.520  -3.275   5.538  1.00  0.00          N  ,
HETATM  238  HN1 NH2 A  15      -4.461  -2.351   5.201  1.00  0.00          H  ,
HETATM  239  HN2 NH2 A  15      -3.975  -3.556   6.306  1.00  0.00          H  ,
TER     240      NH2 A  15                                                    ,
.......................
```

Мы составили список всех таких файлов (их оказалось 604 штуки) и исключили их из нашего списка.

(5) В некоторых местах аминокислоты может располагаться по несколько копий одного и того же атома (так называемые *альтернативные расположения*). Выяснить, про какое расположение идет речь, можно по 5-ой позиции в строке “ATOM...”. Мы решили отбирать лишь те строки “ATOM...”, в которых на 5-ом месте стоит буква “A”. Кстати, если такие *альтернативные фрагменты* слишком длинные, то они идут отдельным списком, уже после терминальной строки. Мы усовершенствовали наш алгоритм, учтя пятую позицию.

(6) В некоторых файлах (1PNJ.pdb, 2PNI.pdb) несколько последовательных аминокислот метится одним и тем же номером, но различаются полем в следующей позиции, которое обозначается iCode и называется *кодом вставки остатка* (Code for insertion of residues). Вот пример из файла 1PNJ.pdb.

```
.......................
ATOM      1  N   GLY A   1A     18.846  -2.578 -21.181  1.00  0.00          N  ,
ATOM      2  CA  GLY A   1A     19.477  -3.527 -22.142  1.00  0.00          C  ,
ATOM      3  C   GLY A   1A     18.401  -4.379 -22.820  1.00  0.00          C  ,
ATOM      4  O   GLY A   1A     18.321  -4.433 -24.032  1.00  0.00          O  ,
ATOM      5  H1  GLY A   1A     17.815  -2.713 -21.185  1.00  0.00          H  ,
ATOM      6  H2  GLY A   1A     19.216  -2.760 -20.227  1.00  0.00          H  ,
ATOM      7  H3  GLY A   1A     19.069  -1.602 -21.461  1.00  0.00          H  ,
ATOM      8  HA2 GLY A   1A     20.022  -2.968 -22.888  1.00  0.00          H  ,
ATOM      9  HA3 GLY A   1A     20.157  -4.175 -21.609  1.00  0.00          H  ,
ATOM     10  N   SER A   1B     17.604  -5.017 -22.000  1.00  0.00          N  ,
ATOM     11  CA  SER A   1B     16.455  -5.834 -22.505  1.00  0.00          C  ,
ATOM     12  C   SER A   1B     15.466  -5.012 -23.344  1.00  0.00          C  ,
ATOM     13  O   SER A   1B     14.698  -5.571 -24.101  1.00  0.00          O  ,
ATOM     14  CB  SER A   1B     15.737  -6.447 -21.291  1.00  0.00          C  ,
ATOM     15  OG  SER A   1B     16.667  -7.386 -20.765  1.00  0.00          O  ,
ATOM     16  H   SER A   1B     17.763  -4.960 -21.035  1.00  0.00          H  ,
```

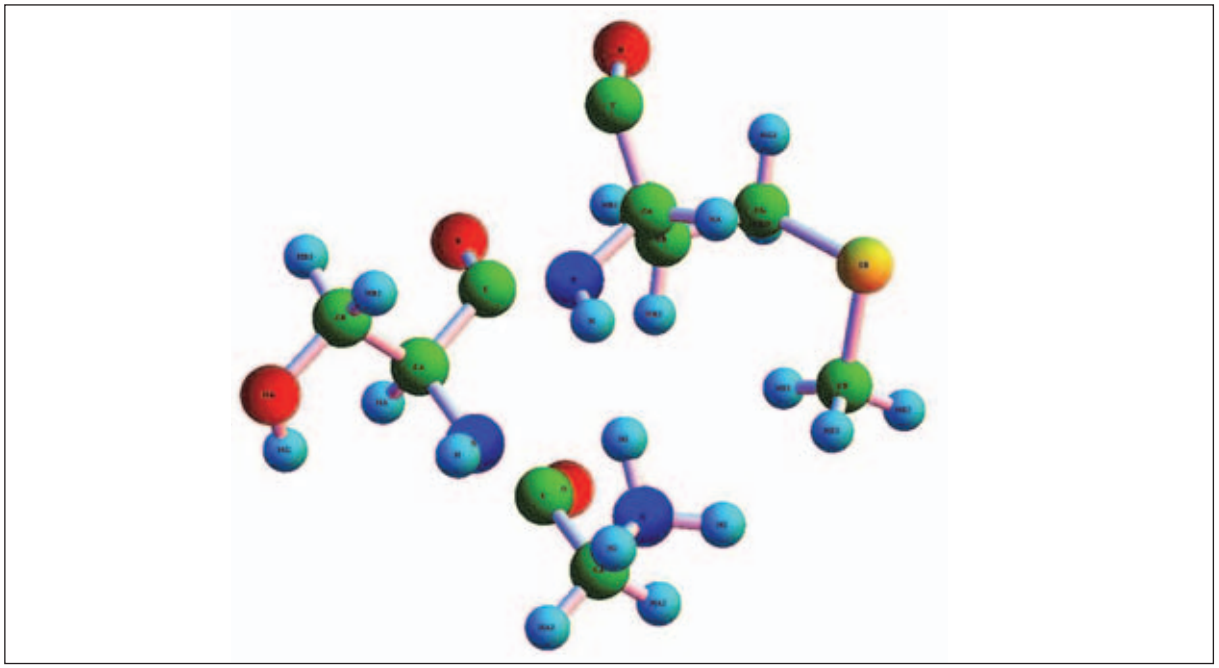

Рис. 4. Три различные аминокислоты, по ошибке распознанные как одна

```
ATOM     17  HA  SER A   1B     16.848  -6.626 -23.126  1.00  0.00          H  ,
ATOM     18  HB2 SER A   1B     15.523  -5.698 -20.541  1.00  0.00          H  ,
ATOM     19  HB3 SER A   1B     14.832  -6.958 -21.575  1.00  0.00          H  ,
ATOM     20  HG  SER A   1B     17.396  -7.490 -21.384  1.00  0.00          H  ,
ATOM     21  N   MET A   1C     15.515  -3.708 -23.185  1.00  0.00          N  ,
ATOM     22  CA  MET A   1C     14.651  -2.769 -23.978  1.00  0.00          C  ,
ATOM     23  C   MET A   1C     13.159  -2.991 -23.648  1.00  0.00          C  ,
ATOM     24  O   MET A   1C     12.312  -3.029 -24.519  1.00  0.00          O  ,
ATOM     25  CB  MET A   1C     14.942  -3.005 -25.496  1.00  0.00          C  ,
ATOM     26  CG  MET A   1C     14.656  -1.721 -26.304  1.00  0.00          C  ,
ATOM     27  SD  MET A   1C     15.658  -0.251 -25.959  1.00  0.00          S  ,
ATOM     28  CE  MET A   1C     17.288  -0.860 -26.468  1.00  0.00          C  ,
ATOM     29  H   MET A   1C     16.138  -3.336 -22.527  1.00  0.00          H  ,
ATOM     30  HA  MET A   1C     14.912  -1.756 -23.706  1.00  0.00          H  ,
ATOM     31  HB2 MET A   1C     15.976  -3.283 -25.625  1.00  0.00          H  ,
ATOM     32  HB3 MET A   1C     14.325  -3.806 -25.874  1.00  0.00          H  ,
ATOM     33  HG2 MET A   1C     14.777  -1.962 -27.348  1.00  0.00          H  ,
ATOM     34  HG3 MET A   1C     13.620  -1.444 -26.162  1.00  0.00          H  ,
ATOM     35  HE1 MET A   1C     17.204  -1.841 -26.912  1.00  0.00          H  ,
ATOM     36  HE2 MET A   1C     17.720  -0.181 -27.189  1.00  0.00          H  ,
ATOM     37  HE3 MET A   1C     17.934  -0.915 -25.603  1.00  0.00          H  ,
.........................
```

Игнорирование этого привело нас к “аминокислоте” с атомным составом

N, CA, C, O, H1, H2, H3, HA2, HA3, N, CA, C, O, CB, OG, H, HA, HB2, HB3, HG, N, CA, C, O, CB, CG, SD, CE, H, HA, HB2, HB3, HG2, HG3, HE1, HE2, HE3.

Мы расширили признак сборки аминокислот из строк “ATOM...” на позицию iCode, в результате чего эта “аминокислота” развалилась на три: GLY, SER, MET (глицин, серин, метионин), см. рис. 4.

(7) Аминокислоты, имеющие одни и те же имена, могут иметь разный атомный состав. Естественно, концевые аминокислоты должны отличаться от неконцевых: на N-конце присутствует пара “лишних” водородов, а на C-конце — “лишний” кислород. Тем самым, мы решили ограничиться *внутренними* (неконцевыми) аминокислотами.

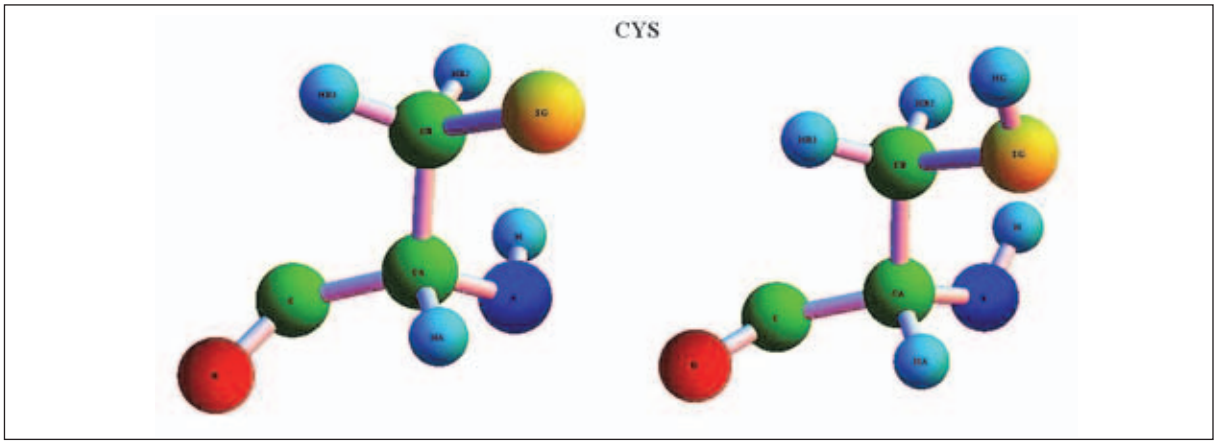

Рис. 5. Различные атомные составы внутреннего цистеина

Однако оказалось, что и внутренние одноименные аминокислоты могут иметь разное атомное строение! Например, у цистеина к атому серы иногда крепится атом водорода, а иногда — нет, см. рис. 5.

Мы прошли по всем файлам, и для каждой аминокислоты определили все возможные *атомные составы* аминокислот, содержащихся в выбранных полипептидах. Параллельно мы вычислили, сколько раз встречается каждая из полученных последовательностей атомов. Вот пример для глицина GLY:

```
{55705,{N,CA,C,O,H,HA2,HA3}},{1576,{N,CA,C,O,H1,HA2,HA3}},
{870,{N,CA,C,O,H1,H2,H3,HA2,HA3}},{399,{N,CA,C,O,OXT,H,HA2,HA3}},
{390,{N,CA,C,O}},{384,{N,CA,C,O,H}},{47,{N,CA,C,O,H,HA2}},
{37,{CA}},{37,{N,CA,C,O,HA2,HA3}},{26,{N,CA,C,O,H2,HA2,HA3}},
{15,{N,CA,C,O,H1,H2,HA2,HA3}},{15,{N,CA,C,O,HA2,HA3,H1,H2,H3}},
{13,{N,CA,C,O,H,HA}},{13,{N,CA,C,O,HA2, HN,HA1}}},{9,{N,CA,C}},
{8,{N,CA,C,O,H1,H2,H3}},{7,{C,CA,H,HA2,HA3,N,O}},
{5,{N,CA,C,O,OXT,H}},{5,{N,CA,C,H,HA2,HA3}},
{4,{N,CA,C,O,H,HA3,HA2}},{2,{N,CA,C,O,H1,H2,H3,HA3}},
{2,{N,CA,C,O,HA2,HA3,H1,H2}},{2,{N,CA,C,O,OXT}},{2,{CA,C}},
{1,{N,CA,C,O,OXT,H,HA2,HA3,HXT}},{1,{N,CA,C,O,H,HA2,HA3,H2,H3}},
{1,{N,CA,C,O,H2,H,H3,HA2,HA3}},{1,{N,CA,C,O,H2,H,HA2,HA3}},{1,{N}},
{1,{N,CA,C,OXT,H,HA2,HA3}},{1,{N,CA,C,O,HA2}},{1,{N,CA,C,O,H,HA3}}.
```

Затем мы выбрали наиболее часто встречающиеся последовательности и назвали их *стандартными*. Вот полученный список стандартных атомных составов аминокислот, вместе с количествами их появлений:

```
(GLY:55705) N,CA,C,O,H,HA2,HA3;
(ALA:50673) N,CA,C,O,CB,H,HA,HB1,HB2,HB3;
(SER:52599) N,CA,C,O,CB,OG,H,HA,HB2,HB3,HG;
(CYS:11784) N,CA,C,O,CB,SG,H,HA,HB2,HB3;
(PRO:32529) N,CA,C,O,CB,CG,CD,HA,HB2,HB3,HG2,HG3,HD2,HD3;
(VAL:48533) N,CA,C,O,CB,CG1,CG2,H,HA,HB,HG11,HG12,HG13,HG21,HG22,HG23;
(THR:39499) N,CA,C,O,CB,OG1,CG2,H,HA,HB,HG1,HG21,HG22,HG23;
(ILE:37003) N,CA,C,O,CB,CG1,CG2,CD1,
                    H,HA,HB,HG12,HG13,HG21,HG22,HG23,HD11,HD12,HD13;
(LEU:60276) N,CA,C,O,CB,CG,CD1,CD2,
                      H,HA,HB2,HB3,HG,HD11,HD12,HD13,HD21,HD22,HD23;
```

```
(ASP:40010) N,CA,C,O,CB,CG,OD1,OD2,H,HA,HB2,HB3;
(ASN:29276) N,CA,C,O,CB,CG,OD1,ND2,H,HA,HB2,HB3,HD21,HD22;
(GLU:53111) N,CA,C,O,CB,CG,CD,OE1,OE2,H,HA,HB2,HB3,HG2,HG3;
(GLN:29973) N,CA,C,O,CB,CG,CD,OE1,NE2,H,HA,HB2,HB3,HG2,HG3,HE21,HE22;
(MET:14570) N,CA,C,O,CB,CG,SD,CE,H,HA,HB2,HB3,HG2,HG3,HE1,HE2,HE3;
(LYS:50343) N,CA,C,O,CB,CG,CD,CE,NZ,
                      H,HA,HB2,HB3,HG2,HG3,HD2,HD3,HE2,HE3,HZ1,HZ2,HZ3;
(ARG:35509) N,CA,C,O,CB,CG,CD,NE,CZ,NH1,NH2,
                    H,HA,HB2,HB3,HG2,HG3,HD2,HD3,HE,HH11,HH12,HH21,HH22;
(HIS: 8040) N,CA,C,O,CB,CG,ND1,CD2,CE1,NE2,H,HA,HB2,HB3,HD1,HD2,HE1;
(PHE:26312) N,CA,C,O,CB,CG,CD1,CD2,CE1,CE2,CZ,H,HA,HB2,HB3,
                                                    HD1,HD2,HE1,HE2,HZ;
(TYR:22457) N,CA,C,O,CB,CG,CD1,CD2,CE1,CE2,CZ,
                                      OH,H,HA,HB2,HB3,HD1,HD2,HE1,HE2,HH;
(TRP: 8998) N,CA,C,O,CB,CG,CD1,CD2,NE1,CE2,CE3,CZ2,CZ3,CH2,
                                    H,HA,HB2,HB3,HD1,HE1,HE3,HZ2,HZ3,HH2.
```

Замечание 1.2. Отметим, что все выбранные типы атомных составов отвечают внутренним аминокислотам, что вполне естественно, так как концевые должны встречаться намного реже.

Отметим также, что "популярности" некоторых типов аминокислот бывают вполне сравнимы. Так, количества появлений типов изображенных на рис. 5 внутренних цистеинов отличаются в нашем эксперименте менее чем в 2 раза (тип, расположенный слева, встречается чаще). Приведем, для сравнения, их последовательности и число встреч:

```
11784, {N,CA,C,O,CB,SG,H,HA,HB2,HB3},
 6254, {N,CA,C,O,CB,SG,H,HA,HB2,HB3,HG}.
```

Замечание 1.3. Легко видеть, что атом H (из пептидного остова) присутствует не во всех отобранных атомных составах аминокислот. А именно, такой аминокислотой является пролин PRO, рис. 3. И это вполне естественно, в силу наличия связи CD→N, которую мы обсуждали выше (на водород уже "не остается места").

Тем не менее, мы решили более внимательно посмотреть на отброшенные атомные составы пролина: может, в некоторых из них все-таки присутствует водород? И мы нашли такие составы:

```
10, {N,CA,C,O,CB,CG,CD,H,H3,HA,HB2,HB3,HG2,HG3,HD2,HD3},
 1, {N,CA,C,O,H,HA,HB2},
 1, {N,CA,C,O,CB,CG,CD,H3,H,HA,HB2,HB3,HG2,HG3,HD2,HD3},
 1, {N,CA,C,O,CB,CG,CD,H,HA,HB2,HB3,HG2,HG3,HD2,HD3},
 1, {N,CA,C,O,CB,CG,CD,HA,HB2,HB3,HG2,HG3,HD2,H}.
```

Здесь число в каждой записи указывает на количество вхождений аминокислот идущего далее атомного состава. Заметим, что таких вхождение совсем мало, а именно 14, против 32529 самых популярных.

В заключение этого этапа, мы прошли по всем pdb-файлам и выделили те, в которых все внутренние аминокислоты, встречающиеся в первой модели, — стандартные. Наша база сократилась примерно в 3 раза (теперь она содержит 2892 файла). После этого мы преобразовали каждый файл в файл "короткого формата", собрав там

именно ту информацию, которая нам нужна для текущих исследований. Теперь в выбранных файлах у нас лежат списки вида

$$
\Big\{
\Big\{\text{ИмяПервойВнутреннейАминокислоты},
\quad \{\text{КоординатыПервогоАтома}, \text{КоординатыВторогоАтома}, \ldots\}\Big\},
\ldots
\Big\{\text{ИмяПоследнейВнутреннейАминокислоты},
\quad \{\text{КоординатыПервогоАтома}, \text{КоординатыВторогоАтома}, \ldots\}\Big\},
\Big\}
$$

Таким образом, если начальная база полипептидов, полученных методом NMR, занимала больше 18 гигабайт, то нынешняя занимает около 300 мегабайт. Конечно же, функции, необходимые для наших дальнейших исследований, с файлами новой базы работают на порядок быстрее.

2. Метрический анализ PDB

Получив наконец одинаковый атомный состав внутренних аминокислот в изучаемых полипептидах, мы вернулись к задаче сравнения геометрии одноименных аминокислот. Для этого мы сначала решили понять, сколь сильно могут варьироваться длины ковалентных связей, и как могут меняться углы между связями.

2.1. Оценка разброса длин ковалентных связей в аминокислотах

Для оценки отклонений длин ковалентных связей в данном наборе полипептидов мы решили поступить так:

(1) выбрать из всех полипептидов внутренние аминокислоты данного типа, например, все глицины;

(2) вычислить средние значения длины каждой ковалентной связи по всем выбранным аминокислотам;

(3) для каждого типа аминокислот (например, для глицинов) и каждой ковалентной связи в аминокислотах этого типа (например для связи N-H) вычислить максимальное, по всем аминокислотам этого типа, относительное отклонение от среднего значения (для удобства, это отклонение мы измеряли в процентах).

Для небольших фрагментов нашей базы эти отклонения были достаточно маленькие (несколько процентов). Однако, когда мы запустили программу на всю базу, то обнаружили огромные отклонения. Самое большое из них было у аланина (705.819% для связи N-H):

$$\frac{454.736}{\text{N}-\text{CA}}, \frac{362.668}{\text{CA}-\text{C}}, \frac{77.8317}{\text{C}-\text{O}}, \frac{244.119}{\text{CA}-\text{CB}}, \frac{328.258}{\text{CA}-\text{HA}}, \frac{105.323}{\text{CB}-\text{HB1}}, \frac{185.751}{\text{CB}-\text{HB2}}, \frac{152.231}{\text{CB}-\text{HB3}}, \frac{705.819}{\text{N}-\text{H}}$$

(числитель каждой дроби — максимальное процентное отклонение ковалентной связи, указанной в знаменателе, от среднего значения).

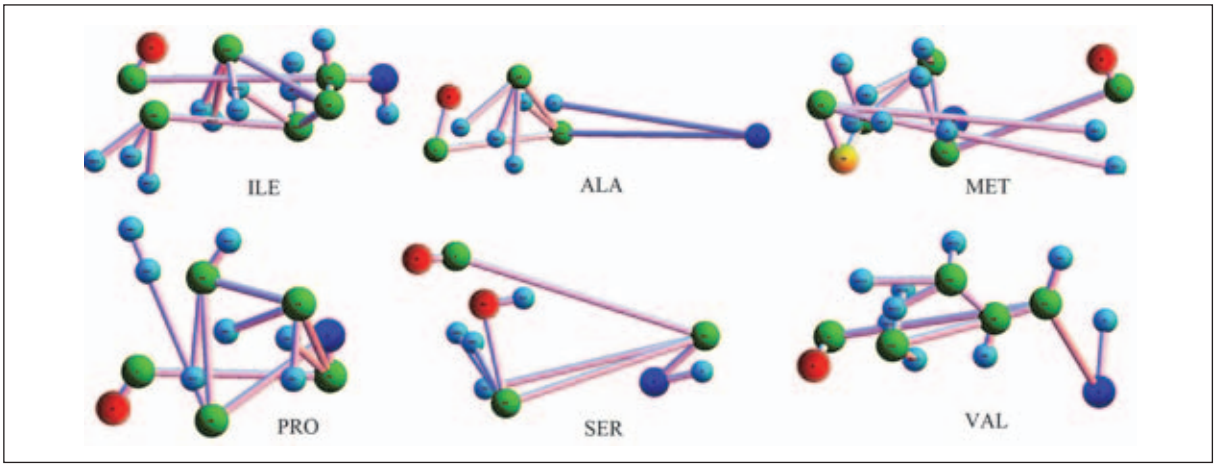

Рис. 6. Некоторые “аминокислоты” из файла 2PDE.pdb

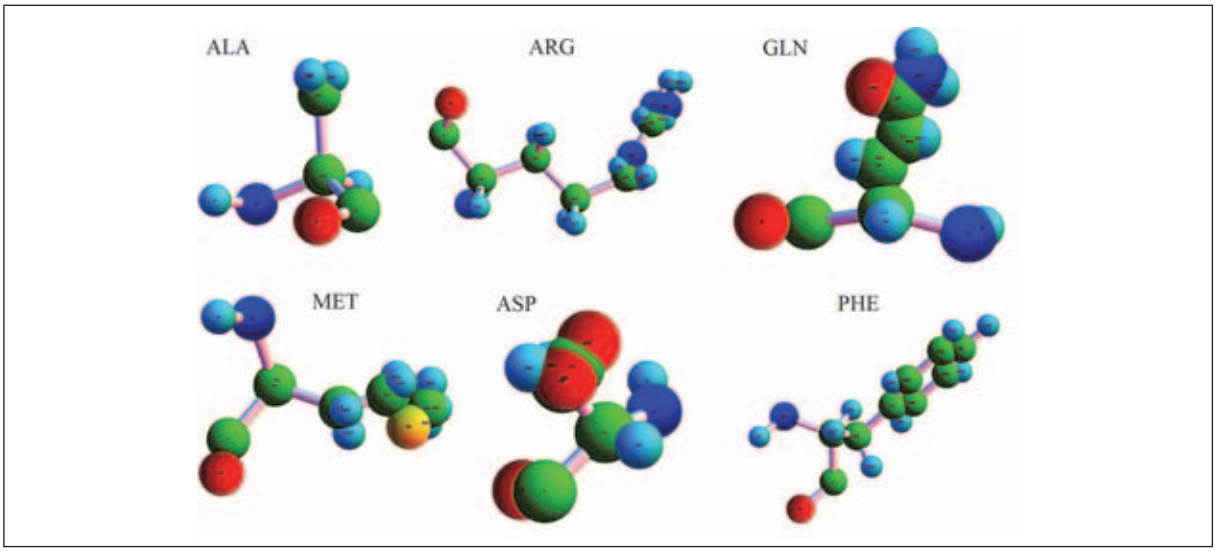

Рис. 7. Некоторые “аминокислоты” из файла 2HQ3.pdb

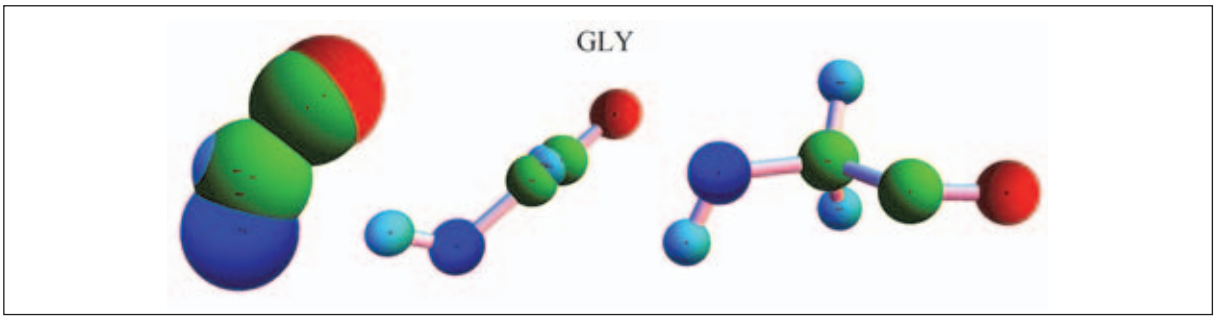

Рис. 8. Слева: два “глицина” из файла 2I2J.pdb. Справа “стандартный” глицин

Мы “отловили” файл, который привел к таким отклонениям. Им оказался 2PDE.pdb. Мы решили выяснить, как выглядят аминокислоты из этого файла. Результаты превзошли наши ожидания, см. рис. 6.

Отметим, что последние 8 аминокислот в этом файле выглядят стандартно.

Следующий “рекордсмен отклонений по аланину” — файл 2HQ3.pdb, см. рис. 7.

Впрочем, после выкидывания файла 2PDE.pdb, отклонения по глицину стали больше, чем по аланину. “Виновником” оказался файл 2I2J.pdb, см. рис. 8.

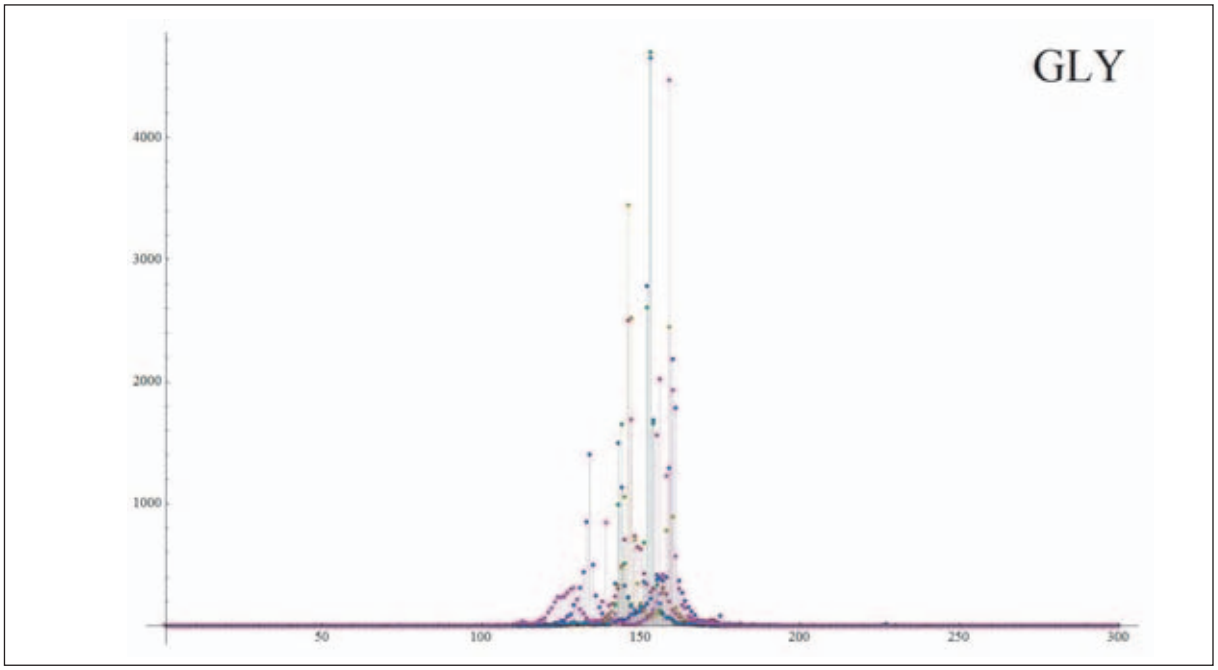

Рис. 9. Распределение количеств файлов с теми или иными максимальными отклонениями от средних значений длин ковалентных связей в глицине

Мы отобрали все файлы, в которых отклонение не только по глицину, но и по остальным аминокислотам, больше 15% — таких файлов оказалось не так много, а именно, 30 штук. Вот список их имен:

1AC0, 1BBA, 1COD, 1DOQ, 1GXX, 1J4M, 1K76, 1KWD, 1KWE, 1L0M, 1L8Z, 1MZI, 1OT4, 1PGY, 1XEE, 1Y7J, 2CYK, 2EVZ, 2HQ3, 2I2J, 2JMM, 2K1W, 2KZG, 2L9V, 2NR1, 2PDE, 3ZGK, 3ZOB, 3ZPK, 4BXU.

Выбросив из базы эти файлы, мы получили максимальное отклонение равным 14.6% (у валина):

$$\overline{\overset{8.89612}{\mathrm{N}\to\mathrm{CA}}},\ \overline{\overset{6.43247}{\mathrm{CA}\to\mathrm{C}}},\ \overline{\overset{4.27047}{\mathrm{C}\to\mathrm{O}}},\ \overline{\overset{8.92551}{\mathrm{CA}\to\mathrm{CB}}},\ \overline{\overset{5.34567}{\mathrm{CB}\to\mathrm{CG1}}},\ \overline{\overset{7.08376}{\mathrm{CB}\to\mathrm{CG2}}},\ \overline{\overset{14.6298}{\mathrm{CA}\to\mathrm{HA}}},$$
$$\overline{\overset{9.57366}{\mathrm{CB}\to\mathrm{HB}}},\ \overline{\overset{7.72194}{\mathrm{CG1}\to\mathrm{HG11}}},\ \overline{\overset{7.7138}{\mathrm{CG1}\to\mathrm{HG12}}},\ \overline{\overset{7.71296}{\mathrm{CG1}\to\mathrm{HG13}}},\ \overline{\overset{7.70222}{\mathrm{CG2}\to\mathrm{HG21}}},$$
$$\overline{\overset{7.69787}{\mathrm{CG2}\to\mathrm{HG22}}},\ \overline{\overset{7.7378}{\mathrm{CG2}\to\mathrm{HG23}}},\ \overline{\overset{10.1827}{\mathrm{N}\to\mathrm{H}}}.$$

Интересно, как распределено количество файлов с теми или иными отклонениями? Чтобы это выяснить, мы разбили отрезок от -15 до $+15$ на малые отрезки длины 0.1 и вычислили, сколько файлов имеют максимальные отклонения, попадающие на тот или иной из малых отрезков (в Mathematica это удобно сделать с помощью функции BinCounts). Результаты мы изобразили на графике (по абсциссе отложены номера малых отрезков, по ординате — количество файлов с максимальными отклонениями, попавшими на соответствующий отрезок). Графики оказались очень похожи друг на друга. На рис. 9 приведен пример для глицина.

Непосредственное рассмотрение всех 20 графиков позволило сделать вывод, что для основной массы отклонения не превышают 5%. Мы решили выяснить, сколько

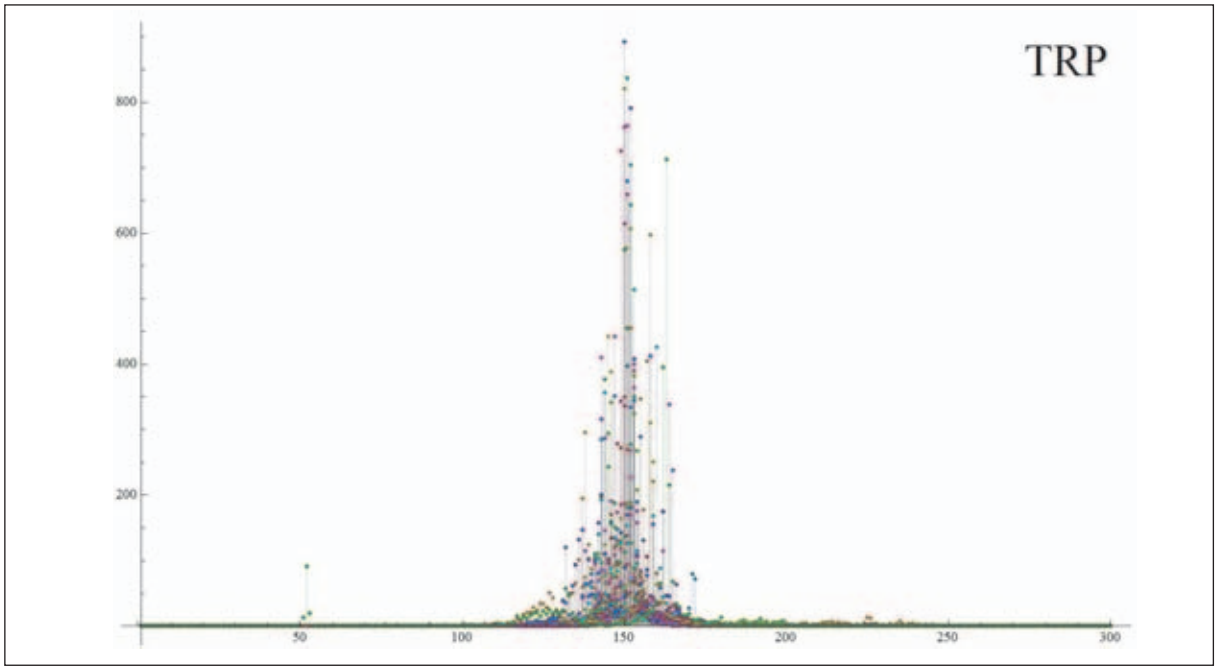

Рис. 10. Распределение количеств файлов с теми или иными максимальными отклонениями от средних значений длин ковалентных связей в триптофане

файлов для каждой аминокислоты не укладывается в 5%-ый интервал. Ниже приведен ответ.

GLY : 28; ALA : 26; SER : 47; CYS : 10; PRO : 157; VAL : 37; THR : 43;
ILE : 31; LEU : 38; ASP : 29; ASN : 33; GLU : 41; GLN : 30; MET : 15;
LYS : 53; ARG : 48; HIS : 122; PHE : 26; TYR : 42; TRP : 183.

Как видно, аминокислоты разбиваются на две группы: 17 штук, у которых "большие" отклонения присутствуют в 10–53 полипептидах, и 3 штуки, у которых отклонения имеются в 122–183 полипептидах (HIS, Pro, TRP). На рис. 10 приведен пример для триптофана.

Отметим, что в приведенной выше таблице количеств файлов с "большими отклонениями" общее количество, а именно, 1039, больше, чем реальное число файлов (некоторые файлы относятся одновременно к нескольким типам аминокислот). Это реальное число файлов равно 442. Так как к настоящему моменту в нашей базе имелось 2862 файлов, число файлов с "большими отклонениями" оказалось достаточно значимым. В следующей таблице приводятся количества файлов с отклонениями большими, чем n%, для $n = 6, \ldots, 12$:

6% : 371; 7% : 279; 8% : 162; 9% : 128; 10% : 20; 11% : 8; 12% : 6.

Так как мы для начала интересуемся типичными метрическими свойствами полипептидов, мы решили поместить в особый список 20 файлов, соответствующих 10%, и исключить их из нашего основного списка. Тем самым, мы удалили файлы с именами

1ALE, 1ALF, 1C7V, 1E0Q, 1E2B, 1IFY, 1LVQ, 1MEQ, 1ODP, 1ODR, 1OEF, 1OEG, 1OPP, 1PFT, 1SOL, 1SVR, 1TXA, 1ZRP, 2BAU, 2KOX

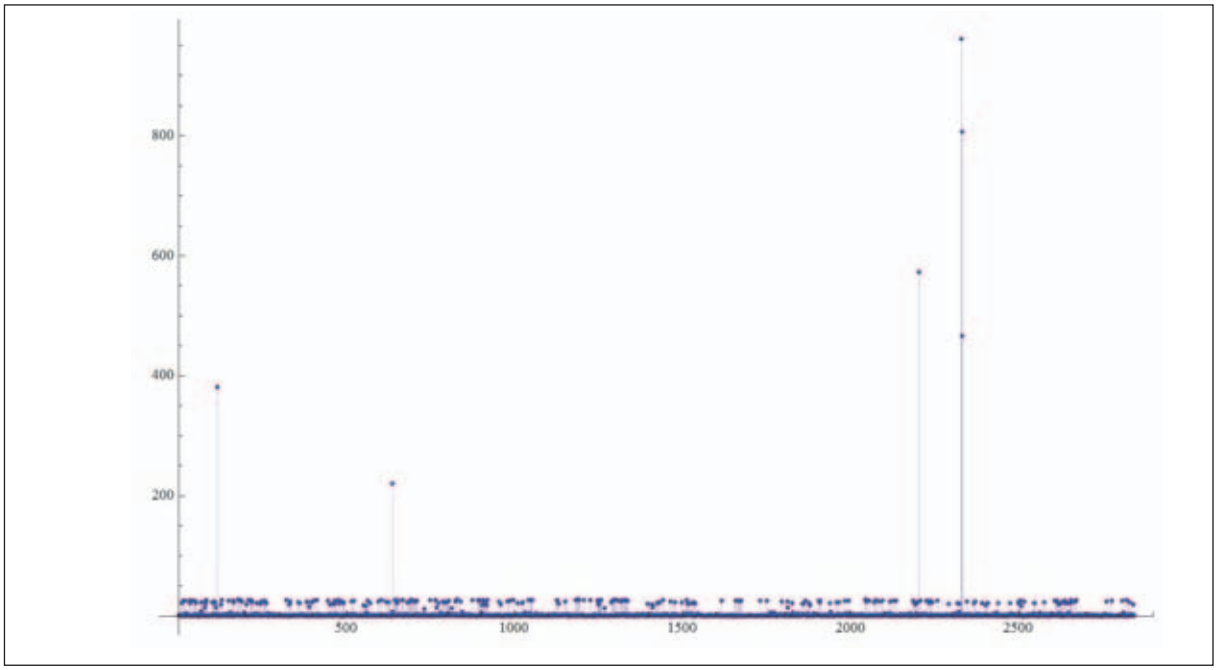

Рис. 11. Максимальные относительные отклонения от среднего расстояний между последовательными α-углеродами в отобранных выше полипептидах

и получили базу данных из 2842 файлов, в которой отклонение длин ковалентных связей в аминокислотах не превышает 10%.

Следующий шаг — изучить отклонение от среднего расстояний между альфа-углеродами последовательных внутренних аминокислот в отобранных выше полипептидах. *В дальнейшем мы будем обозначать альфа-углероды в двух последовательных аминокислотах через* CA-CA.

2.2. Оценка разброса длин расстояний между последовательными альфа-углеродами (начало)

Для каждого полипептида мы вычислили максимальное относительное отклонение расстояний CA-CA и вывели график этих отклонений, см. рис. 11. По абсциссе отложен номер полипептида, а по ординате — максимальное отклонение.

Обратите внимание на небольшое количество "выбросов". Мы нашли соответствующие полипептиды. Вот их список: 1BH7, 1JJS, 2KSE, 2L6F, 2L6G, 2L6H. Причина оказалась простой: некоторые аминокислоты не были распознаны в эксперименте. Впрочем, в этих файлах имеются соответствующие указания. Во-первых, в комментариях можно найти надпись следующего типа (для примера, рассмотрим 1BH7.pdb):

```
REMARK 465 MISSING RESIDUES
REMARK 465 THE FOLLOWING RESIDUES WERE NOT LOCATED IN THE
REMARK 465 EXPERIMENT. (RES=RESIDUE NAME; C=CHAIN IDENTIFIER;
REMARK 465 SSSEQ=SEQUENCE NUMBER; I=INSERTION CODE.)
REMARK 465     RES C SSSEQI
REMARK 465     LEU A    17
REMARK 465     VAL A    28
```

Во-вторых, в списке "ATOM..." 6-ая позиция, отвечающая за номер аминокислоты, также демонстрирует пропущенные аминокислоты. Скажем, в том же файле после номера 16 сразу следует номер 18.

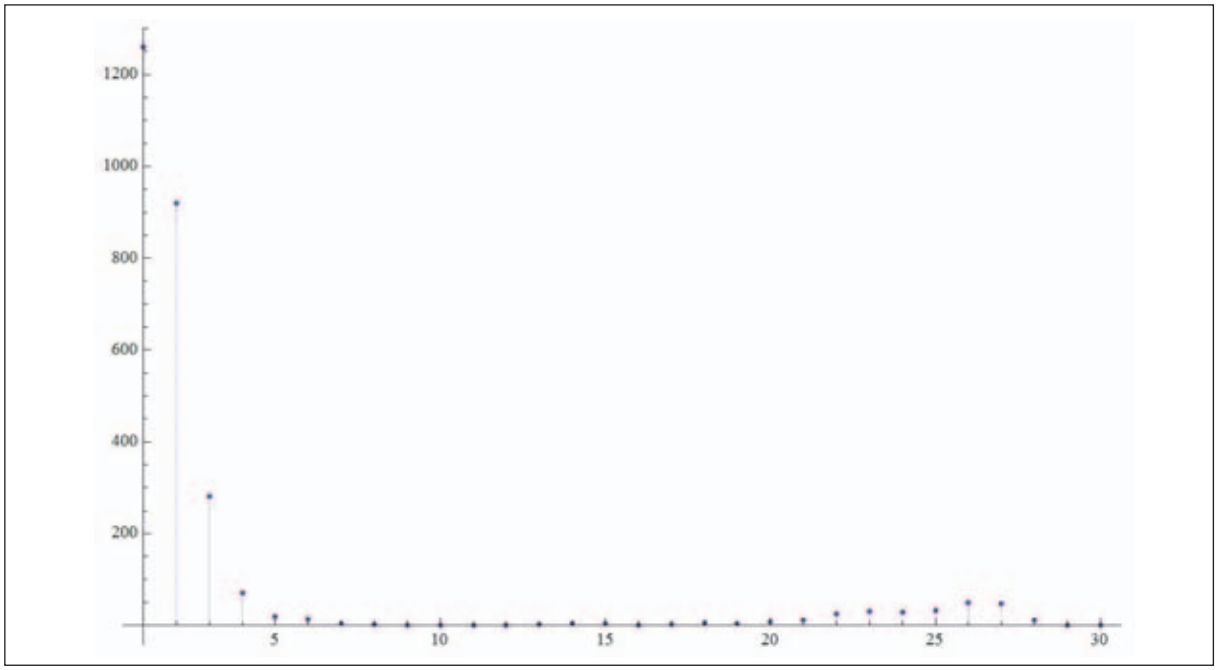

Рис. 12. Количества файлов (ордината) с теми или иными максимальными отклонениями (абсцисса) в CA-CA связях

Выбросив из нашей базы эти 6 файлов, мы получили максимальные относительные отклонения не более 30%. На рис. 12 показано, как распределены количества файлов (ордината) с теми или иными максимальными отклонениями (абсцисса): 1 по абсциссе соответствует изменению величины отклонения на 1%.

А вот соответствующие количества:

$$1260, 920, 282, 70, 20, 13, 4, 2, 1, 0, 0, 0, 2, 3, 3, 1, 2, 5, 3, 8, 12, 24, 30, 28, 33, 49, 48, 12, 1, 0$$

Замечание 2.1. Интересно, что хотя основная масса файлов имеет отклонения не более 5–6%, тем не менее, имеется второе "скопление" в окрестности 25%. Ниже мы объясним, в чем дело.

Конечно, одной из очевидных причин наличия "скопления" мог бы оказаться разброс длин пептидных связей C-N.

2.3. Оценка разброса длин пептидных связей

Мы изучили отклонения в пептидных связях C-N (между последовательными аминокислотами). Эти отклонения оказались не превосходящими 7.8%. На рис. 13 приведено распределение количеств полипептидов с максимальными отклонениями, отложенными по оси абсцисс (шаг по абсциссе равен отклонению на 0.2%).

Таким образом, "скопления" в разбросе расстояний CA-CA обусловлены не этим.

2.4. Закон плоскости

В биохимии традиционно выделяются следующие шестерки атомов пептидного остова: атомы CA, C, O для $(i-1)$-ой аминокислоты и CA, N, H для i-ой аминокислоты. Такие шестерки называются *пептидными группами*. Считается, что каждая

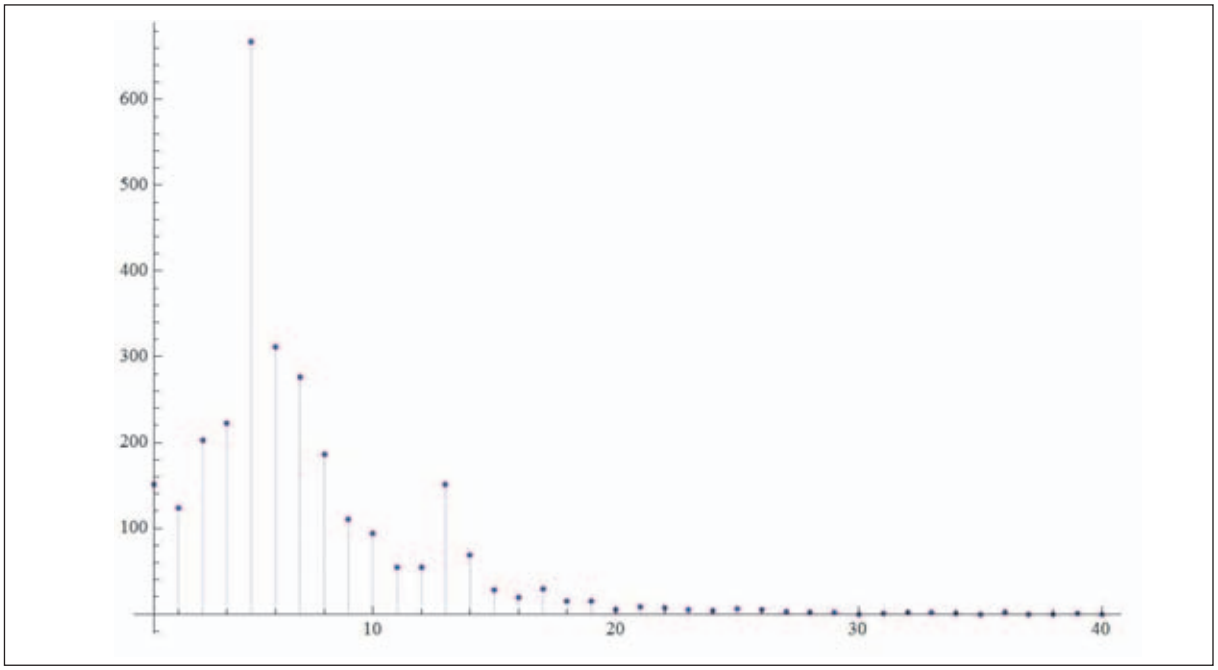

Рис. 13. Количества файлов (ордината) с теми или иными отклонениями (абсцисса) в C-N связях

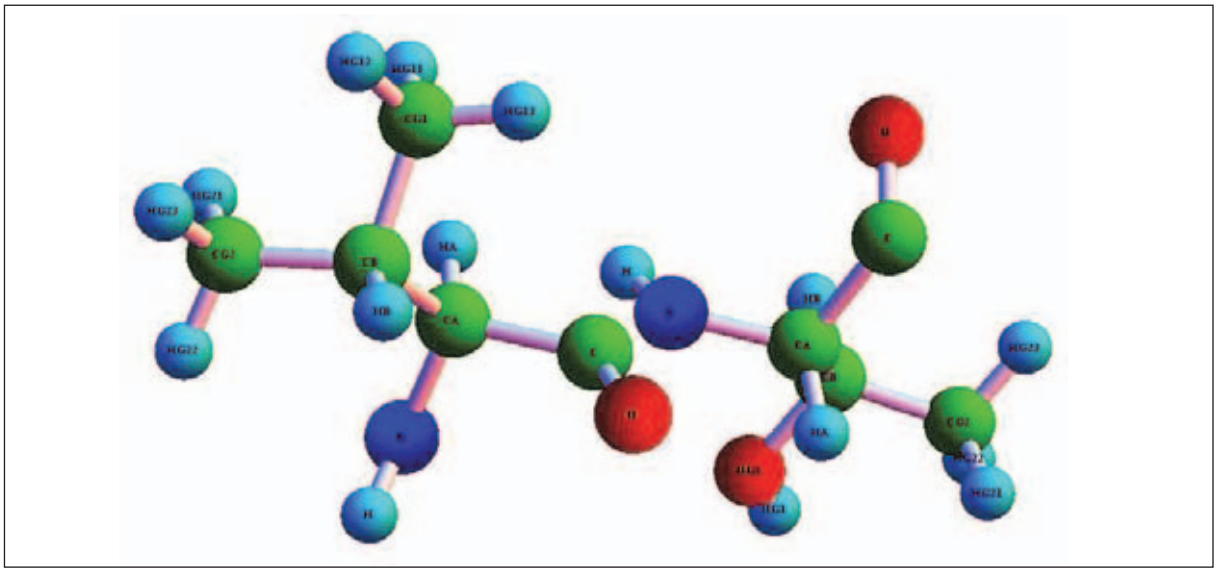

Рис. 14. Иллюстрация утверждения о том, что атомы CA, C, O для (i – 1)-ой аминокислоты и CA, N, H для i-ой аминокислоты лежат в одной плоскости

пептидная группа лежит в одной плоскости. Это утверждение будем называть *законом плоскости*. Естественно, нарушение этого закона также влияет на величину разброса расстояний CA-CA.

Проверим этот закон, взяв случайный полипептид, скажем, 1CKR, и в нем — случайную пару последовательных аминокислот, например, с номерами 91 (VAL) и 92 (THR), см. рис. 14.

Отчетливо видно, что в этом случае пептидная группа образует почти плоскую конфигурацию, т.е. закон плоскости имеет место.

Нам повезло, что второй выбранной аминокислотой не оказался пролин PRO, ведь у пролина нет атома H. Давайте посмотрим, что происходит, если вторая аминокислота — пролин, рис. 15.

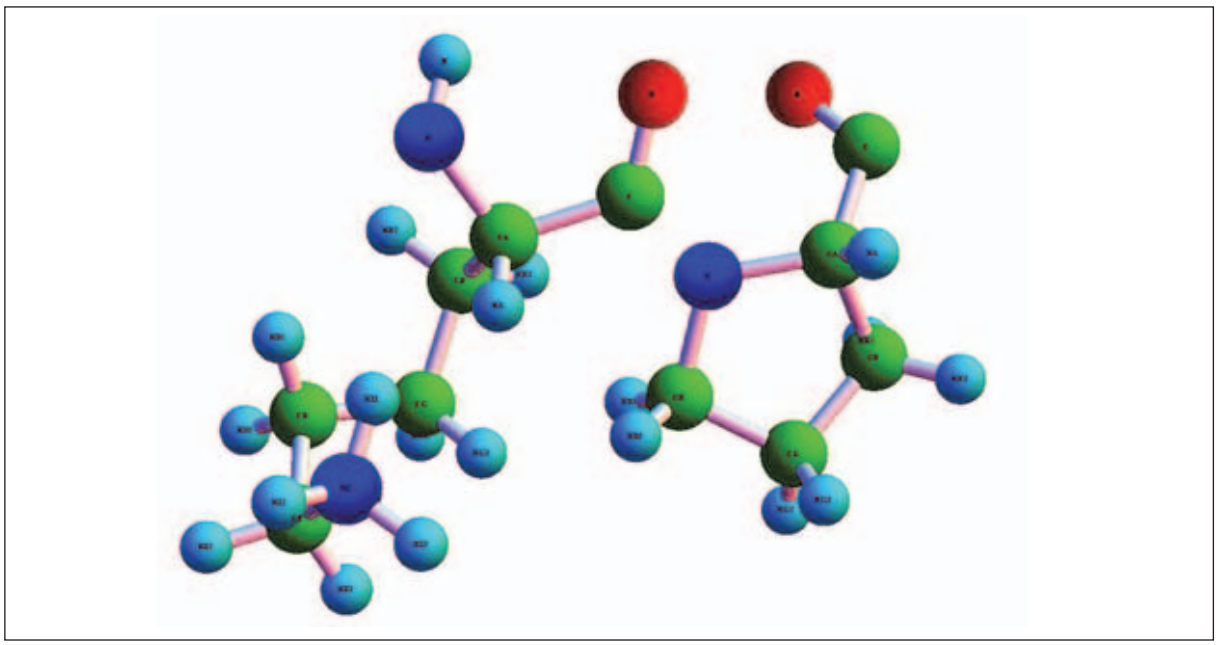

Рис. 15. Последовательные аминокислоты, вторая — пролин

Из рисунка видно, как обобщить закон плоскости и на этот случай: нужно взять вместо атома H атом CD. *В дальнейшем, говоря о законе плоскости, мы будем иметь в виду именно это обобщение.*

Чтобы проверить действие закона плоскости во всех остальных случаях, мы написали программу, которая для каждой пары последовательных аминокислот строит выпуклую оболочку атомов пептидной группы и вычисляет объем этой оболочки. На первых попавшихся полипептидах максимальное значение объема было равно примерно 0.5. Но на всей совокупности полипептидов максимальное значение оказалось равным 4.74761. Мы нашли соответствующие полипептид и пару его последовательных аминокислот. Ими оказались полипептид 1PBA и его 38 (LYS) и 39 (PRO) аминокислоты. И вот что мы увидели, рис. 16.

Приведенный на рис. 16 пример является особым в том смысле, что вторая аминокислота в нем — пролин (напомним, именно такими конфигурациями мы расширили закон плоскости). Отметим, что, вообще говоря, для таких особых конфигураций значения их объемов могут измерять степень неплоскости конфигурации пептидной группы по-другому по отношению к объемам для неособых конфигураций. И это действительно имеет место: объемы для особых конфигураций с аналогичными отклонениями, что и у неособых, меньше. И этот факт хорошо виден на следующем примере неособой конфигурации, на которой достигается максимум объема (4.4387) в классе неособых конфигураций, рис. 17.

Чтобы оценить, как связаны объемы со степенью плоскости, мы рассмотрели выпуклые оболочки пептидных групп с разными объемами, выбрали некоторые из них и постарались сориентировать так, чтобы проекции на фронтальную плоскость имели минимальную высоту (это высота и оценивает то, сколь конфигурация плоская: чем меньше высота, тем более плоская конфигурация). На рис. 18 показаны примеры (для неособых конфигураций) и подписаны объемы. Видно, что даже для объема порядка 1 конфигурация совсем плоской не выглядит.

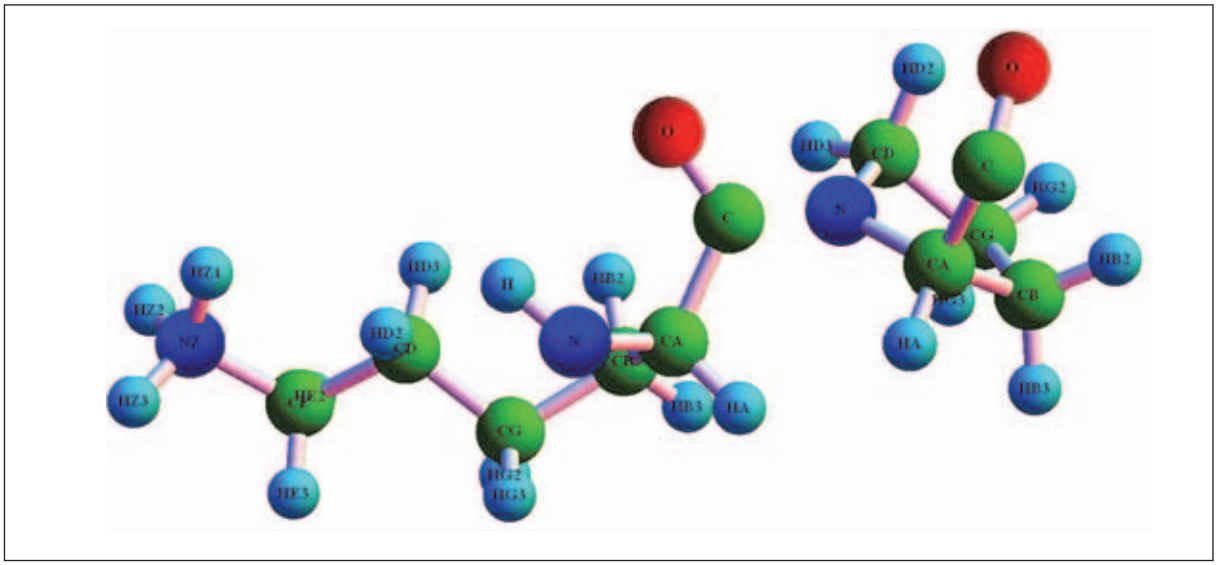

Рис. 16. “Контрпример” к утверждению о том, что атомы CA, C, O для (i – 1)-ой аминокислоты и N, H, CA для i-ой аминокислоты лежат в одной плоскости (“особая” конфигурация)

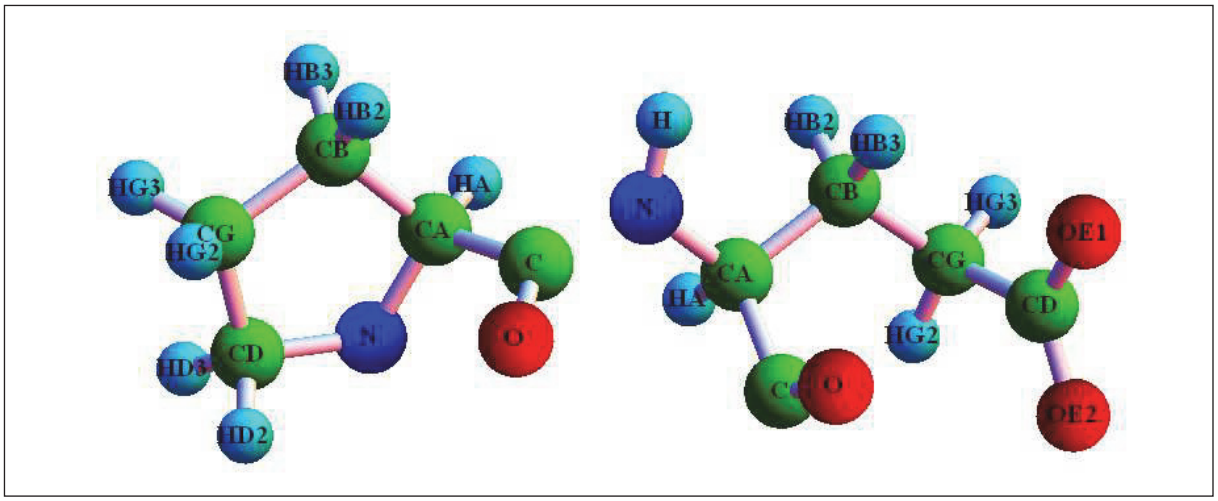

Рис. 17. “Контрпример” к утверждению о том, что атомы CA, C, O для (i – 1)-ой аминокислоты и CA, N, H для i-ой аминокислоты лежат в одной плоскости (“неособая” конфигурация)

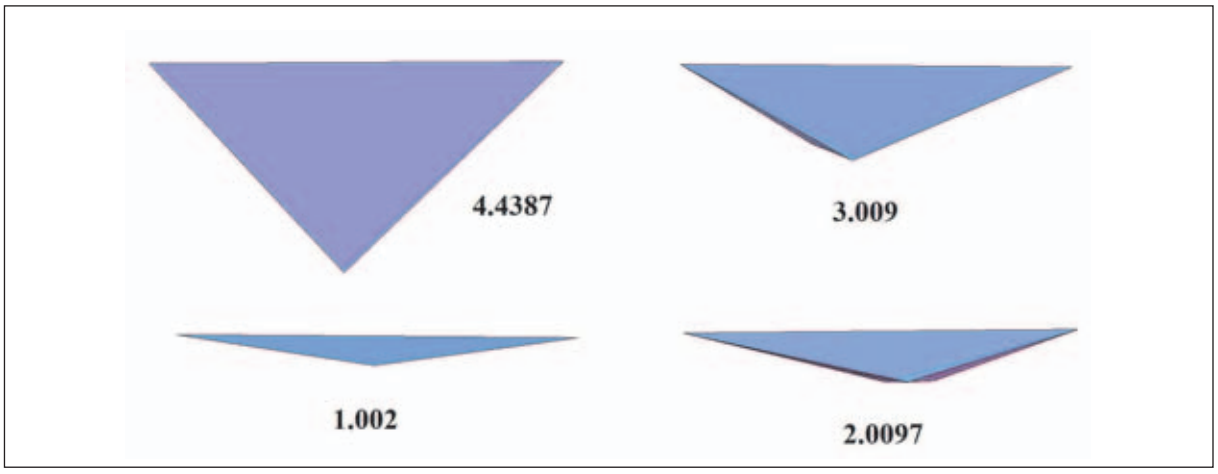

Рис. 18. Проекции выпуклых оболочек на фронтальную плоскость, имеющие наименьшие высоты

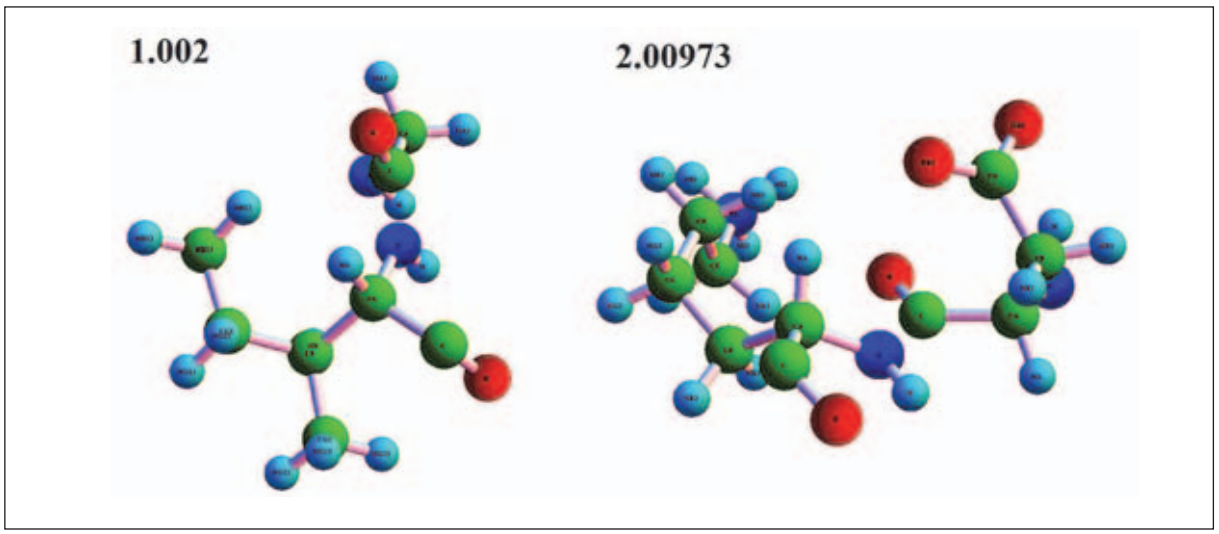

Рис. 19. Пары аминокислот, соответствующие объемам ~ 1 и ~ 2 из рис. 18

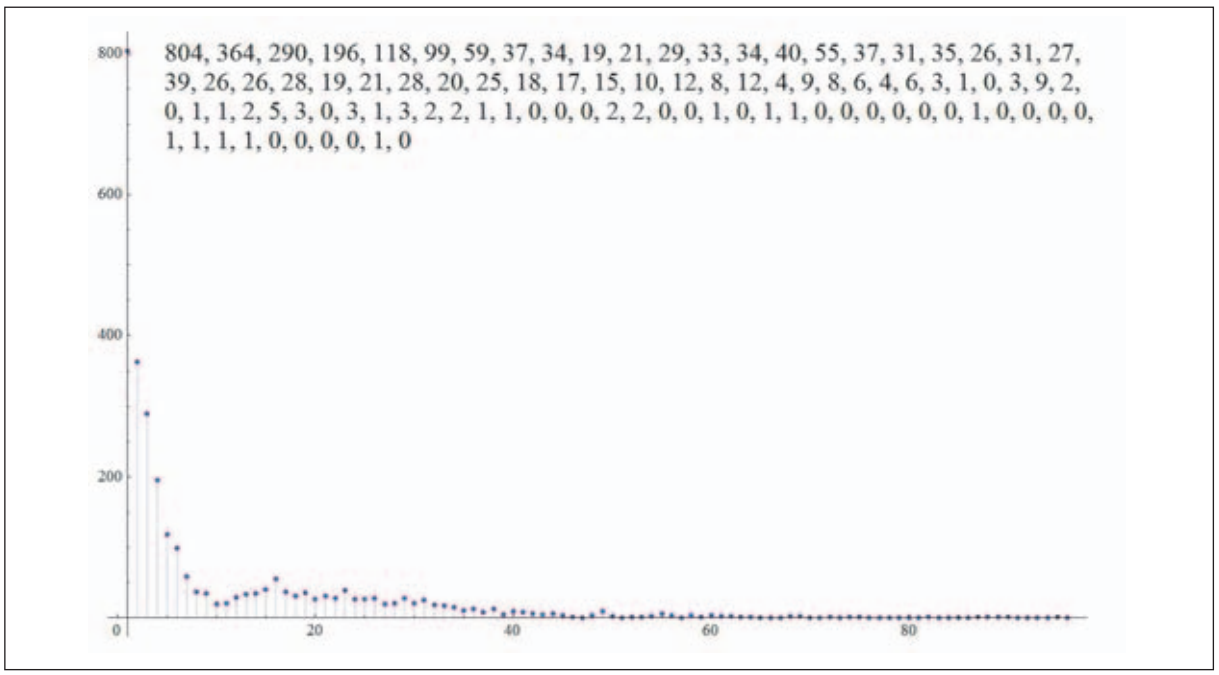

Рис. 20. Распределение количеств полипептидов (ордината) по значениям объемов выпуклых оболочек пептидных групп (абсцисса)

На рис. 19 показано, как выглядят соответствующие пары аминокислот для объемов ~ 1 и ~ 2 из рис. 18.

Мы решили посчитать распределение количеств полипептидов, в которых максимальные отклонения объемов лежат в том или ином интервале. Для этого мы разбили отрезок $[0, 4.8]$ на малые отрезки длины 0.05, и для каждого из полученных малых отрезков посчитали, у скольких полипептидов максимальные объемы выпуклых оболочек рассматриваемых пептидных групп лежат на этом отрезке. На рис. 20 показан соответствующий график.

Замечание 2.2. Несложно вычислить, что максимальный объем больше 1 имеют 475 полипептидов (из 2836), т.е. 16.7%. Как мы уже отмечали выше, при таком объеме хорошо заметно, что конфигурация соответствующей шестерки неплоская. Таким

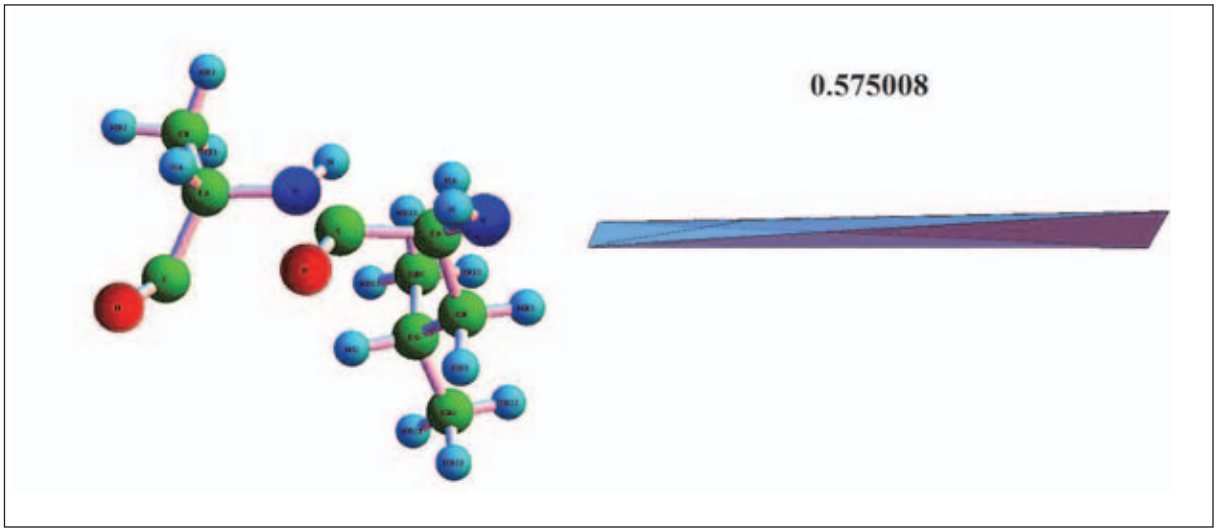

Рис. 21. Пара аминокислот, реализующая максимальный объем 0.575008 для 95% пар последовательных аминокислот

образом, в каждом из 16.7% полипептидов имеется аминокислота, в которой хорошо заметно отклонение от закона плоскости.

С другой стороны, мы рассмотрели все возможные последовательные пары аминокислот в имеющихся полипептидах. Их оказалось 190185 штук. Считая, что закономерность хорошая, если она выполняется для 95% случаев, мы получили следующий результат: количество пар последовательных аминокислот, в которых выпуклые оболочки шестерок имеют объем меньше 0.575, составляет 95% от общего количества пар последовательных аминокислот. На рис. 21 показан пример для максимального объема 0.575008.

Заметим, что здесь закон плоскости выполняется с хорошей точностью. Таким образом, можно считать, что этот закон прошел наш "тест на выживание". Впрочем, ряд полипептидов все-таки стоит исключить из рассмотрения. Если исходить из желания оставить около 95% полипептидов, то нужно исключить все те, у которых объемы превышают 1.67. Тогда останется 95.1%, а исключено будет 138 файлов.

2.5. Оценка разброса длин расстояний между последовательными альфа-углеродами (окончание)

Мы решили сделать следующий "радикальный" эксперимент: исключить все файлы, в которых максимальные объемы превышают 0.5 (в оставшихся файлах закон плоскости выполняется с большой точностью) и посмотреть, как это отразится на разбросе расстояний CA-CA. Исключаемых файлов оказалось 815 (остался 2021 файл). Мы вычислили разброс расстояний CA-CA по этим 2021 файлу. Результат — разброс оказался равным 28%, т.е. причина такого большого разброса устранена не была.

Думаем, многие уже давно догадались, что реальной причиной является наличие как *цис*, так и *транс* конфигураций у пептидных групп, см. рис. 22.

2.5.1. Цис и Транс конфигурации пептидных групп

Чтобы определить такие конфигурации формально, мы рассмотрели векторные произведения [N-CA, N-H] и [C-CA, C-O] для пептидной группы (в случае пролина

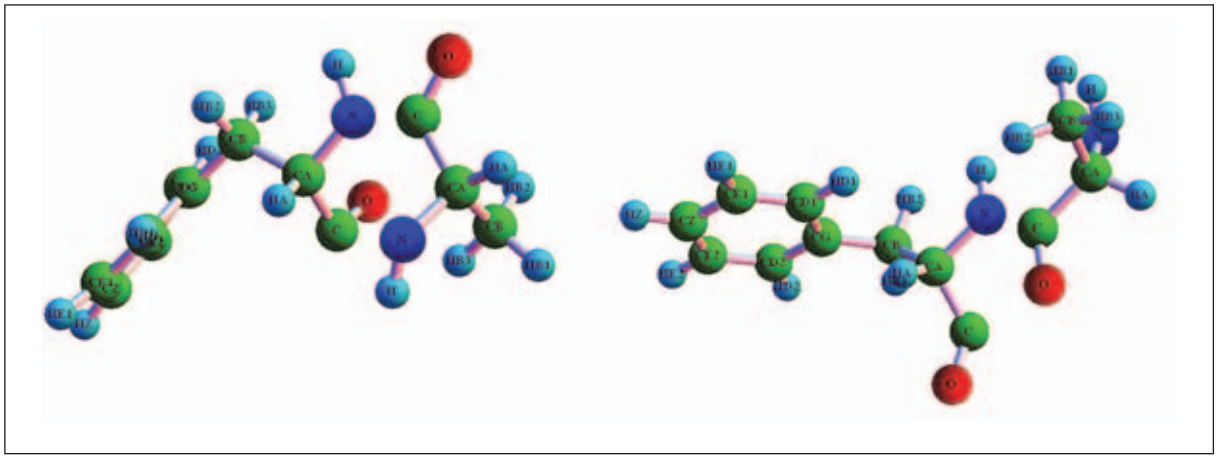

Рис. 22. Цис (слева) и транс (справа) конфигурации пептидных групп. В цис-конфигурации атомы CA находятся “с одной стороны”, а в транс-конфигурации — с разных

вместо H берем CD) и назвали такую шестерку находящейся в *транс конфигурации*, если угол между этими векторами острый (их скалярное произведение положительно). В противном случае, мы отнесли эту шестерку к *цис конфигурациям*.

Мы посчитали количество полипептидов, в которых присутствуют последовательные пары аминокислот, находящиеся в цис-конфигурации. Оказалось, что таких полипептидов относительно мало, а именно, 263. Сначала мы попробовали, не улучшая нашу базу отсеиванием пептидных групп с большими объемами выпуклых оболочек, просто выбросить эти 263 файла и проверить, какой разброс будет у расстояний CA-CA. Оказалось, что разброс стал лучше, но не намного, а именно, он стал равен 17.67%. Тогда мы решили также выбросить полипептиды с наиболее значительными отклонениями от закона плоскости, а именно, все полипептиды с объемами выпуклых оболочек пептидных групп, превышающими 3. Оказалось, что мы должны избавиться от следующих 19 полипептидов:

`1AW3, 1BLV, 1EGO, 1ESK, 1FDF, 1I1S, 1I6C, 1L6H, 1MEA, 1MED,`
`1MNY, 1OZO, 1PBA, 1W9R, 1YUS, 2ADX, 2H3J, 2ITH, 7HSC.`

Результаты нас порадовали: теперь разброс длин CA-CA оказался равен 7.4%. В результате наша база стала содержать 2563 файла.

Наконец, мы добрались до проверки разброса углов между ковалентными связями.

2.6. Оценка разброса углов в полипептидах

Мы очень надеялись, что процедура проверки углов между ковалентными связями будет “чисто формальной”: ведь у нас уже отобраны полипептиды, в аминокислотах которых разбросы длин этих связей достаточно маленькие. Конечно, с формальной точки зрения, отсюда ничего не следует, но все-таки мы существенно “почистили” базу, поэтому есть шанс, что все большие отклонения уже были выкинуты. Тем не менее, нас ждал сюрприз: максимальное отклонение от среднего значения углов оказалось равным 71.1204%. На рис. 23 показаны три примера, соответствующих трем самым большим значениям разброса углов (мы подписали название аминокислоты и содержащего ее полипептида).

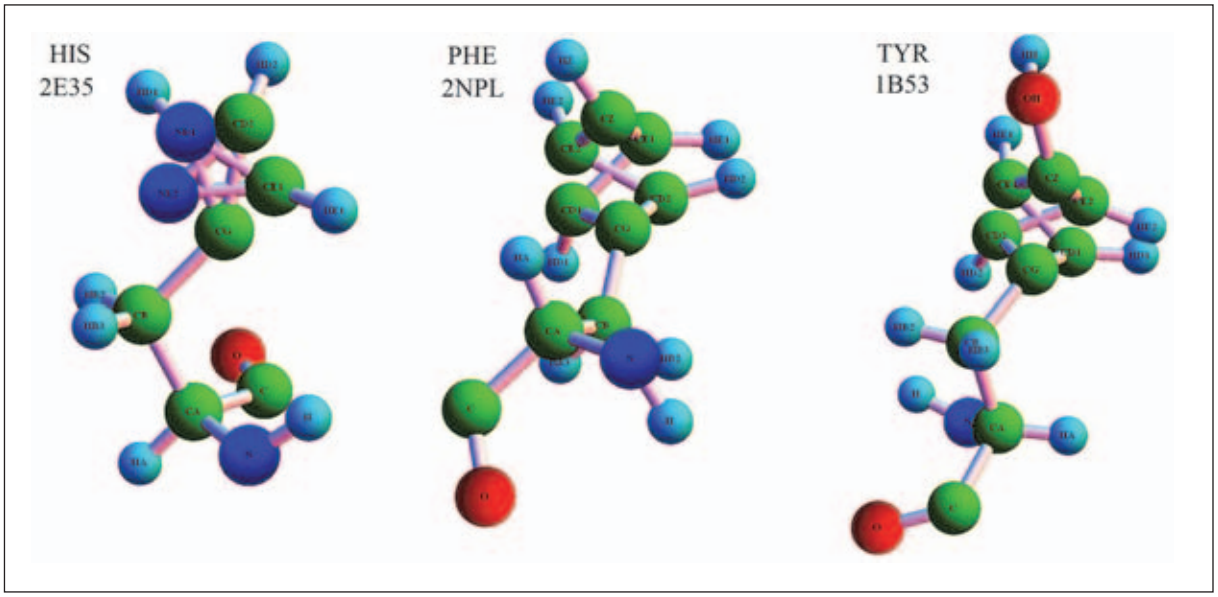

Рис. 23. Аминокислоты, в которых разброс углов принимает три самых больших значения

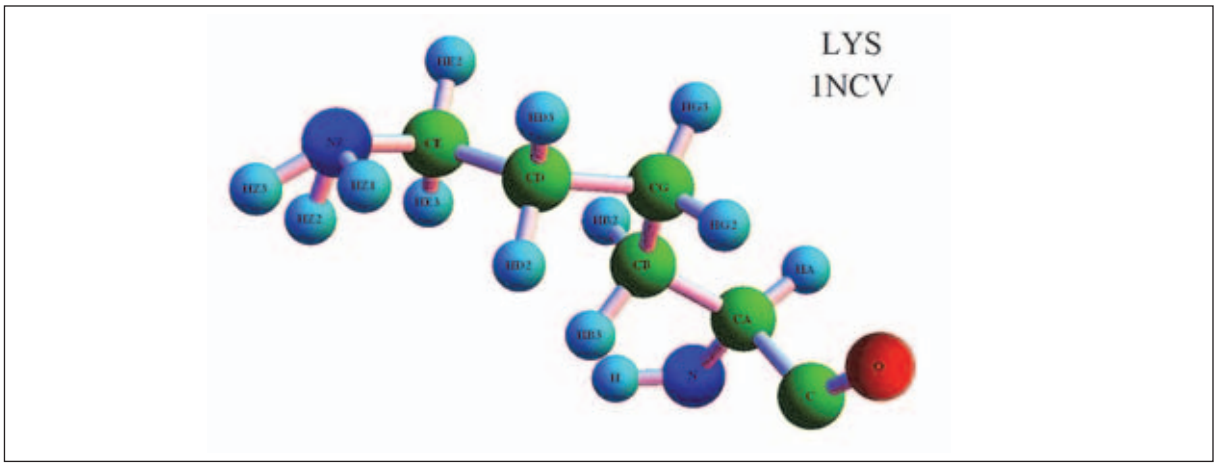

Рис. 24. Атомы, прикрепленные к NZ, лежат в одном полупространстве относительно некоторой плоскости, проходящей через NZ

Отметим, что в остальных аминокислотах, соответствующих меньшим разбросам, отклонения обусловлены существенно неравномерным распределение ковалентных связей для некоторых атомов валентности 4: соседи таких четырехвалентных атомов могут оказаться в одном полупространстве относительно некоторой плоскости, проведенной через этот атом, рис. 24.

Мы выяснили, что подобные сильные отклонения начинаются при разбросах величины около 26%. Всего же полипептидов с максимальным разбросом, превосходящим 26%, имеется 59 штук. Опять же, исходя из соображений “типичности”, мы исключили эти файлы из нашей базы, тем самым, оставив в базе 2532 файлов. Для полученной базы угловые отклонения в аминокислотах не превосходят 26%, метрические отклонения в аминокислотах не превосходят 10%, отклонения в расстояниях между последовательными альфа-углеродами не больше 7.5%. Кроме того, все пептидные группы образуют транс конфигурации, и почти все пептидные группы достаточно точно удовлетворяют закону плоскости. Дальнейшее исследование будем проводить именно на этой базе.

3. Подвижность аминокислот

Рассматривая различные картинки с изображениями одноименных аминокислот в разных местах одного полипептида или в разных полипептидах, мы обнаружили, что эти аминокислоты сильно отличаются друг от друга. Для сравнительного анализа мы решили изобразить одноименные аминокислоты, встреченные в разных местах полипептидов, на одной картинке. Для этого нам нужно было придумать, как совмещать различные, но одноименные кислоты. Опираясь на анализ метрических соотношений, мы ввели понятие *репера аминокислоты*, а именно, в качестве *начала координат* мы выбрали альфа-углерод CA, а в качестве концов базисных векторов — азот N, углерод C и бета-углерод CB — последний во всех случаях, кроме "вырожденной" аминокислоты глицина; в глицине вместо CB мы выбрали водород HA3.

Замечание 3.1. Отметим, что в последовательности атомов аминокислоты, принятой в формате PDB, атомы N, CA, C, CB имеют номера соответственно 1, 2, 3, 5; в глицине же на 5-ом месте стоит совсем "не тот" водород (выбрав автоматически его, мы наблюдали поразительные картинки); нужный же нам водород в глицине расположен на 7-ом месте.

Имея реперы, естественно выбрать в качестве совмещающего преобразования одно из следующих двух: (1) аффинное преобразование, переводящее репер в репер; (2) ортогональное преобразование, располагающее соответствующие атомы реперов наиболее близким образом. Вот пример, рис. 25.

Теперь хорошо видно, что прикрепленный к азоту N водород, а также прикрепленный к углероду C кислород могут вращаться вокруг оси соответствующего вектора репера. Также в случае аланина вращается и водород, прикрепленный к CB. Для серина ситуация более сложная: CB разветвляется на два водорода и кислород OG, к которому также крепится водород. Тройка, на которую разветвляется CB, может вращаться вокруг оси CA-CB; кроме того, прикрепленный к кислороду водород также может вращаться относительно оси CB-OG.

На рис. 26 и 27 приведены еще более сложные картинки.

Возможно, адекватное представление аминокислот, входящих в полипептид, состоит в рассмотрении соответствующих "облаков". Впрочем, водородные связи должны уменьшать *конфигурационное пространство* каждой аминокислоты. Было бы интересно разобраться, как выглядят эти конфигурационные пространства.

Замечание 3.2. При проведения описанного выше компьютерного эксперимента мы первоначально использовали преобразования, располагающие оптимальным образом атомы реперов, но сохраняющие при этом ориентацию этих реперов. В результате мы выяснили, что не *все реперы являются положительно ориентированными*. Мы решили в этом разобраться.

3.1. Ориентация аминокислот

Мы назвали аминокислоту *положительно ориентированной*, если ориентация ее репера совпадает со стандартной, иными словами, если определитель матрицы, строки которой — векторы репера, является положительным числом. В противном случае, назовем аминокислоту *отрицательно ориентированной*.

В последней нашей базе мы нашли все полипептиды, каждый из которых содержит хотя бы одну отрицательно ориентированную аминокислоту. Таких полипепти-

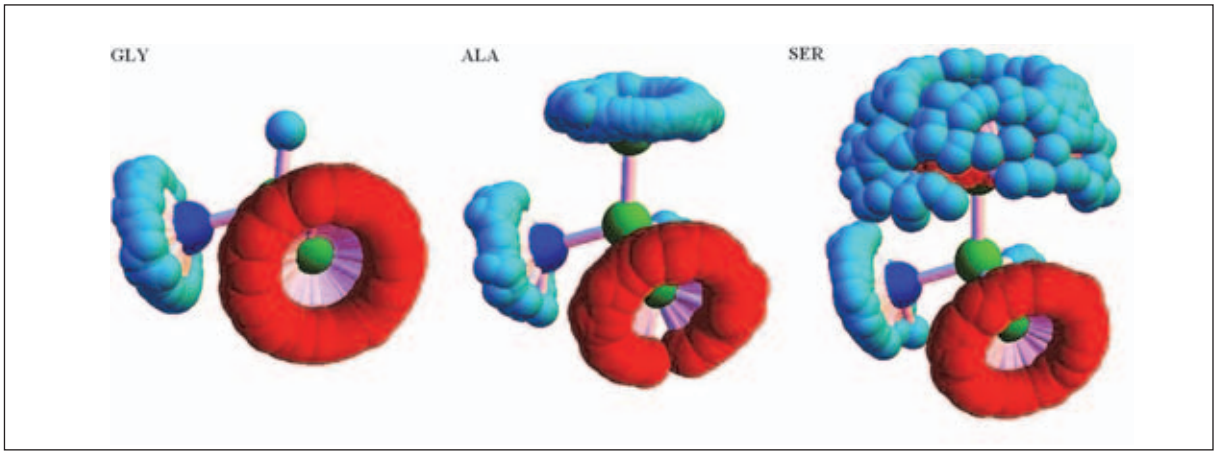

Рис. 25. Глицин, аланин, серин: совмещение реперов одноименных аминокислот

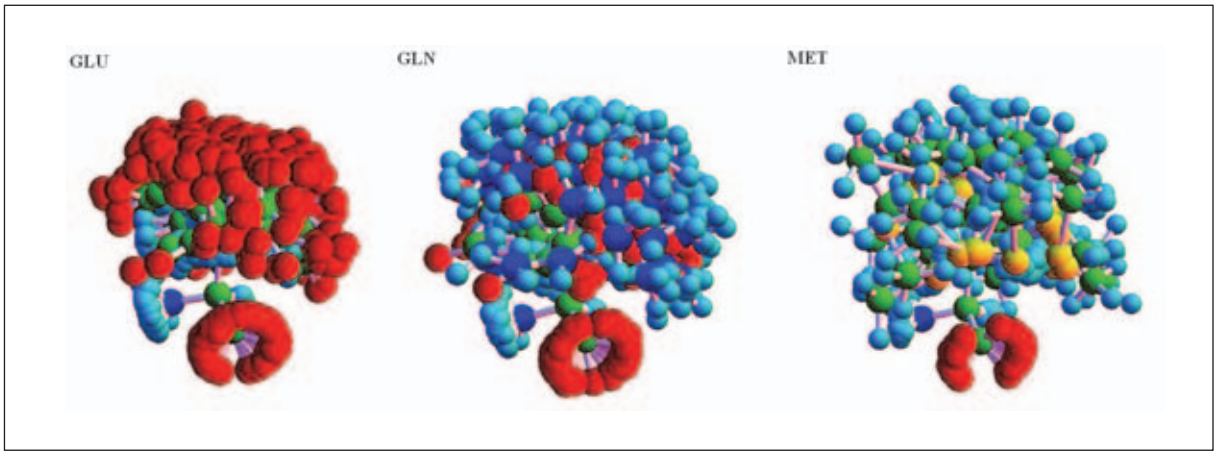

Рис. 26. Глутаминовая кислота, глутамин, метионин: совмещение реперов одноименных аминокислот

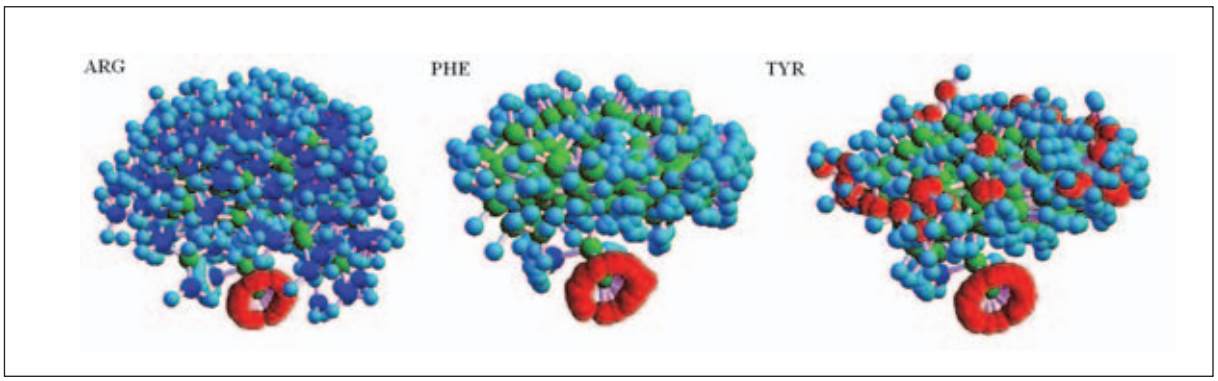

Рис. 27. Аргинин, фенилаланин, тирозин: совмещение реперов одноименных аминокислот

дов оказалось 140. Взяв первый попавшийся из них, мы обнаружили там глицин, изображенный на рис. 28 слева.

Впрочем, если поменять местами названия водородов HA2 и HA2, то левый глицин станет положительно ориентирован, см. рис. 29.

Отметим, что глицинов, у которых репер с водородом HA3 положительно ориентирован, намного больше, чем глицинов, у которых положительно ориентирован репер с водородом HA2.

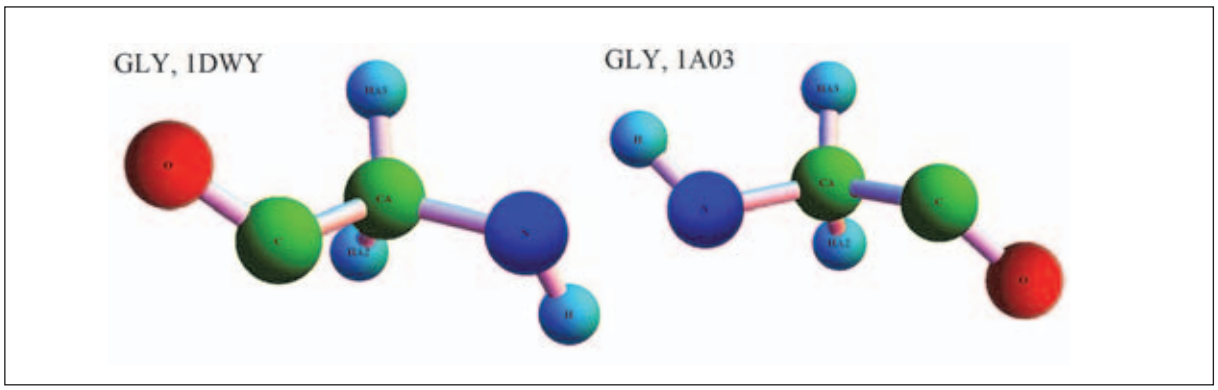

Рис. 28. Отрицательно ориентированный глицин (слева) и, для сравнения, положительно ориентированный глицин (справа)

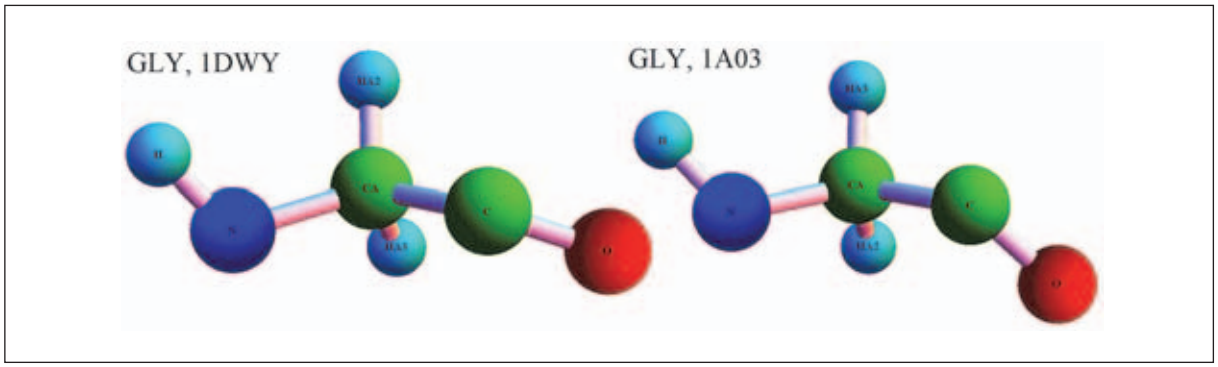

Рис. 29. Отрицательно ориентированный глицин (слева) и, для сравнения, положительно ориентированный глицин (справа)

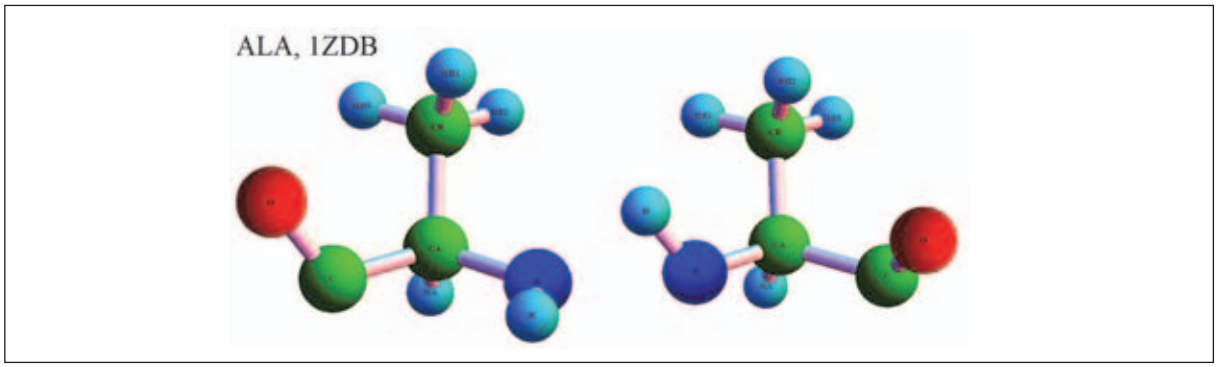

Рис. 30. Отрицательно ориентированный ALA (слева)

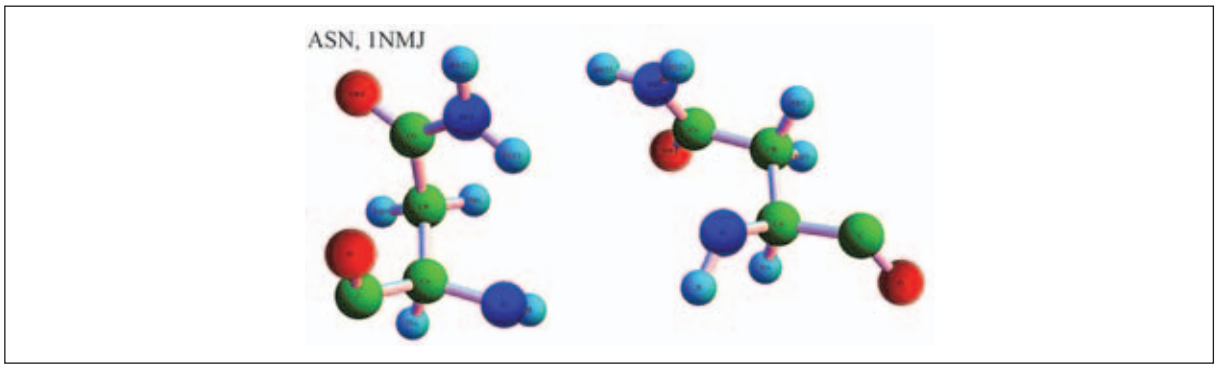

Рис. 31. Отрицательно ориентированный ASN (слева)

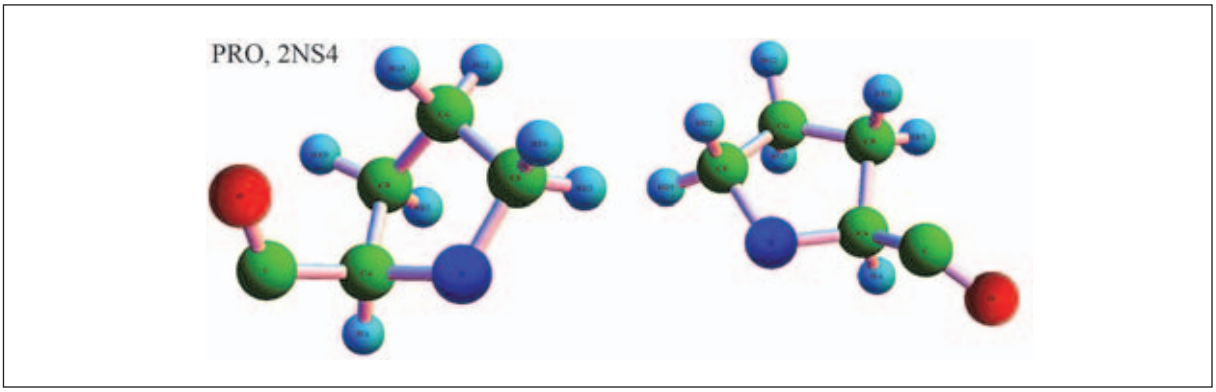

Рис. 32. Отрицательно ориентированный PRO (слева)

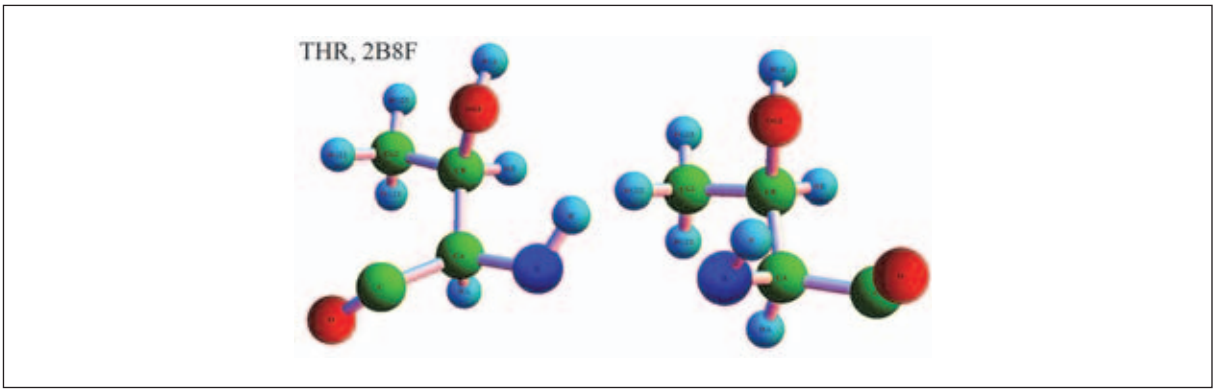

Рис. 33. Отрицательно ориентированный THR (слева).

Таким образом, недоразумение с ориентацией в глицинах, в принципе, легко исправляется. Может, это — единственная причина возникновение неориентированных аминокислот? Ответ оказался отрицательным. Однако, количество полипептидов, содержащих отрицательно ориентированные аминокислоты, отличные от глицина, достаточно мало (равно четырем). Вот список имен этих полипептидов:

`1NMJ, 1ZDB, 2B8F, 2NS4`

Отметим, что в каждом из таких полипептидов имеется ровно одна отрицательно ориентированная аминокислота: в трех их них она располагается на последнем месте, а в одном — на втором. Ниже мы приводим изображения этих аминокислот, вместе с их положительно ориентированными образцами из полипептида 1A03.

4. Выводы

Таким образом, в данной работе демонстрируется, что, прежде чем использовать информацию PDB и подобных баз данных для широкого статистического и геометрического изучения проблемы конформации полипептидов, эти базы данных нужно еще раз глубоко проанализировать и существенно доработать. В работе описан ряд геометрико-статистических методов, которые позволяют находить в базах такого типа явно выходящие за рамки здравого смысла включения.

GEOMETRY OF AMINO ACIDS AND POLYPEPTIDES

A.O. Ivanov, A.S. Mishchenko and A.A. Tuzhilin

Lomonosov Moscow State University

aoiva@mech.math.msu.su, asmish@mech.math.msu.su, tuz@mech.math.msu.su

Received 15.04.2014

In the paper we present several geometrical and statistical methods analyzing information contained in Protein Data Bank and similar data bases for consistency and general theory fitting. Also we include a list of examples demonstrating the necessity of careful analysis and sampling of the PDB items.