ГЕОМЕТРИЯ АМИНОКИСЛОТ И ПОЛИПЕПТИДОВ: СЛУЧАЙ РЕНТГЕНОСТРУКТУРНОГО АНАЛИЗА

Е.А. Вилкул, А.А. Тужилин

Московский государственный университет им. М.В. Ломоносова

tuz@mech.math.msu.su

Поступила 02.10.2014

Настоящая работа продолжает тестирование Protein Data Bank, начатое в [1]: теперь мы интересуемся той частью базы полипептидов, которая была получена с помощью рентгеноструктурного анализа. Мы покажем, что все “патологии”, обнаруженные нами в случае ядерно-магнитного резонанса, встречаются и здесь. Кроме того, мы продемонстрируем, что хорошее разрешение (оно отражается в параметре RESOLUTION) не гарантирует отсутствие “патологий”. Наконец, мы обсудим нарушение закона плоскости Полинга и покажем, что существенные нарушения достаточно часто встречаются как в спиралях, так и в листах.

УДК 514.8+57.087+51-76

Введение

В предыдущей статье [1] было проведено исследование файлов базы Protein Data Bank (PDB), в которых отражены результаты изучения полипептидов, полученные путем ядерного магнитного резонанса. В процессе тестирования базы мы неоднократно сталкивались с различного рода неточностями, которые, как мы думаем, могут быть связаны как с фактическими ошибками, возникшими во время формирования файлов, так и с погрешностями самого метода ядерного магнитного резонанса. В настоящей статье мы расскажем об аналогичном нашем исследовании, проведенном

для полипептидов, полученных путем рентгеноструктурного анализа, который, как считается, дает более точные результаты.

Отметим, что в рассматриваемых нами файлах содержится ровно одна модель, поэтому раздел, посвященный сравнению свойств разных моделей полипептидов будет опущен. Впрочем, ряд полипептидов состоит из нескольких субъединиц. Для начала мы решили ограничиться первой из них.

Как и раньше, в первом разделе статьи будет описана предварительная подготовка базы, после чего мы изложим результаты, связанные с исследованием отклонений длин ковалентных связей, расстояний между последовательными альфа-углеродами, углов между смежными ковалентными связями, а также проверкой "закона плоскости" Полинга [2].

Авторы выражают благодарность профессору В.Л. Голо, научному сотруднику Я.Г. Гурскому, профессору А.О. Иванову, д.ф.-м.н. Ю.Ф. Крупянскому и слушателям его семинара, профессору А.С. Мищенко, д.х.н. В.И. Польшакову и слушателям его семинара, академику РАН, профессору А.Т. Фоменко, к.ф.-м.н. А.К. Шайтану, профессору К.В. Шайтану и слушателям его семинара за внимание к работе и многочисленные полезные обсуждения.

1 Protein Data Bank

В работе [1] мы уже описали, как скачать все необходимые файлы из PDB. Мы не станем повторять это описание. Отметим лишь, что в данном случае наша база оказалась почти в 10 раз больше предыдущей: на момент завершения скачивания она содержала 83194 файла. Дальнейшая работа состояла из двух шагов: предварительной обработки базы (выделение файлов с "однородным" описанием) и метрического анализа. В работе [1] первая часть содержится в разделе "Ряд трудностей, возникающих при работе с pdb-файлами". Оказалось, что почти все проблемы этого раздела имеют место и в случае рентгеноструктурного анализа. Обсудим детали.

1.1 Ряд трудностей, возникающих при работе с pdb-файлами

(1) Хотя строки "ATOM..." присутствуют почти во всех файлах, имеется исключение: файл 4OYW.pdb, после выбрасывания которого наша база стала содержать 83193 файла.

(2) Теперь избавимся от файлов, в которых между первой строчкой "ATOM..." и первой строчкой "TER..." имеются строки, не начинающиеся со слова "ATOM". В результате остается 55024 файла.

(3) На сей раз мы столкнулись с еще одной проблемой: в некоторых из оставшихся файлов (таких файлов 55) на месте аминокислот встречаются обозначения, отличные от 20 стандартных. Имеется четыре таких обозначения: UNK, GLX, ASX и SEC. Первое из них, UNK, содержится в 51 файле, например в 155C.pdb, и говорит о том, что в соответствующей области аминокислотная последовательность определена не была. Второе, GLX, содержится ровно в одном файле, а именно, в 1KP0.pdb, третье, ASX, — в трех файлах: 1KP0.pdb, 2ATC.pdb и 2FMD.pdb, наконец, четвертое, SEC, — ровно в одном файле 2IV2.pdb. После выкидывания всех этих файлов осталось 54969 штук.

(4) Следующий шаг — определить наиболее популярные составы аминокислот. Напомним, что, как было выяснено в нашем предыдущем исследовании, аминокислоты, имеющие одни и те же имена, могут быть представлены разными атомными составами. На сей раз это также было обнаружено. Приведем пример возможностей для глицина.

```
{N, CA, C, O}, {N, CA, C, O, H}}, {CA},
{N, CA, C, O, OXT}, {N, CA, C, O, H, HA2, HA3},
{N, CA, C, O, H1, H2, H3}, {N, CA, C, O, OXT, H}
{N, CA, C}, {N, CA, C, H}, {N, CA}.
```

Мы снова выбрали наиболее часто встречающиеся последовательности и назвали их *стандартными*. Вот полученный список:

```
GLY: N,CA,C,O;
ALA: N,CA,C,O,CB;
SER: N,CA,C,O,CB,OG;
CYS: N,CA,C,O,CB,SG;
PRO: N,CA,C,O,CB,CG,CD;
VAL: N,CA,C,O,CB,CG1,CG2;
THR: N,CA,C,O,CB,OG1,CG2;
ILE: N,CA,C,O,CB,CG1,CG2,CD1;
LEU: N,CA,C,O,CB,CG,CD1,CD2;
ASP: N,CA,C,O,CB,CG,OD1,OD2;
ASN: N,CA,C,O,CB,CG,OD1,ND2;
GLU: N,CA,C,O,CB,CG,CD,OE1,OE2;
GLN: N,CA,C,O,CB,CG,CD,OE1,NE2;
MET: N,CA,C,O,CB,CG,SD,CE;
LYS: N,CA,C,O,CB,CG,CD,CE,NZ;
ARG: N,CA,C,O,CB,CG,CD,NE,CZ,NH1,NH2;
HIS: N,CA,C,O,CB,CG,ND1,CD2, CE1,NE2;
PHE: N,CA,C,O,CB,CG,CD1,CD2,CE1,CE2,CZ;
TYR: N,CA,C,O,CB,CG,CD1,CD2,CE1,CE2,CZ,OH;
TRP: N,CA,C,O,CB,CG,CD1,CD2,NE1,CE2,CE3,CZ2,CZ3,CH2.
```

Замечание 1.1. Отметим, что при разбиении строк “ATOM...” на списки, каждый из которых относится к отдельной аминокислоте, мы использовали не только позиции 23–26, соответствующие номерам аминокислот, но также и позицию 27, в которой располагается код вставки остатка iCode. Тем самым, мы избежали объединения атомов последовательных аминокислот с одними и теми же номерами в списки единых аминокислот.

(5) Полученные в предыдущем пункте наиболее популярные атомные составы аминокислот соответствуют внутренним (не концевым) аминокислотам. На данном этапе мы прошли по всем оставшимся на предыдущих шагах pdb-файлам и выбрали те, в которых все внутренние аминокислоты — стандартные. Теперь наша база содержит 38969 файлов.

(6) Для ускорения расчетов, мы преобразовали каждый файл в файл такого же “короткого формата”, как описано в [1]. Таким образом, если начальная база полипептидов, полученных методом рентгеноструктурного анализа, занимала больше 48 гигабайт, то нынешняя занимает около 2.2 гигабайта.

2 Метрический анализ PDB

В данном разделе будут описаны результаты, связанные с метрическим анализом PDB-базы по таким критериям, как разброс длин ковалентных связей, изменение расстояний между последовательными альфа-углеродами, отклонение от закона плоскости, разброс углов между смежными ковалентными связями. Кроме того, будет исследована зависимость полученных результатов от значения параметра RESOLUTION, указанного к каждом PDB-файле, которое отображает погрешность измерений в методе рентгено-структурного анализа.

2.1 Оценка разброса длин ковалентных связей в аминокислотах

Для оценки отклонений длин ковалентных связей мы решили на сей раз поступить следующим образом.

(1) Для всех аминокислот вычислим среднее значение длин каждой ковалентной связи. В результате мы получили следующую таблицу значений (каждый подсписок соответствует одной из 20 аминокислот, упорядоченных как в приведенном выше списке “стандартных” атомных составов):

```
{1.45521, 1.51779, 1.23268},
{1.46013, 1.52596, 1.23239, 1.52472},
{1.45951, 1.52527, 1.23236, 1.53025, 1.41734},
{1.45905, 1.52396, 1.23188, 1.52988, 1.80867},
{1.46566, 1.52694, 1.23286, 1.53006, 1.49772, 1.50819, 1.47469},
{1.45916, 1.52657, 1.23243, 1.54749, 1.5243, 1.52369},
{1.45899, 1.52566, 1.23236, 1.5438, 1.5244, 1.43266},
{1.45897, 1.52652, 1.23229, 1.54736, 1.5323, 1.52828, 1.51974},
{1.45914, 1.52513, 1.23195, 1.53164, 1.53131, 1.52248, 1.52331},
{1.46025, 1.52626, 1.23203, 1.5325, 1.51964, 1.25094, 1.25116},
{1.45962, 1.5253, 1.23203, 1.53204, 1.51739, 1.32877, 1.23362},
{1.45943, 1.52573, 1.23178, 1.53136, 1.52319, 1.52195, 1.25228,
  1.25268},
{1.45941, 1.52523, 1.23184, 1.53095, 1.52258, 1.51929, 1.23456,
  1.32996},
{1.45908, 1.52481, 1.23175, 1.53067, 1.52073, 1.80633, 1.78985},
{1.45977, 1.526, 1.23202, 1.53135, 1.52309, 1.52524, 1.52574,
  1.49397},
{1.45947, 1.52559, 1.23186, 1.53114, 1.52137, 1.52261, 1.46139,
  1.33145, 1.32968, 1.32832},
{1.45962, 1.52444, 1.23194, 1.53196, 1.49771, 1.3562, 1.37769,
  1.37283, 1.32302, 1.32226},
{1.45906, 1.52449, 1.23197, 1.5327, 1.50417, 1.39119, 1.39097,
  1.3937, 1.39364, 1.38975, 1.38982},
{1.459, 1.52424, 1.23211, 1.53248, 1.51175, 1.3942, 1.39382,
  1.39219, 1.39192, 1.38553, 1.38515, 1.37671},
{1.45931, 1.52436, 1.23219, 1.53222, 1.49989, 1.3675, 1.43262,
  1.37472, 1.41345, 1.40116, 1.39836, 1.36942, 1.39218,
  1.37377, 1.40559};
```

(2) Для каждого полипептида подсчитаем максимальное относительное отклонение от средних значений длин всех ковалентных связей в аминокислотах и сохраним значения в список (на первом месте каждой записи стоит имя полипептида, на втором — значение максимального отклонения в процентах):

```
{"101M", 4.37}
{"102L", 6.1}
{"102M", 4.34}
{"103L", 5.86}
{"103M", 4.32}
{"104L", 6.64}
{"104M", 3.92}
{"105M", 3.61}
{"106M", 2.71}
... ... ... ... ...
```

(3) Построим график распределения количества полипептидов (ордината) с тем или иным максимальным отклонением (абсцисса), см. рис. 1.

Приведем начальный отрезок численных значений этого графика:

```
{327, 4653, 7121, 6267, 5119, 4461, 3199, 2310, 1504, 964, 633,
 404, 368, 244, 174, 163, 163, 103, 93, 69, 53, 76, 32, 37, 49,
 32, 29, 24, 21, 11, 20, 19, 11, 7, 11, 11, 6, 8, 2, 7, 9, 6, 6,
 6, 6, 11, 3, 4, 2, 4, 2, 2, 6, 2, 1, 4, 3, 3, 2, 4, 3, 2, 4, 1,
 4, 1, 0, 2, 0, 2, 2, 2, ...  0, 0, 1}.
```

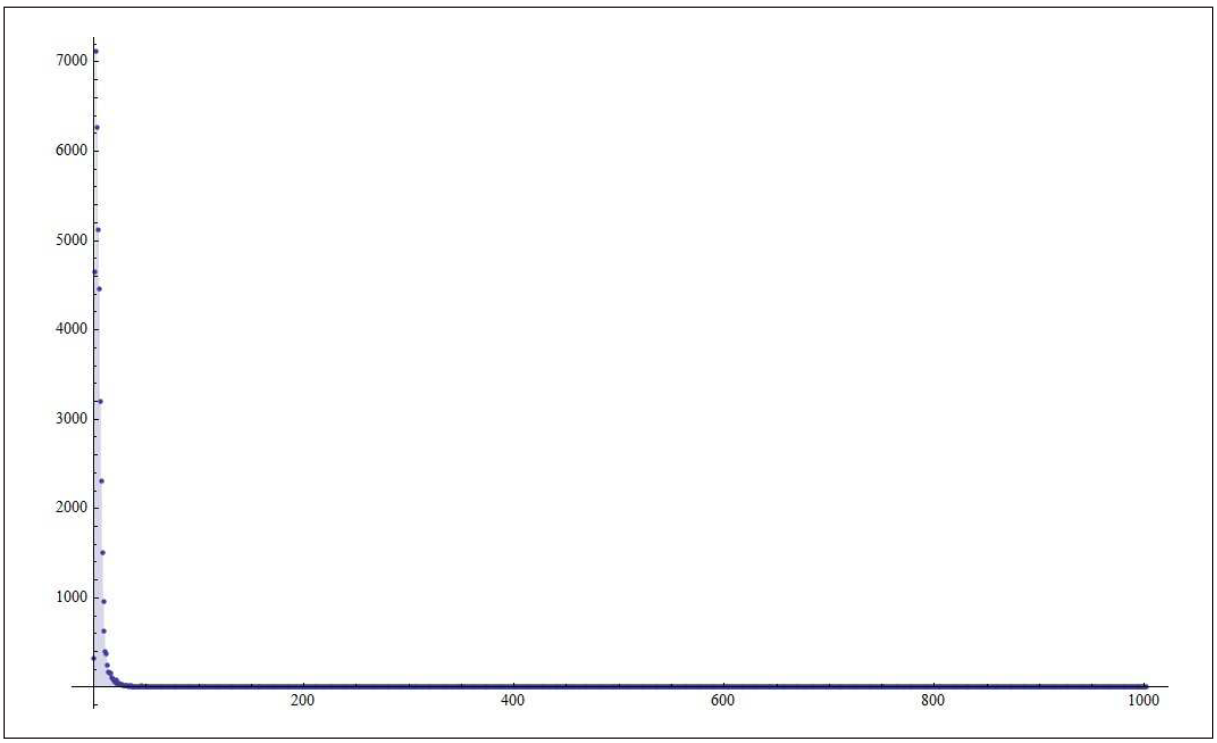

Рис. 1. График распределения количества полипептидов (ордината) с тем или иным максимальным отклонением (абсцисса).

Из рис. 1 видно, что в некоторых файлах отклонение превышает 1000 процентов! Откуда же берутся такие “патологии”? Одна из гипотез состояла в существовании погрешностей при использовании метода рентгеноструктурного анализа. Как было выяснено из бесед со специалистами, погрешности метода отвечает параметр RESOLUTION (или, как его еще называют, *разрешение*), который содержится в каждом pdb-файле: чем меньше значение этого параметра, тем точнее полученные результаты. Мы решили посмотреть, всегда ли “патологии” появляются только при “больших” значениях параметра RESOLUTION.

Для начала мы нашли файл с отклонением, превышающим 1000%: им оказался 1SBT. На рис. 2 приведено изображение аспарагина из полипептида 1SBT, на котором такое огромное отклонение достигается. Отметим, что в этом случае параметр RESOLUTION равен 2.5.

Следующее по величине отклонение возникает в файле 1W5Q, см. рис. 3. Здесь разрешение существенно меньше и равно 1.4, тем не менее, прослеживается явное нарушение длины связи CZ-NH2. Может быть, значение RESOLUTION все еще недостаточно мало?

Чтобы разобраться с этим вопросом, построим график распределения параметра RESOLUTION (ось ординат соответствует количеству файлов, абсцисса — значению параметра), см. рис. 4, и приведем некоторые численные данные:

(1) минимальное разрешение равно 0.2, максимальное — примерно 9;

(2) его среднее значение по всей базе равно 2.165;

(3) количество файлов с RESOLUTION меньшим 1.4 мало, а именно, 585 файла из 38969.

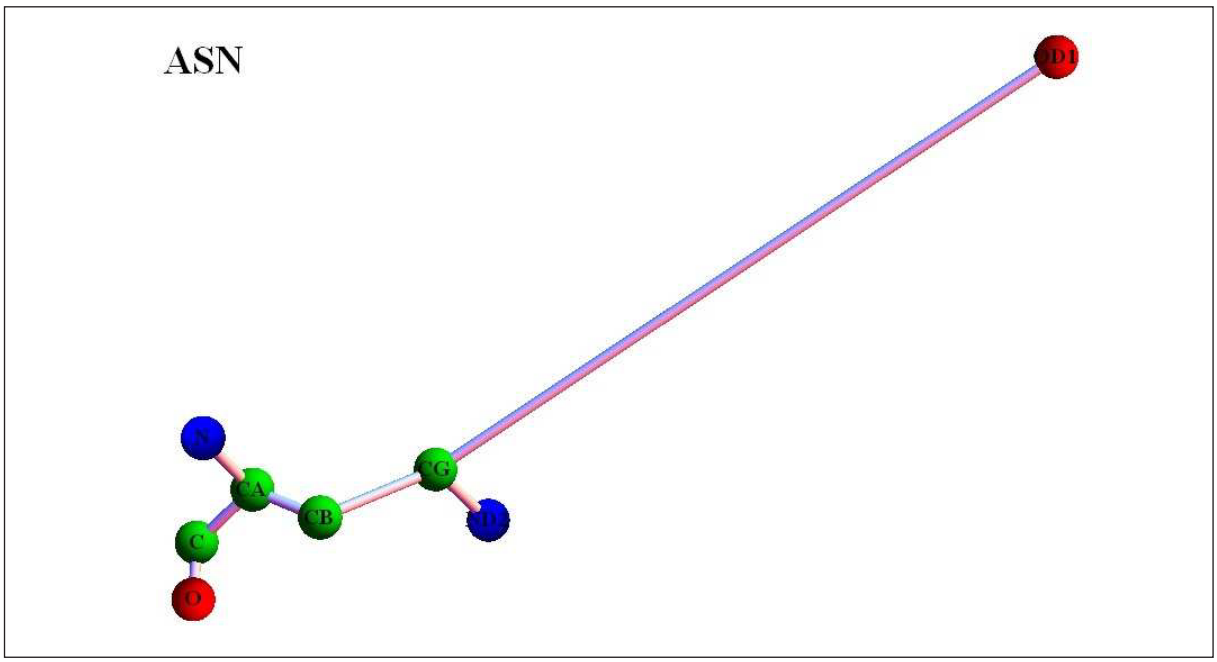

Рис. 2. Аспарагин, на котором максимальное отклонение от средних длин ковалентных связей превышает 1000%.

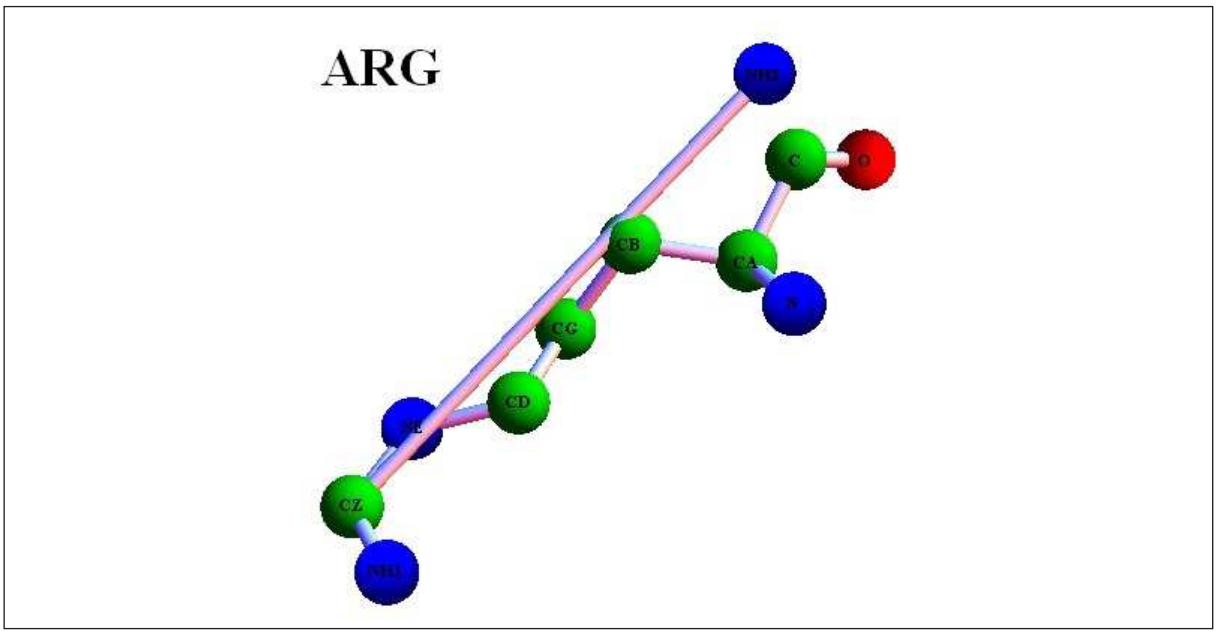

Рис. 3. Аргенин, на котором достигается второе по величине максимальное отклонение от средних длин ковалентных связей (превышает 450%).

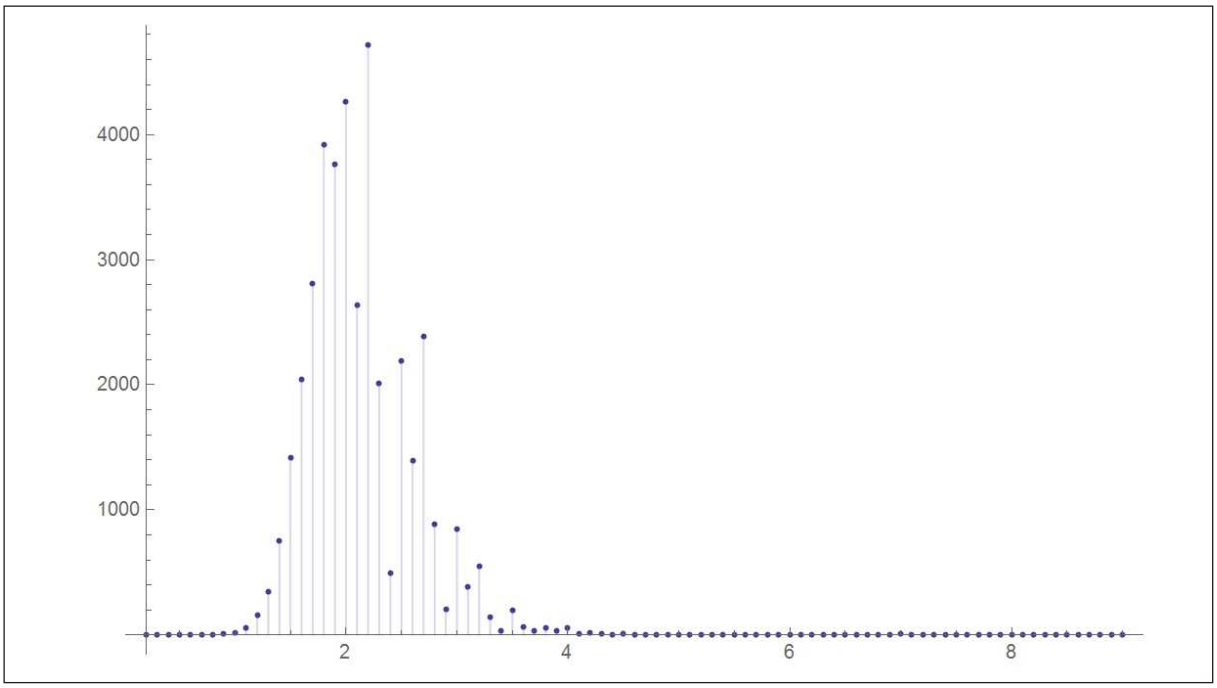

Рис. 4. График распределения параметра RESOLUTION (ордината количество файлов, абсцисса значение параметра RESOLUTION).

Приведенные нами данные показывают, что разрешение 1.4, возникающее в рассмотренном выше примере, является достаточно маленьким относительно остальных. Тем не менее, максимальное относительное отклонение длин ковалентных связей свидетельствует о наличии существенных ошибок даже при таком разрешении.

Приведем список всех файлов со значительными отклонениями от среднего (большими 100 процентов) вместе со значениями их отклонений и разрешений:

```
{1SBT, 1001.55, 2.50}, {1W5Q, 474.02, 1.4},
{2CAB, 359.65, 2.00},  {3CD2, 280.92, 2.50},
{1QB7, 276.43, 1.50},  {1J4X, 264.23, 2.75},
{1OJM, 251.45, 1.78},  {1OJP, 247.44, 1.90},
{2JE1, 246.91, 2.69},  {1OJN, 244.06, 1.60},
{2BV9, 214.91, 1.50},  {1H18, 211.2, 2.3},
{1H16, 195.67, 1.53},  {2J07, 191.16, 1.95},
{1H17, 188.08, 1.75},  {1W5O, 183.2, 1.85},
{2VLR, 170.28, 2.3 },  {2BS2, 168.73, 1.78},
{1KRL, 163.59, 1.90},  {2OI4, 153.36, 2.20},
{1HCH, 152.07, 1.57},  {1MOO, 143.9, 2.40},
{1H5E, 137.91, 1.60},  {1H5D, 137.87, 1.60},
{1H5F, 137.63, 1.6 },  {1E7R, 129.46, 1.6},
{1E6Z, 128.29, 1.99},  {2V28, 128.12, 1.95},
{1LZ9, 127.24, 1.70},  {1W8L, 117.61, 1.8},
{1GQ8, 116.01, 1.75},  {1AZN, 113.76, 2.60},
{1JXB, 109.41, 1.60},  {1UYC, 108.16, 2.0},
{1MC9, 106.14, 1.70},  {2W27, 105.15, 2.80},
{1GTJ, 105.11, 1.75},  {1S6P, 104.44, 2.90},
{1EGD, 100.1, 2.40}.
```

Заметим, что для некоторых из этих файлов разрешение также невелико (не превосходит 1.6), тем не менее, в соответствующих полипептидах наблюдаются аномальные отклонения в длинах ковалентных связей. Для наглядности, приведем изображения шести аминокислот из файлов 2BV9, 1E7R, 1H5E, 1QB7, 1OJN, 1HCH, на которых достигаются максимальные отклонения (см. рис. 5).

Рассмотренные только что примеры опровергают предположение о том, что возникновение патологий напрямую связано с разрешением метода рентгеноструктурного анализа.

После детального анализа было выявлено, что заметными отклонениями от среднего значения длин ковалентных связей можно считать те, которые превосходят 22%. Таких файлов оказалось 511 штук. Как и раньше, мы исключили их из нашей базы.

2.2 Оценка разброса углов в полипептидах

Было решено немного изменить порядок исследования и форму изложения результатов по сравнению с предыдущей статьей. Для оценки разброса углов между смежными ковалентными связями в аминокислотах мы поступили следующим образом.

(1) Для каждой из 20 стандартных аминокислот вычислили среднее значение величин углов между соседними связями, получили таблицу, каждый подсписок которой соответствует своей аминокислоте (порядок аминокислот такой же, как и выше):

```
{113.062, 120.507},
```

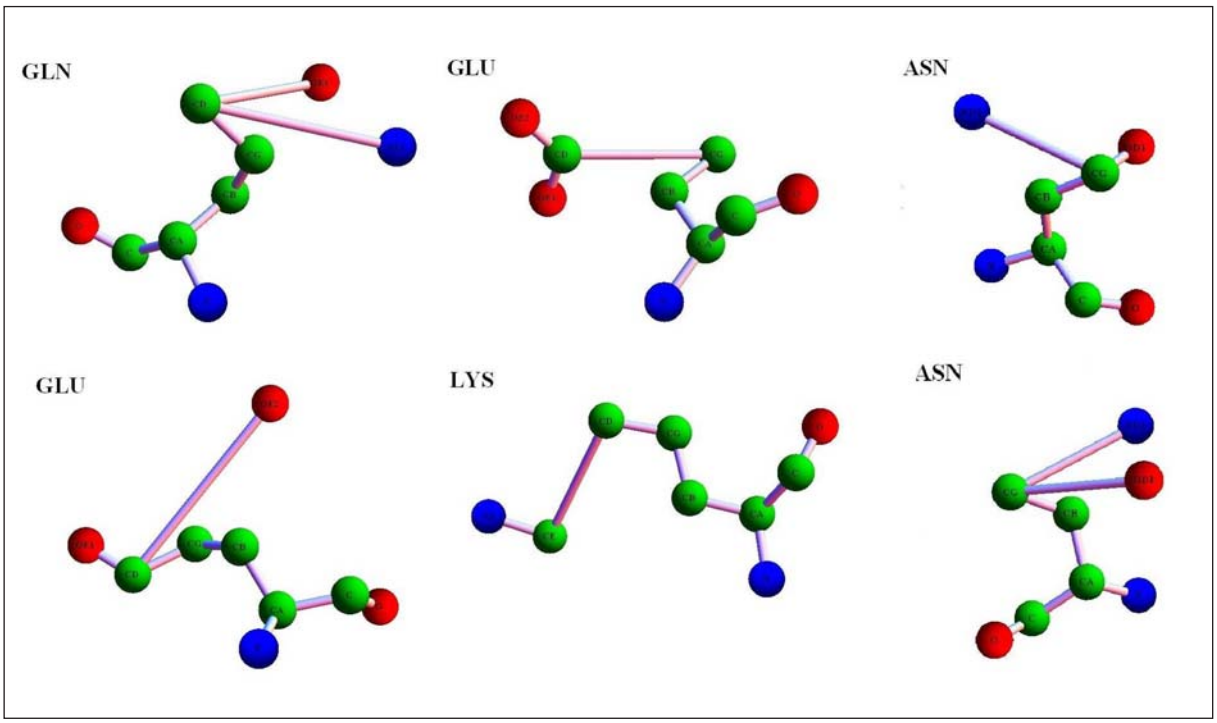

Рис. 5. Изображения “аномальных” аминокислот из файлов 2BV9, 1E7R, 1H5E, 1QB7, 1OJN, 1HCH.

```
{111.078, 110.294, 110.204, 120.56},
{111.313,110.411, 109.772, 120.529, 110.787},
{111.073, 110.454, 109.804, 120.515, 113.921},
{111.37, 112.698, 103.246, 111.176, 120.284, 103.934, 104.367,
  102.715},
{110.102, 111.453, 109.901, 120.576, 110.639, 110.348, 110.407},
{110.816, 111.236, 109.365, 120.536, 109.243, 111.025, 109.058},
{110.041, 111.38, 109.953, 120.567, 110.249, 110.605, 110.618,
  113.947},
{110.964, 110.414, 109.894, 120.525, 116.796, 110.284, 110.37,
  110.57},
{111.113, 110.493, 110.012, 120.544, 112.881, 119.014, 118.389,
  122.572},
{111.634, 110.463, 110.123, 120.541, 112.638, 120.888, 116.561,
  122.521},
{111.27, 110.504, 109.939, 120.547, 114.135, 113.063, 118.832,
  118.243, 122.905},
{111.158, 110.52, 109.909, 120.53, 114.012, 112.712, 120.89,
  116.554, 122.532},
{111.1, 110.503, 109.903, 120.551, 113.988, 112.496, 100.816},
{111.093, 110.498, 109.958, 120.554, 114.175, 111.522, 111.623,
  111.951},
{111.068, 110.507, 109.967, 120.554, 114.063, 111.696, 111.805,
  124.576, 120.537, 119.81, 119.631},
{111.26, 110.454, 109.957, 120.496, 113.631, 122.75, 130.926,
```

```
  106.271, 109.185, 107.126, 108.485, 108.917},
{111.004, 110.546, 109.912, 120.56, 113.828, 120.641, 120.514,
  118.799, 120.806, 120.808, 119.863, 119.868, 119.831},
{111.132, 110.494, 109.847, 120.546, 113.8, 120.888, 120.784,
  118.281, 121.092, 121.108, 119.497, 119.513, 120.485, 119.779,
  119.72},
{111.272, 110.517, 109.88, 120.541, 113.868, 126.896, 126.683,
  106.361, 110.06, 107.094, 133.959, 118.934, 108.981, 107.486,
  122.375, 130.127, 118.675, 117.496, 121.049, 121.449}
```

Заметим, что минимальная величина в этом списке равна 100.82, а максимальная — 133.96.

(2) Для каждого полипептида вычислили максимальное среди отклонений углов от их средних значений, полученные данные сохранили в файл.

Построим график распределения максимальных отклонений углов в полипептидах от их среднего значения, где абсцисса соответствует величине отклонения в процентах, а ордината — количеству файлов (см. рис. 6).

Приведем также список точных численных значений этого графика:

```
{11, 49, 214, 1100, 2786, 4001, 4484, 5048, 5097, 4410, 3240,
2230, 1520, 1043, 687, 489, 353, 307, 237, 190, 158, 148, 96,
83, 51, 39, 60, 31, 27, 23, 30, 27, 19, 24, 17, 12, 10, 21,
10, 13, 5, 9, 8, 14, 4, 3, 3, 4, 5, 2, 2, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1}
```

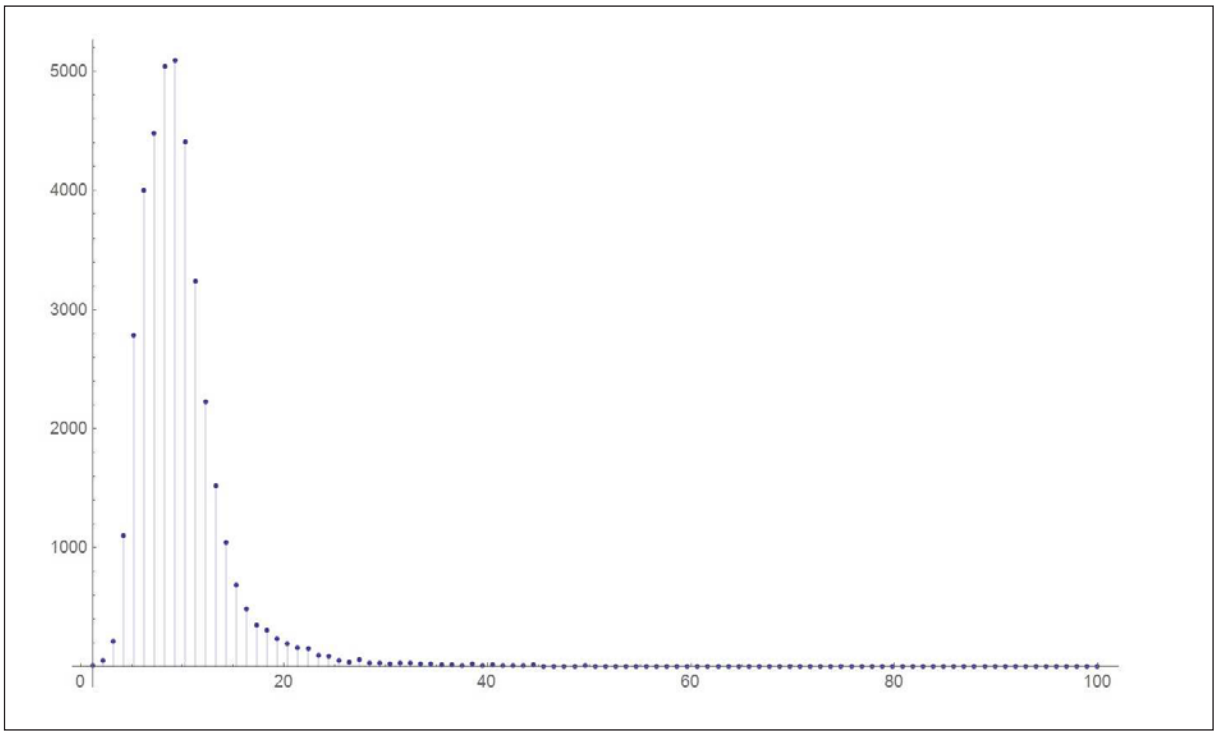

Рис. 6. График распределения максимальных отклонений углов в полипептидах от их среднего значения: абсцисса отклонение в процентах, ордината количество файлов.

Из графика видно, что максимальное относительное отклонение углов может быть более 100%. Как и в предыдущем разделе, проверим, возникают ли большие отклонения в углах только при большом значении разрешения RESOLUTION. Для этого приведем список полипептидов, имеющих максимальные отклонения более 50 процентов, вместе со значениями их отклонений и разрешений:

```
{4ILX, 99.95, 1.6}, {1H55, 96.91, 1.61}, {1K1B, 88.65, 1.9},
{2JIX, 63.74, 3.2}, {1SAC, 51.74, 2.}, {3MCG, 51.24, 2.},
{1CPS, 51.11, 2.25}, {2YWV, 50.45, 1.75}.
```

Заметим, что в списке присутствуют как большие значения разрешений (например, 3.2), так и достаточно маленькие (1.6). Таким образом, прямой зависимости между возникновением “патологий” и параметром RESOLUTION снова не прослеживается.

Построив изображение глутаминовой кислоты из файла 4ILX, на которой достигается отклонение углов от среднего почти в 100 процентов (см. рис. 7), мы столкнулись с интересным явлением: на первый взгляд, аминокислота выглядит абсолютно нормально. Тем не менее, после вывода последовательности атомов в этой аминокислоте

```
N, CA, C, O, CB ,CG, CD, OE1, OE2,
```

а также значений отклонений от среднего каждого из углов было выяснено, что два кислорода OE1 и OE2 “склеились”.

Аналогичная ошибка прослеживается и для полипептида 1H55 со вторым по величине отклонением от среднего: в валине, изображенном на рис. 8, два атома CG1 и CG2 имеют почти одинаковые координаты.

Приведем изображения всех аминокислот, в которых отклонения превышают 50%, см. рис. 9. Помимо вышеописанного эффекта “склеивания” атомов, можно встретить также явные нарушения в бензольном кольце у фенилаланина, а также образование почти развернутого угла CD-CG-CB в некоторых из аминокислот.

Дальнейшее детальное исследование базы показало, что отклонения углов от среднего значения менее 30% можно считать несущественными. После исключения из базы 274 полипептидов с существенными отклонениями осталось 38184 файла.

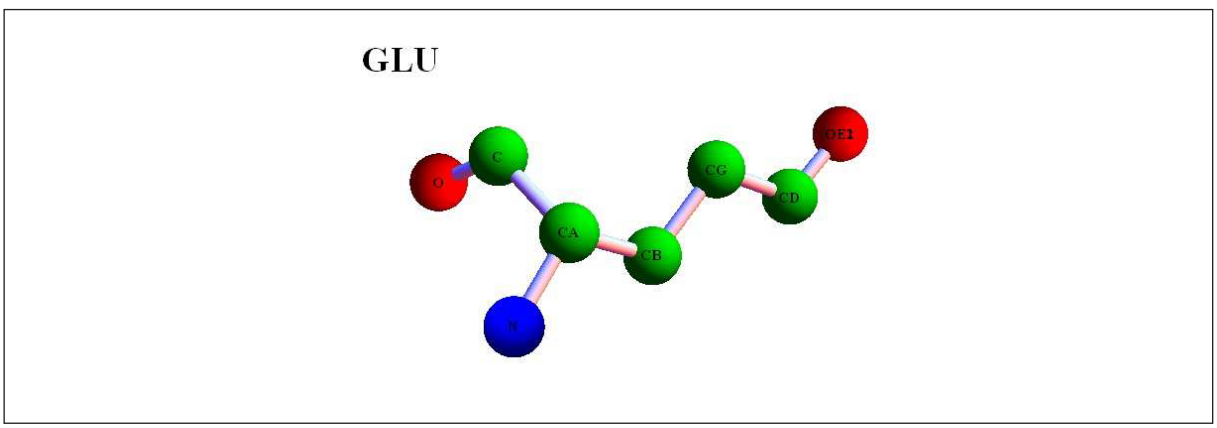

Рис. 7. Глутаминовая кислота из полипептида 4ILX: кислороды OE1 и OE2 ошибочно совмещены.

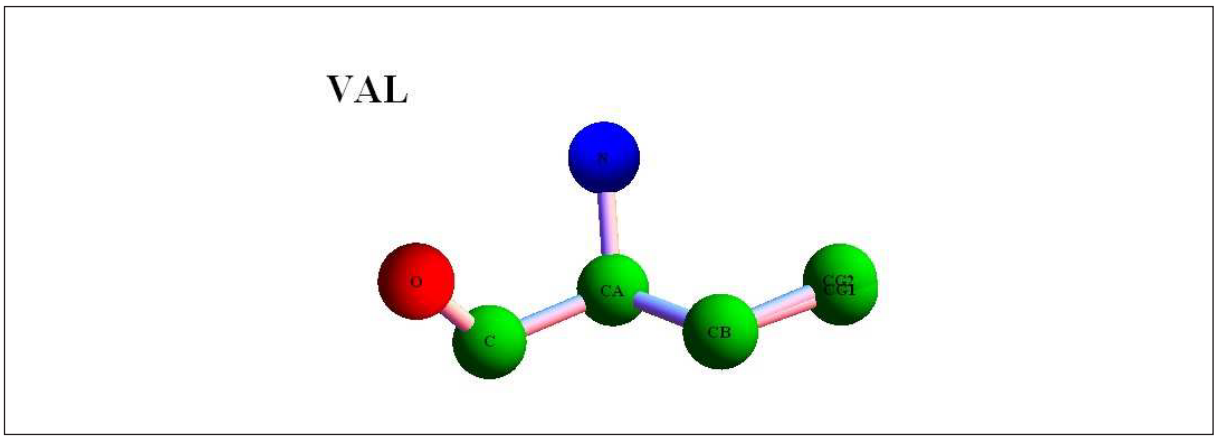

Рис. 8. Валин из полипептида 1H55: углероды CG1 и CG2 ошибочно совмещены.

2.3 Оценка разброса длин пептидных связей

Стоит напомнить, что все вышеописанные исследования мы проводили лишь для внутренних аминокислот каждого полипептида, поэтому, прежде чем начать анализировать базу по критерию разброса длин пептидных связей, было бы целесообразно исключить все файлы с ровно одной внутренней (не концевой) аминокислотой. Таких файлов оказалось 2 штуки: 3IG6 и 4NL8, и мы исключим их из текущего списка.

Как и раньше, измерим среднее расстояние между C и N двух последовательных аминокислот по всем полипептидам. Оно оказалось равным 1.34. Кроме того, измерим максимальное относительное отклонение от среднего значения для каждого из файлов. “Рекордсмен” — полипептид 1B68, чье относительное отклонение достигает 6500 процентов! В чем же причина таких больших отклонений?

Приведем несколько примеров.

Одной из очевидных причин возникающих отклонений может являться, прежде всего, погрешность в измерениях. Например, по этой причине в полипептиде 3EH0 (см. рис. 10) углерод и азот располагаются очень близко друг к другу.

В следующем примере (файл 1B68) ситуация ровно противоположная — C и N находятся очень далеко, и расстояние между ними отклоняется от среднего значения почти в 5 раз (см. рис. 11), что, скорее всего, не может быть вызвано только погрешностью в измерениях. Естественно возникающее предположение подтвердилось после изучения файла 1B68.pdb: оказывается, в этом файле пропущены аминокислоты с номерами 84 и 85.

Есть также промежуточные примеры, когда, глядя на картинку, не очевидно, какой из двух причин вызвана патология. Посмотрим на пример, приведенный на рис. 12: здесь отклонение от среднего значения составляет 100%.

На сей раз было решено считать отклонение более 22% существенным, что привело к исключению 8333 файлов. Оставшийся список составили 29849 полипептидов.

2.4 Цис и Транс конфигурации пептидных групп

Прежде чем перейти к оценке разброса расстояний между последовательными альфа-углеродами, нужно проанализировать базу на наличие цис- и транс-конфигураций, существование которых приводит к большим отклонениям от средних значений. На сей раз количество полипептидов, содержащих цис-конфигурации, оказалось очень

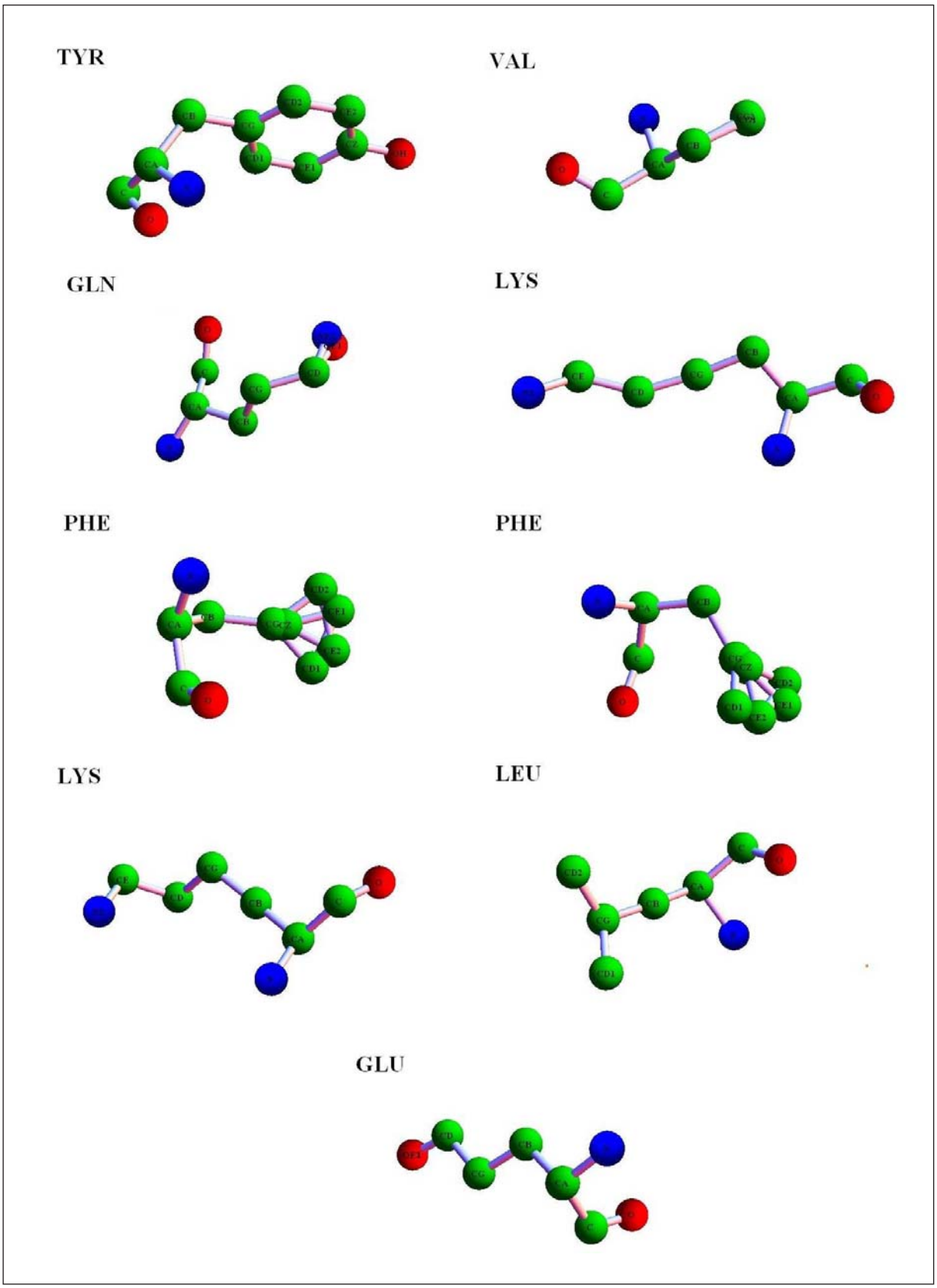

Рис. 9. Все аминокислоты, в которых максимальное отклонение углов от средних превышает 50%.

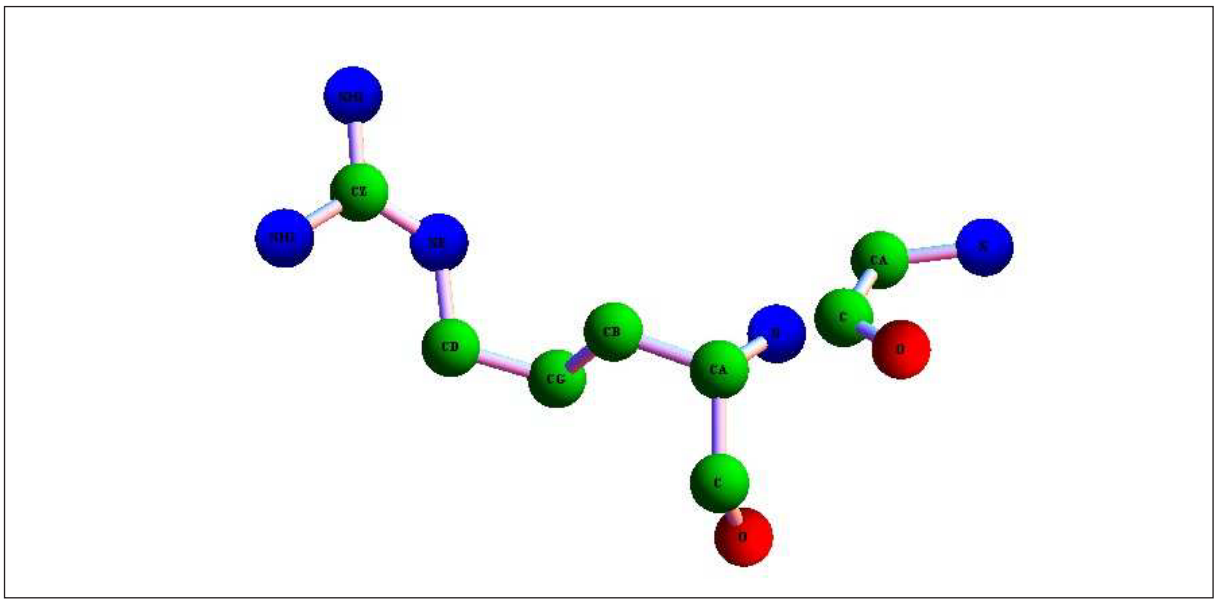

Рис. 10. Аминокислоты из полипептида 3EH0: углерод и азот располагаются очень близко друг к другу.

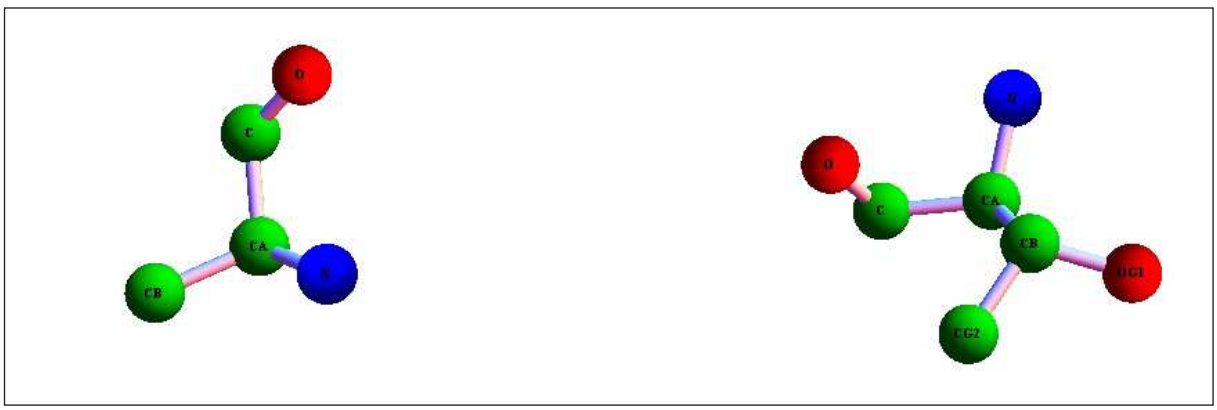

Рис. 11. Аминокислоты из полипептида 1B68: углерод и азот располагаются очень далеко друг от друга (причина пропущены некоторые аминокислоты).

большим, а именно, 13662 файла из 29849 в текущем списке. В основном на каждый такой файл приходится ровно одна подобная конфигурация, о чем свидетельствует следующий список (на первом месте стоит число цис-конфигураций, на втором — количество файлов с таким числом конфигураций):

```
{1, 7583}, {2, 3705}, {3, 1638}, {4, 467}, {5, 107}, {6, 49},
{7, 19}, {8, 72}, {9, 5}, {10, 4}, {11, 2}, {12, 4}, {13, 3},
{16,2}, {19, 1}, {22, 1}.
```

Тем не менее, существует полипептид 2EC5, у которого число цис-конфигураций достигает 22. Приведем, для наглядности, номера пар последовательных аминокислот в этом файле, которые находятся в цис-конфигурации, а также изображения первых четырех из них (см. рис. 13).

```
2, 3, 188, 240, 241, 242, 244, 245, 246, 247, 318, 320, 321, 402,
403, 404, 405, 432, 455, 457, 555, 561.
```

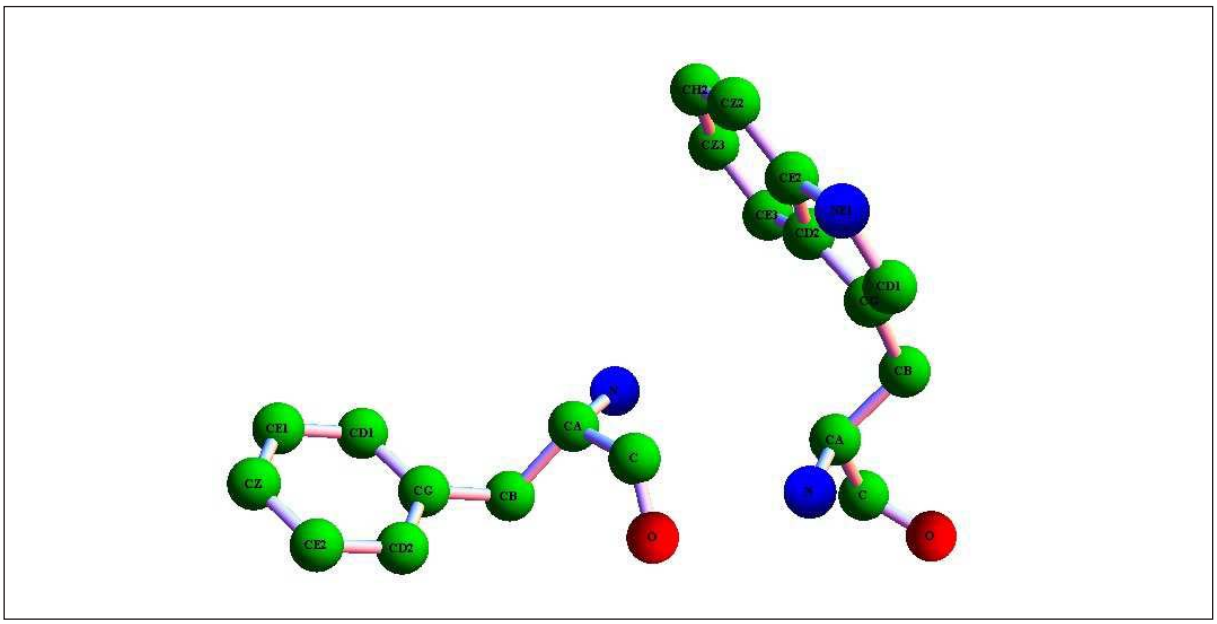

Рис. 12. Последовательные аминокислоты в полипептиде 2QKL. Причина отклонения расстояния между C и N не очевидна.

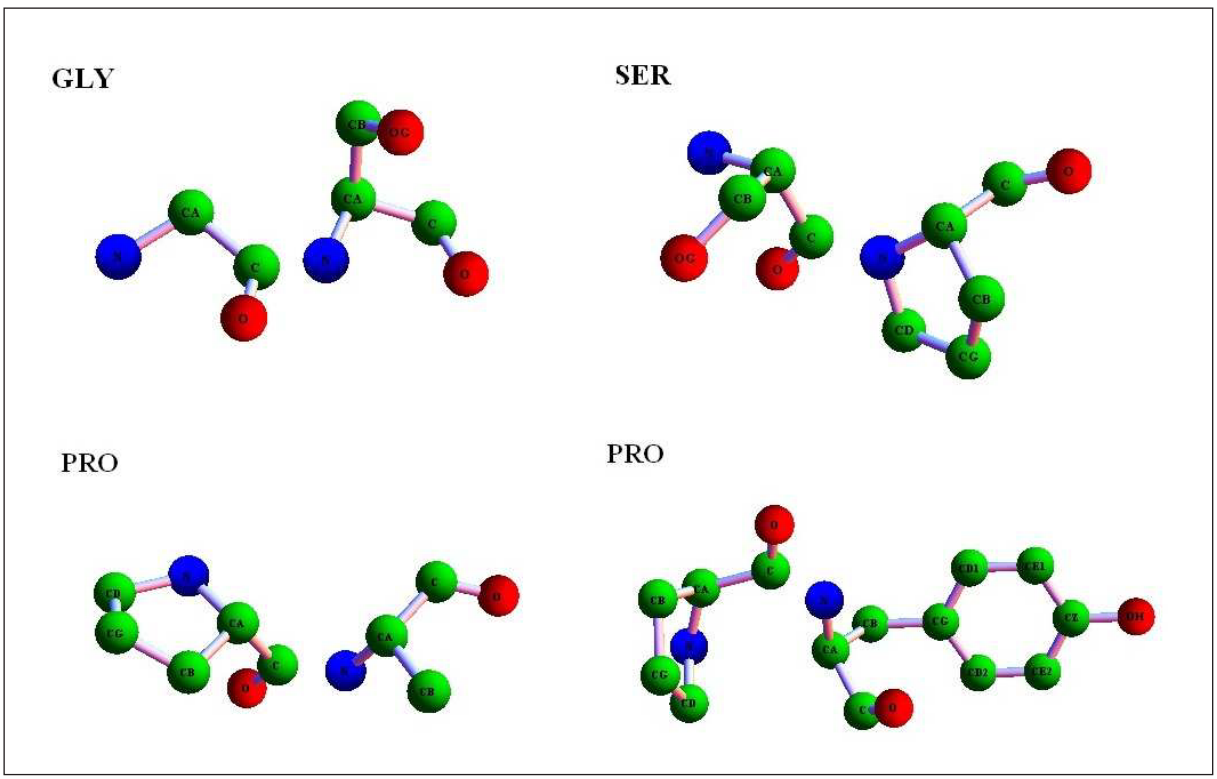

Рис. 13. Четыре примера цис-конфигураций среди пар последовательных аминокислот полипептида 2EC5; этот полипептид рекордсмен: в нем имеется 22 цисконфигурации.

После исключения всех полипептидов, содержащих хотя бы одну цис-конфигураци осталось 16187 файлов.

2.5 Оценка разброса длин расстояний между последовательными альфа-углеродами

После исключения всех цис-конфигураций из базы можно приступить к анализу расстояний между последовательными альфа-углеродами. Его среднее значение по всей базе оказалось равным 3.8, а максимальное отклонение от среднего — 32.6%. На рис. 14 приведен график распределения этих отклонений.

Приведем также список точных значений этого графика:

```
{1134, 8660, 3912, 1409, 580, 241, 109, 57, 35, 14, 11, 8, 5, 3,
3, 1, 1, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1}.
```

В данном случае очень сложно провести границу, начиная с которой отклонение можно считать существенным, так как распределение убывает более или менее равномерно. Внимательно проанализировав изображения полипептидов, мы выяснили, что отклонения от среднего, превышающие 15 процентов, вполне заметны на глаз. Рассмотрим три примера с такими отклонениями (см. рис. 15 – 17).

Анализ данных показал, что существует 6 полипептидов с отклонениями, превышающими 15%:

```
1NRF, 1R59, 1SDL, 1URB, 2HWD, 2W0L.
```

После исключения их из нашего списка остался 16181 полипептид.

2.6 Закон плоскости Полинга

В этой главе будут изложены результаты исследования PDB на наличие отклонений от закона плоскости. По сравнению с ядерно-магнитным резонансом, файлы,

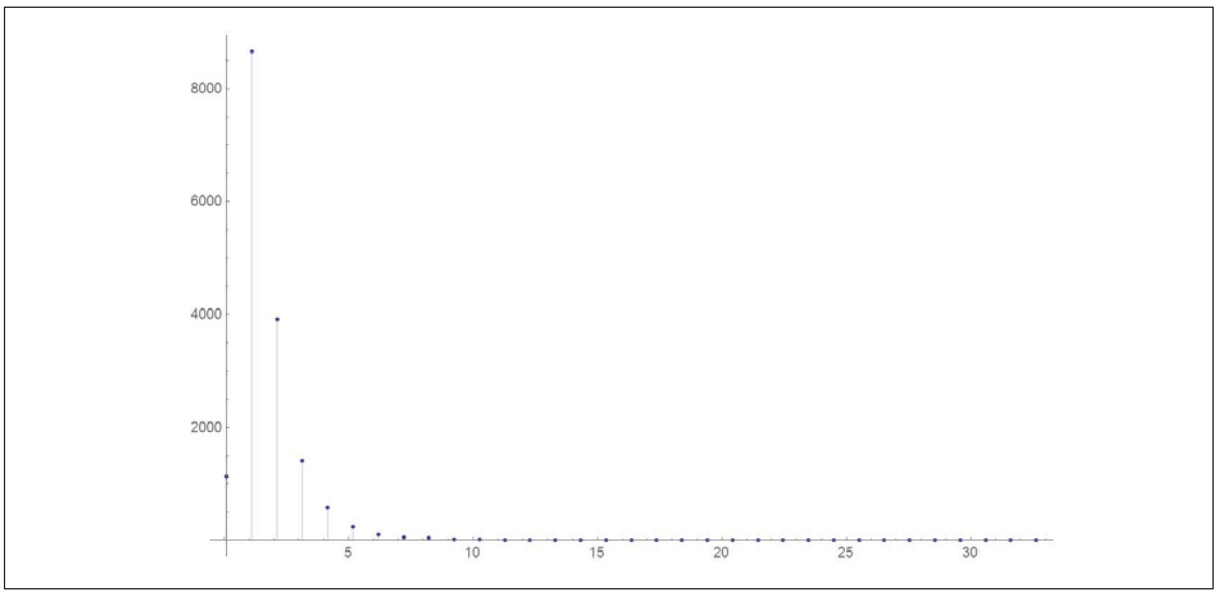

Рис. 14. График распределения максимальных относительных отклонений расстояний между последовательными альфа-углеродами: абсцисса отклонение в процентах, ордината количество файлов.

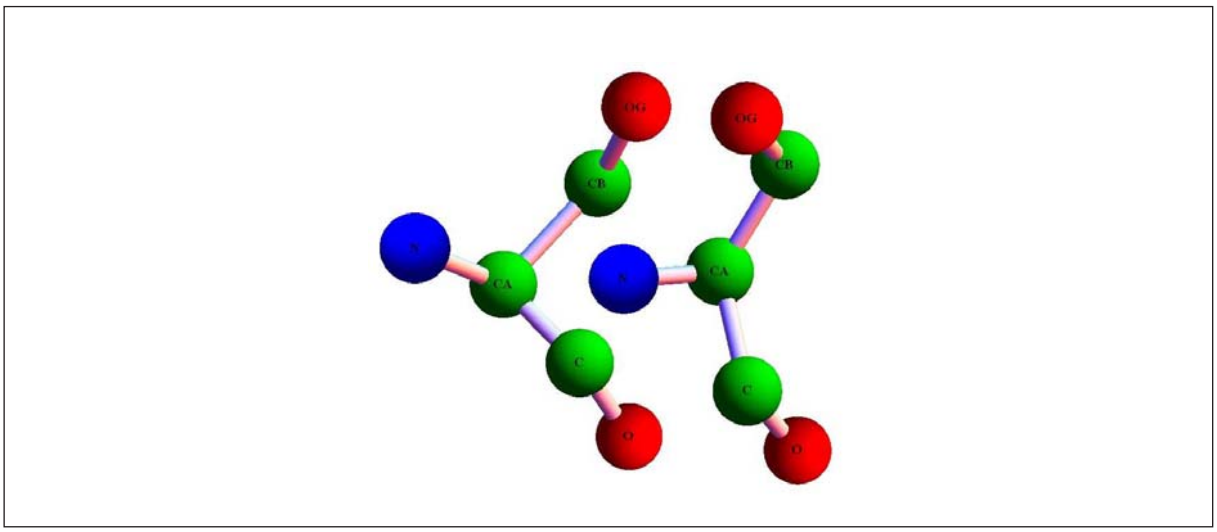

Рис. 15. Файл 2W0L: пара последовательных аминокислот, для которой отклонение расстояния между альфа-углеродами максимальное, равное 32.62% (альфа-углероды находятся слишком близко друг к другу).

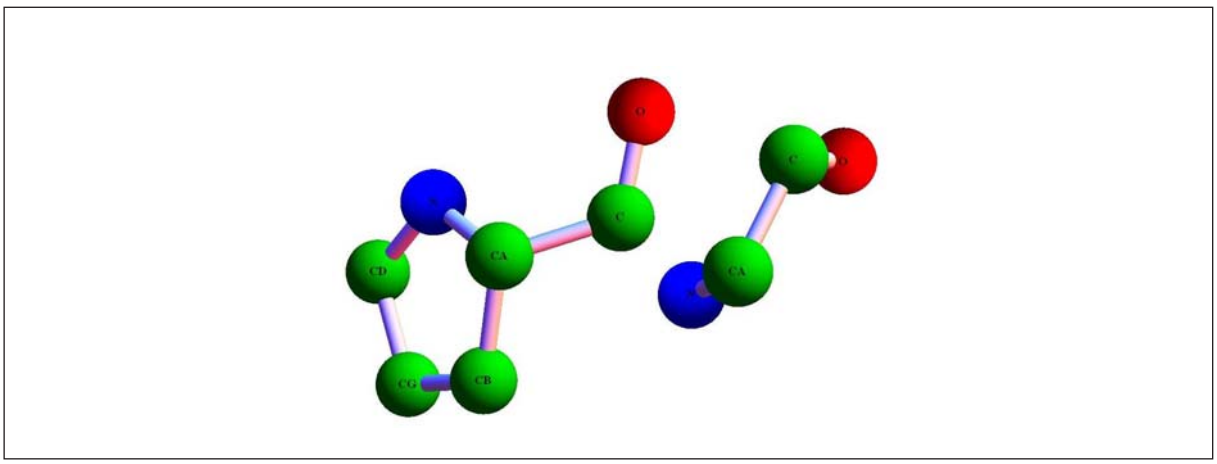

Рис. 16. Файл 2HWD: пара последовательных аминокислот, для которой отклонение расстояния между альфа-углеродами второе по величине, равное 23.06% (альфа-углероды также находятся слишком близко друг к другу).

полученные рентгеноструктурным анализом, не содержат координат водорода. Это немного упрощает нам задачу, так как в предыдущей статье, помимо обычного закона плоскости для шестерки атомов (CA, C, O и CA, N, H), мы были вынуждены вводить еще и обобщенный закон — для случая с пролином.

Сформулируем *закон плоскости* в новых терминах: *пептидная группа*, состоящая из атомов CA, C, O из i-ой аминокислоты и атомов N, CA из $(i+1)$-ой аминокислоты, лежит в одной плоскости.

Раньше для определения степени плоскости пептидной группы мы пользовались объемом выпуклой оболочки, натянутой на нее. Это вызвало некоторые затруднения, связанные с определением граничных значений, начиная с которых отклонения можно считать существенными. Поэтому было решено модифицировать наш алгоритм и использовать в качестве индикатора плоскости пептидной группы максимум из расстояний от N и CA из $(i+1)$-ой аминокислоты до плоскости, проходящей через три оставшихся атома (CA, C, O из i-ой аминокислоты). Если значение максиму-

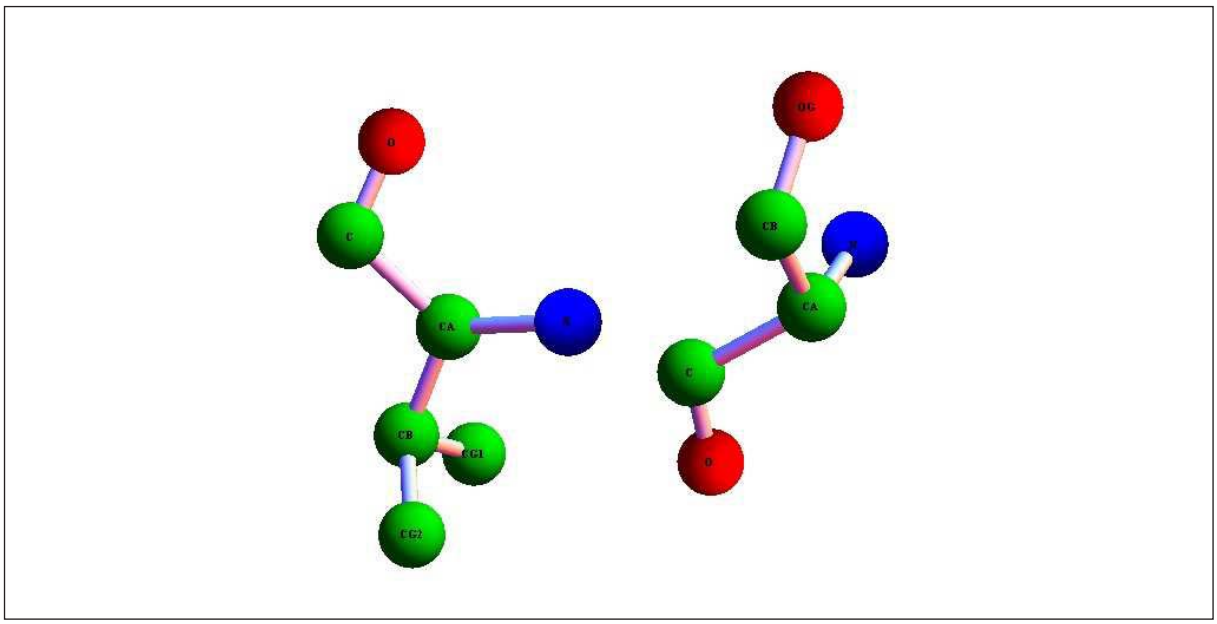

Рис. 17. Файл 1R59: пара последовательных аминокислот, для которой отклонение расстояния между альфа-углеродами четвертое по величине, равное 22.3% (на сей раз альфа-углероды находятся далеко друг от друга).

ма близко к нулю, то соответствующую пептидную группу можно считать плоской, иначе — нельзя.

Приведем, для сравнения, данные для списка аминокислот без цис-конфигураций и вместе с ними:

(1) среднее отклонение от закона плоскости в базе только с транс-конфигурациями равно 0.0758039, в полной базе — 0.0780234;

(2) соответствующие минимальные и максимальные отклонения для первой базы — $\{0.0101321, 2.11352\}$, для второй — $\{0.0101321, 2.34802\}$;

(3) графики распределений для первой и второй баз приведены на рис. 18 и 19 (заметим, что они очень похожи за исключением шкал);

(4) списки точных значений этих графиков:

```
{2362, 3744, 2142, 2863, 2623, 1422, 474, 218, 119, 80, 45, 42,
  22, 16, 5, 3, 0, 0, 0, 0, 0, 1}
{3329, 6748, 3664, 5005, 5371, 2992, 1246, 552, 286, 193, 127,
  133, 89, 51, 29, 14, 7, 4, 4, 0, 2, 1, 1, 1}
```

(5) изображения пар аминокислот, на которых достигается максимальное отклонение от закона плоскости для базы без цис-конфигураций (рис. 20) и вместе с ними (рис. 21).

Рассмотрим только те полипептиды, у которых отклонение от закона плоскости превышает значения 1.4. Таких файлов оказалось немного, а именно, 12 штук:

`1DST,1EGC,1GH4,1HXU,1Q5T,2AQX,2CTS,2W9Q,3LY3,3TPI,4CTS,8API.`

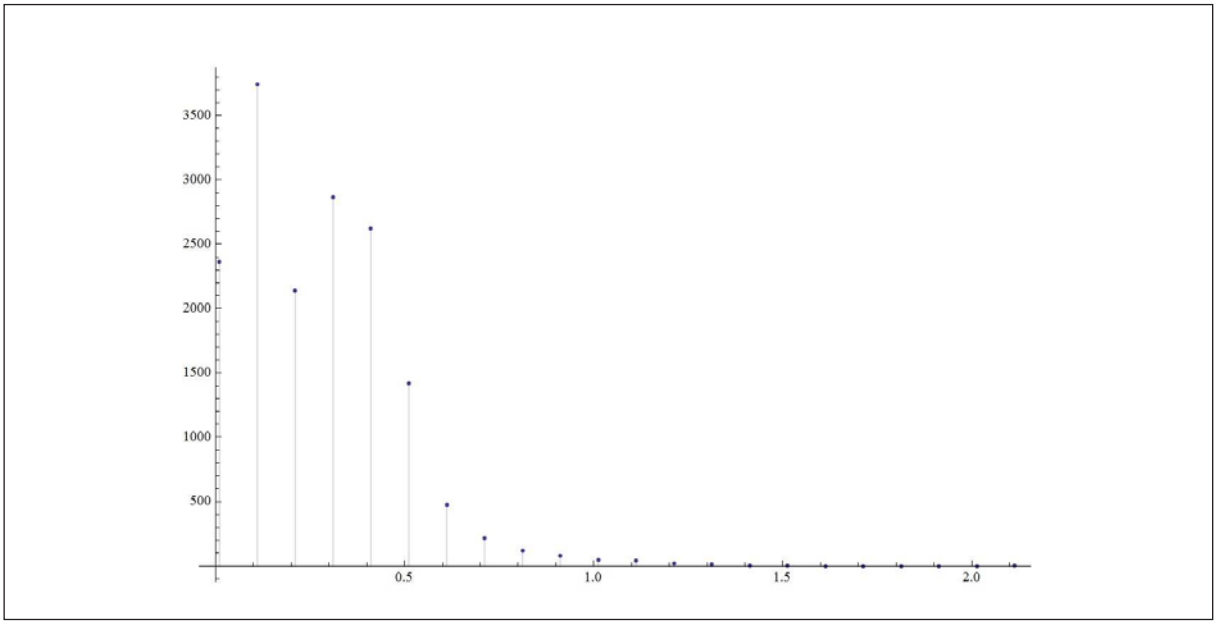

Рис. 18. Распределение отклонений от закона плоскости для базы без цисконфигураций: абсцисса величина отклонения, ордината количество файлов.

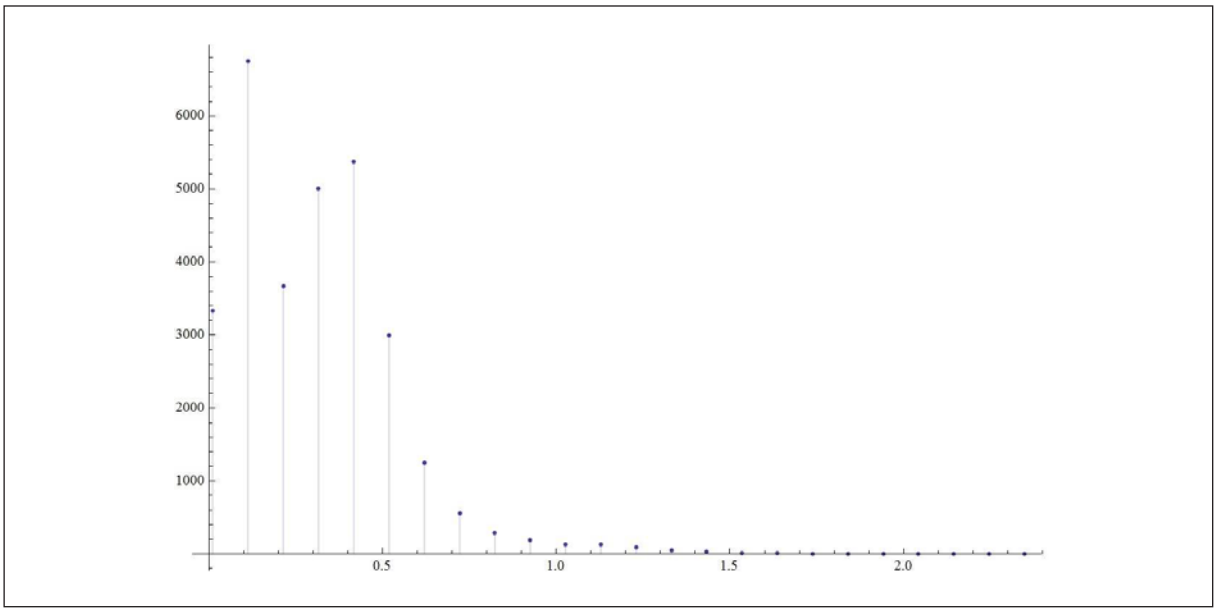

Рис. 19. Распределение отклонений от закона плоскости для полной базы (с цисконфигурациями): абсцисса величина отклонения, ордината количество файлов.

Во время обсуждения со специалистами возникла следующая гипотеза: *наиболее существенные отклонения от закона плоскости происходят на парах последовательных аминокислот, располагающихся либо вне участков спиралей и листов, либо на их границах.*

Чтобы проверить эту гипотезу мы вычислили для полученных 12 файлов процент отклонений, превышающих 1.4 и попадающих либо на спирали, либо на листы. Численные значения оказались равны 33.3333 и 0 соответственно. Для наглядности мы построили графики отклонений от закона плоскости (абсцисса соответствует номеру

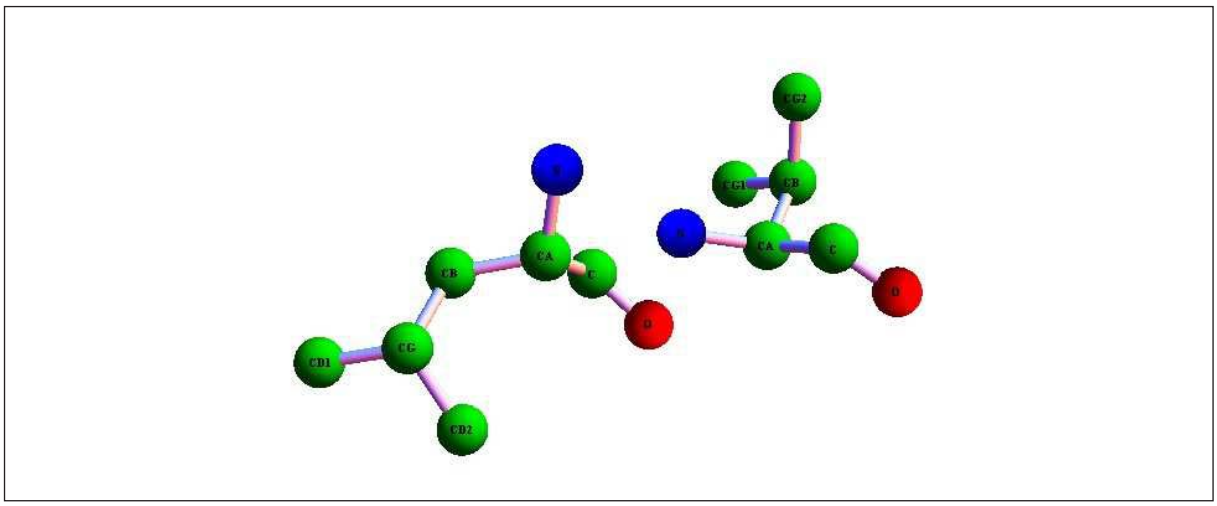

Рис. 20. Пара аминокислот, на которой достигается максимальное отклонение от закона плоскости в базе без цис-конфигураций (файл 1GH4).

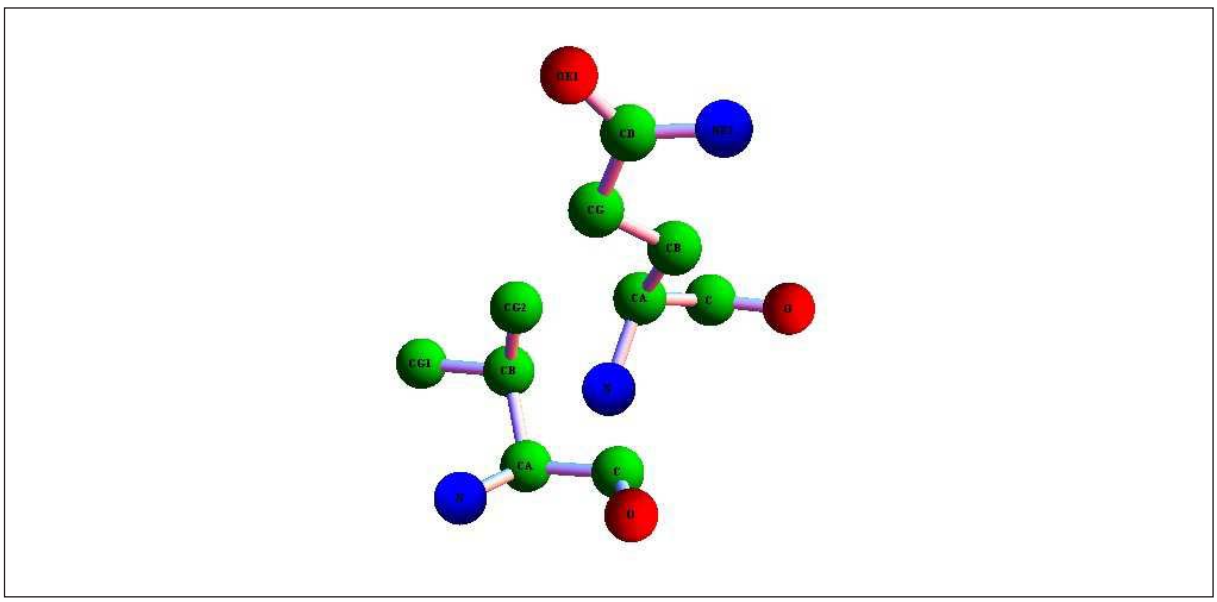

Рис. 21. Пара аминокислот, на которой достигается максимальное отклонение от закона плоскости в полной базе, включающей цис-конфигурации (файл 1SIE).

из вышеуказанных 12 полипептидов и отметили на них участки спиралей красным цветом, а листов — синим, см. рис. 22.

Приведем еще немного данных для проверки гипотезы в следующем формате:

```
{
 граничное значение отклонений;
 количество полипептидов с отклонениями, превышающими граничное;
 {процент отклонений, больших граничного и попадающих на спирали,
  процент отклонений, больших граничного и попадающих на листы}
}:

{1; 137; {16.7883, 10.9489}}
{0.9; 222; {13.0631, 10.8108}}
```

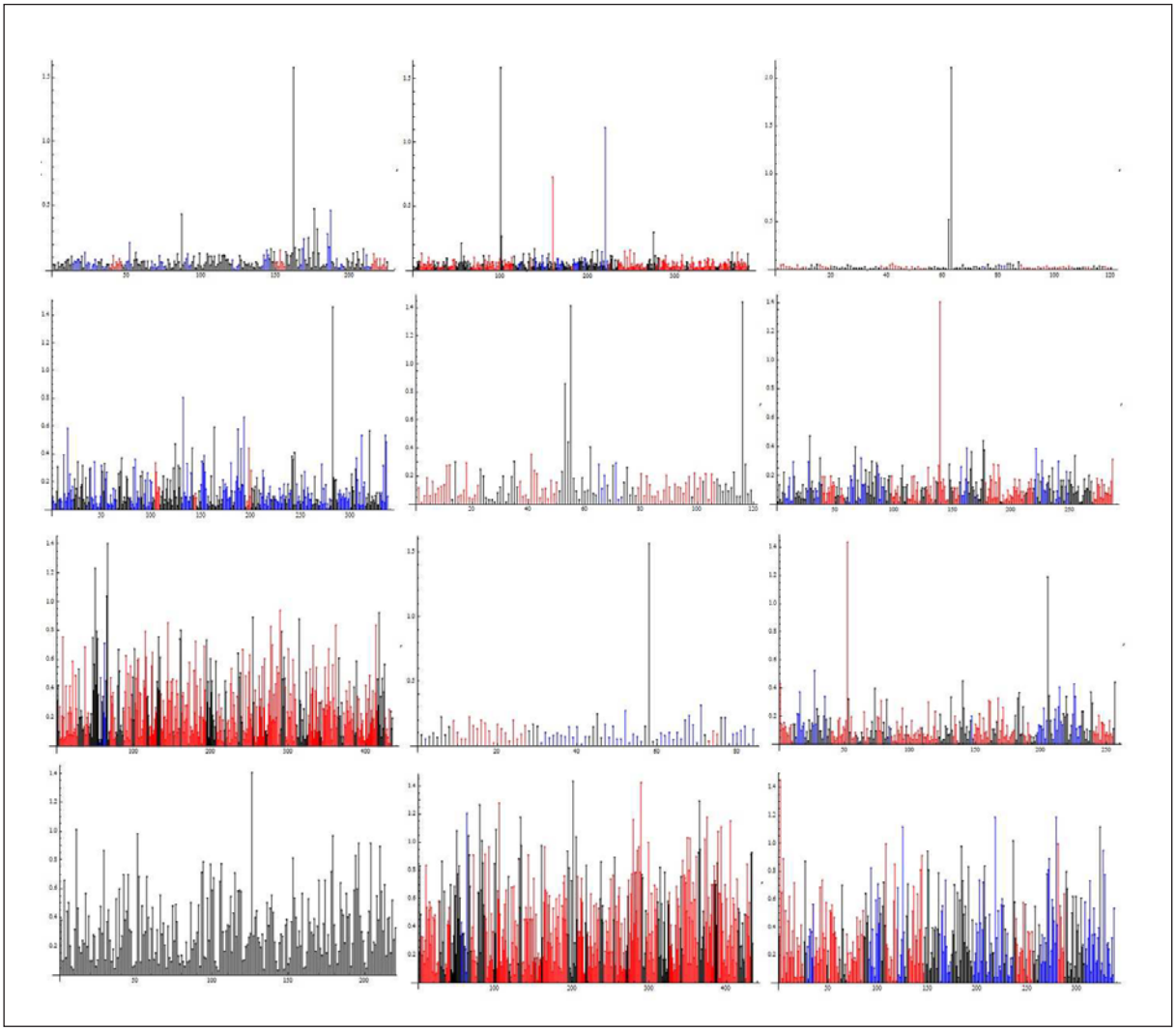

Рис. 22. Графики отклонений от закона плоскости для 12 полипептидов, описанных выше, с указанием участков спиралей (красный цвет) и листов (синий цвет).

```
{0.8; 354; {17.2316, 14.1243}}
{0.7; 585; {20.1709, 14.5299}}
{0.6; 1127; {20.6744, 20.7631}}
{0.5; 2666; {21.2303, 29.4449}}
{0.4; 5365; {29.8788, 43.2992}}
```

Приведенные нами данные показывают, что выдвинутая гипотеза выполняется плохо даже на больших граничных значениях (более 1).

Проанализировав изображения пар аминокислот с отклонениями, приблизительно равными, соответственно, 0.4 и 0.5, было принято решение в качестве существенного значения отклонения взять 0.5, так как оно все еще заметно на глаз, в отличие от 0.4 (см. рис. 23 и рис. 24). В результате, мы исключили еще 2666 файлов.

3 Выводы

Таким образом, в данной работе

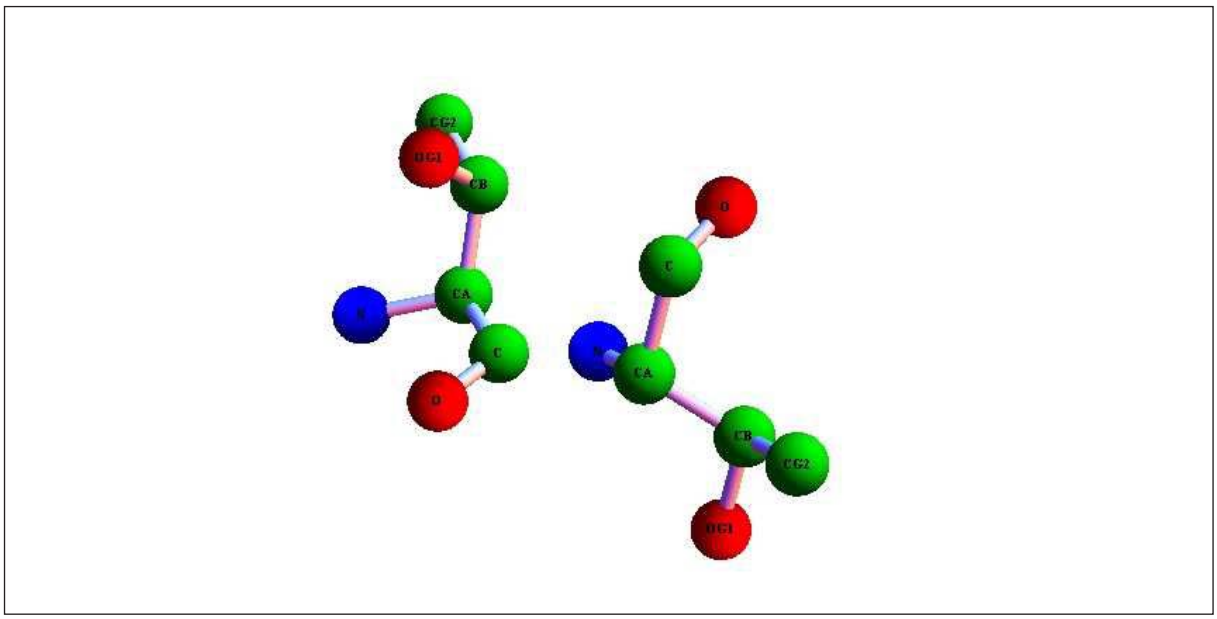

Рис. 23. Пара последовательных аминокислот с отклонением от закона плоскости, приблизительно равным 0:5: непланарность пептидной группы все еще хорошо заметна.

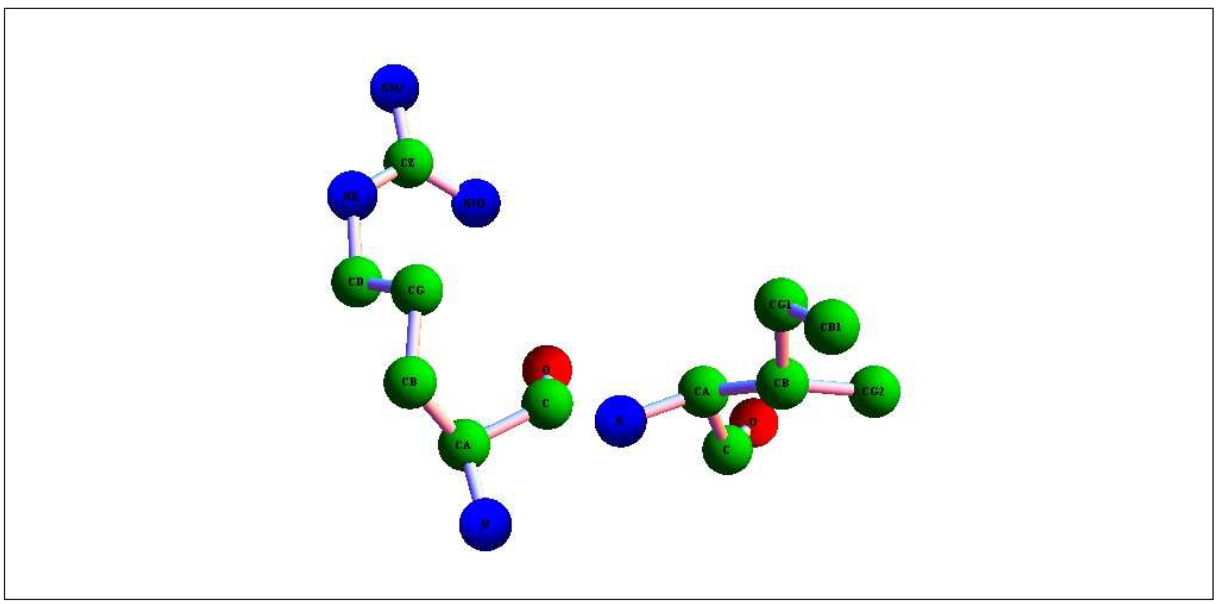

Рис. 24. Пара последовательных аминокислот с отклонением от закона плоскости, приблизительно равным 0:4: пептидная группа выглядит вполне плоской.

(1) демонстрируется, что часть Protein Data Bank, полученная с помощью рентгеноструктурного анализа, так же, как и в случае с ядерно-магнитным резонансом, требует существенной чистки и доработки;

(2) выясняется, что наличие цис-конфигураций не такое уж редкое событие;

(3) показывается, что нарушение закона плоскости, скорее всего, возникает не из-за неточности измерений, а является одним из реальных феноменов.

Список литературы

[1] Иванов А.О., Мищенко А.С., Тужилин А.А. *Геометрия аминокислот и полипептидов*. Наноструктуры. Математическая физика и моделирование, 2014, **10** (1), 49–76, http://nano-journal.ru [in Russian].

[2] Pauling L., Corey R.B., and Branson H.R. *The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain*. Proc. Natl. Acad. Sci., USA, 1951, **37**, 205–211.

GEOMETRY OF AMINO ACIDS AND POLYPEPTIDES: X-RAY CASE

E.A. Vilkul, A.A. Tuzhilin

Lomonosov Moscow State University

tuz@mech.math.msu.su

Received 02.10.2014

The present paper continues testing Protein Data Bank started in [1]: now we are interested in the part of this base which was obtained by means of X-ray analysis. We show that the "pathologies" observed by us in the case of NMR are presented here as well. Moreover, we demonstrate that good resolution (parameter RESOLUTION) does not guarantee the absence of the "pathologies". Finally, we discuss the disturbance of the famous Pauling plane law [2] and show that the high level of disturbance can appear in helixes and sheets.