ПРОДОЛЖАЕМ ЧИСТИТЬ PROTEIN DATA BANK

А.А. Тужилин

Кафедра дифференциальной геометрии и приложений механико-математического факультета МГУ имени М.В.Ломоносова

tuz@mech.math.msu.su

Поступила 14.05.2016

В настоящей работе мы продолжаем исследование геометрии расстояний между атомами белков, начатой в [1] и [2]. В качестве базы данных координат мы используем Protein Data Bank. Мы приведем ряд новых соображений, позволяющих понять, что файл с координатами атомов белка содержит ошибки. В результате мы построим новую базу, которая в следующих работах будет использована для изучения геометрии графа ковалентных связей и его связи с другими экстремальными графами, например, с минимальными остовными деревьями. Эта работа продолжит изучение геометрии ломаных, построенных на последовательных альфа-углеродах, см. [3], [4].

УДК 514.8, 51-76, 57.087

1 Введение

Может, мы обидели кого-то зря,
Календарь закроет этот лист.
К новым приключениям спешим, друзья,
Эй, прибавь-ка ходу, машинист!

Из детской песенки “Голубой вагон”.

В предыдущих работах автора и его коллег [1], [2] был получен ряд результатов, демонстрирующих необходимость очень внимательного отношения к файлам, выложенным в знаменитой базе данных Protein Data Bank (в дальнейшем, для краткости, PDB). Оказалось, что многие файлы содержат как путаницу в обозначениях, так и грубые ошибки в приводимых координатах атомов белков. Примером первого типа недочетов может служить глицин, “атомный состав” которого в файлах из PDB, полученных с помощью ядерно-магнитного резонанса (ЯМР), имеет 31 представление:

```
{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}}.
```

Яркими примерами ошибок второго типа являются файлы 2PDE.pdb и 2I2J.pdb, см. рис. 1 и рис. 2, на которых приведены изображения некоторых аминокислот, построенных по координатам из этих файлов.

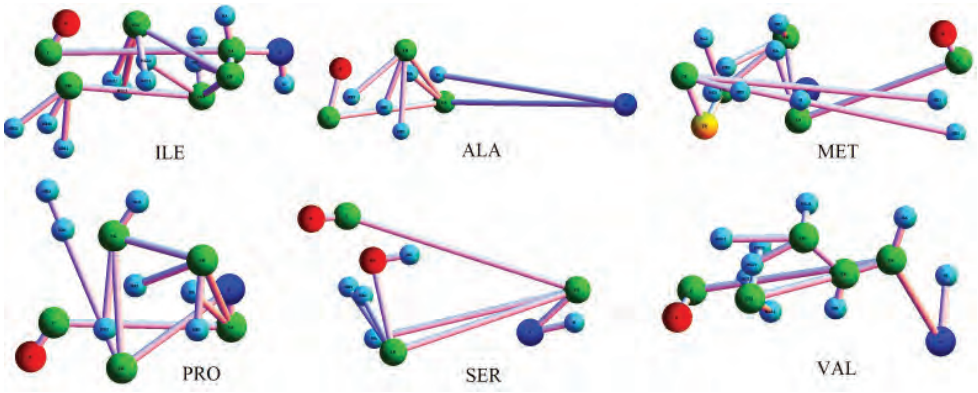

Рис. 1: Некоторые “аминокислоты” из файла 2PDE.pdb (1992 год).

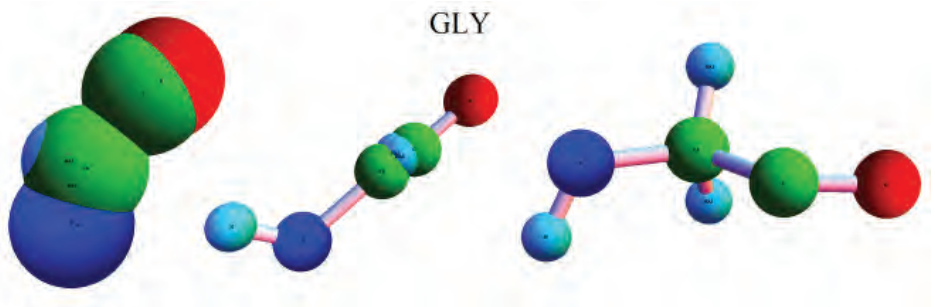

Рис. 2: Слева: два “глицина” из файла 2I2J.pdb (2006 год). Справа — “стандартный” глицин.

Так как первоначальной целью нашего исследования было изучение статистики геометрических свойств конформаций белков, нам нужна была более-менее надежная база данных белковых трехмерных структур. Обнаруженные проблемы с использованием PDB вынудили нас выкинуть примерно 2/3 файлов.

Любопытна реакция, вызванная нашей работой. Одни специалисты, увидев рис. 1, 2 и целый ряд других похожих иллюстраций, мгновенно отреагировали хорошо известной фразой “этого не может быть потому, что этого не может быть никогда” и посоветовали нам пойти искать свои ошибки. На это мы привели следующий фрагмент из файла 2I2J.pdb

```
ATOM     14  N   GLY A   2     -13.525  -1.614  -1.070  1.00 5.70           N
ATOM     15  CA  GLY A   2     -13.174  -1.687  -0.986  1.00 5.18           C
ATOM     16  C   GLY A   2     -12.623  -1.567  -0.783  1.00 4.25           C
ATOM     17  O   GLY A   2     -12.514  -1.508  -0.694  1.00 4.43           O
ATOM     18  H   GLY A   2     -13.535  -1.585  -1.035  1.00 5.83           H
ATOM     19  HA2 GLY A   2     -13.317  -1.721  -1.013  1.00 5.48           H
ATOM     20  HA3 GLY A   2     -13.165  -1.791  -1.040  1.00 5.56           H
```

взяли координаты $(-13.525, -1.614, -1.070)$ азота N, координаты $(-13.535, -1.585, -1.035)$ водорода H и на калькуляторе публично вычислили расстояние между этими атомами: оно оказалось приблизительно равным 0.0465403 ангстрема (эти азот и водород относятся к левой аминокислоте на рис. 2, где они почти совместились).

Кроме того, мы привели пример как полученного нами изображения аминокислот, так и результата визуализации этих аминокислот с помощью программы с сайта PDB, см. рис. 3.

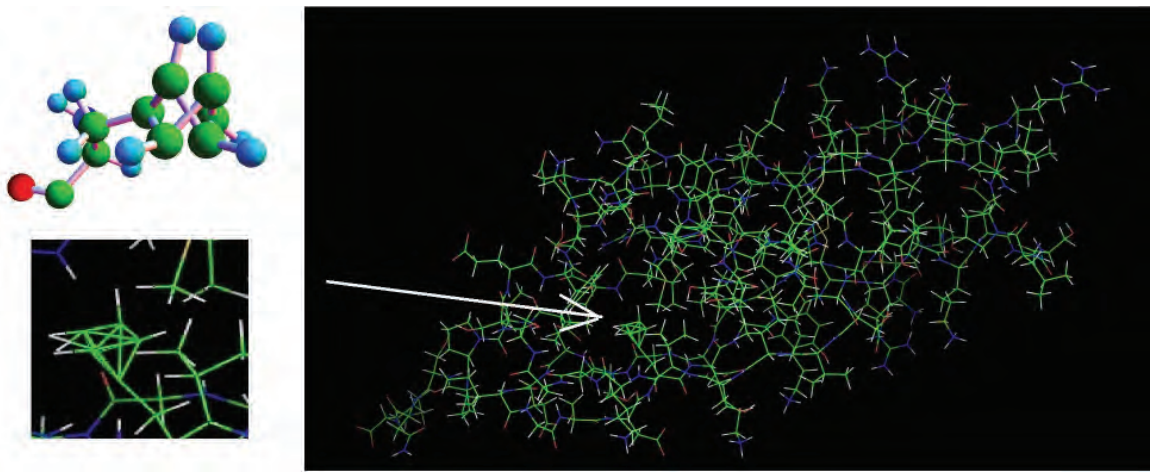

Рис. 3: Файл 2NPL.pdb, 2006 год.

Последнее свидетельствует о том, что обнаруженные нами “недочеты” могли быть выявлены стандартными средствами с PDB.

Другие специалисты сказали, что такие ошибки были лишь на ранней стадии развития, а затем мощные методы верификации все ошибки исправили. Особенно безупречной, по мнению специалистов, является программа CYANA. В ответ на это заявление мы мгновенно привели следующий пример, рис. 4.

Этот же файл может служить иллюстрацией к обсуждению различных моделей (для ЯМР): некоторые специалисты считают, что первая модель получается усреднением, поэтому тут возникновение ошибок вполне понятно. Обратите внимание, что на рис. 4 изображена аминокислота из **4-ой модели**. На рис. 5 приведен пример, в котором все 16 моделей содержат некорректно вычисленную аминокислоту.

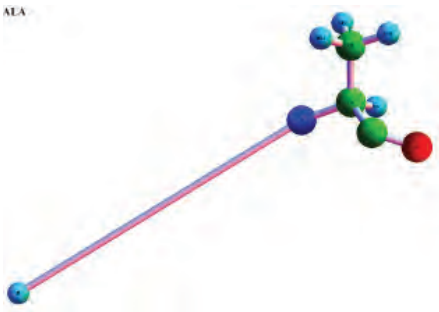

Рис. 4: Файл 2L5R.pdb, 4-ая модель, 19-ая аминокислота, 2010 год, список программ: TOPSPI 2.0, SPARKY 3.114, CYANA2.0.

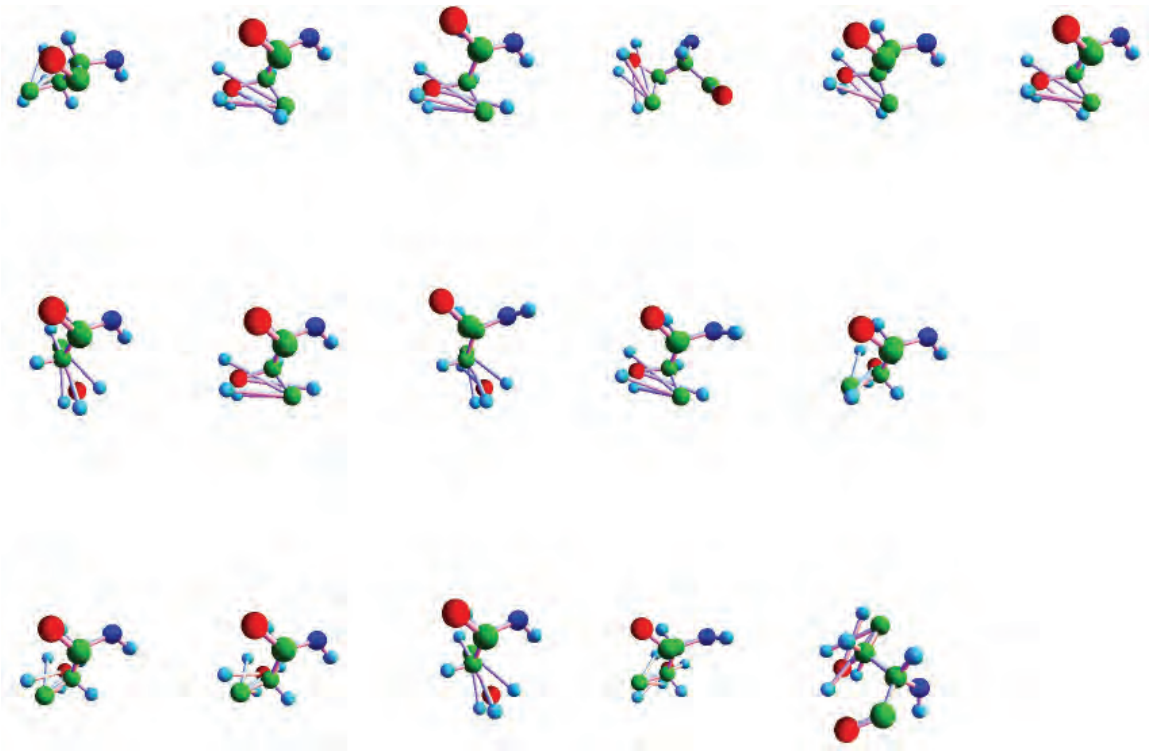

Рис. 5: Файл 1KWD.pdb (2002 год), в котором 10-ая аминокислота имеет некорректный вид во всех 16-ти моделях.

Справедливости ради заметим, что описанная только что сложность с первой моделью иногда действительно имеет место, рис. 6.

Обратим внимание еще на одну забавную особенность: в некоторых случаях визуализация на PDB “скрывает ошибки”, см. рис. 7.

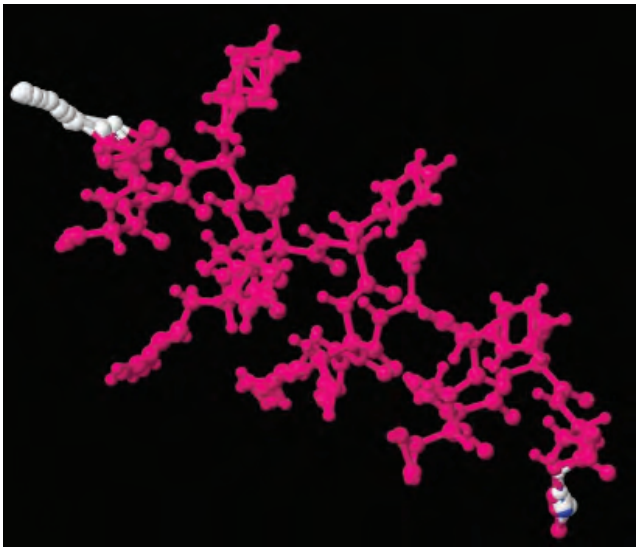

Рис. 6: Файл 2I2J.pdb, 2006 год, “плохая” только первая модель (остальные 33 модели — “хорошие”).

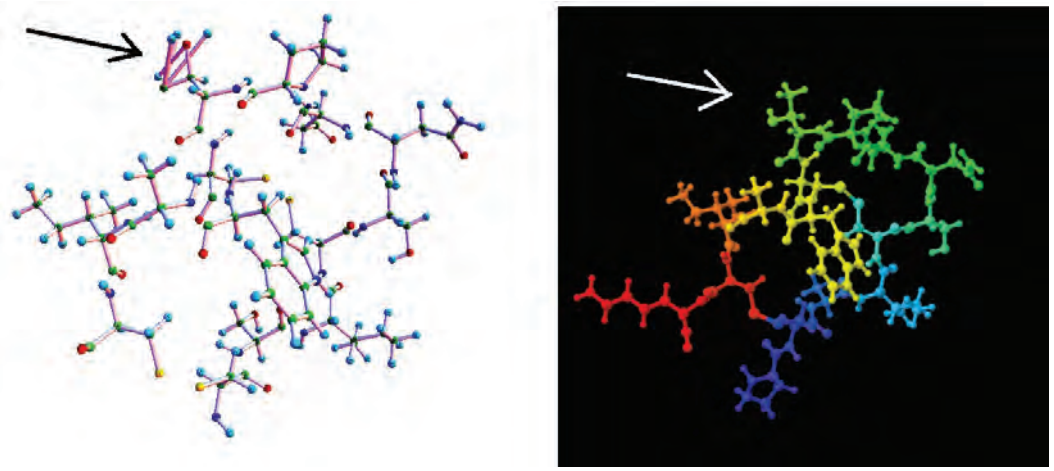

Рис. 7: Файл 1KWD.pdb, 2002 год, 10 аминокислота, неправильное обозначение водородов; на стандартном визуализаторе из PDB ошибка не видна.

Предыдущие обсуждения касались ЯМР. Многие специалисты согласились на том, что рентгено-структурный анализ (РСА) дает более точные результаты, причем тем точнее, чем меньше параметр RESOLUTION (разрешение). Чтобы понять, какое разрешение считать маленьким, мы построили график количеств файлов (ордината), имеющих те или иные значения RESOLUTION (абсцисса), см. рис. 8.

Из графика на рис. 8 видно, что максимальное число файлов соответствует разрешению

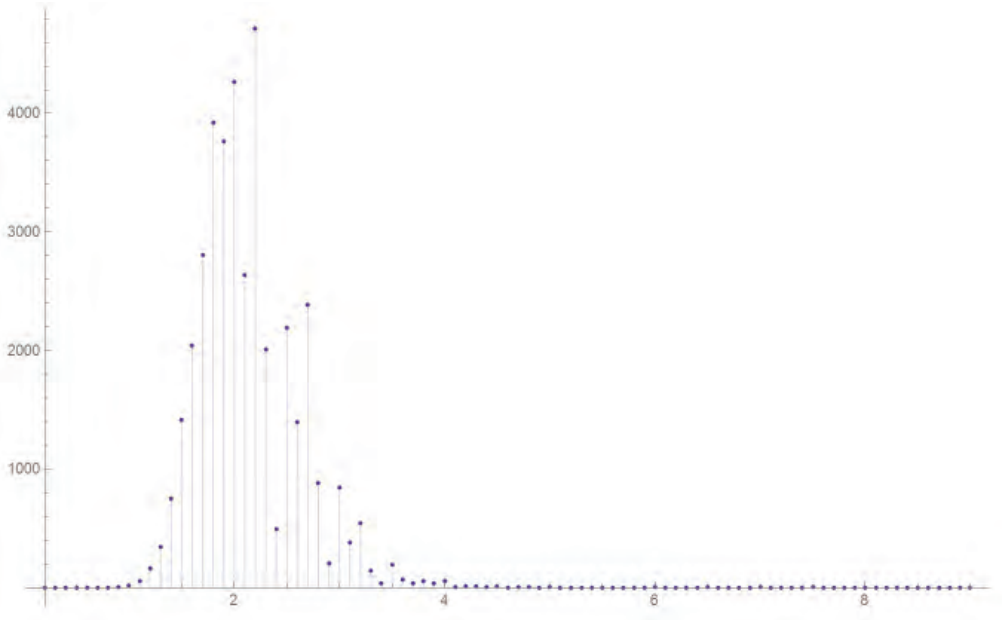

Рис. 8: График количеств файлов (ордината), имеющих те или иные значения RESOLUTION (абсцисса).

2.2, поэтому, учитывая разброс значений, разрешение 1.5 можно было бы отнести к маленьким. Приведенные рассуждения делают следующий пример контрпримером.

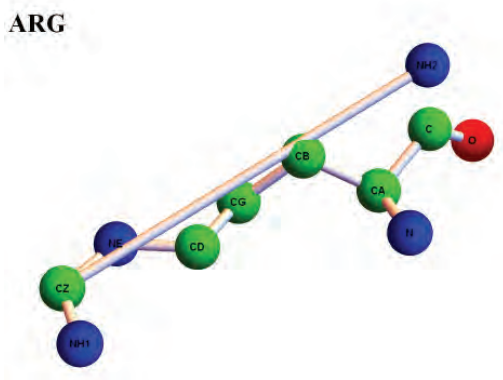

Рис. 9: Файл 1W5Q, 2004 год, RESOLUTION = 1.4.

Интересна также история ряда попыток публикации наших результатов. Мы начали с arxiv, где у нас размещены многочисленные наши математические работы, и где ни разу у нас не было проблем опубликовать очередной математический результат. Однако, к нашему искреннему удивлению, на сей раз статья принята не была: нам посоветовали обратиться в рецензируемый журнал. Когда же мы послали статью в одни из таких журналов, рецензент сообщил нам, что информация о наличии ошибок в файлах из PBD давно и хорошо известна, и даже привел ссылки [5], [6], [7]. В первой из них, самой ранней (1996 год), в разделе "Correspondence", содержится очень краткое, без единой иллюстрации, описание проделанной работы: говорится о проведенном вычислении средних длин ковалентных связей, средних углов между смежными ковалентными связями, отклонений от средних и т.п.; утверждается, что было найдено 76 классов разных проблем; приводится ссылка `http://www.sander.embl-heidelberg.de/johnny/`, видимо, на более подробное изложение; *в настоящее время эта ссылка не работает.*

Как видно из приведенного выше обсуждения, работа [5] осталась незамеченной рядом ведущих специалистов. Тем не менее, публикация такого типа позволяет "сведущему" рецензенту задать вопрос: а что нового есть в Вашей работе? И, естественно, нет никакой возможности объяснить, что вербальное сообщение о 76 проблемах в разделе "Correspondence" имеет меньше шансов на распространение, чем впечатление от рис. 1 и 2.

Итак, в предыдущих работах мы

(1) вычислили наиболее часто встречающиеся "атомные составы" внутренних (неконцевых) аминокислот и отобрали только те файлы, в которых все внутренние аминокислоты такие;

(2) выбросили все файлы с различными включениями (содержащие HETATM), описанными в разделе ATOM–TER;

(3) выбросили все файлы со слишком большими отклонениями от средних длин ковалентных связей;

(4) выбросили все файлы со слишком большими отклонениями от средних углов между смежными ковалентными связями.

Были также проанализированы частоты появления цис-конфигураций, степень плоскости пептидной группы, распределение расстояний между последовательными альфа-углеродами, подвижность аминокислот фиксированного типа.

Как уже было отмечено, мы получили "чистую" базу данных, состоящую для ЯМР примерно из 2500 файлов, а для РСА — из 30000 файлов. Одной из геометрических идей, которую мы хотели проверить на сей раз, является принадлежность ребер минимального остовного дерева, построенного на координатах атомов, графу ковалентных связей. Начав экспериментировать, мы обнаружили некоторые сбои: оказалось, что в ряде случаев ковалентно несвязанные ато-

мы находятся слишком близко друг к другу. Мы провели анализ не только длин ковалентных связей, как в работах [1] и [2], но и расстояний между всеми остальными парами атомов и, в результате, обнаружили еще ряд файлов, которые, скорее всего, следует отнести к содержащим ошибки. Именно об этом исследовании мы расскажем в настоящей статье. В следующей работе мы применим вычищенную с помощью описываемых здесь методов базу к исследованию экстремальной геометрии белков.

Автор выражает глубокую благодарность В.Л.Голо, фактически являющемуся для меня локомотивом написания биологических статей, а также моим коллегам и родственникам Г.А.Армееву, Е.А.Вилкул, А.О.Иванову, Ф.Ю.Попеленскому, Ж.Р.Тужилиной, К.В.Шайтану за многочисленные полезные обсуждения.

Работа выполнена при поддержке РНФ, соглашение № 14-50-00029.

2 Распределение интервалов длин разных типов ковалентных связей

В данной статье мы изучаем файлы, полученные исключительно с помощью ЯМР.

Эксперименты с минимальными остовными деревьями привели нас к необходимости выяснить, в каких пределах может меняться длина той или иной ковалентной связи, и какими бывают длины остальных расстояний (между несвязанными ковалентно атомами). Мы разбили все ковалентные связи на 100 видов, в соответствии с номенклатурой названий атомов в белках, и для каждого вида вычислили интервал между минимальным и максимальным значением длины этой связи по всем белкам. Вот начальный отрезок того, что мы получили:

```
{"NZ" -> "HZ2", {0.9686681578332172, 1.084642337362875}}
{"NZ" -> "HZ3", {0.9689220814905612, 1.0944066885760526}}
{"CZ" -> "OH", {1.3299056357501453, 1.4159099547640697}}
{"OH" -> "HH", {0.9379370981041326, 1.0296382860014484}}
```

Для наглядности, мы решили изобразить эти интервалы отрезками, пометив для каждого из них соответствующей внутренней точкой среднюю длину данной ковалентной связи (сделать последнее предложил К.В.Шайтан). Результаты представлены на рис. 10.

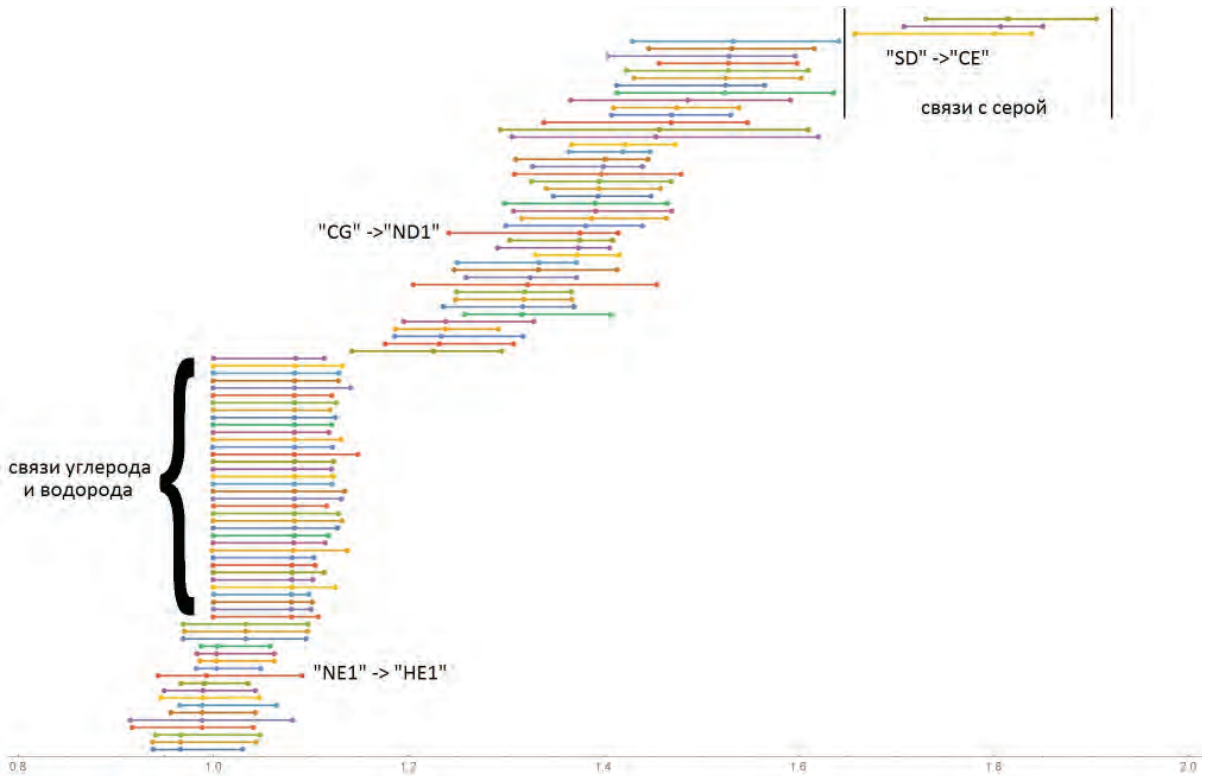

Рис. 10: Распределение интервалов длин ковалентных связей.

Отметим удивительную особенность полученной диаграммы: блок, соответствующий всем связям между углеродом и водородом, выровнен по левому краю. Г.А.Армеев высказал гипотезу, что этот эффект мог быть вызван жестким фиксированием минимальной длины этих связей в алгоритме нахождения конформации (для ряда файлов, в которых встретились самые короткие углеродно-водородные связи). Мы решили посмотреть, в каких файлах достигается эта левая граница.

Заметим, что во всем "углеродно-водородном блоке" левая часть интервала существенно больше правой. Этот эффект опять же может быть легко получен, если в одном из файлов имеются очень короткие углеродно-водородные связи. Мы решили отловить файлы, "виновные" в нарушении симметрии, и написали процедуру, которая выдает список всех файлов, содержащих связи, удлиняющие одну из частей интервала. Вот некоторые результаты для углеродно-водородного блока:

```
связь "CE3" -> "HE3":
"2MEX", {{"CE3" -> "HE3", 0.999462}}
"2JRL", {{"CE3" -> "HE3", 0.999725}}}
"2DJM", {{"CE3" -> "HE3", 1.033645}}}
............

связь "CH2" -> "HH2":
"2MEX", {{"CH2" -> "HH2", 0.999708}}
"2JRL", {{"CH2" -> "HH2", 1.000081}}
"1WRT", {{"CH2" -> "HH2", 1.055652}}
............

связь "CD1" -> "HD1":
"2MEX", {{"CD1" -> "HD1", 0.999255}, {"CD1" -> "HD1", 0.999589}, ...}
"2JRL", {{"CD1" -> "HD1", 0.999373}, {"CD1" -> "HD1", 0.999549}, ...}
"2OYW", {{"CD1" -> "HD1", 0.999702}, {"CD1" -> "HD1", 1.00011}}
"2OYV", {{"CD1" -> "HD1", 0.999925}, {"CD1" -> "HD1", 1.00053}}

связь "CD2" -> "HD2":
"2MEX", {{"CD2" -> "HD2", 0.999551}, {"CD2" -> "HD2", 1.00003}, ...}
"2OYW", {{"CD2" -> "HD2", 0.999709}, {"CD2" -> "HD2", 0.999807}}
"2JRL", {{"CD2" -> "HD2", 0.999774}, {"CD2" -> "HD2", 0.999929}, ...}
"2OYV", {{"CD2" -> "HD2", 0.999862}, {"CD2" -> "HD2", 1.00047}}
"1WM8", {{"CD2" -> "HD2", 1.00014}}
"2CBH", {{"CD2" -> "HD2", 1.02588}}
............

связь "CG" -> "HG3":
"2MEX", {{"CG" -> "HG3", 0.999049}, {"CG" -> "HG3", 0.999448}, ...}
"1WM8", {{"CG" -> "HG3", 0.999391}, {"CG" -> "HG3", 0.999562}, ...}
"2JRL", {{"CG" -> "HG3", 0.999454}, {"CG" -> "HG3", 0.999506}, ...}
"1ZUV", {{"CG" -> "HG3", 0.999535}, {"CG" -> "HG3", 0.999798}, ...}
"2OYV", {{"CG" -> "HG3", 0.999562}, {"CG" -> "HG3", 0.999576}, ...}
"2OYW", {{"CG" -> "HG3", 0.999656}, {"CG" -> "HG3", 0.999974}, ...}
```

Отметим, что для некоторых связей списки оказались очень большими, не смотря на явное смещение среднего значения. Вот, например, начальный отрезок для связи `"CA" -> "HA"`:

```
 "1G10", {{"CA" -> "HA", 0.998062}, {"CA" -> "HA", 0.998759}}
 "1G11", {{"CA" -> "HA", 0.998575}}
 "2MEX", {{"CA" -> "HA", 0.998962}}
 "1X4O", {{"CA" -> "HA", 0.998964}}
 "1WIZ", {{"CA" -> "HA", 0.998968}}
```

```
"1CCV", {{"CA" -> "HA", 0.998969}}
"1M3O", {{"CA" -> "HA", 0.998984}}
"1X6D", {{"CA" -> "HA", 0.998996}}
"1V5J", {{"CA" -> "HA", 0.999057}}
"1V5T", {{"CA" -> "HA", 0.999061}}
.........
```

Мы решили изобразить распределение количеств связей (ордината) с теми или иными длинами (абсцисса) для `"CA" -> "HA"`, рис. 11.

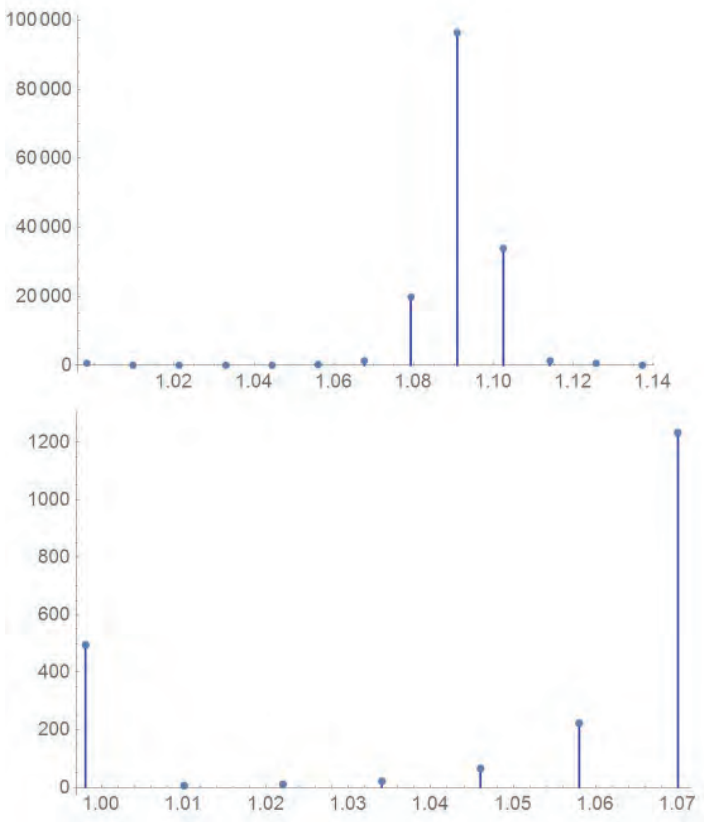

Рис. 11: Распределение количеств связей (ордината) с теми или иными длинами (абсцисса) для "CA"→ "HA".

Из диаграммы 11 видно, что в окрестности “длинного конца” интервала изменения длины связи `"CA" -> "HA"` сосредоточено очень большое число разных файлов, поэтому данный “перекос”, видимо, стоит считать адекватным.

Итак, список претендентов на удаление был получен следующим образом: в него мы поместили все файлы, в которых имеются связи длины не больше 1.01, и около левого конца интервала изменения длины связи сосредоточено “немного” файлов. В результате мы сформировали список из 7 элементов:

```
{"1WM8", "1WT8", "2JRL", "2MEX", "2OYV", "2OYW", "1ZUV"}.
```

Как легко видеть, на диаграмме 10 имеется еще много других типов связей, в которых возникает перекос, например, для связей `"NE1" -> "HE1"`, `"CG" ->"ND1"`, `"SD" -> "CE"`.

Анализ, похожий на тот, который мы провели для `"CA" -> "HA"`, показал, что наиболее подозрительной на наличие ошибок является связь `"SD" -> "CE"`. Ниже мы приводим распределение количеств связей `"SD" -> "CE"` по длинам этих связей, рис. 12 и 13.

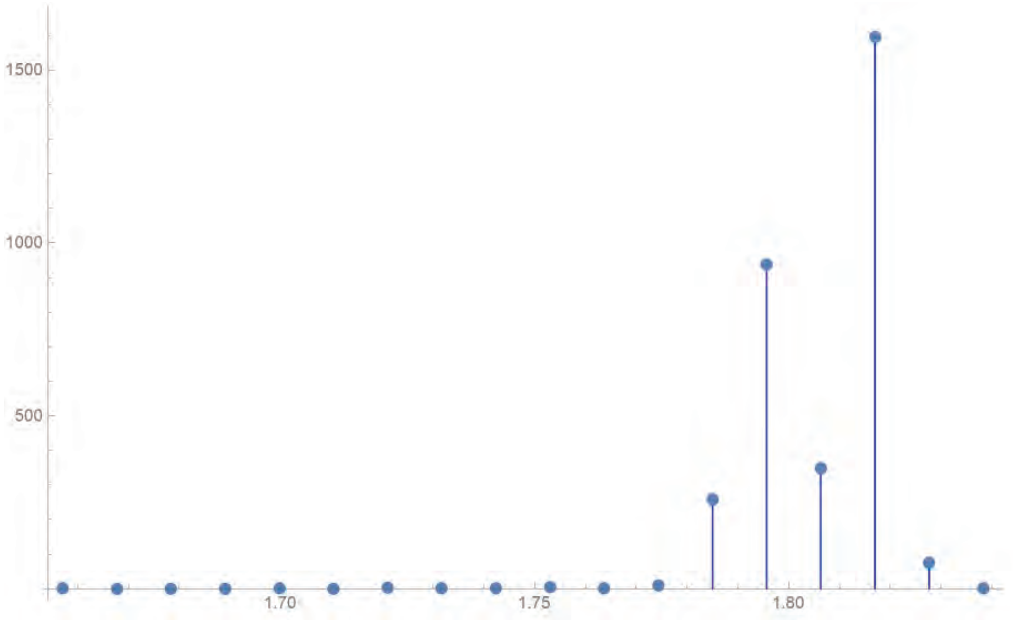

Рис. 12: Распределение количеств связей (ордината) с теми или иными длинами (абсцисса) для "SD"→ "HA".

Из диаграммы 13 видно, что имеется ровно одна кратчайшая связь, относительно изолированная по своей величине. Непосредственное вычисление показывает, что эта связь принадлежит файлу 2LV1.pdb. Впрочем, относительно подозрительными являются также связи, длины которых расположены в пределах между 1.7 и 1.77 ангстрем. Не имея достаточных оснований считать последние файлы ошибочными, мы решили оставить их, а в список для исключения поместить только 2LV1.pdb.

Итак, этот этап “чистки” закончился исключением следующего списка файлов:

```
{"1WM8", "1WT8", "2JRL", "2MEX", "2OYV", "2OYW", "1ZUV", "2LV1"}.
```

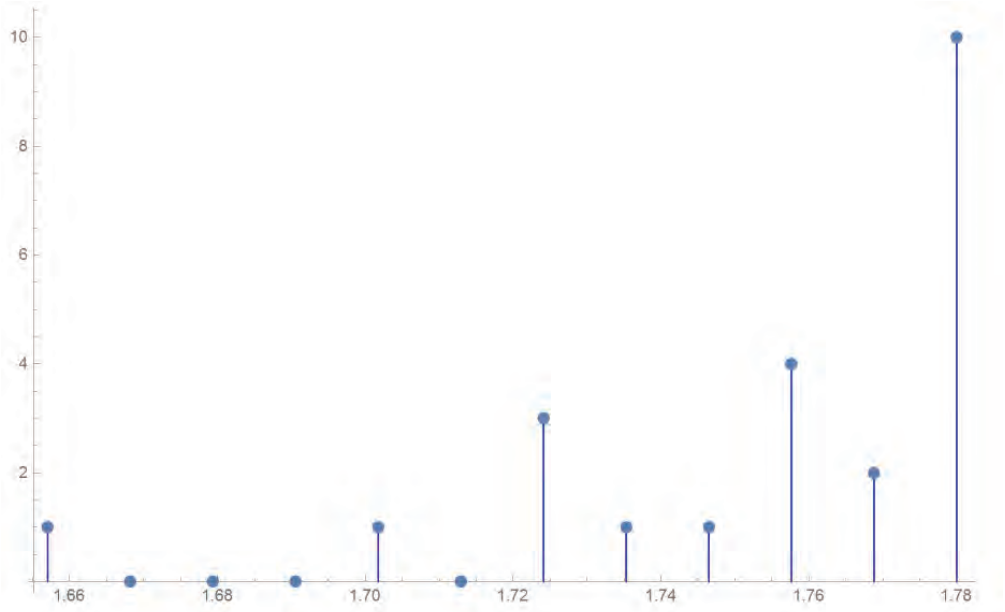

Рис. 13: Распределение количеств связей (ордината) с теми или иными длинами (абсцисса) для "SD"→ "CE".

Мы пересчитали диаграмму распределения интервалов изменения длин связей, см. рис. 14.

Хорошо видно, что в результате мы избавились от выровненных левых концов блока связей углерод-водород, а также сделали распределение длин связи `"SD" -> "CE"` более симметричным.

3 Геометрия пар атомов в одной аминокислоте, не связанных ковалентно

В качестве следующего шага мы решили изучить расстояния между атомами, не связанными ковалентно. Мы начали с пар атомов, принадлежащих одной аминокислоте. На рис. 15 и 16 приведены диаграммы распределений количеств пар по расстояниям между ними.

Из диаграммы 16 видно, что имеется ровно одно очень маленькое расстояние, которое, как оказалось, равно 0.657274 ангстрем, принадлежит файлу 1Q2I.pdb и реализуется в 11-ой аминокислоте (лейцине) между атомами водорода "H"и "HD11".

Исключив из нашей базы файл 1Q2I.pdb, мы добились того, что расстояния между атомами в аминокислоте, не связанными ковалентно, стали не меньше 1.1 ангстрема.

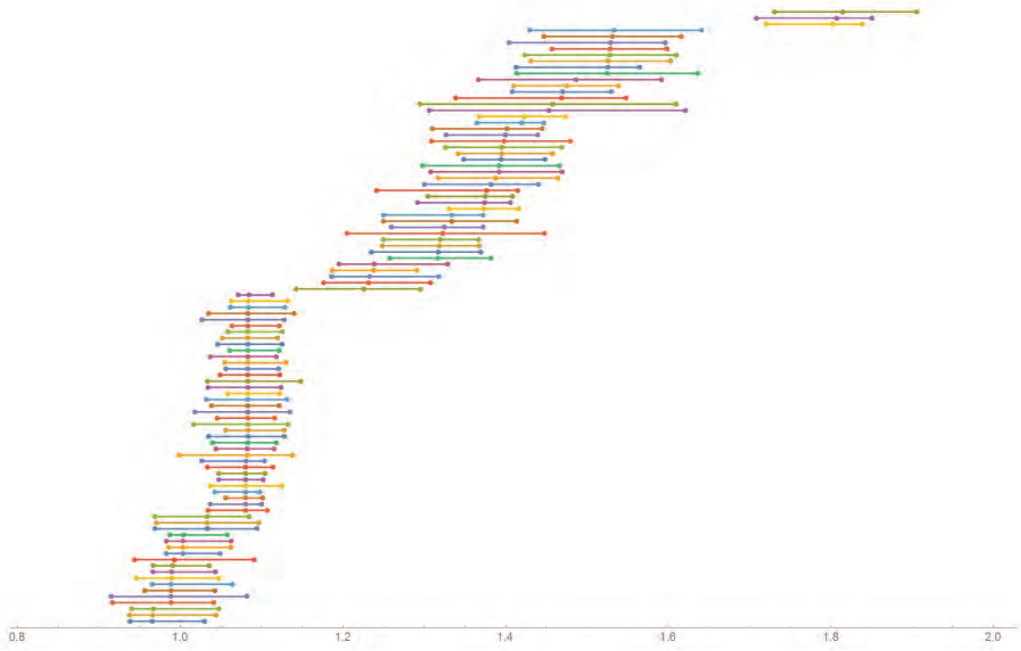

Рис. 14: Распределение интервалов длин ковалентных связей после первой “чистки”.

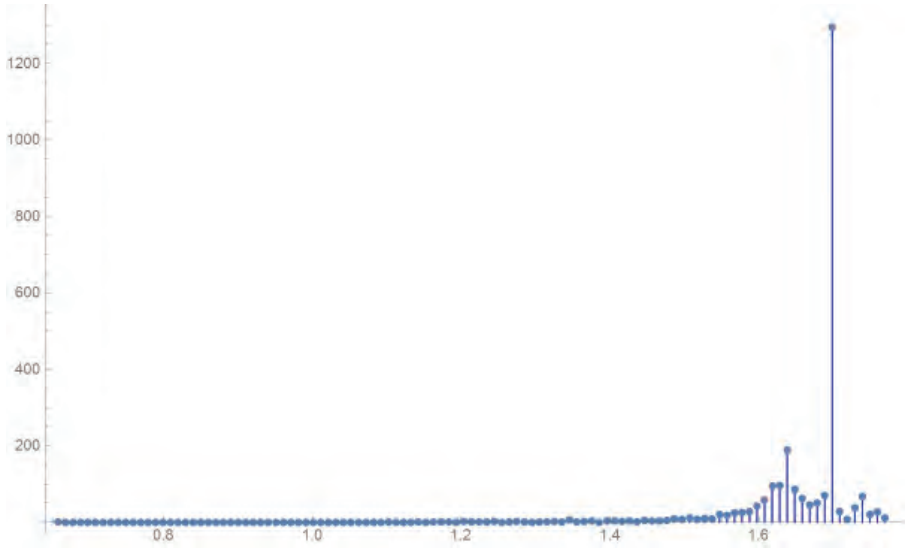

Рис. 15: Распределение количеств пар атомов из одной аминокислоты, не связанных ковалентно, по расстояниям между этими атомами.

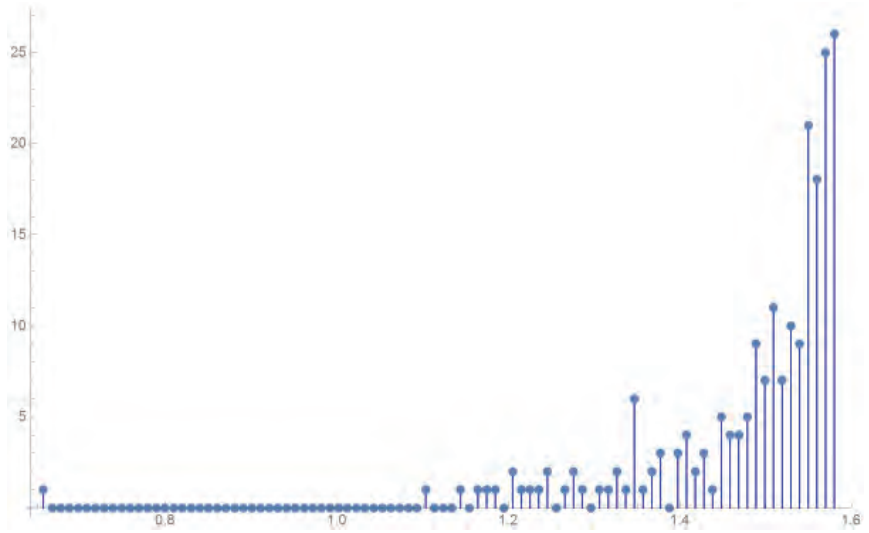

Рис. 16: Распределение количеств пар атомов из одной аминокислоты, не связанных ковалентно, по расстояниям между этими атомами.

4 Геометрия пар атомов в разных аминокислотах, не связанных ковалентно

Заключительный шаг “чистки” состоял в изучении расстояний между атомами, принадлежащими разным аминокислотам и несвязанным ковалентно (единственными ковалентными связями между такими атомами являются пептидные связи `"C" -> "N"`).

Заметим, что таких пар очень много, поэтому прямой перебор всех пар приводит к большим временным затратам. Учитывая нашу дальнейшую цель изучения геометрии минимальных остовных деревьев, нам нужны расстояния, не превосходящие максимальную длину ковалентных связей, которая, как видно из диаграммы 14, не превосходит 2 ангстрем. Следующий трюк позволяет эффективно выделить все пары с расстояниями не больше 2 ангстрем.

(1) Вычислим геометрические центры аминокислот.

(2) Посчитаем “радиусы” аминокислот, определив их как наибольшие расстояния от центров до атомов аминокислоты. Оказалось, что максимальный радиус по всем аминокислотам всех рассматриваемых белков равен примерно 5.575 ангстрем.

(3) Вычислим расстояния между всеми парами центров.

(4) Выберем только те пары центров, которые расположены друг от друга не далее чем на $2 * 5.575 + 2$ ангстрема. Тем самым, мы добьемся того, что каждая пара атомов, расположенных друг относительно друга не далее чем на 2 ангстрема, лежит в некоторой паре аминокислот, чьи центры мы выбрали.

(5) Образуем список всех пар атомов по всем парам выбранных аминокислот.

(6) Выкинем оттуда все пептидные связи.

(7) Выберем из оставшихся пар все пары с расстояниями, не превосходящими 2 ангстрем.

Конечно, этот алгоритм можно улучшить, но для наших целей его оказалось вполне достаточно.

Приведем результаты, демонстрирующие распределения количеств связей по их длинам для выбранных выше пар, рис. 17 и 18.

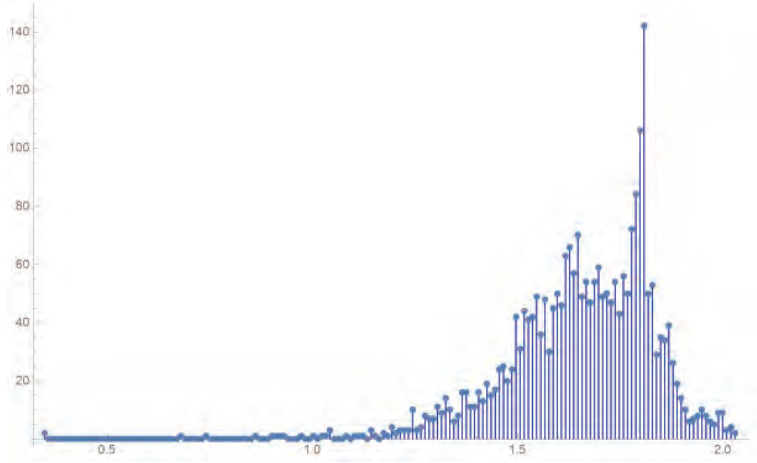

Рис. 17: Распределение количеств пар атомов из разных аминокислот, не связанных ковалентно, по расстояниям между этими атомами.

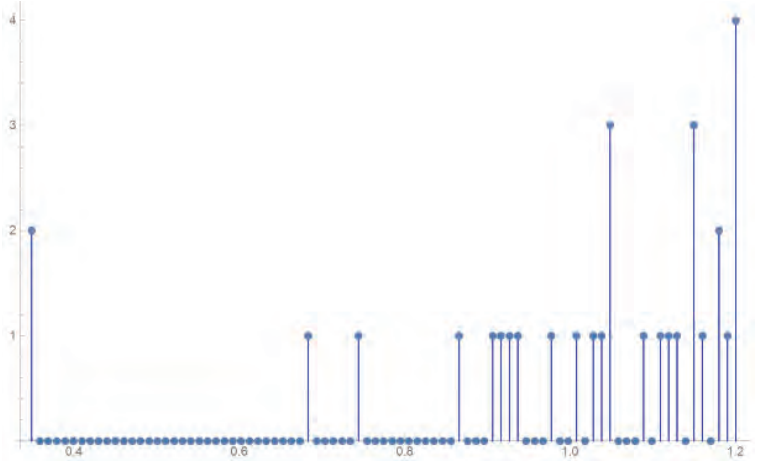

Рис. 18: Распределение количеств пар атомов из разных аминокислот, не связанных ковалентно, по расстояниям между этими атомами.

Из рис. 18 видно, что существуют ровно две совсем короткие связи. Их длины равны 0.356596 и 0.348223 ангстрем, и встречаются они в файлах 1APC.pdb и 2G0K.pdb соответственно. Все остальные расстояния не меньше 0.7 ангстрем.

Итак, окончательный список файлов, который мы решили выкинуть из нашей базы, такой:

```
1WM8, 1WT8, 2JRL, 2MEX, 2OYV, 2OYW, 1ZUV, 2LV1;    1Q2I;    1APC, 2G0K.
```

Так как после итоговой “чистки” диаграмма 14 практически не изменилась, мы не будем приводить ее здесь. Вместо этого мы приведем более компактную диаграмму изменений длин типов ковалентных связей, отождествив друг с другом разные водороды (H, HA, HB2, HB3, и т.д.), разные углероды, разные азоты и т.д. Результат показан на рис. 19.

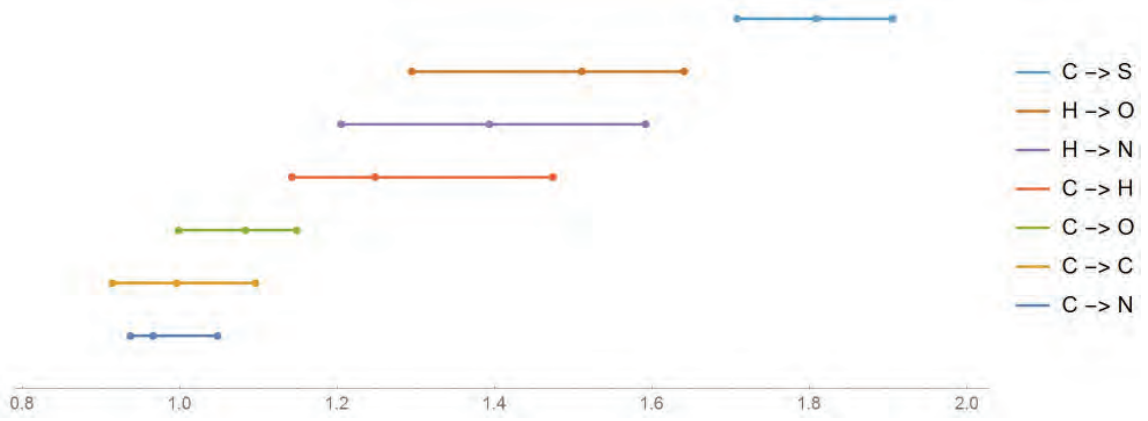

Рис. 19: Интервалы значений разных типов ковалентных связей.

В следующей статье мы расскажем о том, как очищенная нами база была применена для изучения связи экстремальных графов, построенных на координатах атомов белка, и графа ковалентных связей.

ВЫВОД. Для получения более надежной базы пространственных координат белков оказалось недостаточным изучение только средних значений и отклонений от них длин ковалентных связей и углов между смежными ковалентными связями. Предложена и реализована более тонкая методика “чистки”, что привело к более надежной базе, подготовленной для дальнейшего статистического исследования геометрии белков.

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

[2] *Ivanov A.O., Mishchenko A.S., Tuzhilin A.A.* Critical analysis of amino acids and polypeptides geometry // In Continuous and Distributed Systems: Theory and Applications, Springer, 2015, **2**, 29–74; http://springer.com

[3] *Ivanov A.O., Mishchenko A.S., Tuzhilin A.A.* Geometry of space curves and applications to polymers conformations investigation // Educational Internet-Journal “Computer Graphics & Geometry”, 2007, **9**(2), 43–63.

[4] *Иванов А.О., Мищенко А.С., Тужилин А.А.* Геометрия ломаных и полипептидов // Наноструктуры. Математическая физика и моделирование, 2014, **10**(1), 39–48; http://nano-journal.ru

[5] *Hootd R.W.W., Vriend G., Sander Ch., Abola E.E.* Errors in protein structures// Nature, 1996, **381**, 272.

[6] *Nabuurs S.B., Spronk Ch.A.E.M., Vuister G.W., Vriend G.* Traditional Biomolecular Structure Determination by NMR Spectroscopy Allows for Major Errors // PLoS Comput Biol., 2006, **2**(2), 71–79.

[7] *Joosten R.P., Joosten K., Cohen S.X., Vriend G., Perrakis A.* Automatic rebuilding and optimization of crystallographic structures in the Protein Data Bank // Bioinformatics, 2011 **27**(24), 3392–3398.

PROTEIN DATA BANK: NEXT CLEANING

A.A. Tuzhilin

Faculty of Mechanics and Mathematics,
Lomonosov Moscow State University

tuz@mech.math.msu.su

Received 14.05.2016

In the present paper we continue to investigate the distance geometry of proteins atoms. For the atoms coordinates database, we take Protein Data Bank (PDB) files. We present a few new ideas to test PDB-files for errors. As a result, we construct a new database which will be used in our next work for understanding the relation between the covalent bonds graph and some other extreme graphs like minimum spanning tree. This work continues our previous investigation of the geometry of polygonal lines constructed on coordinates of consecutive alpha-carbons.