Отчёт по качеству расшифровки структуры RadA из *Pyrococcus furiosus* (PDB_ID: 4b3b)

Козлова Мария (ФББ МГУ, группа 402)

Содержание

Аннотация

В приводимом ниже отчёте ведётся работа со структурой RadA из *Pyrococcus furiosus* 4b3b. Приводится общая информация о структуре (основные экспериментальные данные и параметры, использованные при улучшении модели), рассматриваются показатели качества модели в целом, выделяются маргинальные остатки на основе различных параметров, разбирается ряд отмеченных на предыдущем шаге остатков с нетипичными параметрами. Работа завершается выводами касательно качества расшифровки структуры. В работе используется ряд интернет сервисов (RCSB PDB, EDS, MolProbity).

Введение

RadA – архейный гомолог RecA, важнейший участника процесса репарации в клетке.

RadA связывается с одноцепочечной ДНК, образуя нуклеопротеиновый филамент, а также связывает двухцепочечную ДНК, сближая их в пространстве и производя поиск гомологичной последовательности, вследствие чего становится возможным образование дуплекса «одноцепочечная ДНК – одна из цепей гомологичной ДНК». Другими словами, белок опосредует однонаправленную миграцию ветвей.

RadA является ДНК-зависимой АТФазой.

В файле с PDB_ID: 4b3b приводится структура димера енолазы из *Pyrococcus furiosus*.

Результаты и обсуждение

1. Общая информация о модели

Целью авторов работы было изучить взаимодействие между человеческим гомологом Rad51 и опухолевым супрессором BRCA2. Для этого был взят термостабильный RadA и промутирован, далее производилось изучение связывания мономера RadA и тетрапептида – консервативного мотива из BRCA2.

Белок была гетерологически экспрессирован в *E.coli* BL21(DE3) и закристаллизован. Помимо тетрапептида структура также содержит низкомолекулярный лиганда: ион фосфата. В структуре определены атомы водорода.

Структура была получена E. C. Schulz и R. Ficner из K в 2013 году.

С использованием информации с сайта EDS (Electron Density Server, [2]) были получены следующие данные об экспериментальных данных:

- Получено 64404 рефлексов, всего измерено 64404 значения структурных факторов, из них для оптимизации были использованы 61153 (95.0%), тестовый набор рефлексов составил, соответственно, 3251 штук (5.0%).
- Заявляемое авторами разрешение структуры составляет **1,19 Å**.
- Полнота данных составляет **93,6%**.
- Диапазон разрешений структурных факторов составляет 26.28 - 1.19 **Å**.

Информация касательно пространственных характеристик ячейки кристалла из PDB-файла (поле CRYST1):

```
CRYST1   40.227   60.589   87.524  90.00  90.00  90.00 P 21 21 21    4
```

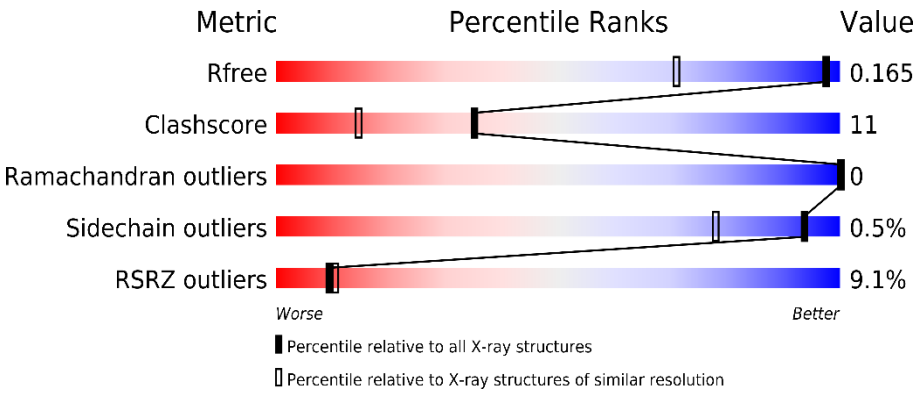
Первые три числа в строке соответствую длинам рёбер кристаллографической ячейки (a, b и c), за ними идут значения углов (α, β и γ), название пространственной группы симметрии P 2_1 2_1 2_1 и число молекул в одной ячейке, равное четырём.

Некристаллографическая симметрия в пределах ячейки отсутствует. Для пространственной группы P 2_1 2_1 2_1 возможны следующие операторы симметрии:

```
X,Y,Z; -X+1/2,-Y,Z+1/2; -X,Y+1/2,-Z+1/2; X+1/2,-Y+1/2,-Z.
```

2. Индикаторы качества модели в целом

Основные характеристики того, насколько хорошо полученная модель описывает реальные данные приведены на сайте PDB [3] в виде поля (см. Рис. 3), на котором происходит сравнение различных параметров (графа Metric) модели с таковыми у других моделей такого же разрешения (белый «ползунок») и с таковыми у всех структур, полученным методами рентгеновской кристаллографии (черный «ползунок»). Сдвиг в синее поле считается хорошим относительно большинства, а в красное, соответственно, плохим. Справа от поля (графа Value) представлено абсолютное значение параметра для рассматриваемой модели. Под визуальным представлением можно видеть таблицу, в которой указано, на основе какого числа моделей было сформировано верхнее представление (каков был объём выборки, с которой сравнивался конкретный параметр).



Metric	Whole archive (#Entries)	Similar resolution (#Entries, resolution range(Å))
R_{free}	66092	1038 (1.26-1.14)
Clashscore	79885	1158 (1.26-1.14)
Ramachandran outliers	78287	1106 (1.26-1.14)
Sidechain outliers	78261	1104 (1.26-1.14)
RSRZ outliers	66119	1038 (1.26-1.14)

Рисунок 1. Параметры модели в сравнении с таковыми для других моделей в базе данных PDB (всех и одинакового разрешения).

Теперь попробуем разобраться с основными параметрами модели по одному и понять с какой стороны они её характеризуют.

R-фактор для модели составляет 0.141. Этот параметр показывает то, насколько рассчитанные уже по построенной модели модули структурных факторов (амплитуды гармоник Фурье) F^{calc} отличаются от таковых полученных в эксперименте F^{obs} при одних и тех же числах (h,k,l). Таким образом, R-фактор является мерой схожести модели с экспериментальными данными и рассчитывается по формуле (1).

$$R = \frac{\sum_{hkl} \left| F_{hkl}^{calc} - F_{hkl}^{obs} \right|}{\sum_{hkl} F_{hkl}^{obs}} * 100\%$$

Хорошими значениями для R-фактора являются значения менее 25%. Наша модель вписывается с неплохим запасом. Однако, при оптимизации модели мы минимизируем именно этот параметр и может возникнуть ситуация, когда происходит переоптимизация (overfitting), и, уменьшая R-фактор, мы всё дальше уходим от реальной структуры белка. Для недопущения такой ситуации служит ещё один параметр: R_{free}фактор.

R_{free}-фактор вычисляется точно так же как и выше, но при этом в качестве рефлексов используется специальная группа рефлексов (в нашем случае 3780 штук (5.0% от всех)), не участвовавших в оптимизации модели. Параметр

рассчитывается по конечной оптимизированной модели. В нашем случае R_{free}-фактор составляет 0.1671, что считается неплохим значением, что разница между R и R_{free} составляет менее 5%, переоптимизации, скорее всего, не произошло.

Ещё одним параметров, приведённым выше является **Clashscore**, который показывает среднее число неблагоприятных перекрываний атомов (более чем на 0.4Å) на 1000 атомов в структуре. В нашей структуре всего 4146 атомов, найдено 40 участков с перекрыванием. Среднее значение составляет около **11**, оно и приведено на рисунке 1.

Ряд далее приведённых характеристик может быть получен при использовании сервиса MolProbity [4], хотя ряд из них можно почерпнуть из отчётов, доступных в PDB на странице структуры.

Хорошим показателем качества структуры является количество остатков, которые выбиваются за допустимые зоны (**Ramachandran outliers**) на карте Рамачандрана (карте значений торсионных углов остова φ и ψ), которая построена по полученной модели. На рисунке 2 представлена карта для всего остова (A), и для отдельных аминокислотных остатков, имеющих отличную от общей карту (B-E). В рассматриваемой модели **98.72%** остатков лежали в благоприятных (**favored**) и **100%** в разрешённых (**allowed**) областях карты.

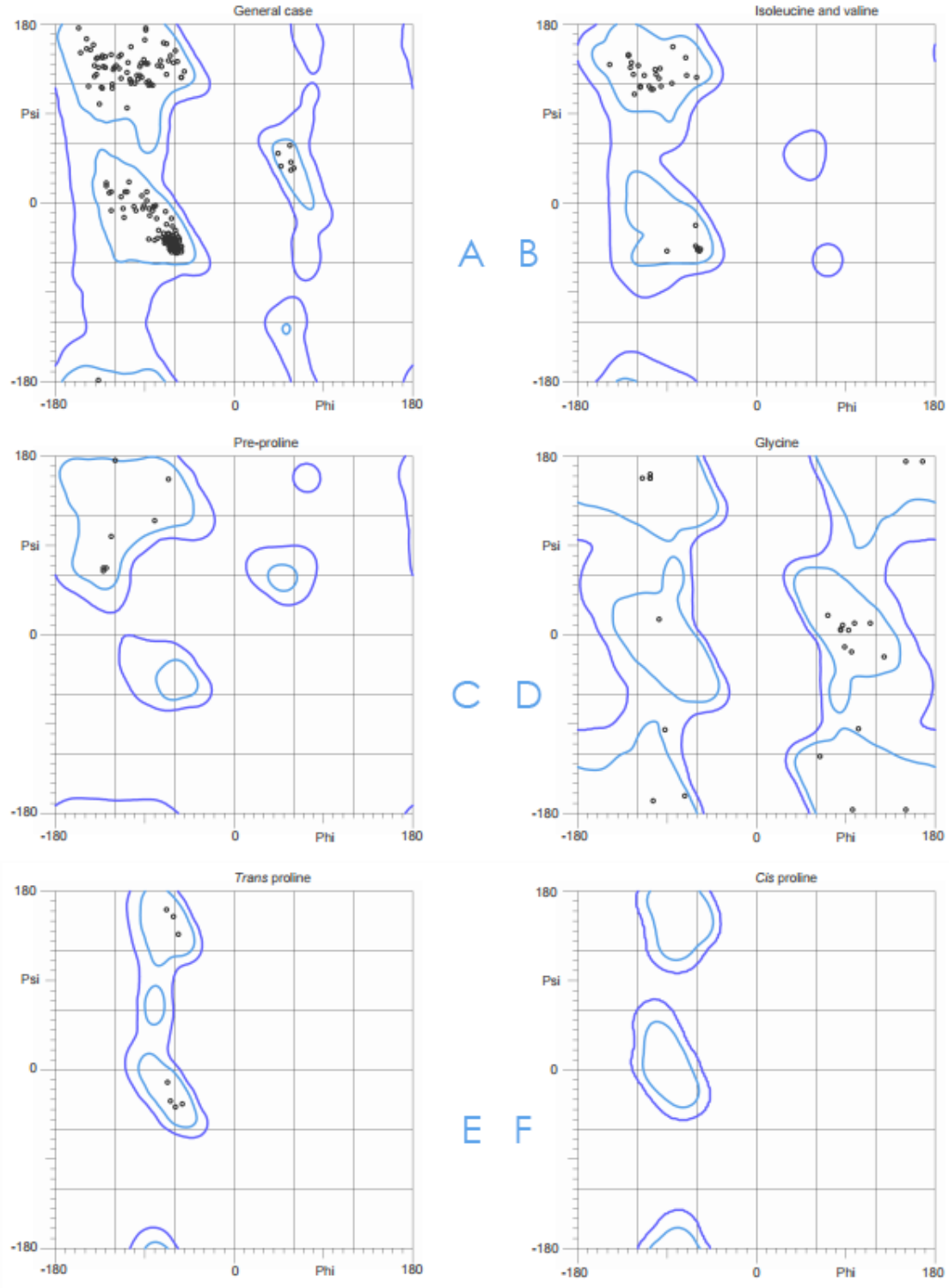


Рисунок 2. Карты Рамачандрана для модели 4B3B: A - общий случай, B - препролиновые остатки, C - изолейцины и валины, D – глицины, E – транс-пролины и F – цис-пролины (0 остатков).

На данных картах голубой контур окружает 98% всех остатков (по множеству структур) – это благоприятная область ограничена голубым, а разрешённая область - фиолетовым.

Стоит отметить, что в структуре присутствует один остаток, не являющийся пролином – Ser239, чья пептидная группа находится в транс-конформации, это больше желаемого порога (0.44% против 0.05%)

Аналогично рассчитывается процент выбивающихся остатков по показателям торсионных углов боковых цепей (**Sidechain outliers**). Это значение в данном случае несколько больше, чем для остова и составляет **0.9%**.

Пространственный R-factor (Real Space R-factor = RSR) является ещё одной мерой соответствия модели реальности, этот параметр рассчитывается по формуле (2) и показывает, насколько реальная электронная плотность согласуется с таковой, построенной в модели.

$$RSR = \frac{\sum_{A \in L} |\rho_{\text{эксп}} - \rho_{\text{модель}}|}{\sum_{A \in L} \rho_{\text{эксп}}} [\cdot 100\%] \quad (2)$$

На основе RSR, рассчитанных для каждого из остатков в цепи белка, можно понять какие остатки выбиваются из общего фона по параметру RSRZ – по сути это Z-score, рассчитанный по RSR с использованием формулы (3).

$$RSRZ = (RSR - \langle RSR \rangle) / Sigma \quad (3)$$

На рисунке 3 представлено распределение значений RSR по цепи белка. Распределение значений RSRZ представлено на рисунке 4. Оба рисунка (3 и 4) получены при помощи соответствующих сервисов EDS [2].

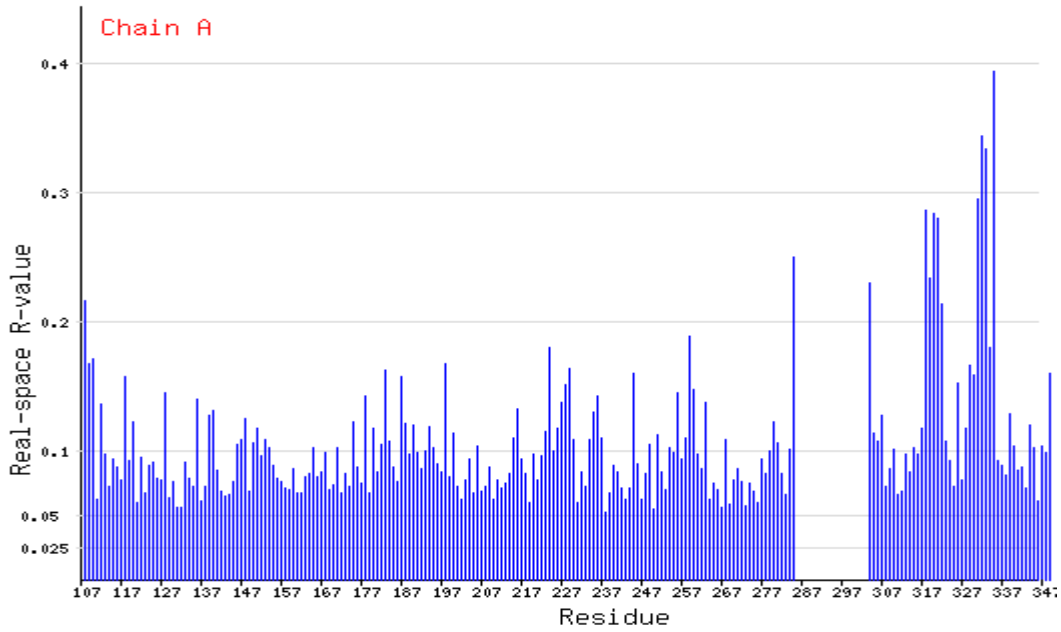


Рисунок 3. Распределение значения RSR от номера остатка в цепи A структуры 4b3b.

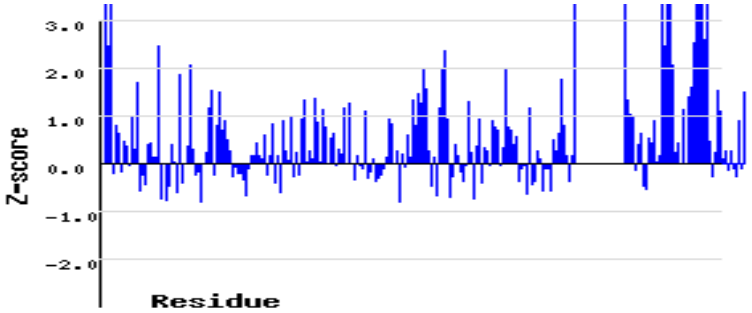


Рисунок 4. Распределение значения RSRZ от номера остатка в цепи A (сверху) и B (снизу) структуры 4b3b.

Маргинальными с точки зрения RSRZ (**RSRZ outliers**) считаются остатки, имеющие значение этого параметра больше двух, то есть отклоняющиеся от среднего качества описания электронной плотности более чем на два стандартных отклонения в «плохую» сторону. Всего таких остатков в обеих цепях белка **21 (9%)** при общем числе остатков равном 349. Эти остатки также будут подробнее рассмотрены в разделе о маргиналах.

В целом электронная плотность модели удовлетворительно соответствует экспериментальной (много остатков с RSRZ<2, но >0 и даже >1). Причина отсутствия оценки для части остатков неясна.

Маргинальные остатки в структуре. Примеры

В таблице 1 приведена информация о нескольких маргинальных остатках, выделенных нами в предыдущем разделе при рассмотрении параметров качества модели в целом. Для каждого остатка приводится трёхбуквенное сокращение, а также причины, заставившие нас обратить внимание на этот остаток. В подборе примеров маргинальных остатков нам также поможет рисунок из PDBReport, который наглядно демонстрирует отклонения (геометрические) и несоответствие электронной плотности (см. Рис. 5).

Любопытно, что маргинальные по разным параметрам остатки часто располагаются рядом один с другим.

Далее приведено более подробное рассмотрение 5 случаев (рис. 6-9).

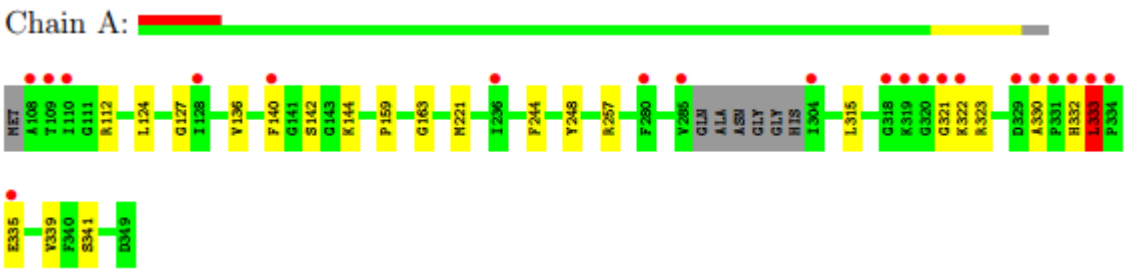


Рисунок 5. Качество отдельных остатков в структуре, маргиналы по одному геометрическому параметру жёлтые, по двум оранжевые. Если остаток плохо описывается исходной электронной плотностью (RSRZ>2.0) рядом с ним ставится красная точка.

Таблица 1. Примеры маргинальных остатков в PCA модели RadA PDB_ID: 4b3b.

№	Остаток	Суть проблемы
1	ASP-238	Маргинал по углам связей worst is C-N-CA: 4.6 σ
1	ARG-257	Маргинал по углам связей (NE-CZ-NH1, NE-CZ-NH2 – 125 и 116 вместо идеальных 120)
2	SER-239	Цис-конформация пептидной группы
3	LYS-319	Маргинал по торсионным углам в боковой цепи остатка chi angles: 234.3,64.3,154.5,122.5
4	LEU-333	Маргинал по торсионным углам в боковой цепи остатка, Маргинал по углам связей (CA-CB-CG – 131.62 против идеальных 115.30), RSRZ outlier (RSRZ - 8.5), перекрывание с HIS-332 на на 0,65 Å.
5	GLU-335	Маргинал по торсионным углам в боковой цепи остатка chi angles: 35.2,288.4,314.9
6	HIS-199	Перекрывание с водой (2092 HOH O) на на 0,56 Å, вероятно, нужна инверсия боковой цепи
7	PRO-331	RSRZ outlier (RSRZ - 6.2)
8	ILE-304	RSRZ outlier (RSRZ - 6.2)
9	GLY-320	RSRZ outlier (RSRZ - 5.7) Ещё 21 таких остатков
10	GLY-127	Перекрывание на 0.72Å с ARG 113

Gly-127 и Arg-113

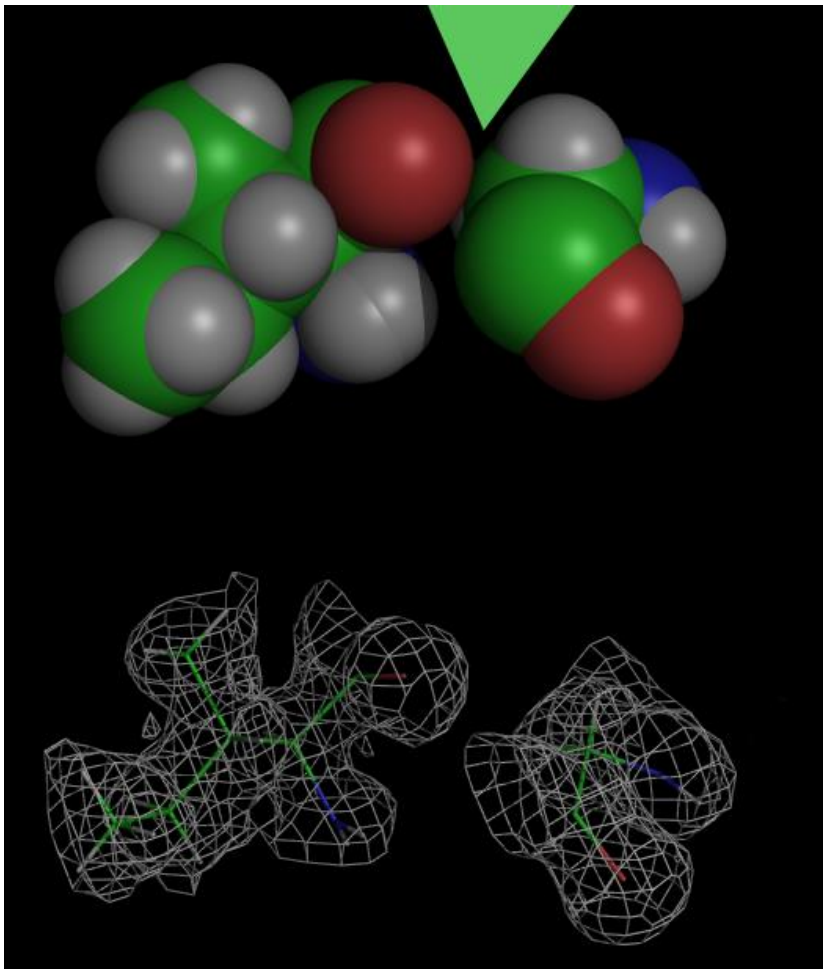


Рис. 6.Перекрывающиеся остатки (на 0.72Å)- Gly 127 справа, Arg-113 справа, область перекрывания показана зелёной стрелкой. Оба остатка хорошо вписаны в ЭП. Такое перекрывание не может объясняться функциональным взаимодействием, это, вероятно, ошибка определения ЭП.

Leu-333

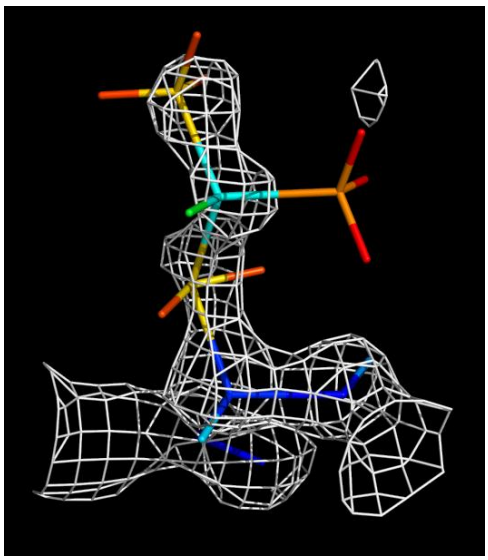


Рис 7. Leu-333, остаток с рекордным числом причин, чтобы быть признанным маргинальным. При рассмотрении ЭП остатка оказалось, что по неизвестной причине для одной из метильных групп ЭП не визуализируется ни при каком пороге. Также оказалось, что у радикала остатка довольно высокий для этого участка цепи B-factor. Причина маргинальности также не функциональная.

Pro-331

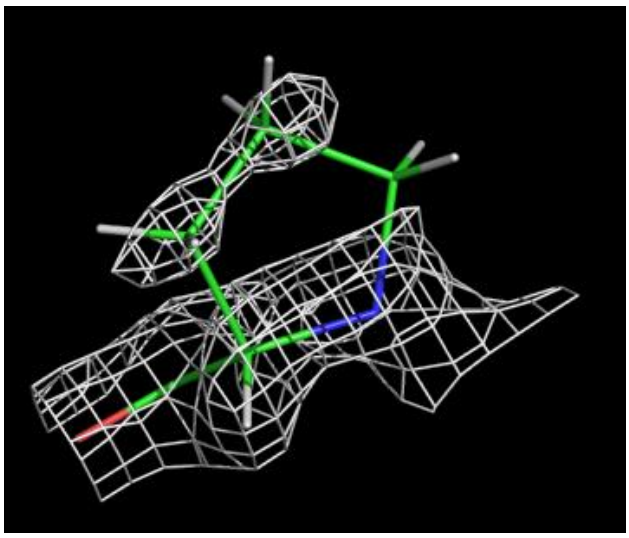


Рис. 8. Pro-331, RSRZ outlier (RSRZ - 6.2), близко расположен к аномальному лейцину 333. При визуализации ЭП видим ту же проблему – «отсутствие» ЭП для части атомов в кольце, это объясняет плохое RSRZ.

Ser-239

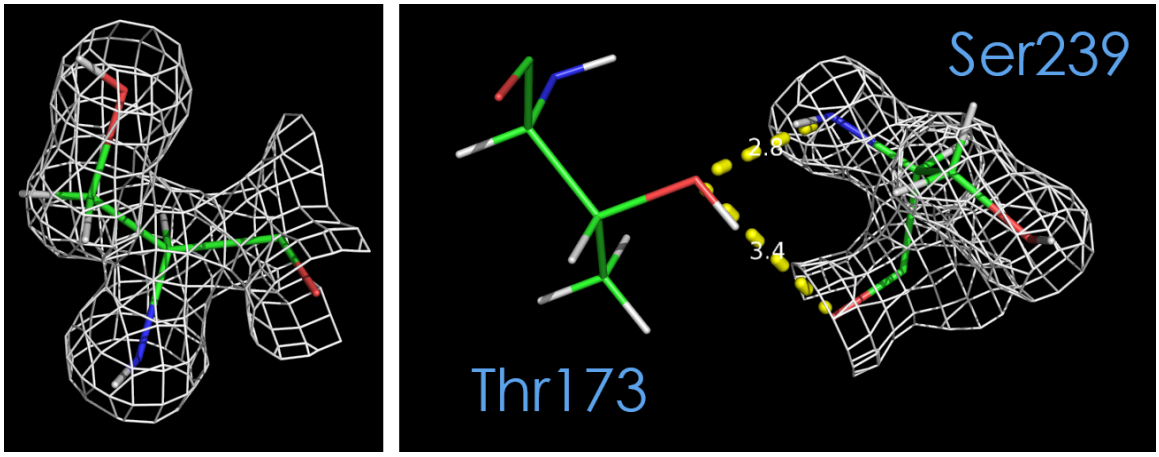


Рис. 9. Ser239 обращает на себя внимание своей пептидной группой в цис-конформации. Остаток хорошо вписан в ЭП. Что же могло заставить его повернуться таким образом? Можно обнаружить водородную связь с радикалом другого остатка, которая могла бы стабилизировать такую конформацию. Это единственный рассмотренный маргинал, свойства которого можно признать

особенностью, а не ошибкой расшифровки/проблемой с экспериментальными данными.

Сравнение модели из PDB с моделью из PDB_redo

PDB_redo представляет из себя сервис оптимизации, который позволяет улучшить модель с привлечением ряда параметров, которые используются в процессе улучшения. Модель была загружена в программу, скачан файл, содержащий новую модель после оптимизации. Программа сама считает основные параметры до и после оптимизации, а также запускает валидацию WhatCheck по основным параметрам. Все это разом вместе с результатами доступно по ссылке: [5].

	From PDB header	Calculated from data	After conservative optimisation	After full optimisation
R	0.1400	0.1417	0.1331	0.1354
R-free	0.1671	0.1678	0.1630	0.1649
σR-free		0.0021	0.0020	0.0020
R-free Z-score		7.24	4.45	4.95

Выше приведены данные из выдачи программы касательно R-фактора и R-free в исходной модели и после двух вариантов оптимизации. Видно, что в результате обработки модели происходит уменьшение обоих параметров, но разница между ними не падает, так что не совсем ясно, происходит ли реальная оптимизация.

На рисунке 10 показано пространственное совмещение двух моделей в двух видах визуализации. Видно, что модели в целом сходны и повторяют одна другую, однако в некоторых местах отличается вторичная структура (рассм. подробнее на рис. 11).

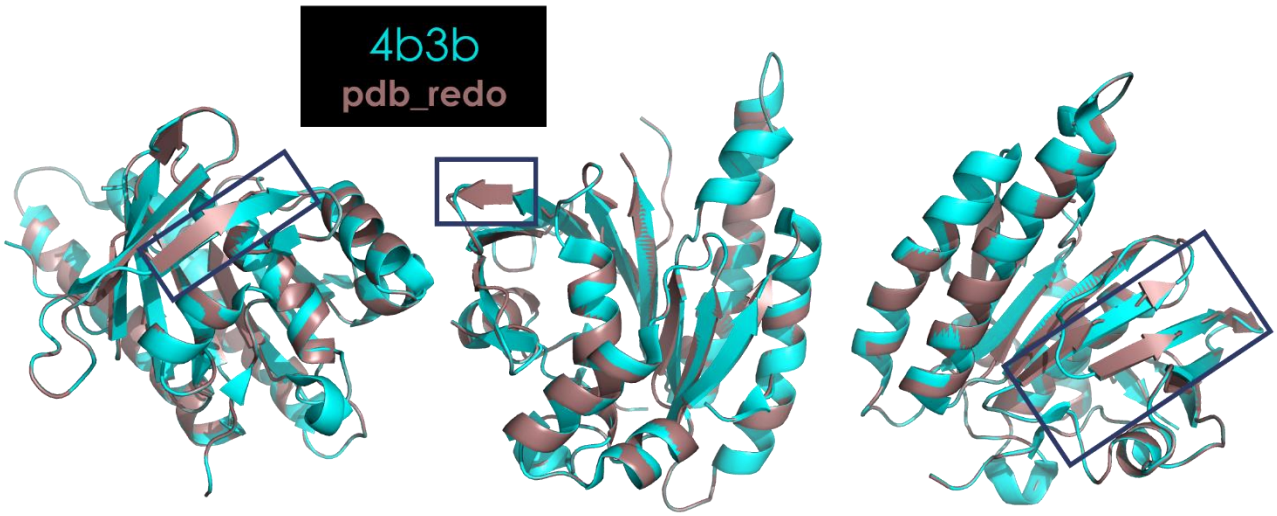


Рисунок 10. Пространственное совмещение структур до и после полной оптимизации программой PDB_redo.

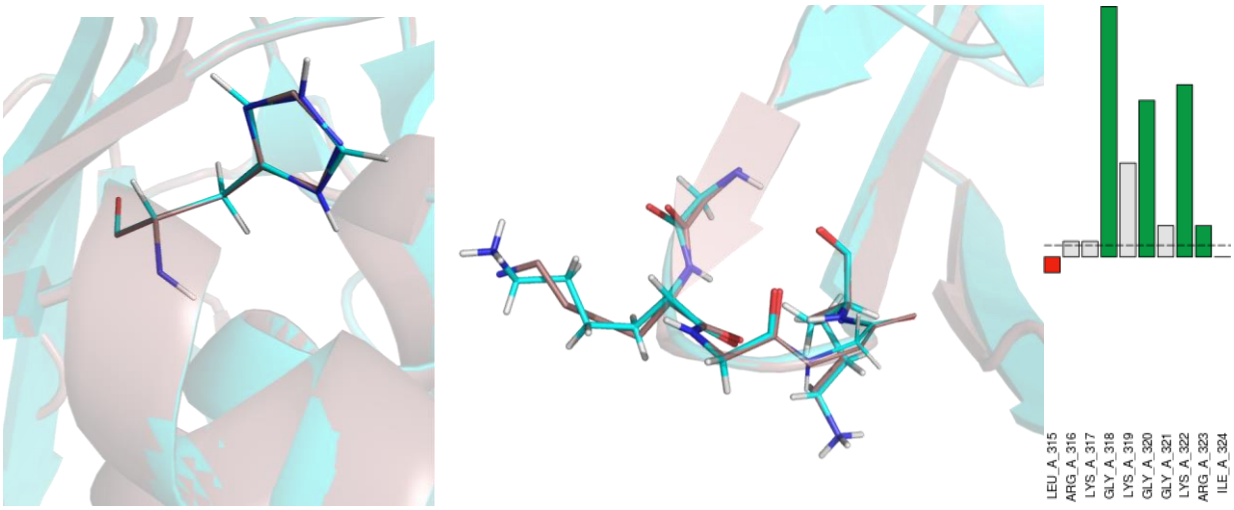


Рисунок 11. Слева: Радикал остатка His-199 (один из маргиналов по перекрыванию) был повёрнут программой оптимизации.
В центре и справа: Один из элементов белка, где различается определение вторичной структуры в исходной модели и оптимизированной. Видно, что оптимизация немного сдвинула несколько последовательно идущих остатков, чего было достаточно, чтобы по другому проинтерпретировать вторичную структуру. Утверждается, что остатки оказались вписаны в ЭП значительно лучше (см график справа).

Заключение о качестве расшифровки структуры

Исходя из вышеприведённой информации, можно сказать, что, несмотря на отличное разрешение, местами структура имеет проблемы с расшифровкой. Некоторые параметры качества структуры имеют более хорошие значения по равнению с другими структурами такого разрешения в базе данных PDB, однако это неверно для RSRZ и clascore.

Повторная оптимизация модели (PDB_redo) внесла ряд изменений в структуру.

Ссылки

[1] Scott DE, Ehebauer MT, Pukala T, Marsh M, Blundell TL, et al. (2013) Using a fragment-based approach to target protein-protein interactions. Chembiochem 14: 332–342

[2] "EDS: PDB entry 4b3b." http://eds.bmc.uu.se/cgibin/eds/uusfs?pdbCode=4b3b.

[3] "RSCB PDB: 4b3b." http://www.rcsb.org/pdb/explore/explore.do?structureId=4b3b.

[4] Сервис MolProbity. http://molprobity.biochem.duke.edu/.

[5] Выдача PDB_redo. http://www.cmbi.ru.nl/pdb_redo/b3/4b3b/index.html.