МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

имени М.В.ЛОМОНОСОВА

ФАКУЛЬТЕТ БИОИНЖЕНЕРИИ И БИОИНФОРМАТИКИ

Отчет по качеству расшифровки структуры белка 4ep4, полученной методом рентгеноструктурного анализа

Работу выполнила:

Фролова А.С.

Москва, 2017 г.

Оглавление

Аннотация 3
Введение 3
Общая информация о модели. 3
Результаты и обсуждение 6
Показатели качества модели 6
Геометрические параметры 7
Карты Рамачандрана 7
Пространственный R-фактор. 9
Оценка пространственного R-фактора – Z-score 10
Температурный фактор 11
Комфортность окружения 12
Лиганды 13
Маргинальные остатки 13
Анализ маргинальных остатков 14
Сравнение модели 4ep4 с моделью из PDB_redo 17
Заключение 20
Литература и сервера 21

Аннотация

В данном отчете приведен анализ модели белка 4ep4 из организма Thermus thermophiles HB8. Были проанализированы различные оценки качества модели, описаны некоторые маргинальные остатки. Кроме того, был проведен анализ данной модели и улучшенной модели, полученной специальным сервисом.

Введение

Структура белка 4EP4 является эндонуклеазой RuvC, или ревертазой. Данная структура была получена Luan Chen и его коллегами из организма Thermus thermophiles HB8 методом рентгеноструктурного анализа[1].

Эта эндонуклеаза распознает структуры Холлидея и разрезает две ее кроссоверные нити поперек точки соединения. По классификации ферментов данный белок имеет номер EC 3.1.22.4.

Структура белка RuvC впервые была получена из Escherichia coli, однако разобраться с механизмом полностью не удалось. Поэтому основной задачей Luan Chen было получить структуру RuvC из другого организма, а именно из Thermus thermophiles HB8 , понять как эндонуклеаза распознает структуры Холлидея, а так же разобраться в механизме симметричного последовательного расщепления последовательностей ДНК в структурах Холлидея.

Общая информация о модели.

В 2012 году Luan Chen, Ke Shi, Zhiqi Yin and Hideki Aihara получили модель белка с PDB ID: 4ep4[4] из организма Thermus thermophiles HB8 с разрешением 1,28 Å.

Фазовая проблема была решена методом молекулярного замещения, где в качестве модели была использована уже известная модель RuvC из E. Coli, с использованием PHASER. Оптимизация модели с помощью REFMAC5 и ее ручное построение в COOT в конечном итоге дали модель RuvC , состоящую из 332 остатков и 414 молекул воды.

Общее число уникальных рефлексов равно 65468, а число рабочих рефлексов составило 63393. Полнота данных из статьи составляет 96,7 (98,4) % для конечной модели, однако в EDS[5] полнота 94%.

Кристалл принадлежит к пространственной группе $P2_12_12_1$. Параметры кристаллографической ячейки a = 36,8 Å, b = 51,2 Å и c = 134,9 Å. Угол α = 90 ° β = 90 °, γ = 90 ° . Асимметрическая ячейка (ASU) содержит две молекулы RuvC.

Финальная модель была оптимизирована с разрешением 1,28 Å и значением R-фактора 19,1% с использованием PHENIX. Диапазон разрешения составляет 50.00 - 1.28 Å, а в EDS оно равно 35.49 – 1.28 Å, так как там, скорее всего, лежит более хорошая модель.

Боковые остатки некоторых аминокислот были смоделированы в альтернативных конформациях. В цепи A это аминокислоты P40, K45, F73, K111, L127, E137, I150, а в цепи B это K45, V67, R76, L127, M128, I150, M160. Всего таких остатков 14. Как видно, некоторые аминокислоты имеют конформации в обеих цепях (F73).

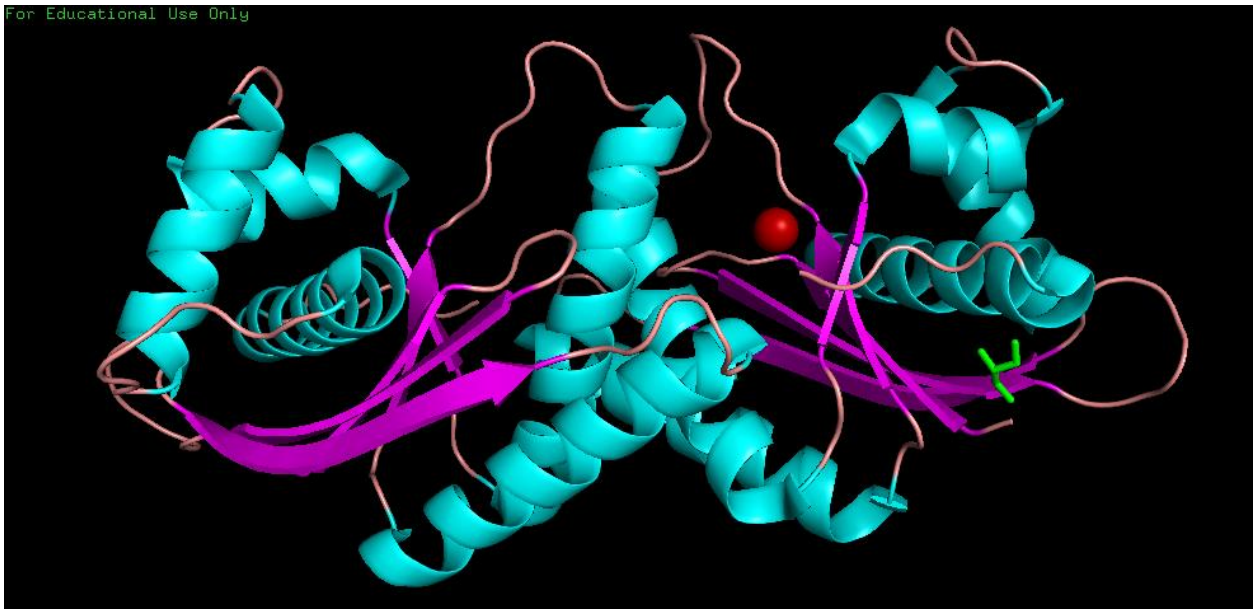


Рисунок 1. Структура белка 4EP4[3]. Окрашивание по вторичным структурам, Mg^{2+} окрашен красным цветом, глицерол – зеленым. Молекулы воды убраны. Визуализация с помощью PyMOL.

Белок представляет собой гомодимер, который состоит из 5 α-спиралей и 1 β-лист в одном мономере (рис.1). В структуре можно найти два лиганда: Mg^{2+} и глицерол, а так же 414 молекул воды. RuvC из этого организма (Thermus

thermophiles) отличается повышенной термостабильностью, скорее всего из-за близкого расположения мономеров в гомодимере.

Результаты и обсуждение

Показатели качества модели

Первым индикатором качества модели является ее разрешение, которое составляет 1,28 Å. Данное разрешение является высоким.

Кроме разрешения, индикаторами качества являются R-фактор и R-free. R-фактор модели характеризует соответствие модели экспериментальным данным. Хорошее значение R-фактора должно быть меньше 0,25, в нашей модели он равен 0,155. R-free фактор показывает насколько наша модель переоптимизована. У полученной модели R-free равен 0,191, что является хорошим показателем, так как он должен быть < 0,2.

Разница между R-free и R-фактором должна быть < 0,1. Такое значение равно 0,036, что говорит о том, что переоптимизации нет.

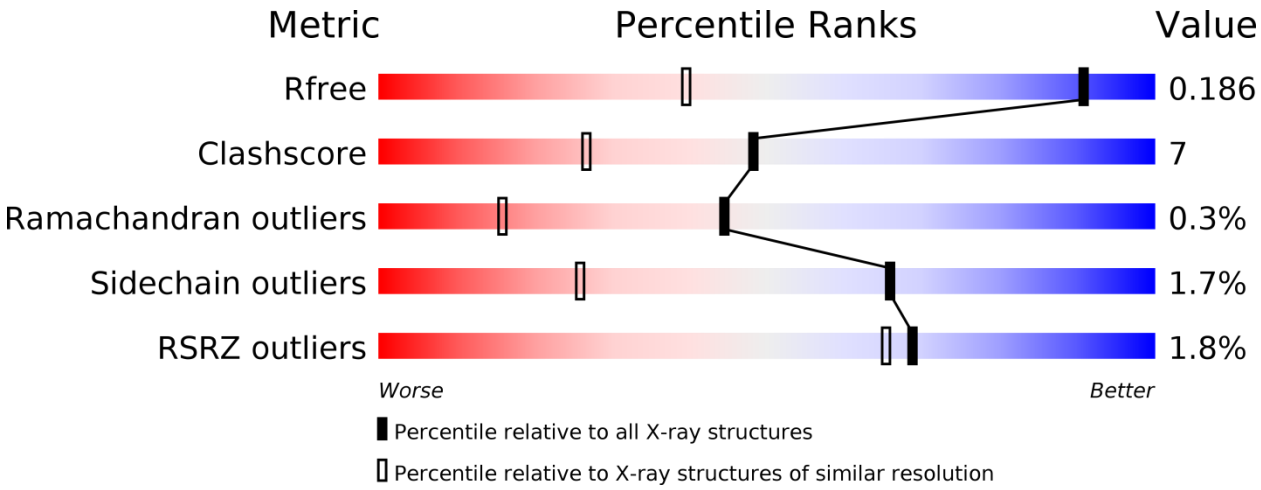


Рисунок 2. Оценка качества модели белка 4ep4 в сравнении с другими структурами с примерно таким же разрешением (1,28 Å).

На рисунке 2 можно видеть различные оценки качества модели 4ep4 (черный) в сравнении с другими моделями такого же разрешения (белый). Чем значение ближе к синей части, тем данные оценки лучше.

Как видно, фактор R-free этой модели намного лучше, чем у других с таким же разрешением. Тоже самое можно сказать и о недопустимых наложениях и маргиналах, однако сами эти значения не очень хорошие (находятся в белой части). Это говорит о том, что есть много перекрытий Ван-дер-Ваальсовых радиусов.

Геометрические параметры

В полном отчете по структуре в PDBe[6] можно найти посмотреть наличие геометрических нарушений в цепях модели белка и их соответствия электронной плотности (рис.3).

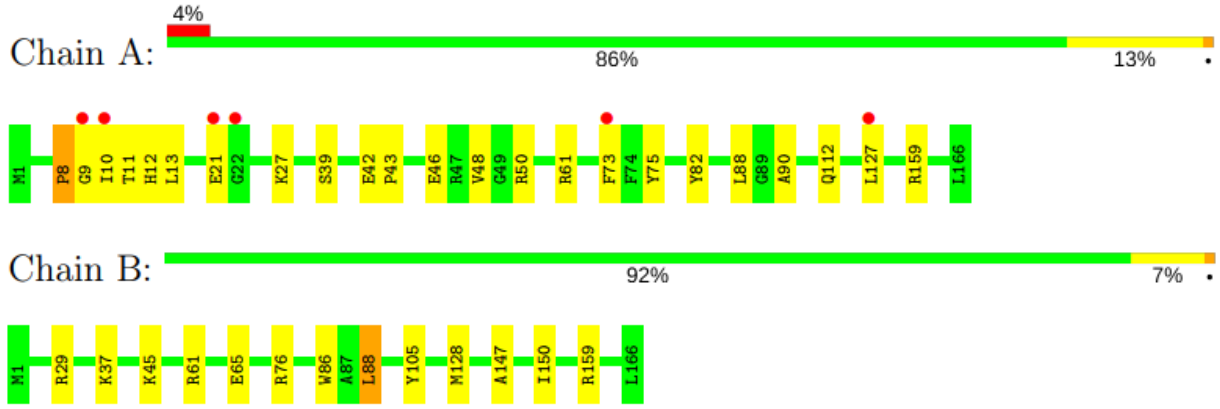


Рисунок 3. Геометрические нарушения в модели белка 4ep4.

Зеленая, желтая, оранжевая и красные полосы показывают долю остатков, которые содержат 0, 1, 2 и >= 3 выбросов по геометрическим критериям качества. У более 85% (зеленые полосы) остатков нет нарушений . Остатков, которые имеют геометрические нарушения, всего 13% и 7% (желтая полоса) в цепях A и B соответственно. В цепи A 4% (красная полоса) остатков, которые плохо вписывается в электронную плотность.

Красные точки над остатками говорят об их плохом соответствии электронной плотности (RSRZ > 2). Это следующие аминокислоты в цепи A: Gly9, Ile10, Glu21, Gly22, Phe73 и Leu127.

В модели нет остатков, которые бы не присутствовали в самом белке (нет остатков с серым фоном).

Карты Рамачандрана

Карты Рамачандрана визуализируют два торсионных угла ψ и ϕ, которые определяют полипептидную цепь белка. Они позволяют оценить конформации остова и найти маргиналов по конформации остова.

Сервис MolProbity[7] позволяет построить карту Рамачандрана с лучшими границами предпочитаемых и допустимых областей, определить маргинальные остатки по отклонению боковых цепей от ротамеров, инверсии боковых цепей Asn, Gln, His, недопустимые наложения атомов.

Для получения данных с этого сервиса необходимо загрузить pdb файл структуры или провести поиск по ID PDB. Далее добавить водороды, которые необходимы для выявления недопустимых наложений атомов и для определения возможных инверсий боковых цепей Asn, Gln, His.

На выходе можно получить таблицу с суммарными характеристиками (Таблица 1). Из нее можно понять некоторую информацию:

- число недопустимых наложений (>0.4 Å) атомов на 1000 (“ClashScore”) равно 7.25;
- 67% структур с примерно таким же разрешением (1.28Å ± 0.25Å) имеет ClashScore больше, чем данная структура;
- 6 (2,34%) остатков с маргинальными по отклонению ротамеров боковыми цепями (“Poor rotamers”); это больше чем 0,3%, что является хорошим значением для этого сервиса;
- 240 (93,75%) ротамеров - боковых цепей в типичных для данного типа остатках конформациях (“Favored rotamers”); при уровне в 98% это не очень хорошо, однако, как мне кажется, это значение завышено;
- 1 (0,3%) полный маргинал по карте Рамачандрана, он лежит вне допустимой области (“Ramachandran outliers”); при высоком разрешении таких остатков должно быть около 0,05%, но так как в данном случае это только один атом из всех, то, наверное, это значение хорошее;
- 326 (99,39%) остатков в предпочитаемой области (“Ramachandran favored”);
- Интегральная оценка структуры по данным этого сервиса равна 1,68 (“MolProbity score”);
- Cβ-атомов с неприемлемым отклонением от ожидаемого положения не обнаружено (“Cβ deviations >0.25 Å”);
- Ковалентные связи и валентные углы, которые существенно отличаются от теории, не обнаружено (“Bad backbone bonds and angles”);
- 2 пролина с cis-конформацией (“Cis prolines”).

Таблица 1. Данные по структуре 4ep4 с сервиса MolProbity.

All-Atom Contacts	Clashscore, all atoms:	7.25		67th percentile* (N=365, 1.28Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	6	2.34%	Goal: <0.3%
	Favored rotamers	240	93.75%	Goal: >98%
	Ramachandran outliers	1	0.30%	Goal: <0.05%
	Ramachandran favored	326	99.39%	Goal: >98%
	MolProbity score^	1.68		60th percentile* (N=2063, 1.28Å ± 0.25Å)
	Cβ deviations >0.25Å	0	0.00%	Goal: 0
	Bad bonds:	0 / 2692	0.00%	Goal: 0%
	Bad angles:	0 / 3665	0.00%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	2 / 19	10.53%	Expected: ≤1 per chain, or ≤5%

Карта Рамачандрана для всех остатков модели 4ep4 представлена на рисунке 4 (левый). Как видно, углы во всех остатках лежат в допустимых областях. На рисунке 4 (правый) показана карта Рамачандрана для транс-пролина, где один из пролинов находится в запрещенной области (Pro8, цепь A; -70.8, -127.3).

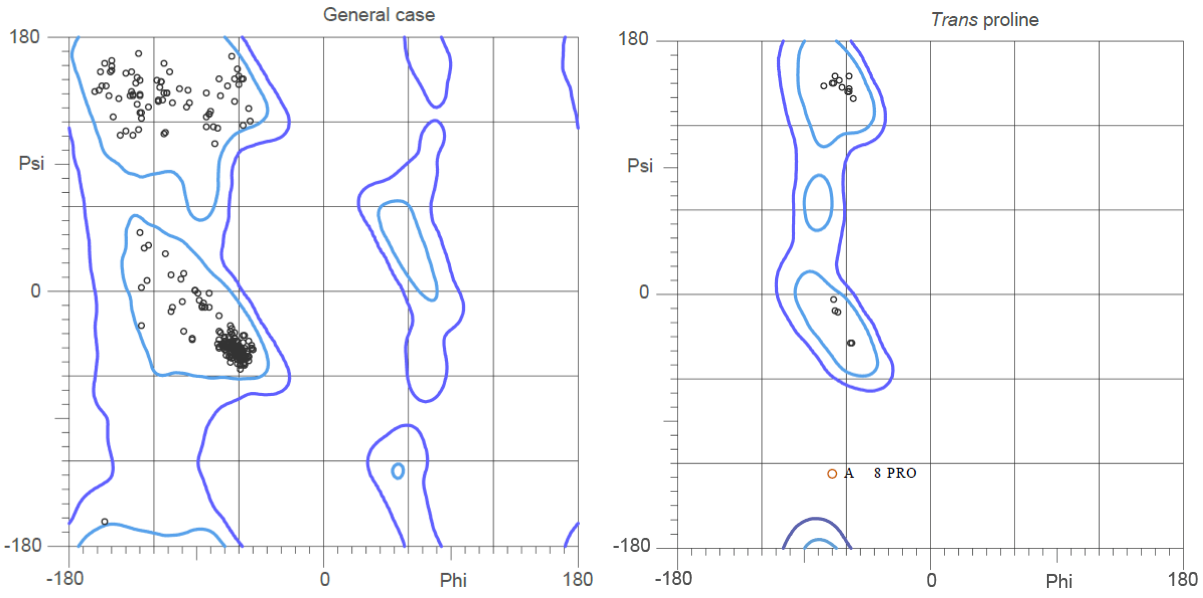


Рисунок 4. Карты Рамачандрана для всей структуры (левый) и для транс-пролина (правый).

Пространственный R-фактор.

Пространственный R-фактор, или RSR, позволяет оценить насколько остаток вписывается в электронную плотность. Он рассчитывается по формуле ниже:

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

где сумма берется по узлам пространственной решетки в окружении группы атомов. Хорошие значения RSR < 10%.

Для нашей модели значение пространственного R-фактора можно посмотреть на сервисе EDS по ссылке “Significant regions”. На рисунке 5 изображены два графика, которые показывают высокие значения RSR для двух цепей модели.

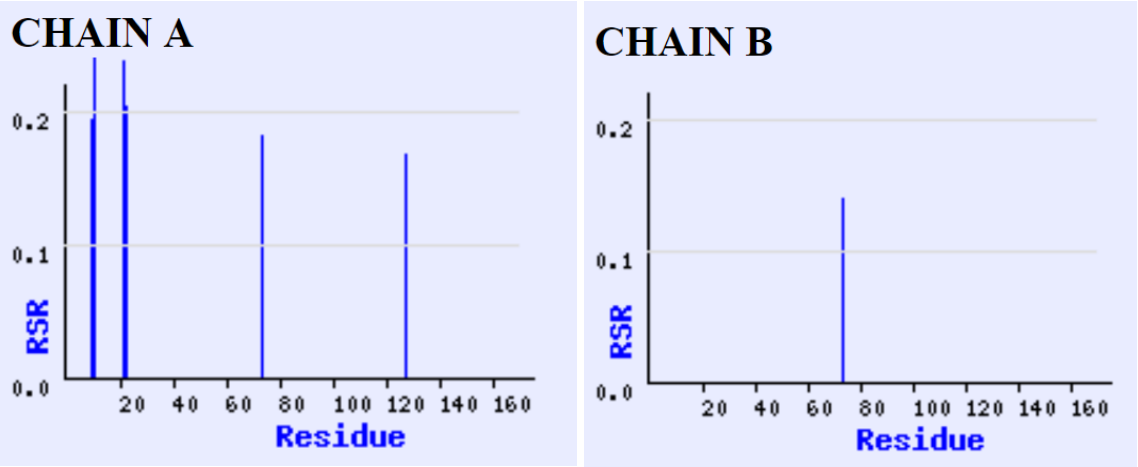


Рисунок 5. Высокие значения RSR фактора для цепей A и B.

Как видно из рисунка 5, в цепи A есть 6 выбросов, которые соответствуют аминокислотам Gly9, Ile10 , Glu21, Gly22, Phe73 и Leu127, а в цепи B только один, Phe73. Можно заметить, что фенилаланин 73 имеет высокое значение RSR в обоих цепях.

Оценка пространственного R-фактора – Z-score

Z-score, или RSRZ, позволяет оценить насколько хорошо остаток вписан в электронную плотность по сравнению с другими структурами такого же разрешения. При высоком положительном значении Z-score > 2 остаток плохо вписан в электронную плотность и он скорее всего является маргиналом. Z-score рассчитывается по формуле:

$$Z = (RSR - \langle RSR_{resolution} \rangle) / Sigma_{resolution}$$

Эта оценка была так же получена сервисом EDS. По рисунку 6 можно определить какие аминокислоты имеют высокие значения этой оценки (граница по Z-score > 2).

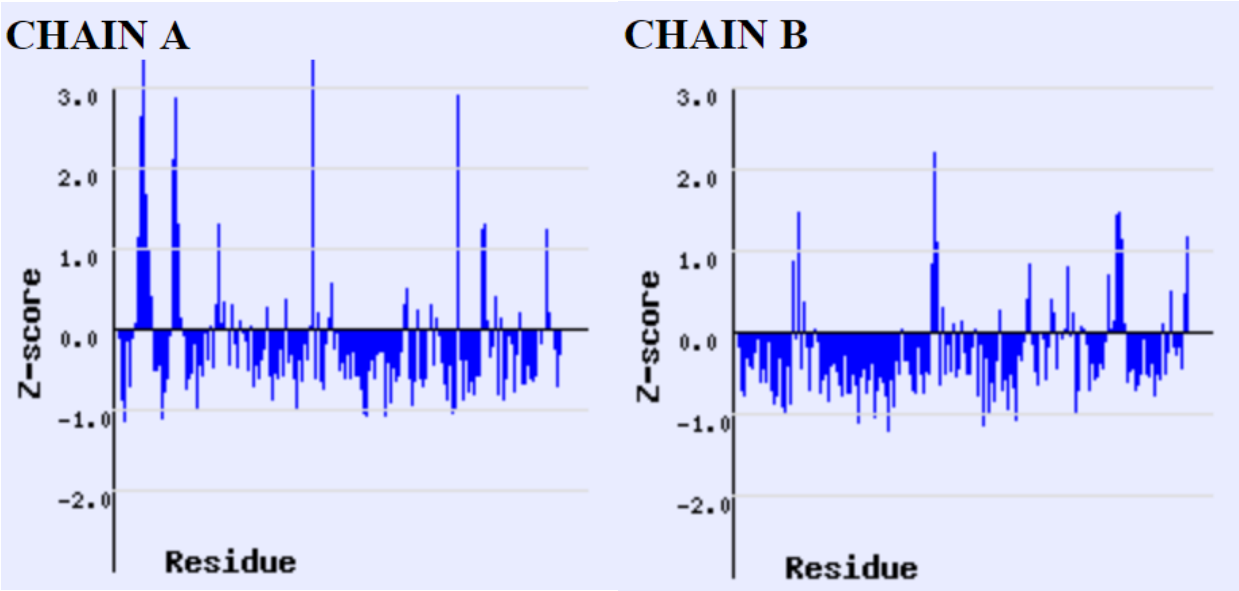


Рисунок 6. Оценка RSR для модели белка 4ep4.

Пики с высокими значениями Z-score соответствуют тем аминокислотам, которые имели высокие значения RSR фактора (Gly9, Ile10, Glu21, Gly22, Phe73, Leu127). Эти аминокислоты являются маргиналами.

Интересно, что в общем отчете с сервиса PDBe Phe73 не имеет высокого значения RSRZ.

Температурный фактор

Температурный фактор, или B-фактор, позволяет узнать насколько атом подвижен. Если атом имеет значение температурного фактора до 30, то это говорит о том, что атом находится в одном месте, а если он больше 60, то атом подвижен и делать важные биологические выводы по положению атома не стоит.

Значения B-фактора для всех остатков можно найти в pdp файле, а общий график, описывающий распределение температурного фактора для модели, через сервис EDS.

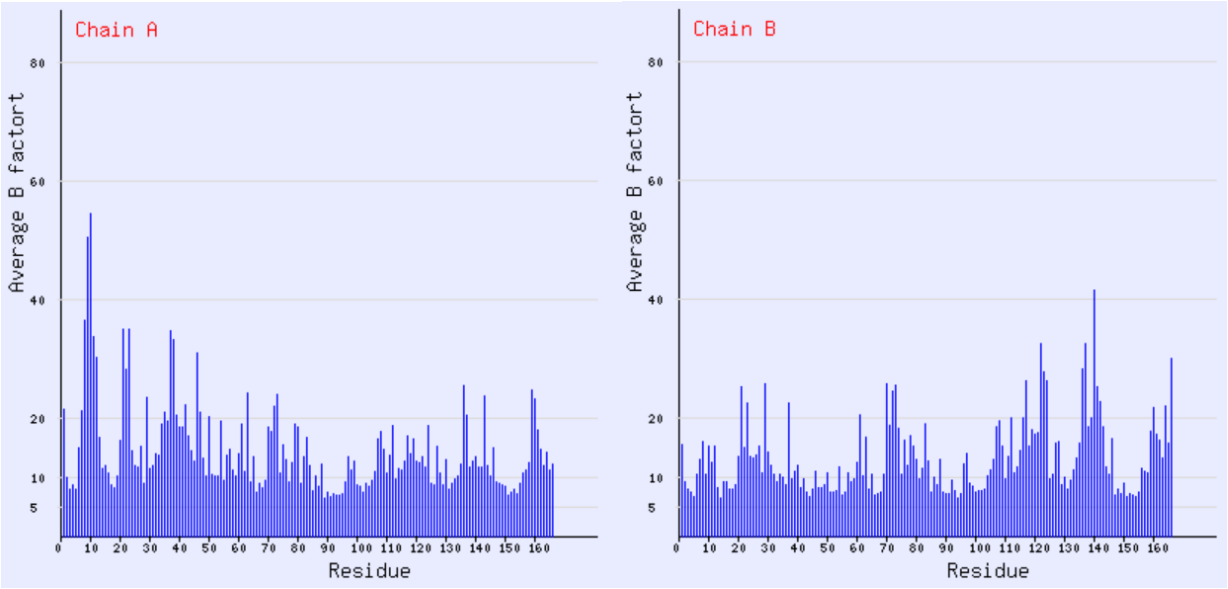


Рисунок 7. Значения температурного фактора для всех остатков модели 4ep4.

Как видно на рисунке 7, нет таких остатков, которые бы имели температурный фактор выше 60. В целом, его значение не более 30. Однако, есть остатки, которые превышают порог в 40, это Gly9 и Ile10 в цепи A и Arg140 в цепи B.

Комфортность окружения

Данный показатель не зависит от процедуры оптимизации и основан на физико-химических ограничениях для взаимодействующих групп атомов.

Физически невозможным окружением для атомов является пересечение их Ван-дер-Ваальсовых радиусов более чем на 0,4 Å (данная ситуация называется “clash’). В этой модели есть 38 пар атомов с clash более чем 0,4 Å.

Таблица 2. Часть таблицы пар атомов с большим перекрытием Ван-дер-Ваальсовых радиусов.

Atom-1	Atom-2	Interatomic distance (Å)	Clash overlap (Å)
1:A:73[B]:PHE:HE2	1:A:75:TYR:CD1	1.91	0.89
1:A:73[B]:PHE:CE2	1:A:75:TYR:CD1	2.71	0.78
1:B:45[B]:LYS:NZ	4:B:340:HOH:O	2.22	0.73
1:A:159:ARG:NH2	4:A:339:HOH:O	2.22	0.72
1:B:65:GLU:OE1	3:B:202:GOL:H32	1.90	0.72
1:B:29:ARG:NH1	4:B:472:HOH:O	2.22	0.71
1:A:12:HIS:ND1	4:A:352:HOH:O	2.24	0.70
1:B:105:TYR:HB2	1:B:150[B]:ILE:HD11	1.73	0.70
1:B:159:ARG:NH1	4:B:406:HOH:O	2.28	0.66

В таблице 2 представлена часть таких пар атомов с межатомным расстоянием (interatomic distance) и перекрытием (clash overlap). Как видно, максимальное перекрытие составляет 0,89 Å.

Через сервис WhatCheck[8] можно получить информацию о локальном окружении атомов. Если остаток имеет значение ниже -5.5, то это говорит о том, что в окружении остатка что-то происходит. Такими остатками, подходящими под это значение, являются Ile10, Lis23, Gln41, Phe73, Arg140, Arg159 в цепи A и lis23, Gln41, Tyr75, Arg10, Arg140 и Arg159.

Лиганды

Дополнительными молекулами в модели является ион Mg и глицерол ($C_3H_8O_3$ – GOL). Параметр LLDF описывает качество электронной плотности группы по отношению к ее соседним остаткам в модели. Если параметр LLDF больше 2, то электронная плотность плохо описывает лиганд. Для глицерола это значение равно 6.71, что явно больше 2 (табл.3).

Таблица 3. Описание лигандов в модели.

Mol	Type	Chain	Res	Atoms	RSCC	RSR	LLDF	B-factors(Å^2)	Q<0.9
3	GOL	B	202	6/6	0.85	0.27	6.71	23,27,31,37	0
2	MG	B	201	1/1	0.96	0.14	-	21,21,21,21	0

Маргинальные остатки

В таблице ниже представлены некоторые маргинальные остатки, которые были отобраны по определенным критериям.

Таблица 4. Маргинальные остатки.

№	Остаток, цепь	Критерий
1	Gly9 (A)	Высокое значение RSR и RSRZ
2	Ile10 (A)	Высокое значение RSR и RSRZ и неправильное окружение
3	Glu21 (A)	Высокое значение RSR и RSRZ
4	Gly22 (A)	Высокое значение RSR и RSRZ
5	Phe73 (A)	Высокое значение RSR и RSRZ
6	Leu127 (A)	Высокое значение RSR и RSRZ
7	Pro8 (A)	Находится в запрещенной области
8	Arg140 (A)	Повышенное значение B-фактора и неправильное окружение

9	Tyr75 (A)	Максимальное значение clash с Phe73 (A)
10	GOL (B)	Высокое значение LLDF

Анализ маргинальных остатков

Для анализа я выбрала маргинальные остатки под номерами 5, 6, 7, 9 и 10 из таблицы 4.

Остаток Phe73 в цепи A:

Значение RSR = 0.181, Z-score = 3.547. Как видно на рисунке 8, данных остаток имеет две альтернативные конформации. При уровне подрезки 0 они оба описываются электронной плотностью, однако не очень хорошо. При уровне подрезки 1 и 2 боковые цепи уже не описываются ЭП. Остаток находится в петле у поверхности белка, так что возможно это сказалось на качестве описания электронной плотности для данного остатка.

Кроме того, Phe73 имеет большое перекрытие с Tyr73, что может говорить о том, что либо одна из конформаций фенилаланина является неправильной, либо сам остаток Tyr73.

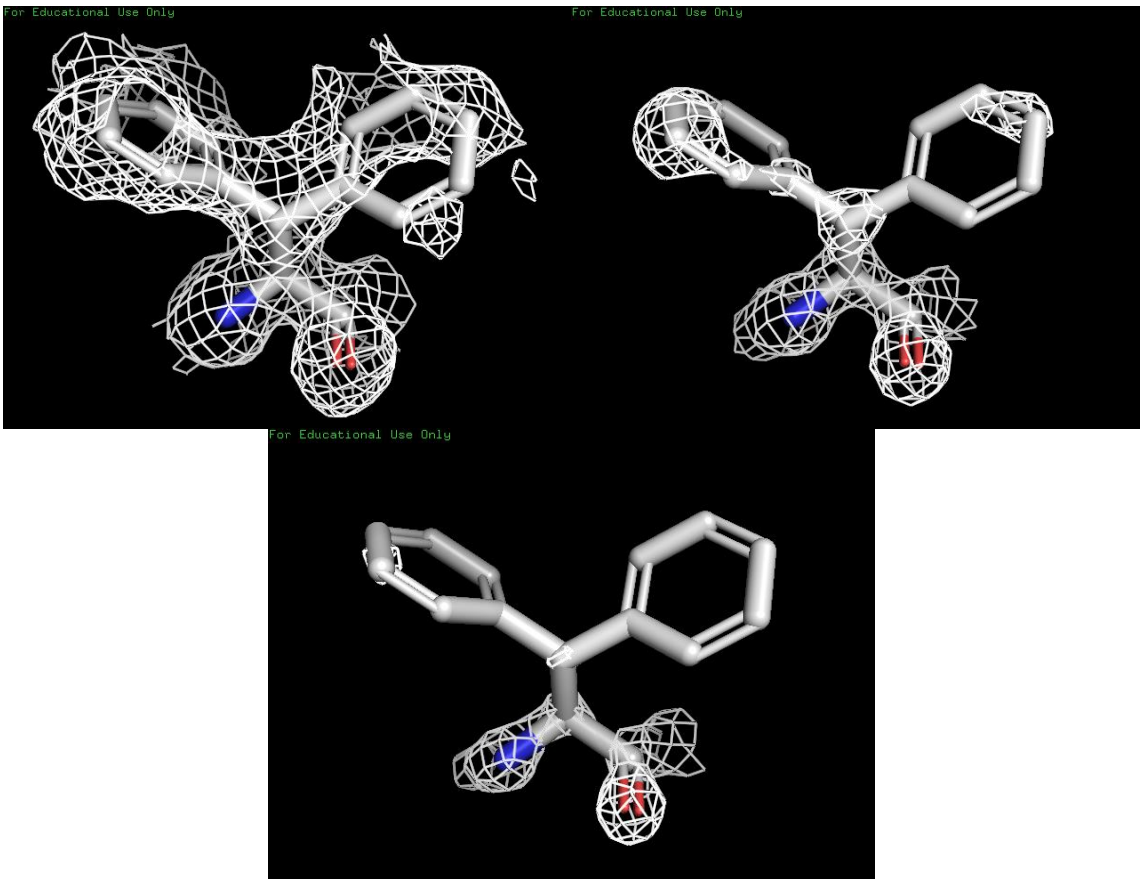


Рисунок 8. Phe73 с ЭП при уровнях подрезки 0, 1 и 2.

Остаток Pro8 в цепи A:

Транс-пролин на карте Рамачандрана находится в запрещенной области и хорошо описывается электронной плотностью (рис.9).

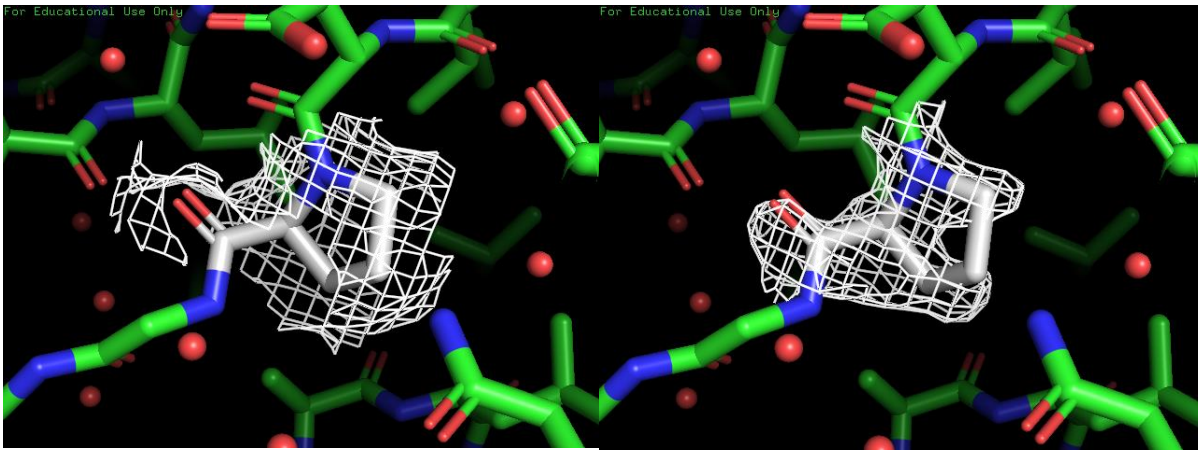
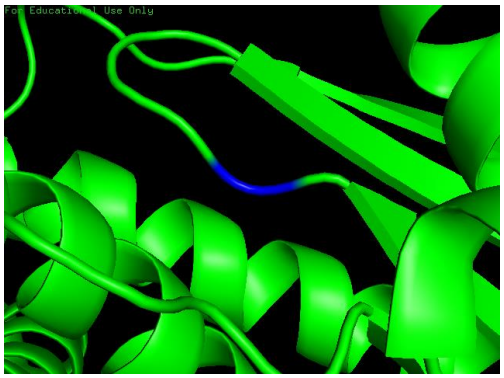


Рисунок 9. Pro8 с ЭП при уровнях подрезки 0 и 1.



Pro8 расположен в петле, но не в перегибе (рис.10). Возможно, наличие таких углов у данного пролина является особенностью структуры.

Рисунок 10. Остаток Pro8 в белке (выделен синим).

Остаток Leu127 в цепи A:

Данный остаток имеет несколько конформаций, так к тому же плохо описывается электронной плотностью. Поэтому он имеет высокое значение RSR и Z-score.

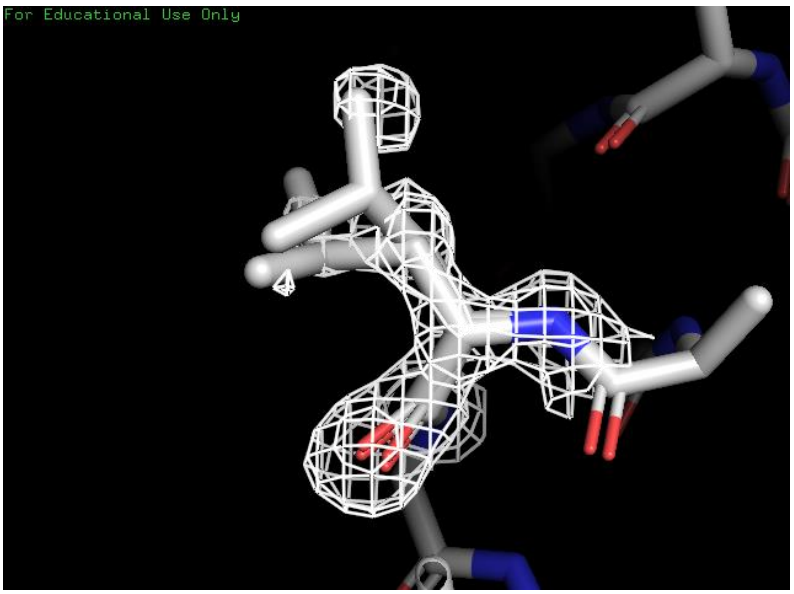


Рисунок 11. Le127, вписанный в электронную плотность, с уровнем подрезки 1.

Остаток Tyr75 в цепи A:

Тирозин имеет перекрытие Ван-дер-Ваальсовых радиусов в 0,89 Å. На рисунке ниже, видно что электронная плотность хорошо описывает остаток, но такое близкое расположение Phe73 дает повод задуматься о правильности данных остатков.

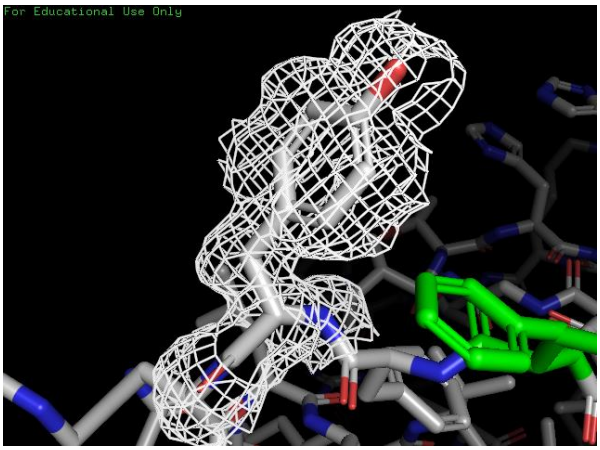


Рисунок 12. Tyr75, вписанный в ЭП, и Phe 73 (зеленый), уровень подрезки 0.

Лиганд Gol 202 в цепи B:

Глицерол имеет высокое значение LLDF и плохо описывается электронной плотностью при уровне подрезки 1. Он образует контакты в белком в двух местах. Вероятно, молекула плохо описывается данной электронной плотностью потому что это маленький лиганд.

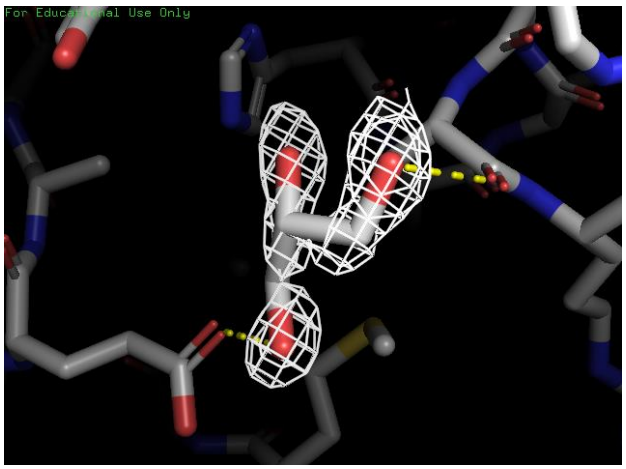


Рисунок 13. Глицерол (Gol202), вписанный в электронную плотность.

Сравнение модели 4ep4 с моделью из PDB_redo

Сервис PDB_redo позволяет получить более точную модель из нашей оригинальной модели. На выходе мы получаем некоторые параметры новой модели, а так же саму структуру.

Сравнительные характеристики приведены в таблице ниже.

Таблица 5. Выходная таблица из сервиса PDB_redo.

Validation metrics from PDB-REDO		
	PDB	PDB-REDO
Crystallographic refinement		
R	0.1491	0.1315
R-free	0.1786	0.1730
Bond length RMS Z-score	0.280	0.847
Bond angle RMS Z-score	0.584	0.956
***Model quality** (raw scores | percentiles)*		
Ramachandran plot appearance	92	92
Rotamer normality	89	97
Coarse packing	N/A	N/A
Fine packing	42	52
Bump severity	93	29
Hydrogen bond satisfaction	96	96

Как мы видим, значение R-фактора немного улучшилось, однако R-free остался почти таким же (измерение на 0,5%). Увеличились S-score для длины и углов связей. Кроме этого, увеличилось количество ротамеров.

Для сравнения, насколько остаток лучше или хуже теперь вписывается в электронную плотность, используется специальная оценка (delta RSCC). Если столбец зеленый, то остаток лучше вписывается; если красных – то хуже.

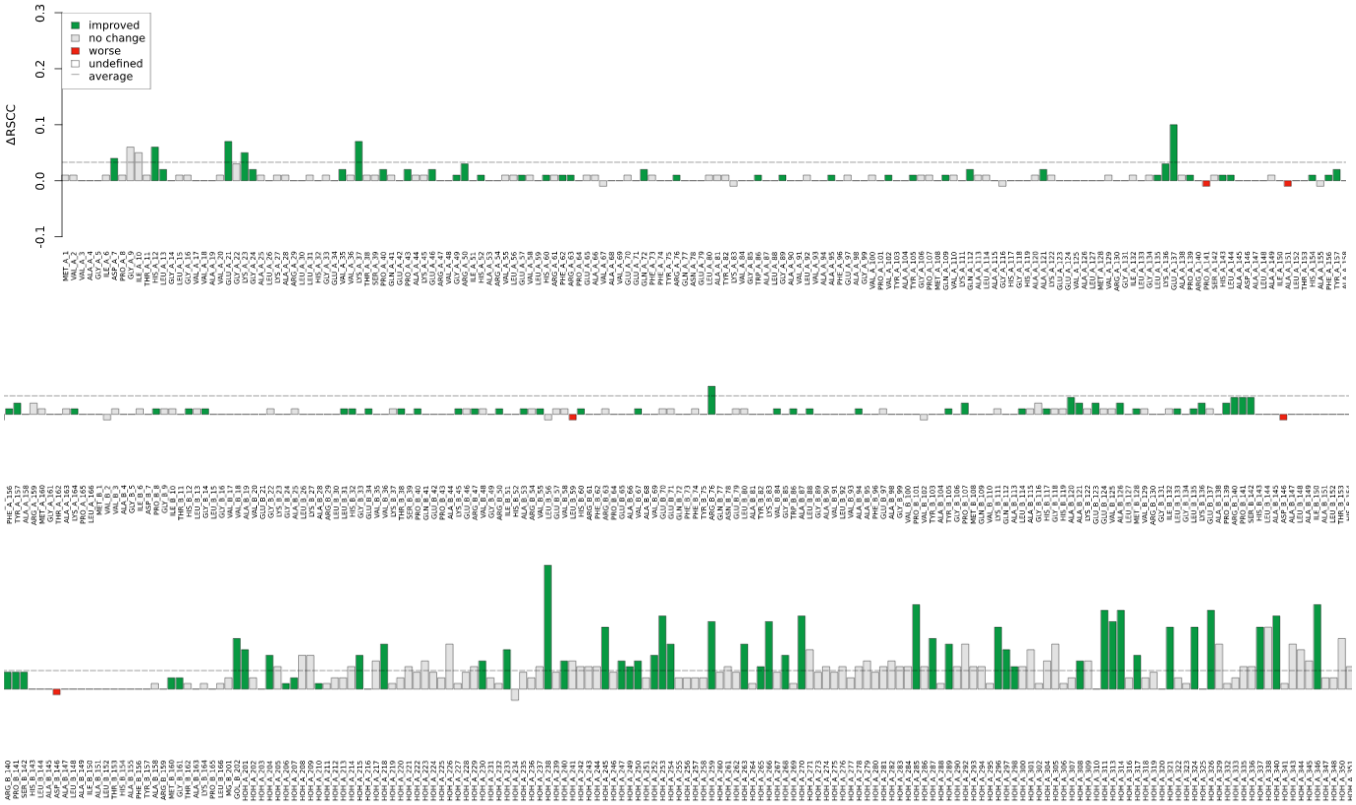


Рисунок 14. Распределение RSCC по остаткам модели.

Как видно из рисунка 14, некоторые аминокислоты лучше вписываются в электронную плотность, а у 4 остатков обратная ситуация. Интересно, что особенно лучше вписываются молекулы воды, которые на данном рисунке находятся в самом конце.

В целом, геометрия остатков улучшилась в сравнении с исходной моделью.

На рисунке 15 ниже представлен Arg76, зеленым цветом обозначена исходная модель, красным – модель, полученная из PDB_redo. Как мы видим, данный остаток имеет немного другое расположение. И по сравнительно таблице, новый остаток лучше вписывается электронную плотность.

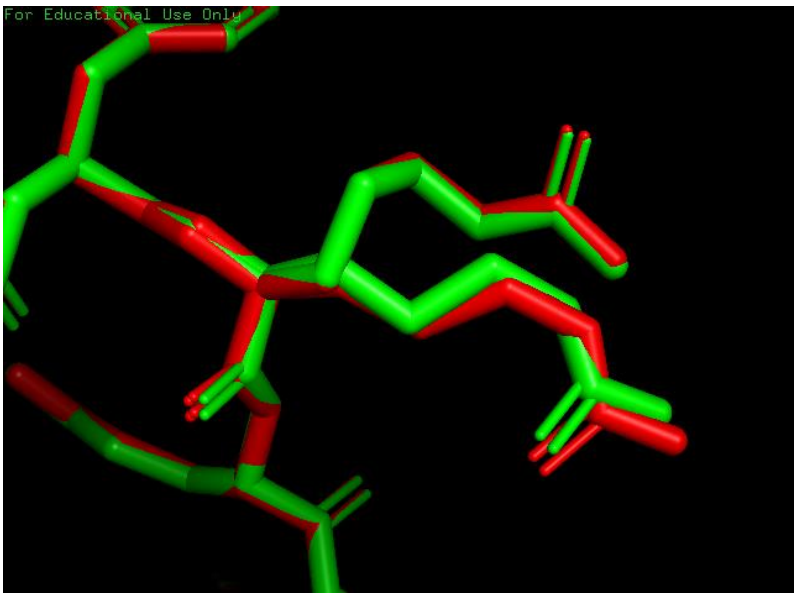


Рисунок 15. Совмещение Arg76 из исходной (зеленый) и новой (красный) моделей.

Много остатков, у которых остатки из этих двух моделей совпадают (рис.16).

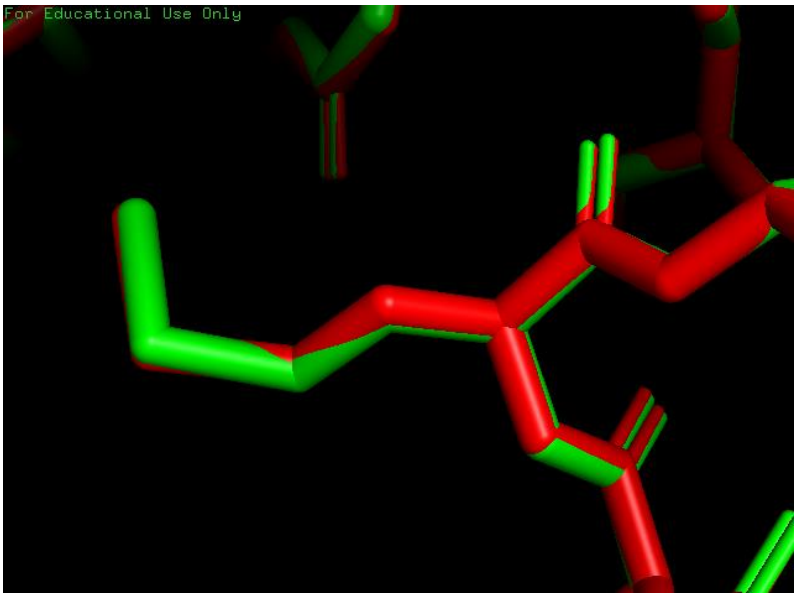


Рисунок 16. Совмещение Met108 из исходной (зеленый) и новой (красный) моделей.

Заключение

Модель белка 4ep4 в целом нормально соответствует известным экспериментальным данным. Высокое значение R-фактора, R-free и их разница это подтверждает.

Однако, наличие маргинальных остатков дает повод задуматься насколько модель все же соответствует реальности.

Новая модель, полученная с помощью сервиса PDB_redo, имеет более лучшие значения основных показателей модели.

Литература и сервера

1. Chen, L., Shi, K., Yin, Z. & Aihara, H. Structural asymmetry in the Thermus thermophilus RuvC dimer suggests a basis for sequential strand cleavages during Holliday junction resolution. *Nucleic Acids Res.* **41,** 648–656 (2013).
2. Chen, L., Shi, K., Yin, Z. & Aihara, H. Structural asymmetry in the Thermus thermophilus RuvC dimer suggests a basis for sequential strand cleavages during Holliday junction resolution. **41,** 648–656 (2013).
3. Thermus thermophilus RuvC structure. Available at: http://www.ebi.ac.uk/pdbe/entry/pdb/4ep4.
4. https://www.rcsb.org/pdb/explore/explore.do?structureId=4ep4
5. https://eds.bmc.uu.se/cgi-bin/eds/uusfs?pdbCode=4ep4
6. https://www.ebi.ac.uk/pdbe/entry/pdb/4ep4/
7. http://molprobity.biochem.duke.edu/
8. http://swift.cmbi.ru.nl/servers/html/index.html