Отчет о качестве расшифровки структуры белка 1VI9 методом рентгеноструктурного анализа

Аннотация

В данном отчете мной была собрана, получена и проанализирована информация о структуре 1VI9, ее качестве, геномном проекте, в рамках которого она была расшифрована, и дополнительных данных, которые, как мне кажется, могут объяснять некоторые качественные особенности этой модели.

Введение и анализ литературы

Структура белка пиридоксаминкиназы 1VI9 была получена методом аномального рассеяния с использованием Se-Met замещения и оказалась доступной в PDB в декабре 2003 года. Публикация в журнале PROTEINS состоялась спустя немалое время в 2005 году, и такая задержка была сделана сознательно. Дело в том, что большой коллектив ученых (указан в списке литературы) в 2000 году запустил крупный проект по решению структур бактериальных белков, которые 1) не имели похожих структур в банках данных, а следовательно не могли быть решены методом молекулярного замещения, 2) были широко распространены в различных геномах, 3) вероятно могли бы иметь широкое применение в качестве мишени для антибиотиков, так как имели высокое значение для роста бактерии.

В рамках проекта ученым удалось решить 80 структур, из которых 49 имели уникальные для PDB модели, а остальные были получены методом молекулярного замещения по появившимся в ходе проекта моделям или (всего несколько структур) по моделям других авторов, возникшим позже 2000 года. Статья с результатами проекта вышла лишь в 2005 году, так как авторы в ней проанализировали, насколько полезным для науки оказались их структуры. По их данным, лишь одна из полученных моделей к 2005 году (~1.5 года разницы) не была использована для молекулярного замещения, что говорит о высокой востребованности данных.

Одной из целей проекта было разработать высокопроизводительный способ расшифровки структур. Для этого совместно с двумя работниками PDB, которым авторы выражают свою благодарность в статье, был специально разработан формат (тоже текстовый), позволивший автоматически и без задержек пополнять базу данных (образец может быть доступен по E-mail одного из авторов и в сопроводительных материалах статьи). С использованием этого формата был создан полуавтоматический конвейер (pipeline) для обработки данных, создания модели и ее анализа. При анализе качества моделей использовались такие критерии, как R-факторы (work & free), карта Рамачандрана, чрезмерно близкие контакты, ротамеры, водородные связи для Asn, Gln и His. В проекте были заданы достаточно жесткие критерии на отбор качества структур (1)-(5), готовых к публикации.

$$\text{Maximum } R_{free} = -0.02d^2 + 0.13d + 0.11 \quad (1)$$

$$\text{Maximum } R_{work} - R_{free} = -0.01d^2 + 0.065d - 0.02 \quad (2)$$

$$\text{Minimum percentage of residues in Ramachandran core} = 100 \times (-0.04d + 0.96) \quad (3)$$

$$\text{Maximum number of abnormal } \chi1 - \chi2 \text{ angles/100 residues} = 0.075d + 0.75 \quad (4)$$

$$\text{Maximum number of short contacts/100 residues} = 2.8571d - 5.5714 \; d < 2.3 \text{ Å} = 1.0 \; d > 2.3 \text{ Å} \quad (5)$$

Теперь подробнее остановимся на самой структуре 1VI9. Это пиридоксамин-, а по данным Uniprot и пиридоксаль-киназа (см. раздел “Catalytic activity”), в кристаллографической ячейке расположено 4 полипептидных цепи, биологическая единица состоит их двух цепей. Структура имеет разрешение 1.96Å, R-фактор 0.198 и R-free 0.256, 0.3% остатков (то есть 4 остатка из 1196 входящих в асимметрическую ячейку) располагаются, по данным PROCHECK, за пределами допустимой области карты Рамачандрана, однако, по данным MolProbity, все остатки данной структуры лежат в пределах допустимой области. Отклонение для длин связей, углов и планарности (CCP4/REFMAC5) равны соответственно 0.012, 2.2 и 0.012, что говорит о среднем качестве модели относительно всего проекта (см. Supplementary Table 2 к статье [1]). Число измеренных рефлексов составило 83507, с полнотой 98.9%. При этом число рефлексов с интенсивностью более трех σ насчитывало 78803 шт. и составило 94.37%. Разрешение гармоник Фурье находилось в пределах 43.03-1.96Å, авторы не выбрасывали рефлексы при создании модели. Для подсчета R-free были использованы 4167 (5%) рефлексов. Более удобно информация скомпанована на странице PDB_redo для данной структуры. Стоит отметить, что оптимизированные там модели (консервативная и полная) по подавляющему числу пунктов были сделаны лучше, чем модель в исходном белке.

Тем не менее, приведенные характеристики указывают на высокое качество структуры, однако, уже при беглом рассмотрении возникают некоторые серьезные замечания и сомнения. Во-первых, в файле .pdb в заголовке и параметрах структуры имеется серьезная путаница, и многие необходимые поля заполнены значением “NULL”, например поле “COMPLETENESS FOR RANGE” и многие другие. Во-вторых, авторы поместили в кристаллографическую ячейку 4 полипептидных цепи из двух гомодимеров, что может свидетельствовать о стремлении получить более хорошие оценки качества структуры. В-третьих, поиск других структур пиридоксалькиназ показал наличие в PDB двух очень похожих структур пиридоксалькиназ овцы (1LHP и 1LHR), опубликованных китайцами в феврале 2003 года (напомню, структура американцев была опубликована 30 декабря 2003 года, а была выложена всего лишь 1 декабря (за месяц до этого), для сравнения, у тех же китайцев с передачи данных до их выхода прошло около 8 месяцев. Очевидно, авторы имели цель уложиться до конца 2003 года, вероятно потому, что их структура оказалась не первой структурой пиридоксалькиназы, и они потенциально могли за 10 месяцев решить ее методом молекулярного замещения, в то же время нет оснований обвинять ученых во лжи, а к их чести скажем, что они ссылаются в своей работе на структуру китайцев (ту из двух, где разрешение ниже), правда в контексте структур, появившихся за период с 2003 по 2005 год, то есть с намеком, что эти структуры сделаны, основываясь на данных их, американского, геномного проекта.

Вскрывшийся факт оставляет неприятное впечатление и требует высказать дополнительные аргументы в пользу структур (1LHP и 1LHR). Они были решены методом множественного изоморфного замещения, белок был очищен из мозга

овцы, а не экспрессирован в плазмиде, одна из структур содержит комплекс с АТФ. По данным Web-страницы PDB этих структур, они имеют разрешение 2.10 и 2.60 Å число перекрываний (clashscore) 8 и 13, R-factor 0.196 и 0.198, R-free 0.224 и 0.222, маргинальных остатков по Рамачандрану 0 и 0.2%, маргиналов по боковой цепи 3.2 и 6.3%, маргиналов по RSRZ 7.8% и 3.9%. Для сравнения в 1VI9 те же параметры составляют 1.96 Å, 7, 0.198 и 0.256, 0%, 3,6%, 3%. То есть структуры получились одинакового качества, при этом первой появилась структура овцы, и авторы еще долго изучали ее и механизм ее работы. Из статьи же с геномным проектом бактерий не следует, что авторы подробно занимались 1VI9, так как они уделили большое внимание нескольким другим структурам, в то же время, они очень старались успеть опубликовать 1VI9 в 2003 году, и вероятно, знали о существовании 1LHP и 1LHR.
Кстати, один из абзацев статьи [1] посвящен как раз рассмотрению возможности создания публичного списка белков, которые ученые расшифровывают различными методами, дабы избежать излишних затрат и «совпадений».
Итак, сравнение и анализ литературных данных показал, что в ходе расшифровки молекул, в том числе 1VI9, ученые

1. имели высокие качественные критерии отбора моделей,
2. стремились наработать много структур нерешенных ранее белков,
3. разработали поток обработки данных и специальный формат, который, по их словам, позволяет не допускать неточности, возникающие при работе вручную,
4. в отношении структуры 1VI9 имели мотив, чтобы как можно скорее выложить ее в PDB,
5. допустили ряд ошибок при оформлении заголовка и заполнении данных файлов .pdb, причем не только в структуре 1VI9 (разрешение 1.96 Å), но и в 1VH9 (разрешение 2.15 Å), и в 1VI1 (разрешение 2.95 Å), и, вероятно, в других структурах проекта,
6. Сделали вывод, что при отсутствии проблем с кристаллизацией и введением Se-Met в белки, хорошем финансировании и правильной организации работы, хороший геномный проект может добиться скорости решения структур в несколько сотен моделей в год, при сборе данных аномального рассеяния около 2 суток и обработке этих данных в течение около 24 часов.

Результаты

Выявленные в ходе анализа литературы и краткого анализа качества структуры особенности и противоречия требуют детального рассмотрения и пояснений.

Во-первых, большинство незаполненных полей файла .pdb, оказались дублированы в различных полях “REMARK ***”, например

REMARK 200 COMPLETENESS FOR RANGE   (%) : 98.9.

Вместе с тем часть полей дублируется, а кое-где наблюдается потеря данных. Так например в поле “REMARK 500” можно прочесть, что авторы собирались привести список атомов, имеющих слишком близкое расположение, однако, они ссылаются на поле REMARK 375, которого в файле попросту нет. Тем не менее, в других подобных примерах списки оформлены правильно, атомы приведены корректно и совпадают с данными WhatCheck для 1VI9.

По тому, что авторы перечисляют недостающие атомы и остатки, которые не удалось определить, а также другие маргинальные остатки и атомы, можно сделать вывод, что они многое учли в своей модели, и грубых ошибок тут быть не должно.

Действительно, атомов, расположенных слишком близко, всего семь, отклоняющиеся от нормы углы между атомами едва ли заметны (Рис. 1А, взят самый грубый пример по данным поля REMARK 500), отклоняющиеся более чем на 10 градусов торсионные углы также слабо видны (в поле REMARK это всего один остаток GLN40 цепи D, Рис. 1Б). Как можно видеть, этот остаток, вероятно, очень важен для димеризации субъединиц, поэтому его небольшие изменения вполне возможны.

Указаны в файле и молекулы воды, удаленные от атомов белка на расстояние более 5 Å, но предположение о возможности неправильного расположения воды относительно белка в кристаллической ячейке пришлось отвергнуть по крайней мере для самой далекой молекулы воды, которая даже при построении соседних молекул белка все-равно не имела с ними общих водородных связей, а соединялась лишь с другой молекулой воды (без Рис.). Из параметров этого атома: (HETATM 8986 O   HOH A 443   16.724 58.813 21.609 1.00 53.85   O) видно, что это нормальная молекула с довольно большим температурным фактором, но электронная плотность в районе атома кислорода действительно сгущалась, так что придраться тут не к чему.

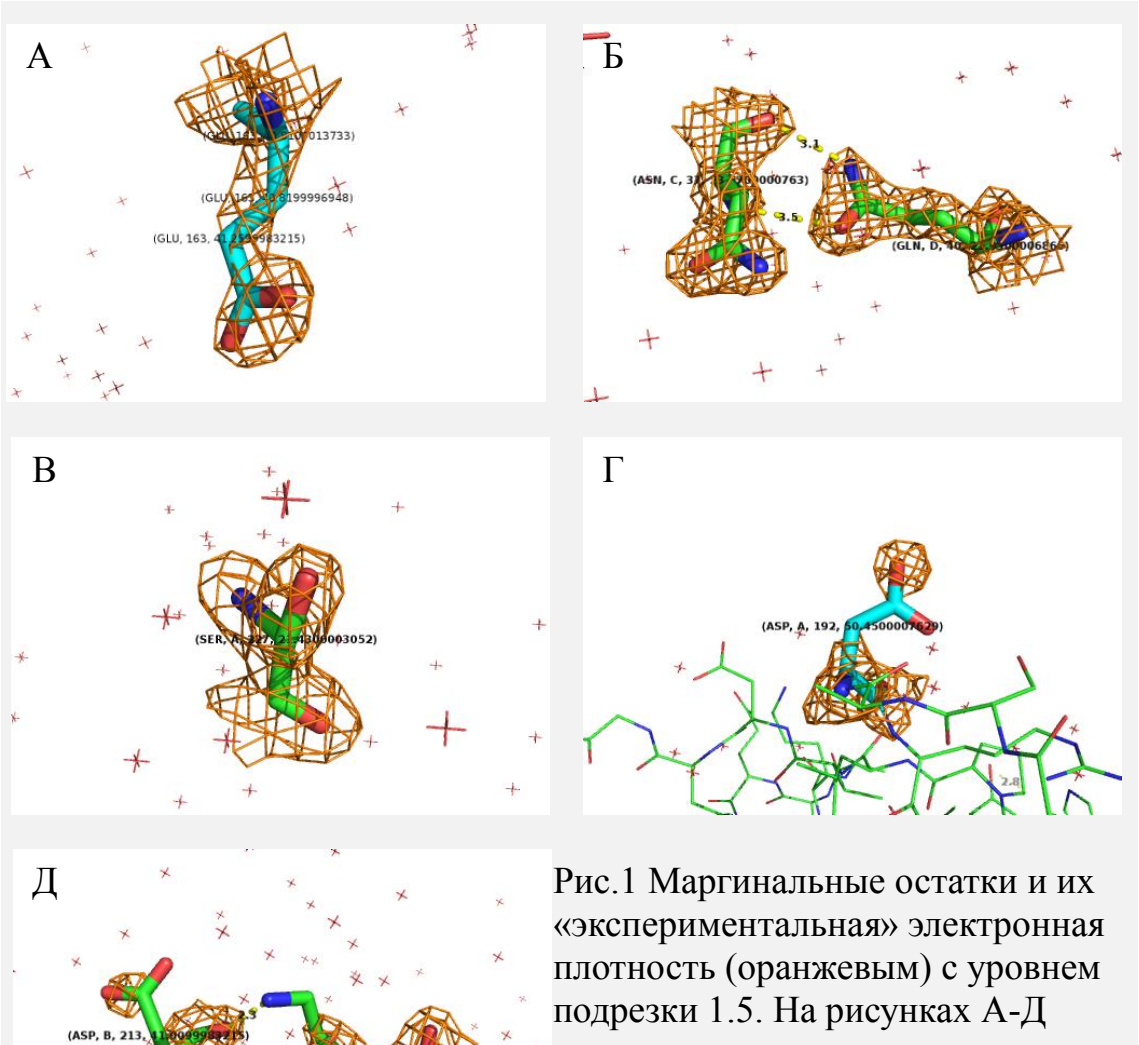

Рис.1 Маргинальные остатки и их «экспериментальная» электронная плотность (оранжевым) с уровнем подрезки 1.5. На рисунках А-Д

По данным WhatCheck, все вышеперечисленные проблемы также присутствовали, но они не носили грубого характера. Стоит отметить один странный ротамер серина 227, любой цепи (Рис. 1В) и другой наиболее нехарактерный угол CA-CB-CG в остатке Asp192 цепи A (Рис. 1Г), а также большое количество нехарактерно близких контактов между атомами, самый близкий из которых приведен на Рис. 1Д. Тут можно сказать, что экспериментальная электронная плотность не слишком хорошо определена для данных атомов, а это вызывает их отклонение от нормы. Несколько десятков остатков имеют, по данным WhatCheck, нехарактерное окружение, действительно, возьмем самый маргинальный по этому признаку остаток и посмотрим. Это Tyr189 любой цепи, который не окружен никакими атомами ближе 3.5 Å, что странно, но судя по электронной плотности, все у модели согласуется с экспериментом. В файле нашлось также три проблемы с названием атомов аргинина, шесть атомов, перенумерация которых улучшила бы водородные связи (надо сделать парные замены углерода и азота), несколько десятков неверно указанных доноров и акцепторов водорода в водородных связях, а также WhatCheck рекомендует проверить, не изменить ли на тысячные доли матрицу движения для кристаллографической ячейки и предлагает матрицу деформации (6)

| 0.998392 -0.000097 -0.000192|
| -0.000097 0.999091 -0.000561| (6)
| -0.000192 -0.000561 0.999386|

Последнее предложение мне кажется излишним, так как модель построена достаточно хорошо.

Очень важными, на мой взгляд, являются графики, так или иначе описывающие качество модели или электронной плотности по каждому из остатков в структуре. На примере Рис. 2 можно увидеть, что очень часто на этих графиках было сильное отклонение по качеству для остатков 115-125 и 190.

Обратим внимание, что

1) Эти участки расположены близко друг к другу
2) Эти участки представлены петлями и небольшим бета-слоем, и прикрывают большую щель, заполненную водой (Рис. 3).

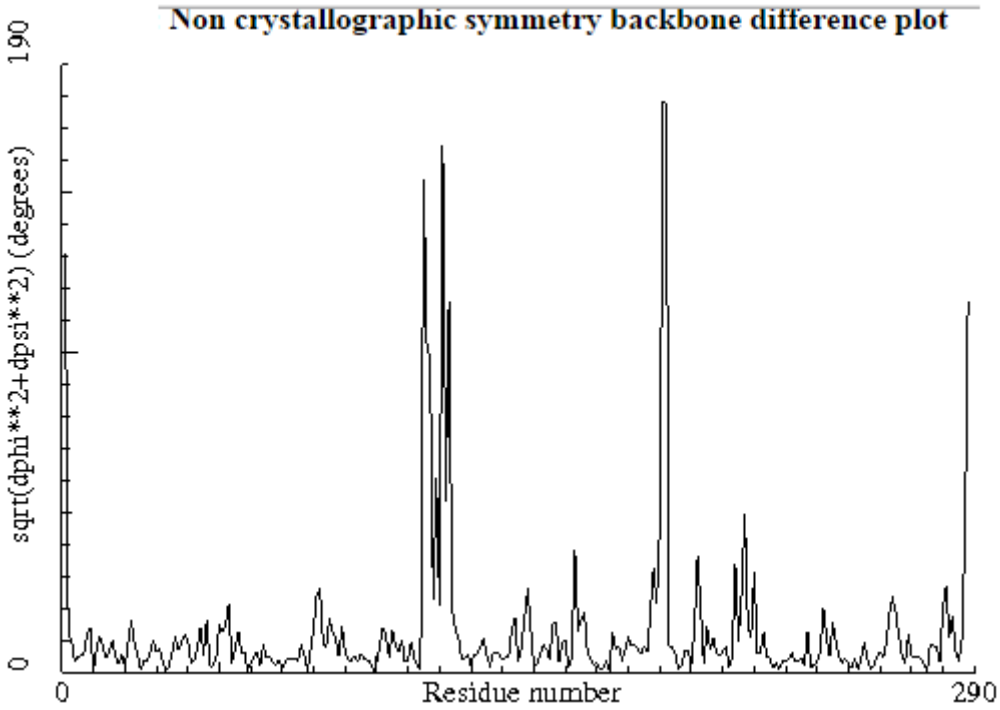

Рис.2. График отличия некристаллографической симметрии по скелету для цепей A и B. Высокие значения говорят о плохой симметрии модели в этих остатках.

Воспользовавшись любой структурой пиридоксалькиназ, содержащих АТФ и пиридоксаль или их аналоги (например 3KEU), мы сможем увидеть, что в этой области и располагается активный центр фермента, который, по-видимому, в отсутствии лигандов был не стабилен, и поэтому структура имела высокий B-фактор и другие проблемы в конкретно этих областях. Таким образом, по графикам из WhatCheck можно предположить, где у молекулы окажется активный центр. В нашем случае это предположение подтвердилось.

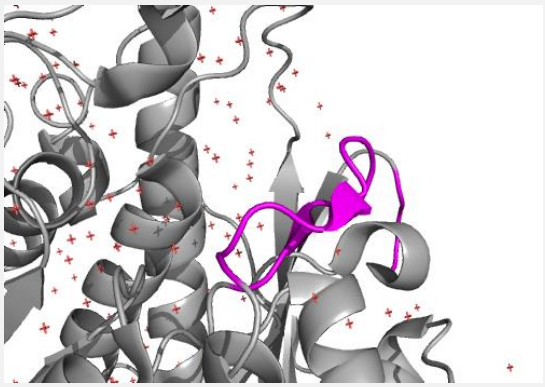

Рис.3 Модель 1VI9 в виде вторичных структур. Розово-фиолетовым показаны остатки 115-125 и 189-171, обладавшие значительно меньшим качеством по ряду критериев, и прикрывающие активный центр. Показаны молекулы воды.

Заключение

Стоит отметить, что на PDB_redo были построены модели немного лучшего качества, где R-факторы, качество упаковки, карта Рамачандрана, ротамеры, общее число сближений, количество неправильных доноров были улучшены на 5-50% в ущерб положению скелета белка и отклонению длин связей. Подсчитано, что число значимо улучшенных остатков составило 109-175, ухудшенных – 10-12. То есть, можно сделать вывод, что обе автоматически перестроенные модели несколько лучше, чем исходная модель. Все-таки, она была получена уже более 10 лет назад, вероятно не учитывала некоторых критериев, была сделана в большой спешке и, вероятно, при не таком большом внимании к деталям, как можно ожидать от работ с отдельными кристаллами, а не в геномном проекте. Тем не менее, построенная модель имеет высокое качество, и даже позволяет косвенно судить о месте расположения активного центра фермента.

Печально, что стремясь наладить конвейер по созданию структур, авторы не заметили ошибки, вносимой в файлы автоматически. Также очевидно, что жесткие критерии и сроки, а также наличие опередившей данную структуру работы [2] могли повлиять на стремление авторов создать модель с более хорошими показателями, что могло привести к подгонке. Частично это видно из отчета WhatCheck и наличия двух одинаковых гомодимеров в кристаллографической ячейке. Также неприятно наблюдать намеренное завышение уникальности работы путем приуменьшения роли других ученых.

Список литературы.

1. **Structural analysis of a set of proteins resulting from a bacterial genomics project.** Badger, J., Sauder, J.M., Adams, J.M., Antonysamy, S., Bain, K., Bergseid, M.G., Buchanan, S.G., Buchanan, M.D., Batiyenko, Y., Christopher, J.A., Emtage, S., Eroshkina, A., Feil, I., Furlong, E.B., Gajiwala, K.S., Gao, X., He, D., Hendle, J., Huber, A., Hoda, K., Kearins, P., Kissinger, C., Laubert, B., Lewis, H.A., Lin, J., Loomis, K., Lorimer, D., Louie, G., Maletic, M., Marsh, C.D., Miller, I., Molinari, J., Muller-Dieckmann, H.J., Newman, J.M., Noland, B.W., Pagarigan, B., Park, F., Peat, T.S., Post, K.W., Radojicic, S., Ramos, A., Romero, R., Rutter, M.E., Sanderson, W.E., Schwinn, K.D., Tresser, J., Winhoven, J., Wright, T.A., Wu, L., Xu, J., Harris, T.J. (2005) Proteins 60: 787-796, PubMed: 16021622

2. **Crystal structure of brain pyridoxal kinase, a novel member of the ribokinase superfamily**. Li, M.H., Kwok, F., Chang, W.R., Lau, C.K., Zhang, J.P., Lo, S.C., Jiang, T., Liang, D.C. (2002) J.BIOL.CHEM. 277: 46385-46390, PubMed: 12235162

3. Сервер pdbreport

4. Сервер PDB_redo

5. Сервер ESD

6. Сервер RCSB PDB

7. Данные MolProbity

8. сайт Uniprot