ЖУРНАЛ ФИЗИЧЕСКОЙ ХИМИИ, 2023, том 97, № 9, с. 1324–1328

ХЕМОИНФОРМАТИКА И КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ

УДК 004.658,577.151.3

БАЗА ДАННЫХ ИНТЕРМЕДИАТОВ ХИМИЧЕСКИХ РЕАКЦИЙ ФЕРМЕНТАТИВНОГО КАТАЛИЗА ENIAD

© 2023 г. А. А. Московский[a], Д. А. Фирсов[a], М. Г. Хренова[a,b], В. А. Миронов[a], Т. И. Мулашкина[a], А. М. Кулакова[a], А. В. Немухин[a,c,*]

[a]*Московский государственный университет им. М.В. Ломоносова, Химический факультет, Москва, Россия*

[b]*ФИЦ Биотехнологии РАН, Москва, Россия*

[c]*ИБХФ имени Н.М. Эмануэля РАН, Москва, Россия*

******e-mail anem@lcc.chem.msu.ru*

Поступила в редакцию 30.03.2023 г.
После доработки 30.03.2023 г.
Принята к публикации 03.04.2023 г.

Для ферментативного катализа характерны многостадийные химические реакции на пути от фермент-субстратных комплексов до продуктов. В ряде случаев в ходе экспериментальных исследований удается характеризовать структуру и свойства интермедиатов сложных химических реакций в белках. Применение современных компьютерных методов моделирования позволяет существенно дополнить знание о механизмах реакций ферментативного катализа и представить подробные данные о реакционных интермедиатах, включая структуры с атомным разрешением. Накопленные к настоящему времени материалы позволяют создать уникальную базу данных, названную ENIAD (ENzyme-In-Action-Data bank). В статье описаны принципы построения базы данных ENIAD, а также мультиплатформенный веб-интерфейс для доступа к данным (https://lcc.chem.msu.ru/eniad/).

Ключевые слова: ферментативный катализ, реакционные интермедиаты, молекулярное моделирование, базы данных

DOI: 10.31857/S0044453723090133, **EDN:** XKQAYW

Изучение каталитических процессов, осуществляемых ферментами, необходимо для развития биотехнологии и биомедицины. Знание детальных механизмов химических преобразований в активных центрах белковых макромолекул является существенным для понимания биологической функции и эволюции ферментов [1]. Для ферментативного катализа характерны многостадийные химические реакции на пути от фермент-субстратных комплексов до продуктов с возможным образованием реакционных интермедиатов. Использование методов кристаллографии, ядерного магнитного резонанса, криоэлектронной микроскопии во многих случаях позволяет характеризовать структуры и свойства молекулярных систем на различных участках реакционного пути. Результаты структурных исследований, как правило, размещаются в наиболее авторитетной в данной научной области базе данных – базе данных белковых структур (Protein Data Bank (PDB)) [2]. Каждая размещенная в этом банке данных структура имеет свой уникальный идентификационный номер (PDB ID). Также известны базы данных кинетических параметров ферментативных реакций, такие как сервис Mechanism, Annotation and Classification in Enzymes (MACiE) [3], содержащий информацию о механизмах и путях реакций, и база данных EzCatDB [4], построенная по результатам анализа литературных данных. Сервис M-SCA [5], комбинирующий возможности MACiE [3] и атласа активных центров ферментов Catalytic Site Atlas [6] был создан как логическое развитие двух проектов. Во всех перечисленных базах данных структуры и свойства интермедиатов ферментативных реакций представлены незначительно, прежде всего, вследствие сложностей экспериментального определения необходимых характеристик участников быстрых химических реакций.

Новые возможности в исследованиях механизмов ферментативных реакций предоставляют современные компьютерные методы молекулярного моделирования, в том числе метод квантовой механики/молекулярной механики (КМ/ММ) [7, 8]. По результатам расчетов энергетических профилей химических реакций в активных центрах можно локализовать стационарные точки на энергетических поверхностях и характеризовать локальные минимумы (интермедиаты) и седловые точки (переходные состояния). При подоб-

ных расчетах генерируется огромный объем данных, который практически не предоставляется для возможного последующего использования научным сообществом. В незначительном числе работ в сопутствующих материалах научных статей публикуются данные по атомным координатам и энергиям в найденных стационарных точках. Учитывая необходимость обеспечить воспроизводимость результатов научных исследований, существенно более объемный массив информации по реакционным интермедиатам ферментативного катализа должен быть доступен пользователям.

База данных и соответствующий веб-сервер, названный ENzymes-IN-Action-Databank (ENIAD), созданы для того, чтобы устранить отмеченные проблемы. Прежде всего, следовало изменить и улучшить формат ранее созданных баз данных. В ходе компьютерного моделирования механизмов ферментативного катализа получается набор структур интермедиатов реакции, соответствующих одной и той же макромолекуле или комплексу белковых молекул. Более того, метаданные моделирования разнообразны и могут не вписываться в специфический PDB формат. Если использовать те же принципы хранения информации, что и в базе данных PDB, то каждой структуре каждого интермедиата необходимо приписывать собственный идентификатор PDB ID, что приведет к потере связи между интермедиатами одной и той же реакции.

База данных ENIAD, в частности, создает платформу для специалистов по компьютерной химии, на которой может происходить хранение, анализ и обмен информацией, полученной в результате дорогостоящих расчетов. Для этой цели применяются средства для размещения и извлечения таких данных, как трехмерные полноатомные структуры стационарных точек вдоль реакционного пути, которые удобно использовать через интернет. Предполагаются две основные стратегии применения ENIAD: (1) загрузка данных о конкретном реакционном пути в базу данных; каждый результат будет доступен с помощью постоянного универсального идентификатора ресурса (uniform resource identifier (URI)); (2) поиск записей в базе данных по ключевым словам, например, по названию молекулы, его PDB ID, названию организма, из которого извлечен фермент, по идентификатору публикации DOI и т.д.

С учетом интересов пользователей мы предлагаем три основных уровня организации данных:

• “Состояние” (“State”). Это один из низших уровней организации данных в ENIAD. Каждое “состояние” соответствует единственной точке в конфигурационном пространстве молекулярной системы. Требуемые данные – энергия “состояния” и трехмерная полноатомная структура (3-D structure). “Состояние” относится либо к точке минимума на энергетической поверхности (реагенту, интермедиату или продукту реакции), либо к седловой точке (переходному состоянию), либо к произвольной структуре. Для каждого “состояния” может быть организован набор метаданных, содержащих детали моделирования – описания методов, протоколов, компьютерных программ.

• “Путь реакции” (“Reaction path”) представляет собой последовательность “состояний”. “Путь реакции” начинается с “состояния” реагента и заканчивается “состоянием” продукта. Все промежуточные “состояния” могут относиться к интермедиатам или переходным состояниям. Для каждого “пути реакции” может быть организован набор метаданных.

• “Реакция” (“Reaction”). Каждая “реакция” определяется ферментом и субстратом. Это высший уровень организации данных в ENIAD, в котором могут комбинироваться теоретические и экспериментальные данные. “Реакция” может быть представлена несколькими “путями реакции”, которые могли быть получены разными научными группами, или с применением разных методов моделирования.

Одной из задач при разработке ENIAD является обеспечение воспроизводимости результатов моделирования, поэтому описание метода расчета, информация о компьютерной программе и полное описание файлов для запуска расчетов составляет важную часть базы данных. В будущем, ENIAD сможет хранить связи с конкретными приложениями, инкапсулированным при помощи технологии контейнеров.

ИЛЛЮСТРАЦИЯ ВОЗМОЖНОСТЕЙ ENIAD

На данный момент база данных ENIAD содержит около полусотни систем, включая результаты моделирования и для нативных ферментов, и для макромолекул с точечными заменами аминокислотных остатков. Для всех систем представлены стационарные точки, отвечающие минимумам и переходным состояниям на энергетических профилях, полученных методами КМ/ММ. Веб-интерфейс базы данных предоставляет возможность поиска по следующей информации: идентификатор белка в PDB (PDB ID), название реакции или общепринятое сокращение (аббревиатура). По результатам поискового запроса будет выдан список подходящих “путей реакции” с указанием DOI публикаций (если имеется), в которых они описаны.

В качестве примера использования базы данных ENIAD приведем информацию, которую можно получить о реакции гидролиза гуанозинтрифосфата (GTP), катализируемой клеточным белком Ras в комплексе с активирующим белком GAP, что может позволить оценить каталитиче-

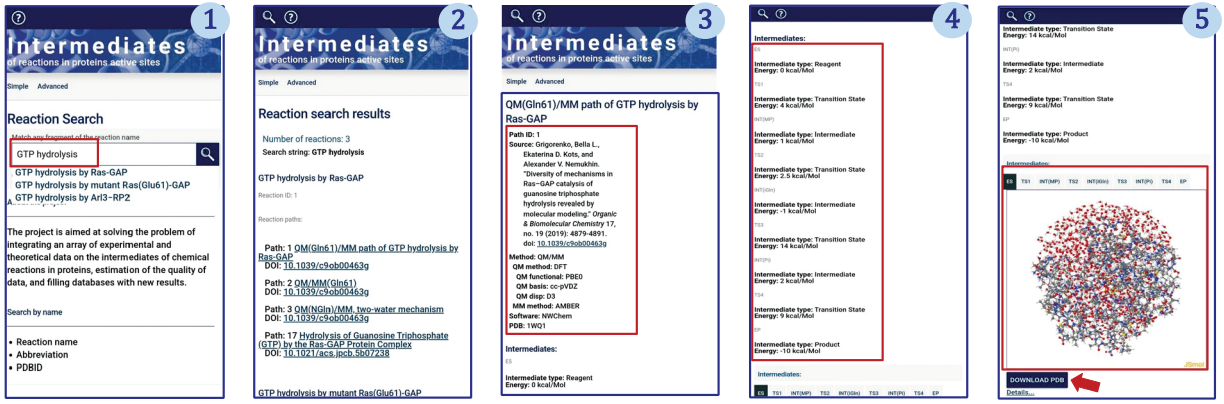


Рис 1. Последовательность действий при работе с базой данных ENIAD. (1) Поиск по ключевым словам и выбор реакции из предложенных вариантов. (2) Список полученных “путей реакции”, относящихся к данной “реакции”. (3) Метаданные “пути реакции”. (4) Информация о наборе “состояний” в выбранном “пути реакции”. (5) Данные о конкретном “состоянии”, включая возможность визуализации трехмерной структуры и ее скачивания на локальное компьютерное устройство.

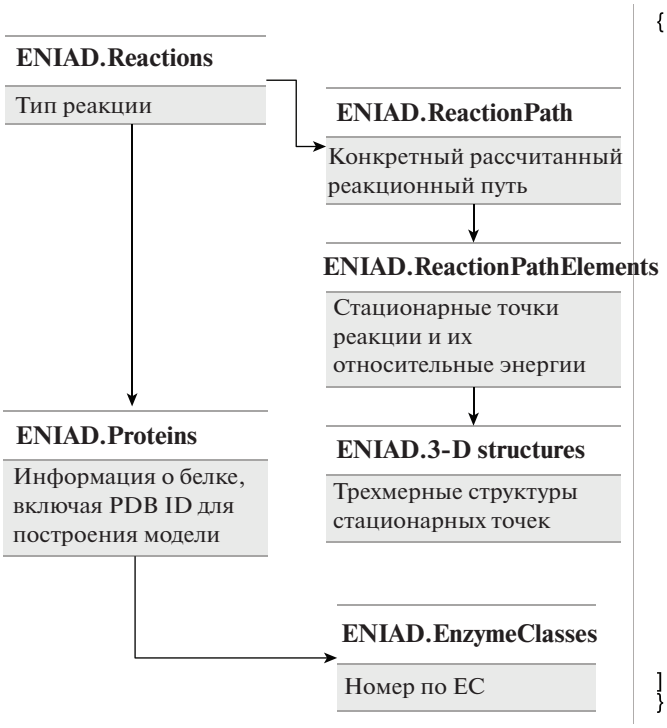
скую роль определенных аминокислотных остатков. Последовательность действий для получения информации представлена на рис. 1 и обсуждается далее. На первом этапе пользователю предлагается ознакомиться со всеми возможными “путями реакций”, удовлетворяющими желаемому запросу. Для этого можно воспользоваться поисковой системой базы данных ENIAD, используя, например, название реакции – GTP hydrolysis. В процессе набора поискового запроса система живого поиска предложит возможные варианты релевантного поискового запроса [9–11] (см. рис. 1 (1)). По результатам поиска по базе данных получаем список “путей реакции”, отвечающих критериям поиска (см. рис. 1 (2)). Здесь могут быть представлены различные механизмы, формы фермента с аминокислотными заменами, способы описания квантовой подсистемы, а также реакции для ферментов из разных организмов. Выбрав результат поиска, можно ознакомиться с источником информации, условиями расчета, т.е. методами описания квантовой и молекулярно-механической подсистем, программным обеспечением, с помощью которого были получены результаты (см. рис. 1 (3)). Как упоминалось ранее в базе данных ENIAD основным уровнем организации данных является “состояние”, поэтому для конкретного выбранного “пути реакции” будут отображены структуры стационарных точек на энергетической поверхности и соответствующие им энергии (см. рис. 1 (4)). Трехмерную структуру выбранного “состояния” можно визуализировать прямо в веб-интерфейсе базы данных (https://lcc.chem.msu.ru/eniad/) с использованием 3D Java-просмотрщика химических структур JSmol или скачать в формате PDB (см. рис. 1 (5)). Для рассмотренного примера в базе данных содержится 6 “путей реакции” и 22 соответствующие им трехмерные структуры.

ТЕХНИЧЕСКИЕ АСПЕКТЫ РЕАЛИЗАЦИИ И РАЗВЕРТЫВАНИЯ СЕРВИСОВ

Система ENIAD реализована как набор независимых подсистем: веб-интерфейс, база данных и файловое/объектное хранилище. Каждый из сервисов запущен в своем виртуальном окружении (Docker-контейнере [12]) и может быть легко перезапущен на доступных вычислительных ресурсах. Мы используем систему Kubernetes [13] для управления комплексом сервисов. Возможность развертывания системы при помощи Kubernetes, а также с использованием географически распределенных ресурсов, была продемонстрирована в работе [14].

Мы предполагаем, что база данных может в будущем содержать до тысяч или десятков тысяч записей о “путях реакции” и структурах интермедиатов. Такой объем данных можно накопить при использовании полуавтоматической загрузки данных при проведении расчетов или же в результате работы программного обеспечения системы извлечения данных из массива научных статей. Оба варианта развития системы представляются технически возможными. С учетом этих обстоятельств мы решили использовать реляционную базу данных для хранения структурированной информации.

База данных создана при помощи системы PostgreSQL [15]. PostgreSQL выбран по соображениям высокой производительности этой системы и возможностей создания специализированных индексов. Схема базы данных приведена на рис. 2 (слева).




```
{
  "Description": {
    "text": "Path is a result of QM/MM simulation with Gln16 in
             QM part",
    "doi": "10.1039/c9ob00463g"
  },
  "Reaction": "GTP hydrolysis by Ras-GAP",
  "PathElements": [
    {
      "Tag": "ES",
      "energy": 0.0,
      "eunits": "kcal/Mol",
      "type": "isReagent",
      "3D_structure": {
        "Description": "ES for 10.1039/c9ob00463g paper",
        "URI": "https://lcc.chem.msu.ru/eniad/file1.pdb",
        "fromPBD": "c9ob000463g3.pdb"
      }
    },
    {
      "Tag": "TS1",
      "energy": 4.0,
      "eunits": "kcal/Mol",
      "type": "isTransitionState",
      "3D_structure": {
        "Description": "First TS for 10.1039/c9ob00463g paper",
        "URI": " https://lcc.chem.msu.ru/eniad/file2.pdb",
        "fromPBD": "c9ob000463g7.pdb"
      }
    },
                                                    ...
]
}
```


Рис. 2. Слева: схема базы данных. Справа: пример поля PathElements с записанными данными о “состояниях” “пути реакции”.

В базе данных в основном используются поля с типом JavaScript Object Notation (JSON) для обеспечения гибкости и расширяемости системы. Только самые необходимые данные хранятся в полях других типов.

Нами была разработана схема JSON для описания данных о путях химических реакций. Корневой элемент включает три поля: “Reaction” (текстовое поле), “Description” (поле типа JSON) и “PathElements”; последнее представляет собой перечисление состояний. Каждое из состояний имеет атрибуты: энергия, тип (реагенты, переходное состояние, интермедиат или продукт), метку с названием стационарной точки и трехмерную полноатомную структуру. Пример такого файла представлен на рис. 2 (справа), полная версия доступна по ссылке https://lcc.chem.msu.ru/eniad/reaction_path_example.json. Также нами разработана схема верификации для предложенного формата, она доступна по ссылке https://lcc.chem.msu.ru/eniad/schema.json.

Веб-интерфейс разработан как простое PHP-приложение. Текущая реализация напрямую обращается к базе данных через SQL-запросы. Страницы динамически генерируются на основе шаблонов и файлов-словарей для обеспечения поддержки интерфейсом нескольких языков. Этот же механизм используется для поддержки различных пользовательских устройств (стационарные компьютеры, ноутбуки, планшетные ПК, смартфоны).

Главная страница веб-интерфейса – страница поиска по ключевым словам, названию фермента или идентификаторам реакций. Страница результатов поиска предоставляет возможность просмотреть список найденных “реакций” и относящимся к ним “путей реакций”, а также ссылки на страницы для просмотра данных “путей реакции”. При поиске по пустому запросу система отображает в результатах поиска все присутствующие в базе данных “пути реакции”.

Пользователи с соответствующими правами доступа имеют возможность загружать и редактировать данные о “путях реакций”. Аутентификация пользователя реализована при помощи механизмов OAuth, что позволяет пользователю выбирать провайдера этого сервиса, например, использовать учетную запись Google.

Данные большого объема, такие как трехмерные полноатомные структуры, хранятся во внешнем сервисе хранения. Ссылки на соответствующие файлы присутствуют в базе данных в виде записей в формате универсальных идентификаторов ресурса (URI). Таким образом, можно легко задействовать облачные услуги хранения внешних провайдеров для обеспечения дополнительного уровня надежности хранения данных. Поля fileURI

в таблице “3-D Structures” содержат именно такие ссылки на внешние сервисы хранения.

Описание “пути реакции”, которое преобразуется к формату JSON, обеспечивает гибкость в представлении информации за счет хранения данных в полуструктурированном виде. Мы планируем использовать эту возможность для хранения ссылок на источники экспериментальных данных о белках и кинетике реакций с их участием.

ЗАКЛЮЧЕНИЕ

Нами разработана база данных, которая облегчает доступ к деталям химических реакций ферментативного катализа широкому кругу специалистов. Важная особенность новой базы данных – предоставление доступа к трехмерным структурам, соответствующих реагентам, продуктам, интермедиатам и переходным состояниям вдоль пути реакции. Накопление подобных данных позволит в будущем вывести исследования эволюции и филогении ферментов на новый уровень, подобно переходу от традиционного схематического представления механизмов реакций к подходам более высокого уровня, используемым в настоящее время в хемоинформатике [16–18].

База данных ENIAD также способствует решению важной проблемы современной вычислительной химии, связанной с воспроизводимостью результатов дорогостоящих сложных компьютерных расчетов. Предоставление доступа к детальным данным подобных расчетов является важной мерой как для верификации результатов компьютерного моделирования, так и для стимулирования дальнейших исследований. Последнее может быть обеспечено включением в базу данных файлов, содержащих входную информацию, а также полной спецификации программного обеспечения. Одним из направлений дальнейшего развития ENIAD является решение этих проблем путем доступа к контейнерам, включающим необходимые программы (например, используя ссылки на репозиторий образов Docker) наряду с файлами входной и выходной информации.

Работа выполнена при финансовой поддержке Российского научного фонда (проект 19-73-20032) с использованием оборудования Центра коллективного пользования сверхвысокопроизводительными вычислительными ресурсами МГУ им. М.В. Ломоносова.

СПИСОК ЛИТЕРАТУРЫ

1. *Варфоломеев С.Д.* Химическая энзимология. М.: Научный мир, 2019. С. 543.
2. *Berman H.M., Henrick K., Nakamura H.* // Nature Structural Biology. 2003. V. 10. № 12. P. 980. https://doi.org/10.1038/nsb1203-980
3. *Holliday G.L., Andreini C., Fischer J.D. et al.* // Nucleic Acids Res. 2012. V. 40. P. D783. https://doi.org/10.1093/nar/gkr799
4. *Nagano N., Nakayama N., Ikeda K. et al.* // Ibid. 2015. V. 43. P. D453. https://doi.org/10.1093/nar/gku946
5. *Ribeiro A.J.M., Holliday J.L., Furnham N. et al.* // Ibid. 2018. V. 46. P. D618. https://doi.org/10.1093/nar/gkx1012
6. *Furnham N., Holliday G.L., de Beer T.A.P. et al.* // Ibid. 2014. V. 42. P. D485. https://doi.org/10.1093/nar/gkt1243
7. *Warshel A., Levitt M.* // J. Mol. Biol. 1976. V. 103. P. 227. https://doi.org/10.1016/0022-2836(76)90311-9
8. *Senn H.M., Thiel W.* // Angew. Chemie Int. Ed. 2009. V. 48. P. 1198. https://doi.org/10.1002/anie.200802019
9. *Grigorenko B.L., Kots E.D., Nemukhin A.V.* // Org. Biomol. Chem. 2019. V. 17. P. 4879. https://doi.org/10.1039/C9OB00463G
10. *Khrenova M.G., Grigorenko B.L., Kolomeisky A.B. et al.* // J. Phys. Chem. B. 2015. V. 119. № 40. P. 12838. https://doi.org/10.1021/acs.jpcb.5b07238
11. *Khrenova M.G., Kots E.D., Nemukhin A.V.* // Ibid. 2016. V. 120. № 16. P. 3873. https://doi.org/10.1021/acs.jpcb.6b03363
12. Docker, Inc. https://www.docker.com, 2019.
13. The Linux Foundation. https://kubernetes.io, 2019.
14. *Brekhov A.T., Mironov V.A., Moskovsky A.A. et al.* // J. Phys.: Conf. Ser. 2019. V. 1392. P. 012049. https://doi.org/10.1088/1742-6596/1392/1/012049
15. PostgreSQL Global Development Group. https://www.postgresql.org, 2019.
16. *Latino D.A.R.S., Aires-de-Sousa J.* // Chemoinf. and Comput. Chem. Biol. 2011. V. 672. P. 325. https://doi.org/10.1007/978-1-60761-839-3_13
17. *O’Boyle N.M., Holliday G.L., Almonacid D.E. et al.* // J. Mol. Biol. 2007. V. 368. P. 1484. https://doi.org/10.1016/j.jmb.2007.02.065
18. *Almonacid D.E., Babbitt P.C.* // Curr. Opin. Chem. Biol. 2011. V. 15. P. 435. https://doi.org/10.1016/j.cbpa.2011.03.008