ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ т. 11, № 2(45), с. 75–92

ББК Г511.4:В192.252
ГРНТИ 31.15.15, 50.07.05
УДК 544.18:004.67

Н. А. Аникин, А. Ю. Мускатин, М. Б. Кузьминский,
С. Н. Леднев, А. В. Смирнов, А. И. Русаков

Кардинальное ускорение расчетов гигантских биомолекул методами квантовой химии, требующими применения суперЭВМ и/или GRID-систем

Аннотация. Расчеты электронной структуры молекул квантовохимическими методами давно проводятся с использованием суперЭВМ. Сегодня они проводятся на лидере суперкомпьютерного списка TOP500 и будут осуществляться на первом в США экзафлопсном суперкомпьютере.

Краткий обзор современных методов квантовой химии и их применения на суперЭВМ для расчетов в первую очередь больших молекул показывает необходимость применения ускоренных аппроксимационных методик для реализации возможностей проведения таких расчетов. Это особенно актуально для массовых расчетов таких гигантских биомолекул, как докинг-комплексы белок-лиганд.

Для этого нами разработаны дающие большое ускорение при приемлемой точности расчетов алгоритмы аппроксимации для вычисления молекулярных интегралов неэмпирических методов квантовой химии. Для массовых расчетов докинг-комплексов полуэмпирическими методами предложена и программно реализована новая методика, базирующаяся на использовании некоторых локализаций взаимодействий лигандов с белком благодаря формированию групп из полного набора лигандов комплекса.

Изложенная методика позволила достигнуть ускорения на порядки и предполагается к использованию в будущих неэмпирических расчетах. Описанные методики и программы для необходимых массовых расчетов докинг-комплексов естественно вписываются в пакетную систему обработки заданий и могут использоваться в GRID-среде. Такая GRID-система создается на вычислительных ресурсах ЯрГУ и ИОХ РАН на базе стандартных в рамках EGI программных средств UMD 4).

Ключевые слова и фразы: быстродействующие квантовохимические методы, докинг-комплексы, GRID.

Работа выполнена при финансовой поддержке РФФИ, проект №18-07-00657.

Введение

Методы квантовой химии, применяемые для расчета молекул, являются традиционной областью использования суперкомпьютеров (см., например, [1]). Квантовохимические расчеты осуществляются на первом в списке TOP500 (на ноябрь 2019) суперкомпьютере Summit [2],[3]. Знаменитый планируемый первым в США экзафлопс-суперкомпьютер Аргоннской национальной лаборатории [4],[5] будет использоваться для решения задач квантовой химии [6].

В неэмпирических и полуэмпирических квантовохимических методах для решения исходных интегродифференциальных уравнений используется аппроксимация волновых функций с применением базисных наборов функций. Далее мы имеем в виду только традиционные для молекул, экспоненциально убывающие с расстоянием от ядер атомов базисные функции — атомные орбитали, (АО) общего функционального вида в сферической системе координат

$$\phi(r,\theta,\varphi) = Y_{lm}(\theta,\varphi) r^l \mathcal{R}(r), \tag{1}$$

где для неэмпирических расчетов применяется линейная комбинация гауссовых экспонент

$$\mathcal{R}(r) = \sum_k \beta_k e^{-\alpha_k r^2}. \tag{2}$$

С ростом числа N используемых АО растет точность расчета, однако сильно возрастает и его ресурсоемкость, особенно время расчета. После расчета необходимых базирующихся на АО молекулярных интегралов, число которых есть $O(N^4)$, решение сводится обычно к некоторым задачам линейной алгебры. В разных типах этих интегралов используется до четырех центрированных на разных атомах АО, причем для больших молекул более всего лимитируют время расчета именно четырехцентровые интегралы.

Основные применяемые сегодня методы квантовой химии требуют для расчета электронного строения, включающего полную энергию E молекулярной системы, процессорного времени от $O(N^3)$ для традиционных быстрых и менее точных полуэмпирических методов до $O(N!)$ для полного метода конфигурационного взаимодействия. Широко используемые методы теории функционала плотности (DFT) требуют $O(N^4)$. Число N линейно пропорциональна числу атомов, и поэтому суперЭВМ требуются для расчета E либо из-за необходимости

использования более точного и ресурсоемкого метода расчета, либо из-за необходимости расчета больших молекул с большим числом атомов и N. Отсюда проблема выбора наиболее быстродействующего метода, достигающего разумной точности.

Часто используемые полуэмпирические методы и методы ХФ/DFT традиционно требуют нахождения собственных значений и векторов получаемой матрицы одноэлектронного гамильтониана (фокиана F) размерности $N \times N$:

$$FC = \varepsilon SC, \tag{3}$$

где S — матрица интегралов перекрывания, а выше речь шла об интегралах, необходимых для расчета матрицы F. Решение этого уравнения требует $O(N^3)$ процессорного времени, дает молекулярные орбитали (МО) как линейную комбинацию АО с коэффициентами из матрицы С, и выполняется итерационно. Для очень больших молекул возможно отбрасывать малые интегралы (их доля существенно возрастает с ростом расстояний между удаленными «концами» молекулы), что дает уменьшение времени расчета нужных интегралов до $O(N^2)$.

Для расчетов наиболее быстрыми полуэмпирическими методами и DFT таких больших молекул появились ускоренные вычислительные методики с отказом от традиционной диагонализации F из уравнения (3) и с линейным $O(N)$ масштабированием времени расчета для гигантских молекул (более тысячи атомов), но с очень большим коэффициентом перед $O(N)$. Для гигантских молекул из многих тысяч атомов для достижения линейного масштабирования необходимо еще применять так называемый метод быстрых мультиполей FMM [7].

Полуэмпирические методы гораздо быстрее (но и менее точны), чем DFT, т.к. не требуют расчета такого большого числа интегралов. Имеется сообщение о проведении полуэмпирического расчета молекулы из 100 тысяч атомов с применением гибридного распараллеливания OpenMP+MPI на 1024 процессорах [8]. Использование суперЭВМ Ломоносов сделало возможным квантовые расчеты полных докинг-комплексов в РФ за разумное время с применением только полуэмпирических методов [9].

Более точные неэмпирические расчеты, в том числе по наиболее быстрому и широко используемому DFT, могут быть в 1000 раз медленнее, чем полуэмпирические [10]. В [11] приведены результаты расчетов методом DFT, адаптированным в программном комплексе

ONETEP на $O(N)$ — расчеты больших молекул. Эти расчеты производились на входившем в TOP500 суперкомпьютере Iridis 4 на базе двухпроцессорных узлов с восьмиядерными Intel Xeon E5-2670/2.6 ГГц и межсоединения Infiniband FDR с использованием гибридного OpenMP + MPI распараллеливания.

Расчет такого вытянутого фибриллярного белка (реальные глобулярные белки будут считаться значительно медленнее) из 13696 атомов занял 7,3 час. При этом DFT-расчет белка на суперЭВМ Cray XC30 [12] проводился вообще не с применением гауссовых АО (соответствующих формулам (1) и (2)), а в базисе плоских волн, используемом обычно для расчетов твердых тел, а не молекул.

Все сказанное относится к так называемым одноточечным расчетам с фиксированной геометрией расположения атомов молекулы, но в ряде случаев требуется еще оптимизация геометрии (декартовых или внутренних координат) с поиском минимума E, что существенно увеличивает время расчета из-за необходимости одноточечных расчетов при разных геометриях, особенно для белков с их тысячами атомов (и соответствующих степеней свободы оптимизации геометрии).

Если же нужен расчет не только E, а и других термодинамических величин, то нужны другие классические для применения суперкомпьютеров методы вычислительной химии — молекулярной динамики, которые и на суперкомпьютерах сейчас основываются обычно на простейших неквантовых расчетах E, а на вышеуказанных квантовохимических уровнях в ближайшем будущем нереальны.

Поскольку полная оптимизации геометрии требует на порядки более высоких времен расчета, когда речь заходит о возможностях таких вычислений по DFT для нескольких сотен (менее тысячи атомов) с прицелом на расчет фрагментов белков [13]. В этой работе оптимизация геометрии по программе Jaguar проводилась для молекулярных систем, содержащих до 600 атомов (до 10 тысяч базисных функций).

Расчеты, проведенные на известной Техасской суперЭВМ Stampede на базе связанных Infiniband FDR-межсоединением двухпроцессорных узлов с восьмиядерными Intel Xeon E5-2680/2,7 ГГц с использованием гибридного распараллеливания OpenMP + MPI, на 256 ядрах требуют около 76 минут, при этом достигается ускорение примерно в 114 раз. На рисунке 1 представлена зависимость времени расчета от числа используемых процессорных ядер, показывающая существенное снижение эффективности распараллеливания на большем их числе.

В некоторых ситуациях возможно проведение квантовохимического расчета не большой молекулы целиком, а только ее реально актуального

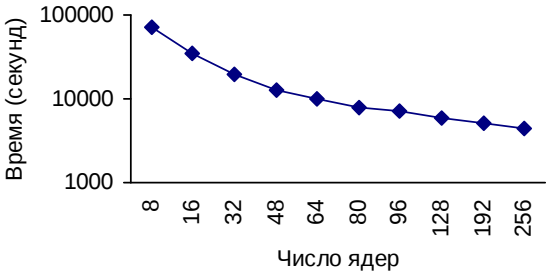


Рисунок 1. Зависимость времени расчета от числа использованных процессорных ядер [13]

для исследования фрагмента. Информация про подобные подходы, в том числе для актуальных докинг-комплексов содержится в [14]. Известной квантовохимической реализацией такого подхода является метод фрагментных молекулярных орбиталей FMO.

Такие расчеты выполняются и на суперкомпьютерах. Например, в расчете [15] на суперЭВМ Ломоносов-2 использовался вариант FMO-TDDFT. Однако мы такое направление в статье не рассматриваем.

Некоторые квантовохимические программные комплексы специально ориентируются на эффективные расчеты больших молекул на суперкомпьютерах, например, NTChem дал возможность считать по DFT молекулы с сотнями атомов [16] на японском K-компьютере, первым в мире достигшем производительности 10 PFLOPS. Едва ли не самыми востребованными сегодня являются расчеты огромных органических молекул, докинг-комплексов белков (протеинов, содержащих обычно больше 1000 атомов, но бывает немало и гигантских белков, содержащих порядка 100 тысяч и более атомов) с лигандами, что актуально, в частности, для задач конструирования лекарств.

При этом в каждом исследовании необходимы расчеты тысяч докинг-комплексов. Даже сейчас на фоне активного развития быстрых полуэмпирических методы квантовой химии и DFT для таких больших молекул некоторые авторы считают квантовохимические расчеты таких молекулярных структур целиком невыполнимыми в течение разумного времени [14].

Нами предложены новые аппроксимационные усовершенствования методов квантовой химии и разработаны соответствующие программные средства на Fortran-95, которые ориентированы исключительно на очень большие молекулы, главным образом массовые расчеты актуальных докинг-комплексов.

1. Кардинальное ускорение расчета интегралов от АО для неэмпирических расчетов больших молекул

Один из современных способов ускорения расчета кулоновских двухэлектронных интегралов больших молекулах (одна из лимитирующих стадий) основан на аппроксимации произведения пар базисных функций АО от координат одного электрона в виде линейной комбинации гауссовых функций, называемых вспомогательными функциями плотности (сокращённо ВФП)

$$\phi_i(\vec{r_1})\phi_j(\vec{r_1}) = f_n(\vec{r_1}) \approx g_n(\vec{r_1}) = \sum_p B_p \chi_p(\vec{r_1}), \tag{4}$$

где ϕ — базисные (атомные) функции (орбитали, АО) отдельных атомов, а χ_p — ВФП. Для двухэлектронных интегралов используются укороченные обозначения:

$$(f_n|f_m) = \iint \frac{\phi_i(\vec{r_1})\phi_j(\vec{r_1})\phi_k(\vec{r_2})\phi_l(\vec{r_2})}{r_{12}} dV_1 dV_2 \tag{5}$$

При этом различных ВФП для хорошей точности аппроксимации надо брать примерно в два раза больше, чем АО, что во много раз меньше квадрата числа произведений пар АО.

Для повышения точности используется прием сведения всех четырехцентровых интегралов от АО (требует $O(N^4)$ вычислений) к трехцентровым интегралам от пар АО, которых $O(N^3)$ с точностью до малых второго порядка по отклонению f от g:

$$\begin{aligned}(f_n|f_m) &= (g_n + (f_n - g_n)|g_m + (f_m - g_m)) = \\ &= (g_n|g_m) + (f_n - g_n|g_m) + (g_n|f_m - g_m) + (f_n - g_n|f_m - g_m) \approx \\ &\approx (f_n|g_m) + (g_n|f_m) - (g_n|g_m)\end{aligned}$$

Здесь g_m — линейная комбинация ВФП χ_p, а $(f_n|g_m)$ — трехцентровые интегралы.

Для ускорения расчета двухэлектронных трехцентровых кулоновских интегралов $(f_n|g_m)$, которые замедляют расчеты больших молекул, мы предлагаем следующее:

- Каждая функция χ для исходного базиса с высокой точностью аппроксимируется комбинацией наших универсальных базовых гауссовых орбиталей (БГО) через аппроксимацию каждой входящей в χ гауссовой экспонентом через комбинацию четырех БГО с показателями экспонент, близких к показателю гауссовой экспоненты из аппроксимируемой функции

$$e^{-br^2} \approx \sum_{n=1}^{4} c_n e^{-a_n r^2}. \tag{6}$$

- Показатели экспонент БГО образуют специально подобранную геометрическую прогрессию, достаточную для хорошей аппроксимации всех актуальных ВФП χ. Для повышения точности этой аппроксимации необходим набор показателей экспонент БГО, близких друг к другу, но при этом стандартная 64-разрядная точность становится недостаточной, что на порядки повышает погрешности округления. Мы преодолели это использованием вместо отдельных БГО наборов их линейных комбинаций с коэффициентами из матрицы $S^{-1/2}$, где S — близкая к вырожденности матрица 4×4 интегралов перекрывания между БГО с разными, но близкими показателями экспонент.

  Все это используется для последующего быстрого вычисления всех интегралов в виде сплайнов от R. Интегралы для любых разнообразных базисов быстро рассчитываются в виде линейный комбинации заранее затабулированных сплайнов от R для интегралов от универсальных БГО из нашей создаваемой БД.
- При расчете трехцентровых интегралов двухцентровое произведение базисных функций аппроксимируется более точно везде, в том числе между соответствующими центрами ϕ. К ВФП на двух центрах добавляется еще необходимый минимум ВФП, центрированных между этими центрами. Все это позволяет с высокой точностью свести все трехцентровые интегралы (и необходимые для расчета вкладов в F их линейные комбинации) к линейным комбинациям новых двухцентровых интегралов от всех ВФП, быстро вычисляемых для любых исходных базисов в виде сплайнов от R.

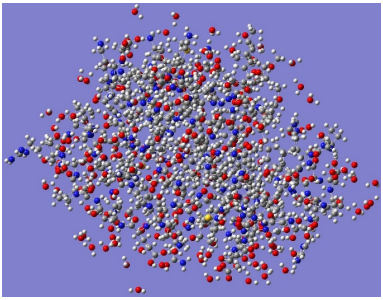
Проведенная нами систематическая сплайн-аппроксимация интегралов перекрывания и кинетической энергии в зависимости от межъядерного расстояния для всех пар БГО с отобранными нами для аппроксимаций показателями экспонент показала вполне достаточный уровень точности расчета: точность интегралов порядка от 10^{-10} до 10^{-8} (чаще всего порядка 10^{-9}, и лишь редко доходит до $4 \cdot 10^{-8}$). Этого достаточно для вполне приемлемой точности квантовохимических расчетов при аппроксимации интегралов перекрывания и кинетической энергии от всех пар АО для широко распространенных расширенных базисов 6-31G* и 6-31++G** для всех атомов 1-3 периодов, образующих ковалентные связи.

Апробация проведена на расчете порядка миллиона интегралов перекрывания и кинетической энергии (исключая пренебрежимо малые, что быстро оценивается еще до этого расчета) для молекулы белка IMMUNOPHILIN FKBP-12 из 1663 атомов.

Расчет всех интегралов перекрывания и кинетической энергии белка FKBP-12 в базисе 6-31G* ускорен более чем в 20 раз. Сейчас нами разрабатывается более сложная аппроксимация для сверхбыстрого вычисления вклада кулоновских трех- и четырехцентровых интегралов в фокиан F — одну из лимитирующих стадий неэмпирческих (в том числе DFT) расчетов гигантских молекул. Метод открывает новые перспективы ускорения вычислений других типов молекулярных интегралов в произвольных центрированных на атомах базисах.

2. Методика ускорения массовых расчетов докинг-комплексов и ее полуэмпирическая реализация

На рисунке 2 приведено изображение всех приблизительно 1700 атомов рассчитанного нами относительно небольшого белка — комплекса человеческого иммуноглобулина с иммунным супрессором IMMUNOPHILIN FKBP-12 из Protein Data Bank [**17**].



- $\sim$ 14000 атомных орбиталей,
- $\sim$ 30000 вспомогательных функций плотности,
- $\sim 2 \times 10^{14}$ непренебрежимых 4-центровых интегралов,
- $\sim 2 \times 10^{11}$ непренебрежимых 3-центровых интегралов.

Рисунок 2. Типичный относительно небольшой протеин ($\sim$ 1700 атомов)

Выше уже отмечено, что многие другие белки еще крупнее. Но даже этот белок значительно крупнее большинства обычных органических молекул. Для наглядности мы привели на рисунке ориентировочное число вышеуказанных двухэлектронных кулоновских интегралов для использованного нами и очень часто применяемого в DFT базиса 6-31G*. Для многих других белков число интегралов на порядки больше. Поэтому для весьма актуальных массовых (от

тысяч и более) расчетов докинг-комплексов белок-лиганд (типичные лиганды содержат от 30 до 60 атомов, располагающиеся, как правило, в полости белка) крайне актуально применение специальных ускоренных методик в выполняемых на компьютерных системах (в первую очередь суперЭВМ) квантовохимических программных комплексах.

Мы разработали новый метод для кардинального ускорения таких массовых квантовохимических расчетов с учетом всех квантовых эффектов всего докинг-комплекса. Он основан на созданном нами базовом методе для ускорения массовых полуэмпирических квантовохимических расчетов докинг-комплексов (тысяч комплексов из тысяч атомов). С использованием геометрии докинг-комплексов из БД PDBbind [18] время расчетов по сравнению обычными расчетами со стандартной диагонализацией матриц сократилось, как правило более чем в 300 раз [19, 20].

Этот наш базовый метод ускоренных расчетов докинг-комплексов основан на выделении активной части атомов белка, близких к каждому атому каждого лиганда, а соответствующая часть электронной волновой функции и матрицы плотности рассчитывается как единое целое вместе с волновой функцией лигандов. Расчет проводится после однократного квантовохимического расчета белка.

Вклады остальной (замороженной, «frozen» на рисунке 3) части белка (от его остальных атомов) явно учитываются в виде фиксированного вклада в фокиан F. Это дает полный квантовомеханический расчет при явном учете только АО активной части белка и лиганда, что сильно ускоряет расчет за счет сильного уменьшения N.

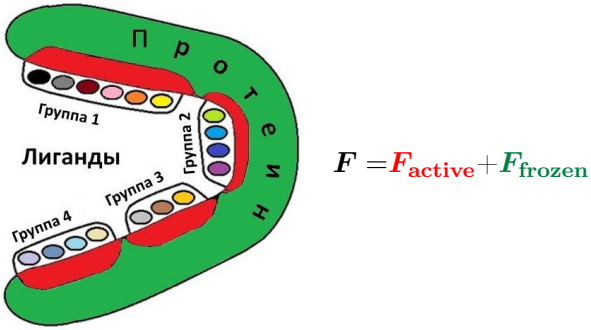


Рисунок 3. Разбиение всех лигандов на группы

Такой упрощенный подход, как было показано нами в [19], не дает сколь-либо существенной ошибки в точности расчета энергии взаимо-

действия ΔE лиганда с белком (численная иллюстрация приведена ниже). Отметим, что в часто используемом сильно упрощенном подходе метода QM/MM влияние почти всех даже немного удаленных от лиганда атомов белка учитывается всего лишь простой молекулярной механикой, что может давать большую ошибку в ΔE.

По сравнению с эффективным линейно масштабируемым (заменяющим диагонализацию матриц) методом CG-DMS время расчета уменьшилось от 5 до 30 раз, по сравнению с другим линейно масштабируемым методом DivCon — от 18 до 120; выигрыш в быстродействии увеличивается с ростом размера белка.

При этом отличие наших рассчитанных ΔE от результатов стандартного расчета всего докинг-комплекса с диагонализацией полной матрицы F составляет лишь около 0,4 ккал/моль (коэффициент линейной корреляции равен 0,997), т.е. все рассчитанные значения ΔE весьма близки к результатам стандартной диагонализации полного фокиана.

В этом нашем базовом вышеописанном методе активная часть формировалась как универсальная (единая) для всех разнообразных лигандов из полного набора, содержащего типично более тысячи лигандов. Из-за этого для лигандов, близких к атомам из, например, «левой» части полости без нужды учитывался вклад многих атомов из других частей полости белка, удаленных от данного лиганда, т.е. N было завышенным.

Индивидуальное выделение активной (и соответственно замороженной) части белка для каждого лиганда приводило бы к чрезмерному росту расхода времени расчета на стадии расчета вклада замороженной части белка в F. Поэтому мы усовершенствовали этот базовый метод разбиением всего набора лигандов на группы лигандов [**21**], где каждая группа лигандов и соответствующая ей активная часть белка локализованы в своей отдельной части полости белка, что уменьшает размер необходимой его активной части.

Это ускоряет расчет, поскольку у каждой группы лигандов теперь имеется своя активная часть белка меньшего размера, чем была общая для всех лигандов, что отображено на рисунке 3.

Так, для вышеупомянутого выше белка FKBP-12 набор из 1308 лигандов был разбит на 17 групп, что уменьшило время расчета в 2,4 раза на одном процессоре AMD Athlon XP 3000+/2,2 ГГц до немного более суток. При этом среднеквадратичное отличие энергий образования докинг-комплексов от величин, полученных по нашему

базовому методу без образования групп, равнялось всего лишь 0,15 ккал/моль, т.е. точность расчета не ухудшилась.

Но мы и далее усовершенствовали (по сравнению с [21]) наш алгоритм образования групп, что ускорило расчет еще в 1,65 раза до 17,5 час, а среднеквадратичное отклонение энергии от не сгруппированного расчета стало 0,08 против 0,15 ккал/моль в предыдущем варианте алгоритма, т.е. расчет стал и быстрее, и точнее.

Указанное время расчетов полуэмпирическим методом AM1 не включает оптимизацию геометрии, и было бы желательно применять более точный, но требующий на несколько порядков больше времени расчета метод DFT. Поэтому наш метод массовых расчетов докинг-комплексов и его усовершенствования далее мы планируем перенести и программно реализовать для расчетов более точным методом DFT. Там ресурсоемкий расчет вклада двухэлектроных кулоновских интегралов предполагается сильно ускорить на основе нашей описанного в предыдущем разделе статьи аппроксимационного метода.

Кроме того, для расчетов десятков и более тысяч докинг-комплексов лимитирующее общее время расчетов будет линейно пропорционально числу докинг-комплексов. Поэтому программные реализации наших вышеуказанных ускоренных методов массовых расчетов докинг-комплексов позволяют естественно использовать GRID-системы с применением пакетной обработки заданий.

Проект РФФИ 18-07-00657 создает такую GRID-систему, включающую вычислительные ресурсы небольшого гетерогенного кластера в ИОХ РАН и часто (в том числе и самой Nvidia) относимого к суперкомпьютерам сервера DGX-1 с 8 GPU V100 (с пиковой DP-производительностью 7,8 Терафопс каждый) и кластера в ЯрГУ.

Кластеры базируются на CentOS 7, а GRID-среда развивается на поддерживаемых EGI программных средствах репозитория UMD 4. GRID-сервер в ИОХ РАН базируется на промежуточном программном обеспечении ARC 15.3.19 из UMD-4.8.2, которое поддерживает все основные языки описания задач в GRID-системах. Служба ARC Resource-coupled EXecution service (A-REX) обеспечивает интерфейс с локальными системами пакетной обработки, в качестве которых у нас применяется Torque 4.2.10. Для обеспечения оптимальной работы с большими наборами данных используются программные средства dCache 3.2.21 из UMD-4.7.0. Данные результатов квантовохимических расчетов хранятся в формате CML версии 3 [22].

Распараллеливание программных комплексов, реализующих наши методические разработки, описанные выше, осуществляется с применением OpenMP, что позволяет не только эффективнее использовать аппаратные средства гетерогенного кластера в ИОХ РАН, но и задействовать в будущем все GPU V100 из DGX-1 в ЯрГУ, поскольку последний стандарт OpenMP 5.0 позволит очень существенно улучшить достигаемую производительность, в том числе за счет эффективной поддержки работы с иерархией памяти в GPU, а реальные реализации OpenMP в компиляторах дают возможность применения нескольких ускорителей на сервере.

Заключение

- Дан краткий обзор современных методов квантовой химии с ориентацией их применения для расчетов больших и гигантских (из тысяч атомов) молекул. Показано, что для актуальных массовых расчетов докинг-комплексов применение даже наиболее быстрых полуэмпирических методов и DFT требует не только применения суперЭВМ, но и разработки и программной реализации дополнительных вычислительных методов для кардинального ускорения расчетов.
- Предложен и программно реализован новый аппроксимационный метод для существенного ускорения расчетов, лимитирующих время вычисления неэмпирическими методами двухэлектронных интегралов. Его эффективность продемонстрирована на примере двухцентровых интегралов, но она планируется к реализации для расчетов трех- и четырехцентровых интегралов.
- Предложен и программно реализован специальный метод для ускорения массовых квантовохимических расчетов докинг-комплексов лигандов с белком на основе образования специальных локальных групп лигандов. Такой подход может применяться в полуэмпирических и неэмпирических методах. Он реализован и показал свою эффективность в ускорении полуэмпирических расчетов, которые в более близком будущем могут стать широко применяемыми на суперЭВМ и в GRID-системах для проведения таких расчетов.
- В этих целях на вычислительных ресурсах ИОХ РАН и ЯрГУ создается GRID-система на основе ставшего современным европейским EGI-стандартом репозитория UMD-4.

Список литературы

[1] P. Jørgensen, T. Kjaergaard, K. Kristensen, P. Baudin, P. Ettenhuber, J. J. Eriksen, Y. M. Wang, D. Bykov. "Quantum chemistry on the supercomputers of tomorrow", Smoky Mountains Computational Sciences and Engineering Conference (August 31–September 2, 2015, Gatlinburg, Tennessee, USA). ↑76

[2] D. Bykov, A. Barnes, D. Lyakh. "Quantum Chemistry on Supercomputers", Southeastern Theoretical Chemistry Association Meeting SETCA 2019 (May 16–18, 2019, The University of Tennessee, Knoxville, USA). ↑76

[3] W. Jia, L. W. Wang, L. Lin. *Parallel transport time-dependent density functional theory calculations with hybrid functional on summit*, 2019. arXiv 1905.01348 ↑76

[4] P. Gwynne. "Exascale supercomputer initiative launched", *Physics World*, **32**:5 (2019), pp. 11. ↑76

[5] P. Cheng, Y. Lu, Y. Du, Zh. Chen. "Tiered data management system: Accelerating data processing on HPC systems", *Future Generation Computer Systems*, **101** (2019), pp. 894–908. ↑76

[6] R. Johnson. *Aurora supercomputer to empower advanced chemistry research*, 2019. ↑76

[7] H. A. Le, T. Shiozaki. "Occupied-orbital fast multipole method for efficient exact exchange evaluation", *Journal of Chemical Theory and Computation*, **14**:3 (2018), pp. 1228–1234. ↑77

[8] M. Hennemann, T. Clark. "EMPIRE: a highly parallel semiempirical molecular orbital program: 1: self-consistent field calculations", *Journal of Molecular Modeling*, **20**:7 (2014), 2331. ↑77

[9] I. V. Oferkin, E. V. Katkova, A. V. Sulimov, D. C. Kutov, S. I. Sobolev, V. V. Voevodin, V. B. Sulimov. "Evaluation of docking target functions by the comprehensive investigation of protein-ligand energy minima", *Advances in Bioinformatics*, **2015** (2015), 126858. ↑77

[10] A. Li, H. S. Muddana, M. K. Gilson. "Quantum mechanical calculation of noncovalent interactions: A large-scale evaluation of PMx, DFT, and SAPT approaches", *Journal of Chemical Theory and Computation*, **10**:4 (2014), pp. 1563–1575. ↑77

[11] J. C. Womack, N. Mardirossian, M. Head-Gordon, Ch.-K. Skylaris. "Self-consistent implementation of meta-GGA functionals for the ONETEP linear-scaling electronic structure package", *The Journal of Chemical Physics*, **145**:20 (2016), 204114. ↑77

[12] E. J. Higgins, M. I. J. Probert, P. J. Hasnip, K. Refson, I. J. Bush. *Hybrid OpenMP and MPI within the CASTEP code*, 2015. ↑78

[13] J. Zhang, A. L. Weisman, P. Saitta, R. A. Friesner. "Efficient simulation of large materials clusters using the jaguar quantum chemistry program: Parallelization and wavefunction initialization", *International Journal of Quantum Chemistry*, **116**:5 (2016), pp. 357–368. ↑78,79

[14] I. Horváth, N. Jeszenői, M. Bálint, G. Paragi, C. Hetényi. "A fragmenting protocol with explicit hydration for calculation of binding enthalpies of

target-ligand complexes at a quantum mechanical level", *International Journal of Molecular Sciences*, **20**:18 (2019), pp. 4384. ↑79

[15] A. V. Nemukhin, I. V. Polyakov, A. I. Moskovsky. "Multi-scale supercomputing of large molecular aggregates: A case study of the light-harvesting photosynthetic center", *Supercomputing Frontiers and Innovations*, **2**:4 (2016), pp. 48–54. ↑79

[16] T. Nakajima, M. Katouda, M. Kamiya, Yu. Nakatsuka. "NTChem: A highperformance software package for quantum molecular simulation", *International Journal of Quantum Chemistry*, **115**:5 (2015), pp. 349–359. ↑79

[17] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne. "The Protein Data Bank", *Nucleic Acids Research*, **28**:1 (2000), pp. 235–242. ↑82

[18] R. Wang, Y. Lu, X. Fang, S. Wang. "An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes", *J. Chem. Inf. Comput. Sci.*, **44**:6 (2004), pp. 2114–2125. ↑83

[19] Н А. Аникин, А. С. Мендкович, М. Б. Кузьминский, А. М. Андреев. «Метод и программа для массовых квантовохимических расчетов докинг-комплексов протеин-лиганд», *Известия Академии наук. Сер. химическая*, 2008, №2, с. 418–420. ↑83

[20] Н. А. Аникин, А. М. Андреев, М. Б. Кузьминский, А. С. Мендкович. «Быстродействующий метод для массовых полуэмпирических расчетов докинг-комплексов», *Известия Академии наук. Сер. химическая*, 2008, №9, с. 1759–1764. ↑83

[21] Н. А. Аникин, А. М. Андреев, М. Б. Кузьминский, А. С. Мендкович. «Новый подход к ускорению массовых квантово-химических расчетов докинг-комплексов», *Известия Академии наук. Сер. химическая*, 2018, №6, с. 1100–1103. ↑84,85

[22] Н. А. Аникин, А. Ю. Мускатин, М. Б. Кузьминский, А. И. Русаков. «GRID-система на основе европейских стандартов EGI для крупномасштабных расчетов по оригинальному ускоренному методу квантовой химии», *Моделирование и анализ информационных систем*, **26**:3 (2019), с. 359–363. ↑85

Поступила в редакцию 17.12.2019
Переработана 16.03.2020
Опубликована 13.06.2020

Рекомендовал к публикации *д.т.н. А. М. Елизаров*

Пример ссылки на эту публикацию:

Н. А. Аникин, А. Ю. Мускатин, М. Б. Кузьминский, С. Н. Леднев, А. В. Смирнов, А. И. Русаков. «Кардинальное ускорение расчетов гигантских биомолекул методами квантовой химии, требующими применения суперЭВМ и/или GRID-систем». *Программные системы: теория и приложения*, 2020, **11**:2(45), с. 75–92. 10.25209/2079-3316-2020-11-2-75-92

http://psta.psiras.ru/read/psta2020_2_75-92.pdf

Об авторах:

Николай Алексеевич Аникин
научный сотрудник лаборатории компьютерного обеспечения химических исследований, кандидат химических наук ИОХ РАН

0000-0002-5724-8969
e-mail: nikan@swf.chem.ac.ru

Александр Юрьевич Мускатин
ведущий инженер лаборатории компьютерного обеспечения химических исследований ИОХ РАН

0000-0002-3596-2782
e-mail: amus74@mail.ru

Михаил Борисович Кузьминский
старший научный сотрудник лаборатории компьютерного обеспечения химических исследований, кандидат химических наук ИОХ РАН

0000-0002-3944-8203
e-mail: kus@free.net

Сергей Николаевич Леднев
старший преподаватель кафедры общей и физической химии ЯрГУ, кандидат химических наук

0000-0002-1041-6603
e-mail: silverpoint07@gmail.com

Александр Валерьевич Смирнов
доцент кафедры теоретической информатики ЯрГУ, кандидат физико-математических наук

0000-0002-0980-2507
e-mail: alexander_sm@mail.ru

Александр Ильич Русаков
ректор ЯрГУ, доктор химических наук, профессор

0000-0001-8893-4577
e-mail: alex@yars.free.net

CSCSTI 31.15.15, 50.07.05
UDC 544.18:004.67

Nikolay A. Anikin, Aleksandr Y. Muskatin, Mikhail B. Kuz'minskiy, Sergey N. Lednev, Aleksandr V. Smirnov, Aleksandr I. Rusakov. *Cardinal acceleration of calculations of giant biomolecules by quantum chemistry methods, requiring the use of supercomputers and / or GRID systems.*

Abstract. Calculations of the electronic structure of molecules by quantum chemical methods long ago are performed using supercomputers. Today they are conducted on the leader of the TOP500 supercomputer list and will be realized on the first exaflops supercomputer in the USA.

A brief review of modern quantum chemistry methods and their supercomputer application for calculations of primarily large molecules shows the need for accelerated approximation techniques to realize the possibilities of such computations. The need is especially urgent for massive calculations of such giant biomolecules as protein-ligand docking complexes. For this, we have proposed and implemented software that gives excellent acceleration at an acceptable accuracy of approximation calculations for calculating the molecular integrals of non-empirical methods of quantum chemistry. For massive calculations of docking complexes using semiempirical methods, we propose and implement a new technique in software. It uses some localizations of ligand-protein interactions due to the formation of groups from a full set of ligands.

This method allows us to achieve acceleration by orders of magnitude and also intended for use in future non-empirical calculations. The proposed methods and programs for the necessary massive calculations of docking complexes naturally fit into a batch job processing system and can be used in a GRID environment. Such a GRID system arises on the computing resources of the Yaroslavl State University and the Institute of Organic Chemistry of the Russian Academy of Sciences on the base of EGI standardized UMD 4 software.

Key words and phrases: high speed quantum chemical methods, docking complexes, GRID.

2010 *Mathematics Subject Classification*: 92C40; 65D30,97M60

References

[1] P. Jørgensen, T. Kjaergaard, K. Kristensen, P. Baudin, P. Ettenhuber, J. J. Eriksen, Y. M. Wang, D. Bykov. "Quantum chemistry on the supercomputers of tomorrow", Smoky Mountains Computational Sciences and Engineering Conference (August 31–September 2, 2015, Gatlinburg, Tennessee, USA). ↑76

[2] D. Bykov, A. Barnes, D. Lyakh. "Quantum Chemistry on Supercomputers", Southeastern Theoretical Chemistry Association Meeting SETCA 2019 (May 16–18, 2019, The University of Tennessee, Knoxville, USA). ↑76

[3] W. Jia, L. W. Wang, L. Lin. *Parallel transport time-dependent density functional theory calculations with hybrid functional on summit*, 2019. arXiv 1905.01348 ↑76

[4] P. Gwynne. "Exascale supercomputer initiative launched", *Physics World*, **32**:5 (2019), pp. 11. ↑76

[5] P. Cheng, Y. Lu, Y. Du, Zh. Chen. "Tiered data management system: Accelerating data processing on HPC systems", *Future Generation Computer Systems*, **101** (2019), pp. 894–908. ↑76

[6] R. Johnson. *Aurora supercomputer to empower advanced chemistry research*, 2019. ↑76

[7] H. A. Le, T. Shiozaki. "Occupied-orbital fast multipole method for efficient exact exchange evaluation", *Journal of Chemical Theory and Computation*, **14**:3 (2018), pp. 1228–1234. ↑77

[8] M. Hennemann, T. Clark. "EMPIRE: a highly parallel semiempirical molecular orbital program: 1: self-consistent field calculations", *Journal of Molecular Modeling*, **20**:7 (2014), 2331. ↑77

[9] I. V. Oferkin, E. V. Katkova, A. V. Sulimov, D. C. Kutov, S. I. Sobolev, V. V. Voevodin, V. B. Sulimov. "Evaluation of docking target functions by the comprehensive investigation of protein-ligand energy minima", *Advances in Bioinformatics*, **2015** (2015), 126858. ↑77

[10] A. Li, H. S. Muddana, M. K. Gilson. "Quantum mechanical calculation of noncovalent interactions: A large-scale evaluation of PMx, DFT, and SAPT approaches", *Journal of Chemical Theory and Computation*, **10**:4 (2014), pp. 1563–1575. ↑77

[11] J. C. Womack, N. Mardirossian, M. Head-Gordon, Ch.-K. Skylaris. "Self-consistent implementation of meta-GGA functionals for the ONETEP linear-scaling electronic structure package", *The Journal of Chemical Physics*, **145**:20 (2016), 204114. ↑77

[12] E. J. Higgins, M. I. J. Probert, P. J. Hasnip, K. Refson, I. J. Bush. *Hybrid OpenMP and MPI within the CASTEP code*, 2015. ↑78

[13] J. Zhang, A. L. Weisman, P. Saitta, R. A. Friesner. "Efficient simulation of large materials clusters using the jaguar quantum chemistry program: Parallelization and wavefunction initialization", *International Journal of Quantum Chemistry*, **116**:5 (2016), pp. 357–368. ↑78,79

[14] I. Horváth, N. Jeszenői, M. Bálint, G. Paragi, C. Hetényi. "A fragmenting protocol with explicit hydration for calculation of binding enthalpies of target-ligand complexes at a quantum mechanical level", *International Journal of Molecular Sciences*, **20**:18 (2019), pp. 4384. ↑79

[15] A. V. Nemukhin, I. V. Polyakov, A. I. Moskovsky. "Multi-scale supercomputing of large molecular aggregates: A case study of the light-harvesting photosynthetic center", *Supercomputing Frontiers and Innovations*, **2**:4 (2016), pp. 48–54. ↑79

[16] T. Nakajima, M. Katouda, M. Kamiya, Yu. Nakatsuka. "NTChem: A highperformance software package for quantum molecular simulation", *International Journal of Quantum Chemistry*, **115**:5 (2015), pp. 349–359. ↑79

[17] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne. "The Protein Data Bank", *Nucleic Acids Research*, **28**:1 (2000), pp. 235–242. ↑82

[18] R. Wang, Y. Lu, X. Fang, S. Wang. "An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes", *J. Chem. Inf. Comput. Sci.*, **44**:6 (2004), pp. 2114–2125. ↑83

[19] N A. Anikin, A. S. Mendkovich, M. B. Kuz'minskiy, A. M. Andreyev. "A method and program for mass quantum chemical calculations of protein-ligand docking complexes", *Russian Chemical Bulletin*, **57** (2008), pp. 428–430. ↑83

[20] N. A. Anikin, A. M. Andreyev, M. B. Kuz'minskiy, A. S. Mendkovich. "A fast method of large-scale serial semiempirical calculations of docking complexes", *Russian Chemical Bulletin*, **57** (2008), pp. 1793–1798. ↑83

[21] N. A. Anikin, A. M. Andreyev, M. B. Kuz'minskiy, A. S. Mendkovich. "A new approach for the acceleration of large-scale serial quantum chemical calculations of docking complexes", *Russian Chemical Bulletin*, **67** (2018), pp. 1100–1103. ↑84,85

[22] N. A. Anikin, A. Yu. Muskatin, M. B. Kuz'minskiy, A. I. Rusakov. "GRID-system based on European EGI standards for large-scale calculations using the original accelerated method of quantum chemistry", *Modelirovaniye i analiz informatsionnykh sistem*, **26**:3 (2019), pp. 359–363 (in Russian). ↑85

Sample citation of this publication:
Nikolay A. Anikin, Aleksandr Y. Muskatin, Mikhail B. Kuz'minskiy, Sergey N. Lednev, Aleksandr V. Smirnov, Aleksandr I. Rusakov. "Cardinal acceleration of calculations of giant biomolecules by quantum chemistry methods, requiring the use of supercomputers and / or GRID systems". *Program Systems: Theory and Applications*, 2020, **11**:2(45), pp. 75–92. (*In Russian*). 10.25209/2079-3316-2020-11-2-75-92
http://psta.psiras.ru/read/psta2020_2_75-92.pdf