

УДК: 004.94

Программный комплекс L-MOLKERN для расчетов разностей свободных энергий с учетом эффектов перераспределения заряда

©2012 Фомин Э.С.* , Алемасов Н.А.

*Институт цитологии и генетики, Сибирское отделение Российской академии наук,
Новосибирск, 630090, Россия*

Аннотация. Представлен программный комплекс L-MOLKERN для расчетов разностей свободных энергий молекулярных комплексов методом λ -динамики. Для уточнения расчетов разностей свободных энергий между состояниями с изменением заряда L-MOLKERN использует оригинальный метод учета эффектов перераспределения заряда Net-q. Показаны результаты выполненных расчетов (на примере мутаций белка барназы), демонстрирующие улучшение точности расчетов. Продемонстрирована высокая масштабируемость разработанного программного комплекса L-MOLKERN.

Ключевые слова: молекулярная динамика, λ -динамика, разность свободных энергий, поляризация, масштабируемость вычислений.

ВВЕДЕНИЕ

Свободные энергии (химические потенциалы) являются основными величинами, которые определяют поведение молекулярных систем около и в точке равновесия, включая связывание лигандов к белкам, конформационные преобразования белков, лежащие в основе механизмов их функционирования, ассоциации совокупности белков и ДНК/РНК в наномолекулярные комплексы и прочее. Прогресс в методах расчета абсолютных и относительных свободных энергий долгое время сдерживался недостаточной мощностью вычислительной техники и неэффективностью (по причине отсутствия сходимости) ранних методов [1–3]. С развитием вычислительной техники и методов расчета в основном удалось решить проблему сходимости свободных энергий даже для относительно сложных гетерогенных систем [4–6]. Основные усилия, позволивших это сделать, связаны с развитием методов “алхимических” преобразований [7], “soft-core” потенциалов [8–9] и методов генерации ансамблей, обеспечивающих широкую выборку точек конфигурационного пространства комплексов [10–12]. Среди методов расчета свободных энергий наибольшей общностью и возможностями применения для широкой области биомолекулярных систем обладают методы, основанные на том или ином варианте метода “алхимических преобразований”, и по этой причине они реализованы во всех наиболее распространенных программах молекулярной динамики (МД), таких как GROMACS [13], AMBER [14], NAMD [15] и прочих.

Решение проблем со сходимостью расчетов свободных энергий позволяет поставить другую задачу в развитии данных методов, связанную с повышением точности вычислений – учет эффектов поляризации. Большинство силовых полей, используемых

* fomin@bionet.nsc.ru

в расчетах свободных энергий, для описания электростатических взаимодействий основываются на фиксированных точечных зарядах для атомов. Точечные заряды обычно локализируются на атомах (или на связях между соседними атомами), и их значения подгоняются в зависимости от типа и локального окружения атомов таким образом, чтобы описать электростатическое поле молекулы в наиболее статистически значимой конфигурации [16]. Приближение фиксированных зарядов не всегда оправдано, поскольку зарядовое распределение в молекуле зависит от ее конфигурации и от внешнего окружения. Общеизвестным примером этого является разный дипольный момент молекулы воды, который меняется от 1.85Д в газовой фазе до 2.5Д в жидкой фазе. Учет эффектов поляризации приводит к существенному увеличению точности расчетов, например при исследовании докинга широкого класса ингибиторов белков, включающих трипсин, сахар-связующие белки и тРНК синтаза [17] получено снижение RMSD от 1-3Å, характерного для стандартных силовых полей с фиксированными зарядами, до 0.2-0.4 Å.

В настоящей работе представлен программный комплекс L-MOLKERN, решающий задачу учета эффектов поляризации при расчетах разностей свободных энергий методом “алхимической” λ -динамики. В L-MOLKERN используется оригинальный Net-q метод учета перераспределения зарядов в процессе динамики, основанный на уравнивании электроотрицательностей [18]. В работе дано описание используемых в L-MOLKERN методов, обсуждены результаты предварительных расчетов, сделано сравнение с экспериментальными данными и расчетами других авторов, проведена проверка эффективности распараллеливания программного комплекса.

ПРОГРАММНЫЙ КОМПЛЕКС L-MOLKERN

Программный комплекс L-MOLKERN построен на базе библиотеки компонент для молекулярного моделирования MOLKERN [19], которая разрабатывается в ИЦиГ СО РАН с 2006 г. с целью обеспечить эффективное выполнение базовых элементов задач моделирования структуры и динамики комплексов биомолекул (белков, ДНК/РНК, лигандов и кофакторов) в явном и неявном водном окружении с использованием силовых полей AMBER и GAFF [20]. Программный комплекс L-MOLKERN позволяет выполнять следующие основные функции:

- считывание файлов (форматов pdb, ent, hin, mol2 и др.), содержащих данные о пространственной структуре молекул;
- восстановление отсутствующих атомов аминокислотных остатков, добавление атомов водорода в соответствии с заданным pH среды, построение дисульфидных связей, выполнение серий одиночных аминокислотных замен в белковой цепи;
- поиск в комплексах молекул водородных связей, солевых и водных мостиков, стэкинг-взаимодействий, определение поверхностных атомов и атомов, находящихся на контактной поверхности белок-лиганд и белок-белок;
- оптимизация структуры комплексов методами L-BFGS и L-MBFGS [21] без и с наличием ограничений;
- аналитический [22] и численный расчет площади поверхности, доступной водному окружению;
- расчет энергии комплексов в вакууме и в водном окружении с учетом энергии поляризации среды обобщенным методом Борна [23];
- оценка свободной энергии сильно связанного лиганда прямым методом [24];
- выполнение молекулярной динамики в NVE, NVT и NVP ансамблях с использованием алгоритма интегрирования уравнений движения с “перешагиванием” (leap-frog) [25] и поддержанием температуры и давления методом “термостата Беренсена” [26];

- построение двойной топологии аминокислотных остатков белков для выполнения расчета разностей свободных энергий методом термодинамического интегрирования в подходе расширенного λ -ансамбля;
- выполнение λ -молекулярной динамики для расчета разностей свободных энергий и сохранение статистических данных в файлах оригинального формата bmm;
- учет перераспределения заряда на основе принципа выравнивания электроотрицательностей методом Net-q при оптимизации, молекулярной и λ -динамике;

L-MOLKERN реализован на языке C++ с использованием библиотек STL и BOOST и может использоваться в операционной системе LINUX. Программный комплекс L-MOLKERN векторизован с помощью sse intrinsics, выполняется в многопоточном режиме с помощью библиотеки boost::thread и использует технологию MPI для счета на многопроцессорных системах.

Эффективность реализаций алгоритмов находится на уровне, сравнимым с общемировым. Многочисленные расчеты с помощью программного комплекса L-MOLKERN показывают, что его производительность достигает 0.1 млрд. взаимодействий в секунду на процессоре Intel® Core™2 Quad CPU Q8400 2.66 GHz в 4-х поточном исполнении [27]. Запросы на объем памяти программным комплексом незначительны, что позволяет считать системы с более чем 3 млн. атомов в пределах оперативной памяти порядка 0.5 Гб. На отдельных тестах полученный код превосходит по эффективности от 1.5 до 2 раз общеизвестные программы AMBER и NAMD, и уступает только GROMACS. Эффективность реализации обусловлена, как алгоритмической оптимизацией – все алгоритмы программы имеют вычислительную сложность, не превышающую $O(N \log N)$, – так и с использованием ряда оригинальных подходов для обеспечения минимизации потока данных между глобальной памятью и CPU, а также автоматической подстройкой параметров в направлении предельной эффективности расчета при различных физических параметрах симуляции [27].

1. Метод расчета свободной энергии и λ -потенциалы

В L-MOLKERN используется метод термодинамического интегрирования [28] для определения изменения свободной энергии между системами А и В:

$$\Delta F = \sum_{j=1}^{L-1} \Delta F(\lambda_j \rightarrow \lambda_{j+1}) = \int_0^1 \left\langle \frac{dH}{d\lambda} \right\rangle d\lambda$$

где состояние расширенной системы А+В при $\lambda = 0$ соответствует системе А, а при $\lambda = 1$ – системе В, а выражение в угловых скобках $\langle dH/d\lambda \rangle$ является усреднением величины по ансамблю:

$$\left\langle \frac{dH}{d\lambda} \right\rangle = \int \rho(p, x) \left(\frac{dH}{d\lambda} \right) dpdx = \frac{1}{Z} \int \exp\left(-\frac{H(p, x)}{k_b T}\right) \left(\frac{dH}{d\lambda} \right) dpdx$$

В L-MOLKERN с каждой частицей i связана величина λ_i : $\lambda_i = \lambda$ для уникальных атомов системы А, $\lambda_i = 1 - \lambda$ для уникальных атомов системы В и $\lambda_i = 1$ для общих атомов расширенной системы А+В, и определены правила комбинирования параметров $\lambda_i, \lambda_j, \dots, \lambda_n$ для взаимодействующих частиц i, j, \dots, n в виде произведения $\lambda_i \lambda_j \dots \lambda_n$.

Несвязующие парные (кулоновский и Ван-дер-Ваальса) потенциалы определены следующим образом:

$$V_{coul}(r_{ij}) = \frac{q_i \lambda_i q_j \lambda_j}{r_{ij}}$$

$$V_{vdw}(r_{ij}) = 4\varepsilon_{ij}\lambda_i\lambda_j \left(\left(\frac{\sigma_{ij}\lambda_i\lambda_j}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}\lambda_i\lambda_j}{r_{ij}} \right)^6 \right)$$

где q_i и q_j - заряды атомов i и j , ε_{ij} и σ_{ij} - стандартные параметры силового поля, определяемыми в соответствии с правилами комбинирования в используемом силовом поле, например для AMBER по правилам $\varepsilon_{ij} = \sqrt{\varepsilon_i\varepsilon_j}$ и $\sigma_{ij} = (\sigma_i + \sigma_j)/2$.

Связующие потенциалы валентных связей $V_b(r, \lambda)$, валентных $V_a(\alpha, \lambda)$ и двугранных $V_t(\varphi, \lambda)$ углов и зависимость кинетической энергии $K_i(\lambda)$ атома i от λ определены следующим образом:

$$V_b(r, \lambda) = \frac{\lambda_i\lambda_j k_{ij}(r - r_0)^2}{2}$$

$$V_a(\alpha, \lambda) = \frac{\lambda_i\lambda_j\lambda_k k_{ijk}(\alpha_{ijk} - \alpha_0)^2}{2}$$

$$V_t(\varphi, \lambda) = \lambda_i\lambda_j\lambda_k\lambda_l \sum_m v_m (1 + \cos(\varphi n_m - \varphi_0))$$

$$K_i(\lambda) = \frac{\lambda_i m_i V_i^2}{2}$$

2. Метод Net-q учета изменения заряда

В L-MOLKERN реализован оригинальный метод расчетов зарядового распределения, называемый Net-q методом [29], основанный на минимизации энергии электростатического поля QEq метода [30] при ограничениях, связанных с уравниванием электроотрицательностей атомов и сохранением полного заряда молекул. Движение заряда в методе разрешено только вдоль ребер графа, представляющего молекулярную систему, при этом учитываются взаимодействия в пределах радиуса обрезания r_{cutoff} , что является одним из условий для обеспечения линейной вычислительной сложности $O(N)$ вычислений. Динамика заряда в методе Net-q описывается подходом расширенного ансамбля (введение фиктивных “масс” и “импульсов” заряда), аналогичному методу fluc-q [31]. Дополнительно в методе Net-q введено ограничение на перенос заряда на расстояниях больших длины химической связи, аналогичное [32]. Следует заметить, что Net-q по способам решения проблем с эффективностью похож на продвинутые методы индуцированных диполей, такие как [33–34], в которых стандартная процедура решения уравнений минимизации энергии электростатического поля, основанная на обращении матрицы заменяется на итерационную, а затем используется подход расширенного ансамбля с целью предсказания начальных значений матрицы либо с целью отказа от итерирования.

Модельный гамильтониан зарядовой подсистемы Net-q метода определен в виде:

$$H(\mathbf{Q}, \mathbf{W}, \mathbf{X}, \mathbf{P}) = \frac{1}{2} \mathbf{W}^T \mathbf{B}^{-T} \mathbf{W} + (\tilde{\chi}, \mathbf{Q}) + \frac{1}{2} \mathbf{Q}^T \mathbf{J} \mathbf{Q} + \frac{1}{2} \mathbf{P}^T \mathbf{M}^{-1} \mathbf{P} + U(\mathbf{X}) \quad (1)$$

где $\mathbf{Q} = \{q_1, q_2, \dots, q_{nq}\}$ и $\mathbf{W} = \{w_1, w_2, \dots, w_{nq}\}$ – вектора зарядов и “импульсов” их движения, $\mathbf{X} = \{\{x, y, z\}_1, \{x, y, z\}_2, \dots, \{x, y, z\}_{na}\}$ и $\mathbf{P} = \{p_1, p_2, \dots, p_{na}\}$ – координаты и импульсы атомов, $\tilde{\chi} = \{\tilde{\chi}_1, \tilde{\chi}_2, \dots, \tilde{\chi}_{na}\}$ – электроотрицательности атомов, $\mathbf{M} = \{m_1, m_2, \dots, m_{na}\}$ – массы атомов, $U(\mathbf{X})$ – потенциал силового поля, который не связан с зарядами атомов (например, Ван-дер-Ваальсовый потенциал). Используемая в (1) матрица $\mathbf{B} = \mathbf{A}^{-T} \tilde{\mathbf{M}} \mathbf{A}^{-1}$ включает в себя матрицу инцидентности графа \mathbf{A} с элементами $\{0, 1\}$ и вектор “масс” токов молекулы $\tilde{\mathbf{M}} = \{\tilde{m}_1, \tilde{m}_2, \dots, \tilde{m}_{nq}\}$. Модифицированная согласно [32] электроотрицательность атома $\tilde{\chi}_i = \sum_j f_k(r_{ij}) (\chi_i^0 - \chi_j^0)$, включает

функцию ограничения переноса заряда на дальних расстояниях $f_k(r_{ij}) = k_{ij}\langle\phi_i|\phi_j\rangle$, где k_{ij} – масштабный параметр, а $\langle\phi_i|\phi_j\rangle$ – интеграл перекрытия Слэтеровских *ns*-атомных орбиталей. Матрица взаимодействий зарядов J на диагонали включает “жесткости” атомов $J_{ij} = \eta_i$, а недиагональные элементы пропорциональны кулоновскому интегралу, вычисляемому как интеграл по *ns*-орбиталям Слэтеровского типа [35]:

$$J_{AB}(R_A, R_B) = \iint \frac{\phi^2(r_A, R_A)\phi^2(r_B, R_B)}{|r_A - r_B|} dr_A dr_B$$

$$\phi_A(r, R) = \frac{(2\zeta)^{n+\frac{1}{2}}}{\sqrt{4\pi(2n)!}} |r - R|^{n-1} \exp(-\zeta|r - R|)$$

Из гамильтониана (1) следуют уравнения движения для заряда и атомов:

$$\dot{Q} = \frac{\delta H}{\delta W} = A \tilde{M}^{-1} A^T W$$

$$\dot{W} = -\frac{\delta H}{\delta Q} = -\{\tilde{\chi} + JQ\}$$

$$\dot{X} = \frac{\delta H}{\delta P} = M^{-1}P$$

$$\dot{P} = -\frac{\delta H}{\delta X} = -\left[\left(\frac{\delta \chi}{\delta X}, Q \right) + \frac{1}{2} \left(W^T \frac{\delta B^{-T}}{\delta X} W + Q^T \frac{\delta J}{\delta X} Q \right) + \frac{\delta U(X)}{\delta X} \right]$$

Поскольку полученные уравнения движения включают топологическую матрицу молекулярной системы A , которая не требует обращения и является разреженной, а матрица взаимодействий ограничена радиусом обрезания r_{cutoff} , то вычислительная сложность метода является линейной. Использование матрицы A означает наличие принудительных запретов на перенос заряда между теми атомами, для которых соответствующий элемент матрицы A равен нулю. Таким образом, матрица A дает возможность компромисса между эффективностью расчетов и детальностью описания процессов переноса заряда. Ненулевые элементы матрицы A разрешают перенос заряда по соответствующим связям, и ограничение переноса в этом случае, если длина связи превышает среднюю длину химической связи, выполняется функцией ограничения $f_k(r_{ij})$.

3. Методика тестовых исследований L-MOLKERN

В качестве объекта для исследования точности расчетов в рамках методов, реализованных в L-MOLKERN, был выбран белок барназа. Барназа (Бактериальная РибоНуклеАза) - небольшой белок (110 аминокислот), который синтезируется в бактерии *Vacillus amyloliquefaciens*. Имеется пространственная структура барназы, наилучшее разрешение которой составляет 1.5Å [36]. Известно свыше 400 мутантных форм белка, образованных, как точечными, так и множественными заменами аминокислот [37].

Белок барназа является хорошо изученным объектом, в том числе и методами λ -динамики. Для большинства (71.6%) изученных мутаций этого белка данным методом расхождение рассчитанных данных с экспериментальными находится в пределах ± 4 кДж/моль, то есть в пределах стандартной ошибки силового поля AMBER. Для мутаций, включающих заряженные аминокислотные остатки, точность рассчитанных данных существенно хуже: только 52% находятся в пределах ± 4 кДж/моль от экспериментального значения, и для 24% мутаций отклонение превышает 10 кДж/моль [38].

Для тестовых исследований L-MOLKERN были выбраны мутации A32G и R72G белка барназы. Для мутации A32G характерно сохранение зарядового состояния белка и отсутствие энтропийного вклада в силу симметрии CH_3 группы в аланине. Для R72G характерно изменение зарядового состояния белка. Кроме того, для данной мутации рассчитанная двойная разность свободных энергий $\Delta\Delta G_{calc} = -57.67$ кДж/моль методом λ -динамики [38] существенно отклоняется от экспериментального значения $\Delta\Delta G_{exp} = -10.45$ кДж/моль [37].

Все расчеты выполнялись со следующими параметрами:

- размер пространственной области моделирования – $60 \times 60 \times 60 \text{ \AA}^3$; число молекул воды – 5577;
- радиус обрезания потенциала – 10 \AA ;
- используемый тип термодинамического ансамбля – NVT с температурой моделирования – 400 К; уравнивание температуры с помощью метода “термостат Беренсена” [26];
- длина расчетных траекторий молекулярной динамики – 20 нс с шагом интегрирования – 2 фс;
- шаг интегрирования расширенной λ -траектории – от 1/8 до 1/10;
- используемый для расчета ΔG диапазон траекторий – от 10 до 20 нс модельного времени.

Расчеты производились с использованием ресурсов ЦКП “Биоинформатика” СО РАН, включающих вычислительный кластер, состоящий из: 64 двухпроцессорных серверов (Intel Xeon E5540), 48 двухпроцессорных серверов (Intel Xeon X5670) и параллельной файловой системы IBRIX для хранения данных: 4 шлюзовых сервера (HP DL380 G6), 4 дисковые полки (2 - HP MSA 2312sa DC, 2 - HP MSA 2000 DC) и сервер управления (HP DL360 G6).

Дополнительно были выполнены испытания уровня масштабируемости вычислений при использовании SSE-векторизации, многопоточности и MPI. При этом проверка уровня масштабирования для векторизации и многопоточности была проведена с использованием одного вычислительного узла кластера G6 ЦКП “Биоинформатика”, а для расчетов, использующих MPI, было выполнено крупномасштабное испытание с использованием 125 вычислительных узлов ($125 \times 8 = 1000$ ядер) кластера, на каждом из которых расчет выполнялся в 8 потоков. Для распределения расчетов по узлам кластера использовалась библиотека Intel MPI. Также была проверена эффективность вычислений в пределе большого числа атомов, при этом их число изменялось в диапазоне от $\sim 10^3$ до $\sim 10^6$.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

1. Расчет разности свободных энергий мутаций A32G и R72G относительно белка барназы в дикой форме

Результаты расчетов разности свободных энергий ΔG в процессе “алхимического” превращения для мутации A32G дали величину $\Delta G = -37.6$ кДж/моль, которая очень хорошо (в пределах точности ± 4 кДж/моль силового поля AMBER) согласуется со значением -36.1 кДж/моль, полученным в работе [38]. Таким образом, для мутаций, в которых отсутствует изменение зарядового состояния показано хорошее согласие получаемых L-MOLKERN результатов с расчетами других авторов.

Результаты расчетов двойной разности свободных энергий $\Delta\Delta G$ для мутации R72G относительно белка барназы в дикой форме дали значение $\Delta\Delta G_{calc} = -19.8$ кДж/моль. Полученная величина $\Delta\Delta G_{calc}$ находится существенно ближе к экспериментальному значению $\Delta\Delta G_{exp} = -10.45$ кДж/моль [37], чем рассчитанная $\Delta\Delta G_{calc} = -57.67$ кДж/моль в [38] с расхождением относительно экспериментального значения в

9.4 кДж/моль и 47.2 кДж/моль, соответственно. Детальное описание выполненных расчетов сделано в работе [39].

Таким образом, учет изменения заряда при расчетах разностей свободных энергий обеспечивает лучшее согласие расчетных данных с экспериментальными значениями.

2. Эффективность распараллеливания

Полное время отработки кода в векторизованном варианте для N атомов включает в себя:

- $T_c \sim 2 \times O(N)$ – время создания SSE-массивов (минимум два массива);
- $T_r \sim O(N^2)/4$ – время векторизованных расчетов с SSE-массивами (в T_r учтено, что за одну операцию выполняются действия для 4 атомов);

Оценка эффективности должна корректироваться на фактор k_f , который отражает то, что не с каждой парой атомов, попавших в SSE-вектора, связано несвязующее взаимодействие. Отсутствие взаимодействия обусловлено двумя факторами: атомы могут быть связаны химической связью прямо или опосредствованное через один или два атома (1-3 и 1-4 взаимодействия); атомы могут быть “случайно” захвачены при расчетах на границе радиуса обрезания r_{cutoff} . Таким образом, при векторизации нельзя получить ускорения счета выше чем:

$$speedup_{sse} \approx \frac{N^2}{2N + \frac{N^2}{4}} k_f = 4N \frac{k_f}{8 + N}$$

Используя в данном выражения значения параметров $k_f \sim 4/5$ (из тестовых расчетов) и $N \sim 30$ (среднее число молекул воды в ячейке с длиной ребра равной $r_{cutoff} = 10 \text{ \AA}$) получим теоретический предел ускорения при SSE-векторизации $speedup_{sse} \approx 2.4$.

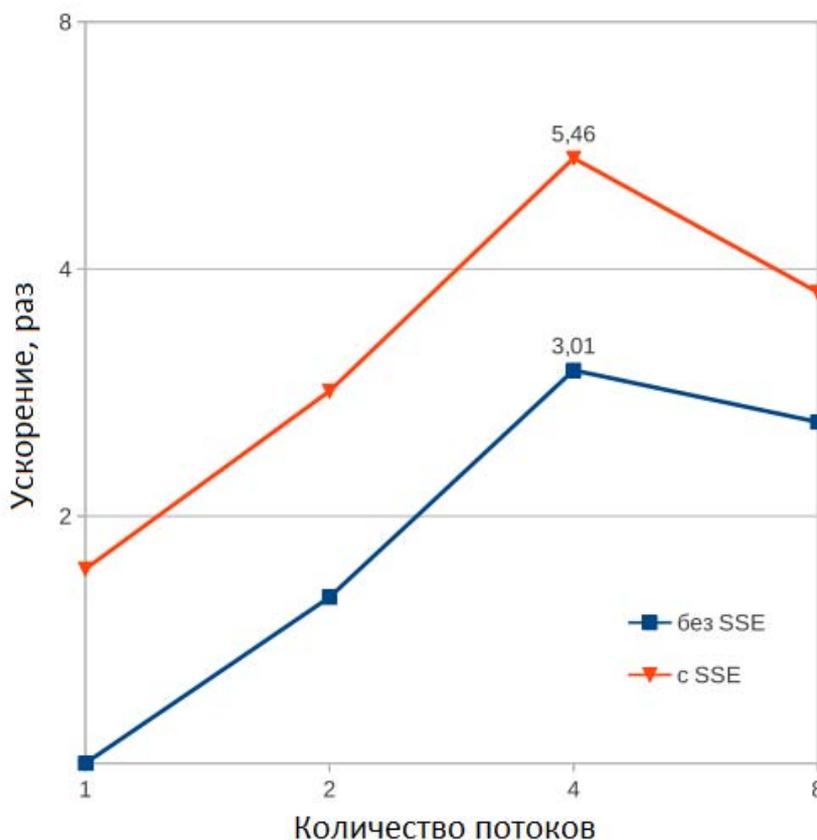


Рис. 1. Ускорение L-MOLKERN без и при использовании SSE-векторизации.

На рис. 1 представлен график ускорения вычислений L-MOLKERN в зависимости от числа использованных потоков с использованием и без использования SSE-векторизации. Все точки графика даны относительно выполнения L-MOLKERN в одном потоке без SSE-векторизации. На графике оси абсцисс и ординат имеют логарифмический масштаб.

Как видно из графика, векторизация увеличивает эффективность выполнения приблизительно в 1.8 раз, что качественно согласуется с теоретическим пределом $speedup_{sse} \approx 2.4$, демонстрируя эффективность использования векторизации на уровне $\sim 75\%$. Отличия от предельного значения обусловлены невозможностью полной векторизации кода L-MOLKERN (например, векторизации не поддается используемый к коду алгоритм сортировок взаимодействий [40]).

На рис. 1 также наблюдается линейный рост производительности L-MOLKERN с увеличением числа потоков выполнения вплоть до значения равного 4, при этом эффективность использования многопоточности достигает $\sim 75\%$. При большем числе потоков производительность падает. Причина данного эффекта детально не исследовалась. Предположительно она связана с тем, что выполнение в 8 потоков требует подключения обоих процессоров вычислительного узла, и, как следствие, существенного обмена данными между ними. При этом данные не могут быть переданы с регистров одного процессора на регистры другого прямым образом, и вынуждены пройти через оперативную память, то есть по пути $\{L1 \rightarrow L2 \rightarrow L3\}_1 \rightarrow memory \rightarrow \{L3 \rightarrow L2 \rightarrow L1\}_2$, где $L1, L2, L3$ кэши 1, 2 и 3 уровней обоих процессоров. Данный процесс передачи данных в силу высокой латентности оперативной памяти, достигающей сотни тактов процессора, может существенно затормозить вычисления. Другой возможной причиной эффекта снижения производительности может быть необходимость синхронизации кэшей, затраты на которую существенно выше для двух процессоров, нежели для одного.

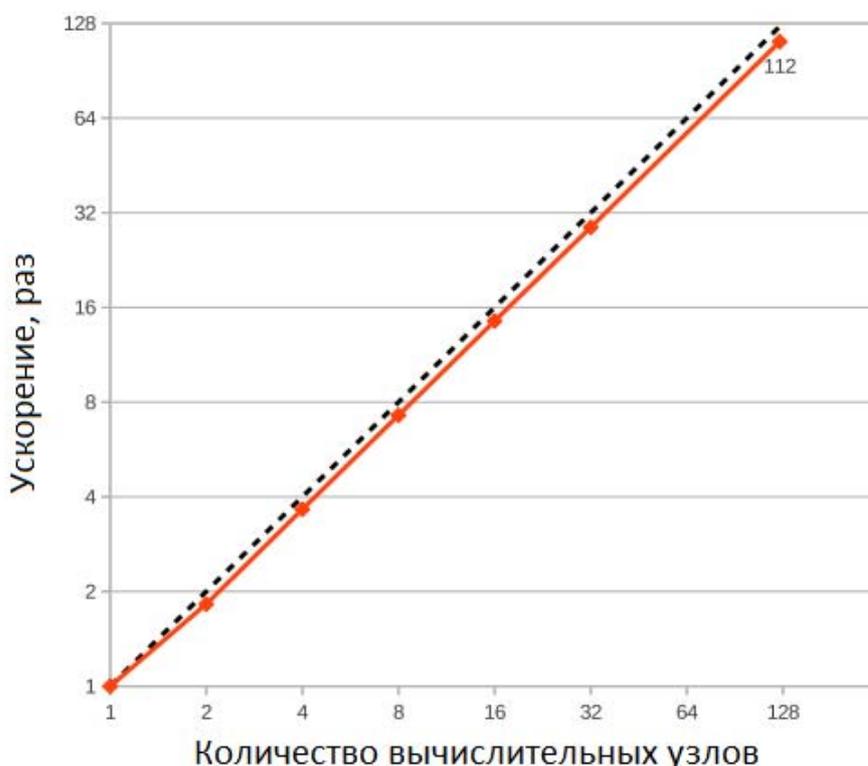


Рис. 2. График ускорения L-MOLKERN в зависимости от числа узлов кластера относительно выполнения на одном узле.

На рис. 2 представлены результаты ускорения расчета L-MOLKERN на вычислительном кластере G6 ЦКП “Биоинформатика” в зависимости от числа использованных узлов кластера относительно вычислений на одном узле кластера. На графике оси абсцисс и ординат имеют логарифмический масштаб. Как видно на рис. 2, наблюдается практически идеальная масштабируемость, что обусловлено практически полной независимостью вычислений по данным (каждый узел кластера выполняет свой расчет МД траектории и накапливает статистику в своей точке λ -траектории).

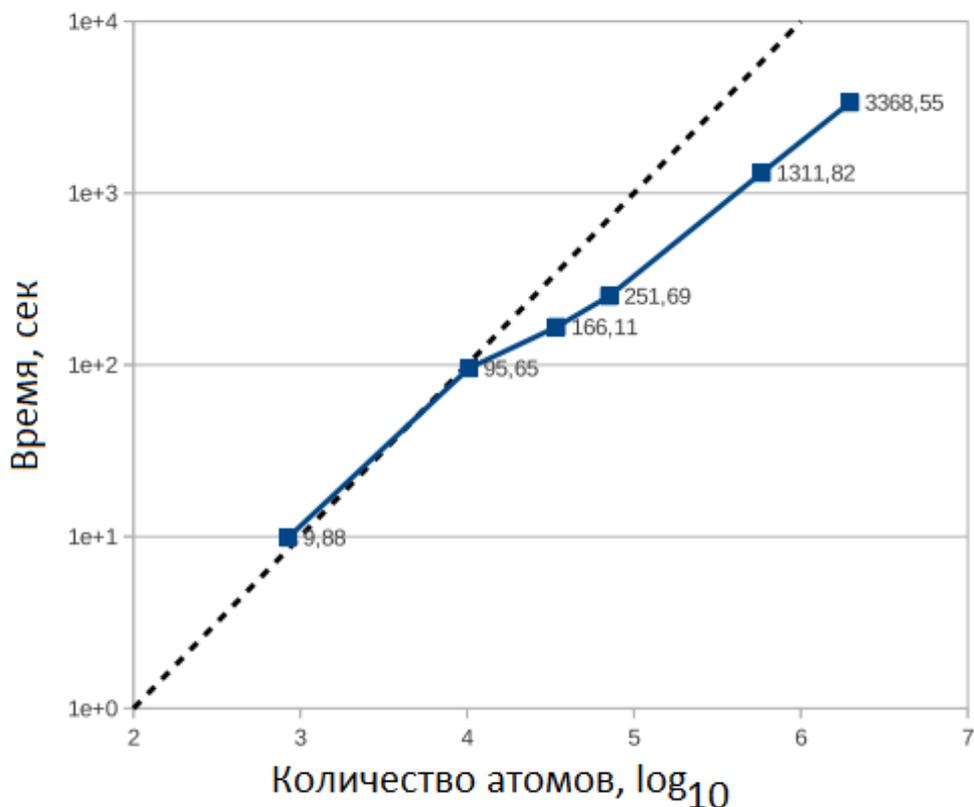


Рис. 3. Зависимость времени расчетов от числа атомов (данные даны в логарифмическом масштабе).

На рис. 3 показана зависимость времени вычислений от размера системы. Обе оси графика даны в логарифмическом масштабе. Для сравнения пунктиром приведена прямая $\log T = const + \log N$, которая представляет “идеальную” ситуацию точной линейной зависимости между размером системы и временем вычислений $T = \alpha N$, где α – некоторый коэффициент пропорциональности. Как видно из графика, рост времени моделирования ниже, чем рост размера системы, то есть эффективность расчетов тем выше, чем больше атомов содержится в моделируемой области. Как результат, средние затраты времени на обработку одного атома падают более чем в 5 раз при увеличении числа атомов в системе от 10^3 до 10^6 . Причина эффекта увеличения эффективности вычислений с размером системы заключается в том, что при $N \rightarrow \infty$ возрастает относительная доля времени вычислений, которые выполняются в параллельном режиме. Кроме того, нужно упомянуть тот факт, что код L-MOLKERN специально оптимизирован для случая $N \rightarrow \infty$.

ВЫВОДЫ

Решение проблем со сходимостью в методах расчета свободных энергий в течении последнего десятилетия позволило поставить новую задачу в развитии этих методов, связанную с повышением точности вычислений – учет эффектов поляризации. В

данной работе представлен программный комплекс L-MOLKERN, который разработан для решения данной задачи. L-MOLKERN позволяет выполнять “алхимическую” λ -молекулярную динамику в рамках метода двойной топологии для различных термодинамических ансамблей с учетом перераспределения заряда на основе принципа выравнивания электроотрицательностей методом Net-q.

В работе предоставлены результаты тестовых испытаний программного комплекса L-MOLKERN. Показано (на примере мутаций белка барназы), что учет зарядового перераспределения существенен для расчетов разностей свободных энергий при мутациях, которые затрагивают аминокислоты с различным зарядовым состоянием. Расчеты двойной разности $\Delta\Delta G$ для мутации R72G белка относительно дикой формы белка в рамках L-MOLKERN дали значение $\Delta\Delta G_{calc} = -19.8$ кДж/моль, которое находится существенно ближе к экспериментальному $\Delta\Delta G_{exp} = -10.45$ кДж/моль, чем значение $\Delta\Delta G_{calc} = -57.67$ кДж/моль, рассчитанное с фиксированными значениями зарядов.

В тестовых испытаниях также продемонстрирована высокая масштабируемость разработанного программного комплекса L-MOLKERN, которая является важнейшим фактором для подобного рода задач, которые для масштабных вычислений требуют вычислительных систем экзафлопного уровня производительности. L-MOLKERN для параллелизации использует все возможности, предоставляемые процессорами и кластерными системами: векторизацию (на уровне ядра процессора), многопоточность (на уровне многоядерного процессора) и MPI (на уровне вычислительного кластера). Тестовые испытания показали, что L-MOLKERN использует ~75% от максимально возможного уровня векторизации, ~75% от максимально возможного уровня многопоточности (при задействовании 4 ядер одного 4-ядерного процессора) и ~90% от максимально возможного уровня при распараллеливании с помощью MPI. L-MOLKERN оптимизирован для расчетов систем с большим числом частиц, и при увеличении числа частиц от 10^3 до 10^6 относительные затраты на обработку одной частицы падают более, чем в 5 раз. Таким образом, вычислительный комплекс L-MOLKERN может быть эффективно использован на высокопроизводительных вычислительных кластерах для больших по размеру молекулярных систем.

БЛАГОДАРНОСТИ

Работа выполнена в рамках Госконтракта № 07.514.11.4011 с Министерством образования и науки РФ.

СПИСОК ЛИТЕРАТУРЫ

1. Berendsen H.J.C. Incomplete equilibration: A source of error in free energy computations. In : *Proteins: Structure, Dynamics and Design*. Eds. Renugopalakrishnan V., Carey P.R., Smith I.C.P., Huang S.G., Storer A.C.: Leiden, ESCOM, 1991. P. 384–392.
2. Chipot C., Pearlman D.A. Free energy calculations: the long and winding gilded road. *Mol. Simulation*. 2002. V. 28. № 1–2. P. 1–12.
3. Reddy M.R., Erion M.D. *Free Energy Calculations in Rational Drug Design*. New York: Kluwer Academic, 2001
4. Raha K., Kenneth M., Merz J. Calculation binding free energy in protein-ligand interaction. *Annu. Rep. Comp. Chem*. 2005. V. 1. P. 113–130.
5. Rodinger T., Romes R. Engancing the accuracy, the efficiency and the scope of free energy simulations. *Curr. Opin. Struct. Biol*. 2005. V. 15. P. 164–170.
6. Chipot C., Pohorille C. *Free Energy Calculations: Theory and Application in Chemistry and Biology*. V. 86. Berlin and Heidelberg: Springer, 2007.

7. Knight J.L., Brooks C.L.J. λ -Dynamics free energy simulation methods. *Comput. Chem.* 2009. V. 30. № 11. P. 1692–1700.
8. Beutler T.C., Mark A.E., van Schaik R.C., Gerber P.R., van Gunsteren W.F. Avoiding singularities and numerical instabilities in free energy calculation based on molecular simulations. *Chem. Phys. Lett.* 1994. V. 222. P. 529–539.
9. Zacharias M., Straatsma T.P., McCammon J.A. Separation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamics integration. *J. Phys. Chem.* 1994. V. 100. № 12. P. 9025–9031.
10. Okamoto Y. Generalized-ensemble algorithms: Enhanced sampling techniques for Monte-Carlo and molecular dynamics simulations. *J. Mol. Graph. Model.* 2004. V. 22. P. 425–439.
11. Woods C.J., Essex J.W., King M.A. Enhanced configurational sampling in binding free energy calculations. *J. Chem. Phys. B.* 2003. V. 107. P. 13711–13718.
12. Christ C., van Gunsteren W.F. Enveloping Distribution Sampling: A method to calculate free energy differenced from a single simulation. *J. Chem. Phys.* 2007. V. 126. P. 185110.
13. Hess B., Kutzner C., van der Spoel D., Lindahl E. Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* 2008. V. 4. № 3. P. 435–447.
14. Case D.A., Cheatham T.E. III, Darden T., Gohlke H., Luo R., Merz K. M. Jr., Onufriev A., Simmerling C., Wang B., Woods R. The Amber biomolecular simulation programs. *J. Comp. Chem.* 2005. V. 26. P. 1668–1688.
15. Phillips J.C., Braun R., Wang W., Gumbart J., Tajkhorshid E., Villa E., Chipot C., Skeel R. D., Kale L., Schulten K. Scalable molecular dynamics with NAMD. *J. Comp. Chem.* 2005. V. 26. P. 1781–1802.
16. Chirlian L.E., Francl M.M. Atomic Charges Derived from Electrostatic Potentials: A Detailed Study. *J. Comp. Chem.* 1987. V. 8. P. 894–903.
17. Cho A.E., Guallar V., Berne B.J., Friesner R.A. Importance of accurate charges in molecular docking: Quantum mechanical/molecular mechanical (QM/MM) approach. *J. Comp. Chem.* 2005. V. 26. № 9. P. 915–931.
18. Sanderson R.T. An Interpretation of Bond Lengths and a Classification of Bonds. *Science.* 1951. V. 114. № 2973. P. 670–672.
19. Fomin E.S., Alemasov N.A., Chirtsov A.S., Fomin A.E. MOLKERN: A Library of Software Components for Molecular Modeling Programs. *Biophysics (rus).* 2006. V. 51. № 7. P. 110–113.
20. Ponder J. W., Case D.A. Force Fields for Protein Simulations. In: *Protein Simulations. Advances in Protein Chemistry.* V. 66. Elsevier, 2003. P. 27–85.
21. Yang Y.T., Xu C.X. A compact limited memory method for large scale unconstrained optimization. *European Journal of Operational Research.* 2006. V. 180. № 11. P. 48–56.
22. Gibson K.D., Scheraga H.A. Surface area of the intersection of three spheres with unequal radii A simplified analytical formula. *J. Mol. Phys.* 1988. V. 64. P. 641–644.
23. Sigalov G., Fenley A., Onufriev A. Analytical electrostatics for biomolecules: beyond the generalized Born approximation. *J. Chem Phys.* 2006. V. 124. № 12. P. 124902.
24. Frenkel D. Free energy computation and first-order phase transitions. In: *Molecular-dynamics simulation of statistical-mechanical systems: proceedings of the 97th international school of physics "Enrico Fermi".* 1986. P. 151–188.
25. Фейнман Р., Лейтон Р., Сэндс М. Фейнмановские лекции по физике. Т. 1. М: Мир, 1967.
26. Berendsen, H.J. C., Postma J.P.M., van Gunsteren W.F., Dinola A., Haak J.R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 1984. V. 81. № 8. P. 3684–3690.

27. Fomin E.S. Consideration of Data Load Time on Modern Processors for the Verlet Table and Linked Cell Algorithms. *J. Comput. Chem.* 2011. V. 32. № 7. P. 1386–1399.
28. Kirkwood J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* 1935. V. 3. № 5. P. 300–313.
29. Фомин Э.С., Васенин А.Е. Net-q модель учета переноса заряда и поляризации в молекулярной динамике. *Вестник НГУ: Математика, механика и информатика.* 2012. в печати.
30. Rappe A.K., Goddard W.A. Charge equilibration for molecular dynamics simulations. *J. Phys. Chem.* 1991. V. 95. № 8. P. 3358–3363.
31. Andersen H.C. Molecular dynamics at constant pressure and/or temperature. *J. Chem. Phys.* 1980. V. 72. P. 2384–2393.
32. Chen J., Martinez T.J. QTPIE: Charge transfer with polarization current equalization. A fluctuating charge model with correct asymptotics. *Chem. Phys. Lett.* 2007. V.438. № 4–6. P. 315–320.
33. van Belle D., Lippens M.F.G., Wodak S.J. Molecular Dynamics Simulation of Polarizable Water by an Extended Langrangian Method. *Mol. Phys.* 1992. V. 77. P. 239–255.
34. Halley J.W., Rustad J.R., Rahman A. A Polarizable, Dissociating Molecular Dynamics Model for Liquid Water. *J. Chem. Phys.* 1993. V. 98. P. 4110–4119.
35. Rosen N. Calculation of Interaction between Atoms with S-Electrons. *Phys. Rev.* 1931 V. 38. P. 255–276.
36. Martin C., Richard V., Salem M., Hartley R., Mauguen Y. Refinement and structural analysis of barnase at 1.5Å resolution. *Acta Crystallographica Section D.* 1999. V. 55. № 2. P. 386–398.
37. Kumar M.D., Bava K.A., Gromiha M.M., Prabakaran P., Kitajima K., Uedaira H., Sarai A. {ProTherm} and {ProNIT}: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.* 2006. V. 34. P. D204–D206.
38. Seeliger D., de Groot B.L. Protein Thermostability Calculations Using Alchemical Free Energy Simulations. *Biophysical Journal.* 2010. V. 98. № 10. P. 2309–2316.
39. Фомин Э.С., Алемасов Н.А. Исследование термостабильности мутантных форм белка барназы на основе программного комплекса MOLKERN. *Вавиловский журнал генетики и селекции.* 2012. Т. 16. № 2. С. 415–426.
40. Gonnet P. A simple algorithm to accelerate the computation of non-bonded interactions in cell-based molecular dynamics simulations. *J. Comp. Chem.* 2007. V. 28. № 2. P. 570–573.

Материал поступил в редакцию 16.07.2012, опубликован 26.07.2012.