МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ имени М.В. ЛОМОНОСОВА

ФИЗИЧЕСКИЙ ФАКУЛЬТЕТ

Курсовая работа по дисциплине
“Параллельное программирование”

Комплексный анализатор результатов DPD-моделирования гомополимера с насыщающимися взаимодействиями

Выполнил
студент 2 курса группы 213
Петров Артём Игоревич
27 апреля 2017

Научный руководитель
Чертович Александр Викторович
27 апреля 2017

Москва, 2017 г.

Оглавление

Аннотация

В данной работе проводится моделирование поведения гомополимерной цепи под действием насыщающихся взаимодействий методом диссипативной динамики частиц (DPD) и обработка результатов моделирования трёхмодульным обработчиком с последовательным выполнением модулей программы. Теоретически обосновывается выбор способа распараллеливания и проверка эффективности подхода на разном количестве обрабатываемых файлов.

Введение

Одной из центральных проблем современной физики биополимеров является проблема описания пространственной укладки хроматина в ядре клетки. Хроматин – это вещество хромосом, являющееся комплексом ДНК, РНК и белков. Известно, что хроматин устроен иерархически, образуя различные пространственные структуры на разных масштабах. На масштабе нескольких пар оснований хроматин – это комплекс ДНК и восьми белков-гистонов, называемый нуклеосомой. Нуклеосома представляет собой компактную структуру, которую схематично можно представить, как шарик из белков, обвитый нитью ДНК. Также известно, что нуклеосомы распределены равномерно по цепи ДНК, что позволяет естественно отождествить хроматин с нитью, на которую «нанизаны» бусы нуклеосом. На большем масштабе образуются структуры, называемые топологически-ассоциированными доменами (TAD – Topologically Associating Domains), представляющие собой клубки из ДНК, скреплённые внутренними связями. Между TAD-ами существуют промежутки, называемые inter-TAD областями. Считается, что активные гены находятся либо в промежутках между TAD-ами, либо на их поверхности. Неактивные гены же находятся внутри этих структур [1]. Так, по современным представлениям, осуществляется регуляция работы генов. То же, как именно образуются TAD-ы, до сих пор неясно. Существуют несколько точек зрения на эту проблему. Одна из них, например, говорит о том, что за образование TAD-ов отвечают специальные белки: когезин, CTCF и другие. Другая же гипотеза говорит о том, что хроматин – саморегулирующаяся структура, и TAD-ы образуются в результате попарного взаимодействия нуклеосом. В пользу и против первой гипотезы существуют экспериментальные факты [1].

Для разработки физической модели саморегулирующегося образования TAD-ов вспомним аналогию ДНК с нитью и «бусами» нуклеосом. Физическая модель ДНК – сферы-мономеры в потенциале, имитирующем факт связанности нуклеосом нитью ДНК. Попарные взаимодействия нуклеосом в модели будем рассматривать как возможность с какой-то вероятностью образовать связь между сферами и с какой-то вероятностью её разорвать. Также нужно учесть то, что некоторые звенья не могут образовывать связь: таким образом будет построена картина с TAD-ами, разделёнными промежутками активных генов. Таким образом, рассматриваемая физическая модель хроматина – мономеры двух типов на «нитке»; мономеры одного типа не могут образовывать связь, мономеры второго типа могут образовывать связь попарно друг с другом с какой-то вероятностью и с какой-то вероятностью её разрывать. Мономеры двух типов расположены на «нитке» с заданной периодичностью. В работе [1] показано, что такая система образует клубки, разделённые промежутками; такая структура напоминает реальный хроматин.

Однако, остаётся непонятным физическое поведение такой модели. Как размер такой цепи зависит от вероятности связывания/разрыва? Как происходит переход из разных состояний цепи друг в друга, являются ли эти переходы фазовыми? Как математически

описать зависимость параметров цепи от задаваемых параметров модели? Чтобы ответить на этот сложный вопрос, в этой работе рассматривалась система, где присутствуют только звенья, способные образовать/разорвать связь, другими словами, рассматривался гомополимер с введённым вышеописанным взаимодействием звеньев.

Теория

Самым изученным видом взаимодействия мономеров в полимерной цепи остаются объёмные взаимодействия. Если представлять мономер, как шарик с конечным объёмом, то сближение центров мономеров ближе, чем на диаметр шарика, запрещено. Это приводит к различным эффектам, и, в первую очередь, к появлению трёх состояний цепи: разбухшего, идеального и глобулярного. Если подобрать растворитель такой, что его молекулы будут больше, чем шарики мономеров, то возникнет осмотическое давление, шарики будут стремиться притянуться друг к другу, система уплотнится. Такое состояние цепи называется глобулярным. Если растворитель подобран так, что притяжение к растворителю полностью устранило возможность мономерам попарно притягиваться, то цепь находится в идеальном состоянии. В таком случае, с точностью до притяжения звеньев третьего и высших порядков, мы можем считать, что объём у звеньев отсутствует. Если же растворитель состоит из молекул, меньших, чем размер мономеров, то растворитель стремится наполнить полимер, и цепь разбухает. В этом состоянии звенья стремятся оттолкнуться друг от друга. Таким видом взаимодействий можно описывать множество реальных процессов: от Ван-дер-Ваальсового притяжения мономеров реальных полимеров, до электростатического взаимодействия заряженных аминокислот в белках. Насыщающиеся же взаимодействия, описанные в предыдущем пункте, качественно иные. Взаимодействие звеньев в них не сводится к притяжению или отталкиванию звеньев, а заключаются в обратимом образовании связей. У звеньев также присутствует объём, то есть, они могут взаимодействовать ещё и объёмно. Похожую систему в приближении самосогласованного поля рассматривали Гросберг, Лифшиц и Хохлов [2]. Однако, в их исследовании звенья обратимо слипались, и систему можно было анализировать в терминах смеси газов разорванных звеньев. В данной же работе между мономерами образуются связи, ничем не отличающиеся от связей вдоль по цепи.

Теоретически полимерные цепи в различных состояниях можно описывать различными методами. Исторически первый метод описания разбухшего состояния цепи состоял в отождествлении этого состояния с хорошо изученным фазовым переходом парамагнетик-ферромагнетик. Для изучения идеальной цепи применяется описание цепи в терминах теории возмущений и вириального разложения (разложения термодинамических потенциалов в ряд по степеням плотности). Глобулярное же состояние традиционно описывается методами теории случайных блужданий во внешнем поле в самосогласованном приближении. Многообещающим общим методом описания является метод ренормализационной группы в реальном пространстве (RSRG – Real-Space Renormalization Group). Детальное описание методов здесь нецелесообразно, поскольку они не используются в численном моделировании и обработке результатов моделирования.

Численные методы

Для моделирования поведения цепи использовался метод молекулярной динамики DPD (Dissipative Particle Dynamics, диссипативной динамики частиц) [3]. Как и в типичной молекулярной динамике, для каждой частицы записываются законы Ньютона:

$\frac{d\vec{r}_i}{dt} = \vec{v}_i, \frac{d\vec{v}_i}{dt} = \vec{f}_i$. Суть метода заключается в определении сил, действующих на частицы. На каждую частицу действует сумма трёх сил: $\vec{f}_i = \sum_{j \neq i} F_{ij}^C + F_{ij}^D + F_{ij}^R$. Сила с индексом «C» - консервативная часть силы, задающаяся следующим образом: $F_{ij}^C = \begin{cases} a_{ij}(1 - r_{ij})\hat{\vec{r}}_{ij}, (r_{ij} < 1) \\ 0, (r_{ij} \geq 1) \end{cases}$. Она реализует мягкое отталкивание: сферы не взаимодействуют при расстоянии большем, чем диаметр сфер, и отталкиваются, если сферы входят друг в друга. Здесь a_{ij} - максимальное отталкивание частиц i и j, r_{ij} - расстояние между центрами i-того и j-того звеньев, $\hat{r}_{ij}$ - единичный вектор, направленный вдоль прямой, соединяющей центры частиц. Это позволяет более быстро моделировать процессы, происходящие на мезоскопическом масштабе, то есть, на масштабе скоплений атомов. В данной работе за мономер принимается нуклеосома, являющаяся комплексом множества атомов. Моделировать систему с атомным разрешением бессмысленно для данной задачи, и, поэтому, был выбран мезоскопический метод. Другим преимуществом данного метода является возможность правильно воспроизводить поведение цепи в растворе благодаря введению диссипативной и случайной сил: $\mathrm{F}_{\mathrm{ij}}^{\mathrm{D}} = -\gamma w^D(r_{ij})(\vec{v}_{ij}\hat{\vec{r}}_{ij})\hat{\vec{r}}_{ij}$, $\mathrm{F}_{\mathrm{ij}}^{\mathrm{R}} = \sigma w^R(r_{ij})\theta_{ij}\hat{\vec{r}}_{ij}$. Здесь w^D, w^R - зависящие от r весовые функции, обнуляющиеся при $r > r_c = 1$, θ_{ij} – случайно флуктуирующая величина с Гауссовой статистикой. γ и σ – амплитуды сил.

Для обработки результатов моделирования в данной работе использовалась программа, результатом работы которой являются характеристики смоделированных структур. Этими характеристиками являются среднеквадратичный радиус цепи, диагональные компоненты тензора инерции, а также корреляционные функции, о которых пойдёт речь ниже. Для выполнения этих задач программа спроектирована следующим образом: первая часть программы (далее: первый модуль) обрабатывает выходные файлы моделирования, в которых записаны начальные условия моделирования, координаты частиц и информация о связях, записывая координаты частицы в вектор класса, описывающий одну частицу («bead»), и связи в другой вектор класса, описывающий связь между частицами («spring»). Всю структуру описывает класс whole_system, который, фактически, является выходным результатом первого модуля. Вторая часть программы осуществляет расчёт среднеквадратичного радиуса инерции полимера и главных компонент тензора инерции на основе координат частиц из первого модуля. Третья часть рассчитывает две корреляционные функции. Эта часть работает с информацией о связях, также поступившей из первого модуля в классе whole_system. Работа программы схематично изображена на блок-схеме.

Вычисление двух корреляционных функций

Для вычисления гидродинамического радиуса цепи была реализована следующая процедура. Сначала проводилась сортировка массива координат, после которой координаты мономеров в цепи оказывались сверху. Далее значения координат суммировались (по каждой координате), после чего суммы делились на число мономеров. Так определились три координаты центра масс молекулы. Далее массив координат проходился ещё раз, и происходило суммирование квадратов разностей координат мономеров и координат центра масс. Разделив сумму на количество звеньев и взяв корень, получим гидродинамический радиус цепи.

Вычисление тензора инерции происходило с использованием библиотеки линейной алгебры alglib. С известными координатами центра масс по известным формулам происходило вычисление компонент тензора инерции, после чего происходила диагонализация тензора методами библиотеки.

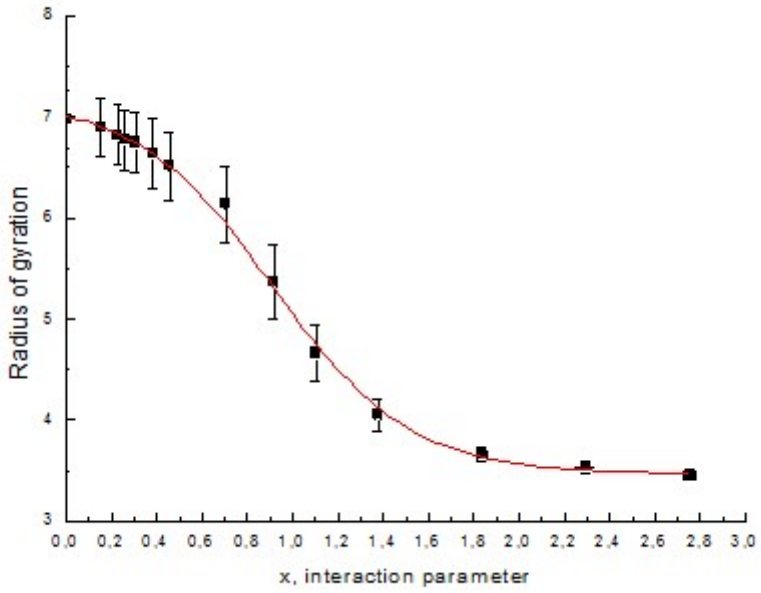
В последнем модуле программы проводится вычисление корреляционных функций. Первая функция – это вероятность того, что если в данном полимере образовал связь один мономер, то соседний с ним также образует связь и вторая – вероятность того, что если в данном полимере образовал связь один мономер, то оба соседних звена образовали связь. Для вычисления этих значений для каждого файла был реализован следующий алгоритм. Используя информацию о связях, строится карта контактов звеньев – двумерный массив с 0 и 1, где 1 показывает, что образовалась связь между i и j звеньями. Далее происходит полное считывание массива, где определяется количество «объединений» звеньев: то есть, последовательности образовавших связь и находящихся по соседству вдоль по цепи звеньев. После этого происходит отсев только «пар» и «троек» из этих последовательностей, после чего количество тех и других делится на число звеньев в цепи: следовательно, для каждого файла мы определили два числа, доли «пар» и «троек» от общего числа звеньев в цепи, которые характеризуют степень скоррелированности образования насыщающихся связей. Эта часть является наиболее ресурсоёмкой последовательной частью кода из-за большого двумерного массива данных (большие длины цепей, массив порядка 10000*10000), который полностью пробегается.

В данной работе было решено распараллелить программу-обработчик, используя технологию MPI, следующим образом: каждый процессор получает свою долю файлов, вычисляемую как результат целочисленного деления количества файлов на количество процессоров +1 файл. Это обеспечивает максимально равномерную загрузку процессоров при невозможности деления количества файлов на количество процессоров без остатка; имена файлов, которые оказались при таком делении несуществующими, игнорируются. Затем, каждый процессор выполняет вторую и третью части программы над своими файлами, записывая каждое числовое значение в отдельный файл. Когда все процессоры выполнили свою работу, параллельная часть заканчивается и процессор с рангом 0 собирает все результаты в один файл. В результате такого распараллеливания ожидается минимальное время, потерянное на простой процессоров: он технически возможен только в конце выполнения третьей части программы. Максимально равномерное статическое распределение нагрузки на процессоры также исключает возможность простоя в конце; учитывая большой объём файлов, а также малое время обработки одного файла, время простоя, неизбежно появляющееся из-за статического распределения нагрузки, не

ожидается выходящим за погрешность измерений времени исполнения программы. За счёт отсутствия обмена информацией между процессорами минимизируются потери на коммуникации. Поскольку количество обрабатываемых файлов порядка нескольких тысяч, то доля нераспараллеленных последовательных вычислений довольно мала; по закону Амдала, ожидается практически линейный рост ускорения от количества процессоров.

Результаты

Для проверки работоспособности программы был выполнен следующий тест. Широко известным результатом физики является сигмоидальная форма кривой зависимости радиуса цепи от параметра взаимодействия Флори объёмных взаимодействий [4]. Был промоделирован коллапс цепи в 1000 звеньев под действием объёмных взаимодействий, результат представлен на графике ниже.



Аппроксимация сигмоидой показала, что коэффициент корреляции R=0.982, что является прямым подтверждением работоспособности программы.

С помощью программы также были получены графики зависимости радиуса цепи от доли насыщающихся связей alpha для цепи с длиной в 1000 звеньев в разных растворителях: атермальном и растворителе, близком к тета-растворителю, но всё ещё остающимся хорошим. Графики полученных зависимостей представлены ниже.

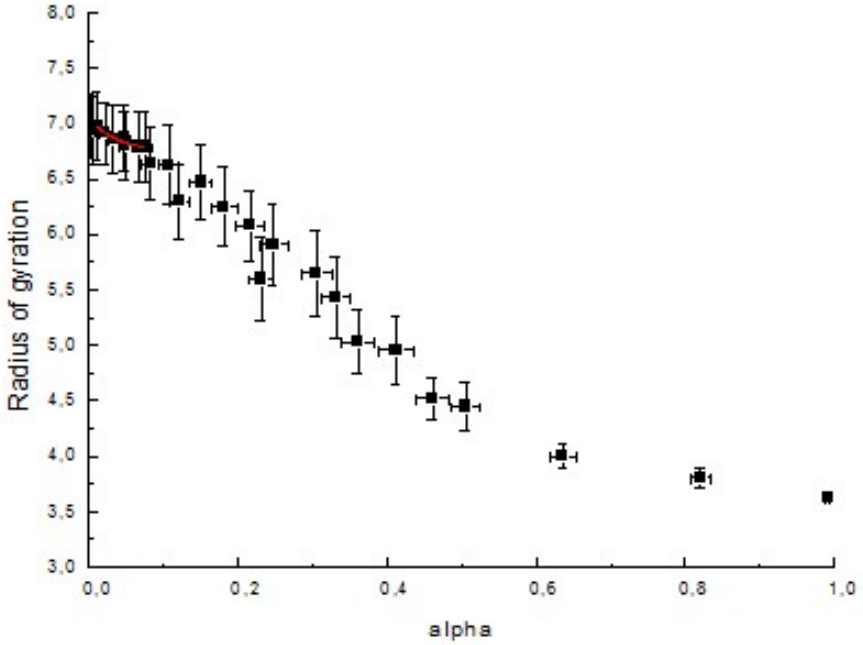


График зависимости радиуса инерции цепи от доли звеньев, участвующих в образовании насыщающихся связей (alpha). Красной линией показана аппроксимация экспонентой в области малой связанности.

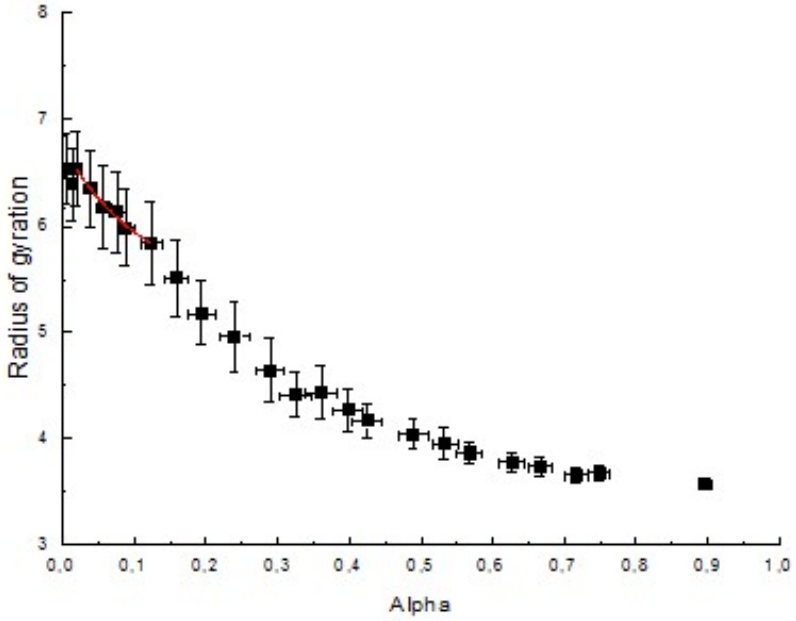


График зависимости радиуса инерции цепи от доли звеньев, участвующих в образовании насыщающихся связей (alpha). Красной линией показана аппроксимация в области малой связанности экспонентой.

Видно, что в области малого связывания, когда количество звеньев, образовавших дополнительную насыщающуюся связь, от общего числа звеньев составляет от 1 до 10 процентов, наблюдается экспоненциальное убывание радиуса инерции при увеличении alpha. Аппроксимация экспоненциальной зависимостью вида $R = Ae^{-\frac{alpha}{t}} + y_0$ даёт коэффициенты, представленные в таблице:

	A	T	y_0	R^2
χ = 0	6,76 ± 0,06	0,31 ± 0,04	0,033 ± 0,019	0.962
χ = 0.46	5,4 ± 0,4	1,4 ± 0,3	0,11 ± 0,06	0.993

Из данных видно, что данные успешно аппроксимировались экспоненциально убывающей зависимостью. При помещении цепи в более плохой растворитель экспоненциальное спадание стало более резким.

В результате моделирования коллапса под действием объёмных взаимодействий мы убедились, во-первых, что система работает правильно, и, во-вторых, процесс коллапса в области малой доли связанных звеньев качественно отличается от процесса коллапса в области малых χ. В области же средней и большой связанности, где мы можем говорить о достижении состояния идеальной цепи и глобулы, соответственно, процессы коллапса оказались качественно схожи как в случае объёмных, так и в случае насыщающихся взаимодействий. Следовательно, мы выяснили, что новым и характерным для системы с насыщающимися связями эффектом является экспоненциально быстрое убывание радиуса инерции с ростом alpha.

Для определения ускорения программы на различном количестве процессоров обработка файлов моделирования проводилась на двух выборках: 100 и 1000 файлов, цепь длиной 1000 звеньев с насыщающимися взаимодействиями. Результаты измерения коэффициентов ускорения на разном количестве процессоров представлены в виде графиков ниже:

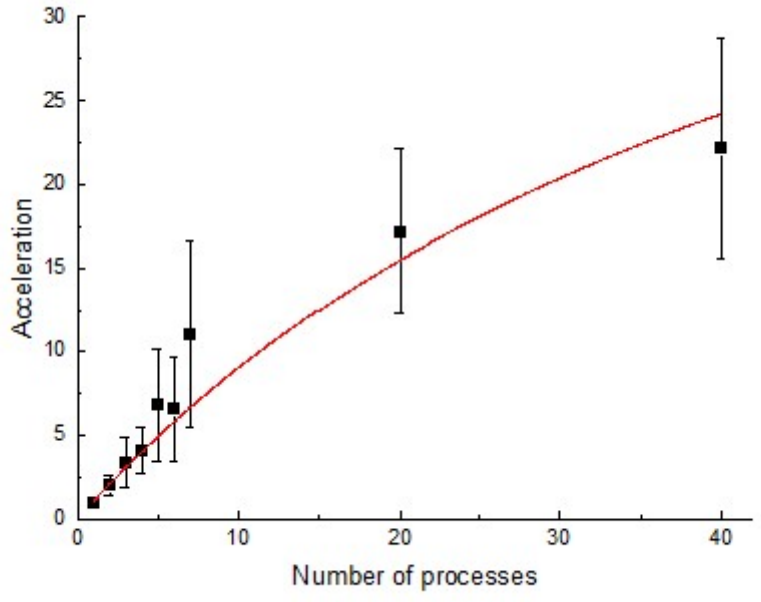


График ускорения для 100 файлов. Уже при таком количестве обрабатываемых файлов наблюдается практически линейный рост ускорения от количества процессоров: доля последовательных вычислений составила 0.018+-0.005, параллельных вычислений 0,93+-0.05.

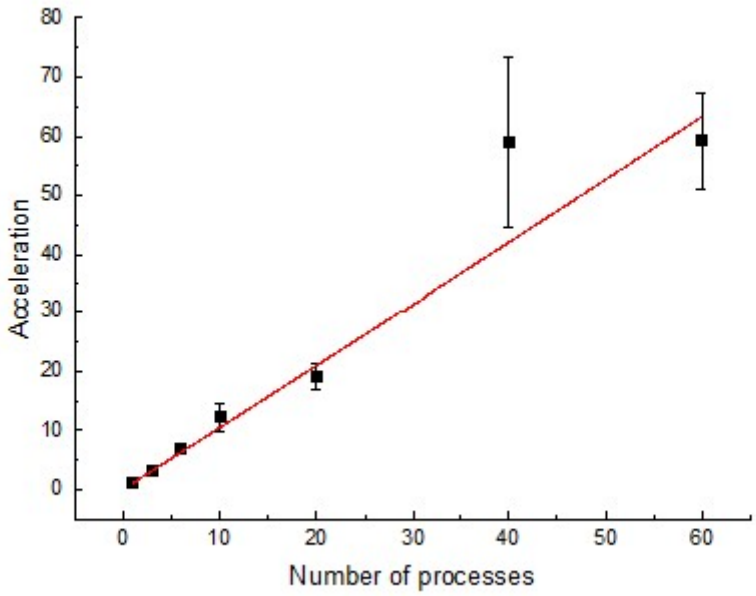


График ускорения для 1000 файлов. Продемонстрирован практически линейный рост ускорения от количества процессоров: доля последовательных вычислений составила 0.00042+-0.00027, параллельных вычислений 0,96+-0.04.

Обсуждение результатов и выводы

Результаты хорошо согласуются с законом Амдала для данной системы: на большом количестве файлов наблюдается практически линейный рост ускорения от количества процессоров. Это говорит о том, что простое статическое распределение нагрузки на процессоры в задачах обработки массива данных, где данные могут обрабатываться независимо, – наиболее оптимальное. Потери на простой и достаточно сложные последовательные вычисления на большом количестве файлов составляют сотые доли процента от общего объёма вычислительной работы.

Также интересным результатом работы является сверхлинейное ускорение в пределах погрешностей на 6 и 40 процессорах в выборке в 1000 файлов. Также программа в среднем, но не в пределах погрешностей, демонстрировала сверхлинейное ускорение на 3 и 12 процессорах. Это может быть связано с флуктуациями загруженности на суперкомпьютере: на время проведения измерений загрузка нод, где проводился расчёт, упала, и обмен данными с памятью происходил быстрее. Это могло существенно повлиять на время последовательных вычислений, где большой объём информации из обрабатываемых файлов записывается в память.

Ссылки

1) Sergey V. Ulianov, Ekaterina E. Khrameeva, Alexey A. Gavrilov et al. “Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains”, *Genome Res.*, **2016 26**: 70-84 (2015).
2) Lifshitz I.M. et al. “Structure of a polymer globule formed by saturating bonds”, *Zh. Eksp. Teor. Fiz.*, **71**, 1634-1643 (1976).
3) Robert D. Groot and Patrick B. Warren “Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation”, *The Journal of Chemical Physics*, **107**, 4423 (1997).
4) А.Ю.Гросберг, А.Р.Хохлов “Статистическая физика макромолекул”, Москва «Наука», главная редакция физико-математической литературы (1989).