Моделирование белков при помощи Blue Gene/L

Практические научные достижения, полученные при моделировании и визуализации данных на суперкомпьютере

Крис Уорд
консультант по программному обеспечению
IBM

Рухон Чжоу
сотрудник исследовательского отдела
IBM

14.12.2010

Суперкомпьютер Blue Gene®/L предоставляет ученым самые передовые вычислительные мощности и продвинутые средства визуализации данных, позволяющие вести исследования на переднем крае науки. Узнайте, как с помощью этой технологии специалисты по вычислительной молекулярной биологии моделируют правильное и неправильное сворачивание белков для улучшения понимания этих сложных молекул.

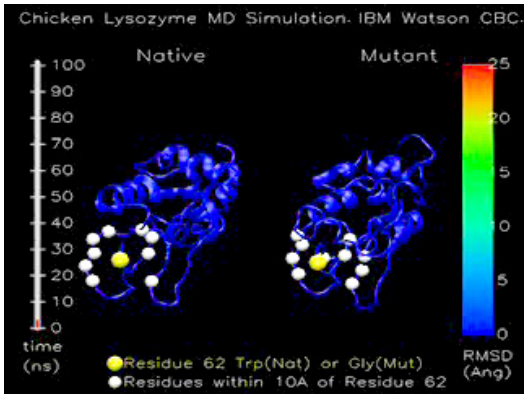
В 2001 г. исследователи из IBM начали разработку нового семейства серверов, в настоящее время поставляемых как IBM System Blue Gene®. Эти серверы были доступны для использования, начиная с 2004 г. — сначала Blue Gene/L (о котором идет речь в этой статье), а затем Blue Gene®/P.

Суперкомпьютеры семейства Blue Gene предоставляют очень высокую производительность в сочетании со стандартной средой программирования и обладают очень высокими показателями эффективности с точки зрения потребляемой мощности, охлаждения и занимаемой площади. Многие университеты, правительственные и коммерческие исследовательские лаборатории используют Blue Gene для вычислительных исследований в радиоастрономии, анализе конформаций белков, климатологии, космологии и разработке лекарств. Эта система значительно, на порядки, меняет организацию научных исследований, предоставляя экономически эффективные средства для разработки и просчета альтернативных версий сложных моделей.

В этой статье мы опишем некоторые достижения одного из проектов по моделированию пространственных конфигураций белков. На рис.1 показан масштаб работы, которую мы теперь можем производить благодаря мощи Blue Gene/L. Отправной точкой для расчета

конфигураций служит кристаллическая структура лизоцима (см. источник в разделе Ресурсы).

Рис. 1. Часть десяти микросекунд жизни белка внутри живой клетки; см. видео



Протеомика: многообразие белков

Белки — это биологические макромолекулы, являющиеся неотъемлемой частью живых организмов и участвующие во всех внутриклеточных процессах. Многие белки являются ферментами, катализирующими биохимические реакции; некоторые принимают участие в сигнальных и иммунных ответах; другие выполняют структурные и механические функции в мышцах и цитоскелетах. Следующие два примера иллюстрируют важность и распространенность белков:

- Один из белков отвечает за красный цвет крови; он переносит кислород из легких по всему телу.
- Другой белок отвечает за реакцию человеческого тела на контакт с ядовитым растением — сумахом, который вызывает сильнейшее раздражение, но, как правило, не приносит вреда.

В процесс жизни на Земле вовлечены сотни и тысячи белков. Протеомика — это наука о том, как работают белки, как они взаимодействуют и как эволюционирует их многообразие и специализация среди живых организмов вокруг нас. Эта статья представляет краткий обзор того, что представляют собой белки, как они производятся и как они затрагивают те системы, в которых существуют.

В каждой клетке каждого растения и животного имеется ДНК — структура, хранящая генетическую информацию. Информация хранится в виде последовательности химических строительных блоков (нуклеотидов), обозначаемых **A**, **C**, **T** и **G** (аденин, цитозин, тимин, гуанин в ДНК, в РНК тимин заменяется урацилом). В целом эти строительные блоки очень похожи, поэтому любой участок ДНК, на который бы вы ни посмотрели, имеет одинаковую общую структуру — знаменитую двойную спираль Уотсона-Крика.

При считывании информации из ДНК она раскручивается, и затем на основе ее внутренней структуры формируется другая молекула, называемая РНК. Вместо того, чтобы делать "слепок ключа", мы получаем «образ ключа». Эта молекула РНК далее доставляется

в качестве сборочного чертежа в рибосому — белковую структуру, которая действует как универсальная фабрика. Рибосома считывает код из букв A-C-T-G группами по три нуклеотида, что дает 64-буквенный "алфавит".

Двадцать из этих "букв" соответствуют аминокислотам — строительным блокам для белков. Эти аминокислоты в основном поступают с употребляемой нами пищей (человек не может синтезировать все необходимые аминокислоты и поэтому вынужден получать недостающие аминокислоты извне). Каждая аминокислота имеет "голову" и "хвост". Рибосома находит подходящую аминокислоту для каждой "буквы" и составляет их в последовательность голова к хвосту; другие "буквы" указывают, где начать и где закончить. Полученная линейная последовательность аминокислот представляет собой свежесозданную белковую молекулу, построенную в точности по коду, записанному в использованном участке ДНК.

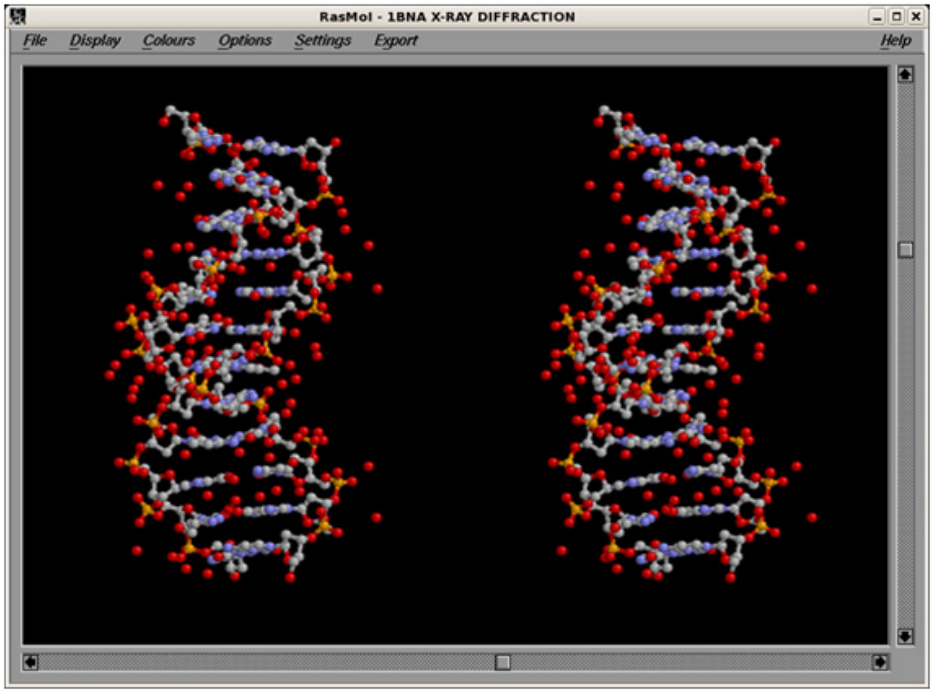
Взаимодействия между атомами внутри белковой молекулы, взаимодействие с немного подсоленной жидкостью в клетке, а также случайные колебания, которые мы бы назвали *тепловым движением*, затем вызывают "сворачивание" белковой молекулы в характерную форму.

Молекулы белков достаточно стабильны, некоторые из них могут существовать без изменений в течение сотен лет и выдерживать температуры в сотни градусов, которые убили бы организм, который их произвел. Они сохраняются в практически неизменном виде до тех пор, пока не будут разрушены сильнодействующими реагентами, высоким давлением, жарой или холодом либо став пищей для другого живого существа.

Форма молекулы и ее изменения во времени, температура, а также окружающие молекулы, определяют, что будет делать белковая молекула — будет ли она переносить кислород, вызывать у вас аллергию на сумах либо производить какие-либо другие действия, происходящие в микромасштабе.

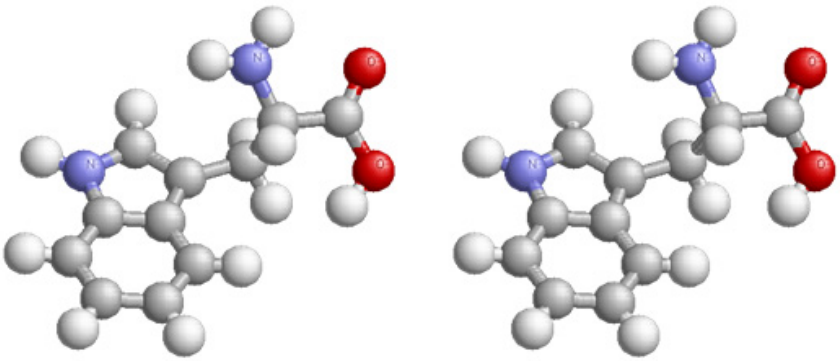
На рис. 2 показана знакомая шаростержневая модель ДНК (это стереопара; источник см. в разделе Ресурсы):

Рис. 2. Шаростержневая модель ДНК



На рис. 3 показан триптофан — одна из 20 стандартных аминокислот (это стереопара; источник см. в разделе Ресурсы).

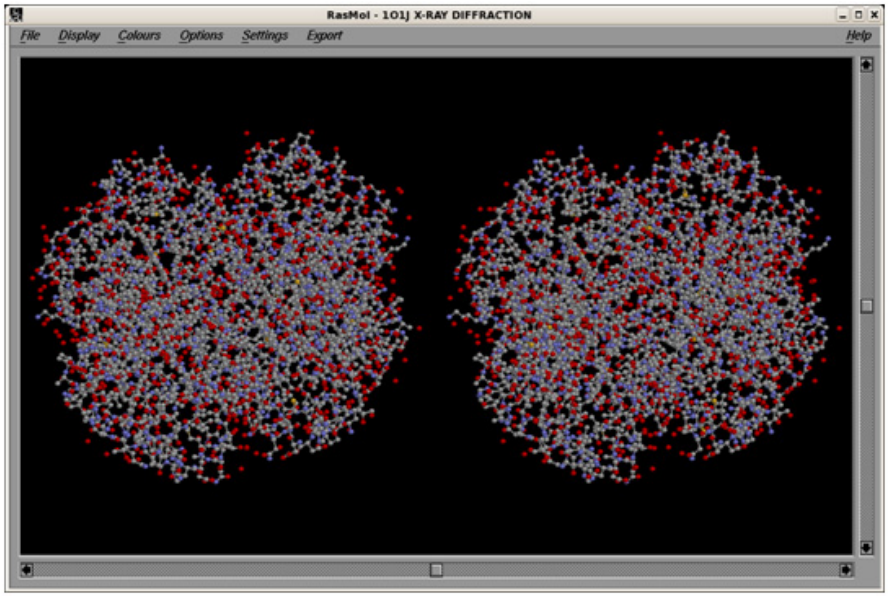
Рис. 3. Триптофан — одна из 20 стандартных аминокислот



Аминокислоты составляются в белки при отсечении группы O-H (правая часть рис. 3) одной молекулы, отсечении H от другой молекулы (верхняя часть рис. 3) и соединении полученных остатков молекул. Оставшиеся атомы H-O-H образуют молекулу воды. Все аминокислоты имеют эту характерную группу атомов (правая верхняя часть рис. 3).

На рис. 4 представлена визуализация белка гемоглобина (это стереопара; источник см. в разделе Ресурсы).

Рис. 4. Белок гемоглобина



Гемоглобин состоит из 574 молекул аминокислот в четырех субъединицах. Гемоглобин вместе с сопутствующими атомами железа (описание того, как они встраиваются в белок, находится за рамками нашей статьи) переносит кислород по току крови. Атомы железа способны переносить кислород и сами по себе, но с белковым "каркасом", который предоставляет структура гемоглобина, это происходит гораздо эффективнее. Если вы поместите это изображение в стереоскоп, то сможете разобраться в трехмерной атомной структуре; визуализация более сложных структур требует другого подхода.

Коммерческие и академические предпосылки

Что такое дикий тип?

Дикий тип — это типичная форма, которую организм, ген, линия либо некоторая характеристика принимает в природе. Применительно к фенотипу (наблюдаемым характеристикам организма, обычно отражению генов и факторов окружающей среды) характеристики дикого типа являются наиболее распространенными в естественной популяции. Применительно к генотипу (ненаблюдаемым генетическим сочетаниям) они определяют аллели в каждом локусе, необходимые для воспроизводства фенотипа дикого типа. Дикие типы не являются ни доминантными, ни рецессивными. Хорошим антонимом к термину «дикий тип» является термин *мутация*.

В последнее время достижения в разработке лекарственных препаратов и здравоохранении всё больше обеспечиваются лучшим пониманием базовых строительных блоков жизни, таких как белки. Один из текущих вопросов —*неправильное сворачивание* и *агрегация белка*: если белок сворачивается в форму, отличную от ожидаемой, то результат часто дает пассивные белки с иными свойствами, которые могут привести к нейродегенеративным заболеваниям, таким как болезнь Альцгеймера, болезнь Крейтцфельдта-Якоба, губчатая энцефалопатия крупного рогатого скота (коровье бешенство), болезнь Хантингтона, болезнь Паркинсона, кистозный фиброз и амилоидозы.

Понимание того, что может заставлять белковые молекулы изменять свою полезную свернутую форму на другую свернутую форму — предмет активных исследований способов лечения этих серьезных заболеваний. Последние эксперименты, начатые Крисом Добсоном и его сотрудниками в Кембриджском университете (см. ссылку в разделе Ресурсы) показали, что амилоиды и фибриллы могут формироваться не только из обычных бета-амилоидных пептидов, но также практически из любых белков (таких как лизоцим), если имеются подходящие условия. Действительно, одна мутация (W62A) белка лизоцима может заставить белок иметь менее стабильное состояние по сравнению с диким типом (см. выноску сбоку); она также может заставить его неправильно свернуться и образовать в растворе мочевины амилоиды благодаря потере ключевых "дальних гидрофобных взаимодействий".

Ученые до сих пор не знают, каким образом всего один остаток W62 может играть ключевую роль в дальних гидрофобных взаимодействиях в процессе сворачивания, а далее по функциональным причинам сдвигаться к поверхности из предполагаемого места нуклеации. Это предоставляет уникальную возможность лучше понять как конкретные эффекты мутации, так и механизм, стоящий за перечисленными заболеваниями, связанными с неправильным сворачиванием и агрегацией белка.

Технология Blue Gene/L предоставляет широкие возможности для изучения таких заболеваний, так как обеспечивает более экономически эффективное (и быстрое) моделирование эффектов сворачивания и неправильного сворачивания белка.

Что мы моделируем?

Видео, из которого был взят фрагмент, представленный на рис.1, является визуализацией части процесса неправильного сворачивания белка лизоцима вследствие одной мутации. Лизоцим — это белок, входящий в иммунную систему человека; при правильном функционировании он пробивает клеточную стенку внедрившейся бактерии и уничтожает ее.

Одиночная мутация, нарушающая последовательность нуклеотидов в ДНК, заставляет рибосому использовать другую аминокислоту при построении молекулы лизоцима. Согласно теории, эта другая аминокислота влияет на форму, в которую сворачивается лизоцим, и свернутая иным образом молекула лизоцима имеет другую эффективность при пробое бактериальных клеточных стенок. Поняв механизм этого изменения, мы могли бы разработать лекарственные препараты или другие виды лечения, которые помогут людям с данной конкретной мутацией излечиваться от бактериальных заболеваний.

В процессе расчета в памяти компьютера хранятся координаты и скорости каждого атома в молекуле лизоцима, а также данные примерно о 10 000 молекулах воды и мочевины (для имитации экспериментов моделирование производится для 8-молярного раствора мочевины). Существует много способов моделирования межатомных сил; мы используем вариант модели шаров и пружин для межатомных связей, а также модель обратных квадратов для электростатических сил между заряженными атомами и модель притяжения-отталкивания для атомов, которые находятся вблизи друг друга, но не являются ковалентно связанными. Расчет модели выполняется шагами по времени. На каждом временном шаге

мы рассчитываем силы, действующие на каждый атом, а затем обновляем скорости и координаты в соответствии со вторым законом Ньютона.

На каждом временном шаге (очень малом, порядка одной фемтосекунды) теоретически необходимо рассчитывать сотни миллионов воздействующих сил. В силу такого большого объёма вычислений, а также того, что для моделирования интересующих нас изменений необходимо выполнять достаточно продолжительные имитационные расчеты (порядка микросекунд), подобный подход стал практически реализуемым лишь в последнее время, даже и с самыми мощными компьютерами, которые только можно построить. Подробные сведения об альтернативных подходах приведены в статье "Destruction of long-range interactions by a single mutation in lysozyme", ссылка на которую есть в разделе Ресурсы.

Оснащение лаборатории

В исследовательской лаборатории IBM им. Т.Дж. Уотсона в Йорктауне, штат Нью Йорк, мы располагаем двадцатью серверными стойками с BlueGene/L. Каждая стойка содержит 1024 двухъядерных микропроцессора PowerPC®, и каждый микропроцессор снабжен 512 МБ оперативной памяти. Для каждых 64 микропроцессоров в этой вычислительной сети имеется дополнительный микропроцессор, подсоединенный к Ethernet-каналу 1 Гбит/с. Эти 320 Ethernet-соединений связаны обычными Ethernet-коммутаторами со стандартными машинами IBM Power с дисками, лентами, компиляторами языков и программным обеспечением по управлению задачами.

В этой работе по моделированию лизоцима для генерирования набора данных молекулярной динамики объёмом более 10 микросекунд в течение нескольких месяцев были задействованы в среднем четыре стойки процессоров BlueGene/L. Периодически приложение выполняет контрольное считывание координат и скоростей всех атомов модели (часть этого информационного потока была использована для создания приведенного выше синтезированного видео). При необходимости перезапуска моделирования подходящие координаты и скорости можно загрузить повторно. Перезапуск может потребоваться после запланированного отключения, случайного сбоя машины либо для просчета интересующего исследователей момента модели с другой величиной шага.

Запуск модели

Приложение загружается на узлы Blue Gene/L при помощи механизма, аналогичного отправке заданий MPICH (MPICH — это бесплатная и переносимая реализация интерфейса передачи сообщений MPI; см. ссылку в разделе Ресурсы). Каждый процессор в кластере предоставляет приложению среду файловой системы POSIX. Считываемые приложением данные могут быть записаны в общую параллельную файловую систему IBM (IBM General Parallel File System, GPFS); когда приложение записывает результаты, они также помещаются туда для последующего использования во внешних системах.

Для пошагового моделирования, подобного нашему, обычной практикой является считывание начальных условий из файловой системы и дальнейшая запись периодических "снимков" состояния модели в файловую систему.

Что всё это нам дает?

Приведенное видео является небольшим окошком в неизведанный мир. Конечно, мы не знаем, представляет ли оно настоящее положение дел, так как ученым всегда необходимо сравнивать то, что показывает модель, с тем, что наблюдается в реальности. Просмотр неправильного свертывания лизоцима в реальных условиях до сих пор является только мечтой; чтобы посмотреть даже часть "фиксированных" конфигураций, необходимо подготовить образцы и поместить их под электронный микроскоп или даже кристаллизовать большое количество молекул лизоцима и затем исследовать кристаллы методом рентгенодифракционной спектроскопии. Однако эти экспериментальные подходы обычно не дают представлений о том, как может двигаться белок.

Таким образом, современные крупномасштабные имитационные модели предоставляют уникальное «окно» для изучения подробностей молекулярных движений и критических изменений, происходящих при неправильном сворачивании, связанном с заболеваниями. Хочется надеяться, что доступность обеспечивающей всё это технологии расширит границы и продвинет дальше передовые достижения в исследованиях амилоидоза. Эта технология также может использоваться для подготовки следующего поколения ученых к применению описанного метода как основного инструмента решения подобных задач.

Прогнозы на будущее

Конечно, мы не берём на себя смелость предсказывать, что будет завтра, но можно предположить, что серия вычислительных машин Blue Gene будет продолжать развиваться по плану (мы использовали версию L; доступная версия Blue Gene/P имеет 4 процессора на чип, Ethernet-соединения 10 Гбит/с, а также несет множество других улучшений). Стоимость более интенсивных расчетов, а также стоимость более объемных и быстрых накопителей (то и другое в значительной мере необходимо для задач визуализации данных вроде той, что мы описали в статье) в ближайшее время скорее всего будут снижаться — как это и необходимо, поскольку существует множество областей, исследование которых требует мощных средств моделирования: и в науке, и для разработки коммерческих продуктов.

Описанная нами модель лизоцима — это лишь ничтожная часть новой области вычислительной биологии. Существует более 50 000 белков, структуры которых описаны в общедоступной базе данных белковых структур (см. ссылку в разделе Ресурсы); нужно подробно исследовать миллионы потенциально полезных лекарственных компонентов; существует большое число человеческих заболеваний, связанных с белками и изменениями в них. И это не учитывая бесчисленное множество других областей исследований, где могло бы быть полезно моделирование в таких масштабах. Работа Blue Gene только начинается.

Ресурсы

Научиться

- Оригинал статьи Protein modeling with Blue Gene/L (EN).
- Обзор исследований, связанных с Blue Gene/L, можно найти на странице проекта IBM Blue Gene. Среди других компонентов решений и ресурсов Blue Gene:(EN)
  - Страница решений IBM Blue Gene/P
  - Общая параллельная файловая система IBMОбщая параллельная файловая система IBM
  - Всё, что вы хотели знать о компиляторе IBM XL C/C++ Advanced Edition for Blue Gene
  - Документы из серии IBM Redbook для технологий Blue Gene
  - Изображение компьютера Blue Gene, на котором работал один из авторов
- База данных белковых структур в RCSB (PDB RCSB) — это архив для изучения биологических макромолекул с информацией об экспериментально определенных структурах белков, нуклеиновых кислот и сложных сборок.(EN) В числе обучающих ресурсов— такие занятные вещи, как рубрика Молекула месяца.
- Исходные данные для рис. 1 взяты из PDB, 1.33: структура тетрагонального лизоцима белка куриного яйца.
- Исходные данные для рис. 2 взяты из PDB: структура додекамера B-формы ДНК, форма и динамика.
- Рис. 3 предоставлен библиотекой MathMol, поддерживаемой в Нью-Йоркском университете.
- Исходные данные для рис. 4 взяты из PDB: дезоксигемоглобин (A-GLY-C:V1M,L29F,H58Q; B,D:V1M,L106W)..
- Группа Криса Добсона публикует дополнительные ссылки на исследования в молекулярной биологии. (EN)
- Статья "Destruction of long-range interactions by a single mutation in lysozyme" (R. Zhou, M. Eleftheriou, A. Royyuru, B. J. Berne; Proc. Natl. Acad. Sci., 2007) содержит дополнительную информацию о подходе к моделированию, использованном в данном исследовании. (EN)
- Статья "Parallel implementation of the replica exchange molecular dynamics algorithm on Blue Gene/L" (M. Eleftheriou, A. Rayshubski, J. W. Pitera, B. G. Fitch, R. Zhou, R. S. Germain; IEEE, 2006) объясняет некоторые математические приемы, использованные для моделирования.(EN)
- MPICH2— следующий уровень развития MPICH, высокопроизводительной, широко переносимой (и бесплатной) реализации стандарта интерфейса передачи сообщений (Message Passing Interface, MPI).(EN)
- У центра Argonne Leadership Computing Facility есть программа по сотрудничеству, в рамках которой время Blue Gene/P предоставляется для вычислений научному сообществу.(EN)
- Доступ к демонстрации моделирующего приложения на IBM Innovation Centers возможен из любой точки планеты. (EN)

- Серия из двух статей "High-performance Linux clustering": справочная информация о высокопроизводительных вычислениях в Linux. Первая часть (developerWorks, сентябрь 2005 г.) описывает основы HPC, доступные типы кластеров, аргументы при выборе конфигурации кластера, а также роль Linux в HPC. Во второй части (developerWorks, октябрь 2005 г.) обсуждается параллельное программирование с использованием MPI, описано управление кластерами и их испытание, а также показано, как настроить Linux-кластер с использованием открытого программного обеспечения.(EN)
- Статья "Port Fortran applications" (developerWorks, апрель 2009 г.) поможет вам преодолеть основные трудности при переносе приложений на Fortran на различные высокопроизводительные компьютерные системы.(EN)
- В разделе Linux на developerWorks вы можете найти дополнительные ресурсы для разработчиков Linux, а также просмотреть наши наиболее популярные статьи и руководства. (EN)
- Оставайтесь в курсе событий, посещая раздел технических мероприятий и Web-трансляций developerWorks. (EN)

Получить продукты и технологии

- Open Discovery— это live-дистрибутив Linux, основанный на Fedora Core и содержащий программные средства по биоинформатике, лицензированные по свободной академической лицензии (Academic Free License, AFL), которые могут работать со всеми задачами, начиная с анализа последовательностей и до задач молекулярной динамики. Дистрибутив может быть загружен с DVD или USB и обеспечивает сохранность данных. Большое спасибо Отделению биоинформатики Университета SRM, кампус Рамапурама, Ченнаи, Индия. (EN)
- Некоторые из инструментов, встроенных в приложения, описанные в этой статье, включают библиотеку трехмерного быстрого преобразования Фурье для Blue Gene/L и написанные Крисом Уордом специальные математические функции для высокопроизводительных вычислений. (EN)
- Ознакомительные версии ПО IBM: используйте в вашем следующем проекте программное обеспечение, которое можно загрузить непосредственно с сайта developerWorks.(EN)

Обсудить

- Присоединяйтесь к сообществу My developerWorks; используя свой личный профиль и специальную домашнюю страницу, вы можете настроить developerWorks в соответствии со своими интересами и общаться с другими пользователями developerWorks.(EN)

Об авторах

Крис Уорд

Крис Уорд поступил в Исследовательскую лабораторию IBM в Херсли (Англия) в 1982 г. после получения степени инженера в Кембриджском университете. Он участвовал в разработке многих проектов для IBM, от дисковых накопителей до фирменного ПО промежуточного уровня. Он имеет доступ к разработке самых важных технологий для будущих заказчиков IBM, какими являются WebSphere и IBM Lotus для сегодняшних заказчиков.

Рухон Чжоу

Рухон Чжоу — научный сотрудник Центра вычислительной биологии, Исследовательского центра IBM Томаса Дж. Уотсона, а также адъюнкт-профессор химического факультета в Колумбийском университете. Он получил степень доктора философии по химическим наукам в Колумбийском университете под руководством Брюса Берна в 1997 г. Поступил в IBM Research в 2000 г. после двух с половиной лет работы с Ричардом Фризнером (Колумбийский университет) и Уильямом Йоргенсеном (Йельский университет) в области полей поляризационных сил и механизмов белково-лигандной связи. Является автором 80 публикаций в журналах и семи патентов, выступал приглашенным докладчиком на крупных конференциях и в университетах, а также был председателем ряда конференций по вычислительной биологии, химии и биофизике. Он получил премию Хаммета за 1997 г. от Колумбийского университета, Премию DEC по вычислительной химии в 1995 г. от Американского химического общества, а также Премию за выдающиеся технические достижения в 2005 и 2008 г. от IBM. Его текущие исследовательские интересы включают разработку новых методов и алгоритмов вычислительной биологии и биоинформатики, крупномасштабное моделирование сворачивания белков, лигандно-рецепторной связи, а также прогнозирование структуры белков.