Тема работы: Разработка системы поддержки высокопроизводительных вычислений для программного комплекса “Гаплоидный эволюционный конструктор”

Состав коллектива:

- Лашин Сергей Александрович, место работы: Институт цитологии и генетики, должность: научный сотрудник, ученая степень: к.б.н., email:lashin@bionet.nsc.ru, обязанности в работе: руководитель
- Клименко Александра Игоревна, место работы: Институт цитологии и генетики, должность: старший лаборант, ученая степень: отсутствует, email:klimenko@bionet.nsc.ru, обязанности в работе: программист
- Чеканцев Антон Дмитриевич, место работы: Институт цитологии и генетики, должность: старший лаборант (студент, 6 курс), ученая степень: отсутствует, email: chekantsev@bionet.nsc.ru, обязанности в работе: программист
- Зудин Роман Константинович, место работы: Институт цитологии и генетики, должность: старший лаборант (студент, 6 курс), ученая степень: отсутствует, email: rockalocular@gmail.com, обязанности в работе: программист

Научное содержание работы:

1. **Постановка задачи:** целью данной работы является организация высокопроизводительных вычислений для программного комплекса «Гаплоидный эволюционный конструктор» для моделирования сложных пространственно-гетерогенных микробных сообществ.
2. **Современное состояние проблемы**: поскольку пространственные факторы являются одними из главных двигателей эволюции живых организмов разных уровней организации, пространственное распределение видов и популяций играет важную роль и в эволюции микробных сообществ, на которую влияют взаимодействия видов между собой и их конкуренция за необходимые ресурсы. В ряде научных исследований, как экспериментальных, так и теоретических, было показано, что в совокупности с другими эволюционными процессами, пространственное распределение влияет на динамику частот аллелей в популяциях сообщества. Также было показано, что независимо от скорости мутаций, рост приспособленности биологических сообществ в пространственно-распределенных средах был меньше, чем в неструктурированных.

   На сегодняшний день существует достаточно большое количество программ по моделированию микробных сообществ и популяций (BacSim, Micro-Gen, COSMIC, RUBAM, INDISIM и др.). Но большинство данных программных решений направлено на моделирование систем с небольшой численностью клеток, при этом такие средства

не учитывают (или учитывают не в полной мере) генетического разнообразия и генетической изменчивости.

Пространственное распределение присутствует во многих существующих решениях, однако, в ГЭК 3D размер моделируемой системы намного больше, что позволяет исследовать более сложные, с точки зрения пространственной организации, системы.

То есть, ГЭК 3D предоставляет наиболее широкий спектр моделируемых в системе процессов, с учетом большой численности клеток (свыше миллиарда) и с учетом моделирования фаговой инфекции и видообразования, что добавляет (по сравнению с другими программными решениями) дополнительной вычислительной сложности. А это, в свою очередь, приводит к тому, что отдельные модели могут считаться до нескольких суток. В связи с чем появляется необходимость в параллельных вычислениях, с целью сокращения временных затрат на моделирование подобных сложных систем.

Ссылка на сайт проекта: http://evol-constructor.bionet.nsc.ru/

Ссылка на отдельные публикации: http://evol-constructor.bionet.nsc.ru/?cat=6

3. Подробное описание работы

3.1 Общая схема работы последовательной версии ГЭК 3D

Для того чтобы реализовать параллельную версию программного комплекса ГЭК – необходимо понять, как работает исходный вариант программы (рис. 1).

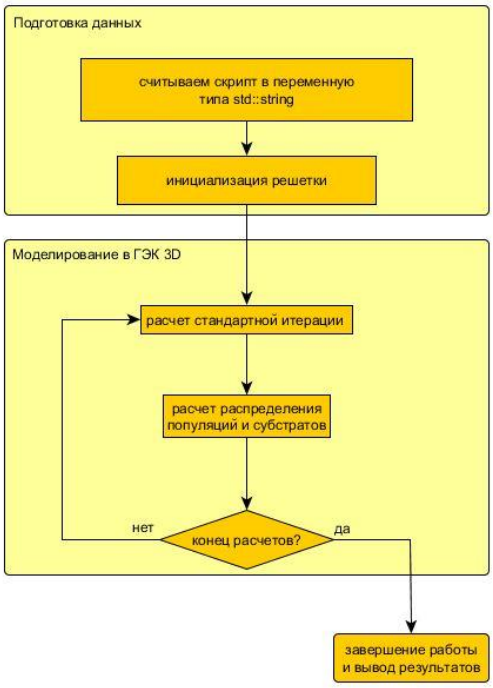

Рисунок 1 Схема работы последовательной версии ГЭК 3D.

В самом начале программы считывается скрипт, в котором задаются все необходимые параметры системы и количество итераций, необходимое для расчета модели. Далее следует инициализация данных в виде одно-, двух- или трехмерной решетки, в зависимости от заданных параметров модели и начинается вычислительная часть, которая состоит из расчета стандартной итерации (учет различных биологических процессов) и расчета перераспределения клеток и веществ под действием протока, диффузии или активного перемещения клеток.

Общая схема работы параллельной версии ГЭК 3D

В самом начале работы главный процесс считывает скрипт и рассылает информацию о нем всем остальным, после чего каждый процесс инициализирует свою область моделируемой среды и начинается расчет модели, состоящий из двух основных этапов: расчет стандартной итерации и расчет перераспределения популяций и субстратов. После этого процессы обмениваются всей необходимой для них информацией с соседними и продолжают расчеты, пока не будет достигнуто заданное в скрипте число итераций.

В конце работы каждый процесс начинает вывод информации в выходные файлы. В случае вывода индивидуальной информации по ячейкам каждый процесс просто создает свой файл и записывает в него необходимую информацию. В случае с информацией о системе в целом главный процесс собирает информацию со всех других и записывает данные о системе самостоятельно.

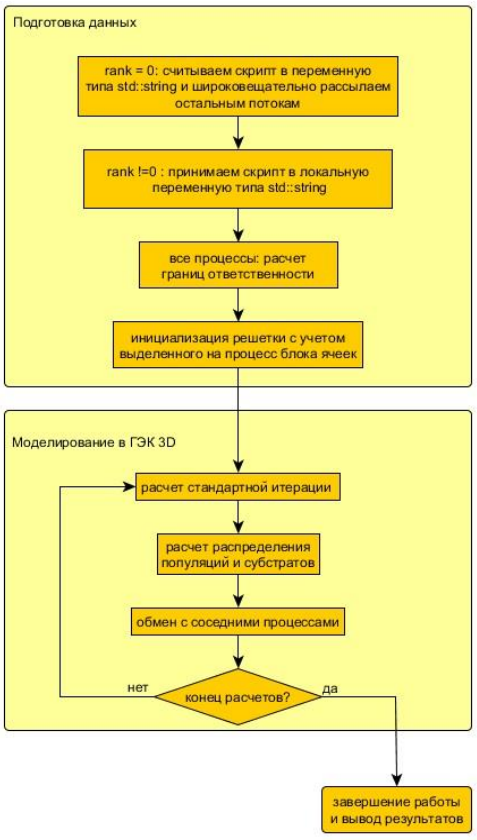

Рисунок 2 Схема работы параллельной версии ГЭК 3D.

4. Полученные результаты, иллюстрация и визуализация результатов

В ходе тестирования параллельной версии программы, реализованной с помощью MPI, были исследованы модели, описанные в таблице 1. Суммарная сложность модели рассчитывалась с учетом каждого коэффициента, указанного в таблице:

Суммарная сложность = **количество ячеек** X **количество популяций** X **Количество генов на популяцию** X **Количество аллельных вариантов на ген**.

Можно заметить, что размер простых задач варьировался от малого числа ячеек к большему при сохранении остальных параметров системы. Сложная же задача рассчитывалась только с использованием 10000 ячеек и указанных в таблице параметров.

Модель	Количество ячеек	Количество популяций	Количество субстратов	Количество генов на популяцию	Количество аллельных вариантов на ген	Суммарная сложность
Простая 1D	10 - 1000	6	7	3	1	18000
Простая 2D	4 - 900	2	3	3	1	5400
Простая 3D	27 - 1000	5	6	3	1	15000
Сложная	10000	56	4	2-6	10	33600000

Таблица 1 Описание тестируемых моделей.

Ниже приводятся сводные таблицы полученных результатов для простых (рис. 3) и сложных (рис. 4) моделей. Данные таблицы представляют не полный спектр полученных данных и содержат значения лишь для некоторого числа используемых процессов/потоков, однако из этих таблиц можно сделать определенные выводы, проанализировав их.

Проанализировав данные таблицы можно увидеть, что:

- Была достигнута эффективность до 180% с использованием небольшого числа процессов на MPI на сложной задаче.
- Было достигнуто ускорение до 40 раз с использованием большого числа процессов MPI.
- Было достигнуто ускорение до 3 раз с использованием QtConcurrent на обычном домашнем компьютере.
- Была достигнута эффективность до 50% с использованием QtConcurrent на обычном домашнем компьютере.
- На задачах с расчетом простых моделей намного эффективнее использовать SMP машину, по сравнению с вычислительным кластером с распределенной памятью.

- На задачах с расчетом сложных моделей стоит использовать высокопроизводительный кластер с распределенной памятью при наличии не более 50 процессов. Если есть возможность использования 50 процессов на SMP машине, то это будет намного эффективнее по сравнению с высокопроизводительным кластером с распределенной памятью.

	ПРОСТАЯ МОДЕЛЬ								
	1D			2D			3D		
	Количество ячеек	Количество популяций	Суммарная сложность	Количество ячеек	Количество популяций	Суммарная сложность	Количество ячеек	Количество популяций	Суммарная сложность
	1000	6	18000 - MPI 9000 - Qt	1000	2	5400 - MPI 9000 - Qt	1000	5	15000 - MPI 9000 - Qt
	эффективность, %		ускорение	эффективность, %		ускорение	эффективность, %		ускорение
MPI 10 потоков	95		9,5	31		3,1	7,3		0,73
MPI 50 потоков	73		36,5	9,2		4,6	2,8		1,4
SMP 10 потоков	110		11	54		5,4	10,6		1,06
SMP 50 потоков	116,8		58,4	21		10,5	6,12		3,06
QT 6 потоков	29,2		1,8	22,5		1,4	22,3		1,3

Рисунок 3 Сводная таблица результатов, простая модель

	СЛОЖНАЯ МОДЕЛЬ								
	1D			2D			3D		
	Количество ячеек	Количество популяций	Суммарная сложность	Количество ячеек	Количество популяций	Суммарная сложность	Количество ячеек	Количество популяций	Суммарная сложность
	10000 - MPI 1000 - Qt	56	33600000 - MPI 3360000 - Qt	10000 - MPI 1000 - Qt	56	33600000 - MPI 3360000 - Qt	10000 - MPI 1000 - Qt	56	33600000 - MPI 3360000 Qt
	эффективность, %		ускорение	эффективность, %		ускорение	эффективность, %		ускорение
MPI 4 потока	180		7,2	145		5,8	120		4,8
MPI 10 потоков	144		14,4	108		10,8	80		8
MPI 50 потоков	48		24	38,4		19,2	30,8		15,4
SMP 4 потока	75		3	60		2,4	47,5		1,9
SMP 10 потоков	66		6,6	60		6	44		4,4
SMP 50 потоков	75,4		37,7	50,2		25,1	17,6		8,8
QT 6 потоков	46,7		2,8	45,0		2,7	43,3		2,6

Рисунок 4 Сводная таблица результатов, сложная модель

Эффект от использования кластера в достижении целей работы

Кластер помог достичь достаточно хорошего ускорения исходной версии программы и сократить время, которое тратится на проведение расчетов, тем самым оставив его на анализ полученных в ходе исследований результатов.

Перечень публикаций, содержащих результаты работы

1. **Зудин Р. К.** Разработка системы поддержки высокопроизводительных вычислений для программного комплекса «Гаплоидный эволюционный конструктор» // Материалы 54-й Международной научной студенческой конференции. – Новосибирск, 2016.
2. **Зудин Р.К.** Поддержка высокопроизводительных вычислений для семейства программ «Эволюционный конструктор» // XII Международная конференция

студентов, аспирантов и молодых ученых «Перспективы развития фундаментальных наук». - Томск, 2015. – с. 692.