Центр высокопроизводительных вычислений в биоинформатике: задачи, вычислительная сложность, программное обеспечение и проблемы конфигурирования системы *

Афонников Д.А., Подколодный Н.Л., Вишневский О.В., Вяткин Ю.В., Стрижак С.В., Федорук М.П. , Чубаров Д.Л., Юдин А.Б.

Среди актуальных прикладных проблем, требующих применения параллельных вычислений одно из первых мест принадлежит биологическим задачам. Это обусловлено стремительным ростом молекулярно-биологических данных, полученных в ходе выполнения проектов по секвенированию геномов, а также необходимостью рассчитывать сложные модели биологических объектов – макромолекул и лекарственных препаратов, генных сетей.

В биоинформатике одними из важнейших являются задачи ассемблирования геномов, сравнения геномных последовательностей, моделирования структуры генетических макромолекул, филогенетический анализ. Необходимо отметить, что биологов интересует комплексный анализ генома, который может быть реализован только в результате использования большого числа биоинформационных методов и программ. Поэтому создание специализированных высокопроизводительных центров обработки данных (ЦОД) в области биоинформационного анализа, которые бы позволяли эффективно интегрировать хранение биологических данных и их высокопроизводительную обработку является актуальной задачей.

Одной из эффективных технологий создания современных ЦОД являются блейд-системы. В работе будут представлены предварительные результаты работы по созданию и тестированию элементов ЦОД на базе кластера Информационно-вычислительного центра Новосибирского государственного университета. Предложен состав программного обеспечения и проведена установка ряда программ, связанных с решением задач биоинформатики. К числу этих программ относятся: ClustalW (выравнивание последовательностей), BLAST (поиск родственных последовательностей), HMMER (идентификация функции белка по последовательности), Celera Assembler (сборка геномных последовательностей), RAxML (построение филогенетического дерева), Gromacs (моделирование пространственной структуры белка), Autodoc (поиск участков взаимодействия белка с лигандом). Предложена схема тестирования производительности этих программ в зависимости от параметров системы и настроек оборудования. Получены предварительные результаты, которые могут быть полезными при создании и оптимизации настроек ЦОД для решения задач биоинформатики.

* Работа поддержана интеграционными проектами СО РАН № 26 и № 113.