

Поиск потенциальных возможностей распараллеливания алгоритма множественного выравнивания нуклеотидных и белковых последовательностей ClustalW2

И.С. Пироженко, А.Н. Сальников

МГУ им. М.В. Ломоносова

1. Введение

Одним из краеугольных камней биоинформатики является сравнение или выравнивание нуклеотидных и белковых последовательностей. Множественное выравнивание биологи используют для поиска диагностических моделей, установления семейств белков, нахождения гомологии между новыми и уже существующими последовательностями, как необходимую часть более сложных анализов. Таким образом, задача создания эффективного и универсального алгоритма множественного выравнивания является важной составляющей развития биологии.

Существует довольно много последовательных алгоритмов множественного выравнивания, таких как: T-Coffee, MUSCLE, MAFFT, ClustalW2. В статье используется ClustalW2 ввиду того что он: наиболее широко используется биологами по всему миру; надежен; обладает простым и понятным C++ кодом, в отличие от конкурентов; работает на всех основных платформах; выдает точные результаты [1].

Однако алгоритм ClustalW2 не имеет параллельной реализации. Активное внедрение многопроцессорных и многоядерных архитектур дает возможность значительно сократить время счета программы, тем самым ускорив решение биологических задач. В статье производится поиск возможностей по распараллеливанию алгоритма ClustalW2, дается теоретическая оценка выигрыша производительности.

2. Алгоритм работы ClustalW2

Алгоритм множественного выравнивания в ClustalW2 состоит из 3 частей: (1) все пары последовательностей выравниваются отдельно для подсчета матрицы расстояний отображающей дивергенцию между последовательностями; (2) по матрице расстояний строится филогенетическое дерево, для отображения эволюционных взаимосвязей между последовательностями; (3) по полученному дереву формируется окончательная последовательность, используется прогрессивное выравнивание с различными профилями [3].

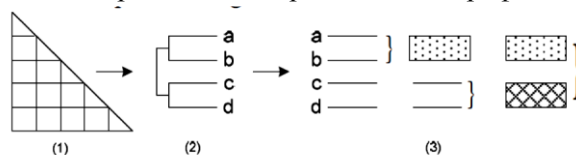


Рис 1. Три этапа работы алгоритма ClustalW2: (1) матрица расстояний; (2) филогенетическое дерево; (3) прогрессивное выравнивание по филогенетическому дереву

3. Метод исследования

Для того чтобы определить потенциальные возможности распараллеливания алгоритма ClustalW2, необходимо определить узкие места в его последовательной версии. Такие работы уже проводились, но для ClustalW версии 1.82. Однако, ClustalW2 содержит много нововведений. Анализы производительности алгоритма на последовательностях из бенчмарков BAliBase, OxBench и базы Pfam были выполнены заново. Для теоретической оценки выигрыша производительности используется закон Амдала.

4. Поиск возможностей распараллеливания алгоритма ClustalW2

4.1 Профайлинг

Чтобы найти узкое место алгоритма ClustalW2 воспользуемся профайлером gprof. Стоит отметить, что в ClustalW2 существует два метода попарного выравнивания: *fast pairwise alignment* и *full pairwise alignment*. В этой статье использовался метод *full pairwise alignment*, т.к. именно он чаще встречается в реальных задачах. Результаты профайлинга на семействе белков из базы Pfam¹ приведены в Таблице 1.

Время в % от общего времени выполнения программы	Время выполнения в сек.	Количество вызовов	Имя функции
29.04	4635.12	510555	FullPairwiseAlign::forwardPass
23.43	3738.76	510555	FullPairwiseAlign::diff
16.32	2603.98	510555	FullPairwiseAlign::reversePass
6.16	983.27	1553007499	FullPairwiseAlign::calcScore

Таблица 2. Результаты выполнения профайлинга ClustalW2 на последовательности Pfam

По результатам профайлинга стало ясно, что больше всего времени программа проводит в первой части алгоритма множественного выравнивания. Следует рассмотреть более подробно этот этап вычислений.

4.1 Полное попарное выравнивание

Полное попарное выравнивание строится из следующих частей: (1) выполняется прямой ход локального выравнивания по алгоритму Смита-Ватермана (Smith - Watermann); (2) выполняется обратный ход; (3) применяется алгоритм Майерса и Миллера (Myers & Miller) для подсчета штрафов за пропуски. Все три этапа могут выполняться независимо для разных пар последовательностей, следовательно, можно организовать параллельную работу первой части алгоритма ClustalW2 [2].

4.2 Теоретическая оценка выигрыша производительности

Очевидно, что доля полного парного выравнивания в алгоритме ClustalW2 занимает около 60%. Следовательно, по закону Амдала максимальный теоретический выигрыш в производительности на 1000 процессорной системе составит 2.5 раза.

5. Заключение и дальнейшие планы

На основе полученных данных, можно сделать вывод о необходимости реализации параллельной версии алгоритма ClustalW2, она сможет увеличить скорость счета в разы. Разработка параллельной реализации, а также анализ второго и третьего шагов алгоритма ClustalW2 часть будущей работы.

Литература

1. Kuo-Bin Li. ClustalW-MPI: ClustalW analysis using distributed and parallel computing. //Bioinformatics Vol. 19, No. 12, 2003, pp. 1585-1586.
2. Yongchao Liu, Bertil Schmidt, Douglas L. Maskell. Parallel Reconstruction of Neighbor-Joining Trees for Large Multiple Sequence Alignments using CUDA. //IEEE Computer Society, pp. 1 - 8.
3. M.A. Larkin, G. Blackshields, N.P. Brown. Clustal W and Clustal X version 2.0. //Bioinformatics Vol. 23, No. 21, 2007, pp. 2947-2948.

¹ Результаты профайлинга на последовательностях из бенчмарков BAliBase, OxBench находятся по адресу <http://dl.dropbox.com/u/12514653/publications/Summary.xlsx>