КОМПАКТНЫЙ ФОРМАТ ХРАНЕНИЯ НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Г.В. Раманчаускайте

Южный федеральный университет, факультет математики, механики и компьютерных наук, студентка. E-mail: galinka@lastbit.com

Стандартным форматом хранения нуклеотидных последовательностей в электронном виде является FASTA [1]. В данном формате нуклеотиды и неопределённости, возникающие при секвенировании, кодируются буквой. В данной работе предлагается более компактный формат – binFASTA, где в каждый байт кодируется не один символ, а два, что осуществляется при помощи побитовой работы с памятью.

Для оценки эффективности работы рассматривалась следующая задача. Заданы последовательность, шаблон (последовательность меньшей длины) и натурально число d. Требуется в исходной последовательности найти все подпоследовательности, которые не совпадают с шаблонам не более чем в d символах. Задача была решена для обоих форматов. В качестве основной последовательности была взята хромосома крысы [2], длина которой составляет – 247 199 719 символов. Количество допустимых несовпадений d = 100. Результаты измерений времени работы программы приведены в таблице 1.

Таблица 1.

	Время (сек)	
Длина шаблона	FASTA	binFASTA
100 000	375	199
300 000	379	194
1 000 000	378	189
2 000 000	375	140
6 500 000	377	127

За счёт компактного хранения данных в кэш памяти размещается большее количество символов и сокращается количество обращений к оперативной памяти, что не только компенсирует затраты на распаковку, но и значительно ускоряет работу. Данный формат актуально использовать для решения задач, требующих больших объёмов памяти и вычислений.

Литература

[1] - http://en.wikipedia.org/wiki/Fasta_format

[2] - ftp://ftp.ncbi.nih.gov/genomes/R_norvegicus/CHR_01/