

Применение графических ускорителей  
для выявления вырожденных олигонуклеотидных  
мотивов в регуляторных районах генов эукариот  
Application of the graphics accelerators for detection  
of degenerate oligonucleotide motifs in the regulatory  
regions of eukaryotic genes

Вишневский О. В.<sup>1</sup>, Лаврентьев М. М.<sup>2</sup>, Романенко А. А.<sup>3</sup>

<sup>1</sup>Институт цитологии и генетики СО РАН, Новосибирский  
государственный университет, Новосибирск, Россия;  
oleg@bionet.nsc.ru

<sup>2</sup>Институт математики им. С. Л. Соболева СО РАН,  
Новосибирский государственный университет, Новосибирск, Россия;  
mmlavr@nsu.ru

<sup>3</sup>Новосибирский государственный университет, Новосибирск, Россия;  
arom@ccfit.nsu.ru

Сборка базального транскрипционного комплекса, ткане- и стадие-специфические особенности транскрипции эукариотических генов зависят от контекстной и структурной организации корового промотора и присутствия в 5'-регуляторном районе гена сайтов связывания транскрипционных факторов. Сайты связывания представляют собой короткие (8–12 пар оснований) участки ДНК, распознаваемые белками — транскрипционными факторами. Связывание транскрипционных факторов с соответствующими сайтами является обязательным условием формирования преинициационного комплекса и инициации транскрипции. Обнаружение сайтов связывания в регуляторных районах генов необходимо как для понимания механизмов регуляции экспрессии генетической информации и выявления структурно-функциональной организации регуляторных районов, так и для разработки методов распознавания регуляторных районов генов в геномных последовательностях. Решение этой задачи осложняется тем, что каждый транскрипционный фактор может связываться с участками ДНК, существенно различающимися по нуклеотидному контексту.

Нами предложен метод выявления вырожденных район-специфичных олигонуклеотидных мотивов — коротких слов фиксированной длины, записанных в 15-буквенном IUPAC коде (A, T, G, C, R=G/A, Y=T/C,

M=A/C, K=G/T, W=A/T, S=G/C, B=T/G/C, V=A/G/C, H=A/T/C, D=A/T/G, N=A/T/G/C), значимых для структурно-функциональной организации регуляторных районов генов и играющих важную роль в регуляции экспрессии генов [1]. Функционально значимыми считаются мотивы, характеризующиеся высокой частотой присутствия в регуляторных районах генов, по сравнению со случайными последовательностями и высокой статистической значимостью, оцениваемой с использованием биномиального критерия. Получаемые таким образом мотивы могут соответствовать как сайтам связывания транскрипционных факторов, так и некоторым структурным и физико-химическим особенностям регуляторных районов.

Одним из наиболее ресурсно-затратных этапов предложенного нами алгоритма является оценка представленности каждого из вырожденных мотивов в выборке анализируемых регуляторных районов и выборке негативных последовательностей. Так, оценка свойств всех мотивов длины 8 требует рассмотрения  $15^8 \sim 2,5 \cdot 10^9$  вариантов мотивов. Для решения этой задачи мы использовали высокопроизводительные графические устройства nVidia, позволяющие проводить параллельные вычисления в многопоточном режиме, используя технологию CUDA. В таблице 1 приведены результаты оценки времени расчета представленности всех мотивов длины 8 в выборке из 1000 последовательностей длины 50 пар оснований на различных устройствах. Полученные результаты демонстрируют высокую эффективность применения графических ускорителей для решения задачи поиска вырожденных сигналов в геномных последовательностях.

Платформа	Затраченное время
GeForce 8800 GTX	9,5 часов
Tesla C1060	5,7 часов
Intel Pentium Dual CPU 1.73GHz	14 дней

Таблица 1. Временные затраты при оценке представленности всех возможных вырожденных мотивов длины 8 в выборке из 1000 последовательностей длины 50 пар оснований на различных платформах

#### ЛИТЕРАТУРА

- Vishnevsky O. V., Kolchanov N. A. ARGO: a web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters // Nucleic Acids Res. 2005. V. 33. P. 417–422.