№11 2000 год
БАЗА ДАННЫХ ТРАНСКРИБИРУЮЩИХСЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ DOTS (БД DOTS)

В настоящее время в университете Пенсильвании (Филадельфия, США, Центр биоинформатики (http://pcbi.upenn.edu)), директором которого является д-р Кристиан Овертон, ведется работа над проектом DOTS (Database on transcribed sequences, http://pcbi.upenn.edu/dots). Целью настоящего проекта является: 1) поддержка аннотации геномов, а именно оценка и пополнение информации о транскрибирующихся генах и 2) анализ совместной экспрессии генов в разных тканях и стадиях онтогенеза. Проект является на 90% инженерной задачей (интеграция баз данных, обеспечение интерфейса пользователя, разработка необходимых программных инструментов) и на 10% – научной (обнаружение открытых рамок считывания (ORF), выяснение клеточной роли того или иного гена и пр.). Таким образом, целью проекта можно считать разработку технологии по получению знаний о геноме путем интеграции всей доступной информации о мРНК.

Мотивацией подобного проекта явился проект “Геном человека”, осуществление которого началось в США в 1990 году. В настоящее время полностью отсеквенированы следующие геномы эукариот: дрожжи (1996 г.) [1], C.elegans (1998 г.) [2], к весне 2000 года планируется отсеквенировать “в основном” (на 98%) геном человека. В процессе развития проект разделился на 2 части: а) секвенирование, картирование последовательностей ДНК; б) функциональная геномика – попытка понять суть функционирования геномов на основе большого числа данных.

В настоящее время проект DOTS направлен на аннотирование последовательностей ДНК геномов мыши и человека, как наиболее полно представлен-ных. Основой для БД DOTS являются:

— экспериментально полученные последовательности мРНК;

— так называемые последовательности EST (Expressed sequence tags) [3];

— предсказанные гены – гены, выявленные с помощью компьютерного анализа.

Самыми многочисленными являются EST последовательности. EST, по сути, являются последовательностями разной длины (в среднем около 500 п.о.), полученными путем обратной транскрипции с 3’ конца мРНК последовательностей. Для общего пользования в настоящее время доступны порядка 2 млн таких последовательностей, из них 1,5 млн – человеческие, которые содержатся в специальной базе UNIGENE (http:// ncbi.nih.gov). Данные последовательности из генома человека сгруппированы на основе схожести (гомологии) в 65 000 кластеров, которые и являются основной частью БД DOTS.

Программное обеспечение DOTS позволяет с помощью программы множественного выравнивания cap2alignment (http://pcbi.upenn.edu/dots) собирать EST последовательности в кластеры (ансамбли) и строить для них обобщенные консенсусы, а также относить вновь поступающие последовательности EST в один из существующих кластеров. В случае отсутствия достаточной гомологии последовательности образуют новые кластеры. В настоящее время БД DOTS содержит 65 000 кластеров мРНК человека и 26 000 – мыши. С помощью другого программного инструмента k2, разработанного в университете Пенсильвании в Центре биоинформатики совместно с отделом систем информатики того же университета, БД DOTS была состыкована с основными существующими в мире базами данных по белкам и генам, с базами данных по картированию генов. В базе данных DOTS имеется средство поиска по гомологии BLAST [4], позволяющее выявить потенциальных “соседей” для интересующей последовательности, а именно: родственные гены, последовательности мРНК или EST, имеющие высокий процент гомологии с данной последовательностью. На основе этой информации с помощью сравнения хромосомной локализации последовательностей, клеточной роли и пр. можно делать выводы о свойствах аннотируемой последовательности ДНК.

Так как центр биоинформатики в основном специализируется на создании баз данных генов, экспрессирующихся в процессе эритропоэза, в текущий момент DOTS ориентирована на аннотирование генов, экспрессирующихся в стволовых кроветворных клетках и выявленных в университете Принстона с помощью технологии вычитания геномов. БД содержит порядка 7 000 генов, проаннотированных экспертами-биологами. Планируется проаннотировать еще порядка 30 000 генов в последующие полгода.

Второй частью проекта DOTS, над которым активно работают коллеги из Филадельфии, является анализ данных по экспрессии генов на основе технологии микрочипов. Кратко данную технологию можно описать следующим образом. Имея микроматрицу зондов приблизительно для 18 000 мРНК на пластине величиной несколько квадратных сантиметров, можно получить моментальный профиль экспрессии мРНК в данном типе клеток, обычно в сравнении с каким-либо контрольным образцом. В Филадельфии разрабатывается БД, к настоящему времени содержащая порядка 20 000 таких профилей для разных тканей, стадий онтогенеза и пр. Стоит сказать, что стоимость получения одного профиля экспрессии в настоящий момент составляет порядка 1,5 тыс. долларов (удовольствие не для бедных), но технология не стоит на месте.

Инкорпорация данных по экспрессии генов эукариот в БД DOTS может обеспечить значительный эффект в процессе аннотирования геномов, а также иметь большое значение для фармакогенетики, медицины и пр.

Список литературы

1. The yeast genome directory // Nature. 1997. № 387 (6632 Suppl.). P. 5–105.

2. Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium // Science. 1998. № 282. P. 2012–2018.

3. Adams M.D., Kerlavage A.R., Fleischmann R.D. et al. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence // Nature. 1995. № 377. P. 3–174.

4. Altschul S.F., Lipman D.J. Protein database searches for multiple alignments // Proc. Natl. Acad. Sci. USA. 1990. № 14. P. 5509–5513.

В.Бабенко, к.б.н., н.с. лаборатории молекулярных основ генетики животных,
ИЦиГ, Новосибирск

Г.Орлова, к.б.н., н.с. лаборатории генетики популяций,
ИЦиГ, Новосибирск