

ПРОГРАММНЫЙ КОМПЛЕКС ДЛЯ МНОГОМЕРНОГО АНАЛИЗА МИКРОЧИПОВЫХ ДАННЫХ. МНОГОМЕРНЫЙ PLS АНАЛИЗ

Д. С. Сулопаров, Д. А. Полунин, И. А. Штайгер, И. А. Агбаш
Институт цитологии и генетики
Новосибирский государственный университет

Технологические достижения последних десятилетий в молекулярной биологии дали исследователям возможность проводить широкомасштабные исследования. Особенно большую популярность приобрел анализ экспрессии генов при помощи микрочипов. Одной из важных целей анализа генетических данных является исследование мультифакторных заболеваний человека. В этой области остается немало нерешенных вопросов и проблем. Анализ этих данных используются для исследования молекулярных механизмов заболеваний, поиска мишеней для терапевтического вмешательства, классификации подтипов заболеваний, которые по-разному реагируют на терапию, поиска биологических маркеров заболеваний.

Объем данных, полученных для различных организмов и фенотипов, растет огромными темпами, и в настоящее время перед исследователями стоит проблема эффективной обработки всех этих массивов информации. Трудности состоят как в технологических особенностях применяемых методов, так и в огромном размере получаемых массивов данных, к примеру, несколько десятков тысяч генов в десятках биологических образцов.

Существующие на данный момент программные средства для многомерного анализа микрочиповых данных на практике не очень удобны, по многим причинам – закрытая реализация, тяжеловесность, либо проприетарность программного средства. По этим причинам разработка такого программного средства является актуальной задачей.

Система будет представлять собой головную программу диспетчер и набор независимых исполняемых модулей, выполняющих различные операции над данными. Моя часть работы заключается в реализации части этих модулей, а именно: стандартные арифметические функции, операции над матрицами, вычисление различных коэффициентов сходства/различия, методы PLS анализа. PLS регрессия (Partial Least Squares) – сравнительно новый статистический метод. С помощью него выделяется небольшое количество латентных переменных, в пространстве которых связь между зависимой переменной и предикторами достигает максимального значения. Он показывает хорошие результаты при анализе генетических данных.

Научный руководитель – д-р биол. наук, доцент В. М. Ефимов