ЛИНГВИСТИЧЕСКИЕ МЕТОДЫ ПОИСКА ВЗАИМОСВЯЗАННЫХ БЕЛКОВ

Пономаренко Е.А., Лисица А.В., Арчаков А.И.

ГУ НИИ Биомедицинской химии им.В.Н.Ореховича РАМН
Россия, 119121, г. Москва, ул. Погодинская, д.10. Тел.: +7(499)-246-37-31,
E-mail: 2463731@gmail.com

В связи с массовым внедрением высокоэффективных экспериментальных технологий перед исследователями в настоящее время существует проблема адекватной интерпретации полученных результатов. Часто итогом таких экспериментов является обширный список названий белков или генов, уровень экспрессии которых меняется под действием задаваемых внешних стимулов.

Целью данной работы является разработка системы, предназначенной для выявления групп взаимосвязанных белков, полученных в ходе проведения высокоэффективных протеомных экспериментов. При разработке системы в качестве исходных данных фигурируют тексты, написанные на естественных языках, т.е. научные публикации. Правомерность такого подхода подтверждают данные о существовании зависимости между частотой встречаемости терминов и смысловым наполнением документа [1].

На вход системы подается список названий белков, полученных в результате проведения эксперимента, дополненный известными синонимами (информация по данным ресурса UniProt [www.uniprot.org]). Каждое название белка из сформированного списка служит поисковым запросом к системе PubMed [www.pubmed.gov], после чего тексты резюме релевантных научных публикаций депонируются на локальный компьютер и представляют собой набор дескрипторов каждого белка из анализируемой группы. Взаимосвязь между двумя белками устанавливается путем выявления близких по смыслу документов из числа дескрипторов. Итогом работы алгоритма является реконструкция семантического графа, узлами которого являются названия белков, а ребрами – число близких по смыслу публикаций. Мера семантической связи между белками является функцией от числа близких по смыслу публикаций для двух белков.

В рамках тестирования работы системы показано, что выявленные группы семантически взаимосвязанных белков в большей степени соответствуют систематике белков согласно ресурсу KEGG [www.genome.ad.jp/kegg], чем классификации белков в терминах GeneOntology [5], а также не зависят от видовой принадлежности белков.

Литература

1. Singhal A., Mitra M., Buckley C. Learning routing queries in a query zone //*Proc.of the SIGIR* 1997, 25-32.
2. Ashburner M., Ball C.A., Blake J.A. et al. Gene Ontology: tool for the unification of biology. //*Nature Genet*. **25**, 2000, 25-29.