Визуализация генетической сети в рамках программы «Passportgen»

Сафарянц Д.С., д.т.н. проф. Софиев А.Э.
Университет машиностроения
89165508877, narpilin@bk.ru

Аннотация. В статье приведены результаты исследований по решению проблемы визуализации данных программного блока. Данный блок является частью информационно поисковой диагностической системы «Passportgen». Этот блок предоставляет врачу–диагносту визуальные данные о генетических взаимосвязях диагностируемого организма.

Ключевые слова *генетические сети, семантическая сеть, ориентированные графы*

Введение

Нами предложен модульный блок, который помогает в оказании медицинской помощи и выявления дополнительных отклонений, а возможно и для назначения лучшего способа лечения пациента. Данный блок разрабатывается в рамках информационно поисковой диагностической системы (ИПДС) «Passportgen». Оператор ИПДС – врач–генетик, в обязанности которого входит интерпретация полученных данных в диагноз и составление рекомендаций по выявленным у пациента заболеваниям, а также вероятным болезням с отсутствующими клиническими признаками. Разработанная методика диагностирования является трудоемкой и при ее использовании существует вероятность пропуска врачом–генетиком необходимой информации. Для ускорения работы и сокращения времени на постановку предварительного диагноза Угаровым И.В. было предложено визуализировать отдел генных сетей, находящихся в базах данных ИПДС «Passportgen». Генетические сети представляют собой группу координировано функционирующих генов, обеспечивающих формирование фенотипических признаков организма (молекулярных, биохимических, физиологических). [1,2]

Постановка задачи

Ссылаясь[3] на утверждение о том, что с помощью глаз воспринимается 90 % всей получаемой информации, было принято решение представить генетическую сеть в виде семантической сети, что является усложненной версией представления ориентированного графа.[4] После обработки информации будут получены связи генов характеризующиеся как:

- Понижение регуляции (Down-regulation);
- Повышение регуляции (Up-regulation);
- Регуляция (Regulation);
- Совместная экспрессия (Co expression);
- Химическая модификация (Chemical Modification);
- Физическая модификация (Physical interaction);
- Предсказанное взаимодействие белков (Predicted Protein Interaction);
- Предсказанное взаимодействие транскрипционных факторов (Predicted TFactor Regulation);
- Другое (Other).

Можно выстроить взаимосвязь понятия и отношения. Понятия в этом частном случае представляют собой базу знаний (полное описание гена – участника результатов исследования) представленную в качестве вершин, а отношения будут отображать характер связи между генами и представлены в виде соединительных дуг. Толщина вершин и длина дуги должна предоставлять информацию о наименовании гена, принадлежности его к определенному участку хромосомы, а вид связи, обозначенный определенным цветом, привязан к характеру связи. После обработки данных программа должна выводить на экран несколько видов полезной информации:

1. Список генетических сетей представленных в виде подвижного изображения для

уточнения вида связи между генами одного заболевания.

2. Всплывающее окно с кратким описанием гена.

Исследования и результаты

На рисунке 1 можно увидеть пример исходных данных по генетическому заболеванию синдрома Альпорта.

Столбик 1	Столбик 2	Столбик 3	Столбик 4	Столбик 5
*edges				
COL4A3	COL4A3	i 11	c 1.0	o 0.0
COL4A4	COL4A4	i 11	c 1.0	o 0.0
COL4A5	COL4A5	i 11	c 1.0	o 0.0
COL4A3BP				
COL4A3BP		i 11	c 1.0	o 0.0
LOC387911				
LOC387911		i 11	c 1.0	o 0.0
NOTCH2	NOTCH2	i 11	c 1.0	o 0.0
COL4A1	COL4A1	i 11	c 1.0	o 0.0
COL4A6	COL4A6	i 11	c 1.0	o 0.0
BUB3	BUB3	i 11	c 1.0	o 0.0
KIF11	KIF11	i 11	c 1.0	o 0.0
ZWINT	ZWINT	i 11	c 1.0	o 0.0
COL4A3BP				
COL4A3		i 8	c 0.6	o -0.5
COL4A3BP				
COL4A3		i 6	c 0.6	o 0.5
COL4A5	COL4A3	i 8	c 0.6	o 0.0
COL4A5	COL4A4	i 8	c 0.6	o 0.0
BUB3	COL4A5	i 5	c 0.6	o 0.0
COL4A5	COL4A6	i 5	c 0.8	o 0.0
COL4A5	KIF11	i 5	c 0.6	o 0.0
COL4A5	ZWINT	i 5	c 0.6	o 0.0
COL4A4	COL4A3	i 4	c -0.6	o 0.0
COL4A1	COL4A4	i 4	c -0.6	o 0.0
LOC387911				
COL4A5		i 4	c -0.6	o 0.0
COL4A5	NOTCH2	i 4	c -0.6	o 0.0

Рисунок 1. Список генов (1, 2 столбик) с видами связей (столбик 3, где i – характер связи между генами)

Выстроим логику, основываясь на принципах ориентированного графа:

1. Для задания множества вершин, непосредственно достижимых из вершины v, используют линейный однонаправленный список. Каждый элемент (ген) такого списка включает данные (некое число характеризующее вид влияния) и указатель на следующий элемент списка. Список в целом задается указателем на его первый элемент (голову списка). Последний элемент списка содержит "пустой" указатель.
2. Задается для вершины v ее список смежности. В отдельном столбце содержаться данные элементов списка номера вершин, для которых в ориентированном графе есть дуга из v в u (v→u). Список смежности вершины (v) обозначают L(v).
3. Отметим, что список смежности вершины может при необходимости дополняться. Для этого в последнем элементе списка "пустой" указатель заменяется указателем на добавляемый элемент, который становится последним элементом списка с "пустым" указателем.
4. Если количество вершин ориентированного графа известно заранее, то ориентированный граф удобно задавать в виде структуры. Подобную структуру в теории графов называют массивом лидеров. Такие структуры максимально эффективны, если рассматривается не итоговый вариант анализов данных, а отдельное заболевания, выбранное из библиотеки знаний программы. Под массивом имеем в виду матрицу-столбец, элементами которой могут быть некоторые объекты (например, гены определенного заболевания). Их называют элементами массива. Число элементов массива лидеров равно числу вершин графа. Элементами массива лидеров являются первые элементы списков смежности вершин ориентированного графа.

В системе программирования Visual Studio Team System был написан код, который позволил получить весьма информативное изображение (рисунок 2) визуализации генов,

участвующих в формировании синдрома Альпорта (наследственного нефрита) − наследственного заболевания, при котором функция почек снижена, в моче присутствует кровь; синдром иногда сопровождается глухотой и поражением глаз [5]. Черными жирными точками изображены гены, имеющие вид связи, замыкающиеся на себя.

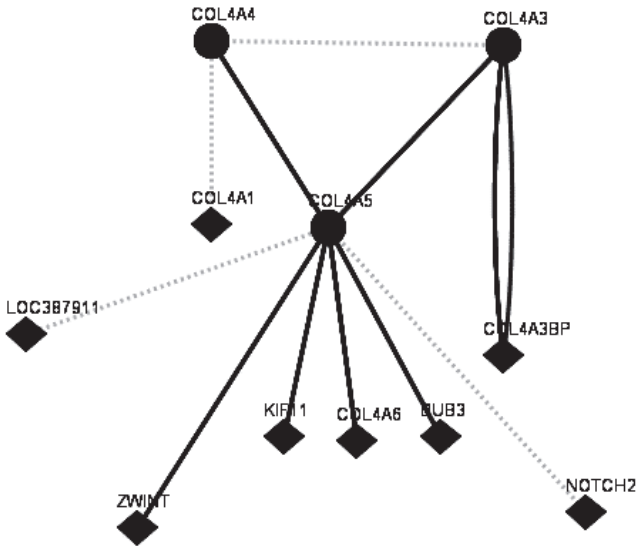

Рисунок 2. Пример визуализации

Плюсы данного метода визуализации: наглядность, информативность, быстродействие системы в целом

Минусы: при попытке отображения больших генетических сетей идет потеря наглядности, обусловленная плотностью расположения дуг с вершинами, а также происходит замедление быстродействия системы.

Примеры подобных визуализаций с исправленными минусами можно увидеть в вебсервисах VisANT [6], GNCPro [7] и в некоторых программах ресурса NCBI [8].

Выводы

Полученные результаты позволяют рекомендовать семантическую сеть в качестве одного из видов решения для визуализации генетической сети.

Благодарность

Выражаем отдельную благодарность к.м.н. Угарову Игорю Викторовичу за предоставленную возможность в участии разработки программного обеспечения «Passportgen».

Литература

1. Генные сети / Колчанов Н.А., Ананько Е.А., Колпаков Ф.А. и др. // Молекулярная биология.- 2000.- т 34.№4-с.617-629
2. Интеграция генных сетей, контролирующих физиологические функции организма / Колчанов Н. А., Подколодная О.А., Игнатьева Е.В. и др. // Вестник ВОГИС.-2005.- Т.9,№2.- С.179-199.
3. Линдгрен Н., 1962
4. Теория графов / Карпов Д.В.
5. http://www.zdorovieinfo.ru/bolezni/sindrom_alporta/
6. http://scolaris.beta.semantico.com/
7. http://gncpro.sabiosciences.com/gncpro/gncpro.php
8. http://www.ncbi.nlm.nih.gov/