OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

ИСПОЛЬЗОВАНИЕ ОДНОРОДНОЙ СЕМАНТИЧЕСКОЙ СЕТИ ДЛЯ КЛАССИФИКАЦИИ РЕЗУЛЬТАТОВ ГЕНЕТИЧЕСКОГО АНАЛИЗА

Куликов А.М.[*], Харламов А.А.[**]

[*]*Институт биологии развития им. Н.К. Кольцова РАН, г. Москва*
amkulikov@gmail.com

[**]*Институт высшей нервной деятельности и нейрофизиологии РАН, г. Москва*
kharlamov@analyst.ru

В работе показано использование механизма сравнения семантических сетей текстов в задаче диагностики заболеваний с использованием сигнальных сетей. Выявление степени пересечения семантических сетей текстов позволяет говорить о степени их смыслового подобия. Однородная семантическая сеть как множество узлов, связанных дугами, имеет численные характеристики – частоты появления слов, а также пар слов в тексте, которые перенормируются с использованием n-граммной модели текста. Такие сети как смысловые портреты текстов могут служить для сравнения (и, следовательно, для классификации) текстов. Генетический квазитекст может быть представлен, в том числе, в виде сигнальной или генной сети. Сигнальные сети разных классов генетических событий могут быть использованы для классификации этих текстов. В этом случае концентрации белков, выявленные в процессе эксперимента, используются для вычисления числовых характеристик узлов сети. Приведены примеры сравнения сетей генетических квазитекстов, соответствующих норме и патологии.

Ключевые слова: однородные семантические сети, сигнальные сети, сравнение текстов, классификация текстов

Введение

Предположение о сходстве текстов естественно-языковых и текстов генетических кодов (в дальнейшем будем их называть генетическими квазитекстами) оказывается правомерным при более подробном сравнении. Семантические сети как смысловые портреты естественно-языковых текстов [Харламов, 2006] имеют свою параллель в виде сетей, представляющих некоторые предметные области в генетике (например, определенную патологию). Рассмотрим это сравнение более подробно.

Семантическая сеть естественно-языкового текста – это граф, вершинами которого являются ключевые понятия этого текста, а дуги описывают взаимосвязи ключевых понятий в тексте. Типичный вид такого графа для, например, текста «Нейросетевая среда (нейроморфная ассоциативная память) для преодоления информационной сложности. Поиск смысла в слабо структурированных массивах информации. Часть I.

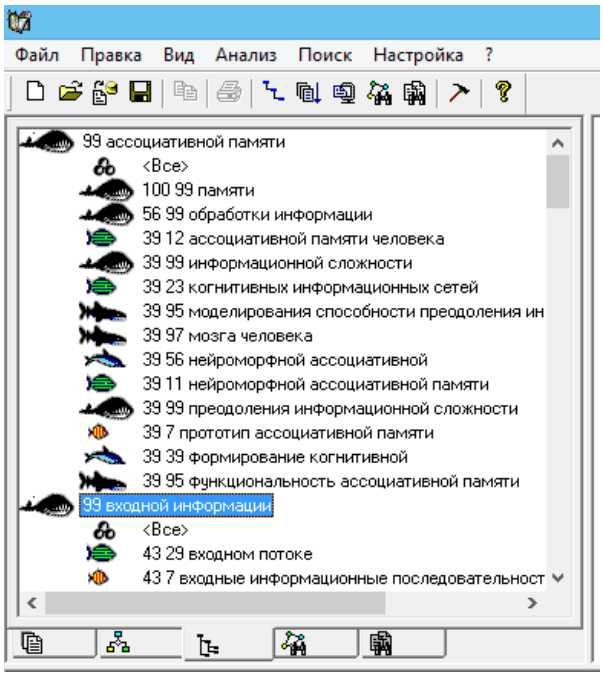

Рисунок 1 – Фрагмент семантической сети текста. Здесь перечислены вершины и показаны их связи. И вершины и связи имеют числовые характеристики

Структурная обработка информации в коре», опубликованный в 11 номере журнала «Информационные технологии» за 2015г. представлен (с использованием графического интерфейса программного продукта для автоматического смыслового анализа текстов TextAnalyst) на рисунке 1.

Такая сеть представляется как множество пар слов, встречающихся в тексте. И сеть они составляют как раз потому, что некоторые пары слов оказываются связанными между собой через промежуточное слово. Другими словами, такая сеть – это перечень пар слов.

Сеть, описывающая генетическое событие (см. как пример сеть, представленную на рисунке 2), как правило, получается при анализе активности работы генов (или экспрессионной активности) клеток, органов, тканей или целого организма на определенной стадии развития и/или под действием тех или иных факторов. В этом случае концентрации тех или других веществ позволяют делать предположение о протекании определенных генетических процессов, каждый из которых имеет свое начало и свой конец. Эти процессы могут иметь общие промежуточные компоненты, то есть суммарный граф, представляющий результаты эксперимента, также разбивается на пары событий (которые можно условно называть словами).

В приведенном примере на схеме изображена сеть межбелковых взаимодействий, представляющая передачу сигналов от толл-подобных рецепторов, участвующих в клеточном иммунном ответе, на определенный набор транскрипционных факторов, активирующих работу соответствующих генов. Набор собственно толл-подобных рецепторов, вспомогательных белков и факторов, участвующих в активации этих рецепторов, расположен в верхней части схемы. В нижней части представлены транскрипционные факторы, т.е. регуляторы активности генов-мишеней. Синие и зеленые стрелки показывают отрицательные обратные связи между регулируемыми транскрипционными факторами и центральными узлами данной сети, желтые и красные – аналогичные положительные обратные связи. Характер связей в представленном графе позволяет сделать вывод, что активация одних узлов сети приводит к формированию циклов последовательного усиления активности генов-мишеней, тогда как активация других приводит к последовательному снижению активности генов-мишеней. Количественные оценки состава белков и РНК в клетке или ткани позволяет, при наложении этих оценок на сигнальную сеть, сделать вывод об относительной активности различных участков такой сети, или подграфов, и об активности биологических процессов, определяемых данными подграфами.

Учитывая последовательность и дискретность актов передачи сигнала между белками-партнерами, весь набор передаваемых сигналов от вершин входа графа до набора вершин – «мишеней» за некоторый промежуток времени можно представить как набор одновременно и/или последовательно идущих предложений, где пары слов представлены парами взаимодействующих молекулярно-генетических объектов.

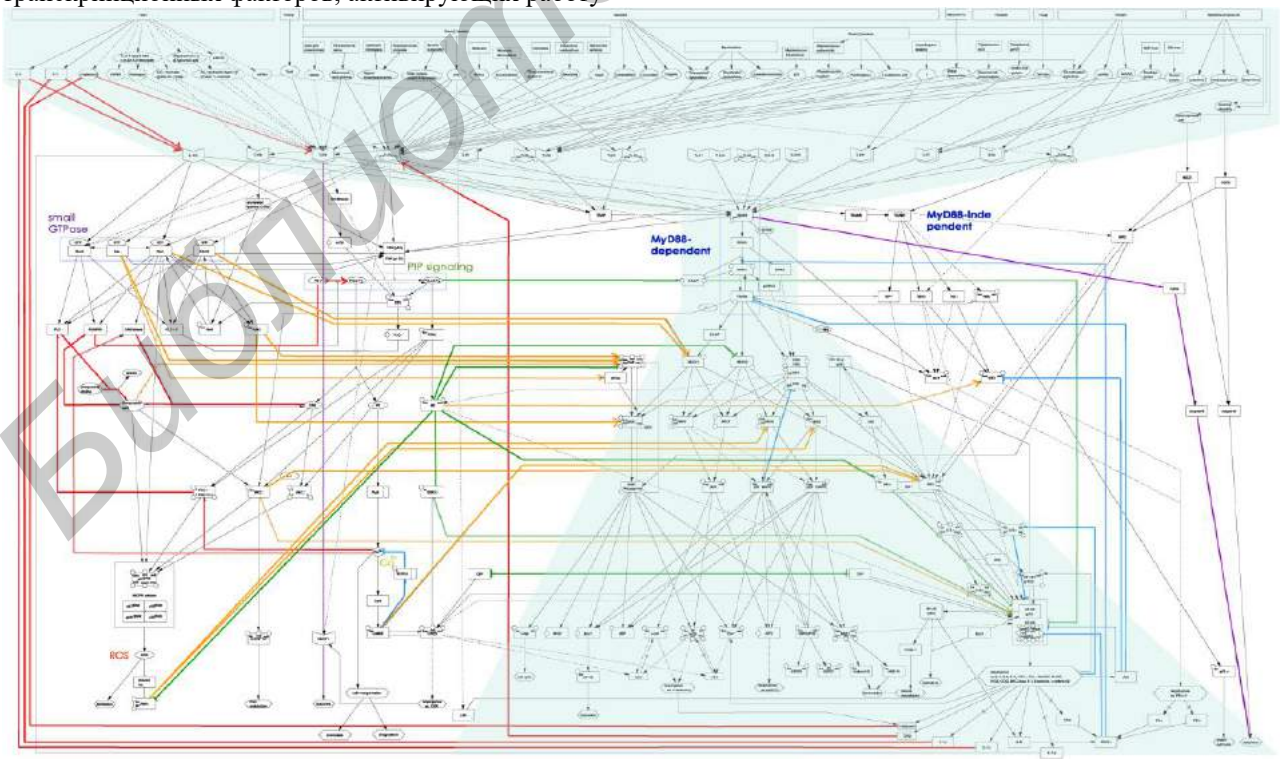

Рисунок 2 – Иллюстрация из статьи K. Oda ·и H Kitano: A comprehensive map of the toll-like receptor signaling network. (Mol Syst Biol , 2006. 2:0015): схема сигнального пути toll-like рецепторов

Анализ семантических сетей – смысловых портретов – естественно-языковых текстов позволяет осуществлять сравнение текстов по структуре (по смыслу) [Харламов, 2006]. Аналогия с генетическими квазитекстами позволяет предположить возможность использования сравнения сетей для выявления степени подобия между ними – для классификации генетических событий.

1. Молекулярно-генетические данные

В настоящий момент экспериментальной биологией для основных модельных объектов – человека, мыши, крысы, дрозофилы и некоторых других видов получены строгие оценки десятков тысяч взаимодействий между веществами белковой и небелковой природы в организме изучаемого объекта. Многие из этих взаимодействий объединены в составе цепей или каскадов взаимодействий, имеющих направление, положительные и отрицательные эффекты взаимодействий, различной степени пересечения. Все вместе эти взаимодействия образуют чрезвычайно сложный высокосвязный граф. Тем не менее, такой граф достаточно строго делится на подграфы, представляющие собой каскады метаболических или сигнальных путей. Это связано с тем, что в клетке передача информации зависит преимущественно от «входных» узлов подграфа, и направлена к ограниченному количеству узлов-«мишеней» на выходе подграфа, а различные латеральные связи узлов подграфа с другими подграфами определяют интенсивность передачи сигнала по каскаду.

Общий граф, определяющий все возможные варианты взаимодействий веществ в данном организме, существует только гипотетически. Реально в разных типах клеток и тканей, на разных стадиях развития организма, в норме и при патологии работают разные наборы генов. В сравниваемых выборках часть генов отличается качественно, т.е. уникальна только для одной из сравниваемых групп, а среди совпадающих генов активность работы может различаться. Такие различия выявляются в количественных показателях состава и/или активности белков или РНК в образце. Суть генетического эксперимента заключается в сравнении наборов данных по альтернативным выборкам, выявление общих для сравниваемых выборок наборов и наборов со значимым изменением активности работы (экспрессии) генов в сторону снижения или увеличения. Анализ структур графов, построенных из таких наборов, позволяет выявить неслучайные процессы роста или падения активности сигнальных или метаболических каскадов.

Упомянутый выше граф, описывающий генетическое событие, может быть представлен перечнем пар «слов» - названий участвующих в эксперименте веществ.

2. Однородная семантическая сеть

Автоматический смысловой анализ естественно-языковых текстов заключается в выявлении ключевых понятий и их взаимосвязей, и ранжировании понятий и связей, то есть в формировании однородной семантической (ассоциативной) сети [Харламов, 2006].

Под семантической сетью $\boldsymbol{N}$ понимается множество несимметричных пар событий $\{<c_i c_j>\}$, где c_i и c_j – события, связанные между собой отношением ассоциативности (совместной встречаемости в некоторой ситуации):

$$\boldsymbol{N} \cong \{< c_i c_j >\} \quad (1)$$

с весовыми характеристиками w_i и w_{ij}, соответственно, ключевого понятия и связи между ключевыми понятиями. В данном случае отношение ассоциативности

несимметрично: $< c_i c_j > \neq < c_j c_i >$.

$$< c_i c_j > \neq < c_j c_i >. \quad (2)$$

Особенностью анализа является итеративная процедура переранжирования частот встречаемости слов текста в их ранг w_i:

$$w_i(t+1) = (\sum_{i \neq j} {}_i \; w_i(t) w_{ij}) \sigma(\bar{E}). \quad (3)$$

Здесь $w_i(0)=z_i$; $w_{ij}=z_{ij}/z_i$ и $\sigma(\bar{E}) = 1/(1+e^{-k\bar{E}})$ – функция, нормирующая на среднее значение энергии всех вершин сети $\bar{E}$; z_i – частота встречаемости i-го слова в тексте, z_{ij} – частота совместной встречаемости i-го и j-го слов в фрагментах текста; t – номер итерации. Полученная числовая характеристика слов – их смысловой вес – характеризует степень их важности в тексте.

3. Классификация текстов с помощью сравнения сетей

Семантическая сеть $\boldsymbol{N}$, описанная таким образом (1), может быть представлена как множество так называемых звездочек $<c_i<c_j>>$ [Харламов, 2015]. Это все пары $\{<c_i c_j>\}$, у которых первое слово одинаковое:

$$\boldsymbol{N} \cong \{z_i\} = \{< c_i < c_j >>\}. \quad (4)$$

Под звездочкой $<c_i<c_j>>$ понимается конструкция, включающая главное событие c_i, связанное с множеством событий – ближайших ассоциантов в сети $<c_j>$, которые являются семантическими признаками главного события, отстоящими от главного события на одну связь. Связи направлены от главного события к событиям-ассоциантам.

Звездочка с единичными значениями весов событий-ассоциантов называется единичной

звездочкой (звездочкой-ортом). Звездочкой-подпространством называется звездочка, полученная на единичной звездочке введением весов событий w_j:

$$Z \cong < c_i < w_j c_j >>. \quad (5)$$

Семантическая сеть в терминах этих определений представляет собой декартово произведение подпространств, порождаемых всеми звездочками, входящими в семантическую сеть, полученными на единичных звездочках за счет введения весовых характеристик понятий-ассоциантов:

$$\boldsymbol{N} = Z_1 \times Z_2 \times \ldots \times Z_I. \quad (6)$$

Введем скалярное произведение на векторах $\overline{c_i}$ и $\overline{c_1}$, где угол между векторами понятий соответствующих c_i и c_j: пропорционален весу связи от c_i к c_{ij} : $w_{ii} \in (0 \ldots 90°)$.

Площадь треугольника s_i, построенного на векторах $\overline{c_i}$, $\overline{c_1}$, развернутых на угол w_{ij} относительно друг друга, будет использована для вычисления степени пересечения сначала звездочек, а потом семантических сетей как совокупностей звездочек.

Под пересечением двух звездочек понимается сумма по всем событиям-ассоциантам данного главного события звездочки пересечений площадей двух треугольников, построенных в плоскости векторов $\overline{c_i}$ и $\overline{c_j}$, один из которых построен на векторах, развернутых на угол, пропорциональный связи $(w_{ij})_1$ между событиями в одной звездочке, а другой – на угол, пропорциональный связи $(w_{ij})_2$ между теми же событиями в другой, сравниваемой с первой, звездочке. В случае если в одной из звездочек пары, для которой считается пересечение:

$$\boldsymbol{s_{12}} = < \boldsymbol{c_{i_1}} < \boldsymbol{c_{j_1}} >> \cap < \boldsymbol{c_{i_2}} < \boldsymbol{c_{j_2}} >> = \sum_{j=1}^{max(N_1 N_2)} (\boldsymbol{s_{j_1}} \cap \boldsymbol{s_{j_2}}). \quad (7)$$

Если не нашлось соответствующего события-ассоцианта, пересечение считается равным 0. Здесь N_1, N_2 – число ассоциантов в звездочках соответственно 1 и 2.

Тогда под пересечением семантических сетей понимается сумма пересечений звездочек, включенных в эти сети (считая по главным понятиям):

$$S_{12} = \sum_{k=1}^{max(M_1 M_2)} (\boldsymbol{s_{j_1}} \cap \boldsymbol{s_{j_2}}), \quad (8)$$

где M_1, M_2 – число звездочек, входящих соответственно в семантические сети $\boldsymbol{N_1}$, $\boldsymbol{N_2}$.

Под классификацией входного текста можно понимать отнесение семантической сети входного текста $\boldsymbol{N}$ к сети $\boldsymbol{N_n}$ (где $n=1..N$ – число предметных областей) одной из предметных областей, описываемых соответствующими корпусами текстов. В идеальном случае семантическая сеть текста вкладывается в сеть соответствующей предметной области.

Используя операцию пересечения сетей $\boldsymbol{N_1}$ и $\boldsymbol{N_2}$, определенную выше, мы можем оценивать степень подобия двух сетей $\boldsymbol{N_1} \cap \boldsymbol{N_2}$ и, тем самым, сравнивать по смыслу (по структуре) тексты. Имея модели предметных областей в виде ассоциативных семантических сетей, мы можем классифицировать входные тексты вычислением степени совпадения (вложения) сети входного текста и сетей предметных областей, относя входной текст к той предметной области, у которой степень совпадения сети с сетью предметной области окажется выше.

Поскольку выше мы показали аналогию между текстами естественного языка и генетическими квазитекстами, представляется возможным использовать приведенный механизм сравнения сетей для классификации генетических квазитекстов.

4. Результаты генетических экспериментов как квазитекст

Аналогия между естественно-языковыми и генетическими текстами позволила предположить возможность использования механизмов анализа естественно-языковых текстов для анализа генетических квазитекстов. Для анализа были использованы так называемые сигнальные или генные сети – сети переходов, описывающие процессы передачи сигналов внутри живой клетки.

Результаты нескольких молекулярно-генетических исследований были использованы для выяснения возможности применения подхода для классификации генетических квазитекстов. В качестве экспериментального материала были использованы результаты анализа экспрессионной активности в ткани опухоли (саркома) и в нормальной ткани.

Так же как и в случае анализа естественно-языковых текстов, представленных однородными семантическими (ассоциативными) сетями, генетические квазитексты, представленные в виде сигнальных сетей – графов, вершинами которых являются названия некоторых веществ, участвующих в генетических процессах, а дуги указывают на взаимосвязи этих веществ в этих процессах – представлялись в виде пар слов $<c_i c_j>$, далее – в виде звездочек $<c_i<c_j>>$.

Сигнальные сети, использованные в эксперименте, были представлены в виде пар «слов», где в качестве слов рассматривались названия белков, концентрации которых анализировались в генетическом эксперименте. Поскольку эти «слова» составляют цепочки при описании генетических процессов (как на рисунке 2), результаты генетического эксперимента могут

быть представлен в виде перечней пар «слов», каждая из которых имеет свой вес – концентрацию, выявленную в результате эксперимента.

Фрагмент одного из таких перечней (саркома) представлен ниже.

(MAPK1 ELK1), (MAPK1 FOS), (MAPK1 MAPK3), (MAPK1 RPS6KA1), (PPP1R3B PPP1R3C), (HDAC9 HSPA5), (TIRAP TRAF6), (STAT5B FOS), (RAP1A BRAF), (RAP1A RASA1), (RAP1A KRIT1), …

Для того чтобы использовать эти перечни в качестве входных данных для анализа, обычно применяемого для естественно-языковых текстов, необходимо осуществить некоторую их интерпретацию, и некоторое преобразование. Основой механизма анализа естественно-языковых текстов является переранжирование весовых характеристик слов, результат которого зависит от внутренней структуры текста, представленной в виде семантической сети.

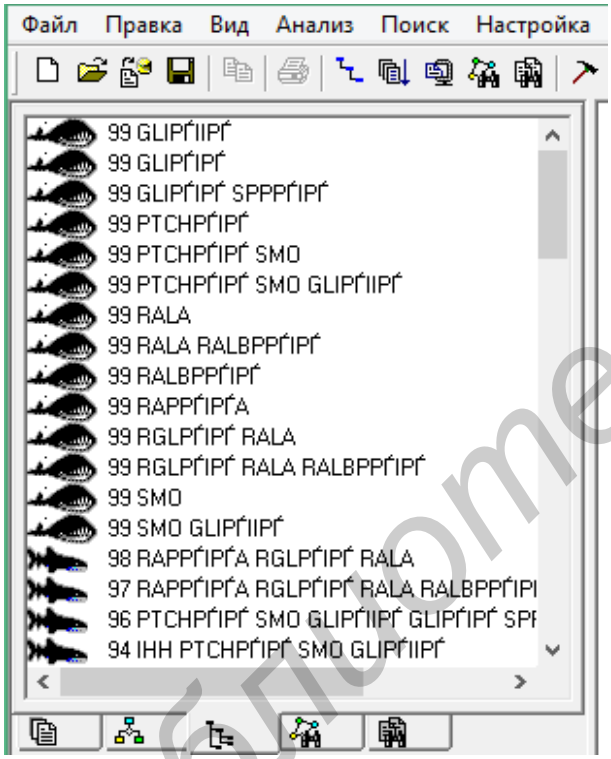

Рисунок 3 – Фрагмент семантической сети генетического квазитекста. Здесь перечислены вершины и показаны их связи. И вершины и связи имеют числовые характеристики. Символы с диакритическими значками кодируют цифровую информацию при обработке в программе TextAnalyst

Другими словами, необходимо привести исходные данные, полученные в генетическом эксперименте к виду, близкому к виду естественно-языкового текста. Для этого полученные в генетическом эксперименте пары «слов» собираются в «предложения» - цепочки пар слов, описывающие некоторые генетические процессы, от их начала – появления на входе рецептора, до их конца – достижения мишени. Такая сборка осуществляется с помощью «шаблонов» - цепочек пар «слов», стандартных для некоторых групп генетических процессов, представленных в подсетях, описывающих норму/патологию. Отнормированная величина концентрации используется как характеристика частоты появления пар «слов» в «тексте» - сигнальной сети. «Текст» представлен выявленными вышеописанным способом «предложениями», перечисленными в случайном порядке.

Тогда мы можем анализировать полученный «текст» как в случае естественно-языкового анализа. При этом выявляется частота встречаемости «слов», частота встречаемости пар «слов», строится семантическая сеть «текста», переранжируются весовые характеристики «слов» «текста».

Ну и далее полученные для трех разных случаев генетического эксперимента такие семантические сети (их пример представлен на рисунке 3) сравниваются между собой с целью выявления степени их пересечения.

5. Обсуждение и анализ результатов

Ниже в таблице 1 представлены результаты сравнения трех семантических сетей, соответствующих трем генетическим «текстам», описывающим соответственно, норму и два вида патологии (саркома), которые были получены в результате генетических экспериментов.

Таблица 1 – Результаты сравнения семантических сетей трех разных генетических квазитекстов, соответствующих один – норме, а два других – патологии (саркома).

	Норма	Патология 1	Патология 2
Норма	1,000	29,144	30,334
Патология 1	29,144	1,000	23,150
Патология 2	30,334	23,150	1,000

Конечно, интерпретация сравнения всего трех текстов не очень убедительна, но результаты сравнения, тем не менее, говорят за себя: мы видим разницу в сравнении сетей нормального и двух патологических случаев, и бóльшую степень пересечения при сравнении сетей, представляющих патологию. Невысокие разницы возникают за счет большой общности так называемого house keeping, имеющего место как в норме, так и патологии.

Заключение

Использование механизма сравнения семантических сетей естественно-языковых текстов позволяет сравнивать и другие похожие сетевые структуры, в том числе – сигнальные сети (генетические квазитексты) различных генетических заболеваний, что можно использовать для классификации таких сетей, а следовательно, и для диагностики заболеваний.

Использование данного подхода для сравнения, следовательно, и классификации генетических «текстов» позволит автоматизировать обработку результатов генетических экспериментов, объем которых в известных хранилищах (например, GeneNet) очень велик. Что, в свою очередь, облегчит и улучшит интерпретацию результатов генетических экспериментов.

На самом деле представление генетических квазитекстов в виде перечней пар слов не совсем корректно, кстати, как и при анализе естественно-языковых текстов. Некоторые пары слов различаются видом связи. В дальнейшем вместо пар слов квазитекстов можно будет использовать тройки, включающие помимо пары слов еще и тип связи между ними.

Библиографический список

[**Харламов, 2006**] Харламов А.А. Нейросетевая технология представления и обработки информации (естественное представление знаний) / А.А. Харламов // М.: Радиотехника, 2006. – 89 с.

[**Харламов, 2015**] Харламов А.А., Ермоленко Т.В. Нейросетевая среда (нейроморфная ассоциативная память) для преодоления информационной сложности. Поиск смысла в слабоструктурированных массивах информации. Часть II. Обработка информации в гиппокампе. Модель мира. / А.А. Харламов, Т.В. Ермоленко // Информационные технологии, 2015, № 12, С. 883—889

HOMOGENOUS SEMANTIC NETWORK FOR GENETIC ANALYSIS RESULT CLASSIFICATION

Kulikov A.M.[*], Kharlamov A.A.[**]

[*]*Koltzov Institute of Developmental Biology of Russian Academy of Sciences, Moscow*

amkulikov@gmail.com

[**]*Institute of Higher Nervous Activity and Neurophysiology of Russian Academy of Sciences, Moscow*

kharlamov@analyst.ru

In the work mechanism of text semantic networks comparison in task of desease diagnostic on signalling network using is shown. One can calculate texts semantic similarity by calculating volume of their networks crossection. Homogenous network consists of nodes and their connections with their weight characteristics. The characteristics may be more exact by their renormalization on the basis of text n-gram model. The examples of such genetic quasi-texts network crossection calculation of two different diseases.

Key words: homogenous semantic networks, signaling networks, texts comparison, text cclassification.

Introduction

There is an analogy between natural language texts and genetic code texts (genetic quasi-texts) because of natural language text semantic network and genetic quasi-text signaling network similarity. Signaling network also can be represented as list of quasi-word pares.

Main Part

Currently, estimations of the interactions between compounds of protein and non-protein nature in the organism of the studied object have been obtained in the field of experimental biology. Many of these interactions are integrated as part of chains or cascades of interactions. These chains have a direction, positive and negative effects of the interactions of varying degrees of intersection. Together, these interactions form a highly complex connected graph, which is divided into subgraphs representing cascades of metabolic and signaling pathways.

A full graph defining all possible interactions of the substances in the organism exists only hypothetically. As a matter of fact, in different types of cells and tissues at different stages of development of an organism, in the normal and pathologic different sets of genes are working. In comparing samples, a part of genes differs qualitatively, i.e. it is unique for only one of the groups being compared. The essence of the genetic experiment is to compare data sets of alternate samples, identification of the common sets for compared samples and sets with a significant change in the activity of gene work (expression). An analysis of the graph structure constructed from such data sets allows revealing nonrandom processes of growing or falling down in activity of signaling or metabolic cascades.

The above mentioned graph describing genetic events may be represented by a list of pairs of "words" - the names of substances participating in the experiment.

Automatical semantic natural language text analysis extract key words and their relationships from the text. That is why the key words can be integrated into homogenous semantic network. The nodes and arcs of the network have their semantic weights.

Such networks can be compared by their crossection calculation – calculation of the nodes and relationships weighting coefficients in the crossection.

Because of natural language text and genetic quasi-text analogy one can use the mechanism for genetic quasi-texts comparing (and for their classification). In the paper results of such comparing signaling networks of two diseases represented.

Conclusion

One can use a natural language text semantic network comparison mechanism and for another network structures comparing also, genetic quasi-texts signaling networks different diseases for example. That is why one can use it and for their diagnostic also. The exactness of the analysis may be can be shifted by using not word pares networks but word triple ones where besides the word pare has a type of relationship too.