Методы автоматического построения специализированного тезауруса

Герасимов М.Б., Пунтиков Н.П., Перегудова М.В., Маленков С.А.,

Цыганков М.А., Евграфов А.А., Виноградов А.Е.

СТАР СПб, Россия

{nick,lloix,vitess,sam,mike,toxa,aevin}@star.spb.ru

Аннотация

В статье описываются методы автоматического построения тезауруса для специализированной предметной области, - молекулярной биологии. В качестве источника терминологической информации использовалась структурированная база данных белков и белковых соединений, SwissProt. Успешная реализация проекта доказала не только возможность полной автоматизации ручного процесса, но и перспективу использования нелингвистических методов наравне с собственно лингвистическими.

Введение

Огромные объёмы информации, которые стали доступны в последние несколько лет в связи с развитием разнообразных информационных и телекоммуникационных технологий ставят принципиально новые проблемы перед поставщиками и потребителями этой информации. Попытка найти информацию, релевантную потребностям рядового пользователя заканчивается тем, что на него обрушивается большое количество различных фактов, так или иначе связанных с запросом. Человек не может за разумное время обработать несколько сотен или тысяч предложенных фактов, пытаясь найти действительно необходимые ему сведения.

В этой связи новые требования предъявляются к системам поиска и навигации в массивах информации (системы Information Retrieval, IR), а следовательно, и к механизмам их построения. Качественно различают две составляющие системы IR: технологическую и концептуальную. К технологическим составляющим можно отнести средства пользовательского интерфейса, новые алгоритмы обработки текста, индексации и поиска, интеграцию информации из различных источников, сложные языки запросов и др. К концептуальным составляющим, прежде всего, относится система представления знания об обрабатываемом материале, лежащая в основе механизма информационного поиска.

Именно второй аспект развития систем поиска является предметом наших исследований. В зависимости от типа этой системы различают несколько технологий поиска.

1. По-символьный поиск, происходящий без привлечения знаний о лексической, грамматической и семантической структуре обрабатываемого материала.
2. Поиск, в ходе которого используется лексико-грамматическая информация. Значит, привлекаются лингвистические словари, программы морфологического анализа текста.
3. Семантический поиск, осуществляющийся на основании знания об отношениях между понятиями предметной области, выраженными посредством слов естественного языка. Носителями такого рода инфорации являются тезаурусы, уже более трех десятилетий использующиеся для информационного поиска.

Разработаны сотни тезаурусов, описывающих понятийные и терминологические системы многих предметных областей. Однако, разработка тезауруса для новой предметной области, равно как и его пополнение все еще остается большой проблемой в силу трудоемкости ручной работы.

С развитием компьютерной техники появилось значительное число разнообразных работ по автоматическому извлечению лингвистических и терминологических знаний из источника. При этом важным остается вопрос неадекватности и неоднозначности получаемой информации, которая должна впоследствии интерпретироваться экспертами. Поэтому при автоматизации сложных интеллектуальных процессов важным является не только разработка автоматических процедур, но и

- обоснованность применения каждой из этих процедур в конкретном случае,

- последовательность их применения,

- относительный «вес» результатов применения каждой процедуры, позволяющий выбрать правильный (из ряда неодинаковых) в случае использования нескольких алгоритмов одновременно, а также

- автоматически контролируемые действия эксперта.

Перед нами была поставлена задача автоматического построения тезауруса для терминов, описывающих белки и белковые соединения, используя данные, содержащиеся в структурированных информационных базах. Необходимо отметить, что эти БД не являются лингвистическими, то есть содержащаяся в них терминологическая информация играет вспомогательную роль. С точки зрения поставленной задачи такое свойство БД является определенным препятствием, поскольку затрудняет привлечение знания о лингвистических закономерностях построения текста, но, с другой стороны, позволяет использовать иные, биологические алгоритмы выявления близких понятий.

В следующих разделах описана специфика терминологии конкретной предметной области, рассмотрена структура использованной базы данных и приведены алгоритмы извлечения терминов из текста записей базы и объединения терминов в синонимические ряды. Особое

внимание уделено результатам применения каждого алгоритма, приемуществам и недостаткам конкретного алгоритма в сравнении с другими.

Специфика биологической терминологии

Ведущей областью биологии в настоящее время является молекулярная биология. Терминология этой области отличается повышенной сложностью, так как здесь пересекаются термины химического и биологического происхождения. Кроме того, существует значительное разнообразие методов и подходов к изучению биологических молекул, что, как правило, дополнительно осложняет работу с терминологией.

Вначале, биомолекулы могут быть охарактеризованы только по общим физико-химическим или иммунологическим свойствам, на основании которых было произведено их выделение (идентификация). (Например, по молекулярному весу - 100 KD PROTEIN, номеру полосы геля электрофореза - 4.7 BAND PROTEIN, константе седиментации при ультрацентрифугировании - 5S RNA, или малоинформативному номеру клона антител к изучаемой биомолекуле - ANTIGEN X1). К этому может быть добавлено указание на источник происхождения биомолекулы (организм, ткань, тип клеток, клеточная органелла, субъединица клеточной огранеллы, стадия онтогенеза, включая даже такие технические характеристики, как, например, номер клона клеточной линии, и т.п.).

Затем, по мере изучения функциональных свойств описание биомолекулы может постепенно конкретизироваться. После выяснения полной структуры молекула может получить дополнительное имя. При этом, хотя окончательное имя и наиболее точно, в силу исторических причин и особенностей человеческой психики в качестве наиболее популярного может закрепиться самое первое, наиболее неопределенное (и иногда даже "неправильное") имя. Например, широко известный (названный в 1993 году "молекулой года") белок-супрессор онкогенеза обычно называют просто "p53", т.е. protein с молекулярным весом 53 kd (при этом интересно, что после выяснения полной первичной структуры оказалось, что его вес составляет всего лишь 44 kd (см. приложение в конце).

Альтернативным (к физико-химическому, иммунологическому и функциональному выявлению/выделению белков) и получающим сейчас все более широкое распространение подходом является тотальное секвенирование (определение структуры) генома с последующим компьютерным нахождением гипотетических генов. При этом функция кодируемых данными генами белков "вычисляется" по их структурному сходству с уже известными белками. В этом случае белок называют по имени наиболее близкого из известных аналогов с уже установленной функцией, добавляя модификаторы типа *hypothetical*, *probable*, *putative*, *homolog*, *-like*, *cognate* и т.п. (Понятно, что при этом в разных случаях степень близости может быть различной, так как она зависит от того, какие белки данного типа были известны в момент описания нового гена.)

Наличие многочисленных дупликаций генов, имеющих различную степень дивергенции, вносит дополнительную путаницу. Терминологические взаимоотношения между гомологичными молекулами в разных организмах также могут быть довольно запутаны. Названия белков одного типа (гомологов) в одном организме обычно различают с помощью добавления к основному термину номеров, которые могут совпадать или не совпадать в разных организмах (т.к. последовательность присвоения номеров может отражать хронологическую последовательность изучения белков этого типа в данном организме).

Кроме того, разные исследователи могут независимо друг от друга описывать одну и ту же (или очень близкую) биомолекулу (причем, изучая ее с помощью разных методов) и давать ей разные имена, степень синонимичности которых выясняется только впоследствии. Ревизии названий, периодически проводимые специальным комиссиями, не успевают за ходом развития науки, и иногда вносят дополнительную путаницу, т.к. в ходу все равно остается и старое, и новое название. Ведь в научной литературе, опубликованной до ревизии, название уже измениться не может, и исследователь все равно вынужден знать все (старые и новые) синонимы для изучаемых им биомолекул.

Дополнительные сложности вносят также:

1) случаи альтернативного сплайсинга, когда один и тот же ген может кодировать различные белковые молекулы;

2) посттрансляционные модификации белков и образование ими комплексов с другими соединениями;

3) четвертичная структура белка, когда одна (комплексная) функциональная единица белка формируется из продуктов разных генов, являющихся ее субъединицами (со своими названиями);

4) полипротеины и мультифункциональные протеины, в которых одна белковая молекула может выполнять функции нескольких белков (и соответственно включает все их названия);

5) "вложение" в название данной молекулы названия другой молекулы, с которым данная молекула взаимодействует (например: *inhibitor of serine protease*, *cytokine receptor*,*p53-binding protein*, *MHC class II regulatory factor*, *myosin heavy chain kinase*, *CD40L* – лиганд рецептора CD40, и т.д.).

Вследствие многоуровневости билогической организации простые термины общего вида могут иметь разные значения (например, слово *region* может означать «регион цепи данного белка», «регион гена/хромосомы, кодирующего данный белок», «регион другого белка, с которым взаимодействует данная биомолекула», и т.д.). Это может иметь место даже в случае более специализированных терминов (например, слово *promotor* обозначает как одну из регуляторных частей гена, так и вещество, способствующее развитию опухоли).

Специфика структуры биологических БД

В настоящее время имеется большое количество баз молекулярно-биологических данных. Они содержат, в основном, сведения о тех биополимерах (нуклеиновых кислотах и белках), для которых определена первичная структура (соответственно, нуклеотидная или аминокислотная последовательность). Кроме первичной структуры, эти базы содержат название биомолекулы, ссылки на научную литературу, в которой впервые была описана данная молекула и ее основные свойства (иногда также краткую аннотацию этих свойств), а также ссылки на другие базы.

Основным методом определения первичной структуры белка сейчас является секвенирование кодирующего его гена (на уровне ДНК или мРНК). Главными центрами

аккумуляции первичных данных по структуре генов являются GenBank (США), EMBL (Великобритания) и DDBJ (Япония). Эти базы имеют одно и то же содержание, так как они ежедневно обмениваются поступающей к ним информацией. (Обязательным условием публикации в научной литературе описания структуры нового гена является депонирование автором полученных данных в одной из этих баз.) Главным достоинством этих баз является их полнота (в настоящее время каждая из них содержит свыше 5 млн записей), главным недостатком – относительно низкий уровень обработанности содержащейся в них информации (мало аннотаций, высокий уровень дублирования – данные разных авторов об одних и тех же генах находятся в разных записях; другими словами, это "сырые" базы). Результаты автоматической трансляции (перекодирования) структуры генов в структуры белков содержатся в производных этих баз – GenPept (GenBank) и TrEMBL (EMBL). Существует также большое количество специализированных баз данных, посвященных отдельным группам белков или модельным организмам (т.е. организмам, геном которых был выбран для тотального секвенирования). Они могут быть лучше аннотированы, но охватывают только небольшую часть предметной области.

Одной из баз, удачно сочетающих полноту охвата с высокой степенью аннотированности, является база белков SwissProt, поддерживаемая швейцарским институтом биоинформатики (Swiss Institute of Bioinformatics, SIB). Она и была выбрана нами в качестве исходного материала для попытки автоматического построения тезауруса названий белков.

База данных SwissProt

SwissProt содержит свыше 80 тыс. записей (версия 38.0). Первичные данные для него берутся из TrEMBL, но дополнительно аннотируются кураторами SwissProt'a. Типичная запись SwissProt'a включает ряд полей, из которых для нас важны следующие:

AC P52439;

DE DNA POLYMERASE PROCESSIVITY FACTOR (POLYMERASE ACCESSORY PROTEIN)

DE (PAP) (PHOSPHOPROTEIN P41) (PP41).

GN U27 OR EPLF1.

KW DNA-BINDING; DNA REPLICATION; PHOSPHORYLATION.

SQ SEQUENCE 393 AA; 44810 MW; 239ADFF63F645D90 CRC64;

MCWSFHLFFK AHKARVGART SFLTEMERGS RDHHRDHRDH REHRETREPP TLAFHMKSWK
TINKSLKAFA KLLKENTTVT FTPQPSIIIQ SAKNHLVQKL TIQAECLFLS DTDRFLTKTI
NNHIPLFESF MNIISNPEVT KMYIQHDSDL YTRVLVTASD CTQASVPCV
HGQEVVRDTG RSPLRIDLDH STVSDVLKWL SPVTKTKRSG KSDALMAHII QVNPPTIKF
VTEMNELEFS NSNKVIFYDV KNMRFNLSAK NLQQALSMCA VIKTSCSLRT VAAKDCKLIL
TSKSTLLTVE AFLTQEQLKE ESRFERMGKQ DGKGDRSHK NDDGSALASK QEMQYKITNY
MVPAKNGTAG SSLFNEKEDS ESDDSMHFDY SSNPNPKRQR CVV

Поле ID содержит идентификатор (в данном случае - VPAP_HSV6U), состоящий из двух частей, первая из которых является идентификатором типа белка (VPAP), а вторая - вида организма, из которого он был выделен (HSV6U). Поле AC содержит идентификационный номер. Поле DE – название белка, причем часто там перечислены синонимы (в скобках). Поле GN включает названия генов, кодирующих данный белок. Поле KW содержит несколько ключевых слов, приписываемых данному белку аннотаторами SwissProt'a. Поле SQ содержит саму аминокислотную последовательность, закодированную с помощью 20-буквенного алфавита. Кроме того, обычно имеется еще поле комментариев (CC), описывающее основные свойства данного белка, что может быть полезно при ручном анализе тестовых наборов для оценки качества автоматических процедур.

Необходимо сказать также несколько слов о структуре терминов в поле DE. Хотя это и не "свободный" текст (как в поле CC), это и не controlled vocabulary (как в поле KW). В терминах поля DE, как правило, можно различить базовую (довольно жесткую и более информативную) часть термина и его дополнительные атрибуты-модификаторы. Они могут быть различного типа: указания на орган, ткань, клеточный компартмент или стадию развития, на которой экспрессируется данный белок (например, *brain, muscle,cardiac, erythroid-specific, acrosomal, embryonic*), его номер (для гомологичных белков), идентификатор белковой цепи в тех случаях, когда полной функциональной единицей является комплекс из нескольких цепей (например, *alpha chain*), возможность посттрансляционных модификаций (*precursor*), полноту определения первичной структуры (*fragment*), степень (не)определенности базовой части термина (*homolog, hypothetical, putative*), район хромосомы, в котором находится ген, кодирующий данный белок (*VNFA 5'region*), другие свойства (*essential, estrogen-preferring, inducible, insoluble*, и т.д.). Формат записи атрибутов и их расположение относительно основной части термина может варьироваться, что затрудняет автоматическое сравнение терминов в разных записях. Поэтому поле DE было предварительно пропущено через программу синтаксического анализа, чтобы, по возможности, выделить основную часть термина.

Последовательность применения алгоритмов описана в следующем разделе.

Процедура извлечения терминологии

Наиболее очевидным подходом было использование поля DE, в котором могут встречаться готовые ряды синонимичных терминов. Идентификация одинаковых терминов в разных записях и последующее включение остального содержимого полей DE этих записей в качестве синонимов могли бы обеспечить построение наиболее полных концептов (наборов синонимов - синсетов). Однако, скоро выяснилось, что вследствие двусмысленности некоторых терминов (в основном сложносокращённых слов[1]) при автоматическом построении происходит слияние совершенно разных понятий в один концепт (синсет). Так, например, аббревиатура PAP (*polymerase accessory protein*), фигурирующая в вышеприведенном примере, является также аббревиатурой для других, совершенно разных белков - *poly(a) polymerase*, *placental anticoagulant protein и*purple acid phosphatase (4 разных значения!), существует также термин PAP-C - *antiviral protein C precursor*. Другие примеры омонимии аббревиатур: MHC - *major histocompatibility complex* и *myosin heavy chain*; LRP - *low-density lipoprotein receptor-related protein* и *leucine-rich protein*. Эти термины являются ложными связующими звеньями между совершенно разными концептами. В результате, гомогенность синсетов, полученных методом автоматической идентификации терминов в полях DE разных записей, достигала всего лишь 33% (т.е. <35% синсетов из

проверенных вручную случайных выборок были "чистыми"). Поэтому пришлось подключить дополнительные методы.

В Табл.1 показаны размеры 10-ти самых больших синсетов, полученных в результате парсирования DE.

N концепта	1	2	3	4	5	6	7	8	9	10
Число терминов	2288	270	168	157	137	123	121	114	95	90

Таблица 1. Размеры 10-ти самых больших концептов.

Степень гетерогенности отдельных синсетов в первом приближении можно оценить, если учесть, что, в среднем, количество терминов на гомогенный ("чистый") синсет сейчас составляет 6-8 терминов.

Другим фактором, который, в противоположность парсированию DE, влиял на неоправданное увеличение количества синсетов, является сильная вариативность терминов, как в их структуре, так и в орфографии. Так, можно перечислить часто встречающиеся случаи вариативности[2]:

1. орфографические ошибки: *grows factor* вместо *growth factor*;
2. замена букв греческого алфавита на букву латинского или на римскую цифру (*alpha* просто буквой *A* или римской цифрой *I)*;
3. Непоследовательное использование пунктуационных знаков (‘,’; ‘-’; ‘/’; ‘(‘; ‘)’ etc) а также символа «пробел»:

Нормализованный термин	Варианты написания в корпусе
(2'-5')oligoadenylate synthetase	2',5'-oligodenylate synthetase
	2', 5'-oligodenylate synthetase
	2', 5-oligodenylate synthetase
	(2'-5')serum oligo(A) synthetase
	2-5 oligo A synthetase
	2'5' oligo (A) synthetase
	(2'-5') oligoadenylate (oligo (A] synthetase

4. Разный порядок следования элементов в термине без изменения смысла:

myosin heavy chain, fast skeletal muscle, embryonic «

muscle embryonic myosin heavy chain

myosin heavy chain, nonmuscle type a «

cellular myosin heavy chain, type a

Интересно, что здесь *nonmuscle* и *cellular* можно считать атрибутами-синонимами, но это корректно только для мышечных белков.

5. Непоследовательное использование римских и арабских чисел:

guanylate cyclase activator 2b « guanylate cyclase activating peptide II

Кроме того, в этом примере присутствуют разные формы выражения одного и того же общего понятия (подтермина) - *activator* и *activating peptide,* а также произвольный пропуск буквенного модификатора числа*.*

Таким образом, синонимичные термины не отождествляются друг с другом, и не могут образовывать правильных синонимичных множеств.

Становится очевидным, что использование только лингвистических методов работы с информацией в базе данных не может привести к удовлетворительным результатам.

Нелингвистические методы образования концептов

Кроме поля DE, информативными являются также поля ID, GN, KW и SQ (поле CC содержит "свободный" текст и поэтому сложно для автоматического анализа). После первых проб использование поля GN, содержащего названия генов, было отвергнуто, так как названия генов – это только аббревиатуры, с вытекающей отсюда высокой степенью омонимии. (К тому же поле GN имеется не во всех записях.)

Наименее омонимичной оказалась первая часть идентификатора в поле ID (в вышеприведенном примере для *polymerase accessory protein* это VPAP). (Вторую часть идентификатора, обозначающую название организма, использовать не было необходимости,

так как названия гомологичных белков из разных организмов должны попадать в один концепт.) Здесь, по-видимому, сказалось стремление составителей SwissProt'a, которые работают со всей базой белковых названий (в отличие от биологов-экспериментаторов, описывающих новые гены/белки только в своей узкой предметной области), избежать двусмысленности при выборе идентификатора. Из полученного множества концептов (а их получается несколько десятков тысяч), сформированных объединением терминов записей с тождественным ID (ID-сеты), были проанализированы несколько выборок. Степень гомогенности концептов существенно возросла и достигла 75%.

Однако, использование поля ID в качестве единственного критерия объединения терминов в синсеты недостаточно. Помимо недостаточно высокой точности, ID-сеты не всегда обеспечивали и необходимую полноту, т.к. некоторые синонимичные термины все-таки были обнаружены в разных ID-сетах. Поэтому мы подключили алгоритмы анализа пересечения разных записей по ключевым словам (точнее – терминам) поля ключевых слов KW и биологический метод – анализ сходства самих аминокислотных последовательностей белков с помощью алгоритма BLAST (Basic Local Alignment Search Tool; Altschul et al. 1990, 1997; Karlin et al. 1990, 1993).

Таким образом, вместе с анализом пересечения терминов в поле DE, у нас было четыре метода, причем биологический метод (BLAST) не зависит от двух лингвистических (ID, DE) и одного онтологического (KW).

Алгоритм вычисления близости синсетов по ключевым словам

Следует отметить, что термины в поле ключевых слов, являясь, как правило, понятиями более высокого/общего уровня, отражающими онтологию предметной области, обычно не совпадают с терминами в DE и идентификатором в ID, так что пересечение записей по ключевым словам является дополнительным методом, не зависящим от двух лингвистических.

Алгоритм пересечения записей по ключевым словам был использован для подтверждения/ опровержения разделения внутри ID-сетов, полученного после отождествления одинаковых терминов поля DE разных записей. Метрика определяла степень пересечения множеств ключевых слов разных DE-субсетов между собой. При этом, все ключевые слова были разбиты на классы, соответственно приписанным им "весам", которые определялись либо относительным местом ключевого слова в иерархии понятий (т.е. если слово более общее, например, *alternative splicing, multifunctional enzyme, signal* – низкий вес, более конкретное *elongation factor* – высокий вес), либо их значимостью (например, *3-D structure, hypothetical protein, multigene family, polymorphism* – не значимые, так как они не связаны со спецификой белка).

На рисунке 1 вверху показано распределение правильно и неправильно разбитых DE-сетов

внутри ID-сетов. Красным цветом показан процент правильно выделенных субсетов, синим – неправильно. Было установлено, что при анализе пересечения по KW оптимальной разрешающей способностью обладало пороговое значение, равное 50% - идентичности значений метрики на наборах ключевых слов сравниваемых записей.

Пример

В нижеприведенном примере ID-сет "CSP" разбит на 5 подмножеств по итогам парсирования терминов в поле DE . На самом деле здесь должно быть выделено только 3 разных подмножества, объединенных в один ID-сет вследствие омонимии аббревиатур, так как CSP является сокращением названий трех разных белков - *cysteine-string protein, circumsporozoite protein,* и *cold shock protein*. *И*з 5 получившихся подмножеств первый должен быть объединен с третьим (*cysteine-string protein*), а второй – с пятым (*circumsporozoite protein*). Они не объединились автоматически из-за вариативности написания терминов в поле DE (например, *cysteine-string protein* vs. *cysteine string protein*). Однако они были объединены автоматически с помощью алгоритма пересечения по KW, который показал совпадение 67% KW для первого объединения и 100% - для второго. Остальные попарные комбинации подмножеств вообще не имеют пересечений по KW (0%). Обращает на себя внимание тот факт, что названия генов (***GSP***) совпадают для *cysteine-string protein* и *cold shock protein*, что привело бы к неправильному объединению этих терминов по полю GN в случае использования такого алгоритма.

========== N 208 contains 25 articles CSP

Split up into 5 pieces.

CSP_DROME

* DE: CSP32/CSP29

* DE: CYSTEINE-STRING PROTEIN

* KW: ALTERNATIVE SPLICING

* KW: LIPOPROTEIN

* GN: CSP

CSP_PLASI CSP_PLARE CSP_PLABA CSP_PLAFO CSP_PLAMA CSP_PLAFT CSP_PLAVI CSP_PLACM CSP_PLACL CSP_PLACG CSP_PLACC CSP_PLACB CSP_PLAFW CSP_PLABE

CSP_PLAYO CSP_PLAKN CSP_PLAKH CSP_PLAFA

********** DE: CS

********** DE: CIRCUMSPOROZOITE PROTEIN PRECURSOR

********** KW: SIGNAL

********** KW: REPEAT

********** KW: SPOROZOITE

********** KW: MALARIA

* GN: CS

CSP_TORCA CSP_MOUSE

* DE: CCCS1

* DE: CSP

** DE: CYSTEINE STRING PROTEIN

** KW: LIPOPROTEIN

* GN: CSP

CSP_ARTGO

* DE: COLD SHOCK PROTEIN

* KW: ACTIVATOR

* KW: DNA-BINDING

* KW: TRANSCRIPTION REGULATION

* GN: CSP

CSP_PLABR CSP_PLAVS CSP_PLAFL

*** DE: FRAGMENT

*** DE: CIRCUMSPOROZOITE PROTEIN

*** DE: CS

*** KW: REPEAT

*** KW: SPOROZOITE

*** KW: MALARIA

KW-similarities between subsets:

1	0	0.666667	0	0
0	1	0	0	1
0.666667	0	1	0	0
0	0	0	1	0
0	1	0	0	1

GN-intersections between subsets:

1	0	1	1	0
0	1	0	0	0
1	0	1	1	0
1	0	1	1	0
0	0	0	0	0

Алгоритм вычисления близости синсетов по BLAST

В алгоритме BLAST критерием выбора является величина вероятности случайного совпадения блоков (участков) аминокислотных последовательностей в сравниваемых белковых цепях, поэтому в качестве порогового значения была принята стандартная для научных работ величина уровня статистической значимости ($P<0.05$). BLAST был использован, вместе с алгоритмом пересечения по KW, при окончательном объединении

фрагментов гомогенных синсетов (в том числе и из разных ID-сетов), которые образовались в результате неправильного первоначального разбиения.

Необходимо также отметить, что в современных базах молекулярно-биологических данных BLAST является единственным способом поиска белков (и генов), одинаковых или близких данному. Однако использование только его одного недостаточно для поиска синонимичных терминов по следующим причинам:

1) часть белков имеет значительные участки аминокислотных последовательностей т.н. "пониженной сложности" (low-complexity regions), которые состоят только из аминокислот определенного типа. Это обусловлено некоторыми физическими особенностями функций этих белков (или их участков). Участки "пониженной сложности" исключаются при BLAST-анализе с помощью специальных фильтров, чтобы предотвратить ложные сигналы сходства из-за случайного (т.е. необусловленного родственностью белков) совпадения аминокислотных последовательностей в них. А оставшиеся "сложные" участки могут иметь недостаточную для статистически достоверного вывода длину. В результате некоторые белки не дают достоверного сигнала совпадения даже сами с собой!

2) Аналогичная проблема возникает с белками, которые секвенированы не полностью (имеют слово fragment в поле DE). Длина таких фрагментов также может быть недостаточна для статистически достоверного вывода, хотя их запись содержит всю соответствующую терминологическую информацию.

3) В ходе эволюции некоторые белки могут утратить свою функцию (или приобрести новую) при лишь незначительном изменении аминокислотной последовательности. Это связано с тем, что ключевым для функциональной активности может быть только небольшой участок данной последовательности (например, каталитический центр фермента). Естественно, что изменение функции должно вести и к изменению названия белка, однако по BLAST-анализу он окажется близким к родственным ему белкам с исходной функцией. В таких случаях могут возникать достаточно противоречивые названия (оксюмороны), например, non-protease homolog of serine protease A, т.е. с помощью BLAST(а) этот белок классифицируется как протеаза, однако он не обладает протеазной активностью, а его новая функция (если она есть) еще не выяснена.

Таким образом, использовать для наших целей только BLAST не представлялось возможным.

Полученные результаты

Полученный в результате тезаурус содержит около 18,000 концептов и около 65,000 терминов. Создание такого тезауруса вручную потребовало бы многие месяцы работы специалистов.

В результате комбинированного применения всех четырех методов при построении тезауруса удалось обеспечить как точность[3], так и полноту свыше 95% (как показали результаты ручного тестирования случайных выборок, а также результаты независимого тестирования заказчика).

Нерешённой в полном объёме проблемой пока остается обработка записей для полипротеинов и мультифункциональных протеинов (>1000 записей), поскольку в этих случаях в поле DE содержатся термины (и синонимы) для всех компонентов таких функционально сложных белков. Эти термины являются практически свободными фразами на английском языке, которые следует анализировать с учетом их синтаксической структуры. Они значительно усложняют обработку записей, загрязняя полученные на их основе концепты (синсеты). В настоящее время разрабатываются методы для обработки и соотнесения частей таких терминов. Однако, количество таких записей относительно невелико (меньше 1,5%), поэтому они не сильно влияют на общее качество результата.

Работа над пополнением тезауруса в настоящее время продолжается. Необходимо отметить, что главным акцентом является, по-прежнему, полностью автоматическая обработка исходного материала. Это позволяет возложить на компьютерные программы большую часть работы, которую раньше делали люди.

Учитывая описанные в статье недостатки лингвистических и биологических методов, следует отметить, что наш уникальный подход с параллельным использованием взаимодополняющей лингвистической (названия белков), отнтологической (KW) и биологической информации (аминокислотные последовательности), является в настоящее время наиболее полным и может быть использован не только для создания тезауруса, но и для усовершенствованного поиска в самих молекулярно-биологических базах. (В настоящее время нами уже разработана такая система, SwissSearch, для поиска в базе SwissProt.)

Из задач, решить которые предстоит в ближайшее время, особого внимания требуют следующие:

синтаксический разбор номенклатуры полипротеинов;

алгоритмы восстановления полной формы сложносокращенных слов из контекста, в котором аббревиатуры вводятся ;

построение таксономии концептов тезауруса: синонимические ряды будут объединены в цепочки отношений вида «IS A» и «часть - целое»;

извлечение информации из CC (комментарии) поля записи. Эта задача подразумевает синтаксический разбор текста на английском языке. В результате будет извлечено дополнительное знание об объектах (белках), что позволит нам производить группирование объектов по нетерминологическим признакам (например, по функциональным, или по месту нахождения в тканях).

Литература

Crawford L (1983) The 53,000-dalton cellular protein and its role in transformation. Int. Rev. Exp. Path. 25: 1-50.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman (1990). Basic local alignment search tool. J. Mol. Biol. 215:403-10.

Karlin, Samuel and Stephen F. Altschul (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA 87:2264-68.

Karlin, Samuel and Stephen F. Altschul (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. Proc. Natl. Acad. Sci. USA 90:5873-7.

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

Приложение 1. Образец записи базы данных SwissProt

ID P53_HUMAN STANDARD; PRT; 393 AA.

AC P04637;

DT 13-AUG-1987 (Rel. 05, Created)

DT 01-MAR-1989 (Rel. 10, Last sequence update)

DT 15-JUL-1999 (Rel. 38, Last annotation update)

DE CELLULAR TUMOR ANTIGEN P53 (PHOSPHOPROTEIN P53).

GN TP53.

OS Homo sapiens (Human).

OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Mammalia;

OC Eutheria; Primates; Catarrhini; Hominidae; Homo.

CC -!- FUNCTION: ACT AS A TUMOR SUPPRESSOR IN MANY TUMOR TYPES. INDUCES

CC GROWTH ARREST OR APOPTOSIS DEPENDING ON THE PHYSIOLOGICAL

CC CIRCUMSTANCES OR CELL TYPE, BUT BOTH ACTIVITIES ARE INVOLVED IN

CC TUMOR SUPPRESSION. IT ACTS IN CELL CYCLE REGULATION, IT IS A

CC TRANS-ACTIVATOR THAT ACTS TO NEGATIVELY REGULATE CELLULAR DIVISION
CC BY CONTROLLING A SET OF GENES REQUIRED FOR THIS PROCESS. ONE OF
CC THE GENES ACTIVATED IS AN INHIBITOR OF CYCLIN-DEPENDENT KINASES.
CC APOPTOSIS INDUCTION SEEMS TO BE MEDIATED EITHER BY STIMULATION OF
CC BAX AND FAS ANTIGEN EXPRESSION, OR BY REPRESSION OF BCL-2
CC EXPRESSION.
CC -!- SUBUNIT: IN VITRO, THE INTERACTION OF P53 WITH CANCER-ASSOCIATED/
CC HPV (E6) VIRAL PROTEINS LEADS TO UBIQUITINATION AND DEGRADATION OF
CC P53 GIVING A POSSIBLE MODEL FOR CELL GROWTH REGULATION. THIS
CC COMPLEX FORMATION REQUIRES AN ADDITIONAL FACTOR, E6-AP, WHICH
CC STABLY ASSOCIATES WITH P53 IN THE PRESENCE OF E6.
CC -!- SUBCELLULAR LOCATION: NUCLEAR.
CC -!- PTM: PHOSPHORYLATED EXCLUSIVELY ON SER RESIDUES IN A CELL CYCLE-
CC DEPENDENT MANNER.
CC -!- PTM: DEPHOSPHORYLATED BY PP2A. SV40 SMALL T ANTIGEN INHIBITS THE
CC DEPHOSPHORYLATION BY THE AC FORM OF PP2A.
CC -!- DISEASE: P53 IS FOUND IN INCREASED AMOUNTS IN A WIDE VARIETY
CC OF TRANSFORMED CELLS. P53 IS FREQUENTLY MUTATED OR INACTIVATED
CC IN ABOUT 60% OF CANCERS.
CC -!- DISEASE: DEFECTS IN P53 ARE ALSO THE CAUSE OF GERMLINE CANCERS
CC SUCH AS LI-FRAUMENI SYNDROME (LFS). LFS IS AN AUTOSOMAL DOMINANT
CC FAMILIAL CANCER SYNDROME THAT IN ITS CLASSIC FORM IS DEFINED BY
CC THE EXISTENCE OF BOTH A PROBAND WITH A SARCOMA AND TWO OTHER
CC FIRST-DEGREE RELATIVES WITH A CANCER BY AGE 45 YEARS. IN THESE
CC FAMILIES THE AFFECTED RELATIVES DEVELOP A DIVERSE SET OF
CC MALIGNANCIES INCLUDING BREAST CARCINOMAS, SARCOMAS, AND BRAIN

CC TUMORS AT UNUSUALLY EARLY AGES.

CC -!- DISEASE: VARIANT ALA-143 IS TEMPERATURE SENSITIVE. AT 32.5 DEGREES

CC CELSIUS IT POSSESSES STRONG DNA BINDING ABILITY, BUT AT 37.5

CC DEGREES CELSIUS ITS TRANSCRIPTIONAL ACTIVITIES ARE GREATLY

CC REDUCED.

CC -!- DISEASE: DEFECTS IN P53 ARE ALSO THE CAUSE OF BARRETT'S

CC ADENOCARCINOMAS (BA). BA IS A CONDITION IN WHICH THE NORMALLY

CC STRATIFIED SQUAMOUS EPITHELIUM OF THE LOWER ESOPHAGUS IS REPLACED

CC BY A METAPLASTIC COLUMNAR EPITHELIUM. THE CONDITION DEVELOPS AS A

CC COMPLICATION IN APPROXIMATELY 10% OF PATIENTS WITH CHRONIC

CC GASTROESOPHAGEAL REFLUX DISEASE AND PREDISPOSES TO THE DEVELOPMENT

CC OF ESOPHAGEAL ADENOCARCINOMA.

CC -!- DISEASE: DEFECTS IN P53 ARE THE CAUSE OF HEAD AND NECK SQUAMOUS

CC CARCINOMAS (HNSC) AND ORAL SQUAMOUS CELL CARCINOMAS (OSCC).

CC CIGARETTE SMOKE IS A PRIME MUTAGENIC AGENT IN CANCER OF THE

CC AERODIGESTIVE TRACT.

CC -!- SIMILARITY: BELONGS TO THE P53 FAMILY.

CC -!- DATABASE: NAME=HotMolecBase; NOTE=p53 entry;

CC WWW="http://bioinformatics.weizmann.ac.il/hotmolecbase/entries/p53.htm".

CC -!- DATABASE: NAME=IARC p53;

CC NOTE=IARC db of somatic p53 mutations;

CC WWW="http://www.iarc.fr/p53/homepage.htm".

CC -!- DATABASE: NAME=Tokyo p53;

CC NOTE=University of Tokyo db of p53 mutations;

CC WWW="http://p53.genome.ad.jp/".

CC -!- DATABASE: NAME=Prague p53;

CC NOTE=University of Prague db of germline p53 mutations;

CC WWW="http://www.lf2.cuni.cz/win/projects/germline_mut_p53.htm".

CC ---

CC ---

DR EMBL; X02469; CAA26306.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]

DR PIR; A25224; A25224.

DR PIR; A25397; A25397.

DR PIR; B25397; B25397.

DR PIR; JT0436; JT0436.

DR PDB; 1A1U; 08-APR-98. [ExPASy / RCSB]

DR SWISS-3DIMAGE; P53_HUMAN.

DR TRANSFAC; T00671; -.

DR SWISS-2DPAGE; P04637; HUMAN.

DR GeneCards; TP53.

DR MIM; 191170; -.

DR MIM; 151623; -.

DR PRINTS; PR00386; P53SUPPRESSR.

DR PROSITE; PS00348; P53; 1.

DR PFAM; PF00870; P53; 1.

DR PRODOM [Domain structure / List of seq. sharing at least 1 domain]

DR BLOCKS; P04637.

DR DOMO; P04637.

DR PROTOMAP; P04637.

DR PRESAGE; P04637.

KW Anti-oncogene; DNA-binding; Transcription regulation; Activator;

KW Nuclear protein; Phosphorylation; Apoptosis; Disease mutation;

KW Polymorphism; 3D-structure.

SQ SEQUENCE 393 AA; *43653 MW*; AD5C149FD8106131 CRC64;

MEEPQSDPSV EPPLSQETFS DLWKLLPENN VLSPLPSQAM DDLMLSPDDI EQWFTEDPGP

DEAPRMPEAA PPVAPAPAAP TPAAPAPAPS WPLSSSVPSQ KTYQGSYGFR LGFLHSGTAK

SVTCTYSPAL NKMFCQLAKT CPVQLWVDST PPPGTRVRAM AIYKQSQHMT EVVRRCPHHE

RCSDSDGLAP PQHLIRVEGN LRVEYLDDRN TFRHSVVVPY EPPEVGSDCT TIHYNYMCNS

SCMGGMNRRP ILTIITLEDS SGNLLGRNSF EVRVCACPGR DRRTEEENLR KKGEPHHELP

PGSTKRALPN NTSSSPQPKK KPLDGEYFTL QIRGRERFEM FRELNEALEL KDAQAGKEPG

GSRAHSSHLK SKKGQSTSRH KKLMFKTEGP DSD

//

[1] Далее все варианты сложносокращённых слов мы будем называть аббревиатурами.

[2] Этот список ни коим образом не претендует на полноту, а лишь иллюстрирует часто встречающиеся случаи вариативности.

[3] Здесь мы используем русскоязычные переводы терминов "recall" (полнота) и "precision" (точность), которые используются для оценки эффективности систем информационного доступа