様式 C-19

科学研究費補助金研究成果報告書

平成２２年　６月　７日現在

研究種目：特定領域研究
研究期間：2005～2009
課題番号：１７０１７０３７
研究課題名（和文）　生命科学辞書とオントロジーの自動構築法の開発

研究課題名（英文）　Statistical method to represent biomedical knowledge
研究代表者
大久保　公策（OKUBO　KOUSAKU）
国立遺伝学研究所・生命情報・DDBJ 研究センター　教授
研究者番号：４０２３３０６９

研究成果の概要（和文）：分野の知識骨格を代表的な教科書が持つ用語およびトピックが作る構造で代表させることで、
１）用語間の意味関係　および２）文書内容の意味関係を統計的に求め　また理解しやすい形で表現する方法を開発しました。
算出された用語の意味関係が全体として表現する知識は
（１）生命科学のwebページ、電子文書やデータベースレコード等の並べ替え
（２）記述的分析データを用いて行う探索的研究での専門知識を動員した解釈の補助
などで利用することを想定しています。応用例として任意の文書の整理や遺伝子特徴のデータの解釈を実行するテストwebサーバーを開発しました。

研究成果の概要（英文）：On assumption that "the distribution of 1) technical terms and 2) enlisted topics in a Textbook" is a measurable representation of a domain knowledge, We have applied latent semantic indexing technique to the index section of a textbook to make a vector space of words and pages.
In this space, the semantic relation of terms and topics are expected to be represented by the cosine between the term vectors as well as page vectors.　The authenticity, sensitivity and specificity of the resulted space was psychologically tested by applying interpretation of gene orders took from genome, expression clusters, as well as in adhoc clustering of PubMed abstracts retrieved by a common key word .
Quantitative evaluation of the results which is essential for tuning parameters is a problem still left behind.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2005 年度	11,800,000	0	11,800,000
2006 年度	12,400,000	0	12,400,000
2007 年度	12,500,000	0	12,500,000
2008 年度	12,700,000	0	12,700,000
2009 年度	12,900,000	0	12,900,000
総　計	62,300,000	0	62,300,000

研究分野：生物学
科研費の分科・細目：基礎生物学　遺伝・ゲノム動態
キーワード：ゲノム、情報工学、マイクロアレイ、遺伝子、生体生命情報学

１．研究開始当初の背景
分子生物学や遺伝学の手法で進める基礎医学領域では「疾患サンプル」の遺伝子発現解析やゲノム解析がよく行われています。ゲノム配列や遺伝子発現などの分子の分析法が機械化されることで速度や網羅性を飛躍的に増大すると、サンプルや遺伝子に関する専門知識を動員したデータの解釈や整理をいかに高速に行うか、すなわちどのように計算機を利用するかが課題になっています。

分野の知識表現のひとつである専門用語の持つ意味を計算機で扱える形に表現する方法には二つのアプローチがあります。
一つはオントロジーに代表されるように人間が宣言的に行う方法です。
宣言的方法では意味関係は概念の上下関係をつないだ人間が理解しやすいグラフ構造（階層分類（木）やDAG）や表現します。一番近い上下関係にある用語同士をリンクさせることを繰り返して得られます。
いまひとつは文書をデータとして用語の分布から統計的に計算する方法です。
統計的方法では用語の文書間などでの出現パターンの距離で用語の意味の関係を表現しようとする方法でパターンを距離に直す手法が様々検討されています。

基礎から臨床までのすべての生命科学領域をカバーした用語集にはNLMのMeSHが知られています。MeSHはNLMが扱う文献に対するインデックス名称として集められた用語であり、実際に文書で使用される用語との重複は必ずしも多くありません。
また基礎生命科学領域では特に遺伝子の機能を表現し生物種を超えた知識の整理を可能にする道具としてGeneOntologyが広く使われています。
すなわち両者とも宣言的な手法であり、広汎な医学知識を有効に表現する統計的手法は見られません。
これら宣言的な手法はいわばマンパワーの投下によって一定の水準での目的を達成しています。
しかし索引用語集と構造の絶え間ない更新と新規文書や新規遺伝子への絶え間ない索引付けを必要とし、文書の爆発と分野の知識の更新や再編成に常に対応できる機械化可能な方法が必ず必要であると考えました。

一般に統計的手法も宣言的手法もそれぞれに長所と短所が知られています。
統計的に行うとデータとして利用した文書に構造が依存することや出来上がった関係が人間に理解しにくい、また手作業で後から修正しにくいことが問題です。
宣言的に行うと当然宣言者に依存し、同じ宣言者でも再現性の低い問題があります。
また枝の距離は対象分野の認識の解像度を表すために出来上がった構造は局所的な意味関係しか表せない（大きな距離には意味がない）という問題があります。一般的には大規模に多様な内容を扱う場合は統計的に、枯れた限定分野を精緻に行うには宣言的に行うことが適当です。生命科学は両者のちょうど中間的な位置にあります。

２．研究の目的
　生命系全体では解析的に研究を進めたために理解が時代とともに知識を表現するための概念が詳細化してゆき構造や概念の名称が階層的になっている特徴があり、このせいで宣言的な用語の構造化になじむと考えられています。しかし近年の分子医学では膨大な数の分子名称から疾患や症状に関わる名称を扱います。これらを同じ宣言ルールですべて宣言的にカバーすることは容易ではないと考えられます。
本計画では用語の意味関係を自動的に構造化する方法の開発をめざし同時にその利用によって評価を試みます。

３．研究の方法
タスクを二つもうけ宣言的分類と統計的手法の両者を行いました。
タスク１．　ESTライブラリ情報からの効率的解剖用語の収集と半自動分類
タスク２．　教科書を利用した専門用語の収集と構造化

タスク１.ESTライブラリ情報からの効率的解剖用語の収集と半自動分類

DDBJ/INSDCのESTデータを対象に数万を数えるライブラリ材料に関する記述を集め、それを自動的に解剖学的に分類するパタン辞書を開発しました。パタン辞書の再生を計算機を用いた再帰的な手作業で半自動的に行いました。
臓器名称が自動的に検出可能でその臓器の系統分類にしたがって離散会合操作することが可能になれば蓄積している多種類の生物の遺伝子発現データを使って異なる生物の遺伝子発現パタンを比較することが可能です。
遺伝子間の関係はオーソログ情報を利用すれば遺伝子の臓器別の発現様式のファミリーメ

ンバー間での発現パタンの変化などを調べることが可能になり、遺伝子発現の進化についての洞察を得ることができます。
たとえば現在広く知られている大野進氏の重複遺伝子生成仮説ではひとつの遺伝子が重複するときに重複と同時に発現パタンも分割されるので重複したコピー遺伝子が生き残るという説明があります。この説明はいくつかの臓器特異的に進化したパラログファミリーをよく説明するが多くの遺伝子の例で確認されてはいません

タスク２　教科書を利用した専門用語の収集と構造化

臨床医学から基礎医学まで医科学系では解剖学および疾患名称や病理学名称が特に多種類の用語を利用しています。近年これらの医学概念が遺伝子やたんぱく質のレベルで詳細に研究されています。
ここでの遺伝子や蛋白質と臓器や疾患との関係は上位下位の概念関係や部分全体の関係ではなく、トピカルリレーションと呼ばれる場面と役者の関係です。
すなわちテニスボールとラケットや野球のボールとバット の関係はオントロジーでよく使われる上位下位（is a）や部分全体（part of）では記述できません。これらの関係がtopical な関係でこれは競技について記述した文書などで簡単に検出可能です。
このように医科学の文書や遺伝子の機能説明の意味的な操作や分類のためには上位下位や部分全体だけでなくトピカルな関係を測れる用語の分布を用いた統計的な分類が適当です。
文書を材料に用語の関係を作る方法は用語の意味を登場する文書の組み合わせで表現するベクトル空間法が代表的です。
文書の内容を全用語数の次元のベクトル（文書ベクトル）と考えれば、同じ文書に共起の多い用語ベクトルが大きな内積（余弦）を与え全く共起のない用語同士は直行する（余弦ゼロ）という直感的な関係を与えます。
本計画でもこのベクトル空間法を用います。

４．研究成果
タスク１．ESTライブラリ情報からの効率的解剖用語の収集と半自動分類

1)ライブラリ情報を集めるために、EST-DDBJエントリーをライブラリ単位にまとめ、それぞれのライブラリの中のから材料に関する記載のあるフィールドを集める
2)分類効果の大きなルールにするためにEST数の大きなライブラリからトップ100のライブラリの該当フィールドを見て、臓器単位に分類する
3)分類の根拠を与えたパターンを集めて正規表現のパターン辞書を作る
4)パターン辞書に同じ１００のライブラリ情報を与えて、誤回答の例につきアドホックな例外集を作る
5)確認後のこりのすべてのライブラリ情報を処理する
6)分類不能であったライブラリから大きなものを100選んで手作業で分類しパタン辞書に加える
7)分類不能が１割以下になるまで３）から６）を繰り返す
分類機を他の哺乳類にあてはめすべてのESTデータを種ｘ臓器分類で整理しなおす。
のステップを繰り返し
http://bodymap.jp/organ_tissue_rule
に記載されているような解剖学名称パタン辞書を得ました。
以降下のような経緯でパタン辞書の公開とデータの更新を行ってきました。
現在は５０種類の生物の６００種類のライブラリ由来の3500万件のESTレコードを４０臓器に分類して公開しています。設計後５年を経過した分類機ですがヒトのライブラリの分類力はいまだに80％を超えておりクリーン数では90％以上が分類できています。開発後の維持修正作業のリストは以下のようです。

2008-05-12　*DDBJ release73 processed
12 new animals　1918583 new ESTs
2007-04-19　*Gene name search: Gene symbol support DDBJ release68 processed
 2 new animals　1196172 new ESTs
2006-09-06 *DDBJ release66 processed
  1 new animal　600597 new ESTs
2006-06-04　*DDBJ release65 processed
  4 new animals 3720010 new ESTs

さらに同じ分類機をSAGE・EST・マイクロアレイ・iAFLPの測定法に適用し、異なるプラットフォームでの発現パタンを比較するビューアーサイトを作成しました。
http://okubolab.genes.nig.ac.jp/bodymap_i/
それぞれの方法が与えるパタンを比較すると相互の不一致に一定の傾向を見出した。
すなわち遺伝子それぞれにライブラリ別の濃度を与えその最大値を使って遺伝子を発現検出のしやすさで容易・中等度・困難3分類すると特に困難な部分で著しくに３つの方法由来の発現パターンに一致が見られないことが明らかになった。
困難部分ではマイクロアレイはハウスキーピングパターンが大半を占めるが、タグ同定法では逆にどこかの組織に偏る傾向が強く、iAFLP法ではその中間でした。

タスク２　教科書を利用した専門用語の収集

と構造化

臨床医学から基礎医学まで医科学系では解剖学および疾患名称や病理学名称が特に多種類の用語を利用しています。近年これらの医学概念が遺伝子やたんぱく質のレベルで詳細に研究されています。
ここでの遺伝子や蛋白質と臓器や疾患との関係は上位下位の概念関係や部分全体の関係ではなく、トピカルリレーションと呼ばれる場面と役者の関係です。
すなわちテニスボールとラケットや野球のボールとバット の関係はオントロジーでよく使われる上位下位（is a）や部分全体（part of）では記述できません。これらの関係がtopical な関係でこれは競技について記述した文書などで簡単に検出可能です。
このように医科学の文書や遺伝子の機能説明の意味的な操作や分類のためには上位下位や部分全体だけでなくトピカルな関係を測れる用語の分布を用いた統計的な分類が適当です。
文書を材料に用語の関係を作る方法は用語の意味を登場する文書の組み合わせで表現するベクトル空間法が代表的です。
文書の内容を全用語数の次元のベクトル（文書ベクトル）と考えれば、同じ文書に共起の多い用語ベクトルが大きな内積（余弦）を与え全く共起のない用語同士は直行する（余弦ゼロ）という直感的な関係を与えます。
本計画でもこのベクトル空間法を用います。

しかしながらこの方法を生命科学のように専門用語が多く同義や類義などの関連にみちた語彙を持つ分野にそのままあてはめると以下のような問題点があります
１）集める文書には同義語や表記ゆれ類義語がみちておりそのまま機械的にインデックスしたのでは類似用語の関係も検出できない。
２）文書の集合によって用語の共起関係などがことなり用語集合の分類に再現性がない
３）得られた類似用語を特徴付ける方法がタググクラウドなどの用語による方法以外に友好なものがない
以上の問題を克服するために
１）データとする文書集合として教科書を利用する
２）教科書内教科書間での類義語関係を検出するためにLatent Semantic Analysis (LSA) を用いる
の二つのアイデアを提案しました。

アイデア１） a. 文書集合として教科書を用いると教科書の内容は非常にトピックの粒度が調整されており、内容の重複や不足など任意の文書集合から出発するよりも有利であること
b. 教科書では既に著者が重要であると考えた用語の選択による索引付けが終わっていること。
これは複数の単語でつくられる用語のどこまでを意味単位であると認めるか、いわゆるチャンキングの問題を回避できる。
c. 教科書は解剖学や生理学など分野によってわかれており、それぞれの小分野の専門用語を分離採取可能でしかもその分野の体系に従った意味関係を検出可能である
d. LSA（後述）では文書（ここではページ）も用語と同じ空間にマップされるので単語の内容が目次を使えば項目見出しでも表現できるなどの利点が考えられます。
アイデア２） LSAは９０年代にベル研究所で開発されたベクトル空間法による高感度な文書の検索方法です。その中心となっているアイデアは用語ｘ文書のスパースなデータに用語間の共起に現れている意味的な類似性や関連性を反映しようというものです。データ行列ではすべて無関係として扱っている用語AとBでも文書中でよく共起すれば意味的に近いと考えAだけが使われている文書にも弱くBでも索引付けしてあげるという方法です。データ内の用語パタンをさらにデータに当てはめるという少し難しい計算を
はじめのスパースなデータ行列を次元を下げて近似した行列に誤差としてあらわれるもともとは０であった多数のセル中の値が数学的にも用語間の共起度数を反映した索引値に等しいことが証明されています。
LSAでは用語の意味も文書の内容も同じ空間のベクトルとして扱うために教科書の索引データによる用語の構造化に利用できるはずです。

以上のアイデアに基づいて以下のような方法を考案しました。
１）分野別に教科書データを用いて十分な数の用語とそれぞれ独立な内容であるはずの教科書の項目によるベクトル空間を作成する。
教科書１５０冊合計５万ページ強の目次および索引データを収集し管理データベース化しました。教科書リストは
http://222.151.240.4/project/vbob/nonsense_dic/bob_textbook.txt
にあります。
２）教科書を分野別に分類しひとつの分野が数冊の教科書の和のデータで構成できるようにします。
教科書索引には同義語を括弧つき表記で表現したものや A See B で表記したものがあり、これらから同義語辞書を作成します。
３）A gene, A protein, A などほぼ同義語として使われる用語バラエティーを作る接尾単語のリストを作成し、教科書間の索引のマージを進めます
４）パッケージプログラムを用いて用語ｘペ

ージ行列を次元下げします。
この際後の計算の負担を減らし実際によく使われている50次元での近似を行う。いわばあらゆるページ内容は独立の50項目の線形和で近似できるという過程である。
５）適当なプログラムライブラリを用いてSingular Value Decomposition を行い、近似行列を得ます
得られた近似行列はそのまま用語とページの関係を与えます。
６）SVDで作った用語ベクトルは任意の文書の内容の表現のためにDB化して保持します。
７）近似行列が与える用語とページの関係をオリジナルの索引と比較して望ましい潜在的関係が得られているかを確認します。
８）似パターンを与える用語を目視でチェックして同義語の検出を行い、十分量検出したのちにステップ３の同義語辞書に加え５以下を繰り返します。
９）５０次元の用語意味ベクトルとページ意味ベクトルが出来上がります。
ここまでで用語の意味の形式表現が完了します。

用語の意味的表現を利用した文書内容の表現
任意の医学文書は教科書用語で索引付けします。
１）文書内容は索引付けされた用語をあらわす用語ベクトルの線形和として表現します。用語の重み付けには標準的なTF/IDF
（あるふれた用語は小さく、その文書だけで繰り返す用語は重く）
２）文書同士の関係は意味ベクトルの余弦で与えられます。
３）文書内容のわかりやすい表現には文書ベクトルと教科書ページベクトルの余弦によって得られる教科書の目次項目との類似度パターンで得られます。
４）文書に与えられたベクトルを利用して文書間のすべての余弦を計算してクラスタリングを行えば文書群の内容によるクラスタリングを行えます。
５）出来上がった文書クラスターの意味づけは教科書項目との類似性パタンを用いて行います。
６）特定の文書に類似する内容の文書は余弦の小さい順に序列化し類似文書をクラスタリングすることで類似する観点別に分解することも可能です。

遺伝子機能の形式表現

遺伝子の機能表現を文書として扱います。ここではEntrez
Geneで遺伝子に付与されている文献アブストラクト、複数あればそれをつないだ文書を作り、使用しました。文書と全く同じ方法で、遺伝子機能をベクトル表現できます。

表現の妥当性の検討（方法全体の評価）
遺伝子の機能表現が妥当であるか否かの、またGOなどに比べ優れたところがあるかの検討は以下のように行いました。
①遺伝子ファミリーの機能表現の共通性およびファミリー内の差異の表現力
②遺伝子発現データの特徴解釈の論文との比較

①はPfamなどのDBで与えられている進化的な遺伝子ファミリーのグループ別に形式表現を受けた遺伝子（教科書のページとの類似パターン）を並べパタンの共通な部分および異なった部分について知られている知識を再現しているかを目視する。GOAの場合にはファミリーはたいてい同じGOが与えられていましたが時に本法では組織特異性が表現できていることがありました。
②は論文などからマイクロアレイデータを取得し遺伝子発現のパタン順に遺伝子列を作り、その順に教科書ページパターンを与えられた遺伝子を並べ替えました。
または遺伝子間の類似性行列を色調の濃淡表示であらわした遺伝子ｘ遺伝子の機能距離行列を作成して、同じように並べ替えると、機能的に類似した遺伝子が発現データによってクラスターされている場合には濃意部分が作る四角形が対角線上に見えました。

著作権上問題のある目次や索引データがそのままDLできないように画像で表現したテストサイトを以下のURLに構築しました。
計算が膨大であり更新やサーバー維持のためのコストも相当額であるために長期維持は不可能です。
BOB（publicly open: http://www.ebob.jp/）
アカウントパスワードは作業の保存の為ですので自由に設定できます。

＜国内外での成果の位置づけ＞
比較すべきものは　文献情報を利用して遺伝子を関係化する様々のサービスであろう。
PubGeneに代表される遺伝子ネットワークの抽出が代表である。一見すると遺伝子を構造にする点で類似しているが動作原理はまったく違う。教科書をフレームワークにして文書を構造化し構造の部分を教科書の表題で表現するような試みは幸いまだない。
一方で検索結果をアドホックにクラスター化してそれぞれを名づけるやりかたは単純な検索結果の序列化や固定的なサイテーション情報などでの関係化の次の方向となりつつある。この点では教科書を選んで好きな視点から無構造の文書をグループ化する本法は類似しています。

宣言的オントロジー：　専門家が一人もしくは小集団で自らの理解に基づいてドメイン知識を概念関係などで表現する手法は過去１０年の間に生命科学系では広く受け入れられてきた。特にテキストマイニングやセマンティックウエブなどの大量の文書を対象に、検索したり操作したりする分野との連携の容易さからほの分野よりも盛んに言及、利用される道具として定着しています。
その記述形式は標準化され、統一され、それを踏まえたフリーの編集ソフトは多数作成されて公共の利用に供されている。また多くのチームが作成したオントロジーは遺伝子オントロジーの創始者であるAshburnerらが主催するOpenBioMedicalOntologieなどのサイトに集積され検索利用が容易になっている。
動的クラスタリング：あらかじめ外部から基準を与えることなく検索で得られた文書群を相互の類似性で分類する方法では得られた文書クラスタの名づけが課題となっている。

＜今後の展望＞
知的構造体を任意の情報整理に利用するアイデアの独創性は特許成立（P2005-259088A）によって認められました。適当な条件で教科書データが利用可能になれば実用化公開は可能です。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）
〔雑誌論文〕（計６件）

①Ogasawara O. and Okubo K.
On theoretical models of gene expression evolution with random genetic drift and natural selection. (peer reviewed)
PLoS One. 4(11), e7943. (2009)

②Hoshino H, Uchida T, Otsuki T, Kawamoto S, Okubo K, Takeichi M, Chisaka O.: Cornichon-like Protein Facilitates Secretion of HB-EGF and Regulates Proper Development of CranialNerves. (peer reviewed) Molecular Biology of the Cell Apr Vol.18 D1143-1152(2007).

③ Ogasawara O, Otsuji M, Watanabe K, Iizuka T, Tamura T, Hishiki T, Kawamoto S, Okubo K.: BodyMap-Xs: anatomical breakdown of 17 million animal ESTs for cross-species comparison of gene expression., Nucleic Acids Res.,34(Database issue):D628-31 (2006)

④Itoh K, Kawasaki S, Kawamoto S, Seishima M, Chiba H, Michibata H, Wakimoto K, Imai Y, Minesaki Y, Otsuji M, Okubo K., Identification of differentially expressed genes in psoriasis using expression profiling approaches. (peer reviewed)
Exp Dermatol., 14(9):667-74, (2005)

⑤Kaimori JY, Takenaka M, Okubo K.
Quantification of gene expression in mouse and human renal proximal tubules. (peer reviewed)
Methods Mol Biol., 293,209-19.(2005)

⑥Tanino, M., Debily, MA., Tamura, T., Hishiki, T., Ogasawara, O., Murakawa, K., Kawamoto, S., Itoh, K., Watanabe, S., de Souza, SJ., Imbeaud, S., Graudens, E., Eveno, E., Hilton, P., Sudo, Y., Kelso, J., Ikeo, K., Imanishi, T., Gojobori, T., Auffray, C., Hide, W., Okubo K. (peer reviewed)
The Human Anatomic Gene Expression Library (H-ANGEL)
Nucleic Acids Res. 2005 Jan 1;33(Database issue):D567-72.

〔学会発表〕（計０件）

ホームページ等
① BOB(試験公開サーバー)
http://222.151.240.6/project/bob/080208-/simple.cgi
②　CGED Source Classification Database
http://cged.genes.nig.ac.jp/scd/
③ H-Angel
http://www.jbirc.aist.go.jp/hinv/h-angel/wge.top
④　BodyMap-Xs
http://bodymap.jp

６．研究組織
(1)研究代表者
大久保　公策（OKUBO KOUSAKU）
国立遺伝学研究所・生命情報・DDBJ 研究センター・教授
研究者番号：４０２３３０６９