計画研究：2000～2004年度

遺伝子発現を基盤とする病態機序の解明法の開発

●大久保公策

国立遺伝学研究所生命情報・DDBJ研究センター

〈研究の目的と進め方〉

発現データの病態解析への利用には①病変と対象の比較を行い病変で何が起こっているのかを推測する。②上記を通じて病変のキーとなる遺伝子を同定する。③QTL解析などゲノム上の位置情報とつき合わせて候補遺伝子の選別を行う　などの方法が考えられる。　本研究では①個別の研究者が独自に生産された物に加えて　公的に存在するデータが利用できる仕組み　②遺伝子発現データの解釈の手助けをしてキー遺伝子の同定や　病態の解釈を助ける仕組み　③既存の知識を動員して候補遺伝子の選別を助ける仕組み　を開発することを目的とした。

〈研究開始時の研究計画〉

実際の病理サンプルを用いてのモデル的な発現解析実験および解釈のための仕組みの設計を計画していた。
病理サンプルは

〈研究期間の成果〉

解釈の仕組みBOB構築

BOBはあらゆる分野の専門的な文書を内容に基づいて高次元に構造化する方法であり、また文書内容を直感的に理解可能に表現するための方法である。論文アブストラクトはいうまでもなく、遺伝子、蛋白質、薬品、疾患など基礎医学を構成する概念の機能や特徴は様々な形でサマリーや関連文献集としてまとめられている。したがってこれらを利用することで遺伝子機能、薬品効能、疾患症状すべてがBOBでの構造化や表現の対象となる。ひとたび内容に形が与えられればマイクロアレイデータの解釈から文献検索結果の複数の文献のアドホックなクラスター化、ゲノム座標と特定の疾患の関係の表現などその様々な応用が期待できる。

考え方

専門性の高い内容を「理解する」という行為を「分野を構成する少数の基本的な話題へのマッピング」でモデル化する。このとき基本話題を当該分野の教科書の内容で代用すると専門性の高い話題の内容は類似する内容の教科書ページの組み合わせ（正確には全ページへの類似度のパターン）として表現される。さらに当該教科書の目次を利用することで類似度のパターンは直感的に理解可能である。すなわち類似ページパターンは機械にとって計算に使える形であるのみならずヒトにとっても理解可能な形をしている。

文書と教科書ページの類似性の測定法

教科書の各ページの内容と与えられた専門文書の内容の類似度測定にはLatent Semantic Indexing（ＬＳＡ）(1990 ベル研)を用いる。LSAはキーワード検索では検出できないような内容の関連が強いが当該用語〔群〕を含まない文書の検索法である。同法は辞書やシソーラスに全く頼らずに検索対象となる文書群中での用語の分布パターンデータだけから背後にある用語同士の意味関係を見出し各文書の索引付けに用いるものである。具体的には各文書中の用語の出現回数で作られる用語 x 文書行列〔通常非常に疎な行列〕を百次元程度の低い次元の行列で近似表現し、そのときに生じる誤差で非出現用語と当該文書との関係を表現する方法である。この方法では全く出現していない用語が他の用語との関連から索引されることや文脈と無関係に出現した用語との関係を低く評価することが期待される。オリジナルデータが用語と文書の表面上の関係(formal relation)であるのに対し、この100次元行列が用語の意味と文書の内容の意味的な(semantic) relation を与えると考えるのである。

今文書同士の類似性を共通な索引用語の数と定義し、用語と文書の意味関係の強さは当該索引用語の文書中での頻度と定義する。すると(m,n)行列ではセルの値が用語と文書の関係、列同士の内積が文書同士の類似性を与える。次元下げによって得られたk次元の(m,n)行列を用いればそれぞれセルの値と列の内積のそれぞれは用語同士の意味が完全に独立でないことを考慮に入れた、より鋭敏な文書と用語の関係、文書同士の関係とみなすことが出来る。

文書の(m,n)次元のデータ行列を固有値分解して得た(m,k)x(k,k)x(k,n)の(m,k)行列で与えられるm個および(k,n)行列であたえられるn個のk-次元ベクトルをそれぞれ“用語の意味ベクトル”、“文書の意味ベクトル”とみなすと、k次元の用語と文書のベクトルの余弦は (m,n)データ行列の当該セルの値、文書ベクトル間の余弦は(m,n)データ行列の列同士の内積と等しいため　このベクトル空間は文書および用語の意味的関係を表現した空間とみなすことが出来る。

ＢＯＢにおいては教科書の索引から作成した用語 x ページ行列をもとに用語の意味およびページの内容を表現した100次元の教科書空間を作成した。

ＬＳＡでは質問（文書）は内包する用語ベクトルの和として空間中にマップされ質問とより小さな余弦を作るベクトルが関連文書として返される。同様にＢＯＢは質問は用語ベクトルの和として教科書空間中に表現され教科書空間の他のベクトルとの意味関係は余弦として与えられる。一方ＬＳＡは検索目的の方法で空間は自由な構成の無構造な文書群で構成されているのに対してＢＯＢでは教科書という内容に重複の少ない分野全体を均等にカバーしある程度階層的に構造化された文書群から空間を作成している。教科書の利用は文書内容が分野全体をカバーすることに加えユーザーに対する意味ベクトルの内容の表現において重要である。つまりＢＯＢ空間の任意のベクトルの意味を教科書の全ページとの余弦として表現すれば教科書の目次を利用してその内容が利用者に簡単に伝達することが可能である。

教科書空間作成作業

基礎医学分野における数千のページの平均的な教科書の索引には数千の用語がリストされている。ＢＯＢでは索引をデータとして使用する。これにより複数の単語で意味をなす用語の検出が不要でより正確なページと用語の意味関係を取り出すことが可能である。教科書の索引では多くの場合大見出しの下に小見出しが複数存在し、小見出しにのみページが与えられている。BOB作成時には扱う用語の意味の粒度があまり細かくならないように小見出しはすべて大見出しに書き換えて使用する。これによって同一ページに複数回同じ大見出しが索引されることが生じるがこれを重みとして扱い用語〔大見出し〕ページ行列を作成する。

基礎医学分野は解剖学、組織学、細胞生物学、生化学、遺伝学、生理学、遺伝学、病理学、薬理学など用語や話題の大きく異なる小さな分野に分割されている。BOBでは用語に対する感度を上げるためにすべての小分野から複数の教科書を選び、すべての索引から大きな用語 x ページベクトルを作成する。　このとき教科書内部では用語の統一が行われているが教科書間では用語の使用が不統一である問題が生じる。そのために複数の索引のマージには注意を要する。具体的には記号の処理、数字の処理、複数形の処理、gene/protein/molecule/factor/disease/syndrome の処理、taxonomic nameの処理などを行い、さらにシソーラスを用いて用語のマージを注意深く行う。シソーラス〔同義語関係〕も索引中にある”See…”や”See also”行を持ちいて教科書から作成する。これらの処理によって出来た用語をbase formと呼び、ひとつのbase form にはそれぞれ索引中に出現したままの用語の形を持たせておいて文書へのマッピング時に使用する。この際に形の異なる索引用語がおおよそ70%のbase formへとコラプスされる。Base from x 教科書のページ行列から　自由な教科書の組み合わせを選択していくつかの小分野特異的な用語 x ページデータを作成することも可能である。現在までに100冊以上の教科書索引をデータ化しており、うち10冊程度について重点的に目視による校正作業が行われている。

用語 x ページ行列(数万 x 数万)は線形台数ライブラリプログラムを使用して100次元の用語行列 x 特異値対角行列 x ページ行列の積に分解する。用語ベクトルの座標とページベクトルの座標はそれぞれ左、右の行列として与えられる。

表 1　BASIC空間作成に利用された教科書

01	Basic Histology, A Lange medical book. 7th ed.
17	A textbook of histology. 10th ed.
05	Harper's Biochemistry. 25th ed.
08	Biochemistry and Molecular Biology.
10	Molecular Cell Biology. 4th ed.
11	The Cell. 2 nd ed.
14	Cell and Molecular Biology. 2 nd ed.
54	Molecular Biology of The Cell. 4 th ed.
07	Developmental Biology. 5 th ed.
12	Principles of Development.
13	Human Embryology & Developmental Biology
02	Genes VII.
09	Human Molecular Genetics. 2 nd ed.
19	Genes and Genomes.
15	Fundamental Neuroscience.
25	GOODMAN & GILMAN'S THE PHARMACOLOGICAL BASIS OF THERAPEUTICS
119	THE BASIC SCIENCE OF ONCOLOGY
03	Robins Pathologic Basis of Disease. 6th ed.
18	Essential Pathology

質問内容のマッピング

出来上がった教科書空間には任意の医学文書の内容がマップ可能である。まず用語には人遺伝子全体やpubMed4年分などの大きな文書群で計測された文書群での分布の程度(document frequency)が与えられる。次にそれぞれの文書中でその頻度もカウントする。カウントには様々な注意を要するが基本的に辞書のマージと同じ方法を用いる。そして頻度を用語の珍しさで補正した値、たとえばTF/log 2 DF が用語の係数となり、それぞれの用語のk-次ベクトル座標に係数をかけた和としてk次空間にマップされる。

遺伝子のマッピング

遺伝子は分子データベースを使用して文書で表現する。たとえばEntrez gene には遺伝子毎に機能のサマリーと根拠論文のリストが掲載されている。サマリーおよび根拠論文のアブストラクトをつなぎ合わせると十分な長さの遺伝子文書が出来上がる。

BOB サーバーの構成は以下のごとくである

用語データベース：教科書の索引から作成する用語ごとのページ情報、用語同士の同義語関係の同義語辞書、アクロニム辞書の更新や修正を管理する用語データベース

教科書空間データ：基礎、臨床、解剖など解釈の視点に対応する教科書セットそれぞれに対して用語関係を計算した空間をあらわす特異値分解後行列データとして得られる100次元の用語座標データおよびページ座標データ

オブジェクトデータベース：NCBIgeneを用いて遺伝子機能を文章で表現したものに対してすべての用語をマップした結果を遺伝子：文章：マップ用語ID、用語頻度、用語ウエイト、すべての空間における座標などを管理するデータベース、PubMedアブストラクト200万件も同様の情報とともに管理している。

BOB web サーバーの利用法

1．空間を選ぶ：〔オブジェクトの関係を図る視点として使用する教科書セット〔空間〕を選ぶ。BASIC空間は教科書25冊37000用語、42000ページがつくる空間。解剖学、生化学、などの空間は25冊のなかからそれぞれの分野の教科書だけを選択して作った空間であり、同じオブジェクトセットでも空間によって当然相互の類似性は大きく異なる。

2．オブジェクト群を与える

遺伝子発現では発現クラスタ化された遺伝子列、pubMedの場合はキーワードなどで検索されたアブストラクトID群

3．オブジェクトの並べ替え

発現解析の場合は不要、無構造なオブジェクト群の場合は視認性を高めるために空間の座標で並べ替える

4．空間を構成する教科書から特定の一冊を選び全ページに対するオブジェクトの余弦を教科書ページの順

に表示する。教科書ページ x オブジェクト余弦の行列の図示

II. 遺伝子発現情報の統合

ESTを自動的に組織別に分類して行く仕組みをつくった比較統合の難しさ、手法間でのデータの食い違いの多さから当初の期待に反して公的マイクロアレイデータの再利用は進んでいない。一方でESTは700万を超える件数の多さや絶対量に関する情報を与えるなどの利点から発現データとして再度注目されている。このESTを発現情報として利用するためには個別エントリーをライブラリ単位にかため　各ライブラリのRNA情報を読み解いて臓器や細胞単位に分類する作業が必要である。NCBIに用意されているUniLiB情報に頼らずにこれが行える仕組みを作成して　常に最新のdbESTを遺伝子 x 臓器のBodyMapフォーマットに整理公開できるようにした。　具体的には、1．エントリーの様式をもとに同じ登録者の同時登録を判別しライブラリ単位にまとめる。2．臓器オントロジーとも呼べる　臓器細胞名称および形容詞の辞書を作成し、エントリ中の特定フィールドの記述中にある用語を40の基本臓器名称にマップする。3．この40の基本臓器の特徴表を作成し自由にライブラリが離散会合できる。の3つの要素機能を作成した。この仕組みには当然SAGEtagやマイクロアレイデータを載せることも可能である。この仕組みを持ちいてマイクロアレイ、SAGE、EST、iAFLPなど正常臓器に対するすべての遺伝子発現データをひとつのテーブルにまとめて表示可能にしたものをH-Angelとして開発し公開した。さらに同じパタン辞書をすべての動物の臓器をおよその臓器の相同関係を用いて同様に分類できるように拡張したAnatomy-taggerを開発し1700万の動物ESTに適用しBodyMap-Xsとして公開中である。これは

〈国内外での成果の位置づけ〉

BOB:「 遺伝子クラスタの自動解釈」はBOB開発開始後ひとつのバイオ情報処理のテーマとなっている。大きくわけて（1）遺伝子機能を木構造で宣言し、それに遺伝子をマップするオントロジーとオントロジーキーワードの遺伝子へのマップというクラシックな専門家による手法　（2）論文アブストラクト中での遺伝子名称の共起をつかって自動的に遺伝子を構造化する方法。（3）遺伝子同士の関係を論文や総説中に探し手作業で結合や促進など限られた関係として表現する方法が存在する。BOBはこのいずれにもあたらない独自の方法である。（1）にはGO,GOA（2）はPuGene　（3）はKEGG、BINDなどが存在する。検出したい遺伝子の関係としては　相互に共同するケース〔遺伝子：遺伝子関係〕、同じ臓器で発現〔遺伝子：臓器〕、同じ疾患に関与〔遺伝子：疾患〕、発生の同じ時期に発現〔遺伝子：発生〕、など多様な関係が期待される、これに対して（1）－（3）の方法では極めて限られた関係以外は表現不可能である。〔表現力〕また（1）（3）では常に専門家による書き足し書き換えが必要である。さらに遺伝子以外の対象の構造化に対してはほとんど無効である。

遺伝子発現統合：遺伝子発現データはNCBIのGEOに集積されているが、生物別、プラットフォーム別に分離されているために異なるプラットフォーム間での遺伝子発現の比較、異なる種類の動物の間でのオーソログ遺伝子の発現パタンの比較を提供しているものはBodyMapのみである。本研究で作成した材料名称自動分類プログラムの類似物や同様の目的の研究も現在は存在しない。このプログラムはマイクロアレイ、EST、SAGE、PCRなど方法によらずにまた正常組織、癌組織などを区別せずにあらゆる遺伝子発現データの記述を自動的に分類整理可能である。遺伝子方向での対応を自由にとれる仕組みと組み合わせればつねに世界中の遺伝子発現データを自動的に整理し続けることが可能である。

〈達成できなかったこと、予想外の困難、その理由〉

病態解析に発現データを利用する場合のサンプル側の問題を正しく評価することが出来なかった。すなわち同じタイプの乳癌と呼ばれるサンプルにも、病理型以外に壊死の割合、浸潤細胞の割合、正常部分の混在の割合など想像以上のばらつきが存在しており、同じ病理診断の病理組織をマスとして用いる意味は極めて疑問である。一方でレーザー顕微鏡などでの切片からの選択的回収も操作上の誤差が相当に大きいことが予想される。癌ばかりでなく代謝疾患や高血圧などでもその患者組織由来のRNAを用いた発現解析はそのデータの厚みを増す形よりは、注意深く集められた一連のサンプルからの発見にのみ期待が持たれる。

BOBの公開が最大の未達成課題である。これはWeb経由での利用目的に作成してきたGUI側での表示の速度の問題および数施設の協力で行っているテスト時に複数ユーザーの同時使用による速度低下の問題による。現在のBOBシステムでは2種類の教科書空間のみに使用を限定して、マップ対象はヒト遺伝子全体(Entrez gene) および2001-2004 PubMed abstractにのみ対応している。これらの200万件の文書オブジェクトはあらかじめ100次元ベクトル化されてそこから選ばれた質問オブジェクトと指定された教科書の全ページへの余弦データをユーザーのブラウザ上で表示しなければならないためにたとえば1000のオブジェクトに対しては1000x3000程度の数値を転送し縮退させて表示しなければならない。またその自由な部分にたいして自由にズームし移動できなければならない。

〈今後の課題〉

遺伝子発現データからの疾患病態の理解については明らかに自らデータを生産するより既存のデータの統合利用が有効と考えれらる。したがって今後データについては公的データの自在な利用を可能にする仕組みの開発を続ける。遺伝子方向の統合はすでに困難ではないが遺伝子発現データはスプライスパタンが複雑な例が多く知られているためにエキソン単位での発現データ比較を可能にしなければならない。これを可能にするためにuniGene単位の遺伝子をrefSeq, cDNA, genome データを用いてexon単位に分解し、SAGEtag　EST、マイクロアレイプローブなどをすべてエキソンにマップし、比較表現できる仕組みを作成中である。遺伝子クラスターの解釈は現状ではパスウェイ単位への変換が最も有効であるために手作業で同定されたKEGGやGOのパスウェイデータが好まれている。しかし遺伝子と疾患の関係にかんしては遺伝病以外に体系的に整理されたものがなく、高感度な関係を扱うような開発を考えている。具体的には疾患関連の教科書を選定することで疾患に対する感度をあげ　Entrezgeneにリストされていない報告を収集することで遺伝子疾患情報を豊富にする計画である。一方でPubMedやwebpageなどの数が膨大で人手で十分な索引付けが不可能な対象は医学的な構造化が自由ではない。BOBはその自動性オントロジー非依存性を生かしこれらの対象へのアプリケーションを開発しその存在を示すこ

とが必須である。現在PubMed検索語のアドホッククラスタリングの仕組みとして再開発をしている。

〈研究期間の全成果公表リスト〉

1：Kawamoto S, Yoshii J, Mizuno K, Ito K, Miyamoto Y, Ohnishi T, Matoba R, Hori N, Matsumoto Y, Okumura T, Nakao Y, Yoshii H, Arimoto J, Ohashi H, Nakanishi H, Ohno I, Hashimoto J, Shimizu K, Maeda K, Kuriyama H, Nishida K, Shimizu-Matsumoto A, Adachi W, Ito R, Kawasaki S, Chae KS. Matsubara K, and Okubo K. BodyMap: A collection of 3' ESTs for analysis of human gene expression information. Genome Res. 2000 Nov;10(11):1817-27

2：Sese J, Nikaidou H, Kawamoto S, Minesaki Y, Morishita S, Okubo K. BodyMap incorporated PCR-based expression profiling data and a gene ranking system. Nucleic Acids Res. 2001 Jan 1;29(1):156-8.

3：Masuda N, Tamaki Y, Sakita I, Ooka M, Ohnishi T, Kadota M, Aritake N, Okubo K, Monden M. Clinical significance of micrometastases in axillary lymph nodes assessed by reverse transcription-polymerase chain reaction in breast cancer patients. Clin Cancer Res. 2000 Nov;6(11):4176-85.

4：Ogasawara O, Kawamoto S, Okubo K. Zipf's law and human transcriptomes: an explanation with an evolutionary model.C R Biol. 2003 Oct-Nov;326(10-11):1097-101.

5：Kawasaki S, Kawamoto S, Yokoi N, Connon C, Minesaki Y, Kinoshita S, Okubo K. Up-regulated gene expression in the conjunctival epithelium of patients with Sjogren's syndrome.
Exp Eye Res. 2003 Jul;77(1):17-26.

6：Sakai R, Kinouchi T, Kawamoto S, Dana MR, Hamamoto T, Tsuru T, Okubo K, YamagamiConstruction of human corneal endothelial cDNA library and identification of novel active genes. Invest Ophthalmol Vis Sci. 2002 Jun;43(6):1749-56.

7：Nakajima H, Takenaka M, Kaimori JY, Nagasawa Y, Kosugi A, Kawamoto S, Imai E, Hori M, Okubo K. Gene expression profile of renal proximal tubules regulated by proteinuria. Kidney Int. 2002 May;61(5):1577-87.

8：Takebayashi H, Ohtsuki T, Uchida T, Kawamoto S, Okubo K, Ikenaka K, Takeichi M, Chisaka O, Nabeshima Y. Non-overlapping expression of Olig3 and Olig2 in the embryonic neural tube. Mech Dev. 2002 May;113(2):169-74.

9：Tanaka S, Tatsumi K, Okubo K, Itoh K, Kawamoto S, Matsubara K, Amino N. Expression profile of active genes in the human pituitary gland. J Mol Endocrinol. 2002 Feb;28(1):33-44.

10：Saito K, Tanaka T, Kanda H, Ebisuno Y, Izawa D, Kawamoto S, Okubo K, Miyasaka M. Gene expression profiling of mucosal addressin cell adhesion molecule-1+ high endothelial venule cells (HEV) and identification of a leucine-rich HEV glycoprotein as a HEV marker. J Immunol. 2002 Feb 1;168(3):1050-9.

11：Soejima H, Kawamoto S, Akai J, Miyoshi O, Arai Y, Morohka T, Matsuo S, Niikawa N, Kimura A, Okubo K, Mukai T. Isolation of novel heart-specific genes using the BodyMap database. Genomics. 2001 May 15;74(1):115-20.

12：Tanino M, ——, Auffray C, Hide W, Okubo K.
The Human Anatomic Gene Expression Library (H-ANGEL)
Nucleic Acids Res. 2005 Jan 1;33(Database issue):D567-72.

13：Tateno Y, Saitou N, Okubo K, Sugawara H, Gojobori T.
DDBJ in collaboration with mass-sequencing teams on annotation. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D25-8.

14：Michibata H, Chiba H, Wakimoto K, Seishima M, Kawasaki S, Okubo K, Mitsui H, Torii H, Imai Y. Identification and characterization of a novel component of the cornified envelope, cornifelin. BBRC. 2004 Jun 11;318(4):803-13.

15：Imanishi T,---- Okubo K, Wagner L, Wiemann S, Strausberg RL, Isogai T, Auffray C, Nomura N, Gojobori T, Sugano S. Integrative annotation of 21,037 human genes validated by full-length cDNA clones.
PLoS Biol. 2004 Jun;2(6):e162. Epub 2004 Apr 20.

16：Chiba H, Michibata H, Wakimoto K, Seishima M, Kawasaki S, Okubo K, Mitsui H, Torii H, Imai Y.
Cloning of a gene for a novel epithelium-specific cytosolic phospholipase A2, cPLA2delta, induced in psoriatic skin.
J Biol Chem. 2004 Mar 26;279(13):12890-7. Epub 2004 Jan 6.

17. DB
http://bodymap.jp
http://www.jbirc.aist.go.jp/hinv/h-angel/wge_top.cgi

18 特許
注目する情報を知識集積物との関係で可視的に処理するためのシステム
特願 2004－73614