3C6-1

生命科学分野の実験手法による論文の分類とオントロジーの利用に関する研究

Study of classification of life science research based on wet laboratory experiment.

川本祥子[*1] 荒木次郎[*2] 藤山秋佐夫[*3] 菅原秀明[*4] 大久保公策[*4]
Shoko KAWAMOTO Jiro ARAKI Asao FUJIYAMA Hideaki SUGAWARA Kousaku OKUBO

[*1] 情報・システム研究機構 新領域融合研究センター
Transdisciplinary Research Integration Center, Research Organization of Information and Systems

[*2] (株)三菱総合研究所 Mitsubishi Research Institute Inc.
[*3] 国立情報学研究所 National Institute of Informatics
[*4] 国立遺伝学研究所 National Institute of Genetics

In the post genomic era, to study gene and protein function and understand the life, the cooperation between wet-lab experiment and *in silico* analysis becomes more important. We therefore propose a new approach to effective use of description about wet method in text mining and database integration.

1. はじめに

2003 年に発表されたヒトゲノムの解読完了後もゲノム解読のスピードは衰えることなく，モデル動物をはじめ，霊長類，植物や家畜，微生物など多数の生物のゲノム解読がなされ，生命の全体像理解への期待が高まっている．しかし，ゲノム配列はこれから解明しようとする遺伝子レベルの問題について研究者が共有する地図がようやく得られたに過ぎず，各地点にマップされた遺伝子を十分に定義するためには，今後も生物を材料とするウェット系の実験が必要である．観察が主体であると言われる生命科学ではあるが，膨大なデータと知見が蓄積されるに至った今、過去の論文やデータベースからのフィードバックをいかにうまく取り入れて実験を実施するかが重要な鍵を握っている．まずは論文やデータベースを検索し，同じような実験結果は無いか，参考になる実験結果は無いか，もしあればそれをもとにより先のモデルや仮説を組み立てられるのである．インターネットの普及によって，誰もが同じように論文やデータベースにアクセス可能な時代になり，生命科学者の研究環境は過去に比べると飛躍的に良くなったと言える．しかし実際にはデータベースを利用しようにも，ゲノム解析や網羅的計測から生み出されるデータは非常に膨大で，不統一かつ難解なデータベースに蓄えられているため，誰もが自在にあやつれるような状況では無い．同じように年間 50 万件以上も発表されるという論文の中から関連論文をくまなく見つけ読解することは容易では無い．このような状況の中，ライフサイエンス分野においても，テキストマイニングの重要性は増すばかりである．実際に遺伝子発現解析に Gene Ontology[1]を用いる有効性が実証されたり，また論文から遺伝子間のネットワークを推定するなど，人手ではできない高度な知識発見が計算機で可能となりつつある．

2. 何故機械解釈は受容されにくいのか

このように情報科学から生命科学へのアプローチとしてデータや論文の機械解釈が先行する反面，一般的な生命科学研究者の率直な感想としては，テキスト及びデータマイニング技術によって自分が読んでいない論文や全体像の見えないデータの固まりから抽出された知識を与えられても，その内容を承諾するには何かが欠けているように思われるというのが実際のところだと思われる．このため，あるはずの分子が登場していないとか，間違った分子が関連付けられているなど，精度上の欠点にばかり注目されてしまうようなことも聞かれる．なぜ現状の機械解釈されたデータが理解されにくいかについて大久保ら[2]の論文にわかりやすく解説されているので，ここでは，辻井[3]によって指摘されている『言語によって表現された知見は，現在のところ，人間が個別の論文を逐一読むことによってしか消費できない．ある論文に現れた知見と別の知見とが関係するとしても，その相互関係は，2 つの論文全体を読んだ読者にしか認定できない』をその障壁の一つとして取り上げるが、これは研究の成果を表現することも，また理解することも，ひとえに論文を中心としてきた生命科学研究者の特質と言えるべきものである。複数の論文を横断的に理解する過程で，知識を蓄え，再利用し、モデルを構築し，実験と関連付けることが研究にとって非常に重要なウェイトを占めているのである．しかしこの従来の方法のみに頼っていては量産されたデータや大量の論文を処理できないことは明らかである．

3. 実験目的とプロセスのデータベース化

そこで我々は，結果の理解のためには個々の知見についての研究目的やプロセスの情報が不足しているのではないかと考え，論文で用いられている実験手法と目的を抽出整理し，実験を中心とした研究の索引付けとモデル化を行うこととした．実験手法に精通した専門家は，研究の目的と，用いた材料，実験手法を提示されれば，どのようなデータが得られるかおおむね判断することが可能であり，目的と実験手法の索引は様々な形式で提供される結果の選別に利用できることが期待される．

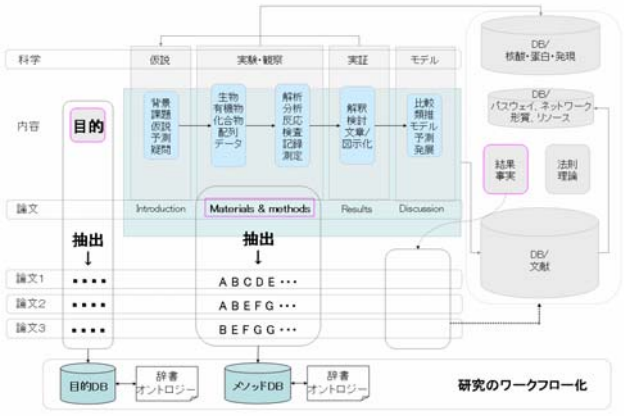

図1 論文からの目的と実験手法の抽出

連絡先：川本祥子，情報・システム研究機構新領域融合研究センター
〒101-8430 東京都千代田区一ツ橋 2-1-2 国立情報学研究所
藤山研究室，skawamot@nii.ac.jp

図1に示すように実験の目的と手法の部分の抽出を，「遺伝子発現」の分野から選んだ数百報の文献に対して手作業で行い，データベースに蓄積されているエントリーとの関連付けを図り，自動化を検討する．目的の抽出は従来どおり論文の要旨を対象とするが，現在，生命科学分野の論文のほとんどは全文の入手が可能となっており，実験手法については，本文中の Materials and Methods に記載される順序にもとづいて，ワークフローとして抽出するものとした．

3.1 実験手法に関する用語辞書の整備

論文から実験手法名を抽出し，相互に参照するためには実験名を標準化しなければならない．しかし実験手法名は Microarray analysis などのように固有の名称がつけられているものもあれば，Preparation of Plasmid DNA by Alkaline Lysis (日本語ではアルカリ溶解法)のようにフレーズで表現されるものもあり，様々な表記ゆれが存在する．MeSH (Medical Subject Heading) 等既存の辞書やシソーラスだけでは不十分なため，実験書の索引などから用語の収集整理を行っている．核酸を対象とする実験分野では，実験書 Molecular Cloning(Cold Spring Harbor Laboratory Press)と，Current Protocols(Wiley)から用語の抜き出しを行った．実験の種類は約 350 種で関連用語は 6000 語程度である．収集した実験名の問題点として，表記ゆれの他に，一つの実験手法が複数の実験手法で組み立てられているという粒度の問題が存在し，後述するような分類体系を利用して階層的に整理していく手段が必要である．今後は核酸以外の実験分野についても同様に収集を行う予定である．

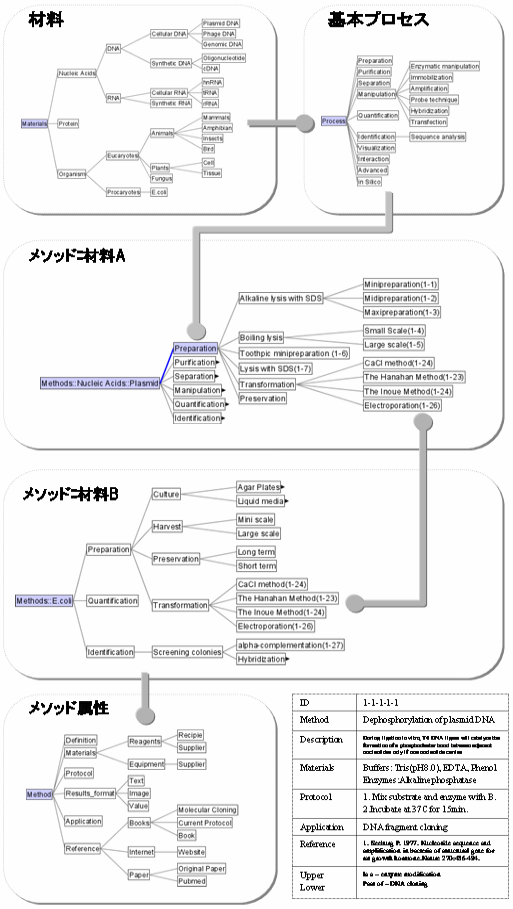

図 2　実験手法名の分類整理体系

3.2 実験手法の分類体系の構築

一般的に実験名の分類は，分野別や対象物別，あるいは解明しようとする生命現象ごとの分類が同一階層に配置されたものが多い．ほとんどの実験は複数のサブ実験の組み合わせで成り立っているため，カテゴリーの作成時に直感的分類が選択されてしまうことが原因であると考えられる．そこで我々は図 2 に示すように，実験手法を材料と基本のプロセス(抽出する，分離する，修飾する，計量するなど)の組み合わせによって表現することで，体系的分類を試みた．辞書作製用に収集した Molecular Cloning の実験手法名をこの体系で分類し，実験手法の属性として材料の他に必要な試薬，手順，結果のタイプ等を記載する．同じ実験が複数のツリーに分配される場合を許し，一つの実験が複数の実験の複合である場合でも，この分類方法で概ね矛盾無く分類することが可能であるが，項目間の関係性などをより整理する必要がある。

4. 今後の課題

辞書等が整備できればある程度自動で論文から実験手法名を抽出することも可能となるが、現在は手作業でそれを行っており、図3に論文から抽出した遺伝子発現解析の代表的なワークフローを示す。最近ではバイオインフォマティクスの解析をもと

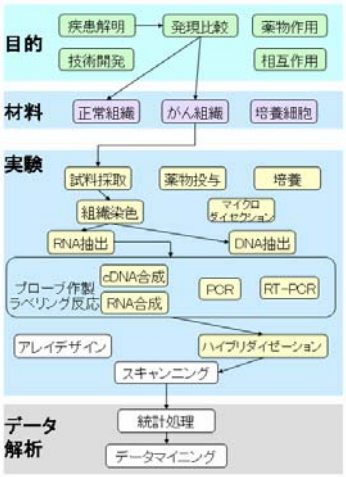

図 3 遺伝子発現解析のワークフロー

にウェットの実験が組み立てられることも多い。荒木ら[4]は論文からバイオインフォマティクスのメソッドを抽出しワークフローを構築、実際のサービスに応用する研究を進めており、今後はドライとウェットを統合したシステムの研究を行っていく予定である．

参考文献

[1] Gene Ontology: http://www.geneontology.org/
[2] 大久保公策，日紫喜光良: ゲノムデータの機械解釈，会誌「情報処理」，社団法人情報処理学会 , 2005.
[3] 辻井潤一: テキストから知識・情報へ:生命科学を題材にして,「2003 年情報学シンポジウム」オンライン論文集，情報処理学会 , 2003.
[4] 荒木次郎，川本祥子，藤山秋佐夫，菅原秀明，大久保公策，武田英明: 文献からのバイオサイエンス研究手法の収集・整理による研究支援セマンティック Web サービスの実現，第 21 回人工知能学会全国大会，1D1-06 , 2007.