知識処理技術を用いた生命システムの再構築とその解析

●高木 利久　◆中谷 明弘
東京大学大学院新領域創成科学研究科

＜研究の目的と進め方＞

生命をシステムとして理解するためには、ゲノム配列や蛋白質立体構造だけでなく、発現、局在、相互作用、パスウェイ、ネットワーク、表現型などに関するデータおよびそれらの間の関係や生物学的な制約や文脈などに関する知識などを計算機上に統合し、その性質、特徴、振る舞い、などを解析することが不可欠である。そこで本研究では、(a) 表現型情報と種々のゲノム情報の統合および知識発見技術の開発による生命システムの構造解明、(b) 医学・生物学文献から知識やその根拠となった実験事実などを抽出し利用する技術の開発、(c) 複雑な生物知識の表現法およびそれらの比較解析や検索のための手法の開発、の３つのテーマについて研究を展開する。進め方は以下の通り。

(a) 本領域研究において、多くの生物について遺伝子破壊実験などにより表現型に関する機能情報の蓄積が行われつつある。当該情報と、下記 (b) において抽出予定の関係性ネットワーク、細胞内局在情報、遺伝子発現情報などを統合したデータベースを構築するとともに、これからの知識発見技術を開発する。これにより遺伝子間の複雑なネットワークの構造を明らかにする。

(b) 蛋白質 / 遺伝子や化合物だけでなく、生物学的機能、疾患、症状、生理学、免疫学などに関する知識（様々な概念に関する情報と概念間の関係性）の効率的な検索、分類、抽出のための情報技術を開発する。また、それらの知識の根拠となる実験事実に関する検索・抽出技術も開発する。さらに、そのための用語辞書等の言語リソースを整備する。これに加え、文献より抽出した断片的な知識を組み合わせて、マイクロアレイ等の膨大な実験データを解釈し、新たな知識発見を支援するシステム開発を行う。

(c) 生命システムのもつ階層性と制約（細胞内局在性など）を考慮したパスウェイ・ネットワークの表現法・解析法および表現型データの表現法・解析法を研究する。また、これらのパスウェイ・ネットワークに潜む規則性やゲノムと表現型の間の関係を明らかにする手法の開発を行う。これに加え、パスウェイやネットワークがどのようにして形作られてきたのか、その進化過程をゲノム情報から明らかにする。

＜ 2007 年度の研究の当初計画＞

(a) 統合データベース構築に向けて、生物学文献のフルペーパー中から実験・観測結果ごとに実験事実（表現型データ）を抽出し検索・閲覧できるシステムを構築する。これを用いて既知の実験事実と新規観測事実を横断的に閲覧可能な環境（実験計画立案支援）を構築する。2007 年度は発生関連文献を対象として半自動的なデータ収集とアノテーションを効率よく行うための語彙セットの整備を行う。一方、表現型データの処理技術開発の一環として、これまでに開発してきた遺伝子座交互作用ネットワーク抽出ツールを実際の問題に適用しその有効性を確認するとともに複数遺伝子座間の関係性を直感的に把握できるツールの開発に取り組む。

(b) 情報抽出の基盤となる疾患名称、パスウェイ名称等の専門用語の整備を引き続き進める。これと並行して、文献アブストラクトからは取得が困難な実験事実などをフルペーパーから抽出することを試みる。さらに、これまではおもに真核生物を対象に情報抽出システムを開発してきたが、近年のメタゲノム解析などの進展に対応するために、原核生物の蛋白質間相互作用を抽出するシステムを開発するとともに、文献情報を利用した原核生物ゲノムアノテーションシステムを構築する。

(c) 研究者は、個々の遺伝子についての情報だけでなく、そのかかわりも含めたパスウェイ情報における種内比較、種間比較を行うことで新たな知見を得ることがしばしばある。そこで、研究者の興味の対象となるパスウェイを質問とし、指定された探索条件によってさまざまな種から抽出したパスウェイを比較するシステムをこれまで構築してきた。2007 年度はパスウェイ比較における類似度評価尺度の検討とその有効性の評価を行う。これまでの研究により再構築された、数百生物種の全遺伝子セットの進化過程をもとに、各種データベースに蓄積された、より高次の情報（相互作用や遺伝子機能など）を統合し、生命システムの進化解析を行う手法を開発する。

＜ 2007 年度の成果＞

(a) Development 誌がオンラインで公開している過去の発表論文中、2001 年以降発表分のうち HTML 形式による公開がなされている 2535 報について本文内容をセンテンス単位に分解し、全文の品詞タグ付けおよび句構造分解を行い併せてデータベース化した。図表の属する論文の PubMedID と図表番号を元に図表と意味的なリンクを持つセンテンス群を一括抽出して閲覧するインターフェースを構築した。図表に関連付けられたセンテンス群の持つ記述内容について各種オントロジーの持つ語彙群、UNIPROT の公開している蛋白質名、遺伝子名セット、NCBI TAXONOMY 等を用いたアノテーションを行った。量的形質に加えて質的形質 (例えば、アルツハイマー病か否かなど）に関連するマーカー（マイクロサテライトや SNP）の組み合わせを網羅的に抽出するツールを作成した。結果は GMM（Genotype Matrix Mapping）と呼ばれる手法で視覚化できるようにした。

(b) 疾患名称、パスウェイ名称の辞書の拡充を行った。フルペーパー中から必要な実験結果を取り出すための要素技術（図の脚注の文章の意味的分割など）を開発した。また、それに必要な辞書を構築した。原核生物の蛋白質間相互作用抽出システムを開発した。また、生物種を限定せずにゲノム配列のアノテーションを文

献ベースで行えるシステムを開発した。

(c) 利用者が指定した質問・探索条件に基づいて、蛋白質間相互作用などからパスウェイ情報を複数の種について抽出するシステムを改良した。質問中の蛋白質と探索対象種の蛋白質を結びつける情報として、配列類似性だけでなく、GO の意味類似性、オルソログ情報等を使えるようにした。数百の生物種のゲノム配列情報をもとに、全遺伝子セットの進化過程を効率的かつ高精度で再構築する手法で得られた結果を相互作用や遺伝子機能と対応付けるための手法の開発を試みた。巨大なネットワークの構造を明らかにするために、ネットワークをノードが互いに良く接続し合っている部分ネットワークに分割する汎用の手法を実装した。これをゲノム配列が決定された 635 生物種の 250 万タンパク質間の配列相同性ネットワークに適用して、網羅的なオーソログ情報を生成した。

＜国内外での成果の位置づけ＞

(a) 現在文献からの知識抽出はもっぱら要旨からの情報抽出が主体であるが、本研究のようにフルペーパーから内容的に互いに関連する情報（図表）を抽出し、知識の検索・閲覧を行うシステムはほとんどない。表現型に関係する複数の遺伝子座の組み合わせを探索する手法として MDR や CPM が知られているが、探索範囲が網羅的でない等のために必ずしも十分ではない。必要な機能を独自に開発する必要があった。本研究はそのためのものである。

(b) 本研究で開発しているような原核生物の網羅的な遺伝子名称の辞書、パスウェイ名称の辞書などは存在しない。また疾患名称についても本研究のように詳細に網羅しているものは数少ない。文献と配列情報を網羅的に結び付けてアノテーションを行えるシステムは世界的にもほとんどない。フルペーパーからの情報抽出は近年大きな注目を集めているが、まだ研究が緒についた段階である。

(c) 蛋白質間相互作用の種間比較について研究が行われているが、種全体にわたって比較し、一定の条件下で有意なパスウェイを抽出する研究が主で、抽出条件・結果が必ずしも生物学研究者の要求を満たすものでない。また、一部、利用者が質問を与えることに対応したものもあるが、パスウェイの構成要素の個々について探索条件を指定できるものはない。数百の生物種のゲノム配列情報をもとに、全遺伝子セットの進化過程を効率的かつ高精度で再構築する手法は世界的に高い評価を受け、世界最大のバイオインフォマティクス国際会議で口頭発表した。規模の大きなオーソログ情報の例として EMBL の eggNOG（NCBI の COG を継承）があるが、370 生物種の 150 万タンパクに留まっており、我々が今回作成したものが網羅性で上回っている。

＜達成できなかったこと、予想外の困難、その理由＞

(a) 量的形質解析の多くの従来手法は、立体構造のように複数の属性値で表される形質（例えば、メダカの顔貌形状など）をそのまま対象とすることができない。そのため、遺伝子座間に加えて形質間の組み合わせを考慮する必要が出てくるが、方法論的に必ずしも確立したものがなく検討が必要。

(b) フルペーパーの図表からの情報抽出においては、図の脚注などの表記にバリエーションが多く、抽出システムの開発が予想外に困難であった。

(c) 質問を与えて検索する仕組みを作成することはできたが、現状では、網羅的な蛋白質間相互作用データ量の不足や偏りによって、種間比較に十分なパスウェイが抽出されなかった。また客観的な有効性評価手法が確立されていないため、我々が提案した類似性評価尺度の評価が困難であった。

＜今後の課題＞

(a) 表現型データの記述のためのオントロジー（キーワード集合）の開発。遺伝子座交互作用解析ツールの実問題での有効性の確認。

(b) 各種専門用語のさらなる充実。原核生物の自動アノテーションの精度向上。

(c) パスウェイ比較においては有効性の評価・確認と利便性のさらなる向上。進化解析においては計算時間の短縮。

＜成果公表リスト＞

1. 0802261020 (論文)
Iwasaki, W. and Takagi, T.,
Reconstruction of Highly Heterogeneous Gene-Content Evolution across the Three Domains of Life, Bioinformatics (ISMB/ECCB2007 proceeding issue), 23, i230-i239 (2007).
2. 0802261021 (論文)
Yamamoto, Y. and Takagi, T.,
OReFiL: an online resource finder for life sciences, BMC Bioinformatics, 8:287 (2007).
3. 0801251406 (論文)
Kurokawa, K. Itoh, T. Kuwahara, T. Oshima, K. Toh, H. Toyoda, A. Takami, H. Morita, H. Sharma, VK. Srivastava, Taylor, T., Noguchi, H., Mori, H., Ogura, Y., Ehrlich, D., Itoh, K., Takagi, T., Sakaki, Y., Hayashi, T. and Hattori, M.,
Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes, DNA Res., 14, 169-181 (2007).
4. 0701050935 (論文)
Yamamoto, Y. and Takagi, T.,
Biomedical knowledge navigation by literature clustering, Journal of Biomedical Informatics, 40(2), 114-130 (2007).
5. 0701051006 (論文)
Koike, A. and Takagi, T.,
Knowledge discovery based on an implicit and explicit conceptual network., J. Am. Soc. Info. Sci. Tech., 58(1): 51-65 (2007).
6. 0602082046 (データベース)
Koike, A. and Takagi, T.,
明示的/潜在的な概念間の関係性のネットワークの可視化
BioTermNet: http://btn.ontology.ims.u-tokyo.ac.jp/
7. 0802261023 (データベース)
Yamamoto, Y. and Takagi, T.,
インターネット上のデータベースやツールなどの検索システム
OReFiL: http://orefil.dbcls.jp/

情報解析および成果公開のための支援活動

●高木 利久[1)]　◆森下 真一[1)]　◆久原 哲[2)]　◆松田 秀雄[3)]　◆金久 實[4)]
1) 東京大学大学院新領域創成科学研究科　2) 九州大学大学院農学研究院　3) 大阪大学大学院情報科学研究科
4) 京都大学化学研究所バイオインフォマティクスセンター

＜目的＞

ゲノム特定4領域の目的は、ゲノムを単位として研究を進めることにより、生命を形つくり働かせる仕組みや生物個体、環境との相互作用により進化・多様性を生み出す仕組みの解明を図ること、および、その成果を健康問題や地球環境問題等の社会的に重要な諸課題の解決に機動的に還元することにある。このような研究プロジェクトにおいては、その成果を論文や特許の形で公表するだけでは十分ではない。成果をデータベースや解析ソフトウェアの形で素早く公開し、我が国の生命科学やバイオ産業にその成果を広く役立てられるようにすることが必要である。その際、生の実験データを単にデータベースとして公開するだけではやはり十分ではない。いろいろな観点から情報解析を行い、データに生物学的医学的な意味付けをして、すなわち、データの付加価値を高めて公開することが欠かせない。これまでの10年を越えるゲノム研究の歴史の中で、データベースの形でその成果を公表することの重要性はゲノム研究者の間で広く認識されるようになり、実行に移されてきた。この意味において、本特定領域研究においても、これまでに引き続きそのような努力が行われるものと期待されるが、以下に述べるような理由から、個々の班員の努力に任せておくだけでは必ずしも十分な成果が得られるとは限らない。

・ゲノム配列のアノテーションなどの配列解析は、そのための情報技術や方法論がある程度確立しつつあるが、それ以外の新しい種類のデータ、例えば、分子間相互作用、パスウェイ、ネットワーク、種々の表現型などのデータについては、まだまだ解析技術や方法論が未成熟である。
・配列解析においても、大規模ゲノム比較、配列の大規模アセンブル、メタゲノム解析などは、高速な計算機とそれを使いこなすための専門的な技術が必要である。また、プロモータ予測などもまだ解析技術が確立していない。
・実験データの情報解析においては、さまざまな観点からデータに解析や解釈を加えることが不可欠である。そのためには、さまざまな分野の専門家集団による総合的な支援が必要である。
・公開に際しては、使いやすい利用者インタフェースやデータの流通性を高めるための標準化などにも十分配慮すべきである。

本支援班は、情報処理やデータ公開の専門家集団による技術的支援を行うことにより、また、データベースの公開や維持のための人的・資金的支援を行うことにより、上に述べたような問題点を解決し、ゲノム特定4領域の研究の成果を素早く、また、十分に利用価値を高めた形で公開するために設けられた。

支援班そのものは、その名の通り、自立的な研究活動を展開するものではないが、ゲノム特定4領域の各研究課題の情報解析、データ公開を支援することにより、また、実験系と情報系の連携を促進・強化することにより、研究成果の価値および国際的な情報発信力を飛躍的に高めるものと期待される。

＜2007年度の活動方針＞

ゲノム関連の4つの特定領域研究には、60件ほどの計画研究課題が設定されており、その中でいわゆる実験系の研究課題は50件弱である。公募班を入れるとその数は軽く100を越え、本支援班でこれらの研究課題すべての情報解析やデータ公開を支援することは、人的にも資金的にも困難である。

そこで、「基盤ゲノム」総括班に設置された情報解析・成果公開支援委員会が、ゲノム特定4領域を統括する領域である「生命システム情報」の総括班の監督のもとに、支援希望課題を募るとともに、それらに対して、国際的な競争の観点からの緊急度、データ公開の波及効果の重要度、支援の必要性などの視点からその優先順位付けを行い、それに従い、支援を実施する。なお、18年度までの応募された課題の中には各領域の運営方針と必ずしも合致しないものやまとめて開発すれば効率のあがるものなどが見受けられた。そこで、平成19年度は、領域代表の意向をより強く反映させるように配慮することとした。また、必要に応じて、複数の支援希望を一つにまとめるなどの調整を行うこととした。これにより、年度あたりの支援件数は10件程度を目安とする。

上記の支援委員会で、支援の方針が決定すると、本支援班において、各分野の専門家の意見や判断も仰ぎながら、どのような方針で情報解析を進めればよいか、企業に外注する必要があるか、その場合どの程度の費用が必要か、また、どのような情報系研究者の協力を仰げばよいか、などの助言を行う。また、（外注する場合その）仕様書作成の支援、研究支援者の派遣、ソフトウェア会社の斡旋、必要な資金的援助、なども併せて行う。

なお、本特定で扱うデータの種類や解析手法は多岐にわたる。そのため、本支援班の代表者、分担者だけでは、多様な需要すべてに適切な対応や助言ができないことが予想される。そのため、各分野の専門家からなる研究協力者を組織し、随時協力を仰ぎながら、本支援班を運営する。

＜2007年度の成果＞

平成19年度は、前期（6月）と後期（10月）の2回支援課題を募集した。これに対して、それぞれ12件と6件の応募があった。これらを基盤ゲノムの総括班会議の下に設けられた情報解析・成果公開支援委員会で審査し、支援課題と支援内容を決定した。審査の結果、本支援班で対応したものは、前期後期あわせて12件であった。

平成19年度に本支援班で対応したおもなものは、遺伝子機能表現による類似トランスクリプトーム種横断検索機能の開発、ビフィズス菌研究基盤データベースの開発、細菌叢の経時的追跡を可能とするアレイデザインプログラムの開発、植物ゲノム転写因子データベースの改良と拡充、大型真核生物ゲノムアノテーション支援システムの改良およびアノテーションパイプラインの開発、ゲノムフィニッシングプラットフォームの構築、マウスMSM系統、および他系統の配列データとの比較による、系統固有変異データベースの構築、公開などである。また、この他に生命システム情報総括班、比較ゲノム支援班で導入の新型シークエンサのデータ解析用計算機を購入し、特定領域全体へのシークエンス支援を行う準備を進めた。

上記の開発ソフトウェアに関しては、平成19年度末に開発を終了し、順次その成果を公開する予定である。