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や、それらの知識をコンパクトな形で表現している論文中の図等に関する検索・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。

＜ 2008 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。2008 年度は発生学関連の検索対象データの拡張および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、単独の量的形質のみではなく、複数の関連した形質（例えば、立体的な形状の各部位の長さ）に関連する遺伝子座間の関係性を抽出し、その結果の直観的な把握を支援するツールを作成するほか、開発した手法をメダカの顔貌形質に関するデータに適用する。また、従来から開発を続けている遺伝子座間交互作用ネットワーク抽出のツールの改善を行う。

(b) 近年の医学生物学分野におけるフルペーパーのオープンアクセス化の流れを受け、論文要旨のみからは取得することが困難なタンパク間相互作用に関する知見の根拠となる実験環境についての情報を抽出するほか、論文中の図をテキスト情報と同様に検索するための分類手法を開発する。

(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ついての種内比較、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類似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。2008 年度は、研究者が質問を与える際の条件設定手法や、抽出された類似ネットワークを表示する際の分類方法や提示すべき付加情報などについて検討、構築を行う。さらに、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ているより高次の情報（相互作用や遺伝子機能など）と組み合わせた進化解析を行う。

＜ 2008 年度の成果＞

(a) 前年度に構築したシステム（HTML 形式による公開がなされているフルペーパー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うとともに、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ためのインターフェース）について、Molecular Cell, Developmental Cell, Cell, Developmental Biology, Genes & Development の各誌のオンラインサイトから新たに HTML データを収集、整形するスクリプトを作成し、これにより 8978 報分のフルテキストデータを DB エントリに追加した。また、図表に関連付けられたセンテンス群へのアノテーション用語彙について、外部オントロジーから転用するのみでは検索効率が悪いため、新たに辞書の整理を行っている。一方、メダカの顔貌形質に関係する量的形質遺伝子座解析に向けて、収集されている複数の近交系の頭部の画像データ（新屋班員・遺伝研）をデータベース化した。また、各近交系の特徴を、単独の「合成形質」として数値化し、その合成形質のマッピングを試みた。その結果、顎付近の比較的少数（3 ～ 5 程度）の要素形質の加重線形和による近交系集団の効率良い分離と、それに関連する候補遺伝子座が確認できた。

(b) オープンアクセス可能でかつ RefSeq 等の公共データベースにおいてアノテーションの根拠文献として比較的多く採用されている論文誌 Journal of Biological Chemistry および Cell について 5 年分の全ての文献データを取得し、実験環境と結果にあたる、細胞名・実験手法、たんぱく質・遺伝子名を抽出した。続いて apoptosis や cell cycle 等幾つかの生物学的現象や概念について文献を整理し、実際に各概念につき、環境や手法にどの程度のばらつきがあるか調査した。また、フルペーパー中の図をテキスト情

報と同様に検索できるように、図の特徴点、勾配ベースの特徴量を抽出した後、クラスタリング・量子化した値をベクトル要素として図分類を行う方法を検討した。特徴点の抽出にHarris-affine、図の特徴量抽出にextended scale-invariant feature transformと図脚注の構成単語の組み合わせの手法により、76%の分類性能を達成した。

(c) 抽出された類似ネットワークの提示において、抽出ノードをクエリ上のノードも含め各類似条件によってクラスタリングしたり、クエリ上のエッジと対応するパスウェイのみ表示したりする機能を付加し、新たな知見を見出す利用環境を検討した。また、ゲノム進化過程の再構築結果を160種の原核生物ゲノムデータに適用し、代謝パスウェイの進化過程の網羅的な再構築･解析を行った。その結果、初期のパスウェイ獲得は異なる系統群間で同時代的に起こったことが示唆された。この結果をもとに、パスウェイの進化が原核生物コミュニティ内での双方向的な遺伝子水平伝播により促進されるという新たな進化モデルを提案した。

＜国内外での成果の位置づけ＞

(a) フルテキストを対象とした類似の検索システムとしてはCalifornia Institute of Technologyの開発によるTextpressoがあるが、現時点の機能は検索対象として個別のセンテンス抽出にとどまっている。本研究で構築するシステムは文献中にある図表単位の実験データ検索とその一覧を目的とし、その過程でフルテキストからの意味抽出をおこなうものである。メダカの顔貌形質の解析に関しては、画像データとして得られる「形状」に関する表現型の汎用的な解析手法の確立は成されておらず（或いは不可能)、既存のツールで活用可能なものは存在していない。

(b) フルペーパーには、論文要旨からでは得られない重要な知識が記述されていることや、オープンアクセスが可能である学術誌あるいは論文が増えていることから、近年テキストマイニングや知識抽出技術を適用する対象として注目されている。図の特徴量とテキスト情報を同時に利用するテキストマイニングは、世界でも数少ない研究である。

(c) タンパク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エ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、クエリを与えることに対応したものもあるが、パスウエ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。パスウェイの進化が原核生物コミュニティ内での双方向的な遺伝子水平伝播により促進されるというモデルは、パスウェイの進化という長く議論されてきた問題に新たな視点を提供するものである。

＜達成できなかったこと、予想外の困難、その理由＞

(a) 複数生物種を対象とする文献を一括して扱おうとする場合、解剖学用語等で同一語でありながら生物種間で意味内容が微妙に異なる場合がある。このためオントロジーのような意味構造を持たせた辞書を一意に構築しようとすると無理が生ずる。そのため対象生物種を代表的なモデル生物種ごとに分割してオントロジーを管理することとし、併せて検索対象データも生物種ごとに分割することを検討している。メダカの顔貌形質の解析に関しては、「何故、似ている / いないのか」という問題には答えられていない。すなわち、形状の特徴量を数値化することができたとしても、人間の直観的な「似ている / いない」の感覚との隔たりは未だ大きい。

(b) 生物学的現象と手法・環境に関連し、実際に生物学的実験が行われている遺伝子・たんぱく質間での相互作用についてのより詳細な調査にまでは至ることができなかった。図の抽出については、解像度の低い図も含まれていることが予測性能劣化に繋がった。

(c) 研究者のさまざまな利用目的に対応できるよう利用環境を整備したが、利用目的に依存する条件を見出すことができず、目的に応じたデフォルト条件の設定には至っていない。また、上記進化モデルの傍証とするため、メタゲノムデータを用いた進化解析を試みたが､有意な知見を見出すことが困難であった。これは、メタゲノム解析において遺伝子の生物種分類およびゲノム再構築が困難であったためであった。

＜今後の課題＞

(a) 意味的に関連する図表を一覧するインターフェースの開発およびデータとオントロジーの対象モデル生物ごとの分割管理への移行を行う。メダカの顔貌形質の解析に関しては、顔貌の「部分」のみでなく､「全体」の数値的な表現方法を検討するほか、得られたデータから、遺伝子座間交互作用ネットワークの抽出を行う。

(b) フルペーパーからの細胞名・実験手法の取得技術を利用し、生物学的現象を構成するパスウェイ等の機構が同じ種類の細胞で生じていることが文献で報告されているかについて、まとまった量の調査を行う。また、図を量子化してテキスト情報と同様にインデクスを張ることが可能となったことを利用し、図とテキスト情報を横断できる検索システムを構築する。

(c) 詳細な探索と探索時間、また、詳細な探索条件の設定と利用の容易さについてのトレードオフについての検討を進め、研究者の目的に応じた利用環境を構築する。真核生物における生命システムの進化解析・種間比較解析を行うために利用可能なデータベース・自然言語処理技術・オントロジー等について調査開発を行う。

＜成果公表リスト＞

1. 0901161637 (データベース)
鈴志野康正, 新屋みのり, 中谷明弘
MCTDB: Medaka Craniofacial Trait Database。メダカの顔貌形質craniofacial trait）の画像データベース。新屋（遺伝研）らによる各個体の頭部画像の検索と閲覧が可能。HNIとHd-rRおよびそれらのF1/F2のデータを含む。任意に選択した制御点間の距離を目的形質として、値の分布の表示や正規性の検定、区間マッピングによる関連遺伝子座の推定が可能。
http://medaka.cb.k.u-tokyo.ac.jp/mctdb/
2. 0902201150（プロシーディングス）
Koike, A. and Takagi, T.,
Classifying biomedical figures using combination of bag of keypoints and bag of keywords, Proc. of 2nd Int. Workshop on Intell. Inform. in Biol. and Med., in press (2009).
3. 0902201151（論文）
Iwasaki, W. and Takagi, T.,
Rapid Pathway Evolution Facilitated by Horizontal Gene Transfers across Prokaryotic Lineages, PLoS Genetics., in press (2009).
4. 0902201152（プロシーディングス）
Ishii, N., Koike, A., Yamamoto, Y. and Takagi, T.,
Figure Classification in Biomedical Literature towards Figure Mining,
IEEE International Conference on Bioinformatics and Biomedicine, 263-269 (2008).

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。

＜2008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、これまで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年度より、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。

＜2008年度の成果＞

平成20年度は、前期（5月）と後期（9月）の2回支援課題を募集した。これに対して、それぞれ7件と9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14件であった。

平成20年度に本支援班で対応したおもなものは、スプライシング異常予測とsiRNA設計ツールの公開システムの開発、疾患関連CNVアレル同定の為のケースコントロール関連検定法の公開支援システム開発、網羅的cDNAアノテーション支援業務及び変異部位スコア化手法の開発、リシークエンシングマイクロアレイの変異検出効率改善の為の支援システム開発、アルツハイマー病のゲノムワイド関連解析のための情報解析、腸管出血性大腸菌臨床分離株のresequencingの情報解析支援などである。また、新型シークエンサーによるゲノムシーケンスデータ保存用ファイルサーバの購入等を行った。

上記の開発ソフトウェアに関しては、平成20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。