論文の内容の要旨

論文題目　Computational analysis of orthologous genes: refined identification, database construction, and functional analysis
（オーソログ遺伝子のコンピュータ解析：精密な同定、データベース構築、および機能解析）

氏 名　　千 葉　啓 和

【概要】

近年では、シーケンシング技術の進歩により、ゲノム配列がより急速に蓄積してきている。これらの大量のゲノムデータから、生物学的知識を発見するための有効なアプローチとして、比較ゲノム解析がある。オーソログ関係は、その比較ゲノム解析の基盤をなすものとして重要である。本論文では、オーソログ遺伝子の解析に対する 3 つのアプローチを述べる。まず、オーソログクラスターのドメインレベルにおける精密化について、次に、統合的データ解析のためのオーソログデータベース構築について、さらに、オーソログ遺伝子のコード領域と制御領域の双方における比較解析について述べる。

【オーソログクラスターのドメインレベルにおける精密化】

急速に蓄積しつつある膨大なゲノムデータを対象に比較解析を進めるためには、多生物種のゲノムを対象としたオーソログクラスターの同定を高い信頼性で行うことが必要となる。多様な微生物のゲノムを比較する際にしばしば問題となるのは、特定の系統において遺伝子融合が生じている場合である。進化的に関連のある配列をグルーピングするためには、これらの融合遺伝子を切断した上で部分配列をクラスタリングすればよい（クラスターに属する遺伝子の部分配列をここではドメインと呼ぶ)。このような、ドメインレベルでのオーソログクラスタリングは、適切なグループを見つけるだけでなく適切なドメイン境界も見つける必要があるため、挑戦的な問題である。この問題に対してこれまでに提案された手法としては、DomClust があるが、この手法はペアワイズアラインメントに基づいてドメイン境界を決定するために、

しばしば不正確な境界を生じることがある。この問題点を解決するため私は、マルチプルアラインメントを用い、オーソログクラスタリングによって生じるドメイン境界を、より信頼性の高いものに修正する手法を開発した。この修正方法の基礎となるのは、ドメインレベルのオーソログクラスターを対象として、信頼性を評価することのできるスコア体系 domain-specific sum-of-pairs (DSP) score である。私は、DSP スコアを定義するとともに、それを最大化するようにドメインレベルのオーソログクラスターを修正することによって、より信頼性の高い結果を得ることのできるパイプラインを構築した。このパイプラインを用いて得られた結果を検証するために、マニュアルキュレーションされたデータベースとの比較を行い、eggNOG データベースや、DomClust などの既存の手法と比べて、より良い結果が得られることを確認した。この手法を用いることによって、より信頼性の高い比較ゲノム解析を行うことが可能となり、生物学的知識発見を促すものと期待される。

【統合的データ解析のためのオーソログデータベースの構築】

ゲノム配列データのみならず、様々な情報をおさめた生物情報データベースが存在し、しかもそれらは増加の一途をたどっている。これらのデータベースからデータを引き出して効果的に利用することで、様々な生物情報解析が可能になるが、現状では、これらの異質なデータを統合的に扱える解析環境を整えるのは、容易なことではない。これに対して私は、統合的解析環境として機能するオーソログデータベースを構築することを目指した。このデータベースの特色は、オーソログ情報を核としている点と、セマンティックウェブ技術を応用している点である。まず、オーソログ情報を利用することによって、多くの生物のデータを統合的に扱うことができる。オーソログ関係にある遺伝子は、機能が保存されていることが多いため、生物種間で知識を移転させることに利用できる。このように、オーソログ関係を、生物情報のハブとして利用することで、データの統合化を促すことができる。さらに、ゲノムデータにとどまらない異質な情報をも統合的に扱うことが求められる。そこで私は、セマンティックウェブ技術を利用した。セマンティックウェブ技術の基礎になるのは、RDF を用いたデータモデルである。まず、RDF モデルの中核として、オーソログ関係の記述に必要な体系的用語集、すなわちオントロジーを構築した。ここで開発した Ortholog Ontology (OrthO) は、多くのオーソログデータベースに共通の概念を体系的にまとめたものになっており、そのため異なるオーソログデータベースの情報を共通の枠組みで表現することを可能にするものである。その一例として、MBGD と eggNOG の二つのオーソログデータベースからデータを取得し、OrthO を用いて RDF 化して、一つの RDF ストアに格納した。RDF ストアに対しては、SPARQL 言語を用

いて、複雑な問い合わせを行うことができる。例えば、MBGD のオーソロググループに着目して、そのメンバーとなっている遺伝子の部分領域を取得し、その領域とある閾値以上のオーバーラップのある eggNOG のエントリーを抽出することができる。さらには、どの程度のメンバーの一致が見られるかを計算することもできる。今回作成したデータベースには、Gene Ontology（GO）のアノテーションや、Taxonomy の情報も格納した。オーソログ情報とこれらの情報を組み合わせ、SPARQL を用いた問い合わせによって、統合的なデータ解析ができることを確認した。例えば、特定の遺伝子の機能を推定したい場合に、オーソロググループに属する遺伝子の中から、GO アノテーションが experimental evidence code のもの、すなわち実験的な確証が得られているもののみを探すことができる。また、特定の GO がアサインされた遺伝子に着目したとき、そのオーソロググループのメンバーがどのような生物に分布しているか（系統パターン）を解析することができる。セマンティックウェブの特徴として、このような複雑な問い合わせを、ネットワーク越しに行うことができるという点がある。したがって、今回開発したようなデータベース作りが普及すれば、ネットワーク上に分散したデータを利用して統合的な解析が進められるようになると期待される。

【タンパク質コード領域と遺伝子制御領域の双方における比較解析】

比較ゲノム解析は、ゲノム配列から機能的な情報、あるいは進化的な情報を抽出するための強力な方法であり、様々な試みがなされてきた。例えば、ヒトとマウスなど、ほ乳類のゲノムを網羅的に比較して、遺伝子の機能カテゴリごとに、配列の保存度に違いがあることなどが見いだされてきた。また、遺伝子の制御領域に関しても、近縁種であれば配列が保存されており、その保存度はやはり遺伝子機能と関係していることが報告されている。ところが、タンパク質をコードしている領域の保存度と、制御領域の保存度との関係については、はっきりしたことが分かっていなかった。私は、ヒトとマウスのオーソログ遺伝子ペアを対象として、タンパク質コード領域の比較と、制御領域の比較を行った。それらの保存度について、機能カテゴリごとに統計的検定を実行して、有意に差があるものを抽出した。その結果、例えば、転写因子などは、タンパク質コード領域の保存度も高く、また制御領域の保存度も高かった。注目に値することとして、リボソームなどのカテゴリについては、タンパク質のコード領域が高度に保存されているにも関わらず、制御領域の保存度が低かった。また、細胞外マトリックスなどのカテゴリについては、タンパク質コード領域の保存度が低いにも関わらず、制御領域の保存度が高かった。これらの結果は、タンパク質コード領域と、遺伝子制御領域には、異なる進化的圧力が存在していることを示している。

【結論】

私は、ドメインレベルでのオーソログクラスタリングを改善させる手法を開発し、また、オーソログデータベースを開発して、オーソログ情報を用いた統合的なデータ解析ができる環境を構築した。これらの研究によって、信頼性の高いオーソログ情報に基づく、幅広い比較ゲノム解析が可能となり、生物学的知識発見を促進すると期待される。また、タンパク質コード領域と、遺伝子制御領域の双方の観点から比較ゲノム解析を行った。この研究は、遺伝子の進化様式に対する知見をもたらし、今後の遺伝子の進化メカニズムの研究において新しい視点を提案するものと考えられる。