生命科学と情報科学の融合：一塩基多型（SNP）データベースの開発とその解析

理化学研究所 播磨研究所 メンブレンダイナミクス研究グループ
石野 洋子

ここ数年来のヒトや微生物も含めた動植物のゲノムプロジェクトの成果として、既に多くのゲノム塩基配列が公開されている。これに伴い、塩基配列の解析手法自体に力点が置かれていたこれまでの生命科学の研究も、得られた塩基配列情報をどう利用するかという方向へと移行しつつある。そこで最近注目を浴びているのが、生命科学と情報科学の融合した学問である、“バイオインフォマティクス”や“計算生物学”という分野である。

バイオインフォマティクスはもともと生物学のデータ管理に情報技術を応用させたもので、主にデータベースに関連した技術を指していた。しかし、最近では、生物システムの特性を数学的あるいは物理的モデルへ抽象化することや、データ解析のための新しいアルゴリズムを導入することなど、より広い範囲を指すようになった。疾病遺伝子の探索や遺伝子ネットワークの推定などは、最近の主なトピックスである。ゲノム塩基配列の大量なデータを解析し、有用な情報や知識を抽出するというバイオインフォマティクスの一連の研究の流れには、近年の計算機技術の進歩やインターネットの普及・発展が寄与するところが大きいのは論を待たない。さらにこのような研究を遂行するためには、情報技術の理解のみならず生物学的そして化学的知識も総合的に必要となる。本講演では、バイオインフォマティクスの一例として、モデル植物であるシロイヌナズナ（*Arabidopsis thaliana*）の一塩基多型（Single Nucleotide Polymorphism: SNP）のデータベース開発とその解析を取り上げる。

ゲノム塩基配列は、同種の生物個体間であっても互いに若干異なっており、この違いが個体間の表現型の差異に反映している。このような塩基配列の違いを一般的に多型（polymorphism）と呼んでいる。遺伝子多型は、ゲノム情報に基づく疾患遺伝子の解明、予防的な遺伝子診断などに重要な役割を果たす。多型にはいくつかの種類があり、次の 3 種類に大別できる。（1）DNA の 1 つの塩基が他の塩基に置き換わったもの（= SNP）、（2）数十から数千の塩基が欠失（または挿入）しているもの、（3）数個から数十個程度の単位の塩基配列が繰り返し存在する部位においてその繰り返し回数が異なるもの。この中でなぜ今 SNP が注目されているのだろうか。 それは第一に、他の多型と比べて出現頻度が高く、多型マーカーとして優れているためである。次に、高速・大量の SNP タイピング技術が実現化されつつあることが挙げられる。最後に、情報処理の観点からは、多型の状態を(0,1)の信号に置き換えることができるため、情報処理操作が容易であるという利点がある。最近では、異なる遺伝子座間の対立遺伝子の非独立性を定式化した連鎖不平衡(linkage disequilibrium)という概念に基づいて、体系的・網羅的に SNP マッピングを行い、原因遺伝子の探索を行う手法の有用性が広く認められるようになってきている。

本プロジェクトは、シロイヌナズナにおける SNP を体系的にサーベイし、その結果をデータベースに格納して Web を通じて世界に公開するというものである。University of Southern California の M. Nordberg と University of Chicago の M. Kreitman および J. Bergelson がリーダーとなり NSF からファンドを得て行っており（http://walnut.usc.edu/2010/）、SNP を利用した網羅的・体系的遺伝子マッピングの先駆けである。このプロジェクトは、生物学者、数学者、コンピュータ科学者たちの連携のもとで遂行され、演者はそのデータベースの設計・実装およびデータ解析部分を担当した。

いしの ようこ

シロイヌナズナはアブラナ科に属し、温帯を中心に世界中に分布している背丈が 20 センチほどの植物である。この植物はゲノムサイズが約 125 Mb と非常に小さいこと、一世代に要する時間が約２ヶ月と短いこと、栽培が容易なこと、多くの種子をつけること、遺伝子導入が容易に行えること、変異体のコレクションが充実しているといった多くの利点があり、植物研究のモデル材料として広く世界的に使用されている。ゲノムの全塩基配列も 2000 年に決定済みである。また、シロイヌナズナで得られた知見はイネなどの有用作物への応用が期待できる。我々の SNP プロジェクト全体の流れは、まず(I)シロイヌナズナの 96 種のアクセション（エコタイプ）を選抜し、(II) ゲノムの約 50 Kb おきに約 700 bp の DNA フラグメントを選択して DNA 配列を調べ（各アクセションにつき 2000 フラグメント）、(III) 各 DNA 配列のアクセション間のホモロジー解析から SNP を検出し、(IV) SNP データをデータベースに格納するというものである。そしてさらに、(V) SNP の解析から有用な知見（例えば、ある形質に大きく関連している遺伝子の予測など）を抽出するのが目的である。

このような学際的なプロジェクトで求められるデータベースの設計方法について、また、留意した点、実際に起こった問題などについて講演では説明する。また、得られた SNP の解析から以下の(1)-(3)の結果が得られたので、それらについても解説する。（1）*A. thaliana* のゲノム塩基配列では、50 Kb の距離で連鎖不平衡の減衰が観察され、これは本プロジェクトで設定した数値の妥当性を示している、(2) SNP の差異を基に 96 アクセション間の系統樹を作成したところ、植物分類上近縁と思われるアクセション間ほど SNP が少ないという妥当な結果であった、(3) 表現型のひとつである開花期と SNP との関連分析の結果、*FLC*, *FRI* など既に開花期との関連が明らかになっている遺伝子近傍の SNP に強い関連が検出された。これらの結果は、プロジェクト初期段階のパイロットデータを分析して得られたものであり、今後はより詳細な結果が得られるものと思われる。

今後、学際的な研究が増えるにつれ、異分野の研究者との連携を上手く図りながら研究を進めていく必要性が増すであろう。本プロジェクトは研究の結果自体に加え、学際的プロジェクトの実例という意味でも興味深いと思われる。

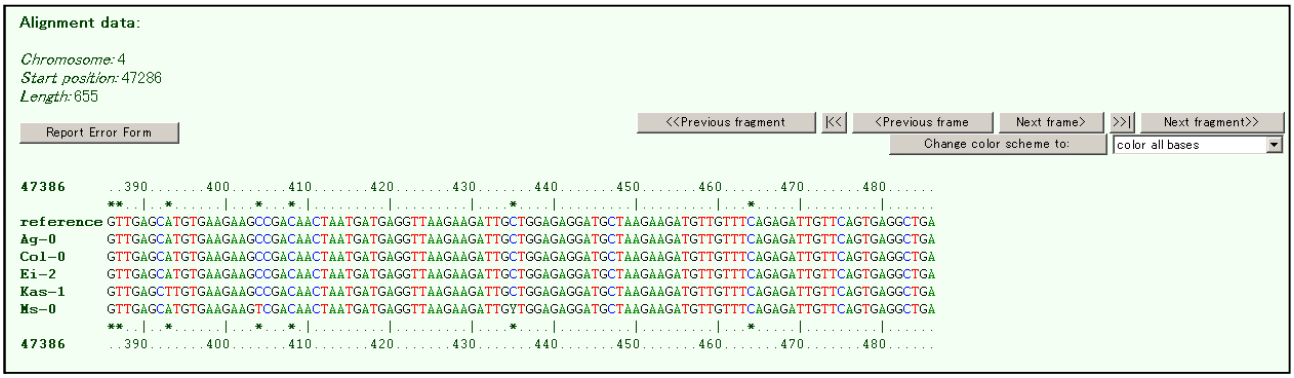

図１． SNP の Alignment Viewer の例（http://walnut.usc.edu/2010/より）

いしの ようこ