課題担当者:
京都大学農学研究科　助教
堺　俊之

内容

バイオインフォマティクス[1]

概要

バイオインフォマティクス（英語：bioinformatics）とは、生命科学と情報科学の融合分野のひとつであり、DNA や RNA、タンパク質をはじめとする、生命が持つ様々な「情報」を対象に、情報科学や統計学などのアルゴリズムを用いた方法論やソフトウェアを開発し、またそれらを用いた分析から生命現象を解き明かしていく(in silico 解析)ことを目的とした学問分野である。そのためバイオインフォマティクスは広義には、生物学、コンピュータサイエンス、情報工学、数学、統計学といった様々な学問分野が組み合わさった学際分野自体を指す。日本語では生命情報科学や生物情報学、情報生命科学などと表記される。

ゲノミクス研究の初期においては、遺伝子予測等のゲノミクスに関する分野がバイオインフォマティクスの主要な対象であった。近年ではゲノムを超えて、ゲノムからの転写物の総体であるトランスクリプトームや、トランスクリプトーム（の一部）が翻訳されたタンパク質の総体であるプロテオーム、更にはゲノムからの直接的に転写・翻訳された実体だけではなく、代謝ネットワーク（代謝マップ）によって生じた代謝産物をも含めた総体を考えるメタボローム、生物個体の表現形の総体であるフェノームなど、バイオインフォマティクスが対象とする研究分野は生物学全体に拡大・発展しつつある。

バイオインフォマティクスの目標

生物学におけるバイオインフォマティクスの主な目的は、他の生物学派生分野と同様に、生物学的プロセスの理解をより深めることにある。ただし、他のアプローチとの違いは、より計算集約的な手法の開発と適用に重点を置いている点である。用いられる技術の例としては、パターン認識、データマイニング、機械学習アルゴリズム、などが挙げられる。また、さまざまなタイプの生物学的データを組み合わせた分析と解釈を行えるように、バイオインフォマティクスの分野は進化してきた。これには、塩基およびアミノ酸配列の他、タンパク質ドメインやタンパク質構造が含まれる(Xiong and Jin 2006)。

[1] 注)この文章は wikipedia のバイオインフォマティクスのページから一部抜き出して修正したものです。ざっと読んだ感じ内容はかなり古いです。内容的にも引用文献はもっと必要だと思います。サイエンスにおいて wikipedia は古い情報/間違った情報であることが多いです。(英語版はまだマシ)。

引用元: https://ja.wikipedia.org/wiki/バイオインフォマティクス, (2021 年 10 月 15 日アクセス)

配列解析

現在、数千の生物の DNA 配列が解読され、データベースに保存されている。この配列情報は、タンパク質、RNA 遺伝子、調節配列、構造モチーフ、反復配列をコードする遺伝子を決定するために分析されている。例えば、種内や種間で遺伝子配列を比較することで、タンパク質機能間の類似性を評価したり、あるいは系統樹を構築することで種間の分子系統学的関係を示すことができる。今日では BLAST などの相同性検索を行うコンピュータプログラムを用いて、例えば GenBank に登録された 260,000 を超える生物から配列を検索することが日常的に行われている。

アセンブリ

DNA 配列は DNA シーケンサーによって読まれる。多くの DNA シーケンス技術は、短い配列フラグメントを生成する。そのため、完全な遺伝子や全ゲノム配列を取得するためには、この配列フラグメントをアセンブルして再構築する必要がある。ゲノム配列をバラバラな短い断片に分断してそれぞれを解読し（シーケンシング技術に応じて、35〜900 ヌクレオチド長)、その後同一の配列を重複する領域として並べ替えることによってゲノム配列を再現する。しかしながら、多くの断片がある中で正しい並び方を決定することはコンピュータの計算能力がなければ不可能である。そのため、高速・高性能なゲノムアセンブリアルゴリズムを開発することは、バイオインフォマティクスの重要な研究領域の一つとなっている。

アノテーション

ゲノミクスの文脈においてアノテーションとは、DNA 配列内の遺伝子領域やその機能、そしてその他の生物学的特徴をマークするプロセスである。ほとんどのゲノムは大きすぎるため、手動で注釈を付けることができない。そのため、このプロセスは自動化する必要がある。さらに次世代シーケンシング技術の登場によって大量のデータが高速に得られるようになっており、大量のゲノムに対して高速にアノテーションを付けたいという研究上の要望は高まっている。

計算進化生物学

進化生物学とは、種の起源と分化、そして系統の経時的な変化を明らかにする学問分野である。バイオインフォマティクスは進化生物学分野においても重要な役割を果たしている。

形態に基づく物理的な分類法や生理学的・生態学的観察のみではなく、ゲノム配列の変化を測定することにより、遺伝学的なアプローチから生物の進化を追跡することができる。

ゲノム全体を比較解析が可能となる。これにより例えば、遺伝子の重複や遺伝子の水平伝達、細菌の種分化に重要な因子の予測など、より複雑な進化的事象の研究が可能になる。

比較ゲノム解析

比較ゲノム解析の目的の一つは、異なる生物における遺伝子(オルソログ遺伝子）や他のゲノム上の特徴の対応関係を明らかにすることである。また例えば、2 つのゲノムが系統上で分岐した際の進化過程は、両ゲノム間の対応関係を取ることで、例えばどのゲノム領域が欠失したり重複したのかを明らかにし、進化過程を追跡することができる。

遺伝子とタンパク質の発現

遺伝子発現解析

多くの場合、遺伝子の発現は RNA-Seq などの手法で mRNA レベルを測定することで決定する。これらの手法はすべて、ノイズが非常に発生しやすく、生物学的な測定バイアスがかかってくるため、ハイスループットの遺伝子発現研究においてこのようなノイズを除去して信頼できる信号を分離する統計ツールの開発が計算生物学の研究分野で重要になっている(Grau et al, 2006)。発現解析の適用例としては、例えば病原菌感染時の植物サンプルの遺伝子発現量を非感染植物のデータと比較して、感染時に特異的に発現上昇あるいは発現抑制される転写産物を決定することで、抵抗性の獲得に重要な遺伝子の予測ができる(図 1)。

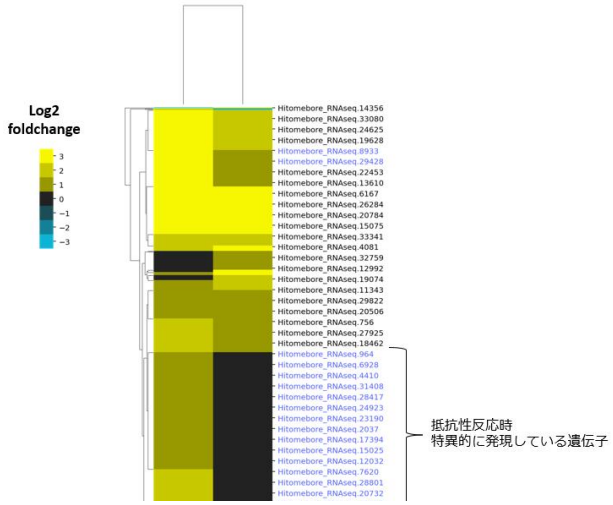

図 1 発現量解析の適用例。

抵抗性反応時にのみ高い発現量となる遺伝子群を特定している。

構造生物学

3 次元タンパク質構造の例。タンパク質立体構造の解析は、バイオインフォマティクス分析の一般的なテーマの一つである。

タンパク質のアミノ酸配列からその高次(2 次、3 次、及び 4 次)構造を予測することは、バイオインフォマティクスの大きな課題の一つである。タンパク質のアミノ酸配列(一次構造)は、それをコードする遺伝子の配列情報から、比較的簡単に決定できる。そして多くの場合、この 1 次構造は実際の細胞内における高次構造を一意に決定する。つまり、同じアミノ酸配列を持つタンパク質は全て同じように細胞内でコンフォメーションをとって折りたたまれ、同じ 2 次構造や 3 次構造を立体構造を作り出す、ということである。高次構造の知識は、タンパク質の機能を理解する上で不可欠である。しかしながら、一次配列からそのような高次構造を予測する一般的な手法は無く、未解決の問題となっている。現在までの多くのこれに関する研究は、ほとんどの場合、ヒューリスティックに焦点が向けられてきた。

タンパク質構造を予測するための他の手法としては、タンパク質のスレッディングや、物理学ベースでゼロからモデリングを行う de novo の手法が提案されている。

近年、深層学習などの機械学習的アプローチからタンパク質の構造予測を行う試みが盛んに行われており、Google 傘下の DeepMind 社が開発した”AlphaFold2”というモデルが非常に高精度でアミノ酸配列からタンパク質の構造予測を可能とした(図 2)。構造生物学における情報科学的アプローチによるブレイクスルーと言える。今後は生物学やゲノミクスの分野でも、情報科学的な知見は非常に重要になってくることが予想される(Jumper et al, 2021)。

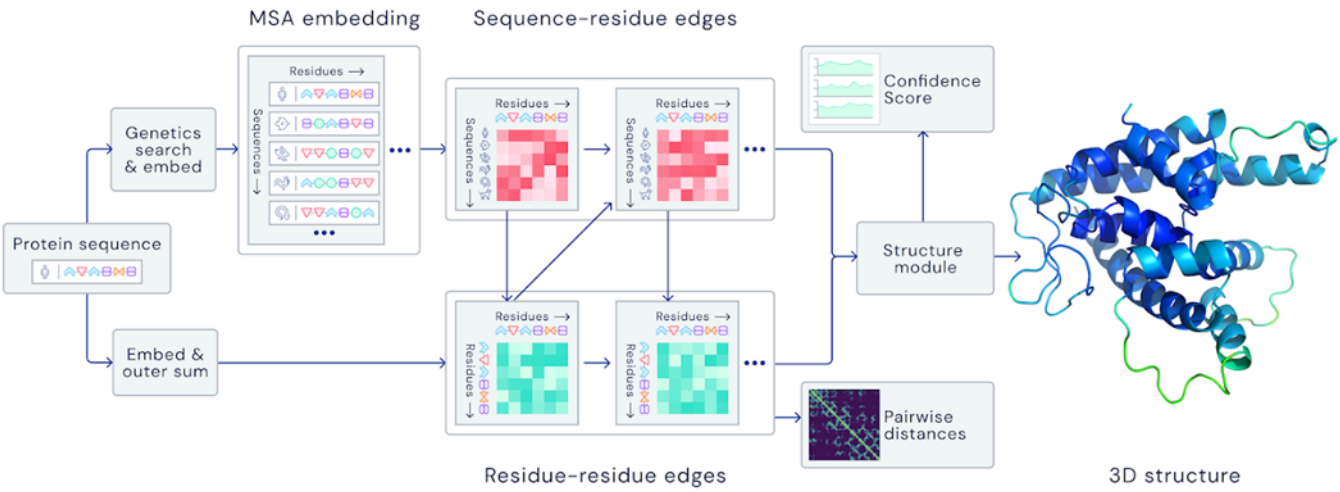

図 2 AlphaFold2 によるアミノ酸配列の立体構造予測の概略図

ネットワークとシステムバイオロジー

タンパク質間の相互作用は、ネットワークによる解析と視覚化が行われる場合が多い。ネットワーク分析は、代謝ネットワークやタンパク質間相互作用ネットワークなどの生物学的ネットワークの関係を理解することを目的としている。生物学的ネットワークは単一のタイプの分子またはエンティティ（遺伝子など）から構築される。また、ネットワーク生物学においてはしばしば、タンパク質や小分子、遺伝子発現データなど、物理的・機能的に関連する様々な異なるデータタイプを統合的に解析することがある。

システム生物学では、細胞内における複雑なプロセスの関係性を分析し視覚化するために、代謝プロセスを担う代謝産物や酵素のネットワークやシグナル伝達経路、遺伝子調節ネットワークといった細胞システムを、コンピューターシミュレーションを用いて解析する研究が進められている。人工生命や仮想進化といった単純な（人工）生命体のコンピューターシミュレーションを介して、進化の過程を理解する試みもなされている。

引用文献

1. Xiong and Jin., 2006, Essential Bioinformatics. Cambridge University Press. pp. 4.

2. Grau et al., 2006, VOMBAT: prediction of transcription factor binding sites using variable order Bayesian trees. Nucleic Acids Research 34: W529-W533.

3. Jumper et al., 2021, Highly accurate protein structure prediction with AlphaFold. Nature volume 596, pages583–589.