

論文の内容の要旨

論文題目 生物学的知識と遺伝子発現情報の融合による生体内システムの解析

氏名 辻 真吾

分子生物学の分野では、細胞の遺伝子発現状態を、ある程度定量的に測定することは古くから行われていたが、昨今の DNA マイクロアレイの誕生によって、大量の遺伝子発現情報が採取されるようになった。これには、様々な生物のゲノム配列が決定されてことも大きな一助になっている。近年、より安価になり、信頼性も向上し続けているこの DNA マイクロアレイを用いた研究は、日々その応用範囲を広げている。一方、このような大規模なデータを解析するための新たな学問分野も誕生した。元は、生物のゲノム配列を扱うために誕生したバイオインフォマティクスの一部に、こうした大規模な遺伝子発現情報を扱うための一分野が形成されつつある。本研究は、この分野に位置するものである。これを、便宜的にアレイインフォマティクスと名付けることにする。アレイインフォマティクスは、DNA マイクロアレイによって採取される数千から数万の遺伝子発現情報を統計的に処理することで、実験間の正規化や、ノイズの除去などに力を発揮する。また、他分野で応用された成果をあげている、クラスタリングの手法や、判別分析（また Support Vector Machine）ニューラルネットワークといった計算方法を取り入れ、DNA マイクロアレイの実験結果から、生物学的な意味を取り出すための研究も精力的に行われている。

DNA マイクロアレイを用いた実験では、一度の実験で注目している生物種のほぼすべての遺伝子発現状態を知ることが出来るので、特定の薬物刺激や、ヒトの場合は癌などで、特異的に発現状態が変化する遺伝子を探し出すのに非常に強力な手法となっている。また、機能がよくわかっていない遺伝子に関しては、遺伝子の発現パターンを時系列や細胞への刺激に応じてクラスタリング（発現状態が似ている遺伝子をいくつかのグループに分類する）することによって、同一クラスターの中に機能が良く知られた遺伝子があれば、機能未知の遺伝子の機能を解明する一助になるだろう。さらには、数千数万の遺伝子の発現パターンを、細胞の状態変化をそのまま表現するマーカーと考えて、ある種の癌のサブタイプを発見するといった研究もなされている。このような遺伝子の発現変化を細かく観測することによって、遺伝子の発現制御ネットワークを解明しようとする試みもある。

こうした背景を踏まえ、本研究では、大規模な遺伝子の発現情報と、これまでの研究から得られている遺伝子の機能に関する情報を、融合することによって解析する手法を提案する。遺伝子の機能に関する情報とは、生体内の特定の機能を一緒になって遂行することがわかっている数十の遺伝子のまとめりである。具体的には、新たなタンパク質の合成に関わる遺伝子群や DNA の損傷を修復するために働く遺伝子群などである。本研究では、これを遺伝子セットと呼ぶことにする。このような知識は、様々な研究グループによって、体系化が試みられており、本研究においてはこれらの知識をそのまま利用する。

本研究は、大きく分けて二つの部分からなる。前半の部分では、遺伝子セットの知識を使って、大規模な遺伝子発現情報を解析する新たな手法を提案する。後半は、遺伝子セットに内包されている情報を使って、化合物の発癌性の有無を高精度で予測するアルゴリズムの開発を目指す。

薬剤に対する細胞の応答や、特定の遺伝子の機能を阻害した細胞など、生体の変化を DNA マイクロアレイで解析する場合、発現パターンの類似度で遺伝子をクラスタリングする方法は頻繁に行われている。この結果得られた遺伝子のリストをどう解釈するかは、非常に難しい問題である。このような、実験結果を解釈するのに、役立つ研究は多くなされている。これらの研究は、既存の生物学的知識と、遺伝子発現情報を、ある程度自動的につなげる役目を果たす。たとえば、実験の結果得られた遺伝子のリストに、有意に出現する Gene Ontology (精神的に研究されている遺伝子に関する機能的まとまりの体系的知識化の一例) のカテゴリを探すのは、こうした分野で普遍的な解析手法になりつつある。

本研究では、これまでの手法では解析が難しかった、時系列・用量依存的に採取された発現データを例に、新たな手法を提案する。

解析に用いたデータは、ラットに 2 種類の構造異性体の物質を経口投与したときの肝臓の発現プロファイルで、投与後 3,6,9,12,24 時間の 5 時点と、投与用量が 4 種類選ばれている。1 つの遺伝子について、20 個のデータが採取されることになる。

このデータを用い、それぞれの化合物で特異的な変化が見られる遺伝子セットの抽出に成功した。さらに、共通の変化を見せる遺伝子セットの情報を使って遺伝子セットの知識を拡張する手法を提案する。

後半は、化合物の発癌性を予測するアルゴリズムの開発を目的とする。我々の身の回りには実に多くの化合物が溢れているが、これらの多くは、生体に悪影響を及ぼすことがわかっている。生物の神経伝達物質を阻害して短期的な毒性を発揮するものをあれば、生体に対する長期的な影響が数年後に発癌という結果をもたらす化合物も多く知られている。このような化合物の毒性を調べる学問は、毒性学として知られているが、近年この毒性学に DNA マイクロアレイを用いて採取された遺伝子発現情報を応用する、トキシコゲノミクスという分野が誕生した。本研究の後半部分は、この分野に位置する。

化合物の発癌の試験は、通常ラットやマウスなどのモデル生物を用いて、2~3 年の歳月をかけて実際に発癌するかを調べるというのが一般的であった。トキシコゲノミクスの誕生によって、遺伝子の発現状態を網羅的に測定することによって、これを数ヶ月にまで短縮できないかという試みが近年なされている。本研究では、90 種類の化合物を 1 ヶ月間ラットに連続投与し、1,3,7,14,28 日目の肝臓の発現プロファイルを採取した。このデータを用いて、これらの化合物の発癌性を予測するアルゴリズムを開発した。また、遺伝子セットに集約された知識が、予測率の向上に寄与することを示す。用いる化合物は、これまでの毒性学の手法によって、発癌性の有無が確認されているものである。アルゴリズムの根幹には、広くデータマイニングの手法として用いられている決定木の方法を用い、このとき

の説明変数に遺伝子セットを用いることで、予測率を向上できることを示す。予測率の見積もりには、Leave one out cross validation 法を用いた。

本研究では、大規模な遺伝子発現情報を、既存の生物学的な知識を応用して解析する為の手法を提案した。前半部分では、時系列・用量依存的に採取された遺伝子発現情報を、遺伝子セットを使って解析する手法を開発した。後半部分では、化合物の発癌性を予測するアルゴリズムを提案し、遺伝子セットの知識を使って、予測率の向上を達成した。

ヒトを始めとして、多くの生物が持つ遺伝子の多くは、その機能が未知であるものが多い。したがって、遺伝子セットの知識としての不完全さは、言うまでも無い。また、遺伝子の機能的なまとめりと言っても、外界からの刺激に対して、発現変動が見られない遺伝子セットでは、解析に使うことが出来ない。今後は、遺伝子セットをより洗練されたものにする必要があるだろう。現実に応じた遺伝子セットを構築することは、生物学的知識が増えることであり、それは取りも直さず、生体システムのより深い理解に他ならない。手法の改良と、知識の拡充という両輪をうまくまわすことで、今後のバイオインフォマティクスの発展に、寄与していくべきだと考えている。