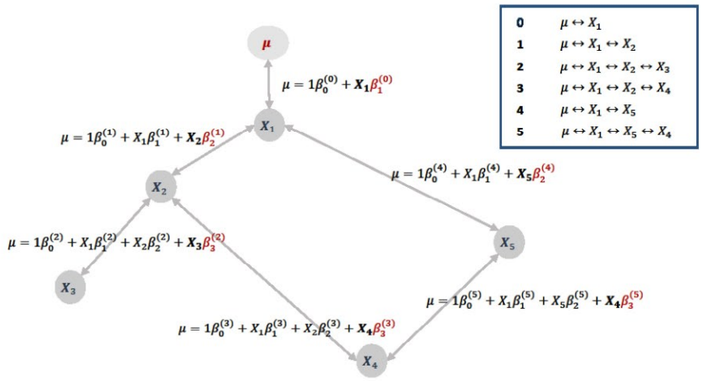
新たな病気治療や予防法の確立を目指す 遺伝子データ解析

精度を高める工夫から新たな解析法を開発

遺伝子データ解析では、一つの遺伝子から病気が発症するという一対一の関係を仮定し、遺伝子ごとに繰り返す解析が標準です。膨大な遺伝子をコンピュータで網羅して一度に調べるため、規模の大きいデータでは計算に数日かかることもあります。それでも、この解析ではデータの一部しか見ることができません。例えば、連鎖不平衡によって、遺伝子間に相関関係が生じますが、標準的な解析では無視されます。複数の遺伝子間の相関関係を利用すれば、より優れた解析ができる可能性があります。私たちは、複数の要因を考慮できる重回帰モデルを用いて、複数の遺伝子を同時に扱う手法の開発に取り組みました。明らかな課題は計算量です。仮に百万個の部位が対象であるとすれば、その二ペアの組み合わせ総数はおよそ五千億、三ペアの組み合わせ総数はおよそ十七京と、総当たりによる検査総数は膨れ上がります。百万部位を一度調べるだけの標準的な解析ですら数日かかるため、二ペアの計算では相当な計算量になり、もはや三ペア以上はスーパーコンピュータを使ったとしても現実的な時間内に計算が終わらないという事態に陥ります。次に、統計学的な面で深刻な課題がありました。ノイズと真のシグナルの切り分けが、検査対象数の増加に応じて難しくなるという多重性の問題です。単純なボンフェローニ補正を用いるとすると、検査総数で割って厳しく設定した有意水準をP値が超えなければいけないので、ハードルが非現実的に高くなってしまい、もし仮に総当たりで計算できたとしても、検出力低下が避けられず検出自体が困難となります。そこで、真正面から取り組むのではなく、手法を工夫することで解決を試みました。連鎖不平衡は、ヒトの場合でしばしば遺伝子間の距離が近い場合に強く、距離が離れるに連れて減衰する、という観察を利用し、近傍に位置する遺伝子セットの相関構造で辺を定義したグラフ（双方向グラフ）の最短経路上で関連性を調べていく新たな解析法を開発しました。技術的には、重回帰モデルとグラフとの対応を与える新たな数学的な結果に基づきます。見る必要のない組み合わせを見いだし検査対象の組み合わせ数を大幅に削減することで、計算量の低下と検出力の向上が同時に達成されて、複数の遺伝子の組み合わせを調べるデータ分析法の開発が実現できました。

遺伝子間の相関関係を利用した解析法



概念図：総当たりで調べず、遺伝子間の相関の強さで定義したグラフの最短経路上の組み合わせに絞って解析。

ビッグデータ活用への期待と展望

がんや糖尿病などのありふれた病気には、体質だけでなく生活習慣などの遺伝要因の関与も無視できません。最近は、数万人以上の人々の生活習慣や健診情報、画像、遺伝子など、なるべく多くデータを集めて網羅的に調べるというプロジェクトが、国内も含めて多くの国々の異なる場所で立ち上がっています。今後はビッグデータを活用して、遺伝子とそれ以外の要因を組み合わせて見ていく必要があるでしょう。例えば、どの遺伝子をもつ人がお酒を飲むと病気になりやすいかといったことです。しかし、遺伝子だけでも相当な数があるのに、それに加えてその他の膨大な要因も考慮しなければいけないとなると、人の手で実践していくのは非常に大変な作業になります。コンピュータにビッグデータを投入すれば、データが精査されて、結果が自動で出てくるようになるのが理想ですが、まだそこまでは至っていません。データに応じて適した方法が異なるということもあるでしょう。幅広いデータに共通して使える手法があれば、自動化に一歩近づきます。最近話題の機械学習は、自由度の高い柔軟なモデルを用いていて、ビッグデータのような複雑なデータ分析に有力な手法として着目されています。一方で、柔軟な反面モデルを注意深く選ばないと、手元のデータに近づきすぎて別のデータでは精度が再現されないという過学習と呼ばれる現象が起こることも知られています。また、機械学習のモデルは複雑なため、高い性能を示したとしても、その理由を説明することが難しいという問題もあります。

未だ実用に向けて課題が山積していますが、今後は機械学習の方法を遺伝子データの解析に取り入れて、より柔軟にビッグデータが解析できる方法の開発に取り組みたいと考えています。

山積する課題を受け止め より柔軟な解析法を見いだす

Text by UEKI Masao

植木優夫 教授
情報データ科学部教授。岡山大学環境理工学部環境数理学科、同大学院を卒業後に、統計数理研究所、山形大学医学部、東北大学東北メディカル・メガバンク機構、久留米大学バイオ統計センター、理化学研究所革新知能統合研究センターにて研究と教育に携わる。専門は数理統計学、遺伝統計学、生物統計学。