表現型関連遺伝子の探索に向けた生物医学ビッグデータ

特　集

大阪大学大学院医学系研究科・ゲノム情報学共同研究講座
特任教授　中　谷　明　弘　氏

●本日の概要

イントロダクション的な内容に引き続いて、前半に作物の育種や品種改良、後半でヒトの疾患に関するお話ができればと思います。対象のデータとして、作物の育種ではマメ科の牧草である赤クローバを例として取り上げます。こちらは、千葉県木更津市にあります、かずさDNA研究所の方々と進めているものです。複数のDNAマーカの組み合わせ最適化による相互作用解析を行っています。ゲノミック選抜と呼ばれる枠組が作物や家畜の育種で行われており、これに関連しても、機械学習や組み合わせ最適化の実例を示しながらお話したいと思います。ヒトの疾患については、新潟大学脳研究所の方々と進めているもので、実際には全国から集めたデータですが、アルツハイマー病の生化学バイオマーカ、つまり、血液等の検査結果を用いた認知症のなり易さの推定、あるいは、実際になった方の特徴の評価のお話になります。また、ゲノムワイドなコピー数変異のデータベースを構築した結果に基づいて、特定の遺伝子のコピー数の変異が疾患に関わっているかも知れないということについてもお話します。

●生物医学データに関わる技術的な背景

生物学や医学の世界で何が「大きなデータ」かと申しますと、高性能シーケンサによるデータはその1つかと思います。これは、NGS（next-generation sequencer・次世代シーケンサ）とも呼ばれている機械でして、機種にも依りますが、机の上に乗るくらいの大きな冷蔵庫程度のものです。データ化自体は、低価格化していると言われており、ヒトの場合、例えば、遺伝子に該当している領域の全てを解析して、一人当たり、大体1,000ドル、約10万円くらいからで読めることになっています。とは申すものの、1,000人やろうとしたら1億円かかることになります。これには、詳細な情報の解析費用などは入っていませんので、基礎的なデータを取得するとして一人当たり20万円くらいでできることになります。これに関連して、大きな転換点になったのはヒトのゲノム配列（ゲノムの塩基配列）の決定で（～2003年）、これをベースにシーケンサの開発が加速度的に進んだわけです。それが10年位前のことで、そこから「ポストゲノム」と呼ばれるゲノム配列が決まった後の今の時代につながっています。また、シーケンサのほかに高密度アレイというものもあり、これは10cmくらいのカートリッジやスライドガラス状のものなのですが、配列の決定はしませんが、ゲノム中の特定の個所がどのような変異をしているか、あるいは、特定の遺伝子がどのくらい発現しているか等を調べることが可能になっています。これらで何が大きかったかと申しますと、データがデジタル化したことだと思われます。恐らくは、生物学や医学の一部が情報科学になってしまったというのが最も大きな点だと思います。また、ゲノム解析の単位も転換してしまって、以前はセンチモルガン（cM）という遺伝的な距離で評価されることが多かったのですが、ベースペア（bp：塩基数）という物理的な距離の単位での解析に移行していったという部分も、大きな点だと考えられます。

講師　中谷　明弘 氏

● NGS 解析パイプライン

さて、そのNGS（次世代シーケンサ）のデータが実際にはどういうものなのかと申しますと、基本的には検体をシーケンサにかけると、例えば、FASTQと呼ばれる形式のファイルが出力されてきます。このFASTQファイルは、ある人のゲノム配列を細かく切断した配列データを大量に含んだものです。先に申しました通り、既に標準的なヒトのゲノム配列が決まっておりますので、それらの断片の一つ一つをその標準的な配列の該当箇所に張り付けていって、個人のゲノム配列を再構築します。ですので、ヒトのゲノム配列が全く無い状態から最初のゲノム配列を決めたというのとは少し違っていて、断片を標準的な配列の該当する場所に張り付けていくマッピングと言われる手順を踏んで、無数の断片を1列に並べることが可能になります。そう致しますと、標準的と言われているヒトのゲノム配列と各個人のそれとではどこが違っているかということが分かってきます。配列の断片は、パーフェクトマッチで張り付いているのでなく、例えば、染色体の特定の位置で標準的な配列ではGですがあなたはAですというような変異の情報が機械的に得られるようになっています。この変異の情報と疾患との関係を探していくことになります。ただし、変異と疾患の関係の解析をどうすれば良いかは単純ではなく、試行錯誤によって探り当てて行っているのが現状かと思います。

●ゲノム配列（変異）の情報だけで十分か

繰り返しになるかもしれませんが、疾患の解析にゲノム配列の情報だけで十分かといえば決してそうではありません。高性能なシーケンサの出現によって、生物種や系統群を問わず、ヒトでも植物でも魚でもバクテリアであっても、FASTQのような汎用的な形式のデータを得ることができるようになりました。標準的な配列との差異は機械的に取得が可能で、例えば、何番染色体の何番塩基目が標準的な配列と違っているというSNP（single nucleotide polymorphism：一塩基多型）の情報は機械的に取れてきます。しかし、そのように染色体が変わっていると言われても、それがどのような影響を及ぼすかについては必ずしも明らかではありません。従いまして、ここで重要なのは、ゲノム配列だけを集めていても十分ではなく、表現型の情報も併せて集めないと、そこからの情報抽出はできないということです。一人ひとりにゲノムと表現型の両方がペアになって紐付いていることがデータ価値として非常に重要になります。その中から知識の推論をして、変異と疾患の関係を探していくことが実際に行われていて、そこでは、機械学習や組み合わせの最適化の手法が必要になってきます。2000年代初めにヒトゲノム、すなわち、生物種としてのゲノム情報が取得されたわけですが、現在は個人としてのゲノム情報が取得されるようになっています。

●表現型に関係する遺伝的要因

ゲノム情報以前に人類は歴史が始まって以来さまざまな生物の品種改良を続けてきています。例えば、イヌとブタはそれぞれオオカミとイノシシを家畜化したものです。他にも、花卉、例えば、ラン科植物のカトレアは、100年以上前から様々な掛け合わせが行われていて、その過程の情報が登録されて残っています。小さな花しか咲かない原種から、巨大で鮮やかな花が咲く品種が作出されています。これらの掛け合わせは、勿論、ゲノムですとかDNAの情報を用いていたわけでなく、花や草の姿といった表現型を目印にして、これとこれを掛け合わせたら良いものができるかもしれないと経験的にやってきたものです。この方法も必ずしも劣っているわけではなく、実際に現在でも野菜や果物などの作物でも行われています。血や血統という言葉を良く使いますが、これは抽象的な遺伝的要因を表しており、直感的に非常に分かり易く、また、概念的には必ずしも間違ってはいません。しかしながら、恐らく多数あるそれらの遺伝的要因を祖先からどのように受け継ぐかは確率的（偶然）になりますので、人間にとって都合の良い個体が得られる確率は非常に低くなります。ですので、より直接的に表現型に関連している遺伝的要因を探し出して、その要因をどのようにもっているかを知った上で積極的に掛け合わせを行っていくことがより良い戦略となるわけです。ゲノム配列が得られることによってそれが可能になってきています。

●遺伝的要因の探索

先ほどの花のような例で申しますと、例えば、大

きな花と小さな花を掛け合わせるとそれらの雑種が得られて、さらにそれら同士を掛け合わせると様々な個体が得られます。その中で大きな花をもつ個体は、元々の個体からいずれの染色体領域を受け継いでいるかを調べることによって、表現型に関係する遺伝的要因を探し出すことが行われています。このプロセスの中で、量として表される表現型をどのように扱うかが非常に重要になってきます。量として表される表現型は、複数の遺伝的要因が関わっていると考えられており、さらには、遺伝的要因だけでなく環境的要因も影響していると考えられています。このような考え方は、表現型を分布としてデータ化し易いモデル生物の場合は良いのですが、ヒトになると少々難しくなります。例えば、人為的な掛け合わせができないですし、我々ヒトはいわゆる純系ではなく万人が雑種です。また、個体数が小さく、世代数も小さいですし、環境も多様で一斉に種を播くようにもいきません。他にも、浸透率（同じ遺伝的要因をもった個人が同じ表現型になる率）の低さや、疾患に関係している遺伝的要因の集団内での発生率の低さ（稀少変異）もあります。また、表現型が複雑で、例えば、認知症ですとか精神の働きというのは、そもそも表現型が数値化し難いため、それらをどうやって解析していくのか自体が明確ではないわけです。他には、個人情報の扱いの問題から情報の入手性が低いかも知れません。私は情報を出したくないというケースは現実に生じてきます。そして、近縁種がありません。我々は基本的にホモ・サピエンスの1種しかなく、似ている生物種から何かを類推するということもできません。

●質的形質と量的形質

表現型を分類する2つのカテゴリーとして、質的なものと量的なものがあります。例えば、質的な表現型で馴染みのあるものとしてABO式の血液型がありますが、これは、A、B、Oの組み合わせで表現できて、表現型、すなわち、血液型は大小関係のない有限個の場合になります。これは、1個の遺伝的要因で説明できます。しかし、体重や身長、血圧や血糖値など、量的な表現型はどのように遺伝的要因が関係しているのかは、単純そうですが実は良く分かっていません。このような量的な表現型が我々の健康には大きく影響しています。体重を重くする遺伝子の働き具合で量的な表現型の値が変わるという考え方はあまりしないことになっています。

量的な表現型については、複数の遺伝的要因の効果の重ね合わせが関与しているという考え方に基づきます。しかしながら、単純な足し算ではないので、構造としては遺伝的要因間のパスウェイ（ネットワークやエピスタシス（交互作用））、例えば、特定の2つが揃うとそれらの足し算以上の効果になるようなモデルも考える必要がありますが、十分に扱い切れていないのが現状かと思います。そのような複数の遺伝的要因を探すため、個体群内の個体のそれぞれのゲノム配列に基づいた遺伝的要因の候補の情報を準備します。この時点ではゲノム配列中のどの位置が表現型に影響しているかは不明です。それと併せて各個体の量的な表現型の情報も準備します。ある特徴の有無（質的な表現型）でなく連続的な量として扱います。これらを突き合わせて、例えば、植物の開花時期ならば、開花日の早い個体や遅い個体はどのような位置にどのような遺伝的要因を持っているかを見つけます。しかしながら、一般的には目で見て簡単に分かるようなものではありません。

●染色体 / ゲノム配列のスキャン

遺伝的要因の探索には、例えば、区間マッピングと呼ばれる手法が1980年代後半から使われてきました。また、ゲノム配列が得られてからも、GWAS（genome-wide association study）と呼ばれている解析も行われてきました。いずれも、染色体やゲノム配列に沿って1カ所ごとにその位置が表現型に関係しているかを評価するものです。複数のマーカを用いてはいますが、各位置での単一の遺伝的要因の評価の繰り返しで多要因的な現象を解析しようとしているものです。これにはもちろん限界があり、もう少し積極的に複数の要因とそれらの間の関連性を考える必要があります。

●赤クローバの開花日の例

実際に解析を行っていると、そのような必要性を示す例に行き当たります。例えば、赤クローバの開花時期と遺伝的要因の解析でも、単独では表現型との関係が薄い遺伝的要因しか見つかってこないのですが、特定の複数の遺伝的要因を組み合わせると、開花時期の遅い個体群を選び出すことができます。

このような遺伝的要因が生物学的にどんな意味があるかは実験的にも検証しなければなりませんが、少なくともそれらが合わさることによって開花時期の遅いものを特徴づけることができる、或いは、開花時期の遅いものはそれらの遺伝的要因をもっているということがデータからは言うことが出来ます。

●全ての場合を数え上げるための探索空間

遺伝的要因の組合せパタンの全てを網羅的に数え上げることになるわけですが、工夫しないとならないところが色々とあります。例えば、候補となる遺伝的要因の候補を100個とすると、それぞれを選ぶ・選ばないを考えることになりますので、その組合せの総数は2の100乗になってしまいます。より一般的には、N個の候補があった場合に、N桁の2進数00・・・0から11・・・1でi桁目をi番目の候補に対応させて、選ぶ・選ばないを1・0に対応させるようにして数え上げていくとします。これらの2進数を最左の桁から伸ばしていく過程を表す探索木を使って数え上げていくことができるのですが、その2進数の総数は2のN乗通りになるので、Nが少し大きくなると途端に計算できなくなってしまいます。探索空間内の全てを数え上げることは現実的でないわけです。場合の数の組み合わせ論的爆発こそが本質的な「ビッグデータ」なのだと個人的には思います。

●形質値の分布の偏りによる評価

そこで、何とか、これを巧く解いてやることが必要になるのですが、まず、特定の組み合わせパタンの善し悪しの評価の方法を決めます。例えば、そのパタンを持っているか否かで集団を2つの部分集団に分割します。もし、その分割に使ったパタンが表現型と関連していれば、2つの部分集団での表現型の分布は互いに偏ってくるはずです。これは、分散分析（ANOVA）と言われる考え方に相当します。もし、ランダムに2つの部分集団に分割した場合、同じ分布でサイズが半分の集団が2つ出てくるだけになります。ですので、分布の偏り具合を指標にして、最も良い組み合わせ、或いは、上位の組み合わせを抽出することで、複数の遺伝的要因の効果を探索することが可能になってくると考えることができます。

●評価関数の上限値による枝刈りの効果

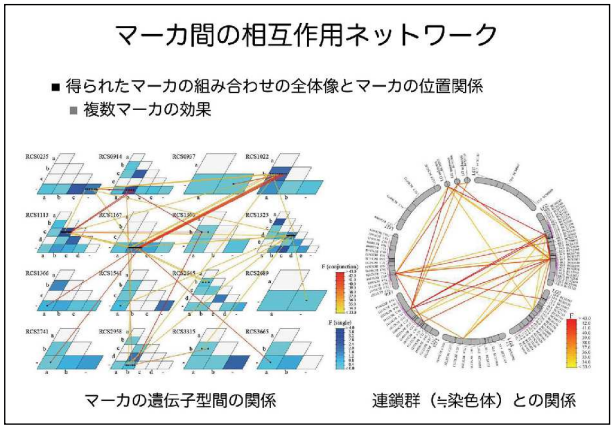
この偏り具合は、分散分析で使われるF値を評価関数として評価することができます。探索木を辿って組み合わせに含める遺伝的要因を増やしていくうちに、ある所からF値がこれ以上は良くならないことがこの評価関数の特性から証明できるので、その位置を根とする探索木の部分木を枝刈りして計算を省略することができます。これはA* と呼ばれる探索アルゴリズムになります。実際に、データ依存な評価ではありますが、枝刈りを行ってあげると、数桁倍の高速化が実現します（※実は候補数が大きくなると枝刈り無しの探索は現実的な時間内には終わりません）。実際に、先程の赤クローバの開花時期に関連した遺伝的要因の組み合わせも見つかってきています。

●マーカ間の相互作用ネットワーク

計算の結果が単に合っているかどうかというだけでなく、もう少し意味的なものを見てみます。赤クローバの開花時期に関連した遺伝的要因の組み合わせの染色体（連鎖群）上での位置を表示してみます。表現型との関連が高い組み合わせの上位のものを選んできて、1つの組み合わせに含まれる位置同士を線分で結んでいきます。そう致しますと、染色体の特定の場所に線分の端点が集中してくることが分かります。そのような領域には開花時期に関連する何らかの遺伝的要因があるのではないかという手がかりが得られてくることになります。計算機処理ではひとまずはここまでの絞り込みになります。

●ゲノミック選抜

このようなデータ中の情報が見えてきますと、実



際にそうした情報を使って次の世代の個体群を作ることが可能になってきます。特定の表現型に関連したゲノム配列中の遺伝的要因、より正確には、遺伝的要因の組み合わせを持っているものを選ぶ手法がゲノミック選抜（GS: genomic selection）といわれているものです。これは、2000年代に入って使われるようになりました。表現型に関連した遺伝的要因の候補をゲノムワイドに多数準備して、重回帰分析的な考え方に基づいてゲノムの情報から表現型を推定する予測式を作成します。その予測式で推定された表現型の情報を元に次の世代の個体群を生成するというのがゲノミック選抜の考え方です。表現型が既知の個体群の情報を用いて予測式を作ってしまえば、表現型が未知な個体でもゲノム配列の情報があれば表現型が推定できることが重要です。

● GS/MAS

ゲノミック選抜は、マーカ選抜（MAS: marker-assisted selection）と呼ばれる手法の一つなのですが、表現型の予測式を作るには、表現型と遺伝的要因の候補の両方が揃っている個体群のデータが必要になります。それらの間の関連を調べて予測式を作成します。遺伝的要因の有無さえ分かっていれば予測式によって表現型の推定値が分かりますので、その予測された表現型が条件を満たしたものだけを選抜対象として、次の世代を作ります。この考え方は、例えば、成熟まで時間がかかる果樹のような個体の表現型を考えると意味があることが分かります。また、雄の産乳量や産卵量のように、その子孫の雌個体を得ない限り評価できないような潜在的な表現型、正確には、雄親としての能力も、この予測式を使えば評価できることになります。

●予測式（単一あるいは複数マーカ）

予測式の作り方は色々と考えられます。例えば、1つの遺伝的要因で予測する時には、その遺伝的要因が特定の型を持っているかどうかで予測を行います。これですと、○か×にしかなりませんので質的な表現型の予測にしか使えません。そこで、予測に用いる遺伝的要因を複数にします。先ほどの1つの遺伝的要因のみの予測が単回帰分析に、こちらは重回帰分析に対応すると言うと分かり易いかも知れません。各要因が特定の型を持っているかどうかで加点していくと、型のパタンに応じてその合計点は色々な値になって、量的な表現型を表すことができることになります。

●解析の対象データ

これまでに、赤クローバの開花時期の、日本、インド、ロシア、スイスで評価したデータを解析しています。それぞれ、表現型と遺伝的要因の両方を含んでいます。一部の国のデータは年度ごとのデータも含んでいます。表現型を個体ごとのデータを使って、いつ花が咲いたのかを評価する予測式を作っています。

● GS にまつわる素朴な疑問

ゲノミック選抜に関してですが、その利点として、例えば、遺伝的要因を同時に多数使用できることや、表現型に関する遺伝的要因を明示的に特定しないまま予測できることが挙げられています。しかし、逆に、これは多数の遺伝的要因の型を決めないとならないことを意味しています。表現型に全く影響がない遺伝的要因の型も決める必要もありますし、また、互いに良く似た挙動の遺伝的要因の型も決めないとなりません。そこで、多数の変数（遺伝的要因の候補）を用いた重回帰分析という枠組みではなく、厳選した変数のみを用いて選抜対象かどうかを判定する手法を用いて次の子孫を作る方法を開発しています。

●機械学習による表現型の推定式の生成

機械学習による表現型の推定式を生成する手法を開発していますが、そこで用いているのは、アダブースト（AdaBoost: adaptive boost）という方法です。この方法は、どの遺伝的要因を使うか、それらを互いにどれくらいの重みで使うかを計算するものです。直感的にいうと多数決で判定します。遺伝的要因の型に基づいた複数の予測式の多数決で選抜対象か否かを決めるという方法です。予測式としては線形和で表される単純なものになります。午前中の講演で51勝49敗という話がありましたが、50%を超えるようなあまり精度が良くない予測式も、たくさん集めて多数決をさせていくと、全体としては文殊の知恵的に精度を上げることができます。そのような特性が数学的に証明されています。

● AdaBoost

例えば、ある遺伝的要因がAAという型を持っているか否かを見て、持っている（イエス）と持っていない（ノー）に応じて加点と減点をします。これは1個の遺伝的要因だけで見ているのですが、さらに2つ目の遺伝的要因を加えると、イエス/イエス、イエス/ノー、ノー/イエス、ノー/ノーの4パタンになります。以下同様にして追加して行き、20個の遺伝的要因と使用すると、高い精度で開花時期の遅い個体を選び出す式を作ることができます。この過程で、どの遺伝的要因を使うか、加点と減点はどのくらいにするかという部分をAdaBoostで決めることができます。この予測式を作る段階では、正解（実際の表現型）を見ていますから、オーバーフィッティング（過剰適合）になってしまっていて、ここで得られた予測精度が未知の表現型の予測精度になるとは限りません。

●複数国間の交差評価

そこで、複数の国と年度のデータを用いて、それらの間で互いに正解を隠して予測し合った結果がこの表です。例えば、日本の2011年と2012年のデータを使って、日本、インド、ロシア、スイスのデータがどの程度まで当てられるか評価したものです。選抜対象と予測して本当に選抜対象であった割合を表しています。ランダムに選択した場合（乱択）に比べると予測精度が上がっていますので、ゲノム情報を加味した選抜は効果があることが示唆されていると考えられます。

●各国年度で用いられたマーカ

こちらも、当たったかどうかという結果だけではなく、実際にどういった染色体領域が花の咲く時期、要は成熟が早い・遅いに関係しているかを見てみることが重要になります。各国、各年度での予測式で使われた位置を集計してみますと、いくつかの位置が共通していることが分かります。そういった領域は、花の咲く時期に関係している何らかの遺伝的要因があるのではないかということが推定されてくるわけです。

●遺伝子要因のデータベース化

このような表現型に関係した染色体領域の情報は遺伝的リソースとしても重要です。そのような情報はデータベースに蓄積されています。例えば、イネ科植物のイネ12本の染色体とソルガム（コーリャン）の10本の染色体のどことどこが対応しているか（シンテニー）はある程度は調べることができます。イネは良く調べられている作物ですが、仮にソルガム側で情報のない状態であるとしても、染色体間の対応関係を辿ることによって、イネ側の知見をソルガム側に対応づけることができます。例えば、イネの花が咲く時期に関する遺伝的要因の位置が分かっていれば、その位置に対応するソルガムの染色体領域に同様の遺伝的要因が存在することが推定できます。こうした情報を蓄積していくことは重要なアプローチになります。従いまして、先ほどの赤クローバのデータで見つかったような遺伝的要因の情報を蓄積していくことによって、実験しなくとも表現型やそれに関係する遺伝的要因を調べられるという可能性が出てきます。

●オルソログDBをハブとしたDBリンク

実際にそのような情報の蓄積はさまざまなプロジェクト、あるいは、省庁や研究機関によってデータベース化されています。我々も独自のデータベースの構築に加えて、データベースのデータベースに相当するものを、種々のデータベースを統合化するプロジェクトとして進めており、ネット上から検索ができるようにもなっています（http://pgdbj.jp）。

●アルツハイマー病の生化学バイオマーカ

ここまでは植物のお話でしたが、ここからはアルツハイマー病（AD: Alzheimer's disease）の生化学バイオマーカとコピー数変異について触れたいと思います。アルツハイマー病は遺伝的な背景を持つ多因子性の疾患といわれ、緩徐進行性、つまり、じわじわと病状が進んでいく神経疾患です。実は、アルツハイマー病は何種類かに分類でき、発症年齢で分けると、早期発症型（EOAD: early-onset AD）と晩期発症型（LOAD: late-onset AD）があります。これらは、遺伝形式も違っていると考えられています。多くの場合、アルツハイマー病というと孤発性のLOADが該当していますが、血液や脊髄液による生化学データを用いてこの疾患の特性を評価できないかということを考えています。実際にアルツハイ

マー病の検体情報を使った予備的な解析を行っているのですが、検体ごとの情報としては年齢や性別などの他に、約50項目の血液や尿のデータ、ApoE遺伝子型や教育期間を用いています。これらを使って病態を推定します。これらで病態を予測できれば、疾患のバイオマーカとして使えるのではないかと想定しているわけです。

●非侵襲な推定指標

ここで、非侵襲というのは大きな意味があります。例えば、脳脊髄液は脳と直結しているためデータとして扱い易いと考えられるのですが、腰椎穿刺をしないと得られないために肉体的な負担が大きくなりますので、可能ならば低侵襲ないし非侵襲な血液や尿で推定できる方が好ましいわけです。健常の方と発症している方のデータをプロットしてみますと、例えば、性別ごとに、健常、軽度認知症、認知症の方々を、それぞれ青、緑、赤と色分けしても、女性と男性での人数比にあまり差があるようには見えません。同じように、年齢に関して見てみますと、高齢側に赤く色分けされた認知症の方が増えてくる傾向があるのが分かります。年齢がアルツハイマー病にとってリスク要因であることがデータからも分かります。このようにして各項目を見ていくと、例えば教育年数で見ると高学歴の方が平均的に発症比率は低いのではないかということも分かってきます。しかしながら、血液検査等による生化学のデータを見ても、そのような傾向があるのかどうか判断するのは簡単そうではありません。

●疾患および健常の推定

実際にこのようなデータから、疾患の方と健常の方の検査値の分布の差から病態が推定できないかということを考えています。ただ、例えば、「血圧135以上だと・・・」という話があると思いますが、そのような1つの指標だけでの判定では難しく、複数の指標を組み合わせる必要がありそうです。そこで、ここでも複数の指標による判定の重み付きの多数決をすることによって推定しようとした結果が、次の例になります。

●閾値判定＋ AdaBoost

検査項目の数値の予め設定した閾値（いきち）との大小比較を行い、複数の検査項目での大小比較の結果の多数決をすることによって健常なのか疾患なのかを分けてみますと、データ依存な結果にはなりますが、10項目程度の多数決で80%くらいは当てることができることを確認しました。どの検査項目をどの位の重みで使うのかはAdaBoostで決定し、検査項目の閾値も結果が最善になるように自動的に設定しています。これは、答えを知っている訓練データでの精度ですので、実際に病態が未知の検体での精度は別に評価しないとなりませんが、複数の検査項目の組み合わせがここでも効果がありそうなことが期待できます。

●訓練データとテストデータ

実際に交差評価（クロスバリデーション）を行うために25検体をランダムに選択して予測式作成用の訓練データとし、それとは別に精度評価用のテストデータ25検体を準備してその病態を当てることができるかを評価したところ、血液など生化学データのみで多数決をさせる検査項目を30項くらいまで増やしていくと、性別ごとに訓練データでの予測精度は98%程度まで向上します。正解を隠したテストデータでどの位の精度が達成できるかの評価ですが、生化学データのみですと、女性の場合は次第に上がって15項目くらいで60%程度、男性はあまり上がらないのですが、50%程度になっています。また、教育期間等の生化学データ以外も含めますと、女性で70%程度、男性で50%強という結果になっています。このデータで罹患率は女性で47%、男性で40%ですので、この罹患率からどれだけ予測精度が向上しているかは評価の1つの目安にはなるかと思います。

●判定ルールに頻出する検査項目

このように見てみますと、まず、特徴的なのは性差があるということです。そして教育期間を入れてやるとさらに上がるということが分かり、そこに関連する要素がやはり疾患の予防に繋がるのではないかということが期待できると思います。生化学関連の検査項目では、肝臓や赤血球、甲状腺、葉酸、尿酸、コレステロールのようなものが決定項目として効いていると思われます。

● AD 関連コピー数変異のデータ

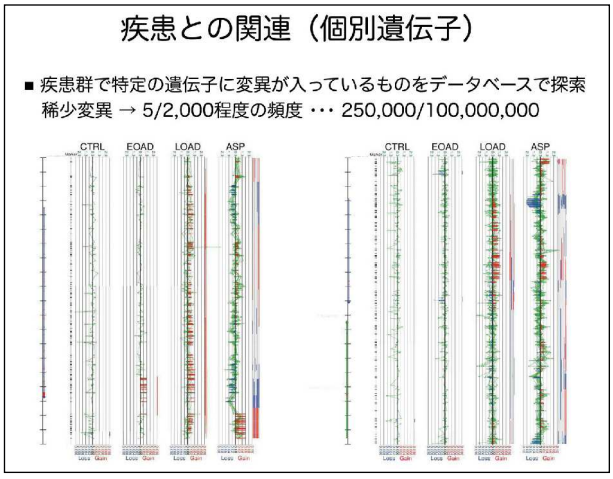
最後になりますが、アルツハイマー病とゲノム配列のコピー数変異について紹介したいと思います。これは、CNV（copy number variation・コピー数変異）と呼ばれるものですが、未発症約 1,000 人、晩期発症型約 1,000 人、早期発症型 250 人程度で調べたデータがありますので、これを紹介します。

●コピー数変異のゲノムワイドな分布

我々は 2 倍体ですので、2 つの対立遺伝子をもっています。血液型が AA や BO という具合に 2 文字で表されるのもそのためです。通常は、この対立遺伝子のコピー数は父方と母方それぞれから由来する 2 個なのですが、中には、3 個や 4 個に増えていたり、1 個や 0 個に減っている領域がゲノム配列中にあることが分かっています。それらを先程の検体群ごとに集計して染色体に沿ってプロットしてみますと、我々のゲノムのコピー数はゲノム配列の多くの領域で多様に増減していることが分かります。このような前提で様々なデータを見る必要があるのかも知れません。例えば、3 コピーある場合に AAC なのか ACC なのか、SNP で A と C を 2 種類もっているとしていても、それらをコピー数を加味しないで見ているのは実は十分ではないのかも知れません。また、公開されている異なる人種のデータを見ていくと、日本人に共通しているものもありますし、中国系、ヨーロッパ系、アフリカ系の人のデータと比べてみても、人類共通でコピー数の変異が多い領域があることも分かってきます。

●疾患との関連

アルツハイマー病の場合は、高い頻度で染色体のどこかが大きく増えていたり抜けていたりするようなことはなく、狭い領域の低頻度なコピー数の変化がゲノム中に散在して見えてきます。健常者と発症者の間で違いがありそうなことが分かってきています。実際に、このデータでもあるアルツハイマー病に関係していることが示唆されている遺伝子の領域が LOAD 検体群でコピー数が増えている例が見つかってきて、2,000 人中 5 人くらいの頻度でコピー数が増えています。この位の頻度になりますと、稀少変異と言われて頻度自体は確かに低いのですが、



2,000 人で 5 人だとしても、1 億人当たりにすると 25 万人くらいになりますので、このような知見をしっかり拾っていくことが重要だと思います。このような変異は他にも見つかってきており、我々は疾患関連の変異情報のデータベースを作っています。

●おわりに

結果の妥当性の評価については、いくつかの段階があると考えています。まずは計算科学的なデータ処理の正しさがあります。次には統計学的な正しさであり、結果が統計的に有意なのかどうかということです。そしてなかなか難しいのが生物医学的な正しさです。先ほどのように、疾患に関連しているかも知れない変異を見つけてきたといっても、データの中に記述された事実としては正しくてもそれが本当に疾患と関係しているかどうかは、実験系や臨床系の方に調べてもらうことが必要となります。その辺りを含めて、うまくコラボレーションしていくことが大事なことだと考えております。