特集

疾患関連遺伝子探索へのIT適用の実際

Example Application of Information Technology in the Search for Disease-Related Genes

あらまし

マイクロサテライトマーカ（microsatellite marker：2～4塩基反復）やSNP（Single Nucleotide Polymorphism：一塩基多型）に代表されるゲノム配列上の多型マーカを用いた遺伝統計解析は，現在，疾患関連遺伝子探索手法の最も有効な研究手法の一つであり，その適用研究が進められている。

本稿では，徳島大学ゲノム機能研究センターとの2年間に及ぶ共同研究において経験した遺伝統計解析の理論と解析手順，IT（Information Technology）適用の実例を紹介し，その有効性および課題を具体的に述べる。また，遺伝子多型解析において重要な多型マーカについても解説する。

Abstract

Genetic statistical analysis of genome sequences using polymorphic markers—among which, microsatellite markers (repetitions of two to four nucleotides) and single nucleotide polymorphisms (SNPs) are typical examples—is one of the most effective investigative procedures in the search for disease-related genes. Significant advances are being made in the study of the application of the genetic statistical analysis procedure. We researched genetic statistical analysis in collaboration with the Institute for Genome Research of the University of Tokushima. This paper gives a detailed description of the theory and procedure of the genetic statistical analysis that we researched and problems encountered during the research and an example application of IT in the search for disease-related genes and the application's effectiveness. This paper also explains polymorphic genetic markers, which are important for polymorphic genetic analysis.

手塚　理（てづか　おさむ）

（株）富士通長野システムエンジニアリングR&Dソリューション部 所属
現在，バイオインフォマティクス，とくに遺伝統計学を用いた疾患関連遺伝子探索研究に従事。

安藤美紀（あんどう　みき）

（株）富士通長野システムエンジニアリングR&Dソリューション部 所属
現在，バイオインフォマティクス，とくに遺伝統計学を用いた疾患関連遺伝子探索研究に従事。

内藤公敏（ないとう　きみとし）

ライフサイエンス推進室 所属
現在，創薬研究ソフトウェアシステムの研究開発に従事。

ま え が き

A（アデニン），T（チミン），G（グアニン），C（シトシン）の四つの塩基から成るゲノム塩基配列（遺伝情報）は，同じ生物種であっても個体間に若干の差異，すなわち多型がある。ゲノム塩基配列の多型（個体間の遺伝形質の差異）と疾患の有無や程度といった個体間における表現形質の差異との関連を解析することにより，疾患に関連する遺伝子座（染色体上の位置）を推定することができる。しかし，個体ごとの全ゲノム塩基配列を完全に調べ，比較するのは現段階において不可能である。そこで，あらかじめ差異があると分かっている部位（多型マーカ）を調べ，その近傍領域における多型の代表として扱う。これは遺伝情報がある領域（塊）として次世代に伝わるという現象を利用しており，その多型マーカの多型パターンと疾患との関連について各種統計学的手法を用いて解析することにより，疾患に関連するゲノム上の領域（遺伝子座）を推定することができる。これを遺伝子多型解析という。このようにして推定された遺伝子座を，より詳細に調べ，疾患の原因となる多型を持った遺伝子を見つけるのが疾患関連遺伝子探索の目的である。

本稿では，著者らが共同研究先である徳島大学ゲノム機能研究センターで行っている遺伝子多型解析を用いた疾患関連遺伝子探索研究について，その探索手順を解説し各手順におけるIT（Information Technology）適用の実際について紹介する。[1), 2)]

多型解析による疾患関連遺伝子探索の手順

以下に遺伝子多型を用いた疾患関連遺伝子探索の手順を示す（**図-1**）。遺伝統計解析手法には種々のものがあるが，マイクロサテライトマーカを用いた連鎖解析を例に解説する。

● 連鎖解析による疾患関連候補領域の同定

連鎖解析は患者家系における各サンプル（個人）のマイクロサテライトマーカによる遺伝子多型と疾患の有無である表現形質の個人間の差異から，疾患と連鎖している遺伝子座を推定するものである。ゲノムワイドで数百か所程度のマイクロサテライトマーカを選定し，患者家系の個人ごとのDNAサンプルを用い，選定したマイクロサテライトマーカの多型パターンを検出する。そのデータをコンピュータに取り込み，連鎖解析アプリケー

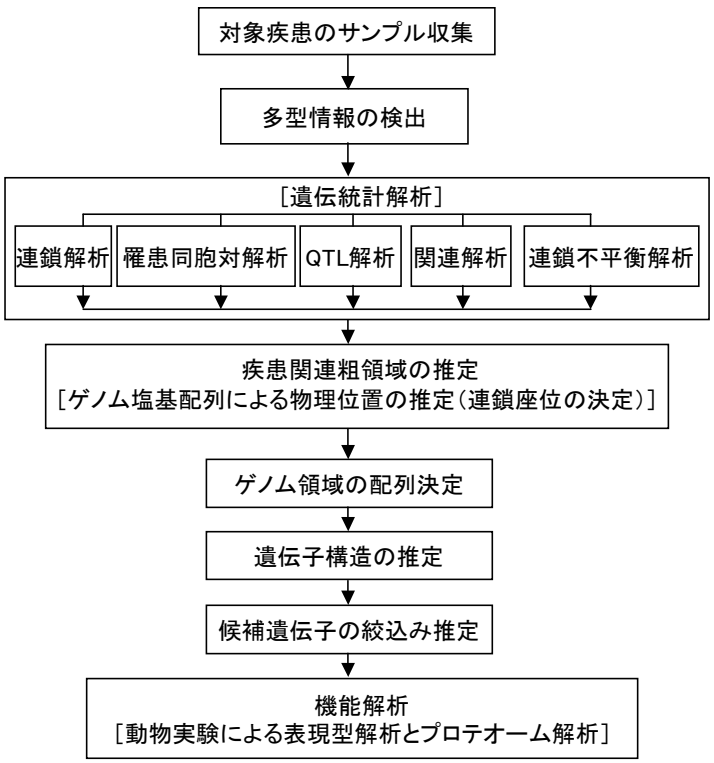

図-1　遺伝子多型を用いた疾患関連遺伝子探索手順
Fig.1-Disease related gene search procedure used gene polymorphism.

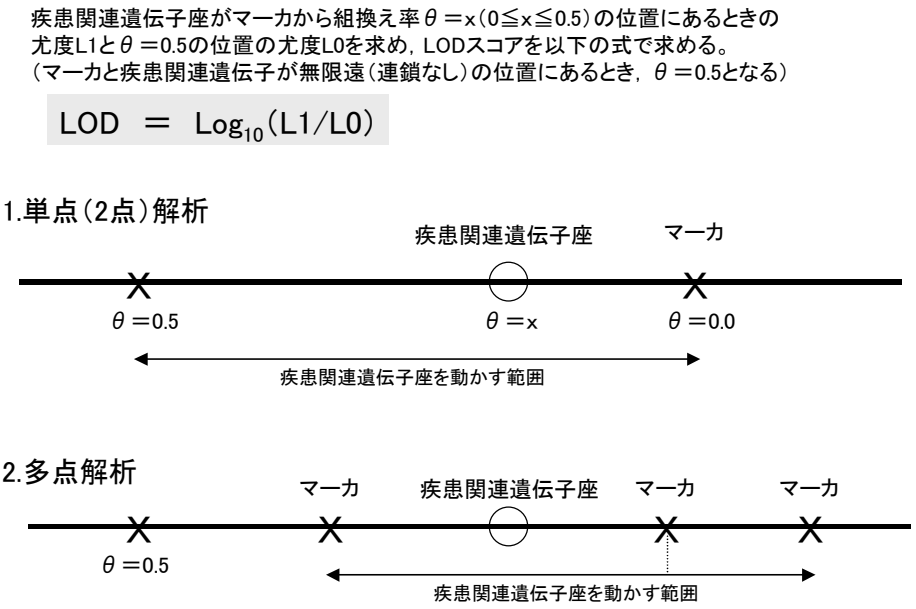

図-2　連鎖解析の概念
Fig.2-Conception of linkage analysis.

ションを用いて解析する。連鎖解析で用いられる計算式を**図**-2に示す。連鎖解析の結果は，遺伝地図上に疾患との連鎖の度合いの分布を与える。遺伝子座と疾患との連鎖の度合いを対数尤度(ゆうど)比，LOD（Logarithm of odds）スコアで示し，この対数尤度比分布の中で高いLODスコア（通常は3以上）を示す領域が疾患関連候補領域である。ゲノムワイドで単点（2点）解析を行い，LODスコアの高かった領域（3以上）について多点解析を行って更に詳しく調べる。

患者家系が大家系であるなど，サンプル条件が良ければ比較的容易にLODスコア3を超える可能性は高いが，小さい家系の場合は適正なピークが容易に得られず苦労する。ただし比較的小さい家系であってもPCR（Polymerase Chain Reaction）の反応条件をサンプルごとに変えるなどして可能な限り欠測値をなくすなどの工夫により，大家系のサンプル条件に近づけた結果，良好なピークを得られることもある。

また同種の解析手法であっても使用するアプリケーションやパラメタによってかなり値が振れることもある。複数アプリケーションでの比較検証や内部的なアルゴリズムを理解した上でのパラメタ設定などが，解析を実施する上で重要である。

●　疾患関連候補領域の正確なマップ作成

疾患関連候補領域の同定後，その領域内の正確なマップ作成がたいへん重要である。まずは候補領域のゲノム塩基配列の取得について説明する。

ヒトの全ゲノム塩基配列は，そのドラフト解読が2001年2月に完了し発表された。疾患関連候補領域の塩基配列について，この研究成果を活用することは非常に有効である。解読されたヒトのゲノム塩基配列データは，公共DNAデータベースであるDDBJ（DNA Data Bank of Japan），EMBL（European Molecular Biology Laboratory），Genbankに蓄積され，上述の公開サイトから提供されている。疾患関連候補領域の塩基配列は，もしその領域が解読済であれば，遺伝子多型解析に用いたマイクロサテライトマーカをキーに，上記の公共DNAデータベースの塩基配列を検索し取得することが可能である。

取得したBAC（Bacterial Artificial Chromosome）クローンなどのゲノム塩基配列を用いてゲノム塩基配列間のホモロジー検索を丁寧に繰り返し，重なりをつなげることにより，可能な限り正確にゲノム塩基配列を連結しコンティグ配列を作成する。コンティグ配列については既に公開提供されているものも多くあるが，ゲノムワイドで機械的に作成されたコンティグだけでは，まだ精度に問題がある場合が多い。狭い範囲（疾患関連候補領域）の可能な限り正確なマップ作成には公開されている複数のコンティグ配列を参考にしながら，ある程度手作業で独自アッセンブリを行う必要がある。

コンティグ配列を更につないでマップの基礎となる疾

患関連候補領域のゲノム塩基配列を作成する。そのゲノム塩基配列に連鎖解析で使用したマイクロサテライトマーカを含めた位置に関係する情報を注意深く詳細に貼り付けていく。

このときにITとして重要なのはスケールの感覚である。例えばゲノム塩基配列と一言で言ってもゲノム全体を扱う場合，ある遺伝子座を扱う場合，1遺伝子のエクソン/イントロン構造も含めた詳細な情報を扱う場合など，扱う領域の大きさによって全く別なものと考えたほうがよい。元となるデータはATGCの4塩基から成る1次元の文字列データで共通であっても必要とされるスケールによって求められる品質やデータ量，表示形式が大きく異なるからである。

● 疾患関連遺伝子探索

疾患関連遺伝子探索は，疾患関連候補領域のゲノム塩基配列をもとにその領域に含まれる遺伝子部位を予測することから始まる。

遺伝子部位の予測には大きく二つの方法がある。一つは，EST（Expressed Sequence Tag）と呼ばれる発現遺伝子の配列タグ情報や発現遺伝子の配列情報を疾患関連候補領域の塩基配列上にアラインメントプログラムを用いてマッピングする方法である。発現遺伝子の塩基配列データは，データベースのUniGeneなどに蓄積し，公開されている。また，ESTデータはデータベースのdbESTあるいは公共データベースのDDBJやEBI，GenbankのEST divisionなどに蓄積し，公開されている。もう一つの方法は，遺伝子部位の塩基の並びに関する種々の規則に基づいて，塩基配列から計算だけで遺伝子部位を予測するものである。これらの遺伝子部位の予測には，Grail，Gene Finderなど，種々のプログラムがある。疾患関連候補領域内の既知遺伝子，予測遺伝子，また更に各遺伝子のエクソン/イントロン構造を含む，詳細な情報，多型マーカ（とくにSNP）などの位置関係を明確にし，一つひとつの遺伝子と疾患との関連を解析していく。ITを活用することよってこれらの作業はかなり軽減できるが，まだコンピュータで完全自動化できるようなルーチンワークにはなっていないため，バイオとインフォマティクス研究者による半自動的な煩雑な作業が発生している。

疾患関連遺伝子探索に用いられる多型マーカ

現在，遺伝子多型解析に用いられている主な多型マーカは，マイクロサテライトマーカ（microsatellite marker）とSNP（Single Nucleotide Polymorphism）の2種類である。徳島大学ゲノム機能研究センターとの共同研究におけるある疾患関連遺伝子探索を進める過程で，従来から一般的に使用されている多型マーカが少ない領域での解析が必要となった。そこで試用し，良い成果を出した新たな多型マーカであるSNR（Single Nucleotide Repeat）についても紹介する（**図-3**）。

● マイクロサテライトマーカ

基本となる2～4の塩基の繰返し（10～30回）から成るマーカで，ゲノム塩基配列上には数万塩基対に1個程度存在すると言われている。多型パターンが多く，突然

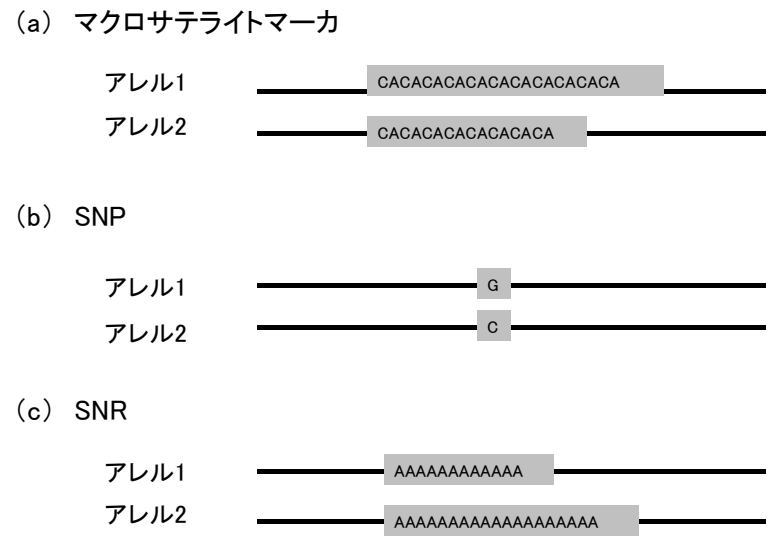

図-3　多型マーカの例
Fig.3-Example of polymorphic marker.

変異が起きやすいと考えられている。ゲノム塩基配列さえあれば，コンピュータを使いテキスト検索を行う簡単なプログラムツールで容易に発見できる。しかし，発見したもの（2～4塩基の繰返し）のすべてが多型解析に使用できるとは限らず，多型解析で使用する前に多型パターンが適当であるかの選別が必要である。

● SNP

点変異（point mutation，点突然変異）の中で，一般集団においてマイナーalleleが1%以上のものをいう。ゲノム上には非常に稠密（ちゅうみつ）に存在しており，約千塩基対ごとに一つ，ゲノム全体では数百万個程度存在すると思われる。実験における多型のパターン分けの自動化が容易で，広範囲かつ大量の解析を行う場合非常に有効である。ただしSNPを発見するにはウェットの実験が必要であり，マイクロサテライトマーカやSNRのようにコンピュータで簡単に見つけることはできない。

● SNR

一つの塩基が10～30回繰り返したもので，ゲノム上では数千塩基対に1個程度存在する。1塩基の繰返しであるSNRは，2～4塩基単位の繰返しであるマイクロサテライトマーカより出現頻度がかなり高い。共同研究者の実験報告によると，多型パターンはマイクロサテライトマーカより少なく，SNPよりは若干多い。マイクロサテライトマーカと同様，ゲノム塩基配列さえあればコンピュータを使いテキスト検索を行う簡単なプログラムツールで容易に発見できる。しかし，発見したものがすべて多型解析に使用できるとは限らず，多型パターンが適当であるかは多型解析で使用する前に選別が必要である。マイクロサテライトマーカと同様，容易に発見でき，多型パターン数はSNPに近く，対象疾患の探索遺伝子座によって非常に有効な多型マーカと成り得る。またSNRの99%以上がAまたはTの繰返しであり，CまたはGの繰返しはほとんど存在しない。これは非常に興味深い事実である。

以上述べた多型マーカについて，どの多型マーカがより有用であるかということではなく，解析条件によって使い分け，あるいは組み合わせて使用することが大切である。

むすび

遺伝子多型を用いた疾患関連遺伝子探索をはじめ，今日のバイオ研究分野では日々膨大な情報が創出されている。これらを有効に活用するためにはコンピュータが不可欠である。外部（インターネット）からの情報や自研究室の実験データなどの膨大な情報を整理し，情報間の関連付けを行ったり，また複雑な統計解析の計算を行ったりするのはITの得意とする分野だからである。今回，徳島大学ゲノム機能研究センターの最先端バイオ研究者と著者らが密接な共同研究を行った結果，遺伝子多型を用いた疾患関連遺伝子探索研究が飛躍的に進んだ。さらにSNRという新しい多型マーカを用いた解析の可能性の発見にもつながった。

その一方，現在のバイオ分野特有の難しい問題にも直面した。インターネットなどを介して膨大なデータが簡単に入手できるのと同時に，それぞれのデータの信憑性（しんぴょうせい）検証や関連付けなどの膨大な作業が発生したのである。少しデータを集めるとすぐにデータ間の矛盾が見つかり，それについての調査検討をしなければならなかった。これは実験解析機器の驚くほどの発達で実験の高速化がなされ，生物のゲノム塩基配列解読と解読されたゲノム塩基配列へのアノテーション（注釈付け）などが急速に進み，今までに経験したことのない大量で多岐にわたるデータが爆発的に発生し，同種のデータであってもデータの生成時期や提供研究機関によって品質や精度が異なり，不確かなデータが大量に存在するようになってしまっているからだと考えられる。この傾向は今後収束されるのではなく，さらに加速されると思われる。バイオ研究分野のデータは，今日までにITが扱ってきた種類のデータとは少し様子が異なることを理解しなければならない。

最後に共同研究先である徳島大学ゲノム機能研究センターの板倉光夫センター長をはじめ，ご指導いただいた多くの先生方に深謝します。

参考文献

(1) 鎌谷直之ほか：ポストゲノム時代の遺伝統計解析．羊土社，2001年10月．

(2) 内藤公敏ほか：遺伝子多型に基づく疾患関連遺伝子探索におけるインターネット利用．細胞工学，Vol.20，No.12，p.1618-1623（2001）．