論文番号： C5-4

SNP および臨床データベースを対象としたハプロタイプ解析による知識発見方式とその実現

小川健二[†1] 吉田尚史[†2] 清木康[†3]
藤島清太郎[†4] 相磯貞和[†4]

本論文では，SNP データベースおよび臨床データベースを対象としたハプロタイプ解析による知識発見方式とその実現について示す．本方式は，個人差を規定する因子として着目されている遺伝子上の多型，特に一塩基多型（Single Nucleotide Polymorphism, SNP）のデータベースと臨床データベースとの組み合わせを対象として，相関ルール抽出アルゴリズムを適用することにより，SNP と臨床情報間の相関ルールを効率的に抽出する方式である．一般に SNP データベースを対象として相関ルールを単純に適用した場合，組み合わせ数が膨大となり，実際的に許容できる時間内に相関ルールを抽出できない．本方式は，全 SNP より遺伝子上に近接して存在する複数の SNP を遺伝子の機能単位で一括抽出するハプロタイプ解析をヒューリスティクスとして用いて，相関ルールを効率的に抽出する方式として位置付けられる．本論文では，実験により本方式の実現可能性および有効性を検証する．

1. はじめに

ヒトゲノムの約 30 億塩基対の配列の決定が報告されて以来，ライフサイエンスは急激な発展を遂げている．とりわけ遺伝子の解析とその遺伝子が人に与える影響は，重要な研究の対象となっている．

本論文では，遺伝子の多型のデータベース (SNP データベース) および臨床情報のデータベース (臨床データベース) を対象としたハプロタイプ解析による知識発見方式とその実現について示す．本方式は，個人差を規定する因子として着目されている遺伝子上の多型，特に 一塩基多型（Single Nucleotide Polymorphism, SNP）のデータベースと臨床データベースとの組み合わせを対象として，相関ルール抽出アルゴリズムを適用することにより，SNP と臨床情報間の有効な相関ルールを効率的に抽出する方式である．

SNP データベースおよび臨床データベースを対象として SNP - 臨床情報間の関連を抽出する場合，相関ルール抽出アルゴリズム[1)] の適用が有効である．しかし，単純に相関ルールを適用した場合，SNP の組み合わせ数が膨大な数となり，現在の計算機で実際的に許容できる時間内に相関ルールを抽出することは困難である．SNP - 臨床情報間の関連を抽出することは医学的に緊急の課題である．計算量を削減し，臨床情報と関連する可能性が高い組み合わせのみに着目して相関ルールを抽出する方法が有効である．

本方式は，全 SNP より遺伝子上に近接して存在する複数の SNP を遺伝子の機能単位で抽出するハプロタイプ解析をヒューリスティクスとして用いて，効率的に相関ルールを抽出する方式として位置付けられる．

本方式では，特に SNP を対象とするが，遺伝子多型を対象とした場合も一般性を失わない．

従来の遺伝子を解析する方法では，主に機能的アプローチおよび遺伝統計的アプローチの２つのアプローチが採用されていた[3)]．前者は，遺伝子に含まれる SNP を分子生物学的に分析することにより SNP の機能を解析するアプローチである．後者は，SNP を臨床情報と

†1 慶應義塾大学 SFC 研究所
Keio Research Institute at SFC
e-mail： ko@mdbl.sfc.keio.ac.jp
†2 慶應義塾大学 政策・メディア研究科
Graduate School of Media and Governance, Keio University
e-mail： naofumi@sfc.keio.ac.jp
†3 慶應義塾大学 環境情報学部
Faculty of Environmental Information, Keio University
e-mail： kiyoki@sfc.keio.ac.jp
†4 慶應義塾大学 医学部
Faculty of Medical Sciences, Keio University
e-mail： {fujishim,aiso}@sc.itc.keio.ac.jp

組み合わせ, 遺伝統計的に関連を抽出するアプローチである．遺伝統計学の分野において，遺伝子データベースおよび臨床データベースを対象として，両データベースに潜在する関連を統計的に抽出する方式が研究されている．遺伝統計学は，両データベースに潜在する関連の傾向を統計的に抽出することを可能としている．

これらの従来のアプローチと比較して，本方式の特徴は，SNP データベースおよび臨床データベースの組み合わせを対象として相関ルールを適用することにより，SNP と臨床情報間の有効なルールを分析的かつ効率的に抽出可能な点である．本方式の特徴は，さらに，分析の方法を解析対象データの属性に応じて段階的に設定し，分析者は分析時間と分析の精度のトレードオフを自由に選択可能である点である．

本論文では，実験により本方式の実現可能性および有効性を検証する．

1.1 遺伝子多型および SNP の定義

人の遺伝子の塩基の配列において，個体間に 1 % 以上存在する変異の事を遺伝子多型 (genetic polymorphism) と呼ぶ．特に，1 個の塩基について存在する多型を一塩基多型（Single Nucleotide Polymorphism, SNP）と呼び，遺伝子多型の中でも特に頻度が高いことから特に着目されている．図 1に, DNA 塩基配列と SNP の関係を示す．人における個人差は，この SNP の違いにより大部分が決定されることが推測されている．

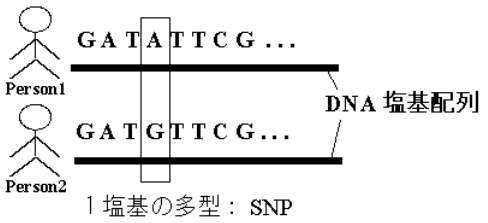

図 1 DNA 塩基配列と SNP

SNP の遺伝子機能や遺伝子発現に与える影響を解明することにより，病気にかかりやすい体質をつきとめたり，個人の体質に合わせたよりよい治療法, 薬剤の選択や医薬品の開発が可能になると考えられている[4),5)]．

本方式は，この SNP に着目し，臨床情報と組み合わせて有効な関連を抽出することを目的としている．

1.2 ハプロタイプの定義

本方式では，全 SNP より遺伝子上に近接して存在する複数の SNP を，遺伝子の機能単位で一括抽出した組をハプロタイプ（haplotype）と呼ぶ．図 2に, DNA 塩基配列, および SNP とハプロタイプの関係を示す．

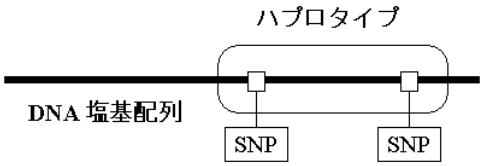

図 2 DNA 塩基配列, および SNP とハプロタイプの関係

本方式は, ハプロタイプ解析を用いて, 相関ルール抽出の対象データの組み合わせ数を削減し, 効率的に有効な相関ルールを抽出することを可能とする．

2. SNP および臨床データベースを対象とした知識発見方式

2.1 概　　要

本方式は，SNP データベースと臨床データベースを対象として知識発見を行い，SNP と臨床情報間に存在する有用なルールの抽出を目的とする．さらに, 知識発見方法において遺伝情報に特有のヒューリスティクスを用いて，すべての組み合わせを対象として網羅的に解析を行った場合に比べて, 段階的に計算量を削減する方式を設定する．これにより，SNP の組み合わせが多い場合においても実際的に許容できる時間内での相関ルール抽出が可能となる．

2.2 SNP および臨床データベースを対象とした知識発見方式

X,Y を塩基とすると，SNP は 1 塩基部位に対して X/X, X/Y, Y/Y の 3 パターンを取る. ただし,X, Y は A（アデニン）,T（チミン）,G（グアニン）,C（シトシン）いずれかの塩基を現す.

表 1 臨床情報および SNP データのデータ構造とその例

臨床情報			SNP データ		
疾患名	···	手術歴	SNP_1	···	SNP_m
疾患 A		有	A/T		A/G
疾患 B		無	A/A		A/A
疾患 A		無	A/T		G/G
···		···	···		···
疾患 C		有	A/A		G/A

本方式では，表 1に示すようなデータ構造を対象として，相関ルール抽出アルゴリズム[1)] を次のように適用する．

A を臨床情報の属性, S_i を SNP データの属性, C_i を属性 i を属性値として持つという条件とする．A と S_1 ～ S_n の相関ルール抽出[1)] において, (1), (2) で定義される Confidence（確信度）を計算する．ただし Support(x) は，データベース中の全属性値のうち条件

x を満たす割合とする． A は, 臨床情報中の 1 属性とする．

$$\text{Confidence}(C_{S_1} \wedge \cdots \wedge C_{S_n}, C_A) = \frac{\text{Support}(C_{S_1} \wedge \cdots \wedge C_{S_n} \wedge C_A)}{\text{Support}(C_{S_1} \wedge \cdots \wedge C_{S_n})} \quad (1)$$

$$\text{Confidence}(C_A, C_{S_1} \wedge \cdots \wedge C_{S_n}) = \frac{\text{Support}(C_{S_1} \wedge \cdots \wedge C_{S_n} \wedge C_A)}{\text{Support}(C_A)} \quad (2)$$

本方式は，SNP のすべての組み合わせのうち，臨床情報と関連する可能性が高い組み合わせのみに着目して，式 (1), 式 (2) で示される値を計算し，最小 Confidence 以上の値を持つ組み合わせをルールとして抽出する．

また，本方式では，SNP について遺伝子情報に特有のヒューリスティクスを用いて, 相関ルール抽出の計算量を削減する方式を示す．SNP データベースおよび臨床データベースを対象とした, 相関ルール抽出アルゴリズムによる知識発見方法において，次の 4 つの性質をヒューリスティクスとして用い, 計算量を削減する．

本方式で用いる４つのヒューリスティクス

(1) 疾患と特に関連が疑われる遺伝子
特定の疾患について, 遺伝子の機能的解析によって特に関連が疑われる遺伝子が存在する．その特定の遺伝子上に存在する SNP を対象に, 優先的に解析を行う．

(2) SNP の存在する塩基配列の構成
遺伝子である DNA 塩基配列の構成は次のように定められている．プローモーター領域とは mRNA への転写を制御する領域, またコード領域（エクソン）とはタンパク質をコードする領域である．コード領域（エクソン）間に存在する非コード領域を，イントロンと呼ぶ．本方式では，プローモーター領域を DNA 塩基配列において, コード領域（エクソン）の約 1.5kb 前までとする．ただし, kb とは千塩基対（kilobase pair）であり, p kb とは, p 千塩基対の距離を示す．一般的にプロモーター領域およびコード領域に存在する SNP は, それ以外の領域にある SNP より疾患などの臨床情報との関連が高いと言える．よって, プローモーター領域とコード領域に存在する SNP を優先して相関ルール抽出を行うことにより, 疾患などの臨床情報との関連性を有効に解析することが可能である．図 3に, SNP の存在する塩基配列の構成を示す．

(3) 近傍の遺伝子群と臨床情報との相関

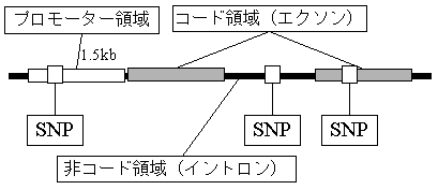

図 3 SNP の存在する塩基配列の構成

一般に，1 つのタンパク質を規定するコード領域およびプロモータ領域は，互いに近傍に存在する (gene)．よって, 近傍に存在する SNP の組み合わせをハプロタイプとして分析対象とすることで, 臨床情報との関連を有効に解析することが可能である．

(4) common disease common variant hypothesis
本方式では, ハプロタイプのうち高頻度のハプロタイプを優先的に相関ルール抽出の対象とする．遺伝子が親から子へ遺伝する際，塩基の配列は，近傍にある塩基ほど関連して伝わる．この現象を連鎖不平衡と呼ぶ．ハプロタイプは，連鎖不平衡により生じる塩基の変異の組である．連鎖不平衡により, それぞれのハプロタイプに頻度の差が生じる．一般に高頻度のハプロタイプは, 特定の臨床情報との相関が高いと言われている．この現象は，「common disease common variant hypothesis」と呼ばれる[3]．

2.3 本方式の構成

2.2 節における (1) 〜 (4) までのヒューリスティクスを利用した, SNP データを対象とした相関ルール抽出アルゴリズムを用いた知識発見の方法を, 下記のように設定する．相関ルールを抽出するための組み合わせを，Method-1 から Method-5 まで段階的に設定し，相関ルール抽出による知識発見方式の実行時間と精度を分析者が自由に設定できる環境を実現する．

また本方式では Method-1 から Method-5 までの各段階において, さらにハプロタイプの性質を 2 つに分類し, それぞれの分類に対応する分析方法を次の Option-1 および Option-2 として設定する．

Method-1: 全 SNP の全ての組み合わせを対象とした相関ルール抽出

Method-2: 特定のタンパク質をコードするコード領域，およびその前 1.5kb 塩基のプロモーターをハプロタイプと見なし，その領域内の k 個の gene 間の SNP の組み合わせを対象とした相関ルール抽出

（この時，複数の gene をまたがった SNP の組み合わせも対象とする．）

Method-3: 特定のタンパク質をコードするコード領域，およびその前 1.5kb 塩基のプロモーターをハプロタイプと見なし，その領域内の k 個の gene 間の SNP の組み合わせを対象とした相関ルール抽出（この時，複数の gene 間の SNP について，一定の距離以内の SNP の組み合わせを対象とする．）

Method-4: 特定のタンパク質をコードするコード領域，およびその前 1.5kb 塩基のプロモーターをハプロタイプと見なし，その領域内の k 個の gene 間の SNP の組み合わせを対象とした相関ルール抽出（この時，SNP の gene をまたがった組み合わせは対象としない．）

Method-5: 疾患との関連が特に疑われる遺伝子上に存在する SNP とその組み合わせを対象とした相関ルール抽出

Option-1: ハプロタイプのうち, 前処理として高頻度のハプロタイプのみを抽出し，分析対象ハプロタイプとして設定する．本方式では, あらかじめ高頻度のハプロタイプの抽出が行われていることを前提とする.

Option-2: ハプロタイプの全ての組み合わせを, 分析対象ハプロタイプと設定する.

2.4 Method-1: 全 SNP の全ての組み合わせを対象とした相関ルール抽出

全 SNP の全ての組み合わせを対象とした相関ルール抽出を行う．解析対象の gene が k 個存在する場合，Method-1 で解析対象とする SNP を図 4に示す.

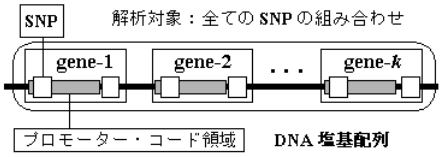

図 4 Method-1 で解析対象とする SNP

この場合，相関ルールを抽出するための計算量は，式 (3) となる．ただし，臨床データベースの属性 i が持つ属性値の種類の数を a_i，属性数を n とし，SNP データの属性数を m とする.

$$\overline{O}(\prod_{i=1}^{n} a_i \sum_{j=1}^{m} {}_m\mathrm{C}_j 3^j) \tag{3}$$

また，式 (3) は，式 (4) と簡略化できる.

$$\overline{O}(\prod_{i=1}^{n} a_i 4^m - 1) \tag{4}$$

Method-1 では, 全 SNP に存在するルールを全て抽出することが可能であるが, 他の Method と比較して計算量が大きい.

2.5 Method-2: 特定のタンパク質をコードするコード領域，およびその前 1.5kb 塩基のプロモーターをハプロタイプと見なし，その領域内の k 個の gene 間の SNP の組み合わせを対象とした相関ルール抽出（この時，複数の gene をまたがった SNP の組み合わせも対象とする．）

特定のタンパク質をコードするコード領域，およびその前 1.5kb 塩基のプロモーターをハプロタイプと見なし，その領域内の k 個の gene 間の SNP の組み合わせを対象とした相関ルール抽出を行う．この時，複数の gene をまたがった SNP の組み合わせも対象とする．Method-2 で解析対象とする SNP を図 5に示す.

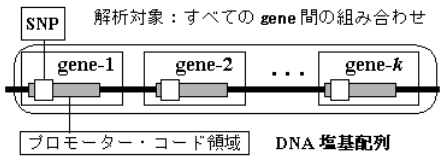

図 5 Method-2 で解析対象とする SNP

この場合， gene でコード領域およびプロモータに存在する SNP の平均塩基数を $l(l < \frac{m}{k})$ とする相関ルールを抽出するための計算量は，式 (5) となる.

$$\overline{O}(\prod_{i=1}^{n} a_i \sum_{j=1}^{l \cdot k} {}_{l \cdot k}\mathrm{C}_j 3^j) \tag{5}$$

Method-2 では, コード領域およびプロモーター領域に存在する全ての SNP の組み合わせを対象とするため, コード領域およびプロモーター領域以外に存在する SNP 間に存在するルールを発見できない可能性がある.

2.6 Method-3: 特定のタンパク質をコードするコード領域，およびその前 1.5kb 塩基のプロモーターをハプロタイプと見なし，その領域内の k 個の gene 間の SNP の組み合わせを対象とした相関ルール抽出（この時，複数の gene 間の SNP について，一定の距離以内の SNP の組み合わせを対象とする．）

Method-2 において，解析対象とする複数の gene 間の組み合わせについて， gene 間の距離が一定個 $w(w \leq k)$ までの gene 間の組み合わせに限定し，解析対象とする． Method-3 で解析対象とする SNP を図6に示す.

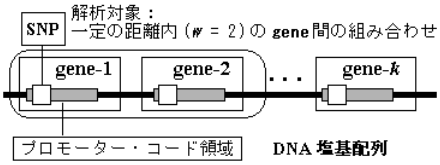

図 6 Method-3 で解析対象とする SNP

この場合, 相関ルールを抽出するための計算量は，式 (6) となる.

$$\overline{O}(\prod_{i=1}^{n} a_i(k-(w-1)) \sum_{j=1}^{l \cdot w} (3^j {}_{l \cdot w}\mathrm{C}_j)) \qquad (6)$$

Method-3 では, 一定の距離内の gene 間の組み合わせのみを対象としているため, 距離の離れた gene 間に存在するルールを発見できない可能性がある.

2.7 Method-4: 特定のタンパク質をコードするコード領域，およびその前 1.5kb 塩基のプロモーターをハプロタイプと見なし，その領域内の k 個の gene 間の SNP の組み合わせを対象とした相関ルール抽出（この時，SNP の gene をまたがった組み合わせは対象としない．）

Method-1 および Method-2 で解析対象とする複数の gene の中で， gene 間の組み合わせを解析対象とせず，各 gene 内での組み合わせを解析対象とする. Method-4 で解析対象とする SNP を図7に示す.

この場合，相関ルールを抽出するための計算量は，式 (7) となる.

$$\overline{O}(\prod_{i=1}^{n} a_i k \sum_{j=1}^{l} (3^j {}_l\mathrm{C}_j)) \qquad (7)$$

Method-4 では, gene 内のみの SNP の組み合わせを対象としているため, 計算量は少ないが, gene 間にまたがった重要なルールを発見できない可能性がある.

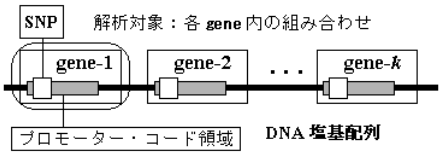

図 7 Method-4 で解析対象とする SNP

2.8 Method-5: 疾患との関連が特に疑われる遺伝子上に存在する SNP とそのハプロタイプを対象とした相関ルール抽出

臨床データベース中の疾患との関連が特に疑われる特定の遺伝子上に存在する SNP を対象とする. Method-5 で解析対象とする SNP を図8に示す.

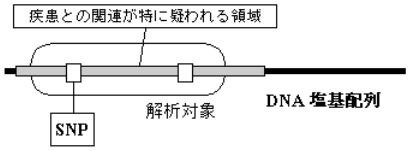

図 8 Method-5 で解析対象とする SNP

疾患との関連性が疑われる特定の遺伝子上に SNP が s 個存在する場合，それらを対象として相関ルールを抽出するための計算量は，式 (8) となる.

$$\overline{O}(\prod_{i=1}^{n} a_i \sum_{j=1}^{s} {}_s\mathrm{C}_j 3^j) \qquad (8)$$

Method-5 では, 特定の SNP データの属性を対象としているため, 計算量は少ないが, 重要なルールを発見できない可能性がある.

2.9 Option-1: ハプロタイプのうち，前処理として行われたハプロタイプ解析により得られた高頻度の組を，分析対象ハプロタイプとして設定した相関ルール抽出.

ハプロタイプのうち, 前処理として行われたハプロタイプ解析により得られた高頻度の組を, 分析対象ハプロタイプと設定する．この場合，分析対象ハプロタイプの数を h, そのハプロタイプを構成する SNP データの属性数を $\#h$, Method-1 〜 Method-5 の各段階において, 対象 SNP データの属性数を x , 式 (3) 〜 式 (8) の $\sum$ の係数を d とすると．相関ルールを抽出するための計算量は式 (9) となる.

$$\overline{O}(\prod_{i=1}^{n} a_i d \sum_{j=1}^{x-\#h} ({}_{x-\#h}\mathrm{C}_j 3^j) 2^h) \qquad (9)$$

2.10 Option-2: ハプロタイプの全ての組を, 分析対象ハプロタイプと設定した場合の相関ルール抽出

この場合の計算量は, Method-1 から Method-5 の各方法における計算量である式 (3) ～式 (8) と同値である.

3. 実　　験

3.1 実 験 方 法

本方式の実現可能性および有効性を示すことを目的として実験を行う. 2 節で示した Method-1 ～ Method-5, および, Option-1 と Option-2 を比較することにより, 相関ルール抽出のための計算の実行時間と, 抽出されたルールの妥当性を検証する.

相関ルール抽出のための実行時間について, Method-1 から Method-5, および Method-2 において Option-1 と Option-2 を実行した時の相関ルールを抽出するための計算の実行時間を計測した.

本実験では, 抽出された相関ルールの精度を次のように設定した. まず相関ルール抽出において抽出されるべき正解ルールを文献[6] を参照して設定した. ただし正解ルールの数は, 次に示す実験 1 で 15 件, 実験 2 で 17 件, 実験 3 で 18 件である. さらに, Method-1 ～ 5 の各方法, および Method-2 における Option-1 と Option-2 において, 2.2 節の式 (1) と式 (2) の Confidence の値が 0.4 以上のものを各 Method において抽出された相関ルールと設定し, 正解ルールと比較することにより, 再現率および適合率を計測した.

ただし, 各 Method において抽出された相関ルールの数を Ra, 各 Method において抽出された相関ルールに含まれる正解ルールの数を Rb, 各実験において設定した正解ルールの数を Rc とすると, 再現率は式 (10), 適合率は式 (11) である.

$$\text{再現率} = \frac{Rb}{Rc} \tag{10}$$

$$\text{適合率} = \frac{Rb}{Ra} \tag{11}$$

3.2 実 験 環 境

実験システムと実験データの作成について述べる. 実験を行う計算機として Sun Enterprise 3500, OS は Solaris 2.6 を使用した. また DBMS として PostgreSQL 7.0.2 を使用し, プログラムの実装には Java1.3.1, および JDBC を使用した.

実験対象データについて述べる. 臨床データベースとして疾患名等の 6 属性を, SNP データベースとして 1 カラムにつき 3 種類のデータを文献[6] を参照して作成し, データ件数が 100 件のデータを作成した.

3.3 実 験 結 果

3.3.1 実　験 1

実験 1 として, 対象 SNP データの属性数について gene 数を固定 k とし, 1-gene に含まれる SNP データの属性数を変化させた場合の相関ルール抽出を行った. ただし, この実験で k は 2 とした.

対象データとして SNP データの属性数が 10,20,30,40,50 のデータを対象とし, Method-1 ～ 5 の方法を適用し, その実行時間および抽出されたルールの精度を検証した. ただし, gene 数 k が 2 と固定であるため Method-2 と Method-3 は同値である. よって Method-2 をもって, Method-3 の実行結果とした. また, Method-1 については, 解析対象 SNP データの属性数 s を 2 とした.

図 9に, 各 Method において 1-gene に含まれる SNP データの属性数を変化させた場合の実行時間に関する実験結果を示す. また, 図 10に各 Method において 1-gene に含まれる SNP データの属性数を変化させた場合の再現率および適合率を示す.

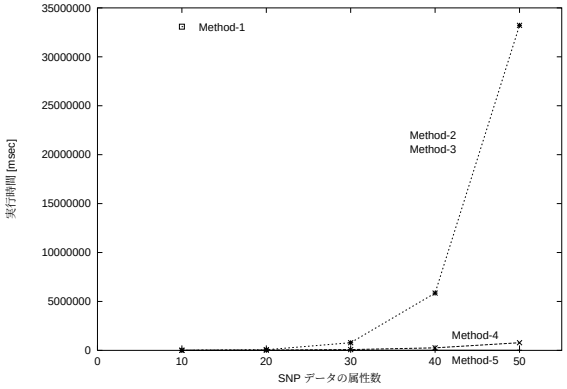

図 9 各 Method において SNP データの属性数を変化させた場合の実行時間（実験 1）

3.3.2 実　験 2

実験 2 として, 対象 SNP データについて 1-gene に含まれる SNP の数を固定 t とし, 含まれる gene 数を変化させた場合の相関ルール抽出を行った. ただし, この実験で t は 4 とした.

対象データとして gene 数が 1 ～ 8 までのデータを作成し, Method-1 ～ 5 の各方法を適用し, その実行時間および抽出されたルールの精度を検証した. また, Method-1 については, 解析対象 SNP データの属性数 s を 2 とした.

図 11に, 各 Method において gene 数を変化させた

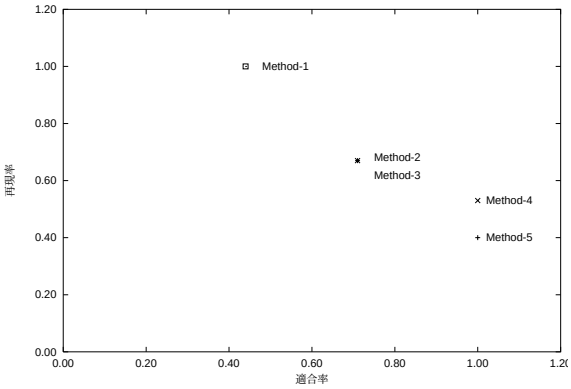

図 10 各 Method において SNP データの属性数を変化させた場合の再現率と適合率（実験 1）

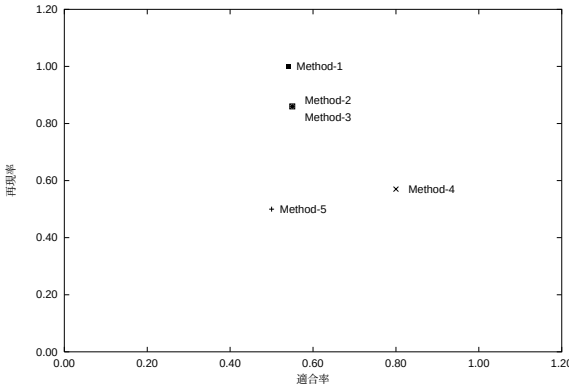

図 12 各 Method において gene 数を変化させた場合の再現率と適合率（実験 2）

場合の実行時間に関する実験結果を示す．また, 図 12に各 Method において gene 数を変化させた場合の再現率および適合率を示す.

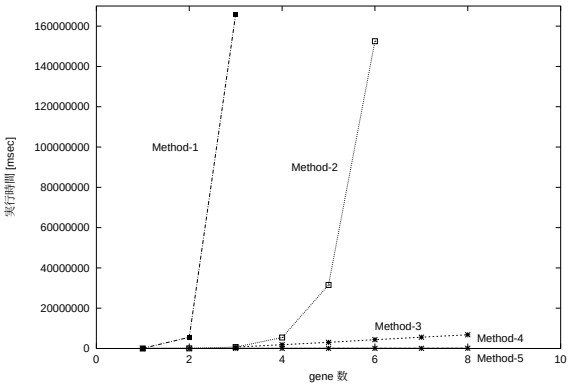

図 11 各 Method において gene 数を変化させた場合の実行時間（実験 2）

3.3.3 実　験 3

実験 3 として, 対象 SNP データについてそのハプロタイプ頻度を考慮した場合の相関ルール抽出を行った. 全てのハプロタイプを分析対象とした場合と, 頻度の高い特定のハプロタイプのみを対象とした場合について相関ルール抽出を行った．具体的には Method-2 について, Option-1 および Option-2 の方法を適用した.

図 13に Method-2 における Option-1 および Option-2 について gene 数を変化させた場合の実行時間に関する実験結果を示す．また, 図 14に Method-2 における Option-1 および Option-2 について gene 数を変化させた場合の再現率および適合率を示す.

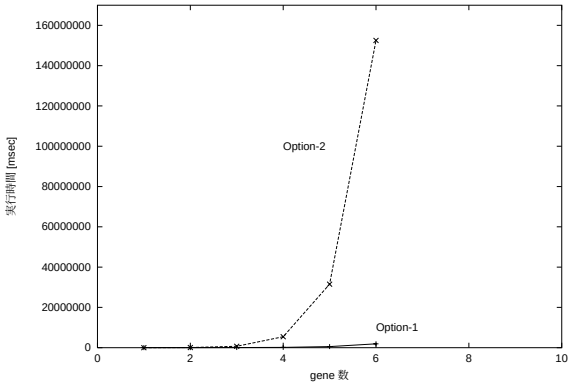

図 13 Method-2 における Option-1 および Option-2 について gene 数を変化させた場合の実行時間（実験 3）

3.4 考　　察

3.4.1 実験 1 についての考察

各 Method において SNP データの属性数を変化させた場合の相関ルールを抽出するための計算の実行時間に関して, 次のように考察できる．図 9より, 網羅的なハプロタイプ解析を行った場合（Method-1）, およびプロモーターとコード領域における gene 間のハプロタイプ解析を行った場合（Method-2, Method-3）では, 解析対象となるハプロタイプ数が指数関数的に増加するため, その実行時間も急激に増加する．そのため, Method-1 では SNP データの属性数が 10 の場合には解析可能であったが, 属性数が 20 以上の場合には実行時間という点において, 知識発見が非常に困難である．同様に Method-2 と Method-3 についても SNP データの属性数が増加した場合には, 実際的に許容できる時間内でのルール抽出が困難である．それに対して, Method-5 においては, 解析対象 SNP データの属性数が一定であるため, 実行時間は SNP データの属性数に

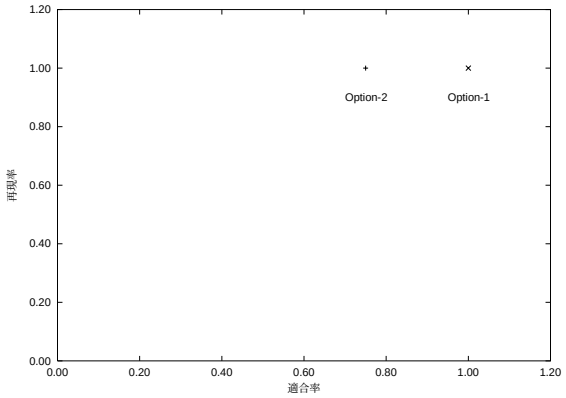

図 14 Method-2 における Option-1 および Option-2 において gene 数を変化させた場合の再現率と適合率（実験 3）

依存しない．また, Method-4 においては, 解析対象とするハプロタイプ数が線形に上昇するため, 実行時間もほぼ SNP データの属性数の増加に比例する．

また, 抽出されたルールの再現率および適合率に関して, 図 10より次のように考察できる．まず, Method-1 では, 抽出されたルールの再現率は 1.0 であり, すべての正解ルールを抽出できている反面, 適合率は低い値を示している．また, Method-2 と Method-3 では, Method-1 と比較して, 抽出されたルールの再現率は低い値を示しているが, 適合率は高い値を示しており, 効率的に正解ルールを抽出できている．また同様に, Method-4 および Method-5 では, Method-2, Method-3 と比較して, 抽出されたルールの再現率は低い値を示しているが, 適合率は 1.0 と高い値を示しており, 効率的に正解ルールを抽出できている．

3.4.2 実験 2 についての考察

各 Method において gene 数を変化させた場合の相関ルールを抽出するための計算の実行時間に関して, 次のように考察できる．図 11より, 網羅的なハプロタイプ解析を行った場合（Method-1）, およびプロモーターとコード領域における gene 間のハプロタイプ解析を行った場合（Method-2）では, 解析対象となるハプロタイプ数が指数関数的に増加するため, その実行時間も急激に増加する．そのため, Method-1 では gene 数が 3 の場合には解析可能であったが, gene 数が 4 以上の場合には実行時間という点において, 知識発見が非常に困難である．同様に Method-2 についても gene 数が増加した場合には, 実際的に許容できる時間内でのルール抽出が困難である．それに対して, Method-5 においては, 解析対象 SNP データの属性数が一定であるため, 実行時間は gene 数に依存しない．また, Method-4 および Method-3 においては, 解析対象とするハプロタイプ数が線形に上昇するため, 実行時間もほぼ gene 数の増加に比例する．

また, 抽出されたルールの再現率および適合率に関して, 図 12より次のように考察できる．まず, Method-1 では, 抽出されたルールの再現率は 1.0 であり, すべての正解ルールを抽出できている反面, 適合率は低い値を示している．Method-2 および Method-3 では, Method-1 と比較して, 再現率は低い値を示しているが, 適合率はわずかに高い値を示している．また, Method-4 では, 抽出されたルールの再現率は, Method-3 と比較して低い値を示しているが, 適合率は高い値を示しており, 効率的に正解ルールを抽出できている．また, Method-5 では, 抽出されたルールの再現率および適合率は, Method-4 と比較して低い値を示しているが, 実行時間は SNP データの属性数に依存しない．

3.4.3 実験 3 についての考察

図 13より, 全てのハプロタイプを対象とした場合, および頻度の高いハプロタイプのみを対象とした場合の実行時間について, 次のように考察できる．Option-2 では実験 2 で考察した通り, gene 数の増加に伴い, 急激な実行時間の増加が見られる．それに対して, Option-1 では gene 数に伴って, 実際的に許容できる実行時間の増加に留まる．

また, 抽出されたルールの再現率および適合率に関して, 図 14より次のように考察できる．Option-1 は Option-2 と比べて, 再現率は同じ値であるが, 適合率は高い値を示している．

以上 2 点は，組み合わせ数が多く実行時間の点において分析が困難な場合，本方式の提供する Option-1 により，困難であった分析が，より高い適合率で可能となることを示している．

3.4.4 実験全体についての考察

網羅的，およびヒューリスティクスを取り入れた場合の相関ルール抽出による段階的な知識発見方法を実験し, 実行時間と抽出されたルールの精度についての両面から, その有効性を実証した．

Method-1 におけるすべての SNP およびそのハプロタイプについての網羅的な知識発見方法では計算量が莫大にかかり, 実際的に許容可能な時間内での解析が困難な場合がある．そのような場合において, Method-2 〜 Method-5 のヒューリスティクスを取り入れた段階的な知識発見方法を適用することにより, 臨床情報および SNP データとの実際的に許容可能な時間内での知識発見が可能であると伴に, 一定の精度の相関ルールの効率的な抽出が可能であることを実証した．

本方式で示している段階的な知識発見方法を用いるこ

とにより, 具体的には次に示すような知識発見が有効であると考えられる. 分析対象の SNP データの属性数および gene 数が大きい場合は, まず実行時間が 分析対象データの属性数に依存しない Method-5 , あるいは実行時間が分析対象データの属性数に比例して, ほぼ線形に増加する Method-4 を適用した知識発見を行う. 分析対象の SNP データの属性数および gene 数が比較的少ない場合は, Method-3 または Method-2 を適用して知識発見を行う. さらに, 分析対象の SNP データの属性数および gene 数が少量の場合, あるいは網羅的な相関ルール抽出を行う必要のある場合には, Method-1 を適用した知識発見を行うことが適切である. また, 各 Method においても, Option-2 に優先して Option-1 を適用した知識発見を行うことにより, 短い実行時間での有効な相関ルール抽出が可能である.

本方法により, SNP および臨床データベースを対象とした知識発見において, 知識発見対象である SNP データの属性数および gene 数に応じて, 知識発見の実行時間および精度を, 分析者が自由に設定する知識発見が実現可能である.

4. おわりに

本論文では, SNP データベースおよび臨床データベースを対象としたハプロタイプ解析による知識発見方式とその実現について示した.

本方式は, 個人差を規定する因子として着目されている遺伝子上の SNP のデータベースと臨床データベースとの組み合わせを対象として, 相関ルール抽出アルゴリズムを適用することにより, SNP と臨床情報間の相関ルールを効率的に抽出する方式である. 本方式は, 全 SNP より遺伝子上に近接して存在する複数の SNP を遺伝子の機能単位で一括抽出するハプロタイプ解析をヒューリスティクスとして用いて, 相関ルールを効率的に抽出する方式として位置付けられる. 本論文では, 実験により本方式の実現可能性および有効性を確認した.

今後の課題としては, 本方式の実際の医療現場での利用が挙げられる. 本論文では, 文献[6] を参照して作成したデータを対象として実験を行ったが, その次の段階として, 実際の医療データである SNP データベースおよび臨床データベースを対象とした知識発見を行い, その実現可能性および有効性を確認することが必要であると考えられる. SNP データや臨床情報などの実際の医療データを対象とする際には, 分析対象とする患者のプライバシー保護に充分留意した対応が必要である. 患者のプライバシー保護に配慮した, SNP データ, 臨床データ等の医療データベースの構築, および知識発見のためのシステムの構築が重要であると考えられる.

本方式は, 従来は医者の経験, 分子生物学, および遺伝統計学により特定していた臨床情報と SNP データとの関連性を, 相関ルール抽出アルゴリズムを使って網羅的に抽出する方式である. 本方式により抽出された臨床情報と SNP データとの関連性は, 直接的に医療に取り入れられるものではなく, 医者の診断の指針となる判断材料として位置付けられる. 抽出されたルールの妥当性を検討するためには, 抽出されたルールについて, その有意性を検定する方式の実現が必要である.

また, 本方式で対象としている密 (dense) なデータについて, 有用なルールを抽出する方式が提案されている. [7] この方式との比較において, 本方式では解析対象を, 解析データ特有のヒューリィスティクスを用いて段階的に設定することにより, 網羅的な相関ルール抽出に比べて有効なルールの抽出を実現している. 本方式への, この方式の導入は今後の課題とする.

参 考 文 献

1) Agrawal, R., and Srikant, R.: "Fast Algorithms for Mining Association Rules," Proc. of the 20th International Conference on Very Large Data Bases, pp.487-489, 1994.
2) Jiawei Han, Micheline Kamber: "Data mining : concepts and techniques", Morgan Kaufmann Publishers, 2001.
3) 鎌谷直之 (編):"ポストゲノム時代の遺伝統計学", 羊土社, 2001.
4) 中村裕輔 (編):"SNP 遺伝子多型の戦略", 中山書店, 2000.
5) 中村裕輔:"先端のゲノム医学を知る", 羊土社, 2000.
6) Matthew Stephens, Nicholas J. Smith, and Peter Donnelly: "A New Statistical Method for Haplotype Reconstruction from Population Data," Am. J. Hum. Genet., Vol. 68, pp.978-989, 2001.
7) Roberto J. Bayardo Jr., Rakesh Agrawal, and Dimitorios Gunopulos: "Constraint-Based Rule Mining in Large, Dense Databases" Proc. of the 15th Int'l Conf. on Data Engineering, 1999.