データベースシステム 128－59
（２００２．７．１９）

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システムの実現方式

吉田　尚史† 　清木　康†† 　藤島清太郎††† 　相磯　貞和†††

† 慶應義塾大学大学院 政策・メディア研究科
†† 慶應義塾大学 環境情報学部
††† 慶應義塾大学 医学部
E-mail: †{naofumi,kiyoki}@sfc.keio.ac.jp, ††{fujishim,aiso}@sc.itc.keio.ac.jp

あらまし　本稿では，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システムの実現方式について示す．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 SNP (Single Nucleotide Polymorphism: 一塩基多型) 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，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用いて，SNP と臨床情報間の関連を効率的に抽出する方式である．本方式は，さらに，大量のデータから短い時間で知識発見を行う必要のある場合，および，精密な知識発見を行う必要のある場合のどちらの場合においても効率的な知識発見を可能とする．ここでは，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．

キーワード　SNP，臨床データベース，ハプロタイプ解析，知識発見

An Implementation Method of a Knowledge Discovery System for SNP and Clinical Databases with Haplotype Analysis

Naofumi YOSHIDA†, Yasushi KIYOKI††, Seitaro FUJISHIMA†††, and Sadakazu AISO†††

† Graduate School of Media and Governance,Keio University
†† Faculty of Environmental Information, Keio University
††† School of Medicine, Keio University
E-mail: †{naofumi,kiyoki}@sfc.keio.ac.jp, ††{fujishim,aiso}@sc.itc.keio.ac.jp

Abstract　In this paper we present an implementation method of a knowledge discovery system for SNP and clinical databases with haplotype analysis. SNPs (Single Nucleotide Polymorphism) are genetic individual variation and will accelerate the identification of disease genes. Our method makes it possible to extract associations between SNP data and clinical data for SNP databases and clinical databases with haplotype analysis, which extracts sets of SNPs for each genetic function from genetic bases. Our method also enables both efficient knowledge discovery for large databases and precise knowledge discovery. We clarify feasibility and effectiveness of our method by several experiments.

Key words　SNP, Clinical Databases, Haplotype Ananysis, Knowledge Discovery

1. はじめに

現在，ゲノム科学が急激な発展を遂げている．人の持つゲノムの全塩基配列の決定が報告され，主要な研究の対象は，ゲノムの塩基配列の決定から，塩基配列が人に与える影響へと移りつつある．

一般に人の持つ約 30 億塩基対のうち 99.9 %は共通であり，0.1 %は個人により差異が存在すると言われている．この差異のうち個人間において 1 %以上の頻度で差異が存在し，かつ，1 塩基のみ差異が SNP (Single Nucleotide Polymorphism: 一塩基多型) と呼ばれており，個人差を決定付ける因子として重要な研究対象となっている．現在では，ポスト・ゲノム・プロジェクトの一つとして，ヒトゲノム上の SNP バンク構築を目指す SNP Consortium [8] や JSNP [3] において， SNP のデータベースが構築されつつある．

本稿では， 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システムの実現方式について示す．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 SNP (Single Nucleotide Polymorphism: 一塩基多型) 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，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用いて， SNP と臨床情報間の関連を効率的に抽出する方式である．

本方式の特徴は，次の 2 点にまとめられる．第 1 に， SNP データベースと臨床データベースの組み合わせを対象として，SNP の膨大な組み合わせからハプロタイプ解析により組み合わせの数を削減し， SNP と臨床情報間の関連を効率的に抽出することを可能とする．第 2 に，本方式は，大量のデータから短い時間で知識発見を行う必要のある場合，および，精密な知識発見を行う必要のある場合のどちらの場合においても効率的な知識発見を可能とする．

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，単純に相関ルール抽出アルゴリズムを適用する場合， SNP の組み合わせ数が膨大な数となり， NP 困難 の問題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用いて，効率的に相関ルールを抽出する方式である．本方式では，特に SNP を対象とするが，他の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 [4].

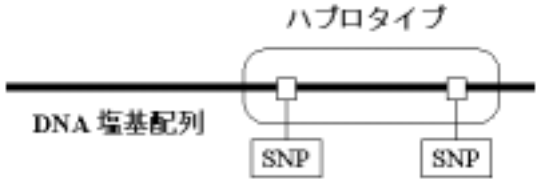

図 1　DNA 塩基配列, および SNP とハプロタイプ

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 を臨床情報と組み合わせ, 統計的に関連を抽出するアプローチであり, 遺伝統計学と呼ばれる．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網羅的に抽出可能な点である．

本稿では，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．

2. 遺伝子多型，SNP，およびハプロタイプ解析

2.1 遺伝子多型および SNP

人の遺伝子の塩基の配列において，固体間に 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 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．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．

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 [5], [6].

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．

2.2 ハプロタイプ

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 1に, DNA 塩基配列, および SNP とハプロタイプの関係を示す.

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.

3. 実 現 方 式

3.1 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．知識発見においてハプロタイプ解析を用いて，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と比較して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．

3.2 SNP および臨床データベースを対象とした知識発見システムの方式

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．

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，臨床データベースの属性数（カラム数）を n とし，SNP データの属性数（カラム数）を m とした場合，SNP および臨床情報間の相関ルール抽出において，SNP すべての組み合わせを反映した相関ルールを抽出する計算量は式 (1) となる．ここで $\overline{O}$ は, 相関ルールを抽出するための計算量を示す．ただし，${}_x\mathrm{C}_y$ は x 個から y 個を選択する組み合わせの数を示す．

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

式 (1) は，m が大きい場合には実際的に許容できる時間内での計算が困難であることを示している．本方式では，SNP の組み合わせを削減するハプロタイプ解析を用いて, 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において，次の 4 種類の性質を用いて, 計算量を削減する [7].

（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．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 2に, SNP の存在する塩基配列の構成を示す．

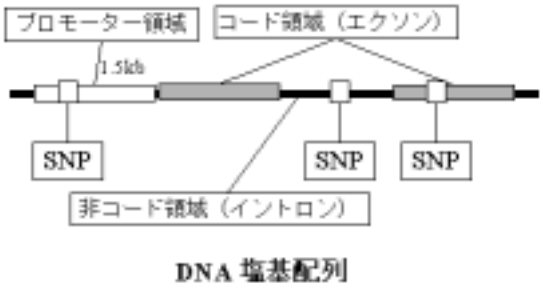

図 2　SNP の存在する塩基配列の構成

（3） 近傍の遺伝子群と臨床情報との相関

一般に，1 つのタンパク質を規定するエクソンおよびプロモータ領域は，同一染色体上に互いに近接して存在する (gene). よって, 近傍に存在する各 gene 上の SNP を一括し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 [4].

3.3 本方式の構成

3.2節における (1) 〜 (4) までの性質を利用した, 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:　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

Method-1 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．

Method-2:　特定のタンパク質をコードするエクソン，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 k 個の 同一 gene 内の SNP を対象とした相関ルール抽出（このとき，それぞれの gene 内の SNP 群を独立と見なし,

SNP の gene をまたがった組み合わせは対象としない．）

Method-2 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.

Method-3: Method-2 において，k が 2 以上の場合の gene 間の SNP の組み合わせを対象とした相関ルール抽出（このとき，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-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.

Method-4: Method-2 において，k が 2 以上の場合の gene 間の SNP の組み合わせを対象とした相関ルール抽出（このとき，複数の gene をまたがった SNP の組み合わせも対象とする．）

Method-4 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.

Method-5: 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-5 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.

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.

3.4 Method-1: 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

臨床データベースの属性 i が持つ属性値の種類の数を a_i，属性数（カラム数）を n とし，SNP データの属性数（カラム数）を m とする.

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. 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 (2) となる.

$$\overline{O}(\prod_{i=1}^{n} a_i \sum_{j=1}^{s} {}_s\mathrm{C}_j 3^j) \tag{2}$$

3.5 Method-2: 特定のタンパク質をコードするエクソン，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 k 個の gene 内の SNP を対象とした相関ルール抽出（このとき, それぞれの gene 内の SNP 群を独立と見なし, SNP の gene をまたがった組み合わせは対象としない.）

この場合， gene 内の平均塩基数を $l(l < m)$ とすると相関ルールを抽出するための計算量は，式 (3) となる.

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

3.6 Method-3: Method-2 において，k が 2 以上の場合の gene 間の SNP の組み合わせを対象とした相関ルール抽出（このとき，それぞれの gene 間において一定の距離以内の組み合わせのみを対象とする.）

Method-2 において，n が 2 以上の場合の gene 間の組み合わせをハプロタイプと見なし, 分析対象とする. この場合、gene 間の距離が一定個 w までの gene をハプロタイプとすると, 相関ルールを抽出するための計算量は，式 (4) となる.

$$\overline{O}(\prod_{i=1}^{n} a_i (k-(w-1)) \sum_{j=1}^{l \cdot w} (3^j {}_{l \cdot w}\mathrm{C}_j)) \tag{4}$$

3.7 Method-4: Method-2 において，k が 2 以上の場合の gene 間の SNP の組み合わせを対象とした相関ルール抽出（このとき，複数の gene をまたがった SNP のハプロタイプも対象とする．）

この場合，相関ルールを抽出するための計算量は，式 (5) となる. ただし式 (5) は, Method-3 の式 (4) において $w = l$ とした場合と同値である.

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

3.8 Method-5: 全 SNP の全ての組み合わせを対象とした相関ルール抽出

臨床データベース中の属性と SNP の全ての組み合わせを対象とした相関ルールを抽出するための計算量は，式 (6) となる. これは式 (1) と同値である.

$$\overline{O}(\prod_{i=1}^{n} a_i \sum_{j=1}^{m} {}_m\mathrm{C}_j 3^j) \tag{6}$$

3.9 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.

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 (2) ～ 式 (6) 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 (7) となる.

$$\overline{O}(\prod_{i=1}^{n} a_i d \sum_{j=1}^{x-\#h} ({}_{x-\#h}\mathrm{C}_j 3^j) 2^h) \quad (7)$$

3.10 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 (2) ～式 (6) と同値となる.

4. 実　　験

4.1 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. 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.

本実験において, 各 Option において抽出する相関ルールを次のように設定する. 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 $S_1 \sim S_n$ の相関ルール抽出 [1] において, (8), (9) 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.

$$\mathrm{Confidence}(C_{S_1} \wedge \cdots \wedge C_{S_n}, C_A) = \frac{\mathrm{Support}(C_{S_1} \wedge \cdots \wedge C_{S_n} \wedge C_A)}{\mathrm{Support}(C_{S_1} \wedge \cdots \wedge C_{S_n})} \quad (8)$$

$$\mathrm{Confidence}(C_A, C_{S_1} \wedge \cdots \wedge C_{S_n}) = \frac{\mathrm{Support}(C_{S_1} \wedge \cdots \wedge C_{S_n} \wedge C_A)}{\mathrm{Support}(C_A)} \quad (9)$$

さらに各 Option において抽出された全相関ルールについて, まず値 (8) による Confidence, 次に値 (9) による Confidence の高さに応じてランキングを行う.

相関ルール抽出のための実行時間について, 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 [9] を参照して設定した. ただし正解ルールの数は, 17 件である. さらに, Option-1 と Option-2 において, 値 (8) と値 (9) の Confidence の値が 0.4 以上のものを各 Option 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 Option において抽出された相関ルールの数を Ra, 各 Option 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.

$$\text{再現率} = Rb/Rc \quad (10)$$

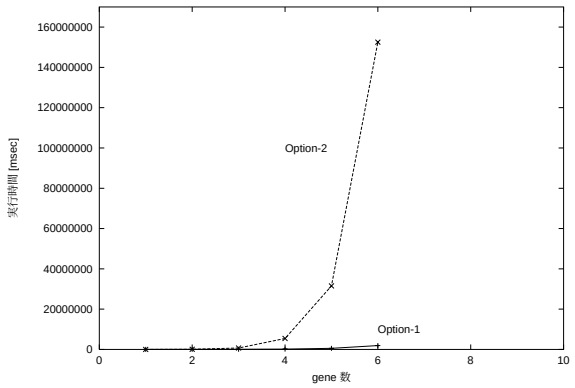

図 3　Option-1 および Option-2 について gene 数を変化させた場合の実行時間

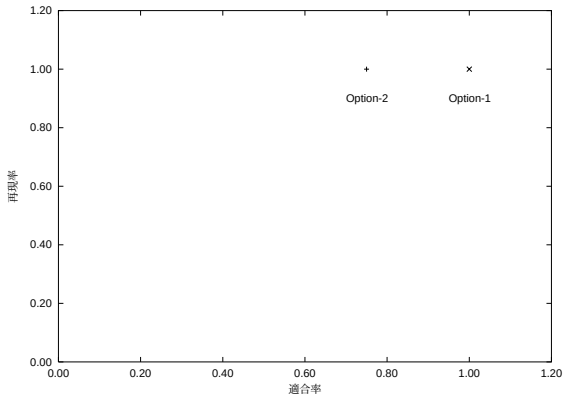

図 4　Option-1 および Option-2 において gene 数を変化させた場合の再現率と適合率

$$\text{適合率} = Rb/Ra \quad (11)$$

本実験では, 3.2節に示した Method のうち, Method-4 を用いて実験を行った.

4.2 実験環境

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.

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 [9] 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.

4.3 実験結果

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 Option-1 および Option-2 の方法を適用した.

図 3に 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 4に 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.

4.4 考　　察

図3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, 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4より次のように考察できる. Option-1 は Option-2 に比べて, 再現率および適合率ともに低い値を示しているが, gene 数の増加に伴う実行時間の増加が小さい. さらに, Option-1 では, Option-2 と比較して, 抽出されている相関ルールにノイズが少ないことを示している.

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可能となり, さらに抽出されている相関ルールにノイズが少なくなることを示している.

SNP の全て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した場合, NP 困難 の問題となり, 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. そのような場合において, Option-1 に示す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, ノイズの少ない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知識発見の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比例して, ほぼ線形に増加する Option-1 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, Option-2 を適用して知識発見を行う. 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, Option-2 を適用した知識発見を行うことが適切である.

本方式により, 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. 以上の結果は, 本方式は大量のデータから短い時間で知識発見を行う必要のある場合, および, 精密な知識発見を行う必要のある場合のどちらの場合においても効率的な知識発見を可能とすることが示された.

5. おわりに

本稿では, 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システムの実現方式について示した.

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用いることにより, 相関ルールを効率的に抽出することを可能とする. 本稿では, 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 SNP データベースおよび臨床データベースを対象とした実験が挙げられる. 本稿では, 文献 [9] 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. また, 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保護に充分留意した対応が必要であるため, 患者のプライバシ保護に配慮した, SNP データ, 臨床情報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.

謝　　辞

実験システムの構築についてご協力頂いた河本穣氏（慶應義塾大学大学院政策・メディア研究科）小川健二氏（慶應義塾大学，現在，京都大学大学院情報学研究科）に感謝致します.

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] JSNP Database: http://snp.ims.u-tokyo.ac.jp/ .
[4] 鎌谷直之 (編):“ポストゲノム時代の遺伝統計学”, 羊土社, 2001.
[5] 中村裕輔 (編):“SNP 遺伝子多型の戦略”, 中山書店, 2000.
[6] 中村裕輔:“先端のゲノム医学を知る”, 羊土社, 2000.
[7] 小川健二, 吉田尚史, 清木康, 藤島清太郎, 相磯貞和: ”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,” 電子情報通信学会 第 13 回データ工学ワークショップ (DEWS2002) 論文集, 2002.
[8] The SNP Consortium: http://snp.cshl.org/ .
[9] 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.