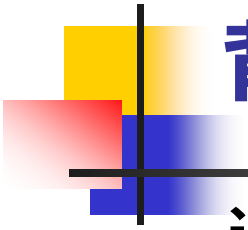
SNPおよび臨床データベースを対象とした ハプロタイプ解析による知識発見方式とその実現

小川健二[†], 吉田尚史[††], 清木康[‡], 藤島清太郎[‡‡], 相磯貞和[‡‡]

[†] 慶應義塾大学 SFC研究所
[††] 慶應義塾大学 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



背景

- 近年のライフサイエンスの急激な発展
  - 2000年：ヒトゲノムの約 30 億塩基配列の解読ほぼ終了
  - 遺伝子解析が人間に与える影響の重要性

- 遺伝子の配列情報から，その機能的意味の解析へ
  - 個人の臨床情報と遺伝子データの関連
  - 個別医療（テイラーメイド医療）や，効果的な医薬品の開発へ

- 対象となる遺伝子データ数は膨大
  - 単純な知識発見方式では，現在の計算機で現実的な時間内に遺伝子，臨床情報間の関連を抽出することは困難（NP完全）
  - 両データ情報間の関連を抽出することは，医学的に急務の課題
  - 計算量を削減し，現実的な時間内に関連を抽出する方法が有効

概要

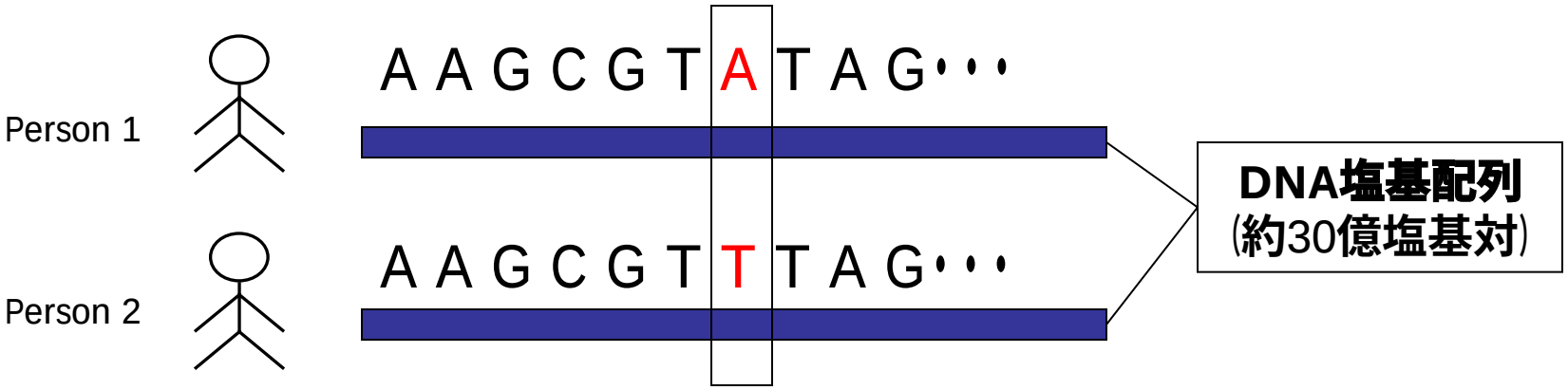
- 遺伝子および臨床データベースを対象とした相関ルール抽出による知識発見方式
  - 遺伝子データ特有のヒューリスティクス (ハプロタイプ解析) を用いて，相関ルールを現実的に許容できる時間内で効率的に抽出する方式
  - 分析時間と精度のトレードオフに基づいて，ヒューリスティクスを選択的に分析方法に適用

従来の方法との位置づけ

- 分子生物学的アプローチ
  - 遺伝子を分子的に分析することにより，その機能を解析
- 遺伝統計学的アプローチ
  - 遺伝子と臨床情報間に潜在する関連の傾向を抽出
  - 少数の候補遺伝子に対する有意性の検定
- 本研究におけるアプローチ
  - 相関ルール抽出によって遺伝子と臨床情報間の関連を体系的かつ直接的に抽出
  - 抽出されたルールは，医療の判断指針を与える位置付け

SNP (Single Nucleotide Polymorphism)

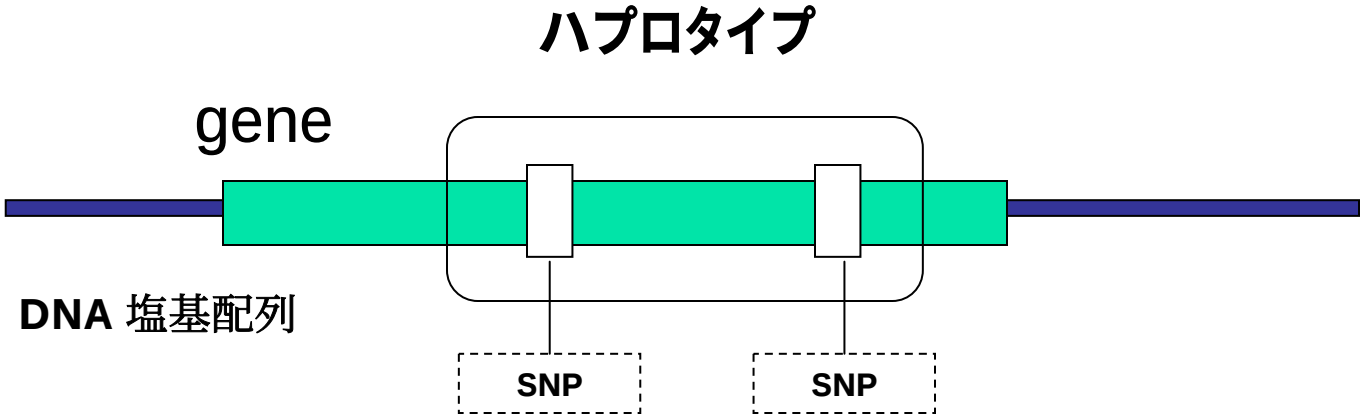
- **人の遺伝子のDNA塩基配列において，個体間に 1% 以上存在する変異**
  ：多型(polymorphism)
- **1塩基について存在する多型**
  ：一塩基多型 (Single Nucleotide Polymorphism, SNP)
- **個人差を規定する因子として着目される．**
  - **体質・病気のかかりやすさ等の個人差は，この SNP の違いにより決定される．**



1塩基の違い：SNP

遺伝子 (gene) , SNPとハプロタイプ

- **1つの遺伝子** (gene) **は, 複数の** SNP **を持つ.**
- **ハプロタイプ** (haplotype) **とは, 遺伝子上に近接して存在する複数の** SNP **の組**
- **本方式では, ハプロタイプを用いて計算量を削減.**



gene, SNPとハプロタイプの関係

対象となるデータ構造とデータ例

- 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.
  - *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

臨床情報			SNP データ		
疾患名	・・・	入院暦	SNP1	・・・	SNPm
疾患A		有	A/T		A/G
疾患B		無	A/A		A/A
疾患A		無	T/T		G/G
・・・		・・・	・・・		・・・
疾患C		有	A/A		A/G

相関ルール

- *A* と *S1* ～*Sn*の相関ルール抽出において，式(1)，(2)で定義される Confidence (確信度) を計算し、その値(最小Confidence)が閾値以上のものをルールとして抽出する．(※)
  - *A* を臨床情報の属性，*Si* をSNPデータの属性，*Ci* を属性 *i* を属性値として持つという条件
  - 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

$$\text{Confidence}\quad (Cs_1 \wedge \ldots \wedge Cs_n, C_A) = \frac{\text{Support}(\quad Cs_1 \wedge \ldots \wedge Cs_n \wedge C_A)}{\text{Support}(\quad Cs_1 \wedge \ldots \wedge Cs_n)} \quad (1)$$

$$\text{Confidence}\quad (C_A, Cs_1 \wedge \ldots \wedge Cs_n) = \frac{\text{Support}(\quad Cs_1 \wedge \ldots \wedge Cs_n \wedge C_A)}{\text{Support}(\quad C_A)} \quad (2)$$

(※)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.

知識発見に用いるヒューリスティクス

- SNP について遺伝子データ特有の，4つのヒューリスティクス(A〜D)を選択的に用いて，相関ルール抽出の計算量を削減
  - ヒューリスティクスA - 疾患と特に関連が疑われる遺伝子
  - ヒューリスティクスB - SNPの存在する塩基配列の構成
  - ヒューリスティクスC - 近傍の遺伝子群と臨床情報との相関
  - ヒューリスティクスD - common disease common variant hypothesis

知識発見に用いるヒューリスティクス [1/4]

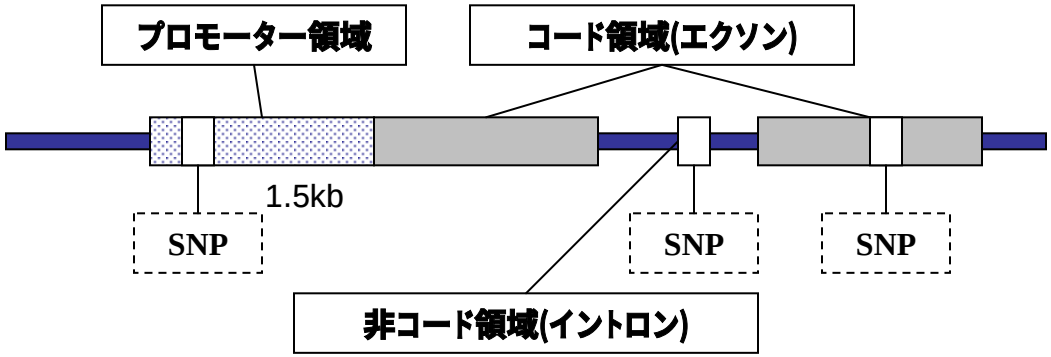
ヒューリスティクス A

- 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
  - その特定の遺伝子上に存在するSNPを対象とすることで，効果的な相関ルール抽出が可能

知識発見に用いるヒューリスティクス [2/4]

ヒューリスティクス B

- SNPの存在する塩基配列の構成
  - プローモーター領域: mRNAへの転写を開始させる領域
  - コード領域 (エクソン) : タンパク質の生成に対応する (コードする) 領域
  - 非コード領域 (イントロン) : 機能的意味のない領域



- プロモータおよびコード領域は, タンパク質の生成に関連する.
- その領域に存在するSNPは, 疾患などの臨床情報との関連が高い.

知識発見に用いるヒューリスティクス [3/4]

ヒューリスティクス C

- **近傍の遺伝子群と臨床情報との相関**
  - **1 つの遺伝子は 1 つのタンパク質の生成に対応しており, 近傍の遺伝子ほど関連して複数のタンパク質の生成に対応している可能性が高い.**
  - **近傍の遺伝子上に存在する SNP の組合せは, 臨床情報との関連が高い.**

知識発見に用いるヒューリスティクス [4/4]

ヒューリスティクス D

- common disease common variant hypothesis
  - **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.**
  - **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.**

本方式の構成

- **ヒューリスティクスを使った知識発見方式を**, Method-1 〜 5 **まで**, **さらに各** Method **につき**Option1, 2**を設定する**.

計算量

大

- Method-1：**全**SNP**が対象**
- Method-2：**プロモーターとコード領域の**SNP**が対象**
  - **全ての** gene **の組み合わせ**
- Method-3：**プロモーターとコード領域の**SNP**が対象**
  - **一定距離までの** gene **の組み合わせ**
- Method-4：**プロモーターとコード領域の**SNP**が対象**
  - **各** gene **内での組み合わせ**

小

- Method-5：**疾患と特に関連が疑われる遺伝子上の**SNP**が対象**

小

大

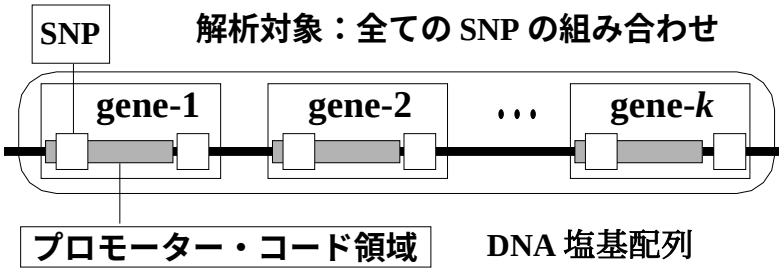
- Option-1:**高頻度のハプロタイプのみが対象**
- Option-2:**すべてのハプロタイプのみが対象**

Method-1

ヒューリスティクスの適用なし
計算量：最大

- **全 SNP の全ての組み合わせを対象とした相関ルール抽出**
  - **全ての gene 間における組み合わせ**
  - **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.**

解析対象とする領域



相関ルールを抽出するための計算量

$$o\left(\prod_{i=1}^{n} a_i \sum_{j=1}^{m} {}_m C_j 3^j\right)$$

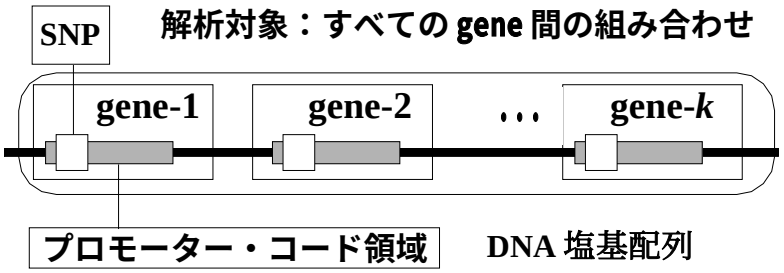
a_i:患者属性 i の属性値の数
n :患者属性の数
m : SNPの数

Method-2

ヒューリスティクス B
計算量：大

- コード領域，およびプロモーター領域をハプロタイプと見なし，その領域内 k 個の gene 内の SNP を対象とした相関ルール抽出
  - 複数の gene をまたがった SNP のハプロタイプも対象とする．
  - 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．

解析対象とする領域



相関ルールを抽出するための計算量

$$o\left(\prod_{i=1}^{n} a_i \sum_{j=1}^{l \cdot k} \left({}_{l \cdot k}C_j \, 3^j\right)\right)$$

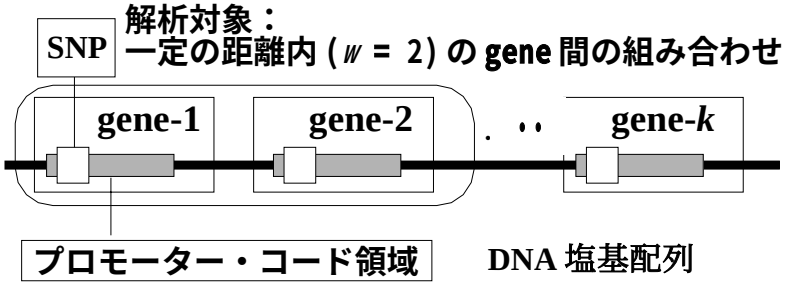
a_i：患者属性 i の属性値の数
n：患者属性の数
k：全 gene 数
l：対象となる gene 内のSNPの平均塩基数

Method-3

ヒューリスティクス B+C

- コード領域，およびプロモーター領域をハプロタイプと見なし，その領域内のk個の gene 内の SNP を対象とした相関ルール抽出
  - 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．
  - 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．

解析対象とする領域



相関ルールを抽出するための計算量

$$o\left(\prod_{i=1}^{n} a_i (k-(w-1)) \sum_{j=1}^{l \cdot w} \left(3^j {}_{l \cdot w}C_j\right)\right)$$

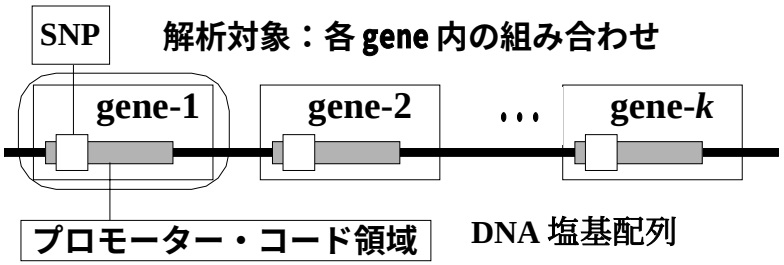
a_i：患者属性 i の属性値の数
n：患者属性の数
k：全 gene 数
w：解析対象とする gene 数
l：対象となる gene 内のSNPの平均塩基数

Method-4

ヒューリスティクス B+C
計算量：小

- コード領域，およびプロモーター領域をハプロタイプと見なし，その領域内 k 個のgene 内の SNP を対象とした相関ルール抽出
  - 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．
  - 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．

解析対象とする領域



相関ルールを抽出するための計算量

$$o\left(\prod_{i=1}^{n} a_i k \sum_{j=1}^{l} \left(3^j {}_l C_j\right)\right)$$

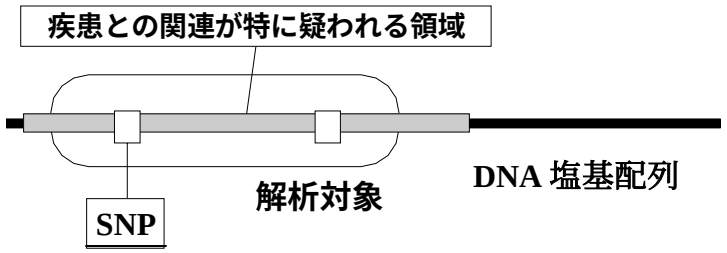
a_i :患者属性 i の属性値の数
n :患者属性の数
k : 全 gene 数
l : 対象となる gene 内のSNPの平均塩基数

Method-5

ヒューリスティクス A
計算量：最小

- **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**
  - **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.**
  - **計算量は少ないが，特定の遺伝子以外に存在する重要なルールを発見できない可能性がある.**

解析対象とする領域



相関ルールを抽出するための計算量

$$o\left(\prod_{i=1}^{n} a_i \sum_{j=1}^{s} {}_{s}C_j 3^j\right)$$

a_i：患者属性 i の属性値の数
n ：患者属性の数
s ：疾患との関連が疑われる遺伝子上のSNPの数

Option1,2

- Option-1：高頻度のハプロタイプのみを分析対象ハプロタイプと設定した場合．
  - 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が前提

ヒューリスティクス D
計算量：小

相関ルールを抽出するための計算量

$$o\left(\prod_{i=1}^{n} a_i d \sum_{j=1}^{x-\#h} \left({}_{x-\#h}C_j 3^j\right) 2^h\right)$$

a_i：患者属性 i の属性値の数
n：患者属性の数
h：分析対象ハプロタイプの数
x：分析対象SNPデータの属性数
d：各Methodでの計算量の式における Σ の係数

- Option-2：全てのハプロタイプを分析対象ハプロタイプと設定した場合．
  - この場合の計算量は，各 Method における計算量と同値

実験方法

- 各 Method, および Option について, 実行速度と精度を検証
  - 実行時間
    - 各 Method, および Option を実行した時の実行時間を計測
  - 精度
    - 各実験について正解ルールを設定
    - 各 Method, および Option について, 再現率と適合率を計算

- 再現率 $= Rb / Rc$
- 適合率 $= Rb / Ra$

Rb：各実験において抽出された正解ルールの数
Rc：各実験において設定した正解ルールの数
Ra：各実験において抽出されたルールの数

- 分析対象データ
  - 公表されている論文を参照して作成したデータ（※）
- 実験環境
  - Solaris 2.6
  - Java 1.3.1
  - PostgreSQL 7.0.2

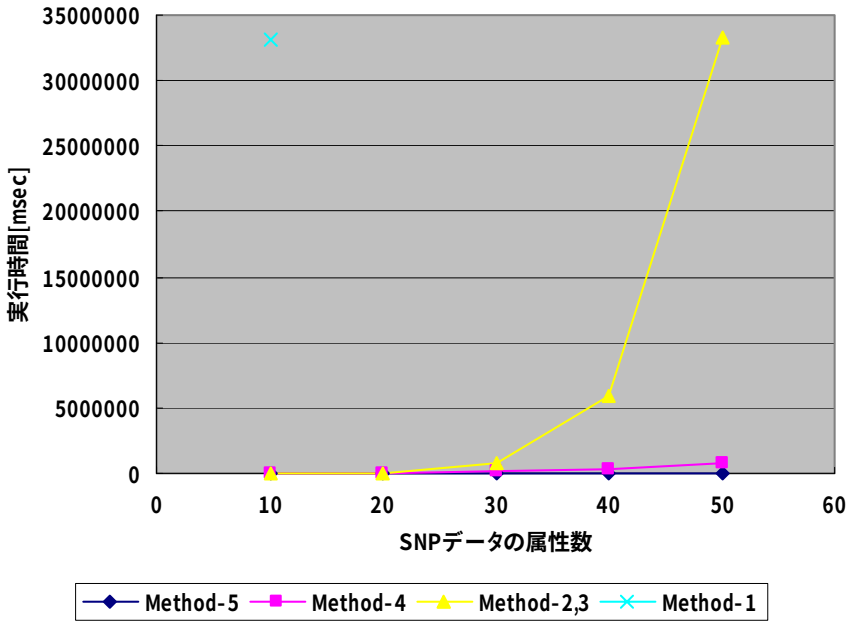
(※) Matthew Stephens, Nicholas J. Smith, and Peter Donnelly:
``A New Statistical Method for Haplotype Reconstruction from Population Data,''
Am. J. Hum. Genet., Vol. 68, pp.978-989, 2001.

実験内容

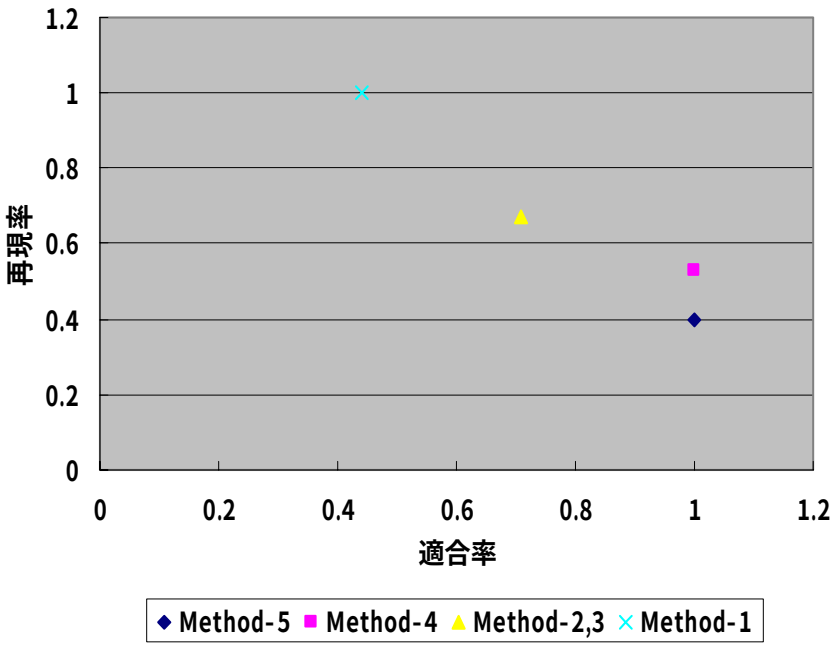
- 実験1～3について，実行時間と精度を検証
- 実験1
  - 各 Method について，gene 数を固定し，各 gene に含まれる SNP 数を変化させる場合の相関ルール抽出
- 実験2
  - 各 Method について，SNP 数を固定し，gene の数を変化させた場合の相関ルール抽出
- 実験3
  - ハプロタイプ頻度を考慮した相関ルール抽出
    - Option-1：頻度の高い特定のハプロタイプのみを分析対象
    - Option-2：全てのハプロタイプを分析対象

実験結果
実験1：各 Method において SNPデータの属性数を変化させた場合

実行時間



再現率・適合率

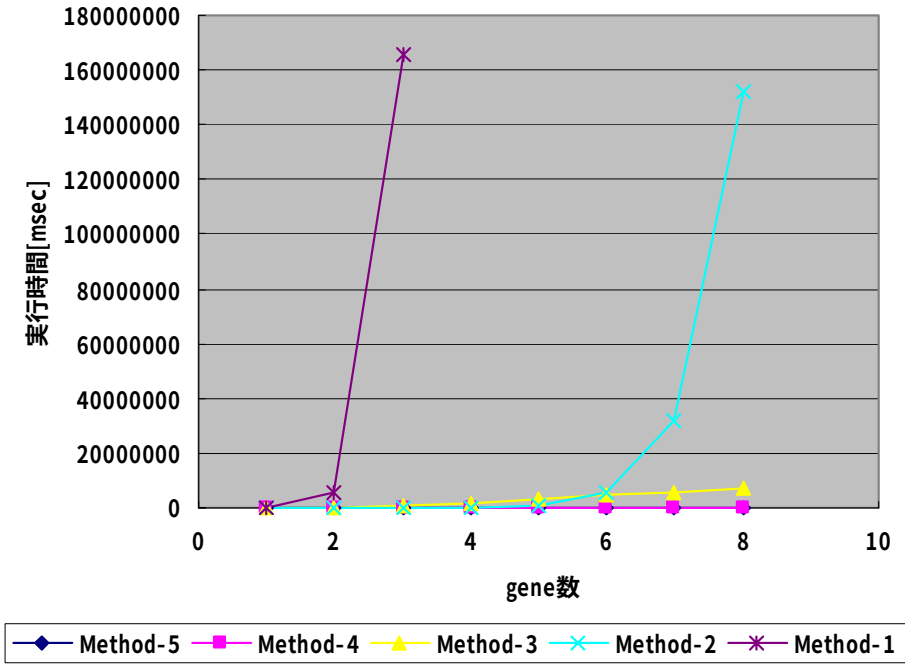


Method-1,2,3：解析対象となるハプロタイプ数が指数関数的に増加するため，その実行時間も急激に増加
Method-4：解析対象とするハプロタイプ数が線形に上昇するため，実行時間もほぼ SNP データの数の増加に比例.
Method-5：解析対象 SNP データの数が一定であるため，実行時間はSNP データの数に依存しない.

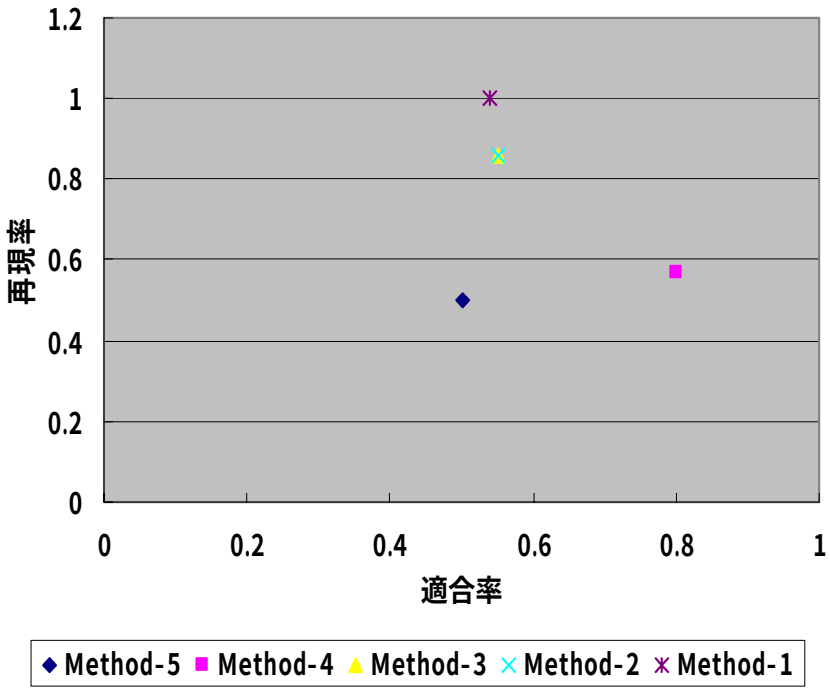
Method-1：再現率は 1.0 であり，すべての正解ルールを抽出できている反面，適合率は低い値を示す.
Method-2,3：Method-1と比べ．抽出されたルールの再現率は低いが適合率は高く，効率的に正解ルールを抽出できている.
Method-4,5：Method-2,3 と比較して，抽出されたルールの再現率は低いが適合率は 1.0 で，効率的に正解ルールを抽出できている.

実験結果
実験2：各 Method において gene 数を変化させた場合

実行時間



再現率・適合率

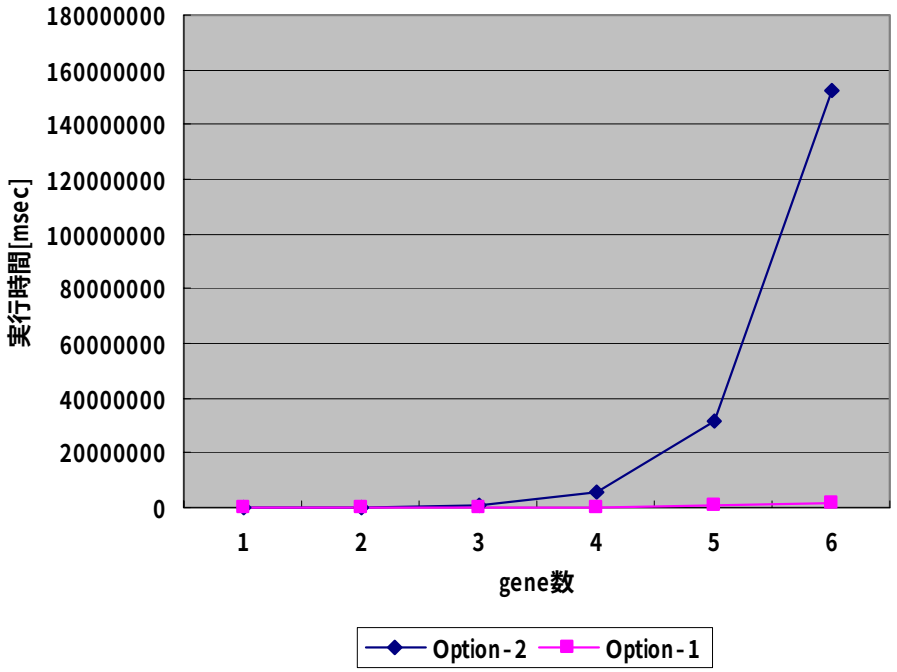


Method-1,2：解析対象となるハプロタイプ数が指数関数的に増加するため，その実行時間も急激に増加．
Method-3,4：解析対象とするハプロタイプ数が線形に上昇するため，実行時間もほぼ SNP データ数の増加に比例する．
Method-5：解析対象 SNP データ数が一定であるため，実行時間はSNP データ数に依存しない．

Method-1：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す．
Method-2,3：再現率は低いが，適合率はわずかに高い値を示す．
Method-4：再現率は低いが，適合率は高く，効率的に正解ルールを抽出できている．
Method-5：抽出されたルールの再現率,適合率は低いが，実行時間は SNP データ数に依存しない．

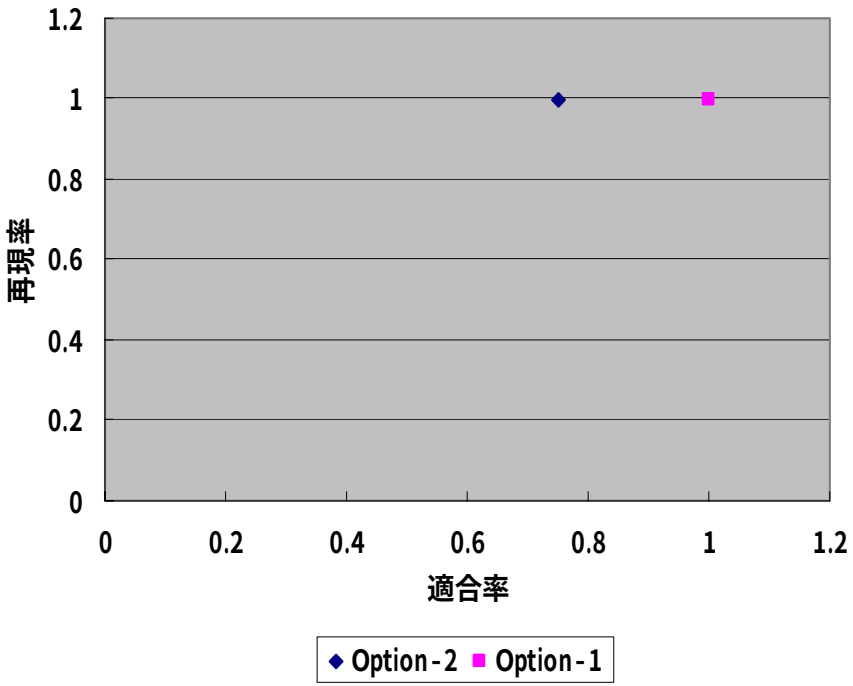
実験結果
実験3：Method-4 における Option-1 および Option-2 について gene 数を変化させた場合

実行時間



Option-2：SNP データ数の増加に伴い急激な実行時間の増加が見られる.
Option-1：SNP データ数に伴って現実的な実行時間の増加に留まる.

再現率・適合率



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.

全体の考察と本方式を使った知識発見の提案

- Method-1 におけるすべての SNP の網羅的な知識発見方法では計算量が膨大にかかり，実行時間という点で解析が困難.
  - Method-2 ～ 5 の選択的にヒューリィスティクスを使った知識発見方法を適用する.
    - 臨床情報および SNP データとの現実的な時間内での知識発見が可能
    - 精度の降下の少ない相関ルール抽出が可能であることを実証した.

- 対象データに応じた知識発見方式の実現
  - 分析対象の SNP データ数および gene 数が大きい場合
    - 実行時間が，対象データ数に依存しない Method-5 ，あるいは比例してほぼ線形に増加する Method-4 を適用することで，効果的な知識発見が可能
  - 各 Method においても，Option-2 に優先して Option-1 を適用することにより，短い実行時間で有効な相関ルール抽出が可能

まとめ

- 本方式は，SNP データ特有のハプロタイプ解析というヒューリスティクスを，選択的に分析方法に適用した知識発見方式
- 実験により，実現可能性および有効性を確認
- 本方法により，対象データ数に応じて，実行時間および精度を任意に設定する知識発見が実現可能

- 今後の課題
  - 実際の医療現場での利用
    - 対象とする医療情報のプライバシーに留意したシステムの構築
    - 抽出されたルールについて，統計による有意性の検定
  - 密(dense)なデータに対する有効なデータマイニング方法の導入
  - 選択的な計算量削減方式の数学的な定式化と分析
    - 臨床情報･遺伝子データに限らず，より一般的な分析対象への適用