

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号
特許第5134397号
(P5134397)

(45) 発行日 平成25年1月30日 (2013. 1. 30)

(24) 登録日 平成24年11月16日 (2012. 11. 16)

(51) Int.Cl.

G O 6 F 19/18 (2011.01)

F I

G O 6 F 19/18

請求項の数 8 (全 33 頁)

(21) 出願番号	特願2008-48748 (P2008-48748)	(73) 特許権者	503359821
(22) 出願日	平成20年2月28日 (2008. 2. 28)		独立行政法人理化学研究所
(65) 公開番号	特開2009-205551 (P2009-205551A)		埼玉県和光市広沢 2 番 1 号
(43) 公開日	平成21年9月10日 (2009. 9. 10)	(74) 代理人	100089118
審査請求日	平成22年11月22日 (2010. 11. 22)		弁理士 酒井 宏明
		(72) 発明者	加藤 護
			神奈川県横浜市鶴見区末広町 1 丁目 7 番 2
			2 号 独立行政法人理化学研究所内 横浜
			研究所内
		審査官	宮久保 博幸

最終頁に続く

(54) 【発明の名称】 ハプロタイプ推定装置、および、プログラム

(57) 【特許請求の範囲】

【請求項 1】

集団における各個体の少なくともコピー数多型を含む実験データからハプロタイプを推定する、制御部と記憶部を少なくとも備えたハプロタイプ推定装置において、

上記記憶部は、

上記個体毎に、上記実験データから得られた、標識によって特定されるマーカー部位に対応付けられた上記コピー数多型のコピー数の総和Nに対する一または複数の条件式を、上記マーカー部位の種類毎に記憶する多型テーブル、

を備え、

上記制御部は、

上記コピー数の総和Nについて、足し合わせた場合に、上記多型テーブルに記憶された上記コピー数の総和Nの上記条件式を満たす任意の2つの整数に分割するコピー数総和分割手段と、

上記コピー数総和分割手段により分割された上記2つの整数をそれぞれ多型識別文字として表し、上記個体において上記多型識別文字を列挙したハプロタイプ文字列の組合せとして格納するハプロタイプ文字列格納手段と、

上記集団において、同一である上記ハプロタイプ文字列の数を集計し、当該ハプロタイプ文字列の上記集団における頻度を求め、当該頻度が所定の条件を満たす上記各個体の上記ハプロタイプ文字列の上記組合せを、上記ハプロタイプの組合せとして推定するハプロタイプ推定手段と、

を備えたことを特徴とするハプロタイプ推定装置。

【請求項 2】

請求項 1 に記載のハプロタイプ推定装置において、

上記条件式は、上記コピー数の総和 N の条件を規定する一または複数の等式または不等式を含むこと、

を特徴とするハプロタイプ推定装置。

【請求項 3】

請求項 1 に記載のハプロタイプ推定装置において、

上記多型テーブルは、

上記個体毎に、塩基多型を更に含む上記実験データから得られた、上記標識によって特定される上記マーカー部位に対応付けられた多型塩基をカウントしたカウント数を、当該多型塩基の種類毎に更に記憶し、

上記ハプロタイプ文字列格納手段は、

上記ハプロタイプ文字列の組合せにおいて、上記多型塩基の種類に対応付けた上記多型識別文字を更に列挙すること、

を特徴とするハプロタイプ推定装置。

【請求項 4】

請求項 1 に記載のハプロタイプ推定装置において、

上記ハプロタイプ推定手段は、

上記ハプロタイプ文字列の頻度を、ハーディ・ワインバークの法則に基づいて算出し、

上記所定の条件を、上記集団におけるハーディ・ワインバーク平衡とすること、

を特徴とするハプロタイプ推定装置。

【請求項 5】

請求項 1 乃至 4 のいずれか一つに記載のハプロタイプ推定装置において、

上記ハプロタイプ文字列格納手段は、

作成した上記ハプロタイプ文字列の組合せにおける上記多型識別文字が表す上記整数の和が、上記多型テーブルにおける上記コピー数の総和 N の上記条件式を満たすか否か確認し、上記条件式を満たさない場合に当該ハプロタイプ文字列を除外し、および／または、作成した上記ハプロタイプ文字列の組合せが他の上記ハプロタイプ文字列の組合せと重複するか否か判断し、重複する場合に当該ハプロタイプ文字列の組合せを除外すること、

を特徴とするハプロタイプ推定装置。

【請求項 6】

請求項 1 乃至 5 のいずれか一つに記載のハプロタイプ推定装置において、

上記ハプロタイプ推定手段は、

EM (Expectation-Maximization) 法を用いて、

上記集団における上記ハプロタイプ文字列の頻度を、当該ハプロタイプ文字列を少なくとも一方に有する上記組合せの頻度により重み付けして算出する M ステップと、

上記組合せの頻度を、当該組合せを構成する上記ハプロタイプ文字列の頻度の積により求め、当該組合せの頻度に基づいて上記重みを算出する E ステップと、

を上記頻度の値が収束するまで交互に繰り返すハプロタイプ頻度算出手段、

を更に備えたことを特徴とするハプロタイプ推定装置。

【請求項 7】

請求項 6 に記載のハプロタイプ推定装置において、

上記 M ステップは、

下記の数式 1 に基づいて、上記ハプロタイプ文字列の頻度を算出し、

10

20

30

40

【数 1】

$$P(h_i) = \frac{\sum_j \sum_k N(c_j) \cdot \delta(h_i, d_{jk}) \cdot w_{jk}}{2n} \dots\dots\dots (数式1)$$

(ここで、 $P(h_i)$ は上記ハプロタイプ文字列の頻度を表し、 h は上記ハプロタイプ文字列を表し、 i は上記ハプロタイプ文字列のインデックスを表す。また、 n は上記集団を構成する上記個体の数、 j は上記多型テーブルにおける上記コピー数の総和 N の上記条件式および／または上記カウント数に関するパターンのインデックス、 k は上記ハプロタイプ文字の組合せのインデックス、 $N(c_j)$ は上記パターン j を持つ上記個体の数を表す。また、 $\delta(h_i, d_{jk})$ は、上記組合せ d_{jk} が一方に当該ハプロタイプ文字列 h_i を有する場合に 1 を返し、両方に当該ハプロタイプ文字列 h_i を有する場合に 2 を返し、当該ハプロタイプ文字列 h_i を持たない場合に 0 を返す関数であり、 d は上記ハプロタイプ文字列の上記組合せを表す。また、 w_{jk} は上記ハプロタイプ文字列の上記組合せの頻度による上記重みである。)

10

上記 E ステップは、

下記の数式 2 に基づいて上記ハプロタイプ文字列の上記組合せの頻度を求め、上記重みとして当該ハプロタイプ文字列の上記組合せの頻度を上記集団における上記組合せの頻度の総和で除して算出すること、

20

【数 2】

$$P(d_{jk}) = P(h_l \oplus h_m) = \begin{cases} P(h_l)P(h_m) & \text{if } l = m \\ 2P(h_l)P(h_m) & \text{if } l \neq m \end{cases} \dots\dots\dots (数式2)$$

(ここで、 $P(d_{jk})$ は、上記ハプロタイプ文字列の上記組合せの頻度を表す。また、 h_l および h_m は当該組合せを構成する 2 つの上記ハプロタイプ文字列を表し、 $P(h_l)$ および $P(h_m)$ は、当該 2 つの上記ハプロタイプ文字列の頻度をそれぞれ表す。)

を特徴とするハプロタイプ推定装置。

【請求項 8】

30

集団における各個体の少なくともコピー数多型を含む実験データからハプロタイプを推定する、制御部と記憶部を少なくとも備えたハプロタイプ推定装置に実行させるためのプログラムであって、

上記記憶部は、

上記個体毎に、上記実験データから得られた、標識によって特定されるマーカー部位に対応付けられた上記コピー数多型のコピー数の総和 N に対する一または複数の条件式を、上記マーカー部位の種類毎に記憶する多型テーブル、

を備えており、

上記制御部において、

上記コピー数の総和 N について、足し合わせた場合に、上記多型テーブルに記憶された上記コピー数の総和 N の上記条件式を満たす任意の 2 つの整数に分割するコピー数総和分割ステップと、

40

上記コピー数総和分割ステップにて分割された上記 2 つの整数をそれぞれ多型識別文字として表し、上記個体において上記多型識別文字を列挙したハプロタイプ文字列の組合せとして格納するハプロタイプ文字列格納ステップと、

上記集団において、同一である上記ハプロタイプ文字列の数を集計し、当該ハプロタイプ文字列の上記集団における頻度を求め、当該頻度が所定の条件を満たす上記各個体の上記ハプロタイプ文字列の上記組合せを、上記ハプロタイプの組合せとして推定するハプロタイプ推定ステップと、

を実行させるためのプログラム。

50

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、ハプロタイプ推定装置、および、プログラムに関する。

【背景技術】

【0002】

生活習慣病のような複雑な病気の原因となる遺伝子の解明や、さらには個別化医療の実現の為に、遺伝子型データなどの実験データから、ヒト等の個体のハプロタイプを推定する必要がある。

【0003】

10

複数座位に渡るハプロタイプを推定する方法には、主に、(1) 1座位当たりの遺伝子型データから複数座位に渡るハプロタイプとその頻度を推定する方法と、(2) 塩基多型を考慮したコピー多型 (Copy Number Polymorphism, あるいは Copy Number Variation、本明細書中で「CNP」と略す場合がある。) に関する実験データから一塩基多型のハプロタイプとその頻度を推定する方法と、が存在する。

【0004】

まず、上述の(1)に示す、1座位当たりの遺伝子型データから複数座位に渡るハプロタイプとその頻度を推定する方法において、「1座位当たりの遺伝子型データ」とは、複数座位に渡る関連が分からない、座位毎に独立した(相(phase)が未知の)遺伝子型データのことである。ここで、図1は、1座位あたりの遺伝子型データの一例を示す図である。図1において、Lは座位(Locus)、Aはアレル(Allele)を表している。

20

【0005】

図1に示すように、1座位当たりの遺伝子型データは、各個体、各座位における、各アレルのカウント数データを有している。ここで、「カウント数データ」とは、個体において、各座位におけるアレルをカウントすることにより得られたカウント数のデータである。例えば、図1においては、個体1における座位(L₁)におけるアレル(A₁)のカウント数は「1」であり、座位(L₃)におけるアレル(A₁)のカウント数は「2」である。

30

【0006】

ここで、上述のように、遺伝子型データは、直接的に相(phase)を特定するものではなく、座位間のアレルの関連は未知であり、図1の例(例えば、個体1)でいえば、カウント数データから、座位L₁と座位L₂の間での相を特定することはできず、座位L₁におけるアレル(A₁/A₂)と座位L₂におけるアレル(A₁/A₃)の関連は未知である。そのため、ハプロタイプを推定する(相を特定する)ための手法が必要となる。

【0007】

ここで、非特許文献1～4に記載のハプロタイプ推定方法は、1座位あたりの遺伝子型データから、複数座位に渡るハプロタイプを推定する。ここで、「複数座位に渡るハプロタイプ」とは、複数座位に渡るアレルの組合せ(相を特定する組合せ)のことである。ここで、図2は、複数座位に渡るハプロタイプの組合せの一例を示す図である。図2において、A(L)は、座位Lに対応するアレルAを表している。

40

【0008】

図2に示すように、例えば、ハプロタイプ1は、座位L₁においてA₁のアレルをもち、座位L₂においてA₁のアレルをもち、座位L₃においてA₁のアレルを持つことが特定されている。このように、従来のハプロタイプの推定方法においては、一般に2種類のアレルを想定し、一塩基多型 (Single Nucleotide Polymorphism、 「SNP」と略す。) などの遺伝子型データから、複数座位に渡るハプロタイプを推定している。また、この方法では、1座位当たりの遺伝子型データさえ分かれば、コピー数多型 (Copy Number Polymorphism、 またはCopy

50

Number Variation、以下「CNP」と略す。）のデータからでもハプロタイプ推定を行うことができる。例えば、1座位あたりの遺伝子型データとして、コピーの数が多いアレル、少ないアレルの2種類を定義し（その数は曖昧ではあるが）、一塩基多型に対してと同じ方法で複数座位にわたるハプロタイプが推定されている（非特許文献5参照）。

【0009】

次に、上述の（2）に示す、塩基多型を考慮したコピー多型に関する実験データから一塩基多型のハプロタイプとその頻度を推定する方法（例えば、本願出願人および発明者による特願2007-237139（未公開であるため先行技術ではない。））に関して、コピー数多型は、ある区間の配列（「コピー単位」と呼ぶ。）が繰り返し現れることがあり、そのコピー数に個体差がある。ここで、「塩基多型を考慮したコピー数多型に関する実験データ」とは、蛍光色素プローブ等によって実験的に識別される異なる塩基の数のデータのことであり、そのデータは生物学的にはコピー単位上にある異なる塩基の数に相当し、例えば図3のように示される。ここで、図3は、塩基多型を考慮した一塩基多型の実験データの一例を示す図である。

【0010】

図3において、Mは、蛍光色素プローブ等の標識で識別される個体間で違いのない配列部位（マーカー部位）、Fは、（異なる蛍光色素等で区別される）個体間で違う塩基（多型塩基）に、それぞれ対応する。換言すれば、図3は、各個体、各マーカー部位における、各多型塩基をカウントしたカウント数の表であり、図1の場合とは異なり、各マーカー部位におけるカウント総数が2以外の数値（0, 1, 2, 3, 4・・・）であることが想定されている。すなわち、多型塩基のカウント数は、個体における2つのハプロタイプ（すなわちディプロタイプ）のコピー数に依存する。ここで「コピー数多型のハプロタイプ」とは、相同染色体の1本上の区画における異なる塩基の組合せで表されたコピー単位の組合せのことであり、例えば、図4のように表現される。ここで、図4は、図3のデータから推定されるハプロタイプの表現例を示す図である。図4において、F（M）は、マーカー部位Mに対応する異なる塩基Fを表し、各コピー単位はMに渡るF（M）の組合せで表され、異なるコピー単位は“,”で区切られている。

【0011】

図4に示すように、ハプロタイプ表現形式は、塩基多型とコピー数多型とを同時に表現され、例えばハプロタイプ1ではコピー数が2であり、それぞれのコピー単位上では $F_1(M_1)F_1(M_2)F_1(M_3)$ と $F_1(M_1)F_2(M_2)F_3(M_3)$ の多型塩基の組合せを有している。

【0012】

【非特許文献1】チアンフア ニウ（Tianhua Niu）著 「アルゴリズムズ フォー インファリング ハプロタイプス（Algorithms for inferring haplotypes）」 Genet Epidemiol.、2004年12月、27巻（4号）334-347頁

【非特許文献2】ジャオフイSキン、チアンフア ニウ、ジュンSリウ（Zhao S. Qin, Tianhua Niu, Jun S. Liu）著 「パーティション-ライゲーション-エクスペクテーション-マキシマイゼーション アルゴリズムズ フォー ハプロタイプ インフェレンス ウィズ シングル-ヌクレオタイド ポリモーフィズムズ（Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms）」 Am J Hum Genet.、2002年11月、71巻（5号）1242-1247頁

【非特許文献3】ローレント エクスコファー、モンゴメリー スラキン（Laurent Excoffier, Montgomery Slatkin）著 「マキシマム-ライクリフッド エスティメーション オブ モレキュラー ハプロタイプ フローク

10

20

30

40

50

エンシーズ イン ア ディプロイド ポピュレーション (Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population)」 Mol Biol Evol.、1995年9月、12巻 (5号)、921-927頁

【非特許文献4】M. E. ホーリー、K. K. キッド (M. E. Hawley, K. K. Kidd) 著 「ハプロ： ア プログラム ユージング ザ EM アルゴリズム トゥー エスティメイト ザ フリークエンシーズ オブ マルチサイト ハプロタイプズ (HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes)」 J Hered.、1995年9-10月、86巻 (5号)、409-411頁

10

【非特許文献5】リチャード レドン、シュンペイ イシカワ、カレンRフィッチ他、全43名 (Richard Redon, Shumpei Ishikawa, Karen R. Fitch, et al.) 著 「グローバル バリエーション イン コピー ナンバー イン ザ ヒューマン ゲノム (Global variation in copy number in the human genome)」 ネチャー (Nature) 出版 2006年11月23日、444巻 (7118号) 444-454頁

【発明の開示】

【発明が解決しようとする課題】

20

【0013】

しかしながら、上述した方法によっては、塩基多型が考慮されていないコピー数多型に関するデータから、コピー数多型のハプロタイプとその頻度を推定することができないという問題があった。

【0014】

例えば、コピー数多型の分析において、効率性を優先する実験プラットフォームによっては、背景技術で示したようなコピー単位上の多型塩基のデータが得られず、一座位当たりの二つの相同染色体に渡る合計のコピー数や曖昧なコピー数しか実験データとして得られない場合があり、上述した方法によってはハプロタイプとその頻度を推定することができないという問題があった。また、コピー数多型に関するデータに1座位当たりの遺伝子型データが加わったデータから、ハプロタイプとその頻度を推定する方法は開発されていないという問題があった。

30

【0015】

本発明は、上記に鑑みてなされたもので、塩基多型が考慮されていないコピー数多型に関する実験データから、曖昧なコピー数の総和しか得られない場合であっても、ハプロタイプとその頻度を高精度で推定することができる、ハプロタイプ推定装置、および、プログラムを提供することを目的とする。また、コピー数多型に関するデータに1座位当たりの遺伝子型データが加わったデータであっても、ハプロタイプとその頻度を推定することができる、ハプロタイプ推定装置、および、プログラムを提供することを目的とする。

【課題を解決するための手段】

40

【0016】

このような目的を達成するため、請求項1に記載のハプロタイプ推定装置は、集団における各個体の少なくともコピー数多型を含む実験データからハプロタイプを推定する、制御部と記憶部を少なくとも備えたハプロタイプ推定装置において、上記記憶部は、上記個体毎に、上記実験データから得られた、標識によって特定されるマーカー部位に対応付けられた上記コピー数多型のコピー数の総和Nに対する一または複数の条件式を、上記マーカー部位の種類毎に記憶する多型テーブル、を備え、上記制御部は、上記コピー数の総和Nについて、足し合わせた場合に、上記多型テーブルに記憶された上記コピー数の総和Nの上記条件式を満たす任意の2つの整数に分割するコピー数総和分割手段と、上記コピー数総和分割手段により分割された上記2つの整数をそれぞれ多型識別文字として表し、上

50

記個体において上記多型識別文字を列挙したハプロタイプ文字列の組合せとして格納するハプロタイプ文字列格納手段と、上記集団において、同一である上記ハプロタイプ文字列の数を集計し、当該ハプロタイプ文字列の上記集団における頻度を求め、当該頻度が所定の条件を満たす上記各個体の上記ハプロタイプ文字列の上記組合せを、上記ハプロタイプの組合せとして推定するハプロタイプ推定手段と、を備えたことを特徴とする。

【0017】

また、請求項2に記載のハプロタイプ推定装置は、請求項1に記載のハプロタイプ推定装置において、上記条件式は、上記コピー数の総和Nの条件を規定する一または複数の等式または不等式を含むこと、を特徴とする。

【0018】

10

また、請求項3のハプロタイプ推定装置は、請求項1に記載のハプロタイプ推定装置において、上記多型テーブルは、上記個体毎に、塩基多型を更に含む上記実験データから得られた、上記標識によって特定される上記マーカー部位に対応付けられた多型塩基をカウントしたカウント数を、当該多型塩基の種類毎に更に記憶し、上記ハプロタイプ文字列格納手段は、上記ハプロタイプ文字列の組合せにおいて、上記多型塩基の種類に対応付けた上記多型識別文字を更に列挙すること、を特徴とする。

【0019】

また、請求項4のハプロタイプ推定装置は、請求項1に記載のハプロタイプ推定装置において、上記ハプロタイプ推定手段は、上記ハプロタイプ文字列の頻度を、ハーディ・ワインバークの法則に基づいて算出し、上記所定の条件を、上記集団におけるハーディ・ワインバーク平衡とすること、を特徴とする。

20

【0020】

また、請求項5に記載のハプロタイプ推定装置は、請求項1乃至4のいずれか一つに記載のハプロタイプ推定装置において、上記ハプロタイプ文字列格納手段は、作成した上記ハプロタイプ文字列の組合せにおける上記多型識別文字が表す上記整数の和が、上記多型テーブルにおける上記コピー数の総和Nの上記条件式を満たすか否か確認し、上記条件式を満たさない場合に当該ハプロタイプ文字列を除外し、および／または、作成した上記ハプロタイプ文字列の組合せが他の上記ハプロタイプ文字列の組合せと重複するか否か判断し、重複する場合に当該ハプロタイプ文字列の組合せを除外すること、を特徴とする。

【0021】

30

また、請求項6に記載のハプロタイプ推定装置は、請求項1乃至5のいずれか一つに記載のハプロタイプ推定装置において、上記ハプロタイプ推定手段は、EM (Expectation-Maximization) 法を用いて、上記集団における上記ハプロタイプ文字列の頻度を、当該ハプロタイプ文字列を少なくとも一方に有する上記組合せの頻度により重み付けして算出するMステップと、上記組合せの頻度を、当該組合せを構成する上記ハプロタイプ文字列の頻度の積により求め、当該組合せの頻度に基づいて上記重みを算出するEステップと、を上記頻度の値が収束するまで交互に繰り返すハプロタイプ頻度算出手段、を更に備えたことを特徴とする。

【0022】

また、請求項7に記載のハプロタイプ推定装置は、請求項6に記載のハプロタイプ推定装置において、上記Mステップは、下記の数式1に基づいて、上記ハプロタイプ文字列の頻度を算出し、上記Eステップは、下記の数式2に基づいて上記ハプロタイプ文字列の上記組合せの頻度を求め、上記重みとして当該ハプロタイプ文字列の上記組合せの頻度を上記集団における上記組合せの頻度の総和で除して算出すること、を特徴とする。

40

【数1】

$$P(h_i) = \frac{\sum_j \sum_k N(c_j) \cdot \delta(h_j, d_{jk}) \cdot w_{jk}}{2n} \dots\dots\dots \text{(数式1)}$$

(ここで、 $P(h_i)$ は上記ハプロタイプ文字列の頻度を表し、 h は上記ハプロタイプ文

50

文字列を表し、 i は上記ハプロタイプ文字列のインデックスを表す。また、 n は上記集団を構成する上記個体の数、 j は上記多型テーブルにおける上記コピー数の総和 N の上記条件式および／または上記カウント数に関するパターンのインデックス、 k は上記ハプロタイプ文字の組合せのインデックス、 $N(c_j)$ は上記パターン j を持つ上記個体の数を表す。また、 $\delta(h_i, d_{jk})$ は、上記組合せ d_{jk} が一方に当該ハプロタイプ文字列 h_i を有する場合に 1 を返し、両方に当該ハプロタイプ文字列 h_i を有する場合に 2 を返し、当該ハプロタイプ文字列 h_i を持たない場合に 0 を返す関数であり、 d は上記ハプロタイプ文字列の上記組合せを表す。また、 w_{jk} は上記ハプロタイプ文字列の上記組合せの頻度による上記重みである。)

【数 2】

10

$$P(d_{jk}) = P(h_l \oplus h_m) = \begin{cases} P(h_l)P(h_m) & \text{if } l = m \\ 2P(h_l)P(h_m) & \text{if } l \neq m \end{cases} \dots\dots\dots \text{(数式2)}$$

(ここで、 $P(d_{jk})$ は、上記ハプロタイプ文字列の上記組合せの頻度を表す。また、 h_l および h_m は当該組合せを構成する 2 つの上記ハプロタイプ文字列を表し、 $P(h_l)$ および $P(h_m)$ は、当該 2 つの上記ハプロタイプ文字列の頻度をそれぞれ表す。)

【0023】

また、請求項 8 に記載のプログラムは、集団における各個体の少なくともコピー数多型を含む実験データからハプロタイプを推定する、制御部と記憶部を少なくとも備えたハプロタイプ推定装置に実行させるためのプログラムであって、上記記憶部は、上記個体毎に、上記実験データから得られた、標識によって特定されるマーカー部位に対応付けられた上記コピー数多型のコピー数の総和 N に対する一または複数の条件式を、上記マーカー部位の種類毎に記憶する多型テーブル、を備えており、上記制御部において、上記コピー数の総和 N について、足し合わせた場合に、上記多型テーブルに記憶された上記コピー数の総和 N の上記条件式を満たす任意の 2 つの整数に分割するコピー数総和分割ステップと、上記コピー数総和分割ステップにて分割された上記 2 つの整数をそれぞれ多型識別文字として表し、上記個体において上記多型識別文字を列挙したハプロタイプ文字列の組合せとして格納するハプロタイプ文字列格納ステップと、上記集団において、同一である上記ハプロタイプ文字列の数を集計し、当該ハプロタイプ文字列の上記集団における頻度を求め、当該頻度が所定の条件を満たす上記各個体の上記ハプロタイプ文字列の上記組合せを、上記ハプロタイプの組合せとして推定するハプロタイプ推定ステップと、を実行させることを特徴とする。

20

30

【発明の効果】

【0024】

本発明によれば、塩基多型が考慮されていないコピー数多型に関する実験データから、曖昧なコピー数の総和しか得られない場合であっても、ハプロタイプとその頻度を高精度で推定することができる。

【0025】

また、本発明によれば、コピー数多型に関するデータに 1 座位当たりの遺伝子型データが加わったデータであっても、ハプロタイプとその頻度を推定することができる。

40

【0026】

また、本発明によれば、文字列による取り得るハプロタイプ文字列の組合せの推定結果が、実験データと一致するか否かを検証し、冗長な文字列を排除するので、精度よく取り得るハプロタイプの組合せを算出することができる。

【0027】

また、本発明によれば、EM法を用いて、高精度にハプロタイプの尤度を検定することができる。

【0028】

また、本発明によれば、個体におけるコピー数多型のコピー数の総和データを用いて、ハプロタイプにおけるコピー数とその頻度を高精度で推定することができる。

50

【発明を実施するための最良の形態】

【0029】

以下に、本発明にかかるハプロタイプ推定装置、ハプロタイプ推定方法、および、プログラム、並びに、記録媒体の実施の形態を図面に基づいて詳細に説明する。なお、この実施の形態によりこの発明が限定されるものではない。

【0030】

[本発明の原理]

【0031】

以下、本発明の原理および概要について説明し、その後、本発明の構成および処理等について詳細に説明する。ここで、図5は、1座位当たりの二つの相同染色体に渡る総コピー数（コピー数の総和）の多型テーブルの一例を表す図である。図5において、Lは座位を表す。

10

【0032】

図5に示すように、実験結果によっては、各個体（個体1、個体2、・・・）、各座位（ L_1 、 L_2 、 L_3 、・・・）における、二つの相同染色体に渡る総コピー数の表（テーブル）が得られる場合（コピー数多型がある場合）があり、総コピー数は2に限られない。

【0033】

また、実験によっては、総コピー数がはっきりと確定されず、例えば“0コピーあるいは1コピー”、“2よりは多いコピー数”など、コピー数の状態が唯一つに決まっていな

20

【0034】

図6において、“0 or 1”は、0コピーあるいは1コピーで表し、“ > 2 ”は、2よりは多いコピー数を表している。実際に、実験で得られるデータは、このような曖昧なデータしか得られない場合があるので、このようなデータから、相同染色体1本上の1座位におけるコピー数と、そのコピー数をアレルと見た時の頻度、あるいはそのようなアレルの複数座位に渡る組合せとしてのハプロタイプとその頻度を推定する必要があった。

【0035】

30

また、そのような総コピー数のデータに、それとは異なる座位における1座位当たりの遺伝子型データが組み合わさったデータから、ハプロタイプを推定する必要があった。ここで、図7は、1座位当たりの二つの相同染色体に渡る総コピー数のデータに、1座位当たりの遺伝子型データが組み合わさったデータの一例を示す図である。図7に示すように、座位 L_1 および L_4 では、1座位当たりの二つの相同染色体に渡る総コピー数を表しており、座位 L_2 、 L_3 、 L_5 では、1座位当たりの遺伝子型データが表現されている。

【0036】

ここで、「ハプロタイプ」とは、二つの相同染色体に渡る総コピー数（曖昧な総コピー数も含む。）が得られた座位に対しては、相同染色体1本上におけるコピー数をアレルとし、遺伝子型が得られた座位に対してはそこで既に定義されたアレルを使った、複数座位に渡るアレルの組合せのことである。ここで、図8は、本発明におけるハプロタイプの表現例を示す図である。図8において、 $N(L)$ は、相同染色体1本上の座位Lにおけるコピー数N（整数値）を表しており、 $A(L)$ は座位Lに対応するアレルAを表している。

40

【0037】

図7において示されるような、二つの相同染色体に渡る総コピー数のデータは、遺伝子型データではないため、従来技術（1）を適用することができないという問題があった。また、その総コピー数データは塩基多型を考慮したコピー数多型に関する実験データとも異なり、さらに今回取り扱うハプロタイプはコピー単位の組合せのことではない為、従来技術（2）も適用することができないという問題があった。

【0038】

50

本発明は、上述した種々の問題に鑑みて本発明者により鋭意検討された結果得られたものであり、一座位当たりの二つの相同染色体に渡る、コピー数の総和（以下、「総コピー数」と呼ぶ。）のデータを処理し、相同染色体1本上の一座位におけるコピー数と、そのコピー数をアレルと見たときの頻度、あるいは、そのようなアレルの複数座位に渡る組合せとしてのハプロタイプとその頻度を推定する。さらには、一座位当たりの二つの相同染色体に渡る総コピー数のデータに、それとは異なる座位における一座位当たりの遺伝子型データが組み合わさったデータから、上述した背景技術のようなハプロタイプとその頻度を推定する。

【0039】

ここでは、本発明の原理を説明するために一例として、総コピー数と遺伝子型が組み合わさったデータからの推定について説明する。なぜなら、総コピー数と遺伝子型が組み合わさったデータからハプロタイプを推定することは、総コピー数データのみからハプロタイプを推定することを含むより一般的な推定であり、かつ、総コピー数データから複数座位に渡るハプロタイプとその頻度を推定することは、一座位におけるコピー数のアレルとその頻度を推定することを含むより一般的な推定であるからである。

【0040】

与えられたデータからハプロタイプとその頻度を推定する原理は、各個体に対しデータと矛盾しないディプロタイプ（2つの相同染色体上それぞれにある2つのハプロタイプの組合せ）を見出し、データを利用してハプロタイプの頻度を計算することである。ここで、二つの相同染色体に渡る（確定された）総コピー数がデータとして得られた場合、「矛盾しない」とは、各座位に対しては、ディプロタイプから数えられるその座位のコピー数アレルのコピー数の和が、そのデータにおける総コピー数に一致するということである。例えば、上述の図7のデータの個体1、座位 L_1 に対しては、図8におけるハプロタイプ1、ハプロタイプ2からなるディプロタイプを考えたとき、 $N_{11}(L_1) + N_{12}(L_1)$ がデータにおける総コピー数2に一致するということである。

【0041】

また、総コピー数が曖昧な総コピー数である場合、「矛盾しない」とは、それが複数の総コピー数が提示されたもの（例えば、不等号を含まず“or”で複数の総コピー数が連結されたもの）であるときは、ディプロタイプから数えられるその座位のコピー数アレルのコピー数の和が、その複数の総コピー数のいずれかに一致するということである。例えば図7のデータの個体1、座位 L_4 に対しては、図8におけるハプロタイプ1、ハプロタイプ2から成るディプロタイプを考えたとき、 $N_{41}(L_4) + N_{42}(L_4)$ が、データ「0 or 1」が提示する複数の総コピー数 $N = 0, 1$ のいずれかに一致するということである。また、曖昧な総コピー数が“ある数より大きい”など不等号で表現される総コピー数であるとき、「矛盾しない」とは、ディプロタイプから数えられるその座位のコピー数アレルのコピー数の和が、不等号で表現される条件を満たすということである。例えば、図7のデータの個体2、座位 L_1 に対しては、図8におけるハプロタイプ1、ハプロタイプ2からなるディプロタイプを考えたとき、 $N_{11}(L_1) + N_{12}(L_1)$ がデータ「 > 2 」によって表現された“2より大きい”を満たす、すなわち $N_{11}(L_1) + N_{12}(L_1) > 2$ を満たす、ということである。

【0042】

また、遺伝子型がデータとして得られた各座位に対しては、上記「矛盾しない」とは、ディプロタイプから数えられるその座位の各アレルの数が、その遺伝子型データにおける各アレルの数（カウント数）と一致するということである。例えば、図7のデータの個体1、座位 L_2 に対しては、図8におけるハプロタイプ1、ハプロタイプ2からなるディプロタイプを考えたとき、ディプロタイプの L_2 における $A_1, A_2, A_3, \dots$ の数それぞれ1, 1, 0, $\dots$ （ディプロタイプ表現の $A_1(L_2), A_2(L_2)$ から得られるカウント数）が、データにおける $A_1, A_2, A_3, \dots$ の数にそれぞれ一致するということである。

【0043】

10

20

30

40

50

以上、まとめると、「矛盾しない」とは、コピー数多型におけるコピー数の総和（2本の染色体に渡るコピー単位の総数）の条件や、塩基多型におけるカウント数（2本の染色体に渡る多型塩基に特異的なマーカー部位の総数）の条件に反しないことである。

【0044】

ここでは、本発明の原理を1個体に対して説明したが、データの各座位における総コピー数及び各アレルの数が同じ個体（すなわち、コピー数データやカウント数データのパターンが同じ個体）は同じように扱えるので、それらの数のパターンでデータをまとめてパターン毎にインデックス（j）を付け、集団における特定のパターンを有する個体数 $N(c_j)$ を保持しておいてもよい。以降、この数のパターンを「カウントパターン」と呼ぶ。（各カウントパターンに対し）与えられたデータと矛盾しないディプロタイプが見出せれば、Expectation-Maximization (EM) 法やGibbsサンプリング法などによって、データからハプロタイプの頻度が計算できる。

10

【0045】

各カウントパターンに対し、データと矛盾しないディプロタイプを見出す方法は種々考えられる。ここではデータと矛盾しない、あらゆる可能なディプロタイプを見出すことを想定して説明する。方法としては、例えば各カウントパターンに対し、まず一座位毎にデータと矛盾しないあらゆる可能な遺伝子型を見出し、それから全座位に渡ってあらゆる可能なディプロタイプを構成する方法が考えられる。ここで、図10は、本発明の概要を模式的に示したフロー図である。

【0046】

20

その第一段階では、二本の相同染色体に渡る（確定された）総コピー数がデータとして得られた座位に対しては、その総コピー数と和が等しい（0を含む正の）整数値2つ一組をあらゆる組合せで作成し、2つの整数値をそれぞれコピー数を表すアレルで、2つ一組を遺伝子型とすればよい。例えば、総コピー数がNのとき、 $[N/0]$ 、 $[N-1/1]$ 、 $[N-2/2]$ 、 $\dots$ 、 $[0/N]$ （ここで、 $[]$ は遺伝子型、あるいはディプロタイプを表し、“/”でアレルまたはハプロタイプを分ける位置を表す。）に対し、重複する冗長な遺伝子型を除いたものが、求める遺伝子型である。総コピー数が曖昧な、“or”で連結された総コピー数である場合は、それぞれの総コピー数に対して、これと同じ方法で遺伝子型を列挙すれば求められる。例えば、0 or 1 or 3 の場合、総コピー数0に対し上と同じ方法で遺伝子型を列挙し（例えば、 $[0/0]$ ）、さらに総コピー数1に対し同じ方法で列挙し（例えば、 $[1/0]$ ）、さらに総コピー数3に対し同じ方法で列挙し（例えば、 $[3/0]$ 、 $[2/1]$ ）、あらゆる組み合わせの遺伝子型（例えば、 $[0/0]$ 、 $[1/0]$ 、 $[3/0]$ 、 $[2/1]$ ）を求める。

30

【0047】

総コピー数が曖昧な、不等号で表現された総コピー数であり、しかも“Nより多いコピー数”を表す“ $>N$ ”である場合は、アレルのコピー数にも“Nより多いコピー数”を表す“ $>N$ ”を用意して $\{0, 1, \dots, N, >N\}$ の各要素をコピー数アレルとし、それら2つのコピー数の和が総コピー数の不等式条件（ $>N$ 、即ちNより大きい）を満たすようなあらゆる2つ一組（同じコピー数アレルからなる組を含む。）を遺伝子型とすればよい（ここで、2つのコピー数の和を出す際、アレルコピー数 $>N$ には、どんな数あるいは $>N$ が足されてもその総コピー数は $>N$ となり、不等式条件は満たされる）。ここで、図9は、2つのコピー数アレルからなる遺伝子型とそのコピー数の和を一例として示す図である。

40

【0048】

例えば、図9のように、 $\{0, 1, \dots, N, >N\}$ から2つのコピー数アレルをあらゆる組合せで作って各組を遺伝子型とし、その2つのアレルのコピー数の和が不等式条件 $>N$ を満たす組だけ選択すればそれが求める遺伝子型である。図9において、（ ）内の数字は2つのアレルのコピー数の和を表す。なお、総コピー数が“Nより少ないコピー数”を表す $<N$ である場合は、0 or 1 or 2 $\dots$ or $N-1$ である場合と同じである。

【0049】

50

遺伝子型がデータとして得られた座位に対しては、データにおける数の分だけアレルを表す文字（多型識別文字）を取り、これを2つに分けて組を作り、遺伝子型とすればよい。遺伝子型データの座位に関しては、一座位当たりのアレルの総数は必ず2であるので簡単にアレルを2つに分けられ、遺伝子型は唯一つ得られる。

【0050】

第二段階においては、各座位の遺伝子型から、全座位に渡るあらゆる可能なディプロタイプを作る。これには、全座位に渡って一つずつ遺伝子型をあらゆる組合せで取り、その各組合せに対し、各座位の各遺伝子型が持つ2つのアレルから、全座位に渡るアレルの組合せとしてのハプロタイプ2つの組合せから構成されるディプロタイプを、 $2^M - 1$ 個（ここでMは全座位数）のあらゆる組合せで作る、このディプロタイプの作成を遺伝子型の組合せ全部に対し行った後、最後に重複する冗長なディプロタイプを除けばよい。

10

【0051】

例えば、座位 L_1 の遺伝子型が $[B_{11}/B_{12}]$ 、 $[B_{13}/B_{14}]$ 、座位 L_2 、 L_3 の遺伝子型がそれぞれ $[B_{21}/B_{22}]$ 、 $[B_{31}/B_{32}]$ であるとき（ここでBは、コピー数アレルあるいは遺伝子型データで定義されているアレルを表す。）、まず全座位に渡る遺伝子型の組合せ、 $[B_{11}/B_{12}]$ 、 $[B_{21}/B_{22}]$ 、 $[B_{31}/B_{32}]$ と $[B_{13}/B_{14}]$ 、 $[B_{21}/B_{22}]$ 、 $[B_{31}/B_{32}]$ を取る。次に、最初の組合せ $[B_{11}/B_{12}]$ 、 $[B_{21}/B_{22}]$ 、 $[B_{31}/B_{32}]$ に対しては、 $[B_{11}B_{21}B_{31}/B_{12}B_{22}B_{32}]$ 、 $[B_{11}B_{21}B_{32}/B_{12}B_{22}B_{31}]$ 、 $[B_{11}B_{22}B_{31}/B_{12}B_{21}B_{32}]$ 、 $[B_{11}B_{22}B_{32}/B_{12}B_{21}B_{31}]$ の $2^3 - 1$ 個のあらゆる組合せを列挙する。さらに、 $[B_{13}/B_{14}]$ 、 $[B_{21}/B_{22}]$ 、 $[B_{31}/B_{32}]$ に対しても、同様に $2^3 - 1$ 個のあらゆる組合せを列挙する。

20

【0052】

最後に重複する冗長なディプロタイプを除く。図10に、あるカウントパターン c_1 に対し、データと矛盾しないディプロタイプを見出す方法の例を示した。その図において、 L_1 、 L_3 は二つの相同染色体に渡る総コピー数がデータとして得られた座位であり、 L_2 は遺伝子型がデータとして得られた座位である。図10<SA-2>の記号（○の中に×）は、全座位に渡って一つずつ遺伝子型をあらゆる組合せで取することを示す。

【0053】

30

こうして得られたディプロタイプを用い、さらに与えられたデータを使って、ハプロタイプの頻度を求める。この方法も幾つか考えられるが、一例として、Expectation-Maximization (EM) 法を用いてもよい。これは、得られたディプロタイプに対し、その存在の重みを割り付け、そのディプロタイプが含むハプロタイプの個数を、その重み分を考慮して数え、ハプロタイプの頻度を計算し（Mステップ）、次にそのハプロタイプ頻度からハーディ・ワインバーグの法則を使って、ディプロタイプの存在の重みを更新し（Eステップ）、さらにその更新された重みから、Mステップ、次にEステップ、さらにMステップ、…と手続きを繰り返して頻度を更新していく方法である。例えば、以下の数式1のようなMステップと、数式2-1の様なEステップを交互に行って、ハプロタイプの頻度の更新していく。

40

【数 3】

<Mステップ>

$$P(h_i) = \frac{\sum_j \sum_k N(c_j) \cdot \delta(h_i, d_{jk}) \cdot w_{jk}}{2n} \quad \dots\dots (数式1)$$

$$\text{ここで } \delta(h_i, d_{jk}) = \begin{cases} 2 & \text{if } d_{jk} \text{ includes two } h_i \\ 1 & \text{if } d_{jk} \text{ includes one } h_i \\ 0 & \text{if } d_{jk} \text{ includes no } h_i \end{cases}$$

10

【数 4】

<Eステップ>

$$w_{jk} = \frac{P(d_{jk})}{\sum_k P(d_{jk})} \quad \dots\dots (数式2-1)$$

【0054】

上記数式2-1において、Pは頻度、nは個体の総数、i, j, kはハプロタイプ、カウントパターン、カウントパターン内でのディプロタイプのインデックス、h, dはハプロタイプ、ディプロタイプ、N(c_j)はカウントパターンc_jを持つ個体の数、wはEM法における、カウントパターン内でのディプロタイプの重みを表す。数2-1におけるディプロタイプの頻度は、ハーディ・ワインバーグの法則から計算する。ハーディ・ワインバーグの法則とは遺伝学における自然法則であり、この法則によって、ディプロタイプを構成する2つのハプロタイプとその確率（あるいは頻度）が分かったとき、そのディプロタイプの確率（あるいは頻度）が計算できる。この法則は、例えば、以下の数2-2の様に表現される。以上のような方法で、ハプロタイプとその頻度を推定する。ここで、頻度の非常に低いハプロタイプは存在しないと解釈してもよい。

20

【数 5】

<ハーディ・ワインバーグの法則>

$$P(d_{jk}) = P(h_l \oplus h_m) = \begin{cases} P(h_l)P(h_m) & \text{if } l = m \\ 2P(h_l)P(h_m) & \text{if } l \neq m \end{cases} \quad \dots\dots (数式2-2)$$

ここで、 $d_{jk} = h_l \oplus h_m$ は、ハプロタイプ文字列の組合せ d_{jk} がハプロタイプ文字列 h_l とハプロタイプ文字列 h_m から構成されることを示す

30

【0055】

[本発明の概要]

まず、本発明は、集団における各個体の少なくともコピー数多型を含む実験データからハプロタイプを推定する、制御部と記憶部を少なくとも備えたハプロタイプ推定装置において実行される。

40

【0056】

そして、図10に示すように、本ハプロタイプ推定装置は、個体毎に、実験データから得られた、標識によって特定されるマーカー部位に対応付けられたコピー数多型の総コピー数Nに対する一または複数の条件式を、マーカー部位の種類毎に記憶する多型テーブルを備える（SA-1）。ここで、コピー数Nに対する条件式は、コピー数の総和Nの条件を規定する一または複数の等式または不等式を含んでもよく、例えば、「N=2」や「N>0」や、「N=1 or 2」や「N≥2」等であってもよい。また、当該多型テーブルは

50

、個体毎に、塩基多型を更に含む実験データから得られた、標識によって特定されるマーカー部位に対応付けられた多型塩基をカウントしたカウント数を、当該多型塩基の種類毎に更に格納されてもよい。

【0057】

そして、本ハプロタイプ推定装置は、コピー数の総和Nについて、足し合わせた場合に、多型テーブルに記憶された総コピー数Nの条件式を満たす任意の2つの整数に分割する(SA-2)。例えば、本ハプロタイプ推定装置は、図5<SA-2>に示すように、総コピー数Nの条件式 $N=2$ の場合に、 $[2/0]$ または $[1/1]$ に分割する。また、例えば、総コピー数Nの条件式 $N \leq 2$ の場合は、 $[2/0]$ 、 $[1/1]$ 、 $[1/0]$ または $[0/0]$ に分割する。

10

【0058】

そして、本ハプロタイプ推定装置は、分割された2つの整数をそれぞれ多型識別文字(例えば、「2」や「3」など)として表し、個体において(複数の座位に渡って)多型識別文字を列挙したハプロタイプ文字列の組合せ(ディプロタイプを表す。)として格納する(SA-3~4)。ここで、図10<SA-3>に示すように、コピー数多型のコピー数を表す多型識別文字は、一例として、アラビア数字により表してもよい。ここで、図10<SA-4>に示すように、多型テーブルに個体毎の標識によって特定されるマーカー部位に対応付けられた多型塩基の種類毎の多型塩基のデータがある場合は、ハプロタイプ文字列の組合せにおいて、多型塩基の種類に対応付けた多型識別文字(例えば、「A₁」や「A₂」など)を更に列挙してもよい(例えば、「2A₁3/0A₂0」)。ここで、本ハプロタイプ推定装置は、まず遺伝子型の組合せを求めてから、次に全ての可能なハプロタイプの組合せ(ディプロタイプ)を求めてもよい。

20

【0059】

すなわち、本ハプロタイプ推定装置は、一例として、図10に示すように、個体毎および座位(マーカー部位)毎の実験データ<SA-1>から、各座位における全ての可能な組合せを求め(SA-2)、次に、複数の座位に渡る全ての可能な遺伝子型の組合せを求め(SA-3)、最後に、遺伝子型の組合せから複数の座位に渡る全ての可能なハプロタイプの組合せ(ディプロタイプ)を表す文字列を求めてもよい(SA-4)。また、ここで、本ハプロタイプ推定装置は、作成したハプロタイプ文字列の組合せにおける多型識別文字が表す整数の和が、多型テーブルにおける総コピー数Nの条件式を満たすか否か(または多型塩基のカウント数と一致するか否か)を確認し、条件を満たさない場合に当該ハプロタイプ文字列を除外してもよい。また、ハプロタイプ推定装置は、作成したハプロタイプ文字列の組合せが他のハプロタイプ文字列の組合せと重複するか否か判断し、重複する場合に当該ハプロタイプ文字列の組合せを除外してもよい(SA-5)。

30

【0060】

そして、本ハプロタイプ推定装置は、集団において、同一であるハプロタイプ文字列の数を集計し、ハプロタイプ文字列の集団における頻度を求め、当該頻度が所定の条件を満たす各個体のハプロタイプ文字列の組合せを、ハプロタイプの組合せとして推定する。ここで、本ハプロタイプ推定装置は、頻度をハーディ・ワインバーグ(Hardy-Weinberg)の法則に基づいて算出し、集団における頻度がハーディ・ワインバーグ平衡となる場合に所定の条件を満たすと判定してもよい。

40

【0061】

また、本ハプロタイプ推定装置は、ハプロタイプの推定処理において、EM(Expectation-Maximization)法を用いて、集団におけるハプロタイプ文字列の頻度を、当該ハプロタイプ文字列を少なくとも一方に有する組合せの頻度により重み付けして算出するMステップと、組合せの頻度を、当該組合せを構成するハプロタイプ文字列の頻度の積により求め、当該組合せの頻度に基づいて重みを算出するEステップと、を頻度の値が収束するまで交互に繰り返すことにより所定の条件を満たす頻度を算出してもよい。EM法の詳細な処理については、後述する。以上で、本発明の概要の説明を終える。

50

【0062】

[ハプロタイプ推定装置の構成]

まず、本ハプロタイプ推定装置の構成について説明する。図11は、本発明が適用される本ハプロタイプ推定装置の構成の一例を示すブロック図であり、該構成のうち本発明に係る部分のみを概念的に示している。

【0063】

図11において、ハプロタイプ推定装置100は、概略的に、ハプロタイプ推定装置100の全体を統括的に制御するCPU等の制御部102、通信回線等に接続されるルータ等の通信装置（図示せず）に接続される通信制御インターフェース部104、入力部112や出力部114に接続される入出力制御インターフェース部108、および、各種のデータベースやテーブルなどを格納する記憶部106を備えて構成されており、これら各部は任意の通信路を介して通信可能に接続されている。

10

【0064】

記憶部106に格納される各種のデータベースやテーブル（多型テーブル106a～実験データファイル106c）は、固定ディスク装置等のストレージ手段であり、各種処理に用いる各種のプログラムやテーブルやファイルやデータベース等を格納する。

【0065】

これら記憶部106の各構成要素のうち、多型テーブル106aは、個体毎に、実験データから得られた、標識によって特定されるマーカー部位に対応付けられたコピー数多型の総コピー数N（二本の相同染色体上のコピー単位の数の総和）に対する一または複数の条件式をマーカー部位の種類毎（すなわちコピー多型の種類毎）にコピー数データとして記憶する多型テーブルである。ここで、多型テーブル106aは、個体毎に、塩基多型を更に含む実験データから得られた、標識によって特定されるマーカー部位に対応付けられた多型塩基をカウントしたカウント数を当該多型塩基の種類毎にカウント数データとして更に記憶してもよい。上述したように、図5や図6や図7や図10<SA-1>は、多型テーブル106aに格納されるコピー数データおよび／またはカウント数データの一例を示す。この多型テーブル106aに格納される情報は、一例として図7等にも示すように、個体毎、座位毎（コピー数多型の種類に対応したマーカー部位毎、および／または、各座位における多型塩基の種類毎）のコピー数（条件式を含んでもよい。）／カウント数を定義している。ここで、コピー数Nに対する条件式は、コピー数の総和Nの条件を規定する一または複数の等式または不等式を含んでもよく、例えば、「 $N=3$ 」や「 $N>0$ 」、「 $N=1 \text{ or } 2$ 」、「 $N \neq 2$ 」、「 $0 < N \leq 2$ 」等であってもよい。

20

30

【0066】

また、ハプロタイプ文字列ファイル106bは、多型テーブル106aに記憶された個体のコピー数データおよび／またはカウント数データに基づいて算出された、取り得るハプロタイプ文字列の組合せを記憶するハプロタイプ文字列記憶手段である。ハプロタイプ文字列ファイル106bは、一例として、ハプロタイプ文字列を、図8に示したようなハプロタイプ表現形式で記憶してもよい。

【0067】

また、実験データファイル106cは、集団における各個体の少なくともコピー数多型を含む実験データを記憶する実験データ記憶手段である。ここで、実験データは、一塩基多型等の塩基多型のデータ（遺伝子型データ等）を含んでもよい。一例として、コピー数多型のコピー単位上の、標識によって特定されるマーカー部位に対応付けられた多型塩基を示す実験データ（DNAチップやPCR等による実験データなど）を記憶してもよい。また、標識としては、蛍光色素プローブの他、蛍光特性を持たない色素や、放射性同位体、GFP・GRPなどのタンパク質、Hisタグ、ビオチン化などによって識別可能なプローブ等を用いてもよい。

40

【0068】

また、図11において、通信制御インターフェース部104は、ハプロタイプ推定装置100とネットワーク300（またはルータ等の通信装置）との間における通信制御を行

50

う。すなわち、通信制御インターフェース部104は、他の端末と通信回線を介してデータを通信する機能を有する。

【0069】

また、図11において、入出力制御インターフェース部108は、入力部112や出力部114の制御を行う。ここで、出力部114としては、モニタ（家庭用テレビを含む。）の他、スピーカ等を用いることができる。また、入力部112としては、キーボード、マウス、およびマイク等を用いることができる。

【0070】

また、図11において、制御部102は、OS（Operating System）等の制御プログラム、各種の処理手順等を規定したプログラム、および所要データを格納するための内部メモリを有し、これらのプログラム等により、種々の処理を実行するための情報処理を行う。制御部102は、機能概念的に、コピー数総和分割部102a、ハプロタイプ文字列格納部102b、ハプロタイプ推定部102c、多型テーブル作成部102eを備えて構成されている。なお、理解の容易のために上述した記号を用いて説明することがある。

10

【0071】

このうち、コピー数総和分割部102aは、コピー数の総和Nについて、足し合わせた場合に、多型テーブル106aに記憶された総コピー数Nの条件式を満たす任意の2つの整数に分割するコピー数総和分割手段である。ここで、コピー数総和分割部102aは、総コピー数Nの条件式が曖昧な数を規定している場合（例えば、「 $N=2$ 」ではなく「 $N>2$ 」など）には、2つの整数の少なくとも一方を曖昧な整数（例えば、 $[>2/0]$ ）に分割してもよい。

20

【0072】

また、ハプロタイプ文字列格納部102bは、コピー数総和分割部102aにより分割された2つの整数をそれぞれ多型識別文字（例えば、アラビア数字の「0」や「1」）として表し、個体において多型識別文字を列挙したハプロタイプ文字列の組合せ（例えば、「 $302/021$ 」）としてハプロタイプ文字列ファイル106bに格納するハプロタイプ文字列格納手段である。ここで、ハプロタイプ文字列格納部102bは、多型テーブル106aに、個体毎に、塩基多型を更に含む実験データ（遺伝子型データ）から得られた、標識によって特定されるマーカー部位に対応付けられた多型塩基が、当該多型塩基の種類毎に格納されている場合（すなわち、カウント数データが格納されている場合）に、ハプロタイプ文字列の組合せにおいて、多型塩基の種類に対応付けた多型識別文字を更に列挙してもよい（例えば、「 $302A_1/021A_2$ 」）。ここで、ハプロタイプ文字列格納部102bは、作成したハプロタイプ文字列の組合せにおける対応する多型識別文字が表す整数の和（二本の相同染色体上のコピー多型のコピー数の和を表している。）が、多型テーブル106aにおける総コピー数Nの条件式を満たすか否か確認し、条件式を満たさない場合に当該ハプロタイプ文字列を除外してもよい。また、ハプロタイプ文字列格納部102bは、作成したハプロタイプ文字列の組合せにおける多型塩基を表す多型識別文字の数が、多型テーブル106aに格納されたカウント数データと一致するか否か確認し、一致しない場合に当該ハプロタイプ文字列を除外してもよい。また、ハプロタイプ文字列格納部102bは、作成したハプロタイプ文字列の組合せが他のハプロタイプ文字列と重複するか否かを判断し、重複する場合に当該ハプロタイプ文字列を除外してもよい。

30

40

【0073】

また、ハプロタイプ推定部102cは、ハプロタイプ文字列ファイル106bを参照して、集団において、同一であるハプロタイプ文字列の数を集計し、当該ハプロタイプ文字列の集団における頻度を求め、当該頻度が所定の条件を満たす各個体のハプロタイプ文字列の組合せを、ハプロタイプの組合せとして推定するハプロタイプ推定手段である。ここで、ハプロタイプ推定部102cは、ハプロタイプ文字列の頻度を、ハーディ・ワインバーグの法則に基づいて算出し、所定の条件を、集団におけるハーディ・ワインバーグ平衡としてもよい。

50

【0074】

ここで、ハプロタイプ推定部102cは、図11に示すように、ハプロタイプ頻度算出部102dを備えて構成される。ハプロタイプ頻度算出部102dは、EM (Expectation-Maximization) 法を用いて、集団におけるハプロタイプ文字列の頻度を、当該ハプロタイプ文字列を少なくとも一方に有する組合せの頻度により重み付けして算出するMステップと、ハプロタイプの組合せの頻度を、当該組合せを構成する2つのハプロタイプ文字列の頻度の積により求め、当該組合せの頻度に基づいて重みを算出するEステップと、を頻度の値が収束するまで交互に繰り返すハプロタイプ頻度算出手段である。ここで、ハプロタイプ頻度算出部102dは、Mステップにおいて算出されたハプロタイプ文字列の頻度と、前回のMステップにおいて算出されたハプロタイプ文字列の頻度と、の対数尤度差を求め、対数尤度差が所定の閾値以下となった場合に、頻度の値が収束したと判定してもよい。これにより、EM法の収束条件を適切に設定することができるので、精度を保証しながら計算時間を節約することができる。

10

【0075】

また、ここで、ハプロタイプ頻度算出部102dは、Mステップにおいて、下記の数式1に基づいて、ハプロタイプ文字列の頻度 $P(h_i)$ を算出し、Eステップにおいて、下記の数式2に基づいてハプロタイプ文字列の組合せの頻度 $P(d_{jk})$ を求め、重み w_{jk} として当該ハプロタイプ文字列の組合せの頻度を集団における組合せの頻度の総和で除して算出してもよい。

【数6】

20

$$P(h_i) = \frac{\sum_j \sum_k N(c_j) \cdot \delta(h_i, d_{jk}) \cdot w_{jk}}{2n} \quad \dots\dots\dots \text{(数式1)}$$

ここで、 $P(h_i)$ はハプロタイプ文字列の頻度を表し、 h はハプロタイプ文字列を表し、 i はハプロタイプ文字列のインデックスを表す。また、 n は集団を構成する個体の数、 j は、多型テーブル106aにおけるコピー数データおよび／またはカウント数データのパターン（以下「カウントパターン」と呼ぶ。）のインデックス、 k はハプロタイプ文字列の組合せのインデックスである。また、 $N(c_j)$ は、カウントパターン j （「 c_j 」と表記する。）を持つ個体の数を表す。また、 $\delta(h_i, d_{jk})$ は、ハプロタイプ文字列の組合せ d_{jk} が一方に当該ハプロタイプ文字列 h_i を有する場合に1を返し、両方に当該ハプロタイプ文字列 h_i を有する場合に2を返し、当該ハプロタイプ文字列 h_i を持たない場合に0を返す関数であり、 d はハプロタイプ文字列の組合せを表す。また、 w_{jk} はハプロタイプ文字列の組合せの頻度による重み（当該ハプロタイプ文字列の組合せ d_{jk} の頻度を、組合せ d_{jk} の頻度の総和で除した数）である。

30

【数7】

$$P(d_{jk}) = P(h_l \oplus h_m) = \begin{cases} P(h_l)P(h_m) & \text{if } l = m \\ 2P(h_l)P(h_m) & \text{if } l \neq m \end{cases} \quad \dots\dots\dots \text{(数式2)}$$

40

ここで、 $P(d_{jk})$ は、ハプロタイプ文字列の組合せの頻度を表す。また、 h_l および h_m は当該組合せを構成する2つのハプロタイプ文字列を表し、 $P(h_l)$ および $P(h_m)$ は、当該2つのハプロタイプ文字列の頻度をそれぞれ表す。

【0076】

また、多型テーブル作成部102eは、実験データファイル106cに記憶された個体毎の実験データを用いて、標識によって特定されるマーカー部位に対応付けられた総コピー数 N （二本の染色体に渡るコピー単位の数の総和）をコピー数多型の種類毎にコピー数データとして、および／または、標識によって特定されるマーカー部位に対応付けられた多型塩基をカウントしたカウント数を多型塩基の種類毎にカウント数データとして、多型テーブル106aに格納する多型テーブル作成手段である。ここで、格納されるカウント

50

数およびコピー数は、一意に特定される数に限られず、曖昧な数（例えば、「 > 5 」、「 $1 \text{ or } 2$ 」、「 $\neq 0$ 」）であってもよい。

【0077】

ここで、本ハプロタイプ推定装置100は、ルータ等の通信装置および専用線等の有線または無線の通信回線を介して、ネットワーク300に通信可能に接続されてもよい。この場合、本システムは、概略的にハプロタイプ推定装置100と、コピー数データやカウント数データ等に関する外部データベースやハプロタイプ推定プログラム等の外部プログラム等を提供する外部システム200とを、ネットワーク300を介して通信可能に接続して構成される。ここで、図11において、ネットワーク300は、ハプロタイプ推定装置100と外部システム200とを相互に接続する機能を有し、例えば、インターネット等である。

10

【0078】

ここで、外部システム200は、ネットワーク300を介して、ハプロタイプ推定装置100と相互に接続され、利用者に対してコピー数データやカウント数データ等に関する外部データベースやハプロタイプ推定プログラム等の外部プログラム等を実行するウェブサイトを提供する機能を有する。ここで、外部システム200は、WEBサーバやASPサーバ等として構成していてもよく、そのハードウェア構成は、一般に市販されるワークステーション、パーソナルコンピュータ等の情報処理装置およびその付属装置により構成していてもよい。また、外部システム200の各機能は、外部システム200のハードウェア構成中のCPU、ディスク装置、メモリ装置、入力装置、出力装置、通信制御装置等およびそれらを制御するプログラム等により実現される。以上で、本ハプロタイプ推定装置100の構成の説明を終える。

20

【0079】

[本ハプロタイプ推定装置100の処理]

次に、このように構成された実施の形態における本ハプロタイプ推定装置100の処理の一例について、以下に図12～図14を参照して詳細に説明する。

【0080】

[ハプロタイプ推定処理]

本実施の形態におけるハプロタイプ推定処理の一例について、以下に図12を参照して説明する。ここで、図12は、本ハプロタイプ推定装置100のハプロタイプ推定処理の一例を示す図である。

30

【0081】

図12に示すように、コピー数総和分割部102aは、コピー数の総和Nについて、足し合わせた場合に、多型テーブルに記憶された総コピー数Nの条件式を満たす任意の2つの整数に分割する（SB-1）。ここで、コピー数総和分割部102aは、総コピー数Nの条件式が曖昧な数を規定している場合（例えば、「 $N > 2$ 」）には、2つの整数を曖昧な整数（例えば、「 $> 2 / 0$ 」）で分割してもよい。

【0082】

そして、ハプロタイプ文字列格納部102bは、コピー数総和分割部102aにより分割された2つの整数をそれぞれ多型識別文字として表し、個体における多型識別文字を複数の座位に渡って列挙したハプロタイプ文字列の組合せを、とりうる全ての組み合わせで算出してハプロタイプ文字列ファイル106bに格納する（SB-2）。なお、多型テーブル106aに、遺伝子型データから得られた、標識によって特定されるマーカー部位に対応付けられた多型塩基のカウント数が、多型塩基の種類毎に格納されている場合には、ハプロタイプ文字列格納部102bは、多型塩基の種類に対応付けた多型識別文字を加えて、ハプロタイプ文字列の組合せに列挙してもよい。ここで、ハプロタイプ文字列格納部102bは、作成したハプロタイプ文字列の組合せにおける対応する多型識別文字が表す整数の和（二本の相同染色体上のコピー多型のコピー数の和を表している。）が、多型テーブル106aにおける総コピー数Nの条件式を満たすか否か確認し、条件式を満たさない場合に当該ハプロタイプ文字列を除外してもよい。また、ハプロタイプ文字列格納部10

40

50

2 b は、作成したハプロタイプ文字列の組合せが他のハプロタイプ文字列と重複するか否か判断し、重複する場合に当該ハプロタイプ文字列を除外してもよい。

【0083】

そして、ハプロタイプ推定部 102 c は、ハプロタイプ文字列格納部 102 b により個体毎に格納されたハプロタイプ文字列の組合せを記憶するハプロタイプ文字列ファイルハプロタイプ文字列ファイル 106 b を参照して、集団において、同一であるハプロタイプ文字列の数を集計する（S B-3）。

【0084】

そして、ハプロタイプ推定部 102 c は、ハプロタイプ文字列の集団における頻度を計算する（S B-4）。ここで、ハプロタイプ推定部 102 c は、ハプロタイプ文字列の頻度をハーディ・ワインバーグの法則により算出してもよい。

10

【0085】

そして、ハプロタイプ推定部 102 c は、ハプロタイプ文字列の頻度が所定の条件を満たす各個体のハプロタイプ文字列の組合せを抽出し、ハプロタイプの組合せとして推定する（S B-5）。ここで、ハプロタイプ推定部 102 c は、所定の条件として、集団におけるハーディ・ワインバーグ平衡を設定してもよい。以上で、本実施の形態におけるハプロタイプ推定処理を終える。

【0086】

[EM法による処理]

ハプロタイプ推定部 102 c の処理による EM 法の詳細な処理の一例について説明する。すなわち、ハプロタイプ推定部 102 c は、ハプロタイプ頻度算出部 102 d の処理により、集団におけるハプロタイプ文字列の頻度を、EM (Expectation-Maximization) 法を用いて、効率よく計算する。

20

【0087】

ここで、EM 法とは、得られたディプロタイプ（本実施の形態においては、ハプロタイプ文字列の組合せとして表す。）に対し、その存在の重みを割り付け、そのディプロタイプが含むハプロタイプの個数を、重み分を考慮して数え、ハプロタイプの頻度を計算し（Mステップ）、次にそのハプロタイプ頻度からハーディ・ワインバーグの法則を使って、ディプロタイプの存在の重みを更新し（Eステップ）、さらにその更新された重みから、Mステップ、次にEステップ、さらにMステップ、・・・と処理を繰り返して、頻度を更新していく方法である。例えば、下記の数式 1 に基づく M ステップと、下記の数式 2（数式 2-1 および数式 2-2）に基づく E ステップを交互に行うことにより、ハプロタイプの頻度の更新していく。

30

【数 8】

<Mステップ>

$$P(h_i) = \frac{\sum_j \sum_k N(c_j) \cdot \delta(h_i, d_{jk}) \cdot w_{jk}}{2n} \quad \dots\dots (数式1)$$

$$\text{ここで } \delta(h_i, d_{jk}) = \begin{cases} 2 & \text{if } d_{jk} \text{ includes two } h_i \\ 1 & \text{if } d_{jk} \text{ includes one } h_i \\ 0 & \text{if } d_{jk} \text{ includes no } h_i \end{cases}$$

40

また、 $P(h_i)$ はハプロタイプ文字列の頻度を表し、 h はハプロタイプ文字列を表し、 i はハプロタイプ文字列のインデックスを表す。また、 n は集団を構成する個体の数、 j はカウントパターン（多型テーブル 106 a におけるコピー数データおよび／またはカウント数データのパターン）のインデックス、 k はハプロタイプ文字の組合せのインデックス、 $N(c_j)$ はカウントパターン (c_j) を持つ個体の数を表す。また、 $\delta(h_i, d_{jk})$ は、ハプロタイプ文字列の組合せ d_{jk} が一方に当該ハプロタイプ文字列 h_i を

50

有する場合に 1 を返し、両方に当該ハプロタイプ文字列 h_i を有する場合に 2 を返し、当該ハプロタイプ文字列 h_i を持たない場合に 0 を返す関数であり、 d はハプロタイプ文字列の組合せを表す。また、 w_{jk} は下記の数式 2-1 に基づくハプロタイプ文字列の組合せの頻度による重みである。

【数 9】

<Eステップ>

$$w_{jk} = \frac{P(d_{jk})}{\sum_k P(d_{jk})} \quad \dots\dots (数式2-1)$$

10

ここで $P(d_{jk})$ は、ハーディ・ワインバーグの法則を表す下記の数式 2-2 に基づいて計算する。

【数 10】

<ハーディ・ワインバーグの法則>

$$P(d_{jk}) = P(h_l \oplus h_m) = \begin{cases} P(h_l)P(h_m) & \text{if } l = m \\ 2P(h_l)P(h_m) & \text{if } l \neq m \end{cases} \quad \dots\dots (数式2-2)$$

ここで、 $d_{jk} = h_l \oplus h_m$ は、ハプロタイプ文字列の組合せ d_{jk} がハプロタイプ文字列 h_l とハプロタイプ文字列 h_m から構成されることを示す

20

【0088】

ここで、上記の数式 2-2 は、ハーディ・ワインバーグの法則を示している。ハーディ・ワインバーグの法則とは遺伝学における自然法則であり、この法則によって、ディプロタイプを構成する 2 つのハプロタイプとその確率（あるいは頻度）が分かった時、そのディプロタイプの確率（あるいは頻度）が計算できる。なお、上記 EM 法においては、頻度の非常に低いハプロタイプは存在しないと解釈される。このように、ハプロタイプ推定部 102c は、ハプロタイプ頻度算出部 102d の処理により、一例として上記の EM 法を用いて、ハプロタイプの組合せとその頻度を推定する。ここで、ハプロタイプ頻度算出部 102d は、下記の数式に基づいて、M ステップにおいて算出されたハプロタイプ文字列の頻度と、前回の M ステップにおいて算出されたハプロタイプ文字列の頻度と、の対数尤度差を求め、対数尤度差が所定の閾値以下となった場合に、頻度の値が収束したと判定してもよい。

30

【数 11】

$$\log \prod_j \left\{ \sum_k P(d_{jk}) \right\}^{N(c_j)}$$

【0089】

以上で、EM 法による処理の一例の説明を終える。

40

【0090】

[実施例]

本実施の形態を、プログラミング言語 Perl で実装した実施例について、以下に図 13～図 14 を参照して説明する。図 13 は、本実装の枠組みを示すフローチャートである。図 14 は、コピー数総和分割処理およびハプロタイプ文字列格納処理の一例を示すフローチャートである。なお、以下の説明において、説明の簡単のために、本実施の形態におけるハプロタイプ文字列を単に「ハプロタイプ」と、ハプロタイプ文字列の組合せを「ディプロタイプ」と述べる場合がある。

【0091】

50

図13に示すように、まず、ハプロタイプ推定装置100は、多型テーブル106aを参照してデータ（例えば、図5、図6、図7等にした多型テーブルのデータ）を読み込む（SC-1）。ここで、ハプロタイプ推定装置は、EM法において収束判定に使われる数値も読み込んでよい。

【0092】

次に、ハプロタイプ推定装置100は、コピー数総和分割処理およびハプロタイプ文字列格納処理として、データと矛盾しないディプロタイプを全て算出する（SC-2）。具体的には、以下のSC-21～SC-25を行う。

【0093】

すなわち、取り得るハプロタイプの組合せを過不足なく求めるため（多型テーブル106aのコピー数データおよび／またはカウント数データと矛盾しないディプロタイプを見出すため）、ハプロタイプ推定装置は、ディプロタイプ（ハプロタイプの組合せ）を構成する。ここで、データと矛盾しないディプロタイプとは、作成したディプロタイプ（2本の相同染色体に渡る文字の集合）において、コピー数多型におけるコピー数の総和（2本の染色体に渡るコピー単位の総数）の条件や、塩基多型におけるカウント数（2本の染色体に渡る多型塩基に特異的なマーカー部位の総数）の条件に反しないことである。具体的には、多型識別文字が表す2本の染色体上のコピー数を足し合わせたとき、多型テーブル106aに記憶されたコピー数データの条件に反しない（各コピー多型の総コピー数Nの条件を満たす）ことである。また、多型識別文字が表す2本の染色体上の多型塩基をカウントしたとき、多型テーブル106aに記憶されたカウント数データのカウント数と一致するということである。ここで、図14は、コピー数データおよび／またはカウント数データの全パターンにおいて、そのカウントパターンと矛盾しないディプロタイプを作成する処理の一例を示すフローチャートである。

【0094】

図14に示すように、まず、ハプロタイプ推定装置100は、多型テーブル106aに記憶されたコピー数データおよび／またはカウント数データから全カウントパターンを算出する（読み出す）（SC-21）。

【0095】

つぎに、ハプロタイプ推定装置100は、カウントパターン（ c_i ）のイテレーション（ i イテレーション）に入る。ここで i イテレーションは、最初 i を1に初期化し、1イテレーション毎に i を1ずつ増加させ、“ $i \leq$ カウントパターンの個数”である限り繰り返すこととする。

【0096】

i イテレーション内において、ハプロタイプ推定装置100は、コピー数総和分割部102aおよびハプロタイプ文字列格納部102bの処理により、各座位に渡って遺伝子型を算出する（SC-22）。コピー数総和分割部102aは、二つの相同染色体に渡る（確定された）総コピー数の座位に対しては、その総コピー数と和が等しい（0を含む正の）整数値2つ一組をあらゆる組合せで作る。ハプロタイプ文字列格納部102bは、その2つの整数値を、アレルを表す文字（多型識別文字）として扱う。多型テーブル106aに格納された総コピー数Nが曖昧な、“or”で連結された総コピー数である場合は、コピー数総和分割部102aは、それぞれの総コピー数に対して、上記と同じ方法でそれぞれ2つ一組を作る。多型テーブル106aに格納された総コピー数が曖昧な、整数Nより少ないコピー数である事を表す“ $<N$ ”である場合は、コピー数総和分割部102aは、“ $<N$ ”を“ $0 \text{ or } 1 \text{ or } 2 \cdots \text{ or } N-1$ ”に変換して、それぞれについて上と同じ方法で2つ一組を作る。多型テーブル106aに格納された総コピー数が曖昧な、Nより多いコピー数である事を表す“ $>N$ ”である場合は、 $\{0, 1, \cdots, N, >N\}$ の各要素をアレルを表す文字（多型識別文字）として、図9に示すように、あらゆる2つ一組を列挙してその2つの和を出し、それが総コピー数の不等式条件（ $>N$ 、即ちNより大きい）を満たす組だけ選択する。ただし和を出す際、アレルコピー数“ $>N$ ”には、どんな数あるいは“ $>N$ ”が足されてもその総コピー数は“ $>N$ ”となり、不等式条件は満たされる。遺

10

20

30

40

50

伝子型の座位に対しては、データにおける数の分だけアレルを表す文字（多型識別文字）を取り、2つ一組を作る（その他、適宜、特願2007-237139号を参照）。

【0097】

そして、ハプロタイプ推定装置100は、各座位に渡って1つずつ遺伝子型を総当たりで取って、全座位に渡る遺伝子型のあらゆる組合せを作る（SC-23）。

【0098】

そして、ハプロタイプ推定装置100は、作成した遺伝子型の組合せに関するイテレーション（ j イテレーション）に入る。ここで、 j イテレーションは、最初 j を1に初期化し、1 イテレーション毎に1ずつ増加させ、“ $j \leq$ 遺伝子の組合せの個数”である限り繰り返すこととする。

10

【0099】

j イテレーション内において、ハプロタイプ文字列格納部102bは、遺伝子型の組合せ j に対し、全座位に渡って各遺伝子型からアレルを1つずつ取って、座位の順に多型識別文字を並べた文字列を作り、その文字列をハプロタイプ文字列とする。このとき、ハプロタイプ文字列格納部102bは、各遺伝子型からアレルを1つずつ取った際残った方のアレルからも座位の順にアレル文字を並べた文字列を作り、その文字列をペア（組合せ）となるもう一つのハプロタイプ文字列として、ディプロタイプ（ハプロタイプ文字列の組合せ）を作る（SC-24）。この、全座位に渡って各遺伝子型からアレルを一つずつ取することは総当たりで行って、あらゆる可能な 2^{M-1} 個（ここで M は全座位数）のディプロタイプを得る。ハプロタイプ推定装置100は、このディプロタイプの作成を遺伝子型の組合せ全部に対して行い、 j イテレーションを終了する。

20

【0100】

そして、ハプロタイプ文字列格納部102bは、得られたハプロタイプ文字列の組合せ（ディプロタイプ）から、重複する冗長なディプロタイプを除外する（SC-25）。すなわち、ハプロタイプ文字列格納部102bは、作成したハプロタイプ文字列の組合せが他のハプロタイプ文字列の組合せと重複するか否か判断し、重複する場合に当該ハプロタイプ文字列の組合せを除外する。ここで、ハプロタイプ文字列格納部102bは、作成したハプロタイプ文字列の組合せにおいて、対応する多型識別文字が表す整数の和が、多型テーブル106aにおけるコピー数の総和 N の条件式を満たすか否か確認し、条件式を満たさない場合に当該ハプロタイプ文字列を除外してもよい。

30

【0101】

ハプロタイプ推定装置100は、上記の処理を各カウントパターンに対して行い、 i イテレーションを終え、最終的に、データと矛盾しない全ディプロタイプを算出し、ハプロタイプ文字列ファイル106bに格納する。以上が、取り得るハプロタイプの組合せを求めるための処理である。

【0102】

再び、図13に戻り、ハプロタイプ推定部102cは、ハプロタイプ頻度算出部102dの処理により、EM法を用いて、ハプロタイプ文字列ファイル106bに記憶されたハプロタイプ文字列の組合せに基づいて、ハプロタイプ文字列の頻度を計算する。

【0103】

40

すなわち、まず、ハプロタイプ頻度算出部102dは、各カウントパターン（ c_i ）に対し、各ハプロタイプの組合せの存在の重みを初期化する（SC-3）。本実施例においては、初期値として重みを平等に割り付けた。すなわち、初期値の重みは、 $w_{j,k} = 1/n_j$ である。ここで w は重み、 j はカウントパターンのインデックス、 k はカウントパターン内でのディプロタイプのインデックスであり、 n_j はカウントパターン内でのディプロタイプの総数である。

【0104】

次にSC-4に移り、ハプロタイプ推定部102cは、EM法のMステップを数式1に基づいて行う。そして、対数尤度を下記の数式に従って計算し、記憶部106に保存する。下記の数式において、記号は上述と同様である。ここで対数尤度とは、計算されたディ

50

プロタイプの頻度がどれくらいデータを説明しているかの指標であり、本実施例においては、これをEM法の収束の判定に用いる。ハプロタイプ推定部102cは、収束判定において、前回のSC-4イテレーションで保存された対数尤度と今回のSC-4イテレーションで計算された対数尤度との差を計算し、その差が一定値以内ならば、もはや対数尤度は改善されないと判定し、SC-5に処理を移す。そうでなければ、SC-4内にとどまり、ハプロタイプ推定部102cは、数式2に従って、EM法のEステップを行う。そしてSC-4のイテレーションを繰り返す。

【数12】

$$\log \prod_j \left\{ \sum_k P(d_{jk}) \right\}^{N(c_j)}$$

10

【0105】

そして、ハプロタイプ推定部102cは、上記条件によってSC-5に処理を移した場合（対数尤度差一定値以内の場合）、ハプロタイプ及びその頻度を結果ファイルに（出力部114などに）出力する（SC-5）。以上で、本実施例の説明を終える。

【0106】

[実証シミュレーション]

本実施の形態がコピー数多型及び一座位当たりの遺伝子型に関するデータから、ハプロタイプとその頻度を推定できるかどうかを確かめる為、シミュレーション実験を行った。実験は四つのタイプからハプロタイプとその頻度を推定できるかどうかシミュレーションを行った。すなわち、（1）一座位における二つの相同染色体に渡る総コピー数のデータから、その座位における相同染色体1本上のコピー数と、そのコピー数をアレルと見た時の頻度を推定できるかどうか、（2）（複数座位に渡る）一座位当たりの二つの相同染色体に渡る総コピー数のデータから、ハプロタイプとその頻度を推定できるかどうか、（3）一座位当たりの二つの相同染色体に渡る総コピー数のデータに、それとは異なる座位において、一座位当たりのSNPの遺伝子型データが組み合わさったデータから、ハプロタイプとその頻度を推定できるかどうか、（4）一座位当たりの二つの相同染色体に渡る総コピー数が曖昧な総コピー数を含む場合のデータに、それとは異なる座位において、一座位当たりのSNPの遺伝子型データが組み合わさったデータから、ハプロタイプとその頻度を推定できるかどうか、についてシミュレーション実験を行った。図15は、シミュレーションの枠組みを示すフローチャートである。

20

30

【0107】

図15に示すように、まず、SD-1において、ハプロタイプ推定装置は、ハプロタイプ（あるいはコピー数アレル）とその確率が書かれたファイルを読み込む。また、ハプロタイプ推定装置は、以下で用いる個体の数も読み込む。曖昧な総コピー数を含むデータを作成する場合は、曖昧でない総コピー数を曖昧な総コピー数に変換する表（例えば総コピー数0と1はいずれも0 or 1に変換する、3以上の総コピー数はどれも>2に変換する、等の対応づけが定義された表）が書かれたファイルも読み込む。

【0108】

40

そして、SD-2において、ハプロタイプ推定装置は、読み込んだハプロタイプから構成しうるあらゆる2つ1組（ハプロタイプ2つから構成される1組）を作成し、1組を1つのディプロタイプとして、ハーディ・ワインバーグの法則（数式2-2参照）に基づいて、読み込んだハプロタイプの確率から全ディプロタイプの確率を計算する。

【0109】

そして、SD-3において、ハプロタイプ推定装置は、ディプロタイプとその確率から、ディプロタイプの多項分布を構成し、与えられた個体の数だけディプロタイプをランダム抽出する。これは、例えば、R言語を使って、`rmultinom(1, size=個体の数, prob=c(ディプロタイプ1の確率, ディプロタイプ2の確率, ディプロタイプ3の確率, ...))`の様なコマンドで簡単に実施できる。抽出されたディプロタイプの

50

1つ1つが、1つ1つの個体に相当する。

【0110】

そして、SD-4において、ハプロタイプ推定装置は、各個体が持つディプロタイプを構成する2つのハプロタイプの各座位に対し、コピー数の座位の時は2つのハプロタイプに渡る総コピー数、SNPの座位の時は各塩基の数を数え、データを作成する。例えば、個体が持つディプロタイプが $[1A_21/1A_12]$ であったら、 L_1 、 L_3 はコピー数の座位であるのでそれぞれ総コピー数は $2(1+1)$ 、 $3(1+2)$ 、 L_2 はSNPの座位であるので、 A_1 、 A_2 の数を数えそれぞれ1、1となる。曖昧な総コピー数を含むデータを作成する場合は、入力で読み込んだ変換表に基づいて総コピー数を曖昧な総コピー数に変換する。

10

【0111】

最後にSD-5において、ハプロタイプ推定装置は、データをファイルに出力する。

【0112】

以上で説明したシミュレーションの枠組みを使って、まず上述の実験(1)の場合について、図16で示されたコピー数アレルとその確率、さらに個体の数500を読み込み、データ図17を作成した。次に、本実施の形態が図17のデータだけから、図16で示されたコピー数アレルとその確率(頻度)を推定(再現)できるか試験した。これは言い換えれば、不完全な観測データからの母集団比率の推定問題である。本実施の形態適用の際、EM法において収束判定に使われる対数尤度差は0.001未満とした。適用の結果、コピー数アレルとその頻度に関し、図18のような結果を得た。図18(再現データ)にあって図16(元データ)にないコピー数アレルの頻度は全て低い。上述のように、これらは存在しないと解釈される。図16にあるコピー数アレルは全て図18に現れており、かつ、それら推定頻度も正解頻度とほぼ等しい。よって、本実施の形態はコピー数アレルとその頻度を推定出来ることが実証された。

20

【0113】

同様に、上述の実験(2)の場合についても、図19で示されたハプロタイプとその確率、さらに個体の数500を読み込み、データ図20を作成した。次に本実施の形態が図20のデータだけから、図19で示されたハプロタイプとその確率を推定できるか試験した。本実施の形態適用の際、EM法において収束判定に使われる対数尤度差は上と同じとした。適用の結果、ハプロタイプとその頻度に関し、図21のような結果を得た。図21(再現データ)にあって図19(元データ)にないハプロタイプの頻度は全て低い。上述のように、これらは存在しないと解釈される。図19にあるハプロタイプは全て図21に現れており、かつ、それら推定頻度も正解頻度とほぼ等しい。よって、本実施の形態はコピー数アレルの組合せとしてのハプロタイプとその頻度を推定出来ることが実証された。

30

【0114】

同様に、上述の実験(3)の場合についても、図22で示されたハプロタイプとその確率、さらに個体の数500を読み込み、データ図23を作成した。次に本実施の形態が図23のデータだけから、図22で示されたハプロタイプとその確率を推定できるか試験した。本実施の形態適用の際、EM法において収束判定に使われる対数尤度差は上と同じとした。適用の結果、ハプロタイプとその頻度に関し、図24のような結果を得た。上述の実験結果と同様、この結果が示す通り、本実施の形態はコピー数アレルとSNPアレルの組合せとしてのハプロタイプとその頻度を推定出来ることが実証された。

40

【0115】

同様に、上述の実験(4)の場合についても、図25で示されたハプロタイプとその確率、さらに個体の数500を読み込み、データ図26を作成した。この際、総コピー数0と1はいずれも曖昧な総コピー数0 or 1に変換し、3以上の総コピー数はどれも曖昧な総コピー数 >2 に変換した。次に、本実施の形態が図26のデータだけから、図25で示されたハプロタイプとその確率を推定できるか試験した。本実施の形態適用の際、EM法において収束判定に使われる対数尤度差は上と同じとした。適用の結果、ハプロタイプとその頻度に関し、図27のような結果を得た。上述の実験結果と同様、この結果が示す通

50

り、曖昧な総コピー数を含むデータの場合においても、本実施の形態はコピー数アレルとSNPアレルの組合せとしてのハプロタイプとその頻度を推定出来ることが実証された。

【0116】

[他の実施の形態]

さて、これまで本発明の実施の形態について説明したが、本発明は、上述した実施の形態以外にも、上記特許請求の範囲に記載した技術的思想の範囲内において種々の異なる実施の形態にて実施されてよいものである。

【0117】

特に、上述の実施の形態において、上述した構成・処理以外にも、適宜、特願2007-237139に記載の事項を補ってもよいものである。

10

【0118】

また、ハプロタイプ推定装置100がスタンドアローンの形態で処理を行う場合を一例に説明したが、ハプロタイプ推定装置100とは別筐体で構成されるクライアント端末からの要求に応じて処理を行い、その処理結果を当該クライアント端末に返却するように構成してもよい。

【0119】

また、実施の形態において説明した各処理のうち、自動的に行われるものとして説明した処理の全部または一部を手動的に行うこともでき、あるいは、手動的に行われるものとして説明した処理の全部または一部を公知の方法で自動的に行うこともできる。

20

【0120】

このほか、上記文献中や図面中で示した処理手順、制御手順、具体的名称、各処理の登録データ等のパラメータを含む情報、データベース構成については、特記する場合を除いて任意に変更することができる。

【0121】

また、ハプロタイプ推定装置100に関して、図示の各構成要素は機能概略的なものであり、必ずしも物理的に図示の如く構成されていることを要しない。

【0122】

例えば、ハプロタイプ推定装置100の各装置が備える処理機能、特に制御部102にて行われる各処理機能については、その全部または任意の一部を、CPU(Central Processing Unit)および当該CPUにて解釈実行されるプログラムにて実現することができ、あるいは、ワイヤードロジックによるハードウェアとして実現することも可能である。尚、プログラムは、後述する記録媒体に記録されており、必要に応じてハプロタイプ推定装置100に機械的に読み取られる。すなわち、ROMまたはHDなどの記憶部106などは、OS(Operating System)として協働してCPUに命令を与え、各種処理を行うためのコンピュータプログラムが記録されている。このコンピュータプログラムは、RAMにロードされることによって実行され、CPUと協働して制御部を構成する。

30

【0123】

また、このコンピュータプログラムは、ハプロタイプ推定装置100に対して任意のネットワーク300を介して接続されたアプリケーションプログラムサーバに記憶されていてもよく、必要に応じてその全部または一部をダウンロードすることも可能である。

40

【0124】

また、本発明に係るプログラムを、コンピュータ読み取り可能な記録媒体に格納することもできる。ここで、この「記録媒体」とは、フレキシブルディスク、光磁気ディスク、ROM、EPROM、EEPROM、CD-ROM、MO、DVD等の任意の「可搬用の物理媒体」、あるいは、LAN、WAN、インターネットに代表されるネットワークを介してプログラムを送信する場合の通信回線や搬送波のように、短期にプログラムを保持する「通信媒体」を含むものとする。

【0125】

また、「プログラム」とは、任意の言語や記述方法にて記述されたデータ処理方法であ

50

り、ソースコードやバイナリコード等の形式を問わない。なお、「プログラム」は必ずしも単一的に構成されるものに限られず、複数のモジュールやライブラリとして分散構成されるものや、OS (Operating System) に代表される別個のプログラムと協働してその機能を達成するものをも含む。なお、実施の形態に示した各装置において記録媒体を読み取るための具体的な構成、読み取り手順、あるいは、読み取り後のインストール手順等については、周知の構成や手順を用いることができる。

【0126】

記憶部106に格納される各種のデータベース等（多型テーブル106a～実験データファイル106c）は、RAM、ROM等のメモリ装置、ハードディスク等の固定ディスク装置、フレキシブルディスク、光ディスク等のストレージ手段であり、各種処理やウェブ

10

【0127】

また、ハプロタイプ推定装置100は、既知のパーソナルコンピュータ、ワークステーション等の情報処理装置を接続し、該情報処理装置に本発明の方法を実現させるソフトウェア（プログラム、データ等を含む）を実装することにより実現してもよい。

【0128】

更に、装置の分散・統合の具体的な形態は図示するものに限られず、その全部または一部を、各種の付加等に応じた任意の単位で、機能的または物理的に分散・統合して構成することができる。

20

【産業上の利用可能性】

【0129】

以上詳述に説明したように、本発明によれば、塩基多型が考慮されていないコピー数多型に関する実験データから、曖昧なコピー数の総和しか得られない場合であっても、ハプロタイプとその頻度を高精度で推定することができる、ハプロタイプ推定装置、および、プログラムを提供することが可能である。

【0130】

また、コピー数多型に関するデータに1座位当たりの遺伝子型データが加わったデータであっても、ハプロタイプとその頻度を推定することができる、ハプロタイプ推定装置、および、プログラムを提供することが可能である。

30

【図面の簡単な説明】

【0131】

【図1】1座位あたりの遺伝子型データの一例を示す図である。

【図2】複数座位に渡るハプロタイプの組合せの一例を示す図である。

【図3】塩基多型を考慮した一塩基多型の実験データの一例を示す図である。

【図4】図3のデータから推定されるハプロタイプの表現例を示す図である。

【図5】1座位当たりの二つの相同染色体に渡る総コピー数（コピー数の総和）の多型テーブルの一例を表す図である。

【図6】曖昧なコピー数を含む、1座位当たりの二つの相同染色体に渡る総コピー数のデータの一例を示す図である。

40

【図7】1座位当たりの二つの相同染色体に渡る総コピー数のデータに、1座位当たりの遺伝子型データが組み合わさったデータの一例を示す図である。

【図8】本発明におけるハプロタイプの表現例を示す図である。

【図9】2つのコピー数アレルからなる遺伝子型とそのコピー数の和を一例として示す図である。

【図10】本発明の概要を模式的に示したフロー図である。

【図11】本発明が適用される本ハプロタイプ推定装置の構成の一例を示すブロック図である。

【図12】本ハプロタイプ推定装置100のハプロタイプ推定処理の一例を示す図である。

50

【図 1 3】本実装の枠組みを示すフローチャートである。

【図 1 4】コピー数総和分割処理およびハプロタイプ文字列格納処理の一例を示すフローチャートである。

【図 1 5】シミュレーションの枠組みを示すフローチャートである。

【図 1 6】実験（１）の場合について、シミュレーション実験に用いたハプロタイプとその頻度のデータ（元データ）の一例を示す図である。

【図 1 7】実験（１）の場合について、読込まれたハプロタイプとその確率についての、作成された個体の数 5 0 0 のコピー数データを示す図である。

【図 1 8】実験（１）の場合について、本実施の形態を適用したシミュレーション結果（再現データ）を示す図である。

10

【図 1 9】実験（２）の場合について、シミュレーション実験に用いたハプロタイプとその頻度のデータ（元データ）の一例を示す図である。

【図 2 0】実験（２）の場合について、読込まれたハプロタイプとその確率についての、作成された個体の数 5 0 0 のコピー数データを示す図である。

【図 2 1】実験（２）の場合について、本実施の形態を適用したシミュレーション結果（再現データ）を示す図である。

【図 2 2】実験（３）の場合について、シミュレーション実験に用いたハプロタイプとその頻度のデータ（元データ）の一例を示す図である。

【図 2 3】実験（３）の場合について、読込まれたハプロタイプとその確率についての、作成された個体の数 5 0 0 のコピー数データおよびカウント数データを示す図である。

20

【図 2 4】実験（３）の場合について、本実施の形態を適用したシミュレーション結果（再現データ）を示す図である。

【図 2 5】実験（４）の場合について、シミュレーション実験に用いたハプロタイプとその頻度のデータ（元データ）の一例を示す図である。

【図 2 6】実験（４）の場合について、読込まれたハプロタイプとその確率についての、作成された個体の数 5 0 0 のコピー数データおよびカウント数データを示す図である。

【図 2 7】実験（４）の場合について、本実施の形態を適用したシミュレーション結果（再現データ）を示す図である。

【符号の説明】

【0 1 3 2】

30

1 0 0 ハプロタイプ推定装置

1 0 2 制御部

1 0 2 a コピー数総和分割部

1 0 2 b ハプロタイプ文字列格納部

1 0 2 c ハプロタイプ推定部

1 0 2 d ハプロタイプ頻度算出部

1 0 2 e 多型テーブル作成部

1 0 4 通信制御インターフェース部

1 0 6 記憶部

1 0 6 a 多型テーブル

1 0 6 b ハプロタイプ文字列ファイル

1 0 6 c 実験データファイル

1 0 8 入出力制御インターフェース部

1 1 2 入力部

1 1 4 出力部

2 0 0 外部システム

3 0 0 ネットワーク

40

【図 1】

	L ₁				L ₂				L ₃			
	A ₁	A ₂	A ₃	A...	A ₁	A ₂	A ₃	A...	A ₁	A ₂	A ₃	A...
個体1	1	1	0	...	1	0	1	...	2	0	0	...
個体2	0	2	0	...	0	1	1	...	2	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...	...

【図 3】

	M ₁				M ₂				M ₃				...
	F ₁	F ₂	F ₃	F...	F ₁	F ₂	F ₃	F...	F ₁	F ₂	F ₃	F...	...
個体 1	2	1	0	...	1	1	1	...	2	0	1	...	...
個体 2	0	0	0	...	0	3	0	...	2	2	0	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...

【図 2】

ハプロタイプ1	A ₁ (L ₁) A ₁ (L ₂) A ₁ (L ₃)
ハプロタイプ2	A ₂ (L ₁) A ₃ (L ₂) A ₁ (L ₃)
...	...

【図 4】

ハプロタイプ1	F ₁ (M ₁) F ₁ (M ₂) F ₁ (M ₃) F ₁ (M ₄) F ₂ (M ₂) F ₃ (M ₃)
ハプロタイプ2	F ₂ (M ₁) F ₂ (M ₂) F ₁ (M ₃) F ₃ (M ₁) F ₂ (M ₂) F ₁ (M ₃) F ₃ (M ₁) F ₂ (M ₂) F ₂ (M ₃)
ハプロタイプ3	F ₃ (M ₁) F ₃ (M ₂) F ₃ (M ₃)
...	...

【図 5】

	L ₁	L ₂	L ₃	...
個体 1	2	0	4	...
個体 2	3	1	5	...
...	...	...	...	...

【図 7】

	L ₁				L ₂				L ₃				L ₄				L ₅				...
	-	A ₁	A ₂	A ₃	A...	A ₁	A ₂	A ₃	A...	-	A ₁	A ₂	A ₃	A...	-	A ₁	A ₂	A ₃	A...	...	
個体 1	2	1	1	0	...	1	0	1	...	0or1	2	0	0	...	...	...	...	...	...	...	
個体 2	>2	0	2	0	...	0	1	1	...	0or1	2	0	0	...	...	...	...	...	...	...	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	

【図 6】

	L ₁	L ₂	L ₃	...
個体 1	2	0or1	>2	...
個体 2	>2	0or1	>2	...
...	...	...	...	...

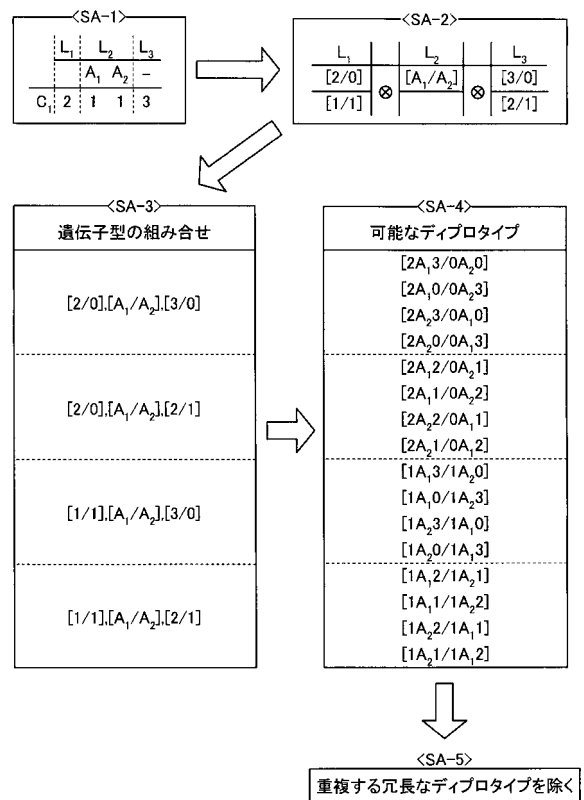
【図 8】

ハプロタイプ1	N ₁₁ (L ₁) A ₁ (L ₂) A ₁ (L ₃) N ₄₁ (L ₄) A ₁ (L ₅)
ハプロタイプ2	N ₁₂ (L ₁) A ₂ (L ₂) A ₃ (L ₃) N ₄₂ (L ₄) A ₁ (L ₅)
...	...

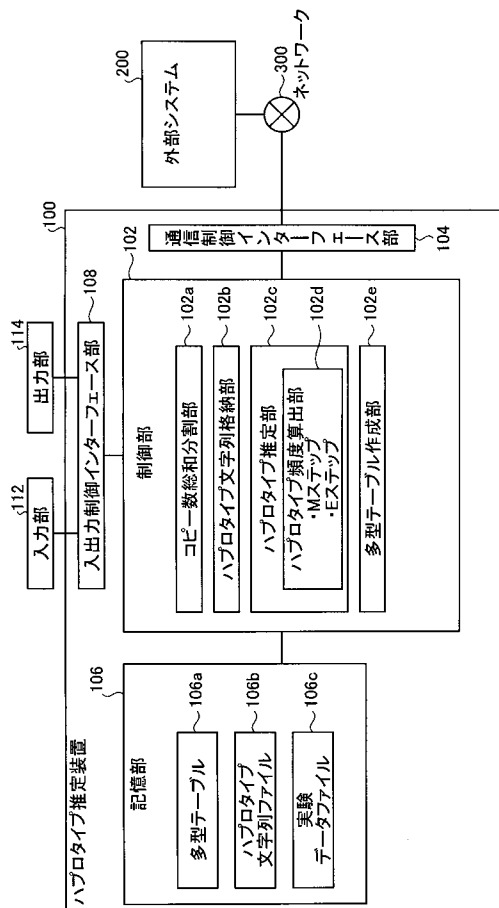
【図 9】

		コピー数アレル2				
		0	1	...	N	>N
コピー数アレル1	0	[0/0] (0)	[0/1] (1)	...	[0/N] (N)	[0/>N] (>N)
	1	-	[1/1] (2)	...	[1/N] (N+1)	[1/>N] (>N)
	...	-	-	...	...	...
	N	-	-	-	[N/N] (N+N)	[N/>N] (>N)
	>N	-	-	-	-	[>N/>N] (>N)

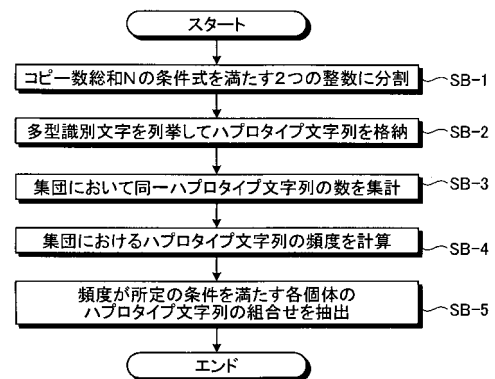
【図 10】



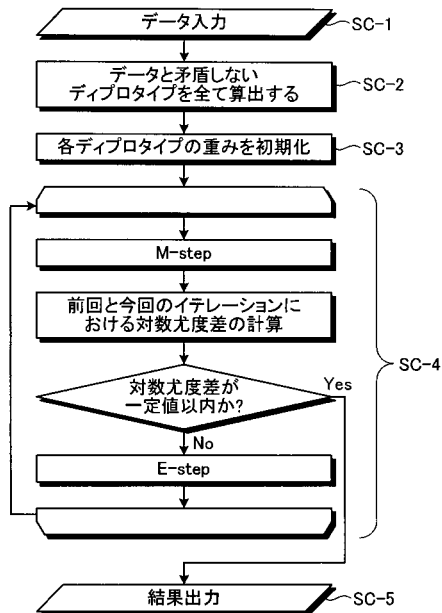
【図 11】



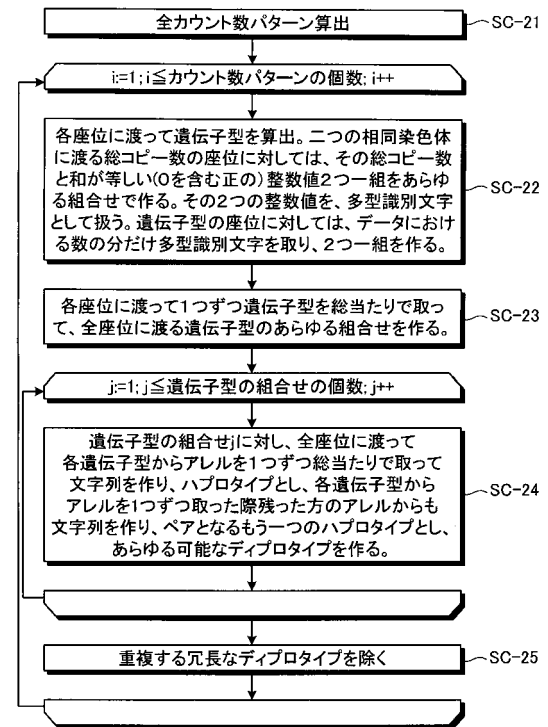
【図 12】



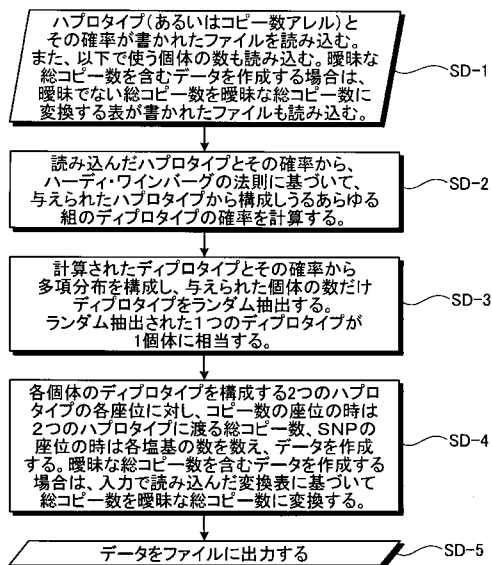
【図 13】



【図 14】



【図 15】



【図 16】

コピー数アレル	確率
0	0.25
1	0.25
2	0.25
3	0.25

【図 17】

個体数	L_1
25	0
53	1
113	2
135	3
80	4
55	5
39	6

【図 18】

コピー数アレル	頻度
0	0.243963423
1	0.252798576
2	0.249512664
3	0.253725252
4	8.51E-08
5	1.42E-15
6	1.61E-29

【図 19】

ハプロタイプ	確率
001	0.2
111	0.2
011	0.2
122	0.2
012	0.2

【図 22】

ハプロタイプ	確率
0A ₂ 2	0.2
1A ₁ 1	0.2
1A ₂ 2	0.2
0A ₂ 1	0.2
2A ₁ 3	0.2

【図 23】

	L ₁	L ₂		L ₃
個体数	-	A ₁	A ₂	-
20	0	0	2	2
42	0	0	2	3
20	0	0	2	4
34	1	1	1	2
88	1	1	1	3
43	1	1	1	4
36	2	1	1	4
39	2	1	1	5
24	2	2	0	2
34	2	2	0	3
23	2	2	0	4
41	3	2	0	4
36	3	2	0	5
20	4	2	0	6

【図 20】

	L ₁	L ₂	L ₃
個体数	-	-	-
29	0	0	2
45	0	1	2
21	0	1	3
13	0	2	2
47	0	2	3
22	0	2	4
39	1	1	2
40	1	2	2
85	1	2	3
43	1	3	3
39	1	3	4
22	2	2	2
39	2	3	3
16	2	4	4

【図 21】

ハプロタイプ	頻度
001	0.206061069
111	0.203938941
011	0.200999988
122	0.196061057
012	0.192938932
022	1.02E-08
121	1.42E-09
023	9.33E-10
000	7.11E-10
221	1.46E-10

頻度 1.00E-10 以下は省略

【図 24】

ハプロタイプ	頻度
0A ₂ 2	0.209929204
1A ₁ 1	0.202929196
1A ₁ 2	0.2010708
0A ₂ 1	0.194070796
2A ₁ 3	0.191999994
2A ₁ 2	5.63E-09
1A ₁ 0	1.94E-09
1A ₁ 3	1.72E-09
0A ₂ 0	5.91E-10
2A ₁ 4	5.07E-10
0A ₂ 3	2.17E-10

頻度 1.00E-10 以下は省略

【図 25】

ハプロタイプ	確率
1A ₁	0.2
0A ₂	0.2
0A ₁	0.2
1A ₂	0.2
2A ₁	0.2

【図 26】

個体数	L ₁	L ₂	
	-	A ₁	A ₂
68	0or1	0	2
103	0or1	1	1
76	0or1	2	0
16	2	0	2
86	2	1	1
51	2	2	0
45	>2	1	1
55	>2	2	0

【図 27】

ハプロタイプ	頻度
1A ₁	0.207615499
0A ₂	0.203562904
0A ₁	0.20340164
1A ₂	0.198436837
2A ₁	0.170847089
>2A ₁	0.016135772
2A ₂	2.59E-07
>2A ₂	3.42E-17

フロントページの続き

- (56)参考文献 特開2004-192018 (JP, A)
国際公開第2008/136210 (WO, A1)
特表2006-519977 (JP, A)
加藤護, コピー数多型に関するハプロタイプ推定のアルゴリズム, 日本人類遺伝学会大会プログラム・抄録集, 2007年 9月12日, 第52回, p. 157
Kato, M., MOCSphaser: a haplotype inference tool from a mixture of copy number variation and single nucleotide polymorphism data, Bioinformatics, 2008年 5月20日, Vol.24, No.14, p.1645-1646
Redon, R., Global variation in copy number in the human genome, Nature, 2006年11月23日, Vol.444, p.444-454
Ogino, S., New insights on the evolution of the SMN1 and SMN2 region: simulation and meta-analysis for allele and haplotype frequency calculations, European Journal of Human Genetics, 2004年, Vol.12, p.1015-1023
Jakobsson, M., Genotype, haplotype and copy-number variation in worldwide human populations, Nature, 2008年 2月21日, Vol.451, p.998-1003

- (58)調査した分野(Int.Cl., DB名)
G06F 19/10
JSTPlus/JMEDPlus/JST7580 (JDreamII)
PubMed