

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2017年7月20日(20.07.2017)



(10) 国際公開番号
WO 2017/122790 A1

- (51) 国際特許分類:
G06F 19/24 (2011.01) **C12N 15/00** (2006.01)
- (21) 国際出願番号: PCT/JP2017/001056
- (22) 国際出願日: 2017年1月13日(13.01.2017)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願 2016-005898 2016年1月15日(15.01.2016) JP
- (71) 出願人: 国立大学法人東北大学(TOHOKU UNIVERSITY) [JP/JP]; 〒9808577 宮城県仙台市青葉区片平二丁目1番1号 Miyagi (JP).
- (72) 発明者: 長崎 正朗(NAGASAKI, Masao); 〒9808577 宮城県仙台市青葉区片平二丁目1番1号 国立大学法人東北大学内 Miyagi (JP). 長谷川 嵩矩(HASEGAWA, Takanori); 〒9808577 宮城県仙台市青葉区片平二丁目1番1号 国立大学法人東北大学内 Miyagi (JP). 小島 要(KOJIMA, Ka-

name); 〒9808577 宮城県仙台市青葉区片平二丁目1番1号 国立大学法人東北大学内 Miyagi (JP). 河合 洋介(KAWAI, Yosuke); 〒9808577 宮城県仙台市青葉区片平二丁目1番1号 国立大学法人東北大学内 Miyagi (JP).

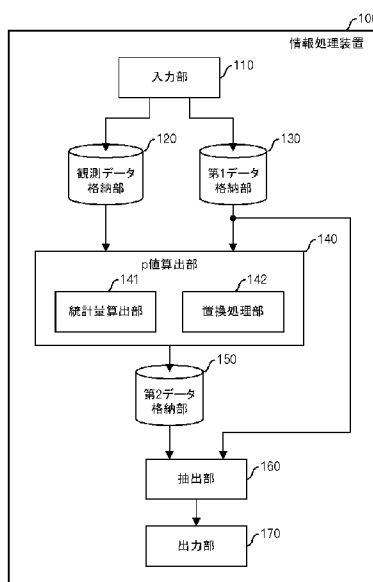
(74) 代理人: 原田 一男(HARADA, Kazuo); 〒1440052 東京都大田区蒲田五丁目49番12号蒲田M&M 1ビル6階 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

[続葉有]

(54) Title: INFORMATION PROCESSING METHOD AND DEVICE

(54) 発明の名称: 情報処理方法及び装置



- 100 Information processing device
110 Input unit
120 Observation data storage unit
130 First data storage unit
140 p-value calculation unit
141 Statistic calculation part
142 Replacement processing part
150 Second data storage unit
160 Extraction unit
170 Output unit

(57) Abstract: The present method includes: a process wherein a statistical amount indicating a genotype-phenotype relationship regarding a single-nucleotide polymorphic set present in a designated part of a genome is calculated on the basis of observation data; a processing execution process wherein processing for calculating a statistical value s to be obtained when the phenotype value in the observation data is replaced is executed until the proportion of times the calculated statistical amount becomes higher than the statistical amount s exceeds a first threshold value or until the proportion of times becomes lower than a second threshold value, or the processing is executed for a predetermined number of times if such events are not detected; and a process wherein the proportion is output as a p-value of the single-nucleotide polymorphic set. At this time, the first threshold value and the second threshold value are dynamically determined, on the basis of an occurrence distribution of cases in which a higher statistical amount than the statistical amount s is calculated and on the basis of predetermined levels of reliability, from values indicating predetermined significant levels.

(57) 要約: 本方法は、ゲノム中の指定部分に存在する一塩基多型の集合についてジェノタイプとフェノタイプとの関係を表す統計量を観測データに基づいて算出する工程と、観測データにおいてフェノタイプの値を入れ替えた場合における統計量 s を算出する処理を、算出した統計量が統計量 s より高くなる回数の割合が第1の閾値を超えるか若しくは第2の閾値を下回るという事象を検出するまで、又は上記事象が検出されなければ予め指定された回数実行するまで、実行する処理実行工程と、上記割合を一塩基多型の集合に対する p 値として出力する工程とを含む。ここで第1の閾値及び第2の閾値は、統計量 s より高い統計量が算出されるケースの出現分布及び予め指定された信頼度に基づき、予め指定された有意水準を表す値から動的に決定されるものである。



(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK,

SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告 (条約第 21 条(3))

明 細 書

発明の名称： 情報処理方法及び装置

技術分野

[0001] 本発明は、Gene-based Testに関する。

背景技術

[0002] GWA S (Genome-Wide Association Studies) 解析は、フェノタイプと変異の関係を調べる方法で、S N P (一塩基多型。Single-Nucleotide Polymorphism) を中心に調べるSNP-based testと、特定遺伝子上に関連するS N P が集合しているという仮定の下でフェノタイプ (形質) とS N P 集合の関連性を検定するGene-based testとがある。

[0003] SNP-based testの場合、特定のフェノタイプとS N P との関連性を検出するためには、そのS N P が集団中で共通であることが要件となる。しかしながら、集団中のレアなS N P に関しては、実際に関連性があったとしても、統計的に有意な水準で関連性を検出することが出来ない。そのため、特定のフェノタイプとレアなS N P との関連性を明確に検出するためには、遺伝子上に存在するレアなS N P を集合として観測することでフェノタイプとの関係を見るGene-based testで行うことがよいとされている。

[0004] しかしながら、既存のGene-based testに用いられる統計手法を用いて正確なp値 (帰無仮説が正しいという条件の下で、実際のデータから計算された統計量よりも大きな統計量が観測される確率) を得るためには、permutation test (並べ替え検定とも呼ぶ) を行うことになる。このpermutation testには、多大な計算コストを要することが分かっている。

[0005] より具体的には、全ゲノム中の遺伝子部分 (エクソン) に存在するレアバリエント (レアなS N P) とフェノタイプの関係を見ようとすると、通常の統計学的な立場からは、そのレアバリエントとフェノタイプとの関係を統計的に処理した場合に、所定の閾値 (例えば0.01や0.05) よりも小さいp値を有するかどうかを判断することになる。

- [0006] この p 値を求める統計的処理を全ゲノムの中に存在する遺伝子部分毎に行うことになり、この遺伝子は全ゲノム中に 20000 個ほど存在している。そのため、多重検定という概念を用いると、遺伝子毎の統計学的処理によりその関係性の確からしさを確認するためには、所定の閾値／遺伝子数を基準とすることになる。そうすると、Gene-based test を行う場合のように permutation test を実行する場合、有意であるかどうかを見極めるために要する試行回数は最低でも $1 / (\text{所定の閾値} / \text{遺伝子数})$ となる。
- [0007] 例えば、所定の閾値 = 0.01、遺伝子数 = 20000 であれば、permutation test をそれぞれの遺伝子に対して少なくとも 200 万回行う必要があり、遺伝子全体ではその 20000 倍である 400 億回行う必要がある。さらに、複数のフェノタイプについて関連性を計算する場合には、その分 permutation test の試行回数は増加することになる。
- [0008] このため、permutation test の試行回数を削減するための技術が幾つか提案されている。
- [0009] 例えば、以下に示す非特許文献 1 では、adaptive permutation test として、最終的に計算される p 値が有意ではないと分かったところでテストを中止することで試行回数を削減する技術が開示されている。
- [0010] さらに、以下に示す非特許文献 2 では、M 回実行する permutation test の途中における p 値が、閾値 α を超える場合にはテストを途中で打ち切ること、また、p 値が、閾値 $\alpha + (\text{p 値の分布の標準偏差 } \sigma_M \text{ の } 6 \text{ 倍})$ を超えるか又は閾値 $\alpha - (\text{p 値の分布の標準偏差 } \sigma_M \text{ の } 6 \text{ 倍})$ を下回る場合にテストを途中で打ち切ることが開示されている。
- [0011] これらの技術でもある程度の高速化がなされるが、特に非特許文献 2 では、閾値 α を中心として「標準偏差 σ_M の 6 倍」の 2 倍という固定的な範囲をテスト継続の条件としているが、このような範囲が適切な範囲であるかどうかは不明である。よって、無駄にテストを継続してしまう若しくは中断してしまう場合もある。

先行技術文献

非特許文献

[0012] 非特許文献1: PLINK/SEQ, PSEQ association analysis [online] [2015年10月1日検索] インターネット<<https://atgu.mgh.harvard.edu/plinkseq/assoc.shtml>>

非特許文献2: SNP&Variation Suite v8.4.1. Manual Collapsing Methods for DNA-Sequence Analysis [Online] [2015年10月16日検索] <http://doc.goldenhelix.com/SVS/latest/svsmanual/collapsing_methods.html>

発明の概要

発明が解決しようとする課題

[0013] 従って、本発明の目的は、一側面によれば、ユーザ指定の信頼度を反映した精度を実現しつつpermutation testを高速化させる技術を提供することである。

課題を解決するための手段

[0014] 本発明に係る方法は、(A) ゲノム中の指定部分に存在する一塩基多型の集合についてジェノタイプとフェノタイプとの関係を表す統計量を観測データに基づいて算出するステップと、(B) 観測データにおいてフェノタイプの値を入れ替えた場合における統計量を算出する処理を、算出した統計量が観測データに基づき算出した統計量より高くなる回数の割合が第1の閾値を超えるか若しくは第2の閾値を下回るという事象を検出するまで、又は上記事象が検出されなければ予め指定された回数実行するまで、実行する処理実行ステップと、(C) 上記割合を一塩基多型の集合に対するp値として出力するステップと、を含む。そして、第1の閾値及び第2の閾値は、観測データに基づき算出した統計量より高い統計量が算出されるケースの出現分布及び予め指定された信頼度に基づき、予め指定された有意水準を表す値から動的に決定されるものである。

図面の簡単な説明

[0015] [図1]図1は、実施の形態に係る情報処理装置の構成例を示す図である。

[図2]図2は、本実施の形態に係る処理フローを示す図である。

[図3]図3は、観測データの一例を示す図である。

[図4]図4は、本実施の形態に係る処理フローを示す図である。

[図5]図5は、置換処理の処理フローを示す図である。

[図6]図6は、コンピュータの機能ブロック図である。

発明を実施するための形態

[0016] 本発明の一実施の形態に係る情報処理装置の一例を図1に示す。

[0017] 本実施の形態に係る情報処理装置100は、入力部110と、観測データ格納部120と、第1データ格納部130と、p値算出部140と、第2データ格納部150と、抽出部160と、出力部170とを有する。

[0018] 入力部110は、ユーザから、各種パラメータ、処理すべき観測データ及びSNPの集合(SNP setと呼ぶ)の指定などを受け付け、第1データ格納部130などに格納する。入力部110は、情報処理装置100がネットワークを介して接続されている他のコンピュータから、それらのデータを受信する場合もある。観測データ格納部120は、複数の被観測者の各々について、フェノタイプのデータと、ゲノムのデータとを対応付けて格納している。これらのデータについては予め格納されている場合もあれば、入力部110から入力される場合もある。

[0019] p値算出部140は、各SNP setについてp値を算出し、算出されたp値をSNP setに対応付けて第2データ格納部150に格納する。このため、p値算出部140は、統計量算出部141と、置換処理部142とを有する。統計量算出部141は、予め定められた統計量を算出する。置換処理部142は、ブートストラップ法などの方法に従って観測データにおける各被観測者のフェノタイプの値を入れ替える処理等を実行する。

[0020] 抽出部160は、第2データ格納部150から、所定の条件を満たすp値が算出されたSNP setを抽出し、出力部170に出力する。出力部170は、ユーザ又は他のコンピュータに対して、抽出部160の抽出結果を出力する。なお、出力部170は、例えば第1データ格納部130に格納され

ているデータに基づき、第2データ格納部150に格納されているデータ等の処理途中のデータを出力するようにしても良い。

[0021] 次に、図2乃至図5を用いて、このような情報処理装置100が行う処理について詳細に説明する。

[0022] 入力部110は、例えばユーザから、各種パラメータの入力を受け付け、第1データ格納部130に格納する（図2：ステップS1）。本実施の形態では、1つのSNP setについて統計量の最大計算回数Bと、第1の信頼度 $\alpha 1$ と、第2の信頼度 $\alpha 2$ と、以下で説明する置換処理において算出する統計量の計算回数Tと、第1の信頼度 $\alpha 1$ に対応する有意水準 p_{sig} とを入力する。なお、 p_{sig} については、多重検定を考慮して設定される。すなわち、 $p_{sig} = (1 - \alpha 1) / \text{SNP set 数}$ となる。また、Bは、少なくとも $1 / p_{sig}$ となる。さらに、第2の信頼度 $\alpha 2$ に対応する有意水準 p_{a2} を入力するようにしても良い。また、第1の信頼度 $\alpha 1$ と第2の信頼度 $\alpha 2$ については同一である場合もある。

[0023] また、入力部110は、例えばユーザから、処理すべき観測データ及びSNP setの指定を受け付け、第1データ格納部130又は観測データ格納部120に格納する（ステップS3）。例えば、観測データ格納部120において、どこからどこまでの観測データを用いるか、例えば遺伝子の位置及びその遺伝子におけるSNPの位置等のデータを指定する。

[0024] 観測データ格納部120に格納される観測データの一例を図3に示す。図3では、被観測者毎に、フェノタイプ（ここではx及びy。例えば身長体重など。）と、ゲノムのデータとが対応付けられて格納されている。例えば、ゲノムには、複数の遺伝子が含まれているが、ここでは遺伝子Hに着目しており、その中で変異がある複数の遺伝子座を処理すべきSNP setとして指定する。図3の例では変異がある2つのSNPが指定される。なお、フェノタイプについても、多数の種類のフェノタイプが観測データに含まれる場合には、いずれのフェノタイプについて処理するのかについて指定する場合もある。

[0025] さらに、遺伝子毎に S N P s e t を指定するのではなく、ある程度の遺伝子のまとまりについて S N P s e t を指定するようにしても良い。また、ここでは、複数の S N P s e t が指定されるものとする。

[0026] そうすると、p 値算出部 140 は、処理すべき S N P s e t のうち、未処理の S N P s e t を 1 つ特定する（ステップ S 5）。例えば、処理すべき S N P s e t のリストを用意して上から順番に特定する。

[0027] また、p 値算出部 140 の統計量算出部 141 は、特定された S N P s e t について、観測データそれ自身に関する所定の統計量 s を算出する（ステップ S 7）。本実施の形態に係る統計量 s は、ゲノム中の指定遺伝子等に存在する一塩基多型の集合に関してジェノタイプとフェノタイプとの関係を表す統計量である。より具体的には、例えば S K A T（Sequential Kernel Association Test）の統計量 s を算出することとする。

[0028] S K A T の統計量 s は、以下のように表される。

$$s = (y - \mu)^T G W G^T (y - \mu) \quad (1)$$

ここで n を被観測者数、m を S N P s e t に含まれる S N P の数として、y は、観測されたフェノタイプの n 次元ベクトルを表し、 μ は、帰無仮説（すなわち着目しているフェノタイプはジェノタイプとの関係を有さないという仮説）の下におけるフェノタイプの予測平均の n 次元ベクトルである。また、行列 G は、以下のように表される。

$$G = (g_1, \dots, g_i, \dots, g_m)^T \quad (2)$$

g_i は、i 番目の S N P について n 人の被観測者のジェノタイプを含む n 次元のベクトルである。なお、 g_i は、離散値（例えば 0, 1, 2）でも連続値（例えば 0 以上 2 以下）でも良い。

行列 W は、以下のように表される。

$$W = \text{diag} (w_1, \dots, w_j, \dots, w_m) \quad (3)$$

W は、 $m \times m$ の対角行列であり、 w_j は、j 番目の S N P についての重み値を表している。

[0029] 本実施の形態で採用可能な統計量は、S K A T の統計量だけに限定されず

、例えば、Burden Testの統計量やc-alpha Test統計量などを用いることができる。これらの統計量については、例えば、Seunggeng Lee et al, "Rare-Variant Association Analysis: Study Designs and Statistical Tests", The American journal of Human Genetics 95, 5-23, July 3, 2014、Benjamin et al. "Testing for an Unusual Distribution of Rare Variants" PLoS Genet 7(3), 2011などを参照のこと。これらの文献を本明細書に取り込む。

[0030] 次に、p値算出部140は、集合 S_{B_c} にNUL Lを代入し、統計量の計算回数を表す変数 B_c に0を代入する（ステップS9）。

[0031] そして、p値算出部140は、 B_c がユーザ指定のBより小さいか否かを判断する（ステップS11）。この条件が満たされる場合には、処理は端子Aを介して図4の処理に移行する。

[0032] 一方、ステップS11の条件が満たされない場合、すなわち計算回数 B_c がユーザ指定の最大計算回数B以上となった場合には、p値算出部140は、ステップS5で特定されたSNP setについての推定p値として、図4の処理で計算される p_{B_c} を、当該SNP setのデータ（例えばSNP setに含まれるSNPの位置データ）に対応付けて第2データ格納部150に格納する（ステップS13）。

[0033] その後、p値算出部140は、未処理のSNP setが存在するか否かを判断する（ステップS15）。存在する場合には、処理はステップS5に戻る。一方、未処理のSNP setが存在しない場合には、抽出部160は、第2データ格納部150において、推定p値がユーザが指定した p_{sig} 未満であるSNP setを有意水準を超えるSNP setとして抽出して、抽出したSNP setを出力部170に出力する。出力部170は、抽出部160の抽出結果を、出力装置（例えば表示装置、印刷装置、他のコンピュータなど）に出力する（ステップS17）。

[0034] 図2に示した処理は、従来とおおよそ変わらない処理である。処理の高速化については図4及び図5を用いて詳しく述べる。

[0035] 端子Aを介して図2のステップS11から移行してくると、置換処理部1

42は、置換処理を実行する（図4：ステップS21）。この置換処理については、図5を用いて詳しく述べる。

[0036] 置換処理部142は、カウンタ t を0に初期化し、集合 S_T にNULLを代入する（図5：ステップS41）。そして、置換処理部142は、観測データ格納部120に格納されている観測データにおいてフェノタイプの入れ替えを実行する（ステップS43）。各被観測者について、復元抽出又は非復元抽出によってフェノタイプの値を抽出して、抽出したフェノタイプの値に入れ替えて、フェノタイプ入れ替え後の観測データを生成する。すなわち、処理対象の被観測者のフェノタイプの値を母集団として、母集団から標本を選ぶ場合に1つの個体を選んだ後にその個体を母集団の中に戻す方法が復元抽出であり、一度選んだ個体を母集団の中に戻さずに標本を選ぶ方法が非復元抽出である。図3の例では、被観測者1001のフェノタイプの値を、被観測者1003のフェノタイプの値として設定するといった処理を行う。

[0037] そして、置換処理部142は、統計量算出部141に、フェノタイプ入れ替え後の観測データに関する所定の統計量 s_t を算出させる（ステップS45）。統計量算出部141は、処理対象の観測データは異なるが、算出方法はステップS7で述べた方法と同じである。

[0038] その後、置換処理部142は、算出された統計量 s_t を、集合 S_T に入れる（ステップS47）。また、置換処理部142は、カウンタ t を1インクリメントする（ステップS49）。

[0039] そして、置換処理部142は、カウンタ t の値が、ユーザ指定の定数 T を超えたか否かを判断する（ステップS51）。カウンタ t の値が、ユーザ指定の定数 T 以下であれば、処理はステップS43に戻る。一方、カウンタ t の値が、ユーザ指定の定数 T を超えた場合には、処理は呼び出し元の処理に戻る。なお、 T は1であってもよいが、ある程度の数値である方が、処理の効率化が図られる。

[0040] このような処理を実行すれば、 T 個の統計量 s_t が集合 S_T に入れられることになる。

[0041] 図4の処理の説明に戻って、p値算出部140は、集合 S_{B_c} に、集合 S_{B_c} 及び集合 S_T を代入する（ステップS23）。また、p値算出部140は、変数 B_c に、集合 S_{B_c} に含まれる統計量の数を代入する（ステップS25）。さらに、p値算出部140は、変数 b に、集合 S_{B_c} 内において統計量 s よりも高い統計量 s_t の数を計数して代入する（ステップS27）。

[0042] そして、p値算出部140は、 b/B_c を算出して、推定p値を表す変数 p_{B_c} に代入する（ステップS29）。

[0043] さらに、p値算出部140は、ガウス分布 $N(0, p_{B_c}(1-p_{B_c})/B_c)$ から、信頼度 α 1の信頼区間の幅 d_a を算出する（ステップS31）。

[0044] 本実施の形態では、推定p値がユーザが指定した p_{sig} 未満（又は以下）であるSNP setを抽出することになるので、明らかに p_{sig} を下回るSNP set及び明らかに p_{sig} を上回るようなSNP setについては、途中で処理を打ち切っても問題が生じない。しかしながら、どのような条件を満たした場合に、途中で処理を打ち切るかという問題が別に生ずる。

[0045] 具体的には、 p_{sig} を中心として一定の幅に入る p_{B_c} が計算された場合には処理を継続し、 p_{sig} を中心として一定の幅に入らない p_{B_c} が計算された場合には処理を途中で打ち切ることとする。そこで、本実施の形態では、ユーザが指定した信頼度 α 1に応じた精度が出るように、当該信頼度 α 1に応じた幅を設定する。

[0046] より具体的には、統計量 s 未満の統計量が算出されると1となり、それ以外であれば0となるバイナリのランダムな変数 X を考え、計算された統計量についての X の期待値 $E x p [X]$ 及び分散 $V a r [X]$ は、以下のように表される。

$$E x p [X] = p_{B_c} = b / B_c$$

$$V a r [X] = E x p [X^2] - E x p [X]^2 = p_{B_c} (1 - p_{B_c})$$

[0047] 大数の法則から期待値の分散は以下ようになる。

$$V a r [p_{B_c}] = p_{B_c} (1 - p_{B_c}) / B_c \quad (4)$$

[0048] さらに中心極限定理から、 p_{B_c} はガウス分布 $N(0, p_{B_c}(1-p_{B_c})/B_c)$

）（平均0で、分散は（4）式のガウス分布）に対応するものとする。そうすると、このガウス分布における信頼度 $\alpha 1$ に対応する信頼区間の幅 d_a を、信頼度 $\alpha 1$ に応じた幅として用いることが適当であることが分かる。

[0049] ガウス分布 $N(0, p_{Bc}(1 - p_{Bc}) / B_c)$ は、統計量 s_t の計算回数が増加すれば動的に変化する。そして、この分布 $N(0, p_{Bc}(1 - p_{Bc}) / B_c)$ が分かれば、ユーザ指定の信頼度 $\alpha 1$ に対応する信頼区間を算出する関数は既知であるから、その幅 d_a の詳細な計算方法については説明を省略する。但し、一般的に、信頼区間の下限及び上限は、以下のように表される。

下限：標本平均 $- t \times$ 標本標準偏差 $/ (\text{標本の数})^{0.5}$

上限：標本平均 $+ t \times$ 標本標準偏差 $/ (\text{標本の数})^{0.5}$

よって、信頼区間の幅は、 $t \times$ 標本標準偏差 $/ (\text{標本の数})^{0.5}$ となる。但し、 t は、既知の t 分布表から得られる標本の数 $- 1$ と信頼度とに対応する値である。

[0050] 一方、一時的に $b = 0$ 又は $b = B_c$ となる状態が生じ得る。このような場合には、分散が0になるので、上で述べたような基準にて処理を途中で終了させるのは妥当ではない。そこで、本実施の形態では、負の二項分布を採用する。ここで、 Y を試行回数を表す正の整数のランダム変数として、 $p_{sig\ single} = p_{sig} * S N P\ s\ e\ t$ 数とする。そうすると、真の p 値は、最も高く p_{sig} であり、最も低く $p_{sig\ single}$ となる。

[0051] よって、 $b = 0$ の場合には、 B_c の確率は負の二項分布 $NB(Y = B_c; B_c, 1 - p_{sig})$ と表され、この確率が、最も高く $p_{Bc} = p_{sig}$ で信頼度 $\alpha 2$ となる有意水準 $p_{a2} (= 1 - \alpha 2)$ より小さければ、処理を途中で停止して、 $p_{Bc} = 1 / B_c$ と設定する。

[0052] すなわち、 p_{Bc} が0である場合には、以下の事象を検出すると、処理を停止する。

予め指定された信頼度 $\alpha 2$ に対応する有意水準 $p_{a2} >$ （算出した統計量 s_t が観測データに基づき算出した統計量 s より低くなる確率を $(1 - p_{sig})$ として、 B_c 回だけ実行すると、 B_c 回だけ、算出した統計量 s_t が観測データに基づき

算出した統計量 s より低くなる確率)

[0053] また、 $b = B_c$ の場合には、 B_c の確率は負の二項分布 $NB(Y = B_c; B_c, p_{sig\ single})$ と表され、この確率が、最も低くて $p_{Bc} = p_{sig\ single}$ で有意水準 p_{a2} より小さければ、処理を途中で停止して、 $p_{Bc} = 1$ と設定する。

[0054] すなわち、 $p_{Bc} = 1$ である場合は、以下の事象を検出すると、処理を停止する。

予め指定された信頼度 α_2 に対応する有意水準 $p_{a2} >$ (算出した統計量 s_t が観測データに基づき算出した統計量 s より高くなる確率を $p_{sig\ single}$ として、 B_c 回だけ実行すると、 B_c 回だけ、算出した統計量 s_t が観測データに基づき算出した統計量 s より高くなる確率)

[0055] このような考え方から、 p 値算出部 140 は、 $p_{Bc} = 0$ 且つ $NB(Y = B_c; B_c, 1 - p_{sig}) < p_{a2}$ という条件を満たすか否かを判断する (ステップ S33)。このような条件を満たす場合には、 $p_{Bc} = 1 / B_c$ として処理を途中で停止しても良いので、 p 値算出部 140 は、 p_{Bc} に対して $1 / B_c$ と設定して (ステップ S34)、処理は端子 C を介して図 2 のステップ S13 に戻る。

[0056] 一方、ステップ S33 の条件を満たさない場合には、 p 値算出部 140 は、 $p_{Bc} = 1$ 且つ $NB(Y = B_c; B_c, p_{sig\ single} = p_{sig} \times SNP\ set\ 数) < p_{a2}$ という条件を満たすか否かを判断する (ステップ S35)。このような条件を満たす場合には、 $p_{Bc} = 1$ として処理を途中で停止しても良い。よって、処理は端子 C を介して図 2 のステップ S13 に戻る。

[0057] 一方、ステップ S35 の条件を満たさない場合には、 p 値算出部 140 は、 $p_{Bc} > p_{sig} + d_a / 2$ 又は $p_{Bc} < p_{sig} - d_a / 2$ という条件を満たしているか否かを判断する (ステップ S39)。すなわち、 p_{Bc} が、 p_{sig} を中心に、信頼区間の幅 d_a に入るか否かを判断している。この条件を満たす場合には、処理を途中で停止して良いので、端子 C を介して図 2 のステップ S13 に戻る。なお、この条件に含まれる不等式は、その関係を維持しつつ変形することが可能である。一方、この条件を満たさない場合には、処理は端子 B を介して図 2 のステップ S11 に戻る。

- [0058] このように、 p_{bc} が、 p_{sig} を中心に、ユーザ指定の信頼度が反映されて決定される信頼区間の幅 d_a に入る場合には、入念に置換処理を実行して正確な p 値を算出するようにする。 d_a は、統計量 s_t の算出状況に応じて変化する分布 $N(0, p_{bc}(1 - p_{bc}) / B_c)$ に基づき動的に変化する。
- [0059] ステップS33及びS35の例外を除けば、 p_{bc} が、 p_{sig} を中心に、信頼区間の幅 d_a に入らなければ、その時点で次のSNP setについての処理に移行するので、それだけ置換処理の実行回数を削減できる。従って、全体の処理速度を高速化できるようになる。
- [0060] 以上述べた実施の形態は一例であって、本発明はこれに限定されるものではない。例えば、図1に示した機能ブロック図は一例であって、プログラムモジュール構成とは一致しない場合もある。処理フローについては、処理結果が変わらない限り、処理順番を入れ替えたり、処理を並列に実行するようにしても良い。本実施の形態の場合、特にSNP set毎に、異なるコンピュータ又はプロセッサコアで並列化しても良い。
- [0061] また、重要なSNP setについては、 $p_{bc} < p_{sig}$ が成り立つ限り置換処理を繰り返し実行することで正確な p 値を算出するようにしても良い。重要か否かについては、フェノタイプの入れ替え無しで算出された統計量 s の値に基づき決定するようにしても良い。特に、フェノタイプの入れ替え無しで算出された統計量 s の値が、所定の閾値以上となる場合には、本処理の対象外とするようにしても良い。
- [0062] さらに、図1の情報処理装置100のように1台のコンピュータで実装するのではなく、クライアント端末から1又は複数の情報処理装置100にネットワークを介してアクセスして処理を行わせるようにしても良い。複数の情報処理装置100を用いる場合には、上で述べた機能を分担すればよい。
- [0063] なお、上で述べた情報処理装置100は、コンピュータ装置であって、図6に示すように、メモリ2501とCPU (Central Processing Unit) 群2503とハードディスク・ドライブ (HDD : Hard Disk Drive) 2505と表示装置2509に接続される表示制御部2507とリムーバブル・ディス

ク2511用のドライブ装置2513と入力装置2515とネットワークに接続するための通信制御部2517とがバス2519で接続されている。オペレーティング・システム（OS：Operating System）及び本実施例における処理を実施するためのアプリケーション・プログラムは、HDD2505に格納されており、CPU群2503により実行される際にはHDD2505からメモリ2501に読み出される。CPU群2503は、アプリケーション・プログラムの処理内容に応じて表示制御部2507、通信制御部2517、ドライブ装置2513を制御して、所定の動作を行わせる。また、処理途中のデータについては、主としてメモリ2501に格納されるが、HDD2505に格納されるようにしてもよい。本技術の実施例では、上で述べた処理を実施するためのアプリケーション・プログラムはコンピュータ読み取り可能なリムーバブル・ディスク2511に格納されて頒布され、ドライブ装置2513からHDD2505にインストールされる。インターネットなどのネットワーク及び通信制御部2517を経由して、HDD2505にインストールされる場合もある。このようなコンピュータ装置は、上で述べたCPU群2503、メモリ2501などのハードウェアとOS及びアプリケーション・プログラムなどのプログラムとが有機的に協働することにより、上で述べたような各種機能を実現する。

[0064] 以上述べた本実施の形態をまとめると、以下のようになる。

[0065] 本実施の形態に係る方法（検定方法、統計処理方法、テスト方法等とも呼ぶ）は、（A）ゲノム中の指定部分に存在する一塩基多型の集合についてジェノタイプとフェノタイプとの関係を表す統計量を観測データに基づいて算出するステップと、（B）観測データにおいてフェノタイプの値を入れ替えた場合における統計量を算出する処理を、算出した統計量が観測データに基づき算出した統計量より高くなる回数の割合が第1の閾値を超えるか若しくは第2の閾値を下回るという事象を検出するまで、又は当該事象が検出されなければ予め指定された回数実行するまで、実行する処理実行ステップと、（C）上記割合を一塩基多型の集合に対するp値として出力するステップと

を含む。そして、上で述べた第1の閾値及び第2の閾値は、観測データに基づき算出した統計量より高い統計量が算出されるケースの出現分布及び予め指定された信頼度に基づき、予め指定された有意水準を表す値から動的に決定されるものである。

[0066] このように、予め指定された信頼度を反映させた精度が実現されるように第1及び第2の閾値が設定され、第1の閾値を超えるか又は第2の閾値を下回る場合には処理が途中で打ち切られるので、permutation testの高速化も実現されている。さらに、第1及び第2の閾値は、観測データに基づき算出した統計量より高い統計量が算出されるケースの出現分布にも基づくものであるため、第1及び第2の閾値は統計量の算出状況に応じて動的に変化する。すなわち、本実施の形態は、統計量の算出方法に依存せずに適用可能である。

[0067] なお、上で述べた第1の閾値が、予め指定された有意水準を表す値に対して、上記出現分布において予め指定された信頼度に対応する信頼区間の幅の $1/2$ の値を加算した値であり、上で述べた第2の閾値が、予め指定された有意水準を表す値から、上記出現分布において予め指定された信頼度に対応する信頼区間の幅の $1/2$ の値を引いた値である場合もある。このように、予め指定された有意水準を表す値を中心に、「予め指定された信頼度に対応する信頼区間の幅」を設定することで、精度からして処理継続が適切な一塩基多型の集合を適切に特定できるようになる。

[0068] 上で述べた処理実行ステップにおいて、(b1)上記割合が0である場合には、(予め指定された第2の信頼度に対応する第2の有意水準) > (算出した統計量が観測データに基づき算出した統計量より低くなる確率を(1 - 予め指定された有意水準)として、上記処理を現在の実行回数だけ実行すると、上記処理の現在の実行回数だけ、算出した統計量が観測データに基づき算出した統計量より低くなる確率)となる第2の事象、又は(b2)上記割合が1である場合には、(上記第2の有意水準) > (算出した統計量が観測データに基づき算出した統計量より高くなる確率を予め指定された有意水準

を表す値に対して処理すべき一塩基多型の集合の数を乗じた値として、上記処理を現在の実行回数だけ実行すると、上記処理の現在の実行回数だけ、算出した統計量が観測データに基づき算出した統計量より高くなる確率)となる第3の事象を検出するまで上記処理を実行するようにしても良い。この場合、上記第2の事象を検出すると、 $1 / \text{処理の現在の実行回数}$ をp値として出力し、上記第3の事象を検出すると、1をp値として出力するようにしても良い。上記割合が0である場合や1である場合等、特別なケースにも対処できる。

[0069] また、本実施の形態に係る方法は、(D)上記p値に基づき、一塩基多型の集合の採否を決定するステップをさらに含むようにしても良い。これによって有意な一塩基多型の集合を得ることができるようになる。

[0070] また、上で述べた処理実行ステップが、上記処理を所定回数実行する毎に、上記事象の発生の有無を判断するステップを含むようにしても良い。このようにすれば、より処理の効率化が図られる。

[0071] なお、上で述べたような処理をプロセッサ又はコンピュータに実行させるためのプログラムを作成することができ、当該プログラムは、例えばフレキシブル・ディスク、CD-ROMなどの光ディスク、光磁気ディスク、半導体メモリ(例えばROM)、ハードディスク等のコンピュータ読み取り可能な記憶媒体又は記憶装置に格納される。なお、処理途中のデータについては、RAM等の記憶装置に一時保管される。

請求の範囲

[請求項1] ゲノム中の指定部分に存在する一塩基多型の集合についてジェノタイプとフェノタイプとの関係を表す統計量を観測データに基づいて算出するステップと、

前記観測データにおいて前記フェノタイプの値を入れ替えた場合における統計量を算出する処理を、算出した前記統計量が前記観測データに基づき算出した前記統計量より高くなる回数の割合が第1の閾値を超えるか若しくは第2の閾値を下回るという事象を検出するまで、又は前記事象が検出されなければ予め指定された回数実行するまで、実行する処理実行ステップと、

前記割合を前記一塩基多型の集合に対するp値として出力するステップと、

を、コンピュータに実行させ、

前記第1の閾値及び前記第2の閾値は、前記観測データに基づき算出した前記統計量より高い統計量が算出されるケースの出現分布及び予め指定された信頼度に基づき、予め指定された有意水準を表す値から動的に決定される

ことを特徴とするプログラム。

[請求項2] 前記第1の閾値が、前記予め指定された有意水準を表す値に対して、前記出現分布において前記予め指定された信頼度に対応する信頼区間の幅の $1/2$ の値を加算した値であり、

前記第2の閾値が、前記予め指定された有意水準を表す値から、前記出現分布において前記予め指定された信頼度に対応する信頼区間の幅の $1/2$ の値を引いた値である

請求項1記載のプログラム。

[請求項3] 前記処理実行ステップにおいて、

前記割合が0である場合には、（予め指定された第2の信頼度に対応する第2の有意水準） $>$ （算出した前記統計量が前記観測データに

に基づき算出した前記統計量より低くなる確率を（1－前記予め指定された有意水準）として、前記処理を現在の実行回数だけ実行すると、前記処理の現在の実行回数だけ、算出した前記統計量が前記観測データに基づき算出した前記統計量より低くなる確率）となる第2の事象、又は

前記割合が1である場合には、（前記第2の有意水準）>（算出した前記統計量が前記観測データに基づき算出した前記統計量より高くなる確率を前記予め指定された有意水準を表す値に対して処理すべき前記一塩基多型の集合の数を乗じた値として、前記処理を現在の実行回数だけ実行すると、前記処理の現在の実行回数だけ、算出した前記統計量が前記観測データに基づき算出した前記統計量より高くなる確率）となる第3の事象

を検出するまで前記処理を実行して、

前記第2の事象を検出すると、 $1 / \text{前記処理の現在の実行回数}$ をp値として出力し、

前記第3の事象を検出すると、1をp値として出力する

請求項1又は2記載のプログラム。

[請求項4] 前記p値に基づき、前記一塩基多型の集合の採否を決定するステップ

をさらに前記コンピュータに実行させるための請求項1乃至3のいずれか1つ記載のプログラム。

[請求項5] 前記処理実行ステップが、

前記処理を所定回数実行する毎に、前記事象の発生の有無を判断するステップを

含む請求項1乃至4のいずれか1つ記載のプログラム。

[請求項6] ゲノム中の指定部分に存在する一塩基多型の集合についてジェノタイプとフェノタイプとの関係を表す統計量を観測データに基づいて算出するステップと、

前記観測データにおいて前記フェノタイプの値を入れ替えた場合における統計量を算出する処理を、算出した前記統計量が前記観測データに基づき算出した前記統計量より高くなる回数の割合が第1の閾値を超えるか若しくは第2の閾値を下回るという事象を検出するまで、又は前記事象が検出されなければ予め指定された回数実行するまで、実行する処理実行ステップと、

前記割合を前記一塩基多型の集合に対するp値として出力するステップと、

を、コンピュータが実行し、

前記第1の閾値及び前記第2の閾値は、前記観測データに基づき算出した前記統計量より高い統計量が算出されるケースの出現分布及び予め指定された信頼度に基づき、予め指定された有意水準を表す値から動的に決定される

ことを特徴とする情報処理方法。

[請求項7]

ゲノム中の指定部分に存在する一塩基多型の集合についてジェノタイプとフェノタイプとの関係を表す統計量を観測データに基づいて算出する算出手段と、

前記観測データにおいて前記フェノタイプの値を入れ替えた場合における統計量を算出する処理を、算出した前記統計量が前記観測データに基づき算出した前記統計量より高くなる回数の割合が第1の閾値を超えるか若しくは第2の閾値を下回るという事象を検出するまで、又は前記事象が検出されなければ予め指定された回数実行するまで、実行する処理実行手段と、

前記割合を前記一塩基多型の集合に対するp値として出力する手段と、

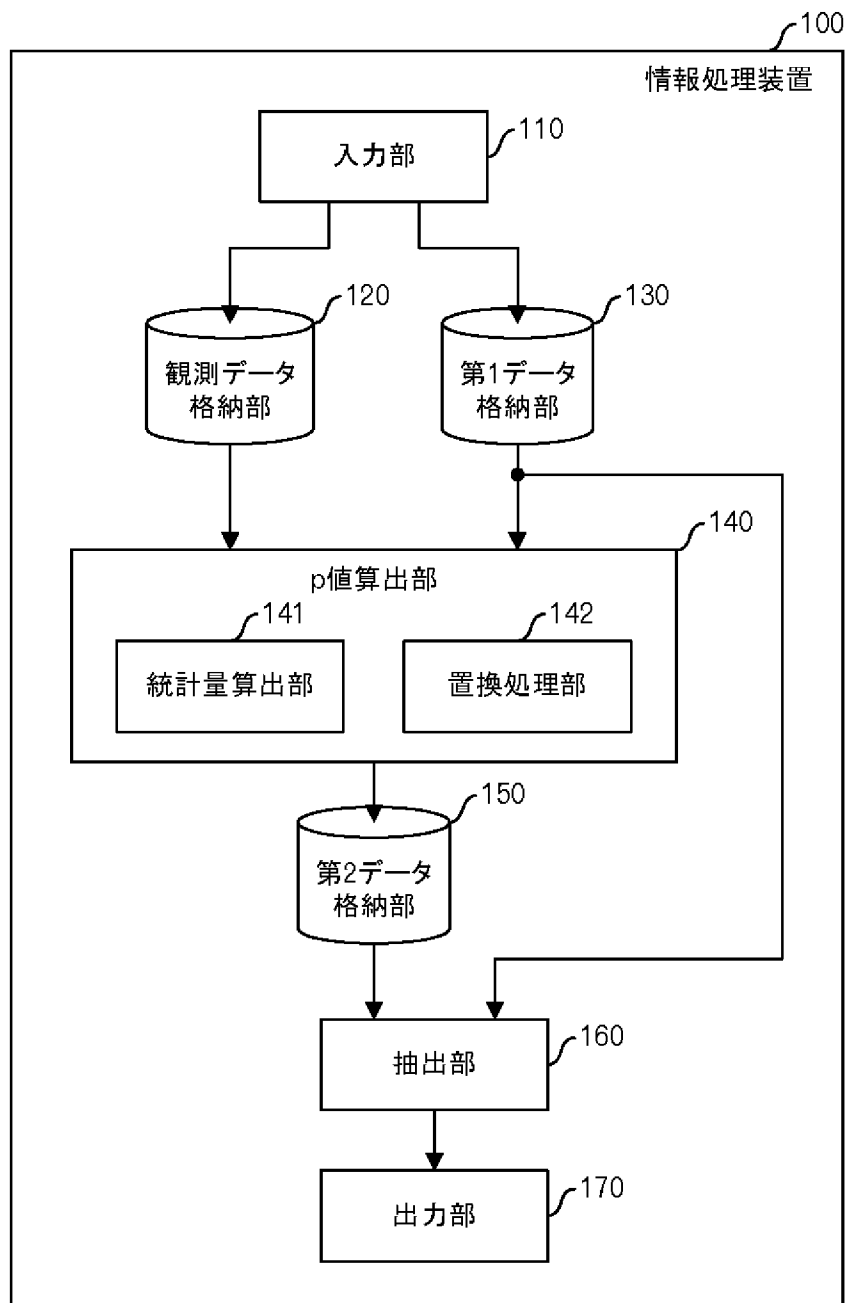
を有し、

前記第1の閾値及び前記第2の閾値は、前記観測データに基づき算出した前記統計量より高い統計量が算出されるケースの出現分布及び

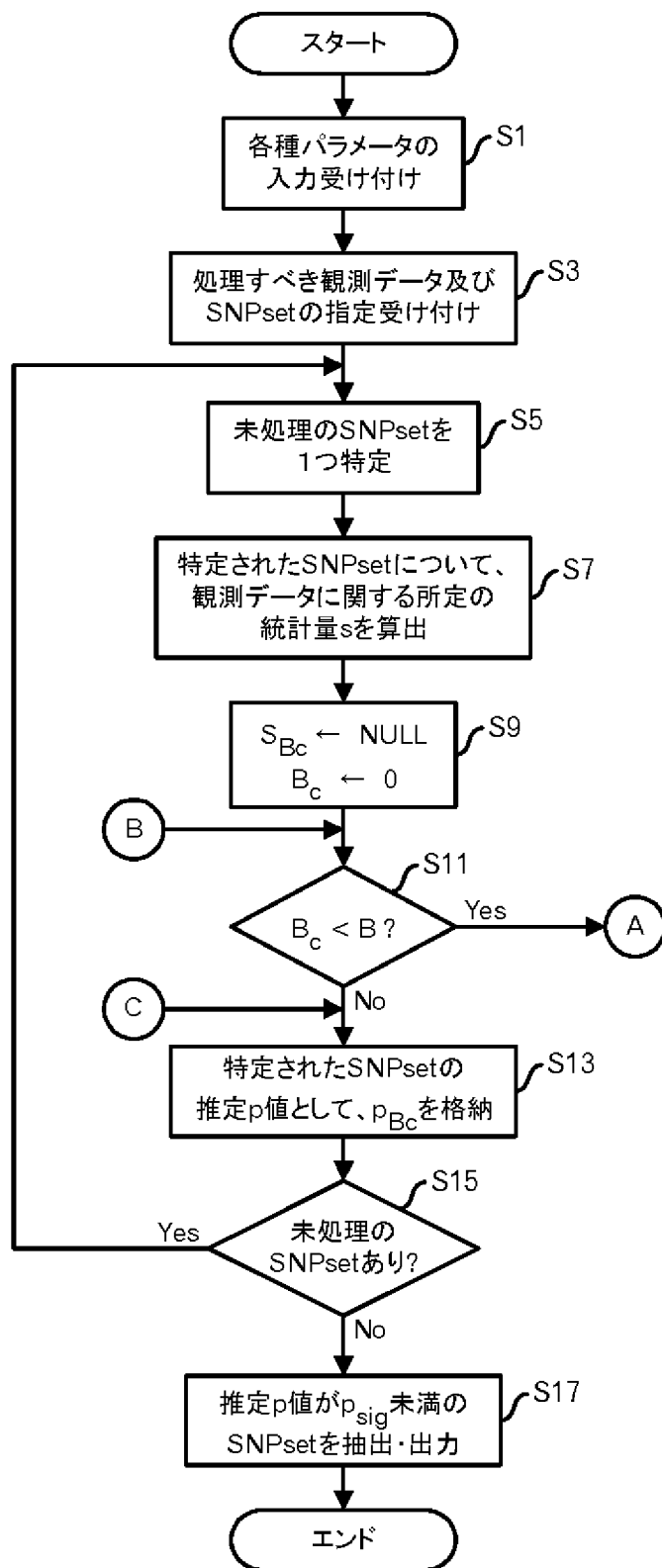
予め指定された信頼度に基づき、予め指定された有意水準を表す値から動的に決定される

ことを特徴とする情報処理装置。

[図1]



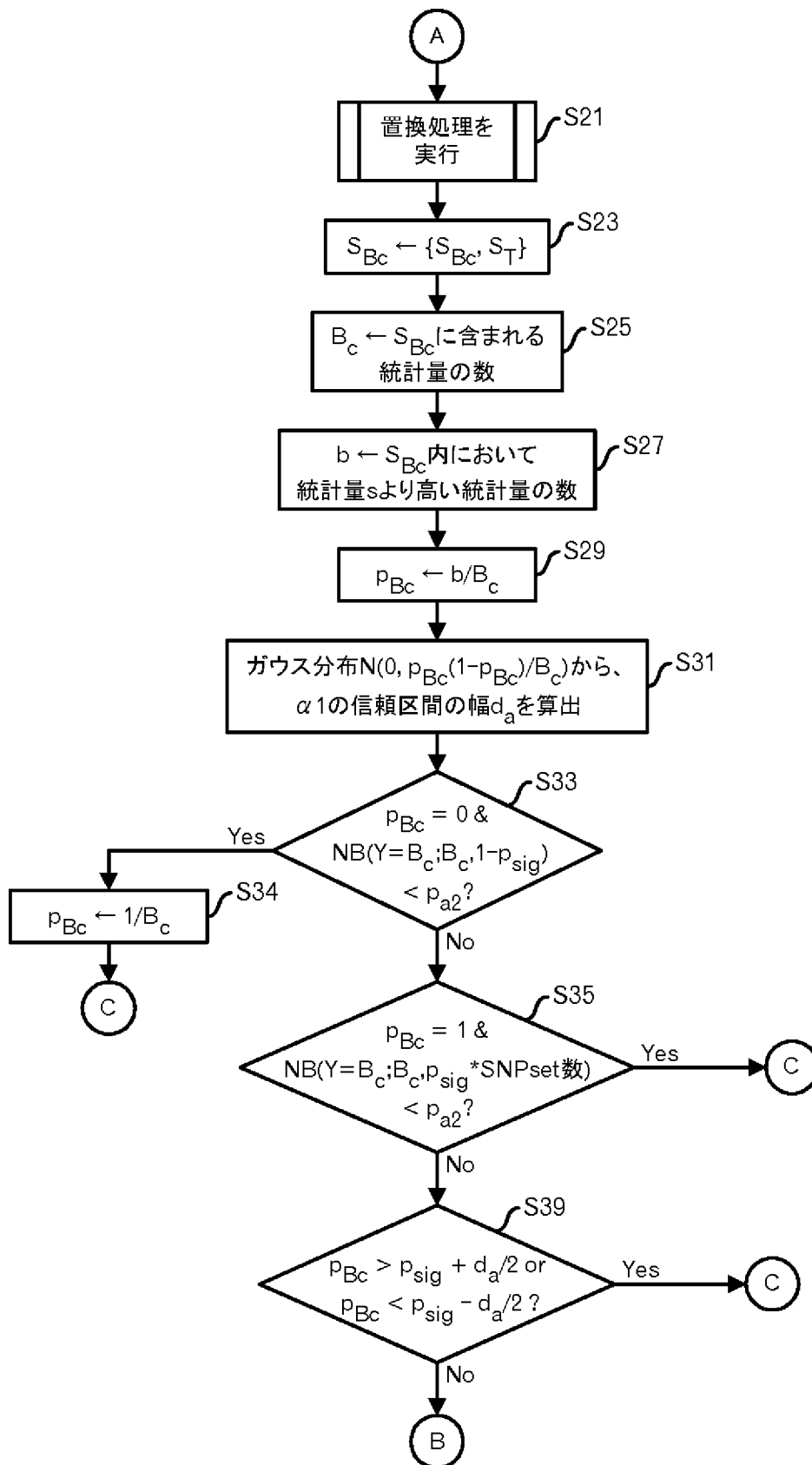
[図2]



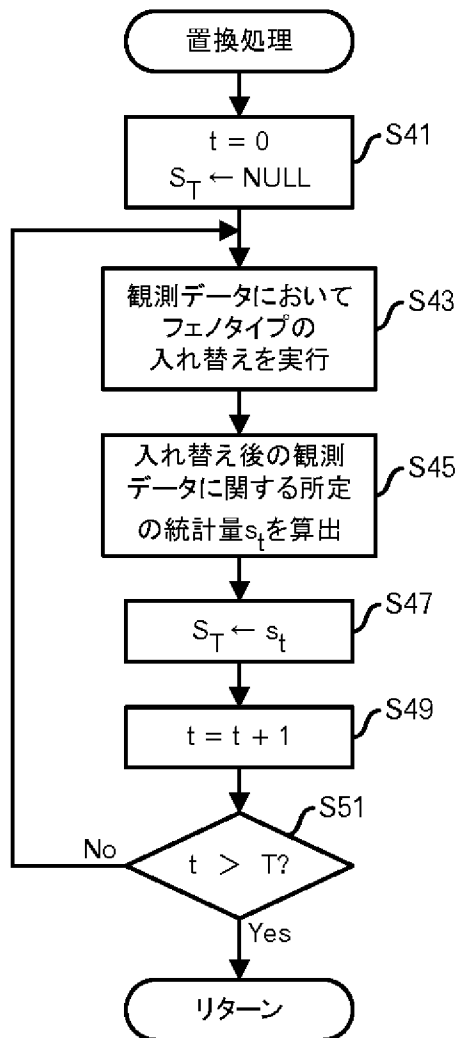
[図3]

		ゲノム		
被観測者No.	フェノタイプ		H	
1001	x1, y1		...A...A...	
1002	x2, y2		...C...C...	
1003	x3, y3		...A...C...	

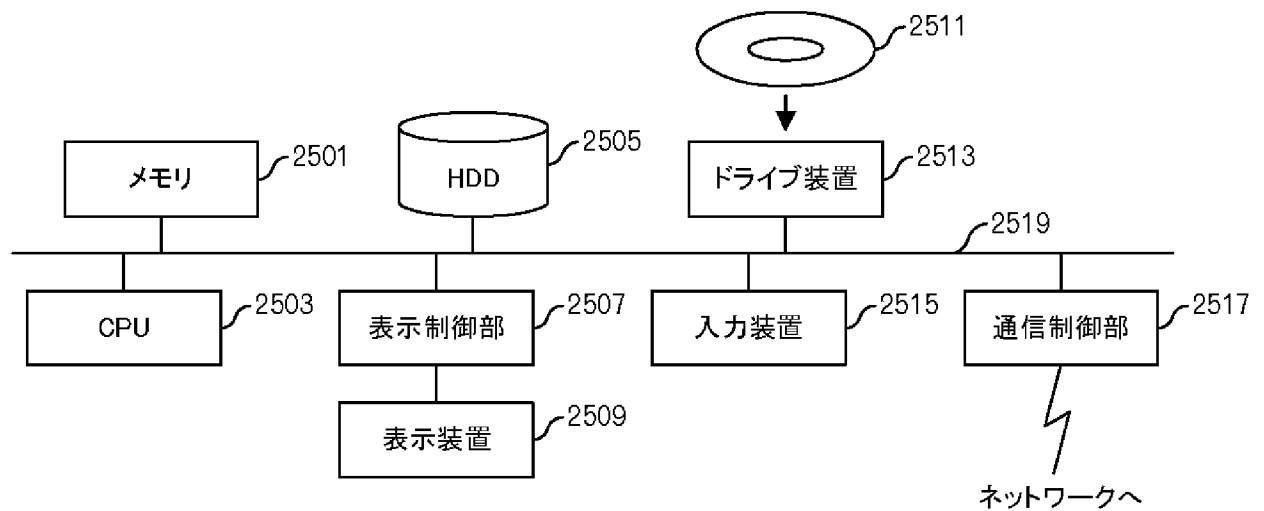
[図4]



[図5]



[図6]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2017/001056

A. CLASSIFICATION OF SUBJECT MATTER

G06F19/24(2011.01)i, C12N15/00(2006.01)n

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F19/24, C12N15/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2017
Kokai Jitsuyo Shinan Koho	1971-2017	Toroku Jitsuyo Shinan Koho	1994-2017

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2009-69911 A (Mizuho Information & Research Institute, Inc.), 02 April 2009 (02.04.2009), entire text; all drawings (Family: none)	1-7
A	JP 2008-262427 A (Niigata University), 30 October 2008 (30.10.2008), entire text; all drawings (Family: none)	1-7
A	JP 2013-175135 A (NTT Data Corp.), 05 September 2013 (05.09.2013), entire text; all drawings (Family: none)	1-7

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
23 February 2017 (23.02.17)

Date of mailing of the international search report
07 March 2017 (07.03.17)

Name and mailing address of the ISA/
Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku,
Tokyo 100-8915, Japan

Authorized officer

Telephone No.

A. 発明の属する分野の分類 (国際特許分類 (IPC))

Int.Cl. G06F19/24(2011.01)i, C12N15/00(2006.01)n

B. 調査を行った分野

調査を行った最小限資料 (国際特許分類 (IPC))

Int.Cl. G06F19/24, C12N15/00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922-1996年
日本国公開実用新案公報	1971-2017年
日本国実用新案登録公報	1996-2017年
日本国登録実用新案公報	1994-2017年

国際調査で使用した電子データベース (データベースの名称、調査に使用した用語)

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
A	JP 2009-69911 A (みずほ情報総研株式会社) 2009.04.02, 全文, 全図 (ファミリーなし)	1-7
A	JP 2008-262427 A (国立大学法人 新潟大学) 2008.10.30, 全文, 全図 (ファミリーなし)	1-7
A	JP 2013-175135 A (株式会社エヌ・ティ・ティ・データ) 2013.09.05, 全文, 全図 (ファミリーなし)	1-7

C欄の続きにも文献が列挙されている。

パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」 特に関連のある文献ではなく、一般的技術水準を示すもの	「T」 国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「E」 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの	「X」 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「L」 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献 (理由を付す)	「Y」 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「O」 口頭による開示、使用、展示等に言及する文献	「&」 同一パテントファミリー文献
「P」 国際出願日前で、かつ優先権の主張の基礎となる出願	

国際調査を完了した日

23.02.2017

国際調査報告の発送日

07.03.2017

国際調査機関の名称及びあて先

日本国特許庁 (ISA/J P)

郵便番号 100-8915

東京都千代田区霞が関三丁目4番3号

特許庁審査官 (権限のある職員)

山内 裕史

5 L

4064

電話番号 03-3581-1101 内線 3562