

(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
2012年3月29日(29.03.2012)

PCT

(10) 国際公開番号
WO 2012/039484 A1

- (51) 国際特許分類:
C12Q 1/68 (2006.01) C12N 15/09 (2006.01)
C12M 1/00 (2006.01)
- (21) 国際出願番号: PCT/JP2011/071731
- (22) 国際出願日: 2011年9月22日(22.09.2011)
- (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 2010-212116 2010年9月22日(22.09.2010) JP
特願 2011-053301 2011年3月10日(10.03.2011) JP
特願 2011-053729 2011年3月11日(11.03.2011) JP
- (71) 出願人(米国を除く全ての指定国について): 独立行政法人産業技術総合研究所(NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY) [JP/JP]; 〒1008921 東京都千代田区霞が関1丁目3番1号 Tokyo (JP).
- (72) 発明者; および
- (75) 発明者/出願人(米国についてのみ): 町田 雅之(MACHIDA Masayuki) [JP/JP]; 〒0628517 北海道札幌市豊平区月寒東2条17丁目2番1号 独立行政法人産業技術総合研究所北海道セン

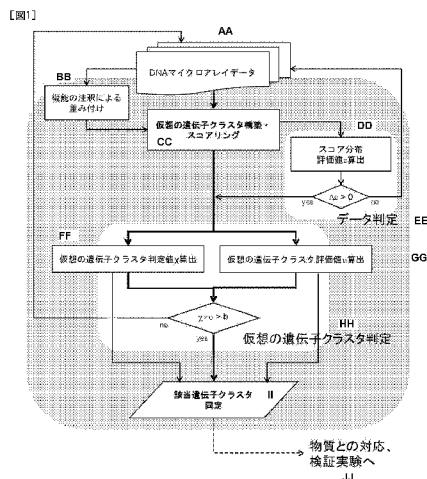
ター内 Hokkaido (JP). 小池 英明(KOIKE Hideaki) [JP/JP]; 〒3058568 茨城県つくば市東1-1-1 中央第6 独立行政法人産業技術総合研究所内 Ibaraki (JP). 梅村 舞子(UMEMURA Maiko) [JP/JP]; 〒0628517 北海道札幌市豊平区月寒東2条17丁目2番1号 独立行政法人産業技術総合研究所北海道センター内 Hokkaido (JP). 浅井 潔(ASAI Kiyoshi) [JP/JP]; 〒1350064 東京都江東区青海2-3-26 独立行政法人産業技術総合研究所内 Tokyo (JP). 堀本 勝久(HORIMOTO Katsuhisa) [JP/JP]; 〒1350064 東京都江東区青海2-3-26 独立行政法人産業技術総合研究所内 Tokyo (JP). 光山 統泰(MITSUYAMA Tōtai) [JP/JP]; 〒1350064 東京都江東区青海2-3-26 独立行政法人産業技術総合研究所内 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, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH,

[続葉有]

(54) Title: GENE CLUSTER, GENE SEARCHING/IDENTIFICATION METHOD, AND APPARATUS FOR THE METHOD

(54) 発明の名称: 遺伝子クラスタ及び遺伝子の探索、同定法およびそのための装置



AA DNA MICROARRAY DATA
BB WEIGHTING BY COMMENTS ON FUNCTIONS
CC CONSTRUCTION/SCORING OF VIRTUAL GENE CLUSTER
DD CALCULATION OF SCORE DISTRIBUTION EVALUATION VALUE
EE JUDGMENT OF DATA
FF CALCULATION OF DETERMINATION VALUE X FOR VIRTUAL GENE CLUSTER
GG CALCULATION OF EVALUATION VALUE Y FOR VIRTUAL GENE CLUSTER
HH JUDGMENT OF VIRTUAL GENE CLUSTER
II IDENTIFICATION OF THE GENE CLUSTER
JJ CORRESPONDENCE WITH SUBSTANCE, VERIFICATION EXPERIMENT

(57) Abstract: Provided is: a method for searching for or identifying a useful gene logically, systematically, within an extremely short time, and with high efficiency, which does not largely rely on the knowledge, experience or the like of a searcher and needs no any gene disruption experiment sequentially in the searching for the useful gene, as not in the case of the conventional techniques; and an apparatus for the method. Pieces of information on the variation of expression of individual genes on a genome, which are obtained on the basis of a microarray or the like, are combined together as the information on the variation of expression of virtual gene cluster units constituted by multiple genes, each of the virtual gene clusters is scored, and a gene cluster containing the useful gene and the useful gene contained in the cluster are searched on the basis of the score for each of the virtual gene clusters.

(57) 要約: 有用遺伝子の探索において、従来技術にみられる研究者の知識や経験等に大きく依存することなく、また、遺伝子破壊実験を逐次行わずとも、論理的、システムチックに極めて短時間で効率的に、有用遺伝子の探索、同定を行う方法及びそのための装置を提供する。マイクロアレイ等に基づくゲノム上の各遺伝子の発現変動情報を、複数の遺伝子により構成される仮定の遺伝子クラスタ単位の発現変動情報として合算して、仮定の各遺伝子クラスタをスコアリングし、この仮定の遺伝子クラスタのスコアに基づき、有用遺伝子を含む遺伝子クラスタ及び該クラスタに含まれる有用遺伝子を探索する。



PL, PT, QA, RO, RS, RU, RW, 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.

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, ML, MR,
NE, SN, TD, TG).

- (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, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア
(AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ
(AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR,

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：

遺伝子クラスタ及び遺伝子の探索、同定法およびそのための装置

技術分野

[0001] 本発明は、遺伝子クラスタを標的として探索し、該遺伝子クラスタ中の有用遺伝子を新たに見いだすことを目的とする、遺伝子クラスタ及び有用遺伝子の探索、同定法、およびそのための探索装置に関する。

背景技術

[0002] 二次代謝物質は、生理活性を有する可能性が高く、医薬のリード化合物として極めて有用である。二次代謝物質は多様で、放線菌、真菌、植物などの様々な生物種から発見されているが、発現する条件が特殊で知られていないことが多く、有用な性質を持つ多数の二次代謝物質が発見されないままに眠っていると考えられている。また発見されたとしても、安定で十分な量の生産が困難であることが利用の際の問題である。

一方、近年、DNAシーケンス技術の革新的な発展により、様々な生物種、特に微生物のゲノム情報の蓄積は加速度的に増加しており、3～5年後には数千種の微生物のゲノム塩基配列が明らかになることは確実である。このようなゲノム中の遺伝子配列と2次代謝産物との対応関係について、詳細かつ膨大な情報を収集してデータベースなどを構築することが可能になれば、これにより、二次代謝物質の構造、多様性、生物界での分布などに関する情報を、遺伝子の配列に基づいて推定することが可能になり、有用な未知の2次代謝物質の発見及び該2次代謝物質の生合成に関与する遺伝子の取得が容易になり、この遺伝子組み換え技術を用いて、2次代謝産物を安定して大量に生産することも可能となる。

[0003] 従来においても、様々な生物種から未知の有用な2次代謝物質を見いだすために、活性スクリーニングによる探索と構造決定が行われており、この際、用いた生物種の形態などの特徴による属の推定やrDNAの塩基配列など

を解析することによって属あるいは種の情報を得ることは試みられてきたものの、2次代謝物質の産生に関与する遺伝子の同定にまで至るケースはまれである。このような方法では、二次代謝物質を生合成する遺伝子は属・種の進化系統樹と矛盾することが多い上、機能が全く解明されていない未知の遺伝子が多数存在することから、二次代謝物質の構造、多様性、生物界での分布などを推定することは到底困難であった。

[0004] また、主として、代謝物質の測定（同定、定量）、ゲノム塩基配列、およびゲノム塩基配列に基づいて作製されたDNAマイクロアレイなどによる遺伝子の発現プロファイルなどの情報を利用して、着目する代謝物質の生合成遺伝子を推定する方法も行われていた。具体的には、着目する代謝物質の生産性が向上する条件（培養条件など）を設定し、この条件においてDNAマイクロアレイなどを用いて遺伝子の発現を測定し、この物質を生産していない条件で同様に測定して得られた遺伝子発現を比較することにより、この物質を生産するときに誘導される遺伝子を推定していた。しかし、培養条件などを変更することによって誘導される遺伝子の数は通常、100～1000以上のことがほとんどであり、遺伝子を特定することは極めて難しい。

[0005] そこで多くの場合、この物質を生産する条件を複数設定し、いずれの条件でも誘導される遺伝子を候補とすることが行われていた。しかし、生物を用いた実験結果は曖昧性が高いこと、測定誤差が大きいこと（DNAマイクロアレイによる遺伝子発現の測定では、一般的に2倍以上の誘導あるいは抑制が見られる場合に、実際に誘導や抑制がかかっていると判断される）、代謝系が複雑に制御されていることなどが理由で、複数の条件で共通に発現が誘導される遺伝子の候補が0（ゼロ）になってしまうことや、多数の遺伝子が候補に挙がって収斂させることができないことが多く、標的とする遺伝子を特定することはほぼ不可能であった。

[0006] そのため、上記の各条件で比較的高い強度で誘導される10～1000程度の遺伝子を候補として選び、必ずしも全ての条件で共通に誘導されていなかったとしても候補として残すなどの工夫をすること、候補とする遺伝子の

中から着目する代謝物質の生産に関与しそうな遺伝子を選択して各条件での誘導性を考慮して候補を絞り込むこと、二次代謝系の遺伝子がクラスタを形成する可能性が高いことなどを指標にして候補遺伝子の中にゲノム上で比較的近傍に位置する遺伝子群を探すことなどによって、可能性の高い遺伝子を絞り込んでいくことが行われていた。これらの「絞り込み」は、主として研究者の知識や経験、他の論文に記載された事実や推定などを参考にして行われた。また、この様な推定過程において、遺伝子破壊などによって、推定された遺伝子が確かに着目する代謝物質の生合成に必須かどうかについて、候補となった遺伝子について逐次検証することにより、標的とする遺伝子を特定することが必要不可欠であった。遺伝子破壊実験は、通常、数個の遺伝子について熟練した技術者が1ヶ月程度以上の時間を費やして行うことがやっとなであり、極めて時間と労力を有するステップである。そこで、通常は10～100位の候補遺伝子に絞り込んだ状態で、優先順位を付けて破壊実験を行うが、10番目以内に正しい遺伝子を候補として絞り込むことができれば、相当に幸運であると言える。また、形質転換系が存在しない場合には、遺伝子破壊実験ができないことから、検証そのものが不可能であり、遺伝子を特定することは困難であった。

[0007] 微生物のゲノム配列から二次代謝関連遺伝子を同定する手法についてはこれまでNRPSおよびPKSについていくつか報告されており(非特許文献1～5)、そのうち数個は検証も行われている(非特許文献3、4、6)。しかしいずれも、これらにおける反応の特殊性に注目して、遺伝子配列情報から特定の反応を行うモチーフを抽出する戦略をとっており、同定される遺伝子の範囲はNRPSおよびPKSに限定されている。すなわち既存の手法は、遺伝子と機能の1対1対応という考え方に基づいたもので、微生物における二次代謝関連遺伝子が集合してゲノム上に位置しているという生物学的知見に基づいた本提案手法とは本質的に異なる。既存の手法に対して本提案手法によって初めて可能となることとして、代表的な微生物の二次代謝経路であるNRPSやPKSだけでなく、その他の反応にかかるモチーフを含む遺伝子群の同定が挙げられる。

また発現情報に基づいて同定するため、休眠遺伝子やPseudo遺伝子等の実際には働いていない遺伝子群を避けることができる。

[0008] また、抗菌物質を生産する遺伝子をゲノム情報に基づいて同定した例もあるが（特許文献1）、この方法は、生産物質として蛋白又はRNAである抗菌物質を想定し、また「クローン・カバレッジ」の低い遺伝子を増殖抑制遺伝子として同定しており、この方法は、それ自体に配列情報が無く、また、極めて多様性に富む2次代謝産物について、その産生に関与する遺伝子を探索するための方法とはなり得ない。

[0009] 特許文献1：WO 2008 / 133479 (Univ. California)

[0010] 非特許文献1：Wilkinson et al., Nat. Chem. Biol., vol. 3-7, 379-386 (2007)

非特許文献2：BMC Bioinform., vol. 10:185, 1-10 (2009)

非特許文献3：Zazopoulos et al., Nat. Biotech., vol. 21, 187-190 (2003)

非特許文献4：Bergmann et al., Nat. Chem. Biol., vol. 3-4, 213-217 (2007)

非特許文献5：Challis et al., FEMS Microbiol. Lett., vol. 187, 111-114 (2000)

非特許文献6：Lautru et al., Nat. Chem. Biol., vol. 1-5, 265-269 (2005)

発明の開示

発明が解決しようとする課題

[0011] 本発明は、上記従来技術における、代謝産物の産生に関与する遺伝子等の有用遺伝子の探索において、上記従来技術にみられる研究者の知識や経験等に大きく依存することなく、また、遺伝子破壊実験を逐次行わずとも、論理的、システマチックに極めて短時間で効率的に、有用遺伝子の探索、同定を行う方法、およびそのための装置を提供することにより、今後増大するゲノム情報を活用して、新たな有用遺伝子の探索を加速させ、ゲノム中の遺伝子配列と有用遺伝子の対応関係について、詳細かつ膨大な情報を収集してデータベースなどを構築することを可能にし、多数の有用な遺伝子産物の発見に資することを課題とするものである。

課題を解決するための手段

[0012] 本発明者は、上記課題を解決するため鋭意研究の結果、従来におけるマイクロアレイのゲノム遺伝子の発現誘導あるいは破壊実験等による有用遺伝子探索にみられるような、ゲノム中の個々の遺伝子についての発現変動情報から直接標的遺伝子を絞りこむのではなく、マイクロアレイ等によるゲノム上の各遺伝子の発現変動情報を、複数の遺伝子により構成される仮想の遺伝子クラスタ単位の発現変動情報として合算して、仮想の各遺伝子クラスタをスコアリングし、この仮想の遺伝子クラスタの中から、有用遺伝子を含む遺伝子クラスタ及び該クラスタに含まれる有用遺伝子を見いだすことにより、上記従来の有用遺伝子探索法に比べ、遙かに正確かつ効率的に有用遺伝子の探索、同定が可能となることを見だし、本発明を完成するに至った。すなわち、本発明は、以下のとおりである。

[0013] 1) 本発明は、以下の有用遺伝子の探索、同定を行う方法を提供する。

(1) 生物ゲノム中の標的遺伝子を含む遺伝子クラスタ及び／または該遺伝子クラスタ中の標的遺伝子を探索する方法であって、生物細胞の生理状態変化を生じる条件とコントロール条件下において生じたゲノム遺伝子の発現量変動比を、ゲノムDNA上に配列する複数の遺伝子により構成される仮想の遺伝子クラスタ単位の発現量変動比として合算することにより、仮想の遺伝子クラスタ単位毎にスコアリングし、得られたスコアに基づき、上記生理状態変化の原因遺伝子である標的遺伝子を含む遺伝子クラスタ及び／または該遺伝子クラスタ中の標的遺伝子を探索する方法。

(2) 生物細胞の生理状態変化を生じる条件とコントロール条件下とを一の対比条件セットとして、該対比条件セットが一種以上設定されていることを特徴とする上記(1)に記載の方法。

(3) 生理状態変化を生じる条件とコントロール条件が、少なくとも代謝産物の産生誘導条件下と非誘導条件下あるいは代謝産物の産生抑制条件下と非抑制条件下との対比条件セットを含むことを特徴とする上記(1)または(2)に記載の方法。

(4) 代謝物産生に関与する遺伝子が2次代謝物産生に関与する遺伝子であ

ることを特徴とする上記（３）に記載の方法。

（５）仮想の各遺伝子クラスタは、ゲノムDNA上に連続する遺伝子を２個から遺伝子数一つずつ増やして、想定される遺伝子クラスタに含まれる最大限のゲノム遺伝子数になるまで抽出し、かつ該抽出において、抽出する遺伝子の各個数毎に、直鎖状DNAからなるゲノムの場合には該DNAのいずれかの末端から、環状DNAからなるゲノムの場合には任意の遺伝子を起点として順にゲノムDNA上に配列する遺伝子一つずつずらしながら抽出された各遺伝子群からなることを特徴とする、上記（１）～（４）のいずれかに記載の方法。

（６）スコアリングされる仮想の各遺伝子クラスタの集合体が、ゲノムDNA上に連続する遺伝子を２個から遺伝子数一つずつ増やし想定される遺伝子クラスタに含まれる最大限のゲノム遺伝子数になるまで抽出し、かつ該抽出において、抽出する遺伝子の各個数毎に、直鎖状DNAからなるゲノムの場合には該DNAのいずれかの末端から、環状DNAからなるゲノムの場合には任意の遺伝子を起点として順にゲノムDNA上に配列する遺伝子一つずつずらしながら抽出された各遺伝子群からなる仮想の各遺伝子クラスタの集合体となり、ゲノム上に存在する遺伝子クラスタの全てが仮想の遺伝子クラスタの集合体中に含まれるように構成されていることを特徴とする上記（１）～（５）のいずれかに記載の方法。

（７）仮想の各遺伝子クラスタのスコアリングが以下の計算式a)によりなされることを特徴とする、上記（１）～（６）のいずれかに記載の方法。

計算式 a)

[数1]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

（式中、Mは仮想の各遺伝子クラスタのスコア、mはスコアリングされる仮想の各遺伝子クラスタに含まれる各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比（m値）の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比（m値）の標準偏差を表す。）

(8) ゲノムDNA上に配列する遺伝子が、標的とする遺伝子機能を有すると推定される場合、あるいは標的とする遺伝子機能を有する可能性が低いか若しくはその可能性がないと推定される場合において、当該ゲノムDNA上に配列する遺伝子については、以下の重み付け計算が適用されることを特徴とする、上記(7)の方法。

[数2]

$$w \frac{m - \bar{m}}{\sigma(m)}$$

(式中、 m は、上記標的とする遺伝子機能を有するか、あるいは標的とする遺伝子機能を有する可能性が低いか若しくはその可能性がないと推定されるゲノムDNA上の遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値) の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値) の標準偏差、 w は重みとしての任意の実数を、それぞれ表す。)

(9) ゲノムDNA上に配列する遺伝子が、標的とする遺伝子機能を有すると推定される場合において、標的とする遺伝子機能を有すると推定された遺伝子を含む仮想の遺伝子クラスタを選出し、選出された仮想の遺伝子クラスタについて、スコアリングすることを特徴とする、上記(7)に記載の方法。

(10) 仮想の遺伝子クラスタが、ゲノムにおいて、近傍に存在することを条件として、以下の1)～3)の内の1以上の遺伝子のみから、あるいは該遺伝子を少なくとも含む1以上の遺伝子から構築されることを特徴とする、上記(4)に記載の方法。

- 1) 2次代謝物産生に関与していると想定される酵素種に属する酵素遺伝子。
- 2) トランスポーター遺伝子
- 3) 転写因子をコードする遺伝子

(11) 仮想の各遺伝子クラスタのスコアリングが以下の計算式a) によりなされることを特徴とする、上記(10)に記載の方法。

計算式 a)

[数3]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、M は各仮想の遺伝子クラスタのスコア、m はスコアリングされる仮想の各遺伝子クラスタに含まれるアノテーション付与に基づき選定された各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれるアノテーション付与に基づき選定された全遺伝子の発現量変動比 (m 値) の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれるアノテーション付与に基づき選定された全遺伝子の発現量変動比 (m 値) の標準偏差を表す。)

(12) 仮想の遺伝子クラスタ全体のスコアの分布から乖離して存在するスコアを有する仮想の遺伝子クラスタを、標的の遺伝子クラスタ候補として選定することを特徴とする、上記 (1) ~ (11) のいずれかに記載の方法。

(13) 仮想の遺伝子クラスタ全体のスコアの分布からの乖離の程度を示す判定値 $I(\chi)$ を、以下の計算式 b) により算出し、算出された該判定値 $I(\chi)$ に基づき仮想の遺伝子クラスタを標的の遺伝子クラスタ候補として選定することを特徴とする、上記 (12) に記載の方法。

計算式 b)

[数4]

$$\chi = -M \log P$$

(式中、 χ は上記仮想の各遺伝子クラスタの乖離の程度を示す判定値 I 、M は仮想の各遺伝子クラスタのスコア、P は、全ての仮想の遺伝子クラスタについての各スコアに対する出現頻度分布において、各スコア M の出現頻度について全体累計を 1 としたときの、上記スコア M の出現頻度の割合を表す。)

(14) 仮想の遺伝子クラスタ全体のスコアの分布からの乖離の程度を示す判定値 $I(v)$ を、以下の計算式 c) により算出し、算出された判定値 $I(v)$ に基づき仮想の遺伝子クラスタを、標的の遺伝子クラスタ候補として選定することを特徴とする、上記 (12) に記載の方法。

計算式 c)

[数5]

$$v = (M - \overline{M})^{d'} / (a\sigma(M))^{d'}$$

(式中、 v は上記仮想の各遺伝子クラスタの乖離の程度を示す判定値 $||$ 、 M は仮想の各遺伝子クラスタのスコア、 $\overline{M}$ は全ての仮想の遺伝子クラスタのスコア (M 値) の平均値、 $\sigma(M)$ は全ての仮想の遺伝子クラスタにおけるスコア (M 値) の標準偏差、 a は任意の正の実数、 d' は次元数となる正の偶数を表す。)

(15) さらに、以下の計算式 $d)$ の算出結果に基づき、 b が 100 未満の仮想のクラスタを少なくとも除外し、標的の遺伝子クラスタ候補をさらに絞り込むことを特徴とする、上記 (13) または (14) に記載の方法。

計算式 $d)$

[数6]

$$\chi \times v > b$$

(式中、 χ は上記 (13) に記載の計算式 $b)$ によって算出した仮想の各遺伝子クラスタの判定値 $||$ 、 v は上記 (14) に記載の計算式 $c)$ によって算出した仮想の各遺伝子クラスタの判定値 $||$ 、 b は閾値としての任意の正の実数を表す。)

(16) 生物細胞の生理状態変化を生じる条件とコントロール条件下において生じたゲノム DNA 上に配列する各遺伝子の発現量変動比を、ゲノム DNA 上に配列する複数遺伝子により構成される仮想の遺伝子クラスタ単位の発現量変動比として合算することにより、仮想の遺伝子クラスタ単位毎にスコアリングし、得られたスコアに基づき、標的とする遺伝子クラスタがゲノム中に存在するか否かあるいは、標的遺伝子クラスタが存在する場合の遺伝子サイズを予測する方法であって、

ゲノム DNA 上に連続する遺伝子を 2 個から遺伝子数を一つずつ増やし想定される遺伝子クラスタに含まれる最大限のゲノム遺伝子数になるまで抽出し、かつ該抽出において抽出する遺伝子の各個数毎に、直鎖状 DNA からなるゲノムの場合には該 DNA のいずれかの末端から、あるいは環状 DNA からなるゲノムの場合には任意の遺伝子を起点として順にゲノム DNA 上に配列する遺伝子の一つずつずらしながら抽出された各遺伝子群から構成された

仮想の各遺伝子クラスタを、以下の計算式 a) によりスコアリングし、この得られた仮想の各遺伝子クラスタのスコアを各遺伝子クラスタに含まれる遺伝子数毎に分け、以下の計算式 e) により、各遺伝子数単位毎に遺伝子クラスタスコア分布判定値 (ε) を求め、該判定値に基づき、予め、標的とする遺伝子クラスタがゲノム中に存在するか否かあるいは、標的クラスタが存在する場合のその遺伝子サイズを予測することを特徴とする、上記方法。

計算式 a)

[数1]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、 M は仮想の各遺伝子クラスタのスコア、 m はスコアリングされる仮想の各遺伝子クラスタに含まれる各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値) の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値) の標準偏差を表す。)

計算式 e)

[数7]

$$\varepsilon = \sum (M - \bar{M})^d / n \sigma(M)^d$$

(式中、 ε は各遺伝子数単位毎に求められた遺伝子クラスタスコア分布判定値、 M は全ての仮想の遺伝子クラスタを遺伝子数単位毎に分けた場合の、各遺伝子数単位に含まれる仮想の各遺伝子クラスタのスコア、 $\bar{M}$ は全ての仮想の遺伝子クラスタのスコアの平均、 n は仮想の遺伝子クラスタの総数、 $\sigma(M)$ は全ての仮想の遺伝子クラスタのスコア (M 値) の標準偏差、 d は任意に設定する次元数としての正の偶数を表す。)

(17) 遺伝子数が k 個のときの ε 値 ($\varepsilon(k)$) と、その前後数のときの ε 値 ($\varepsilon(k-1)$ 、 $\varepsilon(k+1)$) が、以下の関係にあるとき、標的とする遺伝子クラスタがゲノム中に存在すると判定し、標的遺伝子クラスタに含まれる遺伝子数を k 個と予想することを特徴とする、上記 (16) に記載の方法。

[数8]

$$\varepsilon(k) > \varepsilon(k-1) \quad \text{かつ} \quad \varepsilon(k) > \varepsilon(k+1)$$

2) 本発明は、また、以下の有用遺伝子の探索、同定を行うための装置、およびそのためのプログラムを提供する。

(18) 生物ゲノム中の標的遺伝子を含む遺伝子クラスタ及び／または該遺伝子クラスタ中の標的遺伝子を探索する装置であって、a) 生物細胞の生理状態変化を生じる条件とコントロール条件下におけるゲノムDNA上に配列する各遺伝子の発現量データに基づき算出された上記2つの条件下における上記各遺伝子の発現量変動比を記憶する手段、b) ゲノムDNA上に配列する複数の遺伝子を組み合わせて仮想の遺伝子クラスタを構築する手段、c) 該算出され、記憶されたゲノムDNA上に配列する各遺伝子の発現量変動比を複数の遺伝子により構築された上記仮想の遺伝子クラスタ単位の発現量変動比として合算し、仮想の遺伝子クラスタ単位毎にスコアリングし、仮想の各遺伝子クラスタのスコアを記憶する手段、及びd) 得られたスコアに基づき上記生理状態変化の原因遺伝子である標的遺伝子を含む遺伝子クラスタを選定する手段を有するか、あるいはさらにe) 選定された遺伝子クラスタに含まれる遺伝子を表示する手段を有することを特徴とする、上記装置。

(19) 発現量データが、遺伝子発現量測定用DNAマイクロアレイによる蛍光強度情報であることを特徴とする上記(18)に記載の装置。

(20) 蛍光強度情報が、蛍光強度を読み取り、数値化する手段を有する蛍光強度読み取り装置により出力される数値データであることを特徴とする、上記(19)に記載の装置。

(21) 生物細胞の生理状態変化を生じる条件とコントロール条件とを1の対比条件セットとして1以上設定されている場合において、各対比条件セットに含まれる条件毎に各遺伝子の発現量データが入力され、各対比条件セットにおける同一遺伝子の発現量変動比が算出されることを特徴とする、上記(18)～(20)のいずれかに記載の装置。

(22) 標的遺伝子が代謝物産生に関与する遺伝子であることを特徴とする、上記(18)～(21)のいずれかに記載の装置。

(23) 代謝物産生に関与する遺伝子が2次代謝物産生に関与する遺伝子であることを特徴とする、上記(22)に記載の装置。

(24) 設定される対比条件セットが、少なくとも代謝産物の産生誘導条件下と非誘導条件下あるいは代謝産物の産生抑制条件下と非抑制条件下との対比条件セットを含むことを特徴とする上記(22)に記載の装置。

(25) 代謝産物が2次代謝産物であることを特徴とする、上記(24)に記載の装置。

(26) 仮想の各遺伝子クラスタの構築手段が、ゲノムDNA上に連続する遺伝子を2個から遺伝子数を一つずつ増やして、想定される遺伝子クラスタに含まれる最大限のゲノム遺伝子数になるまで抽出し、かつ該抽出において、抽出する遺伝子の各個数毎に、直鎖状DNAからなるゲノムの場合には該DNAのいずれかの末端から、環状DNAからなるゲノムの場合には任意の遺伝子を起点として順にゲノムDNA上に配列する遺伝子一つずつずらしながら抽出した各遺伝子群により構築する手段であることを特徴とする、上記(18)～(25)のいずれかに記載の装置。

(27) 仮想の各遺伝子クラスタのスコアリングが以下の計算式a)によりなされることを特徴とする、上記(18)～(26)のいずれかに記載の装置。

計算式 a)

[数1]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、Mは仮想の各遺伝子クラスタのスコア、mはスコアリングされる仮想の各遺伝子クラスタに含まれる各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の標準偏差を表す。)

(28) ゲノムDNA上に配列する各遺伝子中の特定の遺伝子を選定するためのアノテーション付与手段を有し、上記遺伝子クラスタのスコアリングにおいて、付与されたアノテーションに基づき選定された遺伝子についての発現量変動比計算を以下の重み付け計算式により行うことを特徴とする、上記(27)に記載の装置。

[数2]

$$w \frac{m - \bar{m}}{\sigma(m)}$$

(式中、 m は、上記標的とする遺伝子機能を有するか、あるいは標的とする遺伝子機能を有する可能性が低い若しくはその可能性がないと推定されるゲノムDNA上の遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値) の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値) の標準偏差、 w は重みとしての任意の実数を、それぞれ表す。)

(29) アノテーション付与手段が、それぞれ遺伝子機能の種類毎に異なるアノテーションを付与する手段であることを特徴とする上記(28)に記載の装置。

(30) アノテーションに基づき選定される遺伝子が、1) ~ 3) のうちの1以上の遺伝子であることを特徴とする、上記(29)に記載の装置

1) 2次代謝物産生に関与していると想定される酵素種に属する酵素遺伝子。

2) トランスポーター遺伝子

3) 転写因子をコードする遺伝子

(31) 上記(28) ~ (30) のいずれかに記載のアノテーション付与手段と、構築された仮想の遺伝子クラスタから、アノテーションに基づき選出された遺伝子を含む仮想の遺伝子クラスタを選出する手段を有し、選出された仮想の遺伝子クラスタについてスコアリングすることを特徴とする、上記(27)に記載の装置。

(32) ゲノムDNA上に配列する各遺伝子中の特定遺伝子を選定するため

のアノテーション付与手段を有し、ゲノムDNA上において近傍に位置することを条件として、アノテーションに基づき選定された遺伝子により、あるいは該遺伝子を少なくとも含む1以上の遺伝子から仮想の遺伝子クラスタを構築する手段を有することを特徴とする、上記(18)～(25)に記載の装置。

(33) 上記(32)に記載のアノテーション付与手段が、それぞれ遺伝子機能の種類に応じたアノテーションを付与する手段であることを特徴とする上記(32)に記載の装置。

(34) アノテーション付与に基づき選定される遺伝子が、1)～3)のうちの1以上の遺伝子であることを特徴とする、上記(33)に記載の装置

1) 2次代謝物産生に関与していると想定される酵素種に属する酵素遺伝子。

2) トランスポーター遺伝子

3) 転写因子をコードする遺伝子

(35) 仮想の各遺伝子クラスタのスコアリングが以下の計算式a)によりなされることを特徴とする、上記(32)～(34)のいずれかに記載の装置。

計算式 a)

[数3]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、Mは各仮想の遺伝子クラスタのスコア、mはスコアリングされる仮想の各遺伝子クラスタに含まれるアノテーション付与に基づき選定された各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれるアノテーション付与に基づき選定された全遺伝子の発現量変動比(m値)の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれるアノテーション付与に基づき選定された全遺伝子の発現量変動比(m値)の標準偏差を表す。)

(36) 仮想の遺伝子クラスタ全体のスコアの分布から乖離して存在するスコアを有する仮想の遺伝子クラスタを、標的の遺伝子クラスタ候補として選

定する手段を有することを特徴とする、上記（１８）～（３５）のいずれかに記載の装置。

（３７）標的の遺伝子クラスタ候補として選定する手段として、仮想の遺伝子クラスタ全体のスコアの分布からの乖離の程度を示す判定値 χ を、以下の計算式 b) により算出するプログラムが格納されていることを特徴とする、上記（３６）に記載の装置。

計算式 b)

[数4]

$$\chi = -M \log P$$

（式中、 χ は上記仮想の各遺伝子クラスタの乖離の程度を示す判定値、 M は仮想の各遺伝子クラスタのスコア、 P は、全ての仮想の遺伝子クラスタについての各スコアに対する出現頻度分布において、各スコア M の出現頻度について全体累計を１としたときの、上記スコア M の出現頻度の割合を表す。）

（３８）標的の遺伝子クラスタ候補として選定する手段として、遺伝子クラスタ全体のスコアの分布からの乖離の程度を示す判定値 ν を、以下の計算式 c) により算出するプログラムが格納されていることを特徴とする、上記（３６）に記載の装置。

計算式 c)

[数5]

$$\nu = (M - \overline{M})^{d'} / (a \sigma(M))^{d'}$$

（式中、 ν は上記仮想の各遺伝子クラスタの乖離の程度を示す判定値、 M は仮想の各遺伝子クラスタのスコア、 $\overline{M}$ は全ての仮想の遺伝子クラスタのスコア（ M 値）の平均値、 $\sigma(M)$ は全ての仮想の遺伝子クラスタにおけるスコア（ M 値）の標準偏差、 a は任意の正の実数、 d' は次元数となる正の偶数を表す。）

（３９）さらに、以下の計算式 d) の算出結果に基づき、 b が 100 未満の仮想のクラスタを少なくとも除外し、標的の遺伝子クラスタ候補をさらに絞り込むプログラムが格納されていることを特徴とする、上記（３７）または（３８）に記載の装置。

計算式 d)

[数9]

$$\chi \times v > b$$

(式中、 χ は上記(37)に記載の計算式b)によって算出した仮想の各遺伝子クラスタの判定値I、 v は上記(38)に記載の計算式c)によって算出した仮想の各遺伝子クラスタの判定値II、 b は閾値としての任意の正の実数を表す。)

(40) a) 生物細胞の生理状態変化を生じる条件とコントロール条件下において生じたゲノムDNA上に配列する各遺伝子の発現量を入力する手段、
b) 入力された上記2つの条件下における同一遺伝子の発現量の比を算出する発現量変動比算出手段、c) 該算出されたゲノムDNA上に配列する各遺伝子の発現量変動比を複数の遺伝子により構築された上記仮想の遺伝子クラスタ単位の発現量変動比として合算し、仮想の遺伝子クラスタ単位毎にスコアリングする手段、及びd) 得られた仮想の遺伝子クラスタのスコアから遺伝子クラスタに含まれる遺伝子数単位毎の遺伝子クラスタ分布判定値(ε)を算出する手段を有し、該遺伝子クラスタ分布判定値(ε)から、標的とする遺伝子クラスタがゲノム中に存在する否かあるいは、標的遺伝子クラスタが存在する場合の遺伝子サイズを予測する装置であって、仮想の遺伝子クラスタの構築手段が、ゲノムDNA上に連続する遺伝子を2個から遺伝子数を一つずつ増やし想定される遺伝子クラスタに含まれる最大限のゲノム遺伝子数になるまで抽出し、かつ該抽出において抽出する遺伝子の各個数毎に、直鎖状DNAからなるゲノムの場合には該DNAのいずれかの末端から、あるいは環状DNAからなるゲノムの場合には任意の遺伝子を起点として順にゲノムDNA上に配列する遺伝子の一つずつずらしながら抽出された各遺伝子群を仮想の各遺伝子クラスタとする手段であり、上記仮想の遺伝子クラスタ単位のスコアリング手段は以下の計算式a)による演算手段からなるとともに、上記遺伝子クラスタ分布判定値(ε)の算出手段が、以下の計算式e)によるものであることを特徴とする、上記装置。

計算式 a)

[数1]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、 M は仮想の各遺伝子クラスタのスコア、 m はスコアリングされる仮想の各遺伝子クラスタに含まれる各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値) の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値) の標準偏差を表す。)

計算式 e)

[数7]

$$\varepsilon = \sum (M - \bar{M})^d / n \sigma(M)^d$$

(式中、 ε は各遺伝子数単位毎に求められた遺伝子クラスタスコア分布判定値、 M は全ての仮想の遺伝子クラスタを遺伝子数単位毎に分けた場合の、各遺伝子数単位に含まれる仮想の各遺伝子クラスタのスコア、 $\bar{M}$ は全ての仮想の遺伝子クラスタのスコアの平均、 n は仮想の遺伝子クラスタの総数、 $\sigma(M)$ は全ての仮想の遺伝子クラスタのスコア (M 値) の標準偏差、 d は任意に設定する次元数としての正の偶数を表す。)

(4 1) 遺伝子数が k 個のときの遺伝子クラスタ分布判定値 ε 値 ($\varepsilon(k)$) と、その前後数のときの同 ε 値 ($\varepsilon(k-1)$ 、 $\varepsilon(k+1)$) が、以下の関係にあるとき、標的とする遺伝子クラスタがゲノム中に存在すると判定し、標的遺伝子クラスタに含まれる遺伝子数を k 個とする予想値を出力することを特徴とする、上記 (4 0) に記載の装置。

[数8]

$$\varepsilon(k) > \varepsilon(k-1) \quad \text{かつ} \quad \varepsilon(k) > \varepsilon(k+1)$$

(4 2) 上記 (2 6) に記載の仮想の遺伝子クラスタの構築手段を実行するプログラムであって、ゲノム遺伝子の位置情報に基づき、以下の 1) または 2) の手段を実行することを特徴とする、仮想の遺伝子クラスタ構築プログラム。

1) ゲノム遺伝子が直鎖状ゲノムの場合、

a. ゲノムDNAの一方の末端に位置する遺伝子を起点として、他方の末端方向に、順次、ゲノムDNA上に連続する遺伝子を同一方向に2個から一つずつ増やして想定される遺伝子クラスタに含まれる遺伝子数の最大限になるまで組み合わせ、起点とした遺伝子を含み、かつ遺伝子の個数の異なる複数の遺伝子群を構成する手段。

b. 起点を、順次、他方の末端方向に一遺伝子ずつずらせながら、上記a.と同様の処理を行い、新たな起点遺伝子を含みかつ遺伝子の個数が異なる複数の遺伝子群を構成し、a.の遺伝子群と併せて、複数の遺伝子を組み合わせた遺伝子群からなる仮想の遺伝子クラスタを構築する手段。

2) ゲノム遺伝子が環状の場合、ゲノムDNA上の任意の遺伝子を起点として、上記1) a. 及び b. と同様の処理を順次行い、最初に起点とした遺伝子が起点となる時点で処理を終了する手段。

(43) 上記(42)のプログラムにより構築された仮想の遺伝子クラスタについて、以下の計算式a)によるスコアリングを実行することを特徴とする、仮想の遺伝子クラスタのスコアリングプログラム。

計算式a)

[数1]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、Mは仮想の各遺伝子クラスタのスコア、mはスコアリングされる仮想の各遺伝子クラスタに含まれる各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の標準偏差を表す。)

(44) 上記遺伝子クラスタのスコアリングにおいて、付与されたアノテーションに基づきゲノム遺伝子を選定し、選定された遺伝子についての発現量変動比計算を以下の重み付け計算式により行うことを特徴とする、上記(43)に記載のプログラム。

[数2]

$$w \frac{m - \bar{m}}{\sigma(m)}$$

(式中、 m は、上記標的とする遺伝子機能を有するか、あるいは標的とする遺伝子機能を有する可能性が低い若しくはその可能性がないと推定されるゲノム DNA 上の遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値) の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値) の標準偏差、 w は重みとしての任意の実数を、それぞれ表す。)

(4 5) 上記遺伝子クラスタのスコアリングにおいて、付与されたアノテーションに基づきゲノム遺伝子を選定し、構築された遺伝子クラスタの中から、該選定されたゲノム遺伝子を含む仮想の遺伝子クラスタを選出し、選出された仮想の遺伝子クラスタについてスコアリングを実行することを特徴とする、上記 (4 3) に記載のスコアプログラム。

(4 6) 上記 (3 2) に記載の仮想の遺伝子クラスタの構築手段を実行するプログラムであって、ゲノム DNA 上において近傍に位置することを条件として、アノテーションに基づき選定された遺伝子により、あるいは該遺伝子を少なくとも含む 1 以上の遺伝子から仮想の遺伝子クラスタを構築することを特徴とする、仮想の遺伝子クラスタの構築プログラム。

(4 7) 上記 (4 6) のプログラムにより構築された仮想の遺伝子クラスタについて、以下の計算式 a) によるスコアリングを実行することを特徴とする、仮想の遺伝子クラスタのスコアリングプログラム。

計算式 a)

[数3]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、M は各仮想の遺伝子クラスタのスコア、m はスコアリングされる仮想の各遺伝子クラスタに含まれるアノテーション付与に基づき選定された各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれるアノテーション付与に基づき選定された全遺伝子の発現量変動比 (m 値) の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれるアノテーション付与に基づき選定された全遺伝子の発現量変動比 (m 値) の標準偏差を表す。)

(48) 上記 (43) ~ (45) 又は (47) のいずれかに記載のスコアリングプログラムにより算出された各仮想の遺伝子クラスタのスコアについて、仮想の遺伝子クラスタ全体のスコアの分布からの乖離の程度を算出するプログラムであって、以下の計算式 b) により、判定値 I (χ) を算出することを特徴とする、上記プログラム。

計算式 b)

[数4]

$$\chi = -M \log P$$

(式中、 χ は上記仮想の各遺伝子クラスタの乖離の程度を示す判定値 I、M は仮想の各遺伝子クラスタのスコア、P は、全ての仮想の遺伝子クラスタについての各スコアに対する出現頻度分布において、各スコア M の出現頻度について全体累計を 1 としたときの、上記スコア M の出現頻度の割合を表す。)

(49) 上記 (43) ~ (45) 又は (47) のいずれかに記載のスコアリングプログラムにより算出された各仮想の遺伝子クラスタのスコアについて、仮想の遺伝子クラスタ全体のスコア分布からの乖離の程度を算出するプログラムであって、以下の計算式 c) により判定値 II (ψ) の算出を実行する、上記プログラム。

計算式 c)

[数5]

$$v = (M - \overline{M})^{d'} / (a\sigma(M))^{d'}$$

(式中、 v は上記仮想の各遺伝子クラスタの乖離の程度を示す判定値、 M は仮想の各遺伝子クラスタのスコア、 $\overline{M}$ は全ての仮想の遺伝子クラスタのスコア (M 値) の平均値、 $\sigma(M)$ は全ての仮想の遺伝子クラスタにおけるスコア (M 値) の標準偏差、 a は任意の正の実数、 d' は次元数となる正の偶数を表す。)

(50) 生物細胞の生理状態変化を生じる条件とコントロール条件下におけるゲノムDNA上に配列する各遺伝子の発現量変動比を複数の遺伝子により構築された上記仮想の遺伝子クラスタ単位の発現量変動比として合算し、仮想の遺伝子クラスタ単位毎にスコアリングする手段、及び得られた仮想の遺伝子クラスタのスコアから遺伝子クラスタに含まれる遺伝子数単位毎の遺伝子クラスタ分布判定値 (ε) を算出し、該遺伝子クラスタ分布判定値 (ε) から、標的とする遺伝子クラスタがゲノム中に存在する否かあるいは、標的遺伝子クラスタが存在する場合の遺伝子サイズを予測する手段に用いるプログラムであって、

少なくとも以下 (A) ~ (C) の手段を実行するプログラム。

(A) ゲノム遺伝子の位置情報に基づき、以下の1) または2) の手段により仮想の遺伝子クラスタを構築する手段、

1) ゲノム遺伝子が直鎖状の場合、

a. ゲノムDNAの一方の末端に位置する遺伝子を起点として、他方の末端方向に、順次、ゲノムDNA上に連続する遺伝子を同一方向に2個から一つずつ増やして想定される遺伝子クラスタに含まれる遺伝子数の最大限になるまで組み合わせ、起点とした遺伝子を含み、かつ遺伝子の個数の異なる複数の遺伝子群を構成する手段。

b. 起点を、順次、他方の末端方向に遺伝子一つずつずらせながら、上記 a. と同様の処理を行い、新たな起点遺伝子を含みかつ遺伝子の個数が異なる複数の遺伝子群を構成し、a. の遺伝子群と併せて、複数の遺伝子の組み合わせた遺伝子群からなる仮想の遺伝子クラスタを構築する手段。

2) ゲノム遺伝子が環状の場合、ゲノムDNA上の任意の遺伝子を起点として、上記1) a. 及びb. と同様の処理を順次行い、最初に起点とした遺伝子が起点となる時点で処理を終了する手段。

(B) 上記(A)の手段により構築された仮想の遺伝子クラスタについて、以下の計算式a)により仮想の遺伝子クラスタ単位毎にスコアリングする手段。

計算式a)

[数1]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、Mは仮想の各遺伝子クラスタのスコア、mはスコアリングされる仮想の各遺伝子クラスタに含まれる各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の標準偏差を表す。)

(C) 上記(B)の手段により得られた仮想の遺伝子クラスタのスコアから、以下の計算式e)により仮想の遺伝子クラスタに含まれる遺伝子数単位毎の遺伝子クラスタ分布判定値(ε)を算出する手段。

計算式e)

[数7]

$$\varepsilon = \sum (M - \bar{M})^d / n \sigma(M)^d$$

(式中、 ε は各遺伝子数単位毎に求められた遺伝子クラスタスコア分布判定値、Mは全ての仮想の遺伝子クラスタを遺伝子数単位毎に分けた場合の、各遺伝子数単位に含まれる仮想の各遺伝子クラスタのスコア、 $\bar{M}$ は全ての仮想の遺伝子クラスタのスコアの平均、nは仮想の遺伝子クラスタの総数、 $\sigma(M)$ は全ての仮想の遺伝子クラスタのスコア(M値)の標準偏差、dは任意に設定する次元数としての正の偶数を表す。)

(51) 遺伝子数がk個のときの遺伝子クラスタ分布判定値 ε 値($\varepsilon(k)$)と、その前後数のときの同 ε 値($\varepsilon(k-1)$ 、 $\varepsilon(k+1)$)が、以下の関係にあるとき、標的とする遺伝子クラスタがゲノム中に存在すると判定

し、標的遺伝子クラスタに含まれる遺伝子数を k 個とする予想値を出力することを特徴とする、上記（50）に記載のプログラム。

[数8]

$$\varepsilon(k) > \varepsilon(k-1) \quad \text{かつ} \quad \varepsilon(k) > \varepsilon(k+1)$$

発明の効果

[0014] 従来技術では主にDNAマイクロアレイなどを用いることにより、例えば、代謝産物の産生に関与する遺伝子を探索する場合、着目する化合物が産生される、あるいは着目する活性が観察される条件における発現の誘導や強い発現強度を有することなどを指標として標的となる遺伝子の同定を行ってきた。しかし、生物情報特有のデータの曖昧性、誤り、複雑性などによって正しい遺伝子を高い精度で予測することは極めて困難であった。これに対して、本発明の遺伝子探索方法および装置は、隣接あるいは近傍に位置する複数の遺伝子から、仮想の遺伝子クラスタを構成し、この仮想の遺伝子クラスタをまず探索対象として、有用遺伝子の探索を行うものであり、その手法自体極めて論理的、機械的であり、従来のDNAマイクロアレイによる解析にみられるような研究者の知識や経験等に大きく依存することなく、コンピューターを使用して、迅速、正確に有用遺伝子を特定することが可能となるとともに、同時に該遺伝子を含む遺伝子クラスタも特定できる。

[0015] 一方、本発明の遺伝子探索法においては、探索条件に誤りがある場合には、取得されたデータ自体から把握でき、この場合には探索条件を再設定し、探索をやり直すことができる。これに対し、従来法においては、解析結果が誤りであるか否かの判断には、遺伝子破壊実験等の検証実験を必要とし、膨大な費用、手間をかけざるを得ない。したがって、本発明の遺伝子探索法および探索装置の有利性は明らかである。

[0016] また、本発明の遺伝子探索法および装置は、代謝産物産生遺伝子、とりわけ従来困難であった2次代謝物産生遺伝子の探索に極めて適している。これは、2次代謝物質の産生に関与する遺伝子が遺伝子クラスタを構成している

ことが多いからである。さらに、このようにして探索され、特定された2次代謝物質産生遺伝子等の有用遺伝子等の配列情報を利用すれば、新たな類似遺伝子の取得も可能となる。しかし、本発明の遺伝子探索法および装置によれば、このような代謝産物産生遺伝子の探索のみではなく、広く普遍性を有し、生物の様々な生理状態変化をもたらす原因遺伝子、そのみならず、同時に生理状態変化に関与する遺伝子クラスタも探索可能であり、これにより該原因遺伝子と協働する他の遺伝子も特定することが可能となる。したがって、本発明は、例えば、代謝産物、特に2次代謝物質産生遺伝子、様々な疾病の原因遺伝子、あるいはこれらと共働する遺伝子等の探索に極めて有効であり、新たな有用な化合物の取得、その大量生産あるいは医薬品開発等において、その技術を飛躍的に向上させることができる。

図面の簡単な説明

[0017] [図1]本発明の遺伝子クラスタ及び遺伝子探索法のフローチャートを示す図であり、本発明の手法における解析の流れが示されている。

[図2]本発明装置の概要を示すブロック図である。

[図3]本発明装置における、仮想の遺伝子クラスタを構築する手段のフローチャートを示す図である。

[図4]本発明装置における、仮想の遺伝子クラスタのスコアリング手段のフローチャート示す図である。

[図5]本発明装置における、各遺伝子に付与された該当機能に関するアノテーションに基づいた仮想の遺伝子クラスタの（a）重み付けスコアリングあるいは（b）選出およびスコアリング手段のフローチャートを示す図である。

[図6]本発明装置における、該当機能に関するアノテーションに基づき選定された遺伝子を用いて仮想の遺伝子クラスタを構築する手段のフローチャートを示す図である。

[図7]本発明装置における、仮想の遺伝子クラスタを、スコアの全体の分布からの乖離度を判定する値に基づいて選定する手段のフローチャートを示す図である。

[図8]本発明装置における、仮想の遺伝子クラスタのスコア乖離度判定値から、該当遺伝子クラスタの候補を絞り込む手段のフローチャートを示す図である。

[図9]本発明装置に含まれる、用いる遺伝子発現量変動比データに、標的とする遺伝子クラスタが含まれるか否か、およびそのときの遺伝子クラスタサイズを予測する手段のフローチャートを示す図である。

[図10]遺伝子クラスタスコア分布判定値 ε の挙動を示す例である。

[図11]アスペルギルス・オリゼにおけるコウジ酸産生に必須の3つの遺伝子について、アレイデータの系C 1におけるスコア m 値の全遺伝子中の順位を示した図である。

[図12]アスペルギルス・オリゼにおけるコウジ酸産生に必須の3つの遺伝子について、アレイデータの系C 2におけるスコア m 値の全遺伝子中の順位を示した図である。

[図13]アスペルギルス・オリゼにおけるコウジ酸産生に必須の3つの遺伝子について、アレイデータの系C 3におけるスコア m 値の全遺伝子中の順位を示した図である。

[図14]アスペルギルス・オリゼにおける仮想の遺伝子クラスタの、クラスタサイズを1から30までとったスコアヒストグラムを示した図である。（右）条件、 n_{cl} を変えたときの全体図。行：クラスタサイズ $n_{cl} = 1 \sim 30$ （上から）、列：左から、系C 1、C 2、C 3。（左）系C 2、 $n_{cl} = 5$ のときの拡大図。横軸：発現変動比スコア M 値、縦軸：頻度。

[図15]アスペルギルス・オリゼのアレイデータにおいて、標的とする遺伝子クラスタが含まれるか否かを判定する遺伝子クラスタスコア分布判定値 ε を示した図である。横軸：クラスタサイズ、縦軸：次元数6における ε 値。

[図16]アスペルギルス・オリゼのアレイデータ取得系C 2における、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判定する判定値 χ を示した図である。横軸：クラスタサイズ、縦軸： χ 。コウジ酸産生関連遺伝子の3つを要素に持つ遺伝子クラスタが、 $n_{cl} = 3$ のときに極大か

つ最大値を持つ。

[図17]アスペルギルス・オリゼのアレイデータ取得系C 2における、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判定する判定値 v を示した図である。横軸：クラスタサイズ、縦軸： v 。次元数 d' は2を採用した。コウジ酸産生関連遺伝子の3つを要素に持つ遺伝子クラスタが、 $ncl = 3$ のときに極大かつ最大値を持つ。

[図18]アスペルギルス・オリゼのアレイデータ取得系C 2における、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判断する評価値 $\chi \times v$ を示した図である。横軸：クラスタサイズ、縦軸： $\chi \times v$ 。次元数 d' は2を採用した。コウジ酸産生関連遺伝子の3つを要素に持つ遺伝子クラスタが、 $ncl = 3$ のときに極大かつ最大値を持つ。

[図19]アスペルギルス・オリゼにおける仮想の遺伝子クラスタの、機能注釈に応じてスコアに重み付けを行った後、クラスタサイズを1から30までとったスコアヒストグラムを示した図である。（右）条件、 ncl を変えたときの全体図。行：クラスタサイズ $ncl = 1 \sim 30$ （上から）、列：左から、系C 1、C 2、C 3。（左）系C 2、 $ncl = 5$ のときの拡大図。横軸：発現変動比スコアM値、縦軸：頻度。

[図20]アスペルギルス・オリゼのアレイデータにおいて、機能の注釈による重み付け後、標的とする遺伝子クラスタが含まれるか否かを判定する遺伝子クラスタスコア分布判定値 ε を示した図である。横軸：クラスタサイズ、縦軸：次元数6における ε 値。

[図21]アスペルギルス・オリゼのアレイデータ取得系C 2における、機能の注釈による重み付け後の、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判定する判定値 χ を示した図である。横軸：クラスタサイズ、縦軸： χ 。コウジ酸産生関連遺伝子の3つを要素に持つ遺伝子クラスタが、 $ncl = 3$ のときに極大かつ最大値を持つ。

[図22]アスペルギルス・オリゼのアレイデータ取得系C 2における、機能の注釈による重み付け後の、仮想の各遺伝子クラスタが標的とする遺伝子クラ

スタであるか否かを判定する判定値 v を示した図である。横軸：クラスタサイズ、縦軸： v 。次元数 d' は2を採用した。コウジ酸産生関連遺伝子の3つを要素に持つ遺伝子クラスタが、 $ncl = 3$ のときに極大かつ最大値を持つ。

[図23]アスペルギルス・オリゼのアレイデータ取得系C2における、機能の注釈による重み付け後の、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判断する評価値 $\chi \times v$ を示した図である。横軸：クラスタサイズ、縦軸： $\chi \times v$ 。次元数 d' は2を採用した。コウジ酸産生関連遺伝子の3つを要素に持つ遺伝子クラスタが、 $ncl = 3$ のときに極大かつ最大値を持つ。

[図24]アスペルギルス・オリゼの全ゲノム遺伝子について、クラスタサイズを5としたときに、その仮想の遺伝子クラスタのうち、遺伝子の推定機能注釈に基づいて、該当する機能の注釈を持つ遺伝子が含まれる要素数を記載したベン図である。

[図25]アスペルギルス・オリゼにおける仮想の遺伝子クラスタのスコアM値分布中、クラスタサイズを5としたとき、機能の注釈による重み付けを行うことでコウジ酸産生遺伝子クラスタの順位がどう変化するかを示した図である。(a)全ての仮想の遺伝子クラスタ、(b)膜輸送体、転写制御因子、酸化還元酵素の全てを含む仮想の遺伝子クラスタ。

[図26]アスペルギルス・オリゼにおける仮想の遺伝子クラスタのスコアM値分布中、クラスタサイズを5としたとき、機能の注釈による重み付けの対象を膜輸送体および転写制御因子の2つとしたときに、コウジ酸産生遺伝子クラスタの順位がどこにあるかを示した図である。

[図27]アスペルギルス・オリゼにおける仮想の遺伝子クラスタのスコアM値分布中、クラスタサイズを5としたとき、機能の注釈から1つのキーワードを除いた(膜輸送体を含むが転写制御因子を含まない)ときのスコア分布を示した図である。

[図28]アスペルギルス・フラバスにおける仮想の遺伝子クラスタの、クラス

タサイズを1から30までとったスコアヒストグラムを示した図である。(右)条件、 ncl を変えたときの全体図。行：クラスタサイズ $ncl=1\sim 30$ (上から)、列：左から、系C1、C2、C3。(左)系C2、 $ncl=5$ のときの拡大図。横軸：発現変動比スコアM値、縦軸：頻度。

[図29]アスペルギルス・フラバスのアレイデータにおいて、標的とする遺伝子クラスタが含まれるか否かを判定する遺伝子クラスタスコア分布判定値 ε を示した図である。横軸：クラスタサイズ、縦軸：次元数6における ε 値。

[図30]アスペルギルス・フラバスのアレイデータ取得系C2における、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判定する判定値 χ を示した図である。横軸：クラスタサイズ、縦軸： χ 。

[図31]アスペルギルス・フラバスのアレイデータ取得系C2における、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判定する判定値 ψ を示した図である。横軸：クラスタサイズ、縦軸： ψ 。次元数 d' は2を採用した。

[図32]アスペルギルス・フラバスのアレイデータ取得系C2における、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判断する評価値 $\chi \times \psi$ を示した図である。横軸：クラスタサイズ、縦軸： $\chi \times \psi$ 。次元数 d' は2を採用した。

[図33]アスペルギルス・ニガーにおける仮想の遺伝子クラスタの、クラスタサイズを1から30までとったスコアヒストグラムを示した図である。(右)条件、 ncl を変えたときの全体図。行：クラスタサイズ $ncl=1\sim 30$ (上から)、列：左から、系C1、C2。(左)系C2、 $ncl=5$ のときの拡大図。横軸：発現変動比スコアM値、縦軸：頻度。

[図34]アスペルギルス・ニガーのアレイデータにおいて、標的とする遺伝子クラスタが含まれるか否かを判定する遺伝子クラスタスコア分布判定値 ε を示した図である。横軸：クラスタサイズ、縦軸：次元数6における ε 値。

[図35]アスペルギルス・ニガーのアレイデータ取得系C1およびC2における、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判

定する判定値 χ を示した図である。横軸：クラスタサイズ、縦軸： χ 。(a) C 1、(b) C 2。

[図36]アスペルギルス・ニガーのアレイデータ取得系C 1およびC 2における、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判定する判定値 v を示した図である。横軸：クラスタサイズ、縦軸： v 。次元数 d' は2を採用した。(a) C 1、(b) C 2。

[図37]アスペルギルス・ニガーのアレイデータ取得系C 1およびC 2における、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判断する評価値 $\chi \times v$ を示した図である。横軸：クラスタサイズ、縦軸： $\chi \times v$ 。次元数 d' は2を採用した。(a) C 1、(b) C 2。

[図38]アスペルギルス・オリゼのアレイデータ取得系C 2において、該当する機能の注釈を含む遺伝子を含むように構築した仮想の遺伝子クラスタが、標的とする遺伝子クラスタであるか否かを判定する判定値 χ を示した図である。横軸：クラスタサイズ、縦軸： χ 。

[図39]アスペルギルス・オリゼのアレイデータ取得系C 2において、該当する機能の注釈を含む遺伝子を含むように構築した仮想の遺伝子クラスタが、標的とする遺伝子クラスタであるか否かを判定する判定値 v を示した図である。横軸：クラスタサイズ、縦軸： v 。次元数 d' は2を採用した。

[図40]アスペルギルス・オリゼのアレイデータ取得系C 2において、該当する機能の注釈を含む遺伝子を含むように構築した仮想の遺伝子クラスタが、標的とする遺伝子クラスタであるか否かを判断する評価値 $\chi \times v$ を示した図である。横軸：クラスタサイズ、縦軸： $\chi \times v$ 。次元数 d' は2を採用した。

[図41]アスペルギルス・オリゼのアレイデータ取得系C 2において、該当する機能の注釈を含む遺伝子を含むように構築した仮想の遺伝子クラスタが、標的とする遺伝子クラスタであるか否かを判断する評価値 $\chi \times v$ を、横軸を仮想の遺伝子クラスタ番号として示した図である。横軸：仮想の遺伝子クラスタID、縦軸： $\chi \times v$ 。次元数 d' は2を採用した。

[図42]フザリウム・バーティシリオイデスにおける仮想の遺伝子クラスタの

、クラスタサイズを1から30までとったスコアヒストグラムを示した図である。
 (右) 系C 1、C 2における、 ncl を変えたときの全体図。行：クラスタサイズ $ncl = 1 \sim 30$ (上から)、列：左から、系C 1、C 2。(左)
 系C 2、 $ncl = 14$ のときの拡大図。横軸：発現変動比スコアM値、縦軸：頻度。

[図43]フザリウム・バーティシリオイデスのアレイデータにおいて、標的とする遺伝子クラスタが含まれるか否かを判定する遺伝子クラスタスコア分布判定値 e を示した図である。横軸：クラスタサイズ、縦軸：次元数6における ε 値。

[図44]フザリウム・バーティシリオイデスのアレイデータ取得系C 1およびC 2における、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判定する判定値 c を示した図である。横軸：クラスタサイズ、縦軸： χ 。(左) C 1、(右) C 2。

[図45]フザリウム・バーティシリオイデスのアレイデータ取得系C 1およびC 2における、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判定する判定値 u を示した図である。横軸：クラスタサイズ、縦軸： v 。次元数 d' は2を採用した。(左) C 1、(右) C 2。

[図46]フザリウム・バーティシリオイデスのアレイデータ取得系C 1およびC 2における、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判断する評価値 $c'u$ を示した図である。横軸：仮想の遺伝子クラスタ起点遺伝子ID、縦軸： $\chi \times v$ 。次元数 d' は2を採用した。各仮想の遺伝子クラスタについて、最大の絶対値を採る ncl の値をプロットしてある。

(上) C 1、(下) C 2。

[図47]大腸菌における仮想の遺伝子クラスタの、クラスタサイズを1から30までとったスコアヒストグラムを示した図である。(右) 培養開始後898、908、919分後の各系における、 ncl を変えたときの全体図。行：クラスタサイズ $ncl = 1 \sim 30$ (上から)、列：左から、898、908、919分後の系。(左) 908分後の系における、 $ncl = 4$ のときの拡大図。横軸：発現変動比スコ

アM値、縦軸：頻度。

[図48]大腸菌のアレイデータにおいて、標的とする遺伝子クラスタが含まれるか否かを判定する遺伝子クラスタスコア分布判定値 e を示した図である。横軸：クラスタサイズ、縦軸：次元数6における ε 値。

[図49]大腸菌のアレイデータ取得系における、大腸菌の成長を表す濁度の時系列データである（参考文献11の図1Aより抜粋）。横軸：培養開始後時間、縦軸：濁度

[図50]大腸菌のアレイデータ取得系C2における、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判定する判定値 c を示した図である。横軸：クラスタサイズ、縦軸： χ 。

[図51]大腸菌のアレイデータ取得系における、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判定する判定値 u を示した図である。横軸：クラスタサイズ、縦軸： v 。次元数 d' は2を採用した。

[図52]大腸菌のアレイデータ取得系における、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判断する評価値 $c'u$ を示した図である。横軸：クラスタサイズ、縦軸： $c'u$ 。次元数 d' は2を採用した。

[図53]大腸菌のアレイデータ取得系における、仮想の各遺伝子クラスタが標的とする遺伝子クラスタであるか否かを判断する評価値 $c'u$ を、横軸をゲノム上の起点遺伝子IDとして示した図である。横軸：仮想の遺伝子クラスタ起点遺伝子ID、縦軸： $\chi \times v$ 。次元数 d' は2を採用した。各仮想の遺伝子クラスタについて、最大の絶対値を採る nci の値をプロットしてある。

発明を実施するための最良の形態

[0018] 本発明は、生物細胞の生理状態変化を生じる条件とコントロール条件下において生じたゲノムDNA上に配列する遺伝子の発現量変動比を、ゲノムDNA上に配列する複数の遺伝子により構成される仮想の遺伝子クラスタ単位の発現量変動比として合算することにより、仮想の遺伝子クラスタ単位毎にスコアリングし、得られたスコアに基づき、まず、上記生理状態変化の原因遺伝子である標的遺伝子を含む遺伝子クラスタを特定し、さらに該クラスタ

から標的遺伝子を特定する方法である。

本発明は、また、上記方法を基本原理とし、生物ゲノム中の標的遺伝子を含む遺伝子クラスタ及び／または該遺伝子クラスタ中の標的遺伝子を探索する装置（以下、単に、本発明の遺伝子探索装置という場合がある。）に関するものであり、さらに該装置の一部を応用した遺伝子クラスタの有無及びそのサイズを予測する装置に関する。

本発明の探索法および探索装置においては、真核生物、原核生物を問わず、あらゆる生物種について、ゲノム中の有用遺伝子を含有する遺伝子クラスタを探索対象とすることができる。

また、本発明によれば、ゲノムの配列が明らかになっているものであれば、遺伝子クラスタの境界が明らかになっていない場合であっても本発明の手法および装置を適用でき、遺伝子クラスタ及び該クラスタ中の有用遺伝子を探索することができる。

[0019] 本発明における生理状態変化とは、例えば、生物の代謝物産生量の変化、分泌物質の種類と量の変化、増殖速度など増殖相（グロースフェーズ）の違い、静止期・間期などの細胞の分裂状態の違い、細胞の形態や機能の違い（菌糸、分生子など分化状態の違いを含む）等をいい、本発明においては、これら生理状態変化を生じる条件とコントロール条件とを一の対比条件セットとして、該対比条件セットを1種あるいは2種以上設定し、それぞれの対比条件セットの各条件下におけるゲノム遺伝子の発現量を測定し、その比（発現量変動量比）を求める。

生理状態変化を生じる条件とは、例えば、薬剤の使用、温度、栄養源、培地、培養時間等の調整により、人為的に生理状態変化を誘導する場合の他、特にこのような誘導をせず、経時的に生理状態変化が生じる場合の時間条件も含める。コントロール条件とは、生理状態変化を生じないかあるいは生じてても変化が少なく、生理状態変化を生じる条件下での生理状態変化と対比しうるものをいう。

例えば、2次代謝物の産生に関与する遺伝子クラスタあるいは遺伝子を探

索する場合、2次代謝物産生誘導条件下（あるいは抑制条件下）とコントロール条件としての2次代謝物産生非誘導条件下（あるいは産生条件下）におけるゲノム遺伝子の発現量を測定する。

比較する上記2次代謝物産生誘導条件と2次代謝物産生非誘導条件、あるいは2次代謝物産生抑制条件と2次代謝物産生条件とは、代謝物産生速度、量等に差が生じる条件であればよく、例えば、薬剤の使用、温度、栄養源、培地等の調製の有無等、あるいは特にこのような誘導をせず、経時的に2次代謝物産生量に生じる場合の時間条件も含まれる。

[0020] 本発明における遺伝子クラスタ及び遺伝子探索法の全体的な流れを、図1に示した。このうちグレーで示した大きな四角の内部（白い二つの四角を含む）が、本発明の特徴部分である。

本発明のプロセスにおいては、ゲノムDNA上に配列する各遺伝子の発現量の測定は、例えばマイクロアレイ等により行うが、その他のプロセスは、ゲノムDNA上に配列する遺伝子の発現量データに基づき、数学的データ処理により行うことができ、実験を必要とせず、また、上記発現量測定対象とするゲノム遺伝子の選定等も機械的に、あるいは研究者の特別な知識あるいは勘にほとんど左右されることがなく行うことができる。したがって、本発明の探索法は、コンピューター利用に極めて適しており、本発明によれば、迅速、効率的に有用遺伝子が探索可能となり、従来困難であって、代謝物、とりわけ2次代謝物産生に関与する遺伝子及び該遺伝子が含まれる遺伝子クラスタの探索に特に効力を発揮する。

以下、本発明のプロセスについて、さらに具体的に説明する。

[0021] 本発明における上記仮想の遺伝子クラスタの構成手法としては、例えば、A) ゲノムDNA上に配列する複数の遺伝子を配列順に組み合わせ、各サイズの異なる仮想の遺伝子クラスタを構成する手法、B) 近傍に位置し機能的に遺伝子クラスタを構成する可能性のある複数の遺伝子から構成する手法が挙げられる。この2つの手法は、発現量を測定する遺伝子の対象範囲が異なるため、使用する発現量変動比データ、仮想の遺伝子クラスタの構成ゲノム

遺伝子が異なるが、構成された仮想の遺伝子クラスタ・スコアリング等その他の数学的処理プロセス自体は共通している。

[0022] 以下に、本発明のプロセスについて、順次具体的に説明する（図1参照）。

1) 上記A)の手法による場合の発現量の測定及び発現量変動比データの取得、

A)の手法による場合、原則、ゲノムDNA上に配列する各遺伝子全てについて、生理状態変化を生じる条件とコントロール条件下とにおいて、それぞれ発現量を測定し、両条件下における発現量の比を求め、発現量変動比（生理状態変化条件下での発現量を分子、コントロール条件下での発現量を分母として算出した値）とする。

発現量の測定は、例えば、ゲノムDNA上に配列する各遺伝子に特異的なプローブを有するマイクロアレイを用いてそれ自体周知の方法で行うことができる。

例えば、代謝産物、特に2次代謝物の産生に関与する有用遺伝子を標的とする場合、1以上の2次代謝物産生誘導条件下（あるいは抑制条件下）で細胞を培養し、細胞からゲノムRNAを抽出し、ゲノムDNA上の各遺伝子に特異的なプローブを有するマイクロアレイでゲノムDNA上の各遺伝子の発現量を測定する。一方、コントロール条件として、上記2次代謝物の産生非誘導条件下（あるいは産生条件下）の場合における発現量を測定し、両条件下における発現量の比をとり、これを発現量変動比とする。

各遺伝子発現量の測定は、例えば、上記培養細胞からmRNAを抽出して、色素等でラベリングし、各遺伝子クラスタにおける上記各遺伝子中のDNA配列の一部を有するオリゴDNAをプローブとして基板に固定化したアレイを用い、上記該ラベリングしたmRNAを各オリゴDNAにハイブリダイズさせ、洗浄した後、発光強度等を測定することにより行う。

[0023] 2) 上記A)の手法による仮想の遺伝子クラスタの構築

仮想の各遺伝子クラスタは、ゲノムDNA上に連続する遺伝子を2個から

遺伝子数を1ずつ増やして、想定される遺伝子クラスタに含まれる最大限の遺伝子数になるまで抽出し、かつ該抽出において、抽出する遺伝子の各個数毎に、直鎖状DNAからなるゲノムの場合には該DNAのいずれかの末端から、環状DNAからなるゲノムの場合には任意の遺伝子を起点として順にゲノムDNA上に配列する遺伝子の一つずつずらしながら抽出された各遺伝子群から構成される。

この仮想の遺伝子クラスタの構築手法をより具体的に示すと例えば以下の手法が挙げられる。

[0024] (1) ゲノム遺伝子が直鎖状ゲノムの場合、

a) ゲノムDNAの一方の末端に位置する遺伝子を起点として、他方の末端方向に、順次、ゲノムDNA上に連続する遺伝子を同一方向に2個から一つずつ増やして($N+1$)、想定される遺伝子クラスタに含まれる遺伝子数の最大限(ncl)になるまで組み合わせ、起点とした遺伝子を含み、かつ遺伝子の個数の異なる複数の遺伝子群を構成する。

b) 起点を、順次、他方の末端方向に一遺伝子ずつずらしながら(起点遺伝子の移動)、上記aと同様の処理を行い、新たな起点遺伝子を含みかつ遺伝子の個数が異なる複数の遺伝子群を構成し、a)の遺伝子群と併せて、複数の遺伝子の組み合わせた遺伝子群からなる仮想の遺伝子クラスタを構築する。

(2) ゲノム遺伝子が環状の場合、ゲノムDNA上の任意の遺伝子を起点として、上記(1)a)及びb)と同様の処理を順次行い、最初に起点とした遺伝子が起点となる時点で処理を終了する(最初に起点として遺伝子に基づく仮想の遺伝子クラスタの構築は再度行わない。)。

[0025] 上記仮想の遺伝子クラスタの構築においては、仮想の遺伝子クラスタが複数の遺伝子から構成される点で、遺伝子2個から一つずつ増やす手法が採用されるが、本発明は、遺伝子1個から一つずつ増やす手法を排除するものではない。すなわち、この場合、遺伝子一個の場合が構築される仮想の遺伝子クラスタに混入することになるが、本発明においては、この混入した遺伝子を含む遺伝子2個以上の組み合わせからなる仮想の遺伝子遺伝子クラスタが

必ず構築され、また、仮想の遺伝子クラスタのスコアは、組み合わせた各遺伝子の発現量変動比の合算であるから、ゲノム中に標的遺伝子が存在した場合、この標的遺伝子単独のスコアに比べ、これを含む仮想の遺伝子クラスタのスコアは少なくとも同等以上になり、上記混入は実質的な問題ではない。したがって、仮想の遺伝子構築において、遺伝子2個から一つずつ増やす手法を含む限り、遺伝子1個から一つずつ増やしたとしても、本発明に包含される。

[0026] 例えば、以下のように、ゲノムDNA上に配列する遺伝子がA～Jの10個である場合、構築される仮想の遺伝子クラスタは、表1の示される各遺伝子群から構成される。

[化1]

A B C	D	E F	G H I	J
-------	---	-----	-------	---

[0027]

[表1]

遺伝子数	起点 A	B	C	D	E	F	G	H	I
遺伝子 2 個の各仮想の遺伝子クラスタ(9 個)	AB	BC	CD	DE	EF	FG	GH	HI	IJ
同 3 個の各仮想の遺伝子クラスタ(8 個)	ABC	BCD	CDE	DEF	EFG	FGH	GHI	HIJ	
同 4 個の各仮想の遺伝子クラスタ(7 個)	ABCD	BCDE	CDEF	DEFG	EFGH	FGHI	GHIJ		
同 5 個の各仮想の遺伝子クラスタ(6 個)	ABCDE	BCDEF	CDEFG	DEFGH	EFGHI	FGHIJ			
同 6 個の各仮想の遺伝子クラスタ(5 個)	ABCDEF	BCDEFG	CDEFGH	DEFGHI	EFGHIJ				
同 7 個の各仮想の遺伝子クラスタ(4 個)	ABCDEFG	BCDEFGH	CDEFGHI	DEFGHIJ					
同 8 個の各仮想の遺伝子クラスタ(3 個)	ABCDEFGH	BCDEFGHI	CDEFGHIJ						
同 9 個の各仮想の遺伝子クラスタ(2 個)	ABCDEFGHI	BCDEFGHIJ							
同 10 個の各仮想の遺伝子クラスタ(1 個)	ABCDEFGHIJ								

[0028] すなわち、上記抽出によって構築される仮想の各遺伝子クラスタは、以下の各遺伝子群からなる。

遺伝子の 2 個の各仮想の遺伝子クラスタ（9 個）；AB, BC, CD, DE, EF, FG, GH, HI

, IJ

同 3 個の仮想の各遺伝子クラスタ (8 個) ; ABC, BCD, CDE, DEF, EFG, FGH, GHI, IJK

同 4 個の仮想の各遺伝子クラスタ (7 個) ; ABCD, BCDE, CDEF, DEFG, EFGH, FGHI, GHII

同 5 個の仮想の各遺伝子クラスタ (6 個) ; ABCDE, BCDEF, CDEFG, DEFGH, EFGHI, FGHIJ

同 6 個の仮想の各遺伝子クラスタ (5 個) ; ABCDEF, BCDEFG, CDEFGH, DEFGHI, EFGHIJ

同 7 個の仮想の各遺伝子クラスタ (4 個) ; ABCDEFG, BCDEFGH, CDEFGHI, DEFGH IJ

同 8 個の仮想の各遺伝子クラスタ (3 個) ; ABCDEFGH, BCDEFGHI, CDEFGHIJ

同 9 個の仮想の各遺伝子クラスタ (2 個) ; ABCDEFGHI, BCDEFGHIJ

同 1 0 個の仮想の各遺伝子クラスタ (1 個) ; ABCDEFGHIJ

[0029] したがって、この場合、仮想の各遺伝子クラスタの構築数は 4 5 個であるが、これらの各遺伝子クラスタはデータ上において構築されるだけであって、実験によって実際に構築されるものではない。なお、実際のゲノム DNA 上の遺伝子数は、麹菌の場合、外部データベース DOGAN(<http://www.bio.nite.go.jp/dogan/project/view/A0>)に登録されているもので 1 2 0 8 4 個であり、これよりも遺伝子の定義を緩めて DNA マイクロアレイのプラットフォーム作成に使用されたものの場合 1 4 0 3 2 個である。このうち連続していることが判明しているゲノム上の領域から、仮想の遺伝子クラスタを構築する。

抽出する遺伝子の数の最大限は論理上ゲノム中の遺伝子の数とすることができるが、想定される遺伝子クラスタサイズの最大限の遺伝子数でよく、実際問題として、遺伝子クラスタを構成する遺伝子の数は、最大でも 3 0 個程度であり、これを超える必要は通常ない。

[0030] 1') 上記 B) の手法による場合の発現量の測定及び発現量変動比データの

取得

このB)手法は、上記A)の手法に比べ簡便であり、2次代謝産物の産生に関与する遺伝子クラスタ及び該クラスタ中の2次代謝産物産生遺伝子の探索に特に適している。

この手法は、ゲノムDNAの配列中、(1)2次代謝に関与していると想定される酵素種に属する酵素遺伝子、(2)トランスポーター遺伝子、(3)転写因子をコードする遺伝子の内、1種以上、好ましくは2種以上が近傍に位置する場合、これらの遺伝子から、あるいはこれら遺伝子が含まれるようにゲノム遺伝子を組み合わせて仮想の遺伝子クラスタとするものであり、この場合において、近傍に位置する具体的な条件は、ゲノム上に配列する遺伝子数でいえば、上限30程度以内に存在すればよい。

[0031] 上記遺伝子の発現量の測定は、上記A)の手法と同様に、例えば2次代謝物産生誘導条件下（あるいは抑制条件下）で細胞を培養し、細胞からゲノムRNAを抽出し、ゲノムDNA上の各遺伝子に特異的なプローブを有するマイクロアレイを使用して、ゲノムの各遺伝子の発現量を測定し、上記2次代謝物産生非誘導条件下（あるいは産生条件下）の場合と比較して、発現量変動比を求める。この手法の場合、マイクロアレイにおける発現量測定においては、ゲノムDNA上の全遺伝子について行うが、発現変動量の抽出対象遺伝子が絞られているため、これら遺伝子に対応する配列を有するプローブを用いたマイクロアレイのみを用いても良い。

比較する上記2次代謝物産生誘導条件と2次代謝物産生非誘導条件、あるいは2次代謝物産生抑制条件と2次代謝物産生条件とは、代謝物産生速度、量等に差が生じる条件であればよく、例えば、薬剤の使用、温度、栄養源、培地等の調製の有無等、あるいは特にこのような誘導をせず、経時的に2次代謝物産生量に生じる場合の時間条件も含まれる。

なお、この手法においても、上記A)の手法と同様に発現変動量の測定の他は格別の実験を必要とせず、数学的データ処理によりなされる。

[0032] 一方、ゲノム配列中の(1)2次代謝に関与していると想定される酵素種

に属する酵素遺伝子、(2) トランスポーター遺伝子、(3) 転写因子をコードする遺伝子の判別は、既知の同じ酵素種の遺伝子との相同性あるいはモチーフ等により判別すればよく、例えば、各仮想の遺伝子クラスタ中の遺伝子配列中にこれらの遺伝子が存在するか否かは、上記酵素種に属する酵素、トランスポーター、転写因子の各アミノ酸配列に特有なモチーフと共通するアミノ酸配列をコードする塩基配列が、上記遺伝子クラスタ中に存在するか否かにより識別できる。これらについては市販ソフトウェアを使用することができる。すなわち、上記機能遺伝子の選定、及び以下に示す仮想の遺伝子クラスタのスコアリングにおいて重み付けする遺伝子の選定においては、アノテーション（機能注釈）付与を行い、これに基づき対象遺伝子を選定することが有効である。このようなアノテーション付与は、探索対象ゲノム上の各遺伝子の塩基配列情報等を基に、記憶部に記憶された検索対象のゲノム上の各遺伝子の位置情報中の遺伝子について行うもので、コンピューターにより自動で行うことができる。

[0033] このようなアノテーション付与は、装置使用者が、検索探索対象ゲノム上の遺伝子について予め相同検索あるいはモチーフ検索等の結果に基づき、上記記憶されたゲノム上の各遺伝子の位置情報中の遺伝子を逐一指定し、この指定された遺伝子にアノテーションが付与されるように構成しても良いが、ゲノム上の遺伝子の数は極めて多数であり、上記モチーフ検索を行う市販のソフトウェアを付属のモチーフ情報とともにコンピューターに格納するか本発明装置に格納するか、あるいは該ソフトウェアをモチーフ情報とともに格納した外部コンピューターを利用することが好ましい。これにより、探索対象ゲノム上の各遺伝子の塩基配列情報を上記コンピューターあるいは外部コンピューターに入力することにより、期待される機能に対応したモチーフについて検索を行い、アノテーション付与する遺伝子を自動で選定することができる。また、他のアノテーション付与手段として、上記モチーフ検索により探索対象のゲノム上の全遺伝子にアノテーションを付与した後に、付与されたアノテーションの種類（遺伝子機能）から、期待される機能と一致する

遺伝子を選定してもよい。

[0034] このようにすれば、アノテーション付与を研究者の手を煩わすことなく自動で行うことができる。アノテーション付与は、機能が同様なゲノム遺伝子に付与しても良いし、機能の種類が異なる複数種の遺伝子に付与しても良い。機能が異なる複数種のゲノム遺伝子にアノテーションを付与する場合には、ゲノム遺伝子の機能毎に識別可能なように付与する。アノテーションによる選定の対象となる遺伝子は、例えば、2次代謝物産生に関与する遺伝子クラスタあるいはその中の遺伝子を標的とする場合、ゲノムDNAの配列中、
(1) 2次代謝に関与していると想定される酵素種に属する酵素遺伝子、
(2) トランスポーター遺伝子、
(3) 転写因子をコードする遺伝子を選定可能である。

[0035] 上記(1)の酵素遺伝子の判別において、酵素種は、2次代謝物の化学構造、前駆体、関与しうる補酵素、化学的・物理的性質、既知の酵素反応の事例、生産効率・速度等からその産生反応を推測し、関与する酵素種を想定するが、この酵素種の想定においては、実際にその反応に関与したであろう特定酵素のレベルまで想定しなければならないというわけではなく、該反応に関与することがより確実なレベルの酵素種でよい。例えば、オキシゲナーゼに属する酵素であることは分かるが、その下位概念の酵素種まで特定できないときは、オキシゲナーゼのレベルを酵素種として選定して、ゲノム上の各遺伝子の配列を探索し、その範疇に属する全てのゲノム遺伝子のそれぞれを、各仮想の遺伝子クラスタの構成遺伝子とすればよい。ただし、下位概念の酵素種を選定できた場合には、探索対象とする仮想の遺伝子クラスタの範囲が狭まる可能性があり、その分探索が効率的となる。

また、2次代謝物産生反応において複数の酵素が関与していると想定できる場合には、その複数の酵素種を選定することも可能である。

トランスポーター遺伝子及び転写因子遺伝子においても同様に、標的とする2次代謝物産生に直接関与している遺伝子を特定しなければならないというわけではない。

[0036] 2') 上記B)の手法による仮想の遺伝子クラスタの構築

上記B)の手法による場合、近傍に位置する2次代謝に関与していると想定される酵素種に属する酵素遺伝子、2) トランスポーター遺伝子、3) 転写因子をコードする遺伝子のうち少なくとも1種以上、好ましくは2種以上の遺伝子を抽出し、これらを組み合わせることにより、あるいはこれら遺伝子が含まれるようにゲノムDNA上の遺伝子を抽出して仮想の遺伝子クラスタとする。

例えば、以下のように、ゲノムDNA上に配列する遺伝子がA~Jの10個である場合、

[化2]

A * B C “	D	E F	G * H	I J “
-----------	---	-----	-------	-------

(*は該当する酵素種をコードする遺伝子、“はトランスポーター遺伝子) 前者の場合、仮想の遺伝子クラスタは、AC及びGJとにより構成される。一方、後者の場合は、ABC及びGHIJで構成してもよく、さらにABCDEあるいはFGHIJのように、各仮想の遺伝子クラスタが一定数の遺伝子により構成されるようにゲノムを分割して、各仮想の遺伝子クラスタを構成しても良い。

[0037] 3) 仮想の遺伝子クラスタのスコアリング

上記1)のプロセスにより取得されたゲノムDNA上に配列する各遺伝子の発現量変動比は、各対比条件セット毎に、正規化され、上記2)のプロセスにより構築された仮想の各遺伝子クラスタ単位で、以下の計算式a)により合算され、算出された値を、仮想の各遺伝子クラスタのスコアとする。

[0038] 計算式 a)

[数1]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、 M は仮想の各遺伝子クラスタのスコア、 m はスコアリングされる仮想の各遺伝子クラスタに含まれる各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m 値)の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m 値)の標準偏差を表す。)

なお、上記全ての仮想の遺伝子クラスタに含まれる全遺伝子とは、全ての仮想の遺伝子クラスタを構成するために抽出されたゲノムDNA上の全ての遺伝子をいう。

一方、1')のプロセスにより取得された各遺伝子の発現量変動比も、同様に各対比条件セット毎に、正規化され、上記2)のプロセスにより構築された仮想の各遺伝子クラスタ単位で合算されるが、この手法はアノテーション付与により選定された特定の遺伝子のみの発現量変動比を用いるため、計算式a)の定義が異なる。すなわち上記式中、 M は各仮想の遺伝子クラスタのスコア、 m はスコアリングされる仮想の各遺伝子クラスタに含まれるアノテーション付与に基づき選定された各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれるアノテーション付与に基づき選定された全遺伝子の発現量変動比(m 値)の平均、 $s(m)$ は全ての仮想の遺伝子クラスタに含まれるアノテーション付与に基づき選定された全遺伝子の発現量変動比(m 値)の標準偏差を表す。

[0039] 本発明によれば、このようにして得られた一群の仮想の遺伝子クラスタのスコアに対する出現頻度分布をみる場合、全体としては大凡正規分布となるが、このような全体のスコア分布から離れて存在する仮想の遺伝子クラスタが存在すれば、少なくとも標的の遺伝子クラスタと対応していると判定できる。

すなわち、この仮想の遺伝子クラスタは、該クラスタ中の少なくとも2つの遺伝子が、代謝物産生誘導条件下協働した結果、発現変動量の総量である

スコアが増大したものであり、標的の遺伝子クラスタとみなすことができ、この仮想の遺伝子クラスタ中の遺伝子は少なくとも実際の遺伝子クラスタ中に存在する代謝物産生に関与する遺伝子であると同定することができる。さらに、仮想の遺伝子クラスタ中の遺伝子及び必要に応じ代謝産物の産生機構を検討すれば、直接代謝産物の産生に関与する標的遺伝子のみではなく、未知の機能を有する遺伝子の発見も期待でき、さらに代謝物産生機構の全体像も明らかにすることができる。

[0040] 一方、A)の手法において、ゲノムDNA上に配列する遺伝子が、標的とする遺伝子機能を有すると推定される場合、あるいは標的とする遺伝子機能を有する可能性が低い若しくはその可能性がないと推定できる場合においては、当該遺伝子については、以下の計算式により重み付けをしておくことができる。

[0041] [数2]

$$w \frac{m - \bar{m}}{\sigma(m)}$$

(式中、 m は、上記標的とする遺伝子機能を有するか、あるいは標的とする遺伝子機能を有する可能性が低い若しくはその可能性がないと推定されるゲノムDNA上の遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値)の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値)の標準偏差、 w は重みとしての任意の実数を、それぞれ表す。)

[0042] 重み w の設定は、上記標的とする遺伝子機能を有すると推定される場合、1を超えるように設定し、標的とする遺伝子機能を有する可能性が低い若しくは可能性がないと推定できる場合は、0以上1未満になるように設定する。標的とする遺伝子機能を有するか、あるいはその可能性が低いかの推定は上記と同様に既知の遺伝子との相同性あるいはモチーフ等により判別でき、上記したアノテーション付与手段を利用することができる。

また、ゲノムDNA上に配列する遺伝子が、標的とする遺伝子機能を有すると推定される場合においては、A)の手法により構築された仮想の遺伝子ク

ラストの中から、標的とする遺伝子機能を有すると推定された遺伝子を含む仮想の遺伝子クラスタを選出し、選出された仮想の遺伝子クラスタのみについて、スコアリングすることも可能である。標的とする遺伝子機能を有するか否かについての推定においては、上記したアノテーション付与手段の全てを利用することができる。この手法によれば、スコアリングする対象となる仮想の遺伝子クラスタの数を低減することができる。また、この手法により選定された仮想の遺伝子クラスタは、結果として上記手法B)により構築された仮想の遺伝子クラスタと同様になる場合があるが、この手法による場合、一度A)の手法による網羅的な仮想の遺伝子クラスタ群を構築しておけば、自由に標的とする遺伝子あるいはこれを含む遺伝子クラスタの機能を変更でき、機能選択的な遺伝子解析が容易に行える点で有利である。また、該当するアノテーションが付与されなかった遺伝子のスコアを考慮に含めることができるため、機能未知の遺伝子の影響が大きい場合などに柔軟に対応できる。

[0043] 本発明は、ゲノムDNA上の複数の遺伝子を組み合わせて仮想の遺伝子クラスタを構成し、これら複数の遺伝子の生理状態変化条件下の発現量変動比を合算して各仮想の遺伝子クラスタをスコアリングし、これに基づき、まず、標的の遺伝子クラスタを探索する方法である。スコアリングされてスコアが高いものが得られた場合は、仮想の遺伝子クラスタに含まれる複数の遺伝子が協働した結果であり、各遺伝子単独の発現量変動比スコアをみるよりも、全体のスコア分布に対する特異性がより鮮明となる。一方、従来のように一つ一つの遺伝子の発現変動量のみから有用遺伝子を検出する場合には、正解の遺伝子であっても、全体のスコア分布中に吸収されてしまい、高い順位の遺伝子であっても、目的の遺伝子であるか否かの遺伝子破壊実験等の検証を必要とする。

[0044] これに加え、上記したように重み付けをした遺伝子についての発現量変動比は、A)の手法において構築される仮想の各遺伝子クラスタのスコアリングにおいて、他の遺伝子の発現量変動比と合算され、上記標的とする遺伝子

機能を有すると推定される遺伝子を含む仮想の各遺伝子クラスタのスコアはより高くなり、反対に標的とする遺伝子機能を有する可能性が低い若しくは可能性がないと推定される遺伝子を含む仮想の遺伝子クラスタのスコアはより低くなり、全体のスコア分布との乖離が明瞭となる。したがって、これにより、標的とする遺伝子機能を有する遺伝子あるいはこれを含む遺伝子クラスタの探索がより効率的になる。

[0045] 4) 全体のスコア分布からの乖離の程度の算出

仮想の遺伝子クラスタ全体のスコアの分布からの乖離の程度を示す判定値は、上記3)のプロセスにより算出されたスコアに基づき、例えば、以下の計算式b)あるいはc)から算出される。

[0046] 計算式b)

[数4]

$$\chi = -M \log P$$

(式中、 χ は上記仮想の各遺伝子クラスタの乖離の程度を示す判定値、 M は仮想の各遺伝子クラスタのスコア、 P は、全ての仮想の遺伝子クラスタについての各スコアに対する出現頻度分布において、各スコア M の出現頻度について全体累計を1としたときの、上記スコア M の出現頻度の割合を表す。)

[0047] 上記計算式b)中のスコア M の出現頻度は、仮想の遺伝子クラスタの全てを含む集団における各スコアの出現頻度(P)の累計を1としたときの値であるため、1を超えることはなく、したがって $\log P$ は正になることはない。また、出現頻度が低いものほど $\log P$ は $-\infty$ に近づくため、頻度の低いスコア値を持つ遺伝子クラスタほど $\log P$ の絶対値は大きくなる。したがって、上記計算式b)においては、 $\log P$ と仮想の各遺伝子クラスタのスコアを掛け合わせて -1 を乗算することにより、頻度が低かつスコアの高いものが、より大きな判定値 χ を持つこととなる。

上記計算式b)によれば、判定値 χ が0を超え、高い値を示す仮想の遺伝子クラスタは、仮想の各遺伝子クラスタのスコアに対する出現頻度分布から離れており、高い判定値 χ を示した仮想の遺伝子クラスタを標的の遺

伝子クラスタあるいは標的の遺伝子クラスタに対応する候補として選定することができる。候補の選定は、例えば判定値 l が高い順に仮想の遺伝子クラスタを一定数選定するかあるいは判定値 l が一定値以上を示した仮想のクラスタを選定するか等により行う。

[0048] 計算式 c)

[数5]

$$v = (M - \bar{M})^{d'} / (a\sigma(M))^{d'}$$

(式中、 v は上記仮想の各遺伝子クラスタの乖離の程度を示す判定値 l 、 M は仮想の各遺伝子クラスタのスコア、 $\bar{M}$ は全ての仮想の遺伝子クラスタのスコア (M 値) の平均値、 $\sigma(M)$ は全ての仮想の遺伝子クラスタにおけるスコア (M 値) の標準偏差、 a は任意の正の実数、 d' は次元数となる正の偶数を表す。)

[0049] この判定値 l (v) は、仮想の各遺伝子クラスタのスコアについて、仮想の遺伝子クラスタ全体の平均スコアからのずれを、上記標準偏差の実数倍で割ったものを次元数 (d') 乗したもので、正規分布様のスコアに対する出現頻度分布から乖離するスコアを有する仮想の遺伝子クラスタにおいて大きな値となる。上記式中 d' は任意に設定できる正の整数たる次元数であり、値が大きくなるほど平均スコアからの隔たりが強調されることになる。あまり大きくしすぎると、平均スコアから大きく外れたものの値が強調されて相対的にそれ以外の値が小さくなるため、通常2または4に設定する。外れたものをより鋭敏に検出したい場合は、6以上の偶数とする。また式中の a は外れ度を表す係数で、この値を調節することにより、上記正規分布様分布からどの程度乖離したものをとるかを調節することができる。1を超えて大きく設定するほど、平均スコアから大きく外れたもの以外の v 値はゼロに近付くため、この a 値は通常1～2に設定する。逆に1未満の場合、より外れ方の小さなものも拾うことができる。

この計算式 c) による場合も 上記判定値 l と同様に、 v が0を超え、高い値を示す仮想の遺伝子クラスタを標的の遺伝子クラスタあるいは標的の遺伝子クラスタに対応する候補として選定することができる。候補の選定は、

例えば判定値 l が高い順に仮想の遺伝子クラスタを一定数選定するかあるいは判定値 l が一定値以上を示した仮想のクラスタを選定するか等により行う。

[0050] 5) 遺伝子クラスタ候補の絞り込み

上記計算式 b)、c) により算出された判定値 (χ あるいは v) により、標的の遺伝子クラスタ候補となった仮想の遺伝子クラスタの数が多く、さらに候補を絞り込みたい場合は、以下の計算式 d) の算出結果に基づき、 b が 100 未満の仮想のクラスタを少なくとも除外することにより、標的の遺伝子クラスタ候補をさらに絞り込むことが可能である。

[0051] 計算式 d)

[数10]

$$\chi \times v > b$$

(式中、 χ は【数4】の計算式 b) によって算出した仮想の各遺伝子クラスタの判定値 l 、 v は【数5】の計算式 c) によって算出した仮想の各遺伝子クラスタの判定値 l 、 b は閾値としての任意の正の実数を表す。)

[0052] 上記計算式 d) 中、 b はどの程度の遺伝子クラスタ候補を絞り込むかを決定するための閾値であり、 b を大きくとるほど候補の絞り込み効果がより高くなる。また小さくとるほど多くの候補遺伝子クラスタを選択することができる。 b の値の設定は対象とする生物種や培養条件に依存する。すなわち、候補遺伝子クラスタが強かつ多く発現している系であれば値を高くする必要があるが、逆に発現強度が弱かつ数が少なければ値を低くしなければ候補遺伝子が出現しない。前者の場合、例えば 5000～10000 あるいは 10000～30000 の範囲内の任意の数値に設定し、後者の場合、通常 100 以上、例えば 1000～2000、あるいは 2000～5000 の範囲内の任意の数値に設定する。

[0053] 6) 標的とする遺伝子クラスタの有無及び標的とする遺伝子クラスタが存在する場合のサイズの推定

本発明においては、予めゲノム中に標的の遺伝子クラスタが存在するか否

か及び標的遺伝子クラスタが存在する場合の遺伝子サイズ（クラスタを構成する遺伝子数； n_{cl} ）を推定することができる。

この手法は、まず、生物細胞の生理状態変化を生じる条件とコントロール条件下において生じたゲノムDNA上に配列する遺伝子の発現量変動比を合算し、仮想の遺伝子クラスタのスコアとするが、上記発現量の測定、発現量変動比データの取得、仮想の遺伝子クラスタの構築、及び仮想の各遺伝子クラスタをスコアリングの各プロセスは、上記A)の手法中1)～3)のプロセスと同様なプロセスである。

[0054] すなわち、この手法においては、生物細胞の生理状態変化を生じる条件とコントロール条件下において生じたゲノムDNA上の各遺伝子の発現量変動比を、ゲノムDNA上の複数の遺伝子により構成される仮想の遺伝子クラスタ単位の発現量変動比として合算することにより、仮想の遺伝子クラスタ単位毎にスコアリングするが、この仮想の各遺伝子クラスタは、ゲノムDNA上に連続する遺伝子を2個から遺伝子数を一つずつ増やし想定される遺伝子クラスタに含まれる最大限のゲノム遺伝子数になるまで抽出し、かつ該抽出において抽出する遺伝子の各個数毎に、直鎖状DNAからなるゲノムの場合には該DNAのいずれかの末端から、あるいは環状DNAからなるゲノムの場合には任意の遺伝子を起点として順にゲノムDNA上に配列する遺伝子一つずつずらしながら抽出された各遺伝子群から構成する。

このように構成された各遺伝子クラスタは、上記A)の手法中3)のプロセスと同様に以下の計算式a)により、そのスコアが算出される。

[0055] 計算式a)

[数1]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

（式中、 M は仮想の各遺伝子クラスタのスコア、 m はスコアリングされる仮想の各遺伝子クラスタに含まれる各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比（ m 値）の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比（ m 値）の標準偏差を表す。）

ついで、このスコアを仮想の各遺伝子クラスタに含まれる遺伝子数毎に分け、以下の計算式 e) により、各遺伝子数単位毎に遺伝子クラスタスコア分布判定値 (ε) を求める。

[0056] 計算式 e)

[数7]

$$\varepsilon = \sum (M - \bar{M})^d / n \sigma(M)^d$$

(式中、 ε は各遺伝子数単位毎に求められた遺伝子クラスタスコア分布判定値、 M は全ての仮想の遺伝子クラスタを遺伝子数単位毎に分けた場合の、各遺伝子数単位に含まれる仮想の各遺伝子クラスタのスコア、 $\bar{M}$ は全ての仮想の遺伝子クラスタのスコアの平均、 n は仮想の遺伝子クラスタの総数、 $\sigma(M)$ は全ての仮想の遺伝子クラスタのスコア (M 値) の標準偏差、 d は任意に設定する次元数としての正の偶数を表す。)

[0057] この計算式 e) によれば、仮想の遺伝子クラスタが、実際のゲノム DNA においてクラスタを形成していなければ、仮想の遺伝子クラスタ中に含まれる、標的の生理状態変化に関与せず、発現量変動のない遺伝子の影響を受けるので、仮想の遺伝子クラスタのスコア (M) はサイズ (遺伝子数; n) が大きくなるほど平均化され、すなわちスコアの平均値に近づくためサイズの増大に伴い ε 値は単調減少する (図 2 上から 1 および 3 番目の曲線参照)。しかし、仮想の遺伝子クラスタのあるサイズのものがクラスタを形成している場合、そのサイズにおいて、分布の偏り ε は大きくなり、上記単調減少曲線とはならず、 ε 値はそのサイズにおいて特異点を形成する (図 2 矢印で示した点参照)。したがって、 ε 値が特異点を形成するか否か、及び特異点を形成した遺伝子クラスタのサイズから、遺伝子クラスタの存在及びそのサイズを推定することができる。

[0058] 具体的には、仮想の遺伝子クラスタの該クラスタに含まれる遺伝子数毎の集計において、ある遺伝子数 (k 個) のときの ε 値 ($\varepsilon(k)$) と、その前後数のときの ε 値 ($\varepsilon(k-1)$ 、 $\varepsilon(k+1)$) が以下の関係にあれば、標的とする遺伝子クラスタがゲノム中に存在すると判定し、標的遺伝子クラスタに含まれる遺伝子数を k 個と予想することができる。

[0059] [数8]

$$\varepsilon(k) > \varepsilon(k-1) \quad \text{かつ} \quad \varepsilon(k) > \varepsilon(k+1)$$

[0060] この手法は、本発明による標的の遺伝子クラスタの探索法、特にB)の手法を行うに際し、予め行う手法として有効である。すなわち、遺伝子クラスタが存在し、そのサイズを予想できれば、予想されるサイズ内に、標的とする酵素種に属する酵素遺伝子、(2) トランスポーター遺伝子、(3) 転写因子をコードする遺伝子が存在するゲノム配列のみを対象として、上記仮想の遺伝子クラスタとして探索すればよい。

また、この手法は、ある条件下で細胞が何らかの生理的状态変化を起こす場合においては、どのような生理的状态変化であっても変化を対比する条件が設定できれば、その原因遺伝子はもちろんその変化を生じる機構そのものが全く不明である場合においても、その変化原因が遺伝子クラスタ中の遺伝子の連関にあるのか否か、遺伝子クラスタ中の遺伝子の連関による場合該クラスタの遺伝子サイズも容易に予測できる。すなわちこの手法は、生物の生理的变化が、極めて探索の難しい複数の遺伝子の連関によって生じている場合において、その原因が遺伝子クラスタ中の遺伝子の共働によるものであることを明らかにでき、かつそのサイズも予測できる点で極めて有用である。

[0061] 7) 本発明の手法により解が得られなかった場合

一方、本発明の手法を行った結果、仮に仮想の遺伝子クラスタ全体のスコア分布から乖離したスコアの遺伝子クラスタが見いだされなかった場合、設定する生理状态変化条件、重み付けするゲノムDNA上の遺伝子の選定、あるいは上記B)の手法による仮想の遺伝子クラスタ構築のためのゲノムDNA上の遺伝子の選定等の探索条件設定に問題点がある。したがって、このような場合には、探索条件を再設定して、バックグラウンドの分布から離れたスコアの遺伝子クラスタが見つかるまで、上記した遺伝子クラスタの探索法を繰り返し行えばよい。すなわち本発明においては、得られたデータのみから、探索条件設定の問題点を把握できる。

これに対して、上記したような従来法の場合には、もともと正解の遺伝子であっても、遺伝子全体の発現量についての分布中に埋もれてしまうので、得られたデータからでは正解か否かは不明であり、結果的に無意味かもしれない検証実験を繰り返さなければならない。

[0062] 次に、本発明のプロセスを実施するために用いられる、本発明の遺伝子探索装置について、説明する。

本発明の遺伝子探索装置は、ゲノムDNA上に配列する遺伝子の発現量データに基づき、数学的データ処理を行うもので、研究者の特別な知識あるいは勘にほとんど左右されることがなく、迅速、効率的に有用遺伝子が探索可能となり、従来困難であった代謝物、とりわけ2次代謝物産生に関与する遺伝子及び該遺伝子が含まれる遺伝子クラスタの探索に特に効力を発揮する。

本発明の遺伝子探索装置は少なくとも以下のa)～f)の手段により構成される。

a) 生物細胞の生理状態変化を生じる条件とコントロール条件下におけるゲノムDNA上に配列する各遺伝子の発現量データを入力する手段。

b) 入力された上記2つの条件下における各遺伝子の発現量の比を算出する手段。

c) ゲノムDNA上に配列する複数の遺伝子を組み合わせて仮想の遺伝子クラスタを構築する手段。

d) 該算出されたゲノムDNA上に配列する各遺伝子の発現量変動比を複数の遺伝子により構築された上記仮想の遺伝子クラスタ単位の発現量変動比として合算し、仮想の遺伝子クラスタ単位毎にスコアリングする手段。

e) 得られたスコアに基づき上記生理状態変化の原因遺伝子である標的遺伝子を含む遺伝子クラスタを選定する手段。

あるいはさらに

f) 選定された遺伝子クラスタ中に含まれる遺伝子を表示する手段。

[0063] このような手段を伴う本発明の装置の概要は図2に示される。なお、図2中、点線部は、本発明装置において、さらに記憶されるのが好ましいデータ

及び該データに関連する処理部を示す。

本発明装置は、データの入出力部（キーボード、マウス、ディスプレイ等）、該入出力部の制御を行う入出力制御インターフェース、記憶部（ハードディスク）、主記憶部（メモリ）、制御演算部（CPU）、外部ネットワークと接続する通信制御インターフェースを含む。

本装置の記憶部には、各遺伝子の発現量データ、該発現量変動比データ、ゲノム上の各遺伝子位置データ、及び仮想の遺伝子クラスタのスコアデータが記憶され、さらに必要に応じて、塩基配列に対する遺伝子機能の対応データ、各遺伝子のアノテーションデータ、仮想の遺伝子クラスタのスコア乖離度データが順次格納される。

[0064] また、制御演算部には、ゲノム各遺伝子の発現量変動比の算出部、ゲノム上の遺伝子の位置情報に基づき仮想の遺伝子クラスタの構築を行う仮想の遺伝子クラスタの構築部、及び上記算出された発現量変動比を合算し、仮想の遺伝子クラスタのスコアリングを行う、仮想の遺伝子クラスタのスコアリング部を少なくとも設ける。

また、さらに必要に応じて、各遺伝子へのアノテーション付与部、アノテーションに応じて、仮想の遺伝子遺伝子のスコアリングにおいて重み付けを行う重み付け付与部、仮想の遺伝子クラスタ構築を選定された機能遺伝子に限定して行うための機能遺伝子選択部、仮想の遺伝子クラスタの全体分布からの乖離度を算出する仮想の遺伝子クラスタの乖離度算出部を設け、さらに算出された乖離度では、遺伝子クラスタ候補の選定が充分できない場合に遺伝子クラスタ候補の絞り込みを行う遺伝子クラスタ候補の絞り込み部を設けてもよい。

[0065] 一方、本発明の遺伝子探索装置においては、装置構成はそのまま、標的とする遺伝子クラスタの存在の有無、存在する場合そのサイズを予測する機能をさらに保有させることが可能であり、この場合には、仮想の遺伝子クラスタのサイズ毎にスコアリングするサイズスコアリング部、及び仮想の遺伝子クラスタ分布判定値（ ε ）算出部を設ける。

本装置は、特別なコンピューターを必要とせず、一般的な、制御演算処理装置（CPU）、主記憶装置（メモリ）、記憶装置（ハードディスク）、入出力装置（キーボード、マウス、ディスプレイ）からなるもので構成可能である。オペレーティングシステムは、Linux、Windows、Macのいずれも使用可能であるが、メモリ空間を考慮すると、64bitのものがより望ましい。メモリは、本装置が生物のゲノム全体を対象とすることを考慮して、できれば2GB以上のものが望ましいが、1GB程度であっても、微生物であれば可能である。

[0066] なお、ゲノム上の各遺伝子の位置情報、および機能に対応した塩基配列のデータベースは、NCBI (<http://www.ncbi.nlm.nih.gov/>) やInterproScan (<http://www.ebi.ac.uk/Tools/InterProScan/>) などの、外部データベースを利用することができる。

[0067] 以下に、本発明の装置について、その処理プロセスに従い、具体的に説明する。

A) 遺伝子探索装置

1) ゲノムDNA上に配列する各遺伝子の発現量データ入力及び発現量変動比算出

本発明装置の場合、原則、ゲノムDNA上に配列する各遺伝子全てについて、生理状態変化条件下とコントロール条件下における発現量を測定し、得られた各遺伝子の発現量データを本発明装置の入力手段に入力し、入力された各遺伝子の発現量データに基づき、発現量変動比が算出される。

発現量の測定は、例えば、ゲノムDNA上に配列する各遺伝子に特異的なプローブを有するマイクロアレイを用いてそれ自体周知の手段により行うことができる。

例えば、代謝産物、特に2次代謝物の産生に関与する有用遺伝子を標的とする場合、1以上の2次代謝物産生誘導条件下（あるいは抑制条件下）で細胞を培養し、細胞からゲノムRNAを抽出し、ゲノムDNA上の各遺伝子に特異的なプローブを有するマイクロアレイでゲノムDNA上の各遺伝子の発

現量を測定する。一方、コントロール条件として、上記2次代謝物の産生非誘導条件下(あるいは産生条件下)の場合における発現量を測定し、両条件下における発現量の比をとり、これを発現量変動比とする。

[0068] 各遺伝子発現量の測定は、例えば、上記培養細胞からmRNAを抽出して、色素等でラベリングし、上記各遺伝子中のDNA配列の一部を有するオリゴDNAをプローブとして基板に固定化したアレイを用い、上記該ラベリングしたmRNAを各オリゴDNAにハイブリダイズさせ、洗浄した後、発光強度等を測定することにより行う。

マイクロアレイ中の各遺伝子の発光強度は、例えば、マイクロアレイ読み取り装置における走査手段を伴う画像読み取り手段により読み取り、読み取った発光強度を数値化して、上記a)の入力手段により本発明の装置に入力する。このような画像読み取り装置は、市販されている装置が使用できるが、このような読み取り装置の手段全部あるいは例えば数値化手段等の一部手段を本発明の装置に組み込むか、あるいは該読み取り装置が出力する数値データを介して本発明装置の入力手段に自動で入力可能なように設計しても良い。

[0069] 本発明装置に入力された、上記両条件下に遺伝子の発光強度についての数値化データは、それぞれ本発明装置の記憶部に記憶しておき、この記憶された各条件下における数値化データを各遺伝子について呼び出し、発現量変動比(生理状態変化条件下での発現量を分子、コントロール条件下での発現量を分母として算出した値)算出プログラムを有する発現量変動比算出手段により、各遺伝子(同一遺伝子)毎にその発現量変動比を算出する。この算出には、必要に応じて、各遺伝子の発現強度による歪みの補正も含まれる。すなわち、遺伝子の発現量変動比は、発現の強度に依存して、ノイズの影響で値が強調されることがあるため、発現量変動比の分布がどの発現の強度に対してもほぼ一定となるようなバックグラウンド補正を行う。これらの発現量変動比算出プロセスには、フリーのソフトウェアであるR内のRowessアルゴリズム等が利用できる。算出された各遺伝子の発現量変動比は、本発明装置

の記憶部に記憶される。一方、この各遺伝子の発現量変動比は、上記両条件下の発現量データから予め発現量変動比を求めておき、この発現量変動量を本装置に入力し、本装置の記憶装置に記憶させてもよい。

[0070] 2) 仮想の遺伝子クラスタの構築

a) 本発明の遺伝子探索装置においては、この仮想の遺伝子クラスタの構築手段として、ゲノム上での遺伝子の連続情報及び／又は位置番号を含むゲノム上の各遺伝子の位置情報、及び仮想の遺伝子クラスタの構築を実行する仮想の遺伝子構築プログラムが格納される。

仮想の各遺伝子クラスタは、上記ゲノム上の各遺伝子の位置情報に基づき、上記仮想の遺伝子クラスタ構築プログラムを実行することにより構築される。

すなわち、仮想の遺伝子クラスタは、ゲノムDNA上に連続する遺伝子を同一方向に2個から遺伝子数を一つずつ増やして、想定される遺伝子クラスタに含まれる最大限の遺伝子数になるまで抽出され、かつ該抽出において、抽出する遺伝子の各個数毎に、直鎖状DNAからなるゲノムの場合には該DNAのいずれかの末端から、環状DNAからなるゲノムの場合には任意の遺伝子を起点として順にゲノムDNA上に配列する遺伝子一つずつずらしながら抽出された各遺伝子群からなるが、このような仮想の遺伝子クラスタを構築するため、上記仮想の遺伝子クラスタの構築プログラムは、本発明装置の記憶装置に記憶されたゲノムDNA上の各遺伝子の位置情報に基づき、以下の処理手段を実行する。その手順を図3に示す。なお、図3中、Nは、仮想の遺伝子クラスタを構成する遺伝子数を表す。

[0071] (1) ゲノム遺伝子が直鎖状ゲノムの場合、

a) ゲノムDNAの一方の末端に位置する遺伝子を起点として、他方の末端方向に、順次、ゲノムDNA上に連続する遺伝子を同一方向に2個から一つずつ増やして(N+1)、想定される遺伝子クラスタに含まれる遺伝子数の最大限(nc1)になるまで組み合わせ、起点とした遺伝子を含み、かつ遺伝子の個数の異なる複数の遺伝子群を構成する。

b) 起点を、順次、他方の末端方向に一遺伝子ずつずらせながら(起点遺伝子の移動)、上記 a. と同様の処理を行い、新たな起点遺伝子を含みかつ遺伝子の個数が異なる複数の遺伝子群を構成し、a) の遺伝子群と併せて、複数の遺伝子の組み合わせた遺伝子群からなる仮想の遺伝子クラスタを構築する。

(2) ゲノム遺伝子が環状の場合、ゲノム DNA 上の任意の遺伝子を起点として、上記 (1) a) 及び b) と同様の処理を順次行い、最初に起点とした遺伝子が起点となる時点で処理を終了する(最初に起点として遺伝子に基づく仮想の遺伝子クラスタの構築は再度行わない。)。

[0072] 上記仮想の遺伝子クラスタの構築においては、仮想の遺伝子クラスタが複数の遺伝子から構成される点で、遺伝子 2 個から一つずつ増やす手法が採用されるが、本発明は、遺伝子 1 個から一つずつ増やす手法を排除するものではない。すなわち、この場合、遺伝子一個の場合が構築される仮想の遺伝子クラスタに混入することになるが、本発明においては、この混入した遺伝子を含む遺伝子 2 個以上の組み合わせからなる仮想の遺伝子クラスタが必ず構築され、また、仮想の遺伝子クラスタのスコアは、組み合わせた各遺伝子の発現量変動比の合算であるから、ゲノム中に標的遺伝子が存在した場合、この標的遺伝子単独のスコアに比べ、これを含む仮想の遺伝子クラスタのスコアは少なくとも同等以上になり、上記混入は実質的な問題ではない。したがって、仮想の遺伝子構築において、遺伝子 2 個から一つずつ増やす手法を含む限り、遺伝子 1 個から一つずつ増やしたとしても、本発明に包含される。

[0073] なお、上記ゲノム上の各遺伝子の位置情報は、マイクロアレイによる発現量データにも同様な位置情報を付与することにより、以下の仮想の遺伝子クラスタのスコアリングの際の遺伝子照合に用いられる他、特定の遺伝子の重み付けあるいは特定の遺伝子で仮想の遺伝子クラスタを選定する際の識別手段ともなる。

一方、上記のようにゲノム上の各遺伝子の各位置情報を格納しなくとも、

例えば予めマイクロアレイ上の各DNAをゲノムDNA上の配列順に整列させておくことにより、ゲノムDNA上の遺伝子の配列順に従いそのまま入力して、入力された遺伝子の順序を遺伝子位置番号として記憶し、該位置番号を用いて仮想遺伝子クラスタを構築することもできる。

[0074] この仮想の遺伝子クラスタ構築プログラムは、コマンドに基づき組み合わせる遺伝子数の上限を設定可能なようにしてもよい。上限は検索対象の遺伝子クラスタにもよるが、ほとんどの場合、最大30個で充分である。

このようにして構築された仮想の遺伝子クラスタは記憶部に記憶される。

構築される仮想の遺伝子クラスタは、例えば、次のように、ゲノムDNA上に配列する遺伝子がA～Jの10個である場合、以下の遺伝子群からなる(表1)。

[化1]

A B C	D	E F	G H I	J
-------	---	-----	-------	---

[表1]

遺伝子数	起点 A	B	C	D	E	F	G	H	I
遺伝子 2 個の各仮想の遺伝子クラスタ(9 個)	AB	BC	CD	DE	EF	FG	GH	HI	IJ
同 3 個の各仮想の遺伝子クラスタ(8 個)	ABC	BCD	CDE	DEF	EFG	FGH	GHI	HIJ	
同 4 個の各仮想の遺伝子クラスタ(7 個)	ABCD	BCDE	CDEF	DEFG	EFGH	FGHI	GHIJ		
同 5 個の各仮想の遺伝子クラスタ(6 個)	ABCDE	BCDEF	CDEFG	DEFGH	EFGHI	FGHIJ			
同 6 個の各仮想の遺伝子クラスタ(5 個)	ABCDEF	BCDEFG	CDEFGH	DEFGHI	EFGHIJ				
同 7 個の各仮想の遺伝子クラスタ(4 個)	ABCDEFG	BCDEFGH	CDEFGHI	DEFGHIJ					
同 8 個の各仮想の遺伝子クラスタ(3 個)	ABCDEFGH	BCDEFGHI	CDEFGHIJ						
同 9 個の各仮想の遺伝子クラスタ(2 個)	ABCDEFGHI	BCDEFGHIJ							
同 10 個の各仮想の遺伝子クラスタ(1 個)	ABCDEFGHIJ								

[0075] したがって、この場合、仮想の各遺伝子クラスタの構築数は 4 5 個であるが、これらの各遺伝子クラスタは本発明の装置内でデータ処理に基づき構築されるだけであって、実験によって実際に構築されるものではない。なお、

実際のゲノムDNA上の遺伝子数は、麹菌の場合、外部データベースDOGAN (<http://www.bio.nite.go.jp/dogan/project/view/A0>)に登録されているもので12084個であり、これよりも遺伝子の定義を緩めてDNAマイクロアレイのプラットフォーム作成に使用されたものの場合14032個である。このうち連続していることが判明しているゲノム上の領域から、仮想の遺伝子クラスタを構築する。

抽出する遺伝子の数の最大限は論理上ゲノム中の遺伝子の数とすることができ、想定される遺伝子クラスタサイズの最大限の遺伝子数でよく、実際問題として、遺伝子クラスタを構成する遺伝子の数は、最大でも30個程度であり、これを超えて遺伝子クラスタを構築する必要は通常ない。

[0076] 3) 仮想の遺伝子クラスタのスコアリング

上記のように構築された仮想の各遺伝子クラスタは、本発明装置のスコアリング手段により、スコアリングされる。該スコアリング手段は、本装置の処理演算部に格納されているスコアリングプログラムにより実行される（図4）。

該プログラムは、記憶部に記憶されている、ゲノムDNA上の各遺伝子の発現量変動比データと上記構築された仮想の遺伝子クラスタ情報を呼び出して、各仮想の遺伝子クラスタを構成する遺伝子と各発現量変動比データの遺伝子を照合して、以下の計算式aを使用して各遺伝子の発現量変動比を合算して、各仮想の遺伝子クラスタのスコアを算出する手段を実行する。得られた各仮想の遺伝子クラスタのスコアは、出力されるか及び／又は記憶部に記憶される。

[0077] 計算式a)

[数1]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、 M は仮想の各遺伝子クラスタのスコア、 m はスコアリングされる仮想の各遺伝子クラスタに含まれる各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m 値)の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m 値)の標準偏差を表す。)

[0078] なお、上記式の定義中、全ての仮想の遺伝子クラスタに含まれる全遺伝子とは、全ての仮想の遺伝子クラスタを構成するために抽出されたゲノムDNA上の全ての遺伝子をいう。

本発明によれば、このようにして得られた一群の仮想の遺伝子クラスタのスコアに対する出現頻度分布をみる場合、全体としては大凡正規分布となるが、このような全体のスコア分布から離れて存在する仮想の遺伝子クラスタが存在すれば、少なくとも標的の遺伝子クラスタと対応していると判定できる。

すなわち、この仮想の遺伝子クラスタは、該クラスタ中の少なくとも2つの遺伝子が、代謝物産生誘導などの生理状態変化条件下で協働した結果、発現変動量の総量であるスコアが増大したものであり、標的の遺伝子クラスタとみなすことができ、この仮想の遺伝子クラスタ中の遺伝子は少なくとも実際の遺伝子クラスタ中に存在する代謝物産生などの生理状態変化に関与する遺伝子であると同定することができる。さらに、例えば、仮想の遺伝子クラスタ中の遺伝子及び必要に応じ代謝産物の産生機構を検討すれば、直接代謝産物の産生に関与する標的遺伝子のみではなく、未知の機能を有する遺伝子の発見も期待でき、さらに代謝物産生機構の全体像も明らかにすることができる。

[0079] 4) アノテーション付与

本発明の遺伝子探索装置においては、入力されたゲノム上の各遺伝子にアノテーションを付与する手段を設けることができる。アノテーション付与は

、ゲノム上の遺伝子が、標的とする遺伝子機能を有すると推定される場合、あるいは標的とする遺伝子機能を有する可能性が低い若しくはその可能性がないと推定できる場合において行う。

このようなアノテーション付与は、探索対象ゲノム上の各遺伝子の塩基配列情報等を基に、記憶部に記憶されたゲノム上の各遺伝子の位置情報中の遺伝子について行う。

[0080] このアノテーション付与手段は、装置使用者が、検索探索対象ゲノム上の遺伝子について予め相同検索あるいはモチーフ検索等の結果に基づき、上記記憶されたゲノム上の各遺伝子の位置情報中の遺伝子を逐一指定し、この指定された遺伝子にアノテーションが付与されるように構成しても良いが、ゲノム上の遺伝子の数は極めて多数であり、上記モチーフ検索を行う市販のソフトウェアを付属のモチーフ情報とともに本発明装置に格納するか、あるいは該ソフトウェアをモチーフ情報とともに格納した外部コンピューターと接続可能にすることが好ましい。これにより、探索対象ゲノム上の各遺伝子の塩基配列情報を本発明装置の入力手段に入力するかあるいは外部コンピューターに入力することにより、期待される機能に対応したモチーフについて検索を行い、アノテーション付与する遺伝子を自動で選定することができる。また、他のアノテーション付与手段として、上記モチーフ検索により探索対象のゲノム上の全遺伝子にてアノテーションを付与した後に、付与されたアノテーションの種類（遺伝子機能）から、期待される機能と一致する遺伝子を選定してもよい。

[0081] 選定された遺伝子は、本発明装置の記憶部に記憶されたゲノム上の遺伝子の位置情報における各遺伝子と照合される。

このようなシステムによれば、アノテーション付与を研究者の手を煩わすことなく自動で行うことができる。アノテーション付与は、機能が同様なゲノム遺伝子に付与しても良いし、機能の種類が異なる複数種の遺伝子に付与しても良い。機能が異なる複数種のゲノム遺伝子にアノテーションを付与する場合には、ゲノム遺伝子の機能毎に識別可能なように付与する。アノテ

ションによる選定の対象となる遺伝子は、例えば、2次代謝物産生に関与する遺伝子クラスタあるいはその中の遺伝子を標的とする場合、ゲノムDNAの配列中、(1) 2次代謝に関与していると想定される酵素種に属する酵素遺伝子、(2) トランスポーター遺伝子、(3) 転写因子をコードする遺伝子を選定可能である。

[0082] 5) アノテーションを付与した場合の仮想の遺伝子クラスタ・スコアリング
1

(1) 本発明の遺伝子探索装置においては、各仮想の各遺伝子クラスタのスコアリングにおいて、各仮想の遺伝子クラスタ中に該当する機能に関するアノテーションが付与された遺伝子がある場合、その遺伝子についての発現量変動比に重み付けを実行する重み付けスコアリングプログラム(図5)を格納することができる。これにより、アノテーションに基づき選定されたゲノム遺伝子についての発現量変動比は重み付けがなされ、仮想の各遺伝子クラスタのスコアリングがなされる。この重み付けスコアリングプログラムは、各仮想の遺伝子クラスタのスコアリングにおいて、アノテーションに基づき選定された遺伝子について、以下の計算式による重み付け計算手段を実行する他は、上記3)のスコアリングプログラムと同様の手段を実行する。

[0083] [数2]

$$w \frac{m - \bar{m}}{\sigma(m)}$$

(式中、 m は、上記標的とする遺伝子機能を有するか、あるいは標的とする遺伝子機能を有する可能性が低い若しくはその可能性がないと推定されるゲノムDNA上の遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m 値)の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m 値)の標準偏差、 w は重みとしての任意の実数を、それぞれ表す。)

[0084] (1) 重み w の設定は、上記標的とする遺伝子機能を有すると推定される場合、1を超えるように設定し、標的とする遺伝子機能を有する可能性が低い若しくは可能性がないと推定できる場合は、0以上1未満になるように設

定する。標的とする遺伝子機能を有するか、あるいはその可能性が低いかの推定は上記と同様に既知の遺伝子との相同性あるいはモチーフ等により判別すればよい。

(2) 一方、本発明の遺伝子探索装置においては、上記重み付けの代わりに、構築された仮想の遺伝子クラスタの中から、アノテーションに基づき選定された遺伝子を含む仮想の遺伝子クラスタを選出し、この選出された仮想の遺伝子クラスタについてのスコアリングを実行するプログラムを格納してもよい。このような手段は、上記標的とする遺伝子機能を有すると推定される場合に有効であり、例えば上記した2次代謝物産生に関与する機能遺伝子の探索等においては、特に有効である。これにより、スコアリングする仮想の遺伝子クラスタの数を削減できスコアリング時間を短縮することができる。例えば、上記表1において、アノテーション付与された遺伝子がAとCである場合、遺伝子AとCを含む仮想の遺伝子クラスタのスコアリングは、合計8個ですむ。

また、この手法により選定された仮想の遺伝子クラスタは、結果として、後記する5) アノテーションに基づき遺伝子を選定した場合の仮想の遺伝子クラスタ・スコアリング2において示される、選定された機能遺伝子により構築された仮想の遺伝子クラスタと同様なものになる場合があるが、この手法による場合、後記する一度Aの手法による網羅的な仮想の遺伝子クラスタ群を構築しておけば、自由に標的とする遺伝子あるいはこれを含む遺伝子クラスタの機能を変更でき、種々の機能選択的な遺伝子解析が容易に行える点で有利である。また、該当するアノテーションが付与されなかった遺伝子のスコアを考慮することもできるため、機能未知の遺伝子の影響が大きい場合などに柔軟に対応できる。

[0085] 本発明は、ゲノムDNA上の複数の遺伝子を組み合わせて仮想の遺伝子クラスタを構成し、これら複数の遺伝子の生理状態変化条件下の発現量変動比を合算して各仮想の遺伝子クラスタをスコアリングし、これに基づき、まず、標的の遺伝子クラスタを探索する装置に関する。スコアリングされてスコ

アが高いものが得られた場合は、仮想の遺伝子クラスタに含まれる複数の遺伝子が協働した結果であり、各遺伝子単独の発現量変動比スコアをみるよりも、全体のスコア分布に対する特異性がより鮮明となる。一方、従来のように一つ一つの遺伝子の発現変動量のみから有用遺伝子を検出する場合には、正解の遺伝子であっても、全体のスコア分布中に吸収されてしまい、高い順位の遺伝子であっても、目的の遺伝子であるか否かの遺伝子破壊実験等の検証を必要とする。

[0086] これに加え、上記したように重み付けをした遺伝子についての発現量変動比は、仮想の各遺伝子クラスタのスコアリングにおいて、他の遺伝子の発現量変動比と合算され、上記標的とする遺伝子機能を有すると推定される遺伝子を含む仮想の各遺伝子クラスタのスコアはより高くなり、反対に標的とする遺伝子機能を有する可能性が低い若しくは可能性がないと推定される遺伝子を含む仮想の遺伝子クラスタのスコアはより低くなり、全体のスコア分布との乖離が明瞭となる。したがって、これにより、標的とする遺伝子機能を有する遺伝子あるいはこれを含む遺伝子クラスタの探索がより効率的になる。

[0087] 5) アノテーション付与により遺伝子を選定した場合の仮想の遺伝子クラスタ・スコアリング2

一方、本発明の遺伝子探索装置においては、ゲノム上近傍に存在する遺伝子について、アノテーションの種類毎に、機能遺伝子を一種以上、好ましくは2種以上抽出するか、あるいはこれら遺伝子が含まれるようにゲノムDNA上の遺伝子を抽出して仮想の遺伝子クラスタとする、仮想の遺伝子をクラスタの構築手段を設けることができる。これによればスコアリングの対象となる遺伝子クラスタの数を大幅に減らすことができ、処理データ量が少なく簡便であり、2次代謝産物の産生に関与する遺伝子クラスタ及び該クラスタ中の2次代謝産物産生遺伝子の探索に特に適している。このような処理を実行するプログラム（図6）は、記憶部に記憶されたゲノム上の遺伝子の位置情報に基づき、アノテーションにより選定された遺伝子について、ゲノムD

N A 上において近傍に位置することを条件として、上記選定された遺伝子を 1 種以上、好ましくは 2 種以上抽出し、これら抽出した遺伝子により、仮想の遺伝子クラスタを構築するかあるいは少なくともこれら選定された遺伝子が含まれるようにゲノム遺伝子を抽出して仮想の遺伝子クラスタとする手段を実行する。

例えば、これら仮想の遺伝子クラスタの構築において機能遺伝子のみを組み合わせる場合、ゲノム上に配列する遺伝子数で上限 30 程度の範囲にある遺伝子であり、本発明装置においては、組み合わせる機能遺伝子の範囲の入力、設定手段を設けるとともに、上記プログラムはこれに基づき組み合わせる機能遺伝子を選択する。該プログラムは、上記遺伝子に付与されたアノテーションの種類と上記記憶部に記憶された上記ゲノム上の各遺伝子の位置情報中の位置番号により組みあわせる遺伝子を選択する。

[0088] 二次代謝産物の産生に関与する遺伝子クラスタ及び該クラスタ中の二次代謝産物産生遺伝子を探索する場合、アノテーションによる選定は、例えば、ゲノム DNA の配列中、(1) 2 次代謝に関与していると想定される酵素種に属する酵素遺伝子、(2) トランスポーター遺伝子、(3) 転写因子をコードする遺伝子を対象に行う。

例えば、以下のように、ゲノム DNA 上に配列する遺伝子が A~j の 10 個である場合、

[化2]

A * B C “	D	E F	G * H	I J “
-----------	---	-----	-------	-------

(*は該当する酵素種をコードする遺伝子、“はトランスポーター遺伝子)

仮想の遺伝子クラスタは、AC及びGJとにより構成してもよく、また、これら遺伝子が含まれるように、ABC及びGHIJで構成してもよく、さらにABCDEあるいはFGHIJのように、各仮想の遺伝子クラスタが一定数の遺伝子により構成されるようにゲノムを分割して、各仮想の遺伝子クラスタを構成しても良い。

[0089] ゲノム配列中の（１）二次代謝に関与していると想定される酵素種に属する酵素遺伝子、（２）トランスポーター遺伝子、（３）転写因子をコードする遺伝子の判別は、既知の同じ酵素種の遺伝子との相同性あるいはモチーフ等により判別すればよく、例えば、各仮想の遺伝子クラスタ中の遺伝子配列中にこれらの遺伝子が存在するか否かは、上記酵素種に属する酵素、トランスポーター、転写因子の各アミノ酸配列に特有なモチーフと共通するアミノ酸配列をコードする塩基配列が存在するか否かにより識別でき、それぞれ種類の異なるアノテーションが付与されるが、このような識別、アノテーション付与は、上記４）アノテーション付与において述べた手法を用いればよい。

[0090] 上記（１）の酵素遺伝子の判別において、酵素種は、二次代謝物の化学構造、前駆体、関与しうる補酵素、化学的・物理的性質、既知の酵素反応の事例、生産効率・速度等からその産生反応を推測し、関与する酵素種を想定するが、この酵素種の想定においては、実際にその反応に関与したであろう特定酵素のレベルまで想定しなければならないというわけではなく、該反応に関与することがより確実なレベルの酵素種でよい。例えば、オキシゲナーゼに属する酵素であることは分かるが、その下位概念の酵素種まで特定できないときは、オキシゲナーゼのレベルを酵素種として選定して、ゲノム上の各遺伝子の配列を探索し、その範疇に属する全てのゲノム遺伝子のそれぞれを、各仮想の遺伝子クラスタの構成遺伝子とすればよい。ただし、下位概念の酵素種を選定できた場合には、探索対象とする仮想の遺伝子クラスタの範囲が狭まる可能性があり、その分探索が効率的となる。

また、二次代謝物産生反応において複数の酵素が関与していると想定できる場合には、その複数の酵素種を選定することも可能である。

[0091] また、このような機能遺伝子を組み合わせた仮想の各遺伝子クラスタのスコアリングも、上記１ a) の計算式による計算において、選定された機能遺伝子の発現量変動比のみを用いて行うだけでよく、このような設定を行うだけで、上記３）仮想の遺伝子クラスタのスコアリングにおいて説明したスコ

アリングプログラムを使用できる。すなわち、この場合、計算式 1 a) の定義は、「上記式中、 M は各仮想の遺伝子クラスタのスコア、 m はスコアリングされる仮想の各遺伝子クラスタに含まれるアノテーション付与に基づき選定された各遺伝子の発現量変動比、 $m -$ は全ての仮想の遺伝子クラスタに含まれるアノテーション付与に基づき選定された全遺伝子の発現量変動比 (m 値) の平均、 $s(m)$ は全ての仮想の遺伝子クラスタに含まれるアノテーション付与に基づき選定された全遺伝子の発現量変動比 (m 値) の標準偏差を表す。」になる。

[0092] 6) スコアリング結果の表示

本発明の遺伝子探索装置においては、上記したように仮想の遺伝子クラスタンスコアリングにより算出されたスコアあるいはこれを加工した形態で、画面表示及び／または紙等の表示媒体に出力する手段を設けることができる。表示手段としては、例えば、スコアの高い順に仮想の各遺伝子クラスタを表示したり、あるいは仮想の遺伝子クラスタのスコアの分布状態を表すグラフ等があげられ、さらに仮想の遺伝子クラスタに含まれる遺伝子を表示する手段を設けることもでき、これらに基づき、仮想の遺伝子クラスタを選定することができる。

一方、スコアが高く全体分布と乖離している仮想の遺伝子クラスタは、実際に存在する標的の遺伝子クラスタに一致しないし対応する仮想の遺伝子クラスタの可能性が高い。以下に示す 7) あるいは 8) の手段は、仮想の各遺伝子クラスタのスコアの全体のスコアからの乖離の程度をみることにより、標的の遺伝子クラスタ候補を選定するか、あるいはさらに候補の絞り込みを行うための手段であり、本発明装置にこれら 7) あるいはさらに 8) の手段を設けて、乖離度を示す判定値 I (χ)、判定値 II (v) あるいは絞り込み結果 (b 値) を上記選定された仮想の遺伝子クラスタ及びその中に含まれる遺伝子とともに表示することができる。これらにより、標的の遺伝子クラスタ及び該遺伝子クラスタに含まれる標的の遺伝子を特定できる。

[0093] 7) 全体のスコア分布からの乖離の程度の算出

上記スコアリング結果の表示から、標的とする遺伝子クラスタあるいはその中の標的遺伝子を見いだすことは十分可能と考えられるが、より客観性及び効率性を高めるため、本発明装置においては、さらに仮想の遺伝子クラスタ全体のスコアの分布から乖離して存在するスコアを有する仮想の遺伝子クラスタを、標的の遺伝子クラスタ候補として選定する手段を設けることができる。本発明の装置における、このような、仮想の遺伝子クラスターのスコアの、全体分布からの乖離度を判定する手順について、図7に示す。

この候補選定手段には、仮想の遺伝子クラスタ全体のスコアの分布からの乖離の程度を示す判定値を算出する、乖離度判定プログラムが格納されている。この乖離度判定プログラムは2種あり、上記仮想の遺伝子クラスタのスコアリングプロセスにより算出されたスコアに基づき、例えば、以下の計算式b)あるいはc)に基づき、それぞれ判定値 $I(\chi)$ あるいは判定値 $I(v)$ を算出し、判定値 $I(\chi)$ あるいは判定値 $I(v)$ が、例えば、予め設定した一定値以上を示した仮想のクラスタを標的の遺伝子クラスタの候補として選定する手段を実行する(図7)。選定結果は判定値とともに出力されるが、併せて乖離度の平均値等も出力するようにしてもよい。これら2種のプログラムは、本発明装置にともに格納しても良いが、そのうち1種のみを格納しても良い。

[0094] 計算式b)

[数4]

$$\chi = -M \log P$$

(式中、 χ は上記仮想の各遺伝子クラスタの乖離の程度を示す判定値 I 、 M は仮想の各遺伝子クラスタのスコア、 P は、全ての仮想の遺伝子クラスタについての各スコアに対する出現頻度分布において、各スコア M の出現頻度について全体累計を1としたときの、上記スコア M の出現頻度の割合を表す。)

[0095] 上記計算式b)中のスコア M の出現頻度は、仮想の遺伝子クラスタの全てを含む集団における各スコアの出現頻度(P)の累計を1としたときの値であるため、1を超えることはなく、したがって $\log P$ は正になることはな

い。また、出現頻度が低いものほど $\log P$ は $-\infty$ に近づくため、頻度の低いスコア値を持つ遺伝子クラスタほど $\log P$ の絶対値は大きくなる。したがって、上記計算式 b) においては、 $\log P$ と仮想の各遺伝子クラスタのスコアを掛け合わせて -1 を乗算することにより、頻度が低くかつスコアの低いものが、より大きな判定値 $l(x)$ を持つこととなる。逆に、頻度が低くかつスコアの低いものは、より小さな負の判定値 $l(x)$ を持つ。

上記計算式 b) によれば、判定値 $l(x)$ が 0 を超え、その絶対値が高い値を示す仮想の遺伝子クラスタは、仮想の各遺伝子クラスタのスコアに対する出現頻度分布から離れており、その絶対値が高い判定値 l を示した仮想の遺伝子クラスタを標的の遺伝子クラスタあるいは標的の遺伝子クラスタに対応する候補として選定することができる。

[0096] 計算式 c)

[数5]

$$v = (M - \bar{M})^{d'} / (a\sigma(M))^{d'}$$

(式中、 v は上記仮想の各遺伝子クラスタの乖離の程度を示す判定値 l 、 M は仮想の各遺伝子クラスタのスコア、 $\bar{M}$ は全ての仮想の遺伝子クラスタのスコア (M 値) の平均値、 $\sigma(M)$ は全ての仮想の遺伝子クラスタにおけるスコア (M 値) の標準偏差、 a は任意の正の実数、 d' は次元数となる正の偶数を表す。)

[0097] この判定値 $l(v)$ は、仮想の各遺伝子クラスタのスコアについて、仮想の遺伝子クラスタ全体の平均スコアからのずれを、上記標準偏差の実数倍で割ったものを次元数 (d') 乗したもので、正規分布様のスコアに対する出現頻度分布から乖離するスコアを有する仮想の遺伝子クラスタにおいて大きな値となる。上記式中 d' は任意に設定できる正の偶数たる次元数であり、値が大きくなるほど平均スコアからの隔たりが強調されることになる。あまり大きくしすぎると、平均スコアから大きく外れたものの値が強調されて相対的にそれ以外の値が小さくなるため、通常 2 または 4 に設定する。外れたものをより鋭敏に検出したい場合は、6 以上の偶数とする。また式中の a は外れ度を表す係数で、この値を調節することにより、上記正規分布様分布

からどの程度乖離したものをとるかを調節することができる。1を超えて大きく設定するほど、平均スコアから大きく外れたもの以外の v 値はゼロに近付くため、この a 値は通常1～2に設定する。逆に1未満の場合、より外れ方の小さなものも拾うことができる。

この計算式c)による場合も上記判定値Iと同様に、 v が0を超え、高い値を示す仮想の遺伝子クラスタを標的の遺伝子クラスタあるいは標的の遺伝子クラスタに対応する候補として選定することができる。

[0098] 8) 遺伝子クラスタ候補の絞り込み

上記計算式b)、c)により算出された判定値(χ あるいは v)により、標的の遺伝子クラスタ候補となった仮想の遺伝子クラスタの数が多く、さらに候補を絞り込みたい場合に備えて、本発明装置においては、遺伝子クラスタ候補絞り込み手段として以下の計算式d)による計算を行う、候補絞り込みプログラムを格納することができる(図8)。すなわち、各仮想の遺伝子クラスタについて、判定値IおよびIIの積を取った値について、 b が100未満の仮想のクラスタを少なくとも除外することにより、標的の遺伝子クラスタ候補をさらに絞り込むことが可能である。

[0099] 計算式d)

[数10]

$$\chi \times v > b$$

(式中、 χ は【数4】の計算式b)によって算出した仮想の各遺伝子クラスタの判定値I、 v は【数5】の計算式c)によって算出した仮想の各遺伝子クラスタの判定値II、 b は閾値としての任意の正の実数を表す。)

[0100] 上記計算式d)中、 b はどの程度の遺伝子クラスタ候補を絞り込むかを決定するための閾値であり、 b を大きくとるほど候補の絞り込み効果がより高くなる。また小さくとるほど多くの候補遺伝子クラスタを選択することができる。 b の値の設定は対象とする生物種や培養条件に依存する。すなわち、候補遺伝子クラスタが強かつ多く発現している系であれば値を高くする必要はあるが、逆に発現強度が弱かつ数が少なければ値を低くしなければ候

補遺伝子が出現しない。前者の場合、例えば5000～10000あるいは10000～30000の範囲内の任意の数値に設定し、後者の場合、通常100以上、例えば1000～2000、あるいは2000～5000の範囲内の任意の数値に設定する。

[0101] 9) 本発明の装置を用いて正解が得られなかった場合

一方、本発明の手法を行った結果、仮に仮想の遺伝子クラスタ全体のスコア分布から乖離したスコアの遺伝子クラスタが見いだされなかった場合、設定する生理状態変化条件、重み付けするゲノムDNA上の遺伝子の選定、あるいは上記B)の手法による仮想の遺伝子クラスタ構築のためのゲノムDNA上の遺伝子の選定等の探索条件設定に問題点がある。したがって、このような場合には、探索条件を再設定して、バックグラウンドの分布から離れたスコアの遺伝子クラスタが見つかるまで、上記した遺伝子クラスタの探索法を繰り返し行えばよい。すなわち本発明においては、得られたデータのみから、探索条件設定の問題点を把握できる。

これに対して、上記したような従来法の場合には、もともと正解の遺伝子であっても、遺伝子全体の発現量についての分布中に埋もれてしまうので、得られたデータからでは正解か否かは不明であり、結果的に無意味かもしれない検証実験を繰り返さなければならない。

[0102] B) 遺伝子クラスタ予測装置

一方、本発明における上記仮想の遺伝子クラスタの構築手段及びそのスコアリング手段を用いた他の態様として、標的とする遺伝子クラスタの有無及び標的とする遺伝子クラスタが存在する場合のサイズ（クラスタを構成する遺伝子数； n_{cl} ）を推定する装置（以下、遺伝子クラスタ予測装置という。）を挙げることができる。本発明の装置における、この遺伝子クラスタ予測装置の概要を図9に示す。

この遺伝子クラスタ予測装置においては、まず、生物細胞の生理状態変化を生じる条件とコントロール条件下において生じたゲノムDNA上に配列する遺伝子の発現量変動比を合算し、仮想の遺伝子クラスタのスコアとするが

、ゲノムDNA上に配列する各遺伝子の発現量データの入力、発現量変動比の計算、仮想の遺伝子クラスタの構築、及び仮想の各遺伝子クラスタのスコアリングの各手段は、上記1)～3)に記載した手段と同様である。

[0103] すなわち、この装置は、本発明の上記遺伝子探索装置における、a) 生物細胞の生理状態変化を生じる条件とコントロール条件下において生じたゲノムDNA上に配列する各遺伝子の発現量を入力する手段、b) 入力された上記2つの条件下における同一遺伝子の発現量の比を算出する発現量変動比算出手段、c) ゲノムDNA上に配列する各遺伝子の発現量変動比を複数の遺伝子により構築された上記仮想の遺伝子クラスタ単位の発現量変動比として合算し、仮想の遺伝子クラスタ単位毎にスコアリングする手段を有するものであって、仮想の遺伝子クラスタの構築手段が、ゲノムDNA上に連続する遺伝子を2個から遺伝子数を一つずつ増やし想定される遺伝子クラスタに含まれる最大限のゲノム遺伝子数になるまで抽出し、かつ該抽出において抽出する遺伝子の各個数毎に、直鎖状DNAからなるゲノムの場合には該DNAのいずれかの末端から、あるいは環状DNAからなるゲノムの場合には任意の遺伝子を起点として順にゲノムDNA上に配列する遺伝子一つずつずらしながら抽出された各遺伝子群を仮想の各遺伝子クラスタとする手段であり、スコアリング手段として、以下の計算式a) による計算を行うプログラムを格納している点では、本発明の遺伝子探索装置と共通する。この装置の特徴点は、上記1～3)の手段のプロセスを行い、出力された仮想の各遺伝子クラスタのスコアに基づき、d) 仮想の遺伝子クラスタに含まれる遺伝子数単位毎の遺伝子クラスタ分布判定値(ε)を算出する手段にあり、この手段を実行するプログラムとして遺伝子クラスタ分布判定値(ε 値)算出プログラムを格納した点にある(図9)。

[0104] 計算式a)

[数1]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、 M は仮想の各遺伝子クラスタのスコア、 m はスコアリングされる仮想の各遺伝子クラスタに含まれる各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値) の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値) の標準偏差を表す。)

この遺伝子クラスタ分布判定値 (ε) は、以下の計算式 e) により求められる。

[0105] 計算式 e)

[数7]

$$\varepsilon = \sum (M - \bar{M})^d / n \sigma(M)^d$$

(式中、 ε は各遺伝子数単位毎に求められた遺伝子クラスタスコア分布判定値、 M は全ての仮想の遺伝子クラスタを遺伝子数単位毎に分けた場合の、各遺伝子数単位に含まれる仮想の各遺伝子クラスタのスコア、 $\bar{M}$ は全ての仮想の遺伝子クラスタのスコアの平均、 n は仮想の遺伝子クラスタの総数、 $\sigma(M)$ は全ての仮想の遺伝子クラスタのスコア (M 値) の標準偏差、 d は任意に設定する次元数としての正の偶数を表す。)

[0106] この計算式 e) によれば、仮想の遺伝子クラスタが、実際のゲノム DNA においてクラスタを形成していなければ、仮想の遺伝子クラスタ中に含まれる、標的の生理状態変化に関与せず、発現量変動のない遺伝子の影響を受けるので、仮想の遺伝子クラスタのスコア (M) はサイズ (遺伝子数 ; ncl) が大きくなるほど平均化され、すなわちスコアの平均値に近づくためサイズの増大に伴い ε 値は単調減少する (図 10 上から 1 および 3 番目の曲線参照)。しかし、仮想の遺伝子クラスタのあるサイズのものがクラスタを形成している場合、そのサイズにおいて、分布の偏り ε は大きくなり、上記単調減少曲線とはならず、 ε 値はそのサイズにおいて特異点を形成する (図 10 矢印で示した点参照)。したがって、 ε 値が特異点を形成するか否か、及び特異点を形成した遺伝子クラスタのサイズから、遺伝子クラスタの存在及び

そのサイズを推定することができる。

[0107] 具体的には、仮想の遺伝子クラスタの該クラスタに含まれる遺伝子数毎の集計において、ある遺伝子数（ k 個）のときの ε 値（ $\varepsilon(k)$ ）と、その前後数のときの ε 値（ $\varepsilon(k-1)$ 、 $\varepsilon(k+1)$ ）が以下の関係にあれば、標的とする遺伝子クラスタがゲノム中に存在すると判定し、標的遺伝子クラスタに含まれる遺伝子数を k 個と予想することができる。

[0108] [数8]

$$\varepsilon(k) > \varepsilon(k-1) \quad \text{かつ} \quad \varepsilon(k) > \varepsilon(k+1)$$

[0109] 本発明の遺伝子クラスタ予測装置は、上記 a) ～ d) の手段を具備した独立の装置としても構成しても良いが、上記 a) ～ c) の手段が、本発明の遺伝子探索装置と共通するため遺伝子数単位毎の遺伝子クラスタ分布判定値（ ε ）を算出する手段を、本発明の遺伝子探索装置にさらに設けて、本発明の遺伝子探索装置に標的の遺伝子クラスタの有無および遺伝子クラスタのサイズ予測機能を付加しても良い。このような予測機能は、本発明の遺伝子探索装置を用いて、選定した複数種の機能遺伝子を組み合わせて仮想の遺伝子クラスタを構築し、そのスコアリングを行う際、予め行う手法として有効である。すなわち、遺伝子クラスタが存在し、そのサイズを予想できれば、予想されるサイズ内に、標的とする酵素種に属する酵素遺伝子、（2）トランスポーター遺伝子、（3）転写因子をコードする遺伝子が存在するゲノム配列のみを対象として、上記仮想の遺伝子クラスタとして探索することが可能となる。

[0110] また、この遺伝子クラスタ予測装置によれば、ある条件下で細胞が何らかの生理的状态変化を起こす場合においては、どのような生理的状态変化であっても変化を対比する条件が設定できれば、その原因遺伝子はもちろんその変化を生じる機構そのものが全く不明である場合においても、その変化原因が遺伝子クラスタ中の遺伝子の連関にあるのか否か、遺伝子クラスタ中の遺伝子の連関による場合該クラスタの遺伝子サイズも容易に予測できる。すな

わちこの手法は、生物の生理的变化が、極めて探索の難しい複数の遺伝子の連関によって生じている場合において、その原因が遺伝子クラスタ中の遺伝子の共働によるものであることを明らかにでき、かつそのサイズも予測できる点で極めて有用である。

実施例

[0111] 参考例 1

コウジ酸の産生に必須の遺伝子の同定

本参考例は、本発明による遺伝子の探索、同定の有利性を明らかにするため、まず従来法によるアスペルギルス・オリゼのコウジ酸産生遺伝子の探索、同定手法について示すものである。

アスペルギルス・オリゼ (*Aspergillus oryzae*) の菌株RIB40 (以下、単にアスペルギルス・オリゼと書いた場合にはこの菌株を指す) は、以下の組成の液体培地中で、30℃、150回転/毎分の条件下で生育させた場合、コウジ酸を培地中に産生する。500 mL のこぶつき三角フラスコ中に250 mL の培地を入れ、アスペルギルス・オリゼの孢子懸濁液を $1.05 \sim 1.07 \times 10^7$ / mL になるように接種する。

[0112] (培地組成：以下コウジ酸産生培地と呼ぶ)

10% (W/V) グルコース

0.25% (W/V) イーストエキストラクト (Yeast Extract)

0.1% (W/V) K_2HPO_4

0.05% (W/V) $MgSO_4 \cdot 7H_2O$

pHを6.0に調整後、オートクレーブにより滅菌する。

[0113] アスペルギルス・オリゼが上記の培養によりコウジ酸を産生することは、コウジ酸と塩化第二鉄とのキレート化合物の生成による赤色の発色により検出することが可能である。また培養の上清などを適宜希釈した試料に、最終濃度10 mM程度になるように高濃度の塩化第二鉄溶液を添加した液を作成し、波長500 nmの吸光度を測定することで、コウジ酸量の定量測定が可能である。この波長500 nmの吸光度は、0.1～1.0程度の範囲でコウ

ジ酸濃度に比例する。

このような検出法によれば、接種後3または4日目には産生を検出することが可能であり、少なくとも7日目には十分な速度をもってコウジ酸の産生が行われている。またコウジ酸の産生は、上記の産生培地に0.1% (W/V) 以上の硝酸ナトリウムを加えることで阻害される。この硝酸ナトリウムによる阻害は可逆的である。硝酸ナトリウムの添加によって阻害された菌糸を、培地成分の洗浄後、新たに用意した産生条件を満たす培地に移すことによって、菌はコウジ酸の産生を開始する。

[0114] アスペルギルス・オリゼのコウジ酸の産生量が異なる下記に記載の条件からなるC1～C3の3つの系で、ゲノム中にコードされたほとんどの遺伝子の網羅的な発現の解析を、DNAマイクロアレイを用いた実験により比較した。

C1. 上記コウジ酸産生培地で、4日間および2日間生育させた菌体の遺伝子の発現を比較した(4日目/2日目)。

C2. 上記コウジ酸産生培地で、7日間および4日間生育させた菌体の遺伝子の発現を比較した(7日目/4日目)。

C3. 上記コウジ酸産生培地に0.3% (W/V) の硝酸ナトリウムを添加してコウジ酸産生を阻害した菌体と、上記コウジ酸産生培地で生育させた菌体を比較した。どちらも4日間、30℃、150回転/分の条件で生育させた(NO_3^- なし/あり)。

[0115] 上記各系での菌体の遺伝子発現を、DNAマイクロアレイを用いて解析した結果、系C1～C3のそれぞれにおいて、比較する条件下で培養した2つの菌体での各遺伝子の発現量の比、および発現の強度に相当する値が得られた。各々比較する条件間で、コウジ酸の産生がより顕著な条件で、発現がより顕著になっている遺伝子を抽出するために、以下の手続きにより、候補を抽出した。

発現量の比、および発現の強度に相当する値は、それぞれで正規分布に近い分布をするが、値の絶対値は大きな違いがある。この両者を統合して候補

を抽出するために、発現量の比、発現の強度は、それぞれで値の正規化を実施した後で、比較した。それぞれ正規化した発現量の比、および発現の強度に相当する値の積を作成した。その積が高いほど、コウジ酸の産生に関係する可能性が高いと考え、それぞれの実験で高い積の値をもつもの上位5つを選び出した（表2）。

[0116] [表2]

系	順位	遺伝子ID	発現比 *1	発現量 *2	積	推定された機能
C1	1	AO09000500081 2	4.08	1.98	8.05	1,4-ベンゾキノンレダクターゼ様; Trp リプレッサー結合タンパク質様 /プロトプラスト分泌タンパク質
	2	AO09001000054 1	2.59	2.65	6.86	ニトロレダクターゼに関連する推測 オキシドレダクターゼ
	3	AO09000300109 6	2.86	2.00	5.73	酸化損傷の防御に関与する推測 カルボニックアンヒドラーゼ
	4	AO09000100031 0	5.28	1.07	5.66	フェニルクマランベンジルエーテル レダクターゼ [Populus balsamifera subsp. trichocarpa]
	5	AO09002600013 7	2.53	2.05	5.18	Drosophila melanogaster LD33051p [Plasmodium yoelii yoelii]
C2	1	AO09011300013 6	5.36	2.81	15.03	機能は推定されていない
	2	AO09011300013 8	5.63	1.34	7.54	シナプス小胞トランスポーター SVOPおよび関連トランスポーター
	3	AO09010300002 0	2.40	1.40	3.35	グルコサミン-6-リン酸イソメラーゼ
	4	AO09012000011 2	1.00	3.22	3.24	ペルオキシレドキシン
	5	AO09000300089 5	1.22	2.57	3.13	60s リボソームタンパク質 L15
C3	1	AO09001100001 0	2.49	1.50	3.74	機能未知タンパク質 [Neurospora crassa]
	2	AO09001000022 7	3.78	0.54	2.05	未同定保存タンパク質
	3	AO09001000077 6	1.23	1.20	1.48	液胞ソーティングタンパク質 VPS1, ダイナミンおよび関連タンパク質
	4	AO09001100041 4	0.27	4.09	1.09	グリセルアルデヒド 3-リン酸デヒド ロゲナーゼ
	5	AO09000300087 3	0.55	1.79	0.97	NADH-シトクロム b-5 レダクター ゼ

*1: 発現比は、コウジ酸の生産がより顕著な条件で、遺伝子の発現が高いものを正の大きな数となるようにした。

*2: 発現量は、比較する2つの条件でのシグナルの値を対数で表して足した値と関係する。発現量が高いものが大きな正の数となるようにした。

アスペルギルス・オリゼのDNAマイクロアレイにおける発現比と発現量の積からなるスコア上位遺伝子

[0117] 表2に示した遺伝子は、それぞれの比較する二つの条件下で、コウジ酸の産生条件において顕著に発現が高くなっている遺伝子である。すなわちコウ

ジ酸の産生に必須の遺伝子である可能性が高い遺伝子である。これらの遺伝子について、上位のものから遺伝子欠失破壊実験を行った。

ここで上記C 1～C 3の3つの系は、いずれもコウジ酸の産生量が有意に異なる2つの条件を比較したものである。したがって理想的には、どの系においてもコウジ酸産生に必須の遺伝子が上位に現れると予想した。しかし現実には、3つの系全てにおいて上位にくる遺伝子はなかった。したがっていずれの系においても、上位に来るものはコウジ酸の産生に必須であるか、または各条件に特異的に誘導される遺伝子である可能性の両者を含む。これらの中からコウジ酸の産生に必須の遺伝子を選び出すために、各系において上位に来ている候補遺伝子のいずれかを破壊して変異体を作製し、当該変異体のコウジ酸産生能を解析した。

[0118] その結果、A0090113000136およびA0090113000138の両遺伝子は、破壊によりコウジ酸の産生が著しく低下することが判明した。上記2つの遺伝子は、他の生物種のゲノム中にある機能が既知の遺伝子とのオーソログな関係を持たないため、ゲノム情報から両遺伝子の機能を知ることはできなかった。ただし、該遺伝子のアミノ酸配列には散在する既知の配列モチーフが存在し、機能の概略を予測することが可能であった。A0090113000136の遺伝子は、FAD依存性の酸化還元酵素のモチーフを持っている。これはグルコースからのコウジ酸への変換を考えたとき、変換の過程で複数の酸化還元反応が関係していると予想されていることから、この遺伝子がコウジ酸の生合成における酵素であることを強く示唆している。一方で、A0090113000138は膜輸送に関わる配列モチーフを持っており、Major facilitator superfamilyと分類される。コウジ酸の生合成に伴って産生されたコウジ酸が培地中に分泌されることは明確であり、この遺伝子がコウジ酸の産生に必須であることを示唆している。

[0119] この両遺伝子はゲノム上で近傍に位置する。間には1遺伝子しか存在せず、そのA0090113000137遺伝子のアミノ酸配列は転写因子のモチーフを持つことが判明した。この遺伝子の破壊によってもコウジ酸の産生が著しく低下す

ることが判明した。

以上の解析により、A0090113000136、A0090113000137、A0090113000138の3つの遺伝子がコウジ酸の産生に必須の遺伝子であると同定された。本同定過程には、培養条件の検討などを除いて、およそ1年の時間を要した。

このようにして同定されたコウジ酸の産生に必須の3つの遺伝子について、系C1～C3におけるDNAマイクロアレイの結果において、その発現量変動比m値が全遺伝子中どの位置に来るかを表3にまとめた。

[0120] [表3]

系	A0090113000136	A0090113000137	A0090113000138	最小値	最大値
C1	0.96	0.45	0.92	-9.23	6.59
6692 遺伝子中 ^{*1}	1427位	2311位	1489位		
C2	8.16	3.24	6.97	-10.84	8.16
5595 遺伝子中 ^{*1}	1位	71位	6位		
C3	0.15	0.31	-0.2	-20.32	18.52
6488 遺伝子中 ^{*1}	3070位	2658位	3945位		

^{*1}14032遺伝子の内、シグナルがあるものの数(約半数はシグナルが無く、棄却されている)

アスペルギルス・オリザにおけるコウジ酸産生に必須の3つの遺伝子のスコアm値とその順位

[0121] また系C1～C3における各分布を図11から図13に示した。表3にあるように、系C2では該当する3つの遺伝子は1位、6位、71位と上位にあり、この系のアレイであれば必須遺伝子の同定は比較的容易である。一方系C3では、コウジ酸の産生が顕著に見られるにも関わらず、必須の遺伝子の値は最高でも2658位であるなど、上位には見られない。このアレイをもとにした場合、従来法で遺伝子を特定することは事実上不可能である。その上、必須の遺伝子が分かっていない状況では、どのアレイが正解を与えるものであるかの判断ですら困難である。ここに示した3つのアレイデータだけを用いて3つの遺伝子がコウジ酸の産生に必須であることを同定することは、上記に示した方法で可能ではあったが、偶然的な幸運の要素が大きく一般性は低い。機能の注釈を元に推定する場合でも、100以上の遺伝子を

破壊してみなければ分からない可能性があり、この場合、検証に通常3年程度以上はかかる。

[0122] 実施例1

アスペルギルス・オリゼにおける遺伝子クラスタ・スコアリングによるコウジ酸合成遺伝子同定

本特許で出願する該当遺伝子の同定手法に従い、本発明装置を使用して、アスペルギルス・オリゼのコウジ酸産生関連遺伝子からなる遺伝子クラスタを同定した。

この実験に使用した装置は、データの入出力装置、入出力インターフェース、記憶装置、制御演算装置（CPU）から構成され、上記制御演算装置は、発現量変動比算出部、仮想の遺伝子クラスタ構築部、仮想の遺伝子クラスタのスコアリング部、仮想の遺伝子クラスタの乖離度判定値算出部、遺伝子クラスタ候補絞り込み部、及び遺伝子クラスタ予測部を有し、これら各部には、それぞれ順に、発現量変動比算出プログラム、仮想の遺伝子クラスタ構築プログラム、仮想の遺伝子クラスタ・スコアリングプログラム、乖離度判定値（ χ ）および（ ψ ）算出プログラム、候補絞り込みプログラム並びに遺伝子クラスタ分布判定値（ ε ）算出プログラムが格納されている。

また、これら各部での計算は、Linuxオペレーティングシステム上で、フリーソフトウェアR、およびプログラム言語Perlを用いて行った。

DNAマイクロアレイのデータは、参考例1と同様のものを使用した。すなわち以下の、コウジ酸を産生する培養条件を分子に、コントロールとなる培養条件を分母にして測定した、以下のC1～C3の系における二色法データである。

[0123] C1. 4日目／2日目

C2. 7日目／4日目

C3. NO_3^- なし/あり

これらは各々、産生条件と非産生条件の下で生育させた菌体からmRNAを取り出し、それぞれ色素でラベリングした後でアレイ上のオリゴDNAに

ハイブリダイズすることによりデータを得て、そこから各遺伝子の発現量変動比(m値)を得た。

具体的には、上記系C 1～C 3における各産生条件と非産生条件の下で生育させた菌体からmRNAを取り出し、産生条件と非産生条件から取り出したmRNAをそれぞれ異なる蛍光色素でラベリングした後でアレイ上のオリゴDNAにハイブリダイズさせ、それぞれの検出波長強度情報を入力し、発現量変動比算出部に格納された発現量変動比算出プログラムを適用して各遺伝子の発現量変動比(m値)を得た。

このDNAマイクロアレイの実験においては、14032個のプロブからなるプラットフォームを用いたが、その全てに対応する遺伝子が発現し値を取れるわけではない。そこで本実施例では、3つの系に共通して発現が確認された5179個の遺伝子についての発現強度情報を用いた。

[0124] (A) クラスタ・スコアリング

記憶部に記憶されたアスペルギルス・オリゼのゲノムDNA上の各遺伝子の位置情報に基づき、仮想の遺伝子クラスタ構築部に格納された仮想の遺伝子クラスタ構築プログラムを適用して、遺伝子サイズを1～30と設定し仮想の遺伝子クラスタを構築した。なお、本実施例及び以降の実施例においては、個々の遺伝子を探索する従来法に対する本発明による探索法の有利性を検証するため、仮想の遺伝子クラスタの構築においては、遺伝子サイズを1～30と設定し、順次遺伝子1個から遺伝子数を1つずつ増やしながら30個になるまで行ったが、2個以上の遺伝子の組み合わせからなる仮想の遺伝子クラスタのスコアリングに加え、遺伝子数1の場合のスコアリングも行っている。

[0125] 一方、系C 1～C 3において共通に発現が確認された5179個の遺伝子の発現量変動比を、上記構築された仮想の遺伝子クラスタに含まれる各遺伝子と照合して、仮想の遺伝子クラスタのスコアリング部のスコアリングプログラムを適用して、計算式a)に従って、上記構築された各仮想の遺伝子クラスタをスコアリングし、スコア(M値)を得た。なお、系C 1～C 3にお

いて共通して発現が確認されずシグナルが検出されなかった遺伝子については、仮想の遺伝子クラスタの構成要素としてはカウントするが、値は入れずに計算を行った。また、ゲノム上の末端側に位置する遺伝子については所定個数（1～30個）の遺伝子を組み合わせられないが、この場合においては、組み合わせうる最大個数の遺伝子でスコアリングを行った。このようにしても該遺伝子クラスタの推定には本質的に影響はあたえない。

[0126] 系C1～C3のそれぞれについて、クラスタ・スコアリングを行い、計算式a)に従って、各仮想の遺伝子クラスタのスコア（M値）を得た。得られたスコアは系C1～3のそれぞれについて記憶部に記憶した。

図14はそのヒストグラムである。左の拡大図をみると分かるように、ゼロを中心とした山型の正規分布様集団から外れて高いM値を持つ仮想の遺伝子クラスタがあると、全体を表したヒストグラムにおいて山の中心が左側にずれる。

[0127] (B) データ判定

計算式e)にしたがって、系C1～3における遺伝子クラスタスコア分布判定値 ε を算出した（図15）。

具体的には、本発明の装置に記憶された各仮想の遺伝子クラスタのスコアを呼び出し、遺伝子クラスタ予測部に格納されている遺伝子クラスタ分布判定値（ ε ）算出プログラムを適用し、計算式e)にしたがって、系C1～3における遺伝子クラスタスコア分布判定値 ε を算出した（図15）。該算出に当たり計算式e)における仮想の遺伝子クラスタの数 n は5179とし、上記発現量データを伴う遺伝子5179個中の遺伝子の一つも含まれない仮想の遺伝子クラスタは除外した。また、次元数 d は6を採用した。

図にあるように、C1～3のいずれの系においても ε 値は基本的に単調減少しており、クラスタ・スコアリングによる平均化の影響が見て取れる。しかし系C2において、 $n_{c1}=3$ のとき ε 値はいったん増加に転じており、次の $n_{c1}=4$ において再び減少している。すなわちこの点において、 ε 値は隣り合う二点よりも大きいため、[数6]より、系C2において、標的と

する遺伝子クラスタがゲノム中に存在し、その遺伝子クラスタに含まれる遺伝子数は3個であると推定された。

以上の結果をふまえ、系C 2のDNAマイクロアレイデータを用いて、以下の検証、同定実験を行った。

[0128] (C) 遺伝子クラスタ判定

系C 2のDNAマイクロアレイデータに基づき算出した上記各仮想の遺伝子クラスタのスコア (M値) に基づき、計算式b)にしたがって遺伝子クラスタの判定値 χ を算出した (図16)。

具体的には、本発明の装置に記憶されている系C 2における各仮想の遺伝子クラスタのスコアに、仮想の遺伝子クラスタの乖離度算出部に格納された仮想の遺伝子乖離度判定プログラムのうち χ 値算出プログラムを適用し、計算式b)にしたがって、各仮想の遺伝子クラスタについての判定値 χ を算出した。

なお、図16中の各折れ線は、仮想の遺伝子クラスタ構築において起点となる遺伝子が共通する遺伝子サイズ1～30の各仮想の遺伝子クラスタの判定値を結んだものである (図17、18、21、23、30～32、35～37も同様)。

ここで、 $ncl = 1$ のときの値が $ncl = 2$ のときの値より大きい仮想の遺伝子クラスタは、本手法におけるクラスタ・スコアリングによってスコアを上げているわけではないので、該当しない。また $ncl = 1$ のときの値が負の仮想の遺伝子クラスタは、本手法におけるクラスタ・スコアリングにおいて、そのスコアの上昇に寄与しないため、該当しない。そこで図13においては、これらのものを除外してある。

[0129] 図16にあるように、1つの仮想遺伝子クラスタが $ncl = 3$ で極大かつ最大値を取ることが分かる。この仮想の遺伝子クラスタは、コウジ酸産生に必須の3つの遺伝子、A0090113000136、A0090113000137、A0090113000138のみを含んだものであった。

この結果は参考例の結果と一致し、上記遺伝子クラスタ分布判定値 (ε)

算出による予測結果が正しいことが分かる。また、判定値 χ によって、標的とする遺伝子クラスタ及び該クラスタに含まれる遺伝子も同定可能なことが明らかとなった。

続いてもう一つの遺伝子クラスタ判定値である ψ を、上記と同様の各仮想の遺伝子クラスタのスコアに、各仮想の遺伝子クラスタについて、仮想の遺伝子クラスタの乖離度算出部に格納された遺伝子乖離度判定プログラムのうち ψ 値算出プログラムを適用し、計算式 c) にしたがって算出した (図 17) 。ここでも、 χ 値と同様、 $ncl = 1$ のときの値が $ncl = 2$ のときの値より大きい仮想の遺伝子クラスタは除外してある。ここで次元数 d' は 2、係数 a は 1 を採用した。すると図のように、 $ncl = 3$ のとき極大かつ最大値をとる仮想遺伝子クラスタが一つある。これは χ 値のときと同様、コウジ酸産生に必須の 3 つの遺伝子のみを含む遺伝子クラスタである。したがって、判定値 ψ によっても標的とする遺伝子クラスタ及び該クラスタに含まれる遺伝子が同定可能であることが明らかとなった。

[0130] こうして得られた χ 値および ε 値に対し、遺伝子クラスタ絞り込み部に格納された候補絞り込みプログラムを適用して、計算式 d) に従って二つの値の積から遺伝子クラスタ評価値を算出した (図 18) 。図 18 をみると明らかのように、5000 以上の値で最大値を取る仮想遺伝子クラスタが極めて明確に一つ存在しており、 $ncl = 3$ で極大値をとっている。これがコウジ酸産生に必須の 3 つの遺伝子、A0090113000136、A0090113000137、A0090113000138 のみを含んだものである。このように本発明の手法および装置を用いることで、標的とする生合成遺伝子を同定することができた。また計算式 d) における閾値 b を例えば 2000 とすれば、該当する遺伝子クラスタは 4 個しかなく、実験系による検証する場合においても容易に行える数値である。 χ 値 (図 16) と ψ 値 (図 17) を乗算することで、それぞれにおいて存在していた多くのピークがキャンセルされ、探索対象に該当するもののみ高い値を示している。

以上から、本発明の手法および装置は、DNA マイクロアレイデータのみ

を用いて、ゲノム上に集合して機能を果たす生合成遺伝子の探索、同定を可能とする実効的な手段であることが示された。

[0131] 実施例 2

アスペルギルス・オリゼにおけるアノテーション（機能注釈）による重み付けを行った場合の仮想の遺伝子クラスタ・スコアリングによるコウジ酸合成遺伝子の探索

アスペルギルス・オリゼのコウジ酸産生関連遺伝子からなる遺伝子クラスタを同定することを目的として、予測される機能に関連した注釈のついた遺伝子の m 値に重み付けをした後、該当遺伝子の同定を行った。

この実験に使用した装置は、上記実施例 1 に記載した装置と基本的に同様であるが、アノテーションによる遺伝子選定部、選定された遺伝子に対する発現量変動比の重み付与部を有している点で異なる。

コウジ酸産生に必要な機能は以下の 3 つの機能を選出した。

- ・膜輸送体：transporter または major facilitator
- ・転写制御因子：transcription
- ・酸化還元酵素：oxidoreductase または dehydrogenase

なお、上記英単語は、アノテーションによる遺伝子選定に用いたキーワードである。

[0132] これらを選出した理由として、コウジ酸の生合成がグルコースから酸化により変換されていると推定されること、産生されたコウジ酸の膜輸送による培地中への分泌に膜輸送体が必要と推定されること、および関与する遺伝子の転写制御に転写因子が必要と推定されることが挙げられる。

[0133] (A) アノテーション（機能注釈）による重み付けとクラスタ・スコアリング

一般に利用されているアノテーション推定ソフトウェアシステムの一つである Interproscan (<http://www.ebi.ac.uk/Tools/InterProScan/>) を用いて、アスペルギルス・オリゼのゲノム DNA 上の各遺伝子についてアノテーションを付与し、その結果付与されたアノテーションに基づき、上記 3 つの機

能に該当する遺伝子を選出した。具体的には、該各遺伝子についてのアノテーションデータを本装置の入力装置に入力し、記憶装置に記憶した。記憶したアノテーションデータのデータを呼び出し、上記３種の機能を有する遺伝子を機能遺伝子選択部の選択プログラムを適用して選定した。なお、選定は各遺伝子について付与されたアノテーション中に上記３つの機能群に対応する英単語が含まれるかどうかで行い、その結果、該当した遺伝子は５１７９個のうち７０９個であった。

続いてこれらの該当する注釈のついた遺伝子に対して、実施例１に記載した３つのアレイ測定の系Ｃ１～３のそれぞれについて、その発現量変動比（ m 値）に正規化後重み $w = 2.0$ を積算した後、計算式 $a)$ にしたがって $ncl = 1 \sim 30$ でクラスタ・スコアリングを行い、各仮想の遺伝子クラスタの M 値を得た。

具体的には、このように選定された遺伝子についての発現量変動比は、重み付与部により重み付け（〔数２〕参照）がなされ、この重み付けされた発現量変動比を用いて、各仮想の遺伝子クラスタのスコアが算出される。アノテーションに基づき選定された遺伝子の発現量変動比に重み付けする以外は、仮想の遺伝子クラスタの構築プログラム、スコアリングプログラム自体、実施例１と相違しない。この実験においては、発現量変動比（ m 値）に正規化後、アノテーションにより選定された遺伝子の発現量変動比には、重み $w = 2.0$ を積算し、仮想の遺伝子クラスタのスコアリング部に格納されたスコアリングプログラムを適用し、計算式 $a)$ にしたがって $ncl = 1 \sim 30$ でクラスタ・スコアリングを行って各仮想の遺伝子クラスタのスコア（ M 値）を得た。なお、算出された各仮想の遺伝子クラスタのスコアは、本発明装置の記憶装置に記憶した。

図１９は、上記算出された各仮想の遺伝子クラスタスコアのヒストグラムである。左の拡大図を図１４と比較すると、重み付けによってより高いスコアが出現したために、相対的にゼロを中心とした山型の分布がより尖って見え、かつ山の中心がより左側にずれていることが分かる。

[0134] (B) データ判定

続いて計算式 e) にしたがって、系 C 1 ~ 3 におけるスコア分布評価値 ε を算出した (図 20)。具体的には、上記 (A) により算出され、記憶された各仮想の遺伝子クラスタのスコアを呼び出し、遺伝子クラスタ予測部に格納されている遺伝子クラスタ分布判定値 (ε) 算出プログラムを適用し、計算を行った。ここで実施例 1 と同様に、仮想遺伝子クラスタ数 n は 5179、次元数 d は 6 を採用した。図 20 にあるように、系 C 1 および C 3 では ε 値は基本的に単調減少しているのに対し、系 C 2 では $n_{cl} = 3$ のとき ε 値は大きく増加し極大値を示す。その値は実施例 1 におけるもの (図 15) の 10 倍以上であった。すなわち機能の注釈による重み付けによって、より機能限定的に高精度で該当遺伝子クラスタの存在及びその遺伝子数が予測可能であることを示している。

この実験によって、系 C 2 のマイクロアレイデータ中に標的とする遺伝子クラスタによるものと推定されるデータが存在することが強く示唆されたため、続いて C 2 の DNA マイクロアレイデータを用いて以下の検証、同定実験を行った。

[0135] (C) 遺伝子クラスタ判定

上記 (A) で得られた系 C 2 についての注釈重み付け後の仮想に遺伝子クラスタのスコア (M 値) から、計算式 b) にしたがって遺伝子クラスタ判定値 χ を算出した (図 21)。

具体的には、記憶されている系 C 2 における各仮想の遺伝子クラスタのスコアを呼び出し、仮想の遺伝子クラスタの乖離度算出部に格納された仮想の遺伝子乖離度判定プログラムのうち χ 値算出プログラムを適用し、計算式 b) にしたがって、各仮想の遺伝子クラスタについての判定値 χ を算出した。なお、図 21 においても、実施例 1 の図 16 と同様に、 $n_{cl} = 1$ のときの値が $n_{cl} = 2$ のときの値より大きい仮想の遺伝子クラスタ、及び $n_{cl} = 1$ のときの値が負の仮想の遺伝子クラスタは除外している。

結果は、図 21 に示されるように、実施例 1 と同様、1 つの仮想遺伝子ク

ラストが $ncl = 3$ で極大かつ最大値を示した。これがコウジ酸産生に必須の3つの遺伝子、A0090113000136、A0090113000137、A0090113000138のみを含んだものである点は、実施例1（図16）と同様だが、ここでは上位の χ 値が重み付けによって図16におけるものより2倍程度高くなり、他との差が広がっている。すなわち注釈による重み付けによって、機能に即した該当遺伝子クラスタの検出精度が向上しているといえる。

[0136] 続いて遺伝子クラスタ判定値 v を、各仮想の遺伝子クラスタについて計算式c)にしたがって算出した。具体的には、仮想の遺伝子クラスタの乖離度算出部に格納された v 値算出プログラムを適用し、各仮想の遺伝子クラスタについて計算式c)にしたがって判定値 v を算出した。次元数 d' および係数 a は、実施例1と同様に、それぞれ2および1を採用した。結果は図22に示される。なお、図22においても、実施例1の図17と同様に、 $ncl = 1$ のときの値が $ncl = 2$ のときの値より大きい仮想の遺伝子クラスタは除外している。結果を図22に示す。

実施例1と同様、 $ncl = 3$ のとき極大かつ最大値をとる仮想の遺伝子クラスタが一つあり、これがコウジ酸産生に必須の3つの遺伝子のみを含む遺伝子クラスタである。その他に $ncl = 2$ に小さなピークを持つものが1つ見受けられるが、これはコウジ酸産生関連遺伝子の3つのうちの2つ（A0090113000137、A0090113000138）からなるものである。図22を図17と比較すると分かるように、アノテーション付与により選定された機能を有する遺伝子の発現量変動比に重み付けを行うことで、標的とする遺伝子クラスタのスコアが嵩上げされて浮き彫りになり、探索対象の遺伝子クラスタをより高精度に検出可能となっている。

[0137] 続いて計算式d)に従って、 χ 値および v 値の積から遺伝子クラスタ評価値を算出した（図23）。具体的には、こうして得られた χ 値および v 値に対し、遺伝子クラスタ絞り込み部に格納された候補絞り込みプログラムを適用して、計算を行った。図23をみると明らかなように、結果二つの仮想の遺伝子クラスタが、10000以上の突出して大きな値をとっており、それ

と比較してその他のものは相対的に非常に小さなピークしか示さない。このうち $ncl = 3$ で極大および最大値をとる仮想の遺伝子クラスタが、実施例 1 と同様、コウジ酸産生に必須の 3 つの遺伝子、A0090113000136、A0090113000137、A0090113000138のみを含んだものである。もう一つ顕著なものとして、 $ncl = 2$ で極大値を持つものがあるが、これはコウジ酸産生に必須の 3 つの遺伝子のうちの 2 つ (A0090113000137およびA0090113000138) からなるものである。それ以外の仮想の遺伝子クラスタは、相対的にほぼゼロと見なせる。この結果から明らかなように、アノテーションにより選定された遺伝子の発現量変動比に重み付けを行うことで標的とする遺伝子クラスタに対応する仮想の遺伝子クラスタのスコアをより顕著に高めていることは、重み付けを行わないもの (図 18) と比較すると、明らかである。

以上より、該当アノテーションにより選択されたる遺伝子の発現量変動比に重み付けを行うことによって、より高精度に機能に即した形で該当遺伝子クラスタを検出、同定できることが示された。

[0138] 実施例 3

アスペルギルス・オリゼにおける特定機能を有するゲノム遺伝子により、仮想の遺伝子クラスタを構築し、スコアリングした場合の、コウジ酸生合成遺伝子の探索

本実施例は、アスペルギルス・オリゼのゲノム遺伝子中、特定の機能を持った遺伝子によって仮想の遺伝子クラスタを構築し、仮想の遺伝子クラスタのスコアを解析することにより、コウジ酸産生に必須の遺伝子を探索しうることを検証するための実験である。

本実施例では、仮想の遺伝子クラスタのサイズ (ncl) を 5 として、アスペルギルス・オリゼのゲノム配列より 14032 個の仮想遺伝子クラスタを作成した。実施例 1 と同様、途中抜けているものやゲノム断片の末端に位置する仮想の遺伝子クラスタは、 ncl 個より少ない遺伝子よりなるものとして構成した。

この実験においては、実施例 2 の装置を用いた。ただし、実施例 2 におけ

る実験系C 1 からC 3 の3種のアレイデータについては、足し合わせて一つの発現量変動比（m値）にまとめたものを使用した。また、重み付けの代わりに、仮想の遺伝子クラスタのサイズを遺伝子数で5と設定し、構築された仮想の遺伝子クラスタ中から、アノテーションにより選定された複数種の機能遺伝子が含まれていることを条件として、仮想の遺伝子クラスタを選出し、該選出された仮想の遺伝子クラスタを、スコアリングする対象の仮想の遺伝子クラスタとするように、システムを変更した。その他は実施例2と同様である。

すなわち、ゲノム上近傍に位置する条件として、仮想の遺伝子クラスタのサイズ（n c l）を5と設定して、記憶装置に記憶されたアスペルギルス・オリゼのゲノム上の位置情報に基づき、14032個の仮想遺伝子クラスタを作成した。この場合において実施例1と同様、途中抜けているものやゲノム断片の末端に位置する仮想の遺伝子クラスタは、n c l個より少ない遺伝子で構成した。

[0139] これらの仮想の遺伝子クラスタのうち、実施例2と同様にして、特定の機能の遺伝子を含むものを、機能に即したモチーフとの配列相同性を検索することで選び出した。特定の機能とはすなわち、以下の3つである。

- ・膜輸送体：transporterまたはmajor facilitator
- ・転写制御因子：transcription
- ・酸化還元酵素：oxidoreductaseまたはdehydrogenase

[0140] 続いて、総数で14032個ある仮想の遺伝子クラスタの中から、該当する機能の注釈を持つ遺伝子を含むものを選び出した。その数のベン図を図24に示す。上記3つの因子（膜輸送体、転写制御因子、酸化還元酵素）の全てをもっている仮想遺伝子クラスタは、14032個のうち176個であった。また、上記3つの内、酸化還元酵素を除いて2つの要素（膜輸送体、転写制御因子）をもつものとする、該仮想遺伝子クラスタは636個であった。

上記手順は、具体的には、実施例2と同様にして、記憶装置に記憶された

アノテーションデータの中から以下の3種の機能を有する遺伝子を機能遺伝子選択部の選択プログラムを適用して選定し、さらに、構築された総数で14032個ある仮想の遺伝子クラスタの中から、選定された機能遺伝子を含むものを選出することにより行った。

[0141] 続いて、選り出した各仮想遺伝子クラスタに対して、クラスタ・スコアリングを行った。

なお、アレイのデータは、参考例1および実施例1～2で述べたものと同様、系C1～C3におけるものを二色法によって測定したものであり、産生条件と非産生条件の下で生育させた菌体からmRNAを取り出し、それぞれ色素でラベリングした後でアレイ上のオリゴDNAにハイブリダイズすることによりデータを得て、そこから各遺伝子の発現量変動比(m)を得たものである。

さらに、仮想の各遺伝子クラスタにつき一つのスコアを得るために、それぞれの遺伝子について、3つの系C1～C3から得られたm値を足し合わせて一つの値とした。続いて上記で選出した、該当する機能の注釈を持つ遺伝子を含む仮想の遺伝子クラスタのうち、膜輸送体、転写制御因子、酸化還元酵素の3つ全てを含む176個について、計算式a)にしたがって、スコア(M値)を算出した。

具体的には、上述の手順に従って選出した各仮想遺伝子クラスタ中に含まれる各機能遺伝子の系C1～C3の実験に基づく発現量変動比を記憶部から呼び出し、仮想の遺伝子クラスタ・スコアリング部のスコアリングプログラムを適用し、計算式a)にしたがい、仮想の遺伝子クラスタのスコアリングを行った。

[0142] 図25(a)に全ての仮想の遺伝子クラスタ14032個のスコアM値の分布を示した。また図25(b)には、コウジ酸の産生に関連すると推定される3つの要因(膜輸送体、転写制御因子、酸化還元酵素)の全てをもっている仮想の遺伝子クラスタ176個での点数の分布を示した。さらに両者に、産生に必須の遺伝子3つを含む仮想の遺伝子クラスタのスコアの位置を示

した。本実施例では、仮想の遺伝子クラスタを並びあう5個の遺伝子としたため、並びあう3つの必須の遺伝子（A0090113000136-A0090113000138）を含むクラスタは3個（A0090113000134-A0090113000138、A0090113000135-A0090113000139、A0090113000136-A0090113000140）存在する。よって3つの矢印により位置を示している。

これらは、総数14032の仮想の遺伝子クラスタの中で24、58、59位に位置していた。遺伝子1つ1つで解析した場合は3000位以下であったことを考えれば、正解率は十分に上がっているといえる。しかし、さらに含まれる遺伝子の機能により仮想の遺伝子クラスタを選択する過程を加えることにより、クラスタスコアの順位が2、5、6位と明らかに上位になることが判明した。

[0143] ここで、分布の形にも注目が必要である。総数14032個の仮想遺伝子クラスタのスコア分布では、全体が単一の分布に近い。総数が多くすそ野が広いため（図25（a））、コウジ酸の産生に必須の仮想の遺伝子クラスタより高得点を取るものも存在する。しかし、そこからコウジ酸の生合成経路を仮定し、そこに関係すると考えられる遺伝子をモチーフから同定してコウジ酸産生に関連の深い仮想の遺伝子クラスタを選択して解析することにより、分布の様子が変わった（図25（b））。総数が少なくなることでより単一的なバックグラウンドの仮想の遺伝子クラスタの分布の形は相似なままで小さくなり、その結果すそ野が狭まり、偶然に高得点をとるものが無くなった。一方、コウジ酸の産生に関連の高い仮想の遺伝子クラスタは、バックグラウンドとは関係なく位置するため、結果的に山の頂上を中心とし山形のバックグラウンドの分布とは別の分布が、高スコア側の位置に存在することになる。このように、点数が高いだけでなく、バックグラウンドの分布から外れて高スコアに位置する仮想の遺伝子クラスタが存在することによっても、この解析が正解を含んだものであることが推定できる。

[0144] 実施例4

アスペルギルス・オリゼにおける仮想の遺伝子クラスタ・スコアリングに

よるコウジ酸産生に必須の遺伝子クラスタの選出条件検討

実施例3で得られた結果が、機能注釈による仮想の遺伝子クラスタの選出条件を代えることによって変化するか否かを解析し、方法の検討を行った。

実施例3においては、該遺伝子クラスタの探索対象を、コウジ酸の産生に関連すると推定される3つの要因（膜輸送体、転写制御因子、酸化還元酵素）を含む仮想遺伝子クラスタに限定することにより、産生に必須と判明している3つの遺伝子を含む仮想の遺伝子クラスタが上位に位置することを確認したが、この3つの要因を、2つに減らすことの影響を検討した。機能の注釈による仮想の遺伝子クラスタ選出およびクラスタ・スコアリングに関する手順は、実施例3と同様である。

この実験においては、上記実施例3の装置を用い、機能遺伝子選択部に対する機能遺伝子選択コマンドのみを代えて行った。

[0145] 図26に示すように、膜輸送体および転写制御因子の2つの注釈に該当する遺伝子を含む仮想の遺伝子クラスタ636個のスコア分布をみた場合、実施例3と同様、コウジ酸産生に必須の3遺伝子を含む遺伝子クラスタは、クラスタスコアM値の順位が2、5、6位と上位に位置することが判明した。またスコア分布の形も、関連する遺伝子クラスタはバックグラウンドと考えられる単一の山型の分布とは異なる分布として高スコア側に位置しており、この点でも同様の結果を与えた。機能の注釈による選出条件が多いほどバックグラウンドを減らすことに貢献し、上位に位置する可能性が高くなる。しかし3つの注釈による制限を2つに弱めても、本方法は十分に機能することを確認できた。

[0146] 一方、膜輸送体を含むが転写制御因子を含まない仮想の遺伝子クラスタ2949個のスコア分布を図27に示す。転写制御因子はコウジ酸産生に必須の3つの遺伝子のうち真ん中に位置し、また、この実験の場合、並び合う5個の遺伝子を仮想の遺伝子クラスタの選出条件としているため、転写制御因子を含まないという条件を付けると、コウジ酸産生に必須の3遺伝子を含む仮想の遺伝子クラスタは構築されない。したがってここで示した仮想の遺伝

子クラスタのスコア分布は、バックグラウンドのみの分布に相当する。ここでは、全体の数が増えた分、分布のすそ野が広がり高得点まで分布するが、一方で山の頂上を中心とした単一の山型分布を示している。本分布には、高スコア側に別の分布として位置する仮想の遺伝子クラスタは存在せず、この点でも正解がないことを示していた。

[0147] 実施例 5

アスペルギルス・フラバスにおける仮想の遺伝子クラスタ・スコアリングによる生合成遺伝子同定

本発明の遺伝子の探索、同定方法がアスペルギルス・オリゼのコウジ酸産生に必須の遺伝子クラスタ以外にも適応可能なことを示すため、アスペルギルス・フラバスを対象として二次代謝産物を合成する遺伝子クラスタを同定した。アスペルギルス・フラバスは、二次代謝産物でありマイコトキシンの一つであるアフラトキシンを強く産生することで知られており、その産生至適温度は25℃前後である。この実験に使用した装置は、実施例1の装置と同様である。

[0148] DNAマイクロアレイのデータは、遺伝子発現解析データの公共データベースであるNCBIのGEO(<http://www.ncbi.nlm.nih.gov/geo/>)より、GSE15435のIDで登録されたものの一部を用いた（参考文献1）。すなわちこのデータを、遺伝子発現量入力部を通して記憶部に保存した。このアレイデータは実施例1～4と異なり、一色法で測定されている。そこで各ゲノム遺伝子の発現量変動比m値を得るため、以下のように二次代謝産物をより多く産生すると考えられる条件とそうでない条件を比較し、前者を分子に後者を分母にした値をm値として算出した。検討した系は全部で2つである。

C1: 培養開始後96時間／同18時間

C2: 培養中、育成温度28℃／同37℃

[0149] 以降これら二つの系をそれぞれ系C1、C2とする。遺伝子は、2つの系それぞれにおいて12955個含まれている。

[0150] (A) クラスタ・スコアリング

系C 1、C 2のそれぞれについて、実施例1と同様に計算式a)に従って、仮想の遺伝子クラスタのサイズ $n_{cl} = 1 \sim 30$ でクラスタ・スコアリングを行い、各仮想の遺伝子クラスタのスコア(M値)を得た。図28右は、各仮想の遺伝子クラスタの各サイズ毎にスコアの分布状態を示したヒストグラムである。図28左のグラフはヒストグラムの一部拡大図である。これを見ると分かるように、ゼロを中心とした山型の正規分布様集団から外れて高いスコア(M値)を持つ仮想の遺伝子クラスタがあると、全体を表したヒストグラムにおいて山の中心が左側にずれるが、一見して、系C 2において、 n_{cl} の増加とともに、山の中心が左にずれていくことが分かる。

[0151] (B) データ判定

実施例1と同様にして、計算式e)にしたがって、系C 1およびC 2におけるスコア分布評価値 ε を算出した(図29)。ここで仮想の遺伝子クラスタ数 n は各々のクラスタサイズについて12955、次元数 d は6を採用した。なお実施例1および2同様、 $n_{cl} = 1$ のときの値に応じて該当しない仮想の遺伝子クラスタは除外してある。図29示されるように、系C 1では ε 値はほぼゼロであるのに対し、系C 2では $n_{cl} = 18$ において極大かつ最大値を示す。これは、アフラトキシン産生至適温度25℃であるため、系C 2における温度に関する生理状態変化条件の設定が適切であり、他方系C 1の変化条件設定は適切ではなかったことを反映している。

すなわち系C 2に基づく発現量変動比データを用いたクラスタ・スコアリングによって、 ε 値を増加させる同定対象の遺伝子クラスタが存在すること、及び、そのクラスタサイズは20前後であることが推定することができた。なおアスペルギルス・フラバスが最も強く産生する二次代謝産物の一つは前述のアフラトキシンであり、その生合成遺伝子は29の遺伝子(AFLA_139100-AFLA_139440)からなる遺伝子クラスタを形成していることが知られている(参考文献2)。ただしこの全てが同時に発現しているわけではなく、環境等によってその発現強度は変化する。本結果において $n_{cl} = 20$ 程度の大きなクラスタサイズの位置にピークが存在することは、アフラトキシンの

生合成遺伝子クラスタの発現と対応していると考えられる。また本図では ε 値が 10^4 のオーダーの値を示しているが、二次代謝産物の発現の弱い種であるアスペルギルス・オリゼの ε 値は、図 15 にあるように 10^3 のオーダーである。これは、アスペルギルス・フラバスが同オリゼと比較して二次代謝産物を非常に強く発現する事実と一致する。

以上より、系 C 2 の発現量変動比データを用いた仮想の遺伝子クラスタ中のスコアから、構築された仮想の遺伝子クラスタ中に、標的とする遺伝子クラスタが含まれると予測できたため、系 C 2 の DNA マイクロアレイデータセットを用いて、以下の実験を行った。

[0152] (C) 遺伝子クラスタ判定

系 C 2 に基づく仮想の各遺伝子クラスタのスコア (M 値) から、実施例 1 と同様にして、各仮想の遺伝子クラスタについて、計算式 b) にしたがって遺伝子クラスタ判定値 χ を算出した。なお、この算出においては、実施例 1 および 2 同様、 $n_{cl} = 1$ のときの値に応じて該当しない仮想の遺伝子クラスタは除外してある。結果を図 30 に示すが、この折れ線グラフは、実施例 1 (C) に記載したように、各仮想の遺伝子クラスタの構築において起点とした遺伝子が共通する、遺伝子サイズ 1 ~ 30 の各仮想の遺伝子クラスタの判定値を結んだものである。

図 30 の結果から明らかなように、起点を同じくする仮想遺伝子クラスタの各折れ線グラフは、あるサイズで χ 値の極大値をとっている。この起点を同じくする仮想の遺伝子クラスタの各折れ線グラフにおいて、 χ 値の極大値が高く、150 程度を示すものは、大凡 4 種のサイズに分けられるが、このうち、極大値が高いピークを最も多く含むものは、サイズ (n_{cl}) が 20 付近のものである。アスペルギルス・フラバスのアフラトキシン合成に関与する遺伝子クラスタは既知であり、このサイズ 20 付近のピークの高いもの上位 10 個の各仮想の遺伝子クラスタ中の遺伝子について機能注釈を参照した結果、いずれもアフラトキシン合成に関与する遺伝子を含むことが明らかとなった。この結果は上記 (b) の予測結果と一致し、この χ 値の算出によ

り、アスペルギルス・フラバスにおけるアフラトキシン生合成に関与する遺伝子クラスタとその中に含まれるアフラトキシン生合成遺伝子を、ある程度特定可能であることを示す。一方、図30では他にも大きな値をとる仮想の遺伝子クラスタが複数存在するが、これは、その推定機能の注釈をみても、未知の二次代謝産物合成に関与する遺伝子クラスタである可能性が高い。

[0153] 次に、系C2の仮想の各遺伝子クラスタについて、実施例1と同様にして、遺伝子クラスタ判定値 ψ を計算式c)にしたがって算出した。ここで次元数 d' は2、係数 a は1を採用した。また実施例1および2同様、 $nc1=1$ のときの値に応じて該当しない仮想の遺伝子クラスタは除外してある。結果を図31に示すが、図31も図30と同様に、各仮想の遺伝子クラスタの構築において起点とした遺伝子が共通する、遺伝子サイズ1~30の各仮想の遺伝子クラスタの判定値を結んだものである。

図31に示されるように、多くの仮想の遺伝子クラスタが極大値を示している。このうち ψ 値が200前後を示すものは、4つのサイズに分けられるが、このうち極大値が高いピークを最も多く含むものは、 χ 値の場合と同様にサイズが20付近のものであり、この各ピークの上位10個の仮想の遺伝子クラスタは、いずれも上述のアフラトキシン合成遺伝子を含んでいた。中には上述のアフラトキシン合成遺伝子クラスタを含むものが含まれる。すなわち本評価値 ψ によっても、アフラトキシン生合成に関与する遺伝子クラスタ及びその中に含まれるアフラトキシン生合成遺伝子をある程度特定することができた。

[0154] こうして得られた χ 値および ψ 値を元に、さらに遺伝子クラスタ候補を絞り込むため、実施例1と同様にして計算式d)に従って二つの値の積から遺伝子クラスタ判定評価値を算出した。図32は、この算出結果に基づき、仮想の遺伝子クラスタサイズと $\chi \times \psi$ 値の関係をグラフ化したものである。図32から明らかなように、多くの仮想遺伝子クラスタが特定の $nc1$ において極大値を示すことが分かる。このうち、 $nc1=18$ で最大値を取る仮想の遺伝子クラスタは、AFLA_139150-AFLA_139220、AFLA_139240-AFLA_139280

、AFLA_139300-AFLA_139320からなるもので、そのすべてが既知であるアフラトキシン生合成遺伝子クラスタに含まれる遺伝子である。またその他に25000以上の値を示す仮想の遺伝子クラスタの機能注釈を見てみると、代表的な二次代謝産物関連遺伝子機能であるNRPSやP450といったものが挙げられており、これらも未知の二次代謝産物合成遺伝子クラスタである可能性が高い。さらに値の大きさを実施例1のアスペルギルス・オリゼ(図18)と比較してみると、3倍近く同フラバスが高いことが分かる。これは、アスペルギルス・フラバスが二次代謝産物を非常に活発に産生する種であるという事実と対応している。

以上より、本発明がDNAマイクロアレイデータからゲノム上に集合して機能を果たす生合成遺伝子を同定する有効な手段であることが示された。

[0155] (参考文献1)

Beyond aflatoxin:four distinct expression patterns and functional roles associated with *Aspergillus flavus* secondary metabolism gene clusters

D.RYAN GEORGIANNAら、MOLECULAR PLANT PATHOLOGY(2010)11(2),213-226

(参考文献2)

Genetic regulation of aflatoxin biosynthesis:from gene to genome

D.RYAN GEORGIANNAら、Fungal Genetics and Biology(2009)46(2),113-125

[0156] 実施例6

アスペルギルス・ニガーにおける遺伝子クラスタ・スコアリングによる生合成遺伝子推定

本発明の同定手法に従って、アスペルギルス・ニガーの二次代謝産物を合成する遺伝子クラスタを推定した。この実験の使用装置は、実施例1の装置と同様である。

DNAマイクロアレイのデータは、遺伝子発現解析データの公共データベ

ースであるNCBIのGEO(<http://www.ncbi.nlm.nih.gov/geo/>)より、GSE17329のIDで登録されたものの一部を用いた。すなわち、このデータを、遺伝子発現量データ入力部を通じて記憶部にゲノム遺伝子発現量データとして保存した。このアレイデータは実施例1～4で用いたアスペルギルス・オリゼのものとは異なり、一色法で測定されている。そこで、ゲノム遺伝子発現量変動比 m 値を得るため、遺伝子発現量変動比算出部において、以下のように生理状態が変化する条件として以下の条件を設定し、前者を分子に後者を分母にした値を m 値として算出した。検討した系は以下の2つである。なお、これらの系は、炭素源欠乏条件下で何らかの二次代謝関連遺伝子クラスタが関与していることを期待したもので、例えば上記したコウジ酸あるいはアフラトキシン産生等の特定の機能を標的としたものではない。

[0157] C1:培養中、炭素源枯渇後55.5時間／同5時間

C2:培養中、炭素源枯渇後24時間／炭素源枯渇前3.5時間

以下、上記2つの生理状態が変化する条件をそれぞれ系C1、C2とする。なお、発現量変動比は、2つの系それぞれにおいて14509個の遺伝子について算出した。

[0158] (A) クラスタ・スコアリング

系C1～2のそれぞれについて、実施例1と同様に計算式a)に従って $n_c l = 1 \sim 30$ でクラスタ・スコアリングを行い、各仮想遺伝子クラスタの M 値を得た。図33右は、各仮想の遺伝子クラスタの各サイズ毎にスコアの分布状態を示したヒストグラムである。図33左のグラフはヒストグラムの一部拡大図である。図33左の拡大図をみると分かるように、ゼロ付近を中心とした山型の正規分布様集団から外れて高い M 値を持つ仮想遺伝子クラスタがあると、全体を表したヒストグラムにおいて山の中心が左側にずれるが、系C2において、 $n_c l = 5$ 付近で山の中心が左にずれていることが分かる。

[0159] (B) データ判定

実施例1、2、5と同様にして、計算式e)にしたがって、系C1～2に

おけるスコア分布評価値 ε を算出した（図 3 4）。ここで仮想の遺伝子クラスタ数 n は 1 4 5 0 9、次元数 d は 6 を採用した。図 3 4 に示されるように、系 C 1 は $n_{c1} = 8$ 、系 C 2 は $n_{c1} = 5$ においてそれぞれ極大値を示している。したがって二つの系においてともに、クラスタ・スコアリングによる平均化の方向に反して値を増加させる仮想の遺伝子クラスタが存在することを意味する。すなわちこの二つの系（C 1、C 2）における発現量変動比データを用いた、仮想の遺伝子クラスタのスコアリングによって、 ε 値を増大させる遺伝子クラスタが存在すること、及びその遺伝子クラスタのサイズが 8 前後あるいは 5 前後と推定された。ただし、この実験においては、上記したように、生理状態変化条件として炭素源欠乏条件を設定したものであり、該条件は特定の遺伝子クラスタを標的としたものではないので、極めて多数の遺伝子クラスタが関与していることが予想され、 ε 値によるサイズの予想は確定的なものではない。

この点をふまえ、さらに以下の実験を行った。

[0160] (C) 遺伝子クラスタ判定

系 C 1 および C 2 のそれぞれについて、系 C 1 および C 2 の DNA マイクロアレイデータから、実施例 1 と同様に計算式 b) にしたがって、各仮想遺伝子クラスタについて遺伝子クラスタ判定値 x を算出した（図 3 5 (a) ; C 1、同 (b) ; C 2）。なお実施例 1、2、5 と同様に、 $n_{c1} = 1$ のときの値に応じて該当しない仮想の遺伝子クラスタは除外してある。

図 3 5 の結果から明らかなように、系 C 1、C 2 の双方において、多くの仮想遺伝子クラスタが極大値を示している。これより、アスペルギルス・ニガーには、系 C 2、C 3 の生理状態変化条件で変動する遺伝子クラスタが存在すると考えられ、これは既存の事実と一致する（参考文献 3）。

次に、系 C 1 および C 2 の各仮想遺伝子クラスタについて、実施例 1 と同様にして遺伝子クラスタ評価値 v を計算式 c) にしたがって算出した（図 3 6 (a) ; C 1、同 (b) ; C 2）。ここで次元数 d' は 2、係数 a は 1 を採用した。ここでも実施例 1、2、5 と同様に、 $n_{c1} = 1$ のときの値に応じ

て該当しない仮想の遺伝子クラスタは除外してある。図36の結果に示されるように、系C1およびC2の双方において、複数の仮想の遺伝子クラスタが極大値を示す。ただし ν 値の上位と下位の差は χ 値（図35）に比べて増大しており、本実験においては ν 値の方が、少数の仮想の遺伝子クラスタを抽出するためにはより有利である。例えば、系C1において ν 値が100以上のものとした場合、該当する仮想遺伝子クラスタは1つのみである。系C2の場合、 ν 値が60以上の仮想遺伝子クラスタは3つのみである。

[0161] こうして得られた χ 値および ν 値を元に、実施例1と同様にして、計算式d)に従って二つの値の積から遺伝子クラスタ判定評価値を算出した（図37(a)；C1、同(b)；C2）。図37の結果から明らかなように、系C1において、 $n_{cl}=3$ で極大かつ最大値をとる仮想遺伝子クラスタが一つ存在することが分かる。系C2においてもいくつかの顕著なピークが見られ、例えば値4000以上とした場合、該当する仮想遺伝子クラスタは4個である。これらの仮想の遺伝子クラスタを構成する遺伝子について、その配列に基づきモチーフ検索による推定機能の注釈を見たところ、その多くが機能未知であり、該当する機能遺伝子は見いだせなかった。しかし本評価値において高い値を示すことを考えると、未知の該当遺伝子クラスタである可能性が高い。

（参考文献3）

Review of secondary metabolites and mycotoxins from the *Aspergillus niger* group

K.FOG NIELSENら、Analytical and Bioanalytical Chemistry(2009)395(5), 1225-1242

[0162] 実施例7

アノテーション（機能注釈）に基づき選定された遺伝子を一以上含むことを条件として、仮想の遺伝子クラスタを構築した場合における、コウジ酸合成遺伝子の探索

アスペルギルス・オリゼのコウジ酸産生関連遺伝子からなる遺伝子クラス

タを同定することを目的として、予測される機能に関連した注釈のついた遺伝子を選定したのち、該遺伝子を1以上含むように仮想の遺伝子クラスタを構築し、構築された各仮想の遺伝子クラスタをスコアリングして、該当遺伝子の同定を行った。

この実験に使用した手法は、実施例1と基本的には同様であるが、実施例1においては、仮想の遺伝子クラスタの構築をする際、仮想の遺伝子クラスタのサイズを1～30と設定し、ゲノム遺伝子の配列順に全ての遺伝子が含まれるように仮想の遺伝子クラスタを構築したが、本実施例においては、ゲノム位置情報（配列情報）中に、アノテーション付与に基づき選定された機能遺伝子が出現したとき、その機能遺伝子を起点とする仮想の遺伝子クラスタを構築するように変更した点、構築された仮想の遺伝子クラスタのスコアリングにおいて、選定された機能遺伝子以外の遺伝子についての発現量変動比（m値）は無視し、選定された遺伝子の発現量変動比のみを用いるように変更した点で異なる。なお、遺伝子サイズについては、実施例1と同様にゲノムの遺伝子配列順に1～30個と設定した。

具体的には、この実験に使用した装置は、実施例1に記載した装置と基本的に同様であるが、仮想の遺伝子クラスタ構築プログラムにおいて、ゲノム位置情報（配列情報）中に、アノテーション付与に基づく遺伝子選定部において選定された機能遺伝子が出現したとき、その機能遺伝子を起点とする仮想の遺伝子クラスタを構築するように変更した点、構築された仮想の遺伝子クラスタのスコアリングにおいて、選定された機能遺伝子以外の遺伝子についての発現量変動比（m値）は無視し、選定された遺伝子の発現量変動比のみを用いるように変更した点で異なる。なお、遺伝子サイズについては、実施例1と同様にゲノムの遺伝子配列順に1～30個と設定した。

なお、本実施例では、実施例1のデータ判定結果において該当遺伝子クラスターが含まれていると予測された、系C2（7日目／4日目）のアレイデータのみを用いて実験を行った。また、実施例2と同様、コウジ酸産生に必要な機能として以下の3つの機能を選出した。

- ・膜輸送体：transporterまたはmajor facilitator
- ・転写制御因子：transcription
- ・酸化還元酵素：oxidoreductaseまたはdehydrogenase

[0163] これらを選出した理由として、コウジ酸の生合成がグルコースから酸化により変換されていると推定されること、産生されたコウジ酸の膜輸送による培地中への分泌に膜輸送体が必要と推定されること、および関与する遺伝子の転写制御に転写因子が必要と推定されることが挙げられる。なお、上記の英単語は、アノテーションによる遺伝子選定に用いたキーワードである。

[0164] アノテーション（機能注釈）による遺伝子の選定及び仮想の遺伝子クラスタの構築・スコアリング

一般に利用されているアノテーション推定プログラムの一つであるInterproscan (<http://www.ebi.ac.uk/Tools/InterProScan/>) を用いて、アスペルギルス・オリゼのゲノムDNA上の各遺伝子についてアノテーションを付与し、上記3種の機能を有する遺伝子を選定した。具体的には、各遺伝子についてのアノテーションデータを本装置の入力装置に入力し、記憶装置に記憶した。記憶したアノテーションデータのデータを呼び出し、上記3種の機能を有する遺伝子を機能遺伝子選択部の選択プログラムを適用して選定した。なお、選定は各遺伝子について付与されたアノテーション中に上記3つの機能群に対応するキーワードが含まれるかどうかで行い、その結果、選定された遺伝子は、系C2において有効に遺伝子発現データを取得できた5595個の遺伝子のうち、796個であった。

[0165] 仮想の遺伝子クラスタの構築においては、上記変更プログラムを適用し、ゲノム遺伝子の位置情報に基づき、ゲノムの遺伝子配列中にこのように選定された機能遺伝子が出現した場合、この選定された遺伝子を起点遺伝子として、ゲノム遺伝子の配列順にクラスタサイズを1から30まで変化させて、仮想の遺伝子クラスタを構築した。すなわち、これにより、構築される仮想の遺伝子サイズ中には、必ず付与されたアノテーションに基づき選定された

遺伝子が一つ以上含まれ、選定された機能遺伝子が含まれない仮想の遺伝子クラスタは構築されないが、構築された遺伝子クラスタにおいては、上記選定された機能遺伝子以外の遺伝子も含まれる。このようにした理由は、実施例 1 の装置に格納された仮想の遺伝子構築プログラムの変更を出来るだけ最小限にしたためである。しかし、構築された仮想の遺伝子クラスタのスコアリングにおいては、上記選定された機能遺伝子以外の遺伝子についての発現量変動比については無視し、選定された機能遺伝子の発現量変動比のみを用いて、計算式 a) による計算を行った。これによれば、仮想の遺伝子クラスタのスコアは、仮想の遺伝子クラスタを、上記選定された機能遺伝子のみから構築した場合のスコアと全く同じになる。このようにして、得られた各仮想の遺伝子クラスタのスコアは、本発明装置の記憶部に記憶した。

なお、本実施例では、構築された仮想の遺伝子クラスタに 1 つの遺伝子のみしか含まれない場合が含まれ、また、本実施例においては、実施例 1 ～ 4 と同様に、ゲノム上の末端側に位置する遺伝子については、組み合わせうる最大個数の遺伝子で仮想の遺伝子クラスタを構築したが、クラスタ・スコアリングの性質上、これらによる遺伝子クラスタの探索についての影響はない。このようにして構築された仮想の遺伝子クラスタは、各クラスタサイズについてそれぞれ 796 個である。

つづいて、構築した各々の仮想の遺伝子クラスタについて、計算式 a) にしたがって、 $ncl = 1 \sim 30$ でクラスタ・スコアリングを行って各仮想の遺伝子クラスタのスコア (M 値) を得た。

[0166] 遺伝子クラスタ判定

算出した各仮想の遺伝子クラスタのスコア (M 値) に基づいて、計算式 b) にしたがって、各仮想の遺伝子クラスタについての判定値 χ を算出した。具体的には、記憶されている各仮想の遺伝子クラスタのスコアを呼び出し、仮想の遺伝子クラスタの乖離度算出部に格納された仮想の遺伝子乖離度判定プログラムのうち χ 値算出プログラムを適用し、計算式 b) にしたがって、各仮想の遺伝子クラスタについての判定値 χ を算出した。図 38 は、各仮想

の遺伝子クラスタについて、起点遺伝子を同じくする仮想の遺伝子クラスタの判定値 χ を、横軸をクラスタサイズとして結んで描いたものである。ここで、クラスタ・スコアリングによって絶対値を増加させない仮想の遺伝子クラスタは該遺伝子クラスタではないため、 $n_{c1} = 1$ のときの絶対値が $n_{c1} = 2$ のときの絶対値より大きい仮想の遺伝子クラスタは除外している。

図をみると、多くの仮想の遺伝子クラスタの判定値 χ がゼロ付近に位置するのに対し、起点を同じくする仮想の遺伝子クラスタの3組が、大きな値を取っていることが分かる。その中で最大のものは、 $n_{c1} = 4$ のときに極大かつ最大値を取っている。これは、コウジ酸産生に必須の3つの遺伝子、A0090113000136、A0090113000137、A0090113000138に加えて、その隣に位置し、本実施例において遺伝子選定の対象であるアノテーションとして”major facilitator”（膜輸送体）を持つA0090113000139を含むものである。すなわち、本実施例では、選定対象のアノテーションがついた遺伝子の発現量変動比のみを用いて仮想の遺伝子クラスタのスコアリングを行っているため、スコアリングの対象となる要素が極度にそぎ落とされており、その結果、該当遺伝子クラスタの近傍にアノテーションによって選定された遺伝子がある場合、それを含んだ遺伝子クラスタが高い値をとりうる。しかし、この最大値を示す仮想の遺伝子クラスタが、コウジ酸の産生に必須の3つの遺伝子を含むことから、本手法は遺伝子クラスタの探索手法として有効である。実際、図38において最大値を示す起点遺伝子を同じくする仮想の遺伝子クラスタの組において、コウジ酸産生に必須の3つの遺伝子のみからなる $n_{c1} = 3$ のものと、隣接したA0090113000139を含む $n_{c1} = 4$ のものは、値がそれほど大きく変わらない。

なお、その他にゼロから外れて大きな値を示す2組の仮想の遺伝子クラスタが存在するが、これらはコウジ酸産生に必須の3つの遺伝子のうち、A0090113000136を含まないものである。

[0167] また、同様に、各仮想の遺伝子クラスタについて、計算式c) にしたがって判定値 ψ を算出した。具体的には、仮想の遺伝子クラスタの乖離度算出部

に格納された ψ 値算出プログラムを適用し、各仮想の遺伝子クラスタについて計算式c)にしたがって判定値 ψ を算出した。次元数 d' および係数 a は、実施例1と同様に、それぞれ2および1を採用した。起点遺伝子を同じくする仮想の遺伝子クラスタの判定値 ψ を $ncl = 1 \sim 30$ で結んだ結果を図39に示す。なお図39においても、 $ncl = 1$ のときの値が $ncl = 2$ のときの値より大きい仮想の遺伝子クラスタは除外している。

判定値 χ のときと同様、 $ncl = 4$ のとき極大かつ最大値をとる仮想の遺伝子クラスタが一つあり、これがコウジ酸産生に必須の3つの遺伝子に加えてもう一つの遺伝子A0090113000139を含む遺伝子クラスタである。図17と比較すると分かるように、アノテーションによる遺伝子の選定によって候補となる仮想の遺伝子クラスタの数が大きく減り、該当遺伝子クラスタが存在する場合、他のゼロ付近の値を持つものとの差がより明確になる。

[0168] こうして得られた χ 値および ψ 値に対し、遺伝子クラスタ絞り込み部に格納された候補絞り込みプログラムを適用して、計算式d)に従って二つの値の積から遺伝子クラスタ評価値を算出した(図40)。図をみると明らかなように、一つの仮想の遺伝子クラスタが、 $ncl = 4$ のとき6000以上の大きな値をとっており、これがコウジ酸産生に必須の3つの遺伝子を含むものである。その他にゼロ付近の集団から離れて比較的大きな値を示す2つの仮想の遺伝子クラスタが存在するが、これはコウジ酸産生に必須の3つの遺伝子のうちA0090113000136を含まないものである。図40を図38および図39と比較すると明らかなように、二つの判定値を掛け合わせることで、該当遺伝子クラスタがより明確に大きな値をとり、該当遺伝子クラスタの予測精度が向上することが分かる。

図41は、上記遺伝子クラスタ評価値を、横軸を遺伝子クラスタ番号として、各クラスタサイズについてプロットしたものである。クラスタサイズに対応した各図は、縦軸のスケールを合わせてある。この図で $ncl = 4$ のときに最大値をとり、いずれのクラスタサイズにおいても突出して高い値を示しているのが、コウジ酸産生に必須の3つの遺伝子の3つあるいは2つを含

む遺伝子クラスタである。このように本手法によって、本実施例の予測の対象であるコウジ酸産生遺伝子クラスタは鋭敏に検出されていることが分かる。

以上の実験結果により、アノテーションにより選定されたる遺伝子を含む仮想の遺伝子クラスタ構築し、該選定された遺伝子の発現量変動比を用いてクラスタ・スコアリングを行うことで、高感度に、目的とする遺伝子クラスタ及びその中に含まれる遺伝子を探索できることが示された。また、この実験結果からいえば、アノテーションにより選定された1以上の遺伝子を組み合わせる仮想の遺伝子クラスタを構築し、スコアリングしても同様な結果が得られることは明らかである。

本手法は強いフィルタリング操作を伴うものであり、該当するアノテーションを持つ遺伝子のm値を過度に反映する場合もある。しかし逆に、遺伝子間の発現変動比が比較的小さい場合などには、目的の遺伝子クラスタを鋭敏に予測できる手法である。

[0169] 実施例 8

フザリウム・バーティシリオイデスにおける遺伝子クラスタ・スコアリングによる二次代謝産物生合成遺伝子の予測と検証

本発明の同定手法に従って、菌類であるフザリウム属の一種、フザリウム・バーティシリオイデスの二次代謝産物を合成する遺伝子クラスタを予測した。フザリウム属は、実施例1～6で用いた真菌類アスペルギルス属とは、進化系統樹的に遠い菌類である（参考文献4）。またフモニシンを始めとするマイコトキシンを産生することで知られており、その他多くの二次代謝産物生合成遺伝子クラスタを有すると考えられる（参考文献5）。

DNAマイクロアレイのデータは、米国国立生物工学情報センター（NCBI）が提供する遺伝子発現解析データの公共データベースGEO(<http://www.ncbi.nlm.nih.gov/geo/>)より、GSE16900のIDで登録されたものの一部を用いた。このアレイデータは、フモニシン産生培地における培養時間が24, 48, 72, 96時間である培養条件のそれぞれについて、一色法にて遺伝子の発現量

を測定したものである。そこで、発現量変動比 m 値を得るため、以下のように二次代謝産物をより多く産生すると考えられる条件とそうでない条件を比較し、前者を分子に後者を分母にした値を m 値として算出した。検討した系は2つである。

[0170] C 1 : 培養時間 7 2 時間 / 同 2 4 時間

C 2 : 培養時間 9 6 時間 / 同 4 8 時間

以降これらの系をそれぞれC 1, C 2とする。本発現情報には、遺伝子クラスタを構成するのに用いられる遺伝子が1 2 2 3 0個含まれている。また、元のアレイデータでは各培養時間について3つのデータがとられているため、各遺伝子について3つのデータ間で発現量を平均化した後、以下の手順に進んだ。

[0171] (A) クラスタ・スコアリング

系C 1, C 2のそれぞれについて、計算式a) に従って $n_{cl} = 1 \sim 30$ でクラスタ・スコアリングを行い、各仮想遺伝子クラスタの M 値を得た。図4 2はそのヒストグラムである。左の拡大図をみると分かるように、ゼロを中心とした山型の正規分布様集団から外れて高い M 値を持つ仮想遺伝子クラスタがあると、各 n_{cl} における M 値のヒストグラムにおいて、山の中心が左側にずれる。図右側の各 n_{cl} におけるヒストグラムを上から見ていくと、 n_{cl} の増加とともに山の中心が左にずれていくことが分かる。なおクラスタ・スコアリングを行う際に必要となるゲノム上の遺伝子の連続情報は、アメリカ合衆国の研究機関であるBroad Instituteがウェブ上で公開しているデータベース“Fusarium Comparative Sequencing Project, Broad Institute of Harvard and MIT (<http://www.broadinstitute.org/>)”中の、`fusarium_verticillioides_3_genome_summary_per_gene.txt`を用いた。

[0172] (B) データ判定

計算式e) にしたがって、系C 1, C 2におけるスコア分布評価値 e を算出した(図4 3)。ここで、仮想の遺伝子クラスタ数 n は各 n_{cl} において1 2 2 3 0であり、次元数 d は6を採用した。図にあるように、系C 1は $n_{cl} =$

14、系C2は $n_{c1} = 5$ において極大値を示す。これは、二つの系においてともに、クラスタ・スコアリングによる平均化の方向に反して値を増加させる仮想の遺伝子クラスタが存在することを意味する。すなわちこの二つの系(C1、C2)において、クラスタ・スコアリングによって値を増加させる、本提案の同定対象に該当する遺伝子クラスタが存在すると判定できる。

以上より、系C1、C2双方の遺伝子発現情報を用いて、以下の同定過程へ進む。

[0173] (C) 遺伝子クラスタ判定

系C1およびC2のDNAマイクロアレイデータから、各仮想遺伝子クラスタについて、計算式b)にしたがって遺伝子クラスタ判定値 c を算出した(図44)。ここで、遺伝子クラスタを検出するという目的に鑑み、 $n_{c1} = 1$ のときの絶対値が $n_{c1} = 2$ のときのものより大きな仮想の遺伝子クラスタについては除外してある。すると図にあるように、系C1、C2の双方において、 $n_{c1} = 1$ 以外で複数の仮想遺伝子クラスタが極大値、極小値を示している。これより、フザリウム・バーティシリオイデスには、複数の二次代謝関連遺伝子クラスタが存在すると考えられる。これは既存の事実と一致する(参考文献5)。

次に、系C1およびC2の各仮想遺伝子クラスタについて、遺伝子クラスタ評価値 u を計算式c)にしたがって算出した(図45)。ここで次元数 d' は2、係数 a は1を採用した。ここで c と同様、 $n_{c1} = 1$ のときの値が $n_{c1} = 2$ のときのものより大きな仮想の遺伝子クラスタについては除外してある。すると図のように、系C1およびC2の双方において、複数の仮想の遺伝子クラスタが極大値を示す。ただし極大値における u 値の上位と下位の差は c 値(図44)に比べて増大しており、 u 値によって上位の少数の仮想の遺伝子クラスタを抽出することは c 値よりさらに容易である。例えば系C1において u 値が100以上のものとした場合、該当する仮想遺伝子クラスタは1つのみである。系C2の場合、 u 値が150以上の仮想遺伝子クラスタは3つのみである。このように、評価値 u を用いることで、該当遺伝子クラスタを高く評

価することが出来ると考えられる。

[0174] こうして得られたc値およびu値を元に、計算式d)に従って二つの値の積から遺伝子クラスタ判定評価値を算出した。図46は、アレイデータに含まれるフザリウム・バーティシリオイデスの遺伝子12230個のそれぞれを起点とし、nc1=1~30で構成した30個の仮想の遺伝子クラスタのうち、最大の評価値を取るものを、横軸をゲノム上の起点となる遺伝子としてプロットした図である。図において、C1とC2の縦軸の縮尺は合わせてある。系C1において、突出して高い値をとる3つの仮想の遺伝子クラスタがある。これらはそれぞれ、遺伝子FVEG_00316、FVEG_08708、FVEG_12519を起点とし、クラスタサイズが14, 5, 16のものである。これらの仮想の遺伝子クラスタを構成する遺伝子について、遺伝子配列相同性検索 (blast) を行った結果を表3に示す。データベースは、NCBIが提供するもので、微生物を含む多くの生物種の遺伝子配列を格納するNR (Non-Redundant, <http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/blastdb.html>) を使用した。表では、相同性の高さを評価する値E-valueが10の-100乗以下のもののうち、ベストヒットのものを抜粋してある。既知の二次代謝物質であるフモニシンの生合成遺伝子は、15個からなるクラスタであることが、フザリウム・バーティシリオイデスの完全世代であるジベレリン・モニリフォルミスにおいて報告されている (参考文献6, 7, 8)。また、フザリウム・バーティシリオイデスにおいては、これまでFUM1(5), FUM6, FUM7, FUM8, FUM9の5つがフモニシン生合成遺伝子と同定されている (参考文献: 9)。表をみると、Aとラベルした遺伝子クラスタの構成遺伝子14個が、フモニシン生合成遺伝子15個のうちの14個であることが分かる (ラベルにFumと併記した)。すなわち、本発明手法によって、二次代謝物質フモニシンの生合成遺伝子クラスタをほぼ正確に予測できることが示された。なお今回の結果に含まれなかった残り一つのフモニシン生合成遺伝子は、遺伝子相同性検索の結果によると、FVEG_00316の一つ前のFVEG_00315である。仮想の遺伝子クラスタの本判定評価値をみると、クラスタサイズ15のときのFVEG_00315を起点とするものの値は9242であり、一方クラ

スタサイズ14のときのFVEG_00316を起点とするものの値は9763である。図をみるとわかるようにこれら2点は近接しており（図46のC1、遺伝子クラスタAのピークにおける2点）、僅差でFVEG_00315がもれたことが分かる。従って、同一の遺伝子が含まれる予測されたクラスタにおいては、判定評価値の最大値が近接している場合、仮想クラスタサイズが最大となる物を選択することによって、最も正確な予測結果が得られると考えられる。

系C2においても複数の顕著なピークがみられるが（図46）、なかでもFVEG_03696を起点とするクラスタサイズ4の仮想の遺伝子クラスタが、値10000を超える最大の正のピークを示している。このピークは、培養開始72時間後を24時間後と比較した系C1では見られなかったもので、培養開始96時間後に初めて発現してくる遺伝子クラスタの存在を示唆している。また系C2では、FVEG_08709を起点とするクラスタサイズ4の仮想の遺伝子クラスタが、大きな負の値をとっている。これは、系C1においては正の値を示していたものと、起点は一つずつずれているものの同等であり、培養開始72時間後には発現していたものが、96時間後には発現を止めた遺伝子クラスタであると推測される。このように、本手法を用いる際に、比較する系を目的に応じて選ぶことで、一つの生物種であっても異なる遺伝子クラスタを検出することが可能である。これらの該当遺伝子クラスタ候補の機能については、blast検索の結果をみても判然としないが（表4）、仮想の遺伝子クラスタCに含まれるFVEG_12523は二次代謝物質生合成遺伝子の一つであるpolyketide synthaseと高い配列相同性を示しており、これまでに知られていない新規な二次代謝物質生合成遺伝子が検出されたと期待される。

以上より本提案方法が、アスペルギルス属からは進化系統樹的に遠い菌類であるフザリウム・バーティシリオイデスにおいても、アスペルギルス属のものと同様、全遺伝子の発現情報からゲノム上に集合して機能を果たす生合成遺伝子を同定する実効的な手段であることが示された。

[0175]（参考文献4）

Evolution of the Fot1 transposons in the genus *Fusarium*: dis

continuous distribution and epigenetic inactivation

M.-J. Daboussiら、Molecular Biology and Evolution (2002) 19 (4), 510-520

(参考文献5)

Biochemistry and genetics of Fusarium toxins

A. E. Desjardinsら、Fusarium: Paul E. Nelson Symposium, APS Press (1999)

(参考文献6)

Linkage among genes responsible for fumonisin biosynthesis in *Gibberella fujikuroi* mating population A

Desjardinsら、Applied and Environmental Microbiology (1996) 62, 2571-2576

(参考文献7)

A polyketide synthase gene required for biosynthesis of fumonisin mycotoxins in *Gibberella fujikuroi* mating population A

R. H. Proctorら、Fungal Genetics and Biology (1999) 27, 100-112

(参考文献8)

Co-expression of 15 contiguous genes delineates a fumonisin biosynthetic gene cluster in *Gibberella moniliformis*

R. H. Proctorら、Fungal Genetics and Biology (2003) 38, 237-249

(参考文献9)

Characterization of four clustered and coregulated genes associated with Fumonisin biosynthesis in *Fusarium verticillioides*

J.-A. Seoら、Fungal Genetics and Biology (2001) 34, 155-165

[表4]

本発明手法によって検出されたフザリウム・バーティシリオイデスにおける 遺伝子クラスタ構成遺伝子の相溶性検索結果 (ベストヒット、E-value < 1e-100)				
遺伝子クラスタ (図に記載)	遺伝子名	説明	スコア	E-value
A: Fum	FVEG_00316	Fum1p [Gibberella moniliformis]	1539	0.0
A: Fum	FVEG_00317	Fum6p [Gibberella moniliformis]	2047	0.0
A: Fum	FVEG_00318	Fum7p [Gibberella moniliformis]	860	0.0
A: Fum	FVEG_00319	Fum7p [Gibberella moniliformis]	860	0.0
A: Fum	FVEG_00320	Fum3p [Gibberella moniliformis]	624	1e-176
A: Fum	FVEG_00321	Fum10p [Gibberella moniliformis]	566	2e-163
A: Fum	FVEG_00322	Fum11p [Gibberella moniliformis]	333	3e-139
A: Fum	FVEG_00323	Fum2p [Gibberella moniliformis]	805	0.0
A: Fum	FVEG_00324	Fum13p [Gibberella moniliformis]	678	0.0
A: Fum	FVEG_00325	Fum14 [Fusarium oxysporum]	800	0.0
A: Fum	FVEG_00326	Fum16p [Gibberella moniliformis]	798	0.0
A: Fum	FVEG_00327	Fum17p [Gibberella moniliformis]	731	0.0
A: Fum	FVEG_00328	Fum18p [Gibberella moniliformis]	592	8e-167
A: Fum	FVEG_00329	Fum19p [Gibberella moniliformis]	2076	0.0
B	FVEG_08708	no hits found	-	-
B	FVEG_08709	hypothetical protein [Gibberella zeae]	577	2e-162
B	FVEG_08710	hypothetical protein [Gibberella zeae]	392	6e-149
B	FVEG_08711	hypothetical protein [Gibberella zeae]	1367	0.0
B	FVEG_08712	hypothetical protein [Gibberella zeae]	877	0.0
C	FVEG_12519	hypothetical protein [Gibberella zeae]	306	4e-128
C	FVEG_12520	no hits found	-	-
C	FVEG_12521	aspartate kinase [Glomerella graminicola]	480	4e-143
C	FVEG_12522	no hits found	-	-
C	FVEG_12523	polyketide synthase [Gibberella moniliformis]	2841	0.0
C	FVEG_12524	hypothetical protein [Gibberella zeae]	483	5e-179
C	FVEG_12525	unnamed protein product [Aspergillus oryzae]	440	4e-126
C	FVEG_12526	hypothetical protein [Aspergillus oryzae]	407	1e-111
C	FVEG_12527	unnamed protein product [Aspergillus oryzae]	316	3e-115
C	FVEG_12528	no hits found	-	-
C	FVEG_12529	O-acetylhomoserine (thiol)-lyase-like protein [Chaetomium thermophilum]	375	4e-151
C	FVEG_12530	hypothetical protein [Botryotinia fuckeliana]	272	5e-143
C	FVEG_12531	no hits found	-	-
C	FVEG_12532	no hits found	-	-
C	FVEG_12533	hypothetical protein [Gibberella zeae]	332	2e-149
C	FVEG_12534	predicted protein [Aspergillus terreus]	224	2e-115
D	FVEG_03696	hypothetical protein [Gibberella zeae]	318	7e-157
D	FVEG_03697	predicted protein [Nectria haematococca]	481	0.0
D	FVEG_03698	hypothetical protein [Nectria haematococca]	256	1e-116
D	FVEG_03699	no hits found	-	-
F	FVEG_13461	hypothetical protein [Gibberella zeae]	1531	0.0
F	FVEG_13462	hypothetical protein [Gibberella zeae]	441	1e-121
F	FVEG_13463	no hits found	-	-
F	FVEG_13464	hypothetical protein [Gibberella zeae]	383	2e-104
F	FVEG_13465	no hits found	-	-

[0177] 実施例9

大腸菌における遺伝子クラスタ・スコアリングによるラクトースオペロンの
検出と検証

本発明の同定手法に従って、大腸菌のラクトースオペロンを検出した。大腸菌は原核生物であり、実施例 1 ～ 8 までは本発明手法の検証に用いた真核生物とは生物の分類上大きく異なる。

大腸菌はオペロンの存在が実証された最初の生物である。オペロンとは、ゲノム上に集合して機能を果たす一つの制御単位であり、複数の遺伝子がゲノム上にまとまって存在し高発現して機能するという性質上、本発明の同定対象に該当する。

ここで、本実施例で実証するラクトースオペロンについて説明する。ラクトースオペロンは、リプレッサータンパク質をコードする *lacI* に続いて、プロモーター配列 *lacP*、オペレーター配列 *lacO*、そしてラクトースを代謝する 3 つの遺伝子 *lacZ*, *lacY*, *lacA* (*lacZYA*) から構成される。*lacI* は常時発現しており、*lacO* 領域と強く結合するため、通常はその下流 *lacZYA* は翻訳されない。ところが、*lacI* に翻訳されるリプレッサータンパク質は、ラクトースが異性化したような誘導物質が存在すると、高次構造を変化させて、*lacO* 領域から遊離する。これによってラクトース代謝系である *lacZYA* が翻訳され、ラクトースの代謝が可能となる（参考文献 10）。

[0178] DNA マイクロアレイのデータは、米国国立生物工学情報センター（NCBI）が提供する遺伝子発現解析データの公共データベース GEO (<http://www.ncbi.nlm.nih.gov/geo/>) より、GSE7265 の ID で登録されたものを用いた（参考文献 11, 12）。このアレイデータは、大腸菌 MG1655 株およびその変異株を用いて、グルコースとラクトースの 2 つの栄養源を含む培地上で培養したときの遺伝子発現の変化を、分刻みで追ったものである。この 2 つの栄養源を含む培地上では、大腸菌はまずグルコースを代謝し、グルコースがなくなった後ラクトースを代謝する。すなわち、グルコースからラクトースへと栄養源を切り替える際に、最初に実証されたオペロンであるラクトースオペロンが発現する。ここではこのデータセットのうち、野生株のデータを用いる。野生株のデータには、培養開始後 17 段階におけるデータセットが含まれており、それぞれ、培養開始 780, 830, 861, 869, 878, 888, 898, 908, 919, 929, 939, 969, 9

99, 1035, 1049, 1070, 1089分後に取られたものである。各データは、対数増殖初期（780分後）のデータを分母とした発現誘導比の形で記載されているため、そのまま本手法に適応可能である。ただし、各測定段階について3～4つのデータが取られているため、各遺伝子について3～4つのデータ間で値を平均化した後、以下の手順に進んだ。なおデータに含まれる遺伝子数は4102である。

[0179] (A) クラスタ・スコアリング

17の各測定段階における系のそれぞれについて、計算式a)に従って $n_{c1} = 1 \sim 30$ でクラスタ・スコアリングを行い、各仮想遺伝子クラスタのM値を得た。なおクラスタ・スコアリングを行う際に必要となるゲノム上の遺伝子の連続情報は、公共の学術データベースであるNCBIに登録されている、大腸菌MG1655株のゲノム情報（ID:NC_000913;http://www.ncbi.nlm.nih.gov/nucore/NC_000913）を用いた。ここで大腸菌は環状ゲノムであるため、起点を上記ゲノム情報の中でb0001と名付けられた遺伝子とし、全ての遺伝子は連続しているものとして扱った。また、ラクトースオペロンを構成するlacI, lacZ, lacY, lacAの4つの遺伝子は、本ゲノム情報では向きが逆になっており、lacA, lacY, lacZ, lacIの順に並んでいる。これらの遺伝子IDはそれぞれ、b0342, b0343, b0344, b0345である。図47は、各仮想の遺伝子クラスタのM値のヒストグラムの一部である。左の拡大図をみると分かるように、ゼロを中心とした山型の正規分布様集団から外れて高いM値を持つ仮想遺伝子クラスタがあると、各 n_{c1} におけるM値のヒストグラムにおいて、山の中心が左側にずれる。図右側の各 n_{c1} におけるヒストグラムを上から見ていくと、 n_{c1} の増加とともに山の中心が左または右にずれていくことが分かる。これは、正規分布から外れて高い（低い）値を持つ仮想の遺伝子クラスタの存在を示している。

[0180] (B) データ判定

計算式e)にしたがって、17つの各系におけるスコア分布評価値eを算出した（図48）。ここで、仮想の遺伝子クラスタ数nは各 n_{c1} において4

102であり、次元数dは6を採用した。図にあるように、e値は、培養開始後878, 888, 898分後、および1049, 1070, 1089分後の6つの系において、大きな極大値を示す。ここでこの結果を、大腸菌の成長速度と照らし合わせる。図49は、本アレイデータに関する文献（参考文献11）に記載されている、培養開始後の大腸菌の増殖を表す濁度の時系列変化である。前培養をどこからとるかでアレイデータとの時間のラベルがずれているが、図49での開始点が、アレイデータの780分にあたり、以降の各点がアレイデータの各17段階に順次相当する。すると、図をみると分かるように、スコア分布評価値eが大きな極大値を示す878, 888, 898分後（7, 8, 9点目）および1049, 1070, 1089分後（15, 16, 17点目）のデータはすべて、図49において濁度の上昇が留まっている箇所、すなわち増殖の停滞期にあたる。このうち最初の停滞期は、グルコースを全て消費して栄養源をラクトースに切り替えようとしている段階である。この段階では増殖を一時停止するため、ゲノム上でまとまって存在する増殖に必須のリボソーム遺伝子群が強く抑制される一方、ラクトースを消費するためにラクトースオペロンを発現する。したがって、この段階でe値が大きな極大値を示していることは、リボソーム遺伝子群の抑制およびラクトースオペロンの発現という現象と一致する。二つ目の停滞期はラクトースも枯渇した段階であり、成長そのものが停滞するために、ここでも増殖に必須のリボソーム遺伝子が強く抑制される（参考文献13）。この段階でのe値の大きな極大値は、このリボソーム遺伝子の抑制を検出していると考えられる。

以上より、e値によって、ゲノム上でまとまって発現（または抑制）して機能する遺伝子群の存在を感度よく判定できることが示された。本実施例では、すでに同定されているラクトースオペロンを本手法によって検出できることを示すことが目的であるため、引き続き17段階すべてのデータを用いて以下の手順に進む。

[0181] (C) 遺伝子クラスタ判定

大腸菌MG1655株の培養開始後17段階におけるDNAマイクロアレイデー

タから、各仮想遺伝子クラスタについて、計算式 b) にしたがって遺伝子クラスタ判定値 c を算出した (図 5 0)。図において、一つの仮想の遺伝子クラスタにつき一本の線がグレーで描かれており、このうち黒の太線で描かれた線が、ラクトースオペロンを構成する 4 つの遺伝子のゲノム情報上の最初の遺伝子 $lacA$ (b0342) を起点とする遺伝子クラスタである。この遺伝子クラスタは、869 分の系から徐々に上昇を始め、908,919 分の系で、極大値を示す仮想の遺伝子クラスタの中で最大値を示している。極大値を示す点はクラスタサイズ 3 のときであり、 $lacZYA$ からなるものである。一方この遺伝子クラスタに $lacI$ は含まれないが、これは、 $lacI$ がラクトースオペロンの発現に関係なく常時発現しているという事実と一致する。

次に同様に、17 の各系における各仮想遺伝子クラスタについて、遺伝子クラスタ評価値 u を計算式 c) にしたがって算出した (図 5 1)。ここで次元数 d' は 2、係数 a は 1 を採用した。すると図のように、 c 値のときと同様、黒い太線で示した $lacA$ (b0342) から始まる遺伝子クラスタは、869 分の系から徐々に値の上昇を始め、908,919 分の系で、全ての仮想の遺伝子クラスタの中で最大の極大値をクラスタサイズ 3 で示す。これは、グルコースが枯渇してラクトース代謝系が動き出すときに、ラクトース代謝遺伝子群 $lacZYA$ が発現するという事実と一致する。またこの段階において、黒の太線とその他のグレーで示した線の差は、図 5 0 におけるよりも拡大しており、評価値 u を用いることで、該当遺伝子クラスタを c 値よりもさらに高く評価することが出来ることが示された。

[0182] こうして得られた c 値および u 値を元に、計算式 d) に従って二つの値の積から遺伝子クラスタ判定評価値 $c' \cdot u$ を算出した (図 5 2)。 c および u 値のみの場合 (図 5 0、図 5 1) よりもより鋭敏に、908 分の系で、 $lacZYA$ である黒の太線が検出されていることが分かる。図 5 3 は、評価値 $c' \cdot u$ について、各々の起点遺伝子について、 $n_{c \cdot l} = 1 \sim 30$ で構成した 30 個の仮想の遺伝子クラスタのうちの最大値を、横軸を起点遺伝子 ID として描いた図である。図において、17 の系全ての縦軸の縮尺は合わせてある。黒矢印で示したラク

トースオペロンが、908分の系で最大の値を示している。また、白抜きの矢印で示したリボソーム遺伝子群が、増殖停滞期である878,888,898分、そして1049,1070,1089分の系で、強く負の値を示している。これらより、本評価値が、ゲノム上で集合して機能する遺伝子群を、細胞の状態に応じて正確に検出できることが示された。

以上より本提案方法が、真核生物だけでなく原核生物においても、ゲノム上に集合して機能を果たす遺伝子群を検出する実効的な手段であることが示された。

[0183] (参考文献10)

The lactose repressor system: paradigms for regulation, allosteric behavior and protein folding

C. J. Wilsonら、Cellular and Molecular Life Sciences (2007) 64, 3-16

(参考文献11)

Gene expression profiling of Escherichia coli growth transitions: an expanded stringent response model

Dong-Eun Changら、Molecular Microbiology (2002) 45 (2), 289-306

(参考文献12)

Guanosine 3',5'-bispyrophosphate coordinates global gene expression during glucose-lactose diauxie in Escherichia coli

Matthew F. Traxlerら、Proceedings of the National Academy of Sciences (2006) 103 (7), 2374-2379

(参考文献13)

Control of protein synthesis in Escherichia coli. II. Translation and degradation of lactose operon messenger ribonucleic acid after energy source shift-down

K. C. Westoverら、Journal of Biological Chemistry (1974) 249 (19), 6280-6287

請求の範囲

- [請求項1] 生物ゲノム中の標的遺伝子を含む遺伝子クラスタ及び／または該遺伝子クラスタ中の標的遺伝子を探索する方法であって、生物細胞の生理状態変化を生じる条件とコントロール条件下において生じたゲノム遺伝子の発現量変動比を、ゲノムDNA上に配列する複数の遺伝子により構成される仮想の遺伝子クラスタ単位の発現量変動比として合算することにより、仮想の遺伝子クラスタ単位毎にスコアリングし、得られたスコアに基づき、上記生理状態変化の原因遺伝子である標的遺伝子を含む遺伝子クラスタ及び／または該遺伝子クラスタ中の標的遺伝子を探索する方法。
- [請求項2] 生物細胞の生理状態変化を生じる条件とコントロール条件下とを一つの対比条件セットとして、該対比条件セットが一種以上設定されていることを特徴とする請求項1に記載の方法。
- [請求項3] 生理状態変化を生じる条件とコントロール条件が、少なくとも代謝産物の産生誘導条件下と非誘導条件下あるいは代謝産物の産生抑制条件下と非抑制条件下との対比条件セットを含むことを特徴とする請求項1または2に記載の方法。
- [請求項4] 代謝物産生に関与する遺伝子が2次代謝物産生に関与する遺伝子であることを特徴とする請求項3に記載の方法。
- [請求項5] 仮想の各遺伝子クラスタは、ゲノムDNA上に連続する遺伝子を2個から遺伝子数を一つずつ増やして、想定される遺伝子クラスタに含まれる最大限のゲノム遺伝子数になるまで抽出し、かつ該抽出において、抽出する遺伝子の各個数毎に、直鎖状DNAからなるゲノムの場合には該DNAのいずれかの末端から、環状DNAからなるゲノムの場合には任意の遺伝子を起点として順にゲノムDNA上に配列する遺伝子一つずつずらしながら抽出された各遺伝子群からなることを特徴とする、上記請求項1～4のいずれかに記載の方法。
- [請求項6] スコアリングされる仮想の各遺伝子クラスタの集合体が、ゲノムD

N A上に連続する遺伝子を2個から遺伝子数を一つずつ増やし想定される遺伝子クラスタに含まれる最大限のゲノム遺伝子数になるまで抽出し、かつ該抽出において、抽出する遺伝子の各個数毎に、直鎖状DNAからなるゲノムの場合には該DNAのいずれかの末端から、環状DNAからなるゲノムの場合には任意の遺伝子を起点として順にゲノムDNA上に配列する遺伝子の一つずつずらしながら抽出された各遺伝子群からなる仮想の各遺伝子クラスタの集合からなり、ゲノム上に存在する遺伝子クラスタの全てが仮想の遺伝子クラスタの集合体中に含まれるように構成されていることを特徴とする請求項1～5のいずれかに記載の方法。

[請求項7] 仮想の各遺伝子クラスタのスコアリングが以下の計算式a) によりなされることを特徴とする、請求項1～6のいずれかに記載の方法。
計算式a)

[数1]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、Mは仮想の各遺伝子クラスタのスコア、mはスコアリングされる仮想の各遺伝子クラスタに含まれる各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の標準偏差を表す。)

[請求項8] ゲノムDNA上に配列する遺伝子が、標的とする遺伝子機能を有すると推定される場合、あるいは標的とする遺伝子機能を有する可能性が低いか若しくはその可能性がないと推定される場合において、当該ゲノムDNA上に配列する遺伝子については、以下の重み付け計算が適用されることを特徴とする、請求項7の方法。

[数2]

$$w \frac{m - \bar{m}}{\sigma(m)}$$

(式中、 m は、上記標的とする遺伝子機能を有するか、あるいは標的とする遺伝子機能を有する可能性が低い若しくはその可能性がないと推定されるゲノム DNA 上の遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値) の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値) の標準偏差、 w は重みとしての任意の実数を、それぞれ表す。)

[請求項9] ゲノム DNA 上に配列する遺伝子が、標的とする遺伝子機能を有すると推定される場合において、標的とする遺伝子機能を有すると推定された遺伝子を含む仮想の遺伝子クラスタを選出し、選出された仮想の遺伝子クラスタについて、スコアリングすることを特徴とする、請求項 7 に記載の方法。

[請求項10] 仮想の遺伝子クラスタが、ゲノムにおいて近傍に存在することを条件として、以下の 1) ～ 3) の内の 1 以上の遺伝子のみから、あるいは該遺伝子を少なくとも含む 1 以上の遺伝子から構築されることを特徴とする、上記 (4) に記載の方法。

1) 2 次代謝物産生に関与していると想定される酵素種に属する酵素遺伝子。

2) トランスポーター遺伝子

3) 転写因子をコードする遺伝子

[請求項11] 仮想の各遺伝子クラスタのスコアリングが以下の計算式 a) によりなされることを特徴とする、上記 (10) に記載の方法。
計算式 a)

[数3]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、Mは各仮想の遺伝子クラスタのスコア、mはスコアリングされる仮想の各遺伝子クラスタに含まれるアノテーション付与に基づき選定された各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれるアノテーション付与に基づき選定された全遺伝子の発現量変動比 (m 値) の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれるアノテーション付与に基づき選定された全遺伝子の発現量変動比 (m 値) の標準偏差を表す。)

[請求項12] 仮想の遺伝子クラスタ全体のスコアの分布から乖離して存在するスコアを有する仮想の遺伝子クラスタを、標的の遺伝子クラスタ候補として選定することを特徴とする、請求項1～11のいずれかに記載の方法。

[請求項13] 仮想の遺伝子クラスタ全体のスコアの分布からの乖離の程度を示す判定値 $I(\chi)$ を、以下の計算式b)により算出し、算出された該判定値 $I(\chi)$ に基づき仮想の遺伝子クラスタを標的の遺伝子クラスタ候補として選定することを特徴とする、請求項12に記載の方法。

計算式b)

[数4]

$$\chi = -M \log P$$

(式中、 χ は上記仮想の各遺伝子クラスタの乖離の程度を示す判定値 I 、Mは仮想の各遺伝子クラスタのスコア、Pは、全ての仮想の遺伝子クラスタについての各スコアに対する出現頻度分布において、各スコアMの出現頻度について全体累計を1としたときの、上記スコアMの出現頻度の割合を表す。)

[請求項14] 仮想の遺伝子クラスタ全体のスコアの分布からの乖離の程度を示す判定値 $I(v)$ を、以下の計算式c)により算出し、算出された判定値 $I(v)$ に基づき仮想の遺伝子クラスタを、標的の遺伝子クラスタ候補として選定することを特徴とする、請求項12に記載の方法。

。

計算式c)

[数5]

$$v = (M - \overline{M})^{d'} / (a\sigma(M))^{d'}$$

(式中、 v は上記仮想の各遺伝子クラスタの乖離の程度を示す判定値、 M は仮想の各遺伝子クラスタのスコア、 $\overline{M}$ は全ての仮想の遺伝子クラスタのスコア（ M 値）の平均値、 $\sigma(M)$ は全ての仮想の遺伝子クラスタにおけるスコア（ M 値）の標準偏差、 a は任意の正の実数、 d' は次元数となる正の偶数を表す。)

[請求項15]

さらに、以下の計算式 d) の算出結果に基づき、 b が 100 未満の仮想のクラスタを少なくとも除外し、標的の遺伝子クラスタ候補をさらに絞り込むことを特徴とする、請求項 13 または 14 に記載の方法。

計算式 d)

[数11]

$$\chi \times v > b$$

(式中、 χ は 請求項13 に記載の計算式 b) によって算出した仮想の各遺伝子クラスタの判定値、 v は 請求項14 に記載の計算式 c) によって算出した仮想の各遺伝子クラスタの判定値、 b は閾値としての任意の正の実数を表す。)

[請求項16]

生物細胞の生理状態変化を生じる条件とコントロール条件下において生じたゲノム DNA 上に配列する各遺伝子の発現量変動比を、ゲノム DNA 上に配列する複数遺伝子により構成される仮想の遺伝子クラスタ単位の発現量変動比として合算することにより、仮想の遺伝子クラスタ単位毎にスコアリングし、得られたスコアに基づき、標的とする遺伝子クラスタがゲノム中に存在するか否かあるいは、標的遺伝子クラスタが存在する場合の遺伝子サイズを予測する方法であって、

ゲノム DNA 上に連続する遺伝子を 2 個から遺伝子数を一つずつ増やし想定される遺伝子クラスタに含まれる最大限のゲノム遺伝子数になるまで抽出し、かつ該抽出において抽出する遺伝子の各個数毎に、直鎖状 DNA からなるゲノムの場合には該 DNA のいずれかの末端から、あるいは環状 DNA からなるゲノムの場合には任意の遺伝子を起点として順にゲノム DNA 上に配列する遺伝子一つずつずらしなが

ら抽出された各遺伝子群から構成された仮想の各遺伝子クラスタを、以下の計算式 a) によりスコアリングし、この得られた仮想の各遺伝子クラスタのスコアを各遺伝子クラスタに含まれる遺伝子数毎に分け、以下の計算式 e) により、各遺伝子数単位毎に遺伝子クラスタスコア分布判定値 (ε) を求め、該判定値に基づき、予め、標的とする遺伝子クラスタがゲノム中に存在するか否かあるいは、標的クラスタが存在する場合のその遺伝子サイズを予測することを特徴とする、上記方法。

計算式 a)

[数1]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、M は仮想の各遺伝子クラスタのスコア、m はスコアリングされる仮想の各遺伝子クラスタに含まれる各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値) の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比 (m 値) の標準偏差を表す。)

計算式 e)

[数7]

$$\varepsilon = \sum (M - \bar{M})^d / n \sigma(M)^d$$

(式中、 ε は各遺伝子数単位毎に求められた遺伝子クラスタスコア分布判定値、M は全ての仮想の遺伝子クラスタを遺伝子数単位毎に分けた場合の、各遺伝子数単位に含まれる仮想の各遺伝子クラスタのスコア、 $\bar{M}$ は全ての仮想の遺伝子クラスタのスコアの平均、n は仮想の遺伝子クラスタの総数、 $\sigma(M)$ は全ての仮想の遺伝子クラスタのスコア (M 値) の標準偏差、d は任意に設定する次元数としての正の偶数を表す。)

[請求項17]

遺伝子数が k 個のときの ε 値 ($\varepsilon(k)$) と、その前後数のときの ε 値 ($\varepsilon(k-1)$ 、 $\varepsilon(k+1)$) が、以下の関係にあるとき、標的とする遺伝子クラスタがゲノム中に存在すると判定し、標的遺伝子クラスタに含まれる遺伝子数を k 個と予想することを特徴とする、請求項 16 に記載の方法。

[数8]

$$\varepsilon(k) > \varepsilon(k-1) \quad \text{かつ} \quad \varepsilon(k) > \varepsilon(k+1)$$

- [請求項18] 生物ゲノム中の標的遺伝子を含む遺伝子クラスタ及び／または該遺伝子クラスタ中の標的遺伝子を探索する装置であって、 a) 生物細胞の生理状態変化を生じる条件とコントロール条件下におけるゲノムDNA上に配列する各遺伝子の発現量データに基づき算出された上記2つの条件下における上記各遺伝子の発現量変動比を記憶する手段、 b) ゲノムDNA上に配列する複数の遺伝子を組み合わせて仮想の遺伝子クラスタを構築する手段、 c) 該算出され、記憶されたゲノムDNA上に配列する各遺伝子の発現量変動比を複数の遺伝子により構築された上記仮想の遺伝子クラスタ単位の発現量変動比として合算し、仮想の遺伝子クラスタ単位毎にスコアリングし、仮想の各遺伝子クラスタのスコアを記憶する手段、及びd) 得られたスコアに基づき上記生理状態変化の原因遺伝子である標的遺伝子を含む遺伝子クラスタを選定する手段を有するか、あるいはさらにe) 選定された遺伝子クラスタ中に含まれる遺伝子を表示する手段を有することを特徴とする、上記装置。
- [請求項19] 発現量データが、遺伝子発現量測定用DNAマイクロアレイによる蛍光強度情報であることを特徴とする請求項18に記載の装置。
- [請求項20] 蛍光強度情報が、蛍光強度を読み取り、数値化する手段を有する蛍光強度読み取り装置により出力される数値データであることを特徴とする、請求項19に記載の装置。
- [請求項21] 生物細胞の生理状態変化を生じる条件とコントロール条件とを1の対比条件セットとして1以上設定されている場合において、各対比条件セットに含まれる条件毎に各遺伝子の発現量データが入力され、各対比条件セットにおける同一遺伝子の発現量変動比が算出されることを特徴とする、請求項18～20のいずれかに記載の装置。

- [請求項22] 標的遺伝子が代謝物産生に関与する遺伝子であることを特徴とする、請求項18～21のいずれかに記載の装置。
- [請求項23] 代謝物産生に関与する遺伝子が2次代謝物産生に関与する遺伝子であることを特徴とする、請求項22に記載の装置。
- [請求項24] 設定される対比条件セットが、少なくとも代謝産物の産生誘導条件下と非誘導条件下あるいは代謝産物の産生抑制条件下と非抑制条件下との対比条件セットを含むことを特徴とする請求項22に記載の装置。
- [請求項25] 代謝産物が2次代謝産物であることを特徴とする、請求項24に記載の装置。
- [請求項26] 仮想の各遺伝子クラスタの構築手段が、ゲノムDNA上に連続する遺伝子を2個から遺伝子数を1ずつ増やして、想定される遺伝子クラスタに含まれる最大限のゲノム遺伝子数になるまで抽出し、かつ該抽出において、抽出する遺伝子の各個数毎に、直鎖状DNAからなるゲノムの場合には該DNAのいずれかの末端から、環状DNAからなるゲノムの場合には任意の遺伝子を起点として順にゲノムDNA上に配列する遺伝子を一つずつずらしながら抽出した各遺伝子群により構築する手段であることを特徴とする、請求項18～25のいずれかに記載の装置。
- [請求項27] 仮想の各遺伝子クラスタのスコアリングが以下の計算式a)によりなされることを特徴とする、請求項18～26のいずれかに記載の装置。
- 計算式 a)

[数1]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、Mは仮想の各遺伝子クラスタのスコア、mはスコアリングされる仮想の各遺伝子クラスタに含まれる各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の標準偏差を表す。)

[請求項28] ゲノムDNA上に配列する各遺伝子中の特定の遺伝子を選定するためのアノテーション付与手段を有し、上記遺伝子クラスタのスコアリングにおいて、付与されたアノテーションに基づき選定された遺伝子についての発現量変動比計算を以下の重み付け計算式により行うことを特徴とする、請求項27に記載の装置。

[数2]

$$w \frac{m - \bar{m}}{\sigma(m)}$$

(式中、mは、上記標的とする遺伝子機能を有するか、あるいは標的とする遺伝子機能を有する可能性が低い若しくはその可能性がないと推定されるゲノムDNA上の遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の標準偏差、wは重みとしての任意の実数を、それぞれ表す。)

[請求項29] アノテーション付与手段が、それぞれ遺伝子機能の種類毎に異なるアノテーションを付与する手段であることを特徴とする請求項28に記載の装置。

[請求項30] アノテーションに基づき選定される遺伝子が、1)～3)のうちの1以上の遺伝子であることを特徴とする、請求項29に記載の装置

- 1) 2次代謝物産生に関与していると想定される酵素種に属する酵素遺伝子。
- 2) トランスポーター遺伝子
- 3) 転写因子をコードする遺伝子

[請求項31] 上記請求項28～30のいずれかに記載のアノテーション付与手段と、構築された仮想の遺伝子クラスタから、アノテーションに基づき選出された遺伝子を含む仮想の遺伝子クラスタを選出する手段を有し、選出された仮想の遺伝子クラスタについてスコアリングすることを特徴とする、請求項27に記載の装置。

[請求項32] ゲノムDNA上に配列する各遺伝子中の特定遺伝子を選定するためのアノテーション付与手段を有し、ゲノムDNA上において近傍に位置することを条件として、アノテーションに基づき選定された遺伝子のみから、あるいは該遺伝子を少なくとも含む1以上の遺伝子から仮想の遺伝子クラスタを構築する手段を有することを特徴とする、請求項18～25のいずれかに記載の装置。

[請求項33] 請求項32に記載のアノテーション付与手段が、それぞれ遺伝子機能の種類に応じたアノテーションを付与する手段であることを特徴とする請求項32に記載の装置。

[請求項34] アノテーション付与に基づき選定される遺伝子が、1)～3)のうちの1以上の遺伝子であることを特徴とする、請求項33に記載の装置

1) 2次代謝物産生に関与していると想定される酵素種に属する酵素遺伝子。

2) トランスポーター遺伝子

3) 転写因子をコードする遺伝子

[請求項35] 仮想の各遺伝子クラスタのスコアリングが以下の計算式a) によりなされることを特徴とする、請求項32～34のいずれかに記載の装置。

計算式 a)

[数3]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、 M は各仮想の遺伝子クラスタのスコア、 m はスコアリングされる仮想の各遺伝子クラスタに含まれるアノテーション付与に基づき選定された各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれるアノテーション付与に基づき選定された全遺伝子の発現量変動比 (m 値) の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれるアノテーション付与に基づき選定された全遺伝子の発現量変動比 (m 値) の標準偏差を表す。)

[請求項36] 仮想の遺伝子クラスタ全体のスコアの分布から乖離して存在するスコアを有する仮想の遺伝子クラスタを、標的の遺伝子クラスタ候補として選定する手段を有することを特徴とする、請求項18～35のいずれかに記載の装置。

[請求項37] 標的の遺伝子クラスタ候補として選定する手段として、仮想の遺伝子クラスタ全体のスコアの分布からの乖離の程度を示す判定値 I (χ) を、以下の計算式 b) により算出するプログラムが格納されていることを特徴とする、請求項36に記載の装置。

計算式 b)

[数4]

$$\chi = -M \log P$$

(式中、 χ は上記仮想の各遺伝子クラスタの乖離の程度を示す判定値 I 、 M は仮想の各遺伝子クラスタのスコア、 P は、全ての仮想の遺伝子クラスタについての各スコアに対する出現頻度分布において、各スコア M の出現頻度について全体累計を1としたときの、上記スコア M の出現頻度の割合を表す。)

[請求項38] 標的の遺伝子クラスタ候補として選定する手段として、遺伝子クラスタ全体のスコアの分布からの乖離の程度を示す判定値 II (v) を、以下の計算式 c) により算出するプログラムが格納されていることを特徴とする、請求項37に記載の装置。

計算式 c)

[数5]

$$v = (M - \overline{M})^{d'} / (a\sigma(M))^{d'}$$

(式中、 v は上記仮想の各遺伝子クラスタの乖離の程度を示す判定値、 M は仮想の各遺伝子クラスタのスコア、 $\overline{M}$ は全ての仮想の遺伝子クラスタのスコア（ M 値）の平均値、 $\sigma(M)$ は全ての仮想の遺伝子クラスタにおけるスコア（ M 値）の標準偏差、 a は任意の正の実数、 d' は次元数となる正の偶数を表す。)

[請求項39] さらに、以下の計算式 d) の算出結果に基づき、 b が 100 未満の仮想のクラスタを少なくとも除外し、標的の遺伝子クラスタ候補をさらに絞り込むプログラムが格納されていることを特徴とする、請求項 37 または 38 に記載の装置。

計算式 d)

[数12]

$$\chi \times v > b$$

(式中、 χ は 請求項37 に記載の計算式 b) によって算出した仮想の各遺伝子クラスタの判定値、 v は 請求項38 に記載の計算式 c) によって算出した仮想の各遺伝子クラスタの判定値、 b は閾値としての任意の正の実数を表す。)

[請求項40] a) 生物細胞の生理状態変化を生じる条件とコントロール条件下において生じたゲノム DNA 上に配列する各遺伝子の発現量を入力する手段、b) 入力された上記 2 つの条件下における同一遺伝子の発現量の比を算出する発現量変動比算出手段、c) 該算出されたゲノム DNA 上に配列する各遺伝子の発現量変動比を複数の遺伝子により構築された上記仮想の遺伝子クラスタ単位の発現量変動比として合算し、仮想の遺伝子クラスタ単位毎にスコアリングする手段、及び d) 得られた仮想の遺伝子クラスタのスコアから遺伝子クラスタに含まれる遺伝子数単位毎の遺伝子クラスタ分布判定値 (ε) を算出する手段を有し、該遺伝子クラスタ分布判定値 (ε) から、標的とする遺伝子クラスタがゲノム中に存在する否かあるいは、標的遺伝子クラスタが存在する場合の遺伝子サイズを予測する装置であって、仮想の遺伝子クラスタの構築手段が、ゲノム DNA 上に連続する遺伝子を 2 個から遺伝子

数を一つずつ増やし想定される遺伝子クラスタに含まれる最大限のゲノム遺伝子数になるまで抽出し、かつ該抽出において抽出する遺伝子の各個数毎に、直鎖状DNAからなるゲノムの場合には該DNAのいずれかの末端から、あるいは環状DNAからなるゲノムの場合には任意の遺伝子を起点として順にゲノムDNA上に配列する遺伝子一つずつずらしながら抽出された各遺伝子群を仮想の各遺伝子クラスタとする手段であり、上記仮想の遺伝子クラスタ単位のスコアリング手段は以下の計算式a)による演算手段からなるとともに、上記遺伝子クラスタ分布判定値(ε)の算出手段が、以下の計算式e)によるものであることを特徴とする、上記装置。

計算式a)

[数1]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、Mは仮想の各遺伝子クラスタのスコア、mはスコアリングされる仮想の各遺伝子クラスタに含まれる各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の標準偏差を表す。)

計算式e)

[数7]

$$\varepsilon = \sum (M - \bar{M})^d / n \sigma(M)^d$$

(式中、 ε は各遺伝子数単位毎に求められた遺伝子クラスタスコア分布判定値、Mは全ての仮想の遺伝子クラスタを遺伝子数単位毎に分けた場合の、各遺伝子数単位に含まれる仮想の各遺伝子クラスタのスコア、 $\bar{M}$ は全ての仮想の遺伝子クラスタのスコアの平均、nは仮想の遺伝子クラスタの総数、 $\sigma(M)$ は全ての仮想の遺伝子クラスタのスコア(M値)の標準偏差、dは任意に設定する次元数としての正の偶数を表す。)

[請求項41]

遺伝子数がk個のときの遺伝子クラスタ分布判定値 ε 値($\varepsilon(k)$)

と、その前後数のときの同 ε 値($\varepsilon(k-1)$ 、 $\varepsilon(k+1)$)が、

以下の関係にあるとき、標的とする遺伝子クラスタがゲノム中に存在

すると判定し、標的遺伝子クラスタに含まれる遺伝子数を k 個とする
予想値を出力することを特徴とする、請求項 40 に記載の装置。

[数8]

$$\varepsilon(k) > \varepsilon(k-1) \quad \text{かつ} \quad \varepsilon(k) > \varepsilon(k+1)$$

[請求項42] 請求項 26 に記載の仮想の遺伝子クラスタの構築手段を実行するプログラムであって、ゲノム遺伝子の位置情報に基づき、以下の 1) または 2) の手段を実行することを特徴とする、仮想の遺伝子クラスタ構築プログラム。

1) ゲノム遺伝子が直鎖状ゲノムの場合、

a. ゲノム DNA の一方の末端に位置する遺伝子を起点として、他方の末端方向に、順次、ゲノム DNA 上に連続する遺伝子を同一方向に 2 個から一つずつ増やして想定される遺伝子クラスタに含まれる遺伝子数の最大限になるまで組み合わせ、起点とした遺伝子を含み、かつ遺伝子の個数の異なる複数の遺伝子群を構成する手段。

b. 起点を、順次、他方の末端方向に一遺伝子ずつずらせながら、上記 a. と同様の処理を行い、新たな起点遺伝子を含みかつ遺伝子の個数が異なる複数の遺伝子群を構成し、a. の遺伝子群と併せて、複数の遺伝子を組み合わせた遺伝子群からなる仮想の遺伝子クラスタを構築する手段。

2) ゲノム遺伝子が環状の場合、ゲノム DNA 上の任意の遺伝子を起点として、上記 1) a. 及び b. と同様の処理を順次行い、最初に起点とした遺伝子が起点となる時点で処理を終了する手段。

[請求項43] 請求項 42 のプログラムにより構築された仮想の遺伝子クラスタについて、以下の計算式 a) によるスコアリングを実行することを特徴とする、仮想の遺伝子クラスタのスコアリングプログラム。
計算式 a)

[数1]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、Mは仮想の各遺伝子クラスタのスコア、mはスコアリングされる仮想の各遺伝子クラスタに含まれる各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の標準偏差を表す。)

[請求項44] 上記遺伝子クラスタのスコアリングにおいて、付与されたアノテーションに基づきゲノム遺伝子を選定し、選定された遺伝子についての発現量変動比計算を以下の重み付け計算式により行うことを特徴とする、請求項43に記載のスコアリングプログラム。

[数2]

$$w \frac{m - \bar{m}}{\sigma(m)}$$

(式中、mは、上記標的とする遺伝子機能を有するか、あるいは標的とする遺伝子機能を有する可能性が低い若しくはその可能性がないと推定されるゲノムDNA上の遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の標準偏差、wは重みとしての任意の実数を、それぞれ表す。)

[請求項45] 上記遺伝子クラスタのスコアリングにおいて、付与されたアノテーションに基づきゲノム遺伝子を選定し、構築された遺伝子クラスタの中から、該選定されたゲノム遺伝子を含む仮想の遺伝子クラスタを選出し、選出された仮想の遺伝子クラスタについてスコアリングを実行することを特徴とする、請求項43に記載のスコアリングプログラム。

[請求項46] 上記請求項32に記載の仮想の遺伝子クラスタの構築手段を実行するプログラムであって、ゲノムDNA上において近傍に位置することを条件として、アノテーションに基づき選定された遺伝子のみから、あるいは該遺伝子を少なくとも含む1以上の遺伝子から仮想の遺伝子

クラスタを構築することを特徴とする、仮想の遺伝子クラスタの構築プログラム。

[請求項47] 請求項46のプログラムにより構築された仮想の遺伝子クラスタについて、以下の計算式a)によるスコアリングを実行することを特徴とする、仮想の遺伝子クラスタのスコアリングプログラム。

計算式a)

[数3]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、Mは各仮想の遺伝子クラスタのスコア、mはスコアリングされる仮想の各遺伝子クラスタに含まれるアノテーション付与に基づき選定された各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれるアノテーション付与に基づき選定された全遺伝子の発現量変動比(m値)の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれるアノテーション付与に基づき選定された全遺伝子の発現量変動比(m値)の標準偏差を表す。)

[請求項48] 請求項43～45又は47のいずれかに記載のスコアリングプログラムにより算出された各仮想の遺伝子クラスタのスコアについて、仮想の遺伝子クラスタ全体のスコアの分布からの乖離の程度を算出するプログラムであって、以下の計算式b)により、判定値I(χ)を算出することを特徴とする、上記プログラム。

計算式b)

[数4]

$$\chi = -M \log P$$

(式中、 χ は上記仮想の各遺伝子クラスタの乖離の程度を示す判定値I、Mは仮想の各遺伝子クラスタのスコア、Pは、全ての仮想の遺伝子クラスタについての各スコアに対する出現頻度分布において、各スコアMの出現頻度について全体累計を1としたときの、上記スコアMの出現頻度の割合を表す。)

[請求項49] 請求項43～45又は47のいずれかに記載のスコアリングプログラムにより算出された各仮想の遺伝子クラスタのスコアについて、仮想の遺伝子クラスタ全体のスコア分布からの乖離の程度を算出するプログラムであって、以下の計算式c)により判定値II(ψ)の算出を

実行する、上記プログラム。

計算式 c)

[数5]

$$v = (M - \bar{M})^{d'} / (a\sigma(M))^{d'}$$

(式中、 v は上記仮想の各遺伝子クラスタの乖離の程度を示す判定値、 M は仮想の各遺伝子クラスタのスコア、 $\bar{M}$ は全ての仮想の遺伝子クラスタのスコア (M 値) の平均値、 $\sigma(M)$ は全ての仮想の遺伝子クラスタにおけるスコア (M 値) の標準偏差、 a は任意の正の実数、 d' は次元数となる正の偶数を表す。)

[請求項50]

生物細胞の生理状態変化を生じる条件とコントロール条件下とにおけるゲノムDNA上に配列する各遺伝子の発現量変動比を複数の遺伝子により構築された上記仮想の遺伝子クラスタ単位の発現量変動比として合算し、仮想の遺伝子クラスタ単位毎にスコアリングする手段、及び得られた仮想の遺伝子クラスタのスコアから遺伝子クラスタに含まれる遺伝子数単位毎の遺伝子クラスタ分布判定値 (ε) を算出し、該遺伝子クラスタ分布判定値 (ε) から、標的とする遺伝子クラスタがゲノム中に存在する否かあるいは、標的遺伝子クラスタが存在する場合の遺伝子サイズを予測する手段に用いるプログラムであって、

少なくとも以下 (A) ~ (C) の手段を実行するプログラム。

(A) ゲノム遺伝子の位置情報に基づき、以下の1) または2) の手段により仮想の遺伝子クラスタを構築する手段、

1) ゲノム遺伝子が直鎖状の場合、

a. ゲノムDNAの一方の末端に位置する遺伝子を起点として、他方の末端方向に、順次、ゲノムDNA上に連続する遺伝子を同一方向に2個から一つずつ増やして想定される遺伝子クラスタに含まれる遺伝子数の最大限になるまで組み合わせ、起点とした遺伝子を含み、かつ遺伝子の個数の異なる複数の遺伝子群を構成する手段。

b. 起点を、順次、他方の末端方向に一遺伝子ずつずらせながら、上記 a. と同様の処理を行い、新たな起点遺伝子を含みかつ遺伝子の個

数が異なる複数の遺伝子群を構成し、a. の遺伝子群と併せて、複数の遺伝子の組み合わせた遺伝子群からなる仮想の遺伝子クラスタを構築する手段。

2) ゲノム遺伝子が環状の場合、ゲノムDNA上の任意の遺伝子を起点として、上記1) a. 及びb. と同様の処理を順次行い、最初に起点とした遺伝子が起点となる時点で処理を終了する手段。

(B) 上記(A)の手段により構築された仮想の遺伝子クラスタについて、以下の計算式a)により仮想の遺伝子クラスタ単位毎にスコアリングする手段。

計算式 a)

[数1]

$$M = \sum \frac{m - \bar{m}}{\sigma(m)}$$

(式中、Mは仮想の各遺伝子クラスタのスコア、mはスコアリングされる仮想の各遺伝子クラスタに含まれる各遺伝子の発現量変動比、 $\bar{m}$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の平均、 $\sigma(m)$ は全ての仮想の遺伝子クラスタに含まれる全遺伝子の発現量変動比(m値)の標準偏差を表す。)

(C) 上記(B)の手段により得られた仮想の遺伝子クラスタのスコアから、以下の計算式e)により仮想の遺伝子クラスタに含まれる遺伝子数単位毎の遺伝子クラスタ分布判定値(ε)を算出する手段。

計算式 e)

[数7]

$$\varepsilon = \sum (M - \bar{M})^d / n \sigma(M)^d$$

(式中、 ε は各遺伝子数単位毎に求められた遺伝子クラスタスコア分布判定値、Mは全ての仮想の遺伝子クラスタを遺伝子数単位毎に分けた場合の、各遺伝子数単位に含まれる仮想の各遺伝子クラスタのスコア、 $\bar{M}$ は全ての仮想の遺伝子クラスタのスコアの平均、nは仮想の遺伝子クラスタの総数、 $\sigma(M)$ は全ての仮想の遺伝子クラスタのスコア(M値)の標準偏差、dは任意に設定する次元数としての正の偶数を表す。)

[請求項51]

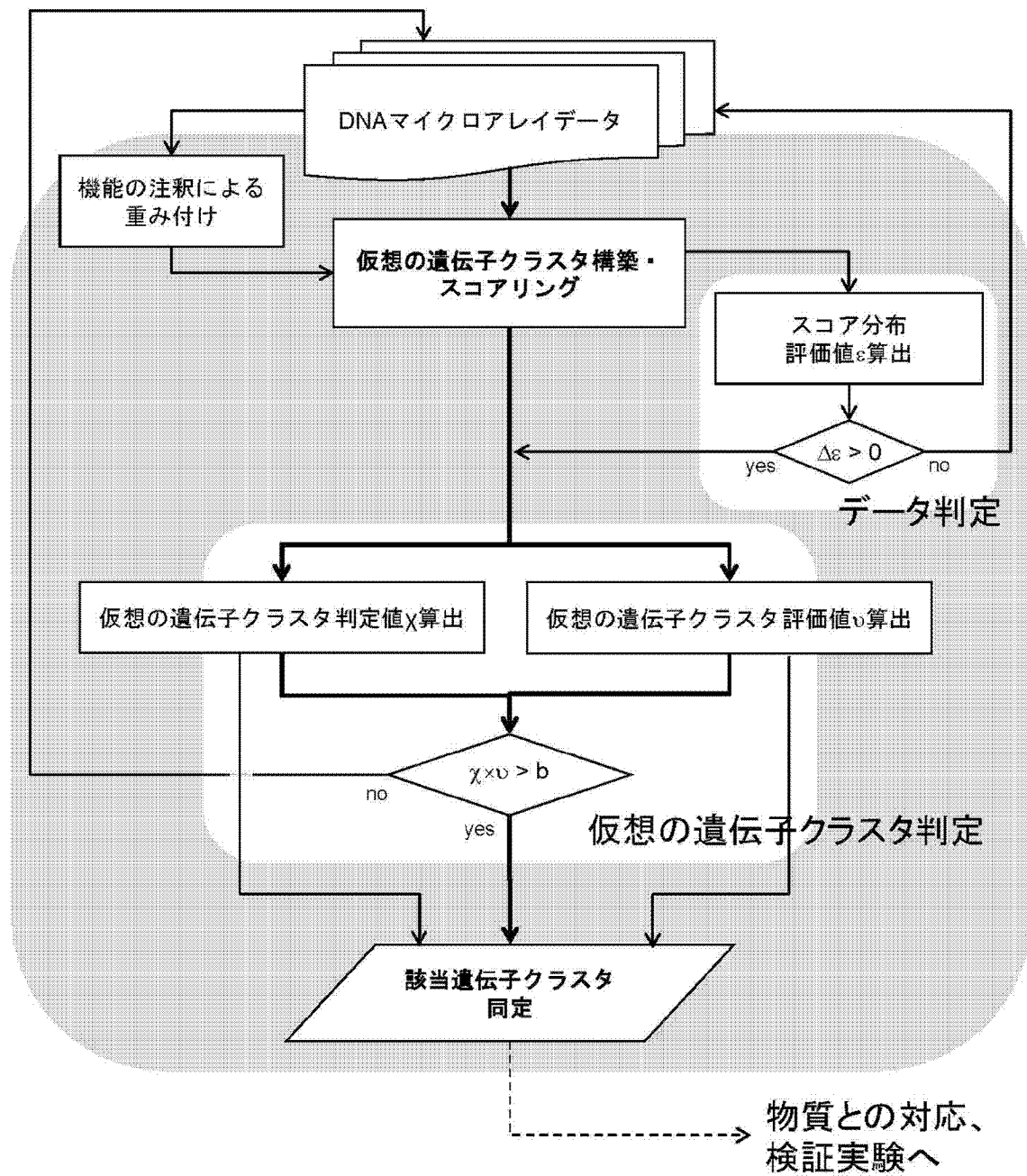
遺伝子数がk個のときの遺伝子クラスタ分布判定値 ε 値($\varepsilon(k)$)

）と、その前後数のときの同 ε 値（ $\varepsilon(k-1)$ 、 $\varepsilon(k+1)$ ）が、以下の関係にあるとき、標的とする遺伝子クラスタがゲノム中に存在すると判定し、標的遺伝子クラスタに含まれる遺伝子数を k 個とする予想値を出力することを特徴とする、請求項 50 に記載のプログラム。

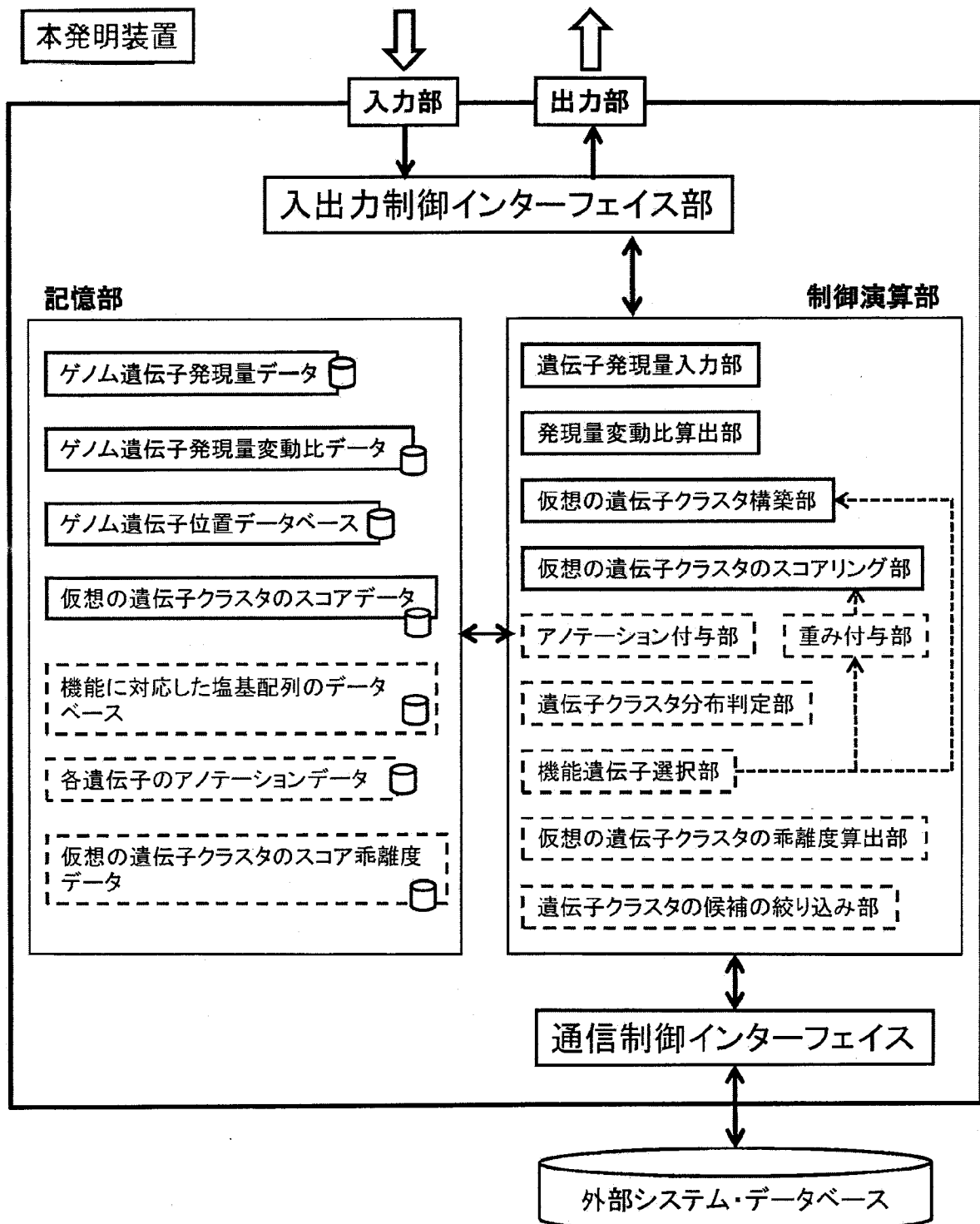
[数8]

$$\varepsilon(k) > \varepsilon(k-1) \quad \text{かつ} \quad \varepsilon(k) > \varepsilon(k+1)$$

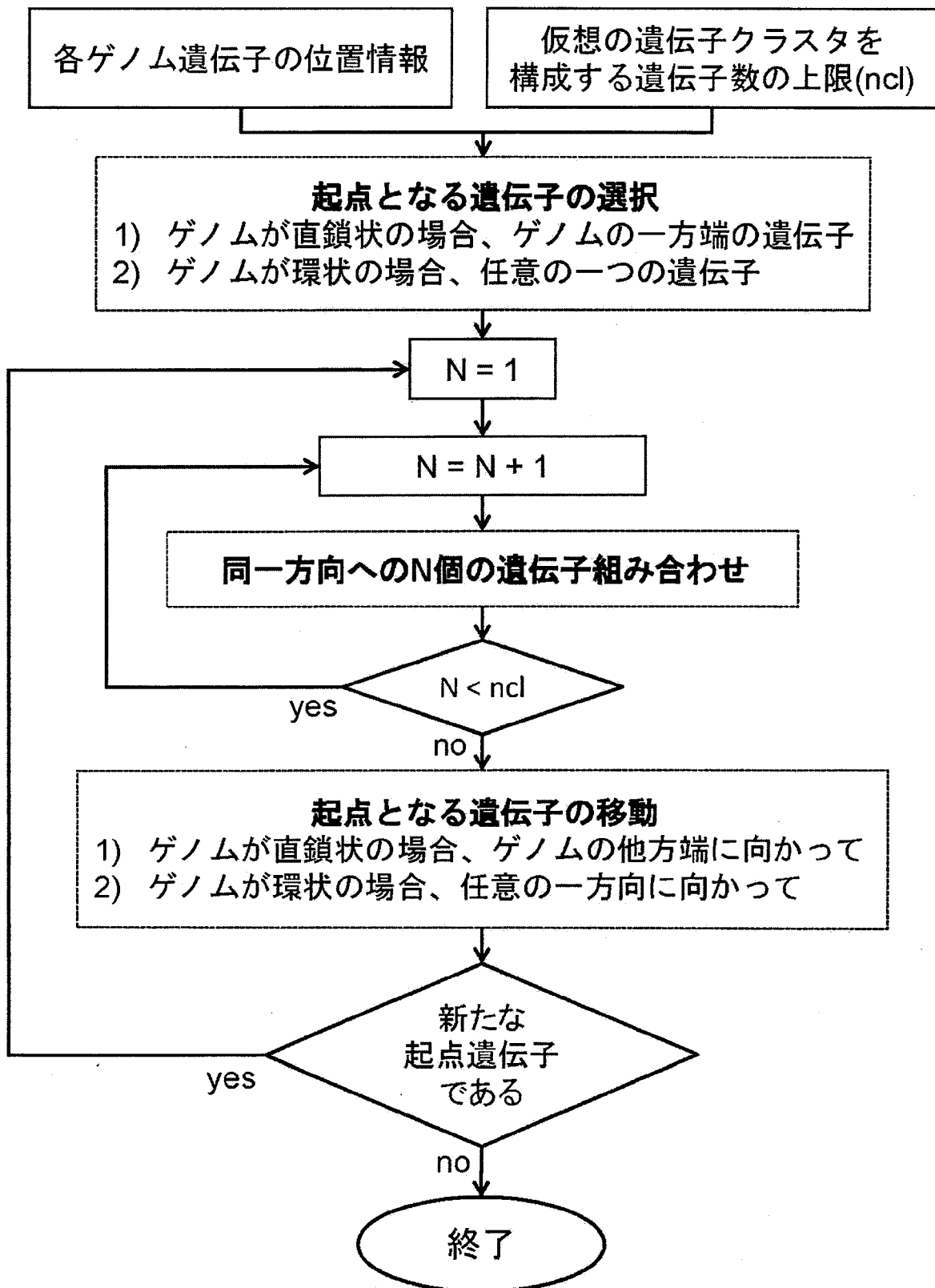
[図1]



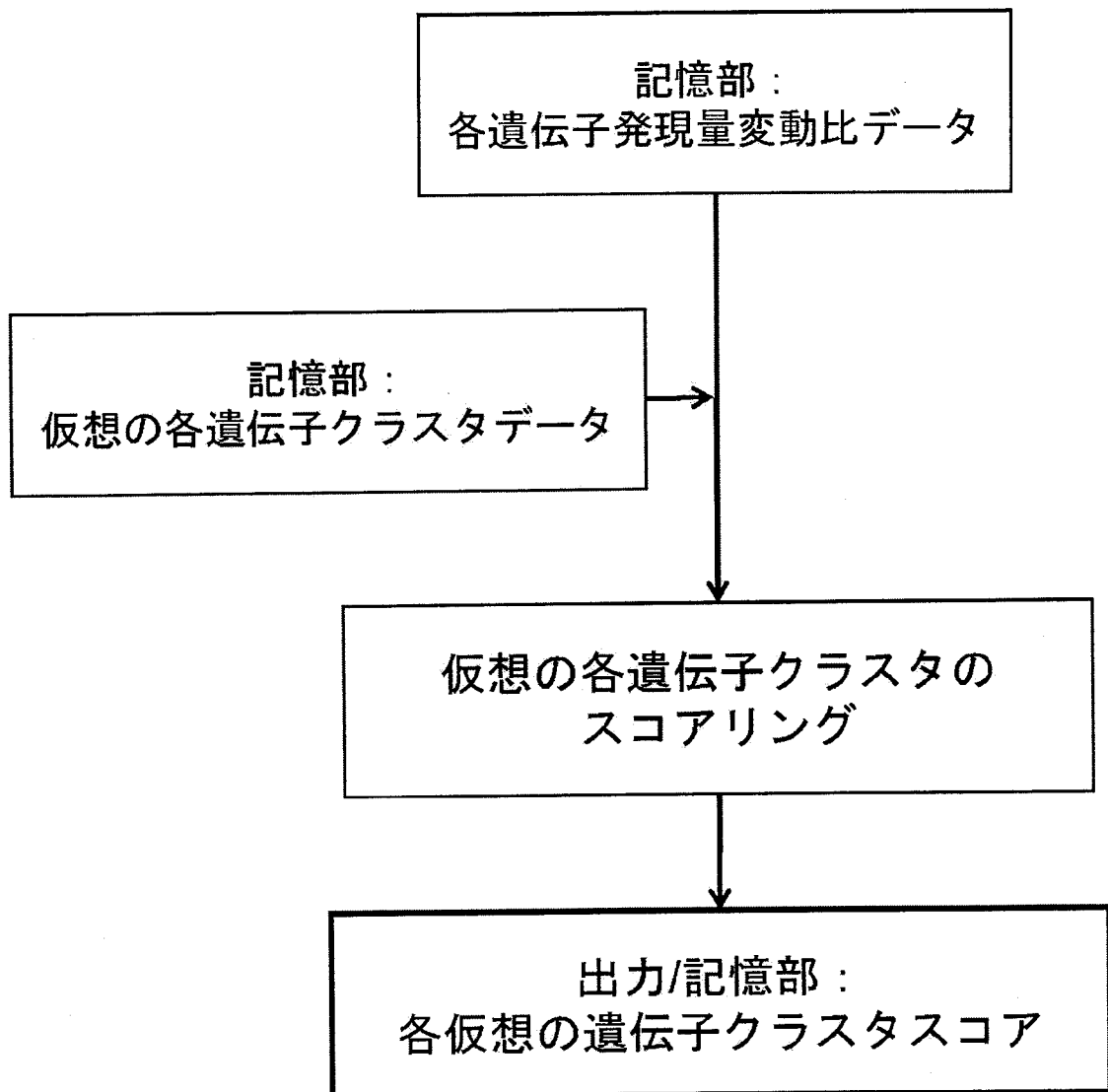
[図2]



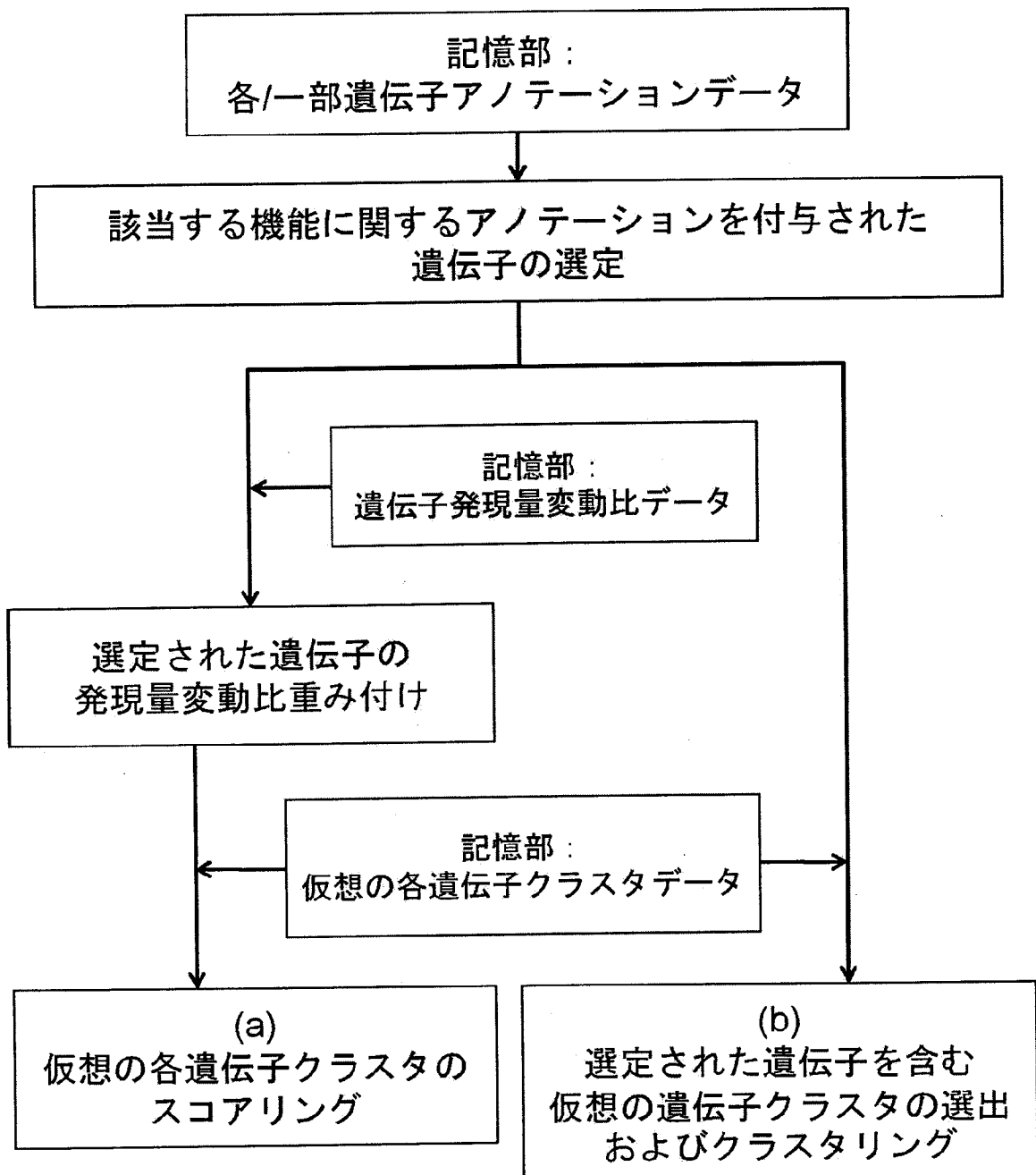
[図3]



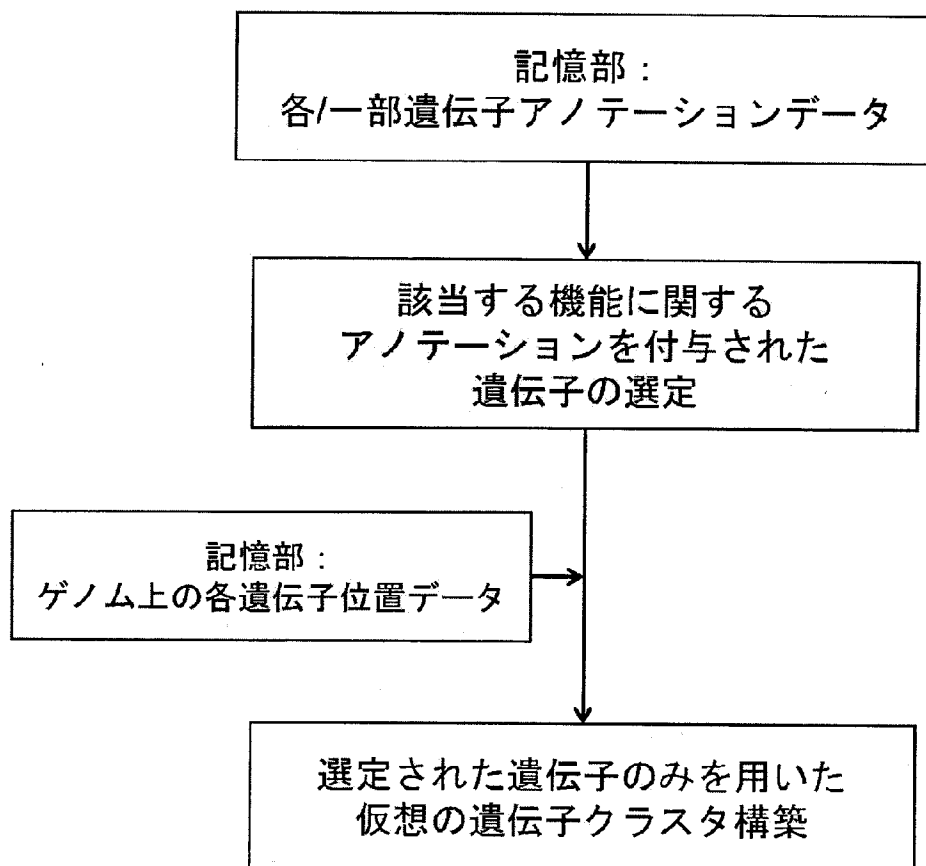
[図4]



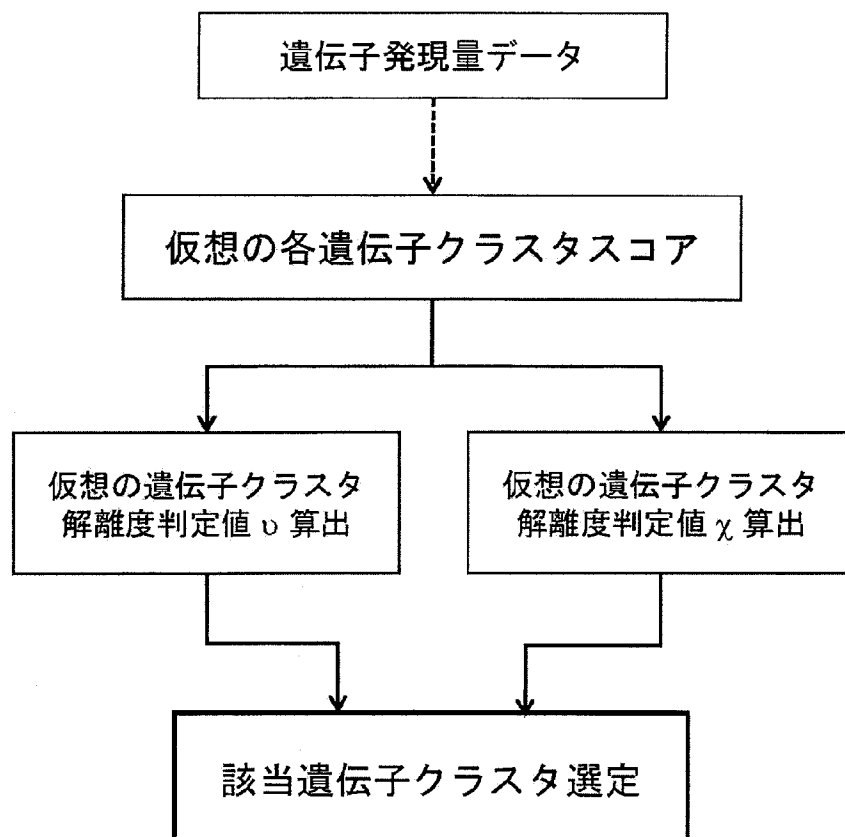
[図5]



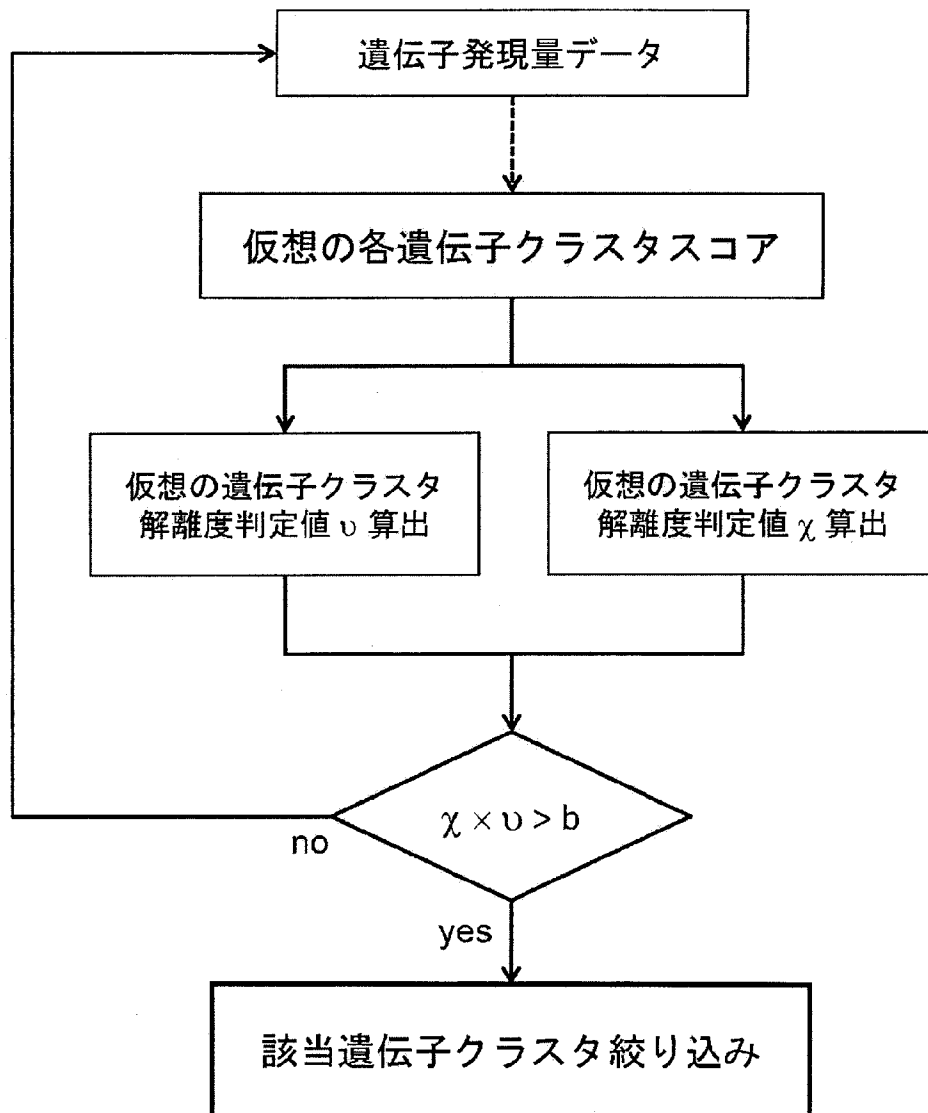
[図6]



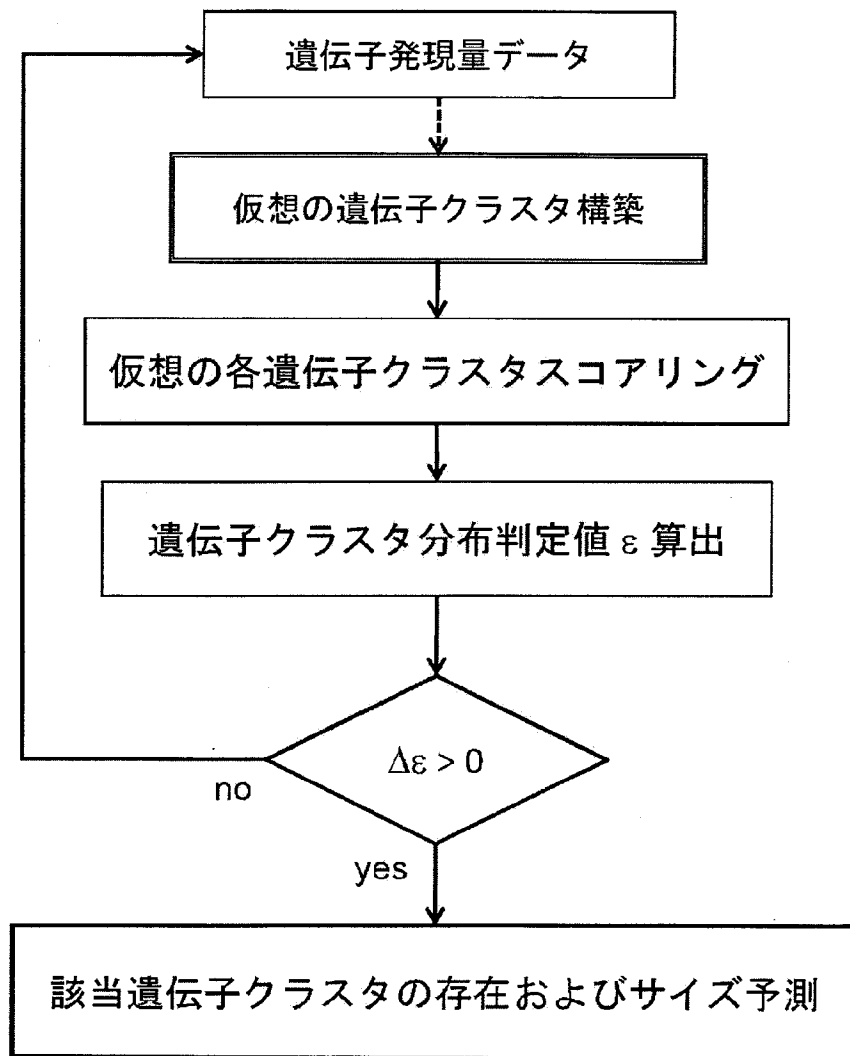
[図7]



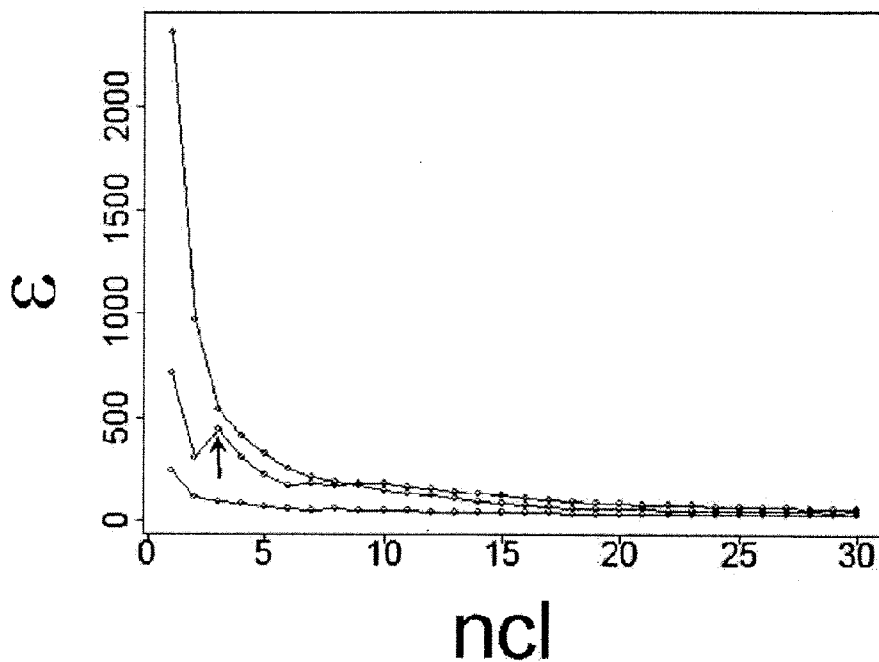
[図8]



[図9]

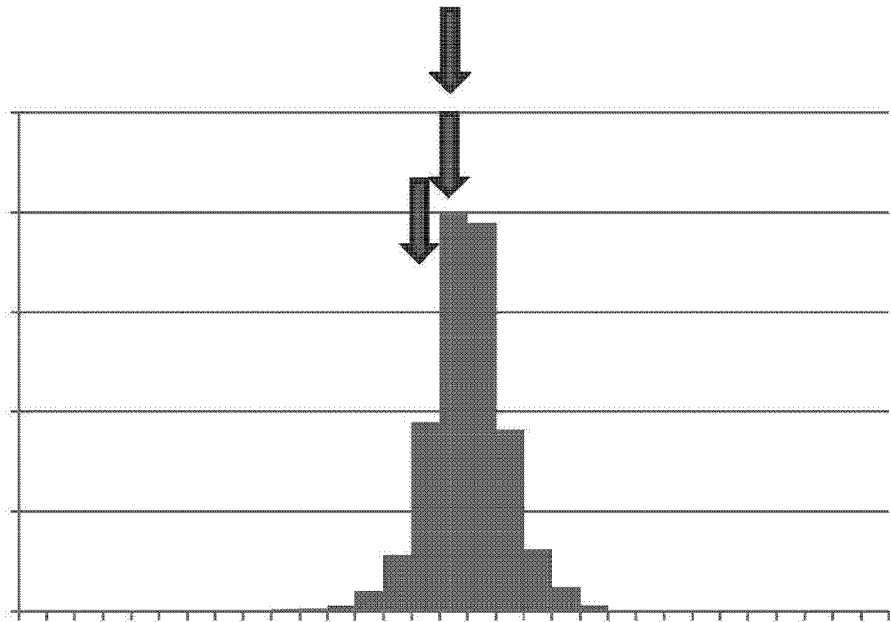


[図10]



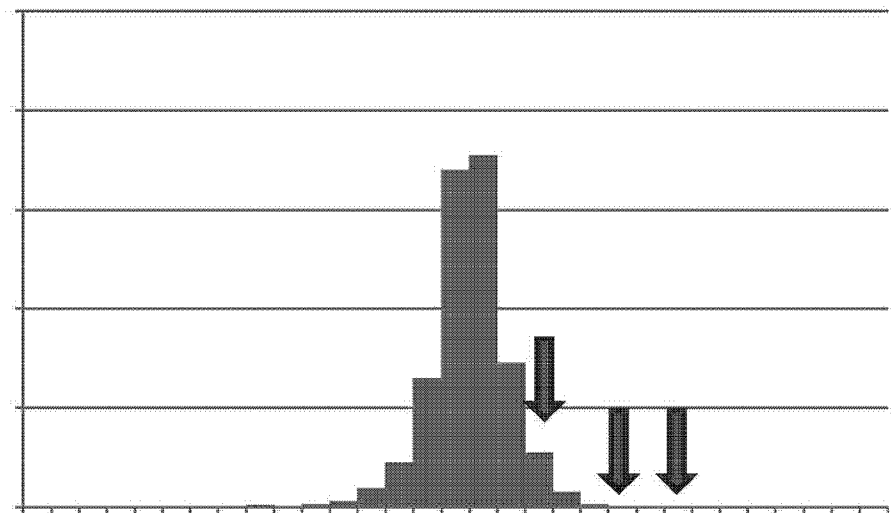
[図11]

データ区 間	頻度
-15	0
-14	0
-13	0
-12	0
-11	0
-10	0
-9	2
-8	0
-7	1
-6	9
-5	15
-4	29
-3	100
-2	281
-1	945
0	1990
1	1941
2	909
3	308
4	118
5	34
6	6
7	4
8	0
9	0
10	0
11	0
12	0
13	0
14	0
15	0
次の級	0

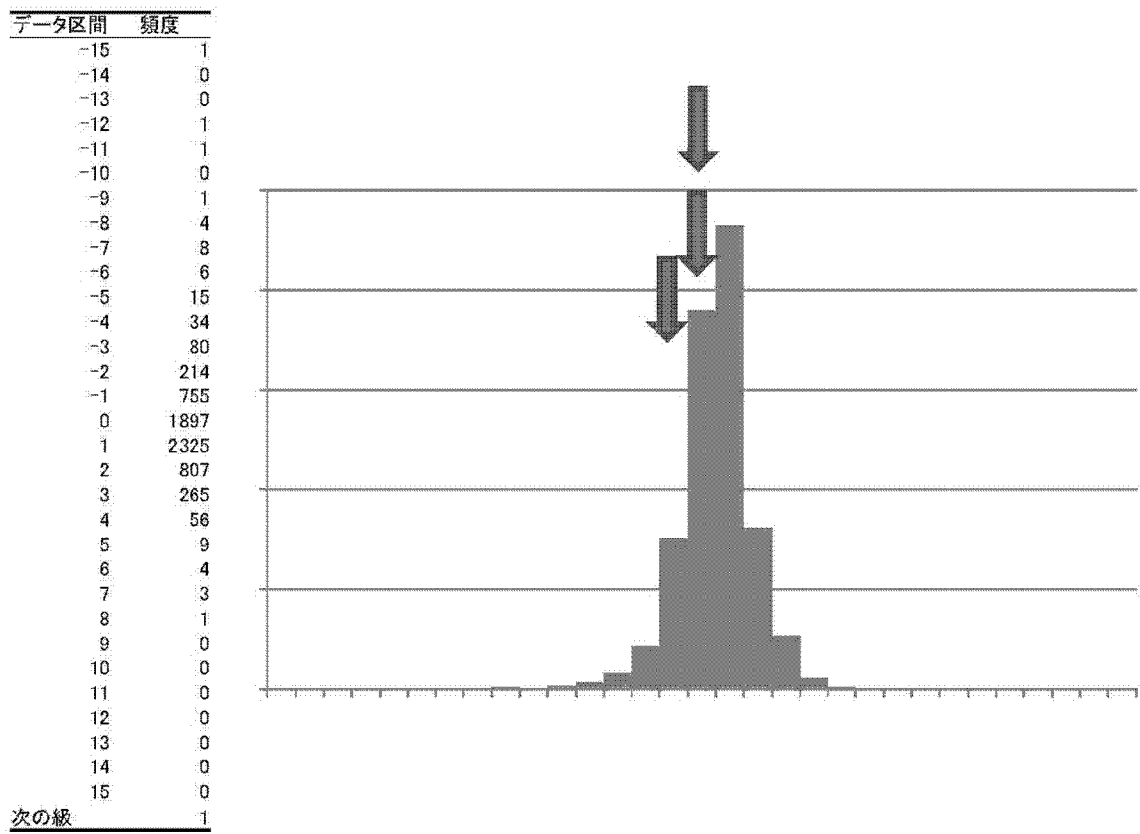


[図12]

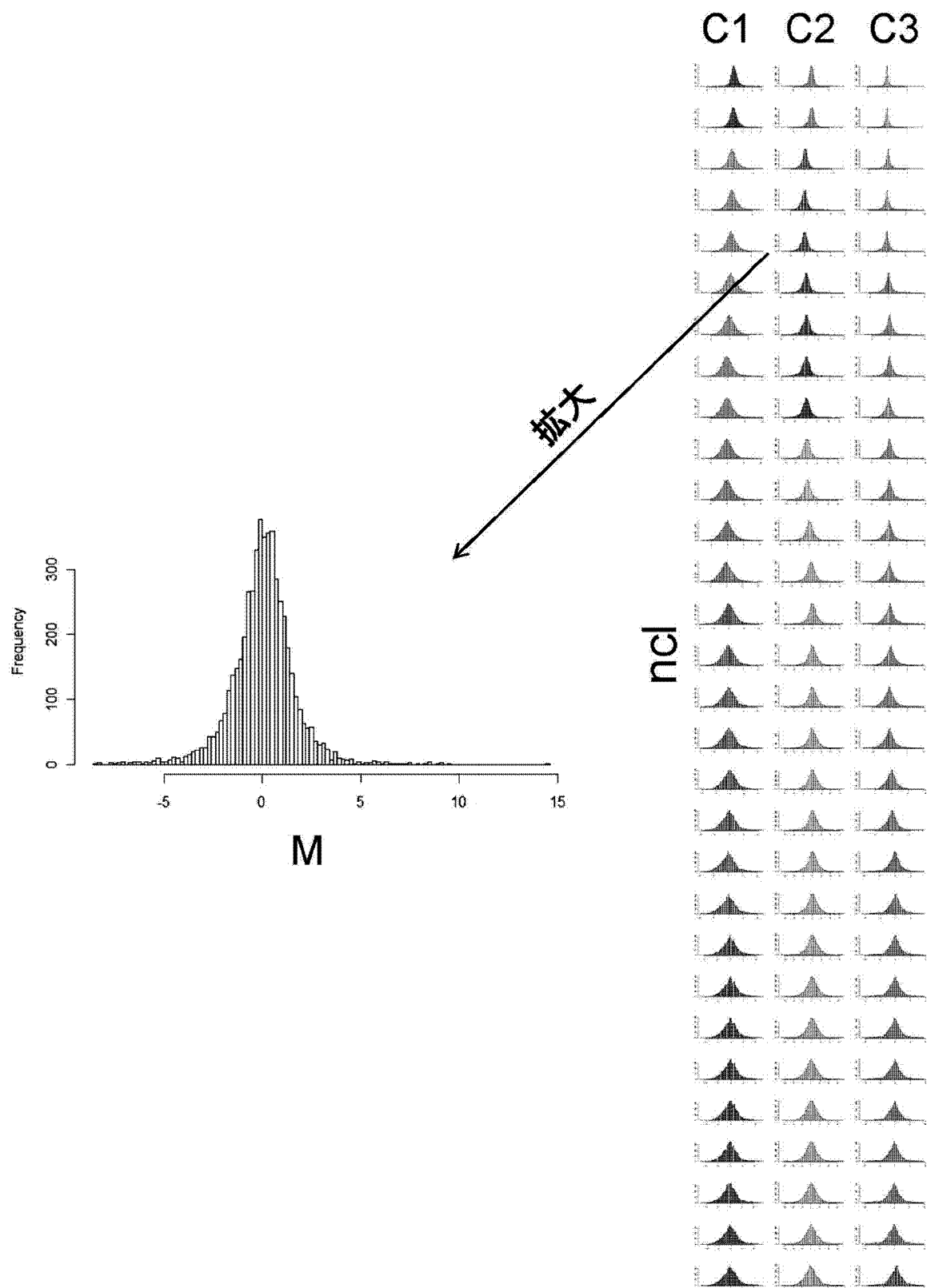
データ区 間	頻度
-15	0
-14	0
-13	0
-12	0
-11	0
-10	2
-9	2
-8	1
-7	9
-6	5
-5	13
-4	29
-3	93
-2	224
-1	650
0	1700
1	1770
2	729
3	273
4	75
5	14
6	1
7	1
8	2
9	2
10	0
11	0
12	0
13	0
14	0
15	0
次の級	0



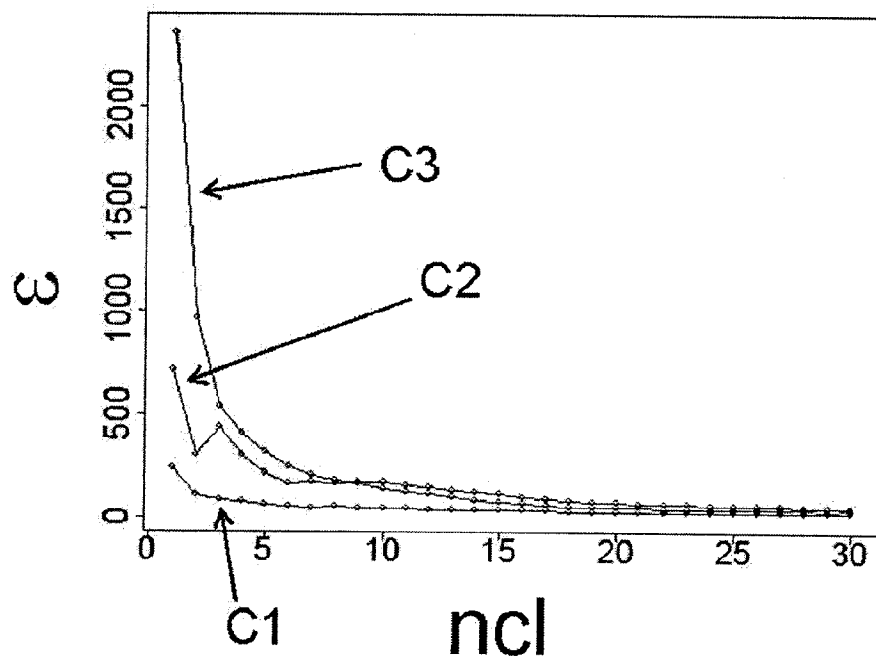
[図13]



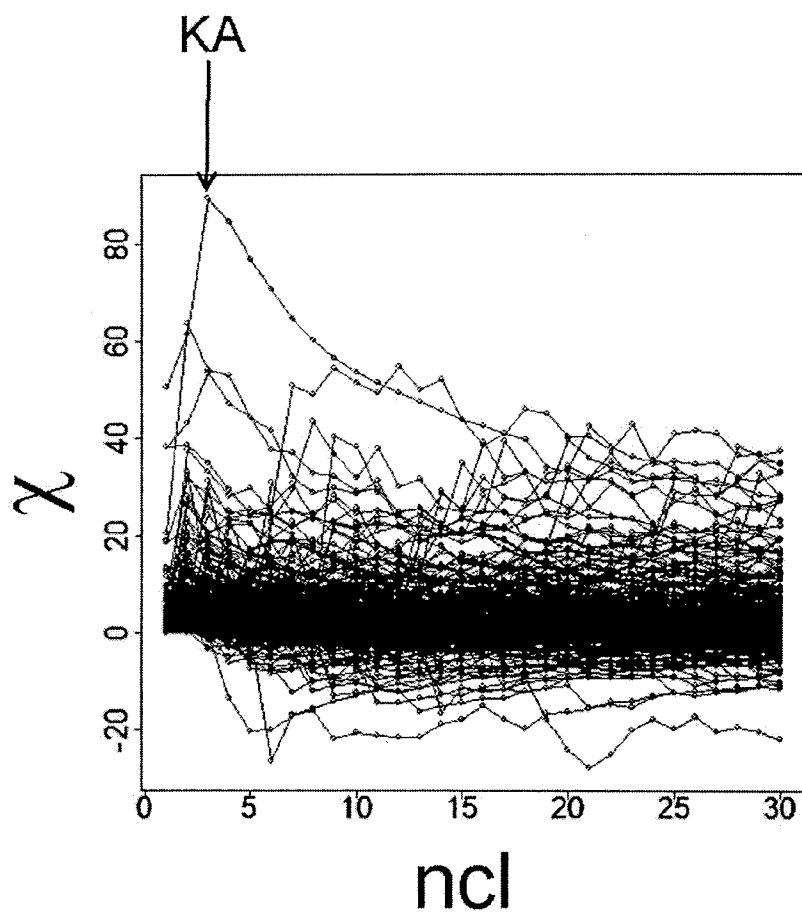
[図14]



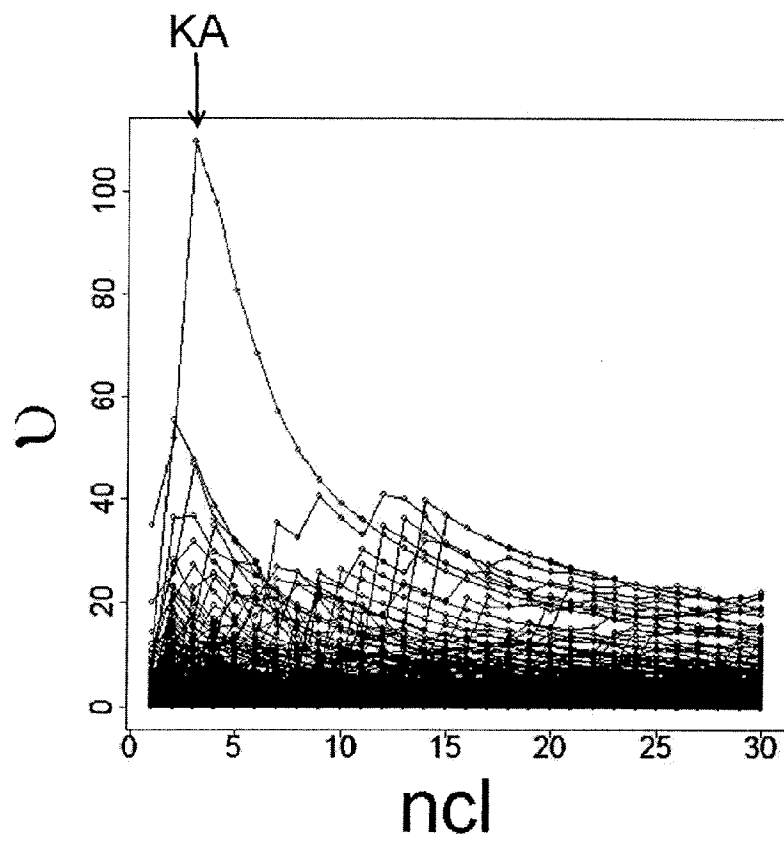
[図15]



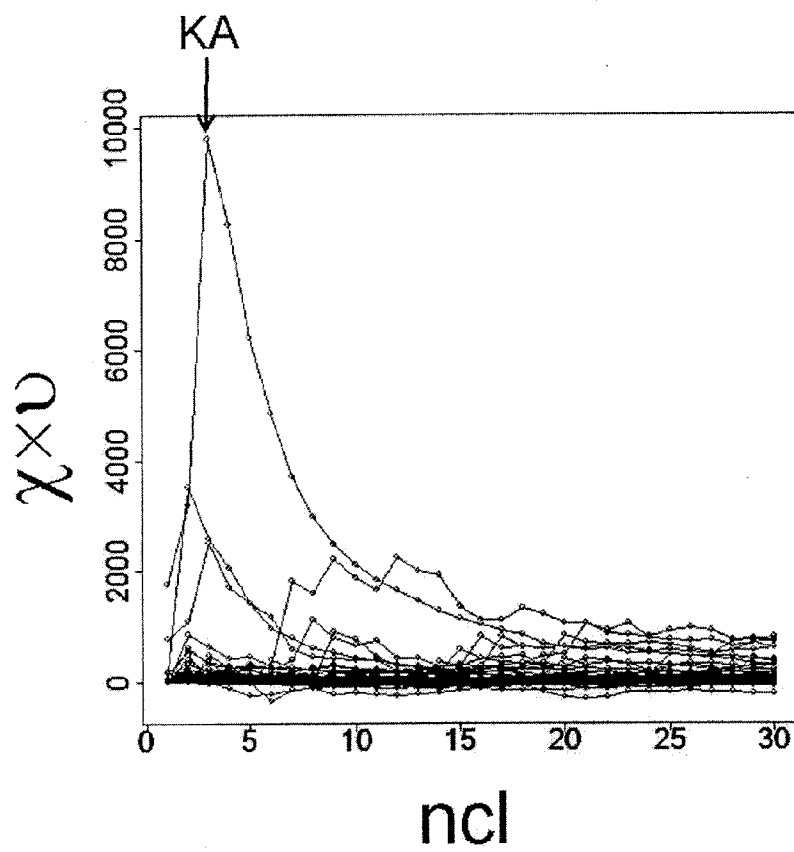
[図16]



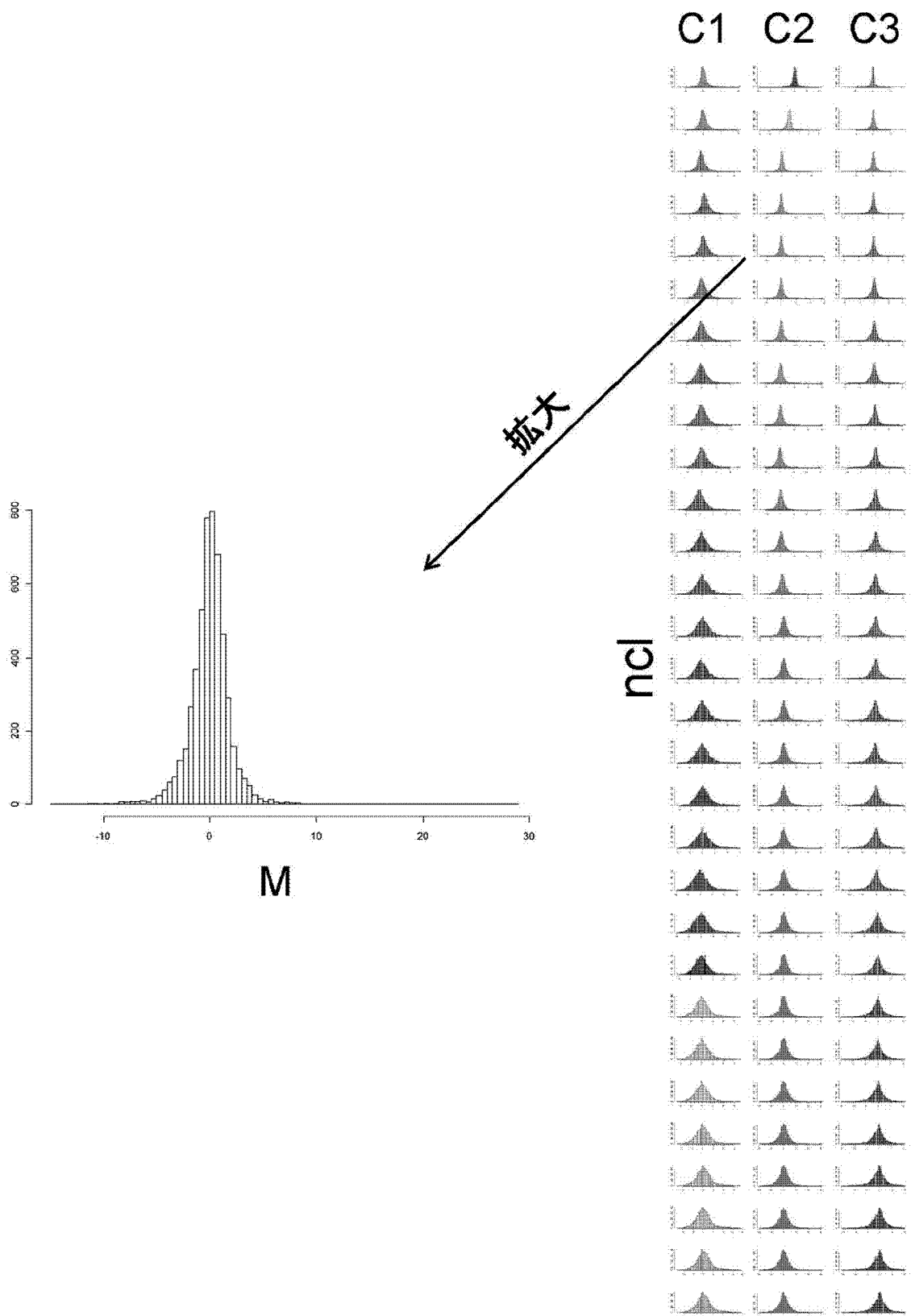
[図17]



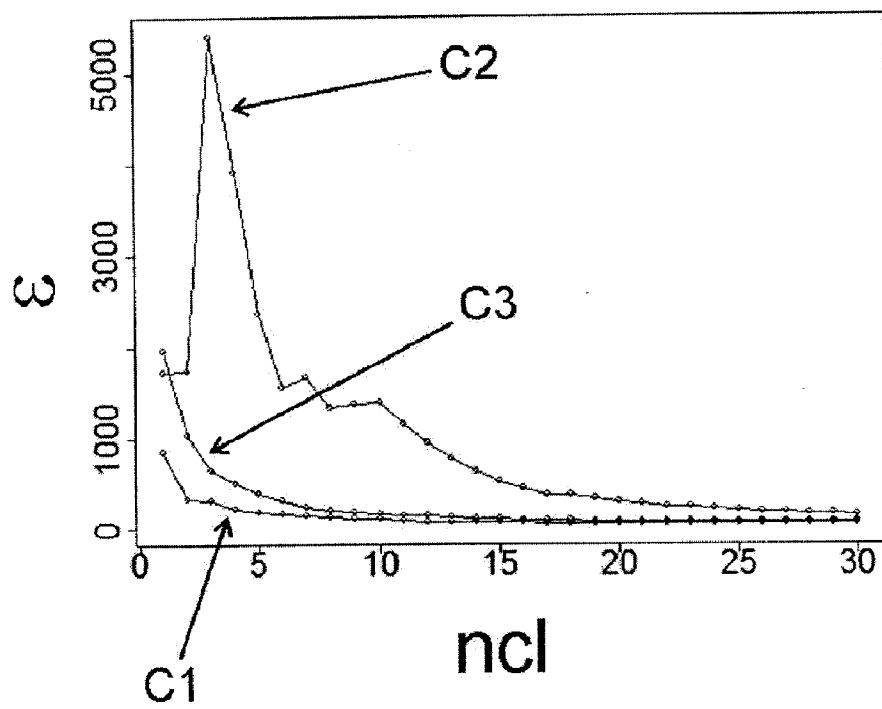
[図18]



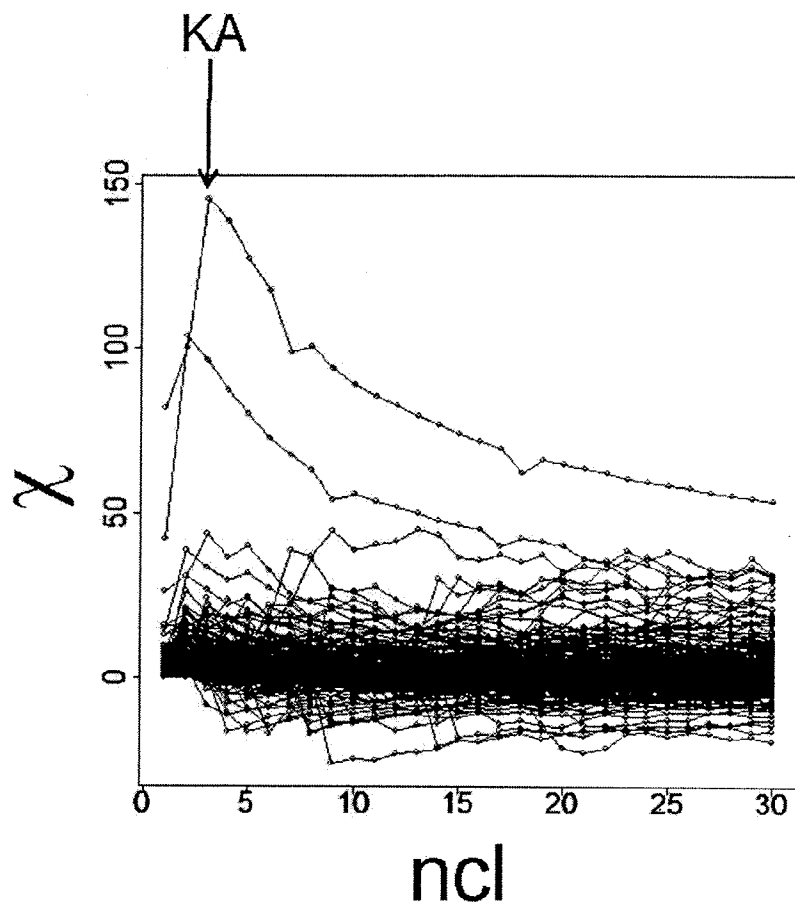
[図19]



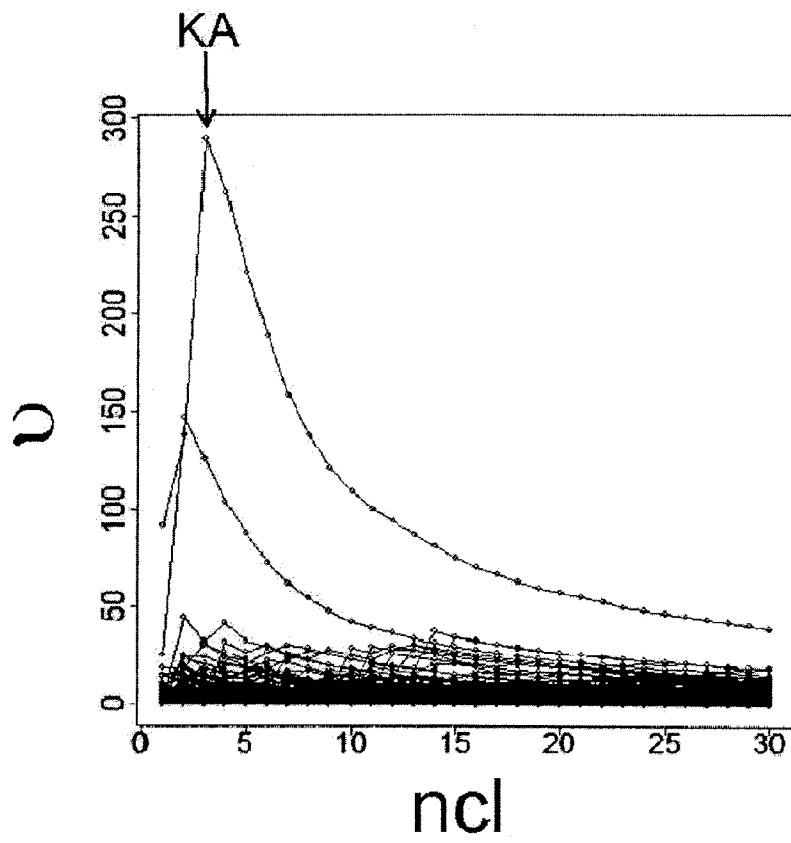
[図20]



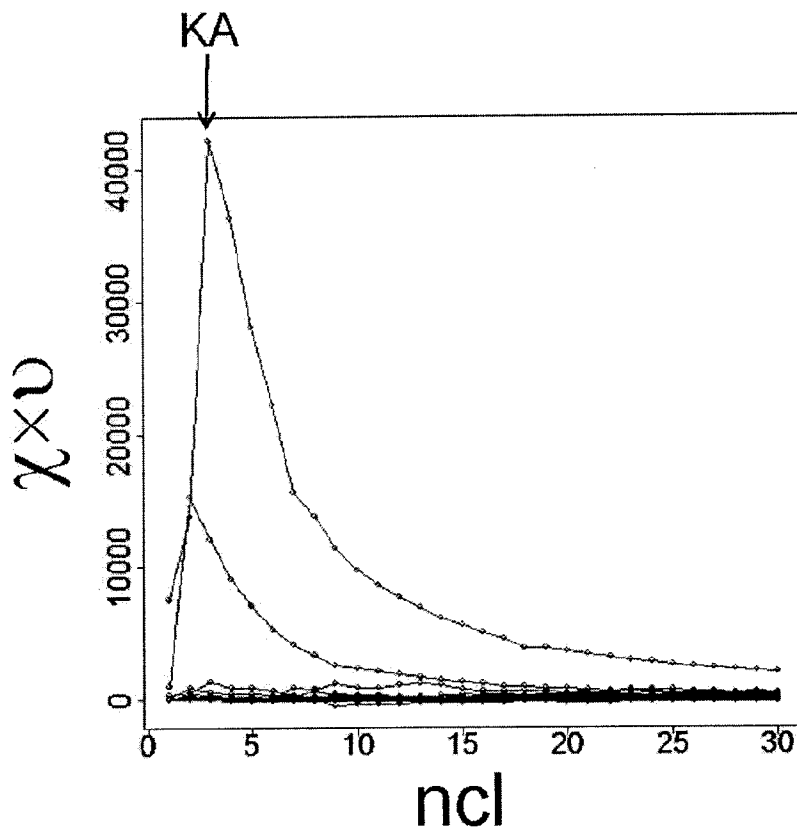
[図21]



[図22]

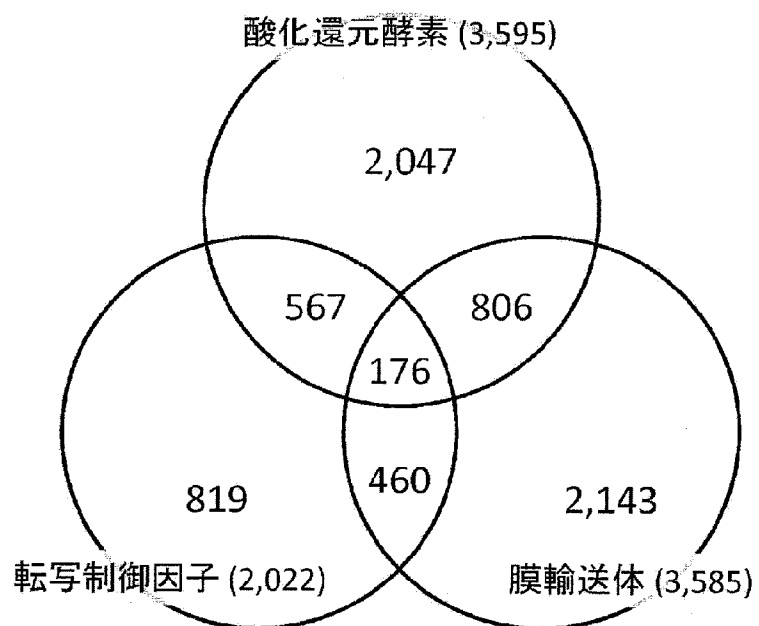


[図23]

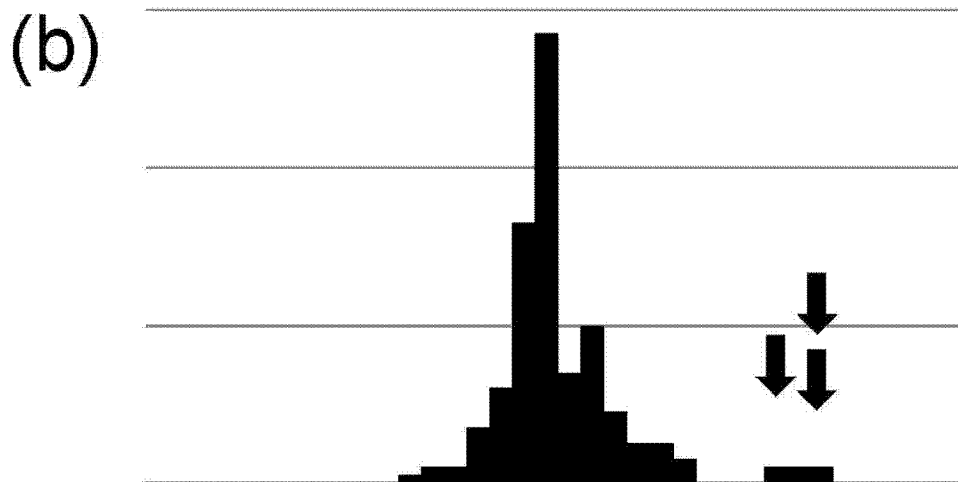
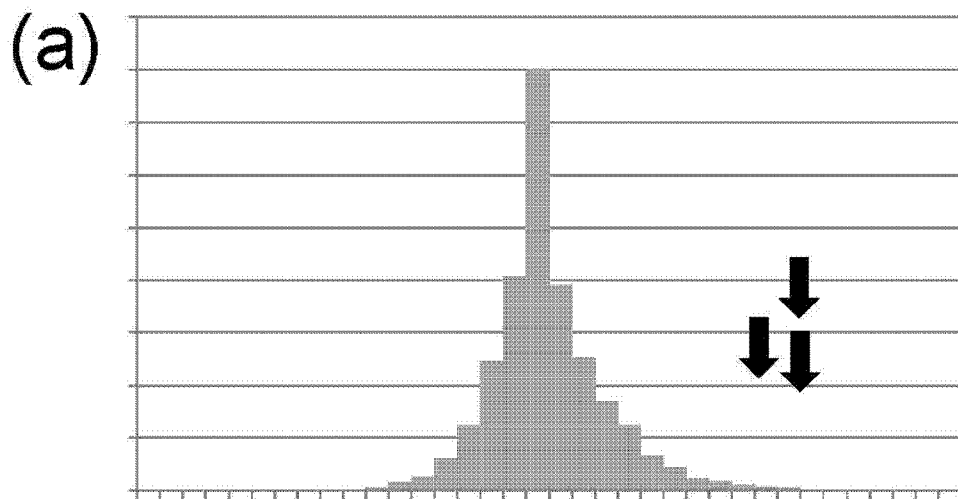


[図24]

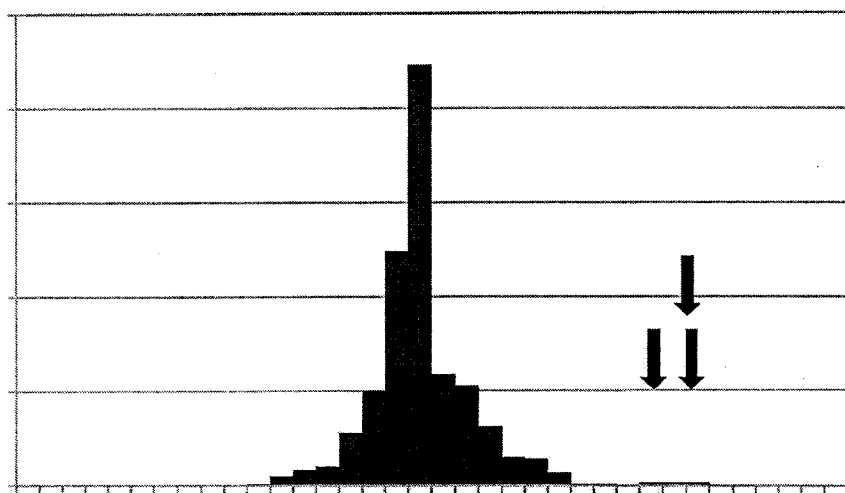
総数 14,032



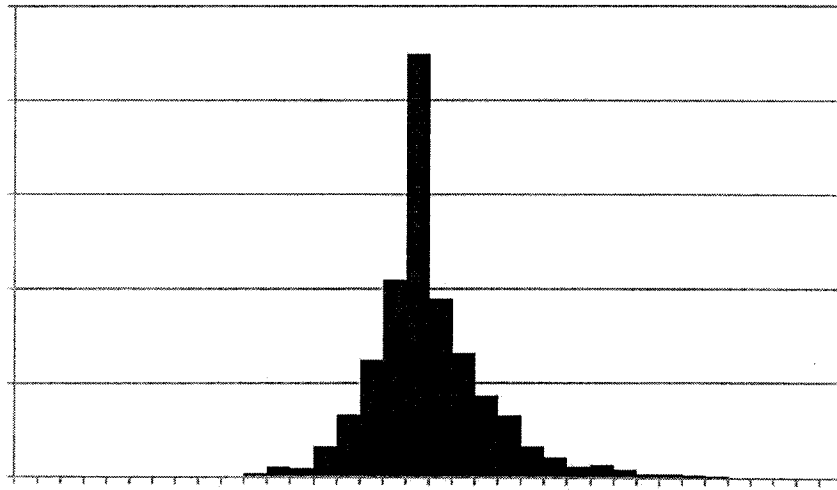
[図25]



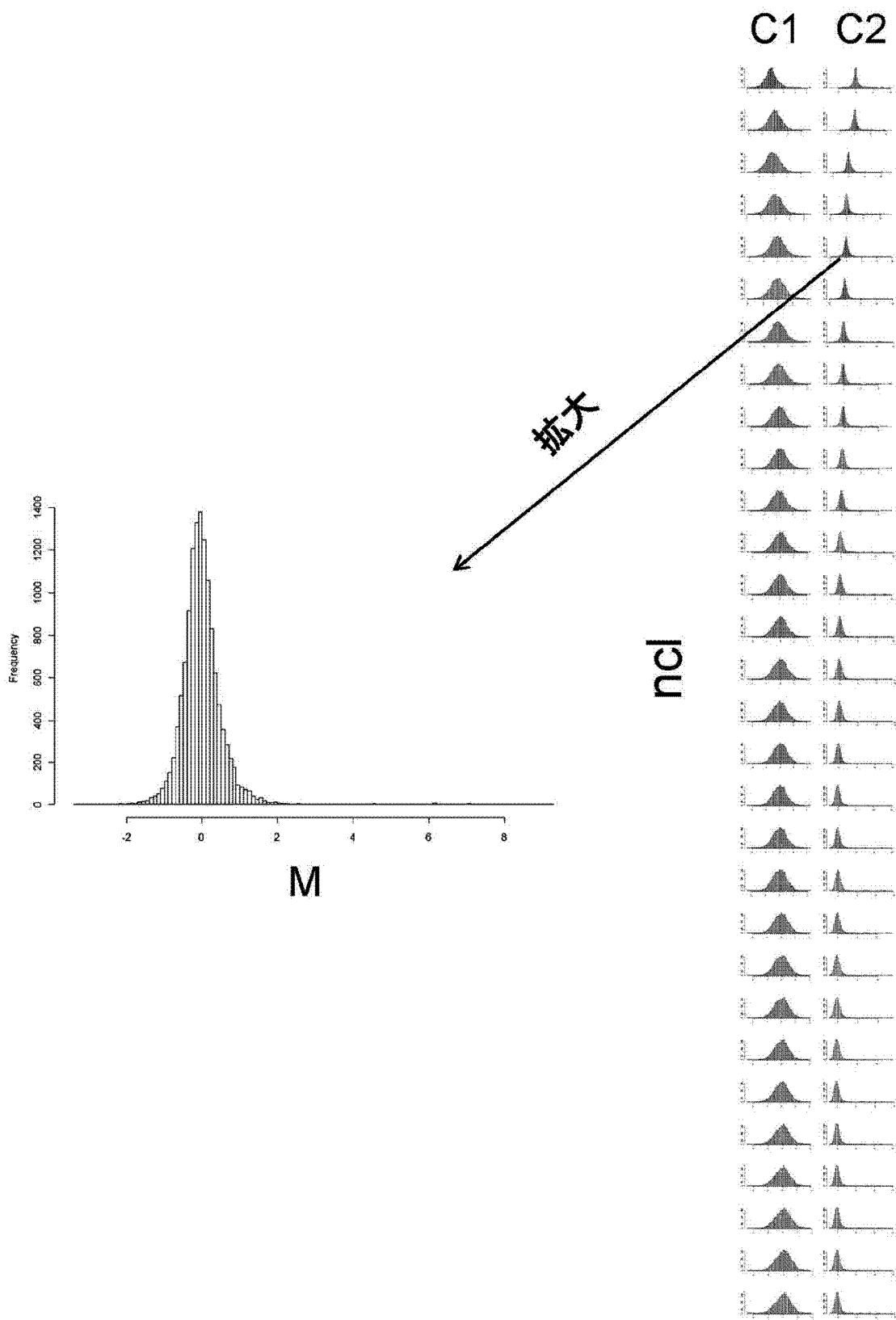
[図26]



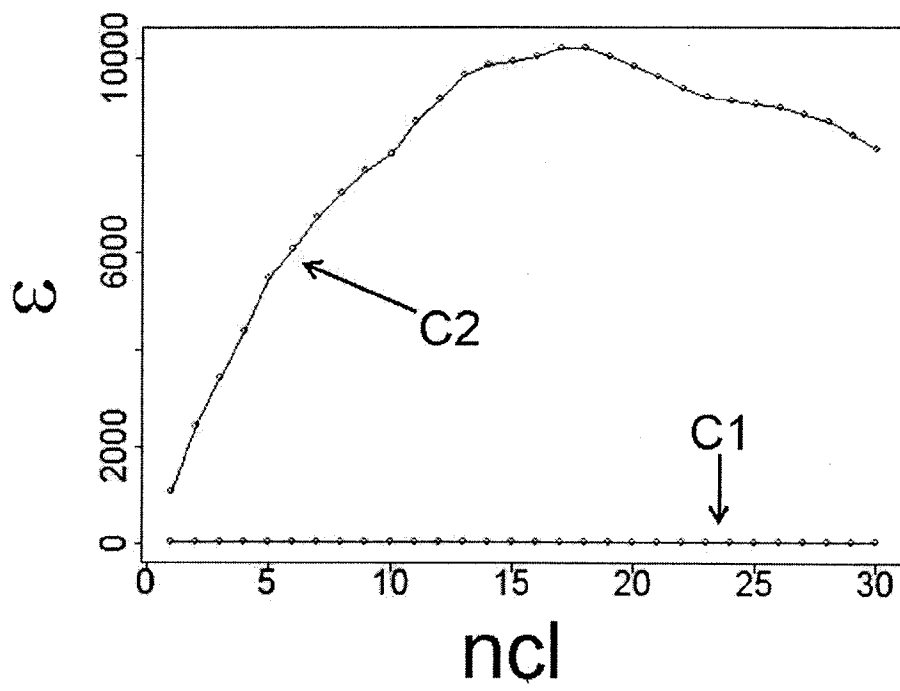
[図27]



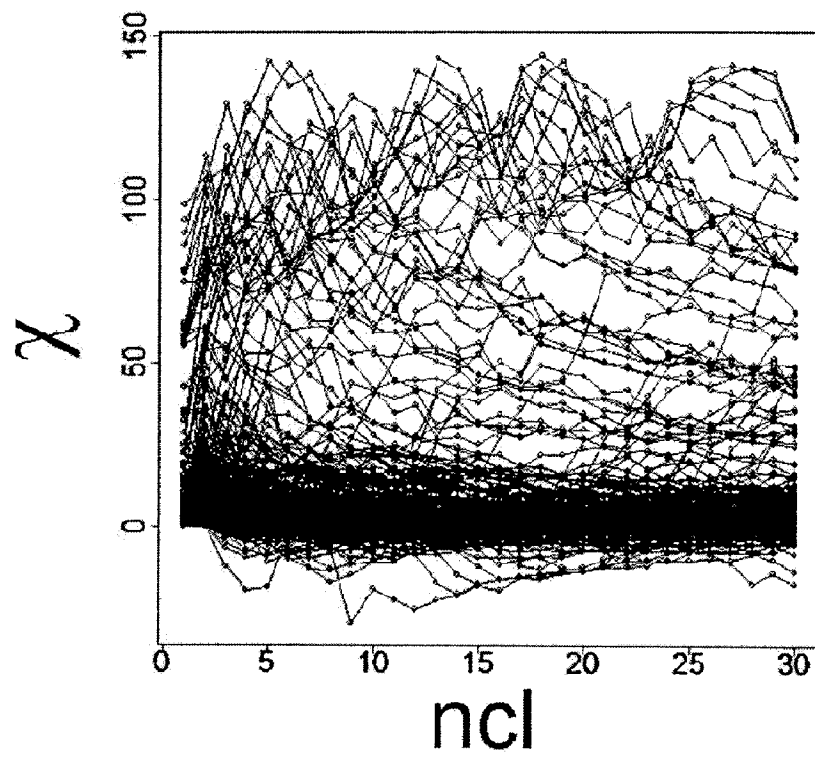
[図28]



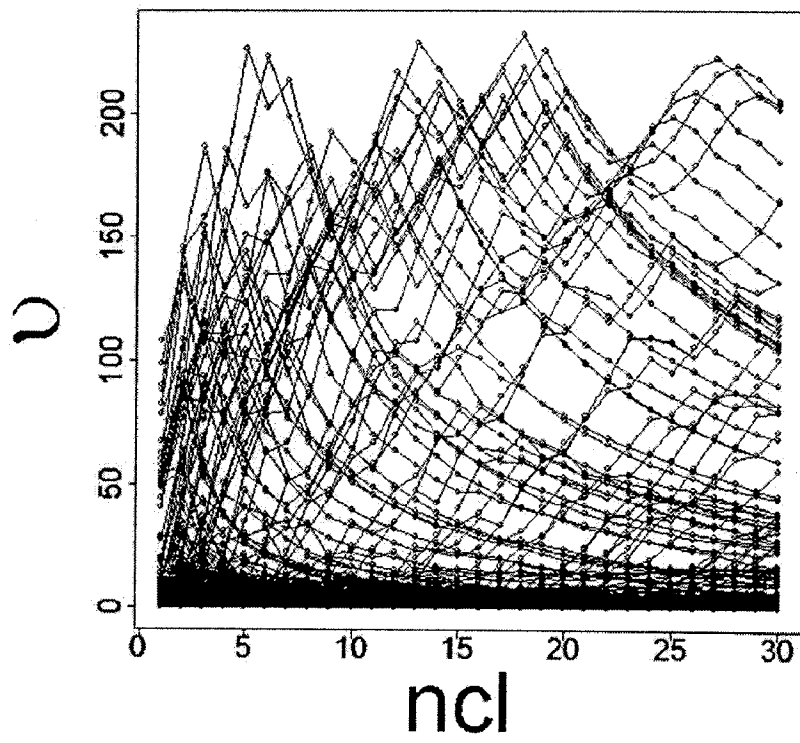
[図29]



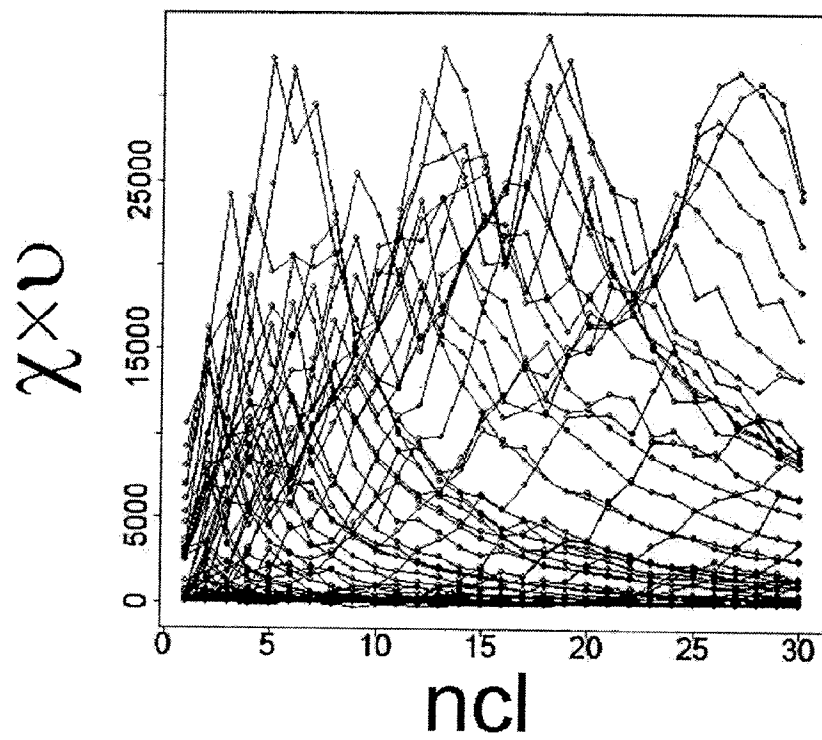
[図30]



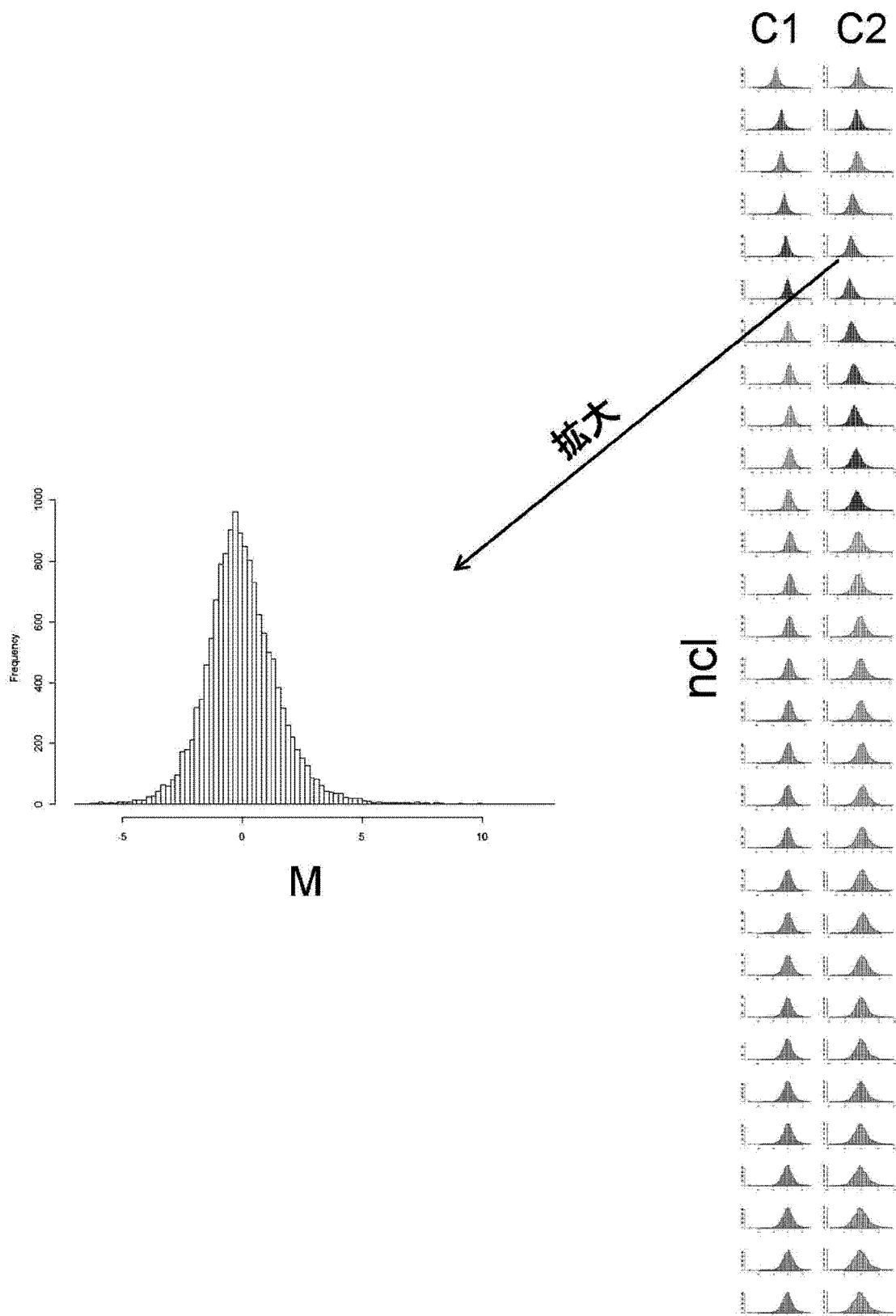
[図31]



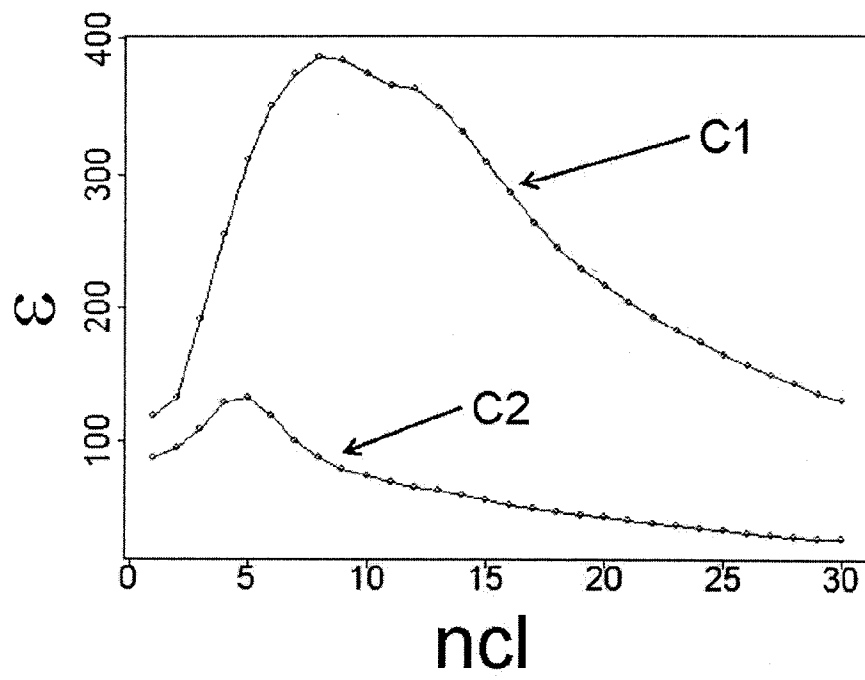
[図32]



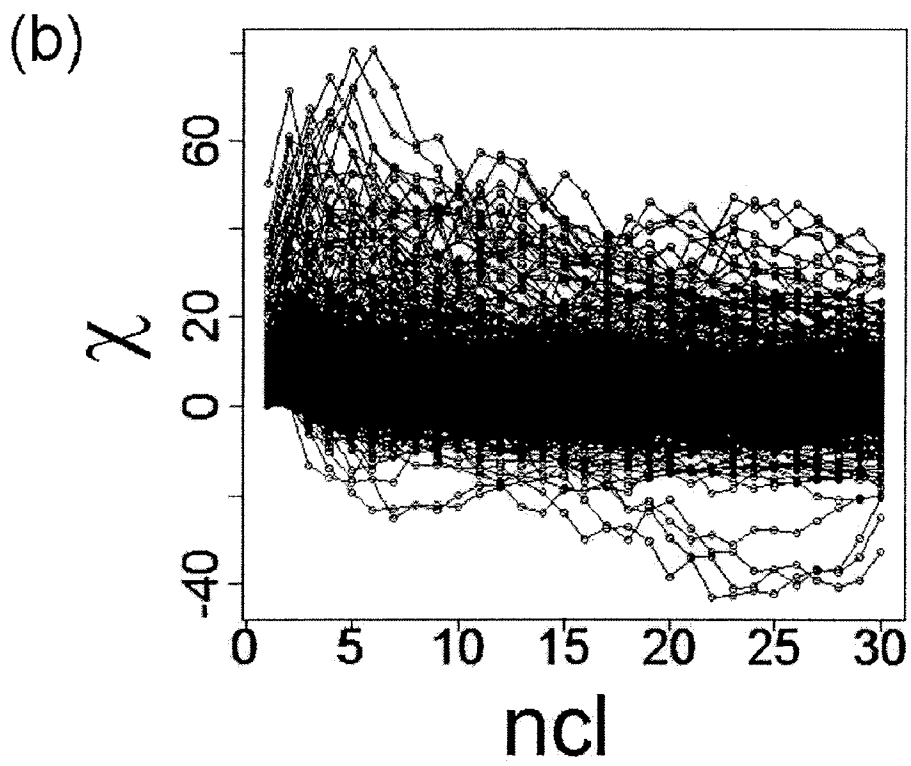
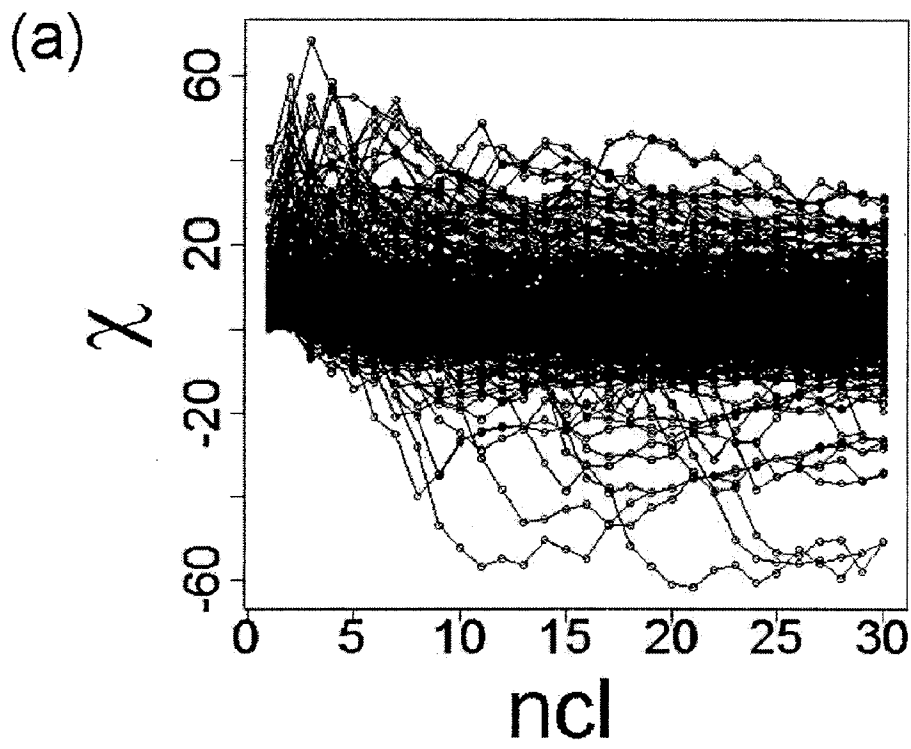
[図33]



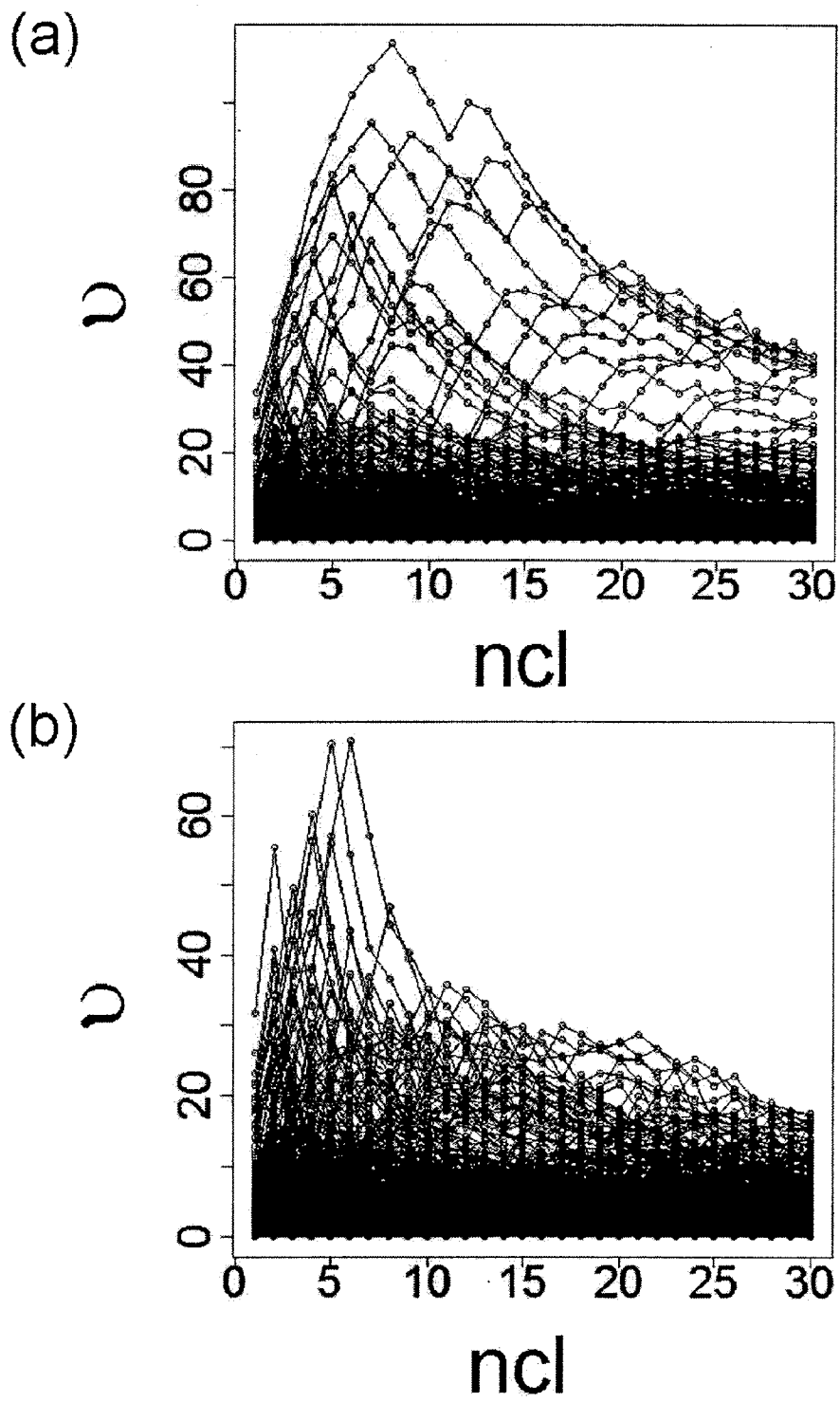
[図34]



[図35]

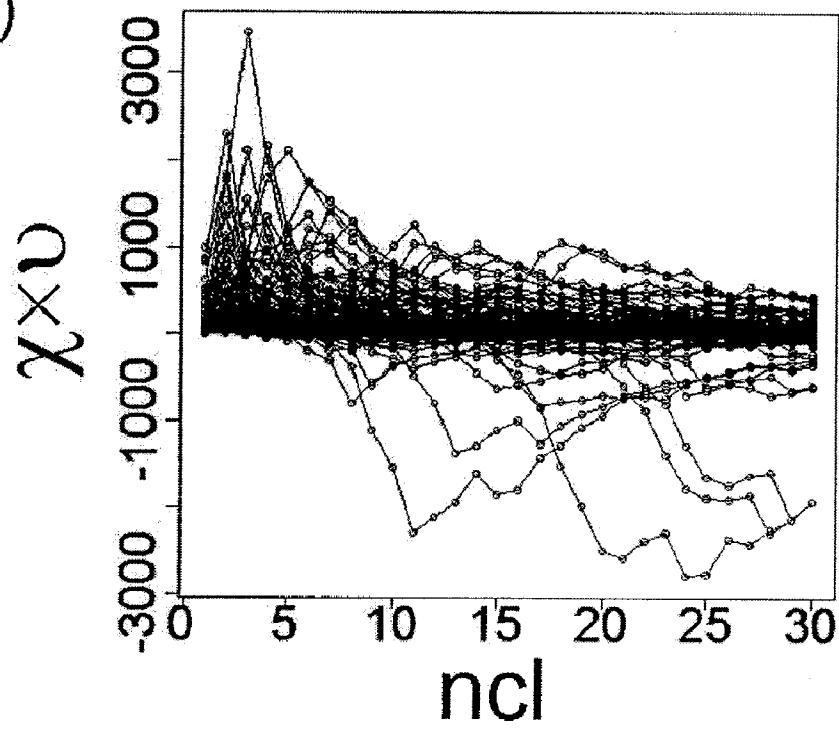


[図36]

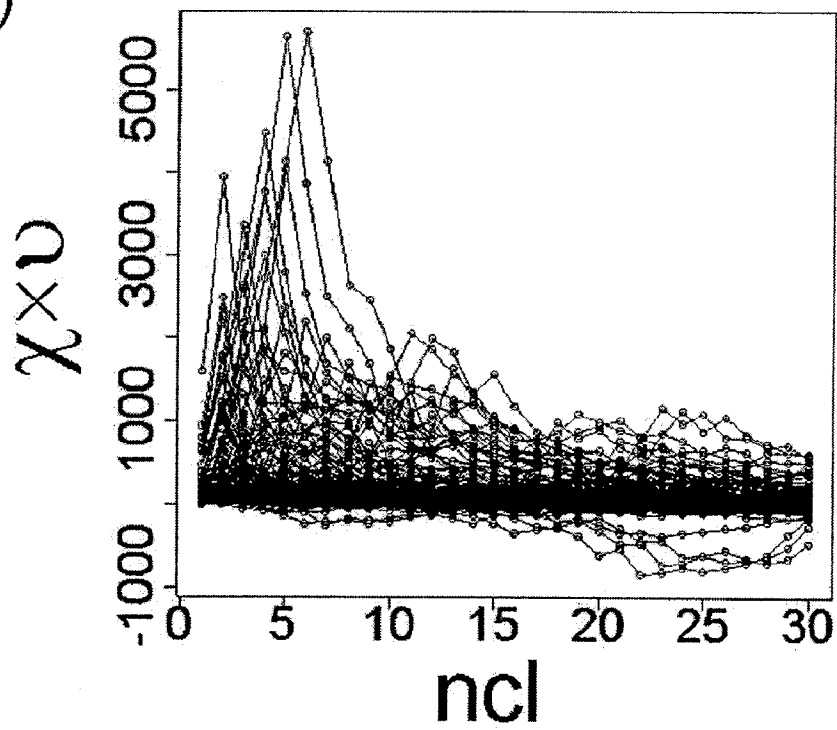


[図37]

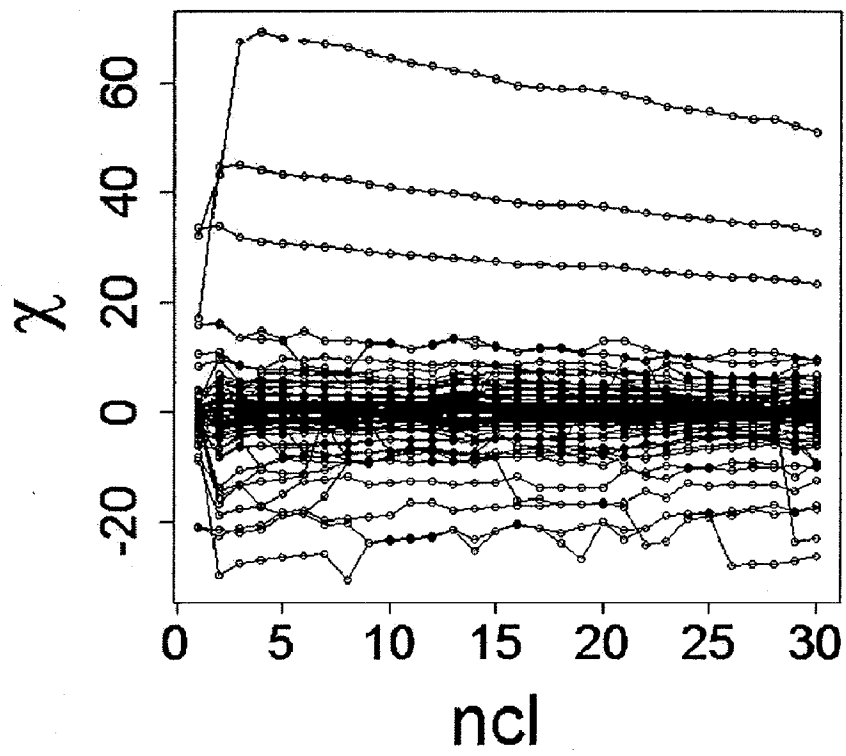
(a)



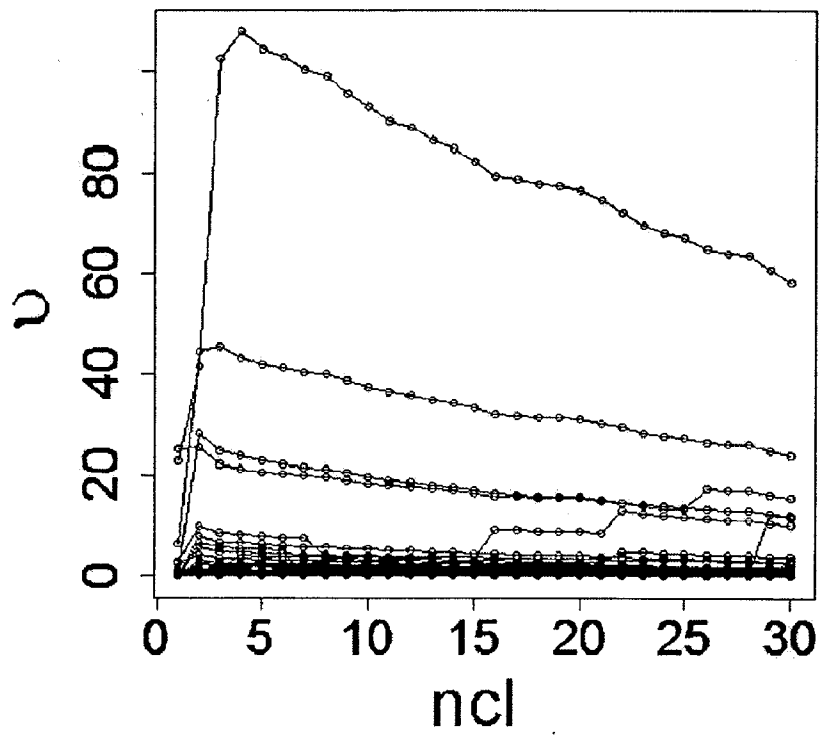
(b)



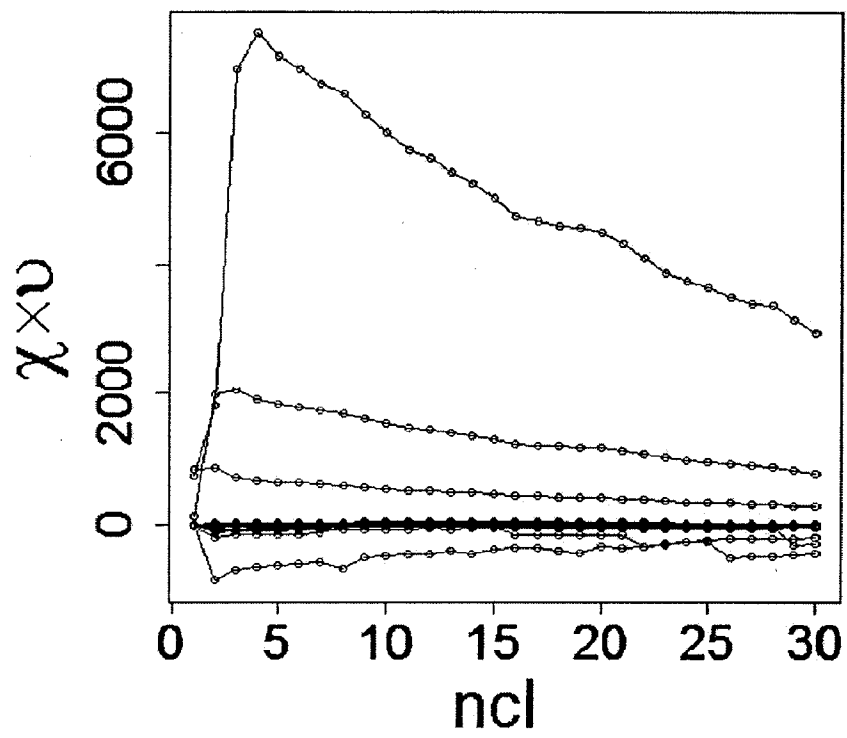
[図38]



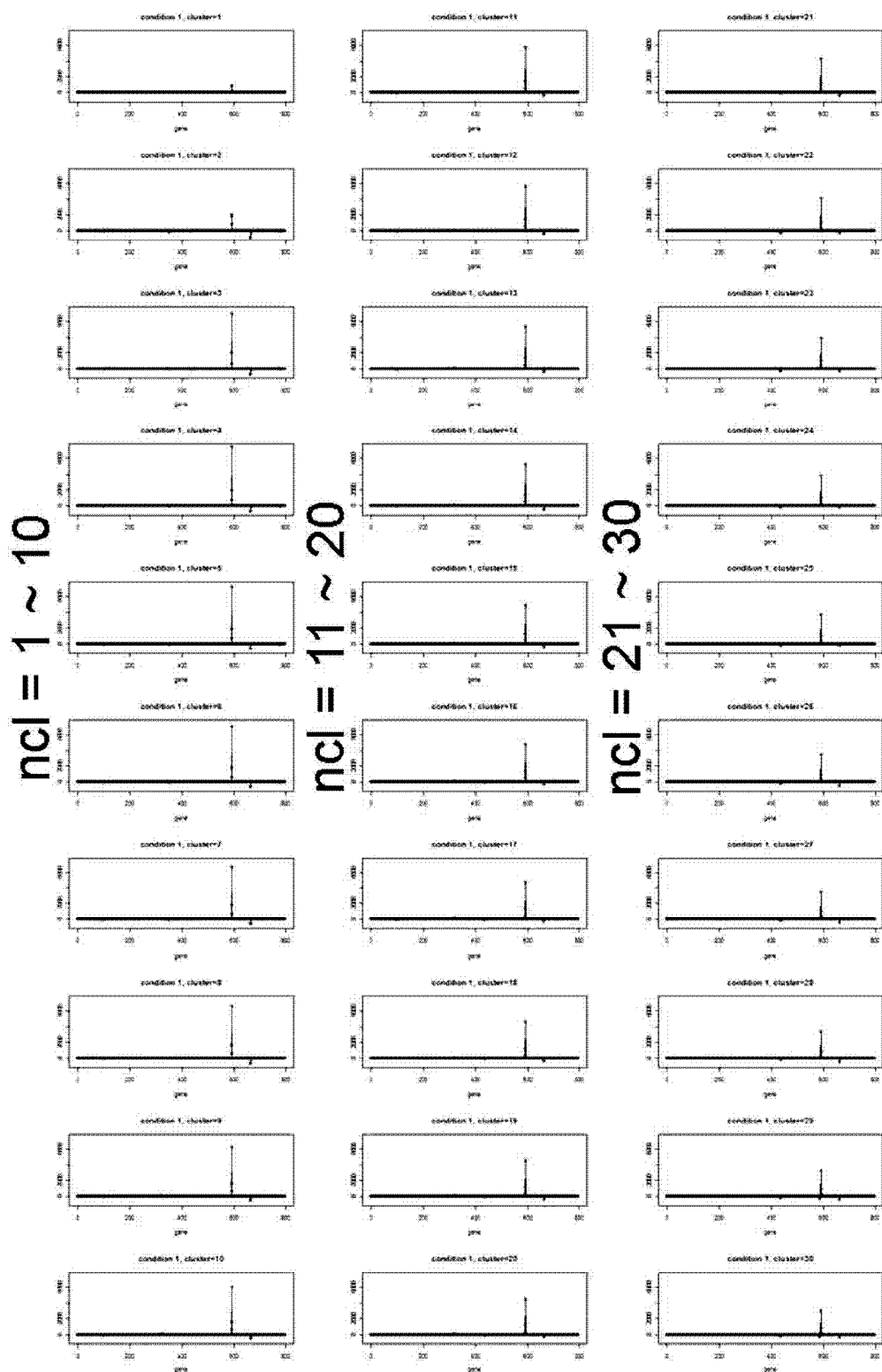
[図39]



[図40]

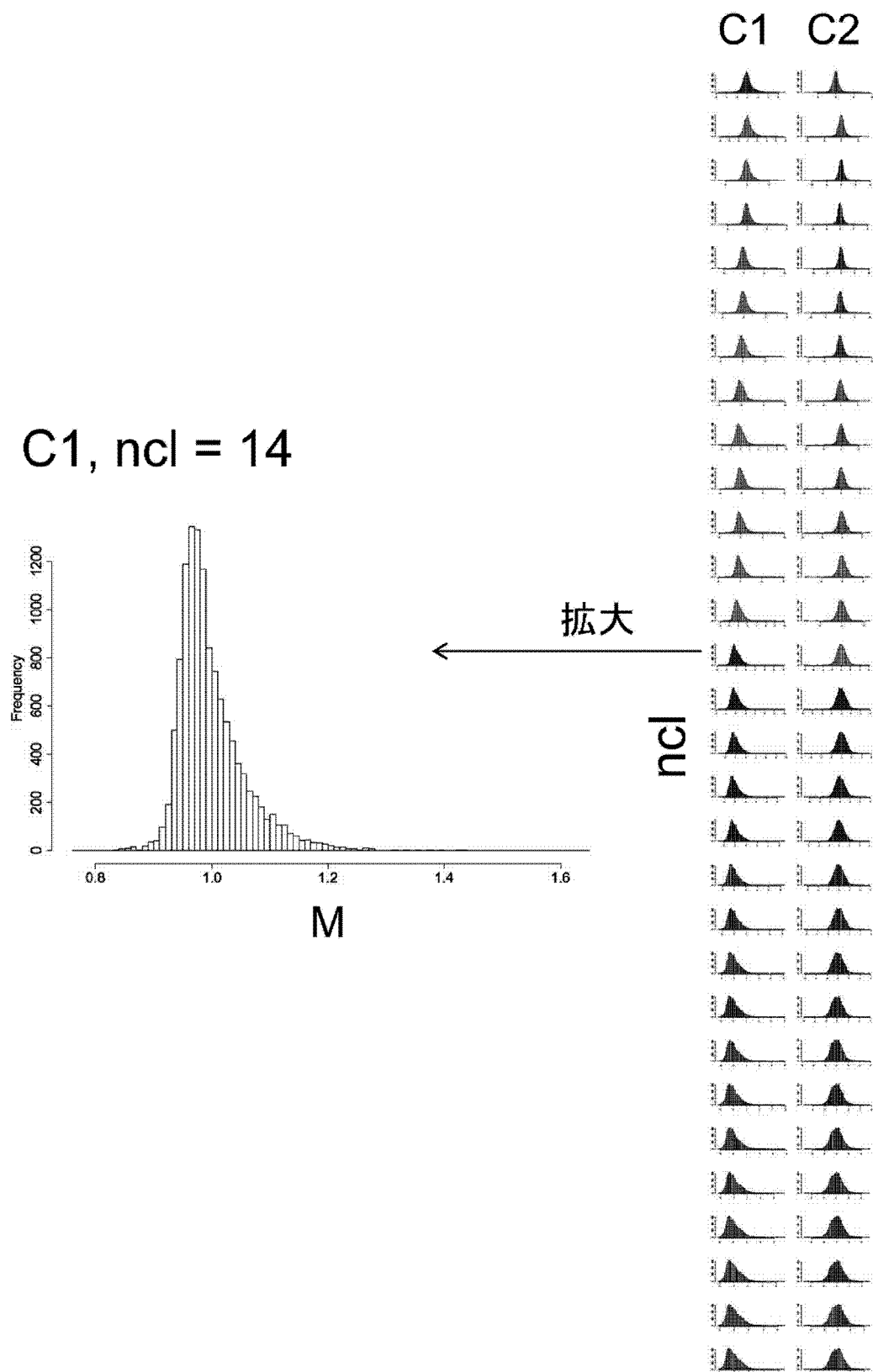


[図41]

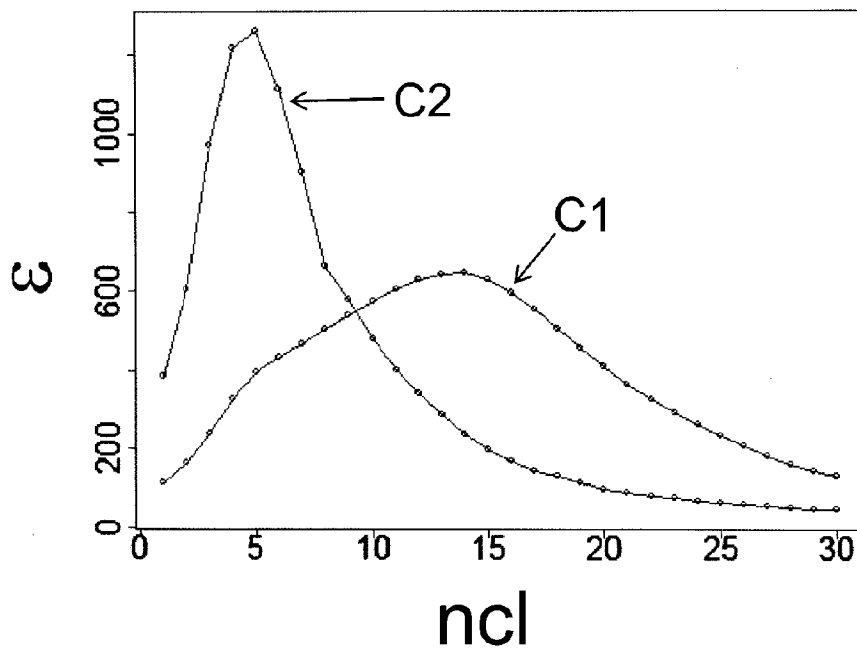


仮定の遺伝子クラスター

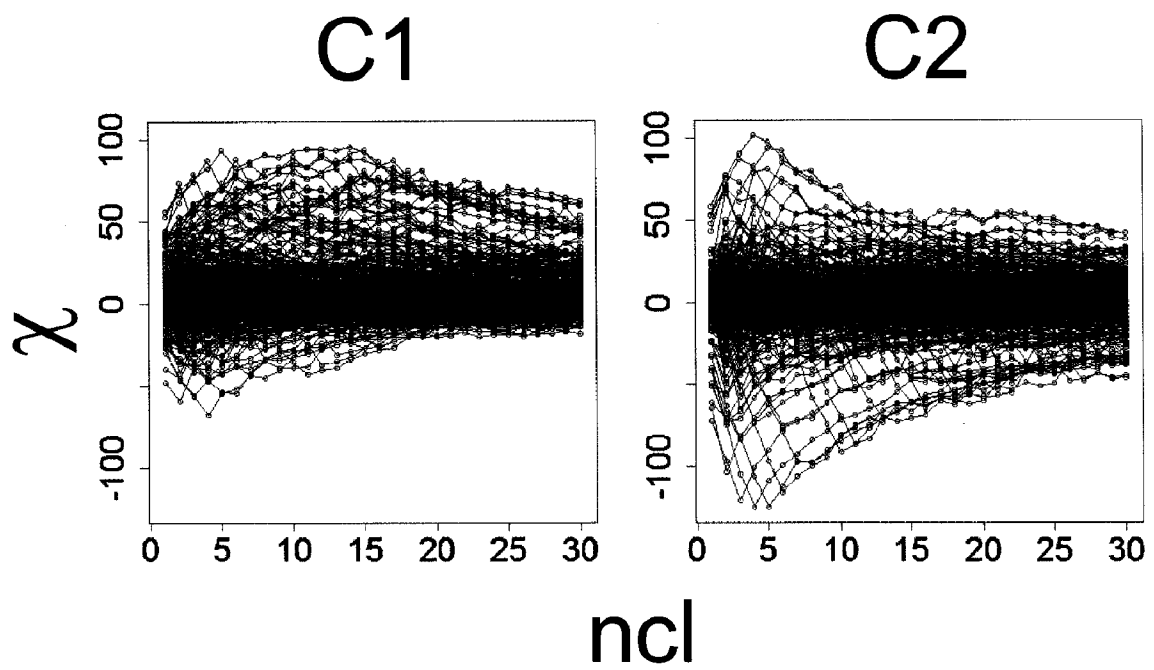
[図42]



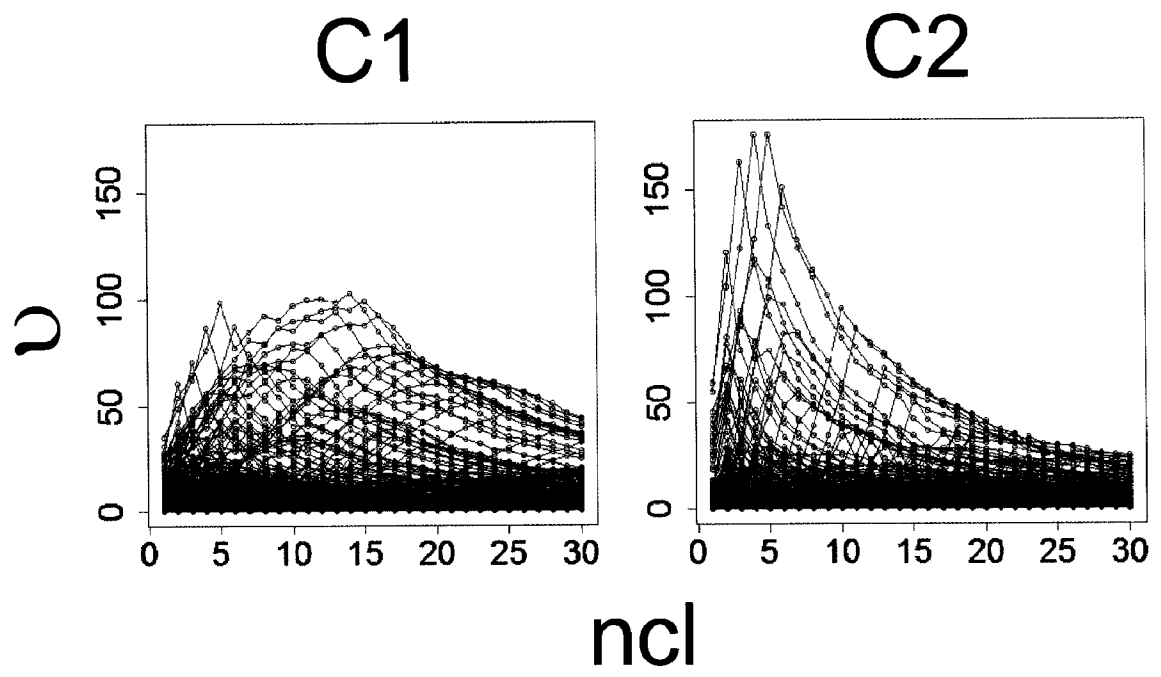
[図43]



[図44]

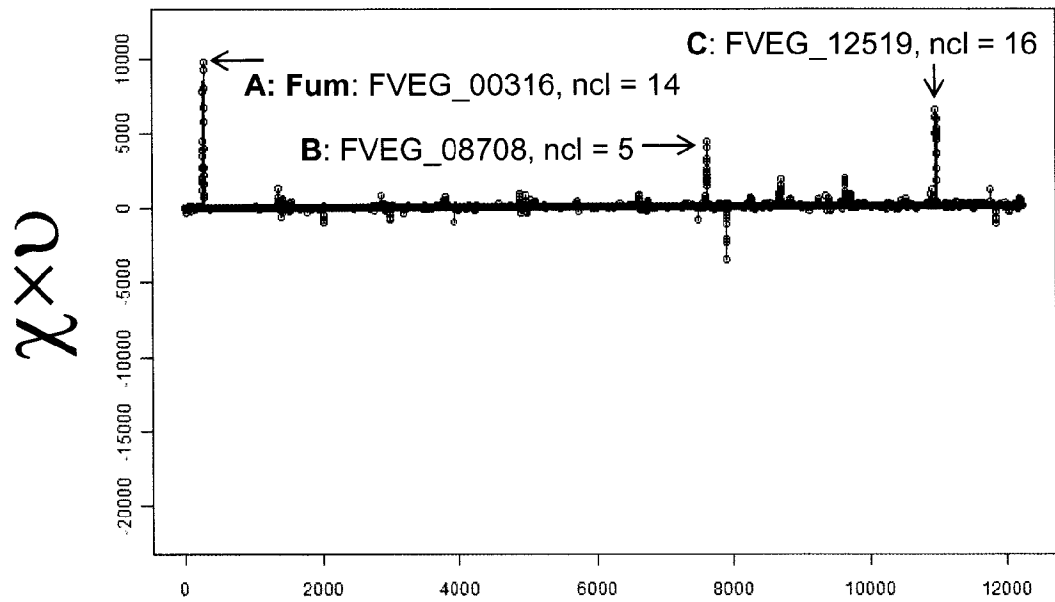


[図45]

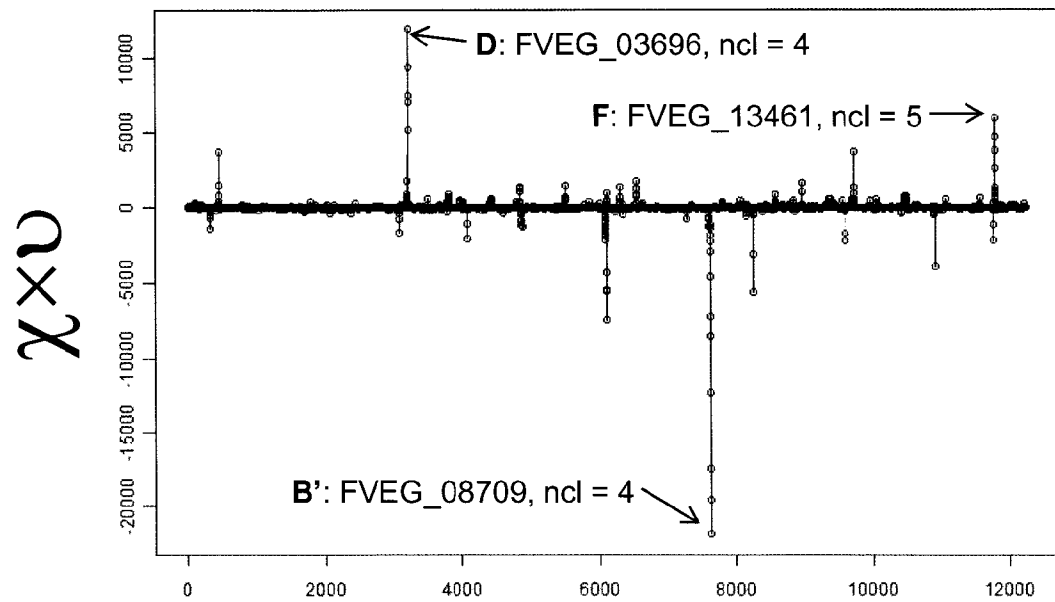


[図46]

C1

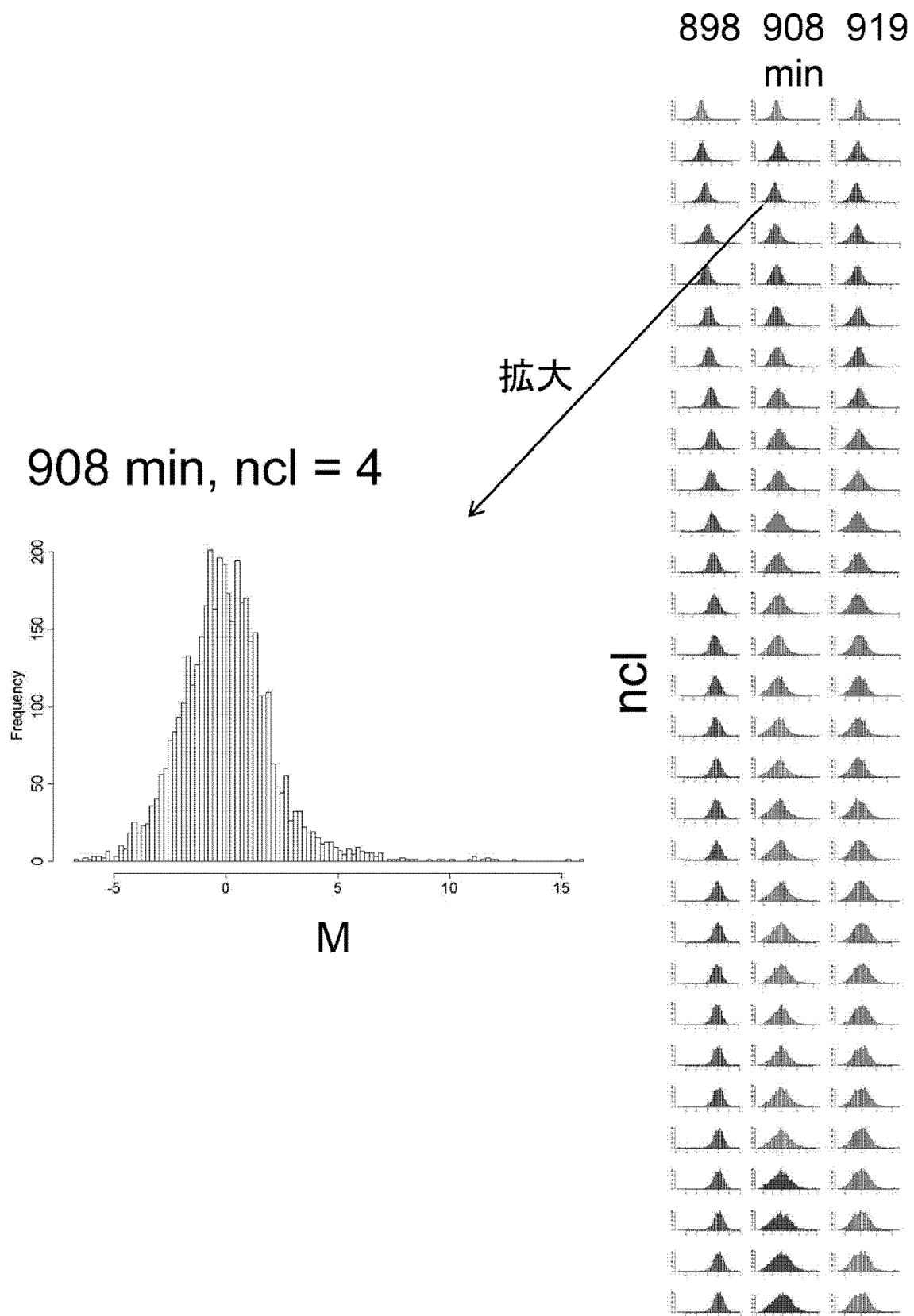


C2

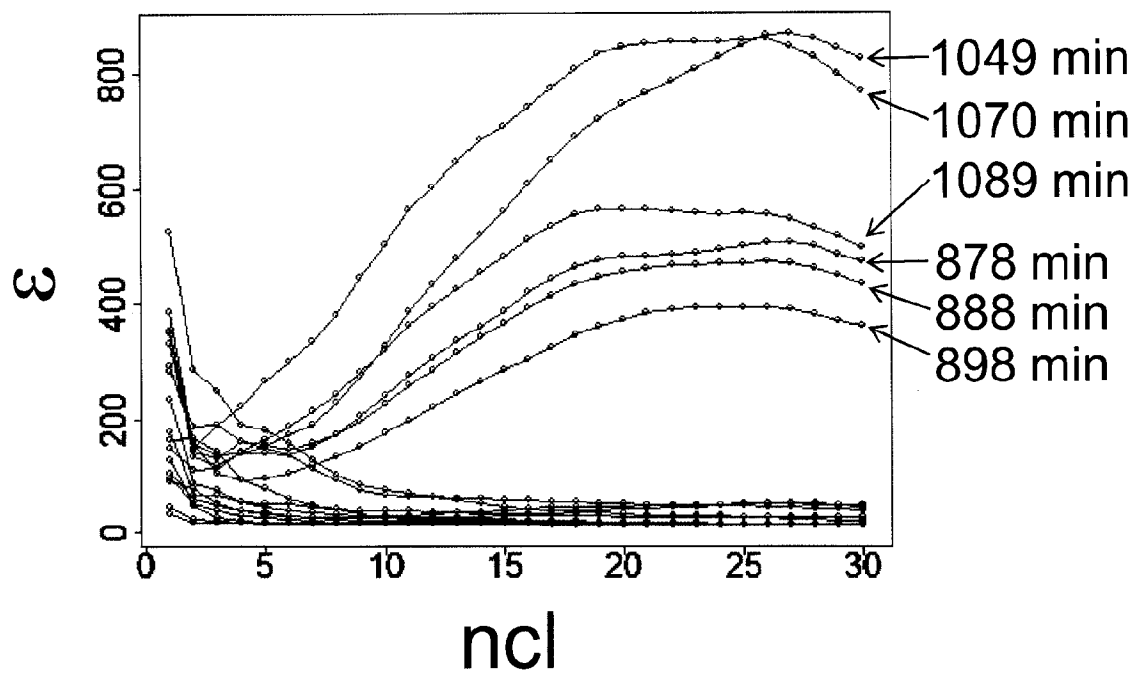


仮想の遺伝子クラスター起点遺伝子ID

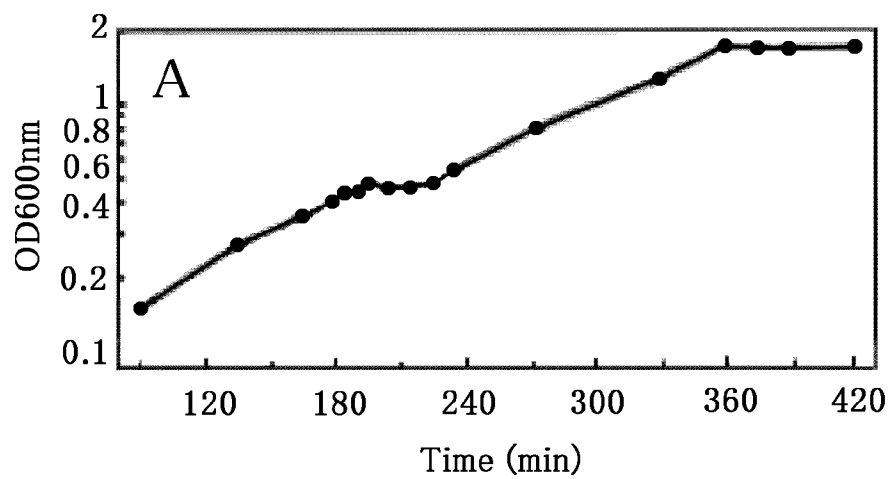
[図47]



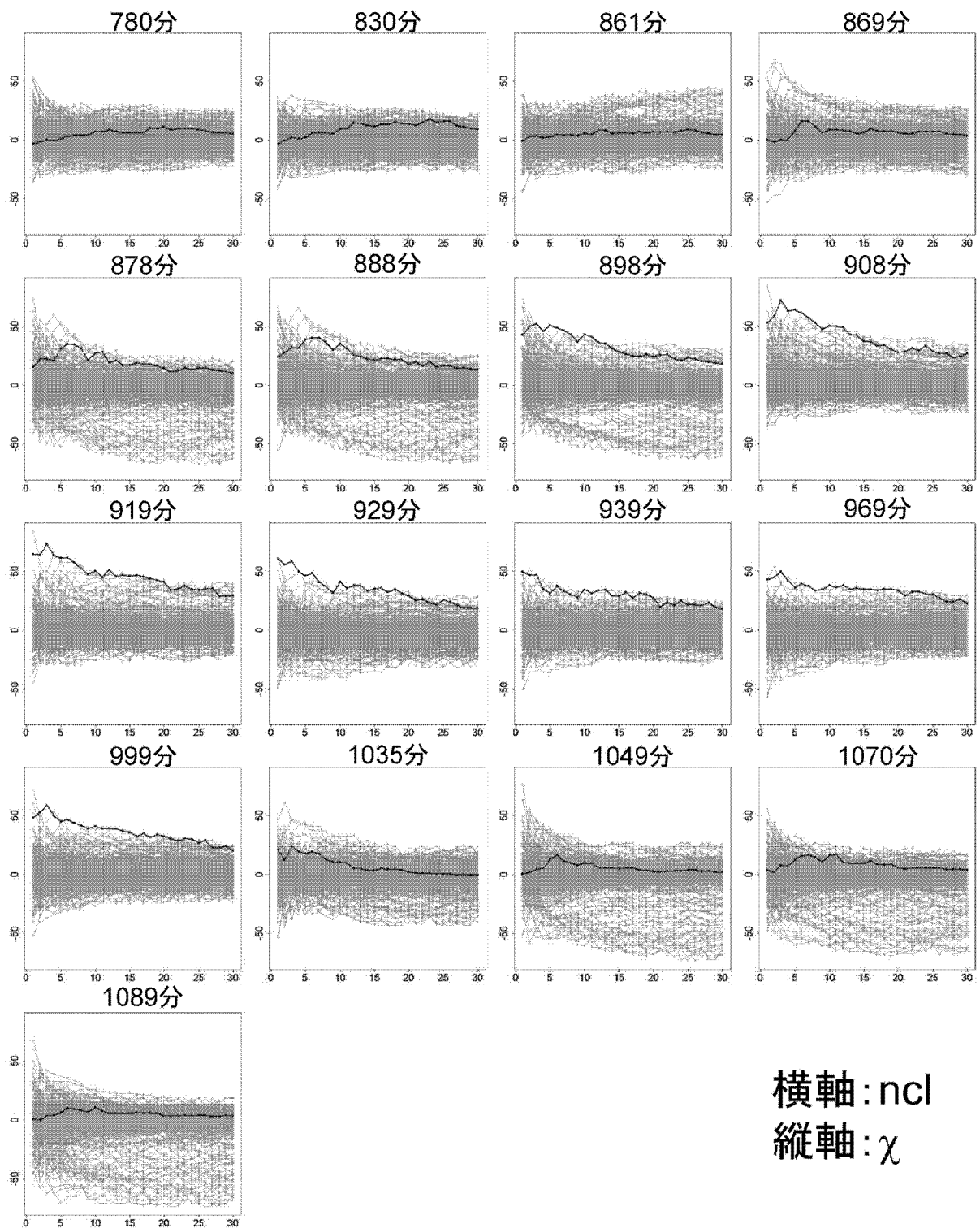
[図48]



[図49]

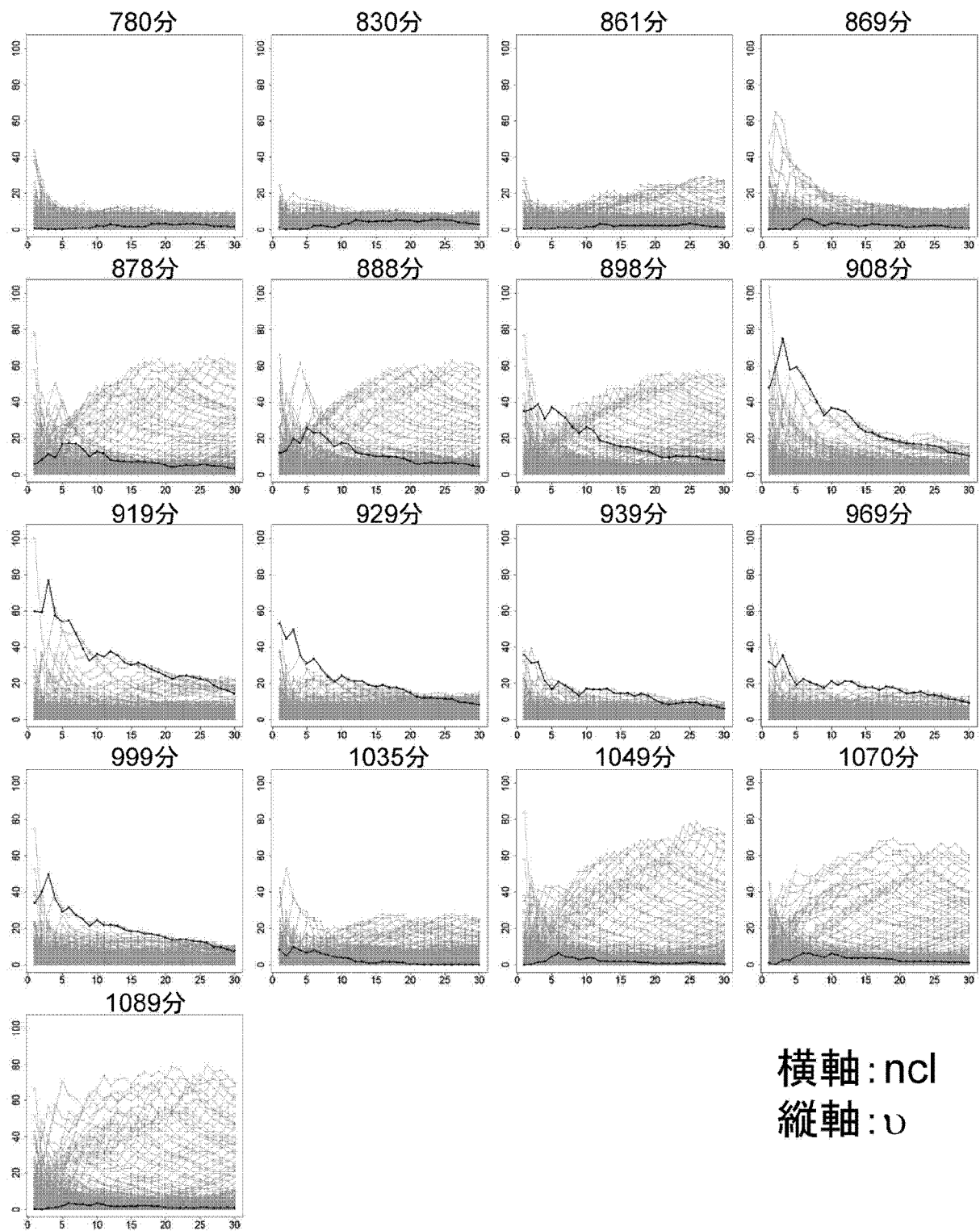


[図50]

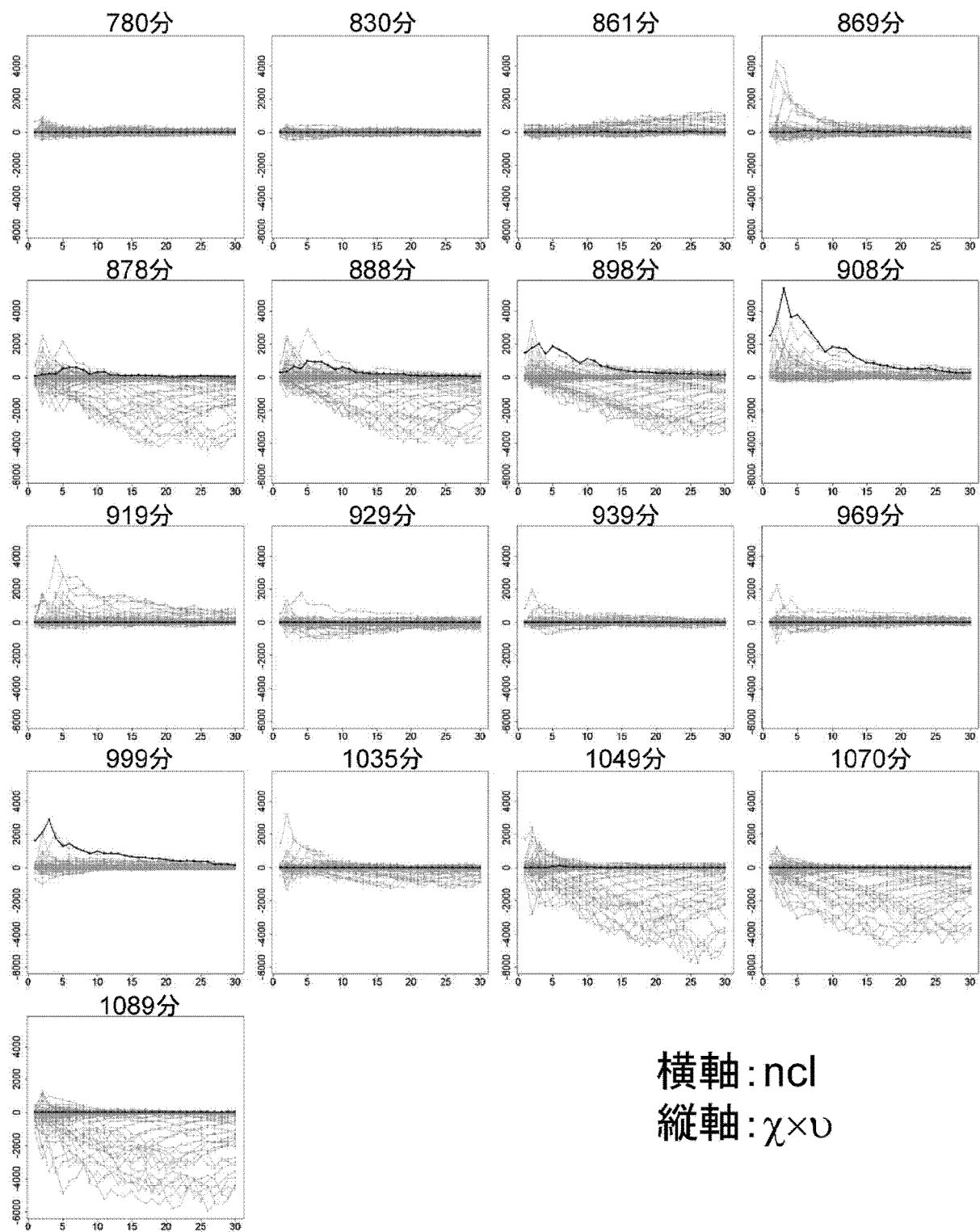


横軸: ncl
縦軸: χ

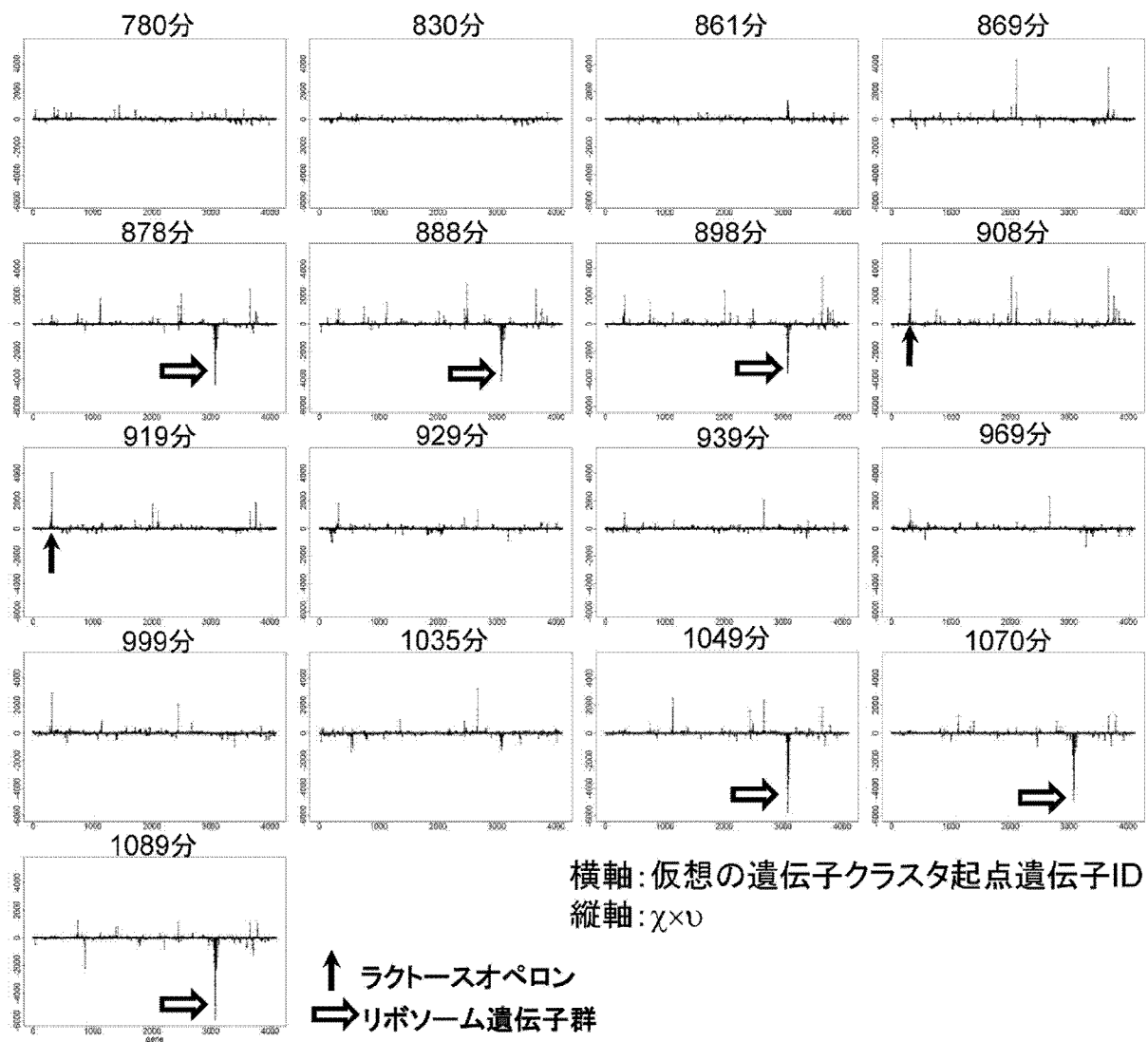
[図51]



[図52]



[図53]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2011/071731

A. CLASSIFICATION OF SUBJECT MATTER

C12Q1/68(2006.01) i, C12M1/00(2006.01) i, C12N15/09(2006.01) i

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)

C12Q1/68, C12M1/00, C12N15/09

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

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

CA/BIOSIS/MEDLINE/WPIDS (STN), JSTPlus/JMEDPlus/JST7580 (JDreamII)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	STARCEVIC A. et al., ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. Nucleic Acids Research, 2008, Vol.36, No.21, pp.6882-6892	1-41
A	ZHAO H. et al., A probabilistic relaxation labeling framework for reducing the noise effect in geometric biclustering of gene expression data. Pattern Recognition, 2009, Vol.42, pp.2578-2588	1-41



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

13 October, 2011 (13.10.11)

Date of mailing of the international search report

25 October, 2011 (25.10.11)

Name and mailing address of the ISA/

Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2011/071731

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Katsuhisa HORIMOTO, Hiroyuki TO, "Inference of a Genetic Network with Use of a Hierarchical Clustering from a Large Amount of Gene Expression Data", Biophysics, 2002, vol.42, no. 3, pages 110 to 115	1-41

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2011/071731

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☒ Claims Nos.: 42-51
because they relate to subject matter not required to be searched by this Authority, namely:
Algorism for analyzing gene cluster is set forth in claims 42-51. Therefore, claims 42-51 are relevant to claims in which scientific and mathematical theory is set forth, and thus relate to a subject matter on which this International Searching Authority is not required to carry out an international search.
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

A. 発明の属する分野の分類（国際特許分類（I P C）） Int.Cl. C12Q1/68(2006.01)i, C12M1/00(2006.01)i, C12N15/09(2006.01)i			
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（I P C）） Int.Cl. C12Q1/68, C12M1/00, C12N15/09			
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			
国際調査で利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） CA/BIOSIS/MEDLINE/WPIDS(STN)、JSTPlus/JMEDPlus/JST7580(JDreamII)			
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	STARCEVIC A. et al., ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. Nucleic Acids Research, 2008, Vol.36, No.21, pp.6882-6892		1 - 4 1
☒ 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 「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 「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） 「O」 口頭による開示、使用、展示等に言及する文献 「P」 国際出願日前で、かつ優先権の主張の基礎となる出願日の後に公表された文献 「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 「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 「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 「&」 同一パテントファミリー文献			
国際調査を完了した日 1 3 . 1 0 . 2 0 1 1		国際調査報告の発送日 2 5 . 1 0 . 2 0 1 1	
国際調査機関の名称及びあて先 日本国特許庁（I S A / J P） 郵便番号 1 0 0 - 8 9 1 5 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） 伊達 利奈 電話番号 0 3 - 3 5 8 1 - 1 1 0 1 内線 3 4 4 8	4 B 3 9 6 0

C (続き) . 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	ZHAO H. et al., A probabilistic relaxation labeling framework for reducing the noise effect in geometric biclustering of gene expression data. Pattern Recognition, 2009, Vol.42, pp.2578-2588	1 - 4 1
A	堀本 勝久, 藤 博幸, 大規模遺伝子発現情報のクラスタリングと遺伝子ネットワークの推定. 生物物理, 2002, Vol.42, No.3, pp.110-115	1 - 4 1

第Ⅱ欄 請求の範囲の一部の調査ができないときの意見（第1ページの2の続き）

法第8条第3項（P C T 17条(2)(a)）の規定により、この国際調査報告は次の理由により請求の範囲の一部について作成しなかった。

1. ☒ 請求項 4 2 - 5 1 は、この国際調査機関が調査をすることを要しない対象に係るものである。
つまり、
請求項 4 2 - 5 1 には、遺伝子クラスタを解析するためのアルゴリズムが記載されている。したがって、請求項 4 2 - 5 1 は、科学及び数学の理論が記載されている請求項であり、国際調査をすることを要しない事項に係るものである。
2. ☐ 請求項 _____ は、有意義な国際調査をすることができる程度まで所定の要件を満たしていない国際出願の部分に係るものである。つまり、
3. ☐ 請求項 _____ は、従属請求の範囲であって P C T 規則6.4(a)の第2文及び第3文の規定に従って記載されていない。

第Ⅲ欄 発明の単一性が欠如しているときの意見（第1ページの3の続き）

次に述べるようにこの国際出願に二以上の発明があるところの国際調査機関は認めた。

1. ☐ 出願人が必要な追加調査手数料をすべて期間内に納付したので、この国際調査報告は、すべての調査可能な請求項について作成した。
2. ☐ 追加調査手数料を要求するまでもなく、すべての調査可能な請求項について調査することができたので、追加調査手数料の納付を求めなかった。
3. ☐ 出願人が必要な追加調査手数料を一部のみしか期間内に納付しなかったため、この国際調査報告は、手数料の納付のあった次の請求項のみについて作成した。
4. ☐ 出願人が必要な追加調査手数料を期間内に納付しなかったため、この国際調査報告は、請求の範囲の最初に記載されている発明に係る次の請求項について作成した。

追加調査手数料の異議の申立てに関する注意

- ☐ 追加調査手数料及び、該当する場合には、異議申立手数料の納付と共に、出願人から異議申立てがあった。
- ☐ 追加調査手数料の納付と共に出願人から異議申立てがあったが、異議申立手数料が納付命令書に示した期間内に支払われなかった。
- ☐ 追加調査手数料の納付はあったが、異議申立てはなかった。