(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
2008 年9 月12 日 (12.09.2008)

PCT

(10) 国際公開番号
WO 2008/108297 A1

(51) 国際特許分類:
G06F 19/00 (2006.01) *G06F 17/30* (2006.01)

(21) 国際出願番号: PCT/JP2008/053647

(22) 国際出願日: 2008 年2 月29 日 (29.02.2008)

(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2007-052583 2007 年3 月2 日 (02.03.2007) JP

(71) 出願人 (米国を除く全ての指定国について): 大学共同利用機関法人 情報・システム研究機構 (RESEARCH ORGANIZATION OF INFORMATION AND SYSTEMS) [JP/JP]; 〒1068569 東京都港区南麻布四丁目６番７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): 五條堀 孝 (GOJOBORI, Takashi). 池尾 一穂 (IKEO, Kazuho). 岡山 利次 (OKAYAMA, Toshitsugu).

(74) 代理人: 辻丸 光一郎, 外(TSUJIMARU, Koichiro et al.); 〒6008813 京都府京都市下京区中堂寺南町１３４ 京都リサーチパーク１号館３０１号室 Kyoto (JP).

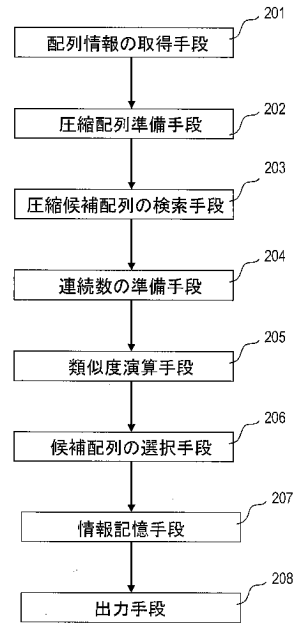
(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG,

[続葉有]

(54) **Title:** HOMOLOGOUS SEARCH SYSTEM

(54) 発明の名称: 相同性検索システム

[図2]



201 UNIT FOR ACQUIRING SEQUENTIAL DATA
202 UNIT FOR PREPARING COMPRESSED SEQUENCE
203 UNIT FOR SEARCHING FOR CANDIDATE COMPRESSED SEQUENCE
204 UNIT FOR PREPARING CONSECUTIVE NUMBER
205 UNIT FOR COMPUTING SIMILARITY
206 UNIT FOR SELECTING CANDIDATE SEQUENCE
207 DATA MEMORY UNIT
208 OUTPUT UNIT

(57) Abstract: In comparing a query sequence with a subject sequence and searching for a similar point in the subject sequence as described above, homologous search can be conducted at a higher accuracy than in the existing methods. After acquiring the sequential data of the query sequence and the subject sequence on the genome scale, these sequences are compression converted into a compressed query sequence and a compressed subject sequence by converting a homopolymer region consisting of two or more consecutive bases of a single kind into a single base of the same kind. Then, these sequences are compared with each other and partial compression subjectsequences in the compressed subject sequence agreeing with the compressed query sequence are narrowed and searched for. For the thus narrowed compressed candidate sequences and the query sequence, the consecutive numbers are compared for each base between both compressed sequences based on the data of the consecutive numbers of a single kind of bases observed in the individual uncompressed sequences. From the degree of agreement or disagreement in the consecutive numbers, a similarity showing the homology of the candidate sequence as described above to the query sequence is computed. Depending on the similarities, an arbitrary number of candidate sequences relatively highly homologous to the query sequence are ranked and selected. Thus, homologous search can be conducted at a high accuracy while avoiding the effect of the consecutive number of a single kind of bases in a homopolymer.

(57) 要約: 問合せ配列を対象配列と対比して前記対象配列における類似箇所を検索する際、従来よりも優れた精度で相同性検索を可能とする。問合せ配列とゲノムスケールの対象配列との配列情報を取得し、同一塩基が２個以上連続したホモポリマー領域を前記塩基１個に置き換えた圧縮問合せ配列

[続葉有]

WO 2008/108297 A1

BH, BR, BW, BY, BZ, CA, CH, 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, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(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, 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** (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:
— **国際調査報告書**

と圧縮対象配列とに圧縮変換し、両者を対比して、圧縮対象配列において圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索する。絞り込んだ圧縮候補配列と問合せ配列とについて、各々の圧縮前配列における同一塩基連続数の情報に基づき、両圧縮配列の間で対応塩基ごとに連続数を対比し、連続数の一致度または不一致度から前記候補配列の前記問合せ配列に対する相同性を示す類似度を演算する。この類似度から問合せ配列と相対的に相同性が高い任意数の候補配列を順位付けて選択することで、ホモポリマーの同一塩基連続数の影響を回避して精度良く相同性検索を行える。

明 細 書

相同性検索システム

技術分野

[0001] 本発明は、相同性検索システム、相同性検索装置、相同性検索方法、および、前記相同性検索方法をコンピュータ上で実行可能なコンピュータプログラムならびにそれを格納した電子媒体に関する。

背景技術

[0002] 近年、生命科学分野において、多くの生物種のゲノム配列全体が解明されている。塩基配列の配列読み取り技術も、オートラジオグラフィにより銀塩フィルムを露光させてラダーパターンを読み取ることによる初期の手法から、電気泳動レーン上の蛍光標識をレーザ光で励起することにより自動で読み取る形式の手法に置き換わり、格段に自動化が推進されている。そして、様々な高感度化、高速化の技術が導入され、スループットも向上している。しかしながら、これらの方法は、全てサンガー法と呼ばれる同じ原理に基づく手法であり、物理的な実泳動時間の制約により、性能に限界がある。そこで、新たにパイロシーケンシング技術が開発され、実用化されるに至っている。この技術は、従来のサンガー法と大きく原理が異なっており、電気泳動ではなく、相補鎖伸長の化学的反応による蛍光強度を、直接読み取る方法である。この原理によって、パイロシーケンシング技術は、サンガー法を遥かに超える配列決定速度を実現した。

[0003] しかしながら、パイロシーケンシング技術では、配列中における同一塩基が複数個繋がった領域(以下、「ホモポリマー領域」という)の配列決定について、次のような問題がある。すなわち、パイロシーケンシング技術においては、配列の情報が、計測時に、ダイナミックレンジの飽和限度のある蛍光強度の比でしか観測されない。このため、同一塩基が連続して繋がっているホモポリマー領域に関しては、同一塩基の数を正確に決定し難く、結果的に、配列決定精度に問題が生じる。このようなホモポリマー領域に関する配列決定精度の問題は、前述のサンガー法においても、同様に技術的限界が潜在していた。しかしながら、パイロシーケンシング技術は、高いスルー

プットであるがゆえに、サンガー法と比較して、前記ホモポリマー領域の問題がより顕著となっている。

[0004] 他方、例えば、ゲノム上の位置が不明な配列、機能や起源等が不明な配列(以下、「問合せ配列」という)について、解読されたゲノム等の配列(以下、「対象配列」という)において相同する部分配列を検索する相同性検索(類似性検索)が、遺伝子解析において行われている。この相同性検索の技術は、前述の配列決定法の飛躍的な進歩と較べて、従来と余り変化がなく、下記の手法が一般的である。

[0005] (1)相同性検索を行う代表的なシステムとしてBLASTがある(非特許文献1)。このシステムは、生命科学分野で配列検索を行う際の標準として、広く普及、定着している。

(2)部分配列の不一致を配列の挿入・削除のスコアリングで最大限に許容する類似度検索法として、動的計画法(ダイナミックプログラミング)によるSmith－Waterman法がある(非特許文献2)。この方式は、複数のシステムの実装に用いられている。

(3)さらに、前記(2)の動的計画法の論理をハードウェアに組み込んで、超並列実行させることで、速度に関する問題の解決を試みる手法が報告されている(特許文献1)。

非特許文献1:Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. (1990) Basic local alignment search tool. J. Mol. Biol. Vol.215, pp.403-410.

非特許文献2:Smith TF, Waterman MS. (1981) Comparison of biosequences. Adv. Appl. Math. 2:482-9.

特許文献1:特開平07－093370号公報

発明の開示

[0006] しかしながら、これらの相同性検索方法は、前述のような塩基配列決定法により決定された塩基配列情報に基づいて行われるが、前記塩基配列決定法によって生じるホモポリマー領域の配列決定精度の問題を回避できるものではない。つまり、相同性検索に使用するゲノム等の対象配列がホモポリマー領域を有する場合、塩基配列決定法によって決定された前記ホモポリマー領域の同一塩基の連続数は、前述のよう

に精度に問題がある。しかし、前述の相同性検索方法は、このような問題に対応した手法とは言い難い。このため、例えば、問合せ配列に対して、ゲノム等の対象配列における部分配列が実際には相同性が高い場合でも、例えば、配列精度の影響によって、結果を抽出できなかったり、類似していないにも関わらず誤って結果を抽出するという問題がある。

[0007] 前記(1)のBLASTは、例えば、長い問合せ配列と対象配列とを対比する場合、前記問合せ配列の内部において同一塩基連続数の不一致があると、この不一致を、塩基の挿入、削除として解析する。これによって、ある程度、完全には一致しないもの同士であっても、対応付けて検索することが可能である。しかし、問合せ配列が短い配列であったり、問合せ配列の両端付近に同一塩基連続数の不一致がある場合、相違を過大評価して、短い問合せ配列全体を不一致と判断したり、前記端の部分を不一致として、相同する配列の候補から除外するケースが多く観察される。

[0008] 前記(2)のSmith－Waterman法は、BLASTと較べて、ホモポリマー領域の場所に依存して不一致を出す可能性は少なくなる。さらに、動的計画法のアルゴリズムによって最適な配列塩基の対応であるアライメントを探し出すことで、BLASTと比較して、良好な探索ができる。しかしながら、ホモポリマー領域における塩基の連続数の不一致と、他の単一塩基の不一致とが、同一尺度で測定されてしまうため、相同性の順序付けの合理性に未だ問題がある。さらに、性能上、基本となっている動的計画法の実行に関して、問合せ配列と対象配列との積のオーダの計算量がかかるため、極端に検索性能が遅いという欠点がある。また、この方法は、例えば、配列決定法の進歩に伴い、問合せ配列として網羅的な量、例えば、1,000,000を超える超大量の数を扱う場合において、実用性に欠ける。

[0009] 前記(3)の手法は、前記(2)と基本的なアルゴリズムが同じであり、動作精度上、同様の問題がある。また、専用のハードウェアを利用する必要があるため、性能的には相当の改善があるものの、計算機ソフトウェアによる方法と較べてコストがかかる。また、汎用計算機上で稼動するシステムと較べて、ハードウェアが固定されることで、信頼性を含む性能仕様が陳腐化するという問題がある。このため、現在では、利用が特定の範囲に限定されている。

[0010] このように、何れの相同性検索の手法も、対象配列と問合せ配列との対比において、ホモポリマー領域における同一塩基の塩基個数に不一致を含む場合、例えば、正確性や、性能上、コスト上等の問題がある。このため、両者の対応するホモポリマー領域において同一塩基の連続数に不一致がある場合に適した相同性検索が求められる。

[0011] そこで、本発明は、問合せ配列について、対象配列において相同する部分配列を検索する際に、前記両者の対応するホモポリマー領域において同一塩基の連続数に差が生じる場合であっても、従来よりも優れた精度で、迅速に、相同性検索を可能とすることを目的とする。

[0012] 前記目的を達成するために、本発明の相同性検索システムは、核酸塩基配列からなる問合せ配列の配列情報を用いて、核酸塩基配列からなるゲノムスケールの対象配列の配列情報から、前記問合せ配列と相同する部分配列を検索する相同性検索システムであって、

前記問合せ配列および対象配列の配列情報を取得する取得手段と;

取得された前記問合せ配列および前記対象配列について、それぞれ、同一塩基が2個以上連続したホモポリマー領域を前記塩基1個に置き換えた圧縮問合せ配列および圧縮対象配列を準備する圧縮配列準備手段と;

前記圧縮問合せ配列と前記圧縮対象配列とを対比し、前記圧縮対象配列において前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索して、絞り込んだ前記圧縮対象部分配列を候補配列の圧縮配列(圧縮候補配列)として選択する検索手段と;

前記圧縮問合せ配列と前記検索手段で選択した前記圧縮候補配列とについて、それぞれの圧縮前の配列における同一塩基の連続数の情報を準備する連続数準備手段と;

前記同一塩基の連続数の情報に基づいて、前記圧縮問合せ配列と前記圧縮候補配列との間で、対応する塩基ごとに、前記塩基の連続数を対比し、前記連続数の一致度または不一致度から、前記候補配列の前記問合せ配列に対する相同性を示す類似度を演算する類似度演算手段と;

前記類似度演算手段により演算した類似度に基づいて、前記問合せ配列と相対的に相同性が高い任意数の候補配列を順位付けて選択する選択手段と;

前記選択手段により選択した前記任意数の候補配列の情報を出力する出力手段とを有する。

[0013] 本発明の相同性検索装置は、核酸塩基配列からなる問合せ配列の配列情報を用いて、核酸塩基配列からなるゲノムスケールの対象配列の配列情報から、前記問合せ配列と相同する対象部分配列を検索する相同性検索装置であって、本発明の相同性検索システムを含む。

[0014] 本発明の相同性検索方法は、核酸塩基配列からなる問合せ配列の配列情報を用いて、核酸塩基配列からなるゲノムスケールの対象配列の配列情報から、前記問合せ配列と相同する対象部分配列を検索する相同性検索方法であって、

前記問合せ配列および対象配列の配列情報を取得する取得ステップと;

取得された前記問合せ配列と前記対象配列について、それぞれ、同一塩基が2個以上連続したホモポリマー領域を前記塩基1個に置き換えた圧縮問合せ配列と圧縮対象配列とを準備する圧縮配列準備ステップと;

前記圧縮問合せ配列と前記圧縮対象配列とを対比し、前記圧縮対象配列において前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索して、絞り込んだ前記圧縮対象部分配列を候補配列の圧縮配列(圧縮候補配列)として選択する検索ステップと;

前記圧縮問合せ配列と前記検索ステップで選択した前記圧縮候補配列とについて、それぞれの圧縮前の配列における同一塩基の連続数の情報を準備する連続数準備ステップと;

前記同一塩基の連続数の情報に基づいて、前記圧縮問合せ配列と前記圧縮候補配列との間で、対応する塩基ごとに、前記塩基の連続数を対比し、前記連続数の一致度または不一致度から、前記候補配列の前記問合せ配列に対する相同性を示す類似度を演算する類似度演算ステップと;

前記類似度演算ステップにより演算した類似度に基づいて、前記問合せ配列と相対的に相同性が高い任意数の候補配列を順位付けて選択する選択ステップと;

前記選択ステップにより選択した前記任意数の候補配列の情報を出力する出力ステップとを有する。

[0015] 本発明のコンピュータプログラムは、本発明の相同性検索方法をコンピュータ上で実行可能なコンピュータプログラムである。

[0016] 本発明の電子媒体は、本発明のコンピュータプログラムを格納した電子媒体である。

[0017] 本発明は、塩基配列の決定において生じるホモポリマー領域の同一塩基連続数の変動による問題を考慮して、まず、同一塩基連続数に影響されない圧縮配列(ホモポリマー領域を塩基1個に置き換えた圧縮配列)として、対象配列と問合せ配列を対比させた上で、つぎに、両者の相同性を、ホモポリマー領域の連続塩基数から判断している。従来法であれば、ホモポリマー領域の同一塩基連続数の変動が原因となり、合理的ではない不適切な相同性順位の結果となったり、または、ホモポリマー領域における同一塩基連続数の変動自体が見落とされるおそれがあったが、本発明は、このような問題を回避して、より正確に問合せ配列と相同する対象配列の部分配列を選択することが可能となった。したがって、本発明によれば、例えば、ホモポリマー領域における同一塩基連続数が、塩基配列の決定法や配列自身の多型のために、誤差や変位を含む場合であっても、その影響を回避して、より正確な相同性検索が可能となる。特に、塩基配列の情報が、従来のサンガー法による場合だけでなく、パイロシーケンシング技術によってハイスループットで決定されている場合であっても、ホモポリマー領域の同一塩基連続数の決定精度が低いことによる影響を回避できる。さらに、このように精度良く相同性検索が行えることから、例えば、問合せ配列と対象配列における部分配列とが、唯一の相同性(類似性)を示すものであるか否かの判断も精度良く行うことが可能になる。また、ホモポリマー領域の同一塩基連続数を考慮しない圧縮配列同士を対比し、一致した対象配列の部分配列を選択するため、従来と比較して、データ処理能力が格段に向上し、低コスト化の実現も可能である。したがって、本発明は、相同性検索(類似性検索)の分野において、これまで回避できなかったホモポリマー領域における同一塩基連続数の変動による影響を解決できることから、特に遺伝子解析の分野において極めて有用な技術であるといえる。

図面の簡単な説明

[0018] [図1]図1は、本発明の一実施形態における相同性検索装置のハードウェア構成の一例を示すブロック図である。

[図2]図2は、本発明のその他の実施形態における相同性検索システムを示す概略図である。

[図3]図3は、本発明のさらにその他の実施形態における相同性検索システムを示す概略図である。

[図4]図4は、本発明のさらにその他の実施形態における圧縮変換ならびに同一塩基の連続数の計数の概略を示す図である。

[図5]図5は、本発明のさらにその他の実施形態における相同性検索方法の処理の流れを示すフローチャートである。

[図6]図6は、本発明のさらにその他の実施形態における類似度の算出方法の概略を示す図である。

[図7]図7は、本発明のさらにその他の実施形態における問合せ配列用ハッシュテーブルの一例を示す概略図である。

[図8]図8は、本発明のシステムを用いたスタンドアローン型の装置の一例の全体構成図である。

[図9]図9は、本発明のシステムを用いたネットワーク利用型の装置の一例の全体構成図である。

[図10]図10は、前記スタンドアローン型の装置の機器構成の一例を示すブロック図である。

[図11]図11は、前記ネットワーク型の装置の機器構成の一例を示すブロック図である。

発明を実施するための最良の形態

[0019] 本発明において「問合せ配列」は、特に制限されず、核酸塩基配列であればよい。前記塩基配列としては、例えば、各種生物のゲノム断片配列やオリゴキャップ法等によって得られた完全長や断片のトランスクリプトーム配列等が含まれる。本発明における問合せ配列の長さは、特に制限されないが、例えば、12～60塩基であり、好ま

しくは18～25塩基である。

[0020] 本発明において「ゲノムスケールの対象配列」は、特に制限されず、例えば、ゲノムとして解読された全核酸塩基配列や、染色体全体の核酸塩基配列、また、それらの1塩基多型やハプロタイプ等の変異配列、さらに、ゲノムからの核酸複製産物であるトランスクリプトームと呼ばれる転写物の網羅的な収集配列等があげられる。また、このようなゲノムスケールの対象配列は、例えば、各種データベース(例えば、DDBJ、EMBL、ENSEMBL、GenBank、UCSC等)に登録されている配列が利用できる。前記対象配列の長さとしては、制限されず、例えば、ヒトゲノム30億塩基対であっても可能であり、特に100万塩基以上の配列について、本発明を適用することが好ましい。

[0021] 本発明において「ホモポリマー領域」とは、核酸塩基配列において、2個以上の同一核酸塩基(例えば、アデニン、グアニン、シトシンまたはチミン)が連続(反復)した領域を意味する。

[0022] 本発明において「圧縮配列」とは、問合せ配列および対象配列の配列情報を、同一塩基が2個以上連続した前記ホモポリマー領域を前記塩基1個に置き換えたそれぞれの配列をいう。つまり、問合せ配列および対象配列に関する、核酸塩基の種別の連なりを示す配列情報である。前記塩基1個への置き換えを、圧縮変換といい、問合せ配列を圧縮変換した配列を、圧縮問合せ配列、対象配列を圧縮変換した配列を、圧縮対象配列という。また、「圧縮対象部分配列」とは、前記圧縮問合せ配列と一致する、圧縮対象配列における部分配列である。

[0023] 本発明において「連続数の情報」とは、問合せ配列および対象配列の配列情報を、核酸塩基の種別の連なりではなく、同一塩基がいくつ連続して存在するのかを示す配列情報である。同一塩基が連続してn個存在する場合は、「n」と計数できる。具体的には、例えば、核酸塩基配列において、同一塩基が1個存在する場合は「1」、同一塩基が連続して2個存在する場合は「2」というように計数できる。

[0024] 本発明において、対象配列は、前述のようにゲノムスケールの核酸塩基配列である。一般的な相同性検索では、ゲノムスケールの対象配列は、通常、部分配列に分解してから検索が行われる。具体的には、対象配列の先頭から、例えば、順次、1塩基

ずつずらして複数の部分配列を作成し、これらの部分配列からなる部分配列群を使用する。本発明においても、対象配列は、部分配列に区切った対象部分配列群を用いた検索であることが好ましい。したがって、本発明において、前記圧縮対象配列は、例えば、対象配列について、同一塩基が2個以上連続したホモポリマー領域を前記塩基1個に置き換える圧縮処理を施し、得られた圧縮後の圧縮対象配列を固定長に区切った圧縮対象部分配列からなる圧縮対象部分配列群であることが好ましい。前記固定長は、制限されず、例えば、1～100塩基であっても本発明を実施でき、特に8～50塩基の長さで、非常に有効な検索が可能である。前記固定長は、例えば、後述するハッシュの対象となる圧縮配列の塩基長（ハッシュ対象塩基長）であり、いわゆる、「ハッシュ幅」である。

[0025] 本発明において、問合せ配列と対比する対象配列の数は、制限されず、例えば、1つの問合せ配列に対して、目的とする1つの対象配列を対比してもよいし、2つ以上の対象配列を対比させてもよい。また、問合せ配列の数も制限されず、例えば、1つの対象配列に対して、1つまたは2つ以上の問合せ配列を対比させてもよい。圧縮後の圧縮対象配列は、前述のように固定長に区切って圧縮対象部分配列群とした後、各圧縮対象部分配列に対して、例えば、後述するハッシュ処理が行われる。この際、前記固定長は、例えば、圧縮した問合せ配列と同じ長さにすることが好ましい。そして、問合せ配列が複数存在する場合には、複数の圧縮問合せ配列の長さのバリエーションだけ、圧縮後の対象配列から圧縮対象部分配列群を生成し、各圧縮対象部分配列群について独立して相同性検索を多重に行うことが好ましい。

[0026] 本発明において、類似度に基づいて選択する候補配列の数は、制限されず、任意数を設定できる。例えば、最も相同性が高い結果を示す候補配列のみを選択してもよいし、相同性が高い順序を決定して、上位数個を前記順序に基づいて選択してもよい。また、相同性を示す候補配列が、設定した任意数に達しない場合、選択する候補配列の数は、任意数未満であってもよい。

[0027] 以下に、本発明の相同性検索システム、相同性検索装置、相同性検索方法、これをコンピュータ上で実行可能なコンピュータプログラム、および、前記プログラムを格納した電子媒体について説明する。本発明の相同性検索方法は、例えば、本発明

の相同性検索システムや本発明の相同性検索装置、本発明のコンピュータプログラムの実行によって実現できる。

[0028] 本発明によれば、前述のように、ホモポリマー領域の同一塩基連続数の変動を考慮した上で、相同性検索を行うため、例えば、従来法では検索できなかった相同する部分配列を検索でき、また、従来法のように誤って相同性が判断されるおそれを回避できる。さらに、ホモポリマー領域の同一塩基連続数の影響をうけない圧縮配列同士を対比するため、データ処理速度を格段に向上できる。これらの効果を、例えば、本発明の相同性検索装置について着目すると、以下のようなことがいえる。問合せ配列をそのまま従来のBLAST検索にかけた場合、特に、問合せ配列および対象配列の少なくとも一方において、複数の個所に、同一塩基が連続したホモポリマー領域が存在すると、ヒットしない場合がある。これを回避して全ての場合で検索を成功させるために、従来は、問合せ配列側で、ホモポリマー領域における同一塩基の連続数の誤差の許容範囲で、全ての組み合わせを生成し、BLASTを実行する必要があった。しかし、この方法では、例えば、4塩基の出現確率が均等かつランダムであるという想定の下、同一塩基連続数の誤りの許容範囲を2倍未満で、モンテカルロ法による1,000,000回シミュレーションをした場合、問合せ配列長25塩基で約135倍、問合せ配列長50塩基で約21，500倍、問合せ配列長100塩基で約84,000,000倍のパタンを検索する必要がある。そして、検索時間もそれにほぼ比例した時間が必要となる。これに対して、本発明によれば、問合せ配列長にかかわらず、何れも1回のBLASTサーチに相当する時間に、本発明独自のスコア計算時間（配列長に対して線形）を加算するだけの性能で検索処理が可能となる。

[0029] ＜相同性検索システム＞

本発明の第1の相同性検索システムは、前述のように、

前記問合せ配列および対象配列の配列情報を取得する取得手段と；

取得された前記問合せ配列および前記対象配列について、それぞれ、同一塩基が2個以上連続したホモポリマー領域を前記塩基1個に置き換えた圧縮問合せ配列および圧縮対象配列を準備する圧縮配列準備手段と；

前記圧縮問合せ配列と前記圧縮対象配列とを対比し、前記圧縮対象配列におい

て前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索して、絞り込んだ前記圧縮対象部分配列を候補配列の圧縮配列(圧縮候補配列)として選択する検索手段と;

前記圧縮問合せ配列と前記検索手段で選択した前記圧縮候補配列とについて、それぞれの圧縮前の配列における同一塩基の連続数の情報を準備する連続数準備手段と;

前記同一塩基の連続数の情報に基づいて、前記圧縮問合せ配列と前記圧縮候補配列との間で、対応する塩基ごとに、前記塩基の連続数を対比し、前記連続数の一致度または不一致度から、前記候補配列の前記問合せ配列に対する相同性を示す類似度を演算する類似度演算手段と;

前記類似度演算手段により演算した類似度に基づいて、前記問合せ配列と相対的に相同性が高い任意数の候補配列を順位付けて選択する選択手段と;

前記選択手段により選択した前記任意数の候補配列の情報を出力する出力手段とを有する。

[0030] 本発明のシステムは、問合せ配列と対象配列とについて、前述のように、圧縮配列および同一塩基連続数を対比して相同性を検索することが特徴であり、その他の条件や構成は特に制限されない。

[0031] 前記取得手段により取得される配列情報は、例えば、圧縮前の配列情報があげられる。また、前記配列情報としては、前記圧縮前の配列に代えて、または、加えて、例えば、圧縮配列、同一塩基の連続数等の情報が含まれてもよい。前記取得手段により、配列情報として圧縮配列が取得される場合、例えば、前記取得手段と、圧縮配列を準備する前記圧縮配列準備手段は、同じ手段ともいえる。また、前記取得手段により、配列情報として同一塩基の連続数の情報が取得される場合、例えば、前記取得手段と、同一塩基の連続数の情報を準備する前記連続数準備手段は、同じ手段ともいえる。また、問合せ配列および対象配列のいずれか一方の配列についてのみ、圧縮配列や同一塩基の連続数等の配列情報が取得され、他方の配列については、圧縮前の配列情報が取得されてもよい。この場合、前記他方の配列についてのみ、圧縮前の配列情報から、後述するような手段により圧縮配列や同一塩基の連続

数の情報を取得してもよい。

[0032] 本発明のシステムにおいて、前記配列情報の取得手段は、特に制限されないが、例えば、対象配列および／または問合せ配列の配列情報を入力する入力手段があげられる。この場合、本発明のシステムは、さらに、入力された問合せ配列の配列情報を記憶する記憶手段（問合せ配列記憶手段）や、入力された対象配列の配列情報を記憶する記憶手段（対象配列記憶手段）を有することが好ましい。前記問合せ配列の配列情報としては、前述のように、圧縮前の配列の他に、例えば、圧縮配列や、同一塩基の連続数、起源等の情報があげられる。また、前記対象配列の配列情報としては、例えば、前述のように、圧縮前の他に、例えば、圧縮配列、同一塩基の連続数、起源、対象配列における各領域の機能の有無や機能の内容等があげられる。

[0033] また、前記配列情報の取得手段としては、例えば、前記問合せ配列の配列情報を入力する入力手段と、前記対象配列の配列情報が記憶されている記憶手段（対象配列記憶手段）とを有してもよい。検索対象とする対象配列は、例えば、対象配列記憶手段に記憶された配列の中から所望の配列を指定することで、対象配列記憶手段から呼び出すことができる。また、さらに、入力した前記問合せ配列の配列情報を記憶する記憶手段（問合せ配列記憶手段）を有することが好ましい。

[0034] 前記対象配列記憶手段としては、例えば、前述のように、対象配列の配列情報が蓄積されたデータベース（対象配列データベース）があげられる。蓄積される対象配列の数は、何ら制限されない。このような対象配列データベースとしては、例えば、各種ゲノムや染色体の核酸塩基配列が蓄積された公知のデータベースが例示できる。また、前記対象配列データベースとしては、制限されず、例えば、通信網を介して接続されるデータベース、対象配列が記憶されたリムーバブル記録媒体等があげられる、前記データベースは、コンピュータシステムの記憶手段（記憶装置）に格納されていてもよい。

[0035] 前記対象配列記憶手段は、対象配列の情報として、例えば、圧縮前の配列情報の他に、さらに、前記圧縮対象配列、前記対象配列における同一塩基の連続数、起源、対象配列における各領域の機能の有無や機能の内容等が記憶されてもよい。このように、圧縮対象配列や同一塩基の連続数等の情報を前記記憶手段に蓄積すれば

、例えば、さらに他の問合せ配列に対する相同性の検索を行う際、前記対象配列記憶手段から、圧縮配列や同一塩基の連続数等の必要な情報を呼び出すことができる。このようにすれば、再度、後述するような圧縮変換や同一塩基数の計数を行う手間を省くことができ、システムの検索能がさらに向上する。また、本発明のシステムにおいて、前記対象配列記憶手段に記憶されていない対象配列について検索を行う場合は、入力手段によって入力された前記対象配列情報を、前記対象配列記憶手段に追加していくことが好ましい。

[0036] また、前記配列情報の取得手段は、例えば、前記対象配列の配列情報を入力する入力手段と、前記問合せ配列の配列情報が記憶されている記憶手段(問合せ配列記憶手段)とを有してもよい。前記記憶手段としては、前記問合せ配列の配列情報が蓄積されたデータベース(問合せ配列データベース)があげられる。検索対象とする問合せ配列は、例えば、前記問合せ配列記憶手段に記憶された配列の中から所望の配列を指定することで、問合せ配列記憶手段から呼び出すことができる。前記指定は、例えば、指定する配列と関連付けた情報(例えば、配列のID等)を前記入力手段で入力することによって行うこともできる。また、さらに、入力した前記対象配列の配列情報を記憶する記憶手段(対象配列記憶手段)を有することが好ましい。

[0037] 前記問合せ配列記憶手段は、問合せ配列の情報として、例えば、圧縮前の配列情報の他に、さらに、前記圧縮問合せ配列、前記問合せ配列における同一塩基の連続数、起源等が記憶されてもよい。このように、圧縮問合せ配列や同一塩基の連続数等の情報を蓄積すれば、例えば、他の対象配列に対する相同性の検索を行う際に、前記問合せ配列記憶手段から、圧縮配列や同一塩基の連続数等、必要な情報を呼び出すことができる。このようにすれば、再度、後述するような圧縮変換や同一塩基数の計数を行う手間を省くことができ、システムの検索能がさらに向上する。また、本発明のシステムにおいて、前記問合せ配列記憶手段に記憶されていない問合せ配列について検索を行う場合は、入力手段によって入力された前記問合せ配列の配列情報を、前記問合せ配列記憶手段に追加することが好ましい。

[0038] また、前記配列情報の取得手段は、例えば、前記問合せ配列が記憶されている記憶手段(問合せ配列記憶手段)と、前記対象配列の配列情報が記憶されている記憶

手段（対象配列記憶手段）とを有してもよい。前記各種記憶手段は、例えば、予め、各配列の配列情報が記憶されていてもよいし、さらに、前述の入力手段により入力された配列情報が記憶されてもよい。そして、検索対象とする問合せ配列および対象配列は、例えば、各記憶手段に記憶された配列の中から検索対象とする所望の配列を指定することで、各記憶手段から呼び出すことができる。

[0039] 前述のように、各種記憶手段（データベース）には、新たな情報を追加することが好ましいことから、本発明のシステムは、さらに、情報を追加するための情報更新手段（データベース更新手段）を有することが好ましい。

[0040] 前記圧縮問合せ配列および圧縮対象配列は、それぞれ、システム内での圧縮変換により得てもよいし、前記入力手段により入力してもよいし、予め配列情報として記憶されている場合には、前述の各記憶手段から呼び出してもよい。すなわち、本発明のシステムにおいて、前記圧縮配列準備手段としては、例えば、取得した問合せ配列および／または対象配列の配列情報に基づいて、圧縮問合せ配列および／または圧縮対象配列に圧縮変換する手段（圧縮変換手段）があげられる。また、前記圧縮配列準備手段は、前述の問合せ配列記憶手段および／または対象配列記憶手段にそれぞれの圧縮配列が記憶されている場合、圧縮問合せ配列が記憶されている問合せ配列記憶手段および／または圧縮対象配列が記憶されている対象配列記憶手段であってもよい。後者の場合、検索対象とする所望の配列を指定することで、各記憶手段からそれぞれの圧縮配列を呼び出すことができる。

[0041] 前記問合せ配列および対象配列における同一塩基の連続数は、それぞれ、システム内での計数処理によって得てもよいし、前記入力手段により入力してもよいし、予め記憶されている場合には、前述の各記憶手段から呼び出してもよい。すなわち、本発明のシステムにおいて、前記連続数準備手段として、例えば、取得した問合せ配列および／または対象配列の配列情報に基づいて、問合せ配列および／または対象配列について、それぞれの圧縮前の配列における同一塩基の連続数を計数（演算）する計数手段（連続数計数（演算）手段）があげられる。また、前記連続数準備手段は、前述の問合せ配列記憶手段および／または対象配列記憶手段にそれぞれの連続数の情報が記憶されている場合、問合せ配列の情報が記憶されている問合せ

配列記憶手段および／または対象配列の情報が記憶されている対象配列記憶手段であってもよい。後者の場合、検索対象とする所望の配列を指定することで、各記憶手段からそれぞれの連続数の情報を呼び出すことができる。

[0042] 本発明において前記検索手段は、特に制限されないが、例えば、ハッシュ検索手段、二分木検索手段、B木(B-tree)検索手段等があげられる。前記ハッシュ検索手段は、例えば、前記圧縮問合せ配列と前記圧縮対象部分配列群の各圧縮対象部分配列とをキーとし、同じハッシュ関数を用いて、ハッシュ検索を行うことにより、前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索する。本発明においてハッシュ検索を行う際には、例えば、未圧縮の問合せ配列と未圧縮の部分配列群の各対象部分配列とをキー(要素)とするのではなく、例えば、前記圧縮問合せ配列および前記圧縮対象部分配列群の各圧縮対象部分配列をキー(要素)とし、同じハッシュ関数を用いてハッシュ検索を行う点がポイントとなる。このポイント以外、すなわち、例えば、ハッシュ関数の設定や、ハッシュ検索の手法自体は、従来公知の方法に基づいて行うことができる。

[0043] 本発明の相同性検索システムによりハッシュ検索を行う場合、例えば、さらに、前記圧縮対象部分配列群の各圧縮対象部分配列をキーとし、同じハッシュ関数を用いて、対象配列用ハッシュテーブルを生成する対象配列用ハッシュテーブル生成手段を有することが好ましい。この場合、前記検索手段は、ハッシュ検索手段であり、例えば、前記圧縮問合せ配列をキーとし、前記対象配列用ハッシュテーブル生成手段と同じハッシュ関数を用いて、前記対象配列用ハッシュテーブル生成手段で生成した前記対象配列用ハッシュテーブルのハッシュ検索を行うことにより、前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索すればよい。このように、対象配列用ハッシュテーブルを生成すれば、これにアクセスすることでハッシュ検索を実行できる。このため、対象配列がゲノムスケールの大規模な核酸塩基配列であっても、計算時間をさらに短縮できる。なお、ハッシュテーブルの生成方法は、圧縮配列をキーとする以外は、従来公知の方法に基づいて行うことができる(以下、同様)。

[0044] また、ハッシュ検索を実行する場合、さらに、2つ以上の前記圧縮問合せ配列をキーとし、同じハッシュ関数を用いて、問合せ配列用ハッシュテーブル生成する問合せ

配列用ハッシュテーブル生成手段を有することが好ましい。この場合、前記検索手段は、ハッシュ検索手段であり、例えば、前記圧縮対象部分配列群の各圧縮対象部分配列をキーとし、前記問合せ配列用ハッシュテーブル生成手段と同じハッシュ関数を用いて、前記問合せ配列用ハッシュテーブル生成手段で生成した前記問合せ配列用ハッシュテーブルのハッシュ検索を行うことにより、前記各圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索することができる。このように問合せ配列用のハッシュテーブルを生成すれば、例えば、各問合せ配列に順次アクセスすることでハッシュ検索が実行できるため、問合せ配列が大量であっても、計算時間をさらに短縮できる。

[0045] 本発明においては、このようにシステム内で各種ハッシュテーブルを生成してもよいが、システム外において予め作成されたハッシュテーブルを入力する形態であってもよい。また、対象配列用ハッシュテーブルは、前述の各種データベースに格納されてもよい。ハッシュテーブルについては後述する。

[0046] 前記類似度演算手段は、前記対応する塩基ごとの連続数の不一致度をペナルティスコアとし、前記対応する塩基ごとのペナルティスコアを加算することによって、類似度を演算することが好ましい。ただし、前記圧縮問合せ配列の上流末端塩基または下流末端塩基における圧縮前の同一塩基連続数が、前記圧縮候補配列の上流末端塩基または下流末端塩基における圧縮前の同一塩基連続数よりも短い不一致を除くことが好ましい。類似度の演算については、後述する。

[0047] 本発明の相同性検索システムは、さらに、前記問合せ配列と前記選択手段により選択した任意数の候補配列との情報を記憶する記憶手段を有することが好ましい。このように、前記問合せ配列とこれに相同性を示す候補配列との情報を順次記憶していくことが好ましい。この記憶手段は、例えば、前述の問合せ配列記憶手段や対象配列記憶手段であってもよい。前者の場合、問合せ配列に関連付けて、前記候補配列の情報を記憶させることが好ましく、後者の場合、対象配列に関連付けて前記問合せ配列の情報を記憶させることが好ましい。記憶する候補配列の数は制限されないが、所望の数(任意数)に設定することが好ましい。そして、記憶した候補配列の個数が前記任意数に達した際には、記憶された候補配列と、新たな候補配列とについて

、類似度を比較して、再度、相同性が高い候補配列の順序を決定し、任意数の候補配列を記憶することが好ましい。したがって、このような場合、前記選択手段は、例えば、前記類似度演算手段によって新たな候補配列の前記問合せ配列に対する類似度が演算された際、前記新たな類似度と、前記記憶手段により記憶された前記任意数の候補配列の前記問合せ配列に対する類似度とに基づいて、前記各候補配列から、再度、任意数の候補配列を選択することが好ましい。そして、前記問合せ配列に相同する候補配列が複数存在する場合には、前記問合せ配列と候補配列の情報を記憶し、さらに、新たな候補配列が検索された場合には、複数の候補配列の中から、再度、任意数の候補配列を選択することが好ましい。このようにすれば、例えば、前記問合せ配列に相同する候補配列が多数検索された場合でも、特に相同する配列を選択していくことができる。

[0048] 本発明の相同性検索システムを用いて、問合せ配列が、対象配列のある部分配列にのみ特異的に相同するか否かを検索する場合には、さらに、問合せ配列用ハッシュテーブルのデータを更新するハッシュテーブル更新手段を有することが好ましい。このハッシュテーブル更新手段は、例えば、前記選択手段により、1つの問合せ配列に対して、最も高い相同性を示す同じ類似度の候補配列が2つ以上選択された際、前記問合せ配列用ハッシュテーブルのデータから、前記問合せ配列と、それに対して選択された前記2つ以上の候補配列を削除する機能を有する。前記問合せ配列について、最も高い相同性を示す同じ類似度の候補配列が2以上選択された場合、この問合せ配列は、これらの候補配列に対して特異的に相同するものではないと判断できる。したがって、対象配列のこれらの部分配列に特異的に相同する他の問合せ配列を検索するには、前述のようにハッシュテーブルから、特異性を示さなかった問合せ配列のデータを削除することが好ましい。これによって、検索効率をさらに向上することができる。

[0049] ＜ネットワーク型相同性検索システム＞

本発明の第2の相同性検索システムは、以下に示す端末とサーバーとを有するシステム、すなわちネットワーク型相同性検索システムであってもよい。なお、特に示さない限りは、前述の相同性検索システムと同様である。

[0050] すなわち、本発明の相同性検索システムは、核酸塩基配列からなる問合せ配列の配列情報を用いて、核酸塩基配列からなるゲノムスケールの対象配列の配列情報から、前記問合せ配列と相同する対象部分配列を検索する相同性検索システムであって、端末とサーバーとを有し、

前記端末およびサーバーは、システム外の通信網を介して接続可能であり、

前記端末は、

前記端末内の情報を前記通信網を介して前記サーバーに送信する端末側送信手段と;

前記サーバーから送信された情報を前記通信網を介して受信する端末側受信手段と;

前記端末内の情報を表示する表示手段と;

前記問合せ配列の配列情報を取得する取得手段とを有し、

前記サーバーは、

前記サーバー内の情報を前記通信網を介して前記端末に送信するサーバー側送信手段と;

前記端末から送信された情報を前記通信網を介して受信するサーバー側受信手段と;

対象配列が蓄積されている対象配列データベースと、

前記対象配列データベースにおける対象配列と、前記サーバー側受信手段により受信した前記問合せ配列とについて、それぞれ、同一塩基が2個以上連続したホモポリマー領域を前記塩基1個に置き換えた圧縮問合せ配列と圧縮対象配列とを準備する圧縮配列準備手段と;

前記圧縮問合せ配列と前記圧縮対象配列とを対比し、前記圧縮対象配列において前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索して、絞り込んだ前記圧縮対象部分配列を候補配列の圧縮配列(圧縮候補配列)として選択する検索手段と;

前記圧縮問合せ配列と前記検索手段で選択した前記圧縮候補配列とについて、それぞれの圧縮前の配列における同一塩基の連続数の情報を準備する連続数準備

手段と;

前記同一塩基の連続数の情報に基づいて、前記圧縮問合せ配列と前記圧縮候補配列との間で、対応する塩基ごとに、前記塩基の連続数を対比し、前記連続数の一致度または不一致度から、前記候補配列の前記問合せ配列に対する相同性を示す類似度を演算する類似度演算手段と;

前記類似度演算手段により演算した類似度に基づいて、前記問合せ配列と相対的に相同性が高い任意数の候補配列を順位付けて選択する選択手段とを有し;

前記問合せ配列の情報が、前記端末側送信手段から前記サーバー側受信手段に送信され、かつ、前記サーバーの前記選択手段により選択した前記任意数の候補配列の情報が、前記サーバー側送信手段から前記端末側受信手段に送信され、前記端末において、受信した前記任意数の候補配列の情報が、前記表示手段により表示される。

[0051] 第2の相同性検索システムにおいて、各手段は、例えば、前述の第1の相同性検索システムと同様である。例えば、前記配列情報の取得手段は、例えば、前述の第1の相同性検索システムと同様に、入力手段でもよいし、問合せ配列が記憶されている記憶手段であってもよい。

[0052] ＜サーバー＞

本発明のサーバーは、本発明の第2の相同性検索システムに用いるサーバーである。なお、特に示さない限りは、前述の第1の相同性検索システムと同様である。

[0053] 本発明のサーバーは、

前記サーバー内の情報を前記通信網を介して端末に送信するサーバー側送信手段と;

前記端末から送信された情報を前記通信網を介して受信するサーバー側受信手段と;

対象配列が蓄積されている対象配列データベースと;

前記対象配列データベースにおける対象配列と、前記サーバー側受信手段により受信した前記問合せ配列とについて、それぞれ、同一塩基が2個以上連続したホモポリマー領域を前記塩基1個に置き換えた圧縮問合せ配列と圧縮対象配列とを準備す

る圧縮配列準備手段と;

前記圧縮問合せ配列と前記圧縮対象配列とを対比し、前記圧縮対象配列において前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索して、絞り込んだ前記圧縮対象部分配列を候補配列の圧縮配列(圧縮候補配列)として選択する検索手段と;

前記圧縮問合せ配列と前記検索手段で選択した前記圧縮候補配列とについて、それぞれの圧縮前の配列における同一塩基の連続数の情報を準備する連続数準備手段と;

前記同一塩基の連続数の情報に基づいて、前記圧縮問合せ配列と前記圧縮候補配列との間で、対応する塩基ごとに、前記塩基の連続数を対比し、前記連続数の一致度または不一致度から、前記候補配列の前記問合せ配列に対する相同性を示す類似度を演算する類似度演算手段と;

前記類似度演算手段により演算した類似度に基づいて、前記問合せ配列と相対的に相同性が高い任意数の候補配列を順位付けて選択する選択手段とを有する。なお、サーバーにおいて、各手段は、前述のシステムにおける手段と同様である。

[0054] ＜端末＞

本発明の端末は、本発明の第2の相同性検索システムに用いる端末である。なお、特に示さない限りは、前述の第1の相同性検索システムと同様である。

[0055] 本発明の端末は、

前記端末内の情報を前記通信網を介して前記サーバーに送信する端末側送信手段と;

前記サーバーから送信された情報を前記通信網を介して受信する端末側受信手段と;

前記端末内の情報を表示する表示手段と;

前記問合せ配列の配列情報を取得する取得手段とを有し、

前記問合せ配列の情報が、前記端末側送信手段から前記サーバー側受信手段に送信され、かつ、前記サーバーの前記選択手段により選択した前記任意数の候補配列の情報が、前記サーバー側送信手段から前記端末側受信手段に送信され、前記

端末において、受信した前記任意数の候補配列の情報が、前記表示手段により表示される。

[0056] ＜相同性検索装置＞

本発明の相同性検索装置は、核酸塩基配列からなる問合せ配列の配列情報を用いて、核酸塩基配列からなるゲノムスケールの対象配列の配列情報から、前記問合せ配列と相同する部分配列を検索する相同性検索装置であって、本発明の相同性検索システムを含む。前記相同性検索装置は、例えば、前記問合せ配列および対象配列の配列情報を取得する取得部と;取得された前記問合せ配列および前記対象配列について、それぞれ、同一塩基が2個以上連続したホモポリマー領域を前記塩基1個に置き換えた圧縮問合せ配列および圧縮対象配列を準備する圧縮配列準備部と;前記圧縮問合せ配列と前記圧縮対象配列とを対比し、前記圧縮対象配列において前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索して、絞り込んだ前記圧縮対象部分配列を候補配列の圧縮配列(圧縮候補配列)として選択する検索部と;前記圧縮問合せ配列と前記検索手段で選択した前記圧縮候補配列とについて、それぞれの圧縮前の配列における同一塩基の連続数の情報を準備する連続数準備部と;前記同一塩基の連続数の情報に基づいて、前記圧縮問合せ配列と前記圧縮候補配列との間で、対応する塩基ごとに、前記塩基の連続数を対比し、前記連続数の一致度または不一致度から、前記候補配列の前記問合せ配列に対する相同性を示す類似度を演算する類似度演算部と;前記類似度演算手段により演算した類似度に基づいて、前記問合せ配列と相対的に相同性が高い任意数の候補配列を順位付けて選択する選択部と;前記選択手段により選択した前記任意数の候補配列の情報を出力する出力部とを有する。また、前述のシステムと同様に、対象配列の配列情報を記憶する、または、記憶している対象配列記憶部、問合せ配列の配列情報を記憶する、または、記憶している問合せ配列記憶部、配列情報の入力部、前記各記憶部に情報を更新するための情報更新部等を備えてもよい。

[0057] ＜相同性検索方法＞

本発明の相同性検索方法は、核酸塩基配列からなる問合せ配列の配列情報を用いて、核酸塩基配列からなるゲノムスケールの対象配列の配列情報から、前記問合

せ配列と相同する部分配列を検索する相同性検索方法であって、

前記問合せ配列および対象配列の配列情報を取得する取得ステップと;

取得された前記問合せ配列と前記対象配列について、それぞれ、同一塩基が2個以上連続したホモポリマー領域を前記塩基1個に置き換えた圧縮問合せ配列と圧縮対象配列とを準備する圧縮配列準備ステップと;

前記圧縮問合せ配列と前記圧縮対象配列とを対比し、前記圧縮対象配列において前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索して、絞り込んだ前記圧縮対象部分配列を候補配列の圧縮配列(圧縮候補配列)として選択する検索ステップと;

前記圧縮問合せ配列と前記検索ステップで選択した前記圧縮候補配列とについて、それぞれの圧縮前の配列における同一塩基の連続数の情報を準備する連続数準備ステップと;

前記同一塩基の連続数の情報に基づいて、前記圧縮問合せ配列と前記圧縮候補配列との間で、対応する塩基ごとに、前記塩基の連続数を対比し、前記連続数の一致度または不一致度から、前記候補配列の前記問合せ配列に対する相同性を示す類似度を演算する類似度演算ステップと;

前記類似度演算ステップにより演算した類似度に基づいて、前記問合せ配列と相対的に相同性が高い任意数の候補配列を順位付けて選択する選択ステップと;

前記選択ステップにより選択した前記任意数の候補配列の情報を出力する出力ステップとを有する。

[0058] 本発明の相同性検索方法は、問合せ配列と対象配列とについて、前述のように、圧縮配列および同一塩基連続数を対比して相同性を検索することが特徴であり、その他の条件や構成は特に制限されない。

[0059] 前記取得ステップにより取得される配列情報は、前述の相同性検索システムの前記取得ステップにより取得される配列情報と同様であり、対象配列および問合せ配列の圧縮前の配列情報があげられる。また、前記配列情報として、前記圧縮前の配列に代えて、または、加えて、例えば、圧縮配列、同一塩基の連続数等の情報が含まれてもよい。前記取得ステップにより、配列情報として圧縮配列を取得する場合、例

えば、前記取得ステップと、圧縮配列を準備する前記圧縮配列準備ステップは、同じステップともいえる。また、前記取得ステップにより、配列情報として同一塩基の連続数の情報が取得される場合、例えば、前記取得ステップと、同一塩基の連続数の情報を準備する前記連続数準備ステップは、同じステップともいえる。また、問合せ配列および対象配列のいずれか一方の配列についてのみ、圧縮配列や同一塩基の連続数等の配列情報を取得し、他方の配列については、圧縮前の配列情報を取得してもよい。

[0060] 前記配列情報の取得ステップは、例えば、配列情報を入力する入力ステップがあげられる。また、配列情報が記憶された記憶手段(例えば、データベース)から配列情報を呼び出す、呼び出しステップであってもよい。対象配列および問合せ配列は、両方が入力ステップにより入力されてもよいし、いずれか一方が入力ステップにより入力されてもよい。また、対象配列および問合せ配列の両方が、呼び出しステップにより記憶手段から呼び出されてもよいし、いずれか一方が呼び出しステップにより記憶手段から呼び出され、他方は、入力ステップにより入力されてもよい。本発明においては、前記取得ステップが、例えば、前記問合せ配列および対象配列の両方を入力する入力ステップである形態、前記問合せ配列を入力する入力ステップと、前記対象配列が記憶されている対象配列記憶手段から前記対象配列の配列情報を呼び出す呼び出しステップとを有する形態、前記対象配列を入力する入力ステップと、前記問合せ配列が記憶されている問合せ配列記憶手段から前記問合せ配列の配列情報を呼び出す呼び出しステップとを有する形態、問合せ配列が記憶されている問合せ配列記憶手段および対象配列が記憶されている対象配列記憶手段から、それぞれ問合せ配列および対象配列を呼び出す呼び出しステップである形態等が例示できる。

[0061] 本発明の相同性検索方法においては、さらに、問合せ配列の配列情報を記憶する問合せ配列記憶ステップおよび／または対象配列の配列情報を記憶する対象配列記憶ステップを有することが好ましい。配列情報は、例えば、前述のような問合せ配列記憶手段や対象配列記憶手段に記憶することが好ましい。記憶する配列情報は、例えば、入力ステップにより入力された配列情報や、後述する圧縮変換ステップにより圧縮変換された圧縮配列の配列情報、連続数準備ステップで準備された同一塩基

の連続数の情報等があげられる。

[0062] 本発明の相同性検索方法においては、前記問合せ配列および対象配列における同一塩基の連続数は、それぞれ、取得された配列情報に基づいて得てもよいし、予め配列情報として前述のような記憶手段に記憶されている場合には、前述の各記憶手段から呼び出してもよい。すなわち、本発明の相同性検索方法において、前記連続数準備ステップとしては、例えば、取得した問合せ配列および／または対象配列の配列情報に基づいて、問合せ配列および／または対象配列について、それぞれの圧縮前の配列における同一塩基の連続数を計数（演算）する計数ステップ（連続数計数（演算）ステップ）があげられる。また、前記連続数準備ステップは、前述の問合せ配列記憶手段および／または対象配列記憶手段にそれぞれの連続数の情報が記憶されている場合、問合せ配列の情報が記憶されている問合せ配列記憶手段および／または対象配列の情報が記憶されている対象配列記憶手段から呼び出すステップであってもよい。後者の場合、検索対象とする所望の配列を指定することで、各記憶手段からそれぞれの連続数の情報を呼び出すことができる。

[0063] 本発明の相同性検索方法においては、前記圧縮問合せ配列および圧縮対象配列は、それぞれ、取得された配列情報に基づいて得てもよいし、予め配列情報として前述のような記憶手段に記憶されている場合には、前述の各記憶手段から呼び出してもよい。すなわち、本発明の相同性検索方法において、前記圧縮配列準備ステップとしては、例えば、取得した問合せ配列および／または対象配列の配列情報に基づいて、圧縮問合せ配列および／または圧縮対象配列に圧縮変換する圧縮変換ステップがあげられる。また、前記圧縮配列準備ステップは、前述の問合せ配列記憶手段および／または対象配列記憶手段にそれぞれの圧縮配列が記憶されている場合、圧縮問合せ配列が記憶されている問合せ配列記憶手段および／または圧縮対象配列が記憶されている対象配列記憶手段から呼び出すステップであってもよい。後者の場合、検索対象とする所望の配列を指定することで、各記憶手段からそれぞれの圧縮配列を呼び出すことができる。

[0064] 前記検索ステップは、特に制限されず、ハッシュ検索ステップ、二分木検索ステップ、B木検索ステップ等があげられる。前記ハッシュ検索ステップは、例えば、前記圧縮

問合せ配列と前記圧縮対象部分配列群の各圧縮対象部分配列とをキーとし、同じハッシュ関数を用いて、ハッシュ検索を行うことにより、前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索する。

[0065] 本発明の相同性検索方法においてハッシュ検索を行う場合、例えば、さらに、前記圧縮対象部分配列群の各圧縮対象部分配列をキーとし、同じハッシュ関数を用いて、対象配列用ハッシュテーブルを生成する対象配列用ハッシュテーブル生成ステップを有することが好ましい。この場合、前記検索ステップは、ハッシュ検索ステップであり、例えば、前記圧縮問合せ配列をキーとし、前記対象配列用ハッシュテーブル生成ステップと同じハッシュ関数を用いて、前記対象配列用ハッシュテーブル生成ステップで生成した前記対象配列用ハッシュテーブルのハッシュ検索を行うことにより、前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索する。

[0066] また、ハッシュ検索を行う場合、さらに、2つ以上の前記圧縮問合せ配列をキーとし、同じハッシュ関数を用いて、問合せ配列用ハッシュテーブル生成する問合せ配列用ハッシュテーブル生成ステップを有することが好ましい。この場合、前記検索ステップは、ハッシュ検索ステップであり、例えば、前記圧縮対象部分配列群の各圧縮対象部分配列をキーとし、前記問合せ配列用ハッシュテーブル生成ステップと同じハッシュ関数を用いて、前記問合せ配列用ハッシュテーブル生成ステップで生成した前記問合せ配列用ハッシュテーブルのハッシュ検索を行うことにより、前記各圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索する。

[0067] 前記類似度演算ステップは、前記対応する塩基ごとの連続数の不一致度をペナルティスコアとし、前記対応する塩基ごとのペナルティスコアを加算することによって、類似度を演算することが好ましい。ただし、前記圧縮問合せ配列の上流末端塩基または下流末端塩基における圧縮前の連続数が、前記圧縮候補配列の上流末端塩基または下流末端塩基における圧縮前の連続数よりも短い不一致を除くことが好ましい。

[0068] 本発明の相同性検索方法は、前記選択ステップにおいて、前記類似度演算ステップにより、新たな候補配列の前記問合せ配列に対する類似度が演算された際、前記新たな類似度と、先立っての前記選択ステップで選択された前記任意数の候補配列の前記問合せ配列に対する類似度とに基づいて、前記各候補配列から、再度、任意

数の候補配列を選択することが好ましい。このような場合、例えば、さらに、前記問合せ配列と前記選択ステップにより選択した任意数の候補配列との情報を記憶する記憶ステップを有することが好ましい。前記記憶ステップにおいて、例えば、前述のような記憶手段に、選択された任意数の候補配列の問合せ配列に対する類似度を記憶しておくことで、先立っての類似度と新たな類似との間で、再度、候補配列の選択を行うことが容易となる。

[0069] 本発明の相同性検索方法を用いて、問合せ配列が、対象配列のある部分配列にのみ特異的に相同するか否かを検索する場合には、さらに、問合せ配列用ハッシュテーブルのデータを更新するハッシュテーブル更新ステップを有することが好ましい。前記ハッシュテーブル更新ステップは、例えば、前記選択ステップにより、1つの問合せ配列に対して、最も高い相同性を示す同じ類似度の候補配列が2つ以上選択された際、前記問合せ配列用ハッシュテーブルのデータから、前記問合せ配列と、それに対して選択された2つ以上の候補配列を削除する。

[0070] ＜コンピュータプログラム＞

本発明のコンピュータプログラムは、本発明の相同性検索方法をコンピュータ上で実行可能なコンピュータプログラムである。

[0071] ＜電子媒体＞

本発明の電子媒体は、本発明のコンピュータプログラムを格納した電子媒体である。前記電子媒体は、コンピュータ読み取り可能な媒体であり、例えば、記録媒体である。

[0072] （実施形態1）

ハードウェアの構成

本発明の相同性検索装置について、ハードウェア構成の概略を説明する。なお、以下に示す構成は、一例であり、本発明はこれに制限されない。

[0073] 図1は、本発明の相同性検索装置のハードウェア構成の一例を示すブロック図である。同図において、相同性検索装置1は、CPU101、RAM102、記憶手段（記憶装置）103、入出力I／F（インターフェイス）105、表示手段（ディスプレイ）106、入力手段（入力装置）107、通信デバイス108、および、ドライブ109を備えている。RAM1

02、記憶装置103および入出力I／F（インターフェイス）105は、通信バス104によって、CPU101に接続されている。入出力I／F（インターフェイス）105には、ディスプレイ106、入力装置107、通信デバイス108およびドライブ109が接続されている。

[0074] CPU101は、相同性検索装置1の全体の制御を担う。RAM102は、コンピュータのメインメモリであり、CPU101のワークメモリである。記憶装置103は、例えば、ROM、HDD、HD等である。ROMは、読み込みのみのメモリーであり、動作プログラムを格納している。HDDは、CPU101の制御下、HDに対するデータの読み込みと書き込みを制御し、HDは、HDDの制御下で書き込まれたデータを記憶する。ドライブ109は、リムーバブル記録媒体用のドライブであり、CPU101の制御下、リムーバブル記録媒体に対するデータの読み込み／書き込みを制御する。リムーバブル記録媒体としては、例えば、FD、CD－ROM（CD－R、CD－RW）、MO、DVD、メモリーカード等が使用でき、これらの記録媒体は、ドライブ109の制御で書き込まれたデータを記憶する。通常、RAM102は、主記憶装置であり、ROM、HDおよびFD等の外部記録媒体が補助記憶装置となる。本発明においては、CPU101により、例えば、本発明のコンピュータプログラムやその他のプログラムが実行され、また、各種情報の読み込みや書き込みが行われる。図1の相同性検索装置1には、一例として、各種ソフトウェア（配列圧縮ソフトウェア111、検索システムソフトウェア112）を格納するプログラム格納部110と、情報を格納する情報格納部113とを備える形態を示す。これらの格納部は、例えば、前述の補助記憶装置に確保された一定の領域に設けられる。図1においては、プログラム格納部110および情報格納部113を、記憶装置103に確保された記憶領域として示す。これらのソフトウェアは、例えば、CPU101によってRAM102上に呼び出され、OS（オペレーションシステム）と共同して実行されることで、それらの機能を実現する。前記配列圧縮ソフトウェア111は、問合せ配列や対象配列を圧縮変換するためのプログラムであり、前記検索システムソフトウェア112は、圧縮変換以外の本発明における処理を実行するプログラムである。また、これらのプログラムは、本発明のプログラムとして単一のソフトウェアであってもよい。

[0075] ディスプレイ106は、文書等の各種情報を表示し、例えば、LEDディスプレイ、液晶ディスプレイ等がある。I／F（インターフェイス）105は、通信デバイス108を通じて、L

ANやインターネット等の外部ネットワークに接続され、前記外部ネットワークを介して、他のサーバーや情報処理装置に接続される。本発明においては、例えば、ゲノムや遺伝子の核酸塩基配列データを備える外部データベース(DB)に接続される。I／F(インターフェイス)105は、前記ネットワークと内部とのインターフェイスを担い、他のサーバー等からのデータの入出力を制御する。通信デバイス108は、例えば、モデム等である。入力装置107としては、例えば、キーボードやマウス等があり、これによって文字、数字、各種指示、カーソルの移動等を行う。また、これらの構成部の他に、例えば、スキャナやプリンタ等を備えてもよい。前記スキャナは、例えば、文章等の画像情報を光学的に読み取り、画像データとして取り込むことができる。また、プリンタを備えてもよく、これは各種情報を印刷する。

[0076] (実施形態2)

本発明の第1の相同性検索システムおよび第2のネットワーク型相同性検索システムの構成について、それぞれ一例を示す。

[0077] 第1のシステム構成例

図8に、本発明のシステムの構成の一例であるスタンドアローン型の全体構成図を示す。図8に示すシステムは、本発明の相同性検索システム1から構成され、相同性検索システム1は、データ入出力部12と相同性検索部13から構成される。相同性検索部13は、例えば、配列情報の取得部、圧縮配列を準備する圧縮配列準備部(例えば、圧縮変換部)、圧縮候補配列の検索部、同一塩基の連続数準備部(例えば、連続数計数部)、類似度演算部、候補配列の選択部を有する。図10に、スタンドアローン型の相同性検索装置のハードウェア構成の一例を示す。図示のように、相同性検索システム1は、データ入出力部12、相同性検索部13および記憶装置37から構成されている。前記データ入出力部12は、各ステップをコンピュータにより実行可能なプログラムを実行するCPU31、入出力I／F(インターフェイス)32、データの入力を行う入力装置33、データの出力を行う出力装置34を有するコンピュータ機器で構成される。相同性検索部13は、プログラムが格納されたプログラム格納部36およびプログラムを実行するCPU35を有するコンピュータ機器で構成される。記憶装置37は、例えば、問合せ配列および対象配列の圧縮前の配列情報、圧縮後の配列情報

、同一塩基の連続数の情報、類似度、候補配列の順位等のデータが記憶される。なお、データ入出力部12、相同性検索部13、記憶装置37は、あくまでも機能上のものであり、例えば、1台のコンピュータ機器で一体に構成してもよいし、複数台のコンピュータ機器で個別に構成してもよい。

[0078] 第2のシステム構成例

図9に、サーバーで処理するネットワーク型のシステムの全体構成図を示す。図9に示すように、本実施形態の相同性検索システム2は、端末21、および、サーバーシステム24から構成される。端末21は、データ入出力部22から構成される。サーバーシステム24は、相同性検索部23と対象配列が蓄積されたデータベース(対象配列DB)25とから構成される。相同性検索部23は、例えば、圧縮配列準備部(例えば、圧縮変換部)、圧縮候補配列の検索部、同一塩基の連続数準備部(例えば、連続数計数部)、類似度演算部、候補配列の選択部を有する。相同性検索部23とサーバーシステム24は、例えば、TCP(Transmission Control Protocol)／IP(Internet Protocol)に基づくインターネットとして機能する公衆網や専用線等の通信回線100を介して接続されている。図11に、前記ネットワーク型システムの装置の構成の一例を示す。端末21は、データ入出力部22および通信インターフェイス47から構成され、通信インターフェイス47を介して通信回線に接続されている。データ入出力部22は、プログラムを実行するCPU41、入出力I／F42、データの入力を行う入力装置43およびデータの出力を行う出力装置44から構成される。前記データ入出力部22および通信インターフェイス47は、あくまでも機能上のものであり、例えば、1台のコンピュータ機器で一体に構成してもよいし、複数台のコンピュータ機器で個別に構成してもよい。サーバーシステム24は、相同性検索部23、対象配列が蓄積されたデータベース(対象配列DB)25および通信インターフェイス48から構成され、通信インターフェイス48を介して通信回線に接続されている。相同性検索部23は、候補配列を選択するための一連のステップを実行可能なプログラムを実行するCPU45および前記プログラムが格納されたプログラム格納部46で構成される。相同性検索部23、対象配列DB25および通信インターフェイス48は、あくまでも機能上のものであり、例えば、1台のコンピュータ機器で一体に構成してもよいし、複数台のコンピュータ機器で個別に構成し

てもよい。

[0079] （実施形態3）

本発明の相同性検索システムの一例について、以下に説明する。図2は、本実施形態における相同性検索システムの構成の概略を示すブロック図である。なお、本発明は、この実施形態には制限されず、発明の要旨を変更しない範囲で種々変形可能である。

[0080] 図2に示すように、本実施形態の相同性検索システムは、配列情報の取得手段（入力手段）201、圧縮配列準備手段202、圧縮候補配列の検索手段203、同一塩基の連続数準備手段204、類似度演算手段205、候補配列の選択手段206、情報記憶手段207および出力手段208を備える。この相同性検索システムは、例えば、前述のようなハードウェア構成のコンピュータシステムによって構築された相同性検索装置があげられる。また、各構成手段は、例えば、コンピュータのCPUが所定のプログラムを実行することによって実現される機能的ブロックであればよい。このため、例えば、各構成手段が、ハードウェアとして実装されてなくともよく、前述のようなネットワークシステムであってもよい。

[0081] 配列情報の取得手段201は、対象配列の核酸塩基配列および問合せ配列の核酸塩基配列の配列情報を取得する機能を有する。この情報の取得は、例えば、前述のような入力装置による入力によって行うことができる。また、前述のように、インターネット等の外部ネットワークにアクセスして、外部データベース等から入手してもよい。外部ネットワークから前記情報を入手する場合、例えば、データベースの情報を、コンピュータの記憶装置（例えば、図1におけるRAM102や情報格納部113等）にダウンロードしてもよいし、通信回線を接続した状態で、前述の情報を利用することもできる。情報元となる外部データベースは、制限されない。また、配列情報が記憶されたリムーバブル記録媒体を利用することもできる。

[0082] 圧縮配列準備手段202は、例えば、核酸塩基配列の配列情報を、同一塩基が2個以上反復したホモポリマー領域を前記塩基1個に置き換えた圧縮配列に変換する機能を有する（圧縮変換手段）。すなわち、この圧縮配列準備手段202により、配列情報の取得手段201によって取得した対象配列および問合せ配列について、同一塩

基が2個以上反復したホモポリマー領域を前記塩基1個に置き換えた圧縮配列の配列情報が作成される。

[0083] ここで、核酸塩基配列の圧縮変換の一例を、図4を用いて説明する。図4は、核酸塩基配列の圧縮変換ならびに後述する塩基の連続数の計数の概略を示す図である。同図において、D1が核酸塩基配列である。この核酸塩基配列においては、同じ塩基が連続して並んでいる領域が存在する。具体的に、左端(5'側)から、6個のアデニンが連続した領域、8個のチミンが連続した領域、7個のグアニンが連続した領域、1個のチミンをおいて、さらに、9個のシトシンが連続した領域、3個のアデニンが連続した領域を有している。このように同じ塩基が連続している領域が、それぞれ、本発明における「ホモポリマー領域」である。圧縮変換では、このように同じ塩基が複数個(2個以上)連続している(反復している)領域については、塩基を1個と見なして、4種類の塩基について種別の並びのみを示す配列とする。図4において、D2で示す配列が、D1で示す核酸塩基配列に対する圧縮配列となる。

[0084] 前述のように、一般的な相同性検索では、ゲノムスケールの対象配列は、通常、部分配列に分解してから検索が行われるため、本発明においても、対象配列は、部分配列に区切って検索にかけることが好ましい。したがって、本発明における対象配列の圧縮配列は、全長の圧縮配列でもよいが、前述のような圧縮対象部分配列群であることが好ましい。前記対象配列からの部分配列の作成は、限定されず、従来公知の手法をとることができる。具体例としては、例えば、圧縮後の対象配列の先頭から、順次、1塩基ずつずらして圧縮対象部分配列群を作成する。すなわち、本実施形態のシステムでは、対象配列の配列情報を取得手段201で取得すると、圧縮配列準備手段202において、圧縮された対象配列の配列情報を作成でき、さらに、これを固定長に区切った圧縮対象部分配列からなる圧縮対象部分配列群の配列情報を作成できる。また、対象配列の相補鎖についても検索を行う場合には、例えば、前記圧縮対象部分配列群の配列情報と、それらの相補鎖の圧縮対象部分配列群の情報を、交互に取得すればよい。後者の圧縮対象部分配列群の配列情報は、前者の圧縮対象部分配列の塩基の並びから、相補塩基について逆列とすることにより容易に決定できる。また、後述する塩基の連続数の情報については、例えば、前者の各ホモポリマ

一領域における同一塩基の連続数を順に示す列を、相補塩基について逆列とすることにより取得できる。例えば、前者の連続数の並びが「6－8－7－1－9－3」であれば、後者の連続数の並びは、「3－9－1－7－8－6」となる。

[0085] 圧縮候補配列の検索手段203は、以下の機能を有する。すなわち、まず、圧縮配列準備手段202において作成した対象配列の圧縮配列（圧縮対象配列）と問合せ配列の圧縮配列（圧縮問合せ配列）とを対比させる。そして、前記圧縮対象配列において、前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索して、絞り込んだ前記圧縮対象部分配列を、候補配列の圧縮配列（圧縮候補配列）として選択する。

[0086] 同一塩基の連続数準備手段204は、例えば、圧縮候補配列の検索手段203で選択した前記圧縮候補配列および前記圧縮問合せ配列について、前記ホモポリマー領域における同一塩基の連続数を計数する機能を有する（連続数計数手段）。なお、前述のように、システム外で計数された連続数の情報を入力する形態であってもよい。

[0087] ここで、同一塩基の連続数の計数の一例を、前述の図4を用いて説明する。前述のように、同図において、D1が圧縮前の核酸塩基配列であり、D2が前記核酸塩基配列を圧縮変換した圧縮配列である。この圧縮配列D2における各塩基の連続回数が、本発明における前記連続数にあたる。なお、本実施形態においては、ホモポリマー領域の塩基数を連続数として計数（1個目の塩基も換算）し、反復していない塩基は1個として計数している。この同一塩基の連続数の情報は、例えば、同図に示すように、同一塩基連続数の列D3「687193」として表すことができる。圧縮配列D2「ATGTCA」の要素数と、同一塩基連続数の列D3「687193」における要素数は共通し、圧縮前の核酸塩基配列D1の要素数は、圧縮配列D2の要素数とを比較すると、ホモポリマー領域が出現する分短くなる。

[0088] 類似度演算手段205は、同一塩基の連続数準備手段204により計数した前記連続数の情報に基づいて、前記圧縮問合せ配列と前記圧縮候補配列との間で、対応する塩基ごとに、前記塩基の連続数を対比し、前記連続数の一致度または不一致度から、前記候補配列の前記問合せ配列に対する相同性を示す類似度を演算する機

能を有する。

[0089] 候補配列の選択手段206は、類似度演算手段205より演算された前記類似度結果を比較することにより、前記候補配列について、前記問合せ配列に対する相同性の順位付けを行い、相対的に相同性が高い任意数の候補配列を選択する機能を有する。

[0090] 情報記憶手段207は、候補配列の選択手段206で選択した前記問合せ配列に対して相対的に相同性が高い任意数の候補配列と、問合せ配列との間における前記類似度とを記憶する機能を有する。特に、圧縮候補配列の検索手段203において複数の候補配列が検索される場合は、例えば、任意数まで相同性が高い候補配列の類似度を記憶し、さらに、新たな類似度が演算された際に、記憶した複数の類似度と新たな類似度とを比較して、再度順位付けを行い、相同性が高い任意数の候補配列を選択し、これらの情報を記憶することが好ましい。これによって、問合せ配列との相同性がより高い対象配列の部分配列をさらに検索していくことが可能となる。

[0091] 出力手段208は、情報記憶手段207に記憶された情報を出力する機能を有する。出力する情報は、例えば、相同性検索システムがディスプレイ等の表示手段(表示装置)を備える場合には、その表示画面に表示されてもよいし、また、プリンタによりプリントアウトすることによって外部に表示されてもよい。また、例えば、コンピュータの記憶装置(例えば、情報格納部)や、リムーバブル記録媒体に出力され、格納されてもよい。

[0092] つぎに、図5を用いて、本実施形態の相同性検索システムにおける処理の流れの一例を説明する。図5は、前記処理の流れを示すフローチャートである。この処理は、本発明の相同性検索方法の一例であり、例えば、本発明の相同性検索システムや本発明のコンピュータプログラムにより実行できる。なお、対象配列の圧縮配列(圧縮対象配列)は、前述した圧縮対象部分配列群として説明する。

[0093] まず、問合せ配列ごとの結果を格納する結果格納領域の初期化(ステップM0)から開始される。ついで、問合せ配列ならびに対象配列の配列情報が取得(入力)される(ステップM1)。例えば、直接的にマウス等の入力装置107から配列情報を入力したり、あるいは、間接的に対象配列が格納されたファイルやネットワーク上の場所を特

定することで、システム（例えば、RAM102や記憶装置103等）に取り込んだ配列情報に対して、後述する処理が開始される。

[0094] つぎに、取得された問合せ配列の配列情報を、圧縮配列に圧縮変換し（ステップM2）、また、取得された対象配列の配列情報を、圧縮配列（圧縮対象部分配列群）に圧縮変換する（ステップM3）。問合せ配列および対象配列の圧縮変換（ステップM2およびM3）は、例えば、順不同で別個に行ってもよいし、並行して行うこともできる。問合せ配列が複数の場合、後述するステップを実行する前に、全ての問合せ配列についての圧縮変換を完了してもよいし、圧縮変換した配列から順次、後述するステップに供することで、このステップM2と後述ステップを並行して行ってもよい。また、対象配列についても、後述するステップを実行する前に、対象部分配列群に含まれる全対象部分配列の圧縮変換を完了してもよいし、圧縮変換した配列から順次、後述するステップに供することで、このステップM3と後述ステップを並行して行ってもよい。なお、問合せ配列が複数の場合（特に大量である場合）には、ハッシュテーブルを作成することが好ましい。ハッシュテーブルを利用すれば、例えば、約1,000,000～5,000,000個の問合せ配列であっても、迅速な処理が可能となる。また、対象配列についても、ゲノムスケールであることから、圧縮対象部分配列群についてハッシュテーブルを作成することが好ましい。これらのハッシュテーブルの作成については、後述する。

[0095] 続いて、問合せ配列の圧縮配列（圧縮問合せ配列）と対象配列の圧縮対象部分配列群とを対比し、圧縮問合せ配列と一致する圧縮対象部分配列の検索を行う（ステップM4）。そして、圧縮問合せ配列と一致する圧縮対象部分配列を検索できれば、その圧縮対象部分配列を候補配列の圧縮対象部分配列（以下、「圧縮候補配列」ともいう）として選択する。他方、圧縮問合せ配列と一致する圧縮対象部分配列が検索できなかった場合は、ステップM11に進む。

[0096] 本発明においては、このように、まず、問題となっていたホモポリマー領域の同一塩基連続数を考慮しない圧縮配列同士を対比し、一致する候補配列を選択している。このため、従来のように、例えば、塩基の連続数が異なるのみで非類似と判断されたり、反対に、塩基の連続数の問題を考慮するあまり、ホモポリマー領域における同一

塩基連続数以外で不一致が生じるものまでも、類似と判断するというような問題が回避できる。少なくとも、塩基の種類の並びに関しては、一致しているため、後は、後述するように、塩基の連続数に関して相同性を判断し、候補配列の相同性の順位(順序)を判断することによって、精度良く相同性検索が行える。

[0097] そして、ステップM4で選択した圧縮候補配列の圧縮問合せ配列に対する相同性を示す類似度を演算する(ステップM5)。類似度の演算にあたっては、圧縮候補配列および圧縮問合せ配列のそれぞれについて、各ホモポリマー領域における同一塩基の連続数の情報が必要である。したがって、このステップM5においては、類似度の演算のために、まず、圧縮候補配列および圧縮問合せ配列について、前記ホモポリマー領域における同一塩基の連続数を計数する。具体的には、前記圧縮候補配列の圧縮前の核酸塩基配列および圧縮問合せ配列の圧縮前の核酸塩基配列から、前記ホモポリマー領域における同一塩基の連続数を計数する。問合せ配列における連続数の計数は、この時点で行ってもよいし、例えば、圧縮変換の際に逐次または並行して行ってもよい。また、対象配列の各部分配列については、前記問合せ配列と同様に、例えば、圧縮変換の際に逐次または並行して行ってもよいが、候補配列についての結果が必要であることから、対象部分配列全てについて計数を行うのではなく、ステップM4で選択された圧縮候補配列についてのみ計数を行ってもよい。

[0098] また、予め、各ホモポリマー領域における同一塩基の連続数を計数しておき、これらの情報を、必要に応じて、例えば、直接入力したり、記憶装置や外部記録媒体等から読み込んでもよい。類似度の演算方法については、後述する。

[0099] つぎに、ステップM5で得られた類似度の結果を、他の類似度と比較する(ステップM6)。相同性検索においては、通常、問合せ配列に対して相同性を示す対象部分配列が複数得られた場合には、相同性についての順序を付けたり、また、最も相同性が高い対象部分配列(すなわち特異性が高い対象部分配列)の選択が行われる。したがって、本発明においても、ステップM4において、圧縮問合せ配列と一致する圧縮候補配列が検索された場合には、各候補配列の問合せ配列に対する相同性を比較することによって、例えば、相同性についての順序付けを行い、相同性が高い任意数の候補配列を選択する。前述のように、選択する候補配列の数は、制限され

ず、任意の数を設定できる。

[0100] このステップM6の次工程は、ステップM6の比較結果に応じて以下のように処理できる。

(1)ステップM6で得られた結果(現類似度)が、初めての類似度の場合は、ステップM8に進み、問合せ配列の結果として、前記現類似度を、初期化された前記問合せ配列用の結果格納領域に記録する(ステップM8)。記録する情報は、問合せ配列に対する候補配列の類似度の他に、例えば、対象配列の種類(例えば、ゲノムの種類、染色体の種類等)、対象配列の順鎖および逆鎖の種類、対象配列における候補配列の座標等があげられる(以下、同様)。

(2)ステップM6で得られた現類似度結果が、2つ目以降に得られた結果であって、候補配列が任意数に到っていない場合、この結果を、さらに前記問合せ配列用の結果格納領域に記録する(ステップM8)。この際、類似度に基づいて、問合せ配列に対する各候補配列の順序付けを行う。

(3)ステップM6で得られた結果(現類似度)が、2つ目以降に得られた結果であって、候補配列が任意数に到っている場合、現類似度とすでに記録された各類似度とに基づいて、問合せ配列に対する各候補配列の順序付けを再度行い、上位から任意数の候補配列を選択し、これらの情報を置換記録する(ステップM8)。選択されなかった候補配列は、ステップM11に進む。

(4)ステップM6で得られた結果(現類似度)が、問合せ配列との相同性を示さない場合、または、極めて相同性が低い結果の場合には、ステップM11に進む。

[0101] また、候補配列が、ある問合せ配列にのみ特異的に相同するか否かを検索する場合には、ステップM6の次工程は、ステップM6の比較結果に応じて以下のように処理できる。

(5)ステップM6で得られた現類似度が、2つ目以降に得られた結果であって、記録された類似度よりも高い相同性を示す類似度の場合(特に、問合せ配列と候補配列とが完全一致)、前記現類似度を、最良の類似度として記録する(ステップM8)。

(6)ステップM6で得られた現類似度が、2つ目以降に得られた結果であって、記録された類似度と同じ類似度の場合、問合せ配列に対して特異性があるか否かの判断

（特異性検索）を行う（ステップM7）。すなわち、現類似度と記録された類似度とが同じ場合、この類似度が最良の類似度（最良の値）か否かを判断する。最良の値とは、例えば、図6を用いて後述するペナルティスコアによる類似度の算出方法に従えば、「0.0（最小値）」であり、これは問合せ配列と候補配列とが完全一致であることを意味する。このように問合せ配列と完全一致である候補配列が2つ以上存在するということは、結果的に、問合せ配列が対象配列に対して特異性を示さないことを示す。したがって、このような場合には、問合せ配列は対象配列に対して特異性なしとの情報を記録する（ステップM10）。問合せ配列はこれらの候補配列に対する特異性がないことが明らかであり、以降の検索にこれを含めることが不要になる。したがって、複数の問合せ配列を検索する場合には、前記問合せ配列のデータを削除してもよい。また、これらの候補配列についても、前記問合せ配列に対する特異性がないことが明らかであり、以降の検索にこれを含めることが不要となる。したがって、特異性がないことを記録した後、前記候補配列を、検索する対象のデータから削除してもよい。このように検索するデータが減少することで、例えば、システムの性能は逆比例してさらに向上する。他方、現類似度と記録された類似度とが同じであっても、前述のような最良の値（最良のスコア）でない場合には、現スコアを、前記問合せ配列用の結果格納領域に記録する（ステップM9）。

[0102] そして、ステップM11において、ある問合せ配列について、準備した対象配列の検索を終了したと判断した場合には、問合せ配列について、記録した類似度やその他の情報を出力し（ステップM13）、対象配列との検索が終了したと判断した場合には、検索を終了する。他方、ステップM11において、ある問合せ配列に対して準備した対象配列との検索が残っている場合には、次の対象配列の圧縮変換（ステップM3）または圧縮変換された圧縮対象部分配列群との対比（ステップM4）に移動する（ステップM12）。

[0103] また、ある問合せ配列について、準備した対象配列の検索が終了した場合、さらに他の問合せ配列が準備されている場合には、他の問合せ配列についても同様の処理を行う。なお、問合せ配列が複数の場合は、ある問合せ配列についてステップM0～M13の一連の処理が終わってから、他の問合せ配列について処理を行ってもよ

いし、逐次または並行して処理を行うこともできる。そして、検索が終了した後に、蓄積した情報（類似度や対象配列における座標等）を、問合せ配列ごとに出力すればよい。

[0104] （実施形態4）

本発明の相同性検索システムのその他の例について、以下に説明する。図3は、本実施形態における相同性検索システムの構成の概略を示すブロック図である。なお、特に示さない限り、前記実施形態3における図2のシステムと同様である。

[0105] この相同性検索システムは、実施形態3のシステムにおける配列情報の取得手段201と圧縮変換手段202に代えて、圧縮配列の取得手段301を備える。このように本実施形態の相同性検索システムでは、予め圧縮変換処理された圧縮問合せ配列や圧縮対象配列の配列情報が取得（入力）されてもよい。

[0106] （実施形態5）

本発明における類似度の算出の一例について、図6を用いて説明する。なお、本発明は、以下の内容には制限されず、本発明の要旨を変更しない範囲で種々変形可能である。

[0107] 図6は、類似度の算出に必要な情報の一例を列挙した図である。同図において、S1～S3は、対象配列に関する情報であり、S4～S6は、問合せ配列に関する情報である。S1は、圧縮前の対象配列であり、S2は、圧縮対象配列、S3は、対象配列の塩基の連続数を示す列である。同様に、S4は、圧縮前の問合せ配列であり、S5は、圧縮問合せ配列であり、S6は、問合せ配列の塩基の連続数を示す列である。

[0108] 本発明の相同性検索システムにおいては、前述のように、まず、対象配列の圧縮配列（圧縮対象配列）と問合せ配列の圧縮配列（圧縮問合せ配列）との対比により、圧縮問合せ配列（S5）と一致する対象配列の圧縮対象部分配列（S2）が検索される。図6に示すように、圧縮問合せ配列（S5）と圧縮対象配列（圧縮対象部分配列S2）は、4種類の塩基が同じ連なりを示している。

[0109] 本実施形態では、類似度の演算方法として、圧縮問合せ配列と圧縮対象部分配列との間で、前記対応する塩基ごとの連続数の不一致度をペナルティスコアとし、前記対応する塩基ごとのペナルティスコアを加算することによって、類似度を演算する例

を説明する。なお、後に説明するが、前記圧縮問合せ配列の上流末端塩基または下流末端塩基における圧縮前の連続数が、前記圧縮候補配列の上流末端塩基または下流末端塩基における圧縮前の連続数よりも短い不一致は除く。このように不一致度を相同性の指標とする場合、例えば、相対的に値が大きいほど相対的に非類似であり、相対的に値が小さいほど相対的に類似であると判断できる。

[0110] 対応する塩基ごとのペナルティスコアを演算するにあたっては、圧縮対象配列(S2)と圧縮問合せ配列(S5)とにおいて、上流側末端塩基のペナルティスコア、下流側末端塩基のペナルティスコア、および、前記両末端以外の内部塩基のペナルティスコアを別途計算する。そして、これらの総和(総和ペナルティー)を、相同性を示す類似度とする。以下、各ペナルティスコアの演算について説明する。

[0111] 図6に示す式S8～S10は、それぞれ、前記上流側末端のペナルティスコアを演算する式(S8)、前記内部の各塩基のペナルティスコアを演算する式(S9)、前記下流側末端のペナルティスコアを演算する式(S10)である。そして、式S7が、ペナルティの総和を演算する式であり、問合せ配列(S4)と対象配列の部分配列(S1)との相同性を示す指標(類似度)が得られる。なお、図6の式S8～S10において、ホモポリマー数とは、一つのホモポリマー領域における同一塩基の連続数を意味する。式S8において、ホモポリマー数$_1$とは、圧縮対象配列(S2)と圧縮問合せ配列(S5)における対応する1番目の塩基種(上流側末端塩基種)の連続数である。式S9において、ホモポリマー数$_i$とは、前記両圧縮配列(S2、S5)における対応するi番目(iは、n－1であり、nは、3以上の整数)の塩基種の連続数である。式S10において、ホモポリマー数$_n$とは、圧縮対象配列(S2)と圧縮問合せ配列(S5)における対応する最後(n番目、nは、3以上の整数)の塩基種(下流側末端塩基種)の連続数である。

[0112] まず、内部の塩基についてのペナルティスコアを求める式S9について説明する。この式の場合、ある塩基について問合せ配列における連続数と対象配列における連続数とが同じである場合(問合せ配列ホモポリマー数＝対象配列ホモポリマー数)、前者を後者で割った値は1となるため、その自然対数は0である。また、問合せ配列における連続数が、対象配列における連続数よりも多い場合(問合せ配列ホモポリマー数＞対象配列ホモポリマー数)、前者を後者で割った値は1以上となるため、その

自然対数は正の値となる。他方、ある塩基について問合せ配列における連続数が、対象配列における連続数よりも少ない場合（問合せ配列ホモポリマー数＜対象配列ホモポリマー数）、前者を後者で割った値は1以下となり、その自然対数は、負の値となる。つまり、問合せ配列における連続数と対象配列における連続数とが近似する程、自然対数は0に近づき、異なる程、自然対数の値は、0から離れる。したがって、式S9においては、前記自然対数の絶対値をペナルティスコアとしている。

[0113] つぎに、式S8および式S10について説明する。末端塩基についても、前記式S9と同様に、問合せ配列における連続数と対象配列における連続数とが同じである場合（問合せ配列ホモポリマー数＝対象配列ホモポリマー数）、前者を後者で割った値は1となるため、その自然対数は0である。また、問合せ配列における連続数が、対象配列における連続数よりも多い場合（問合せ配列ホモポリマー数＞対象配列ホモポリマー数）、前者を後者で割った値は1以上となり、その自然対数は、正の値となる。しかしながら、問合せ配列における連続数が対象配列における連続数より多い場合、すなわち、対象配列における連続数が少ない場合、末端塩基に関しては、以下のように考える。問合せ配列は、ゲノムや染色体における部分配列であるため、末端のホモポリマー領域については、問合せ配列における連続数が対象配列における連続数よりも少ないことのみをもって、相同性が低いという評価を行うことは妥当ではない。そこで、末端塩基の連続数に関しては、式S9のように絶対値はとらず、対象配列の連続数よりも問合せ配列の連続数が少ない場合、同じ連続数の場合には、ペナルティスコアを0とし、問合せ配列の連続数が大きい場合、すなわち自然対数が正の値となる場合のみを演算する式（S8、S10）としている。

[0114] そして、式S7において、式S8で算出した前記上流側末端のペナルティスコア、式S9で算出した前記内部の各塩基のペナルティスコアの和、および、式S10で算出した前記下流側末端のペナルティスコアを加算して、問合せ配列（S4）と対象配列の部分配列（S1）との相同性を示す指標（類似度）を得る。なお、本例においては、「0」が、最大の一致度、すなわち完全一致を表し、相対的に数値が大きくなるほど類似度は、相対的に低くなる。通常は、上限の閾値、例えば、2の自然対数を設け、これ以上は相同性なしとみなすことができる。

[0115] 式S7～S10は、本発明を実現するための一例であり、これには制限されない。一例として、前述のように対数を用いるのは、誤差のペナルティの累積を加算的に表現するためであって、例えば、各式における分子分母の逆転、正負の逆転、自然対数を取るか指数のまま扱うか、自然対数を使うか対数の底として別の値を使うか、絶対値を取るか自乗を取るかなど、等価な式や応用した式に変形することが可能である。また、本発明の用途に応じて、場所による重み付けを付ける等の変更も可能である。また、このような式の変形や構築は、本明細書の記載に基づけば、当該技術分野における当業者であれば実施可能である。

[0116] 具体例として、例えば、式S7において、絶対値を取る前の正負の差によって、重み付けを与えることで、オーバーコール、アンダーコールの評価を組み込むことができる。ホモポリマー領域における同一塩基の塩基数のエラーについては、例えば、塩基配列の決定方法によって、塩基数が多く計数される、または、塩基数が少なく計数される、という傾向があるという可能性を示唆している。したがって、この傾向を考慮することによって、さらに、ペナルティスコアからエラーによる影響を軽減することができる。例えば、問合せ配列側の同一塩基数が多いという傾向が予めわかっている場合には、以下のような処理が可能である。すなわち、例えば、図6に示す式を用いて各塩基についてのペナルティスコアを算出する際、絶対値を取る前の値が正の場合、つまり、問合せ配列側の連続数が対象配列側の連続数よりも多い場合（問合せ配列ホモポリマー数＞対象配列ホモポリマー数）には、例えば、これに1未満の係数をかける。これにより、問合せ配列の同一塩基数が多いという傾向によって大きくなる見かけのペナルティスコアを、軽減することができる。つまり、「オーバーコールの傾向」を反映して、より信頼性の高い類似度を求めることができる。他方、例えば、問合せ配列側の同一塩基数が少ないという傾向が予めわかっている場合には、以下のような処理が可能である。すなわち、例えば、図6に示す式を用いて各塩基についてのペナルティスコアを算出する際、絶対値を取る前の値が負の場合、つまり、問合せ配列側の連続数が対象配列側の連続数よりも少ない場合（問合せ配列ホモポリマー数＞対象配列ホモポリマー数）には、例えば、これに1未満の係数をかける。これにより、問合せ配列の同一塩基数が少ないという傾向によって大きくなる見かけのペナルテ

ィスコアを、軽減することができる。つまり、「アンダーコールの傾向」を反映して、より信頼性の高い類似度を求めることができる。なお、式S8(head__penalty)と式S10(tail__penalty)は、前述のように、ともに正の数だけ積算されるので、常に、前者のケース(絶対値を取る前の値が正)において、1未満の係数を掛けるだけとなる。

[0117] さらに、塩基の種別(A, G, C, T)に関する同一塩基数の傾向も、前述と同様に重み付けを行うことが可能である。つまり、式S7の部分式である式S9により各塩基についてのペナルティスコアを算出する際、絶対値を取る前の値の正負に関して、塩基種別毎に重み付け計数を変えれば、例えば、塩基種別毎にオーバーコール、アンダーコールの評価傾向を反映することも可能となる。

[0118] (実施形態6)

本発明における対象配列用ハッシュテーブルの一例について、説明する。なお、本発明は、以下の内容には制限されず、本発明の要旨を変更しない範囲で種々変形可能である。

[0119] ハッシュテーブルは、データ構造であり、例えば、圧縮対象部分配列群の各圧縮対象部分配列の文字列にハッシュ関数をかけた値によって、テーブル内の要素に直接インデックス付けされている。文字列に対してハッシュ関数をかけた数値を得る方法は、制限されず、従来公知の手法が採用できる。具体例としては、例えば、プログラミング言語のjavaの中にある、標準的パッケージクラスjava. lang. Stringのhash Codeメソッドに定義されているものを使用することで、容易に実現できる。また、整数空間に写像された値をさらにテーブル要素数で除して、正の剰余を算出することで、必要なインデックスを得ることができる。

[0120] 異なる文字列であっても、値の衝突によって、同一のハッシュテーブルのインデックスに割り振られる場合がある。この場合、例えば、オーバーフロー領域を、対象配列の圧縮後の長さ分確保しておくことが好ましい。そして、例えば、衝突した文字列要素同士を次々に順次指し示すデータ構造をとることによって、例えば、あるハッシュインデックスに対する対象配列内の要素を、直接に、次々にアクセスすることができる。ハッシュテーブルの最後の要素には、通常、終了を示す空データを格納する。これは、通常、主記憶装置であるRAM中に置くことが望ましいが、容量が物理的制限を

超える場合には、例えば、外部記憶装置に置いておき、必要に応じて主記憶装置にキャッシュしてもよい。

[0121] このようにオーバーフロー領域を含む対象配列用ハッシュテーブルを予め作成すれば、例えば、圧縮対象配列の取得(図5ステップM3)と、圧縮問合せ配列と一致する圧縮対象部分配列の検索(図5ステップM4)とを組み合わせることができる。そして、対象配列はゲノムスケールであるために、極めて長い配列であるが、ハッシュテーブルを用いたハッシュ検索によれば、圧縮問合せ配列と不一致な圧縮対象部分配列を飛ばして、次の圧縮対象部分配列を一回の論理アクセスで取得することができる。また、ハッシュテーブルを作成すれば、例えば、図5に示すステップM11で、ある対象配列についての検索が終了した場合、対象配列用ハッシュテーブルの次のハッシュエントリーに進むことにより、より高速に次の対象配列(圧縮対象配列群)を得ることができる。

[0122] 前記ハッシュテーブルは、例えば、必要に応じて、圧縮前の対象部分配列の配列や、圧縮前の対象配列における座標、対象部分配列の各ホモポリマー領域における同一塩基の連続数等と、ホモポリマー数との対応が取れるような、データ構造とすることが好ましい。また、システムのライフサイクル上、一度しか検索の必要がない対象配列に関しては、例えば、このようなハッシュテーブルは、特に必要なく、例えば、適宜、圧縮処理のみを行い、圧縮問合せ配列との対応を行うことでも足りる。対象配列用ハッシュテーブルを作成しない場合であっても、後述する問合せ配列用ハッシュテーブルによる性能の向上は阻害されない。

[0123] ハッシュテーブルについては、何ら制限されず、例えば、Donald Knuth. “The Art of Computer Programming” Volume3, Sorting and Searching, second edition, 1998, pp. 513-558, ISBN 0-201-89685-0.を参照できる。また、本発明においては、前述のように、ハッシュ検索以外に、二分木検索、B木検索があげられ、それぞれ、例えば、Donald Knuth. Fundamental Algorithms, Third Edition. Addison-Wesley, 1997. pp.318-348. ISBN 0-201-89683-4.およびR. Bayer and E. McCreight. ″Organization and Maintenance of Large Ordered Indexes,″ Acta Informatica, 1, 1972.等が参照できる。なお、ハッシュ検索、二分木検索、B木検索

は、例示であって、本発明は、これらには何ら制限されない。

[0124] （実施形態7）

本発明における問合せ配列用ハッシュテーブルの一例について、図7を用いて説明する。なお、本発明は、以下の内容には制限されず、また、特に示さない限り、実施形態6と同様にハッシュテーブルを作成できる。

[0125] 図7に、問合せ配列用ハッシュテーブルの構造の一例を示す。このハッシュテーブルは、例えば、各問合せ配列の各文字列にハッシュ関数をかけた値によって、テーブル内の要素に直接インデックス付けされている。前述と同様に衝突に関しては、例えば、次の要素とチェイニングすることでオーバーフロー領域を形成することが好ましい。

[0126] 問合せ配列は、前述のように、検索の進行に併せて、検索結果を記録することが好ましい。したがって、例えば、問合せ配列用ハッシュテーブルとともに、各種検索情報を格納したテーブルを作成することが好ましい。前記検索情報格納テーブルには、例えば、類似度に基づく相同性の順位（U3）、問合せ配列に相同する対象配列の染色体番号（U4）、前記対象配列のストランドの種別（U5）、圧縮前の対象配列（対象部分配列）の染色体上の位置（U6）、類似度（U7）等があげられる。

[0127] また、前述のステップM7（図5）に示すように、特異性検索を行う場合には、適宜、問合せ配列用ハッシュテーブルのデータを削除することが好ましい。前述のように、特異性検索は、例えば、問合せ配列に対して、対象配列の中に最も相同するスコアを示す対象部分配列が一つしか存在しないか否かを調べるものである。このため、対象配列の部分配列群の中で、最も相同性が高いスコア（上位のスコア）が複数検索された時点で、特異性はないことが明確となる。したがって、このように、対象配列において複数の対象部分配列について最も相同性が高いスコアが得られた際には、その要素（問合せ配列）をハッシュテーブルから除外することが好ましい。これによって、例えば、メモリーの逼迫を回避し、オーバーフロー領域の探索時間の短縮を図ることができる。図7のスコアリング体系を例にすると、例えば、類似度（U7）が0．0となる同率一位の要素が複数出現した場合が該当する。このような結果を示す要素を、検索途中に除外すれば、例えば、問合せ配列の数が膨大な場合、特に性能の向上に有

効に寄与できる。

産業上の利用可能性

[0128] 以上のように、本発明によれば、例えば、ホモポリマー領域における同一塩基連続数が、塩基配列の決定法や配列自身の多型のために、誤差や変位を含む場合であっても、その影響を回避して、より正確な相同性検索が可能となる。さらに、このように精度良く相同性検索が行えることから、例えば、問合せ配列と対象配列における部分配列とが、唯一の相同性(類似性)を示すものであるか否かの判断も精度良く行うことが可能になる。また、第一に、ホモポリマー領域の同一塩基連続数を考慮しない圧縮配列同士を対比し、一致した対象配列の部分配列を選択することため、従来と比較して、データ処理能力が格段に向上し、低コスト化の実現も可能である。したがって、本発明は、相同性検索(類似性検索)の分野において、これまで回避できなかったホモポリマー領域における同一塩基連続数の変動による影響を解決できることから、特に遺伝子解析の分野において極めて有用な技術であるといえる。

請求の範囲

[1] 核酸塩基配列からなる問合せ配列の配列情報を用いて、核酸塩基配列からなるゲノムスケールの対象配列の配列情報から、前記問合せ配列と相同する部分配列を検索する相同性検索システムであって、

前記問合せ配列および対象配列の配列情報を取得する取得手段と、

取得された前記問合せ配列および前記対象配列について、それぞれ、同一塩基が2個以上連続したホモポリマー領域を前記塩基1個に置き換えた圧縮問合せ配列および圧縮対象配列を準備する圧縮配列準備手段と、

前記圧縮問合せ配列と前記圧縮対象配列とを対比し、前記圧縮対象配列において前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索して、絞り込んだ前記圧縮対象部分配列を候補配列の圧縮配列(圧縮候補配列)として選択する検索手段と、

前記圧縮問合せ配列と前記検索手段で選択した前記圧縮候補配列とについて、それぞれの圧縮前の配列における同一塩基の連続数の情報を準備する連続数準備手段と、

前記同一塩基の連続数の情報に基づいて、前記圧縮問合せ配列と前記圧縮候補配列との間で、対応する塩基ごとに、前記塩基の連続数を対比し、前記連続数の一致度または不一致度から、前記候補配列の前記問合せ配列に対する相同性を示す類似度を演算する類似度演算手段と、

前記類似度演算手段により演算した類似度に基づいて、前記問合せ配列と相対的に相同性が高い任意数の候補配列を順位付けて選択する選択手段と、

前記選択手段により選択した前記任意数の候補配列の情報を出力する出力手段とを有する、相同性検索システム。

[2] 前記取得手段が、前記問合せ配列の配列情報を入力する入力手段と、前記対象配列の配列情報が記憶されている対象配列記憶手段とを有する、請求の範囲1記載の相同性検索システム。

[3] 前記取得手段により取得される配列情報が、同一塩基が2個以上連続したホモポリマー領域を前記塩基1個に置き換えた前記圧縮問合せ配列および前記圧縮対象配

列である、請求の範囲1記載の相同性検索システム。

[4] 前記圧縮配列準備手段が、取得された前記問合せ配列および前記対象配列について、それぞれ、同一塩基が2個以上連続したホモポリマー領域を前記塩基1個に置き換えた圧縮問合せ配列および圧縮対象配列に圧縮変換する圧縮変換手段である、請求の範囲1記載の相同性検索システム。

[5] 前記連続数準備手段が、前記圧縮問合せ配列と前記検索手段で選択した前記圧縮候補配列とについて、それぞれの圧縮前の配列における同一塩基の連続数を計数する計数手段である、請求の範囲1記載の相同性検索システム。

[6] 前記類似度演算手段は、前記対応する塩基ごとの連続数の不一致度(ただし、前記圧縮問合せ配列の上流末端塩基または下流末端塩基における圧縮前の連続数が、前記圧縮候補配列の上流末端塩基または下流末端塩基における圧縮前の連続数よりも短い不一致を除く)をペナルティスコアとし、前記対応する塩基ごとのペナルティスコアを加算することによって、類似度を演算する、請求の範囲1記載の相同性検索システム。

[7] さらに、前記問合せ配列と前記選択手段により選択した任意数の候補配列との情報を記憶しておく記憶手段を有し、

前記選択手段は、前記類似度演算手段により、新たな候補配列の前記問合せ配列に対する新たな類似度が演算された際、前記新たな類似度と、前記問合せ配列記憶手段により先立って記憶された前記任意数の候補配列の前記問合せ配列に対する類似度とに基づいて、前記各候補配列から、再度、任意数の候補配列を選択する、請求の範囲1記載の相同性検索システム。

[8] 前記圧縮対象配列は、圧縮後の対象配列を固定長に区切った部分配列群について、同一塩基が2個以上連続したホモポリマー領域を前記塩基1個に置き換えた圧縮対象部分配列群である、請求の範囲1記載の相同性検索システム。

[9] 前記検索手段は、ハッシュ検索手段であり、前記圧縮問合せ配列と前記圧縮対象部分配列群の各圧縮対象部分配列とをキーとし、同じハッシュ関数を用いて、ハッシュ検索を行うことにより、前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索する、請求の範囲8記載の相同性検索システム。

[10] さらに、前記圧縮対象部分配列群の各圧縮対象部分配列をキーとし、同じハッシュ関数を用いて、対象配列用ハッシュテーブルを生成する対象配列用ハッシュテーブル生成手段を有し、

前記検索手段は、ハッシュ検索手段であり、前記圧縮問合せ配列をキーとし、前記対象配列用ハッシュテーブル生成手段と同じハッシュ関数を用いて、前記対象配列用ハッシュテーブル生成手段で生成した前記対象配列用ハッシュテーブルのハッシュ検索を行うことにより、前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索する、請求の範囲8記載の相同性検索システム。

[11] さらに、2つ以上の前記圧縮問合せ配列をキーとし、同じハッシュ関数を用いて、問合せ配列用ハッシュテーブル生成する問合せ配列用ハッシュテーブル生成手段を有し、

前記検索手段は、ハッシュ検索手段であり、前記圧縮対象部分配列群の各圧縮対象部分配列をキーとし、前記問合せ配列用ハッシュテーブル生成手段と同じハッシュ関数を用いて、前記問合せ配列用ハッシュテーブル生成手段で生成した前記問合せ配列用ハッシュテーブルのハッシュ検索を行うことにより、前記各圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索する、請求の範囲8記載の相同性検索システム。

[12] さらに、前記問合せ配列用ハッシュテーブルのデータを更新するハッシュテーブル更新手段を有し、

前記ハッシュテーブル更新手段は、前記選択手段により、1つの問合せ配列に対して、最も高い相同性を示す同じ類似度の候補配列が2つ以上選択された際、前記問合せ配列用ハッシュテーブルのデータから、前記問合せ配列と、それに対して選択された前記2つ以上の候補配列を削除する、請求の範囲11記載の相同性検索システム。

[13] 核酸塩基配列からなる問合せ配列の配列情報を用いて、核酸塩基配列からなるゲノムスケールの対象配列の配列情報から、前記問合せ配列と相同する部分配列を検索する相同性検索システムであって、

端末とサーバーとを有し、

前記端末およびサーバーは、

システム外の通信網を介して接続可能であり、

前記端末は、

前記端末内の情報を前記通信網を介して前記サーバーに送信する端末側送信手段と、

前記サーバーから送信された情報を前記通信網を介して受信する端末側受信手段と、

前記端末内の情報を表示する表示手段と、

前記問合せ配列の配列情報を取得する取得手段とを有し、

前記サーバーは、

前記サーバー内の情報を前記通信網を介して前記端末に送信するサーバー側送信手段と、

前記端末から送信された情報を前記通信網を介して受信するサーバー側受信手段と、

対象配列が蓄積されている対象配列データベースと、

前記対象配列データベースにおける対象配列と、前記サーバー側受信手段により受信した前記問合せ配列とについて、それぞれ、同一塩基が2個以上連続したホモポリマー領域を前記塩基1個に置き換えた圧縮問合せ配列と圧縮対象配列とを準備する圧縮配列準備手段と、

前記圧縮問合せ配列と前記圧縮対象配列とを対比し、前記圧縮対象配列において前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索して、絞り込んだ前記圧縮対象部分配列を候補配列の圧縮配列(圧縮候補配列)として選択する検索手段と、

前記圧縮問合せ配列と前記検索手段で選択した前記圧縮候補配列とについて、それぞれの圧縮前の配列における同一塩基の連続数の情報を準備する連続数準備手段と、

前記同一塩基の連続数の情報に基づいて、前記圧縮問合せ配列と前記圧縮候補配列との間で、対応する塩基ごとに、前記塩基の連続数を対比し、前記連続数の一

致度または不一致度から、前記候補配列の前記問合せ配列に対する相同性を示す類似度を演算する類似度演算手段と、

前記類似度演算手段により演算した類似度に基づいて、前記問合せ配列と相対的に相同性が高い任意数の候補配列を順位付けて選択する選択手段とを有し、

前記問合せ配列の情報が、前記端末側送信手段から前記サーバー側受信手段に送信され、かつ、前記サーバーの前記選択手段により選択した前記任意数の候補配列の情報が、前記サーバー側送信手段から前記端末側受信手段に送信され、前記端末において、受信した前記任意数の候補配列の情報が、前記表示手段により表示される相同性検索システム。

[14] 請求の範囲13記載の相同性検索システムに用いるサーバーであって、

前記サーバーは、

前記サーバー内の情報を前記通信網を介して端末に送信するサーバー側送信手段と、

前記端末から送信された情報を前記通信網を介して受信するサーバー側受信手段と、

対象配列が蓄積されている対象配列データベースと、

前記対象配列データベースにおける対象配列と、前記サーバー側受信手段により受信した前記問合せ配列とについて、それぞれ、同一塩基が2個以上連続したホモポリマー領域を前記塩基1個に置き換えた圧縮問合せ配列と圧縮対象配列とを準備する圧縮配列準備手段と、

前記圧縮問合せ配列と前記圧縮対象配列とを対比し、前記圧縮対象配列において前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索して、絞り込んだ前記圧縮対象部分配列を候補配列の圧縮配列(圧縮候補配列)として選択する検索手段と、

前記圧縮問合せ配列と前記検索手段で選択した前記圧縮候補配列とについて、それぞれの圧縮前の配列における同一塩基の連続数の情報を準備する連続数準備手段と、

前記同一塩基の連続数の情報に基づいて、前記圧縮問合せ配列と前記圧縮候補

配列との間で、対応する塩基ごとに、前記塩基の連続数を対比し、前記連続数の一致度または不一致度から、前記候補配列の前記問合せ配列に対する相同性を示す類似度を演算する類似度演算手段と、

前記類似度演算手段により演算した類似度に基づいて、前記問合せ配列と相対的に相同性が高い任意数の候補配列を順位付けて選択する選択手段とを有する、サーバー。

[15] 請求の範囲13記載の相同性検索システムに用いる端末であって、

前記端末は、

前記端末内の情報を前記通信網を介して前記サーバーに送信する端末側送信手段と、

前記サーバーから送信された情報を前記通信網を介して受信する端末側受信手段と、

前記端末内の情報を表示する表示手段と、

前記問合せ配列の配列情報を取得する取得手段とを有し、

前記問合せ配列の情報が、前記端末側送信手段から前記サーバー側受信手段に送信され、かつ、前記サーバーの前記選択手段により選択した前記任意数の候補配列の情報が、前記サーバー側送信手段から前記端末側受信手段に送信され、前記端末において、受信した前記任意数の候補配列の情報が、前記表示手段により表示される、端末。

[16] 核酸塩基配列からなる問合せ配列の配列情報を用いて、核酸塩基配列からなるゲノムスケールの対象配列の配列情報から、前記問合せ配列と相同する部分配列を検索する相同性検索装置であって、請求の範囲1記載の相同性検索システムを含む、相同性検索装置。

[17] 核酸塩基配列からなる問合せ配列の配列情報を用いて、核酸塩基配列からなるゲノムスケールの対象配列の配列情報から、前記問合せ配列と相同する部分配列を検索する相同性検索方法であって、

前記問合せ配列および対象配列の配列情報を取得する取得ステップと、

取得された前記問合せ配列と前記対象配列について、それぞれ、同一塩基が2個

以上連続したホモポリマー領域を前記塩基1個に置き換えた圧縮問合せ配列と圧縮対象配列とを準備する圧縮配列準備ステップと、

前記圧縮問合せ配列と前記圧縮対象配列とを対比し、前記圧縮対象配列において前記圧縮問合せ配列と一致する圧縮対象部分配列を絞り込み検索して、絞り込んだ前記圧縮対象部分配列を候補配列の圧縮配列(圧縮候補配列)として選択する検索ステップと、

前記圧縮問合せ配列と前記検索ステップで選択した前記圧縮候補配列とについて、それぞれの圧縮前の配列における同一塩基の連続数の情報を準備する連続数準備ステップと、

前記同一塩基の連続数の情報に基づいて、前記圧縮問合せ配列と前記圧縮候補配列との間で、対応する塩基ごとに、前記塩基の連続数を対比し、前記連続数の一致度または不一致度から、前記候補配列の前記問合せ配列に対する相同性を示す類似度を演算する類似度演算ステップと、

前記類似度演算ステップにより演算した類似度に基づいて、前記問合せ配列と相対的に相同性が高い任意数の候補配列を順位付けて選択する選択ステップと、

前記選択ステップにより選択した前記任意数の候補配列の情報を出力する出力ステップとを有する、相同性検索方法。

[18] 前記取得ステップが、前記問合せ配列を入力する入力ステップと、前記対象配列が記憶されている対象配列記憶ステップから前記対象配列の配列情報を呼び出す呼び出しステップとを有する、請求の範囲17記載の相同性検索方法。

[19] 前記類似度演算ステップは、前記対応する塩基ごとの連続数の不一致度(ただし、前記圧縮問合せ配列の上流末端塩基または下流末端塩基における圧縮前の連続数が、前記圧縮候補配列の上流末端塩基または下流末端塩基における圧縮前の連続数よりも短い不一致を除く)をペナルティスコアとし、前記対応する塩基ごとのペナルティスコアを加算することによって、類似度を演算する、請求の範囲17記載の相同性検索方法。

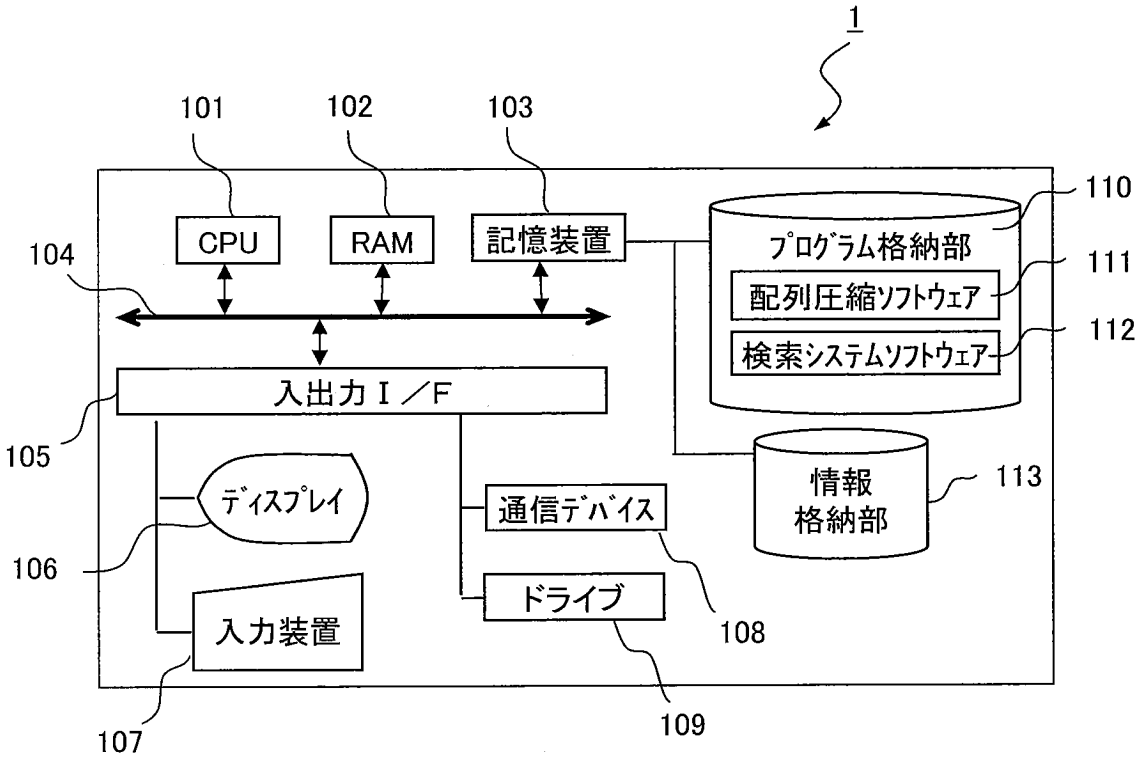
[20] 前記選択ステップにおいて、前記類似度演算ステップにより、新たな候補配列の前記問合せ配列に対する類似度が演算された際、前記新たな類似度と、先立っての前

記選択ステップで選択された前記任意数の候補配列の前記問合せ配列に対する類似度とに基づいて、前記各候補配列から、再度、任意数の候補配列を選択する、請求の範囲17記載の相同性検索方法。

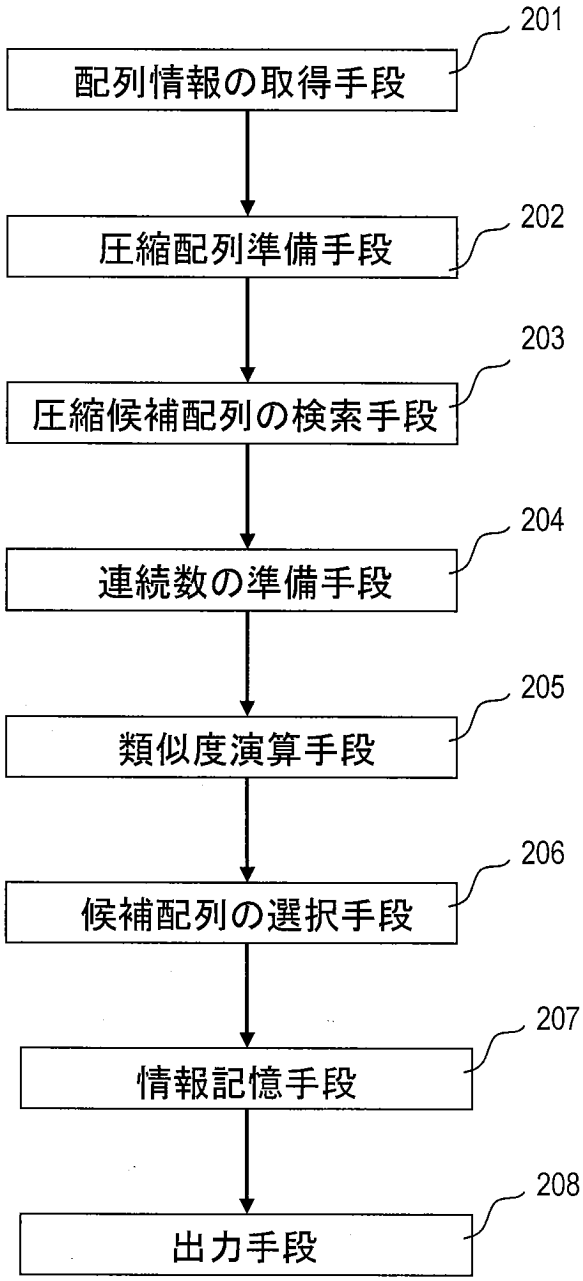
[21] 請求の範囲17記載の相同性検索方法をコンピュータ上で実行可能なコンピュータプログラム。

[22] 請求の範囲21記載のコンピュータプログラムを格納した電子媒体。

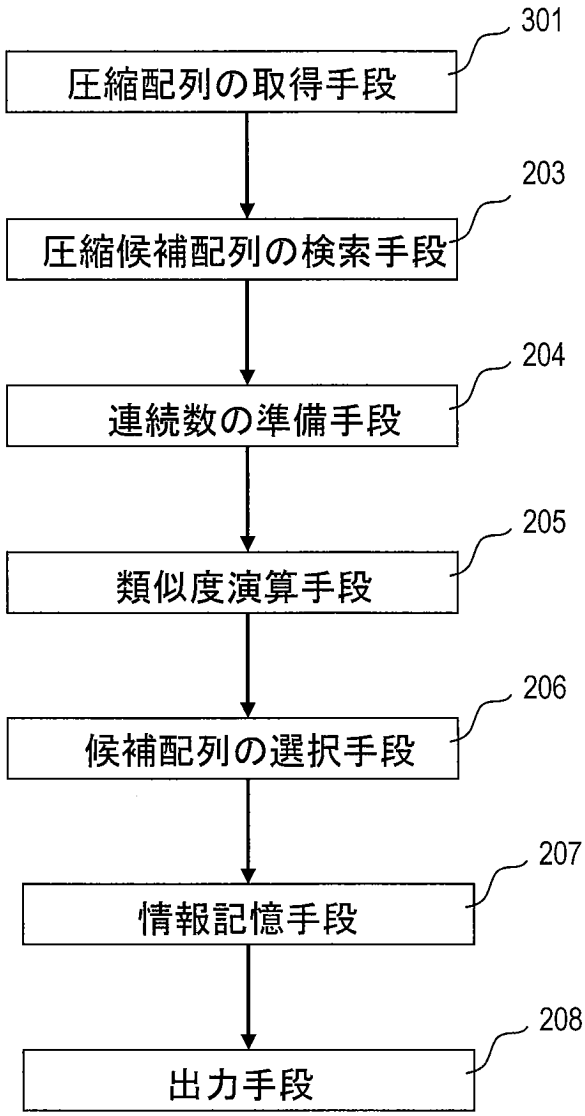
[図1]


1
101
102
103
CPU
RAM
記憶装置
104
プログラム格納部
110
配列圧縮ソフトウェア
111
検索システムソフトウェア
112
入出力Ｉ／F
105
ディスプレイ
106
通信デバイス
情報
格納部
113
入力装置
107
ドライブ
108
109

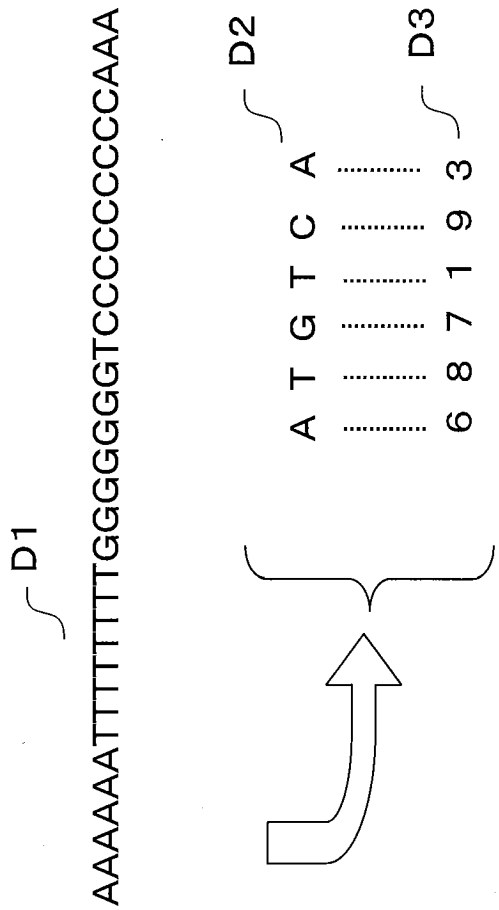
[図2]


201
配列情報の取得手段
202
圧縮配列準備手段
203
圧縮候補配列の検索手段
204
連続数の準備手段
205
類似度演算手段
206
候補配列の選択手段
207
情報記憶手段
208
出力手段

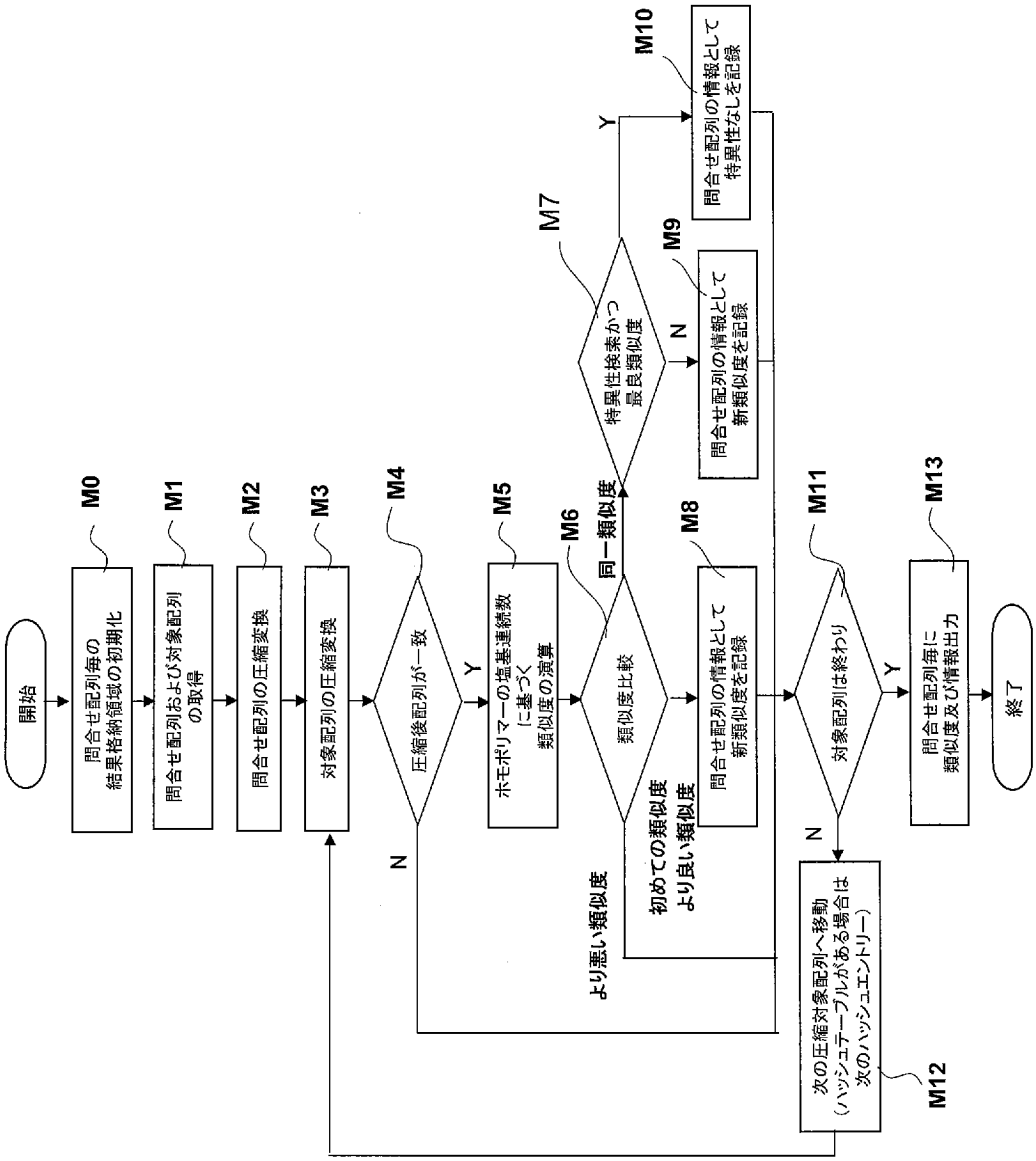
[図3]


301
圧縮配列の取得手段
203
圧縮候補配列の検索手段
204
連続数の準備手段
205
類似度演算手段
206
候補配列の選択手段
207
情報記憶手段
208
出力手段

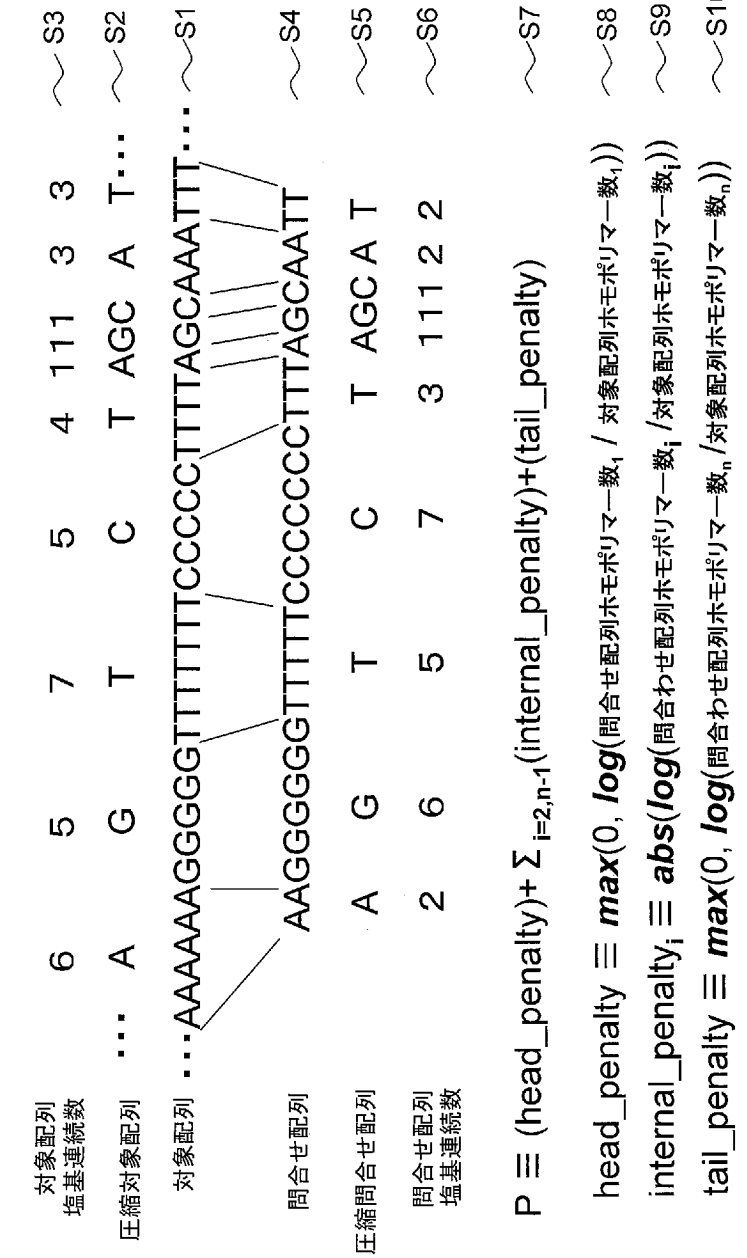
[図4]


D1
AAAAAAATTTTTTTTTGGGGGGGTCCCCCCCCCCAAA
D2
D3
A T G T C A
6 8 7 1 9 3

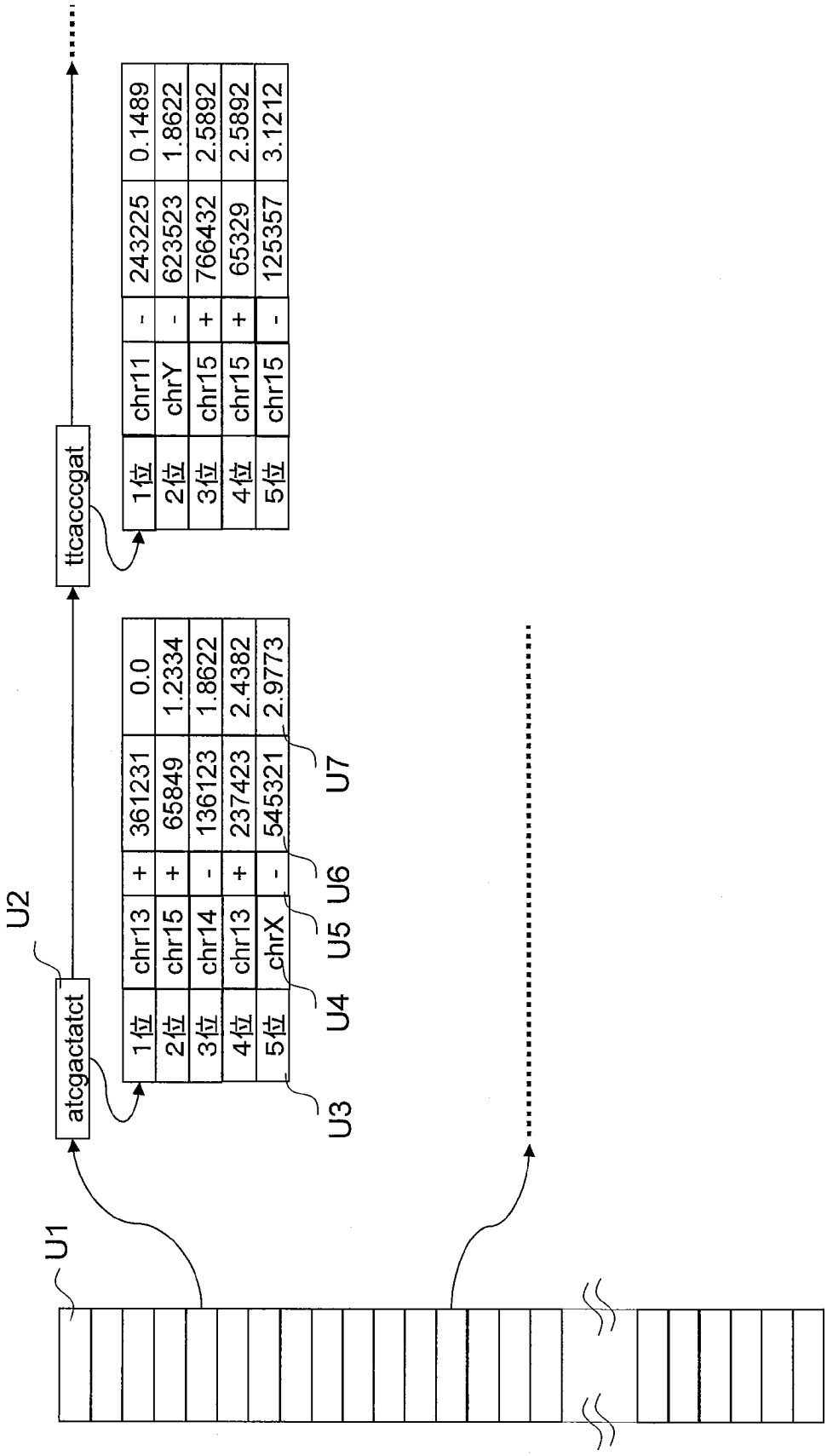
[図5]


開始
M0
問合せ配列毎の結果格納領域の初期化
M1
問合せ配列および対象配列の取得
M2
問合せ配列の圧縮変換
M3
対象配列の圧縮変換
M4
圧縮後配列が一致
Y
N
M5
ホモポリマーの塩基連続数に基づく類似度の演算
M6
類似度比較
同一類似度
より悪い類似度
初めての類似度より良い類似度
M7
特異性検索かつ最良類似度
Y
N
M8
問合せ配列の情報として新類似度を記録
M9
問合せ配列の情報として新類似度を記録
M10
問合せ配列の情報として特異性なしを記録
M11
対象配列は終わり
Y
N
M12
次の圧縮対象配列へ移動（ハッシュテーブルがある場合は次のハッシュエントリー）
M13
問合せ配列毎に類似度及び情報出力
終了

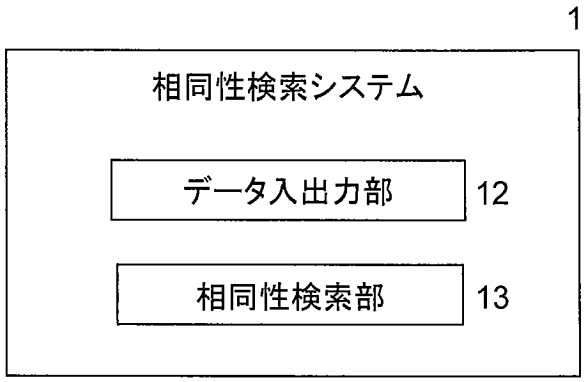
[図6]


対象配列
塩基連続数
6 5 7 5 4 111 3 3
S3
圧縮対象配列
… A G T C T AGC A T…
S2
対象配列
…AAAAAGGGGGTTTTTTCCCCCTTTAGCAAATTT…
S1
問合せ配列
AAGGGGGGTTTTTCCCCCCTTTAGCAATT
S4
圧縮問合せ配列
A G T C T AGC A T
S5
問合せ配列
塩基連続数
2 6 5 7 3 111 2 2
S6
P ≡ (head_penalty)+ Σ i=2,n-1(internal_penalty)+(tail_penalty)
S7
head_penalty ≡ max(0, log(問合せ配列ホモポリマー数1 / 対象配列ホモポリマー数1))
S8
internal_penalty i ≡ abs(log(問合わせ配列ホモポリマー数i /対象配列ホモポリマー数i))
S9
tail_penalty ≡ max(0, log(問合わせ配列ホモポリマー数n /対象配列ホモポリマー数n))
S10

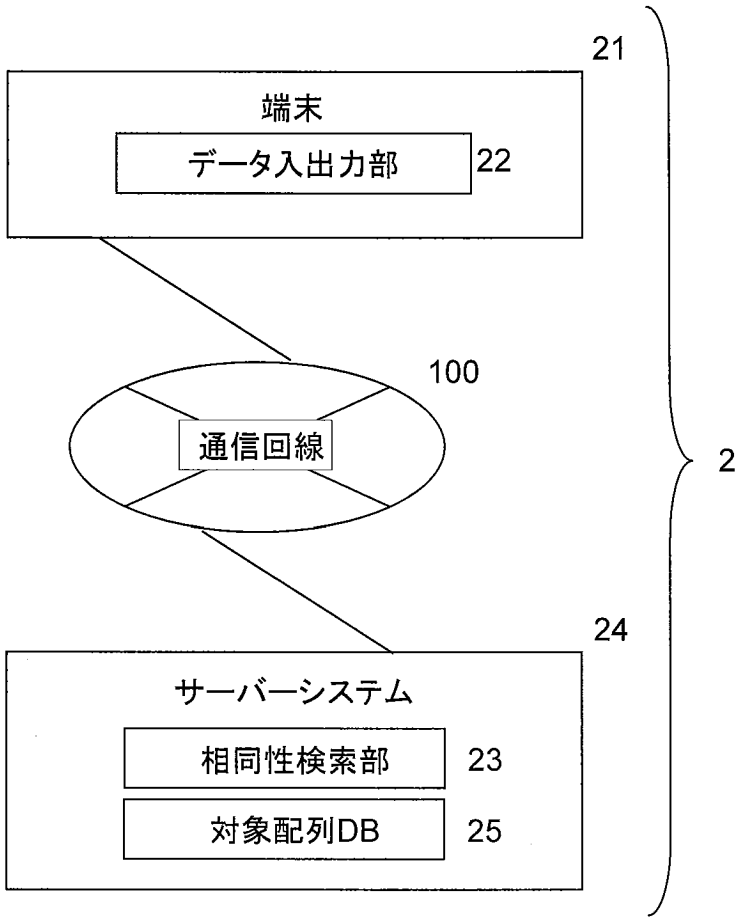
[図7]


U1
U2
atcgactatct
ttcacccgat
1位 chr13 + 361231 0.0
2位 chr15 + 65849 1.2334
3位 chr14 - 136123 1.8622
4位 chr13 + 237423 2.4382
5位 chrX - 545321 2.9773
1位 chr11 - 243225 0.1489
2位 chrY - 623523 1.8622
3位 chr15 + 766432 2.5892
4位 chr15 + 65329 2.5892
5位 chr15 - 125357 3.1212
U3
U4
U5
U6
U7

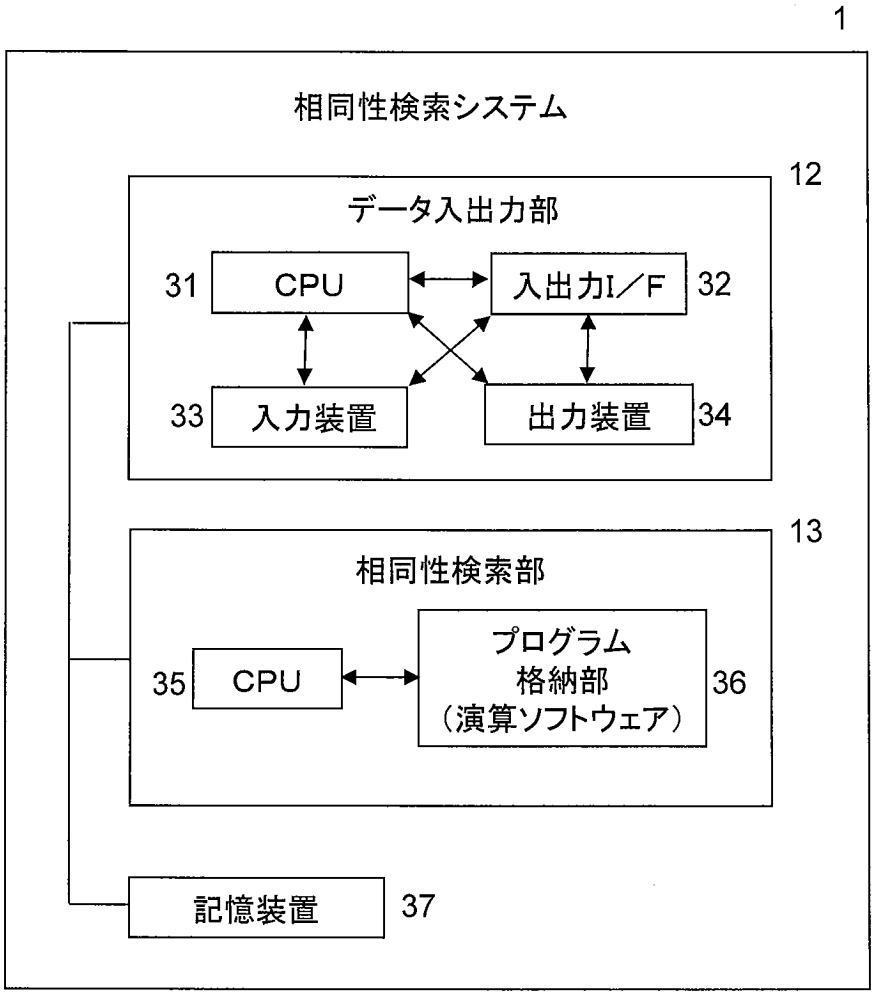
[図8]


1
相同性検索システム
データ入出力部
12
相同性検索部
13

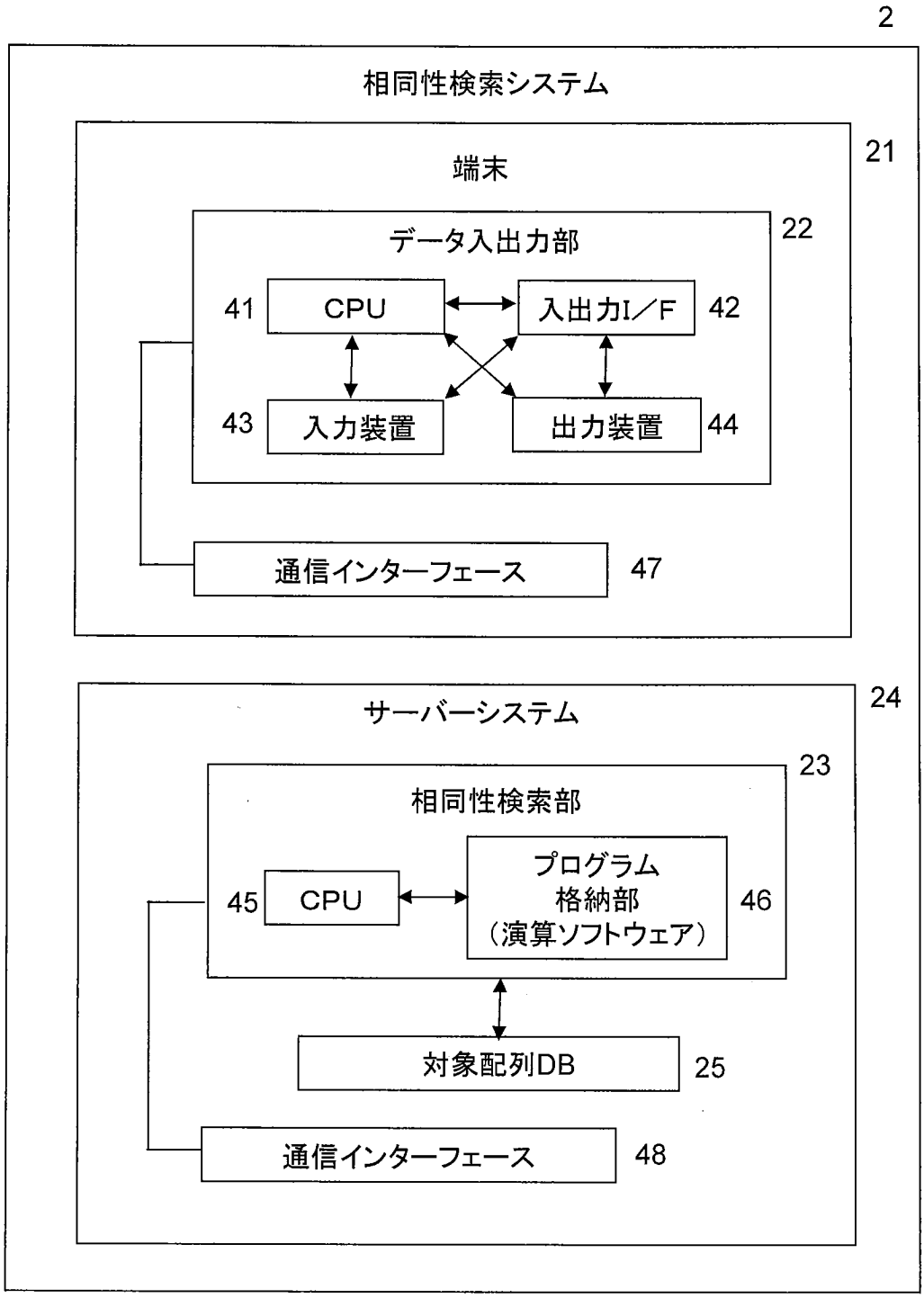
[図9]


21
端末
データ入出力部
22
100
通信回線
2
24
サーバーシステム
相同性検索部
23
対象配列DB
25

[図10]


1
相同性検索システム
12
データ入出力部
31
CPU
入出力I／F
32
33
入力装置
出力装置
34
13
相同性検索部
35
CPU
プログラム
格納部
（演算ソフトウェア）
36
記憶装置
37

[図11]


2
相同性検索システム
端末
21
データ入出力部
22
41
CPU
入出力I／F
42
43
入力装置
出力装置
44
通信インターフェース
47
サーバーシステム
24
相同性検索部
23
45
CPU
プログラム
格納部
（演算ソフトウェア）
46
対象配列DB
25
通信インターフェース
48

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2008/053647

A. CLASSIFICATION OF SUBJECT MATTER
G06F19/00(2006.01)i, *G06F17/30*(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)
G06F19/00, G06F17/30

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

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2008
Kokai Jitsuyo Shinan Koho	1971-2008	Toroku Jitsuyo Shinan Koho	1994-2008

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	FRESCHI, V., Using sequence compression to speedup probabilistic profile matching, Bioinformatics, 2005.05.15, Vol.21, No.10, p.2225-9	1-22
A	MATSUMOTO, T., Biological sequence compression algorithms, Genome Inform. Ser. Workshop Genome Inform., 2000, Vol.11, p.43-52	1-22
A	JP 2006-163734 A (Dainippon Printing Co., Ltd.), 22 June, 2006 (22.06.06), Full text; all drawings (Family: none)	1-22

☒ 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	Date of mailing of the international search report
16 May, 2008 (16.05.08)	27 May, 2008 (27.05.08)
Name and mailing address of the ISA/ Japanese Patent Office	Authorized officer
Facsimile No.	Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2008/053647

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2003-188735 A (NTT Data Corp.), 04 July, 2003 (04.07.03), Full text; all drawings (Family: none)	1-22

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２００８／０５３６４７

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F19/00(2006.01)i, G06F17/30(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（ＩＰＣ））

Int.Cl.　G06F19/00, G06F17/30

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２００８年
日本国実用新案登録公報	１９９６－２００８年
日本国登録実用新案公報	１９９４－２００８年

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

PubMed, JSTPlus(JDreamII), JMEDPlus(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	関連する請求の範囲の番号
A	FRESCHI, V., Using sequence compression to speedup probabilistic profile matching, Bioinformatics, 2005.05.15, Vol.21, No.10, p.2225-9	１－２２
A	MATSUMOTO, T., Biological sequence compression algorithms, Genome Inform. Ser. Workshop Genome Inform., 2000, Vol.11, p.43-52	１－２２

☑　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の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 １６．０５．２００８	国際調査報告の発送日 ２７．０５．２００８
国際調査機関の名称及びあて先 日本国特許庁（ＩＳＡ／ＪＰ） 郵便番号１００－８９１５ 東京都千代田区霞が関三丁目４番３号	特許庁審査官（権限のある職員）　５Ｌ　３１３６ 宮久保　博幸 電話番号　０３－３５８１－１１０１　内線　３５６２

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	関連する請求の範囲の番号
A	JP 2006-163734 A（大日本印刷株式会社）2006.06.22，全文，全図（ファミリーなし）	１－２２
A	JP 2003-188735 A（株式会社エヌ・ティ・ティ・データ）2003.07.04，全文，全図（ファミリーなし）	１－２２