(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2017 年 10 月 5 日(05.10.2017)

WIPO | PCT

(10) 国際公開番号
WO 2017/169736 A1

(51) 国際特許分類:
G06F 19/18 (2011.01) *C12N 15/09* (2006.01)
C12M 1/00 (2006.01)

(21) 国際出願番号: PCT/JP2017/010169

(22) 国際出願日: 2017 年 3 月 14 日(14.03.2017)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願 2016-070976 2016 年 3 月 31 日(31.03.2016) JP

(71) 出願人: 株式会社インテック(INTEC INC.) [JP/JP]; 〒9308577 富山県富山市牛島新町５番５号 Toyama (JP). 国立大学法人　東京大学(THE UNIVERSITY OF TOKYO) [JP/JP]; 〒1138654 東京都文京区本郷七丁目３番１号 Tokyo (JP).

(72) 発明者: 寺井　悟朗(TERAI, Goro); 〒1368637 東京都江東区新砂１－３－３　株式会社インテック内 Tokyo (JP). 浅井　潔(ASAI, Kiyoshi); 〒1138654 東京都文京区本郷七丁目３番１号　国立大学法人東京大学内 Tokyo (JP).

(74) 代理人: ＳＫ特許業務法人，外(SK INTELLECTUAL PROPERTY LAW FIRM et al.); 〒1500012 東京都渋谷区広尾３－１２－４０　広尾ビル４階 Tokyo (JP).

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

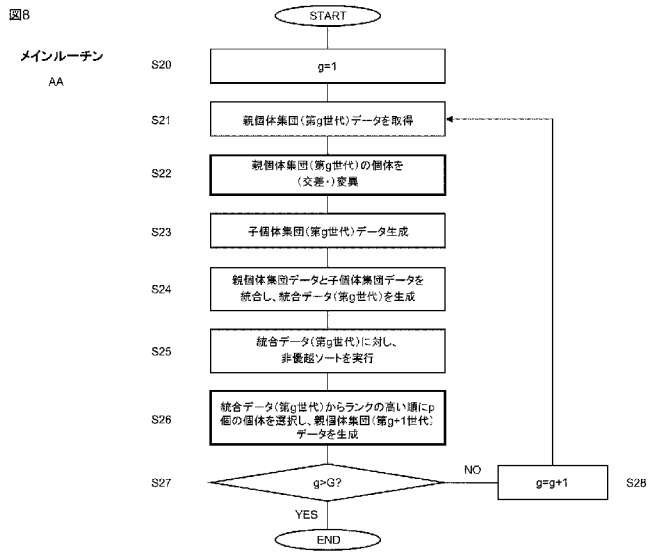
(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK,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

[続葉有]

(54) **Title**: INFORMATION PROCESSING DEVICE AND PROGRAM

(54) 発明の名称 : 情報処理装置及びプログラム

図8



S21 Acquire parent population (g^{th} generation) data
S22 Execute (crossover) mutation of individuals in the parent population (g^{th} generation)
S23 Generate child population (g^{th} generation) data
S24 Generate integrated data (g^{th} generation) by combining the parent population data with the child population data
S25 Apply non-dominated sorting to the integrated data (g^{th} generation)
S26 Select p individuals from the integrated data (g^{th} generation) in descending order of rank, and generate parent population ($g+1^{th}$ generation) data
AA Main routine

(57) **Abstract**: The purpose of the present invention is to provide an information processing device and method capable of increasing the yield of a target protein by suppressing homologous recombination. According to the present invention, the information processing device has: a parent population data acquisition part for acquiring first-generation parent population data representing a first-generation parent population, said parent population data including a predetermined number of individual data generated on the basis of an amino-acid sequence, a number of genes, and data representing a codon frequency table; a mutation treatment part for applying a mutation treatment to the individuals included in the first-generation parent population data; a child population data acquisition part for acquiring first-generation child population data representing a first-generation child population which includes the mutation-treated individuals; a non-dominated sorting execution part for applying non-dominated sorting to first-generation integrated data, which is obtained by combining the first-generation parent population data with the first-generation child population data, on the basis of predetermined evaluation criteria related to codon adaptability and codon base sequences, and classifying all the individual data included in the first-generation integrated data by rank in a Pareto optimal solution; and an individual selection part for selecting a predetermined number of individual data in descending order of rank from all the individual data which have been classified by rank.

(57) 要約:

[続葉有]

WO 2017/169736 A1

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

添付公開書類:

— 国際調査報告（条約第 21 条(3)）

— 明細書の別個の部分として表した配列リスト（規則 5.2(a)）

相同組み換えを抑制し、目的タンパク質の生産量を高めることが可能な情報処理装置及びプログラムを提供するものである。 本発明によれば、アミノ酸配列、遺伝子数及びコドン頻度表を表すデータに基いて生成されたデータであって、予め定められた数の個体データを含む第１世代の親個体集団を表す第１世代親個体集団データを取得する親個体集団データ取得部と、前記第１世代親個体集団データに含まれる個体に対し、変異処理を実行する変異処理部と、前記変異処理が実行された個体を含む第１世代の子個体集団を表す第１世代子個体集団データを取得する子個体集団データ取得部と、予め定められた評価基準であって、コドン適合度及び前記コドンの塩基配列に関する評価基準に基いて、前記第１世代親個体集団データ及び前記第１世代子個体集団データを統合した第１世代統合データに対して非優越ソート処理を実行し、前記第１世代統合データに含まれる全個体データをパレート最適解におけるランク毎に分類する非優越ソート実行部と、前記ランク毎に分類された全個体データから、前記ランクの高い順に予め定められた数の前記個体データを選択する個体選択部と、を有する情報処理装置が提供される。

明　細　書

発明の名称 ： 情報処理装置及びプログラム

技術分野

[0001] 本発明は、目的タンパク質の生産量を高めることができる情報処理装置及びプログラムに関する。

背景技術

[0002] 微生物等に目的タンパク質を生産させる際に、目的タンパク質をコードする遺伝子を複数個導入する手法が知られている。かかる遺伝子は、同じＤＮＡ配列を有するものが利用されることが多い。しかし、同じＤＮＡ配列を有する複数個の遺伝子を導入すると、これらの遺伝子間で相同組み換えが生じ、遺伝子の一部が欠損してしまう。ここで、相同組み換えとは、ＤＮＡの塩基配列がよく似た部位（相同部位）で起こる組み換えのことである。これを概念的に表したのが図１である。図１（ａ）は、同じＤＮＡ配列を有する５個の遺伝子を導入した例を示す。かかる５個の遺伝子のうち、２個目の遺伝子の後半部分～５個目の遺伝子の前半部分において相同組み換えが生じると、図１（ｂ）に示されるように、遺伝子の数が２つまで減少してしまい、目的タンパク質の生産効率が低下してしまう。

[0003] 特許文献１には、合成核酸分子を取得するための方法であって、（ｉ）ポリペプチドのアミノ酸繰り返し領域由来のアミノ酸配列を提供する工程；（ｉｉ）前記アミノ酸配列をそれぞれコードする複数のサンプルコドン最適化核酸配列を推測する工程；（ｉｉｉ）前記複数のサンプルコドン最適化核酸配列を、配列相同性により整列させ、前記複数のサンプルコドン最適化核酸配列を含む近隣結合ツリーを構築する工程；（ｉｖ）前記複数のサンプルコドン最適化核酸配列の１つのみを選択する工程；ならびに、（ｖ）前記選択されたサンプルコドン最適化核酸配列を含む核酸分子を取得する工程を含む、方法が開示されている。

先行技術文献

特許文献

[0004] 特許文献1：特表２０１５-５２４６５８号公報

発明の概要

発明が解決しようとする課題

[0005] 本発明は、相同組み換えを抑制し、目的タンパク質の生産量を高めることが可能な情報処理装置及びプログラムを提供するものである。

課題を解決するための手段

[0006] 本発明によれば、アミノ酸配列、遺伝子数及びコドン頻度表を表すデータに基いて生成されたデータであって、予め定められた数の個体データを含む第１世代の親個体集団を表す第１世代親個体集団データを取得する親個体集団データ取得部と、前記第１世代親個体集団データに含まれる個体に対し、変異処理を実行する変異処理部と、前記変異処理が実行された個体を含む第１世代の子個体集団を表す第１世代子個体集団データを取得する子個体集団データ取得部と、予め定められた評価基準であって、コドン適合度及び前記コドンの塩基配列に関する評価基準に基いて、前記第１世代親個体集団データ及び前記第１世代子個体集団データを統合した第１世代統合データに対して非優越ソート処理を実行し、前記第１世代統合データに含まれる全個体データをパレート最適解におけるランク毎に分類する非優越ソート実行部と、前記ランク毎に分類された全個体データから、前記ランクの高い順に予め定められた数の前記個体データを選択する個体選択部と、を有する情報処理装置が提供される。

[0007] 本発明によれば、異なる２つの評価基準に基づいて、相同組み換えを抑制し、目的タンパク質の生産量を高めることが可能となる。

[0008] 以下、本発明の種々の実施形態を例示する。以下に示す実施形態は互いに組み合わせ可能である。

好ましくは、前記個体選択部は、前記予め定められた数の前記個体データを選択するときに、前記ランクが同じ前記個体データが存在する場合には、

混雑距離が大きいものから順に選択する。

好ましくは、前記親個体集団データ取得部は、前記個体選択部により選択された前記個体データを、第２世代の親個体集団を表す第２世代親個体集団データとし、前記変異処理部、前記非優越ソート実行部及び前記個体選択部による処理を、予め定められた世代数となるまで実行する。

好ましくは、前記コドン適合度に関する評価基準は、各個体が複数有する塩基配列であって、アミノ酸翻訳の対象となる塩基配列を表すＣＤＳのコドン適合インデックスの最小値を基準とする。

好ましくは、前記個体に含まれる前記コドン適合インデックスの最小値が大きいほど、前記個体の評価を高くする。

好ましくは、前記コドンの塩基配列に関する評価基準は、前記各個体に含まれる２つの前記ＣＤＳのうち、互いに一致しない塩基の数を表す不一致塩基数の最小値を基準とする。

好ましくは、前記不一致塩基数の最小値が大きいほど、前記個体の評価を高くする。

好ましくは、前記コドンの塩基配列に関する評価基準は、前記各個体に含まれる前記ＣＤＳのうち、それぞれのＣＤＳ間又は１つのＣＤＳ内部の異なる部位で連続して一致する塩基配列のうち最長の塩基配列である最長共通文字列の長さを基準とする。

好ましくは、前記最長共通文字列の長さが短いほど、前記個体を高く評価する。

好ましくは、前記変異処理部は、第ｇ世代の親個体集団を表す第ｇ世代親個体集団データに含まれる各個体データに対し、第１変異処理及び前記第１変異処理とは異なる第２変異処理を実行する。

好ましくは、前記変異処理部は、前記各個体に含まれる全てのＣＤＳに対し、前記ＣＤＳに含まれる前記コドンを、予め定められた確率で前記コドンより高頻度のコドンに置換する第１変異処理を実行する。

好ましくは、前記変異処理部は、前記各個体に含まれるＣＤＳのうち、そ

れぞれのＣＤＳ間又は１つのＣＤＳ内部の異なる部位で連続して一致する塩基配列のうち最長の塩基配列である最長共通文字と重なる前記コドンを、予め定められた確率で他のコドンに置換する第２変異処理を実行する。

好ましくは、前記第１変異処理又は前記第２変異処理は、ランダムに選択される。

好ましくは、前記第１世代親個体集団データに含まれる個体に対し、交差処理を実行する交差処理部を有し、前記交差処理は、第ｇ世代の親個体集団を表す第ｇ世代親個体集団データから予め定められた偶数個の個体データを抽出し、前記抽出された個体データから２個の個体データを選択し、前記選択された２個の個体データに対して交差処理を実行する。

好ましくは、前記交差処理部は、前記選択された２個の個体データである第１個体データ及び第２個体データに含まれる前記ＣＤＳに含まれる前記コドンの境界から交差ポイントを決定し、前記交差ポイントを境として前記第１個体データと前記第２個体データに含まれる前記コドンを入れ替える。

好ましくは、コンピュータを、アミノ酸配列、遺伝子数及びコドン頻度表を表すデータに基いて生成されたデータであって、予め定められた数の個体データを含む第１世代の親個体集団を表す第１世代親個体集団データを取得する親個体集団データ取得部、前記第１世代親個体集団データに含まれる個体に対し、変異処理を実行する変異処理部、前記変異処理が実行された個体を含む第１世代の子個体集団を表す第１世代子個体集団データを取得する子個体集団データ取得部、予め定められた評価基準であって、コドン適合度及び前記コドンの塩基配列に関する評価基準に基いて、前記第１世代親個体集団データ及び前記第１世代子個体集団データを統合した第１世代統合データに対して非優越ソート処理を実行し、前記第１世代統合データに含まれる全個体データをパレート最適解におけるランク毎に分類する非優越ソート実行部、前記ランク毎に分類された全個体データから、前記ランクの高い順に予め定められた数の前記個体データを選択する個体選択部、として機能させるための情報処理プログラムが提供される。

図面の簡単な説明

[0009] [図1]微生物等に目的タンパク質を生産させる際に、目的タンパク質をコードする遺伝子を複数個導入する従来の手法を表す概念図であり、（ａ）は同じＤＮＡ配列を有する５個の遺伝子を導入した例、（ｂ）は５個の遺伝子のうち、２個目の遺伝子の後半部分～５個目の遺伝子の前半部分において相同組み換えが生じ、遺伝子の数が２個まで減少した結果を表す。

[図2]本発明の一実施形態に係る遺伝子配列設計を表す概念図であり、（ａ）は本発明に係るアルゴリズムに入力データを入力し、出力データとして遺伝子配列を出力する様子を、（ｂ）は導入された５個の遺伝子データに相同組み換えが生じず、全ての遺伝子から目的タンパク質が生産される様子を表す。

[図3]情報処理装置１のハードウェア構成の例を示す図である。

[図4]本発明の一実施形態に係る情報処理装置１の例示的な機能ブロック図である。

[図5]混雑距離を説明するための図であり、（ａ）は混雑距離の概念図、（ｂ）は混雑距離の計算式を表す。

[図6]本発明の一実施形態に係る遺伝子配列設計を実施するためのフローチャートの一例を示す図である。かかる処理は、図８に示されるメインルーチンに先立ち実行される。

[図7]個体データの例を表す図である。本実施形態では、１つの個体を、同じアミノ酸をコードする複数のタンパクコード領域（ＣＤＳ）として表現する。

[図8]本発明の一実施形態に係る遺伝子配列設計を実施するためのフローチャートの一例を示す図である。なお、Ｓ２２において、交差処理は任意であり、必要に応じて省略することができる。

[図9]本発明の一実施形態に係る交差処理を実施するためのフローチャートの一例を示す図である。

[図10]本発明の一実施形態に係る交差処理を表す概念図である。

[図11]本発明の一実施形態に係る変異処理を実施するためのフローチャートの一例を示す図である。

[図12]本発明の一実施形態に係る変異処理を表す概念図であり、（ａ）は第１変異処理、（ｂ）は第２変異処理を表す概念図である。

[図13]本発明の一実施形態に係るコドンの塩基配列に関する評価基準である「不一致塩基数」を説明するための図である。図１３の例では、不一致塩基数は５である。

[図14]本発明の一実施形態に係るコドンの塩基配列に関する評価基準である「最長共通文字列」を説明するための図である。図１４では、実線の下線が付された部分が最長共通文字列、破線の下線が付された部分が共通文字列を表す。

[図15]本発明の実施例における処理結果を表す図である。図１５では、コドン適合度に関する評価基準（各個体が複数有する塩基配列であって、アミノ酸翻訳の対象となる塩基配列を表すＣＤＳのコドン適合インデックス（ＣＡＩ）の最小値）を横軸、コドンの塩基配列に関する評価基準（不一致塩基数の最小値）を縦軸とし、第１世代、第１０世代及び第２５０世代の計算結果をそれぞれプロットしたグラフである。

[図16]本発明の実施例における処理結果を表す図である。図１６では、コドン適合度に関する評価基準（各個体が複数有する塩基配列であって、アミノ酸翻訳の対象となる塩基配列を表すＣＤＳのコドン適合インデックス（ＣＡＩ）の最小値）を横軸、コドンの塩基配列に関する評価基準（最長共通文字列の長さ）を縦軸とし、第１世代、第１０世代及び第２５０世代の計算結果をそれぞれプロットしたグラフである。

発明を実施するための形態

[0010] ＜実施形態＞

以下、図面を用いて本発明の実施形態について説明する。以下に示す実施形態中で示した各種特徴事項は、互いに組み合わせ可能である。

[0011] ＜本発明の一実施形態に係る遺伝子配列設計＞

図２は、本発明の一実施形態に係る遺伝子配列設計を表す概念図である。一実施形態に係る遺伝子配列設計は、相同組み換えを誘発しない遺伝子配列群を設計し、微生物等に導入することで、目的タンパク質の生産量を高めるものである。図２（ａ）に示されるように、目的タンパク質を表すデータ及びＮ個の遺伝子を表すデータを入力データとし、アルゴリズム（以下、本アルゴリズムという）に基いて計算処理し、かかる計算結果である目的タンパク質をコードするＮ個の遺伝子配列群を表すデータを出力する。ここで、本アルゴリズムは、相競合する評価基準を持つ複数の目的関数を同時に最適化することを目的とする多目的遺伝的アルゴリズムを利用する。これにより、図２（ｂ）に示されるように、例えば導入された遺伝子が５個である場合、かかる５個の遺伝子データに相同組み換えが生じず、全ての遺伝子から目的タンパク質が生産される。ここで、図２（ｂ）では、１～５までの遺伝子は、互いに塩基配列が異なる遺伝子である。

[0012] ＜ハードウェア構成＞

次に、本発明の一実施形態に係る情報処理装置１のハードウェア構成の例について、図３を用いて説明する。情報処理装置１は、処理部１０、記憶部２０、操作部３０、表示部４０及び通信部５０を有する。処理部１０は、種々の演算処理を実行するものであり、例えば、ＣＰＵ等により構成される。記憶部２０は、種々のデータやプログラムを記憶するものであり、例えば、メモリ、ＨＤＤ又はＳＳＤ等により構成される。ここで、プログラムは、情報処理装置１の出荷時点においてプリインストールされていてもよく、Ｗｅｂ上のサイトからアプリケーションとしてダウンロードしてもよく、無線通信により他の情報処理装置から転送されてもよい。操作部３０は、情報処理装置１を操作するものであり、例えば、タッチパネル、キーボード、音声入力部、カメラ等を利用した動き認識装置等により構成される。表示部４０は、種々の画像（静止画及び動画を含む）を表示するものであり、例えば、タッチパネルディスプレイ、有機ＥＬディスプレイ、電子ペーパーその他のディスプレイで構成される。通信部５０は、他の情報処理装置と種々のデータ

を送受信するものであり、任意のＩ／Ｏにより構成される。バス１００はシリアルバス、パラレルバス等で構成され、各部を電気的に接続し、種々のデータの送受信を可能にするものである。

[0013] ＜機能ブロック図＞

次に、情報処理装置１の機能について、図４の機能ブロック図を用いて説明する。情報処理装置１は、例えば、多機能情報端末であり、ＰＣ、サーバ、スマートフォン、タブレット端末、スマートウォッチ等である。情報処理装置１は、操作部３０、表示部４０及び通信部５０と、処理部１０と、記憶部２０を備える。処理部１０は、個体生成部１０１、親個体集団データ取得部１０２、子個体集団データ取得部１０３、交差処理部１０４、変異処理部１０５、非優越ソート実行部１０６、個体選択部１０７を備える。また、記憶部２０は、アミノ酸配列データ記憶部２０１、遺伝子数データ記憶部２０２、コドン頻度表データ記憶部２０３、計算データ記憶部２０４、評価基準記憶部２０５を備える。

[0014] 操作部３０、表示部４０及び通信部５０の各機能については、図３の説明を参照されたい。

[0015] ＜処理部１０＞

次に、処理部１０の機能について説明する。個体生成部１０１は、アミノ酸配列、遺伝子数及びコドン頻度表を表すデータをそれぞれアミノ酸配列データ記憶部２０１、遺伝子数データ記憶部２０２及びコドン頻度表データ記憶部２０３から取得し、同じタンパク質のアミノ酸配列をコードするという制約下でランダムに生成した個体を表す個体データをｐ個生成するものである。ここで、ｐは正の数のパラメータであり、任意の数とすることができる。

[0016] 親個体集団データ取得部１０２は、個体生成部１０１が生成したｐ個の個体データを、本アルゴリズムに利用するデータであって、第ｇ世代の親個体集団を表す第ｇ世代親個体集団データとして取得する。さらに、本アルゴリズムにおける計算は、後述するように所定のフローを複数回繰り返し実行す

るループ計算を実行するものであり、親個体集団データ取得部１０２は、第１世代、第２世代、・・・第ｇ世代の親個体集団を表す第１世代親個体集団データ、第２世代親個体集団データ・・・第ｇ世代親個体集団データを取得する。ここで、ｇは正の数であり、本アルゴリズムにおける計算のループ数を表す。

[0017] 交差処理部１０４は、第ｇ世代親個体集団データからｅ個（予め定められた偶数個）の個体データを抽出し、抽出された個体データから２個の個体データを選択し、選択された２個の個体データに対して交差処理を実行するものである。そして、まだ交差処理が行われていない個体データの中から２個の個体データを選択し、交差処理を実行する。かかる処理を、抽出されたｅ個の個体データの全てに対して繰り返す。具体的には、選択された２個の個体データである第１個体データ及び第２個体データに含まれるＣＤＳに含まれるコドンの境界から交差ポイントを決定し、交差ポイントを境として第１個体データと第２個体データに含まれるコドンを入れ替える。ここで、２個の個体データの選択は、例えば乱数表等を利用してランダムに実行される。ここで、ｅは「ｐ×Ｐｃ」を超えない最大の偶数である。なお、Ｐｃはパラメータであり、０より大きく１より小さい任意の値とすることができる。ｐ個の個体データを含む第ｇ世代親個体集団データからｅ個の個体データを抽出する手法は特に限定されないが、例えば「ｂｉｎａｒｙ　ｔｏｕｒｎａｍｅｎｔ　ｓｅｌｅｃｔｉｏｎ法」を用いることができる。

[0018] 変異処理部１０５は、第ｇ世代親個体集団データに含まれる全ての個体に対し、変異処理を実行するものである。本実施形態では、第ｇ世代親個体集団データに含まれる各個体データに対し、第１変異処理及び前記第１変異処理とは異なる第２変異処理を実行する。具体的には、各個体データに対し、第１変異処理及び第２変異処理をランダムに決定する。そして、第１変異処理と決定された場合、各個体データに含まれる全てのＣＤＳに対し、ＣＤＳに含まれるコドンを、予め定められた確率Ｐｍでかかるコドンより高頻度のコドンに置換する。また、第２変異処理と決定された場合、各個体データに

含まれるＣＤＳのうち、それぞれのＣＤＳ間又は１つのＣＤＳ内部の異なる部位で連続して一致する塩基配列のうち最長の塩基配列である最長共通文字と重なるコドンを、予め定められた確率Ｐｍで他のコドンに置換する。これらの処理の詳細については後述する。

[0019] 子個体集団データ取得部１０３は、変異処理部１０５による変異処理が実行された個体を含む第ｇ世代の子個体集団を表す第ｇ世代子個体集団データを取得する。ここで、第ｇ世代子個体集団データに含まれる個体データの数は、第ｇ世代親個体集団データに含まれる個体データの数と等しく、ｐ個である。これは、変異処理部１０５による変異処理が、第ｇ世代親個体集団データに含まれる全ての個体に実行されたためである。

[0020] 非優越ソート実行部１０６は、予め定められた評価基準であって、コドン適合度及び前記コドンの塩基配列に関する評価基準に基いて、第ｇ世代親個体集団データ及び第ｇ子個体集団データを統合した第ｇ世代統合データに対し、非優越ソート処理を実行するものである。第ｇ世代統合データに含まれる個体データの数は、２ｐ（＝ｐ＋ｐ）個である。そして、全個体をパレート最適解におけるフロント毎（ランク毎）に分類する。

[0021] 個体選択部１０７は、パレート最適解におけるフロント毎（ランク毎）に分類された第ｇ世代統合データから、ランクの高い順に定められた数の個体データを選択するものである。例えば、予め定められた数として、ｐを採用することができる。そして、親個体集団データ取得部１０２は、個体選択部１０７により選択されたｐ個の個体データを、第ｇ＋１世代の親個体集団を表す第ｇ＋１世代親個体集団データとして取得する。

[0022] ここで、個体選択部１０７は、定められた数の個体データを選択するときに、ランクが同じ個体データが存在する場合には、混雑距離（Ｃｒｏｗｄｉｎｇ　Ｄｉｓｔａｎｃｅ）が大きいものから順に選択することとしてもよい。ここで、混雑距離とは、ある解の両側にある２つの解の平均距離である。これを概念的に表したのが図５（ａ）である。そして、図５（ｂ）の計算式により、混雑距離が計算される。ここで、混雑距離は、図５（ａ）において

破線で示される四角形の周囲の長さの平均に相当する。

[0023] ＜記憶部２０＞

次に、記憶部２０の機能について説明する。アミノ酸配列データ記憶部２０１は、アミノ酸配列を表すデータを記憶するものである。アミノ酸配列は、タンパク質中のアミノ酸の配列を表すものである。

[0024] 遺伝子数データ記憶部２０２は、遺伝子数を表すデータを記憶するものである。ここで、本実施形態では、遺伝子数は、個体データに含まれるＣＤＳの数を表すものとする。

[0025] コドン頻度表データ記憶部２０３は、コドン頻度表を表すデータを記憶するものである。コドン頻度表は、宿主細胞におけるコドンの使用頻度をまとめた表である。

[0026] 計算データ記憶部２０４は、個体生成部１０１、親個体集団データ取得部１０２、子個体集団データ取得部１０３、交差処理部１０４、変異処理部１０５、非優越ソート実行部１０６、個体選択部１０７等による種々の処理における計算結果を記憶するものである。

[0027] 評価基準記憶部２０５は、予め定められた評価基準であって、コドン適合度及びコドンの塩基配列に関する評価基準を記憶するものである。具体的には、コドン適合度に関する評価基準は、各個体が複数有する塩基配列であって、アミノ酸翻訳の対象となる塩基配列を表すＣＤＳのコドン適合インデックスの最小値を基準とする。以下、かかる基準を第１評価基準という。第１評価基準においては、個体に含まれるコドン適合インデックスの最小値が大きいほど、個体が高く評価される。そして、コドンの塩基配列に関する評価基準のうちの１つ目は、各個体に含まれる２つのＣＤＳのうち、互いに一致しない塩基の数を表す不一致塩基数の最小値を基準とする。以下、かかる基準を第２評価基準という。第２評価基準においては、不一致塩基数の最小値が大きいほど、個体が高く評価される。コドンの塩基配列に関する評価基準のうちの２つ目は、各個体に含まれるＣＤＳのうち、それぞれのＣＤＳ間又は１つのＣＤＳ内部の異なる部位で連続して一致する塩基配列のうち最長の

塩基配列である最長共通文字列の長さを基準とする。以下、かかる基準を第３評価基準という。第３評価基準においては、最長共通文字列の長さが短いほど、個体が高く評価される。

[0028] 次に、以上説明した種々の機能、処理及び基準の詳細について、図６～図１３を用いて説明する。

[0029] ＜前処理＞

図６は、本発明の一実施形態に係る遺伝子配列設計を実施するためのフローチャートの一例を示す図である。図６に示される処理は、図８に示されるメインルーチンに先立ち実行される処理である。以下、図６に示される処理を前処理という。

[0030] まず、Ｓ１１において、処理部１０は、アミノ酸配列データ記憶部２０１及び遺伝子数データ記憶部２０２から、アミノ酸配列データ及び遺伝子数データを取得する。そして、図示しないキャッシュメモリ等の記憶部にデータを記憶する。

[0031] 次に、Ｓ１２において、処理部１０は、コドン頻度表データ記憶部２０３からコドン頻度表データを取得する。そして、図示しないキャッシュメモリ等の記憶部にデータを記憶する。

[0032] 次に、Ｓ１３において、個体生成部１０１は、同じタンパク質のアミノ酸配列をコードするという制約下でランダムに生成した個体を表す個体データをｐ個生成する。例えば、ランダムな個体データを１００個生成してもよい。

[0033] （個体データ）

ここで、図７を用いて個体データについて説明する。図７に示されるように、本実施形態では、１つの個体を、同じアミノ酸をコードする複数のタンパクコード領域（ＣＤＳ）として表現する。図７に示される個体データでは、ＣＤＳがｘ個である。これは、図６のＳ１１において処理部１０が遺伝子数データ記憶部２０２から取得した遺伝子数データが表す遺伝子の数である。各ＣＤＳはそれぞれ同じアミノ酸をコードする。ここで、図７に示される

Ｇ，Ｉ，Ｖ，Ｅ，Ｑは、図６のＳ１１において処理部１０がアミノ酸配列データ記憶部２０１から取得したアミノ酸配列データが表すアミノ酸配列である。また、各ＣＤＳは、それぞれ塩基配列が異なっている。

[0034] 図６に戻り、前処理についてさらに説明する。Ｓ１４において、親個体集団データ取得部１０２は、Ｓ１３において個体生成部１０１がランダムに生成したｐ個の個体データを第１世代の親個体集団を表す第１世代親個体集団データとして取得する。親個体集団データは、本アルゴリズムにおける処理において保存されるアーカイブ母集団である。そして、第１世代親個体集団データを取得すると、前処理を終了する。

[0035] ＜メインルーチン＞

次に、図８を用いて、本アルゴリズムにおけるメインルーチンについて説明する。まず、Ｓ２０において、処理部１０は、変数ｇを１にセットする。ここで、ｇは第ｇ世代の親個体集団を表す符号である。ｇは、１～Ｇ（後述する予め定められた世代数Ｇ）までの値をとる。

[0036] 次に、Ｓ２１において、処理部１０は、親個体集団データ取得部１０２から第１世代親個体集団データを取得する。

[0037] 次に、Ｓ２２において、交差処理部１０４及び変異処理部１０５は、第１世代親個体集団データに含まれる個体データに対して交差処理及び変異処理を実行する。なお、交差処理は任意であり、必要に応じて省略することができる。以下、図９～図１２を用いて交差処理及び変異処理について説明する。

[0038] ＜交差処理＞

まず、図９及び図１０を用いて交差処理について説明する。図９は、本発明の一実施形態に係る交差処理を実施するためのフローチャートの一例を示す図である。まず、Ｓ３２１において、交差処理部１０４は、変数ｉを０にセットする。

[0039] 次に、Ｓ３２２において、交差処理部１０４は、処理部１０又は親個体集団データ取得部１０２から、第ｇ世代親個体集団データ（図８におけるメイ

ンルーチンでｇ＝１の場合は第１世代親個体集団データ）を取得する。そして、第ｇ世代親個体集団データに含まれるｐ個の個体データから、（ｅ－ｉ）個（現時点ではｉ＝０のためにｅ個）の個体データをランダムに抽出する。かかる抽出に利用する手法は特に限定されないが、例えば「ｂｉｎａｒｙ　ｔｏｕｒｎａｍｅｎｔ　ｓｅｌｅｃｔｉｏｎ法」を用いることができる。ここで、ｅは「ｐ×Ｐｃ」を超えない最大の偶数である。なお、Ｐｃはパラメータであり、０より大きく１より小さい任意の値とすることができる。

[0040] 次に、Ｓ３２３において、交差処理部１０４は、（ｅ－ｉ）個の個体データからランダムに２個の個体データを選択する。２個の個体データの選択は、例えば乱数表等を利用してランダムに実行される。

[0041] 次に、Ｓ３２４において、交差処理部１０４は、Ｓ３２３にて選択された２個の個体データに対して交差処理を実行する。ここで、交差処理について、図１０を用いて具体的に説明する。

[0042] 図１０に示されるように、Ｓ３２３にて選択された２個の個体データをそれぞれ第１個体データ及び第２個体データとする。図１０の例では、第１個体データ及び第２個体データはそれぞれ３つのＣＤＳを有し、異なる塩基配列を有する。これらの個体データから、交差ポイントを決定する。交差ポイントは、コドンとコドンの境界から１箇所選ばれる。かかる決定はランダムに行われてもよい。本実施形態では、第１個体データと第２個体データにおける交差ポイントは同じ場所とする。そして、交差ポイントを境として、第１個体データと第２個体データに含まれるコドンを入れ替える。本実施形態では、かかる処理を交差処理という。

[0043] 図９に戻り、交差処理についてさらに説明する。Ｓ３２５において、交差処理部１０４は、変数ｉを２増やす。

[0044] 次に、Ｓ３２６において、交差処理部１０４は、変数ｉ＝ｅであるか否かを判定する。そして、判定結果がＮＯであれば、再びＳ３２３に戻る。一方、判定結果がＹＥＳであれば、交差処理を終了し、かかる計算結果を計算データ記憶部２０４へ出力する。ここで、現時点ではｉ＝２であり、ｅが２よ

りも大きいとすると、Ｓ３２６からＳ３２３へ戻ることになる。そして、まだ交差処理が実行されていない（ｅ－２）個の個体データからランダムに２個の個体データを選択する。かかる処理を、Ｓ３２６における判定結果がＹＥＳ、つまり、ｅ個の個体データ全てに対して交差処理が実行されるまで繰り返す。なお、前述のとおり、かかる交差処理は任意であり、必要に応じて省略することができる。

[0045] ＜変異処理＞

次に、図１１及び図１２を用いて、変異処理について説明する。変異処理は、第ｇ世代親個体集団データに含まれる全ての個体に対して実行される。ここで、Ｓ２２において交差処理が実行されていない場合には、第ｇ世代に含まれるｐ個の個体データに対して変異処理を実行する。一方、Ｓ２２において交差処理が実行された場合には、交差処理が実行されたe個の個体データと、交差処理が実行されていないｐ－ｅ個の個体データを合わせた計ｐ個の個体データに対して変異処理を実行する。

[0046] 図１１は、本発明の一実施形態に係る変異処理を実施するためのフローチャートの一例を示す図である。まず、Ｓ２２１において、変異処理部１０５は、第ｇ世代親個体集団データに含まれる各個体データに対し、第１変異処理又は第２変異処理のいずれを実行するかをランダムに決定する。本実施形態では、第ｇ世代親個体集団データに含まれるｐ個の個体データの全てに対して変異処理を実行するものとする。ここで、第２変異処理は、第１変異処理とは異なる変異処理である。

[0047] 次に、Ｓ２２２において、変異処理部１０５は、Ｓ２２１における決定結果が第１変異処理であるか否かを判定する。そして、判定結果がＹＥＳであれば、Ｓ２２３ａに進み、第１変異処理を実行する。一方、判定結果がＮＯであれば、Ｓ２２３ｂに進み、第２変異処理を実行する。

[0048] （第１変異処理）

次に、Ｓ２２３ａにおいて、変異処理部１０５は、個体データに対して第１変異処理を実行する。具体的には、個体データに含まれる全てのＣＤＳに

対し、各コドンを予め定められた確率Ｐｍでかかるコドンより高頻度のコドンに置換する。ここで、より高頻度のコドンは、図６の前処理におけるＳ１２でコドン頻度表データ記憶部２０３から取得したコドン頻度表データより得る。ここで、図１２（ａ）を用いて第１変異処理について説明する。

[0049] 図１２（ａ）は、個体データに３つのＣＤＳが含まれる例を示す。図１２（ａ）に示されるように、第１変異処理では、個体データに含まれる３つのＣＤＳについて、全てのコドン（５×３＝１５個のコドン）に対して確率Ｐｍで変異処理を実行する。なお、図１２（ａ）中の破線は、確率Ｐｍで変異処理が実行される対象となるコドンの範囲を表すものである。一例として、３つ目のＣＤＳであるＣＤＳ－３に含まれる最初のコドンである「ＧＧＣ」を、確率ＰｍでＧＧＣより高頻度なコドンに置換する。ここで、より高頻度なコドンは、コドン頻度表データから得る。図１２（ａ）の例では、「ＧＧＣ」より高頻度なコドンは、「ＧＧＴ」及び「ＧＧＡ」が存在する。このように、より高頻度なコドンが複数ある場合には、いずれか１つのコドンをランダムに選び、「ＧＧＣ」と置換する。なお、「ＧＧＣ」より高頻度なコドンが存在しない場合、かかる置換はされない。このような置換を、個体データに含まれる全てのコドンに対して実行する。本実施形態では、このような処理を第１変異処理という。ここで、第１変異処理は、後述する第１評価基準に係る最小ＣＡＩ値を大きくすることを意図するものである。

[0050] （第２変異処理）

図１１に戻り、変異処理についてさらに説明する。Ｓ２２３ｂにおいて、変異処理部１０５は、個体データに対して第２変異処理を実行する。具体的には、個体データに含まれるＣＤＳのうち、それぞれのＣＤＳ間又は１つのＣＤＳ内部の異なる部位で連続して一致する塩基配列のうち最長の塩基配列である最長共通文字と重なるコドンを、予め定められた確率Ｐｍで他のコドンに置換する。ここで、図１４を用いて、最長共通文字列について説明する。

[0051] 「最長共通文字列」

図１４に示される個体データは、一例として３つのＣＤＳを含むものである。ここで、各ＣＤＳに含まれる５個のコドンを表す文字列（３個の塩基（＝文字）×５＝１５文字）を、他のＣＤＳ又は１つのＣＤＳ内部の異なる部位に含まれる文字列と対比して、連続して一致する文字列の中で最も長いものを最長共通文字列という。図１４の例では、「ＧＧＣＡＴＣＧＴＣＧＡ」（実線の下線が付された部分）が最長共通文字列となり、その長さ（文字数）は１１である。なお、「ＧＴＣＧＡＧＣＡＧ」（破線の下線が付された部分）も共通文字列であるが、長さが９であり、最長ではないので最長共通文字列とならない。なお、最長共通文字列は、計算機科学における最長共通部分文字列(Ｔｈｅ　ｌｏｎｇｅｓｔ　ｃｏｍｍｏｎ　ｓｕｂｓｔｒｉｎｇ)と呼ばれている概念に相当する。

[0052] 図１２（ｂ）は、個体データに３つのＣＤＳが含まれる例を示す。図１２（ｂ）に示されるように、第２変異処理では、個体データに含まれる３つのＣＤＳについて、ＣＤＳに含まれる５個のコドンを表す文字列（３個の塩基（＝文字）×５＝１５文字）のうち、最長共通文字列と重なるコドンに対して確率Ｐｍで変異処理を実行する。ここで、図１２（ｂ）の例では、最長共通文字列は「ＧＧＣＡＴＣＧＴＣＧＡ」（実線の下線が付された部分）である。図１２（ｂ）の例では、２つ目及び３つ目のＣＤＳであるＣＤＳ－２及びＣＤＳ－３に含まれるコドンのうち、１～４つ目のコドンが最長共通文字列と重なるコドンである。なお、図１２（ｂ）中の破線は、確率Ｐｍで変異処理が実行される対象となるコドンの範囲を表すものである。一例として、ＣＤＳ－３に含まれる最初のコドンである「ＧＧＣ」を、確率Ｐｍで他のコドンに置換する。図１２（ｂ）の例では、「ＧＧＣ」とは異なるコドンとして、「ＧＧＴ」、「ＧＧＡ」及び「ＧＧＧ」が存在する。このように、他のコドンが複数ある場合には、いずれか１つのコドンをランダムに選び、「ＧＧＣ」と置換する。なお、「ＧＧＣ」以外のコドンが存在しない場合には、かかる置換はされない。例えば、特定のアミノ酸をコードするコドンが１種類しか存在しないときには、置換ができない場合があるためである。このよ

うな置換を、最長共通文字列と重なるコドンに対して実行する。本実施形態では、このような処理を第２変異処理という。ここで、第２変異処理は、後述する第２評価基準に係る不一致塩基数を大きくし、最長共通文字列を小さくすることを意図するものである。

[0053] そして、第１変異処理及び第２変異処理が終了すると、かかる計算結果を計算データ記憶部２０４へ出力する。

[0054] 図８に戻り、メインルーチンについてさらに説明する。Ｓ２２において、第ｇ世代親個体集団データに対して変異処理部１０５による変異処理、必要に応じて、交差処理部１０４による交差処理が実行された後、Ｓ２３に進む。

[0055] 次に、Ｓ２３において、子個体集団データ取得部１０３は、第ｇ世代子個体集団データを生成する。以下、交差処理の実行の有無毎に、第ｇ世代子個体集団データの生成の仕方について説明する。

[0056] １．Ｓ２２において変異処理のみが実行された場合

子個体集団データ取得部１０３は、第ｇ世代親個体集団データに含まれるｐ個の個体データが全て変異処理されたｐ個の個体データを、新たに第ｇ世代子個体集団データとする。

[0057] ２．Ｓ２２において変異処理及び交差処理が実行された場合

子個体集団データ取得部１０３は、第ｇ世代親個体集団データに含まれるｐ個の個体データのうち、交差処理が実行されたe個の個体データと、交差処理が実行されていないｐ－ｅ個の個体データを合わせた計ｐ個の個体データが全て変異処理されたｐ個の個体データを、新たに第ｇ世代子個体集団データとする。

[0058] 次に、Ｓ２４において、処理部１０は、第ｇ世代親個体集団データ及び第ｇ世代子個体集団データを統合し、第ｇ世代統合データを生成する。これにより、第ｇ世代統合データには２ｐ個の個体データが含まれることとなる。

[0059] 次に、Ｓ２５において、非優越ソート実行部１０６は、予め定められた評価基準であって、コドン適合度及び前記コドンの塩基配列に関する評価基準

に基いて、第ｇ世代統合データに対して非優越ソートを実行する。そして、２ｐ個の個体データをパレート最適解におけるフロント毎（ランク毎）に分類する。

[0060] 次に、Ｓ２６において、個体選択部１０７は、パレート最適解におけるフロント毎（ランク毎）に分類された第ｇ世代統合データから、ランクの高い順に定められた数の個体データを選択する。なお、個体選択部１０７は、定められた数の個体データを選択するときに、ランクが同じ個体データが存在する場合には、混雑距離が大きいものから順に選択することとしてもよい。ここで、予め定められた数として、ｐを採用することができる。そして、親個体集団データ取得部１０２は、個体選択部１０７により選択されたｐ個の個体データを、第ｇ＋１世代の親個体集団を表す第ｇ＋１世代親個体集団データとして生成し、取得する。以下、図１３及び図１４を用いて、かかる評価基準について説明する。

[0061] ＜評価基準＞

本実施形態では、非優越ソートを実行した２ｐ個の個体データからｐ個の個体データを選択するに際し、２つの観点の評価基準を利用する。かかる観点は、相同組み換えを抑制し、目的タンパク質の生産量を高めることを目的として導き出された観点である。１つ目の観点は、「コドン適合度」に関するものである。具体的には、各個体が複数有する塩基配列であって、アミノ酸翻訳の対象となる塩基配列を表すＣＤＳのコドン適合インデックスの最小値を基準とする。これが第１評価基準である。そして、２つ目の観点は、「コドンの塩基配列」に関するものである。さらに、２つ目の観点は、「不一致塩基数」及び「最長共通文字列」に分かれる。そして、個体データに含まれる２つのＣＤＳのうち、不一致塩基数の最小値を基準とするのが第２評価基準である。また、個体データに含まれるＣＤＳのうち、最長共通文字列の長さを基準とするのが第３評価基準である。以下、これら３つの評価基準の意義について、それぞれ説明する。

[0062] （第１評価基準：コドン適合度）

第１の観点である第１評価基準は、「コドン適合度」に関するものである。ここで、「コドン適合度」とは、個体データに含まれるＣＤＳ中に利用頻度の高いコドンが多く含まれているほど高くなるものとする。具体的には、各個体データに含まれるＣＤＳのコドン適合インデックス（Ｃｏｄｏｎ　Ａｄａｐｔａｔｉｏｎ　Ｉｎｄｅｘ（以下、ＣＡＩという））の最小値（以下、最小ＣＡＩ値という）を基準とする。ＣＡＩは、例えば以下の式で求めることができる。

[0063] [数1]

$$\mathrm{CAI} = \exp\left(1/L \sum_{l=1}^{L} \log\left(w_i(l)\right)\right)$$

$$w_i = \frac{f_i}{\max(f_j)} \qquad i, j \in [\text{synonymous codons for amino acid}]$$

Ｌ：個体データに含まれるコドンの数

ｆｉ：ｉ番目のコドンの頻出度

ｍａｘ（ｆｊ）：最頻出である同義コドン（ｊ番目のコドン）の頻出度

ここで、同義コドンとは、同じアミノ酸をコードするコドンであって、異なる配列を持ったコドンのことである。

[0064] そして、上記の式で求めたＣＡＩを用いて、以下の式で最小ＣＡＩ値を求めることができる。

[0065] [数2]

$$\min_{1 \le i \le x} CAI(C_i)$$

ｘ：個体データに含まれるＣＤＳの数

Ｃｉ：ｉ番目のＣＤＳ

ＣＡＩ（Ｃｉ）：ＣｉのＣＡＩ

[0066] ここで、あるＣＤＳのＣＡＩが高いほど、そのＣＤＳには利用頻度の高い

コドンが多く含まれている（逆に言うとＣＤＳに含まれるレアコドンの数が少ない）ことを示す。そして、ある個体データが、ＣＡＩが極端に低い（換言すると、レアコドンが多く含まれた）ＣＤＳを持っていると、そのＣＤＳは効率的に翻訳されない可能性がある。したがって、最小ＣＡＩ値を第１評価基準として用い、最小ＣＡＩ値が大きいほど、かかる個体データの評価を高くすることにより、ＣＡＩが極端に低いＣＤＳを持つ個体データを最適化の過程で取り除くことが可能になる。したがって、第１変異処理により、最小ＣＡＩ値を大きくすることで、より好ましいシミュレーション結果を得ることができる。

[0067]（第２評価基準：不一致塩基数）

次に、図１３を用いて第２評価基準について説明する。第２の観点のうちの１つ目である第２評価基準は、「不一致塩基数」に関するものである。具体的には、不一致塩基数の最小値（以下、最小不一致塩基数という）を評価基準に用いる。ここで、不一致塩基とは、個体データに含まれるｘ個のＣＤＳのうち、２つのＣＤＳ（以下、ＣＤＳペアという）Ｃｉ及びＣｊを対比して、コドンを構成する塩基が不一致となる塩基のことである。図１３の例では、Ｃｉ及びＣｊを構成する塩基のうち、不一致塩基の数が５個となっている。したがって、かかるＣＤＳペア（Ｃｉ及びＣｊ）の不一致塩基数は５となる。最小不一致塩基数は、以下の式で求めることができる。

[0068] [数3]

$$\min_{1 \le i < j \le x} NN(C_i, C_j)$$

ｘ：個体データに含まれるＣＤＳの数

Ｃｉ：ｉ番目のＣＤＳ

Ｃｊ：ｊ番目のＣＤＳ

ＮＮ（Ｃｉ，Ｃｊ）：ＣｉとＣｊの不一致塩基数

[0069] ここで、ある個体データが、不一致塩基数が極端に低い（換言すると、塩

基配列がよく似た）ＣＤＳペアを持っていると、そのＣＤＳペアの間で相同組み換えが生じる可能性が高くなる。これは、相同組み換えは、塩基配列がよく似た部位（相同部位）で生じるためである。したがって、最小不一致塩基数を第２評価基準として用い、最小不一致塩基数が大きい（換言すると、塩基配列が異なる割合が大きい）ほど、かかる個体データの評価を高くすることにより、塩基配列がよく似た個体データを最適化の過程で取り除くことが可能になる。したがって、第２変異処理により、不一致塩基数を大きくすることで、より好ましいシミュレーション結果を得ることができる。

[0070]（第３評価基準：最長共通文字列）

次に、図１４を用いて第３評価基準について説明する。第２の観点のうちの２つ目である第３評価基準は、「最長共通文字列」に関するものである。すでに述べたように、「最長共通文字列」とは、各ＣＤＳ又は１つのＣＤＳ内部の異なる部位に含まれるコドンを表す文字列を、他のＣＤＳに含まれる文字列と対比して、連続して一致する文字列の中で最も長いもののことである。

[0071] ここで、「全く同じ塩基配列」がゲノム近傍にあると、相同組み換えが生じる可能性が高くなる。これは、前述の通り、相同組み換えは、塩基配列がよく似た部位（相同部位）で生じるためである。したがって、「最長共通文字列」の長さを第３評価基準として用い、最長共通文字列の長さが短いほど、かかる個体データの評価を高くすることにより、「全く同じ塩基配列」が高い割合で含まれる個体データを最適化の過程で取り除くことが可能になる。したがって、第２変異処理により、最長共通文字列を小さくすることで、より好ましいシミュレーション結果を得ることができる。

[0072] 以上説明したように、第１の観点である第１評価基準を用いることにより、利用頻度の高いコドンが多く含まれるＣＤＳを有する個体データを選択することが可能となる。また、第２の観点である第２評価基準及び第３評価基準を用いることにより、塩基配列が異なる割合が大きい個体データを選択し、相同組み換えの発生を抑制することが可能となる。

[0073] 図８に戻り、メインルーチンについてさらに説明する。Ｓ２６において、個体選択部１０７は、パレート最適解におけるフロント毎（ランク毎）に分類された（２ｐ個の個体データを含む）第ｇ世代統合データから、ランクの高い順にｐ個の個体データを選択する。そして、親個体集団データ取得部１０２は、選択されたｐ個の個体データを新たに第ｇ＋１世代の親個体集団データとし、第ｇ＋１世代親個体集団データを生成する。

[0074] 次に、Ｓ２７において、処理部１０は、変数ｇが予め定められた世代数Ｇを超えるか否かを判定する。そして、かかる判定結果がＮＯであれば、Ｓ２８に進む。一方、Ｓ２７における判定結果がＹＥＳであれば、メインルーチンを終了し、かかる計算結果を計算データ記憶部２０４へ出力する。

[0075] Ｓ２７における判定結果がＮＯであれば、Ｓ２８に進み、変数ｇをインクリメントし（つまり、変数ｇに１を加え）、再びＳ２１に戻る。ここで、現時点では変数ｇ＝２であるので、親個体集団データ取得部１０２は、Ｓ２６において生成された第２世代親個体集団データを取得する。かかる処理を、変数ｇが予め定められた世代数Ｇとなるまで繰り返し実行する。換言すると、Ｓ２１～Ｓ２６における処理を２５０回繰り返し実行する。

[0076] 以上説明したメインルーチンを繰り返し実行することにより、３つの評価基準に基いて選択されたｐ個の個体データは、繰り返し回数が増えるほど、遺伝子配列群として好ましいものとなっていく。

[0077] ＜実施例＞

以下、本アルゴリズムを用いた遺伝子配列設計につき、実施例について説明する。かかる実施例では、シミュレーションとして、ヒトのインスリンＡ鎖（アミノ酸配列：ＧＩＶＥＱＣＣＴＳＩＣＳＬＹＱＬＥＮＹＣＮ）をコードする１０個のＣＤＳを設計した。種々のパラメータについては、以下の通りである。

予め定められた確率Ｐｍ（変異率）＝０．０５

Ｐｃ（交差率）＝０．５

第ｇ世の個体集団データ（親個体集団データ、子個体集団データ）に含まれ

る個体データの数ｐ＝１００

予め定められた世代数Ｇ（最大世代数）＝２５０

[0078] 以下、図１５及び図１６を用いて、本シミュレーションにおける計算結果について、第１評価基準を横軸に、第２評価基準を縦軸にとってプロットしたグラフと、第１評価基準を横軸に、第３評価基準を縦軸にとってプロットしたグラフについて説明する。

[0079] 図１５は、第１評価基準を横軸に、第２評価基準を縦軸にとってプロットしたグラフである。ここで、グラフ中にて丸で表されるプロットは第１世代、四角形で表されるプロットは第１０世代、三角形で表されるプロットは第２５０世代における計算結果を示す。なお、１つのプロットは１つの設計結果（＝個体データ）に対応する。すでに述べたように、第１評価基準は最小ＣＡＩ値が大きいほど評価が高いので、グラフ中では横軸の右側にプロットされた点ほど評価が良く、横軸の左側にプロットされた点ほど評価が悪いといえる。また、第２評価基準は、最小不一致塩基数が大きいほど評価が高いので、グラフ中では縦軸の上側にプロットされた点ほど評価が良く、縦軸の下側にプロットされた点ほど評価が悪いといえる。図１５に示されるように、世代数が大きくなるにしたがって（換言すると、図８におけるメインルーチンの繰り返し回数が増えるにしたがって）、個体集団データ全体として好ましいものとなっていることが読み取れる。

[0080] 図１６は、第１評価基準を横軸に、第３評価基準を縦軸にとってプロットしたグラフである。丸、四角形及び三角形で表される各プロットの意味は、図１５と同様である。ここで、第３評価基準は、最長共通文字列の長さが短いほど評価が高いので、グラフ中では縦軸の下側にプロットされた点ほど評価が良く、縦軸の上側にプロットされた点ほど評価が悪いといえる。図１６に示されるように、世代数が大きくなるにしたがって（換言すると、図８におけるメインルーチンの繰り返し回数が増えるにしたがって）、個体集団データ全体として好ましいものとなっていることが読み取れる。

[0081] 以上、種々の実施形態について説明したが、本発明はこれらに限定されな

い。

[0082] 例えば、図８におけるメインルーチンのＳ２６における選択は、第１評価基準及び第２評価基準、又は、第１評価基準及び第３評価基準のいずれか一方を用い、図１５及び図１６に示されるグラフの一方を得ることとしてもよい。また、第１評価基準及び第２評価基準、及び、第１評価基準及び第３評価基準の両方を用いる場合は、世代毎に図１５におけるグラフと図１６におけるグラフからそれぞれ評価の高い個体データを特定し、任意の基準でポイントを付与し、これら２つのグラフにおけるポイントの合計が高い個体データを選択してもよい。もしくは、図１５及び図１６のように２次元のグラフではなく、第１評価基準をｘ軸に、第２評価基準をｙ軸に、第３評価基準をｚ軸にして、３次元のグラフを作成することにより３つの評価基準のそれぞれについて高い評価を得た個体データを選択してもよい。

[0083] また、記憶部２０は、情報処理装置１の内部に設けずに、外部のＰＣ又はサーバ等の情報処理装置に設けるクラウドコンピューティングの態様とすることができる。この場合、計算の度に必要なデータを外部の情報処理装置が情報処理装置１に送信する。

[0084] また、情報処理装置１の機能を実装したＡＳＩＣ（ａｐｐｌｉｃａｔｉｏｎ ｓｐｅｃｉｆｉｃ ｉｎｔｅｇｒａｔｅｄ ｃｉｒｃｕｉｔ）、ＦＰＧＡ（ｆｉｅｌｄ－ｐｒｏｇｒａｍｍａｂｌｅ ｇａｔｅ ａｒｒａｙ）、ＤＲＰ（Ｄｙｎａｍｉｃ ＲｅＣｏｎｆｉｇｕｒａｂｌｅ Ｐｒｏｃｅｓｓｏｒ）として提供することもできる。また、コンピュータに、情報処理装置１の機能を実現するためのプログラムとして提供することもできる。この場合、かかるプログラムをインターネット等を介して配信することもできる。

[0085] さらに、本アルゴリズムとして、多目的遺伝的アルゴリズムである「ＮＳＧＡ－II」を利用することもできる。これは、本アルゴリズムと同様に、ｐ個の最適解をまとめて得ることができるためである。また、組み合わせ最適化アルゴリズムの一種である「シミュレーテッドアニーリング」や「（単目的の）遺伝的アルゴリズム」を利用してもよい。ただし、この場合には、ｐ

個の最適解をまとめて得ることができないので、計算を少なくともｐ回以上繰り返し、ｐ個の最適解を得る必要がある。さらに、これら２つ以上のアルゴリズムの計算結果を混合してもよい。この場合、ｐ以下の任意の整数αを設定し、あるアルゴリズムによる計算結果からα個の個体を選択し、他のアルゴリズムによる計算結果からｐ－α個の個体を選択し、これらを結合したｐ個の個体を用いることとしてもよい。

[0086] さらに、本発明は、

アミノ酸配列、遺伝子数及びコドン頻度表を表すデータに基いて生成されたデータであって、予め定められた数の個体データを含む第１世代の親個体集団を表す第１世代親個体集団データを取得する親個体集団データ取得ステップと、

前記第１世代親個体集団データに含まれる個体に対し、変異処理を実行する変異処理ステップと、

前記変異処理が実行された個体を含む第１世代の子個体集団を表す第１世代子個体集団データを取得する子個体集団データ取得ステップと、

予め定められた評価基準であって、コドン適合度及び前記コドンの塩基配列に関する評価基準に基いて、前記第１世代親個体集団データ及び前記第１世代子個体集団データを統合した第１世代統合データに対して非優越ソート処理を実行し、前記第１世代統合データに含まれる全個体データをパレート最適解におけるランク毎に分類する非優越ソート実行ステップと、

前記ランク毎に分類された全個体データから、前記ランクの高い順に予め定められた数の前記個体データを選択する個体選択ステップと、

を有する遺伝子配列設計方法

として捉えることもできる。

符号の説明

[0087] １：情報処理装置、１０：処理部、２０：記憶部、３０：操作部、４０：表示部、５０：通信部、１００：バス、１０１：個体生成部、１０２：親個体集団データ取得部、１０３：子個体集団データ取得部、１０４：交差処理

部、１０５：変異処理部、１０６：非優越ソート実行部、１０７：個体選択部、２０１：アミノ酸配列データ記憶部、２０２：遺伝子数データ記憶部、２０３：コドン頻度表データ記憶部、２０４：計算データ記憶部、２０５：評価基準記憶部２０５

請求の範囲

[請求項1] アミノ酸配列、遺伝子数及びコドン頻度表を表すデータに基いて生成されたデータであって、予め定められた数の個体データを含む第１世代の親個体集団を表す第１世代親個体集団データを取得する親個体集団データ取得部と、

前記第１世代親個体集団データに含まれる個体に対し、変異処理を実行する変異処理部と、

前記変異処理が実行された個体を含む第１世代の子個体集団を表す第１世代子個体集団データを取得する子個体集団データ取得部と、

予め定められた評価基準であって、コドン適合度及び前記コドンの塩基配列に関する評価基準に基いて、前記第１世代親個体集団データ及び前記第１世代子個体集団データを統合した第１世代統合データに対して非優越ソート処理を実行し、前記第１世代統合データに含まれる全個体データをパレート最適解におけるランク毎に分類する非優越ソート実行部と、

前記ランク毎に分類された全個体データから、前記ランクの高い順に予め定められた数の前記個体データを選択する個体選択部と、

を有する情報処理装置。

[請求項2] 前記個体選択部は、前記予め定められた数の前記個体データを選択するときに、前記ランクが同じ前記個体データが存在する場合には、混雑距離が大きいものから順に選択する、

請求項１に記載の情報処理装置。

[請求項3] 前記親個体集団データ取得部は、前記個体選択部により選択された前記個体データを、第２世代の親個体集団を表す第２世代親個体集団データとし、

前記変異処理部、前記非優越ソート実行部及び前記個体選択部による処理を、予め定められた世代数となるまで実行する、

請求項１又は請求項２に記載の情報処理装置。

[請求項4] 前記コドン適合度に関する評価基準は、各個体が複数有する塩基配列であって、アミノ酸翻訳の対象となる塩基配列を表すＣＤＳのコドン適合インデックスの最小値を基準とする、

請求項１～請求項３のいずれか１項に記載の情報処理装置。

[請求項5] 前記個体に含まれる前記コドン適合インデックスの最小値が大きいほど、前記個体の評価を高くする、

請求項４に記載の情報処理装置。

[請求項6] 前記コドンの塩基配列に関する評価基準は、前記各個体に含まれる２つの前記ＣＤＳのうち、互いに一致しない塩基の数を表す不一致塩基数の最小値を基準とする、

請求項１～請求項５のいずれか１項に記載の情報処理装置。

[請求項7] 前記不一致塩基数の最小値が大きいほど、前記個体の評価を高くする、

請求項６に記載の情報処理装置。

[請求項8] 前記コドンの塩基配列に関する評価基準は、前記各個体に含まれる前記ＣＤＳのうち、それぞれのＣＤＳ間又は１つのＣＤＳ内部の異なる部位で連続して一致する塩基配列のうち最長の塩基配列である最長共通文字列の長さを基準とする、

請求項１～請求項７のいずれか１項に記載の情報処理装置。

[請求項9] 前記最長共通文字列の長さが短いほど、前記個体を高く評価する、

請求項８に記載の情報処理装置。

[請求項10] 前記変異処理部は、

第ｇ世代の親個体集団を表す第ｇ世代親個体集団データに含まれる各個体データに対し、第１変異処理及び前記第１変異処理とは異なる第２変異処理を実行する、

請求項１～請求項９のいずれか１項に記載の情報処理装置。

[請求項11] 前記変異処理部は、

前記各個体に含まれる全てのＣＤＳに対し、前記ＣＤＳに含まれる

前記コドンを、予め定められた確率で前記コドンより高頻度のコドンに置換する第１変異処理を実行する、

請求項１０に記載の情報処理装置。

[請求項12] 前記変異処理部は、

前記各個体に含まれるＣＤＳのうち、それぞれのＣＤＳ間又は１つのＣＤＳ内部の異なる部位で連続して一致する塩基配列のうち最長の塩基配列である最長共通文字と重なる前記コドンを、予め定められた確率で他のコドンに置換する第２変異処理を実行する、

請求項１０又は請求項１１に記載の情報処理装置。

[請求項13] 前記第１変異処理又は前記第２変異処理は、ランダムに選択される、

請求項１０～請求項１２のいずれか１項に記載の情報処理装置。

[請求項14] 前記第１世代親個体集団データに含まれる個体に対し、交差処理を実行する交差処理部を有し、

前記交差処理は、

第ｇ世代の親個体集団を表す第ｇ世代親個体集団データから予め定められた偶数個の個体データを抽出し、前記抽出された個体データから２個の個体データを選択し、前記選択された２個の個体データに対して交差処理を実行する、

請求項１～請求項１３のいずれか１項に記載の情報処理装置。

[請求項15] 前記交差処理部は、

前記選択された２個の個体データである第１個体データ及び第２個体データに含まれる前記ＣＤＳに含まれる前記コドンの境界から交差ポイントを決定し、

前記交差ポイントを境として前記第１個体データと前記第２個体データに含まれる前記コドンを入れ替える、

請求項１４に記載の情報処理装置。

[請求項16] コンピュータを、

アミノ酸配列、遺伝子数及びコドン頻度表を表すデータに基いて生成されたデータであって、予め定められた数の個体データを含む第1世代の親個体集団を表す第1世代親個体集団データを取得する親個体集団データ取得部、

前記第1世代親個体集団データに含まれる個体に対し、変異処理を実行する変異処理部、

前記変異処理が実行された個体を含む第1世代の子個体集団を表す第1世代子個体集団データを取得する子個体集団データ取得部、

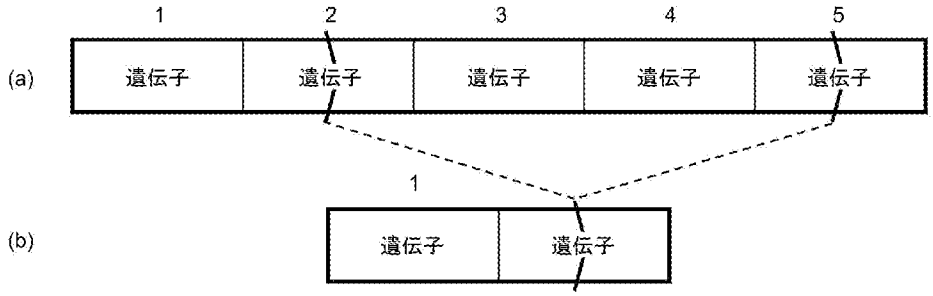
予め定められた評価基準であって、コドン適合度及び前記コドンの塩基配列に関する評価基準に基いて、前記第1世代親個体集団データ及び前記第1世代子個体集団データを統合した第1世代統合データに対して非優越ソート処理を実行し、前記第1世代統合データに含まれる全個体データをパレート最適解におけるランク毎に分類する非優越ソート実行部、

前記ランク毎に分類された全個体データから、前記ランクの高い順に予め定められた数の前記個体データを選択する個体選択部、

として機能させるための情報処理プログラム。

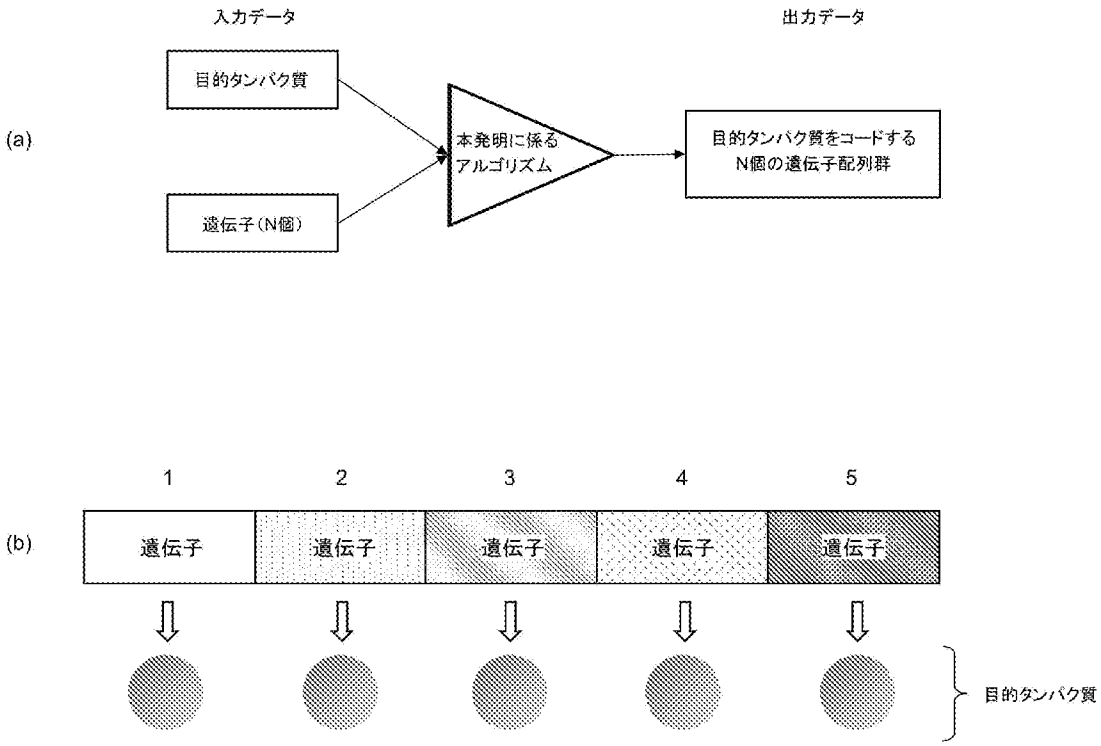
[図1]

図1


1
2
3
4
5
(a)
遺伝子
遺伝子
遺伝子
遺伝子
遺伝子
1
(b)
遺伝子
遺伝子

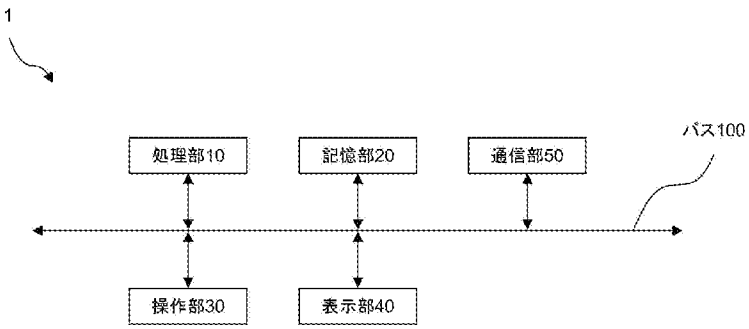
[図2]

図2


入力データ
出力データ
(a)
目的タンパク質
遺伝子(N個)
本発明に係る
アルゴリズム
目的タンパク質をコードする
N個の遺伝子配列群
1
2
3
4
5
(b)
遺伝子
遺伝子
遺伝子
遺伝子
遺伝子
目的タンパク質

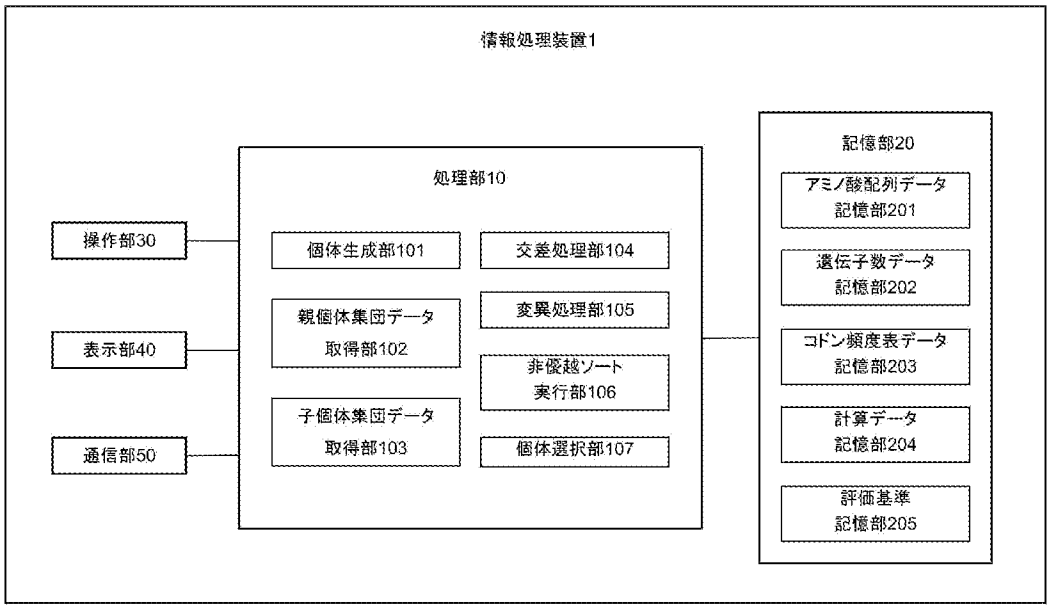
[図3]

図3


1
処理部10
記憶部20
通信部50
バス100
操作部30
表示部40

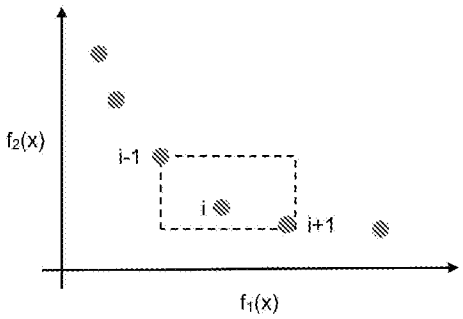
[図4]

図4


情報処理装置1
処理部10
操作部30
表示部40
通信部50
個体生成部101
親個体集団データ取得部102
子個体集団データ取得部103
交差処理部104
変異処理部105
非優越ソート実行部106
個体選択部107
記憶部20
アミノ酸配列データ記憶部201
遺伝子数データ記憶部202
コドン頻度表データ記憶部203
計算データ記憶部204
評価基準記憶部205

[図5]

図5



(a) 混雑距離の概念図

$$CD(x^{(i)}) = \frac{1}{k}\sum_{j=1}^{k}|\tilde{f}_j(x^{(i+1)}) - \tilde{f}_j(x^{(i-1)})|$$

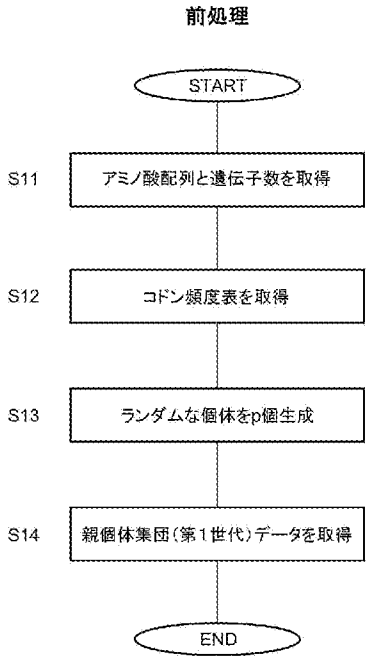
$f_j(x)$:j番目の目的関数

$\tilde{f}_j(x)$:$f_j(x)$をその値域幅で正規化した目的関数

(b) 混雑距離の計算式

[図6]

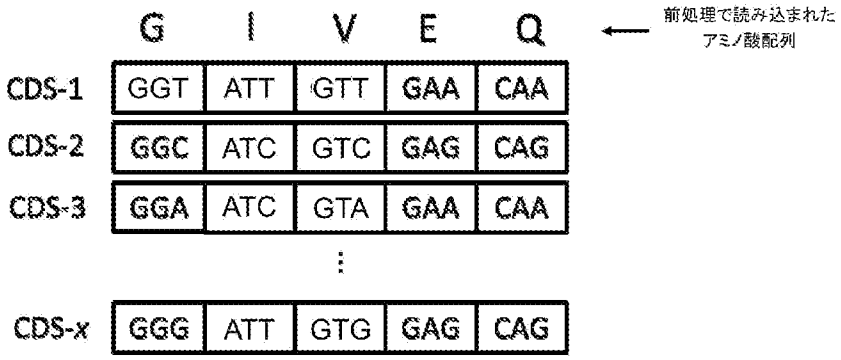
図6



[図7]

図7

個体データの例

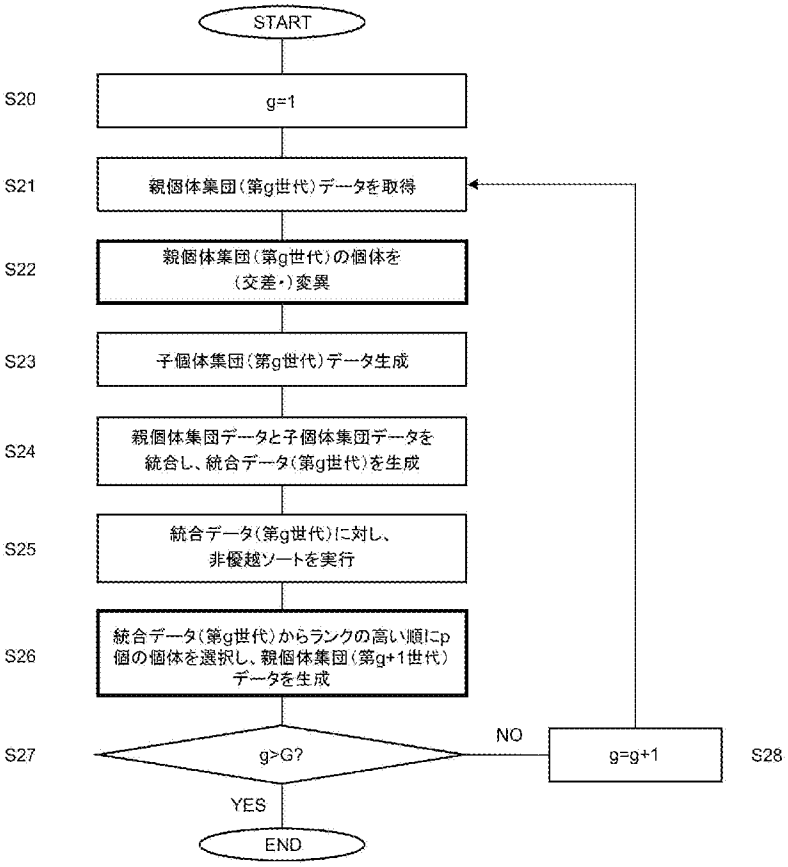


X個のCDSを有する個体
X:前処理で読み込まれた遺伝子数

[図8]

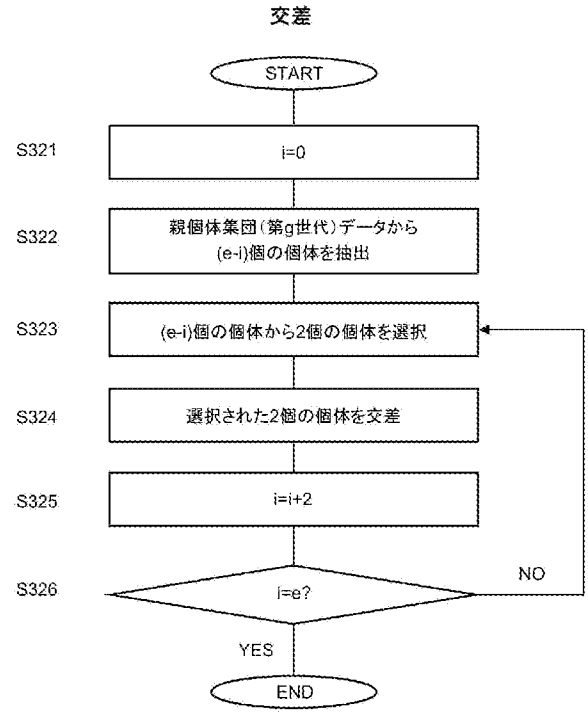
図8

メインルーチン



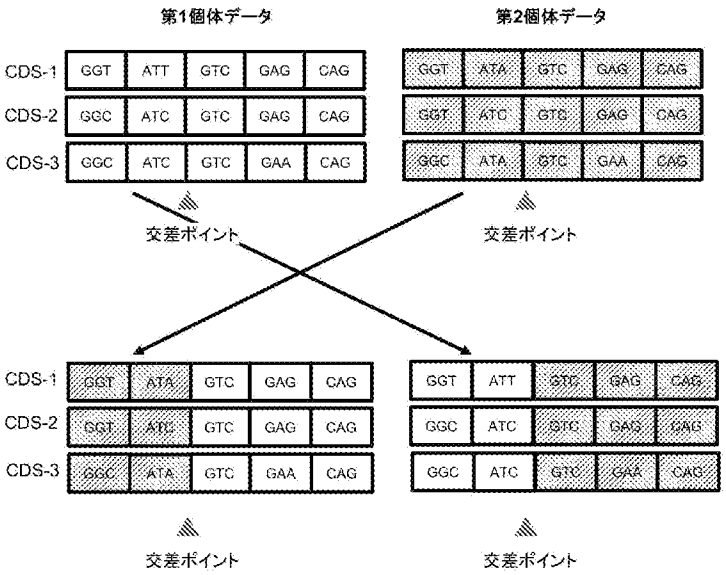
[図9]

図9


交差
START
S321
i=0
S322
親個体集団(第g世代)データから
(e-i)個の個体を抽出
S323
(e-i)個の個体から2個の個体を選択
S324
選択された2個の個体を交差
S325
i=i+2
S326
i=e?
NO
YES
END

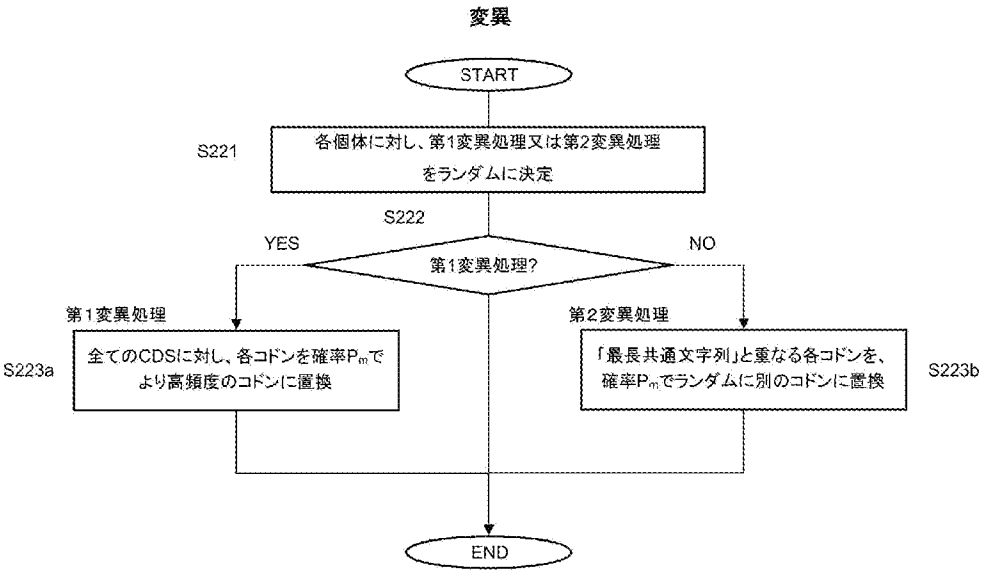
[図10]

図10


第1個体データ
第2個体データ
CDS-1
CDS-2
CDS-3
GGT ATT GTC GAG CAG
GGC ATC GTC GAG CAG
GGC ATC GTC GAA CAG
GGT ATA GTC GAG CAG
GGT ATC GTC GAG CAG
GGC ATA GTC GAA CAG
交差ポイント
交差ポイント
交差ポイント
交差ポイント

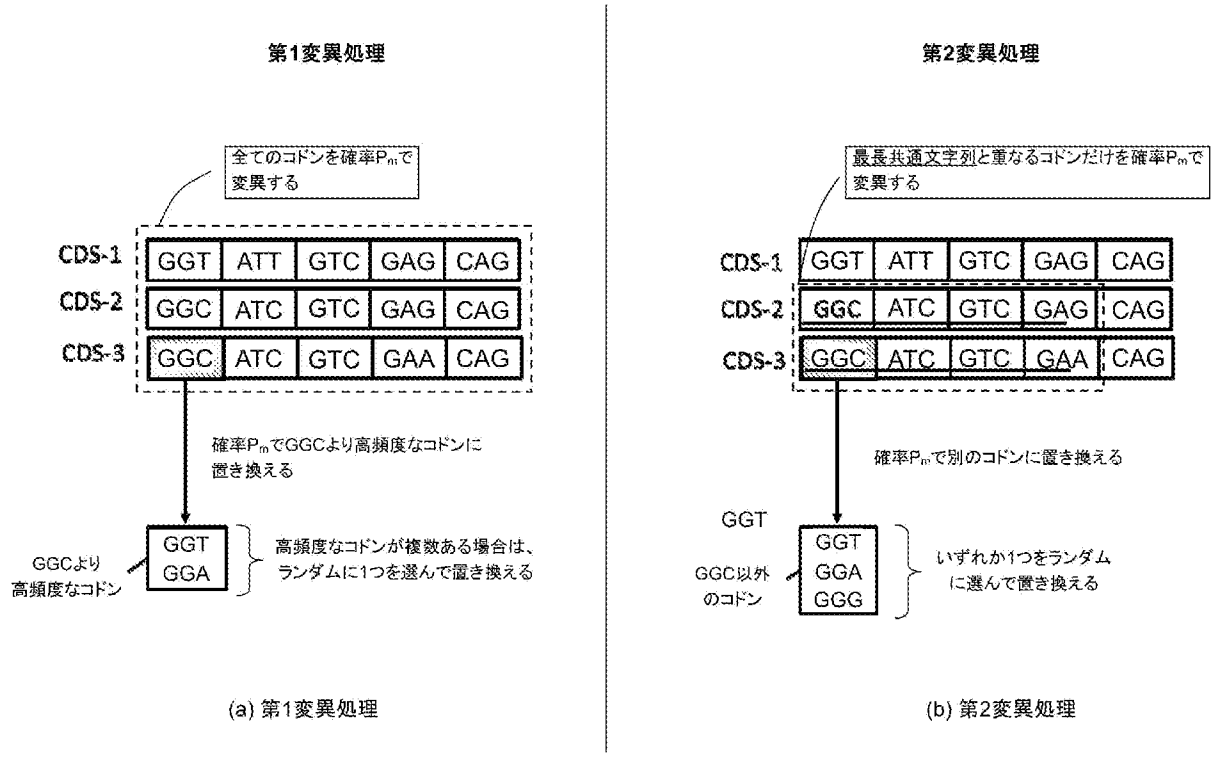
[図11]

図11



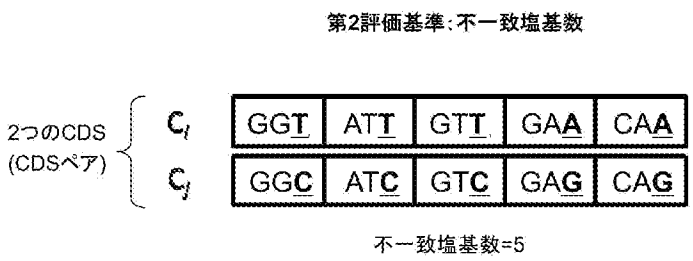
[図12]

図12



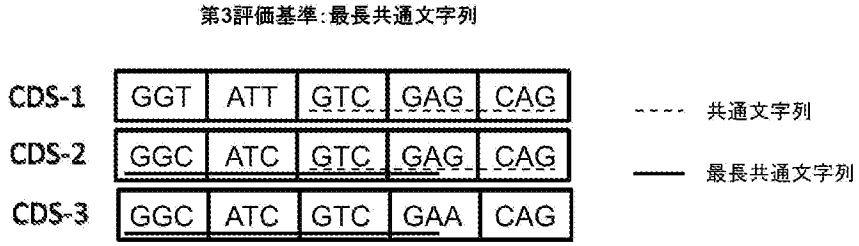
[図13]

図13



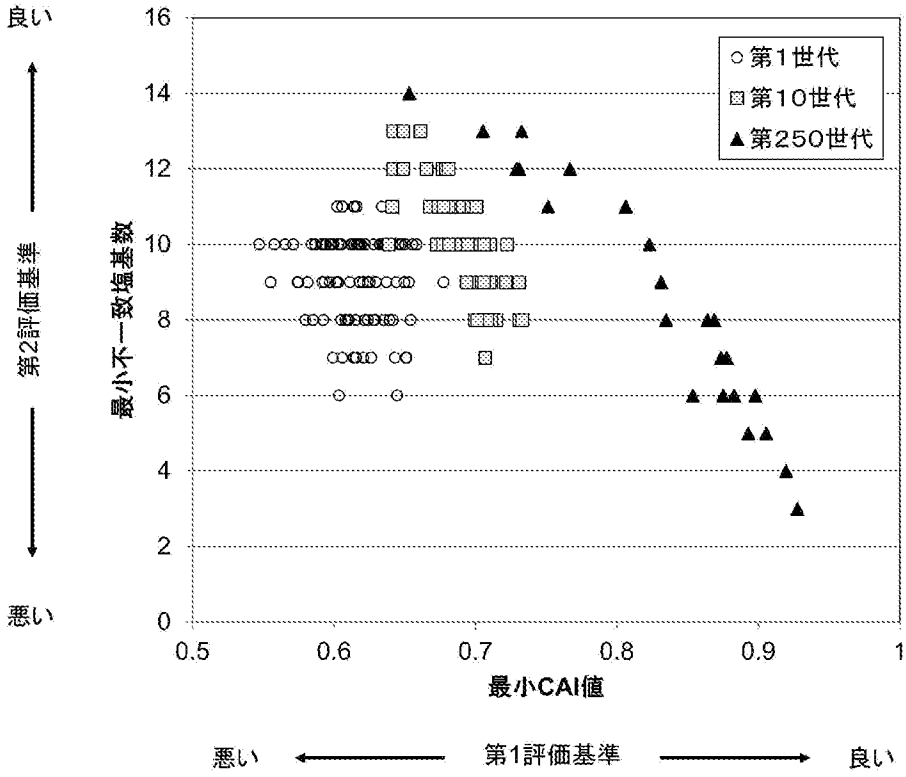
[図14]

図14



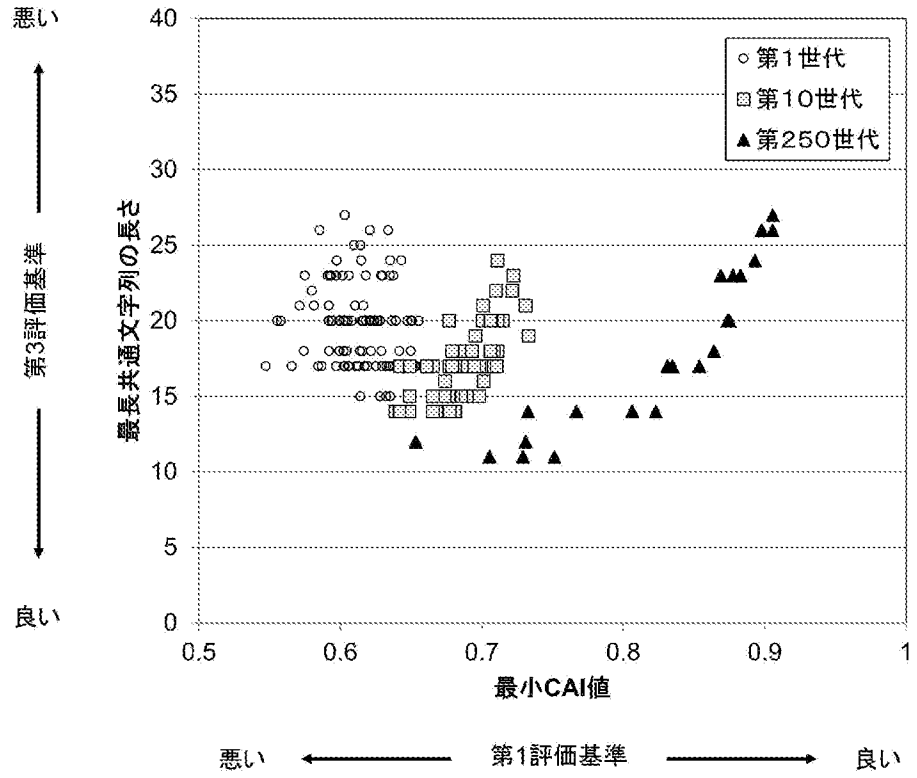
[図15]

図15



[図16]

図16


悪い
第3評価基準
良い
最長共通文字列の長さ
40
35
30
25
20
15
10
5
0
0.5
0.6
0.7
0.8
0.9
1
最小CAI値
第1世代
第10世代
第250世代
悪い
第1評価基準
良い

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2017/010169

A. CLASSIFICATION OF SUBJECT MATTER
G06F19/18(2011.01)i, *C12M1/00*(2006.01)n, *C12N15/09*(2006.01)n

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F19/18, C12M1/00, C12N15/09

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

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2007-172306 A (Yamaha Motor Co., Ltd.), 05 July 2007 (05.07.2007), entire text; all drawings (Family: none)	1-16
A	JP 6-38772 A (P.C.C. Technology Inc.), 15 February 1994 (15.02.1994), entire text; all drawings (Family: none)	1-16
A	Luyi WANG, "Tamokuteki Identeki Algorithm ni Okeru Kai no Seido to Habahirosa no Kojo no Kento", Shushi Ronbun, [online], 23 January 2010 (23.01.2010), pages 1 to 24, [retrieval date 23 May 2017 (23.05.2017)], Internet<URL: http://www.is.doshisha.ac.jp/academic/papers/pdf/09/2009mthesis/2009luy.pdf>	1-16

☐ 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
23 May 2017 (23.05.17)	06 June 2017 (06.06.17)

Name and mailing address of the ISA/	Authorized officer
Japan Patent Office 3-4-3,Kasumigaseki,Chiyoda-ku, Tokyo 100-8915,Japan	Telephone No.

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１７／０１０１６９

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F19/18(2011.01)i, C12M1/00(2006.01)n, C12N15/09(2006.01)n

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06F19/18, 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

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

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

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	JP 2007-172306 A（ヤマハ発動機株式会社）2007.07.05, 全文, 全図（ファミリーなし）	1-16
A	JP 6-38772 A（株式会社ピーシーシーテクノロジー）1994.02.15, 全文, 全図（ファミリーなし）	1-16

☑　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	国際調査報告の発送日
２３．０５．２０１７	０６．０６．２０１７

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　３９９０
毛利　太郎
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	関連する請求項の番号
A	王　路易，多目的遺伝的アルゴリズムにおける解の精度と幅広さの向上の検討，修士論文，［online］，2010.01.23，p.1-24，［２０１７年５月２３日検索］，インターネット<URL:http://www.is.doshisha.ac.jp/academic/papers/pdf/09/2009mthesis/2009luy.pdf>	1-16