(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2005-135154
(P2005-135154A)
(43)公開日　平成17年5月26日(2005.5.26)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 19/00	G06F 19/00	600	4B024
C12N 15/09	G06F 17/30	170F	5B075
G06F 17/30	G06N 3/08	Z	
G06N 3/08	C12N 15/00	A	

審査請求 未請求　請求項の数 8　OL　（全 13 頁）

(21)出願番号　特願2003-370572 (P2003-370572)
(22)出願日　平成15年10月30日 (2003.10.30)

（出願人による申告）平成１５年度、経済産業省、独立行政法人　新エネルギー・産業技術総合開発機構（再）委託研究、産業再生法第３０条の適用を受ける特許出願

(71)出願人　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　100091096
弁理士　平木　祐輔
(72)発明者　上地　潤一
埼玉県比企郡鳩山町赤沼２５２０番地　株式会社日立製作所基礎研究所内
(72)発明者　木村　宏一
東京都国分寺市東恋ヶ窪一丁目２８０番地　株式会社日立製作所中央研究所内

Fターム(参考)　4B024 AA20　CA02　HA11
5B075 ND02　PR06　QM05　UU19

(54)【発明の名称】配列類似性に基づく遺伝子オントロジーターム予測方法

(57)【要約】
【課題】　配列類似性に基づく遺伝子オントロジーターム予測を精度よく行う。
【解決手段】　遺伝子オントロジータームが既知の遺伝子配列を用い、予測条件を変えつつ遺伝子オントロジーターム予測と予測精度の計算を行い、最適予測条件を探索・決定し、この最適予測条件を用い、遺伝子オントロジーターム予測を行う。
【選択図】　図１

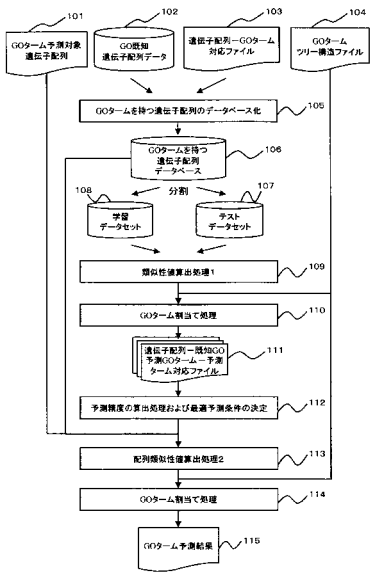


図１

【特許請求の範囲】
【請求項１】
遺伝子オントロジータームの割り当てられた第１の複数の遺伝子配列と、遺伝子オントロジータームの割り当てられた第２の複数の遺伝子配列との間の配列類似性値を用い、遺伝子オントロジー予測の予測精度が十分に高くなる条件を決定する工程と、
第３の遺伝子配列と、遺伝子オントロジーの割当てられた第４の複数の遺伝子配列との配列類似性値を計算し、前記配列類似性値と前記工程で決定された条件に従い、前記第３の遺伝子配列に遺伝子オントロジータームを割り当てる工程とを含むことを特徴とする遺伝子オントロジーターム予測方法。
【請求項２】
遺伝子オントロジータームの割当てられている第１の複数の遺伝子配列各々と、遺伝子オントロジータームの割当てられている第２の複数の遺伝子配列各々との配列類似性値を計算する第１の工程と、
前記工程で計算された配列類似性値を用いて、前記第１の複数の遺伝子配列各々について、前記第１の遺伝子配列との配列類似性値が第１の閾値を超える遺伝子配列を前記第２の複数の遺伝子配列中から選択し、前期選択した遺伝子配列に割当てられている遺伝子オントロジータ－ム中から出現頻度が第２の閾値を超える遺伝子オントロジータームを、前記第１の遺伝子配列の遺伝子の特徴として予測する第２の工程と、
前記第１の閾値および前記第２の閾値を別の値に設定しつつ前記第２の工程を繰り返す第３の工程と、
前記第２～第３の工程で用いた前記第１の閾値と前記第２の閾値毎に、前記第１の閾値と前記第２の閾値を用いた予測結果について予測精度を各々求め、前記予測のうち、十分に高い予測精度に対応する前記第１の閾値と前記第２の閾値を決定し、前記第１の閾値を最適配列類似性閾値とし、前記第２の閾値を最適ターム出現頻度閾値とする第４の工程と、
第３の遺伝子配列と、遺伝子の特徴として遺伝子オントロジータームの割当てられている第４の複数の遺伝子配列各々との配列類似性値を各々求め、前記第３の遺伝子配列との前記配列類似性値が前記最適配列類似性閾値を超えた複数の遺伝子配列を選択し、前記複数の遺伝子配列に割り当てられている複数の遺伝子オントロジータームのうち、出現頻度が前記最適ターム出現頻度閾値を超えた遺伝子オントロジータームを選択し、該遺伝子オントロジータームを前記第３の遺伝子配列の遺伝子の特徴として予測する第５の工程とを含むことを特徴とする遺伝子オントロジーターム予測方法。
【請求項３】
請求項２記載の遺伝子オントロジーターム予測方法において、第１の複数の遺伝子配列および第２の複数の遺伝子配列および第３の遺伝子配列および第４の複数の遺伝子配列は、核酸塩基配列もしくは蛋白質アミノ酸配列であることを特徴とする遺伝子オントロジータ－ム予測方法。
【請求項４】
請求項２～３のいずれか１項記載の遺伝子オントロジータ－ム予測方法において、前記配列類似性値として、遺伝子配列間アライメントから得られる変数又は前記変数の関数を用いることを特徴とする遺伝子オントロジータ－ム予測方法。
【請求項５】
請求項２～４記載のいずれか１項記載の遺伝子オントロジータ－ム予測方法において、前記複数の第２の遺伝子配列および前記複数の第４の遺伝子配列に予め割当てられている遺伝子オントロジータームだけではなく、該遺伝子オントロジータームの上位概念にあたる遺伝子オントロジータ－ム各々についても前期出現頻度を算出し予測対象とすることを特徴とする遺伝子オントロジータ－ム予測方法。
【請求項６】
請求項２～５のいずれか１項記載の遺伝子オントロジータ－ム予測方法において、前記第１の配列との前記配列類似性値が前記第１の閾値を超えた複数の遺伝子配列のうち、前

記第１の遺伝子配列との前記配列相同性値が比較的高いn本の遺伝子配列を選択し、前記n本の遺伝子配列に割り当てられている複数の遺伝子オントロジータームに含まれるある同一の遺伝子オントロジータームの総数をmとしたとき、m/nを前記遺伝子オントロジータームの前期出現頻度とすることを特徴とする遺伝子オントロジーターム予測方法。

【請求項７】

請求項２～６のいずれか１項記載の遺伝子オントロジーターム予測方法において、ある遺伝子配列について予測された第１の遺伝子オントロジータームと前記遺伝子配列に予め割り当てられていた第２の遺伝子オントロジータームが同一である場合と、前記第１の遺伝子オントロジータームが前記第２の遺伝子オントロジータームの下位概念もしくは上位概念に位置する場合に、前記２の遺伝子オントロジータームは正しく予測された遺伝子オントロジータームであるとすることを特徴とする遺伝子オントロジーターム予測方法。

【請求項８】

請求項２～７のいずれか１項記載の遺伝子オントロジーターム予測方法において、前記複数の第１の遺伝子配列に予め割り当てられている遺伝子オントロジータームの総数をAとし、前記複数の第１の遺伝子配列各々について予測された遺伝子オントロジータームの総数をBとし、前記予測された遺伝子オントロジータームのうち正しく予測された遺伝子オントロジータームの総数をCとし、次式（１）を満足する実数をRとし、次式（２）を満足する実数をPとしたとき、実数Rと実数Pの関数を前期予測精度とすることを特徴とする遺伝子オントロジーターム予測方法。

$R = C/A$ ……（１）

$P = C/B$ ……（２）

【発明の詳細な説明】

【背景技術】

【０００１】

本発明は遺伝子配列の情報解析に係わり、配列類似性検索により、遺伝子機能に対応する遺伝子オントロジータームを推定する遺伝子オントロジーターム予測方法に関する。

【０００２】

従来、遺伝子配列の特徴を表す遺伝子オントロジータームを予測する方法として、遺伝子配列がもつ機能モチーフを抽出し、その機能モチーフに対応する遺伝子オントロジータームを選択する方法があった（下記非特許文献１）。配列類似性による方法としては、遺伝子オントロジーを採用した遺伝子配列のデータベースを用いて、配列類似性値がある閾値をこえる遺伝子配列の遺伝子オントロジータームを予測結果としてそのままもちいる方法がある（下記非特許文献２）。

【０００３】

しかし、従来方法は予測精度をより高くすることを意識した方法とはなっておらず、従来方法で用いる閾値は、予測精度をより高くするように決められた値ではない。予測精度には、予測の正解率（Precision値）と、予測の回収率（Recall値）の二つの要素があり、Recall値とPrecision値が共に十分に高くなるような方法が望ましい。

【０００４】

【特許文献１】Apweiler, R., et al., Nucleic Acids Res., 29: 37-40, 2001

【特許文献２】The FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team, nature, 420:563-573, 2002

【発明の開示】

【発明が解決しようとする課題】

【０００５】

本発明が解決しようとする課題は、配列類似性に基づいた遺伝子オントロジーターム予測において、従来方法に比べRecall値とPrecision値が共に十分に高くなるような方法を提供することである。

【課題を解決するための手段】

【０００６】

本発明は、遺伝子オントロジーターム予測のRecall値とPrecision値を共に十分に高めるため、以下の処理工程から構成される方法によって遺伝子オントロジーターム予測を行う。

【０００７】

すなわち、本発明による遺伝子オントロジーターム予測方法は、遺伝子オントロジータームの割り当てられた第１の複数の遺伝子配列と、遺伝子オントロジータームの割り当てられた第２の複数の遺伝子配列との間の配列類似性値を用い、遺伝子オントロジー予測の予測精度が十分に高くなる条件を決定する工程と、第３の遺伝子配列と、遺伝子オントロジーの割当てられた第４の複数の遺伝子配列との配列類似性値を計算し、前記配列類似性値と前記工程で決定された条件に従い、前記第３の遺伝子配列に遺伝子オントロジータームを割り当てる工程とを含むことを特徴とする。

【０００８】

第１の複数の遺伝子配列と第２の複数の遺伝子配列を準備するために、まず、遺伝子の特徴がすでに知られており、その特徴を表す遺伝子オントロジータームが各エントリーに付与されている遺伝子配列データベースを用い、このデータベースを、乱数を用いて２分することにより、第１の複数の遺伝子配列と第２の複数の遺伝子配列を決定することが好ましい。

【０００９】

また、本発明による遺伝子オントロジーターム予測方法は、遺伝子オントロジータームの割当てられている第１の複数の遺伝子配列各々と、遺伝子オントロジータームの割当てられている第２の複数の遺伝子配列各々との配列類似性値を計算する第１の工程と、前記工程で計算された配列類似性値を用いて、前記第１の複数の遺伝子配列各々について、前記第１の遺伝子配列との配列類似性値が第１の閾値を超える遺伝子配列を前記第２の複数の遺伝子配列中から選択し、前期選択した遺伝子配列に割当てられている遺伝子オントロジーターム中から出現頻度が第２の閾値を超える遺伝子オントロジータームを、前記第１の遺伝子配列の遺伝子の特徴として予測する第２の工程と、前記第１の閾値および前記第２の閾値を別の値に設定しつつ前記第２の工程を繰り返す第３の工程と、前記第２～第３の工程で用いた前記第１の閾値と前記第２の閾値毎に、前記第１の閾値と前記第２の閾値を用いた予測結果について予測精度を各々求め、前記予測のうち、十分に高い予測精度に対応する前記第１の閾値と前記第２の閾値を決定し、前記第１の閾値を最適配列類似性閾値とし、前記第２の閾値を最適ターム出現頻度閾値とする第４の工程と、第３の遺伝子配列と、遺伝子の特徴として遺伝子オントロジータームの割当てられている第４の複数の遺伝子配列各々との配列類似性値を各々求め、前記第３の遺伝子配列との前記配列類似性値が前記最適配列類似性閾値を超えた複数の遺伝子配列を選択し、前記複数の遺伝子配列に割り当てられている複数の遺伝子オントロジータームのうち、出現頻度が前記最適ターム出現頻度閾値を超えた遺伝子オントロジータームを選択し、該遺伝子オントロジータームを前記第３の遺伝子配列の遺伝子の特徴として予測する第５の工程とを含むことを特徴とする。

【００１０】

前記第１～４の工程の目的は、実際の予測に先立ち、予測精度が十分に高くなるような予測の配列類似性閾値とターム出現頻度閾値を探索することである。また、予測精度を求めるために、前記第１の配列として、実際の遺伝子の特徴が知られており、その特徴として遺伝子オントロジータームの割当てられている前記第１の複数の配列を用いている。

【００１１】

また、本発明による遺伝子オントロジーターム予測方法は、第１の複数の遺伝子配列および第２の複数の遺伝子配列および第３の遺伝子配列および第４の複数の遺伝子配列は、核酸塩基配列もしくは蛋白質アミノ酸配列であることを特徴とする。

【００１２】

本発明で用いる遺伝子配列は、すべて核酸塩基配列であるか、あるいはすべて蛋白質アミノ酸配列であることが好ましい。

【００１３】
また、本発明による遺伝子オントロジー予測方法は、前記配列類似性値として、遺伝子配列間アライメントから得られる変数あるいは前記変数の関数を用いることを特徴とする。
【００１４】
本発明では配列類似性値として、配列相同性検索ツールBLASTで用いられるE-value値、またはBLAST以外の配列相同性検索ツールで用いるE-value値に相当する値、またはアライメントのアイデンティティ、またはアライメント長、またはアライメントする双方の配列全長に占めるアライメント長の割合を用いることが好ましい。
【００１５】
また、本発明による遺伝子オントロジータームの予測方法は、前記複数の第２の遺伝子配列および前記複数の第４の遺伝子配列に予め割当てられている遺伝子オントロジータームだけではなく、該遺伝子オントロジータームの上位概念にあたる遺伝子オントロジータームの各々についても前期出現頻度を算出し予測対象とすることを特徴とする。
【００１６】
遺伝子配列に予め割当てられている遺伝子オントロジータームの上位概念にあたる遺伝子オントロジータームも予測対象する目的は、上位概念の遺伝子オントロジータームも予測できるようにし、予測のRecall値を高めることである。
【００１７】
また、本発明による遺伝子オントロジータームの予測方法は、前記第１の配列との前記配列相同性値が前記第１の閾値を超えた複数の遺伝子配列のうち、前記第１の遺伝子配列との前記配列相同性値が比較的高いn本の遺伝子配列を選択し、前記n本の遺伝子配列に割り当てられている複数の遺伝子オントロジータームに含まれるある同一の遺伝子オントロジータームの総数をmとしたとき、m/nを前記遺伝子オントロジータームの前期出現頻度とすることを特徴とする。
【００１８】
第１の閾値（配列類似性値の閾値）により選ばれた遺伝子オントロジータームをそのまま予測結果としてしまうと、誤った遺伝子オントロジータームを多く含んでしまうためPrecision値が低下することが分かっている、そこで第１の閾値で選ばれた遺伝子オントロジー各々について、出現頻度を計算し、出現頻度が高い遺伝子オントロジーを選択することにより、より配列類似性値の高い配列に割当てられ、かつ、より高い出現頻度で現れる遺伝子オントロジータームを選別することを目的とする。
【００１９】
また、本発明による遺伝子オントロジータームの予測方法は、ある遺伝子配列について予測された第１の遺伝子オントロジータームと前記遺伝子配列に予め割り当てられていた第２の遺伝子オントロジータームが同一である場合と、前記第１の遺伝子オントロジータームが前記第２の遺伝子オントロジータームの下位概念もしくは上位概念に位置する場合に、前記２の遺伝子オントロジータームは正しく予測された遺伝子オントロジータームであるとすることを特徴とする。
【００２０】
また、本発明による遺伝子オントロジータームの予測方法は、前記複数の第１の遺伝子配列に予め割り当てられている遺伝子オントロジータームの総数をAとし、前記複数の第１の遺伝子配列各々について予測された遺伝子オントロジータームの総数をBとし、前記予測された遺伝子オントロジータームのうち前記正しく予測された遺伝子オントロジータームの総数をCとし、次式（１）を満足する実数をRとし、次式（２）を満足する実数をPとしたとき、実数Rと実数Pの関数を前期予測精度とすることを特徴とする。

$$R = C/A \quad \cdots\cdots (1)$$

$$P = C/B \quad \cdots\cdots (2)$$

【００２１】
実数Rは予測のRecall値に対応し、実数Pは予測のPrecision値に対応する。さらに、前

記予測精度は、実数Rの値と実数Pの値が共に高いほど高い値となる関数を用いる。したがって、予測精度がより高くなる予測条件を探索することにより、Recall値およびPrecision値が共により高くなる予測条件を探索することが可能となる。

【発明の効果】

【００２２】

本発明によれば、配列類似性検索に基づく遺伝子オントロジーターム予測において、十分に高い精度で遺伝子オントロジータームを予測することが可能となる。

【発明を実施するための最良の形態】

【００２３】

以下発明の実施の形態を、図を用いて詳細に説明する。

図１に与えられた遺伝子配列の特徴を遺伝子オントロジータームとして予測する方法において、予め最適予測条件を求めておくことにより高い予測精度で予測を行うことを目的とした、本発明の一実施例における処理の流れを示す。本実施の形態としては、遺伝子配列として、蛋白質アミノ酸配列を用い、蛋白質データベースとしてSIB（Swiss Institute of Bioinformatics）がインターネット上で公開しているSWISS-PROTおよびTrEMBLを用いた場合について説明する。

【００２４】

まず、図１において、101は遺伝子オントロジーターム（以下省略のためGOタームと呼ぶ）を予測する対象である遺伝子配列である。また、102は遺伝子の特徴が既知であり、その特徴に対応するGOタームがエントリーに割当てられている遺伝子配列データである。また、103は102の遺伝子配列データベースの各エントリーとGOタームとを対応付けたファイルであり、SIBがインターネット上で公開しているものを用いる。104は概念の上下関係によりGOタームをツリー状に構造化したデータのファイルであり、Gene Ontology Consortiumがインターネット上で公開しているものを用いる。Gene Ontology Consortium の開発したGOタームには、molecular function、biological process、cellular componentの３つのカテゴリーがあるが、ここでは104のファイルとしてmolecular functionのタームからなるものを用いる。次に、103のファイルを参照し、GOタームを持つエントリーを102中から検索し、GOタームを持つ遺伝子配列からなる105のデータベースを作成する。このとき、実験による証拠に基づいて割当てられたGOタームのみを扱うようにするため、103のファイルを参照し、GOタームのevidence codeがIEA（Inferred from electronic annotation：コンピュータによる機械的なアノテーションによる予測）となっているGOタームは除外するのが好ましい。次に、106のデータベースをテストデータセット107と学習データセット108に分割する。５回のクロスバリデーションテストを行う場合は、106のデータベースを５分割することで、テストデータセット107と学習データセット108を作成する。このとき乱数を用いてランダムに分割するのがよい。次に109の工程で、テストデータセット107の遺伝子配列各々と学習データセット108の遺伝子配列各々との配列類似性値を計算する。配列類似性値の計算には、米国NCBI（National Center for Biotechnology Information）がインターネット上で公開しているプログラムBLAST、あるいは類似のプログラムを用いる。110において、予測条件を変化させながら、予測条件各々について、GOタームを予測し、予測結果を遺伝子配列－既知GOターム－GOターム対応ファイル111に出力する。この110の工程では、学習データセット108の遺伝子配列に割当てられている遺伝子オントロジータームのうち、ある予測条件を満たしたものを選択する。この工程では、108に予め割当てられているGOタームのみではなく、このGOタームの全ての上位概念に位置するGOタームも予測条件を満たしていれば、予測する。あるGOタームの上位概念に位置するGOタームの選択は、GOタームのツリー構造ファイル104を参照して行う。110の工程で得られる111のファイルには予測条件各々について、（１）各遺伝子配列、（２）各遺伝子配列に予め割当てられている既知GOタームおよび（３）予測されたGOタームとの対応関係が記述されている。111のファイルを用い、112の予測精度の算出処理および最適予測条件の決定を行う。この工程では、110で用いた予測条件毎に予測精度を算出し、予測精度が最大となる予測条件（最適予測条件）を決定する。ここまでの工程は、最適予測条件を求め

るためのものである。以降はこの最適予測条件を用い、予測対象遺伝子配列101に対しGOターム予測を行う。113の工程では、GOターム予測の対象となる101の遺伝子配列各々と106のデータベースの遺伝子配列各々との配列類似性値を算出する。114の工程では、得られた配列類似性値と、112の工程で得られた最適予測条件を用いてGOタームの割当てを行い、割当てられたGOタームを予測結果として115に出力する。114の工程ではまた、GOタームツリー構造ファイル104を用いて、110の工程同様、上位概念のタームでも最適予測条件を満たせば予測結果として115に出力する。

【００２５】

図２は、109における、テストデータセットの配列と学習データセットの配列との配列類似性値算出処理１の結果から得られるデータのデータ構造を表す。201は、１本のテストデータセットの遺伝子配列に対応するデータであり、全体のデータはこの繰り返し構造を含む。201は少なくとも、テストデータセットの遺伝子配列を識別する名前及び、そのテストデータセットの配列に予め割当てられている202のGOタームの繰り返し構造及び、テストデータセットの配列と類似性のあった学習データセットの遺伝子配列に関する情報203の繰り返し構造を含む。202はGOタームの情報であり、GOタームとGOタームを識別するGOタームIDを含む。203は少なくとも、学習データセットの遺伝子配列の配列名及び、その配列割当てられているGOタームに関する204の繰り返し構造を含む。204はGOタームの情報であり、少なくともGOタームもしくはGOタームを識別するIDを含む。

【００２６】

図３は、110における、GOターム割当て処理を説明するためのフローチャートである。301の終了判定を含む繰り返し処理により、全ての配列類似性閾値とターム出現頻度閾値の組み合わせについて、以下の処理を行う。302で、配列類似性閾値とターム出現頻度閾値の組み合わせを設定する。303の終了判定を含む繰り返し処理により、テストデータセットの全ての遺伝子配列について以下の処理を行う。303で処理中のテストデータセットの遺伝子配列に対する201に示す情報を読み込む。この中には、202に示す配列類似性の見られた学習データセットの遺伝子配列の情報が複数含まれる。305で、203の複数のデータのうち、配列類似性値が配列類似性閾値を超えていないデータを削除する。ただし、配列類似性値としてBLASTツールのE-value値を用いた場合は、配列類似性値が配列類似性閾値以下のときにデータを削除する。306で、203のデータ各々の持つ204の複数のGOターム各々について出現頻度を算出する。307で、出現頻度が出現頻度閾値以上のGOタームを選択する。308で、選択されたGOタームを予測GOタームとし、テストデータセット配列名と、予測GOタームに関する情報と、202のGOタームに関する情報とを対応付けて109のファイルに出力する。

【００２７】

図４は、109のファイルのデータ構造を表す。401は302で設定した配列類似性閾値およびターム出現頻度閾値に対応するデータであり、全体のデータはこの繰り返し構造を持つ。401は、少なくとも、302で設定した配列類似性閾値および、ターム出現頻度閾値のデータおよび、テストデータセットの全ての配列に関するデータ402の繰り返し構造を含む。402は、少なくとも、テストデータセットの配列名、および402の配列に対して110で予測されたGOタームに関する繰り返し構造403のデータと、402の配列に予め割当てられているGOタームに関する繰り返し構造404のデータを含む。ただし、あるテストデータセットの配列についてＧＯタームが予測できなかった場合は、402は、403のデータを含まない。403および404はGOタームあるいはGOタームID、あるいはその両方の情報を含む。

【００２８】

次に110の工程に含まれる各処理の手順を、具体例を用いて説明する。これらの工程ではある201のデータを扱うが、このデータはすでに205の工程により、加工されている。この201のデータに含まれる、学習データセットの遺伝子配列各々について、その遺伝子の持つGOタームと配列類似性値を取り出し、図５に示す表を作成する。この表に含まれる各GOタームについて、そのＧＯタームの上位概念に位置するターム（親ターム）をすべて選択し、図５の表に追加することで、図６に示す表を作成する。親タームの選択は104のGO

タームツリー構造ファイルを参照し行う。この104のファイルは、ＧＯターム同士の概念の上下関係を図７で示すようなツリー構造で記述している。図７は矢印の矢の向きに下位概念のタームが位置するように表記しており、702のタームＢは701のタームＡにとって下位概念に位置するターム（子ターム）である。次に、図６のような表を参照し、比較的配列類似性の高い（Ｅ-value値の小さい）遺伝子配列に多く出現するＧＯタームを選択する。そのために、配列類似性の高い順（E-value値の小さい順）に上位Ｎ位までのグループに注目し、そのグループ中で出現頻度が出現頻度閾値を超えるＧＯタームを選択する。たとえば、701の上位３位までの配列グループ中でＧＯタームＢは67%の出現頻度で現れている（３配列中２配列）。このように1からnまでの各Ｎの値（nは配列類似性閾値により選ばれた遺伝子配列の本数）において各タームの出現頻度を計算し、図８のような表を得る。そして、307の工程で、この表中で出現頻度がターム出現頻度閾値に満たないＧＯタームを削除し、残ったＧＯタームを予測ＧＯタームとして選択する。ターム出現頻度閾値が７０%の場合、図８の表を参照し、ＧＯタームＡ、Ｂ、Ｄ、Ｅが予測ＧＯタームとして選択される。このようにタームの出現頻度を考慮することでタームＥのように配列類似性値がもっとも高い（E-value値が最も低い）配列に割り当てられていないタームでも、全体的にみて出現頻度の高いタームであれば予測ＧＯタームとして選択することができる。こうして選ばれた予測ＧＯタームを、402で示すデータ構造を持つデータに加工して出力していき、303および301の繰り返し処理を経ながら追加出力することにより、図４全体のデータ構造を持つファイル111を生成する。

【００２９】

次に、予測精度の算出および最適予測精度の決定に関する112の処理内容を詳細に説明する。この工程で現れる情報はすべて、図４に示したデータ構造を持つ111のファイルのデータから得られる。このファイルに含まれる配列相同性閾値およびタームの出現頻度閾値各々についてRecall値およびPrecision値を算出する。Recall値は、テストデータセットの配列全てが持つＧＯタームの総数のうち、予測されたＧＯタームの総数の割合であり、予測すべきＧＯターム全体のうち、とりこぼしなく予測できたＧＯタームの割合である。また、Precision値は予測したＧＯタームの総数に占める正しく予測されたＧＯタームの総数であり、予測の正解率を意味する。ＧＯタームが正しく予測されたか否かは次のように判断する。予測されたＧＯタームＡとテストデータセット配列に予め割り当てられているＧＯタームＢが同一のＧＯタームであれば、ＧＯタームＡは正しく予測されたとする。また、ＧＯタームＡとＧＯタームＢが同一でなくとも、互いに概念の上下関係にあればＧＯタームＢは正しく予測されたとする。次にRecall値とPrecision値を用い、

(F-measure値) = 2(Recall値)(Precision値)／((Recall値)+(Precision値))

によりF-measure値を算出し、この値を予測精度とする。このF-measure値はRecall値とPrecision値が共に高い値になるほど大きな値となる評価尺度である。F-measure値以外にも、Recall値とPrecision値が共に高い値になるほど大きな値となるような評価尺度が他にあれば、その評価尺度を予測精度としてもよい。

【００３０】

以上説明したRecall値、Precision値、F-measure値を、配列類似性閾値およびターム出現頻度閾値毎に計算し、図９で示すような表を得る。この表を参照し、F-measure値が最大となる配列類似性値およびターム出現頻度閾値を求め、それぞれを最適配列類似性値、最適ターム出現頻度閾値とし、この二つの閾値の組み合わせを最適予測条件とする。また、105のデータベースを５分割し、５回のクロスバリデーションテストを行った場合は、F-measure値が最大となる配列類似性値およびターム出現頻度閾値はそれぞれ５つ求められるので、この５つの値の平均値を最適配列類似性値、最適ターム出現頻度閾値とするのが好ましい。

【００３１】

次に、最適配列類似性値と最適ターム出現頻度閾値を用い、101の遺伝子配列に対しGOタームを予測する113～115の処理手順を説明する。この手順は、109～110の処理により111のファイルを出力する手順と基本的に同じである。ただし、109の処理で用いる学習デー

タセット108の代わりに、ＧＯターム予測対象遺伝子配列101を用い、また、テストデータセット107の代わりに105のデータベースを用いる。さらに、最適配列類似性値と最適ターム出現頻度閾値を用いた予測を１回だけ行うため、201のような繰り返し処理は行わない。このような処理により、101の各配列についてＧＯタームを予測し、その結果を115のファイルに出力する。

【００３２】

上記の方法を用いて、最適配列類似性閾値と最適ターム出現頻度閾値の決定を試みた。102の遺伝子オントロジーターム既知の遺伝子配列のデータとしては、公共のタンパク質データベースであるSWISS-PROTのタンパク質アミノ酸配列を用いた。105の工程により、7830の配列が得られた。さらに、これら配列を1560配列ずつ５分割し、5回のクロスバリデーションテストを行った。配列類似性閾値は1から1E-80まで段階的に変化させ、ターム出現頻度閾値は０％から１００％まで段階的に変化させた。そして、これら配列類似性閾値およびターム出現頻度閾値についてそれぞれRecall値、Precision値、およびF-measure値を計算した。５回のクロスバリデーションテストを行ったので、Recall値とPrecision値は５回のテストの平均値を用いた。その結果、配列類似性閾値が0.01、ターム出現頻度閾値60%のときF-measure値が0.63であり最大であった。また、このときのRecall値とPrecision値はそれぞれ0.55、0.75であった。また比較のため、遺伝子の機能予測でよく用いられる配列類似性閾値１Ｅ-10を用い、タームの出現頻度による選択を行わずに予測を行った。その結果、Recall値0.60、Precision値0.39、F-measure値0.47であった。したがってF-measure値の点において本手法の予測精度は比較の方法を上回っており、本手法の有効性が実証された。

【００３３】

以下に、図１０を用いて、本願発明の手順を説明する。

＜手順１＞

GOターム既知遺伝子配列データ1011、遺伝子配列－GOターム対応データ1012を入力し、1005のGOターム既知遺伝子配列をデータベース化するプログラムにより、1014のGOターム既知遺伝子配列データベースを出力する。

【００３４】

＜手順２＞

1006のGOターム既知遺伝子配列をデータベース化するプログラムにより、1014のGOターム既知遺伝子配列データベースを入力し、学習データセット1015とテストデータセット1016を出力する。

【００３５】

＜手順３＞

1006の学習データセットの遺伝子配列とテストデータセットの遺伝子配列との配列類似性値を求めるプログラムにより、学習データセット1015とテストデータセット1016を入力し、1017の学習データセットの遺伝子配列とテストデータセットの遺伝子配列との配列類似性値データを出力する。

【００３６】

＜手順４＞

1007の複数の予測条件で学習データセットの遺伝子配列のGOタームを予測するプログラムを用い、1017の学習データセットの遺伝子配列とテストデータセットの遺伝子配列との配列類似性値データ、および、1013のGOタームツリー構造ファイルを入力し、1018の各予測条件における学習データセットの遺伝子配列の既知GOタームと予測されたGOタームの対応データを出力する。

【００３７】

＜手順５＞

1008の各予測条件での予測精度を算出し最適予測条件を決定するプログラムを用い、1018の各予測条件における学習データセットの遺伝子配列の既知GOタームと予測されたGOタームの対応データを入力し最適予測条件を決定する。

【００３８】
＜手順６＞
1009のGOターム予測対象遺伝子配列とGOターム既知遺伝子配列との配列類似性値を求めるプログラムを用い、1019のGOターム予測対象遺伝子配列、および、1014のGOターム既知遺伝子配列データベースを入力し、1020のGOターム予測対象遺伝子とGOターム既知遺伝子配列との配列類似性値データを出力する。
【００３９】
＜手順７＞
1010の最適予測条件にもとづきGOターム予測対象遺伝子配列に対しGOタームを予測するプログラムを用い、1020のGOターム予測対象遺伝子とGOターム既知遺伝子配列との配列類似性値データ、および、1013のGOタームツリー構造ファイルを用い、1021のGOターム予測結果を出力する。
【００４０】
＜手順８＞
1002ディスプレイ、1003ポインティングデバイスを用い、1002ディスプレイに、1021GOターム予測結果を出力数する。
【００４１】
手順は上記の通りであるが、補助記憶装置に予め記憶しておくべきデータは、少なくとも、GOターム既知遺伝子配列データ1011、遺伝子配列－GOターム対応データ1012、GOタームツリー構造ファイル1013、GOターム予測対象遺伝子配列1019である。他の補助記憶装置の残りのデータは、計算過程で生成することができる。
【図面の簡単な説明】
【００４２】
【図１】本発明の遺伝子オントロジーターム予測方法の全体的な流れを説明するためのフローチャート。
【図２】配列類似性値計算の結果から得られるデータのデータ構造。
【図３】遺伝子オントロジーターム割り当て処理の流れを説明するためのフローチャート。
【図４】予測精度を計算するために用いるデータのデータ構造。
【図５】配列類似性閾値により選ばれた配列に関するデータを表す表。
【図６】配列類似性閾値により選ばれた配列の遺伝子オントロジータームに上位概念の遺伝子オントロジータームが追加されたデータを表す表。
【図７】遺伝子オントロジーのツリー構造を表す説明図。
【図８】最適予測条件の探索による得られるデータを表す表。
【図９】Recall値、Precision値、F-measure値を、配列類似性閾値およびターム出現頻度閾値毎に計算した表。
【図１０】本発明の手順を説明するシステム構成図。
【符号の説明】
【００４３】
101：予測の対象となる遺伝子配列データ、102：遺伝子オントロジータームが既知の遺伝子配列データ、103：遺伝子オントロジータームが既知の遺伝子配列名と遺伝子オントロジータームとの対応ファイル、104：遺伝子オントロジータームのツリー構造ファイル、105：配列類似性値算出処理のために遺伝子オントロジータームが既知の遺伝子配列に関するデータベースを作成する処理、106：配列類似性値算出処理のために用いる遺伝子オントロジータームが既知の遺伝子配列データベース、107：最適予測条件の探索のために用いる遺伝子配列のテストデータセット、108：最適予測条件の探索のために用いる遺伝子配列の学習データセット、109：配列類似性値算出処理、110：遺伝子オントロジータームの割り当て処理、111：遺伝子オントロジー割り当て処理により得られたファイル、112：予測精度の算出および最適予測条件の決定処理、113：遺伝子オントロジー予測対象配列と遺伝子オントロジータームが既知遺伝子配列データベースの配列との配列類似性値算出処理

、114：遺伝子オントロジーターム予測対象遺伝子配列への遺伝子オントロジーターム割り当て処理、115：遺伝子オントロジーターム予測結果データファイル。
202：テストデータセットの配列に予め割り当てられている遺伝子オントロジータームに関するデータ、203：テストデータセットと配列類似性の見られた学習データセットに関するデータ、204：テストデータセットと配列類似性の見られた学習データセットに予め割り当てられている遺伝子オントロジータームに関するデータ。
403：テストデータセットの配列について予測された遺伝子オントロジータームに関するデータ、404：テストデータセットの配列に予め割り当てられている遺伝子オントロジータームに関するデータ。
601：上位３位までの配列グループ。
701～702：遺伝子オントロジーのツリー構造中に含まれるある遺伝子オントロジーターム。

【図１】

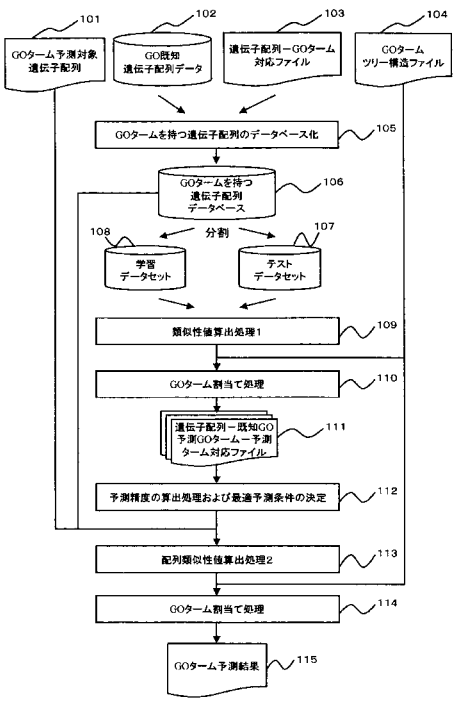


図1

【図２】

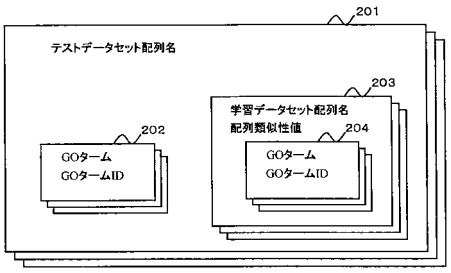


図2

【図３】

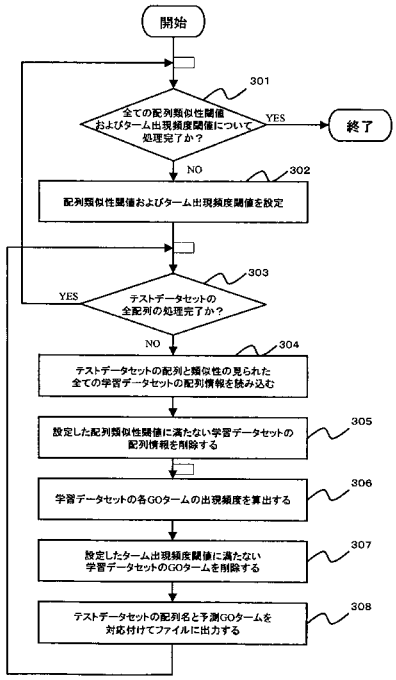


図3

【図４】

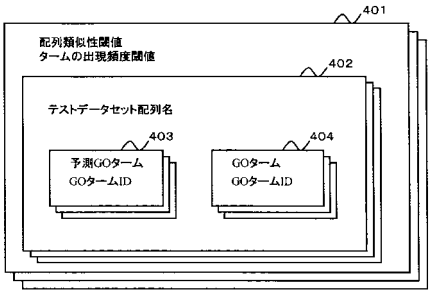


図4

【図５】

配列名	GOターム	E-value値
配列1	D	8E-59
配列2	B, F	1E-45
配列3	C, F	4E-30
配列4	G	8E-23
配列5	F	5E-22

図5

【図６】

配列名	GOターム	E-value値
配列1	A, B, D	8E-59
配列2	A, B, E, F	1E-45
配列3	A, C, E, F	4E-30
配列4	E, G	8E-23
配列5	E, F	5E-22

601

図6

【図７】

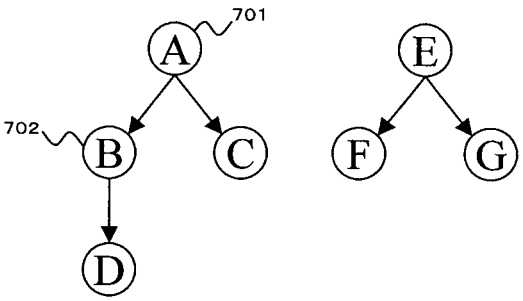


図7

【図８】

グループ＼ターム	A	B	C	D	E	F	G
上位1位までのグループ	100	100	0	100	0	0	0
上位2位までのグループ	100	100	0	50	50	50	0
上位3位までのグループ	100	67	33	33	67	67	0
上位4位までのグループ	75	50	25	25	75	75	25
上位5位までのグループ	60	40	20	20	80	80	20

図8

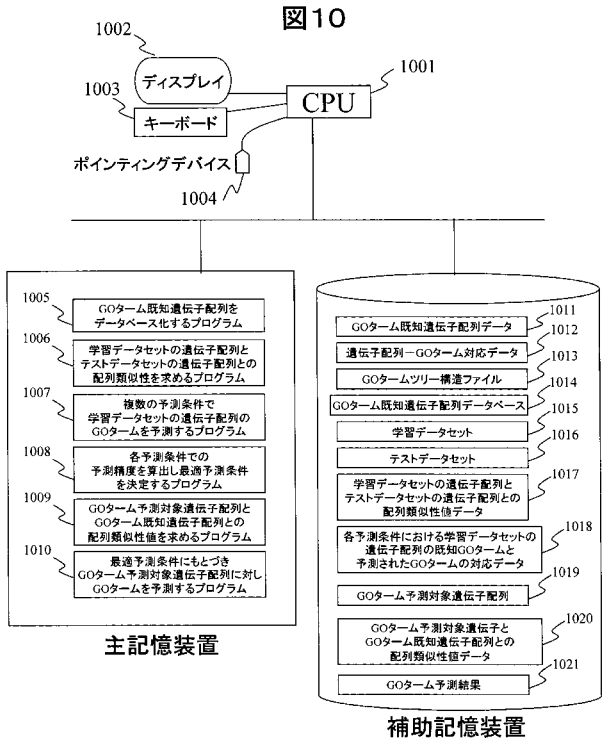
【図９】

配列類似性閾値（E-value値）	ターム出現頻度閾値（%）	Recall値	Precision値	F-measure値
10E-10	40	0.57	0.71	0.63
10E-20	40	0.50	0.73	0.59
10E-30	40	0.45	0.78	0.57
10E-10	60	0.52	0.79	0.62
10E-20	60	0.45	0.81	0.58
10E-30	60	0.40	0.86	0.54

図9

【図１０】

図10


1002
ディスプレイ
1001
CPU
1003
キーボード
ポインティングデバイス
1004
1005
GOターム既知遺伝子配列をデータベース化するプログラム
1006
学習データセットの遺伝子配列とテストデータセットの遺伝子配列との配列類似性を求めるプログラム
1007
複数の予測条件で学習データセットの遺伝子配列のGOタームを予測するプログラム
1008
各予測条件での予測精度を算出し最適予測条件を決定するプログラム
1009
GOターム予測対象遺伝子配列とGOターム既知遺伝子配列との配列類似性値を求めるプログラム
1010
最適予測条件にもとづきGOターム予測対象遺伝子配列に対しGOタームを予測するプログラム
主記憶装置
1011
GOターム既知遺伝子配列データ
1012
遺伝子配列－GOターム対応データ
1013
GOタームツリー構造ファイル
1014
GOターム既知遺伝子配列データベース
1015
学習データセット
1016
テストデータセット
1017
学習データセットの遺伝子配列とテストデータセットの遺伝子配列との配列類似性値データ
1018
各予測条件における学習データセットの遺伝子配列の既知GOタームと予測されたGOタームの対応データ
1019
GOターム予測対象遺伝子配列
1020
GOターム予測対象遺伝子とGOターム既知遺伝子配列との配列類似性値データ
1021
GOターム予測結果
補助記憶装置