遺伝子探索支援システム：GeneDiscovery

特集

Biological Information Annotation System: GeneDiscovery

あらまし

GeneDiscoveryは，cDNAやゲノム配列について，機能解析を支援するシステムである。機能アノテーション情報の自動収集，独自技術の解析手法の搭載，および解析作業の自動化によって，バイオインフォマティクス分野の研究者の作業負担を軽減する。ゲノム創薬において，研究の大半を占める目的の化合物の探査作業に有用なシステムである。

DNA配列の解析として，ホモロジー検索およびLocusLink-dataをもとにした機能アノテーション情報の自動収集，アラインメント解析による機能保存部位の抽出，モチーフの検索が可能である。アミノ酸配列の解析として，アミノ酸変換のあと二次構造予測・柔軟性予測・疎水性予測・抗原決定基予測を利用し，蛋白質の抗原決定部位の予測が可能である。

本稿では，GeneDiscoveryシステムの役割と特徴について紹介する。

Abstract

The GeneDiscovery system supports functional analysis of cDNA and genome sequence data. It helps bioinformatics researchers to reduce their workloads with automatic collection of function-related annotated information, original analytic methods, and automation of analysis work. The system is useful for searching for object compounds, which is the major task in the search for gene-based drugs. For DNA sequence analysis, the GeneDiscovery system enables researchers to collect function-related annotated information automatically on the basis of homology search and LocusLink data, extract meaningful subsets of genes by alignment analysis, and search for motifs. For amino-acid sequence analysis, this system enables researchers to predict the antigen determinant of a protein through prediction of secondary structure, flexibility, hydrophobicity, and antigen determinant base on a converted amino acid. This paper describes the roles and features of the GeneDiscovery system.

児玉貞夫（こだま　さだお）

（株）富士通九州システムエンジニアリングCAD/CAM統括部応用システム部 所属

現在，バイオ系システム開発・商談推進業務に従事。

酒井広太（さかい　こうた）

（株）富士通九州システムエンジニアリングCAD/CAM統括部応用システム部 所属

現在，バイオ系システム開発に従事。

ま え が き

生物の遺伝情報は，DNA（デオキシリボ核酸）の塩基配列に記憶されている。DNA内の遺伝子情報は，細胞内でいったんmRNA（メッセンジャーRNA）として取り出され，アミノ酸配列への翻訳（転写）を経て，アミノ酸配列が立体構造を取り，蛋白質となる。mRNAの相補的配列をcDNAと呼ぶ。

バイオインフォマティクス分野では，遺伝子を構成する核酸配列のシークエンサ（注1）が高速化・高機能化し，大量のゲノム配列（注2）情報が容易に解析されるようになった。ゲノム配列から遺伝子探索（注3）を行うには，cDNA（Complementary DNA）やゲノムデータについて，各種の外部データベースより様々な機能アノテーション情報（注4）を収集し，これらに対して詳細な解析を行う必要がある。通常，これらの情報は，世界中のインターネットサイトに散在し，種々の情報を収集する作業は利用者にとって負担となっている。

GeneDiscoveryはこれらの解析作業の自動化と，独自技術の解析手法の搭載によって，支援機能を中心とした，研究者の作業負担を軽減するためのシステムである。本稿ではGeneDiscoveryシステムの役割と特徴について紹介する。

ゲノム配列と公共データベース

● 公共データベースとホモロジー（相同性）検索

バイオインフォマティクス分野では，過去数十年にわたる配列の解析結果とアノテーション（注釈付け）された膨大なデータが公共データベースとしてインターネット上に公開されている。一般的に遺伝子配列（ゲノム・蛋白質）の解析は，過去に類似するものが公開されているかどうかの検索からスタートする。この検索にはホモロジー（相同性）（注5）検索がよく利用される。

ホモロジー検索により相同性を求めることは，遺伝子の進化と密接な関係がある。遺伝子は長い間にいろいろな変異を経て進化を遂げている。この変異を探し出すために，完全に一致していないが類似している配列を探すことで，同一の祖先を持つであろうと推定されるほかの生物の遺伝子を探し出し，そのアノテーション情報によって機能推定が可能となる。

● 公共データベースの問題点

2002年3月現在，公共データベースGenBankの登録件数は1,600万件，50 Gバイトに達する。加えて，約10万件/日のペースでデイリーな情報が追加されている。この膨大なデータに対しホモロジー検索をかける場合，公共サイトに検索を依頼するか，自研究室にミラー（注6）を構築するしかない。

公共サイトを利用する場合はシークエンスごとに検索指示を行う必要があり，自研究室で検索システムを構築する場合は，ミラーを維持・更新するのに，毎日のデイリー情報のダウンロードと2カ月に一度のリリース情報のダウンロード（インターネット経由でのFTP取得）が必要であり，小規模の研究室ではデータベースの維持は非常に困難である。

GeneDiscoveryの機能

GeneDiscoveryは，小規模の研究室を対象としたシステムである。利用者の作業負担の軽減，維持・更新作業の撲滅，安価なシステム構築を主眼におき設計・開発されている。ホモロジー検索を基本としたアノテーション情報の自動収集機能と，二次構造予測までの機能を実装し，研究の前段の自動化を図っている。また，要望があればカストマイズにより大規模システムにも対応が可能な設計としている。GeneDiscoveryの特徴と機能を以下に示す。

(1)　特徴

・煩わしいホモロジー検索などはプレ処理として自動化する。
・公共データベースを持たずに各種検索を可能とする。
・cDNAに対し，外部DBをホモロジーベースで検索し，様々なアノテーション情報を自動収集する。
・常に最新のデータに対応するよう，定期的な検索を実行可能とする。
・解析したデータはローカルシステムに保存し，いつでも参照・変更を可能とする。

(注1) DNA配列（染色体）を解読する装置。1,000個ほどの長さのDNA配列を読むことができる。
(注2) 個々の生物が持つ遺伝情報全体を示すDNA配列全体のこと。染色体全体の配列。
(注3) 染色体上に散在する遺伝子領域を見つけること。ヒトの場合，染色体上の3～5%が遺伝子をコードする領域であり，残りは機能しないジャンク領域と言われている。
(注4) 発見した遺伝子の機能をいろいろな手法を使って推定し，詳細を記録（コメント付け）する行為，もしくは付加された情報自身を示す。
(注5) 二つのDNA配列を一定の基準に従って比較し，類似性を判定する。通常の文字列比較に，部分的にギャップを入れたり，DNAの組み合わせに得点を与えるなどして類似性にスコアを付けて算出する。
(注6) インターネット上に公開されているWebサイトやデータと同一の物をローカルな環境に構築すること。

(2)　機能
・プレ処理（バッチ処理のホモロジー検索・情報収集）
・解析およびアノテーション付加・編集
・再検索依頼（ホモロジー再検索）
・アミノ酸配列機能予測（二次構造予測など）
・システム管理

cDNA配列の解析

GeneDiscoveryは，とくにcDNA配列解析を強力にサポートするシステムである。現状，全ゲノム配列を決定してから遺伝子産物を決めていくのは困難である。そこで，遺伝子から蛋白質への翻訳過程で生成されるmRNAが注目を浴びている。実際は，不安定なmRNAでなく保存可能な形に変換したcDNAを用いての同定解析が注目を浴びている。網羅的に完全長cDNAの配列決定を行うプロジェクトも推進されている（かずさDNA研究所[1]ほか）。

●　プレ処理の機能

GeneDiscoveryのプレ処理は，配列の解析を行う上で，最低限必要となる検索・解析処理を自動化するものである。利用者がcDNAをQueryとしてシステムに投入すると，システムはバックグラウンドで一連の処理を実行する。Queryをもとにホモロジー検索をかけ，相同性が高い既存の配列群を探し出す。さらに得られた配列群の詳細情報とLocusLink[2]データから得られる情報をもとに，アノテーション情報を収集し，結果をオリジナルGeneデータベースへ書き込む（**図**-1）。プレ処理は以下の特徴を持つ。

(1)　外部サイト（GenomeNet[3]）を利用することで，自研究室内に公共データベースを持つ必要がない。

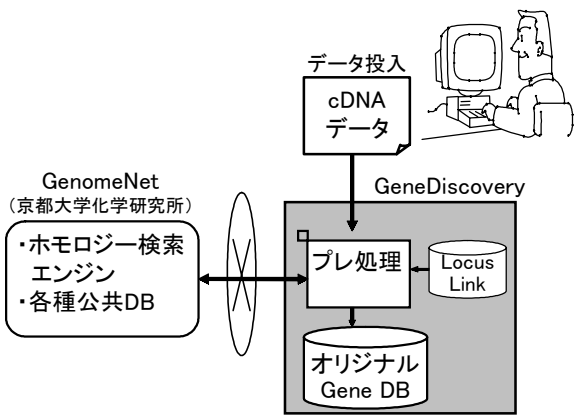

図-1　プレ処理の概要
Fig.1-Outline of pre-processing.

(2)　公共機関を利用するので，大量のデータ処理依頼を行わないよう，10分程度の間隔で一件ずつ依頼する。

検索処理は，公開されているGenomeNetのHTMLインタフェースを利用し，結果はE-mailベースで受け取る。LocusLinkは，容量が小さいことと処理速度の観点から内蔵している。アノテーション情報の収集は，洗い出された情報をもとにGenomeNetから収集している。

なお，ホモロジー検索には，FASTA/BLASTのどちらかが選択できる。

図-2にプレ処理で収集した結果と，**表**-1に収集する情報の一覧を示す。

●　cDNA配列の解析処理

cDNAシークエンスをQueryとして，プレ処理で収集された情報は，オリジナルGeneデータベースに保管されている（**図**-3）。利用者は，この情報と内蔵の解析エンジン（アラインメント解析，アミノ酸部位変換など）を利用しシークエンスの解析とアノテーション付けを行う。

(1)　アラインメント解析とモチーフ検索

相同性と遺伝子の進化とは密接な関係があると述べたが，生物間をまたがって相同性を示す配列部位は何らかの（遺伝的）機能を保存している可能性が高いと考えられる。プレ処理で得られた配列群のマルチプルアラインメントを取ると複数の配列で相同性が高い部位が見いだ

図-2　プレ処理の結果の例（表示形式）
Fig.2-Result of pre-processing.

表-1　プレ処理で収集するアノテーション情報の一覧

項　目	意　味
Locus ID	LocusLinkのID
Locus organism	LocusLinkでの生物種
Locus product	生成される蛋白質
Locus gene_symbol	遺伝子シンボル
Locus gene_name	遺伝子名
Locus phenotype	表現型
Unigene ID	UnigeneへのリンクID
OMIM ID	OMIMへのリンクID
Map	染色体の遺伝的地図位置
Ecnumber	EC番号
dbSNP	dbSNPへのリンクID
Homologene	HomologeneへのリンクID
PubMed ID	PubMed[(4)]へのリンクID
GenBank Definition	GenBankのDefinition情報
GenBank organism	GenBankの生物種
GenBank chr	GenBankの染色体番号
GenBank map	GenBankの遺伝的地図位置
GenBank note	GenBankのNote情報
GenBank tissue_type	GenBankの組織（臓器）

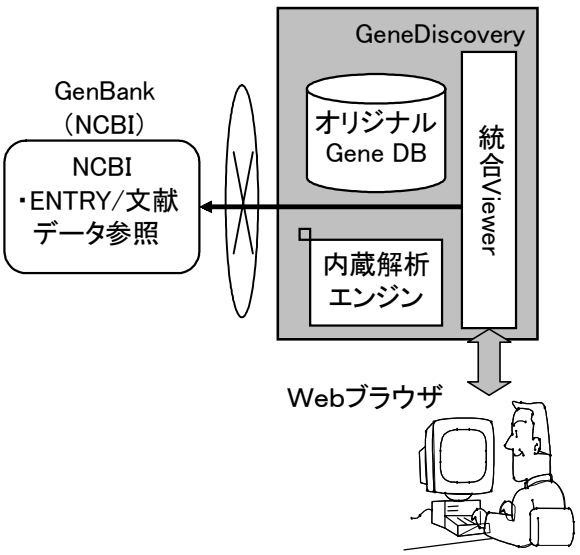

図-3　解析処理の概要
Fig.3-Outline of analytical processing.

される（**図**-4）。

この部位は，生物種間で何らかの機能が保存されている可能性が高いと考えられ，既知の蛋白質に対しモチーフ（機能部位）検索を行うことで，同様の機能を持つ既知蛋白質が検索でき，対象配列の機能推定に利用できる。

図-4　マルチプルアラインメントの結果
Fig.4-Result of multiple-alignment.

図-5　モチーフ検索で得られた蛋白質（PIRの検索）
Fig.5-Result of motif search (retrieval of PIR).

図-5の例では，period clock proteinに関係する蛋白質が多数検索されていることから，同様の機能を持つ配列であると推定できる。

（2）　アミノ酸部位変換

遺伝子が記述されているゲノム配列上には，核酸3個の組み合わせで1個のアミノ酸を表すコードが記述されている。また，アミノ酸変換の開始・終了の印である開始コドン，終了コドンが存在する。GeneDiscoveryは，6枠（1塩基ごとにずらした3パターンの読み枠と順方向・逆方向）と開始・終了コドンで，遺伝子領域を表示し選択・決定する（**図**-6）。

なお，開始コドンは生物種に合わせて選択可能である。またcDNAを対象としているため，イントロンはないと仮定している。

一連の核酸シークエンス解析の完了後は，アミノ酸の

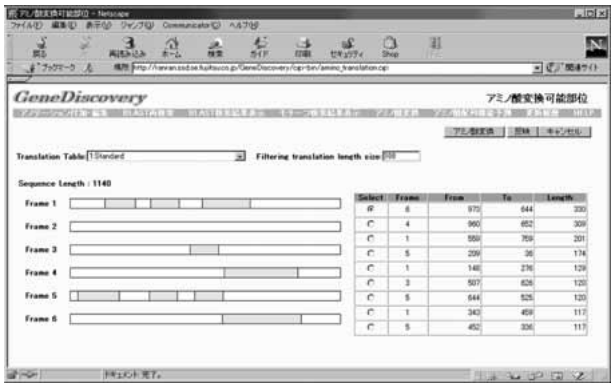

図-6　アミノ酸部位変換の結果
Fig.6-Result of amino acid part conversion.

解析に移る。cDNA配列のアミノ酸部位変換を行った後，アミノ酸の機能の予測に移る。

●　アミノ酸配列の解析処理

アミノ酸配列が決定した後，このアミノ酸配列を持つ蛋白質の機能予測を行う。GeneDiscoveryは，二次構造予測・柔軟性予測・疎水性予測・抗原決定基予測を実装している。本解析手法は富士通の独自技術であり，予測精度もほかに引けを取らない。

本解析には以下の機能予測がある（**図-7**）[5]。

（1）　二次構造予測

アミノ酸配列より蛋白質がとる断片的な構造を予測する。αヘリックス，βシートなどがある。この構造は蛋白質の立体構造と密接な関係を持つ。

（2）　柔軟性予測

蛋白質はそれぞれ固有な立体構造を持つ。その構造を保持するため，機能に関与しない硬い部分と，関与する可能性のある柔らかい部分に分かれる。本機能はこの蛋白質の柔らかい部分を予測する。

（3）　疎水性予測

蛋白質，とりわけ膜蛋白は膜に物質を通すため，水と親和性が高い部分（親水性）と，構造を保持するための水に溶けない部分（疎水性）に分かれる。本機能は蛋白質のこの疎水性部位を予測する。

（4）　抗原決定基予測

抗体は抗原抗体反応において，抗原のどの部分でも攻撃するというわけではなく，抗原の特定の部分を攻撃しやすいという傾向がある。本機能は抗原中で抗体が攻撃しやすい部分（抗原決定基）を予測する。

そのほかの機能

前章までで，cDNAシークエンスからアミノ酸に至る解析の手順を説明した。利用者は，これらの解析結果をもとにcDNAに対してアノテーションを行い，オリジナルGeneデータベースへ保存する。

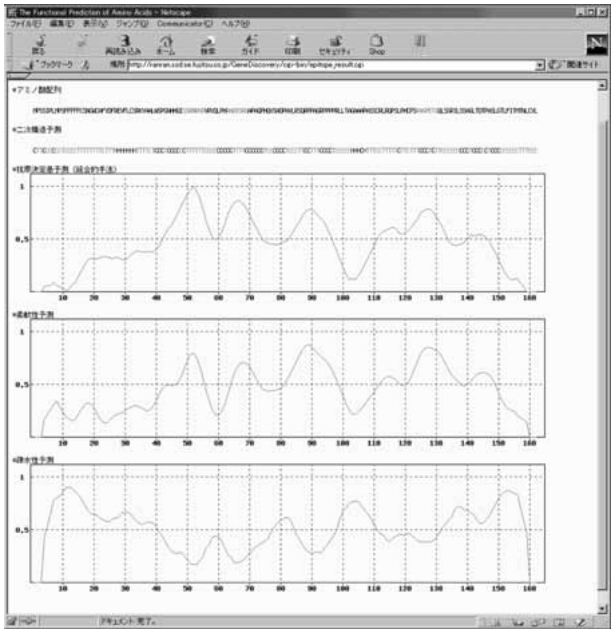

図-7　二次構造予測，抗原決定基予測，柔軟性予測，疎水性予測の結果
Fig.7-Prediction results for second structure, antigenic determinants, chain flexibility and hydrophobicity.

GeneDiscoveryは，アノテーション付加のための様々な機能を用意している。また，システムを維持するための管理機能も実装している。

（1）　スケジュール検索と履歴管理

公共データベースは，日々発見された情報のアップデートがなされている。注意すべき点は最新のデータが常に正しいとは限らない点である。数日たって元に戻されている場合も多々ある。GeneDiscoveryは常に最新のデータに対応するよう，定期的に再検索を行う設定が可能である。また，過去の解析結果の履歴を管理し，情報の喪失を防いでいる。

・指定したcDNAデータの指定日時・指定間隔での自動ホモロジー再検索指示・設定機能。
・検索した結果，データ変更があった場合のみデータの更新を行い，かつ，利用者にE-mailで通知する機能。
・データ更新時は履歴管理を行い，情報を保存する機能。

（2）　アノテーション付加・編集機能

アノテーション付加・編集機能では，以下の項目をcDNAに付加・編集することができる。また，アノテーション情報も履歴管理の対象であり，過去の情報に戻すことが可能となっている。

Definition，Gene Name，Gene Symbol，Gene Family，Organism，Chr. Localization，Category，Medline，OMIM[6] Note，Motif，Comment，EC-Number，Group，Project-Name

（3） ジョブ状態監視機能

GenomNetへの検索依頼は，キュー管理を行い，順次行う。また，依頼中のジョブに対し状況監視・取り消し機能を持つ。

（4） システム管理機能

システム管理（管理者向け）として以下の機能を提供する。

・データ投入（アップロード）機能
・データ削除機能
・オリジナルGeneデータベースの取り出し機能（ダウンロード：GenBank[7]形式）
・システムバックアップ機能
・ユーザ管理機能（登録・削除・変更）

む　す　び

ゲノム・蛋白質の解析ツールの進化は目覚ましく，またヒトゲノムの解析も2003年には完了する予定である。

GeneDiscoveryは，ゲノムを中心に据え，ゲノムから蛋白質までの，一連の解析を可能とするように進化する必要がある。追加機能としては，

・cDNA中心から，ゲノム中心としたシステムへの転換
・SNPs-DBなどの新規データベースとの連携強化
・新規の解析ツールの取込み
・二次構造予測の機能強化
・立体構造のラフな予測と蛋白質シミュレーションへの連携

などを予定している。遺伝子情報からラフな立体構造を予測し，蛋白質構造シミュレーションアプリへデータが渡された時，新たな世界が開かれると信ずる。

参 考 文 献

（1） かずさDNA研究所：http://www.kazusa.or.jp/
（2） LocusLink：Focal point for genes and associated information.，http://www.ncbi.nlm.nih.gov/LocusLink/index.html
（3） GenomeNet：http://www.genome.ad.jp/
（4） PubMed：Public Medline，http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
（5） 中村春木ほか：バイオテクノロジーのためのコンピュータ入門．コロナ社，1995年3月20日初版．
（6） OMIM：Online Mendelian Inheritance in Man.，http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
（7） GenBank：Genetic Sequence Data Bank，http://www.ncbi.nlm.nih.gov/Genbank/index.html