1I1030

創薬設計基盤としての受容体情報システム

○中田 琴子 (国立衛研・情報)

受容体に関する情報を一箇所に集め、創薬設計のための基盤研究に利用し易い形に編集し、WWWを介して提供できるデータベースを開発／更新した。基本となる受容体タンパク質の構造情報（アミノ酸配、二次構造予測、三次構造イメージ）、機能部位情報、遺伝子情報に加え、細胞内信号伝達系（CSNDB）情報、転写因子（Transfac）や転写領域（TRRD）データベース、結合親和力データベース（BADB）情報、および配列類似性情報等が階層的に表示できる。 既存データの拡充のみでなく、データ構造を含めた更新にも容易に対応できるオブジェクト指向データベース管理ソフトウェア ACEDB（A Caenorhabditis elegans Database）を用いてデータベース構築／拡充を行った。Perlで作成したプログラムを ACEDB に組み込み、データの取り込みや編集、解析システムへの投入等機能を拡充した。大型データベースおよび解析システムについては、インターネット上の発信源にシステム内で直接接続することにより最新情報を取り入れることができる。構造および機能情報、受容体とリガンドとの結合親和力情報等の利用に加えて、日本や米国の医薬品データベースと連携することにより、有用性を増している。SNP（Single Nucleotide Polymorphism）情報を取り入れることにより、個々の患者に適した医薬品開発に貢献する情報提供を目指している。URL: <http://impact.nih.go.jp/RDB.html>

K.Nakata : Receptor Information System as the Infrastructure of Drug Design.

1I1100

2種マイコプラズマに見られるゲノム塩基組成の変異の方向性

○中島 広志¹, 山下 紗代², 西川 建³ (¹金沢大・医・保健, ²科技研, ³遺伝研)

生物種によりDNAの塩基組成は異なっており、GC量に依存する。ゲノム上の遺伝子を2塩基あるいは3塩基組成空間に一点で表わしたとき、遺伝子は生物種それぞれの平均近くに密に分布し、距離により遺伝子を定量的に判別できた。この現象は原核生物、古細菌、真核生物において、コーデイング領域、非コーデイング領域に関係なくゲノム全体に見られた。ここではゲノムの全塩基配列が決定されている2種のマイコプラズマ *Mycoplasma genitalium* (Mg) と *Mycoplasma pneumoniae* (Mp) に注目し、以下の解析を行った。まず、2組に共通する294組のオルソログス遺伝子について、配列アライメントをおこない塩基置換パターンを解析した。また、これらのオルソログス遺伝子を10次元の2塩基組成空間にプロットし、得られたMgとMpそれぞれの分布に対して主成分分析を行った。Mg、Mpそれぞれの第一主成分を単位ベクトルとする2次元平面を定義し、その平面上にオルソログス遺伝子を投影した。2つの主軸において2塩基AG/CT成分の相違が大きかった。それぞれの遺伝子は配列の保存度（塩基配列の一致度）に対応して分布の主軸（第一主成分）方向に整列する傾向が認められた。すなわちホモロジーの高い（塩基配列の一致度80％以上）遺伝子から低いもの（60％以下）の順に主軸方向に沿って分布するような傾向である。主軸方向の座標値と配列の一致度（％）の相関係数は、Mgでは0.43、Mpでは0.29であった。以上のことから、2種のマイコプラズマは共通先祖種から分岐した過程で、塩基組成が異なるように変異したこと（GC含量の差は約8％）、また塩基組成の変異の方向が異なっていることから、それぞれの種は種特異的な方向への定向的な淘汰圧を受けていると結論できる。

H.Nakashima,S.Yamashita and K.Nishikawa : Mutational directions of genomes of two mycoplasma species

1I1045

ヒト遺伝子発見プログラムの統合

○矢田 哲士¹、十時 泰¹、高枝 佳男²、榊 佳之¹、高木 利久³ (¹ 理研・G S C、²三菱総研、³東大・医科研)

これまで、数多くのヒト遺伝子の発見プログラムが開発されてきたが、ゲノムスケールの配列から高い信頼性で遺伝子を発見するプログラムは未だ開発されていない。一方、これらのプログラムは、しばしば別々のエキソンを正しく予測することが報告されている。このことは、エキソンの予測傾向が相補的な複数のプログラムを利用することによって、より高い信頼性の遺伝子発見プログラムが開発可能であることを示している。我々が現在進めている研究の目的は、複数の遺伝子発見プログラムの予測結果を統合することで、遺伝子をより高い信頼性で発見する計算機アルゴリズムを開発することである。この研究は以下の手順で進められている。(1) 統合すべき既存の遺伝子発見プログラムの選定、(2) 選定された遺伝子発見プログラムの予測結果を統合するアルゴリズムの開発。(1) では、予測精度が高く、互いに相補的なエキソンを予測するプログラムを選定しなければならない。(2) のアルゴリズムは、選定されたプログラムが各予測エキソンに与えるスコアを考慮するとともに、予測エキソンを統合した結果が遺伝子としての読み枠の一貫性を保っていなければならない。まずはじめに、我々は、比較的最近まとめられた GeneSet'98 データを利用して、20 種類以上の既存の遺伝子発見プログラムについて、それぞれの予測精度と予測の傾向を調べた。その結果、FGENES と FGENESH の予測精度が高く、それらが正しく予測するエキソンが相補的であることが明らかとなった。次に、我々は、HMM の枠組みを応用して、予測エキソンを統合するアルゴリズムを開発した。現在、我々は、いくつかのデータセットを利用して、この HMM の予測精度を検証している。

T.Yada, Y.Totoki, Y.Takaeda, Y.Sakaki and T.Takagi : Integrating human gene finding programs

1I1115

タンパク質ドメインリンカーの配列解析に基づくドメイン予測法の構築

○宮崎 哲^{1,2}、黒田 裕³、横山 茂之^{1,2,3} (¹ 東大・院理・生化、² 理研・細胞情報伝達、³ 理研・G S C)

タンパク質のドメイン単位を知ること、は、構造・機能解析を行う上で大きな意味をもっている。例えば、高分子量タンパク質の構造決定を円滑に進めるためには、タンパク質を構造ドメインごとに断片化し、その各々について解析を行うのが現実的な方法となるが、その際、切断すべきドメイン境界の決定が構造解析の律速段階となっている。タンパク質の機能解析においては、機能単位を同定する際に構造ドメインの情報が重要な鍵となり得る。また、相同性検索や高次構造予測においては、ドメインの正しいアサインメントがよりよい結果を得るためにも必須である。こうした場面において、配列から直接ドメイン単位を同定するための方法が必要となる。そこで我々は、ドメイン同士をつなぐリンカー領域に着目し、その配列上の特徴抽出を通して配列からドメイン単位の予測を試みた。本研究では、ドメインリンカーの特徴抽出に際してパターン認識の一手法であるニューラルネットワークを用いた。SCOPのドメイン定義に基づきドメインリンカーのデータセットを準備した後、ドメインリンカーを含む配列とそれ以外の配列の識別をニューラルネットワークに学習させた。その結果、ドメインリンカーにおいて識別可能な配列上の特徴があることがわかった。次にこのニューラルネットワークを使って、アミノ酸配列からドメインリンカー領域の予測を試みた。各予測領域にスコアに応じて順位をつけたところ、順位1で予測された領域のうち約50%が正解であった（そのとき約35%のドメインリンカーが予測された）。この結果は、単に長いループ領域をドメインリンカーと予測した場合に比べて良い結果であった。このことは先の特徴抽出の結果からも示唆されたようにドメインリンカーが単なるループ領域ではなく、それとは異なった情報も持っていることを示している。

S.Miyazaki, Y.Kuroda, S.Yokoyama : Prediction of domain structures using a neural network analysis of domain linkers