3H0945

相同配列解析に基づく自動的なドメイン推定

○黒田 裕¹、松尾 洋²、谷 一寿²、横山 茂之¹ (¹ 理化学研究所・ゲノム科学総合研究センター・タンパク質構造機能研究グループ、² 理化学研究所・ゲノム科学総合研究センター・ゲノム情報科学研究グループ)

最近、NMRは、X-線結晶解析と並んでタンパク質の構造解析の有力な手法になっている。しかし、NMRで容易に構造解析を行えるタンパク質の分子量の限界は2万前後である。そこで、我々はNMRを中心とした構造ゲノム科学プロジェクトを支援するために、安定に構造を形成する蛋白断片を自動的に検出する手法を開発している。一般的に保存されている配列はドメイン単位を形成すると考えられ、Pfamなど配列比較を基にしたドメインライブラリーが発表されている。これらのドメインの多くは、配列上のモチーフから推定され、機能的にも保存された重要なドメインであると考えられる。我々は、単独で構造形成する蛋白断片を同定する手がかりが、配列保存性に基づいて得られないかと考えた。そこで、query タンパク質に相同な配列を調べて、各残基で検出されていた相同配列の数を調べた。すると、同じタンパク質のアミノ酸配列でも相同配列が多く検出される領域（よく保存されている）と少ない領域（あまり保存されていない）があった。我々はテストデータとしてPDBに構造が登録されているタンパク質の断片を含む全長のアミノ酸配列を解析し、周囲より相同配列の多い領域を選んで構造の有無を調べた。その結果、相同配列の多い領域は約50%の確率で他と独立した構造を形成していることが明らかになった。即ち保存性の高い領域は、単独でも構造を形成するドメイン単位との相関が高く、それらを同定する手法になりうることが明らかになった。

Y.Kuroda, Y.Matsuo, K.Tani, S.Yokoyama : Automated prediction of structured protein domains based on sequence similarities

3H1015

構造ゲノム科学における構造解析の対象となるタンパク質ドメイン領域の自動推定システム

○柳 在圭、黒田 裕、松尾 洋、谷 一寿、河合 純、林崎 良英、横山 茂之（理化学研究所 ゲノム科学総合研究センター）

構造ゲノム科学プロジェクトはゲノムにコードされる全ての蛋白質ファミリーから代表構造を決定することを目指している。そのため、蛋白質構造決定の行程を自動化し、機械的な作業にすることが必要とされる。しかし、現在の構造決定技術、特にNMRを用いた場合、構造決定の自動化が可能なものは分子量が2万以下の小さなタンパク質に限られる。分子量の大きなタンパク質の構造も構造ドメイン単位に分解することにより迅速な構造決定に対応が可能になる。そこで、本研究では、構造ゲノム科学をサポートするものとして、構造決定の対象となるメイン領域を自動推定するシステムの開発を提案する。手順としては、まず、本研究で開発した手法を用いてアミノ酸配列の相同性に基づいて核酸配列におけるFrameshift errorを検出する。次に、予測された真のアミノ酸配列を基に個々の残基における相同配列に関するスコアを与えることによりドメイン領域を予測する。さらに、予測したドメイン領域から構造既知の領域や膜通関領域を除き、類似した配列を同じファミリーにまとめる（Clustering）。全ての蛋白質ファミリーから代表配列を選び、その構造解析の適性を評価し、構造決定のターゲットとしての順位を付ける。実践的な応用例として、本手法を用いてmouse cDNAプロジェクトで得られた11163の遺伝子配列のドメイン解析を行った。その結果、NCBI/nrデータセットに相同配列が存在する遺伝子が7602個あり、本手法で6443の構造ドメイン単位を予測した。その中から、既知の構造領域と膜通関領域を除いた70aa以上の5340個ドメインを2440のファミリーにまとめた。これらのドメイン配列を構造解析適用性に従って順位付けた。

J.K.Yoo, Y.Kuroda, Y.Matsuo, K.Tani, J.Kawai, Y.Hayashizaki and S.Yokoyama : A structural genomics targets selection procedure based on protein dissection

3H1000

構造ゲノミクスにおける立体構造決定ターゲット選択支援のためのシステム開発

○谷 一寿、長内 隆、黒田 裕、松尾 洋、横山 茂之（理研・ゲノム科学総合研究センター）

理化学研究所ゲノム科学総合研究センター（GSC）で取り組まれている構造ゲノミクスプロジェクトを推進するために、立体構造決定ターゲット蛋白質選択支援のためのシステムを開発している。

マウス cDNA プロジェクトなどから得られる新規遺伝子配列及び公共データベースから得られる遺伝子配列を収集し、全体及び局所的な類似性に従ってそれらをファミリーに分類する。そして、それらの配列データを用いて、ホモロジー検索・立体構造予測・機能推定・ホモロジーの出現する生物種の系統パターンの推定などの様々な情報解析を行う。その結果に基づいて、蛋白質ファミリーの間に立体構造決定に関する優先順位を与える。各ファミリーからは代表蛋白質が選ばれ、その構造決定が試みられることになる。また、立体構造決定プロジェクトなどから得られる立体構造情報の結果を逐次取り込むことにより、ファミリー間の優先順位評価の過程へフィードバックさせる。

このような構造決定ターゲット選択のために必要な情報システムを設計運用するために、データベース技術の導入を行い、このデータベースを容易に利用できるような Web インターフェースを作成した。今後、どのようなターゲット蛋白質の構造決定がどの段階にまで進んでいるかに関する情報を、各国のプロジェクトから収集するとともに、たとえば米国の PRESAGE プロジェクトなどに協力して、我々のプロジェクトからも発信する予定である。

K.Tani, T.Osanai, Y.Kuroda, Y.Matsuo and S.Yokoyama : Development of a system for selection and prioritization of structure-determination targets in structural genomics.

3H1030

二次構造予測および相同配列のアラインメントを利用したタンパク質ドメインリンカーの同定

○田中 剛範、黒田 裕、横山 茂之（理研ゲノム科学総合研究センター・タンパク質構造機能研究グループ）

分子量の大きなタンパク質ほど、実験的にその立体構造を決定することは困難である。そのため大きなタンパク質については、一旦適当な大きさのドメインに分割した後、それぞれの立体構造を解くという手順がよく用いられる。しかしそのような部分配列からなるペプチドに、構造解析に適した安定な立体構造を形成させるには、そのタンパク質におけるドメイン間のリンカー領域を正しく同定し、切断することが重要である。簡便なリンカー予測手法の一つとして二次構造予測が考えられるが、それによって提示される多数のループ構造候補のうち、どれが正しいリンカー領域であるかを直ちに判断することは困難である。そこで我々はこれに相同タンパク質から得られる情報を付加し、リンカー領域の持つ特徴によるふるい分けを組み合わせて、正確なドメインリンカー予測を行なう事を目指した。まず、構造が既知であるマルチドメインタンパク質の配列集合から、ループ構造をとっている配列を抽出した。その中における各アミノ酸残基の出現頻度を調べたところ、リンカー領域と非リンカー領域の間で明らかに出現頻度の異なっていた残基を見出した。さらに、複数の相同タンパク質とのアラインメントを行なった上で、ループ部分における各残基の保存性を調べたところ、やはり特定の残基でリンカー／非リンカーの間に差が見られた。そこで、二次構造予測で得られたループ候補に対し、これらの解析から得られた特徴を指標にして取捨選択を行なった結果、リンカーの予測精度を実験上有効なレベルにまで向上させることが出来た。

T.Tanaka, Y.Kuroda and S.Yokoyama : Identification of domain linker in proteins by using secondary structure prediction and aligned homologous sequences