完全長 cDNA 間の共通ドメイン検索システムの開発

大阪大学大学院情報科学研究科　　○松田秀雄

Development of a conserved domain retrieval system between full-length cDNA sequences

Hideo Matsuda
Department of Informatics and Mathematical Science,
Graduate School of Engineering Science,
Osaka University.

Development of a conserved domain retrieval system between full-length cDNA sequences
Hideo Matsuda, Graduate School of Information Science and Technology, Osaka University
The project intends to analyze mouse full-length cDNA (complementary DNA) sequences determined by RIKEN Genome Exploration Research Group, which provide important means to identify biological mechanisms of higher order mammals. We have developed a system that can systematically retrieve novel conserved domains between protein sequences translated from the cDNA sequences. In the system, we took a graph theoretic approach that identifies conserved domains as highly-connected components among possible subsequences expressing similarity to each other.

１．はじめに

高等動物の遺伝子は、それらの配列が複雑なドメイン構造を有し、その組合せにより発生・分化・情報伝達などの多様な機能を実現していると考えられている。このように多様なドメイン構造を計算科学技術により明らかにするため、本プロジェクトでは、タンパク質配列の間に共通して存在するドメイン(共通ドメイン)を検索するシステムである共通ドメイン検索システムの開発を行った。本プロジェクトと並行して、参加研究グループの一つである理化学研究所のグループが 2002 年までに総計で約 6 万個という大量のマウス完全長 cDNA 配列の決定に成功したため、本プロジェクトではこの配列データを利用することができた。完全長 cDNA 配列は、タンパク質をコードするための完全な配列情報を持っており、コード領域を予測できれば容易にタンパク質配列を推定できるという長所を持っている。マウス機能アノテーション(FANTOM)プロジェクト[1]により、上記の完全長 cDNA 配列においてコード領域の予測[2]が行われており、これにより得られた信頼性の高いタンパク質配列を共通ドメイン検索システムの開発過程で利用することができた。

２．共通ドメイン検索システムの開発

共通ドメイン検索システムの開発を始めた当初、大阪大学グループはグラフ理論に基礎をおいた最大密度部分グラフ(MDS)法により、微生物ゲノムプロジェクトから得られたタンパク質配列間の共通ドメインを検出しており、ここで使われたシステム[3]を原型に共通ドメイン検索システムの初期バージョンを開発した。これを、マウス cDNA 配列から推定されたタンパク質配列に予備的に適用した結果、微生物ゲノムでのドメイン解析ではそれほど必要性のなかった、類似配列による冗長性の除去のための配列クラスタリング[4,5]、cDNA 配列のゲノムへのマッピング[6,7]、cDNA 配列中に挿入された繰返し配列のマスク、膜貫通領域、コイルドコイル領域、細胞内局在シグナル配列領域などの領域予測などの前処理が必要であることがわかり、これらが行えるよう共通ドメイン検

索システムを拡張した[8]。また、マウスのタンパク質配列中には、既に多くのドメインの存在が知られており、既知のものを検索するよりは、今までに知られていない新規の共通ドメインの検索に集中すべきであると考えられた。そこで、EBI の InterPro プロジェクトの協力を得て既知ドメインのデータを入手し、共通ドメイン検索システムでは検索されたドメインからこれらの既知ドメインを除く処理を加えることにした。

新規ドメイン候補として検索された共通ドメインの妥当性の評価では、そのドメインを持つ遺伝子の機能情報、遺伝子が発現している組織と時期の情報や、当該遺伝子産物がどのタンパク質と相互作用を示すかの情報が有用であることがわかった。システム評価のための遺伝子機能情報としては、FANTOM アノテーション会議において研究者により注意深く付加されたアノテーション情報をもとに、それぞれの遺伝子を機能により分類したものを使った。得られた共通ドメインをもつ複数の遺伝子が､機能により分類されたひとつの遺伝子ファミリーに属することが示された場合には、その共通ドメインが実際に機能をもつ意義のあるものと判定できる。一方、同一の共通ドメインを有する複数の遺伝子が、同一の遺伝子ファミリーに属さない場合は、その共通ドメインが偶然の類似による偽物である可能性が高いと判定でき、排除できることが示唆された。

さらに、この遺伝子機能情報を格納したデータベースを開発した。遺伝子機能情報データは、実験の未実施などによる情報の欠落や新たな知見によるデータ構造の変化等が起きる可能性があるため、このデータベースでは、柔軟なデータ構造を表現することができる XML (extensible Markup Language)を使用してデータを表現した[9]。また、公共配列データベースの更新に伴う遺伝子機能情報の見直しを組織間で効率的に行うため、配列相同性検索の結果を XML を用いて整形するシステムを開発した。さらに、共通ドメイン検索システムから得られた共通ドメイン情報を視覚的に閲覧するためのデータベースを開発した。

３．得られた新規ドメインの概要

本システムを使ってマウス完全長 cDNA 配列から得られた 7 種類の新規ドメイン[8]が、前述の共通ドメイン情報データベースを通じて公開されている（http://motif.ics.es.osaka-u.ac.jp/MDS/）。このうち 4 個については種々の予測や配列機能情報の解析により、ドメインの機能についてある程度予測できているが、残りの 3 個については機能がまだ不明である。ドメインの機能を予測した例を図１に示す。このドメインは配列中でロイシン（L で表記）が周期的に並んでおり、ロイシンジッパーと呼ばれる既知のドメインに良く似た配列パターンを持つ。しかし、ロイシンジッパードメインはコイルドコイル構造を持つ DNA 結合ドメインであるが、見つかったドメインは DNA 結合に必要な塩基性アミノ酸が周囲に少なく、またコイルドコイル領域とは予測されない。種々の解析結果から、このドメインについてはタンパク質結合ドメインと予測された[8]が、別の研究グループによって実際にこのドメインがタンパク質結合ドメインであることが報告された[10]。

４．ネットワークの活用について

本プロジェクトの実施過程では、特に大阪大学と理化学研究所の間でデータ交換のためにネットワークを活用した（IMnet と SINET により接続)。理化学研究所から大阪大学には、マウス完全長 cDNA 配列データおよびそれに付随した遺伝子機能情報、遺伝子発現情報、タンパク質間相互作用情報などが送られた。また、大阪大学から理化学研究所には、マウス完全長 cDNA 配列間で新規とみられる共通ドメインのデータ、そのドメインを持つタンパク質の 2 次構造、細胞内局在、膜貫通領域、コイルドコイル領域などのコンピュータ予測結果が送られた。遺伝子機能情報などは、SWISS-PROT などの配列データベースをもとに付けられているものが多く、データベースの更新

に伴い、機能情報にも更新が発生するため、それに応じた情報の転送やそれによる再計算結果の転送が生じた。以上のデータ通信では、配列データが未発表のものであるため、機密性には特に留意して、少なくとも SSL による認証と暗号化を行うようにした。

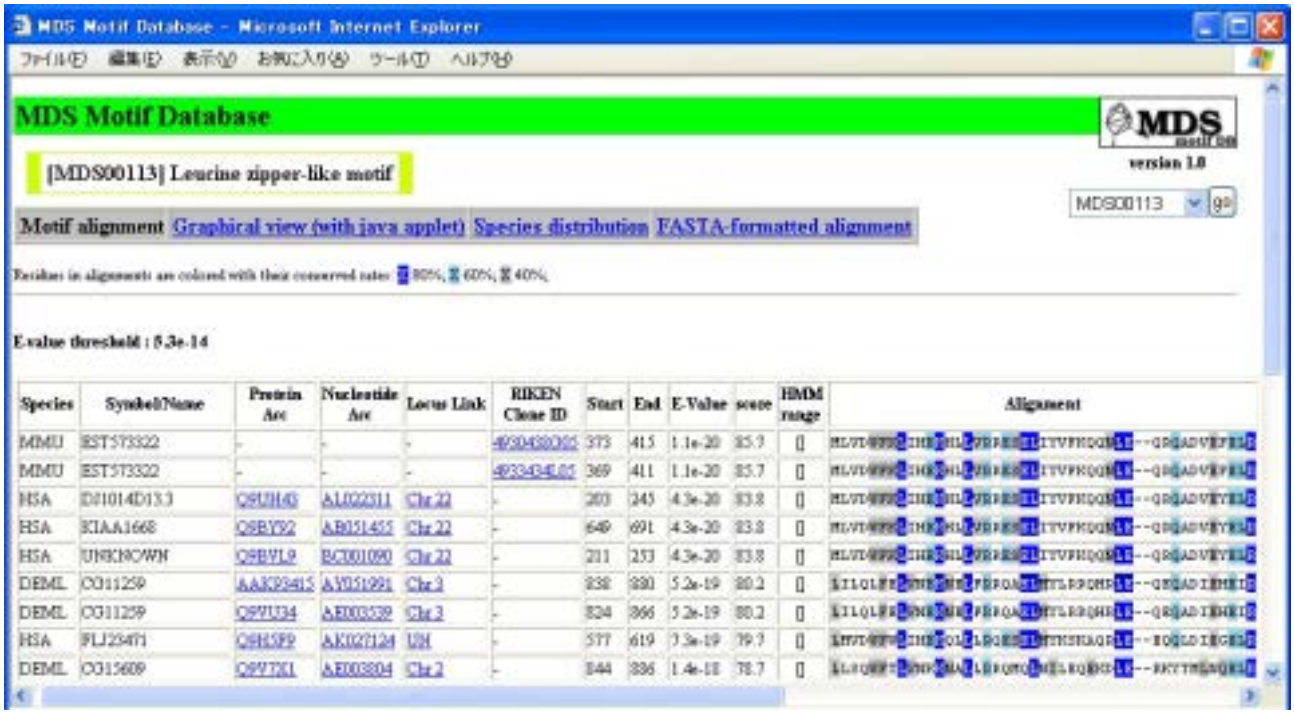

図１．新規ドメインの例(擬似ロイシンジッパードメイン）

５．まとめ

本プロジェクトでは、何よりも配列データを実際に生産しているグループから直接フィードバックを受けられたことが重要であった。例えば、同じ遺伝子座から取られた複数の cDNA 配列で、配列の領域ごとにシーケンシングの精度が異なり、たまたまある領域が精度が高くその周辺では精度が低いようなことが起こると、精度の高い領域が周囲と比べて配列の類似性が高くなり見かけ上はそこが共通ドメインとして検出されてしまう。これを防ぐためには、配列の精度のデータが必要であるが、通常はこれは入手困難であり、本プロジェクトのようにデータの生産者との直接の共同研究プロジェクトにより始めて可能となったと考えられる。さらに、本プロジェクトの開始から数ヵ月後から同じマウス完全長配列の機能アノテーション会議(FANTOM)が始まったが、そこで世界のトップクラスの研究者と共同で同じ配列データの解析に取り組むという機会に恵まれた。これらの中で特にドメイン解析では世界で最も優れた研究グループの一つである EBI の InterPro プロジェクトのグループとの共同の解析により、既知のドメインについてのデータと共に数多くの専門的知識の供与を受けることができたことは、本プロジェクトの発展に大きな助けとなった。

今後の課題としては、まず入力となる配列データ中にある冗長配列の除去があげられる。前述のように、cDNA 配列は一つの遺伝子座から複数個取られることがあり、もともと冗長性が高い。解決策は、同じ遺伝子座の配列をクラスタリングまたはゲノムとのマッピングによりまとめ、各クラスタからは１個の代表配列だけを取って、他は入力として使わないことが考えられる。しかし、配列をクラスタとしてまとめる基準が明確ではなく、さらなる改良が必要と考えられる。

次の課題としては、検索された新規ドメインと思われる領域の生物学的な機能予測法の開発がある。現在は、配列機能情報とともに、タンパク質２次構造予測、膜貫通領域・コイルドコイル領域・

細胞内局在シグナル配列領域などの予測結果をまとめて表示することによるドメイン機能予測の支援にとどまっている。しかし、配列の機能が予測できたとしても、それがそのドメインの機能とは限らない（同じ配列上の他のドメインがその機能を担っているかも知れない）ところに難しさがある。高等動物遺伝子のドメイン解析については世界中で多くの研究者が取り組んでおり、これらの成果報告の論文からテキストマイニングなどにより情報を得る仕組みも必要と考えられる。

６．研究開発実施体制

代表研究者　松田 秀雄（大阪大学大学院情報科学研究科）

研究開発題目

(1)マウス完全長 cDNA 配列間の共通ドメイン検索システムの設計・開発

松田 秀雄、竹中 要一、伊達 進（大阪大学大学院情報科学研究科）

研究協力者：　橋本 昭洋、川本 芳久、西田 知博（大阪学院大学情報学部情報学科）

(2) マウス完全長 cDNA 配列による共通ドメイン検索システムの評価

林崎 良英、河合 純、今野 英明、近藤 伸二、品川 朗、斎藤 輪太郎、足立 淳、福田 史郎（理化学研究所ゲノム科学総合研究センター遺伝子構造・機能研究グループ）

(3) 共通ドメイン検索システムの開発支援

村田 賢太郎、香月 祥太郎、中野 健司、粕川 雄也、川路 英哉、日比谷 尚武（エヌ・ティ・ティ・ソフトウェア株式会社技術開発部）

７．参考文献

[1] J. Kawai et al., “Functional Annotation of a Full-Length Mouse cDNA Collection”, Nature, 409(6821), 685-690 (2001).

[2] Y. Fukunishi and Y. Hayashizaki, “Amino-acid translation for cDNA with frame-shift error”, Physiological Genomics, 5(2), 81-87 (2001).

[3] H. Matsuda, “Detection of Conserved Domains in Protein Sequences using a Maximum-Density Subgraph Algorithm”, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, E83-A(4), 713-721 (2000).

[4] H. Konno, et al., “Computer-based methods for the mouse full-length cDNA encyclopedia: real-time sequence clustering for construction of a nonredundant cDNA library”, Genome Research, 11(2), 281-289 (2001).

[5] H. Kawaji et al., “A Graph-based Clustering Method for a Large Set of Sequences using a Graph Partitioning Algorithm”, Genome Informatics, 12, 93-102 (2001).

[6] S. Kondo et al., “Computational analysis of full-length mouse cDNA compared with human genome sequences”, Mammalian Genome, 12(9), 673-677 (2001).

[7] I. Yamanaka et al., “Mapping of 19032 mouse cDNAs on the mouse chromosomes”, Journal of Structural and Functional Genomics, 2(1), 23-28 (2002).

[8] H. Kawaji et al., “Exploration of Novel Motifs derived from Mouse cDNA sequences”, Genome Research, 12(3), 367-378 (2002).

[9] T. Kasukawa et al., “MaXML: Functional Annotation of Mouse cDNA Sequences in XML”, Proceedings of NETTAB Workshops, 52-56 (2001).

[10] J.R. Terman et al., “MICALs, a family of conserved flavoprotein oxidoreductases, function in plexin-mediated axonal repulsion”, Cell, 109(7), 887-900 (2002).