タンパク質の構造・機能予測法の開発とヒトゲノム配列への適用

名古屋大学大学院情報科学研究科
太田　元規

Research and development on prediction methods of protein structure and function, and their application to Human proteins

Motonori Ota
Graduate School of Information Science, Nagoya University

For the functional analyses of genomic sequences, it is necessary to develop methods to predict protein structures from their sequences, and to infer protein functions from their structures. We developed methodologies to address these issues on both fundamental and application bases, and integrated our results into the annotation pipeline of protein sequences. The data of proteins coded in the human genome was analysed by the pipeline and deposited in the SAHG database.

1. はじめに

DNA の担う遺伝情報はアミノ酸配列に翻訳され、それが生状態で固有の立体構造を形成した結果、分子機能という形で実体化する。配列から立体構造を知ること、つまり構造予測は、フォールド認識法という実用的な立体構造予測法が発展したことなどから既に確立されたように思われているが、ニーズが高いと思われるヒト・マウスといった高等生物由来のタンパク質はマルチドメイン構造や天然変性領域を保持するために立体構造予測が困難なものが多い。また、配列や立体構造から分子機能を知ること、つまり分子機能予測は分子機能記述の多様性を反映して、構造変化、活性部位など個別領域での解析に留まっている。その結果、高等生物由来のゲノム決定を受け、構造を通して機能についての知見を得る理論的枠組みは依然として未完のままである。

本研究開発ではタンパク質の「配列→立体構造→分子機能」というつながりを一層強固にするため基礎的な解析研究を推し進めるとともに、「高等生物由来のタンパク質にも適用可能で自動化されたフォールド認識法」および「多様な分子機能表現に対応した総合的な分子機能予測システム」からなるタンパク質の構造・機能アノテーションシステムを開発する。これらの研究開発過程では、構造予測、機能予測結果について検証実験を実施し予測法へのフィードバックを行う。開発した構造・機能予測アノテーションシステムをヒトゲノムへ適用し、結果を「ヒトゲノム構造・機能アノテーションデータベース」としてまとめ公開する。このデータベースでは立体構造を中心に、構造変化、相互作用などが一目で概観できるよう、様々な工夫を施す。

2. 研究開発の成果

2.1 構造予測

産総研生命情報工学研究センター（当時は生命情報科学研究センター）のグループは、2004 年に行われたタンパク質立体構造予測法のコンテスト（CASP6）で好成績を残した。その際に予測法として開発・

利用したフォールド認識法：FORTE を中心技術として、構造予測法の研究開発を行った。ドメイン単位でタンパク質の配列を処理する場合は、FORTE で鋳型構造を検索した後に、MODELLER によるホモロジーモデリングで立体構造を多数構築し、最適モデルを選択する。選択時に利用する関数としてタンパク質の構造安定性評価で実績のある STABILITY（LIBRA_Rotamer）を新たに採用した。この方法で 2006 年に行われた CASP7 に臨んだ結果、優秀チームとしては選定されなかったが、ある程度の実績（例えば、「ベストな鋳型より良いモデルの提出数」で、出場 189 チーム中 16 位）を残すことができた。

高等生物由来のタンパク質に多く含まれる天然変性領域についても独自の取り組みを行った。これまでに、機械学習データに含まれる天然変性領域の長さによって個別のパラメータを持つ予測法：POODLE シリーズ（POODLE-S、 POODLE-L）を開発してきたが、本研究課題として新たに POODLE-W を開発した[1]。3 つの POODLE を利用して臨んだ CASP7 では、天然変性領域予測のカテゴリで優秀チームとして認定された（総合評価 2 位）。これら POODLE について、予測サービスを公開した。

次世代のフォールド認識法として確率的アラインメントを用いた方法を開発した[2]。確率的アラインメントとは最適アラインメント以外の準最適解も評価するような方法で、統計力学の枠組みを利用して類似性スコアを確定する。この方法をプロフィール－プロフィール比較に適用したところ、従来法でアラインメントをしたプロフィール－プロフィール比較を上回る精度を発揮した（**図 1**）。

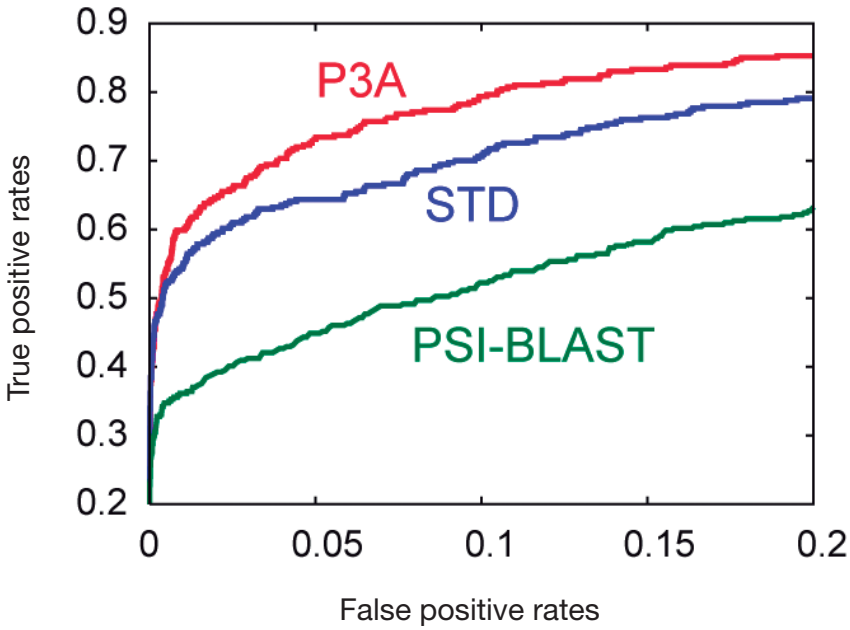

図 1 フォールド認識法の性能比較（ROC カーブ）
確率的アラインメント法（P3A）は従来法（STD）や PSI-BLAST より精度が高い

2.2 機能予測

分子機能は様々なイベントが集積した結果、実現していると考えられる。例えば酵素の場合、構造変化がおきて基質と結合し、活性部位が反応を実行した後にまた構造変化がおきて基質が開放される。活性に金属結合や補酵素が必須の場合もある。酵素複合体の場合は単体のタンパク質ではなくサブユニット構造が重要となる。つまり、構造変化、結合、反応、複合体、などについての知見が揃って初めて、そのタンパク質の担う分子機能が表現される。機能予測法の開発においては、分子機能予測がまだ未成熟な分野であることを鑑み、基礎的な知見を得る研究から予測法の適用研究まで、幅広く研究開発を実施した。

構造変化についてはまず、酵素について種別にその様子を調べた。アポ構造、ホロ構造ともに立体構造が決定されている酵素について調査したところ、転移酵素の立体構造変化は大きく、加水分解酵素の立体構造変化は小さいことを見出した[3]。転移酵素の反応は水をさけるため、構造が基質を覆うように構造変化のおきることが推察された。また、これまでに線形応答理論に基づく構造変化予測法を開発してきたが、この方法の適用範囲を知るために基質結合に伴う 240 の既知構造変化を調査した。ホロ構造を隠し

て、アポ構造からの構造変化を予測したところ、169例について線形応答理論は有効であった。よって、構造変化のほぼ7割が予測可能と判定された。

タンパク質の中には、生理的環境に応じて相互作用の相手を変えるハブタンパク質がある。このような社会的なハブタンパク質がどういう構造的な特徴によって実現しているかを調べた。タンパク質のハブ性は天然変性領域との関連性が議論されていたが、本研究では天然変性領域の含有量ではなく、立体構造の柔軟性が社会的なハブ性を特徴付けることを発見した[4]（**図2**）。

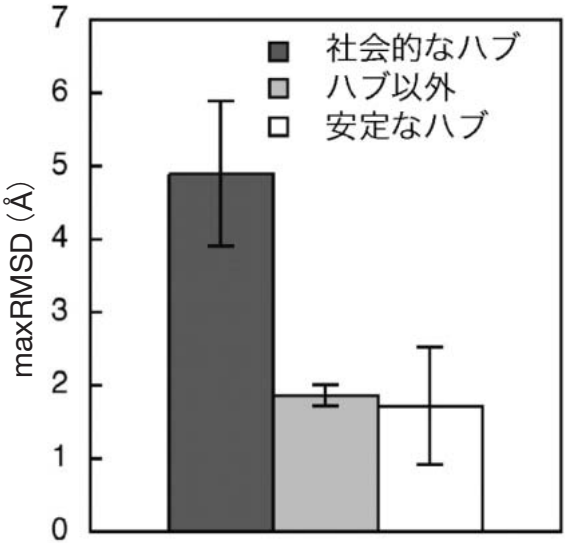

図2　タンパク質間相互作用によって生じる構造変化の大きさ
社会的ハブタンパク質は柔らかく、構造変化が大きい。安定なハブとは、相互作用の相手を変えない、複合体の要のようなハブをさす

タンパク質の会合状態を調べると、同一ファミリであっても会合状態が一致するとは限らない。同一ファミリに属するが会合状態の異なるタンパク質の立体構造を収集し、相互作用面でのどういうアミノ酸置換が会合状態変化をもたらすかを調べた。相互作用面が形成される時、露出していたアミノ酸残基はより疎水的になる、タンパク質内部に埋もれていた疎水性残基がより大きな残基に置換されて外に露出するようになる、という2つの一般側が、31例中24例について認められた。

タンパク質複合体予測においては、独自に開発した既存方法を利用してコンテスト（CAPRI）に取り組み、優秀チームとして認定された[5]。酵素反応データベース（EzCatDB）開発においては、170件のアノテーションを新規に追加した。タンパク質の立体構造データベース（PDB）を自動で解析し、構造変化、リガンド結合などについて調査を行うシステムを開発した。

2.3 実験検証

高速に発現系構築が可能な特殊なベクター：PRESAT-vectorを利用して、FORTE予測を検証するために古細菌由来タンパク質の立体構造をNMRで2件決定した。ヒト・マウス由来の17配列についてもNMRを測定し、2件について立体構造を決定した。49の配列についてはドメイン境界の決定を行い、49組のタンパク質間相互作用について実験検証を行った。また、天然変性領域予測の結果を検証するために改良型PRESAT-vectorとNMRを利用した実験系を開発した。この実験系ではNMRシグナルが既知の膜貫通領域と構造ドメインの間に、測定対象のタンパク質配列を挿入して利用する。測定対象のタンパク質が天然変性状態である場合、挿入配列は自由度のあるリンカーとなるので、膜貫通領域と構造ドメインのNMRシグナルは変化しない。一方測定対象のタンパク質が構造をとる場合、NMRシグナルは消失する。この方法を利用して、POODLEが天然変性領域と予測した40配列について発現を確認し、NMR測定を進行させている。

2.4 構造・機能アノテーションシステム

これまで本研究グループの内外を問わず開発されてきた方法および本研究開発で開発された方法などを総動員して、タンパク質の立体構造・分子機能アノテーションシステムを開発した（**図3**）。システム開発を集中的に行うため、ヒトの22番染色体を初期ターゲットとし、予測・解析法のパイプライン化に務めた。

構造予測パイプライン：天然変性領域を多く含むマルチドメインタンパク質に対応するには、構造ドメ

イン部分を同定して処理を行う必要がある。しかし構造ドメイン予測を先に実施してしまうと、それが間違っていた場合に正解は得られない。そこで、ドメイン同定は行うが、それをなるべく先送りにするようなプロセスを構築した。まずは問題配列についてBLASTサーチ、PSI-BLASTサーチを順次実施し、構造ドメインが該当した場合はそれを正解とする。残った部分は構造ドメイン部分が射貫かれているので、いくつかのセグメントに分割されている場合もある。それらのうち、POODLEで長大な天然変性領域と予測された箇所を除いて50残基以上のものをFORTEによるフォールド認識の対象とする。FORTEの結果に基づきMODELLERでモデルを多数発生し、STABILITY関数で良い値をとるモデルを選定する。

基質結合、構造変化モデル作成パイプライン：2.2で述べたPDB自動解析システムを利用し、PDBに登録されているアポ体、ホロ体のペアリストを作成する。これを基に、配列検索やフォールド認識でペアのいずれかの鋳型が選定された場合に、もう片方の鋳型についてもMODELLERでモデリングを行う（モデリング時に基質も含む）。2つのモデルを利用することで、基質結合に伴う構造変化を容易に提示することが可能となる。

基質結合、構造変化予測：上記パイプラインでアポ体、ホロ体の同時モデル作成ができなかった対象に対して、結合基質予測システム：eF-Seekで基質複合体の予測を行う。高い信頼度で予測された基質複合体に対し、線形応答理論に基づく構造変化予測を実施し、ホロ体モデルを作成する。

構造・機能アノテーション：パイプラインを経由しないアノテーションとして、天然変性領域、タンパク質複合体、酵素反応を用意した。天然変性領域はPOODLEの予測結果およびPRESAT-vectorによる実験結果である。タンパク質複合体モデルは、PDBに登録された複合体を鋳型として、MODELLERで作成する。酵素反応はEzCatDBに登録されたデータに基づくアノテーションで、酵素反応の種類と活性部位の位置情報である。解析結果は全てXML形式で扱うこととし、XML管理ツールXindiceで受け付けた後にリレーショナルデータベース：MySQLで管理する（**図3**）。

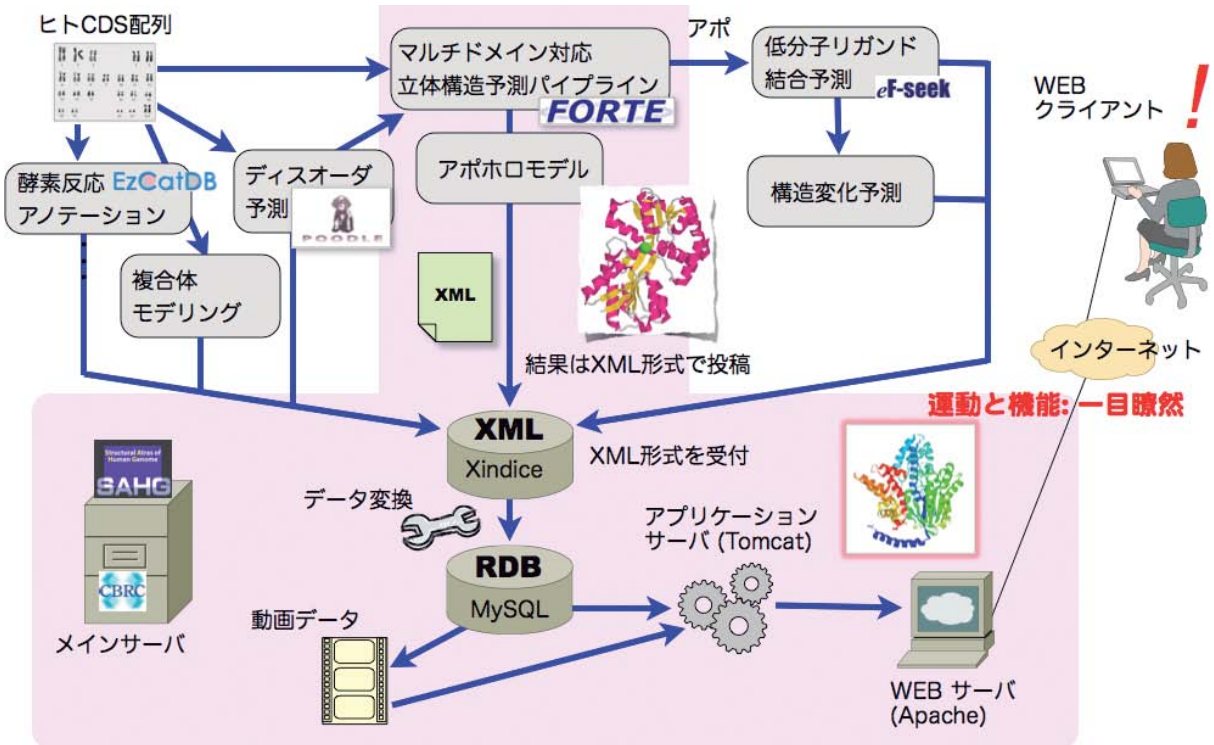

図3　構造・配列アノテーションシステムの概略図

2.5 ヒト由来タンパク質のアノテーションデータベース：SAHG

開発したタンパク質の構造・機能アノテーションシステムをヒトゲノム由来の配列に適用し、アノテーションデータベース：SAHG（Structural Atlas of Human Genome）を作成した。パイプラインとデータベースを評価するために、現在は22番染色体にコードされている600本ほどのタンパク質配列に対し、アノテーションを実施している。SAHGの開発に当たって、タンパク質の立体構造を中心に据え、その動きと働きが視覚的に認識できるよう配慮した（**図4**）。これは読むよりも見ることで理解するデータベースを目指すこと、マウスクリックによるブラウジングを主なインターフェースにすること意味する。SAHGの初期画面では、24本のヒト染色体アイコンが表示され、それをクリックすることで染色体にコードされたタンパク質一覧の画面に移動する。タンパク質は構造予測に従ってタンパク質単位、もしくはドメイン単位で表示されるが、動きと基質結合についてアノテーションがある場合はモーフィングによってそれが表示される。個々のタンパク質はアイコンであり、それをクリックすると詳細なアノテーション画面に移動する。ここでは拡大表示されたタンパク質の予測構造をマウスで自由に回転させて眺めることが可能である（**図4**）。またタンパク質配列を帯で表した上にドメイン構造、天然変性領域、活性部位、基質結合部位などがマップされており、複合体モデルへのリンクもある。つまり、構造・機能アノテーションを一次元的、三次元的、四次元的に把握することができる。染色体をクリックする以外に、配列検索、GeneInfo、RefSeqの該当エントリ情報を取り込んだキーワード検索も可能なので、ユーザは目的とするタンパク質のアノテーションに容易に辿り着くことができる。方法はほぼ自動化されているので22番染色体で実現されたことは、全配列で実現可能である。

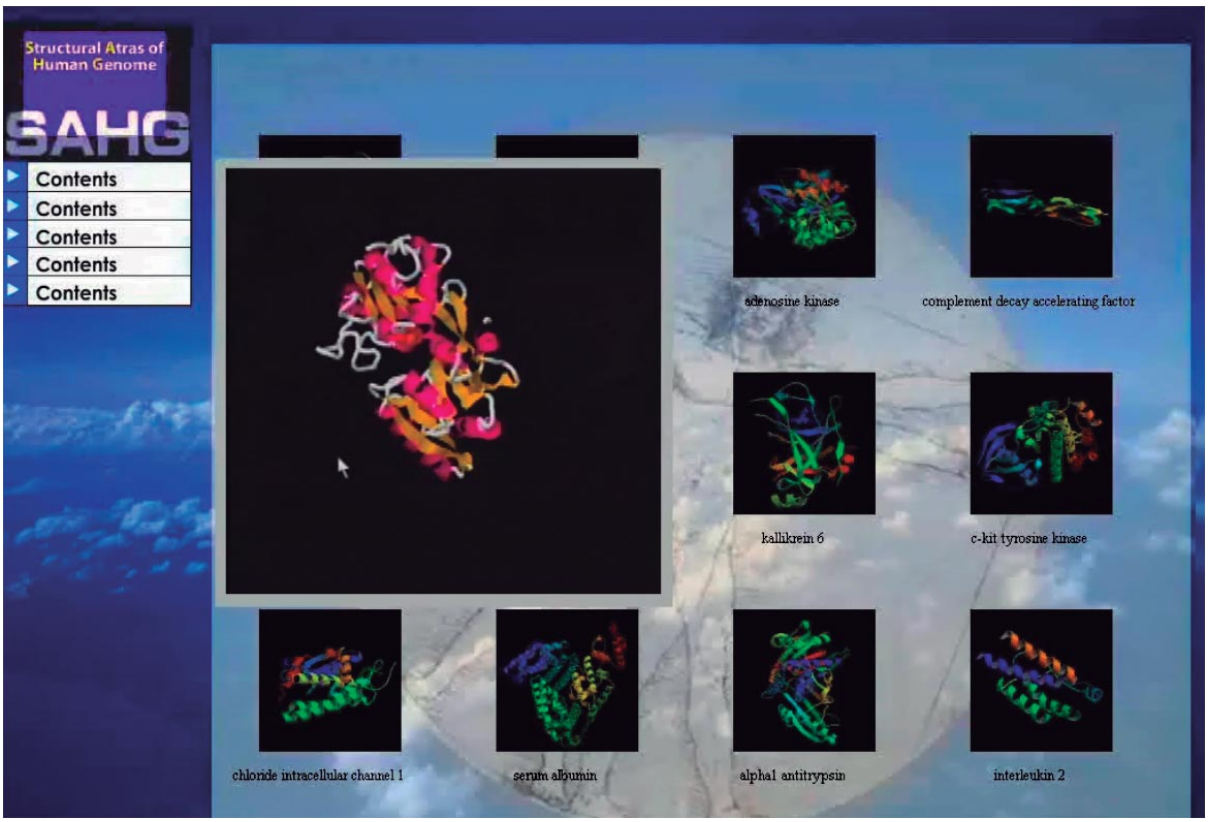

図4　SAHGによるタンパク質の構造表示のプロトタイプ画面
画面左の鉄結合タンパク質のオープン・クローズ運動はマウスで構造を回転させながら観察することができる

3. まとめ

タンパク質の構造予測法、分子機能予測法に関して最新のデータに基づく基礎的な解析研究から予測法開発、その応用研究を広範に実施した。それらの成果物とこれまでに開発された既存手法などを組合せ、タンパク質の構造・機能アノテーションシステムを開発し、ヒト22番染色体由来の配列に適用した。解析結果をまとめてアノテーションデータベース：SAHGとした。SAHGはタンパク質の構造、動き、分子機能を視覚情報として提供するので、ヒトの機能ゲノミクスを実施する際に有用な初期情報を与える。また現SAHGを基に、相互作用、複合体情報を豊富に含む、タンパク質ネットワークや細胞機能レベルのアノテーションに拡張することも可能である。構造から分子機能、さらに細胞機能の記述と理解に向け、開発したアノテーションパイプラインとデータベースは1つの端緒を提供している。

4. 研究開発実施体制

代表研究者　太田　元規（名古屋大学大学院情報科学研究科）

研究開発題目

（1）タンパク質複合体の網羅的モデリング法の開発と適用
　　グループリーダー　太田　元規（名古屋大学大学院情報科学研究科）

（2）タンパク質の構造変化予測法の開発と適用
　　グループリーダー　木寺　詔紀（横浜市立大学大学院国際総合科学研究科）

（3）タンパク質複合体の立体構造予測法の開発と適用
　　グループリーダー　木下　賢吾（東京大学医科学研究所）

（4）タンパク質立体構造予測法の開発、適用と酵素反応データベースの作成
　　グループリーダー　野口　保（産業技術総合研究所生命情報工学研究センター）

（5）タンパク質の構造・機能予測結果の実験検証法の確立と実施
　　グループリーダー　廣明　秀一（神戸大学大学院医学研究科）

5. 参考文献（成果発表の一部）

[1] K. Shimizu et al. Predicting mostly disordered proteins by using structure-unknown protein data. *BMC Bioinformatics*, **8**, 78, 2007

[2] R. Koike et al. Probabilistic alignment detects remote homology in a pair of protein sequences without homologous sequence information. *Proteins*, **66**, 655-663, 2007

[3] R. Koike et al. Protein structural change upon ligand binding correlates with enzymatic reaction mechanism. *J. Mol. Biol*. **379**, 397-401, 2008

[4] M. Higurashi et al. Identification of transient hub proteins and the possible structural basis for their multiple interactions, *Protein Sci*., **17**, 72-78, 2008

[5] E. Kanamori et al. Docking of protein molecular surfaces with evolutionary trace analysis, *Proteins*, **69**, 832-838, 2007