計画研究：2000～2004年度

ホモロジーモデリングによるゲノム規模の蛋白質構造データベースの構築と機能推定

●梅山 秀明[1)]　◆岩舘 満雄[1)]　◆西川 建[2)]　◆深海 薫[3)]

1)　北里大学　薬学部　2) 国立遺伝学研究所　生命情報・DDBJ研究センター　3) 理化学研究所　バイオリソースセンター

〈研究の目的と進め方〉

「ゲノム情報学」（２０００～２００２年の３年間）
２０００年（西川が研究代表者、分担者は西川研の深海、梅山、早稲田大学の輪湖、）
目的は、ゲノム上のタンパク質全ての立体構造を構築することである。進め方は、ホモロジーモデリングの手法を用いる。具体的には、タンパク質の立体構造完全自動モデリングソフトFAMSの開発・改良を行う。

ゲノムにコードされた全タンパク質を対象に、既存の配列データ解析ツールを動員した自動解析を行い、個別タンパク質の立体構造予測とファミリー分類を行う。その解析結果をデータベースとして構築したGTOPを更新・維持する。GTOPの解析結果を用いた比較ゲノム研究を展開する。

さらにGTOPのRPS-BLASTのアライメントに基づいてFAMS計算を行いモデル立体構造データベースであるFAMSBASEを構築する。

２００１年（梅山が研究代表者、分担者は梅山研の岩舘）ゲノムから得られるタンパク質のアミノ酸配列情報からその機能と関連し非常に重要である。
任意のアミノ酸配列情報が与えられた場合、立体構造データベースＰＤＢから類似性の高いタンパク質を選び出しアライメントを得て、さらにホモロジーモデリングの手法を開発してゆくことが目的である。

モデリングソフトとしてはＦＡＭＳがある。これをより改良してゆく。ソフトウェアとしての優秀さは国際コンテスト等に出場することによって証明してゆくのが基本方針である。

２００２年（梅山が研究代表者、分担者は梅山研の岩舘）目的は２００１年と略同じなので略。

「ゲノム生物学」（２００３～２００４年の２年間）
２００３年（梅山が研究代表者、分担者は梅山研の岩舘、遺伝研の西川、理研の深海）

ゲノムにコードされた全タンパク質を対象として、ホモロジーモデリングの手法により立体構造を作成し、データベースとして構築することを目的とする。まず、ゲノム中のそれぞれのタンパク質について、立体構造データベース（PDB）に対するホモロジー検索を行い、その結果をGTOPデータベースに格納する。既知構造とのホモロジーが検出されたタンパク質については、配列同士のアラインメントを入力情報として、立体構造自動モデリングソフト（FAMS）に掛け、モデリング計算を行う。その結果はデータベース FAMSBASE として構築する。

本研究課題において、モデリング手法としてのFAMSの完全自動化を目指して、手法の開発を完成させる。

GTOPでは、PDBの他にSCOP、Pfam、Swiss-Prot などの公共データベースに対するホモロジー検索も行いすべての解析結果を公開しているが、新規にゲノムの解明された生物種にも拡張していきたい。また、GTOPの解析結果を用いたゲノム情報解析研究を展開する。

２００４年（梅山が研究代表者、分担者は梅山研の岩舘、遺伝研の西川、理研の深海）
目的は２００３年と略同じなので略。

〈研究開始時の研究計画〉

「ゲノム情報学」（２０００～２００２年の３年間）
２０００年（西川が研究代表者、分担者は西川研の深海さん、梅山、早稲田大学の輪湖、）

ゲノム上の全ＯＲＦを対象に、コンピューター解析による個別タンパク質の分類と予測を行う。立体構造が予測される場合は全原子モデルを構築し、モデル構造に基づいて機能推定を行う。また、機能予測への試みとして、モチーフと呼べるような、頻繁に出現する共通の局所構造を同定・解析する。あるいは分子進化的な観点から構造の変化と機能の多様化を整理し、一般的な傾向を明らかにする。

２００１年（梅山が研究代表者、分担者は梅山研の岩舘）

二次構造の情報、疎水性残基が球状タンパク質形成のために疎水コアを形成しているか否か、プロファイルアライメントにおけるアミノ酸の類似性が高い領域か否か等を考慮して、アライメントを修正する。

ゲノムに対してこの操作を網羅的に行うために、あらゆる状況に対応できるよう力を注ぐ。

また２０００年に出場した国際コンテストにおける反省から、立体構造に基づきアライメントを変化させる方法を考えた。二次構造の情報、疎水性残基が球状タンパク質形成のために疎水性コアを形成しているか否か、プロファイルアライメントにおけるアミノ酸の類似性が高い領域か否かを考慮してアライメントを修正する。ゲノムに対しこの操作を網羅的に行うために、この手法を自動化するソフトウェアの開発を行う。

２００２年（梅山が研究代表者、分担者は梅山研の岩舘）

国際コンテストCASP5、CAFASP3へ参加しホモロジーモデリングの実力を示しさらにソフトとしての価値を証明する。
1) CAFASPへの参加を通じたFAMSの改良。
2) CASPの参加を通じて得た経験則の完全自動化への試み。
3) 1000台のPCクラスターを用いて1,2等で確立した方法論の適用したモデル構築。
3)で作成のモデルのデータベース化を中心に行う。

「ゲノム生物学」（２００３～２００４年の２年間）
２００３年（梅山が研究代表者、分担者は梅山研の岩舘、遺伝研の西川、理研の深海）
１）　CASP5（あらゆる手法がゆるされる総合部門）のホ

モロジーモデリング部門には、CHIMERAというグループ名で参加した（ソフトCHIMERAおよびFAMSを利用）が、ホモロジーモデリング対象51ドメインのGDT_TSの合計点（本研究室試算）で２位の成績だった。

2） CAFASP3（完全自動部門）のホモロジーモデリング部門には、FAMSDおよびFAMSサーバーが参加した（FAMSを利用）が、メタサーバーを除くと側鎖の確度において1位と2位だった。

3） ドッキングコンテストでもあるCAPRIにも出場した(1)(9)。

4） GTOPに収録された生物種は昨年の７０種から１１８種へと拡大した。なかでも真核生物では新たに分裂酵母、フグ、マウス（cDNA）を追加した。これら以外に４３種のファージも加えた。

5） GTOPの構造予測結果をもとに、原核生物が持つSCOPドメインの組合せパターンの種間比較を網羅的に行った結果、それぞれの種が持つ組合せパターンの違いを利用して系統関係を推定できることが分かった。

6） GTOP更新に伴い、FAMSBASEを更新した。結果総遺伝子数669980個に対して、その50.5%に相当する個の338020遺伝子についてモデリングを行い、総モデル数は126万個となった。

２００４年（梅山が研究代表者、分担者は梅山研の岩舘、遺伝研の西川、理研の深海）

梅山・岩舘

１．FAMSのモデリングサービスを再開し以前とアドレスが変わりhttp://www.pharm.kitasato-u.ac.jp/fams/となった。学術界に広く利用いただきたいと考えている。

２．国際コンテストを通じて新たなアライメントソフトの開発を行った。このソフトはホモロジー検索プログラムでありながらCASPの新規フォールド部門でもトップ30の成績を収めたことから遠縁の構造の検出能力があると考えている（論文準備中）。

３．GTOPの結果を受けてFAMSBASEの更新を行った。GTOP277生物種の736230遺伝子に対して2003年11月14日のPDBを参照し、RPS-BLASTのe値が10-3以下の50%に相当する369964遺伝子に対して最大5個の計140万個をモデリングした。

西川・深海

タンパク質の構造ドメインは基本的な構造単位であり、進化的にも安定に保存される。また構造ドメインは構造形成（フォールディング）の単位でもあり、フォールディングにおける協同性効果のため「半分のドメイン」などは許されない。言いかえると、ドメインはデジタル的であり、ドメインの全体があるか、または無いか、のどちらかであり、部分的に存在するというケースは（経験的にも）ない。このような「ドメインの一体性」という特性に注目しながら、GTOPに収録されたゲノム情報の解析を行う。具体的な課題として、１つは、選択的スプライシングの産物（ASバリアント）に焦点を当てる。ASバリアントは配列の一部がごっそり入れ替わるような変化を含んでいるはずであるが、このとき立体構造はどうなっているか、という問題である。もう１つは、上記の「ドメイン一体性」という特性を活かして、タンパク質のドメイン構成に基づいたゲノム系統樹の作成を試みる。

〈研究期間の成果〉

「ゲノム情報学」（２０００～２００２年の３年間）

２０００年（西川が研究代表者、分担者は西川研究室の深海さん、北里大学の梅山、早稲田大学の輪湖、）

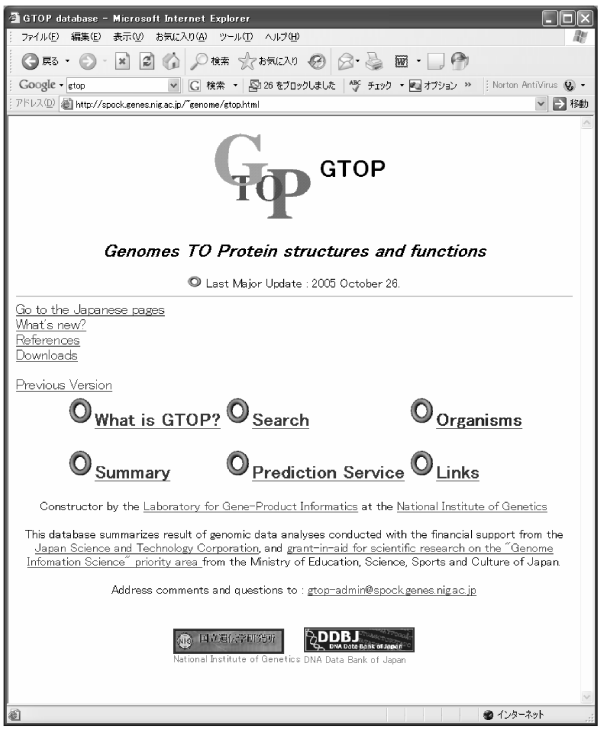


図1 GTOPトップページ

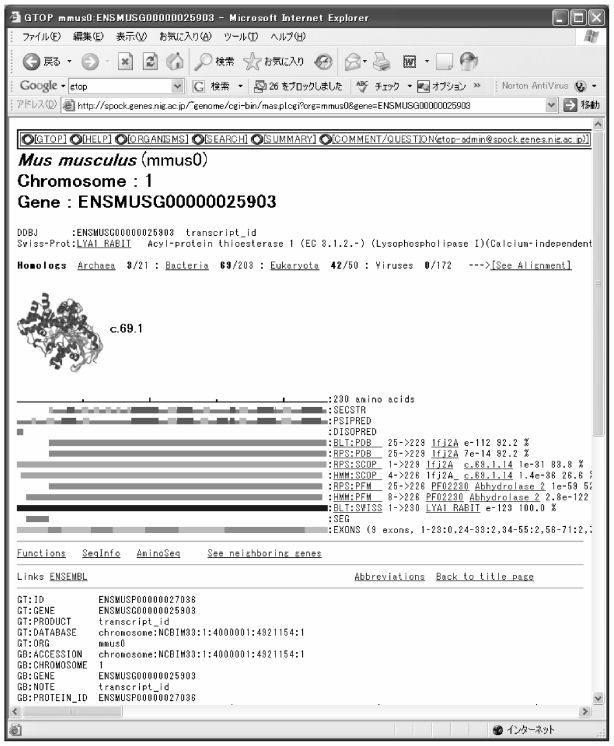


図2 GTOPの個別ORFに関する解析結果を示したトップページ

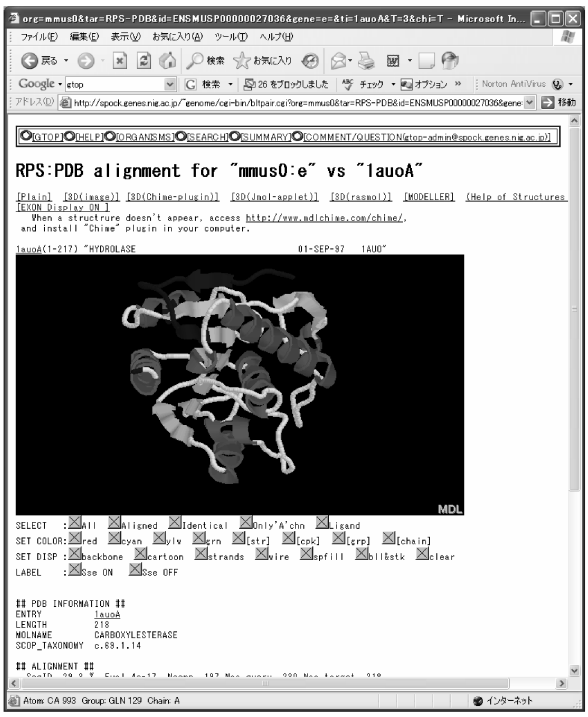


図3　GTOP の予測された立体構造に関する詳細情報

ゲノム上の全タンパク質に対し自動解析が行えるよう、既存の配列ホモロジー検索ツール（ＦＡＳＴＡ、ＢＬＡＳＴ、ＰＳＩ－ＢＬＡＳＴ）を動員して体制作りを行い、これまでに２０種以上の生物種について自動解析を終了した。得られた全ての解析データを格納し、検索・表示するＷＷＷサーバー（ＧＴＯＰデータベース）を立ち上げ、公開している。図１にＧＴＯＰトップページ、図２に個別ＯＲＦに関する解析結果を示したトップページ、図３に予測された立体構造に関する詳細情報の表示画面の例を示す。

自動解析の結果、立体構造／機能が新規に予測されるＯＲＦの割合は、予想以上に多い（バクテリアでは全ＯＲＦの４～５割にのぼる）ことがわかった。それら予測可能なＯＲＦについては実験データと照合して確認したり、生物学的意味、重要性について専門家の判断を仰ぐため、協力体制を作る必要がある。現在も何人かの専門家と協力関係にあるが、今後はさらに広い範囲で築く努力をしたい。また立体構造が予測される場合について、リガンドとの結合部位を調べることで機能予測を試みたい。

全原子モデルの構築については、タンパク質立体構造モデリングソフトであるＦＡＭＳを開発・改良した。図４にアルゴリズムを示す。さらに、同ソフトの優秀さを証明するために、国際コンテストであるＣＡＳＰ４、ＣＡＦＡＳＰ２に参加した(11)。ＣＡＦＡＳＰ２の参加にはインターネットのサイトを公開することが必須であり、北里大学内にウェブページ（http://physchem.pharm.kitasato-u.ac.jp/　このページは現在はhttp://www.pharm.kitasato-u.ac.jp/fams/ に変更）を開設して２０００年５月からのＣＡＦＡＳＰ２で行ったことと同じモデリングサービスを開始している。

局所的な構造の同定については、これまでのタンパク質の立体構造をDelaunay四面体で分割し分類する方法を発展させ、四面体の集合体を考えた。すなわち、対応する四面体それぞれが同じコードを持つより大きな局所構造モチーフを見出すためのアルゴリズムを新たに考え、解析プログラムの開発を行った。

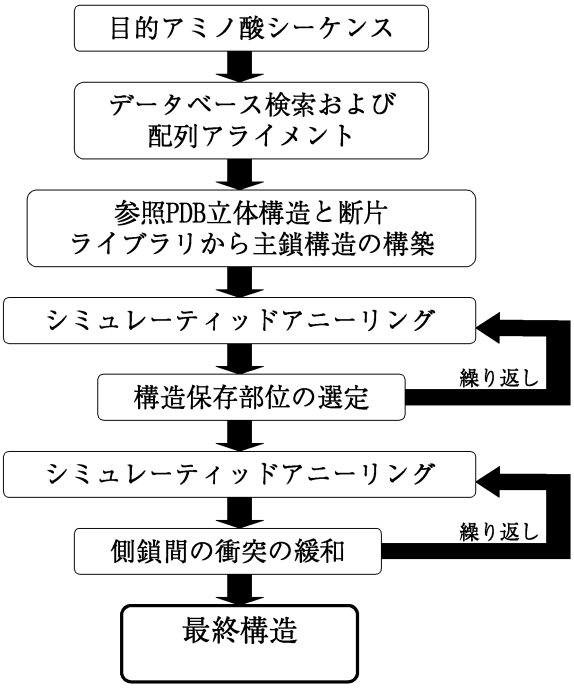


図4　モデリングソフトFAMSのアルゴリズム
モデリングは Ca 原子から始まり、主鎖原子、側鎖原子と順に構築される。特に最後の側鎖構築段階を繰り返しやることによって側鎖の精度を上げることが出来る。

４面体を用いた局所構造の同定法は、相同タンパク質間で立体構造の相違が比較的小さい、いわばＲｉｇｉｄな局所構造の同定や共通局所構造領域の構造アライメントに利用することも可能である。特に後者については、ペプチド鎖に沿った通常のアライメントとは異なり、四面体の対応関係から直接構造アライメントが出来るところに特徴がある。これらの点について、機知のモチーフに適用し、検証を行ったところ、その有用性が確かめられた。

分子進化的な観点からの解析は、ペリプラズム結合タンパク質（ＰＬＢＰ）を対象に行った。ＰＬＢＰの系統樹と立体構造比較の結果をもとに、二次構造のトポロジーが異なる三つのタイプがどのように出現したか、また、ＬｅｃＩ、ＰｕｒＲといったＰＬＢＰと相同な構造をもつリプレッサーがＰＬＢＰからどのように進化したかを解析した。

ＰＬＢＰの分子進化学的解析では、全ゲノム配列が決定されている生物種を解析対象にすることにより、リプレッサー・ＰＬＢＰの遺伝子が頻繁に重複／消失を起こしていたことも明らかになった。

２００１年（梅山が研究代表者、分担者は梅山研の岩舘）

計画に述べた研究よりはむしろ下記のデータベースの作成に力を注いだ。国際コンテスト等に出場し、自動ホモロジーモデリングを行うソフトとして十分な能力を証明できたと考え、これを利用したモデリングデータベースを作成した。具体的には遺伝学研究所のＧＴＯＰのア

図5　1000台のPCクラスター構築（味の素ライフライエンスセンター内北里大学分室）

ライメントに従ってＦＡＭＳモデリングを行った。ＧＴＯＰでは41生物種のＰＳＩ－ＢＬＡＳＴと61生物種のＲＰＳ－ＢＬＡＳＴのアライメントがあり、２１２，７６９種の遺伝子の約４５％に相当する９４．９７３種の遺伝子について立体構造を構築した。

モデルは遺伝子あたり上位５位までモデリングを行い、ＰＳＩおよびＲＰＳ－ＢＬＡＳＴの両方のアライメントを合わせて、５２２，６４４個のモデリングを行った。また、株式会社味の素ライフサイエンス研究所内に北里大学生物分子設計学教室の分室を設け、この計画研究を遂行するために１０００台のＰＣクラスターを構築してＦＡＭＳが稼動し立体構造を計算するシステムを構築した。

２００２年（梅山が研究代表者、分担者は梅山研の岩舘）

1) モデリングの国際コンテストであるCASP5およびCAFASP3に参加した。
2) CASP5（あらゆる手法がゆるされる総合部門）のホモロジーモデリング部門には、CHIMERAというグループ名で参加した（ソフトCHIMERAおよびFAMSを利用）が、ホモロジーモデリング対象51ドメインのGDT_TSの合計点（本研究室試算）で２位の成績だった(2)。
3) CAFASP3（完全自動部門）のホモロジーモデリング部門には、FAMSDおよびFAMSサーバーが参加した（FAMSを利用）が、メタサーバーを除くと側鎖の確度において1位と2位だった（下写真参照）。

4) 国立遺伝学研究所のGTOPで公開している99生物種の遺伝子アミノ酸配列合計322432のRPS-BLAST出力を用いてe値が1e-3以下の163382配列について上位5位のモデリングを行った。これは、遺伝子全体の50.7%に相当しており、先に行った2001年3月の41生物種、や2001年9月の時点の61生物種の時と比べて立体構造を構築できる割合が増えてきていることが見て取れ（図１３参照）、今後ますます重要になって行くことは想像に難くない。立体構造モデル数は約54万である。

5) またドッキングコンテストでもあるCapriにも出場している。

２００３年（梅山が研究代表者、分担者は梅山研の岩舘、遺伝研の西川、理研の深海）

梅山・岩舘

上記のCASP、CAFASP、CAPRI等の国際コンテストに参加することにより国内外での位置づけは明確だと考えており、本年度も参加の予定である。

特にCASP、CAFASPにおいてはホモロジーモデリングの手法において、手動、自動の両方において国際的に競争力があることを示す良い機会であると考えている。

タンパク質モデルデータベースFAMSBASEについては、アメリカはロックフェラー大学のSaliのグループが発表しているMODBASEが主なる競争相手と言える。彼らの発表によると126万の座標データを収録しているとのことで、本データベースとしてはその数字的同等と言える。

西川・深海

１） 選択的スプライシング（AS）によって生じるバリアントの構造安定性

エキソンの組み合せの異なるASバリアントについて、配列と構造の関係を次にように解析した。ヒト脳由来の完全長cDNAデータからASバリアントの組を選び出し、それぞれの配列をGTOP解析にかけた。既知構造（PDB）とのホモロジーが検出される配列の領域で、かつスプライシングの変化が同時に起きている領域に注目すると、配列変化と立体構造の関係は次の３つの場合に分けることができた。第１は、配列の変化を起こすエキソン境界と構造ドメインの境界がほぼ一致する場合であり、このときは構造ドメインが丸ごと挿入されたり、欠失することになるので、構造的にみて問題は起きない（つまり、安定なバリアントを生じる)。第２は、構造ドメインの途中に（短い）エキソンが挿入される場合であり、構造との関係を調べてみると、挿入点は常に分子表面に相当しフォールディングを妨げないことがわかった。このような、分子表面に挿入ループを形成するバリアントは構造的に安定であると考えられる。第３は、エキソンの組み替えによって構造ドメインの大きな部分が欠落する場合であり、このようなときはフォールディングに障害を及ぼすため構造的に不安定になると考えられる。かずさDNA研究所との共同研究により、RT-PCRによるcDNAの相対存在量を調べたところ、第１、第２のケースでは確かにバリアントペアの量比はほとんど変わらなかったのに対し、第３のケースでは、不安定と推定されたバリアントは常に極端に存在量が低い（元々cDNAが採れているので存在することは間違いないが）ことが確かめられた。我々の見積りでは、このような不安定なASバリアントは全体の１割強を占めると推定された。また、ヒトとマウスの間の保存性をみると、安定なASバリアントはマウスにも同じバリアントが見出されるのに対し、不安定なASバリアントの対応物はマウスに見つからなかった。このことから、構造的に安定なASバリアントは進化的にも安定である（また、その逆も真）ことが示唆された。詳細は発表論文１５を参照。

２） タンパク質のドメイン構成に基づくゲノム系統樹の作成

一般にタンパク質はいくつかのドメインからなるが、個々のドメインは構造単位として安定であり、デジタルな単位（あるか、ないか）として取り扱うことができる。さらに、ドメインの組み合せ（ドメイン構成）をデジタ

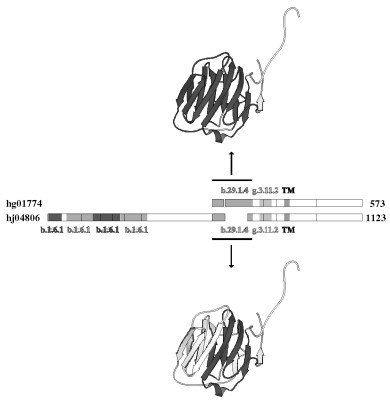


図６．２つのASバリアント（上と下）。上はドメインの全長配列をもつので安定。下は中央部の欠失のために不安定。

ルパターンと見なせば、通常の配列比較に比べて、より粗視化したレベルでの比較解析が可能になる。このようにタンパク質を粗視化した上で、ゲノム同士の比較を行い、ゲノム情報に基づく生物種の系統樹を作成する方法を開発した。

個々のタンパク質のドメイン（SCOPまたはPfam）の同定はすでにGTOPで行われているので、ゲノム中のすべてのタンパク質のドメイン構成をSCOP ID（または、Pfam ID）の組み合せとして表すことができる。さらに簡略にゲノムを特徴づけるために、あるドメイン構成がゲノム中に「あるか、ないか」だけを問題とし、タンパク数は問題にしないことにして、そのゲノムにおけるドメイン構成のレパートリと呼ぶ。２つのゲノム（A, B）を比較するには、レパートリ同士を比較することにより、A, Bに共通するドメイン構成の数、Aだけにある数、Bだけにある数、を用いて「距離」を定義することができる。任意のゲノム間の距離が定義できれば、それを用いて系統樹を描くことができる。我々はこの方法をGTOP中のすべての生物種（真正細菌136種、古細菌17種、真核生物14種）に適用し、全生物界の系統樹（Tree of Life）を得た（図７）。

その結果は、rRNA塩基配列などに基づく従来の系統樹と大まかな点では変りはないが、細かい点ではいくつかの違いが見られた。例えば、古細菌の２つの分類群（Crenarchaea, Euryarchaea）が完全に分離せず、前者が後者に含まれる関係になること。また、超高温好熱菌が系統樹のルート近くに集まることはなく、真正細菌のT. maritimaはむしろFirmicues類に近い位置にくる（高いbootstrap値）こと、などである（論文投稿中）。

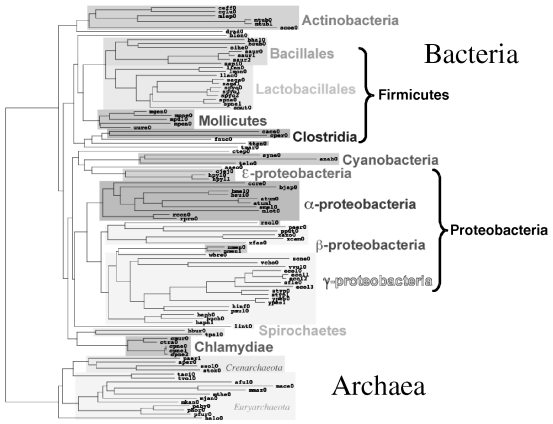


図７．真正細菌と古細菌ゲノムの系統樹

２００４年（梅山が研究代表者、分担者は梅山研の岩舘、遺伝研の西川、理研の深海）

梅山・岩舘

本研究室のモデリングソフトは国内外での位置づけのために国際コンテストCASP、CAFASP、CAPRI等に参加することによってた研究者と競っている(3)。FAMSはホモロジーモデリングのソフトとしては国際競争力を示してきている。

立体構造を構築するためのアライメントの部分では、新たに作ったソフトの優劣はやはり国際コンテストを通じて調べており十分に競争力のあるソフトウェアになりうると考えている。

FAMSBASE277生物種の計140万個のモデルはロックフェラー大学の126万個を超えて現在世界最大であると言える。

西川・深海

昨年報告したように、我々は構造ドメインの組み合せパターンの種間比較を行うことにより、ゲノム全体の情報を用いて系統関係を推定しうることを明らかにした。このことは比較ゲノムの方法として、タンパク質の構造ドメイン（の有る無し）を基本情報として用いることの有効性を意味している。特定のタンパク質群を対象とした 比較ゲノム解析の手始めとして、本年度は二成分制御系の解析を行った。二成分制御系の 比較ゲノム解析はすでにいくつかの報告があるが、遺伝子／タンパク質の全長配列の比較では部分的なホモロジーと全長でのホモロジーの区別が明確でないため、クリアカットな結果は報告されていない。真正細菌とアーケアとの間で明瞭なタイプの差があると いう発見は、我々の方法論の有効性を示しているといえる。

〈国内外での成果の位置づけ〉

「ゲノム情報学」（２０００～２００２年の３年間）

２０００年（西川が研究代表者、分担者は西川研究室の深海さん、北里大学の梅山、早稲田大学の輪湖、）

ＦＡＭＳの優秀さはＣＡＦＡＳＰのＣＭ部門で最優秀の結果であったことから明らかである。詳細な結果は、

ＣＡＦＡＳＰ２査定員Ｒｏｌａｎｄ Ｌ．Ｄｕｎｂｒａｃｋ（http://www.fccc.edu/research/labs/cafasp-results.html）からも報告されているが以下に詳細を示す。

CASP (Critical Assessment of Techniques for Protein Structure Prediction) がどの研究者が優秀かを競うものであるとすれば、CAFASP(Critical Assessment of Fully Automated Structure Prediction)はどのプログラムが優秀かを競う。

バイオインフォマティックス分野において、以下の理由により優秀なプログラムが必要なはずであると考えている。

１．構造生物学への踏み込みに関して、多くの生物学に携わる研究者に抵抗感を持たせない簡便さ。

２．全ゲノムが解明された生物種がいくつか存在する現状でタンパク質の立体構造をアミノ酸シーケンス情報のみから生み出す必要性。

現在ＦＡＭＳは、ＦＡＳＴＡ、ＰＳＩ－ＢＬＡＳＴ等

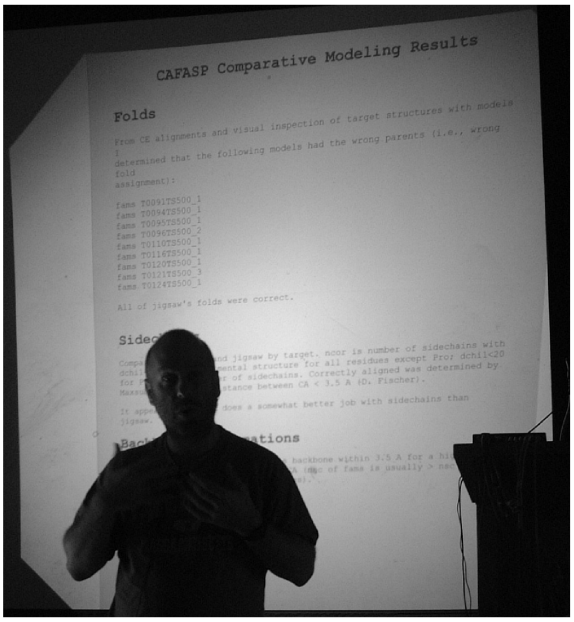


図8 CAFASP2での結果を述べているDunblack氏。

のアライメントソフトの出力結果をほぼ完全にタンパク質の立体構造へと変換することが可能な状態にあり、今後ゲノム情報からタンパク質立体構造をそのままモデリングすることも可能な状態にある。

CASPとCAFASPについて

CASPは1994年に始まったCASP1から2年ごとに開催され、1996年にCASP2、1998年にCASP3、そして2000年CASP4として開催された。また、1998年には「全てを人間の手を介さずにコンピューターのプログラムのみで予測を行う」という理念からCAFASP1が生まれ、2000年CAFASP2として開催される運びとなった。
CAFASP2は、CASP4の一部門として扱われており、CAFASP2への参加のためにはCASP4への参加が必須である。そのため、必然的にCAFASP2の参加チーム全てはCASP4にも参加している。

本研究室のホモロジーモデリングソフトFAMSも
http://physchem.pharm.kitasato-u.ac.jp/にwebページを開設して2000年5月からCAFASP2で行ったことと同じモデリングのサービスを開始した。

出場チーム

CMはホモロジーの高い参照タンパク質が存在する場合にいかに主鎖および側鎖において精度の高い座標を算出出来るかが争点となる。出場した4チームは以下のとおり
1. FAMS（北里大学生物分子設計学教室）
2. 3D-JIGSAW（イギリスのSternberg研究室）
3. SDSC1（米国サンディエゴのスーパーコンピュータセンター）
4. msi_GeneAtlas（米国、MSI社）

この内、msi_GeneAtlasはT0086－T0128のターゲットのうち、T0112、T0113、T0123の三つしか構造を提出しておらず、しかもいずれも提出期限である「出題から48時間以内」を守っておらず評価からはずされている。また、SDSC1は主鎖のみしか提出しておらず、側鎖のない座標であったためこれも評価対象外とされた。

CMの本文「ホモロジーの高い参照タンパク質が存在

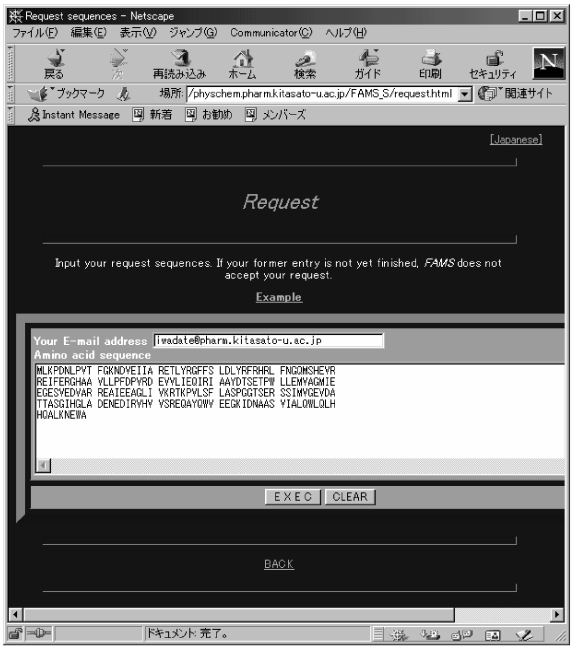


図9 国際コンテストに出場した当時のウェブページ

する場合に以下に主鎖および側鎖において精度の高い座標を算出する」に遵守したサーバーは
1. FAMS
2. 3D-JIGSAW
のみということである。

Fold認識について

まず、本研究室のサーバーであるFAMSもJIGSAWらのグループもアライメントソフトとしてPSI-BLASTを採用しており用いているアライメントが似ているせいか、サブミットの状況もSDSC1と比べるとよく似ていた（http://cafasp.bioinfo.pl/server/参照）。

しかしながら、FAMSはT0091、T0094、T0095、T0096_2、T0110、T0116、T0120、T0121、T0124
でフォールド認識を間違えていると指摘された。
それは、主に次のような規則によるものの為であった。

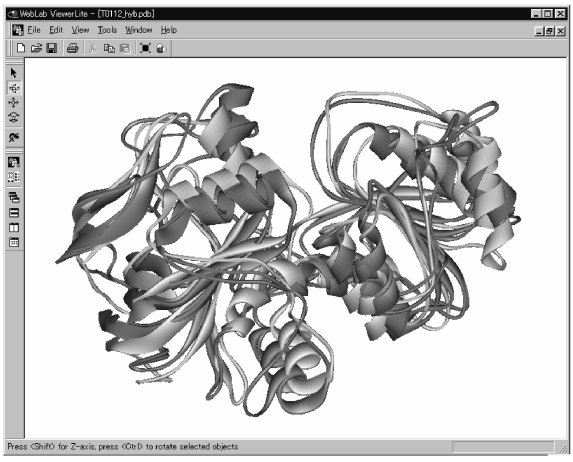


図10 立体構造の例
T0112（DHSO, Ketose Reductase / Sorbitol Dehydrogenase）アライメントに同じPSI-BLASTを用いているため、立体構造のFittingには、FAMSとJIGSAWには有意な差は無い。

1、CAFASP2で査定官の審査対象となるためには、CASP4とCAFASP2の両方への提出が必要となり、片方でも欠けるとそのターゲットについてはCAFASP2の審査対象外となる（もちろんCASP4のみに提出すればCASP4には参加したことになる）。
2、CAFASP2へは出題後４８時間以内且つ自動的な提出が求められるが、CASP4への提出は、手作業でCASP4のいう規則に則った書式の提出が求められる。

CAFASP2のみの提出はweb上に公開されておりhttp://cafasp.bioinfo.pl/target/を見れば各チームのサーバーが提出している内容が参照できる。

従ってCAFASP2のみの提出後、間違った構造であるとそのチームの人間が判断すれば、CASP4へ提出しないという選択を取り得る。従って、CAFASP2は構造計算については完全自動であるが、提出するか否かについては各チームの人間の判断に委ねられる。

チームFAMSはFold認識を全てPSI-BLASTの判断に則っており、e値が0.1以下であれば自動的にサブミットしていた。このことは完全自動であると言えるが、人為的な判断をせずにコンピュータプログラムFAMSが作成したタンパク質立体構造モデルを全てCAFASP2およびCASP4にサブミットしている。

表1　主鎖原子 Cαの RMSD

	fams	jigsaw	sdsc1
T0089	17.1773	21.8053	25.9982
T0092	4.22258		
T0094	14.2657		
T0099	4.90849	4.82491	4.76515
T0101			1.60513
T0103		20.3975	13.1962
T0110	8.70028		
T0111	1.73862	2.3315	2.73261
T0112	4.11556	4.11373	5.64019
T0114		13.3494	
T0115			34.9115
T0116	43.6884		27.372
T0120	23.6583	71.2137	37.306
T0121	3.30324	3.23358	3.45288
T0122	2.26866	2.52171	2.16785
T0123	3.59055	4.17979	3.59161
T0124	77.3679	104.299	
T0125		4.70701	
T0127	14.5661		30.0161
T0128	1.33571	1.20781	0.99166

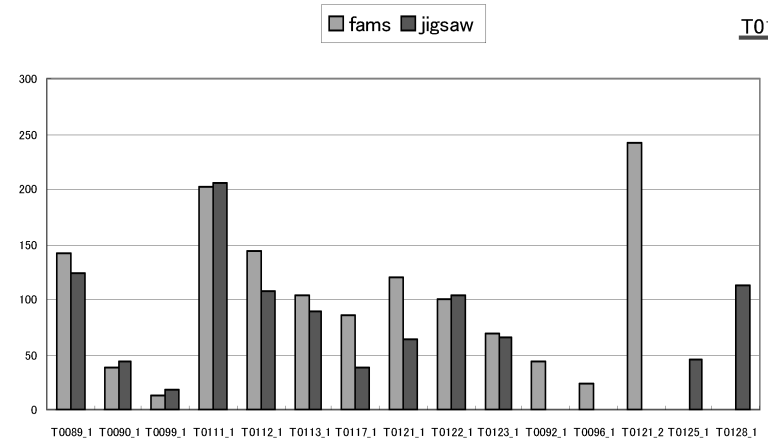
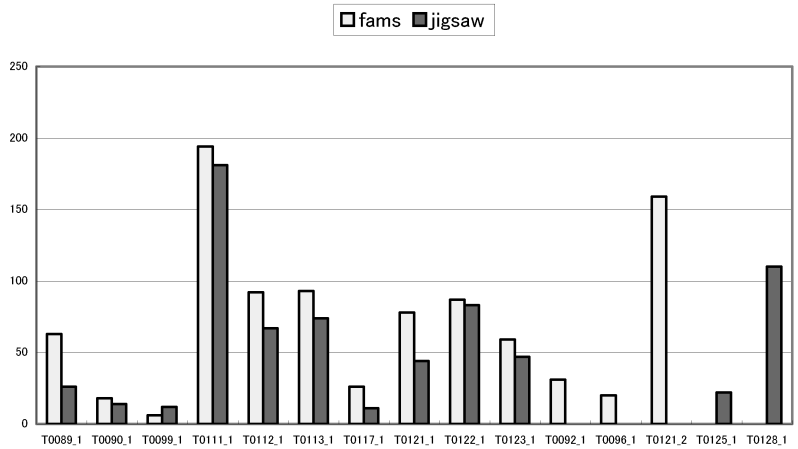


図11　正解構造の側鎖の内部回転角χ_1とモデルのχ_1が 40 度以内に収まっている数



競争の相手であるJIGSAWのグループでは、出題後48時間以内にCAFASP2へ提出後、T0096_2、T0120、T0121、T0124においては人為的にCASP4へ提出しておらず、その行為自体は完全自動と言えない。

また、FAMSがFold認識を間違えていると指摘されているT0120は、CASP4全体でFold認識部門も含めて評価しても、Fold認識を間違っているとは考えにくく、単にCMのターゲットとするには難易度が高いターゲットであったというだけの話でチームFAMSにとってなんらネガティブな結果はではない。ちなみにT0120における一致残基数の最高は同じく日本の松下技研チーム（CAFASP2ではなくCASP4のみにFR部門で参加）の９７残基、FAMSは２位の８９残基とCASP4の公式記録に掲示されており、参照タンパク質は松下技研チームと同じPDB ID: 1CIIを選んでいる。

CAFASP2に提出したターゲットについて、正解構造との最小自乗フィッティングを行い、RMSDを算出した(表1)。その際、JIGSAWがCAFASP2へ提出し、CASP4へ提出していないものもあわせて評価した。

これを見ると、確かにT0116、T0124などは、RMSDが大きく、正解構造から離れていることを意味しており、Fold認識を誤っていると、指摘を受けて然るべきである。が、その他については、Fold認識部門においても、なかなか正解の出辛いターゲットであり、ネガティブな結果とはなっていないような印象である。

側鎖の内部回転角の確度について

続いて、査定官のDunbrackは、側鎖の内部回転角の確度について両者を比較した結果を示している。

これは、具体的には正解構造の側鎖の内部回転角c1とモデルのc1とを比べて40度以内に収まっている数をカウントしている。

この結果、FAMSはJIGSAWよりもより幾分正確に側鎖の構築が出来（図11）、さらにその傾向はアライメントが正しいとされる領域においてより顕著であるといえる(図１２)。

アミノ酸ごとの側鎖内部回転角の確度について

さらに査定官は側鎖の評価をアミノ酸の種類ごとに行った。

そこでFAMSはJIGSAWよりも有意に正しい側鎖を予測できることを示している。ここでも、アライメンとが正しいとされる立体構造フィッティングで3.5Å以内にCa原子がある残基のみを評価した結果はFAMSがJIGSAWよりも正しく側鎖を予測できている傾向がより顕著である事を示していた。

主鎖の構築について

続いて査定官は主鎖の評価を行った。これは、正解構造とモデル構造の主鎖の内部回転角ϕ、ψを算出しそのRamachandran Plot上での直線距離が60度以内にある残基の数をカウントしたものである。

そこではFAMSはJIGSAWよりも有意に正しい主鎖の内部回転角角ϕ、ψを予測できることを示している。ここでも、アライメンとが正しいとされる立体構造フィッティングで3.5Å以内にCa原子がある残基のみを評価した結果はFAMSがJIGSAWよりも正しく主鎖の内部回転角角ϕ、ψを予測できている傾向がより顕著である事を示している。

Fold認識、側鎖の確度、主鎖の内部回転角の確度を比べた結果を示してきたが、今後モデル構造を、創薬等に利用して行く(4)際に特に必要とされるのは、「生理活性に重要な部位が側鎖を含めて正確に構築できているか？」であると考えられる。タンパク質の生理活性は、アミノ酸側鎖が担っており、その意味でいくら正確に主鎖を作れても生物学上は、FR部門と何ら変わりない情報しか与えていないと言える。

その意味において側鎖を確度高く構築する技術においてFAMSというソフトウェアの力が証明された結果となった。このことは、今後数年続くと思われるバイオインフォマティックスがサイエンス界をリードして行く時代において、同ソフトの担う役割が非常に大きいことを示している。

主鎖及び側鎖で高精度の構造が構築できるFAMSは、Fold認識を現在PSI-BLASTに頼っていて、20世紀が終ろうとする時点において世界トップレベルの争いをするCASP、CAFASPのような会においても比較的遜色ない結果を提供してくれている。しかしながら、FR部門のみに集中すればより優秀なチームがいくつも存在し、それらのアライメント情報をPSI-BLAST情報の変わりに用いれば、より精度の高い構造を算出できることになるだろう。

近い将来、タンパク質の高次構造形成問題を解くかぎは、間違いなくNew Fold部門もしくはその関連の研究から解決されるであろうと考えられる。しかしながら、ゲノムが解析を終えた生物種がいくつか存在するような現在、ゲノム情報が経済的価値を生むために、CM部門が非常に大切であると考えられる。

２００１年（梅山が研究代表者、分担者は梅山研の岩舘）

タンパク質モデルのデータベースとしては、収録した５２万は当時の世界最大である。しかしながら遺伝子数としては、当研究計画が９４９７３種であるのに対し、ロックフェラー大学のA. SaliがModellerを用いてTrEMBL６０万配列に対して３０万のセグメントの立体構造を得たと報告しており、遺伝子数において差がついていた。

２００２年（梅山が研究代表者、分担者は梅山研の岩舘）

国際コンテストにおいて上記の成績をあげたことから国内外での位置は明確であろうと考えられる。

作成した約54万の立体構造を持つデータベースは、世界最大級だと考えられる。

達成できなかったこと、予想外の困難、その理由

上記の通りCASPで優秀な成績をあげた経験を完全自動にする試みに対しては、今後頑張りたい。

２００３年（梅山が研究代表者、分担者は梅山研の岩舘、遺伝研の西川、理研の深海）

梅山・岩舘

上記のCASP、CAFASP、CAPRI等の国際コンテストに参加することにより国内外での位置づけは明確だと考えており、本年度も参加の予定である。

特にCASP、CAFASPにおいてはホモロジーモデリングの手法において、手動、自動の両方において国際的に競争力があることを示す良い機会であると考えている。

タンパク質モデルデータベースFAMSBASE(9)については、アメリカはロックフェラー大学のSaliのグループが発表しているMODBASEが主なる競争相手と言える。彼らの発表によると126万の座標データを収録しているとのことで、本データベースとしてはその数字的同等と言える。

西川・深海（含む2004年の内容）

高等生物における選択的スプライシングはホットな問題であり、国内外の研究は多いが、立体構造との関係に焦点を当てた論文は今のところ我々のもの以外にはない。

一方、ゲノム系統樹に関しては、我々が研究に着手した時点（約３年前）ではドメイン、またはドメイン構成を用いた方法は皆無であったが、その後、１年前にドメインを考慮した論文がYangらによって発表された。方法と結果は大筋において我々のものと変りないが、ばらばらのドメインを用いた彼らの方法よりもドメインの組み合せを考慮した我々の方法が系統樹としての精度（bootstrap検定による）が良いこと、またドメインレパートリの違いを「距離」に換算するときの定義が我々とは違っているが、質的にはこちらの方が優れていると考えている。

２００４年（梅山が研究代表者、分担者は梅山研の岩舘、遺伝研の西川、理研の深海）

梅山・岩舘

本研究室のモデリングソフトは国内外での位置づけのために国際コンテストCASP、CAFASP、CAPRI等に参加することによってた研究者と競っている。FAMSはホモロジーモデリングのソフトとしては国際競争力を示してきている。

立体構造を構築するためのアライメントの部分では、新たに作ったソフトの優劣はやはり国際コンテストを通じて調べており十分に競争力のあるソフトウェアになりうると考えている。

FAMSBASE277生物種の計140万個のモデルはロックフェラー大学の126万個を超えて現在世界最大であると言える。

西川・深海

2003年分の内容が2004年も含むので割愛

〈達成できなかったこと、予想外の困難、その理由〉

２０００～２００２年（梅山が研究代表者、分担者は梅山研の岩舘）

１０００台のパソコンを稼動したときの電力量は相当なものであり、パソコンの規格どおりの電源を用意してあってもブレーカが何度も落ちてしまい、その都度の復旧は大きく計画の進行を遅らせた。

２００３年～２００４年（梅山が研究代表者、分担者は梅山研の岩舘、遺伝研の西川、理研の深海）

梅山・岩舘

CAFASPでは通常サーバーの答えを見て解答することが許されるメタサーバーの台頭が著しく、通常サーバーであるFAMSDおよびFAMSサーバーはメタサーバーの代表格であるワシントン大Robettaサーバーに負けている。モデリングの前段階のアライメント段階に既に差がありそれが結果として現れたと考えられる。

西川・深海

ゲノム系統樹の研究は１年半以上前に完了していたに

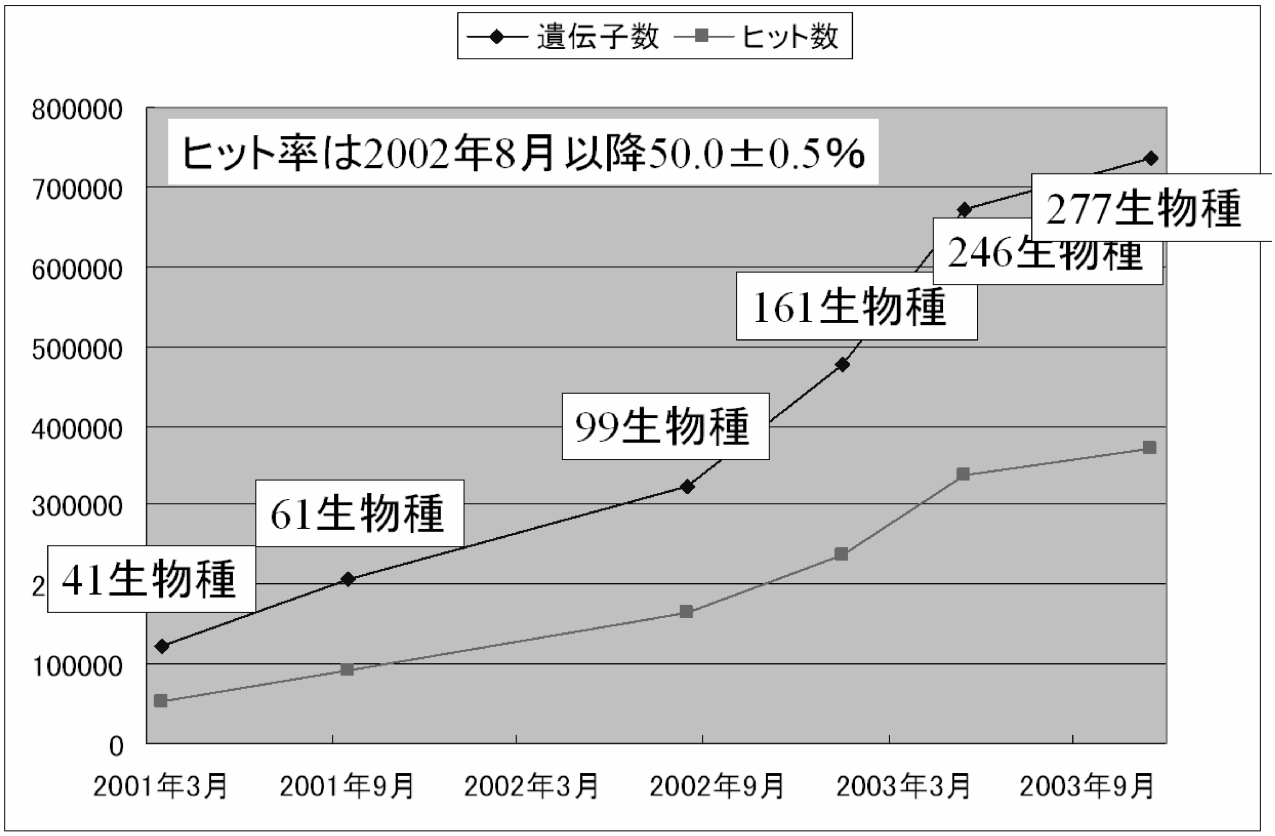


図13 FAMSBASE モデル構築数推移

もかかわらず、論文を書くのが遅れてしまい、その間に類似の論文（Yangら）が出てしまった。原因は、ちょうどその時期に筆頭著者（深海）の異動が急に決まり、論文をまとめるタイミングを逸したからである（現在、あらためて投稿中）。

〈今後の課題〉

梅山・岩舘

更に完全なホモロジーモデリングソフトとして簡便化かつ、高精度化を目指す。

西川・深海

当初設定した「ドメインの一体性」という観点は、タンパク質レベルでゲノム情報を解析するのに非常に有効であることが分かったので、今後とも同様の視点に立ってゲノム情報解析の具体的な課題に追求していきたい。

〈研究期間の全成果公表リスト〉

1) Mayuko Takeda-Shitaka, Genki Terashi, Chieko Chiba, Daisuke Takaya and Hideaki Umeyama. FAMS Complex: A fully automated homology modeling system for protein-protein complex structure. Medicinal Chemistry, in press.
2) Mayuko Takeda-Shitaka, Genki Terashi, Daisuke Takaya, Kazuhiko Kanou, Mitsuo Iwadate and Hideaki Umeyama, Protein structure prediction in CASP6 using CHIMERA and FAMS, Proteins, Published online: 26 Sep, 2005.
3) Terashi G, Takeda-Shitaka M, Takaya D, Umeyama H. "Searching for protein-protein interaction sites and protein-protein docking by the methods of molecular dynamics, grid scoring and the pair-wise interaction potential of amino acid residues." Proteins, 60(2): 289-95, (2005).
4) 著者18名, Matsumoto M(1), Takeda-Shitaka M(10), Iwadate M(11), Umeyama H(12), Miyata T(17), "Molecular characterization of ADAMTS13 gene mutations in Japanese patients with Upshaw-Schulman syndrome", Blood. 2004; 103(4):1305-10. (2003)
5) Adachi M, Kurihara Y, Nojima H, Takeda-Shitaka M, Kamiya K, Umeyama H., "Interaction between the antigen and antibody is controlled by the constant domains: normal mode dynamics of the HEL-HyHEL-10 complex", Protein Sci.; 12(10):2125-31, (2003)
6) Nojima H, Takeda-Shitaka M, Kurihara Y, Kamiya K, Umeyama H., "Dynamic flexibility of a peptide-binding groove of human HLA-DR1 class II MHC molecules: normal mode analysis of the antigen peptide-class II MHC complex", Chem Pharm Bull (Tokyo); 51(8):923-8. (2003)
7) Kurihara Y, Watanabe T, Nojima H, Takeda-Shitaka M, Sumikawa H, Kamiya K, Umeyama H., "Dynamic character of human growth hormone and its receptor: normal mode analysis", Chem Pharm Bull (Tokyo); 51(7):754-8. (2003)
8) Katsuichiro Komatsu, Youji Kurihara, Mitsuo Iwadate, Mayuko Takeda-Shitaka, and Hideaki Umeyama, "Third Solvent Clusters Fitting for Protein-Protein Interaction Prediction", PROTEINS, Structure, Function and Genetics,; 52(1):15-18. (2003)
9) Yamaguchi A, Iwadate M, Suzuki E, Yura K, Kawakita S, Umeyama H, Go M., "Enlarged FAMSBASE: protein 3D structure models of genome sequences for 41 species", Nucleic Acids Res,; 31, 1, 463-8 (2003)
10) Hiroyuki Nojima, Mayuko Takeda-Shitaka, Youji Kurihara, Masaaki Adachi, Shigetaka Yoneda, Kenshu Kamiya and Hideaki Umeyama, "Dynamic Characteristics of a Peptide-Binding Groove of Human HLA-A2 Class I MHC Molecules: Normal Mode Analysis of the Antigen Peptide-Class I MHC Complex", Chem. Pharm. Bull., 50, 1209-1214 (2002).
11) Iwadate M., K. Ebisawa, Umeyama H., Comparative Modeling of CAFASP2 Competition, Chem-Bio Informatics Journal, 1(4); 136-148. (2001)
12) 0305311726 Nakashima, H., Fukuchi, S. and Nishikawa, K.: J. Biochem. 133, 507-513 (2003). Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures.
13) 0404061035 Fukami-Kobayashi K., Tateno, Y., and Nishikawa, K.: Mol. Biol. Evol. 20, 267-277 (2003). Parallel Evolution of Ligand Specificity Between LacI/GalR Family Repressors and Periplasmic Sugar-Binding Proteins
14) 0411091200 Fukukuchi, S. and Nishikawa, K.: DNA Res. 11, 219-231 (2004). Estimation of the number of authentic orphan genes in bacterial genomes
15) 0411111135 Homma, K., Kikuno, R.F., Nagase, T., Ohara, O., and Nishikawa, K.: J. Mol. Biol. 343, 1207-1220 (2004). Alternative splice variants encoding unstable protein domains exist in the human brain
16) 0211071506 GTOPデータベース
http://spock.genes.nig.ac.jp/~genome/gtop.html