

膜貫通トポロジー情報を利用した膜貫通タンパク質機能の網羅的分類・予測

●清水 俊夫

弘前大学理工学部

＜研究の目的と進め方＞

膜貫通(TM)タンパク質はプロテオームの20～30%を占めているにもかかわらず、その機能分類は水溶性タンパク質のそれに比べて立ち遅れている。TMタンパク質の構造決定が進んでいないことが最大の原因であるが、TMタンパク質の配列比較が簡単ではないことも原因の一つとなっている。TMセグメントを繋ぐループ部分の長さが保存されていない場合が多くあって、そのことがアミノ酸配列の保存性を低くし、アライメントを難しくしているからである。この傾向は真核生物由来のTMタンパク質でより顕著である。一方、TMタンパク質の機能はTMトポロジーと密接に対応しており、正確なTMトポロジーデータが得られれば、機能予測が十分可能であることが分かってきている。本研究では、配列の類似性に加えてトポロジーの類似性に着目して、TMタンパク質の網羅的機能分類・予測をより正確に行う方法の開発を目指している。このためには正確なTMトポロジーデータを得ることが必要になるが、本研究では、いくつかのTMトポロジー予測法の組み合わせによる予測精度向上を利用した。最終的には、このようにして得られた機能分類結果とTMトポロジーとの関係を詳細に解析することによって、TMタンパク質のドメイン・アーキテクチャー、それと機能との関係、それらの進化的成り立ちなどを明らかにする。さらに、これらの結果に基づいた、TMタンパク質構造・機能の分類・予測システムの構築を目指すものである。

＜研究開始時の研究計画＞

- ①実験によって決定されたTMトポロジーデータ収集を行ない、それをデータベースとして構築する。それらのデータを用いてTMトポロジー予測法の予測性能の客観的評価を行う。収集するTMトポロジーモデル数400程度を目標に設定する。
- ②収集したTMトポロジーデータを用いて、多くのTMトポロジー予測法の性能評価を実施する。さらに、トポロジー予測精度を向上させるための当面の策として、既存の予測法の最適な組み合わせとしてのコンセンサスTMトポロジー予測法の開発を行う。既存の予測法の最良のものより少なくとも10%程度の精度向上を目指す。また、予測法によって予測されにくい埋没型、折れ曲がり型などの変則的TMヘリックス、TMセグメントとして予測される非TMセグメント領域などの詳細な分析を行い、その結果を反映させた、隠れマルコフモデルなどに基づく新たなTMトポロジー予測法の開発を目指す。
- ③シグナル・ペプチド (SP)領域の適切な処理を含めた、信頼度の高いTMトポロジーデータを獲得するための予測手順を確立する。
- ④全ての解読済みゲノムから、予測によるTMタンパク質の網羅的抽出を行ない、それらのプロテオーム解析、TMトポロジー解析を行う。
- ⑤原核87ゲノムのTMプロテオームの配列類似性比較に

よって可能な機能同定のレベルを調べる。さらに、それに加えてTMトポロジー類似性を利用した場合の機能分類・同定によってそのレベルをどこまで上げられるかについて調べる。

- ⑥TMタンパク質構造には、TMヘリックス領域を含む繰り返し構造が認められるものが多くある。それらは、遺伝子内重複によって生じたものと理解されている。そこで、配列内部分配列比較によって、87原核TMプロテオームの中からそのような配列を網羅的に探し出し、それらの詳細な解析から、TMトポロジーの進化的道筋の解明を目指す。その際、"positive-inside rule"に基づき、特に、2倍体型重複によって形成されたと考えられる8本型、10本型、12本型のTMタンパク質の構造・機能の進化的成り立ちの理解に焦点をあてる。
- ⑦原核87ゲノムからのTMタンパク質配列を対象とした、ドメイン・アーキテクチャーの網羅的解析を行う。まず、Pfamドメイン、SCOPドメインのアライメントを行ない、TMタンパク質に対する水溶性ドメイン定義の現状を押さえ、未定義の水溶性ドメインがどの程度存在するかの見積もりを行う。TMタンパク質のドメイン・アーキテクチャーは、長いループ上に存在する水溶性ドメインと、短いループによって繋がったTMヘリックスの束 (TMモジュール) (さらにそれらが集まってTMドメインを形成) の組み合わせとして捉えることが可能である。そのような描像が妥当であるかどうかの検証も同時に行う。

＜研究期間の成果＞

- ①実験によって決定されたTMトポロジーデータを多数の文献から収集し、TMトポロジーデータベース (TMPDB) として構築し、インターネット上に公開した。TMPDBには、真核131、原核217の348のTMトポロジーモデルが含まれている。(論文17、22)
- ②TMPDB中のTMトポロジーモデルをトレーニングデータとして決定される、既存のTMトポロジー予測法の最適な組み合わせとしてのコンセンサスTMトポロジー予測法 (ConPred_all) を開発した。これによって、予測精度を既存のものよりも10～20%向上させることができた。さらに、予測できる配列の割合を小さくする(20～30%) 代わりに、予測精度をほぼ100%にするというコンセンサス予測法(ConPred_elite)を新たに開発した。これらを統合することによって、統合化TMトポロジー予測法ConPred IIを構築し、ウェブ上に公開した。(論文8、11、22)
- ③TMトポロジー予測法の多くが、シグナル・ペプチド (SP) をTMセグメントとして予測すること、SPの存在がTMトポロジー予測を大きく狂わせることを明らかにした。この結果を考慮に入れて、まずSP予測を行い予測されたSP領域を取り除いた後に、TMトポロジー予測法にかけるという、信頼度の高いTMトポロジー予測の手順を確立した。(論文19、20)
- ④予測によって、87原核ゲノム、12真核ゲノムからTM

タンパク質を網羅的に抽出し、ホモロジー検索による機能アノテーションを行った後、それらのプロテオーム中のTMタンパク質の割合、TMセグメント本数などのTMトポロジー分布、SPを持つものの割合、TMトポロジータイプ (I, II, III) ごとの割合などの解析を行った。TMタンパク質の割合は、真核、原核の違い、ゲノムによらずに、20~25%程度であることを明らかにした。(論文16)

- ⑤配列類似性のみを利用した原核87ゲノムのTMプロテオームの機能同定の現状を調べ、それに加えてTMトポロジー類似性を利用した機能分類・同定を行った。その結果、例えば6本型TMタンパク質では、配列類似性だけでは34.4%であった機能分類・同定率を63.2%まで向上させられた。(論文7)
- ⑥配列内部分配列比較を行い、遺伝子内重複によって形成されたと考えられるTMトポロジーの配列を網羅的に探し出し、TMトポロジーの進化的道筋を明らかにした。特に、"positive-inside rule"を検証することによって、2倍体型重複によって形成されたと考えられる8本型、10本型、12本型のTMタンパク質の構造・機能の進化的成り立ちを詳しく調べた。(論文9)
- ⑦原核87ゲノムを対象として、TMタンパク質のループ上にPfamドメイン、SCOPドメインのアサイメントを行ない、ループ上のドメイン・アーキテクチャーの網羅的解析を行った。その結果、長さが60残基未満のループ上にはドメインが存在しないこと、TMタンパク質特有のドメイン組み合わせが多数存在すること、まだ定義されていない新規のドメインが数多く存在していることなどを明らかにした。(論文5)

〈国内外での成果の位置づけ〉

TMタンパク質の網羅的機能分類の試みが国内外で盛んになってきている。しかしながら、立体構造決定が進んでいないことも一因となって、機能分類もままならず、その結果、構造決定の対象も絞りにくいという悪循環に陥っている。この構造決定に代わる当面の手段はTMトポロジー予測法である。我々が開発した統合化予測法(ConPred II)の予測精度は既存のものより10%以上も高く(予測対象を絞ることによってほぼ100%の予測精度も可能)、有力な武器になるものである。TMタンパク質のドメイン・アーキテクチャーを網羅的に明らかにしようとする研究は現時点でもまだない。

現在、TMタンパク質の構造決定が進んでおり (non-redundant な完全長構造は現時点で45程度)、それらの構造の中にTMタンパク質の特徴的なドメイン・アーキテクチャーが浮かび上がって来ており、Engelmanらのグループなどがそのことを指摘し始めている。

また、遺伝子内重複によるTMタンパク質の進化的成り立ちについての研究も内外で進展しているが、TMトポロジーの基本アーキテクチャーとの関わりで調べている例はまだない。

TMPDBのトポロジーデータを用いて、提案されている多くのTMトポロジー予測法の精度評価を行った論文は多くの論文で引用されており (正確な数字は定かではないが)、現時点でのTMトポロジー予測精度の水準を知りためのスタンダードの一つとして位置づけられていると見られる。また、ConPred II のサービスサイトには多い月には1千件を超えるアクセスがあることがある程に広く利用されており、例えば、PSORTdb (Nucleic Acid Res., 33, D164-D168, 2005)中のトポロジーデータとして取り込まれることになっている。さらに、ConPred IIのプ

ログラムソース入手を求めるメールなども多数届いている。

〈達成できなかったこと、予想外の困難、その理由〉

今後のさらなる追究が必要ではあるが、原核生物由来のTMタンパク質を対象とした研究については目指したことの多くをほぼ達成することができた。しかしながら、真核生物由来のTMタンパク質については、TMトポロジー予測の精度が原核由来のものに比べて低く、TMトポロジー情報を利用した網羅的機能分類に至っていない。

コンセンサス予測法、ConPred IIの開発によって、他の方法に比べて10~20%増の予測精度の達成を実現でき、当面のTMタンパク質の網羅的解析のための使用には役割を果たしうが、さらに高い予測精度・信頼度が必要である。HMMなどを用いた、さらに高精度の本格的予測法の開発が求められている。そのためには、従来の生化学的実験によって得られたTMトポロジーデータだけでは不十分であり、さらに数多くの立体構造データが必要である。しかしながら、この1、2年、急増しているとはいえ、TMタンパク質の立体構造数は絶対的に少なく、そのことが高性能のTMトポロジー予測法の開発を妨げている最大の要因となっている。特に真核由来のTMタンパク質に対してはTMトポロジー予測の精度は低く、例えば、ループ上に存在する水溶性ドメインのアサイメント、新規ドメインの定義をも困難にしている。このように、TMタンパク質のプロテオーム・スケールでの研究において、TMトポロジー予測法の予測精度の向上が大きな鍵を握っていると言える。

〈今後の課題〉

上述のように、決定されるTMタンパク質立体構造数の増大を待ちながらも、新たな高精度TMトポロジー予測法の開発を急ぐことが最大の課題である。そのためには当面、これまでに得られている構造データの詳細な解析を行い、予測を困難にしている問題点などを洗い出し、それを解決することによって精度向上を目指す戦略が有効であろう。現在このアプローチによる試みを遂行しているところである。

原核ゲノムからのTMタンパク質に対しては、ドメイン・アーキテクチャー解析、クラスタリングによる網羅的機能分類・同定を行うことができてきているが、真核TMタンパク質については、前述のTMトポロジー予測精度の問題があって、成功裏には達成されていない。真核TMタンパク質については、原核とは少し違ったアプローチが必要であると思われる。それは、配列類似性比較に立ち戻って機能分類を行うことである。しかしながら、all-against-allのDP配列比較をプロテオーム・スケールで行うことは、計算コストの点からも実現が難しい。一方、多くのタンパク質はマルチドメイン構造を持っており、BLAST等による局所配列比較による分類は正確さを欠いている。そこで、現在、「分割・併合法」を適用したクラスタリング (TMトポロジー情報と配列情報を組み合わせて行う、詳述は省略) によって、真核TMタンパク質配列の網羅的機能分類を試みているところであり、良好な結果が得られつつある。このようにして得られたクラスターを基にして機能グループごとにプロフィールHMMを求め、それらのコレクションとして、TMタンパク質の機能分類・同定のためのデータベース・予測システムを構築する。

また、TMタンパク質におけるドメインの定義を行い、TMタンパク質のドメイン・アーキテクチャーを明らか

にする。TMタンパク質におけるドメイン定義は、TMセグメントとそれらをつなぐ短いループから構成される「TMドメイン・モジュール」と長いループ上の水溶性の構造・機能ドメインの二種類のドメインを区別して行われる必要がある。分類された機能ファミリーを対象として、短いループによって繋がれたTMセグメントを「TMドメイン・モジュール」とし、「水溶性ドメイン」は長いループ上に割り当てられるSCOPドメイン、Pfamドメイン、新規ドメインとして定義することができる。「TMドメイン・モジュール」は、それ自身あるいはそれらの組み合わせによって「TMドメイン」を形成することになる。このような「TMドメイン」と「水溶性ドメイン」の組み合わせとして、TMタンパク質のドメイン・アーキテクチャーを取り出すことができるであろう。このような方法で、ドメイン・アーキテクチャーの網羅的定義を完了させる。さらに、その結果をTMトポロジー予測法と組み合わせることによって、TMタンパク質機能分類・予測システムを構築する。

〈研究期間の全成果公表リスト〉

1) 論文

1. Kimura, T., Ohta, T., Nishio, T., and Shimizu, T., How to obtain multiple alignments in which transmembrane segment regions are well aligned without inserted, *Chem-Bio Informatics Journal*, under revision.
2. Mizuta, S., Munakata, H., Aimaiti, A., Oya, I., Oosawa, K., and Shimizu, T., Analysis of tandem repeats found in 44 prokaryotic genomes, *In Silico Biology*, in press.
- 3.0602081608 Inoue, Y., Yamazaki, Y., and Shimizu, T., How accurately G-protein-coupled receptors can be discriminated as 7-tms TM protein sequences from others?, *Biochem. Biophys. Res. Commun.*, 338, 1542-1546 (2005).
4. Nishio, T., and Shimizu, T., Model analysis of surfactant-polymer interaction as cooperative ligand binding to linear lattice, *Biophysical Chemistry*, 117, 19-25 (2005).
- 5.0602081556 Arai, M., Fukushi, T., Satake, M., and Shimizu, T., A proteome-wide analysis of domain architectures of prokaryotic single-spanning transmembrane proteins, *Comput. Biol. Chem.*, 29, 379-387 (2005).
- 6.0602081429 Mizuta, S., Munakata, H., Aimaiti, A., Oosawa, K., and Shimizu, T., Evaluation of the color-coding method for searching tandem repeats in prokaryotic genomes, *Chem-Bio Informatics Journal*, 4, 133-141 (2004).
- 7.0408231716 Arai, M., Okumura, K., Satake, M., and Shimizu, T., Proteome-wide functional classification and identification of prokaryotic transmembrane proteins by transmembrane topology comparison, *Protein Science*, 13, 2170-2183 (2004).
- 8.0408231722 Arai, M., Mitsuke, H., Ikeda, M., Xia, J.-X., Kikuchi, T., Satake, M., and Shimizu, T., ConPred II: a consensus prediction method for obtaining transmembrane topology models with high-reliability, *Nucleic Acid Res.*, 32, W390-W393 (2004).
- 9.0408231727 Shimizu, T., Mitsuke, H., Noto, K., and Arai, M., Internal gene duplication in the evolution of transmembrane proteins, *J. Mol. Biol.*, 339, 1-15 (2004).
- 10.0403101243 Inoue, Y., Ikeda, M., and Shimizu, T., Proteome-wide classification and identification of mammalian-type GPCRs by binary topology pattern, *Comput. Biol. Chem.*, 28, 39-49 (2004).
- 11.0403101246 Xia, J.-X., Ikeda, M., and Shimizu, T., ConPred_elite: a highly reliable approach to transmembrane topology prediction, *Comput. Biol. Chem.*, 28, 51-60 (2004).
- 12.0403101240 Azumi, R., De Santis, A., De Tomaso, I., Rigoutsos, F., Yoshizaki, M., Pinto, R., Marino, R., Shida, K., Ikeda, M., Arai, M., Inoue, Y., Shimizu, T., Satoh, N., Rokhsar, D. S., Du Pasquier, L., Kasahara, M., Satake, M., and Nonaka, M., Genomic Analysis of Immunity in a Basal Chordate and the Emergence of the Vertebrate Immune System: Waiting for GodotK, *Immunogenetics*, 55, 570-581 (2003).
13. Nishio, T., Shimizu, T., Kwak, J. C. T., and Minakata, A., The cooperative binding of large ligands to a one-dimensional lattice: the steric hindrance effect, *Biophysical Chemistry*, 104, 501-508 (2003).
14. Shimizu, T., The Cooperative Binding of Surfactant Ions by Small Oligomers of Opposite Charge, *J. Phys. Chem. B*, 107, 8228-8231 (2003).
- 15.0309012034 Sugiyama, Y., N. Polulyakh, N., and Shimizu, T., Identification of Transmembrane Protein Functions by Binary Transmembrane Topology Pattern, *Protein Engineering*, 16, 479-488 (2003).
- 16.0303241936 Arai, M., Ikeda M., and Shimizu, T., Comprehensive analysis of transmembrane topologies in prokaryotic genomes, *Gene*, 304, 77-86 (2003).
- 17.0303241933 Ikeda, M., Arai, M., Okuno, T., and Shimizu, T., TMPDB: a database of experimentally-characterized transmembrane topologies, *Nucleic Acids Res.*, 31, 406-409 (2003).
18. Miura, N., Shida, K., Kawashima, R., Kawazoe, Y., Fukuda, H., and Shimizu, T., Automatic Brain Region Extraction Method from 3-D MRI-T1 Weighted Image, *Journal of Computer Assisted Tomography*, 26, 927-932 (2002).
- 19.0304251718 Lao, D. M., Okuno, T., and Shimizu, T., Evaluating transmembrane topology prediction methods for the effect of signal peptide in topology prediction, *In Silico Biol.*, 2, 485-494 (2002).
- 20.0303241929 Lao, D. M., Arai, M., Ikeda, M., and Shimizu, T., The presence of signal peptide significantly affects transmembrane topology prediction, *Bioinformatics*, 18, 1562-1566 (2002).

21.0303241926

Shimizu, T., and Minakata, A., Effect of divalent cations on the volume of a maleic acid copolymer gel examined by incorporating lysozyme, *European Polymer Journal*, 38, 1113-1120 (2002).

22.0309012027

Ikeda, M., Arai, M., Lao, D.M., and Shimizu, T., Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies, *In Silico Biol*, 2, 19-33 (2002).