NAIST-IS-DD0361035

博士論文

アミノ酸配列からの膜タンパク質機能予測に関する研究

村松 孝彦

2006年3月24日

奈良先端科学技術大学院大学
情報科学研究科 情報生命科学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
博士 (理学) 授与の要件として提出した博士論文である。

村松 孝彦

審査委員：

石井 信 教授　(主指導教官)
諏訪 牧子 教授　(委員)
箱嶋 敏雄 教授　(委員)
金谷 重彦 教授　(委員)
川端 猛 助教授　(委員)

アミノ酸配列からの膜タンパク質機能予測に関する研究*

村松 孝彦

内容梗概

膜タンパク質は生体膜に局在しているタンパク質であり、全タンパク質の約20-40%占めているといわれており、生体機能において情報伝達や物質輸送等の重要な役割を担っている。ゲノム解析によって新規に同定された多くの膜タンパク質の機能解析の観点から実験を支援する計算機を用いた手法の開発は不可欠であるが、膜タンパク質に焦点をあてた計算機手法は少ない。本論文では、膜タンパク質のアミノ酸配列からの計算機を用いた機能予測について提案する。

はじめに配列を用いた機能解析法において重要な膜タンパク質アミノ酸配列のペアワイズアラインメント法について述べる。ペア隠れマルコフモデルを用いて、膜貫通領域予測を考慮しながら配列アラインメント法を行うことができる方法を開発し、その手法が従来手法より高い精度であることを示す。次にGタンパク質共役型受容体 (GPCR) という特定のファミリーに対するアラインメント法及び機能予測法について議論する。汎用的なアラインメント手法よりも高精度のアラインメントを作成するために、プロファイル隠れマルコフモデル使用し、さらにモデルに対して立体構造に基づく改良を行った。その手法で作成されたアラインメントを基に2種類の機能予測法を提案する。1番目の機能予測法として、GPCRのペプチド・タンパク質系リガンドの残基長を予測する手法を提案する。アラインメントを基にアミノ酸指標を用いた特徴量を作成し、サポートベクター回帰を用いた。リガンド既知データを用いた評価実験による予測精度とリガンド未知の

* 奈良先端科学技術大学院大学 情報科学研究科 情報生命科学専攻 博士論文, NAIST-IS-DD0361035, 2006年3月24日.

GPCR に適用した結果を示す。2 番目の機能予測法として、GPCR と G タンパク質の共役選択性についての予測方法を提案する。アラインメント情報から共役選択性と関連があると予測される部位を探す手法を提案し、予測部位と共役選択性の関係について考察した。考察から得られた情報と他の物理化学的な特徴を特徴量として定め、サポートベクター分類を用いて共役選択性を予測する手法を提案する。さらに本手法を共役選択性既知の GPCR に適用した評価実験による予測精度を示す。

キーワード

膜タンパク質, 配列アラインメント, GPCR, 機能予測, G タンパク質

Studies on prediction of membrane protein function from protein sequence*

Takahiko Muramatsu

Abstract

Membrane proteins constituting about 20-40% of all proteins are located in biological membrane, and have an important role in biological function, such as cell communication and molecules transporter. It is important to develop computational methods in order to support functional analysis of membrane proteins identified from genome analysis. However, computational methods focused on them have not been sufficiently developed. In this thesis I propose computational methods of protein function prediction from membrane protein sequence.

First, I propose pairwise alignment method for all types of alpha helical membrane proteins. Pairwise sequence alignment method is important for protein functional analysis. The method can solve sequence alignment and predict membrane topology at the same time. Next, I discuss sequence alignment method and protein function method for a particular membrane protein family, G-protein coupled receptor (GPCR). We propose to align GPCR sequences using profile hidden markov mode (HMM). This profile HMM is improved by using GPCR structure information. I propose prediction method for two GPCR functions using the GPCR alignment. First function is about GPCR-ligand interaction. I propose the method to predict ligand residue length of peptide/protein-activated GPCR

* Doctoral Dissertation, Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0361035, March 24, 2006.

using support vector regression (SVR). In SVR amino acid indices are used as feature vector. I evaluated the algorithm using known peptide/protein-activated GPCR sequences, and applied this algorithm to peptide/protein-activated orphan GPCR sequences. Second function is about GPCR-G-protein coupling sensitivity. I propose to predict sites related to G-protein coupling sensitivity from alignment information and discuss relationship between these predicted sites and the function. I propose a system for predicting GPCR-G-protein coupling sensitivity using support vector classification (SVC). In SVC the predicted sites and the other phisico-chemical factors are used as feature vectors. I evaluated the algorithm using known GPCR sequences.

Keywords:

membrane protein, sequence alignment, G-protein coupled receptor, functional prediction, G-Protein

目 次

図 目 次

表 目 次

第1章　序論

1.1. はじめに

膜タンパク質は全タンパク質の 20-40%を占めると言われており [66]、生体膜に局在するタンパク質で細胞内と細胞外との間における情報や物質のやり取りをするタンパク質など細胞に対して重要な役割をするタンパク質が多くみられる。このような細胞内で重要な役割を果たしているものが多い膜タンパク質の機能を解析することは重要である。ヒトをはじめとする様々な生物種でのゲノム解析の結果、膨大な新規の膜タンパク質が同定された。これらの膨大な膜タンパク質をすべて実験的に機能解析を行うためには、多大な時間と人員が必要であると予想される。そのため、生物学実験による機能解析研究を計算機を用いた手法によって支援することで、膜タンパク質機能解析の研究スピードを高めることが必要である。

膜タンパク質はアミノ酸配列の観点からも脂質二重膜に囲まれた立体構造の観点からも水溶性タンパク質と大きく異なっている。それにもかかわらず膜タンパク質に特化した十分な計算機を用いた研究が行われてきていないのが現状である。その理由の 1 つには実験的な難しさにより膜タンパク質の決定されている立体構造が水溶性タンパク質に比べるとかなり少ないことがあげられる。しかしながら、実験手法の発展により、ここ数年は水溶性タンパク質に比べると少ないものの徐々に立体構造データが増えつつあり、今後数年でさらに増える可能性がある。そのような状況をふまえると膜タンパク質の計算機を用いた機能解析では限られた立体構造データと膨大なアミノ酸配列データをうまく組み合わせていかに機能に迫っていくかが重要である。立体構造データと似た構造を持つと推測されるアミノ酸配列の間の橋渡しをする技術が配列アラインメントである。立体構造

を持つタンパク質のアミノ酸配列とターゲットとなる機能未知のアミノ酸配列のアラインメントを構築することにより立体構造データとアミノ酸配列データとの対応関係がとれ、そこから様々な立体構造データを基にしたアミノ酸配列解析が可能となる。

本研究では膜タンパク質アミノ酸配列と立体構造の橋渡しの技術であるアラインメントについて、膜タンパク質に汎用的に使用できる手法と膜タンパク質の特定のファミリーへ適用する手法を提案する。さらに、構築されたアラインメントを基に構造データの一部を参考にしながら機能予測を行う方法を提案する。このような立体構造とのアラインメントを基にした機能予測方法が今後、膜タンパク質の機能解析のスピードをさらに高めていくことが期待できる。

1.2. 本論文の構成

本論文の構成について述べる。まず始めに、第 2 章では研究背景として膜タンパク質の機能解析、アミノ酸配列アラインメント法の現状と関連研究について述べる。また、第 4 章から第 5 章においてモデルケースの膜タンパク質ファミリーとして研究を行った G タンパク質共役型受容体 (GPCR) の現状及び関連研究についてもあわせて述べる。

第 3 章では膜タンパク質全般にし様出るアミノ酸配列のペアワイズアラインメント法について述べる。配列アラインメントは配列と構造、機能を結びつける上で重要な解析手法の 1 つである。ここでは、ペア隠れマルコフモデルを用いて、膜貫通領域予測のモデルとアラインメントのモデルを統合して、膜貫通領域予測を考慮しながら配列アラインメント法を行うことができる方法を開発した。構造アラインメントに基づく正解アラインメントをもとにした評価を行うことで、既存の標準的なグローバルアラインメントよりも性能が向上することを示す。

第 4 章、第 5 章においては膜タンパク質の中からその代表的なタンパク質ファミリーである G タンパク質共役型受容体 (GPCR) に関する研究について述べる。GPCR は主に 5 つのクラス (Class A, B, C, D, E) に分けられるが、本論文では GPCR の大部分が属する Class A に関する 2 つの機能予測法を開発した。

第 4 章ではペプチド・タンパク質系のリガンドを持つ GPCR に関して、アミノ酸配列からそのリガンドの残基長を予測する手法を提案する。GPCR はリガンドと結合することによって活性化し、細胞内にシグナルを送るため、GPCR がどのリガンドと結合するかということは GPCR の生体内の役割を知ることや創薬の観点から重要である。GPCR のリガンドは非常に多様であり、リガンドそのものを直接予測することは難しいため、ここではペプチド・タンパク質系のリガンドの残基長を予測することで、大量のリガンド候補から絞込みを行うための手法を開発する。まず始めに GPCR 専用の配列アラインメント法について述べる。アラインメントの手法としては GPCR 特有の保存配列に着目し、プロファイル隠れマルコフモデルを用いた。さらに、立体構造がわかっているロドプシンの二次構造情報を用いて、膜貫通領域をより正確にアラインメントを行えるように改良した。作成したアラインメントに関して、各カラム及び各構造領域に関してアミノ酸指標を用いた特徴量を作成した。その特徴量に対してサポートベクター回帰を用いてリガンドの残基長を予測した。リガンド既知の GPCR 配列を用いて評価実験を行った。さらにリガンド未知の GPCR に適用し、リガンド未知の GPCR の残基長の分布を予測した。

第 5 章では GPCR の G タンパク質共役選択性予測手法を提案する。G タンパク質は主に 3 種類のグループ ($G_{i/o}$, $G_{q/11}$, G_s) に分かれており、それぞれ細胞に与える影響が違うことが知られている。この共役選択性を機能未知の GPCR に対する予測は、機能未知 GPCR の機能解析のためのアッセイ系構築に役立つ。まず始めに、GPCR の共役選択性に関係する配列的特徴を調べるために、第 4 章で作成したアラインメント法を用いて、各 GPCR と立体構造のわかっているロドプシンとの残基間の対応関係を推定した。そして、ロドプシンの各残基において機能の異なる GPCR 間で出現するアミノ酸の違いを評価した。特に出現するアミノ酸の違う部位について立体構造にマッピングを行い、共役選択性に関するアミノ酸配列に関する特徴を考察した。次に共役選択性を予測するために、アミノ酸出現頻度の実験からの考察から得られた情報とその他の GPCR の特徴をあらわす特徴量を検討し、共役選択性と相関性のあるものを抽出した。それらを特徴量としてサポートベクター分類を用いて共役選択性を予測し、その結果を示す。

第2章　研究背景と関連研究

この章では本論文で行われている研究の背景と現状について説明する。まず始めに研究対象である膜タンパク質について、その特徴と研究の現状について説明する。次に、配列アラインメントについて概説するとともに、膜タンパク質のアラインメントする上での問題点について説明する。最後に膜タンパク質の配列アラインメントを利用した応用例として、第 4 章と第 5 章で述べるタンパク質共役型受容体 (GPCR) についての概要と機能解析における課題について説明する。

2.1. 膜タンパク質

膜タンパク質は細胞膜やオルガネラ膜など生体膜に局在するタンパク質である。計算機による推定ではタンパク質の 20-40%ぐらいが膜タンパク質だといわれている。ここでは、今回の研究対象とする膜タンパク質について述べ、その後膜タンパク質に関する配列データと立体構造データがどのような状況にあるかについて説明する。

2.1.1 研究対象とする膜タンパク質

ここで研究対象となる膜タンパク質の対象について説明する。本研究では膜貫通型タンパク質を研究対象としている。生体膜（細胞膜、オルガネラ膜など）に貫通はしないが、脂質修飾などで膜と結合しているタンパク質もあり、広義の膜タンパク質はそれらも含むが今回の研究対象としては、生体膜内に局在し、膜を貫通する領域を持つものを対象とする。膜貫通タンパク質には α ヘリックス型と β バレル型に分かれ、その構造及び配列の特徴は大きく異なる。α ヘリックス型は

多くの生物種で見られるのに対して、βバレル型はグラム陰性のバクテリア [98] もしくはオルガネラ膜 [8][34] でしか見つかっていない。ここでは膜貫通タンパク質の大部分を占めるαヘリックス型膜貫通タンパク質を研究対象の膜タンパク質として定め、解析を行う。

2.1.2 膜タンパク質に関するデータの現状

計算機を用いた解析においてはどのようなデータ数がの規模と、今後のデータ数の推移に関する情報が、研究の方向性を決める上で重要である。ここでは、アミノ酸配列及び立体構造のデータの現状と今後の動向について簡単に説明する。

アミノ酸配列データ　アミノ酸配列データベース Swiss-Prot [11] 及び TrEMBL [11] に格納されているアミノ酸配列の中で膜タンパク質の数を図 2.1 に示す。キーワードに「Transmembrane」が指定されているものを膜タンパク質としている。ただし、TrEMBL のデータには必ずしもキーワードに関するアノテーションがついているとは限らないため、データベース中の膜タンパク質の総数とは必ずしも一致しない。図 2.1 を見るとわかる通り、配列数は年々指数的に増加している。この配列数の増加の背景には、近年急速に進んでいるゲノム解読の結果であると予想される。ここ数年で多くのゲノムが解かれ、ゲノムベースによるタンパク質の同定が進みタンパク質の配列データ数の増加をさせた。また、計算機研究の側面でも多くの膜タンパク質予測プログラム [57][114][105][47] が開発されて、それらの利用が同定されたタンパク質の機能解析を容易にされたことも膜タンパク質の配列データの増加につながっている。なお、2002 年でデータが増加していないのは、2002 年にデータベースのメジャーバージョンアップが行われていないためである。

このグラフから読み取れることは、膜タンパク質のアミノ酸配列データは現時点でも多数存在していて、今後も多数発見されることが予想される。また、計算機による解析に関しても膜タンパク質か水溶性タンパク質かどうかというレベルの機能予測は多くの方法開発され、一定の成果が得られている。今後は同定され

た新しい膜タンパク質に対してさらに細かな機能予測（ファミリー予測や相互作用予測など）が計算機レベルの研究の課題であると考える。

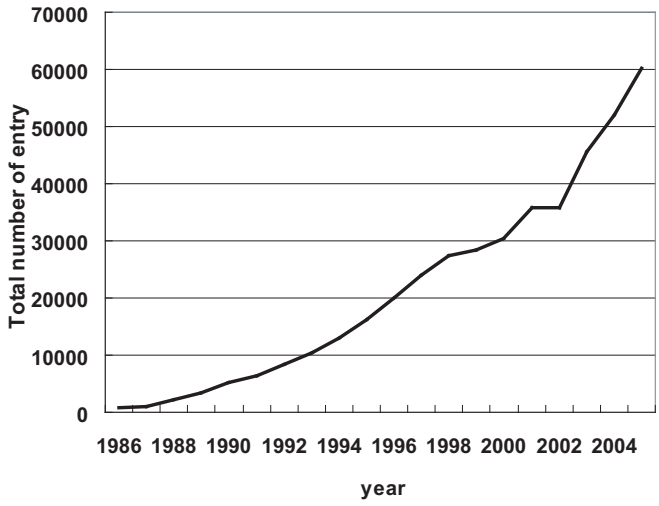

図 2.1 Swiss-ProtS 及び TrEMBL 内の膜タンパク質の総数の遷移。横軸はデータが登録された年数、縦軸はデータの累積数を表す。

立体構造データ　立体構造分類データベース SCOP (1.69 release) において膜タンパク質と分類されているエントリの総数の遷移を図 2.2 に示す。解析対象としている SCOP (1.69 release) 内の総エントリ数が 25,973 個であることを考えると、タンパク質の約 20-40%を占めるといわれる膜タンパク質の立体構造の総数は 243 個とかなり限られている。これは膜タンパク質の立体構造の決定は発現、精製、結晶化の過程が難しいため、構造決定は現在でもハイスループットに決定することは難しいからである [67]。しかしながら、図 2.2 を見る限り、少ないながらも着実に増加しており、今後はこの決定された限られた立体構造を利用して、機能解析につなげていくかが重要である。

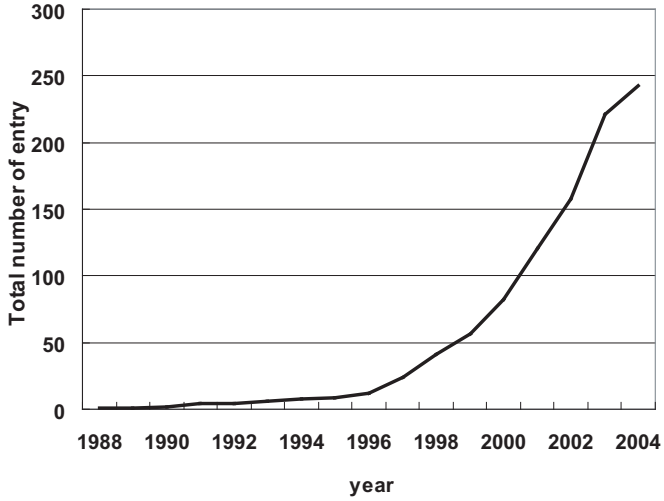

図 2.2 SCOP で膜タンパク質として分類されいている PDB 内の膜タンパク質の総数の遷移。横軸はデータが登録された年数、縦軸はデータの累積数を表す。

2.2. 配列アラインメントと膜タンパク質配列アラインメントの概要

ここでは配列アラインメントの概要と現状を説明し、その後膜タンパク質における配列アラインメントの問題点について説明する。

2.2.1 アミノ酸配列アラインメント

アミノ酸配列アラインメントは進化をモデルとし、置換 (match)、挿入 (insertion)、欠損 (deletion) の 3 つの操作により、2 つのアミノ酸配列を比較し、並べる手法である。一般的に使われるアラインメント手法では、置換の操作にはアミノ酸置換マトリックスという 2 つのアミノ酸の類似度をあらわすパラメータとギャップ開始時に与えるギャップ開始ペナルティ、ギャップが連続する時に与えるギャップ拡張ペナルティの 3 種類のパラメータによって求められる。これらのことは図

2.3 のような有限状態オートマトンで記述することができ、動的計画法を用いて求められる [29]。

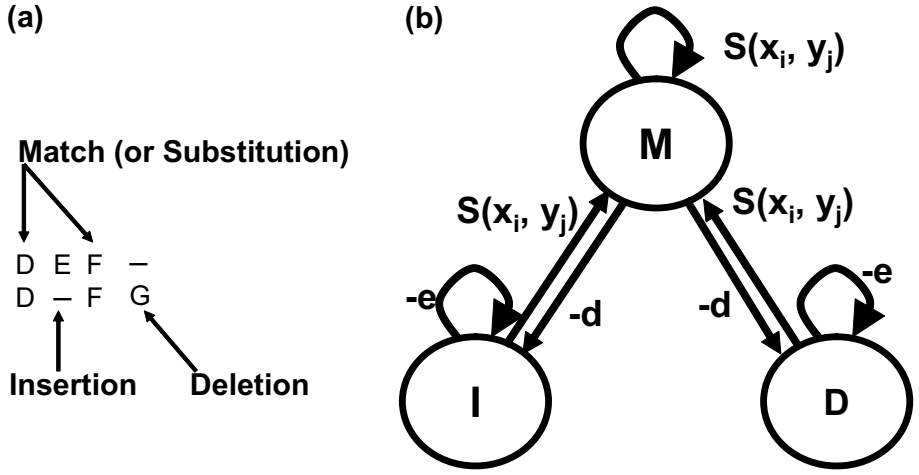

図 2.3 配列アラインメントの概要。(a) 置換、挿入、欠損による配列アラインメントの例 (b) 配列アラインメントの有限状態オートマトン。S(x_i, y_j)、-d、-e はそれぞれアミノ酸置換マトリックス、ギャップ開始ペナルティ、ギャップ拡張ペナルティを表す。

2.2.2 アミノ酸配列アラインメントの現状

配列アラインメントはバイオインフォマティクス研究の中でも古くから研究が行われてきた研究分野の 1 つである。1970 年に Needleman&Wunsch がグローバルアラインメントに関する研究を発表し [87]、その後 1981 年に Smith&Waterman がローカルアラインメントに関する研究を発表した [104]。これらの 2 つのアルゴリズムが基本となり、ローカルアラインメントに関しては主にデータベース検索目的に FASTA [93]、BLAST [4] などの高速なアルゴリズムが開発された。一方、グローバルアラインメントに関しては、複数配列をアラインメントするマルチプルアラインメントの開発がさかんに行われ、CLUSTAL W [110] や T-COFFEE [89]、MAFFT [60]、MUSCLE [31]、PROBCONS [26] など開発され、特にここ

数年の研究の発展が目覚しい。ここであげた手法は全てのタンパク質を扱えることを目標に開発されている。

さらに同様の進化モデルを基にした手法としては、プロファイル隠れマルコフモデル（付録 A.2 参照）もあげられる。これは特定の配列パターンを確率モデルの 1 つである隠れマルコフモデルで記述したものであるが、特定の配列パターンに対しての置換、挿入、欠損という 3 つの操作が可能なモデルとなっている。このように配列アラインメント及びそこで使用されている進化モデルはアミノ酸配列に対する計算機を用いた多くの手法でとりいれられている。

2.2.3 膜タンパク質の配列アラインメントの問題点

一般的に配列一致度が 30%以下となるような配列ペアは「twilight zone」と呼ばれ、正確なアラインメントが難しくなると言われている。こういった配列類似性の低いアラインメントをいかに正確にアラインメントできるかは現在においてもアラインメント研究における重要な課題の 1 つである。膜タンパク質のアラインメントに関しても配列一致度の高い配列ペアでは高い精度でアラインメントを作成できるが、配列一致度の低い配列ペアではアラインメントが困難なケースも見られる。特に膜タンパク質に関しては水溶性タンパク質に比べて立体構造データが限られているため、立体構造予測を行う時に、ターゲットの配列との配列一致度が低いテンプレート構造を基に立体構造を行われるケースは少なくない [63][32][19]。このような配列類似性の低いケースにおいて、高精度な立体構造予測を行うためには膜貫通領域を中心とした膜タンパク質の構造形成に重要な部位について正しくアラインメントを行うことが必要である。

膜タンパク質において配列アラインメントを困難にしている大きな原因は大きく異なる構造領域にある。膜タンパク質の構造領域は大きく分けると膜貫通領域とループ領域（N 末端ループ、C 末端ループも含む。また、ここでのループは膜貫通領域ではない全領域の意味で使用する。）に分けられる。ループ領域は様々な構造をとる可能性が領域である。例えば G タンパク質共役型受容体の 1 つであるロドプシン構造 [92] を見ると、C 末端ループ領域に膜と平行な方向にヘリックス構造である部位が存在する。一方、細胞外ループには β シート構造もみれる。

さらに、3 番目の細胞内ループには特定の構造をとらない、所謂ディスオーダー領域と予想される領域も見られる。このようにループ領域では様々な構造を持つ可能性がある。また、ループ領域では水溶性タンパク質と同様に領域の周辺の環境は水分子で囲まれている。このようなループ領域では、水溶性タンパク質配列のアラインメントと同じようなパラメータでアラインメントをすることが求められる。一方、膜貫通領域では今回ターゲットとしている α ヘリックス型膜タンパク質では一部の例外を除くとヘリックス構造となっている。ヘリックス構造では基本的には規則正しい構造をとっているため、配列アラインメントの観点からはギャップは入らないと考えるのが妥当である。よって、ギャップペナルティは大きな値であることが望ましい。また、膜貫通領域の周辺の環境を考えると脂質分子に囲まれている。そのため、ループ領域で出現するような強い極性を持つアミノ酸の出現は低く、全体的に疎水性の高いアミノ酸の出現が多い。このようにアミノ酸の出現頻度大きく異なり、アミノ酸置換マトリックスに関しても一般で使われるようなマトリックスと異なるものが必要になる。すなわち、構造領域ごとに最適なパラメータが大きく異なっている。特殊な膜貫通領域のパラメータに対応するために、いくつかの膜貫通領域用のアミノ酸置換マトリックスが提案されている [58][88]。しかしながら、膜貫通領域用のパラメータを用意することだけでは本質的な解決にはならない。なぜならば、アラインメントを作成したい配列に対して、立体構造情報や膜貫通領域に関する実験情報が何も入手できない場合、どの領域が膜貫通領域かがわからず、適切な領域で適切なパラメータを使用することが難しいからである。

膜タンパク質専用のアラインメントの先行研究としては Shafrir&Guy が STAM (simple transmembrane alignment method) というマルチプルアラインメント法を提案している [101]。この手法ではまず始めに疎水性指標とヘリックス指標を用いて各配列個別に膜貫通領域を予測する。次に予測された膜貫通領域同士を膜貫通領域用のパラメータを用いてアラインメントを行う。この手法では膜貫通領域には膜貫通領域用のパラメータを用いることができるため、膜貫通領域のギャップの問題も解消できる。しかしながら、膜貫通領域予測が外れた場合アラインメントに多大な影響を与える可能性がある。例えば 2 本貫通の膜タンパク質の配列

A, B をアラインメントする場合に、A が 1 本目の膜貫通領域を予測できず、もう片方が 2 本目の膜貫通領域が予測できなかった場合、この手法では A の 2 本目の膜貫通領域と B の 1 本目の膜貫通領域をアラインメントすることになるため、まったく見当違いのアラインメントを作成することになる。すなわち、STAM は膜貫通領域予測の精度に大きく依存している。Möller らによって様々な膜貫通領域予測法の性能を評価した論文 [82] によると、膜貫通領域予測で一番高い予測精度で予測できる TMHMM [105] でも、膜タンパク質の全ての膜貫通領域を正確に予測できる割合は 60%程度であると述べている。すなわち、膜貫通領域予測の現状では、膜タンパク質に関してすべての膜貫通領域を正確に予測できるレベルにはないため、膜貫通領域予測に強く依存したような手法を使うことは難しい。

第 3 章では膜貫通領域とアラインメントを同時に行う手法を提案する。膜貫通領域とアラインメント構築の両方を同時に行うことで、膜貫通領域を考慮しつつ各構造領域のアラインメントパラメータを用いてアラインメントを行うことができる。さらにアラインメントを考慮しつつ膜貫通領域を行うため、対応関係にない膜貫通領域をアラインメントする可能性を大幅に減らすことができると推測される。

2.3. G タンパク質共役型受容体 (GPCR)

ここでは第 4 章、第 5 章では膜タンパク質の特定のファミリーに限定したアラインメント法及び機能予測法の開発を行っている。具体的には G タンパク質共役型受容体 (GPCR) を取り上げている。以下の理由で GPCR を選択した。

- ヒトで約 1000 個と大きな膜タンパク質ファミリーである。
- 立体構造がわかっているロドプシンが属する。
- 多様なリガンドの受容体ファミリーで生体内で重要な役割をしている。
- 創薬の重要なターゲットである（市場の薬の約半分が GPCR 関連 [27]）。

本節では GPCR についてその概要及び関連研究について述べる。

2.3.1 GPCR の概要

GPCR の機能メカニズム　GPCR の機能メカニズムを図 2.4 に示す。不活性状態の GPCR に対して細胞外からのリガンドと呼ばれる物質が結合すると、GPCR は活性化される。活性化された GPCR は G タンパク質と相互作用し、G タンパク質が活性化される。活性化された G タンパク質は α サブユニットと $\beta\gamma$ サブユニットに分かれて、それぞれのサブユニットが細胞内に様々な影響を与える。G タンパク質は α サブユニットの種類により主に 3 つのグループに分けられる [13]。それぞれのグループで細胞内に異なる影響を与える。グループごとの機能は表 2.1 に示す。

表 2.1　三量体 G タンパク質の 3 つの主要なファミリーとその機能 [13]

major group	minor group	機能
$G_{i/o}$	G_i	アデニリル・シクラーゼを抑制
	G_t	脊椎動物の網膜細胞で細胞でサイクリック GMP を活性化
$G_{q/11}$	G_q	ホスホリパーゼ Cβ を活性化
G_s	G_s	アデニリル・シクラーゼを活性化、カルシウムチャネルを活性化
	G_{olf}	嗅覚神経細胞でアデニリル・シクラーゼを活性化

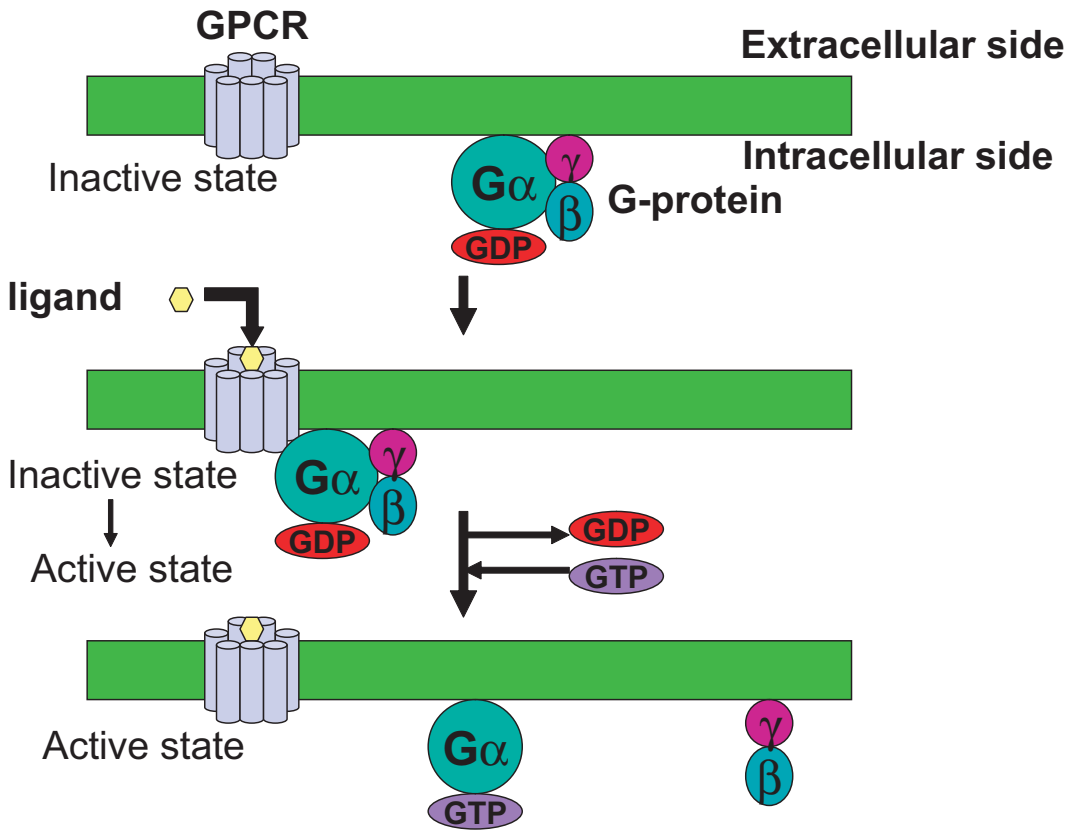

図 2.4　GPCR の機能メカニズムの概要。Gα、β、γ はそれぞれ三量体 G タンパク質の α、β、γ サブユニットを表す。

GPCR の構造領域と立体構造情報　現在 GPCR の中で立体構造が決定されているのは bovine のロドプシン [92] のみである。bovine のロドプシンの簡単な概略図を図 2.5 に示す。GPCR は 7 つの膜貫通領域と N 末端ループを含む 4 つの細胞外ループ、C 末端ループを含む 4 つの細胞内ループで構成されている。本論文では細胞内ループについては配列の N 末端側から細胞内ループ 1、細胞内ループ 2、細胞内ループ 3、C 末端ループもしくは IL1、IL2、IL3、CTL と呼ぶ。同様に細胞外ループは N 末端ループ、細胞外ループ 1、細胞外ループ 2、細胞外ループ 3 もしくは NTL、EL1、EL2、EL3 と呼ぶ。膜貫通領域は膜貫通領域 1-7 もしくは TM1-7 と呼ぶ。bovine のロドプシンに関しては、C 末端ループに膜に平行な 8 つ目のヘリックスも存在する。

bovine のロドプシンの立体構造を図 2.6 に示す。ロドプシンの構造は細胞外側から視点では、反時計回りで 1 本目の膜貫通領域、2 本目の膜貫通領域、$\cdots$ 、7 本目の膜貫通領域となっており、7 本目の膜貫通領域と 1 本目の膜貫通領域が互いに隣接している。なお、この立体構造は不活性状態のロドプシンを表し、リガンド結合後の活性化状態では別の構造となっていると推測されるが、活性化状態の立体構造はまだ決定されていない。

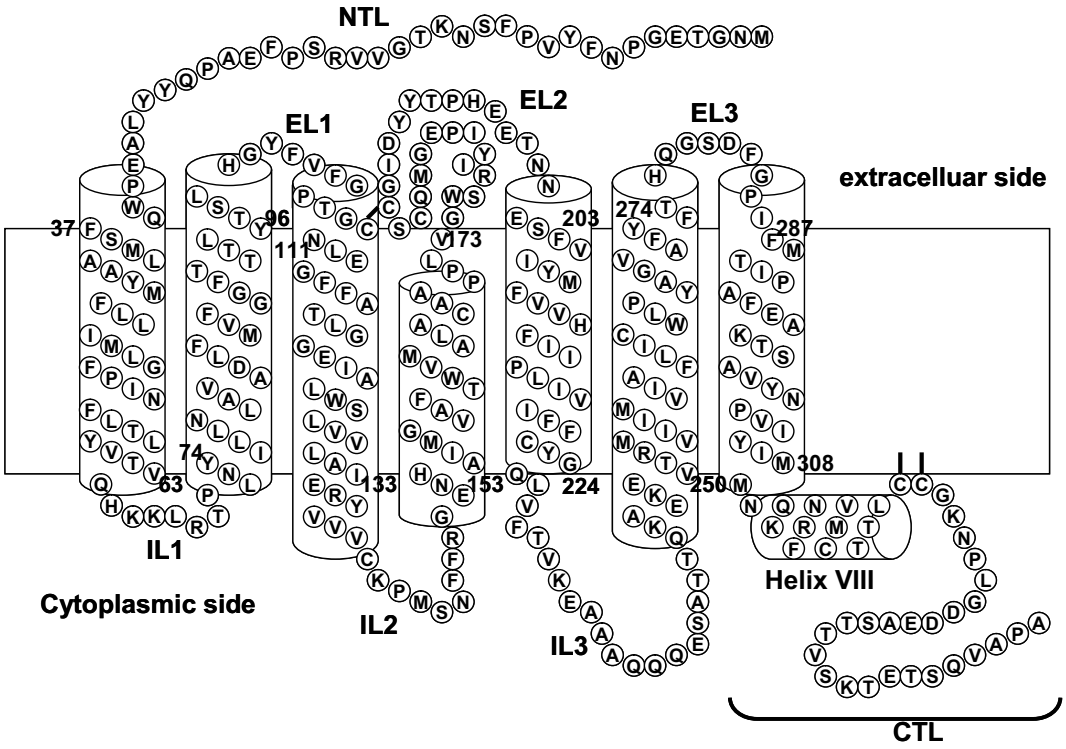

図 2.5 bovine のロドプシンの概略図。下側が細胞内側、上側が細胞外側を表す。NTL、CTL はそれぞれ N 末端ループ、C 末端ループ、EL1、EL2、EL3 はそれぞれ細胞外ループ 1、細胞外ループ 2、細胞外ループ 3、IL1、IL2、IL3 はそれぞれ細胞内ループ 1、細胞内ループ 2、細胞内ループ 3 を表す。ロドプシンの場合は 7 本の膜貫通ヘリックスに膜に平行な 8 本目のヘリックスも存在する。

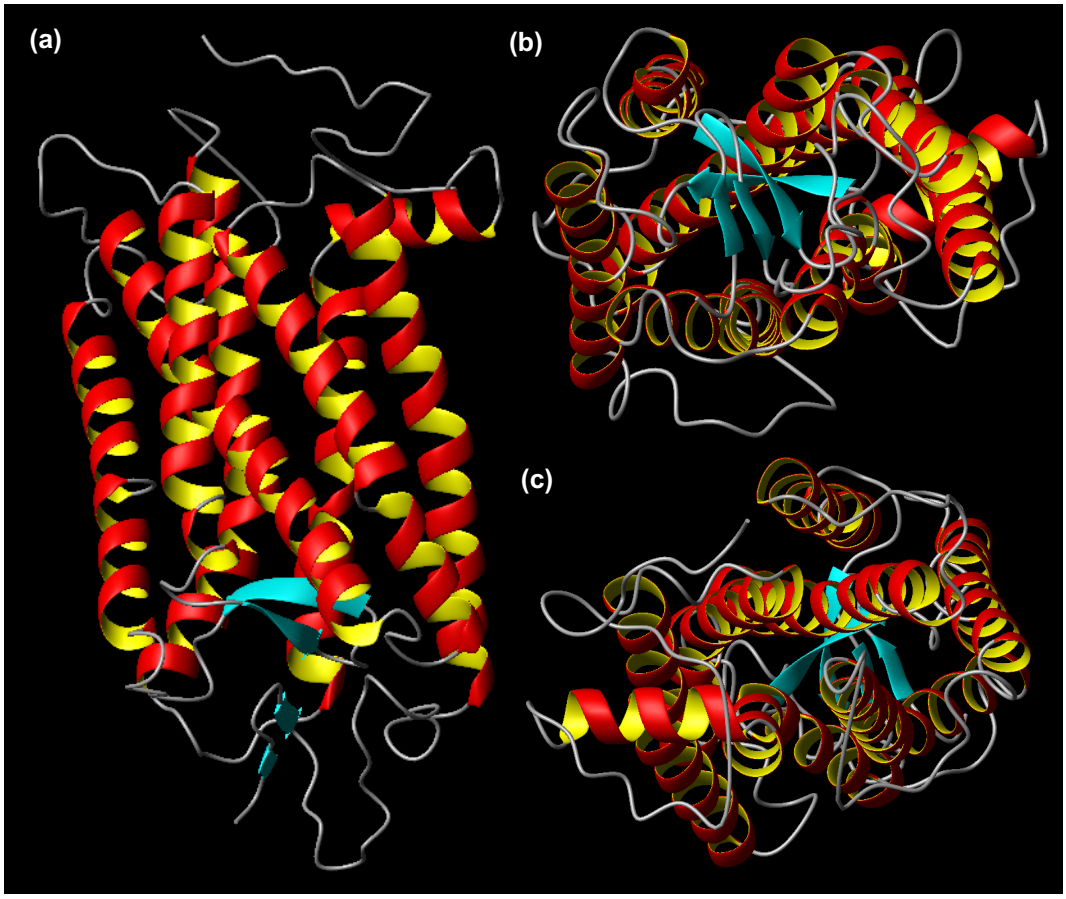

図 2.6　bovine のロドプシンの立体構造のリボンモデル (PDB code 1L9H chain A) [90]。(a) 横から見た構造。上側が細胞内側、下側が細胞外側 (b) 細胞外側から見た構造 (c) 細胞内側から見た図。MOLMOL [64] にて描画。

GPCRの分類　GPCRに関する階層的な分類が知られている。本論文ではGPCRデータベースGPCRDB [50][48] における分類を使用している。図2.7に分類の一部を示す。まず始めにGPCRはClass AからClass Eとその他に大きく分類されている。これは、アミノ酸配列の特徴（配列類似性、モチーフ、N末端の長さなど）によって決定されている。図2.8にヒトGPCRにおける各Classの配列数の割合を示す。ヒトにはClass DとClass Eに属するGPCRは存在しない。918配列が分類されているClass Aが全Class中で大部分を占める。otherの中には「Putative / unclassified Class A GPCRs 」というカテゴリーも含むため、将来的に分類されるClass Aはさらに多いと予想される。このようにヒトのGPCRではClass Aに属するものがほとんどである。本論文の第4章、第5章ではClass A GPCRのみを研究対象としている。その理由は、図2.8で示すようなデータ数の多さ、立体構造が決定されているロドプシンがClass Aに属しているといった理由でClass Aが選ばれている。

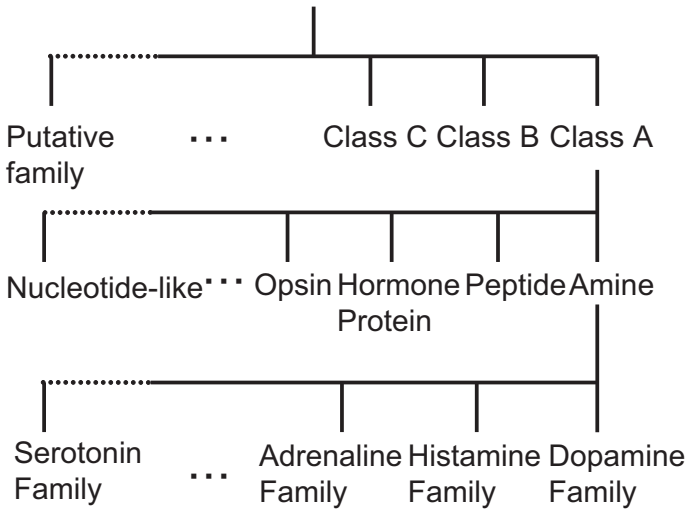

図 2.7　GPCRはクラスにまず分類され、Class A GPCRはさらにサブクラス、ファミリーと階層的に分類される。サブクラスはリガンドの種類によって分類されている。その他のクラスのGPCRはファミリーに分類される。

次にClass Aの下の階層（本論文ではサブクラスと呼ぶ）について説明する。

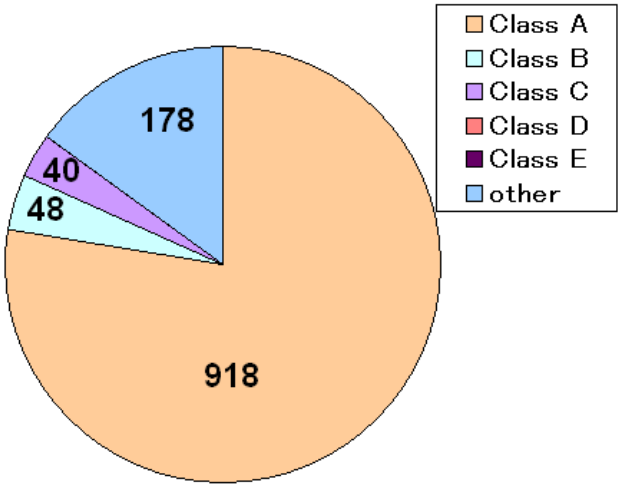

図 2.8 GPCRDB 内のヒト GPCR における各クラスの配列数の割合

ここでは Class A GPCR を相互作用するリガンドの大まかなタイプによって分類されている。図 2.9 はヒトのサブクラスの分類ごとの配列数の割合を示す。図 2.8 を見るとわかるように、Class A の中では Olfactory（嗅覚受容体）が 534 配列と大多数を占める。嗅覚受容体はヒトでは 11 番染色体をはじめ、様々な染色体で多くの遺伝子が発見されている [130]。次に多いのが Peptide で 170 配列存在している。GPCR の中には様々なペプチドと結合するものが存在する。数残基のペプチドリガンドから 100 残基を越えるペプチドリガンドまで存在する。その次に多いのは Amine で 66 配列存在している。Amine は Adrenaline や dopamine などのサブクラスであり、そのほとんどが神経伝達物質で構成されている。なお Orphan はオーファン GPCR であり、機能未知の GPCR の総称である（以下、機能未知 GPCR のことをオーファン GPCR と呼ぶ。）その他は 30 配列以下とあまり多くはない。サブクラスの中で Peptide は Olfactory と Orphan を除けば、54%もしめており Class A の中ではかなりのデータ数が存在している。第 4 章でペプチドリガンドに焦点をあてている理由の 1 つには、サブクラスの中でかなりの割合を占めることが大きな理由である。

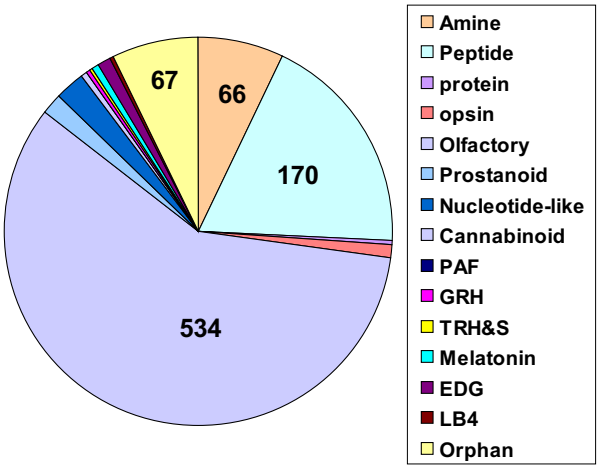

図 2.9 GPCRDB 内でのヒトの Class A GPCR のサブクラスによる分類ごとの配列数の割合

サブクラスの下の階層（本論文ではファミリーと呼ぶ）について説明する。図 2.7 に例示してある Dopamine, Adrenaline のような特定のリガンドと結合する GPCR のグループをファミリーとしている場合や特定のリガンド群をもとにファミリーとしている場合がある。GPCR の多くはリガンドに基づいてファミリーを決めている。

2.3.2 GPCR の機能解析の課題とそれに対する計算機による解析

ゲノム情報から GPCR を発見する計算機を用いたいくつかの手法が提案されており [4][91][36]、それを基にたくさんのオーファン GPCR が同定され、その機能解析が重要となっている。GPCR の機能解析といっても様々な課題が存在するが、その中で最も重要視されている課題はオーファン GPCR のリガンド決定である。GPCR のリガンドを決定することによって、GPCR を活性化に導くアゴニスト（作用薬）や機能を阻害するアンタゴニスト（阻害薬）の設計にもつながり、GPCR の生理的機能の解析も容易になる。すなわち、GPCR の機能メカニズムの解明や創薬の観点からはオーファン GPCR のリガンド同定が研究の鍵であり、

この課題を支援することが計算機による研究においても重要なことである。

体内に実際に存在する内因性リガンドを同定を考えた場合、実際その組織からリガンド候補物質を抽出し、抽出物をオーファン GPCR に相互作用するか試してみるアプローチが一般的な方法である [21]。しかしながら、このようなアプローチは非常に時間のかかる作業であり、オーファン GPCR のリガンド同定における作業でボトルネックとなっている。ペプチドリガンドに限ればゲノムからペプチドリガンド候補を探索するという新しいアプローチが存在する [46][56]。ペプチドリガンドやタンパク質リガンドは基本的にはゲノム上に遺伝子としてコードされており、ゲノムデータベースやアミノ酸配列データベースにデータとして格納されている可能性が高い。今後はこれらのデータベースから様々な予測法を用いて GPCR のリガンド候補を収集し、それらからリガンドを探索するというようなアプローチが増えてくると考えられる。そのような中で重要な研究は大量に集まると推測されるリガンド候補をどのように優先順位をつけて実験を行っていくかである。優先順位付けのための手法の 1 つとしてリガンドの残基長に着目した手法を第 4 章で述べる。

ここまでは GPCR とリガンドとの相互作用に関する先行研究をみてきたが、リガンド同定の実験を行う上では、G タンパク質側との相互作用も重要である。上述したアプローチでリガンドを決定する場合、リガンド候補とオーファン GPCR との相互作用を直接実験的に観察することは難しく、実際には G タンパク質を通じて送られる下流のシグナル（cAMP の濃度変化など）を検出する必要がある。本章の 2.3.1 で見たように、GPCR は共役する G タンパク質の種類によって下流に送られるシグナルが異なり、それぞれ別の測定方法（アッセイ系）が必要である。もちろん、全ての G タンパク質の種類に対するアッセイ系を構築すればよいが、その分大きなコストがかかるため、可能ならば優先順位がつけられることが望ましい。そこで、オーファン GPCR がどのような G タンパク質共役選択性を持つか解析することは有用である。

共役選択性の解析ではキメラ実験や変異実験など生物実験による多くの解析が行われているが、未だに共役選択性のメカニズムはよくわかっていない [121]。これまでの研究によると細胞内ループ 2(IL2)、細胞内ループ 3(IL3) の N 末端側と

C 末端側が重要な働きをしているといわれている [121]。しかしながら、細胞内ループ 1(IL1) や C 末端ループ (CL)、細胞外ループや膜貫通領域も共役選択性に重要であるといった実験データ [123] もあり、結局のところどの部位が重要であるかということはよくわかっていない。

計算機による解析については GPCR の配列に関して同じファミリー内でのマルチプル配列アラインメントによる異なる共役選択性を持つもの同士の比較 [49] が行われているが、特に共役選択性に対して、新しい知見を発見できていない。また、共役選択性予測手法について一般的な機能予測手法としては BLAST [4] などによるホモロジー検索があげられるが、配列類似性が高い配列ペアても必ずしも共役選択性が一致しない。例えば、ムスカリン性アセチルコリンファミリーではサブタイプ 2 とサブタイプ 3 では配列一致度が 45%と高いが、異なる共役選択性を表す。このように配列類似性が高いからいって同じ共役選択性であるとは限らないことを考えると、配列類似性以外の情報による予測方法が必要である。ホモロジー検索以外のすでに提案されている予測手法としてはいずれも細胞内ループの配列パターンのみを考慮している。既存の予測手法としては、弱いモチーフをたくさん探索し、集めたモチーフの有無で判別する手法 [83]、細胞内ループの配列をその配列類似性を基にいくつかのグループにわけそのグループ内で配列情報からベイズモデルを用いて判別する手法 [15]、プロファイル隠れマルコフモデルでループのモデルを作成して判別する方法 [106][99]、さらにループごとにプロファイル隠れマルコフモデルを用いてそこから得られるスコアを基にニューラルネットワークを用いて判別する手法 [100] などが提案されている。現在のところ共役選択性に重要な配列パターンというものは発見されていないのが現状である [120][83]。このように配列パターンが共役選択性に関与している知見がない上でプロファイル隠れマルコフモデルなどを用いた手法を用いると本当に共役選択性に関係する配列パターンによって判別しているのか、近縁の配列の中で共役選択性の同じ配列情報の影響で判別しているのかがわからず、既知の GPCR 配列との配列類似性が低いオーファン GPCR へ適用可能か疑問が残る。そこで、配列パターン以外の視点を加えた情報基に予測を行う手法の開発が必要であると考える。そのような手法の 1 つを第 5 章で述べる。

第3章　膜タンパク質のペアワイズアラインメント法

本章では汎用的な膜タンパク質の配列アラインメント法について述べる。配列アラインメントは第 2 章 2.2 で述べたように、アミノ酸配列を用いた構造解析もしくは機能解析において重要な役割を果たしている。しかしながら、膜タンパク質は他のタンパク質とは違い、構造領域によってその配列的構造が異なる。そこで、各構造領域において最適なアラインメントパラメータを用いて配列アラインメントを構築する手法を本章で提案する。まず始めに、その手法について述べ、その後構造アラインメントを基にして作成した配列アラインメントの正解データを下にその手法の評価を行った。

3.1. データセットと方法

3.1.1 ペア隠れマルコフモデル (PHMM)

膜タンパク質について膜貫通領域の位置及び各ループが細胞内ループか細胞外ループかという情報を膜貫通トポロジーと呼ぶ。これまでたくさんの膜貫通トポロジー予測法が開発されてきたが、その中で高精度で予測できている方法では隠れマルコフモデル (HMM) を使用している [114][105]。ここでは HMM の拡張であるペア隠れマルコフモデル (PHMM) を使用して、膜貫通トポロジー予測の先行研究に対して拡張を加えることによって、膜貫通領域トポロジーモデルとアラインメントを同時に行えるような手法を提案する。PHMM は Durbin [29] により提案された確率モデルで、スタンダードなグローバルアラインメントとほぼ同等であることがわかっている。PHMM はこれまでは遺伝子発見 [79][3] やノンコー

ディング RNA 予測 [97] の手法に取り入れられている（HMM 及び PHMM の理論に関する説明は付録 A を参照）。PHMM を用いることによって、各構造領域（膜貫通領域、ループ領域）に対して、それぞれ固有の出力確率及び遷移確率を割り当てることができる。これは各構造領域に対してそれぞれ固有の置換マトリックスやギャップペナルティを割り当てていることとほぼ同等の意味を持つ。このようにして、膜貫通領域を予測しながら、各構造領域固有のアラインメントパラメーターを用いてアラインメントが可能となる。

3.1.2 膜タンパク質のアミノ酸配列の特徴

膜タンパク質配列アラインメントのモデルを説明する前に、膜タンパク質のアミノ酸配列の特徴を説明する。アミノ酸配列の特徴を図 3.1 に示す。まず大きくは膜貫通領域とループ領域に分かれるが、膜貫通領域ではループ領域に比べて疎水性アミノ酸の出現頻度が全体的に多いことが知られている [68]。ループ領域は細胞内ループと細胞外ループに分かれるが、細胞内ループは細胞外ループに比べ、膜電位の関係から塩基性アミノ酸（特に Arg, Lys）の出現頻度が多くなっている [115]。膜貫通領域を細かくみると膜貫通領域の中央部と膜貫通領域の両端（キャップと呼ぶ）に分かれるが、このキャップ部分では芳香族アミノ酸や側鎖が長い極性残基（特に Lys, Arg）の出現頻度が中心部に比べ相対的に多い [81]。現在開発されている多くの膜貫通領域予測手法は採用するアルゴリズムは違うものの、基本的にはここでの述べた特徴の一部もしくは全部を用いて予測している [115][114][105][47]。

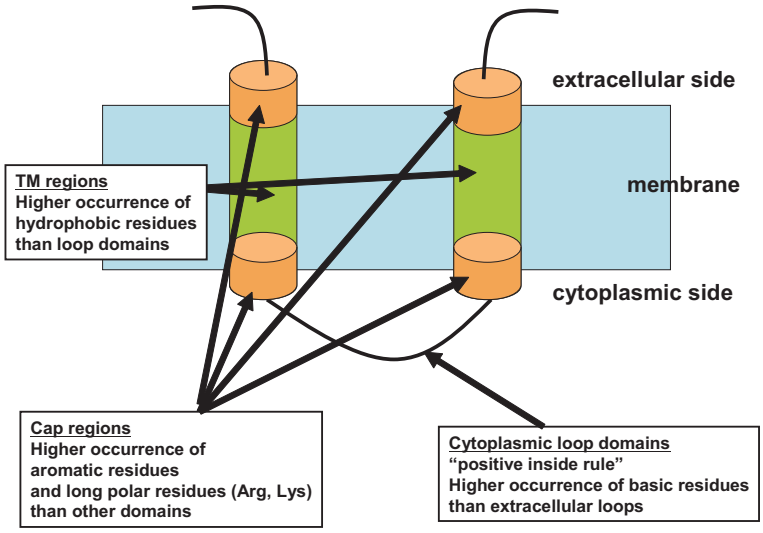

図 3.1 膜タンパク質のアミノ酸配列の特徴。下側が細胞内側、上側が細胞外側。配列のアミノ酸出現に特徴が見られるのは膜貫通領域、膜貫通領域の両端のキャップ領域、細胞内側のループの 3 つの領域である。

3.1.3 PHMM のアーキテクチャー

PHMM のモデルのレイアウトを図 3.2 に示す。図 3.2(a) が示す通り、このモデルは 7 つの構造セグメント（長い細胞内ループ、短い細胞内ループ、長い細胞外ループ、短い細胞外ループ、膜貫通ヘリックス、N 末端ギャップ領域、C 末端ギャップ領域）から構成されている。

ループには細胞内側、細胞外側の両方のループについて、長いループと短いループが存在している。長いループの方が短いループに比べて様々な長さをとり、2 つの配列についてループの長さの差が大きくなる可能性が高くなるので、長いループは短いループに比べて、ギャップが出現しやすくなっている。
図 3.2(b) は細胞内ループと細胞外ループの詳細である。このループ領域では置換、挿入、欠損の 3 つの状態が存在する。細胞内側と細胞外側のループではアミノ酸の出現頻度が違うことが知られており [115]、そのためそれぞれの出力確率は異なっている。一方長いループと短いループの出力確率は同じとしている。図 3.2(c) は両末端のギャップの領域を示している。この領域は挿入と欠損の状態で構成されていて、アラインメントの各両端に配置されており、連続した挿入もしくは欠損のみしか出力しない。これは、N 末端や C 末端の長さが大きく違い、多くのギャップを出力するときに有効である。図 3.2(d) は膜貫通領域のモデルであり、膜貫通コアとキャップ領域で構成されている。キャップ領域は膜貫通コアの両端に配置されている。膜貫通領域の長さはキャップ領域を含めて 15 から 31 残基となっている。キャップ領域の長さは 5 残基である。キャップ領域は細胞内側と細胞外側でわかれ、それぞれ出力確率は異なっている。基本的には膜貫通領域はヘリックス構造をとるため、ギャップは存在しえないが、ごくまれに特殊な構造をとり、ギャップをとりうるが、今回は簡単のため、ギャップは出力しないようにした。

3.1.4 パラメータ推定

モデルのパラメータは既知構造の配列を含むマルチプル配列アラインメントから推定する。まずはじめに、立体構造から膜タンパク質を自動的に収集したデー

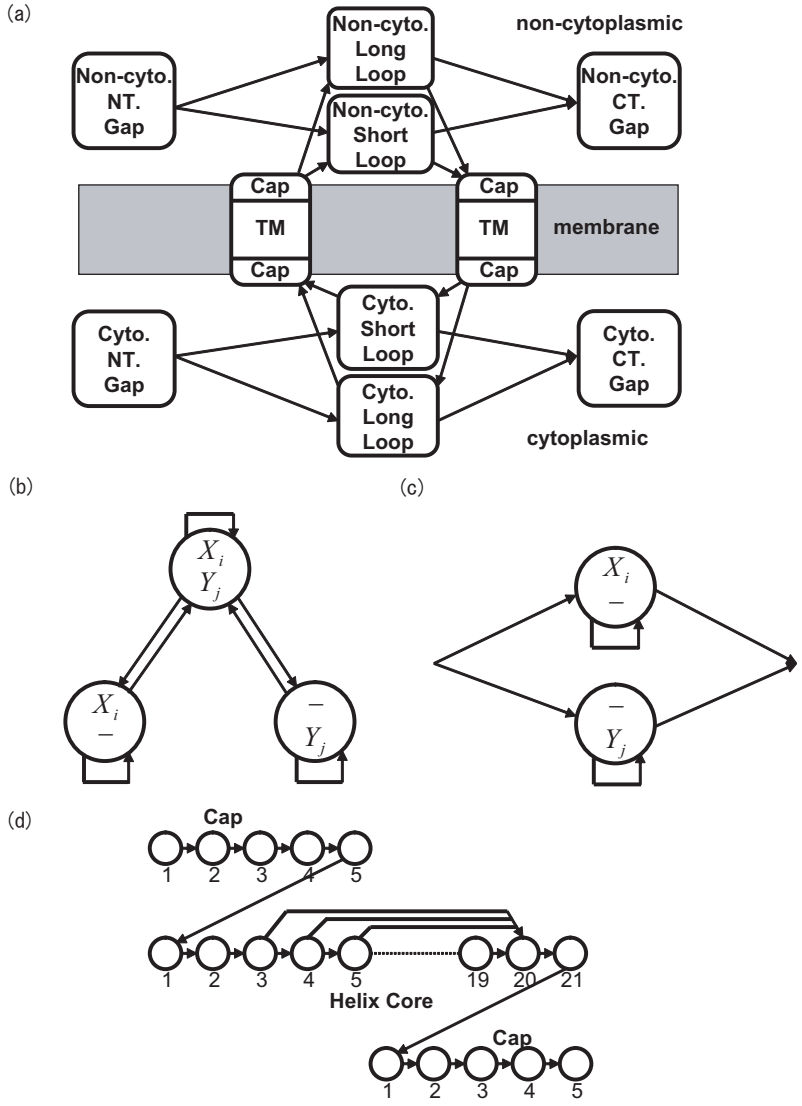

図 3.2 膜タンパク質配列アラインメントの PHMM のレイアウト。(a) はモデルの全体のレイアウトを示す。non-cyto. は細胞外、CT は C 末端ループ、cyto. は細胞内、それぞれの構造セグメントは (b)、(c)、(d) のいずれかの複数の状態から構成されている。(b) は通常のアラインメントを行う領域を表し、置換、挿入、欠損のいずれの操作もとりうる。ループ領域はこの状態群で構成されている (c) この領域は挿入もしくは欠損のいずれかのみを連続してとりうる。配列の両端に配置されている。(d)（キャップを含む）膜貫通領域の状態群である。膜貫通領域は 15-31 残基の置換状態を出力する。

タベース PDB_TM [113] から 62 個の構造データを収集した。PDB_TM は PDB [12][25] 内の立体構造について、その側鎖の疎水性などから膜貫通領域だと予想される領域を推定し、その結果を格納してあるデータベースである。ただし、評価セットや理論モデル、ごくわずかしかわかっていない構造などの不適切な構造については除いてある。次にその構造のホモログ配列を見つけるために BLASTP [4][5] を用いて、タンパク質の配列データベース Uniprot 5.0 [7] に対して検索を行った。E-value のカットオフ値は 10^{-5}とし、他のオプションはデフォルトである。次に発見したホモログ配列群と対象としている構造のアミノ酸配列とをマルチプル配列アラインメントプログラム $MUSCLE$[31] を用いてマルチプル配列アラインメントを生成した。これらのマルチプルアラインメントの膜貫通トポロジーはその中に含まれる立体構造既知の配列の膜貫通トポロジーと同じであると仮定した。なお、立体構造の膜貫通トポロジーは PDB_TM から収集した。パラメータについて、出力確率はアミノ酸置換マトリックスである $BLOSUM$ で用いられた計算方法 [43] と同じ計算方法で推定した。置換確率はN個のマルチプル配列アラインメントをn(n-1)/2 個のペアワイズアラインメントに分割し、膜貫通領域のアノテーションが正しいと仮定し、状態が既知として推定した。なお、ループには長いループと短いループがあるが（3.1.3 参照)、30 残基より長いループを長いループと定義し、30 残基以下のループを短いループとしてパラメータを推定した。このように推定したモデルに対して、Viterbi アルゴリズムを用いて最適な PHMM のパスを求め、そのパスが出力する残基ペア列をアラインメントとした。PHMM のアルゴリズムについては付録 A.3 に書かれている。

3.1.5 評価方法

構造アラインメントを利用した評価セットを用意した。これらの評価セットを作成するに当たり、SCOP1.65 データベース [84] のファミリーの分類を使用した。SCOP データベースは専門家によって手動で作られた構造分類データベースであり、その分類は自動的に作られるものよりも信頼できる。また、各構造分類における代表構造は ASTRAL [17] から抽出した。評価セットは同じファミリーに分類されているタンパク質生物種レベル（同一のタンパク質でも、生物種が違うも

のはそれぞれ異なる構造とする）代表構造ペアについて、構造アラインメントの配列類似性が 50%以下のもので構成され、そのデータ数は 50 ペアである。これらのペアについて構造アラインメントを基にした配列アラインメントを生成することによって、信頼できる正解アラインメントを構築する。構造アラインメントには CE [102] と MATRAS [61] を使用して、両方の構造アラインメントにおいて一致しているところをコアブロックとし、この領域が評価対象のアラインメントとどの程度一致しているかによってアラインメントの精度を評価する。一部の構造ペアに関しては、構造的類似度が低くなっているため、CE と MATRAS の構造アラインメントの結果がほとんど一致しない場合がある。このような構造ペアは信頼できるコアブロックを形成しているか疑問が残るため、以下ように定義されるカバレッジを使用して、このような構造ペアを取り除いた。

$$\text{カバレッジ} \quad = \quad \frac{L_{core}}{min(L_A, L_B)} \tag{3.1}$$

L_{core} はコアブロックの長さ、$L_A L_B$ はそれぞれ配列 A,B の長さを表す。評価セットではカバレッジが 0.5 以上のもののみを使用する。また、上述の配列類似性には CE によって算出したものを使用する。

ペアスコア (PScore) はコアブロックのアラインメントの精度を決めるために計算する。配列 A,B のコアブロックが M カラム（残基ペア）あるとすると、アラインメント内の各々の残基ペア (A_i, B_i) について、A_i と B_i が正解アラインメントでもペアとなっている場合に $p_i = 1$、そうでない場合に $p_i = 0$ となるように p_i（i はカラム番号）を定義する。配列 A と B 間でのスコア S_j は以下のように表す。

$$S_j \quad = \quad \sum_{i=1}^{M_j} \frac{P_i}{M_j} \tag{3.2}$$

M_j は評価セット j でのカラム数を表す。

評価セット全体での PScore は以下のようになる。

$$PScore \quad = \quad \sum_{j=1}^{N} \frac{S_j}{N} \tag{3.3}$$

N は評価セットの配列ペア数を表す。

3.2. 結果と議論

3.2.1 PScoreの比較

今回提案した手法を評価するために標準的なグローバルアラインメントであるALIGN [85] を用いて比較評価した。一般的にアラインメントにおけるアミノ酸置換マトリックスに何を使用するかがALIGNに対して影響を与えると考える。膜貫通領域用のアミノ酸置換マトリックスもJTT [58] やPHAT [88] が提案されている。ALIGNのパフォーマンスを計るにあたって、BLOSUM50、BLOSUM62、BLOSUM80、PAM250、PHAT70、PHAT75、PHAT80、PHAT85、JTT250のすべてを使用した。また、ギャップパラメータについてギャップ開始ペナルティとギャップ拡張ペナルティの2つについて合計150の組み合わせについて試した。結果としてはBLOSUM62でギャップ開始ペナルティ20、ギャップ拡張ペナルティ1が最適な結果を出したので、今後はその結果のみを示す。PHATやJTT250はBLOSUM62よりよいスコアを出すことはできなかった。これはおそらく膜タンパク質は膜貫通領域だけではなく、ループ領域も存在し、ループ領域のアラインメントの精度を下げる原因となるからだと推測される。

表3.1はPScoreでのPHMMとALIGNとの比較を示す。PHMMはALIGNよりよい精度を出している。PScoreは0.014改善している。このスコアの改善が統計的有意であるかを調べてるためにウィルコクソンの符号付順位和検定 [122] を用いて検定した。ここでは有意確率（以下、P-valueと呼ぶ）が0.05以下をここでは有意とする。配列全体についてのPHMMとALIGNとの間のP-valueはP-valueは0.016となり、PScoreに関してPHMMとALIGNは有意に差がみられた。

表 3.1 評価セットについてのPScoreの比較

手法	
ALIGNのPScore	0.806
PHMMのPScore	0.820
ALIGNとPHMM間のP-value	0.016

表 3.2 は PHMM と ALIGN の PScore の観点から評価セット全体の各配列ペアについて改善している割合か、あるいはは悪化している割合を示している。結果をみると、PHMM は ALIGN に比べて半数弱が改善していることがみてとれる。一方、逆に PScore が悪化している例は 2 割弱存在する。

表 3.2 PScore の大小による配列ペア数の比較

	配列ペア数 (全体の割合)
PHMM の PScore > ALIGN の PScore	24 (48 %)
PHMM の PScore = ALIGN の PScore	17 (34 %)
PHMM の PScore < ALIGN の PScore	9 (18 %)

次に各構造領域ごとの PScore の改善について評価する。アラインメント内のカラムを膜貫通領域とループ領域に分ける。膜貫通領域とループ領域は PDB_TM [113] データベースからから決定される。表 3.3 は各領域について PHMM と ALIGN を比較を示す。膜貫通領域についても PHMM の PScore は ALIGN に比べて改善している。膜貫通領域に関して PScore は 0.039 改善している。P-value は 0.002 と統計的有意に精度が改善していることがわかる。一方、ループ領域では PScore は 0.008 改善している。その P-value は 0.058 と統計的有意差はみられなかった。この結果は本手法は主に膜貫通領域のアラインメントの精度を改善していることを示している。

表 3.3 各構造領域についての PScore の比較

手法	膜貫通領域	ループ領域
ALIGN の PScore	0.793	0.789
PHMM の PScore	0.833	0.797
ALIGN と PHMM 間の P-value	0.002	0.058

異なる配列類似性のレベルに PScore を用いて評価を行った（表 3.4)。評価セッ

トをさらにCEの配列類似性によって、0-30%（低類似性）と30-50%（高類似性）の2つに分けた。低い類似性の配列ペアについてはPScoreは0.017改善した。そのP-valueは0.036であり、その差は統計的有意に差がある。一方、高類似性の配列ペアでは、PScoreは0.005と改善しているが、そのP-valueは0.084となっており、統計的有意に改善しているとはいえなかった。よって、本手法は低類似性の配列ペアについて特に改善が見られることがわかった。

表 3.4 配列類似性のレベルによるPScoreの比較

手法	0-30%	30-50%
ALIGNのPScore	0.747	0.864
PHMMのPScore	0.764	0.875
ALIGNとPHMM間のP-value	0.036	0.084

3.2.2 改善したアラインメントの具体例

ここではアラインメントが大きく改善した具体例について考察する。今回とりあげる例は*E. coli*由来のsuccinate dehydrogenaseのサブユニットC (SdhC) (PDB: 1NEK chain C) [127]と*Wolinella succinogenes*由来のFumarate reductase respiratory complex II-like fumarate reductaseのサブユニットC (FrdC) (PDB: 1QLA chain C) [71]の結晶構造はどちらもSCOPデータベース [84]においてfumarate reductase respiratory complex TM subunitスーパーファミリーに属している。図3.3はSdhCとFrdCのアラインメントについてALIGNとPHMMとの比較を示している。ボックスはコアブロックに対して各手法のアラインメントが正しい箇所を示す。コアブロック内の膜貫通領域に相当する箇所はグレーで表現している。これらの配列のアラインメントはCEでの配列類似性が13.5%と非常に低いため正確にアラインメントすることは難しい。なお、このアラインメントにおけるPScoreはALIGNが0.180、PHMMが0.696である。SdhCはPDB_TMによると3つの膜貫通領域を、FrdCは5つの膜貫通領域をもっている。構造ア

ラインメントの結果は SdhC における 1 本目、2 本目、3 本目の膜貫通領域がそれぞれ FrdC の 3 本目、4 本目、5 本目の膜貫通領域にアラインされている。構造アラインメントからのコアブロックはこの 3 つのアラインされた膜貫通領域とその付近に割り当てられている。各膜貫通領域ごとにみていくと、最初の膜貫通領域付近のコアブロックは ALIGN はすべての残基を正確にアラインメントをとることに失敗している。すなわち最初の膜貫通領域が相手方の適切な膜貫通領域にアラインメントできていない。一方 PHMM では最初のコアブロックはすべて正しくアラインメントができている。次に 2 つ目のコアブロックについてみると、ALIGN は膜貫通領域を含むコアブロックの N 末端側の半分はうまくアラインメントを生成できていない。一方で PHMM は 2 本目の膜貫通領域を正確にアラインメントできている。3 つ目のコアブロックは双方の手法でアラインメントはうまくできていない。膜貫通領域とその付近を高精度でアラインメントを生成することは比較モデリングなどで立体構造を予測することなどにおいて非常に重要であり、この例では PHMM の方が ALIGN より優れていることがわかる。

図 3.3 SdhC と FrdC の配列アラインメント。アラインメントの上からコアブロック部分のアラインメント、ALIGN によるアラインメント、PHMM のアラインメントを表す。コアブロックで四角で囲われている色付き部分は膜貫通領域部分を示す。ALIGN 及び PHMM で四角で囲われている部分はコアブロックの結果と一致している部分である。この 2 つの配列一致度は CE の結果では 13.5%である。また、ALIGN の PScore は 0.180、PHMM を用いた手法では 0.696 である。

3.3. まとめと今後の課題

PHMM を用いた新しい膜タンパク質ペアワイズアラインメント法を提案した。膜貫通領域とアラインメントの両方をモデル化することで、膜貫通トポロジー予測と配列アラインメントを同時に行うことが可能となった。これまでのグローバルアラインメント法は配列全長に対してすべて同じパラメータ（アミノ酸置換マトリックス、ギャップペナルティ）を使用しているものがほとんどであった。本手法では各構造領域の配列はそれぞれの構造領域固有のアラインメントパラメータ（出力確率、遷移確率）を用いてアラインメントができるようになり、膜タンパク質のような構造領域間で大きくことなるアミノ酸置換頻度、ギャップ生成頻度のような場合でも対応できるようになった。提案した手法を評価したところ標準的なグローバルアラインメントに対して、精度を向上されることができた。特に膜貫通領域におけるアラインメントの精度の向上が顕著にみられた。また、SdhC と FrdC のアラインメントの例のように配列類似性が低い配列ペアに対してその差が大きいことがわかった。このような構造領域によってアラインメントパラメータを変更してアラインメントを行う方法は膜タンパク質に限らず、構造領域ごとにパラメータが大きく変化するようなケースについても応用可能であると考える。これらの改善が比較モデリング法を用いた立体構造予測などをはじめとする様々なアラインメントを用いた手法に対してその精度を向上することが期待できる。

第4章　GPCRのペプチド・タンパク質系リガンドの残基長予測方法

本章では配列アラインメント法の応用例として、ペプチド・タンパク質系リガンドによって活性化されると予想されるオーファン GPCR のリガンド残基長予測法について説明する。第 2 章 2.3.2 で述べた通り、オーファン GPCR の機能解析において主要な課題の 1 つは内在性リガンドの決定である。ここでは、今後ゲノム情報から同定されていくと予想されるペプチドやタンパク質リガンドの候補に対してリガンド候補を絞り込むための新しいアプローチを提案する。今回はリガンド候補の絞り込むための特性としてリガンドの残基長に焦点あてる。

本章の構成としてはまずはじめに、GPCR のアラインメント法について述べる。次に残基長予測のための特徴量について述べ、最後に結果と考察について述べる。

4.1. GPCRのアラインメント法

ここでは、次節で述べる残基長予測法に使われる GPCR のアラインメント法について述べる。なお、本手法は次章においても同じ手法でアラインメントを作成している。

4.1.1 手法概要

手法のフローを図 4.1 に示す。全体としては、マルチプルアラインメントプログラム CLUSTAL W [110] を用いて、プロファイルの基となるアラインメント（シードアラインメント）を構築し、そのシードアラインメントを基にプロファイル隠れマルコフモデルのプロファイルを構築する。ここまではプロファイルを構築する一般的流れであるが、さらにそのプロファイルに対してロドプシンの立体構造を基に修正を加える。まず始めに修正前のプロファイルを作る過程を簡単に説明し、その後構造によるプロファイルの修正について述べる。

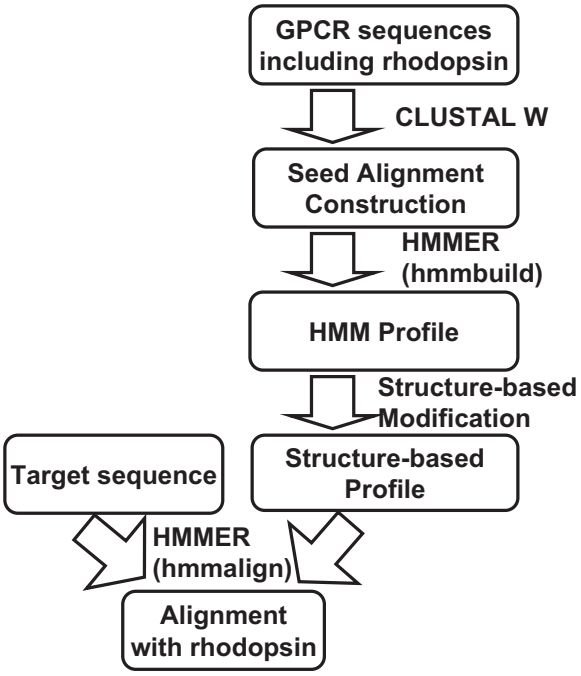

図 4.1 Class A GPCR のアラインメントのためのプロファイルの構築フロー

4.1.2 プロファイル隠れマルコフモデルの構築

プロファイルを構築するために、GPCRDB [50][48] から Class A GPCR の全エントリ 3482 配列を取り出した。このエントリは全生物種のデータを含んでい

る。ただし、配列内に Asx、Glx、Xaa といった特定のアミノ酸を表さない残基が含まれている場合はその配列を取り除いた。さらに全体の配列の冗長性を減らすため、最短距離法のクラスタリングアルゴリズムを利用して各配列ペアの配列類似性が 30%となるように代表配列を決めた（ただし、bovine のロドプシンは立体構造情報を利用するため意図的に代表配列として選択した）。配列冗長性を減らす理由は、CLUSTAL W において経験的に配列の偏りがある場合にはうまくアラインメントを構築できないことが多いためである。次に HMMER [30] プログラムを用いてプロファイル隠れマルコフモデルのプロファイルを構築した。構築する上で、各 Match 状態にロドプシンの各残基がアラインされているカラムを指定した（指定は HMMER のオプションにて設定した）。

4.1.3 立体構造を用いたプロファイルの修正

ここでは 4.1.2 で構築したプロファイルに構造情報を基に修正を加える方法について述べる。ロドプシンとのアラインメントでは置換、挿入、欠損の各操作が考えられるが、このうち置換に関しては、プロファイル隠れマルコフモデルの各 Match 状態の出現確率が置換の役割をしており、特に構造領域ごとに変更を加える必要がない。構造領域を考えた上で修正が必要なのは挿入、欠損すなわちギャップをどう考慮するかである。膜貫通領域はαヘリックス型膜タンパク質に関してはおおむねヘリックス構造という規則正しい構造をとっているため、ギャップが入る余地がない。しかしながらごく稀に膜貫通領域においてもギャップが入る場合があるため、アラインメントを構築する上ではそのような場合も想定する必要がある。そこでまず始めに立体構造の修正について述べる前に膜タンパク質の膜貫通領域についてのギャップについて考察する。

2.6 Å の解像度の bovine のロドプシン (PDB code 1L9H chain A) の立体構造 [92][90] に対して、二次構造決定プログラム DSSP [59] を用いて、ロドプシンの各残基は立体構造のヘリックス構造であるかそれ以外であるかを決めた。プロファイルの修正の方針は以下のようである。

- 膜貫通領域でヘリックス構造の場合には、欠損及び手前での挿入によるギャップをなくす。

- 膜貫通領域でヘリックス以外の構造の場合には、手前に一残基のみ挿入を許す。

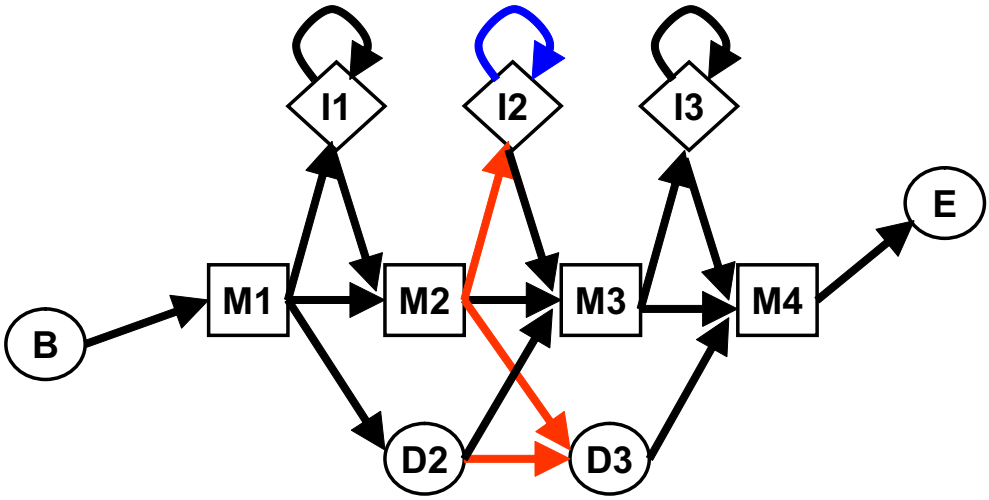

図 4.2 構造によるプロファイルの修正方法に関する参考図。B は開始状態、E は終端状態、正方形は一致状態、菱形は挿入状態、円形は欠損状態を表す。一致状態が膜貫通領域の場合は、その構造にしたがって、青線もしくは赤線の遷移確率値を変更する。

プロファイルの修正方法を図 4.2 を使って説明する。(プロファイル隠れマルコフモデル自体の説明は付録 A.2 参照) 図 4.2 における状態 M3 に対応するロドプシンの残基が DSSP で'H'（ヘリックス）であり、かつ膜貫通領域である場合には図 4.2 での赤線の遷移をなくす。これによって、状態 M3 に対応するロドプシンの残基が欠損する場合及び状態 M3 の手前でギャップが挿入されることはなくなる。次に状態 M3 におけるロドプシンの残基が DSSP で'H'（ヘリックス）以外である場合、すなわち、プロリンやその他の残基の影響によってヘリックス構造が壊されている場合、青線の遷移をなくす。これによって、状態 M3 の手前には挿入によるギャップは最大 1 だけ入れられる。このような方法でプロファイル隠れマルコフモデルのプロファイルを修正したものを用いて、ロドプシンとターゲット配列とのアラインメントを構築する。

4.2. 残基長予測のデータセットと方法

4.2.1 データセット

ヒトの GPCR は Class A, B, C に属するが、その中でペプチドもしくはタンパク質をリガンドとする既知の GPCR は Class A の一部と Class B のすべてである。Class B の GPCR はデータ数が非常に限られており、評価を行うだけのデータ数が確保できなかったため、今回は Class A GPCR に焦点を絞った。学習及びテストに使用した GPCR は GPCR の分類データベース GPCRDB [50][48] において Class A に属していて、かつ文献情報 [2] よりペプチドもしくはタンパク質と結合するものを抽出した。抽出した GPCR 配列について BLASTCLUST [5] を用いて各配列ペアの配列類似性が 35%以下になるようにデータセットの配列冗長性を減らした。1 つの GPCR がいくつかの内在性リガンドと相互作用する場合には、高いアフィニティを持つリガンドの残基長の平均を用いた。リガンドが複合体を作る場合には、すべてのサブユニットの残基長の和をリガンドの残基長とした。

4.2.2 回帰分析と特徴量

特徴量　回帰分析を用いてリガンドの長さを予測するために、GPCR の配列を表す特徴量を定義した。本手法ではロドプシンとターゲット配列とのアラインメント（4.1 参照）からの位置特異的特長量とドメインベース特徴量を使った。

アミノ酸の物理化学的特性はタンパク質の構造や機能に重要な役割をしており、そのような効果はリガンドの相互作用に直接もしくは間接的に影響を与えている。アミノ酸の物理化学的特性としてはアミノ酸指標データベース AAindex [86][112][62] のデータを用いた。位置特異的特徴量ではターゲット GPCR 配列の各（ロドプシン上の）位置のアミノ酸に対応するアミノ酸指標を特徴量として使う。すなわち、疎水性のアミノ酸指標があり位置 i のアミノ酸が a だとすると、アミノ酸 a の疎水性アミノ酸指標の値を特徴量として使う。ただし、Asx、Glx、Xaa など特定のアミノ酸を示さないものやギャップ領域ではアミノ酸指標の平均値を

代用して使う。また、各アミノ酸指標は以下のような変換を行って標準化する。

$$v' = \frac{v - v^{min}}{v^{max} - v^{min}} \tag{4.1}$$

v はアミノ酸指標、v^{max}、v^{min} はそれぞれ対象としているアミノ酸指標の最大値と最小値を表す。このようにして、各アミノ酸指標を 0 から 1 までの範囲のスケールに変換する。

次にドメインベース特徴量としては、GPCR の 15 ドメイン（7 つの膜貫通ドメイン、N 末端ループを含む 4 つの細胞外ループ、C 末端ループを含む 4 つの細胞外ループ）について、それぞれのドメイン内の配列のアミノ酸指標の平均値を求め、その値を使用した。ドメインの領域の定義はアラインさらたロドプシンのドメインを利用して決めた。アミノ酸指標としては位置特異的特長量についてもドメインベース特徴量についても AAindex 内の全 494 エントリを使用した。位置特異的特長量と同様にアミノ酸指標は式 4.1 でスケールを変換した。このように求めた位置特異的特長量とドメインベース特徴量の双方を要素とした特徴量ベクトルを作成した。

特徴量選択　上述の方法で特徴量ベクトルを作成したが、その次元 N （= 494 アミノ酸指標 * 348（位置） + 494 * 15（ドメイン））を得たが、その次元は非常に膨大である。これらの特徴量の中にはリガンドの相互作用とは関係ない位置やドメインが存在すると予想される。また、似たアミノ酸指標も多数 AAindex には存在しており、そのようなアミノ酸指標に対しては代表的な指標 1 つで十分である。このような特徴量を削り、次元を落とすことを考える。

まずはじめに、リガンドの相互作用と関係ありそうな特徴量を選択するために、各特徴量とリガンドの長さとの間の類似度 $T_{i,p}$ を求める。類似度を求める類似度関数については後述する。位置特異的特長量及びドメインベース特徴量について $T_{i,p}$ が $T^{Threshold}$ 以下である場合、その特徴量は使用しない。

次に AAindex 内の多くの似たようなアミノ酸指標があることによる、アミノ酸指標の冗長性を減らすことについて述べる。AAindex 内のアミノ酸指標は大きく分けると 6 個のグループに分かれることが階層的クラスタリングを用いた解析により知られている [112]。そこで、同様の手法を用いてアミノ酸指標を 6 個の

グループに分類する。アミノ酸指標間の距離 d を以下のように定義する。

$$d = 1 - |r| \tag{4.2}$$

r は Peason の相関係数で以下のように定義される

$$r = \frac{\sum_i (a_i^x - \bar{a^x})(a_i^y - \bar{a^y})}{\sqrt{\sum_i (a_i^x - \bar{a^x})^2 \sum_i (a_i^y - \bar{a}^y)^2}} \tag{4.3}$$

a_i^x、a_i^y はそれぞれアミノ酸 i のアミノ酸指標 x、y の値を表し、$\bar{a^x}$、$\bar{a^y}$ はそれぞれアミノ酸指標 x、y の平均値を表す。この距離を用いて Ward 法 [119] を利用して AAindex 内の 494 アミノ酸指標を 6 つのグループに分類した。そして分類されたグループについて 1 つのグループから 1 つの特徴量を選択した。

$$Feature(G, p) = \max_{i \in G}(T_{i,p}) \tag{4.4}$$

G はアミノ酸指標グループ内のアミノ酸指標の集合、p は位置もしくはドメインを表す。これらの手法によって選択された特徴量を用いて回帰分析を行って予測する。

類似度関数　本研究では以下の 5 つの類似度関数を特徴量選択のために使用し、比較検討を行った。一般的な Peason の相関係数に加え、順位相関係数である Spearmann 順位相関係数、Kendall 順位相関係数も検討した。また、そのほかにも代表的な類似導関数として、Tanimoto 係数及び Cosine 係数についても検討した。

1 つ目は Peason の相関係数で以下の式から求める。

$$Peason \text{ 相関係数} = \frac{\sum_i (a_{i,p} - \bar{a}_p)(l_i - \bar{l})}{\sqrt{\sum_i (a_{i,p} - \bar{a}_p)^2 \sum_i (l_i - \bar{l})^2}} \tag{4.5}$$

$a_{i,p}$ は GPCR 配列 i 内での位置 p でのアミノ酸指標の値、l_i は GPCR 配列 i のリガンドの長さ、$\bar{a}_p$、$\bar{l}$ はそれぞれ $a_{i,p}$、l_i の平均を表す。

$a_{i,p}$、l_i をそれぞれ小さい順に順位をつける。どう順位がある場合には平均順位をつける。両者の順位の差を d_i、$a_{i,p}$、l_i に関して、同順位の個数をそれぞれ n_x、n_y、同順位の大きさを t_i、t_j $(i = 1, 2, \ldots, n_x; j = 1, 2, \ldots, n_j)$ とする。すると、

Speamann 順位相関係数は以下のように求める。

$$
\begin{aligned}
Speamann\,\text{順位相関係数} &= \frac{T_x + T_y - \sum_i d_i^2}{2\sqrt{T_x T_y}} \\
T_x &= \frac{(n^3 - n) - \sum_i^{n_x} (t_i^3 - t_i)}{12} \\
T_y &= \frac{(n^3 - n) - \sum_j^{n_j} (t_j^3 - t_j)}{12}
\end{aligned}
\tag{4.6}
$$

$a_{i,p}$、l_i について小さいほうから順位をつけ、$a_{i,p}$ について小さい順に並べ変える。同順位の場合には平均順位をつける。l_i について、$l_i < l_j$ の個数を P_i、$l_i > l_j$ の個数を Q_i とする。すると、Kendall の順位相関係数は以下のように求める。

$$
\begin{aligned}
Kendall\,\text{順位相関係数} &= \frac{\sum_i P_i - \sum_i Q_i}{\sqrt{\frac{n(n-1)}{2} - T_x}\sqrt{\frac{n(n-1)}{2} - T_y}} \\
T_x &= \sum_i^{n_x} \frac{t_i(t_i - 1)}{2} \\
T_y &= \sum_j^{n_y} \frac{t_j(t_j - 1)}{2}
\end{aligned}
\tag{4.7}
$$

残りの 2 つは Tanimoto 係数と Cosine 係数で以下の式から求める。

$$
Tanimoto\,\text{係数} = \frac{\sum_i (a_{i,p} l_i)}{\sum_i a_{i,p}^2 + \sum_i l_i^2 - \sum_i (a_{i,p} l_i)} \tag{4.8}
$$

$$
Cosine\,\text{係数} = \frac{\sum_i (a_{i,p} l_i)}{\sqrt{\sum_i a_{i,p}^2 \sum_i l_i^2}} \tag{4.9}
$$

回帰分析　回帰分析手法としては、重回帰分析といった線形モデルの最小自乗法を用いた手法が知られている。本研究では前述の通り、高次元の特徴量として用いることに対処するため、特徴量選択法を述べた。しかしながら、前述した特徴量選択法は特徴量としてあまり有効でない可能性が高い特徴量を削減することを目的としており、少数の必要最小限の特徴量まで選択することを目的としていない。一般に、学習データの数に対して特徴量の次元が大きい場合には、必要とするパラメータが増大するため、高い精度で予測できないことが知られている。これらの問題に対処するため、本研究では回帰分析の手法として、Support vector regression (SVR) を用いた（SVR は付録 B を参照）。SVR はカーネルトリックと

組み合わせることで、マージン最大化という基準から少数の学習データに対応するカーネルのみが選択され、最適な非線形回帰モデルが求められる。また、特徴量次元が大きくなるが原因で汎化能力が低下するという問題を回避できるような高い汎化能力を持つ。SVR には LIBSVM [18] を用いて実装した。

4.2.3 手法の評価方法

予測の精度は Leave-one-out クロスバリデーション法 [70] を用いて評価した。あるテストデータを 1 つ選び、残りの評価セットで特徴量選択及び学習を行い、学習を基に作成したモデルでテストデータのリガンドの残基長を予測する。予測精度は正解のリガンドの残基長と予測したリガンドの残基長間の誤差の絶対位置の平均 (Mean of absolute Error, Mean AE)、誤差の絶対位置の中央値 (Median of absolute Error, Median AE)、誤差の自乗平均のルート (Root mean square error, RMSE)、Peason の相関係数 (Peason's Correlation Coefficient, PCC) を用いて測定した。Mean AE、Median AE、RMSE は小さいほど予測精度がよく、一方 PCC は大きいほど予測精度がよい。

$$MeanAE = \frac{1}{N}\sum_i \left| l_i - l_i^{pred} \right| \tag{4.10}$$

$$MedianAE = \begin{cases} \left| l_m - l_m^{pred} \right| & n \text{が奇数}, m = (n+1)/2 \\ \frac{\left(\left| l_m - l_m^{pred} \right| + \left| l_{m+1} - l_{m+1}^{pred} \right|\right)}{2} & n \text{が偶数}, m = n/2 \end{cases} \tag{4.11}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_i (l_i - l_i^{pred})^2} \tag{4.12}$$

$$PCC = \frac{\sum_i (l_i - \bar{l})(l_i^{pred} - \bar{l}^{pred})}{\sqrt{\sum_i (l_i - \bar{l})^2 \sum_i (l_i^{pred} - \bar{l}^{pred})^2}} \tag{4.13}$$

l_i は GPCR 配列 i の正解リガンド残基長、l_i^{pred} は予測リガンド残基長、$\bar{l}$、$\bar{l}^{pred}$ はそれぞれ正解リガンド残基長と予測リガンド残基長の平均を表す。

4.3. 結果

4.3.1 類似度関数の決定と考察

類似度関数の比較　5つの類似度関数を提示したが、その中でどの類似度関数がよいかについて評価を行った。評価方法としては、各類似度関数について類似度の上位200を特徴量として選択し、それを基に、予測を行った。その結果を表4.1に示す。結果が示す通り、すべての評価指標においてTanimoto係数が一番よい結果をえている。この結果より以後は類似度関数としてはTanimoto係数を用いて予測を行う。

表 4.1　類似度関数の評価結果

類似度関数	Mean AE	Median AE	RMSE	PCC
Peason 相関係数	15.18	12.90	19.66	0.916
Speamann 順位相関係数	19.46	16.97	24.00	0.883
Kendall 順位相関係数	18.67	15.87	23.02	0.891
Tanimoto 係数	14.01	10.14	19.10	0.931
cosine 係数	15.93	15.05	20.64	0.910

類似度スコアについての考察　上述の結果より特徴量選択にはTanimoto係数を使用することにした。ここでは、Tanimoto係数の類似度がリガンドとの関連性がある特徴量を選択できているかについて考える。この妥当性を評価することは難しい。なぜならば、リガンドの残基長と関連がある残基という情報は生物実験などで調べることは難しく、データが存在しないためである。そこでここでは、Tanimoto係数の類似度で上位となっている位置がロドプシン上のどのような部位にみられるかを参考に簡易的な評価をする。

概要　表4.2にTanimoto係数による類似度スコアの上位25位までの位置とアミノ酸指標のペアを示す (位置としては複数のアミノ酸指標があるため16箇所)。mutationは該当する位置においてペプチド・タンパク質をリガンドとするGPCRの該当する残基を変異実験によって残基置換をさせた時、リガンド結合に大きな影響を与えるものという結果を示した文献への参照である。上位25位までの位置に相当する多くの残基は膜貫通領域にあることがわかる。その他では細胞外ループが2例あり、細胞内ループもわずか2例である。次に、それらの位置が実際にロドプシン上のどこにあるかについて、図4.3に示す。まず、細胞膜と平行にみた図では多くが膜貫通領域の中心より細胞外側にあることがわかる。細胞内側には先ほどあげた細胞内ループの2例が存在するだけである。次に細胞外側から見た図について、多くの位置がロドプシンの内部側にあることがみれる。これらの残基は各膜貫通ヘリックスの間もしくはリガンド結合ポケット付近に配置されており、膜貫通ヘリックス間の相互作用やリガンドとの相互作用に影響を与えていると考えられる。

膜貫通領域　膜貫通領域に関して、7本の膜貫通領域に高い類似度スコアを持つ位置が存在している。特に集中している部位は膜貫通領域3、膜貫通領域5、膜貫通領域6の3つの膜貫通領域がヘリックス間で相互作用している領域である。

GPCR内でのリガンド結合のための空洞に関わる残基が知られている。SurgandらはGPCRのリガンド結合のための空洞に関わる残基として30個の位置を提案している [109]。提案された30個の位置の他にもリガンド結合に重要な位置があるかなどは議論の余地があるが、これらの位置とその周辺がリガンド結合に関わっ

表 4.2 類似度スコアの上位 25 位までの位置とアミノ酸指標

$T_{i,p}$	位置	ドメイン	AAindex ID	mutation
0.834886	105	細胞外ループ 2	ZIMJ680103	[41]
0.807895	160	膜貫通領域 4	AURR980117	
0.7811	265	膜貫通領域 6	QIAN880109	[126]
0.771444	121	膜貫通領域 3	CHOP780206	[35]
0.766402	146	細胞内ループ 2	FINA910101	
0.757274	105	細胞外ループ 2	FAUJ880111	[41]
0.75326	215	膜貫通領域 5	FUKS010107	
0.751514	171	膜貫通領域 4	ONEK900102	[40]
0.747854	212	膜貫通領域 5	MAXF760105	[40]
0.743588	212	膜貫通領域 5	KOEP990102	[40]
0.742799	212	膜貫通領域 5	BUNA790101	[40]
0.74251	261	膜貫通領域 6	HOPA770101	[54]
0.74017	296	膜貫通領域 7	MUNV940104	[69]
0.734724	261	膜貫通領域 6	FAUJ880108	[54]
0.73456	265	膜貫通領域 6	PONP800106	[126]
0.727254	87	膜貫通領域 2	MAXF760104	
0.713795	122	膜貫通領域 3	CHOP780204	[94]
0.713689	31	N 末端ループ	RACS820104	[42]
0.711485	41	膜貫通領域 1	EISD860102	
0.7076	204	膜貫通領域 5	MITS020101	[96]
0.702689	160	膜貫通領域 4	CHOP780215	
0.702542	261	膜貫通領域 6	KOEP990102	[54]
0.701812	160	膜貫通領域 4	FUKS010101	
0.701345	122	膜貫通領域 3	HOPA770101	[94]
0.699632	136	細胞内ループ 2	PONP800107	[9]

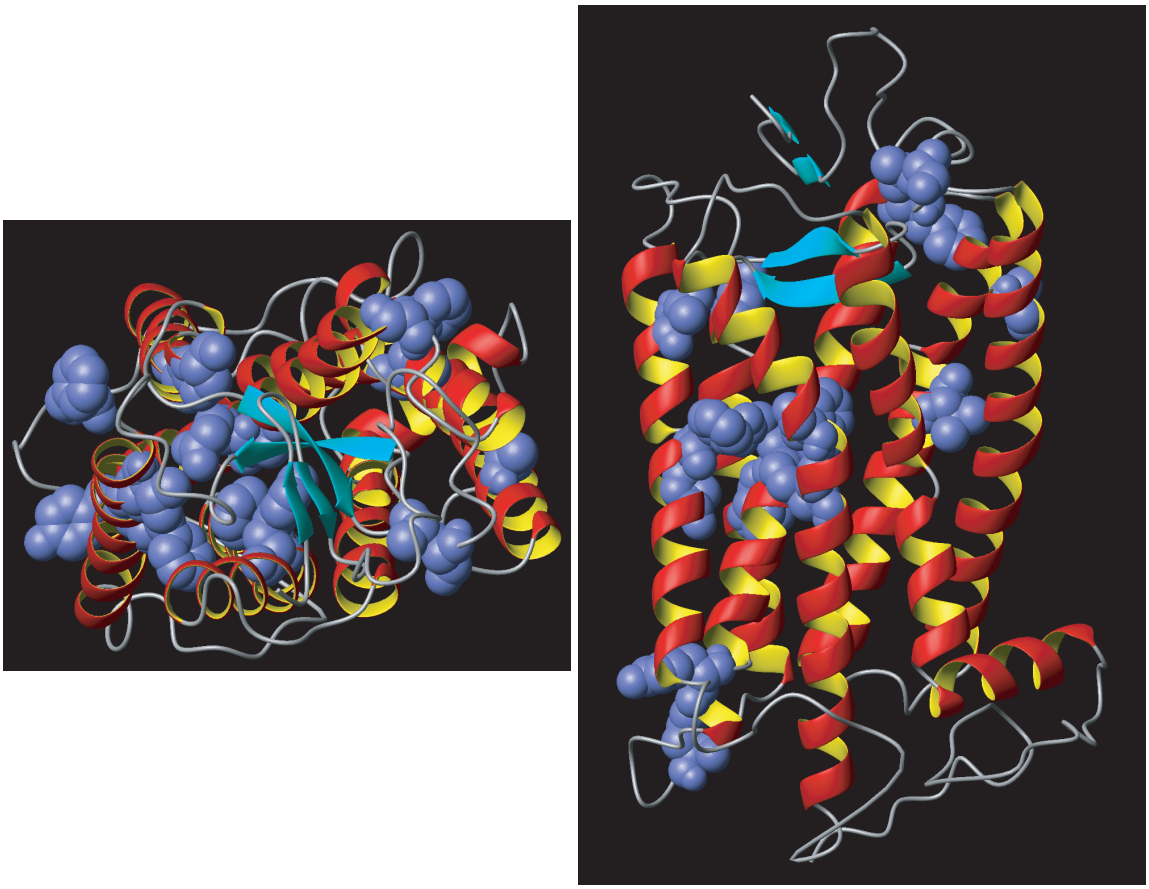

図 4.3 Tanimoto 係数による類似スコア上位 25 位までの位置のロドプシン上の残基。左側が細胞外側から見た図、右側が細胞膜に平行な視点で見た図で上側が細胞外側、下側が細胞内側である。側鎖を青色の CPK で表示している部分がスコア上位 25 位までの位置に対応するロドプシンの残基である。MOLMOL [64] にて描画。

ているということは様々な変異実験の結果からわかっている [54][126][69]。空洞にかかわる残基を図 4.4 に示す。リガンド結合に関わる空洞は膜貫通領域 1、膜貫通領域 2、膜貫通領域 3 の一部、膜貫通領域 7 の間の空洞であるサブグループ 1 と膜貫通領域 3 の一部、膜貫通領域 4、膜貫通領域 5、膜貫通領域 6 の間の空洞であるサブグループ 2 の 2 つに分かれる [109]。類似度スコアが高い位置でかつ空洞に関わる位置は 5 箇所あるが、そのうち 4 箇所がサブグループ 2 に存在する。また、図 4.3 を見ればわかるように、類似度の高い位置はサブグループ 2 の空洞に関わる位置の周辺に多く存在する。これらの位置に相当する残基がいくつかの GPCR のリガンドとの相互作用に影響を与えていることは、表 4.2 で示した。これらの結果を考慮すると、サブグループ 2 の空洞に接するもしくは、その周辺の残基を中心にサブグループ 1 やループ領域の一部の残基によって大きなリガンドと小さなリガンドとの違いの認識に影響を与えている可能性を示唆している。

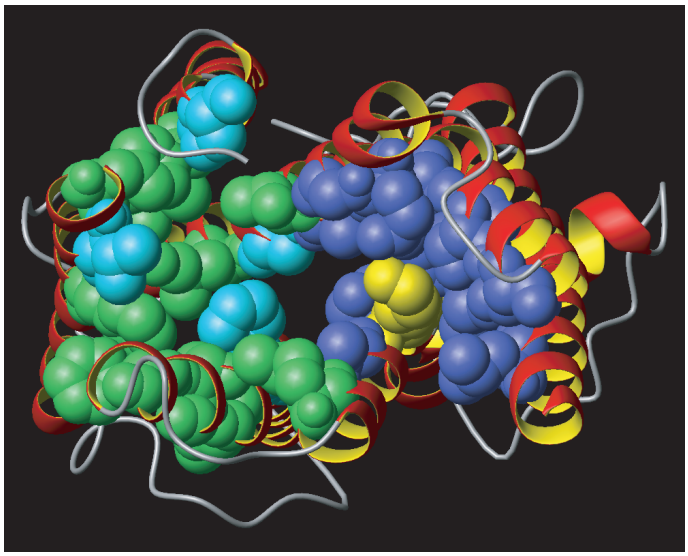

図 4.4 Surgand らのリガンド結合のための空洞の残基 [109]。提案された 30 個の位置に相当するロドプシンの残基を CPK で表示した。これらの残基は 2 つのサブグループに分かれる。サブグループ 1(膜貫通領域 1、膜貫通領域 2、膜貫通領域 3 の一部、膜貫通領域 7) に相当する残基を青色、サブグループ 2(膜貫通領域 3 の一部、膜貫通領域 4、膜貫通領域 5、膜貫通領域 6) に相当する残基を緑色、サブグループ 1 の中で類似度スコアの上位 25 位までのものを黄色、サブグループ 2 の中で類似度スコアの上位 25 位までのものを水色で表示した。MOLMOL [64] にて描画。

細胞外ループ　細胞外ループはわずか 2 例しかない。細胞外ループがいくつかの GPCR においてリガンドとの相互作用にかかわっていることは知られている。その点を考えると、細胞外ループに高い類似度スコアの位置が少ないことは、ループ領域は保存性が低いため仮にリガンドとの相互作用に重要な残基でもファミリーの垣根を越えて保存するような残基は少ないということが理由ではないかと考えられる。

類似度スコアの上位 25 位までに入っている細胞外ループにある 2 つの位置は N 末端と細胞内ループ 1 に位置している。これらに相当する位置の残基はいくつかの GPCR ではリガンド結合の残基と変異実験などから推定されている [41][42]。なんらかのリガンドのとの相互作用と関連がありそうではあるが、今回の解析ではその残基周辺に他の類似度スコアの高い位置が発見できていないため、その位置の役割とリガンドの残基長との関係について考察することは難しい。

細胞内ループ　細胞内ループには 2 例の高い類似度スコアの位置が発見された。2 例とも細胞内ループ 2 であった。細胞内ループがリガンドとの相互作用に直接的に関わる可能性は低いので、これらの位置はノイズ（リガンド相互作用とは関係ない）可能性も存在する。しかしながら、位置 136 は DRY モチーフと呼ばれるものの Tyr に相当する部分で、保存性は約 70%程度である。いくつかのリガンドサイズが大きい GPCR にはこの保存残基が別の残基に置換していることを考えると、何らかの構造的役割を担っている可能性も存在する。

4.3.2 予測結果

3 つのカーネルと多くのパラメータを試し、リガンドの予測結果について評価指標が最大とするものを表 4.3 に示す。カーネルはどちらも RBF カーネルである。上のパラメータでは Mean AE、Median AE、RMSE が最小であった。一方、下のパラメータでは PCC が最大であった。上のパラメータの PCC の値は下のパラメータに比べそれほどの差がない一方で、下のパラメータは上のパラメータに比べ、Mean AE、Median AE、RMSE ともに大きく劣っている。そこで、ここでは上のパラメータを採用した。

表 4.3 予測結果

C	γ	Mean AE	Median AE	RMSE	PCC
7.8×10^6	5.86×10^{-11}	11.78	8.31	16.31	0.942
1.0×10^7	6.24×10^{-11}	17.33	13.56	22.37	0.943

次にリガンドの残基長を 3 つのグループに分けて、各残基長グループでの予測精度について考える。グループとしては”small petide” （20 残基以下)、”middle peptide” （20-50 残基)、”Large peptide and Proteins” （50 残基以上）に分けた。残基長グループごとの予測結果を表 4.4 に示す。”middle peptides”の予測精度がどの評価指標でも一番よかった。それに次いで、”small peptides”の評価指標がよかった。”Large peptide and Proteins”に関しては予測精度があまりよくない。

表 4.4 リガンドサイズごとの予測精度

	データセット数	Mean AE	Median AE	RMSE
Small Peptides	25	9.85	7.31	12.85
Middle Peptides	19	6.60	5.60	7.49
Large Peptides or Proteins	12	25.12	29.55	27.98

さらに詳細にリガンドの残基長と予測精度の関係について見るために、予測結果と正解データに関する散布図を図 4.5 に示す。リガンドの残基長が 50 残基以下の時と 150 残基以上の時はこれらの結果はおおよそ正解の値付近にある。しかしながら、60-100 残基のリガンドでは予測リガンド残基長は過小評価されており、その結果が全体的な予測精度を大きく押し下げている。

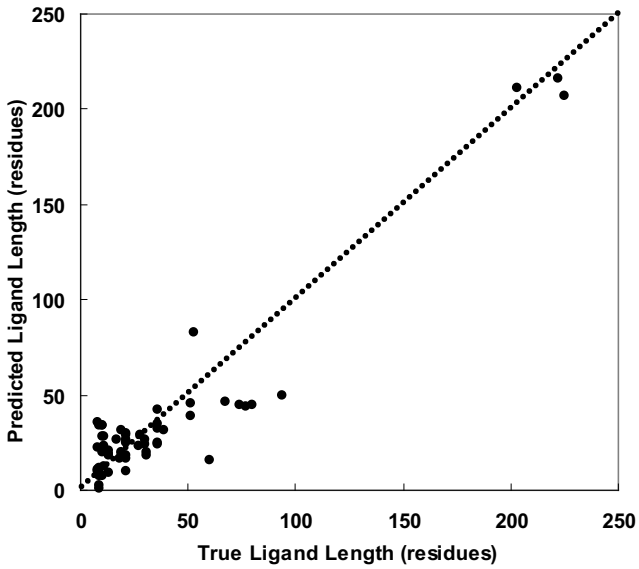

図 4.5 予測される残基長と正解である残基長の比較

4.3.3 オーファン GPCR に関する予測される残基長の分布

オーファン GPCR のリガンドの残基長の分布を推定するために、112 のペプチド・タンパク質によって活性化されると予測される GPCR に対して、本手法を適用した。その結果を図 4.6 に示す。結果をみると 30-50 残基ぐらいに多くのオーファン GPCR が存在することがわかる。ここ数年に同定された新規リガンド（例えば、Neuropeptide W、Prolactin-releasing peptide、neuropeptide QRFP など）は 20-50 残基のものが多く [20]、この結果と併せて考えるとオーファン GPCR の多くのペプチドリガンドは 20-50 残基付近の大きさのものが多く存在することが示唆される。

60-90 残基付近には今回のデータからでは観察されない。60-90 残基の既知リガンドは走化性リガンド（例えばケモカインや C5a など）が存在している。クロスバリデーションテストの結果をみると、60-90 残基のリガンドの長さは過小評価しているものが見られるので、もしかしたら、50 残基近辺でみられる GPCR のデータの中に 60-90 残基ぐらいの GPCR も含まれているのかもしれない。

200 残基以上のリガンドは観察されなかった。ホルモンタンパク質（例えば、follicle stimulating hormone、lutropin-choriogonadotropic hormone、thyrotropin）は 200 残基以上のリガンドと知られている。今回のオーファン GPCR のデータセットの中にはホルモンタンパク質のリガンドが含まれていないことが今回の結果から示唆される。

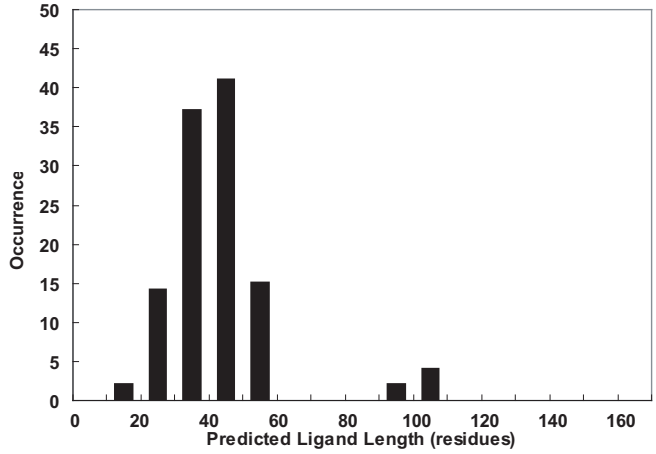

図 4.6 オーファン GPCR に関する予測される残基長の分布

4.4. 議論とまとめ

本研究では位置特異的特徴量とドメインベース特徴量を配列アラインメントから求め、それを基にペプチド・タンパク質リガンドの残基長を予測可能であることを示した。また、ペプチド・タンパク質リガンドによって活性化すると予測されるオーファン GPCR のリガンドの長さの分布を示した。

本手法を利用することで予測リガンド残基長を使ってペプチドリガンドスクリーニングライブラリの中の内在性リガンド候補の優先順位をリガンドの残基長によって優先順位を付けることが可能である。もちろん、リガンドの残基長の情報だけからは少数のリガンド候補に絞ることは難しいが、他の付加的なリガンド

の特性（たとえば、電荷を持つアミノ酸の出現頻度や疎水性など）も同様の手法で予測し、それらを組み合わせれば、リガンド候補をさらに絞り、ハイスループットリガンドスクリーニング技術に対するペプチドライブラリ内のリガンド候補を効率的に選ぶことができ、新規リガンド発見に役に立つ。この新しいアプローチは計算機によるオーファン GPCR の新規リガンド予測の可能性を示し、オーファン GPCR のリガンド発見研究のスピードを加速することが期待できる。

第5章　GPCRのGタンパク質の共役選択性予測

ここではアラインメント法の応用方法として、GPCR配列からのGタンパク質の共役選択性予測について述べる。第2章2.3.2で述べた通り、オーファンGPCRのGタンパク質共役選択性を事前に予測することはオーファンGPCRのリガンド同定においてのアッセイ系の構築に重要である。しかしながら、Gタンパク質の共役選択性の問題はその原理はあまり理解できていない。また、同じ共役選択性を持つGPCRでも必ずしも配列類似性は高くはない。そのような状況であるので、第4章のような自動的な特徴選択による手法ではなく、様々な特徴量候補を検討して、それを予測の特徴量として使用するというアプローチを行う。まずはじめに、GPCRの配列のどの部分がGタンパク質の共役選択性に関係しているか位置に基づく解析を行った。次に、位置以外の特徴量で重要なものを探索し、最後にそれらを特徴量としてSVMを用いた予測方法を提案する。

5.1. 特定の位置のアミノ酸出現と共役選択性の関係ついての解析

まず始めにGPCRの配列上のどの部位（位置）がGタンパク質共役選択性と関係していそうか、アミノ酸の出現頻度の観点から解析を行う。この解析を通じて、アミノ酸の出現と共役選択性との関係を考察するとともに、どのような配列情報を特徴量に入れるべきかについての指針を得る。

5.1.1 データセットと方法

データセット　GPCR の分類データベース GPCRDB [50][48] より Class A の GPCR を抽出し、その中から文献データ [1] より共役選択性が既知のものを選択した。トータルのデータ数は 111（$G_{i/o}$ 共役 GPCR: 55 $G_{q/11}$ 共役 GPCR: 34 G_s 共役 GPCR: 22）である。配列情報は Swiss-Prot [11] と一部 Entrez Protein [75] から抽出した。近年 G タンパク質共役選択性の多様性（複数のタイプの G タンパク質と共役する) が知られるようになったが、その原因は受容体のリン酸化による切り替えや一部には GPCR の過剰発現などの人為的な影響なども指摘されており [45]、そのメカニズムはまだまだ理解できない部分が多い。今回は簡単のため、複数の G タンパク質と共役する GPCR はすべて除いて解析を行った。

方法　テンプレートであるロドプシンとターゲット配列とのアラインメントは 4.1 と同じ方法で生成した。位置に関する定義も 4.1 と同じである。各位置について 10%以上ギャップとなっているところに関しては解析対象として除外した。解析対象の位置を”有効な位置”と呼ぶ。

機能残基を推定するためのスコアとしてグループ間のアミノ酸出現確率についての Peason の相関係数 (Peason's Correlation Coefficient, PCC) を使用した。アミノ酸出現確率の比較についてのスコア関数はいくつか提案されているが、PCC はプロファイル-プロファイルアラインメントの精度の評価でよい成績を収めており [76]、そのアミノ酸出現確率の比較のためのスコア付けとして信頼できる。PCC は以下のように表せる。

$$PCC(c) \quad = \quad \frac{\sum_i (p^c_{gi} - \hat{p}^c_g)(p^c_{\bar{g}i} - \hat{p}^c_{\bar{g}})}{\sqrt{\sum_i (p^c_{gi} - \hat{p}^c_g)^2}\sqrt{\sum_i (p^c_{\bar{g}i} - \hat{p}^c_{\bar{g}})^2}} \tag{5.1}$$

p^c_{gi} は G タンパク質のタイプ g ($G_{i/o}$, $G_{q/11}$, G_s) のデータセットについてロドプシン上での位置 c ののでのアミノ酸グループ i が出現する確率を表し、$p^c_{\bar{g}i}$ は G タンパク質のタイプ g と共役しないデータセットについての確率を表す。アミノ酸は 20 種類もあり、データ数が少ない場合にはアミノ酸出現確率 p^c_{gi} を正確には推定できない。そこで事前知識を加えた Dirichlet 混合分布による推定法 [103] を使用してアミノ酸出現頻度を推定した。また、配列の類似性による偏りを減らすため

に、配列類似度による重み付き出現頻度推定法 [44] もあわせて用いた。さらにアミノ酸をそのまま使うのではなく、アミノ酸をその特性ごとにグルーピングを行いその各グループに含まれているアミノ酸の出現確率の合計を求め、アミノ酸の代わりにアミノ酸グループを基にして PCC を求める。グループとしては Mirny [80] が提案したものを修正して使う。7 つのグループがあり、それぞれ脂肪族アミノ酸 AVLIMC、芳香族アミノ酸 FWYH、電荷のない極性アミノ酸 STNQ、正電荷を持つアミノ酸 KR、負電荷を持つアミノ酸 DE、プロリン P、グリシン G である。Mirny のグルーピングでは特殊 PG としていたが、膜たんぱく質の膜貫通領域ではグリシンとプロリンは異なる構造特性を持つため、別々のグループとした。

PCC のランダム分布の生成　PCC の統計的有意性を評価するためにランダムサンプリング法を用いて PCC のランダム分布を推定した。既知共役選択性のデータセットからランダムに 55 配列、34 配列、22 配列を選択した。選択した配列に対してロドプシンとのアラインメントを形成した。さらに有効な位置の中から 1 つだけを選択した。その位置について上記の方法と同様にアミノ酸グループの出現確率を推定し、PCC を求めた。この操作を乱数生成器として”Mersenne Twister” [78] を用いて、1,000,000 回上記の手続きを行って PCC のランダム分布を求めた。このランダム分布からある PCC 以上の累積密度を計算することで PCC 以上の値が出現する確率を求めそれを P-value とした。

5.1.2 結果

特定の G タンパク質の共役選択性に関係する残基を予測するために、特定の G タンパク質と残りの共役既知データとの間での PCC を計算し、そのスコアによってアミノ酸出現頻度の共役選択性との関連度合いを測定し、重要な残基を予測した。さらにその予測した位置をアラインメントを用いてロドプシンにマッピングして観察する。もし特定の G タンパク質に対する PCC が有意な値をとる位置があるならば、その位置はその特定の G タンパク質の共役に重要であるもしく

は、特定のGタンパク質との共役について阻害する働きがある位置であると推定される。

PCCの有意性　この解析のPCCの有意性を調べるために各Gタンパク質についてPCCのランダム分布を作成した。図5.1を見ればわかる通り、ランダム分布の形は$G_{i/o}$タイプ、$G_{q/11}$タイプ、G_sタイプそれぞれについて異なっている。これはPCCのランダム分布はデータセットのサイズに依存すると考えられる。PCCの値そのままでは、異なるGタンパク質タイプ間での比較が難しいため、ランダム分布よりPCCの値をP-valueに変換する。この研究では、P-valueが0.01以下となる位置を予測されたGタンパク質共役選択性と関係される位置（予測機能位置と呼ぶ）とし、表5.1にまとめた。

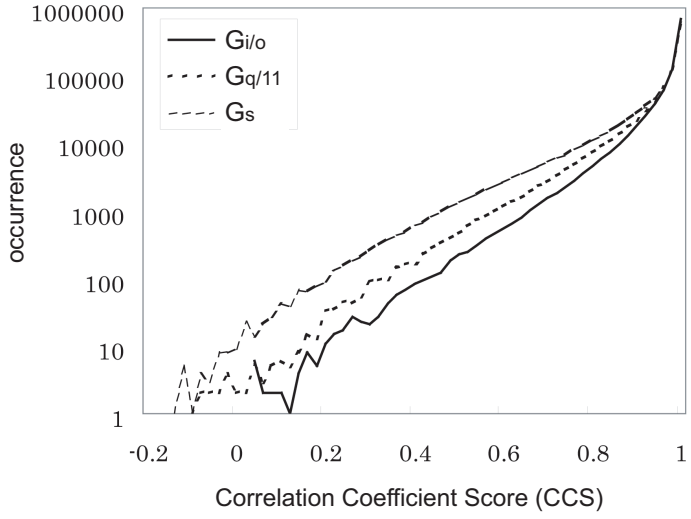

図 5.1　ランダムに生成したCCSの分布

予測機能位置を含むドメイン　データセットとテンプレートとのアラインメントを用いて予測機能位置の残基をロドプシンの構造上にマッピングした（図5.2、図5.3、表5.1）。図5.3では予測機能位置の残基はロドプシン内で側鎖を青い球体で

表している。左側の図が細胞内側からの視点で、右側の図は膜に対して平行な視点から見ている。図 5.2、図 5.3 及び表 5.1 では細胞内ドメインでの予測機能位置が他のドメインより多く見られる。GPCR は細胞内側で G タンパク質と相互作用するため、この結果は予想された結果と一致する。そのほかにも 3 本目の膜貫通ヘリックスの細胞外にも予測機能位置を見つけた。

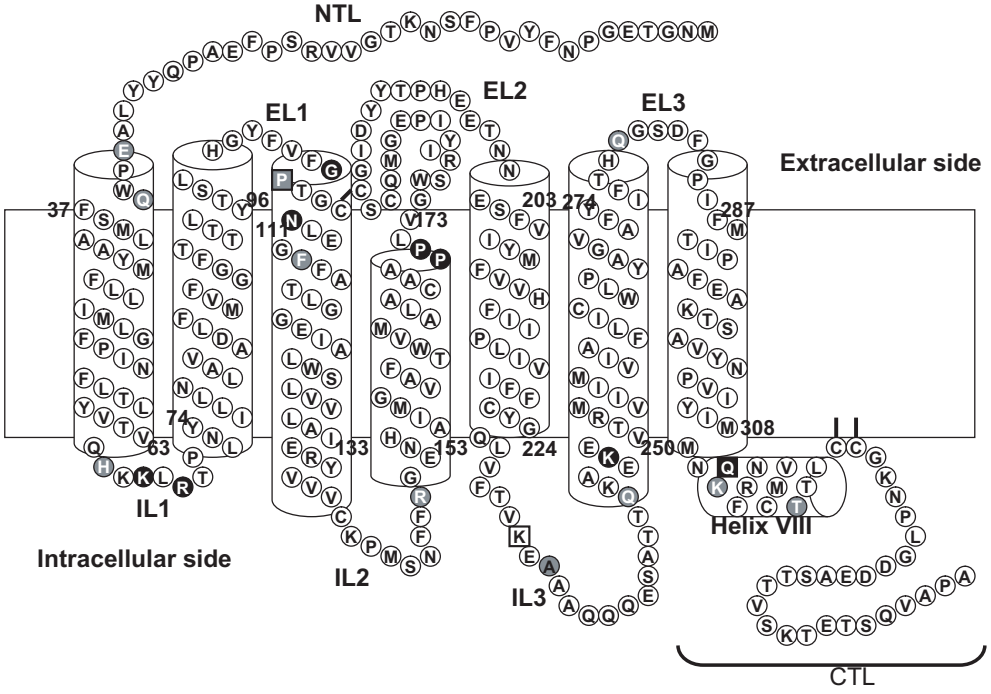

図 5.2　ロドプシンの概略図と予測機能位置に該当する残基。グレーの円形の残基は $G_{i/o}$ タイプとの共役選択性と関連すると予測された残基、白色四角は $G_{q/11}$ タイプ、黒色円形は G_s タイプ、グレー四角は $G_{i/o}$ と $G_{q/11}$ の双方、黒色四角は $G_{q/11}$ と G_s の双方、グレー円形は $G_{i/o}$ と G_s の双方と関連すると予測された残基を表している。

図 5.2 と図 5.3(a) は $G_{i/o}$ タイプの G タンパク質と共役する GPCR の予測機能位置の残基を示している。膜貫通ヘリックスの細胞外側には位置 33, 36, 107, 279 にあり、膜貫通領域では 3 本目の位置 115 にあり、細胞外側には位置 65, 148, 233,

表 5.1 特定の G タンパク質タイプ ($G_{i/o}$、$G_{q/11}$、G_s) に共役する GPCR の予測機能位置 (P-value < 0.01)

G タンパク質	位置	ドメイン	PCC (P-value)	mutation data
$G_{i/o}$	33	N 末端ループ	0.605 (0.0033)	
	36	N 末端ループ	0.640 (0.0045)	
	65	細胞内ループ 1	0.610 (0.0029)	[124][129]
	107	膜貫通領域 3	0.698 (0.0081)	
	115	膜貫通領域 3	0.620 (0.0037)	
	148	細胞内ループ 2	0.195 (0.00004)	
	233	細胞内ループ 3	0.607 (0.0033)	[16]
	244	細胞内ループ 3	0.684 (0.0071)	
	279	細胞外ループ 3	0.594 (0.0030)	
	311	C 末端ループ	0.402 (0.0004)	
	319	C 末端ループ	0.665 (0.0058)	[77]
$G_{q/11}$	107	N 末端ループ	0.525 (0.0033)	
	231	細胞内ループ 3	0.525 (0.0045)	[16]
	312	C 末端ループ	0.749 (0.0050)	
G_s	67	細胞内ループ 1	0.476 (0.0071)	
	69	細胞内ループ 1	0.445 (0.0054)	[6][38]
	106	膜貫通領域 3	0.416 (0.0043)	
	111	膜貫通領域 3	0.213 (0.0006)	
	170	膜貫通領域 4	0.344 (0.0022)	[74][73]
	171	膜貫通領域 4	0.375 (0.0029)	[107]
	233	細胞内ループ 3	0.233 (0.0007)	[16]
	248	細胞内ループ 3	0.459 (0.0061)	[23][72][117]
	312	C 末端ループ	0.366 (0.0022)	

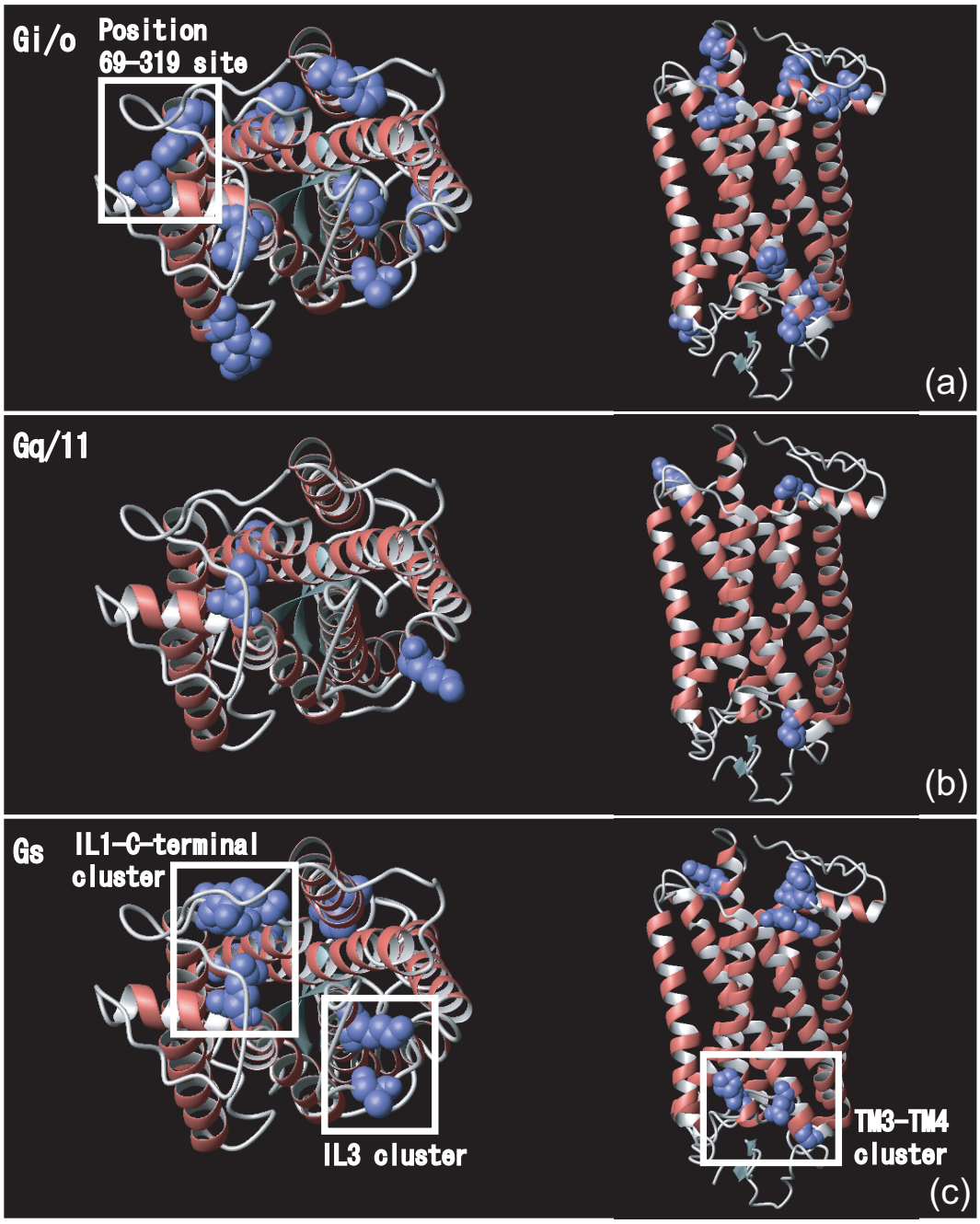

図 5.3 共役選択性に関する予測機能位置にあるロドプシン構造上の残基。MOLMOL [64] にて描画。

244, 311, 319 に存在した。細胞内側の位置 65 と 319 は構造上は隣接している位置にあるが、他の位置は特にお互いに構造上では隣接していなかった（図 5.3(a)）。次に図 5.2 と図 5.3(c) が示すには、G_s タイプと共役する GPCR に対する予測機能位置は 3 つのクラスターを形成しており、それらはそれぞれ 3 つ目の細胞内ループクラスター、1 つ目の細胞内ループと C 末端ループのクラスター、3 本目と 4 本目の膜貫通領域のクラスターに分かれる。3 つ目の細胞内ループクラスターは 6 本目の細胞内側に位置する位置 248 と細胞内ループ 3 の中間付近にある位置 233 に存在する。1 つ目の細胞内ループと C 末端のクラスターはロドプシンの構造上で細胞内で非常に近い位置に存在する（図 5.3(c)）。$G_{q/11}$ タイプと共役する GPCR に関しては有意水準を越える位置はあまりなかった（図 5.2、図 5.3、表 5.1）。C 末端にある膜貫通領域に隣接する 8 本目のヘリックス（図 5.2）にある位置 312 と 3 つ目の細胞内ループにある位置 231、膜貫通領域の細胞外側にある位置 107 のみが有意水準を超えている。この結果は $G_{q/11}$ タイプに共役する GPCR は他のタイプに比べて配列上での特徴が乏しいことを示唆している。Cao らの予測手法 [15] では $G_{q/11}$ タイプの予測精度が極端に低い結果となっているが、その理由は今回の解析で $G_{q/11}$ において配列上の特徴が少ないことに関連している可能性がある。

表 5.2 では細胞内ループに正電荷もしくは負電荷を持つアミノ酸の出現について示している。各位置について正電荷もしくは負電荷を持つアミノ酸の分布は G タンパク質の共役選択に関係する特徴的な傾向を示す。特に 3 つ目の細胞内ループでは $G_{i/o}$ タイプや $G_{q/11}$ タイプの GPCR では 20 配列以上と多くのものに見られるが、G_s タイプと共役する GPCR ではわずか 1 配列のみである。負電荷を持つアミノ酸の場合、位置 311 では $G_{i/o}$ タイプの GPCR とその他の GPCR では出現頻度の差が 30%以上と大きく差がある。また、位置 312 では G_s タイプと共役する GPCR とそれ以外とでは出現頻度の差が 20%もある。

これらの結果は予測機能残基の大部分は細胞内ドメインに位置していることを示唆している。しかしながら、表 5.1 と図 5.2 では細胞外側の残基のいくつかも予測機能残基であると予測されている。例えば、位置 33、36、106、107、170、171、279 はヘリックスの細胞外側付近に位置している。これらの位置のうち、170 番目の位置ヘリックスの細胞側にあり、特定のアミノ酸について特徴的な出現頻度

表 5.2 細胞内ループ内の予測機能位置での正電荷/負電荷を持つアミノ酸の出現頻度

		正電荷を持つアミノ酸			負電荷を持つアミノ酸		
位置	ドメイン	$G_{i/o}$	$G_{q/11}$	G_s	$G_{i/o}$	$G_{q/11}$	G_s
65	細胞内ループ 1	3 (0.06)	6 (0.18)	7 (0.32)	6 (0.11)	4 (0.12)	0 (0.00)
67	細胞内ループ 1	29 (0.53)	18 (0.53)	4 (0.18)	3 (0.05)	1 (0.03)	0 (0.00)
69	細胞内ループ 1	36 (0.65)	15 (0.44)	4 (0.18)	1 (0.02)	0 (0.00)	0 (0.00)
148	細胞内ループ 2	34 (0.62)	6 (0.18)	0 (0.00)	0 (0.00)	1 (0.03)	1 (0.00)
231	N 末端ループ	30 (0.55)	8 (0.24)	15 (0.68)	0 (0.00)	3 (0.09)	1 (0.00)
233	細胞内ループ 3	16 (0.29)	9 (0.26)	0 (0.00)	0 (0.00)	2 (0.06)	0 (0.00)
244	細胞内ループ 3	16 (0.29)	4 (0.12)	0 (0.00)	6 (0.11)	0 (0.00)	0 (0.00)
248	細胞内ループ 3	34 (0.58)	22 (0.65)	1 (0.05)	1 (0.02)	0 (0.00)	0 (0.00)
311	C 末端ループ	10 (0.18)	18 (0.53)	9 (0.41)	19 (0.35)	5 (0.15)	0 (0.00)
312	C 末端ループ	18 (0.33)	8 (0.24)	0 (0.00)	17 (0.31)	4 (0.12)	14 (0.64)
319	C 末端ループ	28 (0.51)	12 (0.35)	4 (0.18)	3 (0.05)	6 (0.18)	6 (0.27)

表内部の数字は特定の位置について正電荷 (もしくは負電荷) を持つアミノ酸が観察される特定の G タンパク質に共役する GPCR の個数、() 内はその割合、下線部は特定のタイプの G タンパク質と共役する GPCR が他のタイプに比べて割合が 0.2 以上大きいもしくは小さいことを表す。

が見られる。表5.3に示す通り、プロリン (170P) は$G_{i/o}$タイプ、$G_{q/11}$タイプではたくさんの配列で観測されるが、G_sタイプと共役する受容体では5-HTタイプ7受容体、アドレメデュリン受容体、バソプレッシン受容体V2でしか観測されない。位置171では、G_s受容体内でプロリンが出現する頻度は高いが、その他のタイプではあまり観測されない。

表 5.3 予測機能位置でのプロリンの出現頻度

位置	$G_{i/o}$	$G_{q/11}$	G_s
170	45 (0.82)	25 (0.74)	3 (0.14)
171	15 (0.27)	7 (0.21)	14 (0.64)

表内部の数字は特定の位置についてプロリンが観察される特定のGタンパク質に共役するGPCRの個数、() 内はその割合を表す。

5.1.3 議論

予測機能残基をロドプシン構造にマッピングする解析はGタンパク質の共役選択性に関係する重要だと予想される残基を明らかにする。これらの残基は2つの位置に分かられる (1)Gタンパク質と直接相互作用する細胞内ドメインに位置する残基 (2)GPCRとGタンパク質のインターフェイスから離れている膜貫通領域や細胞外ループなどの領域に位置する残基。ここではこの2つについて分けて予測機能残基とGタンパク質共役選択性との関係について述べる。

細胞内ドメイン　細胞内ドメインの残基の多くはGタンパク質と直接相互作用することから、細胞内ドメインに位置する残基がGタンパク質共役選択性を決定する上で重要な役割を果たしていると考えられる。いくつかの変異実験は細胞内ドメイン、特に2つ目の細胞内ループ及び3つ目の細胞内のN末端側とC末端側の重要性を示唆している [121]。しかしながら、Gタンパク質の相互作用の決定にかかわる共通の配列パターンはいまだに見つかっていない [121]。

過去の変異実験によると細胞内ループの電荷を持つ残基がGタンパク質の認識と活性化の鍵となる役割を果たしている [117]。そのため、予測機能残基と電荷を持つアミノ酸との関係について調べた。表5.2は細胞内ドメインでの正電荷もしくは負電荷を持つアミノ酸の出現を示している。つ目の細胞内ループに存在する位置248ではG_sタイプと共役するGPCRに比べて、その他のタイプのGPCRは正電荷を持つアミノ酸の出現頻度は非常に大きい。いくつかの変異実験が位置248において正電荷を持つアミノ酸の存在が$G_{i/o}$もしくは$G_{q/11}$との共役に重要であることを示唆ししている [23][72][117][116]。それらの結果は位置248での正電荷を持つアミノ酸が出現しないことは$G_{i/o}$もしくは$G_{q/11}$との共役を妨げているという可能性を示唆している。G_sタイプと共役するGPCRと他のGPCRとの間で位置247、248、251での正電荷を持つアミノ酸の出現頻度の差が大きい(>0.2)(ただし、位置247と251はPCCの有意水準に達していない)。いくつかの変異実験は位置247と位置252も$G_{i/o}$もしくは$G_{q/11}$との共役に重要であることを示している [117][118]。これらの結果は細胞内ループ3のC末端のいくつかの位置での電荷を持つ残基の存在が共役選択性に影響を与えていることを示唆している。

細胞外ドメイン及び膜貫通領域　細胞内ドメインがGタンパク質共役選択性に最も重要な部位の1つであることはよく知られているが、LH/CG受容体の2つ目の細胞外ループなどのような他のドメインの残基がGタンパク質の共役選択性について重要な役割を果たしているという実験結果も報告されている [39]。これらの残基はGタンパク質との相互作用をする細胞内ドメインから離れている。表5.1が示すように、いくつかの予測機能位置はGPCRの細胞外側に位置している。細胞外ドメインと膜貫通領域の細胞外側は一般には主にリガンド結合にかかわっていると考えられている [108][120]。これらの残基はGPCRの活性化における構造変化に影響を与えるもしくは制限するかもしれなく、特定のGタンパク質との相互作用を阻害する可能性もある。ヘリックスの外側にある位置170ではのGPCRではプロリン残基が$G_{i/o}$タイプと$G_{q/11}$タイプよく出現するが、G_sタイプではほとんど出現しないという特徴的な出現が見られることを前述した。ヘリックス構造内でのプロリン残基は一般的にヘリックスを壊したり、ヘリックスを曲げたりすることが知られており、構造フォールディングにおいて重要である。また、

4 本目の膜貫通ヘリックスの細胞外側での位置 170 のプロリンはリガンド結合やいくつかの GPCR については活性化について重要であることが示唆されている [55][74][73]。しかしながら同じリガンドと結合する特定の GPCR ファミリー（例えばセロトニン受容体ファミリーやヒスタミン受容体ファミリー、ドーパミン受容体ファミリーなど）において必ずしも保存されていないことを考えるとリガンド結合以外の別の機能がある可能性がある。変異実験による裏づけはないが、これらの解析は位置 170 のプロリンはなんらかの形で G タンパク質共役選択性の決定に対して直接もしくは間接的に影響を与えているかもしれない。

何よって G タンパク質共役選択性は制御されているのか？ GPCR と G タンパク質との直接的な相互作用に関係する残基すなわち細胞内ループ内の残基で G タンパク質共役選択性が決定されると仮定する。すると、同じ受容体の異なるドメインのいくつかの部位で G タンパク質共役選択性に影響を与える。G タンパク質共役選択性の変化に影響のある一残基置換実験は 2 つ目の細胞内ループ、3 つ目の細胞内ループ、C 末端ループといったさまざまなドメインで発見されている [120][123]。それらのドメインに存在するいくつかの予測機能位置が今回の解析で発見された。これらの残基は異なる位置に広く散らばっており、これらの残基から共役選択性に関係するモチーフパターンのような局所的な共通配列領域をみつけることはできない。これらの発見は G タンパク質と接触する領域付近に位置されている一残基もしくは複数の残基だけではなく、他の離れた残基が組み合わさって共役選択性を実現していることが示唆される。これらの残基は一次配列上では離れているかもしれないが、三次元構造上では近い領域に存在しうる（図 5.3）。G タンパク質と直接相互作用しないいくつかの残基もまた GPCR の構造変化を通して G タンパク質の活性化に影響を与えるかもしれない。

細胞内ドメインの立体構造が G タンパク質共役選択性を決定するという仮説は G タンパク質の共役選択性の変化がわずかの細胞内ループの残基や細胞内ドメインと離れた残基が関係している [123] ということも説明できる。この研究では表 5.1、図 5.2、図 5.3 で示したように G タンパク質活性化に影響を与えるいくつかの膜貫通領域もしくは細胞外ドメインに存在する予測機能残基を示した。これらの予測機能位置の残基が細胞内ループドメインの構造変化に直接もしくは間接的

に影響を与えている。変異実験ではそれらの中のいくつかはGタンパク質の活性化に重要であると示している [39][74]。

GPCRとGタンパク質の相互作用している高解像度の立体構造データの不足のため、Gタンパク質共役選択性を理解するのに必要なだけのGPCRの構造の知識はまだない。変異実験はGタンパク質共役選択性についての情報を提供するけれども、この選択性を考えるためにすべてのGPCRについて実験データを得ることは難しい。インシリコによる分析がGPCRの既知配列を基にしたGPCRとGタンパク質の関係についての情報を提供できる。本研究は変異実験のデータを補完し、共役選択性に対して新しい視点を提供している。

5.2. その他の特徴量候補の探索

ここではアミノ酸出現頻度以外の観点から共役選択性予測の特徴量の候補となるものを探す。特定のドメインの長さやドメイン内のアミノ酸の各種指標（疎水性値、側鎖の体積など）の平均値、GPCRのリガンドの分子量など様々な観点から特徴量候補を探索した。その中で有効な特徴量候補として発見したドメインの長さ、リガンドの分子量について説明する。

ドメインの長さに関する特徴量　GPCRは7本の膜貫通領域を持つタンパク質である。よって、ドメインは7本の膜貫通領域、N末端ループと3つの細胞外ループ、3つの細胞内ループとC末端ループと計15のドメインを持つ。各ドメインは4.1で説明した手法でロドプシンの配列とターゲット配列とのアラインメントをとり、ロドプシンの膜貫通領域とアラインされた領域をターゲット配列における膜貫通領域とした。この手法ではほとんど膜貫通領域ではギャップは挿入されない。よって、膜貫通領域の長さはどのターゲット配列もロドプシンの膜貫通領域の長さとほぼ等しく、データセット中のGPCR配列の膜貫通領域の長さはほとんど差がない。よって、ここではループ領域の長さに着目する。解析するデータセットとしては文献データ [1] やアノテーション情報 [11] から抽出した133個のヒトのGPCR配列 ($G_{i/o}$：61, $G_{q/11}$：47, G_s：24) を用いた。

表 5.4 133 の GPCR 配列に対する各ループの長さの最頻値とその値でのデータ数

ドメイン	最頻値	最頻値でのデータ数
N 末端ループ	50	8
細胞内ループ 1	10	103
細胞外ループ 1	14	62
細胞内ループ 2	20	100
細胞外ループ 2	28	19
細胞内ループ 3	19	15
細胞外ループ 3	15	23
C 末端ループ	48	5

表 5.4 はデータセット内の各ループの長さの最頻値とその最頻値でのデータ数を表す。この図が示す通り、細胞内ループ 1、2 と細胞外ループ 1 は 4 割以上のデータが同じループの長さになっていることがわかる。このようにほとんどループの長さが変動しないものは各共役選択性ごとに差が出ないので、残りの 5 つのループの長さについて検討する。

その 5 つのループの共役選択性別のループの残基長の分布を表したものを図 5.4 に示す。N 末端ループ、細胞内ループ 2 及び細胞外ループ 3 についてはどの共役選択性も分布の形に違いがあるものの判別分析の特徴量として使えるような特徴的な傾向は見られない。細胞外ループ 2 については一部の G_s タイプと共役する GPCR にループの長さが 5 残基のものが集まっているが、これらは Melanocortin 受容体ファミリーが集まっているだけで共役選択性に関係していると考えるのは難しい。次に細胞内ループ 3 について見ると、どの共役選択性についても 10-40 残基ぐらいに多くの GPCR が存在する。しかしながら、100 残基以上では G_s タイプの GPCR は存在せず、$G_{i/o}$ タイプと $G_{q/11}$ タイプの GPCR のみが観察される。一方、C 末端ループを見ると、多くの GPCR は 90 残基以下である。90 残基以上に注目してみると、$G_{i/o}$ タイプは一番データ数が多いにもかかわらず Somatostatin 受容体タイプ 3 のみが 90 残基以上であり、他の GPCR は 90 残基以下である。一

方、$G_{q/11}$ タイプと G_s タイプでは 90 残基以上の GPCR が観察される。

細胞内ループ 3 と C 末端ループの長さについて詳細に検討ため、この 2 つのループの長さの散布図を作成した（図 5.5)。散布図で見ると細胞内ループ 3 が 80 残基以上で C 末端ループが 50 残基以下の領域に $G_{i/o}$ タイプと $G_{q/11}$ タイプの GPCR が集積しており、G_s タイプは観察されない。また、細胞内ループ 3 が 50-100 残基で、C 末端ループが 80 残基以上の領域には $G_{q/11}$ タイプと G_s タイプの GPCR が集積しており、$G_{i/o}$ タイプの GPCR は観察されない。他の多くの GPCR は細胞内ループ 3 が 50 残基以下で C 末端ループが 100 残基以下の領域に観察される。細胞内ループ 3 と C 末端ループの残基長が長い GPCR の一部に関して共役選択性との関連が予想される。さらにデータをサブクラス Amine に限定した散布図を図 5.6 に示す。第 2 章 2.3 でも述べたが、Amine は感覚系 GPCR を除くと Class A GPCR の中では Peptide に次いで数が多い。散布図を見ればわかるように、先ほど指摘した細胞内ループ 3 が長く C 末端ループが短い領域と細胞内ループが 50-100 残基程度で C 末端ループが長い領域に存在する GPCR の多くはサブクラス Amine の GPCR である。よって、このループの長さによる特性はサブクラス Amine に特に強く見られる傾向であることがわかる。これらのループの長さの違いが共役選択性にどのような影響があるか現在の知見からはわからないが、共役選択性になんらかの関係がある可能性がある。

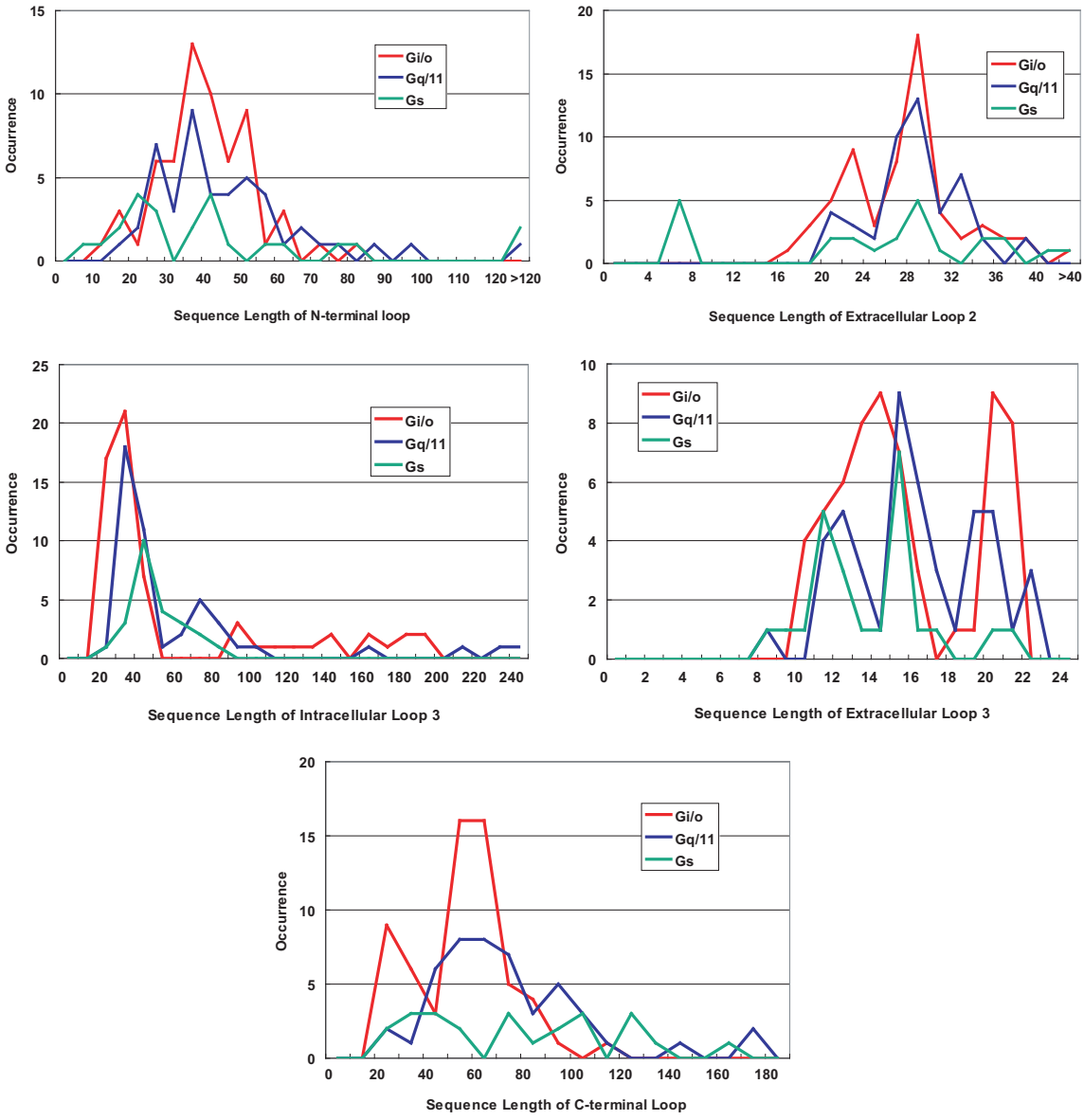

図 5.4 共役選択性別ループの残基長の分布。左上：N 末端ループ 右上：細胞外ループ 2 右中：細胞内ループ 3 左中：細胞外ループ 3 下：C 末端ループ

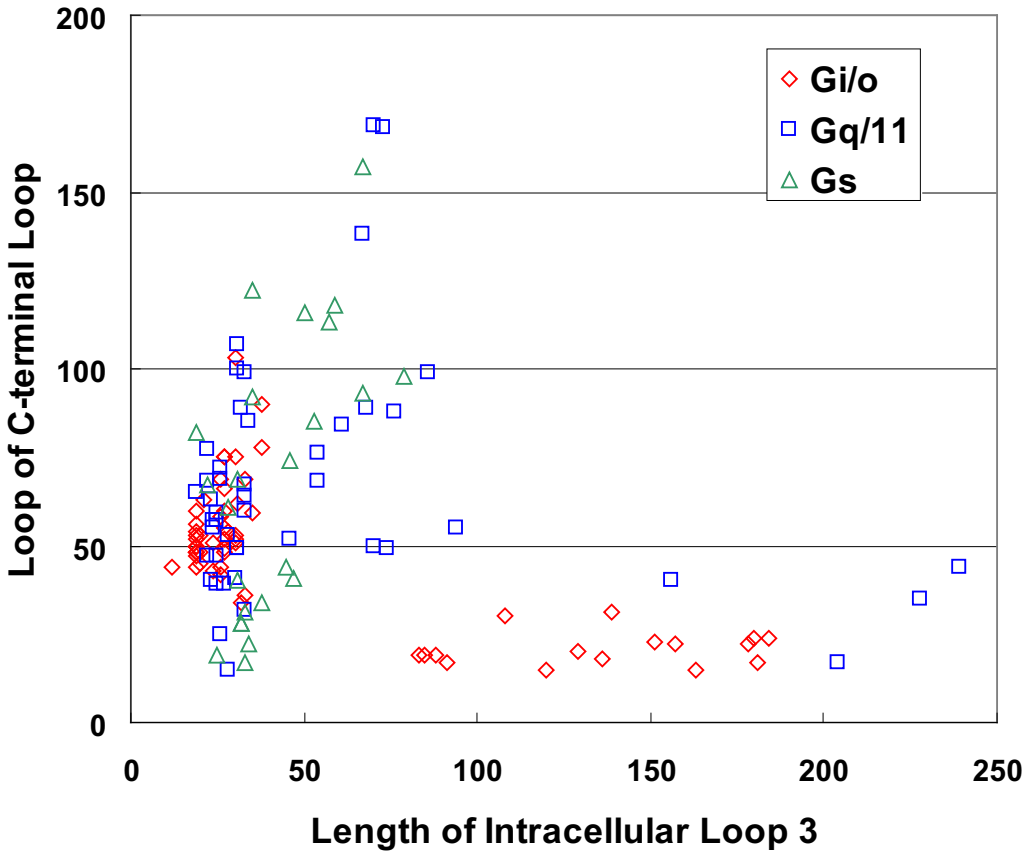

図 5.5　細胞内ループ 3 と C 末端ループの残基長についての散布図。縦軸は C 末端ループの残基長、横軸は細胞内ループ 3 の残基長を表す。

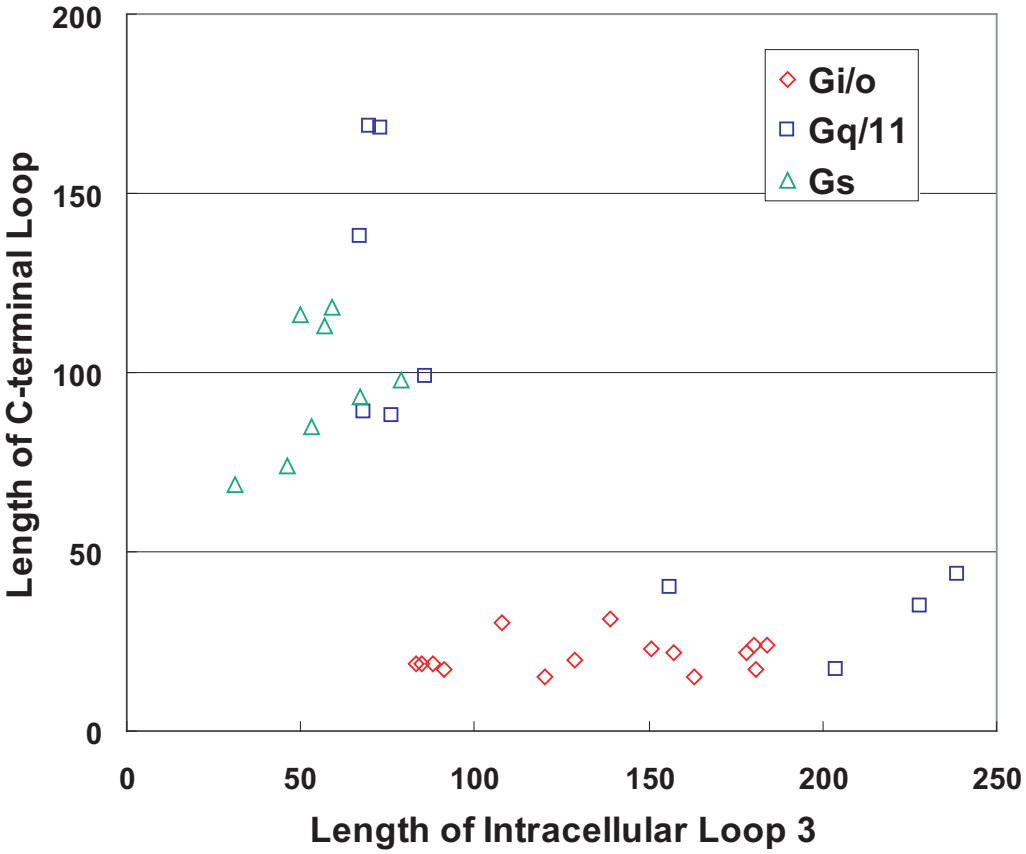

図 5.6 細胞内ループ 3 と C 末端ループの残基長についての散布図 （サブクラス Amine の GPCR のみ） 縦軸は C 末端ループの残基長、横軸は細胞内ループ 3 の残基長を表す。

リガンドの大きさに関する特徴量　リガンドは直接 G タンパク質を相互作用するわけではないので、共役選択性に関係ないとも考えられる。しかし、GPCR はリガンドにより活性化され、その活性化構造が G タンパク質と相互作用するため、リガンドと GPCR との相互作用が共役選択性とまったくかかわりないと決め付けることができない。そこで、リガンドと共役選択性の関係について検討してみる。GPCR のリガンドは Amine （Dopamine, Adreneline など）のような低分子化合物もあれば、脂質、ペプチド、タンパク質など非常に多様である。このようなリガンドすべてで計測できる尺度を考えると、単純なものでは分子量が存在する。そこで共役選択性とリガンドの分子量との関係を図 5.7 に示す。おおよその目安としては分子量 500 以上はペプチドもしくはタンパク質をリガンドとするものであり、その他のリガンドは分子量 500 未満である。半分以上の GPCR が分子量が 500 未満であり、500 未満の分子量ではどの共役選択性の GPCR も観察される。serotonin 受容体ファミリー（serotonin の分子量:170）や adreneline 受容体ファミリー（adreneline の分子量:183）に属する GPCR は多くあり、そのファミリー内では同じリガンドと結合するが、共役選択性に関しては $G_{i/o}$ タイプと共役する GPCR、$G_{q/11}$ タイプと共役する GPCR、G_s タイプと共役する GPCR が全て含まれており、これらのファミリーではリガンドの分子量と共役選択性は関係がないということは明白である。図 5.7 と前述の serotonin ファミリーの例などを勘案すると、低分子において分子量と共役選択性との関係は見出せない。次に分子量 500 以上を見ると、$G_{q/11}$ タイプの GPCR の多くは 1000-2000 ぐらいに集中している。一方 $G_{i/o}$ タイプは分子量 2500 以上に偏在している。G_s は分子量 500 以上のデータ数が少ないが、存在するものを観察すると広く偏在している。分子量 500 以上では $G_{i/o}$ タイプと $G_{q/11}$ タイプについては分布の偏りがみられる。これらのリガンド分子量の違いが共役選択性にどのように関係しているか、現在の知見からは判断はできないが、何らかの関係がある可能性がある。

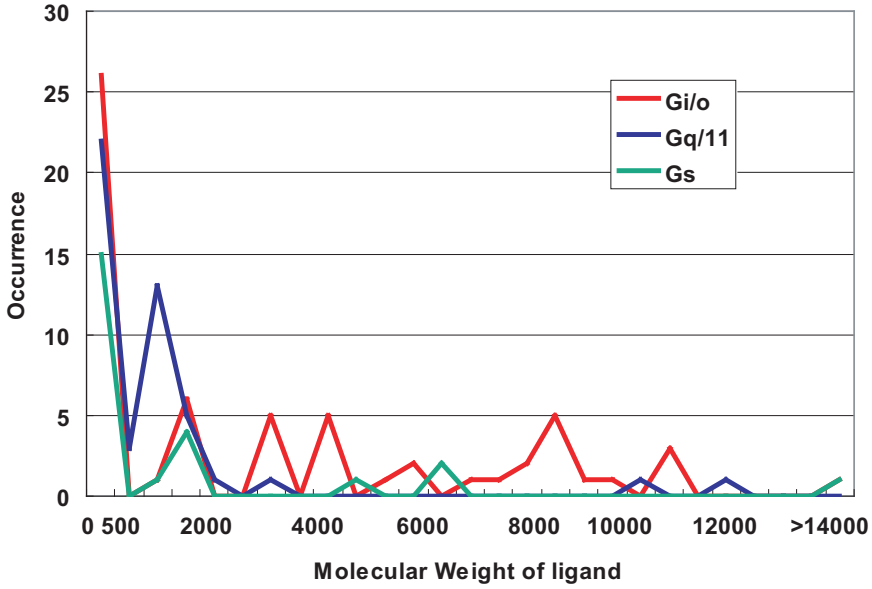

図 5.7 共役選択性別のリガンドの分子量の分布

5.3. 予測方法

共役選択性に関連する要因をいくつかの観点からみてきた。ここでは、これらの要因を統合して、ひとつの予測システムを考える。作成した予測システムの全体的なフローは図 5.8 のようになっている。予測システムは特定ファミリー判別と G_s 判別、$G_{i/o}$-$G_{q/11}$ 判別の全部で 3 つのステップからなる。

まずはじめのステップではプロファイル隠れマルコフモデルを用いて今回の解析しているデータセットに含まれていないような GPCR は前処理として分類される。ここまで、検討してきた共役選択性はあくまでデータセットで使われているものの傾向である。データセットとしては、特定の細胞にのみ発現する GPCR や Class A 以外の GPCR は除外してきた。Opsin ファミリーに属する GPCR は視覚関連の細胞にのみ発現し、視覚関連の細胞に特異的に発現する G_t タイプの G タンパク質と共役する。Olfactory ファミリーに属する GPCR は嗅覚関連の細胞にのみ発現し、嗅覚関連の細胞に特異的に発現する G_{olf} タイプの G タンパク質と共役する。Class B GPCR は現在分かっている GPCR はすべて G_s タイプと共役する。Class C GPCR は $G_{i/o}$, $G_{q/11}$ タイプのどちらかの G タンパク質と共役する。Frizzled, Smoothened ファミリーの GPCR は共役 G タンパク質は不明である。これらの特定のファミリーや特定の Class に属する GPCR は互いに配列の類似性が高く、プロファイル隠れマルコフモデルなどの既存の配列手法で十分分離可能である。そこでここでは各ファミリーについて HMMER [30] を用いて、ファミリー分類を行い。その結果を基に G タンパク質の共役選択性を予測する。ただし、Class C GPCR は 2 つのタイプの G タンパク質と共役する可能性があるので、各共役選択性に対してプロファイルを作成する。すなわち、$G_{i/o}$ 共役型 Class C GPCR プロファイルと $G_{q/11}$ 共役型 Class C GPCR プロファイルを作成して予測を行う。

残りの 2 つのステップでは前述したファミリーではない Class A GPCR に対して 3 つの共役選択性のどれかを予測する。ただし、本章の 5.1 や 5.2 の結果を見る限り、各共役選択性に対して、それぞれで重要な特徴量が違う可能性が高い。そこで、ここではまずはじめに G_s タイプの共役選択性を判別し、最後に $G_{i/o}$ と $G_{q/11}$ タイプの GPCR を判別する。

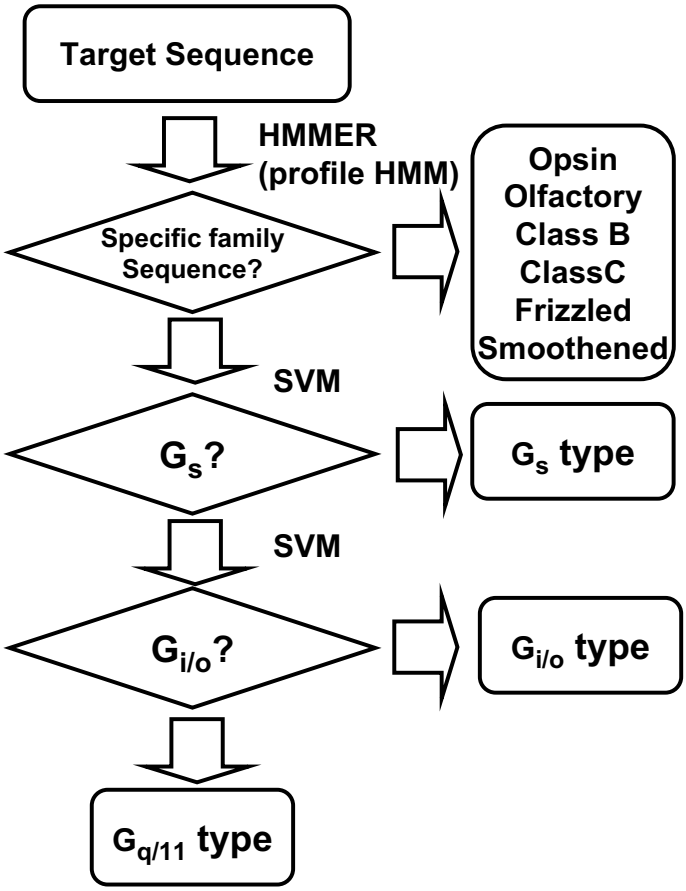

図 5.8 共役選択性予測システムの全体のフロー

特定のファミリーの共役選択性予測　特定のファミリーに対するプロファイル隠れマルコフモデルのプロファイルを作成した。Class A GPCR に属する Opsin ファミリー及び Olfactory ファミリーに対しては、4.1 で作成した Class A GPCR 用プロファイルを用いてデータセット全体のマルチプル配列アラインメントを作成し、そのアラインメントをシードアラインメントとして HMMER の hmmbuild コマンドを用いてプロファイルを作成した。Class A GPCR ではない Class B GPCR、Class C GPCR、Frizzled、Smoothened に関しては、CLUSTAL W [110] を用いてデータセットのマルチプル配列アラインメントを作成し、そのアラインメントをシードアラインメントとしてプロファイルを作成した。

G_s タイプの判別分析のための特徴量　G_s タイプの判別分析のために、細胞内ループ 3 の長さ、C 末端ループの長さ、Amine プロファイル、3 番目の細胞内ループの C 末端の塩基性残基の数、位置 170 のプロリンの有無を使用した。

細胞内ループ 3 と C 末端ループの長さは前述の通り共役選択性に関係性が示唆されている。特にその傾向はサブクラス Amine に顕著に見られる。そこで、これらのループの長さと Amine に関するプロファイル隠れマルコフモデルを用いたプロファイルの bit score の値を利用して、サブクラス Amine らしさとループの長さを考慮した判別を考えた。Amine プロファイルは特定のファミリーの共役選択性予測で述べたプロファイルの作成方法と同じ方法で作成した。残りの 2 つの特徴量は立体構造へのマッピングを行った研究から作成した。立体構造へのマッピングの研究ではいくつかの重要と予想される残基を提示したが、この中で構造への影響が大きそうな位置 170 とプロリン、細胞内ループ 3 の C 末端側の塩基性残基 (Arg, Lys) を採用した。プロリンに関しては位置 170 にプロリンがあるかないかについて 1 か 0 で表現した。細胞内ループ 3 の C 末端の塩基性残基については位置 244、247、248、251 について Lys、Arg の出現数を特徴量として使用した。これらの位置は G_s タイプの GPCR において他のタイプと共役する GPCR と比較して相対的に正電荷を持つアミノ酸の出現頻度が低い位置をピックアップした。これらの位置のロドプシン上の残基を図 5.9 に示す。これらの位置は膜貫通ヘリックス 6 の膜から飛び出た部分と推測されるヘリックス領域に存在しており、さらに膜貫通領域 3 と膜貫通領域 5 と接している位置に存在する。この膜貫通へ

リックス 3、膜貫通ヘリックス 5、膜貫通ヘリックス 6 の細胞内側の領域は G タンパク質との相互作用に重要な領域といわれており、活性化構造ではこの 3 つのヘリックスの配置の変化することが知られている。この塩基性アミノ酸の出現に特徴がある位置がこの領域に存在していることは共役選択性との関連も大いに考えられる。

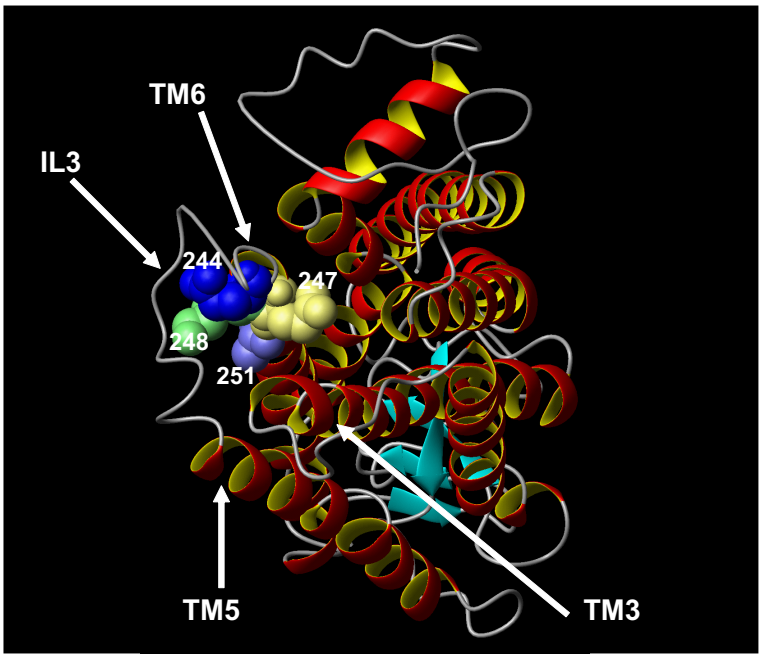

図 5.9　塩基性アミノ酸の特徴量として使用した位置の残基。ロドプシン立体構造を細胞内側から見た図。該当する残基は側鎖を CPK で表示。

$G_{i/o}$ タイプの判別分析のための特徴量　残る $G_{i/o}$ タイプと $G_{q/11}$ タイプの判別分析のために、G_s タイプの判別に使用した細胞内ループ 3 の長さ、C 末端ループの長さ、Amine プロファイルの 3 つの特徴量に加え、リガンドの特徴量と Peptide プロファイルを使用した。

リガンドの特徴量は前述の通り、分子量 500 以上のペプチドリガンドでその傾向が見られる。分子量 500 未満では関係性がみられないため、リガンドの分子量

とともにサブクラス Peptide の配列プロファイルを作成した。作成方法は Amine プロファイルの場合と同じである。このようにループの長さとリガンドの長さを特徴量として予測を行った。

予測したリガンド分子量　この研究ではリガンドの分子量を特徴量として組み入れている。しかしながら、G タンパク質の共役選択性予測をリガンド未知のオーファン GPCR に適用するという観点からは望ましくない。そこで第 4 章で開発したペプチド性リガンドの予測値を使う。第 4 章ではペプチド、タンパク質リガンドの残基長を予測した。この予測残基長とアミノ酸 1 残基当たりの分子量の期待値の積を求めるという簡易的な手法で分子量を推定した。

$$\text{一残基当たりの分子量の期待値} \quad = \quad \sum_{a} f_a * m_a \tag{5.2}$$

f_a はペプチド・タンパク質リガンドにおけるアミノ酸 a の出現頻度、m_a はアミノ酸の分子量（PeptideMass [37] から取得）を表す。一残基当たりの分子量の期待値は 113.70 であった。

リガンドの予測長としては、ターゲット配列がリガンド予測の学習データに含まれている場合にはそのデータを除いて学習して予測した値を、含まれていない場合は第 4 章での学習データをそのまま学習に使い予測値を求めた。

判別手法　判別手法としては線形判別法はじめ多くの手法が提案されている [28]。今回の特徴量を観察する限りでは線形判別を用いた平面での予測は難しく、非線形判別手法が必要であると考える (例えば、図 5.4 など参照)。そこで、本研究ではサポートベクターマシーン (SVM) を用いた（SVM の概要は付録 B を参考）。SVM はカーネル関数を利用することで、特徴空間において判別分析を行うことが可能となり、カーネル関数によっては非線形判別も可能となる。一般的に非線形の写像によって変換した特徴空間の次元は大きくなる傾向になり、膨大な計算量を必要とし、また汎化能力の低下にもつながる。しかしながら、SVM では内積が計算できれば最適な識別関数を構成することができ、計算も容易であり、パラメータ次第では高い汎化能力を持つことができる。今回は SVM には LIBSVM [18] を用いて実装した。

予測の評価方法と評価指標　カーネル関数としては linear、RBF、polynomial の 3 つを使い一番結果のよいものを採用した。パラメータ C、γ はそれぞれ C は 2^{-5} から 2^{15}、γ は 2^{-13} から 2^{3} の範囲で最適なものを探した。評価方法は 4-fold クロスバリデーションテスト（データを 4 分割し、3/4 を学習データとして、1/4 をテストデータとして使用する）を用いた。なお、リガンドの予測値を使うときは学習ステップから予測値を使って学習した。

評価指標としては以下のような Specificity と Sensitivity を用いた。

$$Specificty = \frac{TP}{TP+FP} \tag{5.3}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{5.4}$$

TP (True Positive) は正解データを正しく正解だと予測した数、FP (False Positive) は不正解データを誤って正解だと予測した数、FN (False Negative) は正解データを誤って不正解だと予測した数を表す。これらの評価指標について 4-fold クロスバリデーションテストを 10,000 回行い、その平均値を求めた。

5.4. 予測結果

予測システムの予測結果の評価について述べる。評価データベースとしては、プロファイル隠れマルコフモデルのステップでは、GPCR 分類データベース GPCRDB [50][48] に格納される GPCR を対象とし、特定のファミリーに属するデータとそれ以外のデータとで判別できるかを評価した。データ数は全 GPCR データ 5895 個、Opsin 170 配列、Olfactory 394 配列、Class B 34 配列、Class C $G_{i/o}$ 20 配列、Class C $G_{q/11}$ 9 配列、Frizzled 40 配列、Smoothened 5 配列であった。また、SVM を用いるステップでは文献データ [1] やアノテーション情報 [11] から抽出した 133 個のヒトの GPCR 配列 ($G_{i/o}$: 61, $G_{q/11}$: 47, G_s : 24) を用いた。複数のタイプの G タンパク質と共役する GPCR は除いてある。

プロファイル隠れマルコフモデルステップの結果　前処理であるプロファイル隠れマルコフモデルの精度を表 5.5 に示した。結果をみてわかるようにどのファミ

リーについてもほぼ完璧に近い精度で予測できている。これは各ファミリーに属するGPCRの配列類似性が互いに高く予測が容易であることがわかる。OpsinとClass C $G_{i/o}$ファミリーについて一部誤判別が見られるが、一部の例外的な配列を除けば、前処理としての機能は十分発揮していると判断できる。

表 5.5 前処理における隠れマルコフモデルによる予測精度

ファミリー	Gタンパク質タイプ	Sensitivity (%)	Specificity (%)	閾値
Opsin	G_t	99.7	100.0	153.9
Olfactory	G_{olf}	100.0	100.0	151.2
Class B	G_s	100.0	100.0	68.0
Class C	$G_{i/o}$	93.5	100.0	1054.6
	$G_{q/11}$	100.0	100.0	1325.3
Frizzled	不明	100.0	100.0	168.7
Smoothened	不明	100.0	100.0	627.6

SVM ステップの結果　ここでは、SVM を用いた予測の評価の結果を表 5.6 示す。G_s タイプの判別では RBF カーネルが、$G_{i/o}$ タイプと $G_{q/11}$ タイプの判別では polynomial カーネルが最適であった。それぞれのステップでの判別結果を評価した。結果をみると、G_s タイプ判別ステップでも $G_{i/o}$ タイプと $G_{q/11}$ タイプの判別ステップでもどの共役選択性のタイプも Specificity、Sensitivity ともに 80-90%程度の値で予測できていることがわかる。

表 5.6　132GPCR について SVM 部分での G_s タイプに関する共役選択性の予測精度

G タンパク質タイプ	データ数	Sensitivity (%)	Specificity (%)
G_s	24	84.97	88.74

表 5.7　108GPCR について SVM 部分での $G_{i/o}$ タイプと $G_{q/11}$ タイプに関する共役選択性の予測精度

G タンパク質タイプ	データ数	Sensitivity (%)	Specificity (%)
$G_{i/o}$	61	87.07	88.32
$G_{q/11}$	47	84.97	83.62

リガンド予測との組み合わせ　次にリガンドデータを実際の値ではなく、第 4 章で求めた予測値を使い、それを基に共役選択性予測を行った。G_s タイプの判別にはリガンドデータを使用していないため $G_{i/o}$ タイプと $G_{q/11}$ タイプの判別のみを行っている。ここでは両方の判別で RBF カーネルを用いている。パラメータの探索は前述の範囲と同じである。結果を表 5.8 に示す。表 5.6 と比較すると、Specificity、Sensitivity ともに 6-7%程度精度が落ちる結果となっている。今後は第 4 章の手法のさらなる改良と共役選択性に関する他の特徴量の探索が必要である。

表 5.8　リガンド予測と組み合わせた場合の G タンパク質共役選択性の予測精度

G タンパク質タイプ	Sensitivity (%)	Specificity (%)
$G_{i/o}$	81.72	81.66
$G_{q/11}$	75.86	76.54

5.5. 議論とまとめ

本章では G タンパク質共役選択性予測法について述べてきた。まずはじめに、共役選択性に関係する残基を評価するために、ロドプシンとデータセットの配列とのアラインメントを基にどの位置の残基がアミノ酸出現頻度の観点から重要であるかを評価した。その手法により同定された予測機能部位をロドプシンの立体構造にマッピングし、どのような部位が重要かを評価した。それらの残基の一部は配列上は隣接していないが、構造上はクラスターを形成していることを発見した。また、ループ領域の電荷を持つアミノ酸の出現及び膜貫通ヘリックス上のプロリン残基の出現に特徴があることを発見した。

次に、この考察を基に共役選択性に特に関係していそうな要因として、細胞内ループ 3 の C 末端側の正電荷を持つアミノ酸の出現頻度と位置 170 におけるプロリンの有無の予測への利用を考えた。さらに、他の要因を解析した結果、細胞内ループ 3 と C 末端のループの長さとリガンドの分子量という要因を発見した。そ

れらの要因を特徴量としてサポートベクター分類を用いて判別した。その結果、specificity 及び sensitivity ともに 80-90%で予測できることがわかった。

ここまで解析を行ってきたが、依然として共役選択性がどのようなメカニズムで実現しているか理解できていない。今後の課題としてまず構造的なアプローチで共役選択性の解析があげられる。今回の解析はアミノ酸配列をベースとして、構造情報としてはロドプシンの構造情報を利用した二次構造予測とロドプシンの構造を利用した可視化による考察のみにとどまっている。そこで、次のアプローチはとしては実際に比較モデリング法により立体構造予測を行い、予測構造をもとにした解析が考えられる。

近年の研究の進展により、GPCR のリン酸化 [24] が共役選択性に変化を与えることが知られている。また、他のタンパク質との相互作用や GPCR の二量体化による影響についてもまだ理解できていない。このように様々な要因と共役選択性の関係が示唆されており、他にも関連する可能性要因も多く、ますますメカニズムの解明を困難にしている。今回の特徴量ではこれらの問題に対して何も対応できていない。そこで、予測の観点からは新しい実験の知見に関して共役選択性との関連が示唆されるような知見に関して大量に収集し、知識ベースを蓄積し、予測したい配列に対してこれらの知見があてはまりそうか検索する（例えば、リン酸化部位の有無やどの位置かを予測など）ようなシステムを作り、その結果を提示する、もしくは結果を基に今回のような多変量解析手法を用いて予測するなどの解決法が考えられる。

第6章　結論

本論文ではアミノ酸配列からの機能予測法について議論した。

第3章では汎用的な膜タンパク質のアラインメント手法を提案した。ペア隠れマルコフモデルを用い、膜貫通領域とアラインメントの両方を同じモデルで記述することによって、膜貫通領域予測とアラインメント作成を同時に行う手法を開発した。構造アラインメントを用いた評価実験では、配列類似性が低い配列ペア及び膜貫通領域において提案手法が既存のグローバルアラインメントより高い精度であることを示した。

第4章、第5章ではGPCRという特定の膜タンパク質ファミリーに関して機能予測法を提案した。

第4章ではGPCRのペプチド・タンパク質系リガンドの残基長をGPCR配列から予測する手法を提案した。まず始めに、ロドプシンとターゲット配列とのアラインメントを作成法について述べた。Class A GPCRの配列アラインメントを行うために、ロドプシンの各残基を一致状態に対応したプロファイル隠れマルコフモデルを構築した。さらに既知立体構造における膜貫通領域のギャップについて考察し、ロドプシンの構造ベースとして、膜貫通領域のモデルパラメータを変更した。そのアラインメントを用いて、ターゲット配列に対して位置とドメインを推定した。位置に対してはその位置のアミノ酸に対応するアミノ酸指標を特徴量とした。ドメインに関してはドメイン内のアミノ酸に対するアミノ酸指標の平均を特徴量とした。アミノ酸指標としてはアミノ酸指標AAindex内のすべての指標を利用した。さらに特徴量選択を行うためにTanimoto係数を用いて、類似アミノ酸指標による冗長性を減らし、また残基長との類似度の低い特徴量を取り除いた。選択された特徴量に対してサポートベクター回帰を用いて残基長を予測した。その結果、平均12残基、中央値8残基程度の誤差で予測可能であることを

示した。また、ペプチド・タンパク質系リガンドを持つと予想されるオーファンGPCRに対して適用し、どのような残基長のリガンドを持つGPCRが潜在的に存在するか推定した。

第5章ではGPCRのGタンパク質共役選択性をGPCR配列から予測する手法を提案した。4.1の手法を用いて、ロドプシンとターゲット配列とのアラインメントを作成した。そのアラインメントを用いて各位置に対してアミノ酸出現頻度と共役選択性の関係性について評価した。その結果、アミノ酸出現頻度に特徴がある部位をいくつか発見した。それらの残基をロドプシン構造にマッピングして観察すると、いくつかのクラスターが存在することがわかった。そのほかにも細胞内ループには電荷を持つアミノ酸の出現頻度に特徴がある位置が多いこと、4本目の膜貫通領域にはプロリン残基に特徴がある位置があることがわかった。

アミノ酸出現頻度に関する解析によって得られた電荷を持つアミノ酸とプロリンの出現頻度に関する情報とループの長さ、リガンドの分子量の情報を用いて特徴量ベクトルを作成した。その特徴量を用いてサポートベクター分類を行った。その結果specificity及びsensitivityともに80-90%程度の予測精度で分類できた。

本論文では特にGPCRに関する機能予測法を中心に議論を展開してきたが、今回開発した手法はGPCR以外の膜タンパク質ファミリーにも適用できると考えている。本論文はGPCR特有の機能について、GPCRの特徴を考えながら機能予測法について議論してきたが、アラインメントを基にした位置情報及び構造領域情報を利用した予測方法であるため、これらの部分は今回の第3章や4.1の方法でアラインメントを構築することで他の膜タンパク質ファミリーにも適用可能である。もちろん個別ファミリーに関する情報を活用し、より正確なアラインメントを基に第4章のようなアラインメントのベースとした機能予測法や、より詳細に特徴量を検討した第5章のような手法を検討する必要がある。別のファミリーに対してファミリー特有の機能について予測方法を開発する場合に本論文で述べた手法は重要なヒントを与えるものである。

膜タンパク質には実験の困難な場合が多く、まだ多くの機能未知遺伝子が存在している。これらの遺伝子の機能解析は今もなお重要な課題であるが、現在の実験手法から考えるとすぐにこの課題が解決する可能性は低い。そこで、本研究を

基にした膜タンパク質の各ファミリーに対する機能予測法の開発し活用することが膜タンパク質の機能解析研究を加速させることができると期待している。

謝辞

本論文をまとめるに当たって多くの皆様のご支援とご教授を承りました。研究活動全般に対して様々なサポートをしていただいた方々に感謝の意をここで表したいと思います。諏訪牧子先生には研究活動に対するアドバイスなど研究全般についてご指導ありがとうございました。

また、産業技術総合研究所生命情報科学センター (CBRC) の生体膜情報チームの Dr. Michael Gromiha、向井有理博士、池田修己博士、Dr. Xavier Suresh、矢葺幸光さん、小野幸輝さん、新居真吏さん、数理モデルチームの浅井潔先生、金大真博士、木立尚孝博士、加藤毅先生、分子設計チームの広川貴次先生、富井健太郎博士、アルゴリズムチームの山田真介さんをはじめとする CBRC の研究員及びスタッフ、実習生の皆様には研究活動及び研究生活の両面で多くの議論の機会やアイディアを頂き、様々なご支援、ご教授を承りました。皆様の支援がなければ本論文の完成はなかったと思っております。

また、石井信先生には様々なご支援をいただき、また、本論文の主査を引き受けていただきありがとうございました。最後になりましたが有益なアドバイスとともに本論文の審査を引き受けてくださった審査員の先生の皆様に深く感謝申し上げます。

付録

A. 隠れマルコフモデル (HMM)

隠れマルコフモデル (HMM) は確率モデルの 1 つであり、システムがパラメータ未知のマルコフ過程であると仮定し、観測可能な情報からその未知のパラメータを推定する。配列解析分野では遺伝子発見 [125][14]、膜貫通領域予測 [114][105]、プロファイル検索 [30] など様々な応用分野に適用されている手法である。

本節では本研究で使用したペア隠れマルコフモデル (PHMM) [29] 及び HMMER タイプのプロファイル隠れマルコフモデル (Profile HMM) [30]、その前提となる隠れマルコフモデル (HMM) [95][29] について簡単に説明する。

A.1 隠れマルコフモデル (HMM)

N 個の状態 $\{s_1, s_2, \ldots, s_N\}$ を考える。プロセスはある状態から次の状態を移りながら状態の配列 $(s_{i1}, s_{i2}, \cdots, s_{ik})$ を生成する。マルコフチェインの特性として、各部分配列の状態の確率は 1 つ前の状態にのみ依存する。すなわち、以下のことのようになる。

$$P\left(s_{ik}|s_1, s_2, \ldots, s_{ik-1}\right) \quad = \quad P\left(s_{ik}|s_{ik-1}\right) \tag{6.1}$$

状態はわからないが、各状態は M 個の観測シンボル（例えば、アミノ酸）$\{v_1, v_2, \ldots, v_M\}$ の 1 つをランダムに生成する。

隠れマルコフモデルを定義するために、遷移確率、出力確率を特定する必要がある。

遷移確率：$a_{kl} = P\left(\pi_i = l|\pi_{i-1} = k\right)$

出力確率：$e_k\left(b\right) = P\left(x_i = b|\pi_i = k\right)$

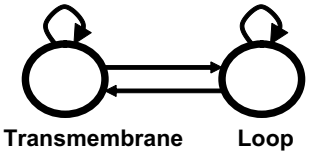

図 6.1 簡単な膜タンパク質の隠れマルコフモデル

遷移確率、出力確率の行列を $A = (a_{kl})$、$B = (e_i(v_m))$ とすると、モデルは $M = (A, B)$ として表せれる。

ここで簡単な膜タンパク質を表す HMM の例を図 6.1 に示す。状態は膜貫通領域とループの 2 つだけである。ある膜タンパク質のアミノ酸配列とモデルが与えられた時、その配列上の各残基が膜貫通領域かループであるかを推定する場合、以下の式のような最適パスを探さなければならない。

$$\pi^* = arg\max_{\pi} P(x, \pi) \tag{6.2}$$

この最適パスは動的計画法を使って、Viterbi アルゴリズムという方法で求める。

繰り返し $(i = 1 \ldots L) : v_l(i) = e_l(x_i) \max_k (v_k(i-1) a_{kl})$

$$ptr_i(l) = arg\max_k (v_k(i-1) a_{kl})$$

最終処理 : $P(x, \pi^*) = \max_k (v_k(L))$

$$\pi^*_L = arg\max_k (v_k(L))$$

トレースバック : $\pi^*_{i-1} = ptr_i(\pi^*_i)$

なお確率値は非常に低くなるため、実装上は確率を log スケールに変換して各種アルゴリズムを実行する。

次に、アミノ酸配列とモデルが与えられた時、その配列が膜タンパク質であるかその確率を求める場合、すなわち以下の式の値を求めたい場合を考える。

$$P(x) = \sum_{\pi} P(x, \pi) \tag{6.3}$$

この場合は以下の Forward アルゴリズムを使って求める。

繰り返し $(i = 1 \ldots L) : f_l(i) = e_l(x_i) \sum_k (f_k(i-1) a_{kl})$

最終処理 : $P(x) = \sum_k (f_k(L))$

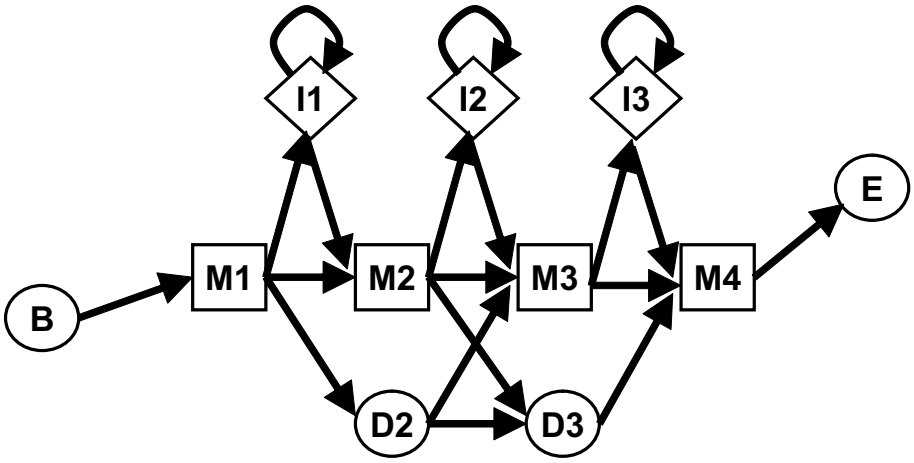

図 6.2 HMMER Plan7 のプロファイル HMM アーキテクチャーの一部

A.2 プロファイル隠れマルコフモデル (HMM)

プロファイル隠れマルコフモデルは隠れマルコフモデルの 1 種であり、生物学の配列解析の分野では、モチーフなどの特定の配列パターンをモデル化することに利用される。モデルのアーキテクチャーは研究ごとに若干異なるが [30][52]、ここでは今回使用した HMMER [30] をベースとして話を進める。

HMMER でプロファイル HMM のコアとなる部分のモデルのアーキテクチャーを図 6.2 に示す。B と E の状態は開始状態、終端状態を表し、シンボルは何も出力しない。M1-M4 は（一般的には）配列のコンセンサスパターンに対応し、そのパターンにマッチした場合にこの状態を通る。M1-M4 ではアミノ酸残基を 1 つだけ出力する。そのに大して D2-D3 の状態はコンセンサスパターンの欠損のための状態であり、この状態では何も出力せず、例えば M1 → D2 → M3 というパスを通った場合コンセンサスパターンである M2 は出力されない。残りの I1-I3 はコンセンサスパターンに対する挿入を表し、コンセンサスパターン内でパターンとは関係ないアミノ酸を出力される。関係ないアミノ酸が出力される長さは遷移確率によって定義されている。また、HMMER におけるプロファイル隠れマルコフモデルでは図 6.3 のような NULL モデルが定義されている。状態 F でのみシンボル

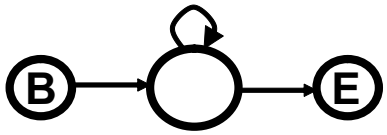

図 6.3 HMMER Plan7 の NULL モデル

を出力し、ランダムにアミノ酸配列を生成するモデルとなっている。状態 F でアミノ酸 a を出現する確率（バックグランド確率とも呼ぶ）を q_a とする。

ある配列にコンセンサスパターンがあると分かっている場合に、どの残基がコンセンサスパターンに当てはまるかという問題を解くには、配列の各残基に対する状態系列を求めればよい。状態系列を求めるアルゴリズムは Viterbi アルゴリズムに対応する。マッチ状態の数が m、配列長が n である場合のプロファイル隠れマルコフモデルにおける Viterbi アルゴリズムは以下のようになる。

繰り返し：$i = 1, \ldots, n$

$$
\begin{aligned}
v_j^M(i) &= \frac{e_{M_j}(x_i)}{q_{x_i}} max \left\{ \begin{array}{l} a_{M_{j-1}M_j} v_{j-1}^M (i-1) \\ a_{I_{j-1}M_j} v_{j-1}^I (i-1) \\ a_{D_{j-1}M_j} v_{j-1}^D (i-1) \end{array} \right. \\
v^I(i) &= \frac{e_{I_j}(x_i)}{q_{x_i}} max \left\{ \begin{array}{l} a_{M_j I_j} v_j^M (i-1) \\ a_{I_j I_j} v_j^I (i-1) \\ a_{D_j I_j} v_j^D (i-1) \end{array} \right. \\
v^D(i) &= max \left\{ \begin{array}{l} a_{M_{j-1}D_j} v_{j-1}^M (i) \\ a_{I_{j-1}D_j} v_{j-1}^I (i) \\ a_{D_{j-1}D_j} v_{j-1}^D (i) \end{array} \right.
\end{aligned}
$$

終端処理：$v^E = max\left(v^{M_m}(n), v^{I_m}(n), v^{D_m}(n)\right)$

次にある配列が特定のコンセンサスパターンを持っているかを表すスコア（コンセンサスパターンを持つ確率と NULL モデルの確率との比の対数）を考える。プロファイル隠れマルコフモデルにおける bit score とはこのスコアを示してい

る。bit score は以下のように Forward アルゴリズムで求めることができる。

$$
\begin{aligned}
F_j^M(i) &= log\frac{e_{M_j}(x_i)}{q_{x_i}} \\
&+ log[a_{M_{j-1}M_j} exp\left(F_{j-1}^M(i-1)\right) \\
&+ a_{I_{j-1}M_j} exp\left(F_{j-1}^I(i-1)\right) \\
&+ a_{D_{j-1}M_j} exp\left(F_{j-1}^D(i-1)\right)] \\
F^I(i) &= log\frac{e_{I_j}(x_i)}{q_{x_i}} \\
&+ log[a_{M_jI_j} exp\left(F_j^M(i-1)\right) \\
&+ a_{I_jI_j} exp\left(F_j^I(i-1)\right) \\
&+ a_{D_jI_j} exp\left(F_j^D(i-1)\right)] \\
F^D(i) &= log[a_{M_{j-1}D_j} exp\left(F_{j-1}^M(i)\right) \\
&+ a_{I_{j-1}D_j} exp\left(F_{j-1}^I(i)\right) \\
&+ a_{D_{j-1}D_j} exp\left(F_{j-1}^D(i)\right)]
\end{aligned}
$$

終端処理： $v^E = v^{M_m}(n) + v^{I_m}(n) + v^{D_m}(n)$

Viterbi アルゴリズムによって求めらるスコアは最適パスにおけるスコアであるのに対して、Forward アルゴリズムではすべての考えられるパスについてその確率値を加算して、さらに NULL モデルの確率で割っている値を対数化したものに相当している。

A.3 ペア隠れマルコフモデル

ここまで述べてきた隠れマルコフモデルは 1 つ 1 つのシンボル（例：残基、塩基）を出力することで 1 つの配列を生成するもでるであった。それに対してここで述べるペア隠れマルコフモデル (PHMM) は 2 つの配列ペア X, Y に関してシンボルのペアを出力し、アラインメントを生成する手法である。標準的なグローバル配列アラインメントと同等なペア隠れマルコフモデルを図 6.4 に示す。配列 X, Y の両方からの残基ペアを出力する状態（一致・置換状態)、配列 X からの残基

とギャップを出力する状態（挿入状態)、配列 Y からの残基とギャップを出力する状態（欠損状態）が存在している。

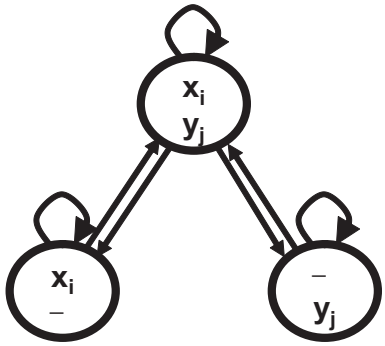

図 6.4 シンプルな PHMM のアーキテクチャー

2 つの配列 X, Y が与えられた時、その配列のアラインメント（ペア隠れマルコフモデルの最適な状態系列）を求める場合には以下の Viterbi アルゴリズムを用いる。

繰り返し：$i = 1, \ldots, n, j = 1, \ldots, m$

$$v^M(i,j) = p_{x_i y_j} max \begin{cases} a_{MM} v^M(i-1, j-1) \\ a_{MI} v^M(i-1, j-1) \\ a_{MD} v^M(i-1, j-1) \end{cases} \tag{6.4}$$

$$v^I(i,j) = q_{x_i} max \begin{cases} a_{IM} v^M(i-1, j) \\ a_{II} v^I(i-1, j) \end{cases} \tag{6.5}$$

$$v^D(i,j) = q_{y_j} max \begin{cases} a_{DM} v^M(i, j-1) \\ a_{DD} v^D(i, j-1) \end{cases} \tag{6.6}$$

終端処理：$v^E = \max\left(v^M(n,m), v^I(n,m), v^D(n,m)\right)$

B. サポートベクターマシーン (SVM)

本節では本研究で使用した LIBSVM [18] におけるサポートベクターマシーン (SVM) についての概要の説明する。パターン認識分野における代表的な手法の 1 つで、主に分類問題と回帰分析問題に適用される。配列解析分野では分類問題ではアミノ酸配列からの遠縁のホモログ配列予測 [53]、細胞局在予測 [51] などに利用されていて、回帰分析問題ではタンパク質の各残基の接触可能表面の面積予測 [128] などに利用されている。ここでは LIBSVM で実装されている分類問題のために使用した C-SVC と回帰分析問題のために使用した ϵ-SVR について説明する。

B.1 サポートベクター分類 (C-SVC)

ここでは第 5 章で使用したサポートベクター分類の手法の 1 つである C-SVC について説明する。

2 つのクラスについて学習ベクトル $x_i \in R^n$、$i = 1, \ldots, l$、$y_i \in \{1, -1\}$ となるようなベクトル $y \in R^l$ が与えられた時、C-SVC における解くべき問題は以下のようになる。

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{制約条件 } y_i\left(\mathbf{w}^T\phi\left(x_i\right) + b\right) \geq 1 - \xi_i, \\ \xi_i \geq 0, i = 1, \ldots, l \end{aligned} \tag{6.7}$$

その双対問題は以下のようになる。

$$\begin{aligned} \min_{\alpha} \quad \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \\ 0 \geq \alpha_I \geq C, i = 1, \ldots, l \\ \text{制約条件 } y^T\alpha = 0, \end{aligned} \tag{6.8}$$

$C > 0$ は上限、Q は非負定値行列で $Q_{ij} \equiv y_i y_j K(x_i y_j)$ であり、$K(x_i y_j) \equiv \phi(x_i)^T\phi(x_j)$ はカーネルを表す。ここでは訓練ベクター x_i は関数 ϕ によってより高い次元空間に飛ばされる。

なお、本研究で使用したカーネル関数 $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_i)$ は以下の 3 種類である。

$$
\begin{aligned}
linear \quad kernel : K(x_i, x_j) &= x_i^T x_j && (6.9)\\
polynomial \quad kernel : K(x_i, x_j) &= \left(\gamma x_i^T x_j + r\right)^d, \gamma > 0 && (6.10)\\
RBF \quad kernel : K(x_i, x_j) &= exp\left(-\gamma ||x_i - x_j||^2\right), \gamma > 0 && (6.11)
\end{aligned}
$$

この二次計画問題は SMO-type 分解法 [33] によって解かれる。

その識別関数は

$$
f(x) = sgn\left(\sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b\right) \tag{6.12}
$$

B.2 サポートベクター回帰 (ϵ-SVR)

ここでは第 4 章で使用したサポートベクター回帰の手法の 1 つである ϵ-SVR について説明する。

$x_i \in R^n$ が入力、$z_i \in R^1$ がターゲット出力とし、データ点の集合 $\{(x_1, z_1), \ldots, (x_l, z_l)\}$ が与えられた時、Support vector regression の標準形は次のようになる。

$$
\begin{aligned}
\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i + C\sum_{i=1}^{l}\xi_i^* && (6.13)\\
\mathbf{w}^T\phi(x_i) + b - z_i &\geq 1 - \xi_i,\\
z_i - \mathbf{w}^T\phi(x_i) - b &\geq 1 - \xi_i,\\
\xi_i, \xi_i^* \geq 0, i &= 1, \ldots, l
\end{aligned}
$$

その双対問題は

$$
\min_{\mathbf{w},b,\xi} \frac{1}{2}(\alpha - \alpha^*)^T Q(\alpha - \alpha^*) + \epsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} z_i(\alpha_i - \alpha_i^*)
$$

$$
\text{制約条件} \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0, 0 \geq \alpha_i, \alpha_i^* \geq C, i = 1, \ldots, l
$$

ここで $Q_{ij} = K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ を表す。

近似関数は以下のようになる。

$$f(x) = \sum_{i=1}^{l} (-\alpha_i + \alpha_i^*) K(x_i, x) + b \tag{6.14}$$

参考文献

[1] S. Alexander, A. Mathie, J. Peters, G. MacKenzie, and A. Smith. Tips receptor and ion channel nomenclature supplement 2001. *Trends Pharmacol. Sci.*, 19:1–106, 2001.

[2] S.P. Alexander, A. Mathie, and J.A. Peters. 7 tm receptors. *Br. J. Pharmacol.*, 144:S4–S62, 2005.

[3] M. Alexandersson, S. Cawley, and L. Pachter. Slam: cross-species gene finding and alignment with a generalized pair hidden markov model. *Genome Res.*, 13:496–502, 2003.

[4] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[5] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.

[6] K.K. Arora, L.Z. Krsmanovic, N. Mores, H. O'Farrell, and K.J. Catt. Mediation of cyclic amp signaling by the first intracellular loop of the gonadotropin-releasing hormone receptor. *J. Biol. Chem.*, 273:25581–25586, 1998.

[7] A. Bairoch, R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, and L.S. Yeh. The universal protein resource (uniprot). *Nucleic Acids Res.*, 33:D154–159, 2005.

[8] R. Benz. Permeation of hydrophilic solutes through mitochondrial outer membranes: review on mitochondrial porins. *Biochim. Biophys. Acta.*, 1197:167–196, 1994.

[9] M.M. Berglund, R. Fredriksson, E. Salaneck, and D. Larhammar. Reciprocal mutations of neuropeptide y receptor y2 in human and chicken identify amino acids important for antagonist binding. *FEBS Lett.*, 518:5–9, 2002.

[10] C. Bissantz, A. Logean, and D. Rognan. High-throughput modeling of human g-protein coupled receptors: amino acid sequence alignment, three-dimensional model building, and receptor library screening. *J. Chem. Inf. Comput. Sci.*, 44:1162–1176, 2004.

[11] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic. Acids. Res.*, 31:365–370, 2003.

[12] P.E. Bourne, K.J. Addess, W.F. Bluhm, L. Chen, N. Deshpande, Z. Feng, W. Fleri, R. Green, J.C. Merino-Ott, W. Townsend-Merino, H. Weissig, J. Westbrook, and H.M. Berman. The distribution and query systems of the rcsb protein data bank. *Nucleic Acids Res.*, 32:D223–D225, 2004.

[13] A. Bruce, J. Alexander, L. Julian, R. Martin, R. Keith, and W. Peter. *Molecular Biology of the Cell, Fourth Edition.* Garland Science Publishing, 2001.

[14] C. Burge and S. Karlin. Prediction of complete gene structures in human genomic dna. *J. Mol. Biol.*, 268:78–94, 1997.

[15] J. Cao, R. Panetta, S. Yue, A. Steyaert, M. Young-Bellido, and S. Ahmad. A naive bayes model to predict coupling between seven transmembrane domain receptors and g-proteins. *Bioinformatics*, 19:234–240, 2003.

[16] S.A. Carlson, T.K. Chatterjee, K.P. Murphy, and R.A. Fisher. Mutation of a putative amphipathic alpha-helix in the third intracellular domain of the platelet-activating factor receptor disrupts receptor/g protein coupling and signaling. *Mol. Pharmacol.*, 53:451–458, 1998.

[17] J.M. Chandonia, G. Hon, N.S. Walker, L. Lo Conte, P. Koehl, M. Levitt, and S.E. Brenner. The astral compendium in 2004. *Nucleic Acids Res.*, 32:D189–192, 2004.

[18] C. Chang and C. Lin. Libsvm: a library for support vector machines. *National Taiwan University, Department of Computer Science and Information Engineering, Taiwan*, 2004.

[19] C. Chang, A. Ray, and P. Swaan. In silico strategies for modeling membrane transporter function. *Drug Discov. Today*, 10:663–671, 2005.

[20] O. Civelli. Gpcr deorphanizations: the novel, the known and the unexpected transmitters. *Trends Pharmacol. Sci.*, 26:15–19, 2005.

[21] O. Civelli, H.P. Nothacker, Y. Saito, Z. Wang, S.H. Lin, and R.K. Reinscheid. Novel neurotransmitters as natural ligands of orphan g-protein-coupled receptors. *Trends Neurosci.*, 24:230–237, 2001.

[22] F.S. Cordes, J.N. Bright, and M.S. Sansom. Proline-induced distortions of transmembrane helices. *J. Mol. Biol.*, 323:951–960, 2002.

[23] S. Cotecchia, S. Exum, M.G. Caron, and R.J. Lefkowitz. Regions of the alpha 1-adrenergic receptor involved in coupling to phosphatidylinositol hydrolysis and enhanced sensitivity of biological function. *Proc. Natl. Acad. Sci. USA*, 87:2896–2900, 1990.

[24] Y. Daaka, L.M. Luttrell, and R.J. Lefkowitz. Switching of the coupling of the beta2-adrenergic receptor to different g proteins by protein kinase a. *Nature*, 390:88–91, 1997.

[25] N. Deshpande, K.J. Addess, W.F. Bluhm, J.C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, L. Xie, L. Chen, Z. Feng, R.K. Green, J.L. Flippen-Anderson, J. Westbrook, H.M. Berman, and P.E. Bourne. The rcsb protein data bank: a redesigned query system and relational database based on the mmcif schema. *Nucleic Acids Res.*, 33:D233–D237, 2005.

[26] C.B. Do, M.S. Mahabhashyam, M. Brudno, and S. Batzoglou. Probcons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.*, 15:330–340, 2005.

[27] J. Drews. Drug discovery: a historical perspective. *Science*, 287:1960–1964, 2000.

[28] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification.* Wiley-Interscience, 2000.

[29] R. Durbin, S.R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, 1998.

[30] S.R. Eddy. Profile hidden markov models. *Bioinformatics*, 14:755–763, 1998.

[31] R.C. Edgar. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, 2004.

[32] M. Ernst, D. Brauchart, S. Boresch, and W. Sieghart. Comparative modeling of gaba(a) receptors: limits, insights, future developments. *Neuroscience*, 119:933–943, 2003.

[33] R.E. Fan, P.H. Chen, and C.J. Lin. Working set selection using the second order information for training svm. *Technical report, Department of Computer Science, National Taiwan University*, 2005.

[34] K. Fischer, A. Weber, S. Brink, B. Arbinger, D. Schunemann, S. Borchert, H.W. Heldt, B. Popp, R. Benz, and T.A. Link. Porins from plants. molecular cloning and functional characterization of two new members of the porin family. *J. Biol. Chem.*, 269:25754–25760, 1994.

[35] T.M. Fong, H. Yu, and C.D. Strader. Molecular basis for the species selectivity of the neurokinin-1 receptor antagonists cp-96,345 and rp67580. *J. Biol. Chem.*, 267:25668–25671, 1992.

[36] R. Fredriksson and H.B. Schioth. The repertoire of g-protein-coupled receptors in fully sequenced genomes. *Mol. Pharmacol.*, 67:1414–1425, 2005.

[37] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, R.D. Appel, and A. Bairoch. *Protein Identification and Analysis Tools on the ExPASy Server; (In) John M. Walker (ed): The Proteomics Protocols Handbook,*, pages 571–607. Humana Press, 2005.

[38] L. Geng, J. Wu, S.P. So, G. Huang, and K.H. Ruan. Structural and functional characterization of the first intracellular loop of human thromboxane a2 receptor. *J. Biol. Chem.*, 273:25581–25586, 2004.

[39] R.L. Gilchrist, K.S. Ryu, I. Ji, and T.H. Ji. The luteinizing hormone/chorionic gonadotropin receptor has distinct transmembrane conductors for camp and inositol phosphate signals. *J. Biol. Chem.*, 271:19283–19287, 1996.

[40] A. Giolitti, P. Cucchi, A.R. Renzetti, L. Rotondaro, S. Zappitelli, and C.A. Maggi. Molecular determinants of peptide and nonpeptide nk-2 receptor antagonists binding sites of the human tachykinin nk-2 receptor by site-directed mutagenesis. *Neuropharmacology*, 39:1422–1429, 2000.

[41] B. Han and A.H. Jr. Tashjian. Importance of extracellular domains for ligand binding in the thyrotropin-releasing hormone receptor. *Mol. Endocrinol.*, 9:1708–1719, 1995.

[42] S. Hemmerich, C. Paavola, A. Bloom, S. Bhakta, R. Freedman, D. Grunberger, J. Krstenansky, S. Lee, D. McCarley, M. Mulkins, B. Wong, J. Pease, L. Mizoue, T. Mirzadegan, I. Polsky, K. Thompson, T.M. Handel, and K. Jarnagin. Identification of residues in the monocyte chemotactic protein-1 that contact the mcp-1 receptor, ccr2. *Biochemistry*, 38:13013–13025, 1999.

[43] S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U. S. A.*, 89:10915–10919, 1992.

[44] S. Henikoff and J.G. Henikoff. Position-based sequence weights. *J. Mol. Biol.*, 243:574–578, 1994.

[45] E. Hermans. Biochemical and pharmacological control of the multiplicity of coupling at g-protein-coupled receptors. *Pharmacol. Ther.*, 99:25–44, 2003.

[46] S. Hinuma, Y. Shintani, S. Fukusumi, N. Iijima, Y. Matsumoto, M. Hosoya, R. Fujii, T. Watanabe, K. Kikuchi, Y. Terao, T. Yano, T. Yamamoto, Y. Kawamata, Y. Habata, M. Asada, C. Kitada, T. Kurokawa, H. Onda, O. Nishimura, M. Tanaka, Y. Ibata, and M. Fujino. New neuropeptides containing carboxy-terminal rfamide and their receptor in mammals. *Nat. Cell Biol.*, 2:703–708, 2000.

[47] T. Hirokawa, S. Boon-Chieng, and S. Mitaku. Sosui: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, 14:378–379, 1998.

[48] F. Horn, E. Bettler, L. Oliveira, F. Campagne, F.E. Cohenm, and G. Vriend. Gpcrdb information system for g protein-coupled receptors. *Nucleic Acids Res.*, 31:294–297, 2003.

[49] F. Horn, E.M. van der Wenden, L. Oliveira, A.P. IJzerman, and G. Vriend. Receptors coupling to g proteins: is there a signal behind the sequence? *Proteins*, 41:448–459, 2000.

[50] F. Horn, J. Weare, M.W. Beukers, S. Horsch, A. Bairoch, W. Chen, O. Edvardsen, F. Campagne, and G. Vriend. Gpcrdb: an information system for g protein-coupled receptors. *Nucleic. Acids. Res.*, 26:275–279, 1998.

[51] S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17:721–728, 2001.

[52] R. Hughey and A. Krogh. Hidden markov models for sequence analysis: extension and analysis of the basic method. *Comput. Appl. Biosci.*, 12:95–107, 1996.

[53] T. Jaakkola, M. Diekhans, and D. Haussler. Using the fisher kernel method to detect remote protein homologies. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, pages 149–158, 1999.

[54] A. Jagerschmidt, N. Guillaume, B.P. Roques, and F. Noble. Binding sites and transduction process of the cholecystokininb receptor: involvement of highly conserved aromatic residues of the transmembrane domains evidenced by site-directed mutagenesis. *Mol. Pharmacol.*, 53:878–885, 1998.

[55] J.A. Javitch, L. Shi, M.M. Simpson, J. Chen, V. Chiappa, I. Visiers, H. Weinstein, and J.A. Ballesteros. The fourth transmembrane segment of the dopamine d2 receptor: accessibility in the binding-site crevice and position in the transmembrane bundle. *Biochemistry*, 39:12190–12199, 2000.

[56] Y. Jiang, L. Luo, E.L. Gustafson, D. Yadav, M. Laverty, N. Murgolo, G. Vassileva, M. Zeng, T.M. Laz, J. Behan, P. Qiu, L. Wang, S. Wang, M. Bayne, J. Greene, F. Jr Monsma, and F.L. Zhang. Identification and characterization of a novel rf-amide peptide ligand for orphan g-protein-coupled receptor sp9155. *J. Biol. Chem.*, 278:27652–27657, 2003.

[57] D.T. Jones, W.R. Taylor, and J.M. Thornton. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, 33:3038–3049, 1994.

[58] D.T. Jones, W.R. Taylor, and J.M. Thornton. A mutation data matrix for transmembrane proteins. *FEBS Lett.*, 339:269–275, 1994.

[59] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[60] K. Katoh, K. Kuma, H. Toh, and T. Miyata. Mafft version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, 33:511–518, 2005.

[61] T. Kawabata. Matras: a program for protein 3d structure comparison. *Nucleic Acids Res.*, 31:3367–3369, 2003.

[62] S. Kawashima and M. Kanehisa. Aaindex: amino acid index database. *Nucleic Acids Res.*, 28:374–374, 2000.

[63] S.B. Kirton, C.A. Baxter, and M.J. Sutcliffe. Comparative modelling of cytochromes p450. *Adv. Drug Deliv. Rev.*, 54:385–406, 2002.

[64] R. Koradi, M. Billeter, and K. Wuthrich. Molmol: a program for display and analysis of macromolecular structures. *J. Mol. Graph.*, 14:51–55, 1996.

[65] N.A. Kratochwil, P. Malherbe, L. Lindemann, M. Ebeling, M.C. Hoener, A. Muhlemann, R.H. Porter, M. Stahl, and P.R. Gerber. An automated system for the analysis of g protein-coupled receptor transmembrane binding pockets: alignment, receptor-based pharmacophores, and their application. *J. Chem. Inf. Model.*, 45:1324–1336, 2005.

[66] A. Krogh, B. Larsson, G. von Heijne, and E.L. Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol.*, 305:567–580, 2001.

[67] Y. Kyogoku, Y. Fujiyoshi, I. Shimada, H. Nakamura, T. Tsukihara, H. Akutsu, T. Odahara, T. Okada, and N. Nomura. Structural genomics of membrane proteins. *Acc. Chem. Res.*, 36:199–206, 2003.

[68] J. Kyte and R.F. Doolittle. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, 157:105–132, 1982.

[69] N.E. Labrou, N. Bhogal, C.R. Hurrell, and J.B. Findlay. Interaction of met297 in the seventh transmembrane segment of the tachykinin nk2 receptor with neurokinin a. *J. Biol. Chem.*, 276:37944–37949, 2001.

[70] P. Lachenbruch and M. Mickey. Estimation of error rate in discriminant analysis. *Technometrics*, 10:1–111, 1968.

[71] C.R. Lancaster, A. Kroger, M. Auer, and H. Michel. Structure of fumarate reductase from wolinella succinogenes at 2.2 a resolution. *Nature*, 402:377–385, 1999.

[72] N.H. Lee, N.S. Geoghagen, E. Cheng, R.T. Cline, and C.M. Fraser. Alanine scanning mutagenesis of conserved arginine/lysine-arginine/lysine-x-x-arginine/lysine g protein-activating motifs on m1 muscarinic acetylcholine receptors. *Mol. Pharmacol.*, 50:140–148, 1996.

[73] M.Y. Leung, O. Al-Muslim, S.M. Wu, A. Aziz, S. Inam, M. Awadh, O.M. Rennert, and W.Y. Chan. A novel missense homozygous inactivating mutation in the fourth transmembrane helix of the luteinizing hormone receptor in leydig cell hypoplasia. *Am. J. Med. Genet. A.*, 130:146–153, 2004.

[74] Z.L. Lu, J.W. Saldanha, and E.C. Hulme. Transmembrane domains 4 and 7 of the m1 muscarinic acetylcholine receptor are critical for ligand binding and the receptor activation switch. *J. Biol. Chem.*, 276:34098–34104, 2001.

[75] D. Maglott, J. Ostell, K.D. Pruitt, and T. Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Res.*, 33:D54–58, 2005.

[76] M.A. Marti-Renom, M.S. Madhusudhan, and A. Sali. Alignment of protein sequences by their profiles. *Protein Sci.*, 13:1071–1087, 2004.

[77] D.A. Mason, J.D. Moore, S.A. Green, and S.B. Liggett. A gain-of-function polymorphism in a g-protein coupling domain of the human beta1-adrenergic receptor. *J. Biol. Chem.*, 274:12670–12674, 1999.

[78] M. Matsumoto and T. Nishimura. Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. acm trans. on modeling and computer simulation. *Protein Sci.*, 8:3–30, 1998.

[79] I.M. Meyer and R. Durbin. Comparative ab initio prediction of gene structures using pair hmms. *Bioinformatics*, 18:1309–1318, 2002.

[80] L.A. Mirny and E.I. Shakhnovich. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.*, 291:177–196, 1999.

[81] S. Mitaku, T. Hirokawa, and T. Tsuji. Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. *Bioinformatics*, 18:608–616, 2002.

[82] S. Möller, M.D. Croning, and R. Apweiler. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, 17:646–653, 2001.

[83] S. Möller, J. Vilo, and M.D. Croning. Prediction of the coupling specificity of g protein coupled receptors to their g proteins. *Bioinformatics*, 17:S174–S181, 2001.

[84] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of protein database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.

[85] E.W. Myers and W. Miller. Optimal alignments in linear space. *CABIOS*, 4:11–17, 1988.

[86] K. Nakai, A. Kidera, and M. Kanehisa. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.*, 2:93–100, 1998.

[87] S.B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453, 1970.

[88] P.C. Ng, J.G. Henikoff, and S. Henikoff. Phat: a transmembrane-specific substitution matrix. predicted hydrophobic and transmembrane. *Bioinformatics*, 16:760–766, 2000.

[89] C. Notredame, D.G. Higgins, and J. Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, 302:205–217, 2000.

[90] T. Okada, Y. Fujiyoshi, M. Silow, J. Navarro, E.M. Landau, and Y. Shichida. Functional role of internal water molecules in rhodopsin revealed by x-ray crystallography. *Proc. Natl. Acad. Sci. USA*, 99:5982–5987, 2002.

[91] Y. Ono, W. Fujibuchi, and M. Suwa. Automatic gene collection system for genome-scale overview of g-protein coupled receptors in eukaryotes. *Gene*, 2005.

[92] K. Palczewski, T. Kumasaka, T. Hori, C.A. Behnke, H. Motoshima, B. A. Fox, I. Le Trong, D.C. Teller, T. Okada, R.E. Stenkamp, M. Yamamoto, and M. Miyano. Crystal structure of rhodopsin: A g protein-coupled receptor. *Science*, 289:739–45, 2000.

[93] W.R. Pearson and D.J. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, 85:2444–2448, 1988.

[94] J.H. Perlman, L. Laakkonen, R. Osman, and M.C. Gershengorn. A model of the thyrotropin-releasing hormone (trh) receptor binding pocket. evidence for a second direct interaction between transmembrane helix 3 and trh. *J. Biol. Chem.*, 269:23383–23386, 1994.

[95] L.R. Rabiner. A tutorial on hidden markov models and selected applications in sppech recognition. *Proc. IEEE*, 77:257–286, 1989.

[96] A.R. Renzetti, R.M. Catalioto, M. Criscuoli, Cucchi P., C. Ferrer, A. Giolitti, M. Guelfi, L. Rotondaro, F.J. Warner, and C.A. Maggi. Relevance of aromatic residues in transmembrane segments v to vii for binding of peptide and nonpeptide antagonists to the human tachykinin nk(2) receptor. *J. Pharmacol. Exp. Ther.*, 290:487–495, 1999.

[97] E. Rivas and S.R. Eddy. Noncoding rna gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:8–8, 2001.

[98] G.E. Schulz. beta-barrel membrane proteins. *Curr. Opin. Struct. Biol.*, 10:443–447, 2000.

[99] N.G. Sgourakis, P.G. Bagos, P.K. Papasaikas, and S.J. Hamodrakas. A method for the prediction of gpcrs coupling specificity to g-proteins using refined profile hidden markov models. *BMC Bioinformatics*, 6:104–104, 2005.

[100] Hamodrakas SJ. Sgourakis NG, Bagos PG. Prediction of the coupling specificity of gpcrs to four families of g-proteins using hidden markov models and artificial neural networks. *Bioinformatics*, 21:4101–4106, 2005.

[101] Y. Shafrir and H.R. Guy. Stam: simple transmembrane alignment method. *Bioinformatics*, 20:758–769, 2004.

[102] I.N. Shindyalov and P.E. Bourne. Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Eng.*, 11:739–747, 1998.

[103] K. Sjolander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I.S. Mian, and D. Haussler. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homolog. *Comput. Appl. Biosci.*, 12:327–345, 1996.

[104] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *J. Mol. Biol.*, 147:195–197, 1981.

[105] E.L. Sonnhammer, G. von Heijne, and A. Krogh. A hidden markov model for predicting transmembrane helices in protein sequences. *Int. Conf. Intell. Syst. Mol. Biol.*, 6:175–182, 1998.

[106] K.R. Sreekumar, Y. Huang, M.H. Pausch, and K. Gulukota. Predicting gpcr-g-protein coupling using hidden markov models. *Bioinformatics*, 20:3490–3499, 2004.

[107] J. Stitham, K.A. Martin, and J. Hwa. The critical role of transmembrane prolines in human prostacyclin receptor activation. *Mol. Pharmacol.*, 61:1202–1210, 2002.

[108] C.D. Strader, T.M. Fong, M.R. Tota, D. Underwood, and R.A.F. Dixon. Structure and function of g protein coupled receptors. *Annu. Rev. Biochem.*, 63:101–132, 1994.

[109] J.S. Surgand, J. Rodrigo, E. Kellenberger, and D. Rognan. A chemogenomic analysis of the transmembrane binding cavity of human g-protein-coupled receptors. *Proteins*, 62:509–538, 2006.

[110] J.D. Thompson, D.G. Higgins, and T.J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence

weighting, position-specific gap penalties and weight matrix choice. *Nucleic. Acids. Res.*, 22:4673–4680, 1994.

[111] D.P. Tieleman, I.H. Shrivastava, M.R. Ulmschneider, and M.S. Sansom. Proline-induced hinges in transmembrane helices: possible roles in ion channel gating. *Proteins*, 44:63–72, 2001.

[112] K. Tomii and M. Kanehisa. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.*, 9:27–36, 1996.

[113] G.E. Tusnady, Z. Dosztanyi, and Simon I. Pdb_tm: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.*, 33:D275–278, 2005.

[114] G.E. Tusnady and I. Simon. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.*, 283:489–506, 1998.

[115] G. von Heijne. Membrane protein structure prediction. hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.*, 225:487–494, 1992.

[116] S.M. Wade, W.K. Lim, K.L. Lan, D.A. Chung, M. Nanamori, and R.R. Neubig. G(i) activator region of alpha(2a)-adrenergic receptors: distinct basic residues mediate g(i) versus g(s) activation. *Mol. Pharmacol.*, 56:1005–1013, 1999.

[117] H.L. Wang. Basic amino acids at the c-terminus of the third intracellular loop are required for the activation of phospholipase c by cholecystokinin-b receptors. *J. Neurochem.*, 68:1728–1735, 1997.

[118] H.L. Wang. A conserved arginine in the distal third intracellular loop of the mu-opioid receptor is required for g protein activation. *Mol. Pharmacol.*, 56:1005–1013, 1999.

[119] J.H. Ward. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, 58:236–244, 1963.

[120] J. Wess. Molecular biology of muscarinic acetylcholine receptors. *Crit. Rev. Neurobiol.*, 10:69–99, 1996.

[121] J. Wess. Molecular basis of receptor/g-protein-coupling selectivity. *Pharmacol. Ther.*, 80:231–264, 1998.

[122] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83, 1945.

[123] S.K. Wong. G protein selectivity is regulated by multiple intracellular regions of gpcrs. *Neurosignals*, 12:1–12, 2003.

[124] S.V. Wu, M. Yang, D. Avedian, M. Birnbaumer, and J.H. Walsh. Single amino acid substitution of serine82 to asparagine in first intracellular loop of human cholecystokinin (cck)-b receptor confers full cyclic amp responses to cck and gastrin. *Mol. Pharmacol.*, 55:795–803, 1999.

[125] T. Yada and M. Hirosawa. Detection of short protein coding regions within the cyanobacterium genome: application of the hidden markov model. *DNA Res.*, 3:355–361, 1996.

[126] Y.K. Yang, T.M. Fong, C.J. Dickinson, C. Mao, J.Y. Li, M.R. Tota, R. Mosley, L.H. Van Der Ploeg, and I. Gantz. Molecular determinants of ligand binding to the human melanocortin-4 receptor. *Biochemistry*, 39:14900–14911, 2000.

[127] V. Yankovskaya, R. Horsefield, S. Tornroth, C. Luna-Chavez, H. Miyoshi, C. Leger, B. Byrne, G. Cecchini, and S. Iwata. Architecture of succinate dehydrogenase and reactive oxygen species generation. *Science*, 299:700–704, 2003.

[128] Z. Yuan and B. Huang. Prediction of protein accessible surface areas by support vector regression. *Proteins*, 57:558–564, 2004.

[129] L. Zhang, G. Huang, J. Wu, and K.H. Ruan. A profile of the residues in the first intracellular loop critical for gs-mediated signaling of human prostacyclin receptor characterized by an integrative approach of nmr-experiment and mutagenesis. *Biochemistry*, 44:11389–11401, 2005.

[130] S. Zozulya, F. Echeverri, and T. Nguyen. The human olfactory receptor repertoire. *Genome Biol.*, 2:RESEARCH0018, 2001.

研究業績

論文

- Muramatsu, T. and Suwa, M. Statistical Analysis and Prediction of Functional Residues Effective for GPCR-G-Protein Coupling Selectivity. 2006. *Protein Engineering, Design, and Selection* (*in press*)

- Yabuki, Y., Muramatsu, T., Hirokawa, T., Mukai, H. and Suwa, M. GRIFFIN: a system for predicting GPCR-G-protein coupling selectivity using a support vector machine and a hidden Markov model. 2005. *Nucleic Acids Res.* **33**:W148-53.

査読なし国際学会

- Suwa, M., Yabuki, Y., Muramatsu, T., Hirokawa, T. and Mukai, H. GPCR and G-protein Coupling Selectivity Prediction Based on SVM with Physico-Chemical Parameters. 2004. *International Conference on Genome Informatics.* P056.

- Muramatsu, T. and Suwa, M. Sequence Alignment Tool of Membrane Protein. 2004. *International Conference on Genome Informatics.* P083.

ワークショップ・国内学会など

- 村松孝彦, 諏訪牧子. G タンパク質結合選択性に相関性のある GPCR の配列及び構造的特徴 2002. 第 40 回生物物理学会年会.

- 村松孝彦, 諏訪牧子. 位置特異的配列プロファイルを利用した GPCR の G タンパク質に対する共役特異性に関する解析. 2003. 第 41 回生物物理学会年会.

• Muramatsu, T. and Suwa, M. Bioinformatics Analysis for the Receptor-G-protein coupling Selectivity. 2004. *Frontiers in Bioinformatics: Structure, Interaction and Function.*