2007－BIO－10 (5)
2007／9／13

SLCトランスポータータンパク質の網羅的データベースの構築と基質特異的モチーフによる機能部位予測

荒川 貴行、幸 賢二郎、宮崎 智
東京理科大学大学院 薬学研究科 生命情報科学研究室

SLC(solute carrier)ファミリー（以下SLCとする）は様々な臓器に発現し、栄養素を細胞内への取り込みに関与するだけでは無く、薬物の体内動態にも深く関与しているトランスポーターである。

トランスポーターをターゲットとした薬物をデザインする場合、輸送基質を認識する部分、つまり機能部位の情報が必要不可欠となる。機能部位情報を得るためにはトランスポーターの立体構造情報が有用であるが、膜タンパク質の立体構造解析は技術的に難しいため、既知の立体構造情報が少なく、トランスポーターをターゲットとした薬物を設計することは困難となっている。そこで、本研究ではSLCをターゲットに、立体構造情報に依存しない機能部位予測法を考案した。この手法で予測した結果と実験データを併せた解析により、トランスポーターをターゲットとした創薬に貢献できると考えられる。

キーワード：SLC、機能部位、BABMS、基質特異的モチーフ、ベストモチーフ

Construction of exhaustive database of SLC transporters and prediction functional site by using substrate specific motifs

Takayuki Arakawa, Kenjiro Yuki, and Satoru Miyazaki
Department of Pharmaceutical research, Tokyo University of Science

SLC (solute carrier) family is transporter, expressing in various organ, which has close relation to not only delivering nutrient into our body but also disposition of drug.

In case of drug design targeting on transporters, it is essential to get information of the functional sites, in other words, the sites to recognize transport substrates. To get information of the functional sites, known 3D structures of transporters are useful. But, structural analysis of membrane proteins is technically difficult. It leads to the lack of 3D structures and makes it hard to design a new drug targeting on transporters.

In this study, we propose a novel strategy of predicting functional sites not dependent on structural data. This study is useful to reveal a functional site and will help transporter targeting drug discovery.

1．はじめに

近年、薬物の体内動態を制御する重要な膜タンパク質として、薬物トランスポーターの研究が盛んに行われている。薬物トランスポーターは小腸、腎臓、肝臓、血液脳関門などに発現し、生体に投与された薬物のADME（Absorption, Distribution, Metabolism, Excretion：吸収、分布、代謝、排泄）を制御している膜タンパク質の総称である。栄養物質の摂取に関わるものから異物の解毒に関わるも

のまで、多様なトランスポーターが薬物トランスポーターとして機能している。薬物トランスポーターの特徴として重要なのは基質特異性の広さである。酵素を例に挙げると、ある酵素は特定の基質にのみ作用すると考えられており、それらは鍵と鍵穴の関係として例えられる(一酵素一基質説)。一方、薬物トランスポーターでは1つのトランスポーターで様々な薬物を輸送することが知られている[1]。

この薬物トランスポーターの立体構造を解明し、作用機序や化学構造の異なる多種多様な薬剤をどのような仕組みで基質として認識しているのかを理解することにより、薬物トランスポーターに的確に認識される薬剤設計が可能となり、的確に薬効発現させることが出来ると考えられる。また、薬剤を適切な薬剤を、適切な場所に、適切な量だけ輸送するシステムであるDDS (Drug Delivery System)にも応用することが可能であると考えられる。

今回、研究のターゲットとなるSLCファミリーはグルコース、アミノ酸、カルニチンや金属イオンなどの生体内栄養素のほか、コリンやドパミンなどの神経伝達物質、さらに、β-ラクタムや非ステロイド抗炎症薬 (NSAIDs) などの外因性物質を輸送するため、薬物トランスポーターの中でも重要なファミリーである。

一般的にタンパク質の機能部位を予測する方法としては、モチーフによる機能解析がある。既存のモチーフ探索方法として、マルチプルシークエンスアラインメンを行うソフトウェアであるClustaW(X)などから、配列間で保存された領域を見出す方法がとられてきた。しかし、この方法では比較する配列群の中に長さが大きく異なる配列や、繰り返し配列が含まれる場合にうまくアラインメントできず、モチーフが抽出できなくなってしまうという欠点が挙げられる。また、SLCはその疎水性という性質の為にX線立体構造解析が困難であり、機能部位に関する情報をほとんど得ることができないのが現状である。

そこで、我々は立体構造情報に頼らない機能部位予測を実現するために、アミノ酸1次配列から機能部位を予測する方法の開発を目指している。

その為、本研究では、配列長の影響を受けないアミノ酸配列モチーフの探索アルゴリズムを開発と、これによって得られた輸送基質に特異的なモチーフ情報から機能部位を予測する方法について検討した。

2. 研究方法　—SLCの網羅的データベースの構築とSLCの基質別機能部位の予測—

SLCの機能部位を予測するため、まず始めにSLCの網羅的なデータベースを構築した。このデータベースにはヒトを含めた多くの生物種のエントリーが含まれており、今後のSLCの解析の基盤としていく。次に、SLCをこのデータベースの輸送基質ごとにグループ分けを行い、グループ毎に基質特異的モチーフを探索した。そして、探索したモチーフの機能部位予測への応用について検討した。

2.1. SLCトランスポーターの網羅的データベースの構築

2.1.1. SLCのアミノ酸配列、立体構造データの取得

今回取得するデータはSLCのアミノ酸配列データベースであるUniProt(UniProtKB/Swiss-Prot、UniProtKB/TrEMBL)[3]より取得した。UniProtでは1タンパク質1レコードとして、EMBL形式と呼ばれるファイル形式でデータを公開されている。取得方法は、DDBJのデータ検索システムである

All-round Retrieval of Sequence and Annotation（ARSA）より、「solute carrier family」をクエリとしたキーワード検索を行った。その結果、2,280 件（Swiss-Prot：653 件、TrEMBL：1,627 件）のフラットファイルデータを取得した。このデータに含まれる生物種は、全部で 263 種類である。これらのフラットファイルデータを 1 レコードにつき 1 データとしてリレーショナルデータベースとして加工・整理した。項目は、ID、アクセッション No、定義名、遺伝子名、生物種、文献情報、特徴、配列長、残基情報、膜貫通回数、PDBID である。

2.1.2 輸送基質データの取得

輸送基質データは、先ほどの UniProt エントリーの情報と、遺伝子名データベースである HUGO Gene Nomenclature Committee（HGNC）の定義名より取得し、リレーショナルデータベースとして整理した。基質データは、SLC の 356 遺伝子中 305 件取得した。残りの 51 件は基質情報が入手できないので、モチーフ探索の対象外とした。

2.2. SLC の基質特異的モチーフ探索

SLC ファミリーはグルコース、アミノ酸、カルニチンや金属イオンなどの生体内栄養素のほか、コリンやドパミンなどの神経伝達物質、さらに、β－ラクタムや非ステロイド抗炎症薬（NSAIDs）などの外因性物質を輸送する。各 SLC トランスポーターが輸送する基質は一種類ではなく多選択性ではあるが、例えば SLC22A1（OCT1）の「有機カチオン輸送」という様に、輸送基質の特徴が共通であるものが多い。このことから、特定の輸送基質を認識するサイトには輸送基質毎の特徴があるものと考えられる。　SLC ファミリーを輸送基質毎に分類し、各グループで基質特異的モチーフを探索することができれば、その基質を輸送するのに重要なサイトを予測することに繋がる。しかし、既存のモチーフ探索プログラムは、精度の高いマルチプルアラインメントを必要とするため、配列が長く、多様性のあるトランスポーターの解析には向いていない。そこで我々は、これらの問題を克服した独自のモチーフ探索プログラムによって、輸送基質ごとに分類したグループに対してモチーフ探索を行うことで、そのグループに特異的な部位を予測した。

2.2.1. モチーフ探索アルゴリズムの開発

配列長に依存せずに、計算量を抑えたアミノ酸配列モチーフ探索方法として、モチーフ探索アルゴリズムを開発した。これは DNA モチーフ探索アルゴリズムである Branch And Bound Median Search（BABMS）[4]アルゴリズムを応用し、アミノ酸の配列モチーフを探索できるように改良した。開発言語は、Perl 5.6.1 を用いた。

通常、20 種類のアミノ酸の配列パターンは長さ l のとき、20^l パターン存在するが、これら全てのパターンについてモチーフの可能性評価を行うと膨大な計算量となってしまい、現実的に計算は不可能である。そこで、すでにモチーフを探索した配列群に存在するパターンのみをモチーフ候補とすることで、結果の質を落とさずに計算量のみを減らすことができる。（図 1）

モチーフ探索方法は、次の通りである。

1．モチーフを探索したい配列群（query sequences）を用意する。
2．探索したいモチーフの長さ *l* を指定
3．モチーフ候補（pre-motif）を query sequences からピックアップする。
4．各 pre-motif 毎に作業5．、6．を行う
5．各 query sequences 毎に、その pre-motif と最も一致する場所を探し、その一致度を pre-motif と、ある query sequence 間の Hamming 距離として求める。Hamming 距離は一致度が高いほど小さな値を示す。
6．pre-motif は各 query sequence との間で Hamming 距離を求めたら、それらを全て足し合わせた値を、その pre-motif の持つスコア（TotalDistance：TD）として求める。つまり、TD が 0 である pre-motif の場合、全く同じパターンが全ての query sequence 中に出現していることになる。
7．全ての pre-motif について TD を計算し、一致度に換算して、80%以上の pre-moti-f を、基質特異的モチーフとして選出する。これは、後述の**研究方法 2.3.1.**で用いる。
8．基質特異的モチーフのうち、最も TD が小さいものをベストモチーフとして選出する。最小 TD を持つものが２つ以上ある場合、それら全てを best motif とする。best motif、後述の**研究方法 2.3.2.**で用いる。
8．探索したいモチーフの長さを変えて再度実行する。

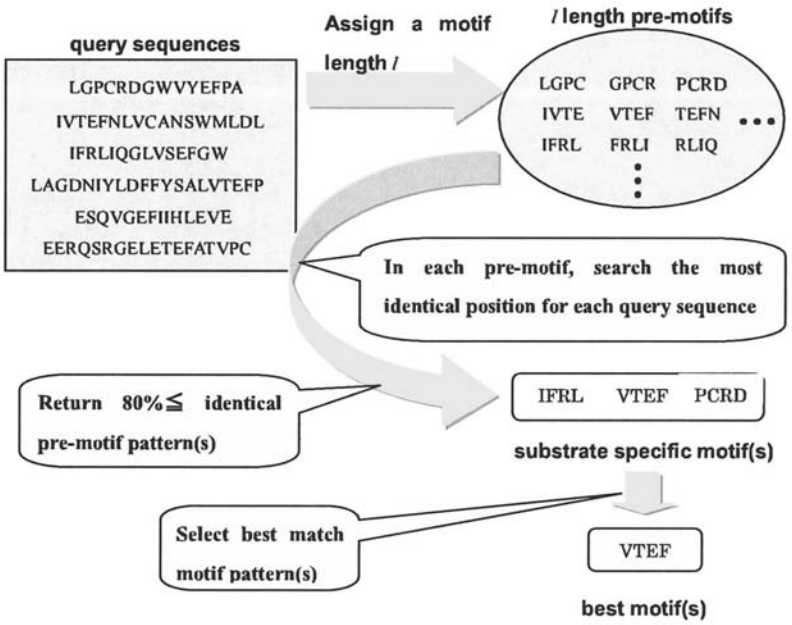

図1. モチーフ探索方法

2.2.2. SLC の基質特異的モチーフの探索

SLC の 1 次配列群を輸送基質ごとにグループ分けし、それぞれのグループに対してモチーフ探索行い、基質特異的なモチーフを探索する。探索するモチーフの長さは 3〜10 で行った。最終的には各モチーフ長で最も保存されている、つまり TD が最も小さいパターンを出力する。

2.3. 基質特異的モチーフによる機能部位予測

一般的に、配列モチーフは活性部位または活性部位付近に存在することが多く、類似の機能を持つ相同なタンパク質の間で共有されるので、その相同タンパク質ファミリーを特徴づけるものとして利用できる。そのため、共通の機能を持ち、同じファミリー内に存在するタンパク質でよく保存されているモチーフは、そのタンパク質の機能発現にとって重要な配列であると考えられる。そこで、輸送基質ごとに探索したモチーフの持つ情報をタンパク質の機能情報に適応することが出来ないか考えた。その方法として、モチーフの出現頻度や、出現位置、進化的関係から、SLC の機能に重要と思われる部位を予測することを試みた。今回は、有機カチオンを輸送することで知られる SLC22A1-3 の配列グループで解析を行った。

2.3.1. 一致度 80%以上のモチーフによる機能部位予測

探索したモチーフの持つ情報が機能部位に関連しているかどうかを検証するため、モチーフを全長配列上にマッピングし、頻度分布を作成した。マッピングするモチーフの基準は、モチーフ長が 3-10 残基で、一致度の平均値が 80%以上とし、これらのモチーフをモデル配列上にマッピングした。モデル配列は、query sequences 中に含まれるヒトの配列とした。今回はヒトの SLC22A1(554aa、UniProt ID：O15245)のアミノ酸配列上にマッピングした。このモチーフ頻度分布を他の論文で報告されている実験データと比較することで、モチーフと機能部位の関連性を検証した。

2.3.2. ベストモチーフの Evolutionary Trace[5]法を利用した比較

探索されたモチーフに進化的保存性と多様性を考慮して機能部位関連性を考察する方法を提案する。選出されたベストモチーフを各生物の query sequence にマッピングし、生物の進化系統関係樹と重ね合わせて比較する。マッピングは各モチーフ長ごとに行い、ヒトと近縁種のものから順に並べて進化的に辿る事で、それらのモチーフが基質特異的モチーフの特徴を見る。

3. 結果と考察

3.1 モチーフの探索結果

SLC22A1-3（11 query sequences）に対してモチーフを探索した結果、上の表の通りとなった。11本の query sequence から、長さ7の pre-motif は 3918 パターン存在した。通常、長さ7のアミノ酸配列のモチーフは $20^7=1.28\times10^9$ パターン存在するので 3.27×10^5 倍計算効率が良い。

	TD	motif	position
	13	LFLLYYW	274 275 310 275 254 275 280 275 275 257 182:271:292:298:301:541
	13	VGIVFLG	37 37 32:159:276 38 17 37 37 37 37 144 266:387
	12	LDLVRTP	335 336 18 336 315 336 339 334 334 322 71:164:201:247:311:499:508:541
	11	VPESPRW	282 283 276 283 262 283 288 283 283 265 305
	9	ESPRWLI	284 285 122 285 264 285 290 285 285 267 307
➡	7	PESPRWL	283 284 121 284 263 284 289 284 284 266 306

表1. SLC22A1-3 (11 query sequences)の一致度 80%以上のモチーフ探索結果 （モチーフ長 7、一部抜粋）
position の数字は、各 query sequence 上でのモチーフの開始位置を表す。position が複数あるものは x:y というように記してある。ベストモチーフは、黒矢印の行。

3.2. 一致度 80%以上のモチーフによる機能部位予測

SLC22A1～3 の各生物種のデータから、頻度分布を作成した（図2）。頻度が相対的に大きい領域のうち、丸枠で囲われた領域は 283PESPRWL289 であるが、輸送活性について報告されている実験データを見てみると、Inui.K らの研究によって Pro283Leu と Arg287Gly（図1の下線付き残基）の置換により、SLC22A1 の基質輸送活性が消失することが分かっている[6]。これらの事から、今回の基質特異的モチーフ探索による機能部位予測が有用であると考えられる。

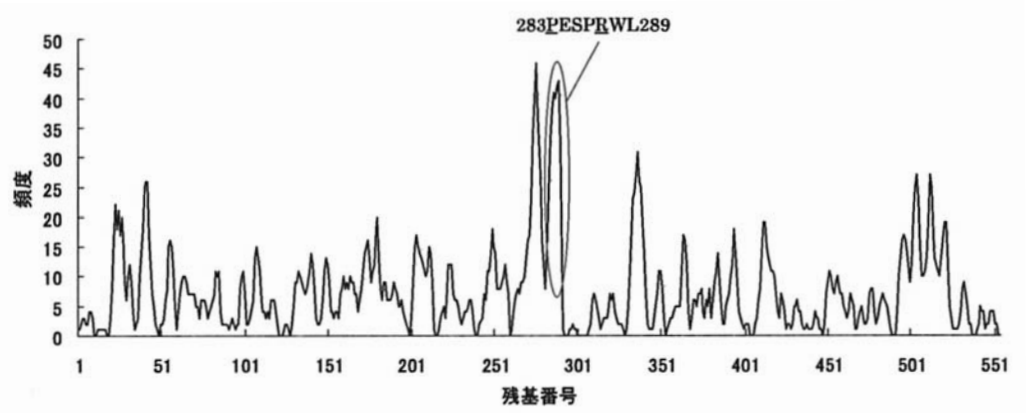

図2. モチーフの頻度分布
x 軸はモデル配列(Human SLC22A1)の全長配列、y 軸はマッピングの頻度を表す。

3.3. ベストモチーフの Evolutionary Trace 法を利用した比較

SLC22A1-3 の各生物種のモチーフ探索で得られたベストモチーフパターン PESPRWL を、SLC22A1-3 の各配列上にマッピングした（図３）。このパターンは各配列でほぼ同じ位置に存在し、１箇所のみに現れていることが分かる。実際に各生物種の配列パターンを見てみると、哺乳類のモチーフパターンはよく保存されているのに対し、魚類（SLC22A3_Brachydanio rerio）と、植物（SLC22A3_Oryza sativa）のパターンには多様性が見られた。

また、SLC22A サブファミリーで同じようにモチーフ探索を行った結果、SLC22A1-3 と異なるパターンがベストモチーフとして選ばれた（図４）。SLC22 には大きく分けてアニオン輸送、カチオン輸送、両性イオン輸送のものからなるが、このモチーフは SLC22 サブファミリーを特徴付けるモチーフになり得ると考えられる。今後、詳しい検証を行いたいと思う。

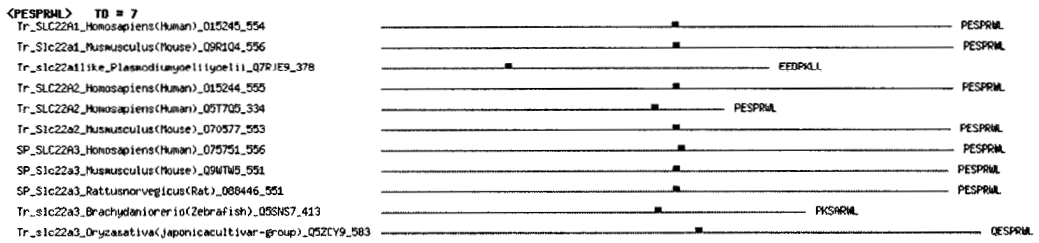

図3. ベストモチーフを SLC22A1-3 (11 query sequences)の配列上にマッピングしたドットマップ
左上の太字は左から、ベストモチーフ、TD。左側の各名称は、データベース名_遺伝子名_生物種_UniProtID_配列長を表す。塗りつぶされた BOX がモチーフの出現位置と長さを、右端の文字列は各 query sequence の実際のモチーフパターンを表す。

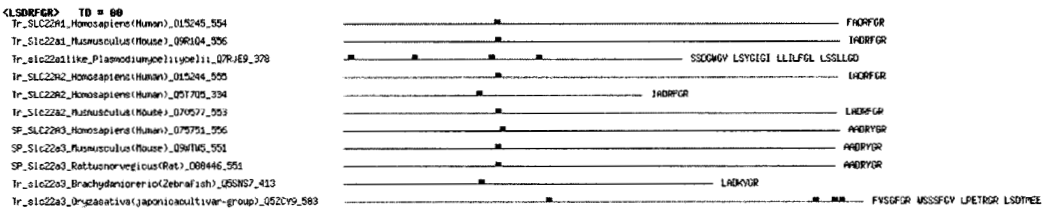

図4. SLC22A1-21(59 query sequence)でベストモチーフを探索したドットマップ （SLC22A4-21 は省略）

4. 終わりに

今回の研究は、SLC の網羅的なデータベースを構築し、そこから基質特異的モチーフを探索し、モチーフの持つ情報から機能部位を予測した。

基質特異的モチーフの探索に限らず、バイオインフォマティクスの解析において非常に重要なのが、データベースの質である。データが間違っていたり、ある情報に偏っていれば、解析手法が正しくてもよい結果が得られない可能性がある。特に、アノテーションがしっかりしていないデータを扱う場合、データ取得方法は適切か、関係ないデータは含まれていないか、重複データを含んだまま解析していないか、などを考える必要がある。

モチーフの頻度分布作成では、今回はモチーフの閾値を経験的に適当と思われる一致度 80%以上と設定している。今後は閾値を経験的な値ではなく、何らかの仮説や統計値に基づいた上で設定し、本手法の予測と既知情報との比較によって、予測精度について検証していきたい。

また、予測された機能部位が仮に正しかったとしたら、それがどんな機能を担っているか（基質センサー、基質を引き寄せる、基質を輸送する、ハッチのような役割、など）を特徴づけすることも重要な課題である。

SLC ファミリーは今後もファミリーのメンバー数が増加していくと思われるが、今回の基質特異的モチーフ情報を、各輸送基質グループを表す特徴量として捉え、マシンラーニング手法により、機能未知トランスポーターの機能の特徴付けができると考えられる。

今回は配列から基質特異的モチーフを予測したが、さらに、基質特異的な残基の予測を行うことが出来るようにしていきたい。

5. 謝辞

本研究を遂行するにあたり、熱心にご指導くださいました、東京理科大学薬学部　生命情報科学研究室の宮崎智教授ならびに鈴木智典助教に謹んで感謝申し上げます。

また、研究で世話になりました、東京理科大学薬学部　生命情報科学研究室の皆様方に心より感謝申し上げます。

6. 参考文献

1) 村上聡, 蛋白質 核酸 酵素　Vol.52, No.5, 406-414 (2007)
2) 藤博幸, 「はじめてのバイオインフォマティクス」, 講談社サイエンティフィック, 40-41 (2007)
3) The UniProt Consortium, Nucleic Acid Res., Vol.35, Database issue, D193-197 (2007)
4) Neil C. Jones, Pavel Pevzner, An Introduction To Bioinformatics Algorithms, Bradford Books (2004)
5) O.Lichtarge et al., J. Mol. Biol., 257, 342-358 (1996)
6) Inui K et al., Drug Metab. Pharmacokin., 18(6): 409-412 (2003)