Identification of Glycosyltransferases Focusing on Golgi Transmembrane Region

ゴルジ膜貫通領域に着目した糖転移酵素の判別

Mukai, Yuri¹; Hirokawa, Takatsugu¹; Tomii, Kentaro¹; Asai, Kiyoshi^{1,2}; Akiyama, Yutaka¹; and Suwa, Makiko^{1,*}¹Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST),
2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan²Department of Computational Biology, Graduate School of Frontier science, The University of Tokyo, 1-5-1 Kashiwanoha,
Kashiwa-shi, Chiba 277-8562, Japan

FAX:81-3-3599-8081, E-mail: m-suwa@aist.go.jp

*Correspondence should be addressed.

Key words: Golgi-localized signal-anchor-type glycosyltransferase, plasma-membrane-localized signal-anchor-type membrane protein, protein with signal peptide, hydropathy profile, position-specific score matrix

Abstract

To uncover novel glycosyltransferase (GT) genes from the entire human genome, we have developed a computational method for determining the transmembrane region of Golgi-localized signal-anchor-type GT. The characteristics of GT were extracted by comparison with those of the plasma-membrane-localized signal-anchor-type membrane protein (PM) and the protein with signal peptide (SP), whose topologies are similar to that of GT. The Golgi-membrane-spanning region of GT was distinguished from PMs and SPs with sensitivity and specificity exceeding 95% using the position-specific score matrix (PSSM) of amino acid propensity based on hydropathy alignment. Using our algorithm for determining the Golgi-membrane-spanning region of GTs, an improvement in the efficiency of GT gene detection is expected.

要 約

ヒトゲノムからの糖転移酵素遺伝子の網羅的発見を目指し、我々は糖転移酵素のほとんどを占めるゴルジ膜局在シグナルアンカー型の糖転移酵素 (GT) における膜貫通領域を判別する計算機的手法の開発を行った。糖転移酵素のアミノ酸配列が持つ物理化学的性質を、膜存在形態の似た細胞膜局在シグナルアンカー型膜タンパク質 (PM) やシグナルペプチド (SP) を持つタンパク質と比較した結果、糖転移酵素の特徴を抽出することに成功した。疎水性アラインメント位置におけるアミノ酸出現頻度から求めた位置特異的スコアマトリックスから、GT のゴルジ膜貫通領域は 95% 以上の感度と選択性で PM や SP から高精度に判別することができた。糖転移酵素のゴルジ膜貫通領域判別アルゴリズムを用いることにより、糖転移酵素遺伝子発見の効率化が期待される。

A. Significance and Approach to Discovering Novel Glycosyltransferases Using Bioinformatics

Most proteins contain oligosaccharides that are modified by glycosyltransferase (GT). The variation in oligosaccharide modification depends on the localization and type of GT. Therefore, the identification and classification of GTs are indispensable for the clarification of the oligosaccharide modification mechanism. To this end, the development of a computational method for discovering novel GTs from the human genome is desired.

The sequence similarity search is a general method for finding a useful gene from open reading frames (ORFs) in human genome. According to CAZy (<http://afmb.cnrs-mrs.fr/CAZY/>), known human GTs comprise 42 families based on sequence similarity classification, although there is a low similarity among the 42 families. Even if the sequence similarity search is conducted for a known GT, only known GTs with high similarities are found, and novel ones are hardly

A. 糖転移酵素遺伝子発見の意義とバイオインフォマティクスによるアプローチ

生体内のほとんどのタンパク質は、糖転移酵素による糖鎖修飾を受けており、糖鎖修飾のパリエーションは糖転移酵素の局在位置や糖転移酵素の種類に依存する。したがって、糖鎖修飾メカニズムの解明には、糖転移酵素の同定や分類が必須である。そのため、膨大なヒトゲノム情報に潜む未知の糖転移酵素遺伝子を計算機により見つけ出す方法の開発が求められている。

ヒトゲノム ORF から有用な遺伝子を見つける一般的な方法には配列類似性検索があるが、糖関連酵素データベース CAZy (<http://afmb.cnrs-mrs.fr/CAZY/>) によると、配列類似性で分類したヒトの既知糖転移酵素は 42 ファミリーにまたがり、ファミリー間の配列類似性は低い。既知の糖転移酵素をもとに配列類似性検索を行っても、配列類似性の高い既知の糖転移酵素しか見出すことができず、他のファミリーにたどりつ

Fig. 1. Structure model of Golgi-membrane-localized signal-anchor-type glycosyltransferase (GT).

found. In this regard, the development of a GT detection method that does not rely on sequence similarity is required for identifying new GTs.

In Fig. 1, a Golgi-localized signal-anchor-type GT, which is the most common of the GTs known to date, has one transmembrane helix near its N-terminus, which spans the Golgi membrane, and a functional domain on its C-terminus, which is found in aqueous environments as “nucleotide-diphospho-sugar transferases” or “P-loop-containing nucleoside triphosphate hydrolases” (1-3).

From the Golgi membrane localization characteristic, ORFs with Golgi membrane localization are treated as candidates for GT on the basis of a subcellular localization prediction program. Many subcellular localization prediction systems are available; NNPSL (4) and SubLoc (5) was developed using artificial neural network (ANN) and support vector machine (SVM), respectively, on the basis of composition of amino acids. PSORT II (6) uses k-nearest neighbor method on the basis of amino acid sequence and organisms, iPSORT (7) is for detection of sorting signal in N-terminus such as signal-peptide, mitochondrion or chloroplast targeting sequence, and PLOC (8) predicts by amino acid and its pair composition. However, they are weak predictors of Golgi membrane localization because they have few membrane proteins in their training data sets and there is no clear motif of Golgi transport.

Therefore, we extracted the characteristics of the human Golgi-membrane-spanning region of GT and developed an original computational method for detecting the GT transmembrane region.

けない。ましてや新規の発見は非常に困難である。以上のような背景から、新規糖転移酵素を同定するためには、配列類似性を用いない糖転移酵素判別技術の開発が必要となる。

糖転移酵素のほとんどを占めているゴルジ膜局在シグナルアンカー型糖転移酵素（以下 GT とする）では、図 1 に示した模式図のように、ゴルジ膜を貫通するヘリックスを N 末端近傍に 1 本だけ持ち、C 末端側は “Nucleotide-diphospho-sugar transferases” や “P-loop containing nucleoside triphosphate hydrolases” に代表される機能ドメインが水溶性環境の中に存在することが知られている (1-3)。

ゴルジ膜局在という特徴から、細胞内局在予測プログラムでゴルジ体局在と予測される ORF を糖転移酵素の候補とする方法も考えられる。細胞内局在予測プログラムには、アミノ酸組成を基にアーティフィシャルニューラルネットワーク (ANN) を用いた NNPSL (4) やサポートベクターマシン (SVM) を用いる SubLoc (5)、アミノ酸配列と生物種の情報から k 最近接法を用いて予測を行う PSORT II (6)、シグナルペプチドやミトコンドリアや葉緑体への移行シグナルを検出する iPSORT (7)、アミノ酸組成やアミノ酸ペアの構成比に基づいて予測を行う PLOC (8) などがある。しかし、これら既存のシステムでは、膜タンパク質がデータセットから排除されていること、局在化モチーフがないゴルジ体の局在予測が不得手であることなどから、現存するシステムではゴルジ膜局在膜タンパク質の予測は困難であった。

そのような理由から、我々は GT からヒトのゴルジ膜貫通領域の特徴を抽出し、独自のヒト・ゴルジ膜局在判別方法の開発を行った。

B. Extraction of Golgi-membrane-Spanning Region Characteristics and Identification of Golgi-Localized Signal-Anchor-Type GT

We compared the characteristics of GT with those of plasma-membrane-localized signal-anchor-type membrane protein (PM) and protein with signal peptide (SP), whose topologies are similar to that of GT, and extracted the characteristics of the Golgi-membrane-spanning region of GT. All the data sets were from SWISS-PROT 47, TrEMBL, and Genbank, and contain sequences with less than 20% identity to each other. We used 58 entries of human GT data sets and two human protein data sets (PM, 46 entries; SP, 388 entries) as negative controls. Fig. 2 shows the flowchart of characteristics extraction for identification of Golgi-membrane-spanning region and Golgi-localized signal-anchor-type GT described in the paragraphs below.

B. ゴルジ膜貫通領域の特徴抽出とゴルジ膜局在シグナルアンカー型糖転移酵素の判別

我々は、糖転移酵素の構造的特徴と類似した細胞膜局在シグナルアンカー型膜タンパク質（以下 PM とする）やシグナルペプチドを持つタンパク質（以下 SP とする）との比較を行い、GT のゴルジ膜貫通領域の特徴抽出を試みた。データセットはすべて SWISS-PROT 47, TrEMBL, Genbank に含まれるヒト既知タンパク質のアミノ酸配列を用い、お互いに 20% 以上の類似性を持つ配列は含んでいない。データセットは GT : 58 エントリー、ネガティブコントロールとして 2 種類のタンパク質（PM : 46 エントリー、SP : 388 エントリー）を用いている。以下で述べる GT の特徴抽出方法と判別アルゴリズム作成手順を図 2 に示した。

Fig. 2. Flowchart of characteristics extraction for identification of Golgi-membrane-spanning region and Golgi-localized signal-anchor-type GT.

The hydropathy profiles of each protein were estimated by average hydropathy calculation with the moving-average method slide window (Fig. 2-(1)). Fig. 3 shows a representative example. <tm> or <sp> indicates the transmembrane region and the signal peptide region annotated in SWISS-PROT, respectively. The position with the highest average hydropathy in the region spanning 100 a.a. from the N-terminus was considered to be the transmembrane region of GT. In addition, each sequence was aligned by superpositioning the most hydrophobic point in the area (Fig. 2-(2)). On calculating the average of the hydropathy profile of each subset based on the alignment position, the difference between three subsets was determined; the hydropathy profile of PM shows the highest peak and the largest width in the area where residues -15 to +20, around the standard point, and these characteristics of GT, SP decrease in order (Fig. 4).

Next, we picked up and analyzed amino acid propensity to the alignment position in the region of -20 to +20 a.a.

Fig. 3. Average hydropathy plots of the region spanning 100 a.a. from the N-terminus by window size 11 of (a) Golgi-membrane-localized signal-anchor-type glycosyltransferase, GT (SWISS-PROT ID: FUT3_HUMAN), (b) plasma-membrane-localized signal-anchor-type membrane protein, PM (AMPE_HUMAN) and (c) the protein with signal peptide, SP (WNT1_HUMAN). The arrows with <tm> or <sp> indicate the transmembrane or the signal peptide region, respectively.

まず移動平均法で配列に沿ってある残基幅ごとに平均疎水性値を算出し、疎水性プロファイルを求めた (図 2-(1))。代表例を図 3 に示す。図 3 で <tm> や <sp> と示された領域はそれぞれ SWISS-PROT における膜貫通部分とシグナルペプチドのアノテーション領域を示している。N 末端から 100 残基以内の領域で最も疎水性の高い位置は、GT の膜貫通領域の範囲に必ず含まれていた。さらに、GT の膜貫通領域がすべて含まれる N 末端から 100 残基以内の領域の配列を、平均疎水性値が最も高い位置を基準点にしてアラインメントを行った (図 2-(2))。アラインメント位置による疎水性値の平均をサブセットごとに算出したところ、疎水性プロファイルに差異が見られた (図 4)。基準点前後 -15 ~ +20 残基における疎水性プロファイルは PM で最も高く幅広い山を示し、次いで GT、SP の順であった。

次に、疎水性アラインメントの基準点から前後 20 残基

Fig. 4. Average hydropathy profiles (window size 19) of all entries of GT (thick line), PM (small dotted line) and SP highest average hydropathy.

through the region that included the transmembrane region or the signal peptide regions of GT, PM and SP (Fig. 2-(3)). On the basis of the propensity of each amino acid shown in Fig. 2-(4), the position-specific score matrix (PSSM, shown in Fig. 2-(5)) was calculated to the alignment position using the proportional mode of pseudo-count (9) in order to avoid setting a denominator to 0, and the total score of each subset was estimated using PSSM (Fig. 2-(6)). According to the distribution of total score (Fig. 5), the detected region was set on the basis of the Mahalanobis distance and identification accuracy was evaluated in terms of sensitivity and specificity (Table I). GT could be distinguished from all the data sets, including those of the negative controls with a high accuracy, with both sensitivity and specificity exceeding 95%.

C. GT Gene Detection from Entire Human Genome

Our algorithm for detecting the transmembrane region of GT can efficiently distinguish GT from proteins of the signal-anchor or signal peptide type with similar topology. Yuan and Teasdale have reported an identification program

を膜貫通領域（あるいはシグナルペプチド領域）が必ず含まれる配列領域と考えてこの領域のアミノ酸配列を抽出し（図 2-(3)）、GT、PM、SP それぞれについて、アミノ酸出現頻度解析を行った（図 2-(4)）。アライメント位置特異的に算出されたおのおののアミノ酸の出現頻度から位置特異的スコアマトリックス (PSSM, Position Specific Score Matrix) を求め（図 2-(5)）、PSSM を用いて各データセットの積算スコアを算出した（図 2-(6)）。その際、スードカウント (9) を導入してスコア算出式の分母が 0 になるのを避けた。積算スコアの分布（図 5）から、判別領域をマハラノビス半径により設定し、判別精度を感度と選択性で評価した（表 1）。ネガティブコントロールを含む全データセットの中から、GT を感度、選択性ともに 95% 以上の精度で判別することができた。

C. ヒトゲノムからの糖転移酵素遺伝子の網羅的発見に向けて

我々が提案する GT の膜貫通領域判別アルゴリズムは、膜存在形態の似た PM や SP から GT を効率よく判別することが

Fig. 5. Total score distribution as determined by PSSM of GTs (gray circles), PMs (open squares) and SPs (open triangles). The X-axis shows the score obtained with PSSM from the amino acid propensity of GT and PM, and the Y-axis shows that obtained with PSSM from GT and SP.

Table. I. Assignment accuracy of proposed algorithm.

Actual (entry)	Predicted result				
	GT	PM	SP	Sens. (%)	Spec. (%)
GT (58)	56	1	1	96.6	100
PM (46)	0	46	0	100	97.9
SP (388)	0	0	388	100	99.7

for Golgi-localized proteins from other signal-anchor-type proteins (10). Compared with their algorithm, ours is thought to be more suitable for identifying GTs from human genome ORFs, because we focused on human GTs and we can clearly distinguish GTs from SPs.

Actually, the identification of the GT gene from the entire human genome follows several stages (Fig. 6). First, ORFs are predicted by GeneDecoder (11), GeneScan (12) and other gene-finding tools (13, 14) or functional unknown ORFs found by the human cDNA project, and GTs and other proteins with similar topology are selected on the basis of having one transmembrane in the neighborhood of the N-terminus. Some transmembrane region prediction programs are available, including SOSUI (15), ConPred II (16), TMHMM (17), and HMMTOP (18). However, careful examination is necessary because some of them may mistake a signal peptide for a transmembrane, while others are trained by the data sets of only plasma-membrane-localized proteins. Second, our detection system distinguishes the candidate ORFs of GT after the selection of proteins with similar topology using the transmembrane region prediction program. Finally, we confirm whether such soluble domain as "nucleotide-diphospho-sugar transferases" or "P-loop-containing nucleoside triphosphate hydrolases" at the C-terminus of GT exist using such fold recognition tools as Bioinfo.PL MetaServer (19), thereby decreasing the possibility of an incorrect prediction of

できる。シグナルアンカー型膜タンパク質のゴルジ局在予測ツールはYuanとTeasdaleからも以前に報告がある(10)が、我々は対象をヒトのGTに絞っていること、SPとの判別も強化していることなどの観点から、ヒトゲノムからの糖転移酵素遺伝子の発見には我々の方法がより適切であると思われる。

しかし、実際にヒトの全ゲノムから糖転移酵素を網羅的に発見する場合、いくつかの段階(図6)が必要になる。まず、GeneDecoder(11)、GeneScan(12)や他の遺伝子予測ツール(13, 14)で予測した新規遺伝子や、ヒトcDNAプロジェクトで見出された機能未知ORFから、膜貫通領域1本をN末端近傍に持つORFを選択し、糖転移酵素と膜存在形態の似たタンパク質群に絞り込む。膜貫通領域予測プログラムにはSOSUI(15)、ConPred II(16)、TMHMM(17)、HMMTOP(18)などがあるが、膜貫通領域予測の際にシグナルペプチド領域を膜貫通領域1本分と誤予測するケースや、細胞膜局在の膜タンパク質のみでトレーニングされたアルゴリズムであるケースなどがあり、糖転移酵素遺伝子の発見に適したプログラムの慎重な検討を要する。膜貫通領域判別プログラムで糖転移酵素と膜存在形態の似たタンパク質を絞り込むと、いよいよ我々の判別システムの出番となり、糖転移酵素ORFの候補が挙がる。最終段階では、Bioinfo.PL MetaServer(19)に代表されるフォールド認識法を用いて糖転移酵素に特徴的な"Nucleotide-diphospho-

Fig. 6. Strategy for detecting Golgi-membrane-localized signal-anchor-type glycosyltransferase from entire human genome.

candidates for the GT gene. 9% of human Golgi-localized signal-anchor-type proteins exclude GT such as mannosidase or UDP-glucuronate decarboxylase in SWISS-PROT. These are predicted GT by using only the algorithm described in section (B). The last step including fold recognition in Fig. 6 can identify GT from other Golgi-localized signal-anchor-type proteins. The functional analysis and classification of known and novel GTs are advancing at a rapid pace, and are expected to contribute to furthering our understanding of the mechanism of oligosaccharide modification, which is dependent on the function of many proteins.

Acknowledgements

We thank Professor Hisashi Narimatsu, Dr. Yeon-Dae Kwon and Mr. Norihiro Kikuchi (RGC, AIST) for providing GTs data sets and helpful discussions.

sugar transferases”や“P-loop containing nucleoside triphosphate hydrolases”などの水溶性ドメインを持つかどうかを確認し、糖転移酵素遺伝子候補の誤予測を低減することができる。SWISS-PROTでは、ヒトのゴルジ局在II型膜タンパク質のうち9%がmannosidaseやUDP-glucuronate decarboxylaseなど糖転移酵素以外のタンパク質に相当し、膜貫通領域判別アルゴリズムだけではこれらも同時に判別されてしまうが、この段階でゴルジ局在II型膜タンパク質を糖転移酵素とそれ以外とに区別することができる。こうして集められた糖転移酵素の機能解析や分類が精力的に進められることにより、多くのタンパク質の機能を司る糖鎖修飾のメカニズムへの理解につながる事が期待される。

References

1. Breton, C., Mucha, J., and Jeanneau, C. (2001) *Biochimie*, **83**, 713–718
2. Vrielink, A., Ruger, W., Driessen, H. P., and Freemont, P. S. (1994) *EMBO J.* **13**, 3413–3422
3. Morera, S., Imberty, A., Aschke-Sonnenborn, U., Ruger, W., and Freemont, P. S. (1999) *J. Mol. Biol.* **292**, 717–730
4. Reinhardt, A., and Hubbard, T. (1998) *Nucl. Acids Res.* **26**, 2230–2236
5. Hua, S., and Sun, Z. (2001) *Bioinformatics* **17**, 721–728
6. Nakai, K., and Horton, P. (1999) *Trends. Biochem. Sci.* **24**, 34–36
7. Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S. (2002) *Bioinformatics* **18**, 298–305
8. Park, K. J., and Kanehisa, M. (2003) *Bioinformatics* **19**, 1656–1663
9. Claverie, J. M., and Audic, S. (1996) *Comput. Appl. Biosci.* **12**, 431–439
10. Yuan, Z., and Teasdale, R. D. (2002) *Bioinformatics* **18**, 1109–1115
11. Asai, K., Itou, K., Ueno, Y., and Yada, T. (1998) *Pac Symp. Biocomput.* 228–239
12. Burge, C., and Karlin, S. (1997) *J. Mol. Biol.* **268**, 78–94
13. Solovyev, V., and Salamov, A. (1997) *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5**, 294–02
14. Yada, T., Takagi, T., Totoki, Y., Sakaki, Y., and Takaeda, Y. (2003) *Pac. Symp. Biocomput.* 375–387
15. Hirokawa, T., Boon-Chieng, S., and Mitaku, S. (1998) *Bioinformatics* **14**, 378–379
16. Ikeda, M., Arai, M., Lao, D. M., and Shimizu, T. (2002) *In Silico Biol.* **2**, 19–33
17. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001) *J. Mol. Biol.* **305**, 567–580
18. Tusnady, G. E., and Simon, I. (2001) *Bioinformatics* **17**, 849–850
19. Ginalski, A., Elofsson, D., Fischer, L., and Rychlewski, L. (2003) *Bioinformatics* **19**, 1015–1018

Received on May 30, 2006, accepted on September 5, 2006

Profile of the Author

Yuri Mukai received her B.Sc. degree in 1995 and M.Sc. degree in 1997, both from Tokyo University of Agriculture and Technology, and obtained her Ph.D. degree in pharmaceutical science from Hokkaido University in 2000. Currently she is working as an AIST research staff at the Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST). Her main research interests are focusing on the structure and function of membrane proteins.