BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 16 2008, pages 1779–1786
doi:10.1093/bioinformatics/btn309

Sequence analysis

Predicting sub-Golgi localization of type II membrane proteins

A. D. J. van Dijk[1], D. Bosch[2], C. J. F. ter Braak[3], A. R. van der Krol[2] and R. C. H. J. van Ham[1,*]

[1]Applied Bioinformatics, [2]Metabolic Regulation, PRI, Wageningen UR, Droevendaalsesteeg 1, 6708 PB Wageningen and [3]Biometris, PRI, Wageningen UR, Bornsesteeg 47, 6708 PD Wageningen, The Netherlands

Received on March 28, 2008; revised on June 9, 2008; accepted on June 11, 2008

Advance Access publication June 18, 2008

Associate Editor: Burkhard Rhost

ABSTRACT

Motivation: Recent research underlines the importance of fine-grained knowledge on protein localization. In particular, sub-compartmental localization in the Golgi apparatus is important, for example, for the order of reactions performed in glycosylation pathways or the sorting functions of SNAREs, but is currently poorly understood.

Results: We assemble a dataset of type II transmembrane proteins with experimentally determined sub-Golgi localizations and use this information to develop a predictor based on the transmembrane domain of these proteins, making use of a dedicated protein-structure based kernel in an SVM. Various applications demonstrate the power of our approach. In particular, comparison with a large set of glycan structures illustrates the applicability of our predictions on a 'glycomic' scale and demonstrates a significant correlation between sub-Golgi localization and the ordering of different steps in glycan biosynthesis.

Contact: roeland.vanham@wur.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

A crucial aspect in protein functioning is correct localization within the cell, and a variety of mechanisms, often based on special targeting sequences, exist to ensure proper delivery of proteins. The initial development of predictors such as PSORT I (Nakai and Kanehisa, 1991) has resulted in a wide array of methods to predict protein sub-cellular compartment localization on the basis of sequence information (Emanuelsson, 2002; Gardy and Brinkman, 2006; Schneider and Fechner, 2004). Although the usefulness of general localization predictors is without doubt, fine-grained models of cellular processes require insight into sub-compartmental localization of proteins, e.g. in plasma membrane microdomains (Buist *et al.*, 2006) or sub-nuclear compartments (Lei and Dai, 2005).

The biologically important secretory pathway consists of among others ER and Golgi, the latter of which is compartmentalized into *cis*-, *medial*- and *trans*-Golgi and *Trans Golgi Network* (*TGN*). Sub-Golgi localization is important for various processes. In particular, *N*-glycans influence protein conformation, stability and biological activity (Lehle *et al.*, 2006) and the sequence of additions and trimmings of *N*-glycans to and from glycoproteins is governed by the sub-Golgi location of glycosyltransferases and glycosidases (Colley, 1997). Most of these enzymes are type II transmembrane proteins, which contain a single pass transmembrane domain (TMD); in addition, they consist of a cytoplasmic N-terminus and a luminal stem domain which links the TMD to the luminal C-terminal catalytic domain (Breton *et al.*, 2006). Various regions have been implicated specifically as localization signal sequence, depending on the enzyme being investigated: the length of the TMD (Saint-Jore-Dupas *et al.*, 2006), the cytoplasmic N-terminal domain (Zerfaoui *et al.*, 2002), and the combined cytoplasmic, transmembrane and stem (CTS) regions (Grabenhorst and Conradt, 1999). Experimental exchange of the CTS domains has been shown to result in altered enzyme locations with which the glycosidase assembly line could be redesigned for the production of specific glycan structures (Czlapinski and Bertozzi, 2006).

A second type of biologically important type II proteins for which sub-Golgi localization is highly relevant are Golgi SNAREs, which are involved in vesicle trafficking; the localization of t-SNAREs determines which vesicles (containing corresponding v-SNAREs) can dock (Puthenveedu and Linstedt, 2005). Experimental localization of these and other proteins in the Golgi is relatively easy but sub-Golgi localization requires complicated experimental approaches and consequently, many proteins are annotated as Golgi-localized without known sub-Golgi localization.

Different mechanisms have been proposed for Golgi and post-Golgi localization: (1) the TMD length can play a role according to the bilayer thickness model for Golgi retention, which proposes that the shorter TMD of Golgi proteins prevents them from entering cholesterol-rich transport vesicles destined for the plasma membrane (Webb *et al.*, 1998); (2) alternatively, the formation of oligomers within the Golgi may prevent protein movement into transport vesicles (Colley, 1997). A predictor has been previously developed for Golgi versus post-Golgi localization, based on hydrophobicity and frequency of different residues within the TMD of type II membrane proteins (Yuan and Teasdale, 2002). Here we develop a predictor specifically for the sub-compartments within the Golgi part of the secretion pathway: *cis*-, *medial*- and *trans*-Golgi and the TGN. It is known that the composition of ER, Golgi and plasma-membranes differs (Mitra *et al.*, 2004; van Meer *et al.*, 2008) and likewise the membrane composition of the different sub-Golgi compartments may vary; we therefore reasoned that the TMD of type II membrane proteins may contain signatures that help identify

*To whom correspondence should be addressed.

their preferred location, even though these TMD characteristics themselves may not be the single driving force that targets these proteins to their location.

Our approach takes the following steps: (1) assembly of an exhaustive dataset of experimentally determined sub-Golgi localization for type II transmembrane proteins; (2) development of a predictor based on this dataset; we specifically assess whether the TMD contains enough information to develop such a predictor; and (3) application of the predictor to several glycosylation-related enzymes and SNAREs. Because the available amount of type II transmembrane proteins with known sub-Golgi localization is small, we followed an innovative approach where we use additional homologous sequences in the training set of our predictor. As prediction algorithm we chose the support vector machine (SVM) algorithm which can incorporate sequence information via the kernel function. A variety of string-based kernel functions exist, ranging from bag-of-words kernels (mainly useful for document classification) to dedicated kernels for protein sequence analysis, most notably substring spectrum kernels (which count the occurrence of substrings) and variations e.g. including gaps (Lodhi *et al*., 2002). We tailor the definition of the kernel such that it fits the nature of our biological problem, i.e. prediction based on TMD sequence. Since the TMD has a well-defined alpha-helical structure it is rational to take this into account when defining a kernel. This approach is an example of applying structure-based features in kernel functions, which to our knowledge has not previously been explored in depth. Our results show that this approach indeed can be used to predict sub-Golgi localization, and we provide several examples of valuable applications of our predictions.

2 METHODS

2.1 Localization dataset and prediction of type II membrane proteins

Based on literature search and on entries in the general localization database Locate (http://locate.imb.uq.edu. au/cgi-bin/display.cgi) (Aturaliya *et al*., 2006; Sprenger *et al*., 2008) a dataset of protein sequences with known sub-Golgi localization was assembled. Because our main interest for sub-Golgi localization prediction is related to type II transmembrane proteins, a first filtering step was to apply transmembrane (TM) helix prediction to these sequences using TMHMM (Krogh *et al*., 2001). Only sequences predicted to be of the type II transmembrane signature (i.e. with exactly one TM helix predicted and of 'N-term in / TM / C-term out' topology) were retained. Since the number of type II transmembrane proteins with experimentally verified sub-Golgi localization information is rather small, no distinction was made between different species. Subsequently, these sequences were aligned with Muscle (Edgar, 2004) and clustered based on sequence identity, in order to prepare a non-redundant dataset for training and testing. Clustering was performed using the minimum variance method implemented in the R function hclust (Murtagh, 1983). The number of clusters was selected by assessing the inter-cluster sequence identity, which was found to rise sharply above 31 clusters.

For each cluster, additional sequences were obtained based on ENSEMBL families (Flicek *et al*., 2008) if available, and otherwise via BLAST. In both cases, matching over the full-length amino-acid sequence was required, as well as a minimum sequence identity of 70%. On these additional sequences, the same TM-prediction-based filter as described above was applied, but these predicted TMD sequences were only used for training, not for testing or validation.

2.2 Amino acid grouping

Different groupings of amino acids were tested in the definition of kernel features: (1) amino acids with similar dipole and volumes of side chains were clustered following (Shen *et al*., 2007) into the following 7 groups: AGV, ILFP, YMTS, HNQW, RK, DE and C; (2) we also tested using all 20 amino acids separately; and (3) a grouping of most similar amino acids based on a transmembrane substitution matrix (Ng *et al*., 2000), resulting in the groups VIM, YF, DE, TS with all other amino acids separately. Because the grouping based on dipole and volume resulted in the best predictor performance, only SVM results using this approach are presented here.

2.3 Kernel function and Support Vector Machine (SVM)

String-based triads: An important choice when applying SVM is how to define the string kernel that describes the protein sequence. We took as a starting point the conjoint triad string kernel, as recently proposed in the context of protein interaction prediction (Shen *et al*., 2007). To adapt this to the biological problem at hand, we reasoned that in addition to sequential triplets, triplets consisting of residues with a fixed linear spacing between one another might be important, because this determines alignment of such triplets to specific sides of the transmembrane helix. Thus, we redefined triads to accommodate a fixed spacing of either 0 (the original triad definition) or 1, 2 or 3 (non-sequential triads).

3D-structure based triads: Because the results of applying the above defined string-based kernels indicated the importance of taking into account structural features of the TMD (see Results), a final kernel was designed based on observed residue–residue contacts in 3D models of the TMD helix. These models were obtained by carrying out structure calculations in CNS (Brunger *et al*., 1998). CNS topologies were generated with the CNS script generate_seq.inp. Dihedral angle restraints were defined for backbone angles phi, $-65^\circ \pm 20^\circ$ and psi $-40^\circ \pm 20^\circ$, respectively, and hydrogen bond restraints were defined between each O(i)-N(i+4) pair (lower and upper bound 2.3 and 3.5 Å, respectively) and O(i)-HN(i+4) pair (lower and upper bound 1.7 and 2.5 Å, respectively). The anneal.inp CNS-script was used, which applies a high-temperature torsion-angle dynamics phase (1000 steps of 15 fs integration time step at 50 000K) followed by a torsion angle dynamics cooling phase from 50 000K to 0K (1000 steps of 15 fs) and a second cartesian dynamics cooling phase from 2000K to 0K (3000 steps of 5 fs). Ten structures were calculated for each TMD and sorted according to the total energy, and the lowest energy structure was used to obtain the kernel-features. Side-chain–side-chain contacts were counted using a distance cutoff of 3.5Å and each triplet of amino acids within this distance cutoff was counted as one occurrence of a triad, again using the amino acid grouping defined above. No sequential spacing effects were applied a priori in this case.

As SVM implementation SVMlight (Joachims, 1999) was applied. Using the observed string- or structure-based triads, for each type of triad v_i (i ranging from 1 to 343 in the case of the grouping of the amino acids into 7 groups) a normalized count was defined as $d_i=(f_i-\min)/\max$, where f_i is the raw count and min (max) is the minimum (maximum) over all f_i. Since the number of training examples was relatively small compared to the dimension of the feature space, a linear kernel was expected to be powerful enough. Leave-one-out cross validation was applied to optimize the parameter C (trade-off between training error and margin), for which a grid [1,2,3,4,5,6,7,8,9,10,15,20,25,30] was used. For the sake of completeness, the radial basis function (RBF) kernel was also tested, where the additional γ parameter was optimized on a grid [500,200,100,50,10,5,1,0.1,0.01,0.001,0.0005,0.0001]. To obtain an unbiased performance estimation, an independent 'leave-one-out' loop, following the nested cross-validation setup, as described previously, was used (Varma and Simon, 2006). Note that this setup avoids erroneously optimistic estimates obtained by simply using cross-validation to optimize the SVM parameters. The 'leave-one-out' was performed cluster-wise, meaning that all sequences in one cluster were removed simultaneously.

Table 1. Test data sets

Dataset	Nr.	Reference / source
Human-mouse orthologs	2×81	http://www.genome.jp/kegg/glycan/GT.html ftp://ftp.informatics.jax.org/pub/reports/HMD_HumanSequence.rpt
Arabidopsis-rice orthologs	2×35	(Dunkley *et al*., 2006); Inparanoid (Remm *et al*., 2001)
Sialyltransferases	134	(Harduin-Lepers *et al*., 2005)
SNAREs	145	(Yoshizawa *et al*., 2006)

2.4 Test sets and applications

The predictor was applied to several small-scale (proteins with multiple localization, type I membrane proteins, and a multi transmembrane domain protein) and large-scale protein test sets (see Table 1). The set of predicted SNAREs from 40 genomes was classified previously based solely on their SNARE domain (Yoshizawa *et al*., 2006). From these, we obtained the 145 cases that are predicted orthologs to human and yeast SNAREs with known *cis*-Golgi or TGN localization (these include some syntaxins, sec22 and vamp4). Approximately half of these sequences have only one of the predicted localizations (*cis*-Golgi or TGN) among the top 5 homologous sequences according to the analysis in Yoshizawa *et al*., whereas the other half have both *cis*-Golgi and TGN among the top 5 predictions.

In addition, reaction patterns for 97 glycosyltransferases were obtained as described previously (Kawano *et al*., 2005). These proteins were filtered using TMHMM and WoLF PSORT (Horton *et al*., 2007) (to obtain type II Golgi-localized proteins). Glycan structures were obtained from KEGG (Hashimoto *et al*., 2006). For each of those, a penalty score was calculated based on the reaction patterns observed for those structures and the predicted localizations for the associated enzymes. To calculate this score, for each two subsequent edges connecting monosaccharides in the glycan structure, the predicted localization of the associated enzymes was mapped, and a penalty score of +1 was assigned each time the second enzyme was predicted in an earlier compartment than the first one. The penalty score for each structure was normalized using the total number of edge-pairs, giving a value between 0.0 and 1.0. These penalties were compared with the scores for random enzyme-localizations, where the randomization was constructed such that for enzymes catalyzing the same reaction the number of different localizations was the same as in the real predictions.

3 RESULTS

3.1 Localization dataset and TMD prediction

In total, a dataset of 102 proteins with known sub-Golgi localization was obtained. For 64 of these, the relevant type II transmembrane topology was predicted by TMHMM. Five of those were discarded based on multiple locations reported in the literature. The remaining 59 sequences were clustered in order to remove redundant sequences prior to training, which would otherwise result in unjustified high performance of the resulting predictor. Figure 1A shows the inter-cluster sequence identity as a function of the number of clusters; based on this analysis 31 clusters were selected (Table 2 and Supplementary Table S1), because the maximum similarity between clusters rises sharply when using more clusters. The number of clusters with *cis*, *medial*, *trans* and TGN localization was 12, 6, 4 and 9, respectively (Fig. 1B). Of the 31 clusters, 18 clusters had only one entry and 13 clusters had multiple entries with consistent localization, indicating that within these limited sets the available data is consistent. This is reassuring given the fact that experimental

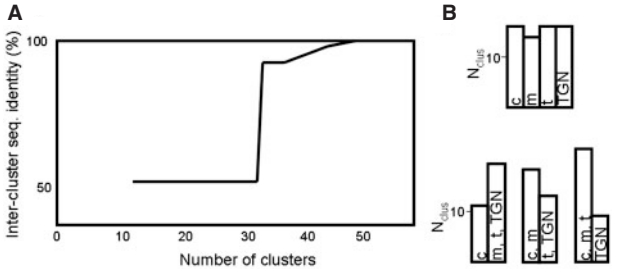

Fig. 1. Clustering type II transmembrane sequences with known sub-Golgi localization. **(A)** Maximum value for sequence identity between different clusters versus number of clusters. **(B)** Distribution of the various cases over *cis* (c), *medial* (m), *trans* (t) and TGN, for the full dataset (top) and the clustered dataset (bottom panel); for the latter, the combinations of cases used in the three separate predictors are shown. N_{clus}, number of clusters.

Table 2. Sub-Golgi localization dataset[a]

Species	Cluster members
Cis (107)	
Glycine max	α-1,2 mannosidase I (0; 0)
Human	BCMA peptide (0; 2); β-1,4 N-acetylgalactosaminyltransferase 2 (1; 21)
Mouse	Tmed10 (0; 1)
Pig	UDPN- acetyl-D- galactosamine:polypeptide N-acetylgalactosaminyltransferase (0; 41)
Rat	alpha-mannosidase IB (0; 17)
Yeast	anp1p (1; 0); hoc1p (0; 2); mnn10p (0; 2); mnn11p (0; 2); mnn9p (0; 4); och1p (0; 0)
Medial (117)	
Arabidopsis	β1,2-xylosyltransferase (0; 0)
Mouse	Lactosylceramide α-2,3-sialyltransferase (1; 13); mannoside acetyl-glucosaminyltransferase 1 (1; 20); fukutin related protein (0; 12)
Human	Fukutin (1; 19); β-3GalT6 (1; 5)
Trans (86)	
Guinea pig	nucleoside diphosphatase (0; 20)
Human	SGalT (1; 10); heparan sulfate 6-O-sulfotransferase (1; 17)
Rat	GD3 synthase (0; 12)
TGN (89)	
Arabidopsis	Syp42 (1; 2); AtVTI1a (0; 7)
Human	STX6 (0; 7)
Mouse	mannoside acetylglucosaminyltransferase 3 (1; 16); GES30 (1; 0); Vti1a (1; 5); Ndst2 (1; 0); vamp4 (1; 7); syntaxin 16 (0; 30)

[a]For each cluster, one entry is listed; additional number of proteins with experimental localization information in cluster, and additional number of training sequences obtained from ENSEMBL or via BLAST are indicated in brackets. For full table including references see Supporting Information Table 1.

determination of the sub-Golgi localization is difficult. For most of the clusters, additional training sequences were obtained based on sequence similarity (see Methods).

We determined the distribution of TMD length for the type II TMD's in our dataset and compared these with results for all plasma-membrane located type II proteins from the Locate database. This confirmed the suggestion in the literature that Golgi localized proteins have shorter TMD lengths than plasma proteins (Fig. 2).

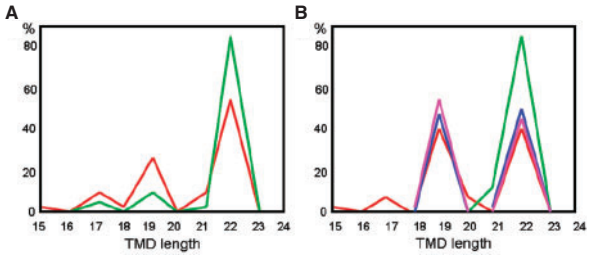

Fig. 2. Histogram of occurrence of predicted TMD lengths for type II transmembrane proteins, for **(A)** Golgi (red) and plasma membrane (green) and **(B)** 4 different sub-Golgi localizations: *cis* (red), *medial* (green), *trans* (blue), TGN (purple; colour online).

In the sub-Golgi set, however, there is no clear distinction between the various compartments. Differences may be artifacts of TMHMM (Cuthbertson *et al.*, 2005), although a comprehensive comparison between various TM-predictors showed that TMHMM was among the best performing tools (Moller *et al.*, 2001). Note that the length of the TMD-helix may not be directly related to the thickness of the lipid bilayer as the TMD may be inserted in the membrane under an angle (Killian and Nyholm, 2006).

3.2 Support Vector Machine (SVM)

To obtain a multiclass classification, three separate predictors were built: one for *cis* versus the other three localizations, one for *cis* or *medial* versus *trans* or *TGN*, and one for *TGN* versus the other three localizations. This particular ordering was chosen because *cis* and *TGN* are the locations with the highest number of cases, and this ordering coincides with the biologically relevant order *cis*−*medial*−*trans*−*TGN*. In addition, it requires only three separate predictors, in contrast to the approach of testing each location separately versus the other locations. The *cis/medial* versus *trans/TGN* predictor was used to test the performance of the various string kernels. Importantly, the error estimates that we provide are based on a nested cross-validation for the SVM parameter optimization, which precludes over-optimistic estimates based on a 'single' cross-validation-based optimization. As shown in Table 3, the highest accuracies were obtained for the linear string kernel using triads with a spacing of 3 or 2 between the subsequent residues in the triad (69 and 73%, respectively). Notably, this spacing coincides with the average rise in one helical turn, which is 3.6 amino acid (spacing of 2.6). In Figure 3, we illustrate the relation between the best performing kernel and the structure of an α-helical TMD region, by indicating for a helical structure some triplets based on a spacing of 2 for which the constituent residues are indeed proximal in 3D space.

Because the linear kernels suggested a relationship between amino acids on the same face of the α-helical TMD, we also defined a kernel based purely on observed side-chain–side-chain contacts in modeled structures of the TMD regions. Figure 4 illustrates how the features for this kernel were obtained. This final predictor improved overall performance to 76%, while in comparison to the string kernels its accuracy was also more balanced in the sense that accuracy for *cis/medial* and for *trans/TGN* was comparable (Table 3). Importantly, we tested that this spatial structure-based kernel captures more than only linear sequence similarity by analyzing a simple predictor based on sequence identity between the TMD regions. This predictor uses the localization of the protein whose TMD has the highest sequence identity to the TMD of the test protein, after removing all sequences that are in the same cluster as the test protein (an approach similar to the

Table 3. Performance for cis/medial versus trans/TGN prediction

Kernel	Accuracy (%)		
	Cis/medial	Trans/TGN	All
Sequence-based			
Spacing 1	64.0	43.0	55.2
Spacing 0	73.4	58.5	67.2
Spacing 3	64.2	76.6	69.4
Spacing 2	64.3	84.3	73.0
Structure-based	78.5	72.6	76.1

Sequence-based kernels are designated by the spacing used to defined the triads (0–3); the structure-based kernel is defined using observed contacts in modeled transmembrane helices (see Methods for details). Accuracies are reported for cis/medial, trans/TGN and for all sequences.

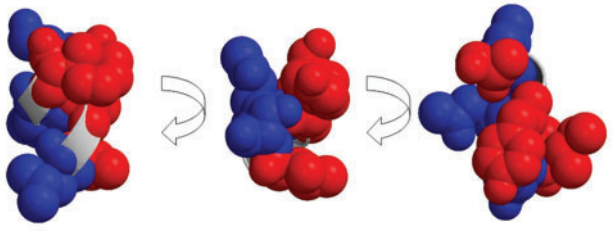

Fig. 3. String kernel based on triads with spacing of two amino acids in-between captures structural effects. Two triads (blue and red, respectively; colour online) formed by residues with two residues in-between in the amino acid sequence are shown here for one particular α-helical structure with sequence SLLYQLIS, using three different points of view for the same structure. For both triads the constituent residues are visibly close in 3D-space.

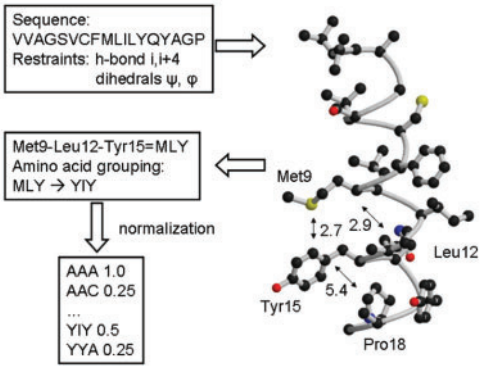

Fig. 4. Flow diagram illustrating the construction of the contact-based kernel (for details, see Methods). Observation of one specific triplet is indicated, consisting of Met9, Leu12 and Tyr15. As indicated, the distance between Tyr15 and Pro18 is larger than the cutoff of 3.5 Å, meaning that Met9, Tyr15 and Pro18 do not form a triplet (colour figure online).

cross-validation procedure applied to the SVM-predictor). When applying this simple sequence-based predictor to the *cis/medial* versus *trans/TGN* prediction, the performance is only 42% (13 out of 31), and especially '*medial*' turns out to be difficult to predict based on sequence identity. Note that the average TMD-sequence-identity within the clusters is 79% (+/−20%), whereas the average highest sequence-identity among clusters is only 35% (+/−3%).

In addition, randomly assigning class labels to each set of clustered sequences and retraining the SVM-predictor resulted in much lower performance (47% accuracy), and the same was true for randomly shuffling the features for every protein, while keeping the labels to their experimental values (49% accuracy). This demonstrates that the performance obtained by the SVM-predictor is non-trivial.

Based on its superior performance, the contact-based kernel was subsequently applied to develop the two other predictors, *cis* versus *medial/trans/TGN* and *TGN* versus *cis/medial/trans*. Table 4 shows that the performance of those predictors is comparable to that of the *cis/medial* versus *trans/TGN* predictor. For each of these three predictors we also tested an RBF instead of linear kernel, which gave comparable results (data not shown).

The simplest way to combine these three predictors is by using combinatorial logic, e.g. if for a given sequence the *cis/medial* versus *trans/TGN* predictor returns '*cis/medial*' and the *cis* versus the rest predictor returns 'not *cis*' then the prediction would be '*medial*'. This procedure could lead to inconsistencies, if, for example, in the above case also, '*TGN* versus the rest' would be predicted '*TGN*'. Such inconsistency occurred only once within our data set; for the remaining cases, Table 5 shows the confusion table. Overall, 19 cases out of 31 are correctly predicted, leading to a (cross-validated) prediction accuracy of 61%, which is a reasonable performance for a four-class classification problem.

Table 4. Prediction results for three predictors

Predictor	Accuracy				
	Cis	Medial	trans	TGN	All
Cis vs. other	48.1	91.3	76.0	92.0	72.8
Cis/medial vs. trans/TGN	75.0	85.6	86.1	66.7	76.1
TGN vs. other	83.5	92.5	94.4	61.3	80.2

Table 5. Confusion table for combined predictor

Experimental	Predicted			
	Cis	Medial	Trans	TGN
Cis	6	3	1	2
Medial	0	5	0	1
Trans	0	0	3	0
TGN	1	2	1	5

3.3 Applications

3.3.1 Small scale test sets

Proteins with multiple localizations As a small-scale test, the predictor was applied to the 5 cases out of the dataset that have multiple localizations according to the literature (and which were not used for training or cross-validation). This resulted in the prediction '*medial*' for three of them (human α1,3-1,6 mannosidase II, human *N*-acetyl glucosaminyltransferase I and human α2,3sialyltransferase) and '*cis*' for *Arabidopsis* xyloglucan galactosyltransferase, which are all among the experimentally found localizations. For only one case, the yeast guanosine diphosphatase (*cis/medial*), the predictor incorrectly predicted '*trans*' localization.

Type I membrane proteins In contrast to type II membrane proteins, the N-terminal domain of type I proteins is located in the Golgi, and as a consequence, the helical structure of their TMD has an opposite orientation in the membrane. If the lipid bilayers of the different sub-Golgi compartments are symmetrical, then the orientation of the TMD helix in the membrane will not make any difference for the interactions between the helix and membrane lipids. In that case our predictor should function equally well for type II and type I proteins. Of the 14 type I transmembrane proteins for which we could find an experimentally verified sub-Golgi localization, 7 were correctly placed by our predictor, giving 50% accuracy. In only one of the 7 mispredicted cases *cis*-Golgi and *TGN* were swapped by the predictor. Compared with an expected accuracy of 25% for a random 4-class predictor, this shows that our predictor is reasonably accurate for type I proteins as well, indicating at most a minor role for membrane asymmetry.

Multiple transmembrane domain proteins It is not obvious how to handle proteins (15 in total) with multiple transmembrane domains in our dataset. For one protein, however, protein M from the avian coronavirus infectious bronchitis virus, it is known that the first of its membrane spanning domains is involved in its targeting to the *cis*-Golgi (Machamer *et al*., 1993). Our model correctly predicts this *cis*-Golgi localization. Interestingly, mutagenesis experiments indicated that several polar residues lining one face of the helix would be important for the Golgi localization (Machamer *et al*., 1993), which is in line with the idea behind our structure-based kernel, and indeed most of the triplets found for protein M contain one or more of those previously identified residues.

3.3.2 Large scale applications

Human glycosyltransferases On a larger scale, the predictor was applied to a set of human glycosyltransferases and their mouse orthologs. This resulted in a consistent prediction across human–mouse ortholog pairs in over 70% of the cases (57 out of 81 cases). In addition, differences between ortholog pairs were 'intermediate' in the sense that in only 7 of the mismatch cases one of the pair was predicted '*cis*' and the other '*trans*' or '*TGN*'. Most mismatches involved directly adjacent compartments, e.g. '*cis*' was mostly mixed with '*medial*' and not with '*trans*'. Of course, not all cases where the prediction was consistent over the two of a pair are necessarily correct, but because functional differences between orthologous human and mouse glycosyltransferases can be expected to be small, it is reassuring that the predictor is reasonably consistent; the 30% inconsistent predictions would be expected to reflect the error rate of the predictor. Importantly, comparison with a simple sequence-similarity-based predictor again showed that this is not trivial: when for each mouse sequence a prediction is obtained by looking for the human TMD with the highest sequence identity, in only 23% of the cases this is identical to the prediction for the real

human ortholog of that mouse sequence. The predictions for the human and mouse glycosyltransferases and glycosylhydrolases are shown in Supplementary Table S2.

Plant type II membrane proteins A similar test was performed for *Arabidopsis* – rice orthologs. Of the 35 pairs of type II transmembrane proteins, 19 pairs had a consistent predicted localization, again indicating a reasonable accuracy for our predictor.

Animal sialyltransferases To illustrate how our approach could be helpful in function prediction, it was applied to a set of putative animal sialyltransferases (Harduin-Lepers *et al.*, 2005). For 55 out of a total of 136 sequences that passed the type II transmembrane-filter, the predicted localization was '*trans*-Golgi' or '*TGN*' and for 52 cases a '*medial*' localization was predicted. This overall pattern is in accordance with the expected late localization of these enzymes. Again, the predicted localizations were reasonably consistent between orthologs (in 23 out of 26 families the same localization was predicted for the majority of sequences in that family). Note that the set of enzymes used here were predicted to be sialyltransferases based on sequence similarity only. Our sub-Golgi localization prediction might be taken as indication that the *cis*-predicted proteins (29×) have other enzymatic activities.

SNAREs SNAREs provide specificity in transport vesicle–target membrane fusion. They contain an N-terminal SNARE domain and a C-terminal TMD and the predicted function of Golgi-SNAREs depends critically on their localization (ER-Golgi transport or post-Golgi transport). We compared our TMD-based predictions with previous predictions based on homology of the SNARE domain and found exact overlap in 49% of the cases. This number raised to 70% when only requiring matching of early (*cis*/*medial*) versus late (*trans*/*TGN*). Our predictions turn out to match much better with the unambiguous SNARE-domain-based predictions as compared to the ambiguous SNARE-domain-based predictions (57% exact correspondence versus 34% exact correspondence); for details, see Methods. This suggests that our predictor is helpful in discriminating between those ambiguous cases.

3.3.3 *Consistency of predicted localizations with glycan structures*

As 'glycomic' type of application of our predictor, we assessed the consistency of predicted localizations of glycosyltransferases with available glycan structures in KEGG (Hashimoto *et al.*, 2006). The link between the glycan structures and the predictor is formed by 'reaction patterns' that reflect the substrate specificity of glycosyltransferases which can be mapped onto the glycans. Sequences and reaction patterns were obtained as described previously (Kawano *et al.*, 2005) and after filtering with TMHMM and WoLF PSORT 37 enzymes remained.

Out of 10460 glycan structures, we could analyze 3651 with at least one pair of connected edges that both were associated with one of the 37 enzymes. Our assumption then was that the observed order of synthesis of the glycan (subsequent addition of monosaccharide units) should match the predicted localization of the glycosyltransferases involved. A penalty score was calculated for each glycan to reflect whether the predicted ordering of those enzymes indeed matches the ordering observed in the structure (with value 0.0 for perfect ordering throughout the structure and 1.0 for consistent wrong ordering). The average penalty score of the predicted localizations (0.116) was much lower than the

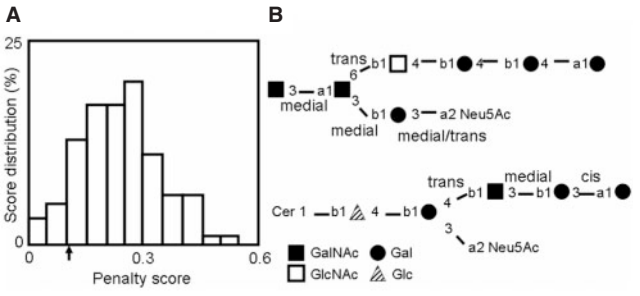

Fig. 5. Assessing consistency of predicted localization with observed glycan structures in KEGG. **(A)** Distribution of penalty scores for random location assignments (histogram); arrow indicates the value for the predicted localizations. **(B)** Examples of glycan structures with low penalty score (top panel) and high penalty score (bottom panel). Order of synthesis is from left to right.

value of 0.375 one would expect in the case of a random homogenous distribution of localizations (0.25*0.0 (*cis*) + 0.25*0.25 (*medial*) + 0.25*0.5 (*trans*) + 0.25*0.75 (*TGN*)). A total of 91 out of 100 random location assignments gave a higher average penalty score than the actual predictions, indicating that with $p \sim 0.1$ our predictions matched better to the observed glycan structures in KEGG. Note that simplifications in our approach are that edges are only considered pairwise, and that we do not explicitly deal with retrograde Golgi-transport.

Figure 5 shows the observed average penalty scores for random location assignments and for our predictor. In addition, two examples of predictions are shown, one with low and one with high penalty score. We also assessed for each of the enzymes whether it consistently was found in pairs of correctly ordered edges. For 24 out of 37, this was the case, giving an accuracy estimate of 65% which is close to the 61% found by cross-validation. Only in 8 of the 100 random trials the number of consistently correctly ordered enzymes is higher than 24, and on average this number is only 17 (+/−5). Note that the penalty score reflects both the accuracy of our predicted localizations and the more general fact of how well the glycan structures reflect the specific ordering of synthesis steps. Thus, our results here both provide additional validation for our sub-Golgi localization predictions as well as additional evidence for an 'assembly line' concept of glycan biosynthesis.

4 DISCUSSION AND CONCLUSION

Attention in experimental protein localization studies is shifting towards increasing resolution. Here we aim to support this by a computational analysis and show that our contact-based predictor confirms that the TMD of type II transmembrane proteins influences sub-Golgi localization. In contrast to a clear correlation between TMD-length and Golgi versus post-Golgi localization, no relation was observed between TMD-length and sub-Golgi localization. However, the form of the linear sequence-based kernel function that gave best prediction results coincided with predicted properties of the TMD in its α-helical conformation, where residues separated in sequence by 2-3 amino acids are proximal in 3D space. Indeed, by directly defining the kernel based on observed contacts in modeled 3D-structures, an even better performance was obtained. We propose

that such structure-based features can be of more general use in protein classification predictions.

There are some reports in the literature about the importance of combinations of residues at specific sides of the transmembrane domain for Golgi localization (Machamer *et al*., 1993; Sousa *et al*., 2003). Based on the properties of our best performing kernel function, we expect these to be of general importance and to reflect a physical mechanism of interactions with lipids or other membrane proteins. Note that we currently used results from one particular transmembrane domain predictor, TMHMM, and it might be possible to improve somewhat the performance of our method by incorporating results from additional predictors such as the recently developed consensus approach MemO (Davis *et al*., 2006).

A possible extension of our work is to also consider other domains in the sequence, using e.g. motif-occurrence or HMMs. In addition, it is clear that the currently available amount of data is very small, preventing a full assessment of our approach. In particular, it is difficult to give a precise estimate of its predictive performance. For this reason, our work is best seen as the first of a kind, providing directions for further research. Extending the dataset would enable further analysis aimed at understanding the mechanism behind sub-Golgi localization. Our current results already indicate that the transmembrane domain contains information that is not captured simply by sequence similarity, but by using a larger dataset it might be possible to directly extract those features that are most important for the localization. This could be performed using feature-selection algorithms (Saeys *et al*., 2007) and might result in insight into the roles of, for example, specific protein–lipid or protein–protein interaction sites. One would also expect that the current analysis, which only focuses on the TMD, is related to the stability of type II transmembrane proteins within their correct membrane environment, whereas analysis of, for example, the cytoplasmic N-terminal region might result in further understanding of the localization mechanism itself.

An interesting extension and application of our prediction algorithm is to combine it with existing approaches for analyzing the reaction paths leading to specific glycans (Hossler *et al*., 2006; Kawano *et al*., 2005), which currently do not consider localization information. Here, we already showed that there is a significant correlation between localization and ordering of the different steps in glycan biosynthesis. This demonstrates the potential of our predictor and paves the way towards further exploration and prediction of glycosyltransferase pathways.

ACKNOWLEDGEMENTS

We thank G. van Meer en J.C. Holthuis (Utrecht University) for helpful discussions.

Funding: This work was supported by the BioRange programme(SP 2.3.1) of the Netherlands Bioinformatics Centre (NBIC), which is supported through the Netherlands Genomics Initiative (NGI).

Conflict of Interest: none declared.

REFERENCES

Aturaliya,R.N. *et al*. (2006) Subcellular localization of mammalian type II membrane proteins. *Traffic*, **7**, 613–625.

Breton,C. *et al*. (2006) Structures and mechanisms of glycosyltransferases. *Glycobiology*, **16**, 29r–37r.

Brunger,A.T. *et al*. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. Section D-Biol. Crystallogr.*, **54**, 905–921.

Buist,G. *et al*. (2006) Different subcellular locations of secretome components of Gram-positive bacteria. *Microbiology-Sgm*, **152**, 2867–2874.

Colley,K.J. (1997) Golgi localization of glycosyltransferases: More questions than answers. *Glycobiology*, **7**, 1–13.

Cuthbertson,J.M. *et al*. (2005) Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Eng. Design Selection*, **18**, 295–308.

Czlapinski,J.L. and Bertozzi,C.R. (2006) Synthetic glycobiology: exploits in the Golgi compartment. *Curr. Opin. Chem. Biol.*, **10**, 645–651.

Davis,M.J. *et al*. (2006) MemO: a consensus approach to the annotation of a protein's membrane organization. *In Silico Biol.*, **6**, 387–399.

Dunkley,T.P.J. *et al*. (2006) Mapping the Arabidopsis organelle proteome. *Proc. Natl Acad. Sci. USA*, **103**, 6518–6523.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Emanuelsson,O. (2002) Predicting protein subcellular localisation from amino acid sequence information. *Brief Bioinform.*, **3**, 361–376.

Flicek,P. *et al*. (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.

Gardy,J.L. and Brinkman,F.S.L. (2006) Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol.*, **4**, 741–751.

Grabenhorst,E. and Conradt,H.S. (1999) The cytoplasmic, transmembrane, and stem regions of glycosyltransferases specify their in vivo functional sublocalization and stability in the Golgi. *J. Biol. Chem.*, **274**, 36107–36116.

Harduin-Lepers,A. *et al*. (2005) The animal sialyltransferases and sialyltransferase-related genes: a phylogenetic approach. *Glycobiology*, **15**, 805–817.

Hashimoto,K. *et al*. (2006) KEGG as a glycome informatics resource. *Glycobiology*, **16**, 63r–70r.

Horton,P. *et al*. (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–587.

Hossler,P. *et al*. (2006) GlycoVis: Visualizing glycan distribution in the protein N-glycosylation pathway in mammalian cells. *Biotechnol. Bioeng.*, **95**, 946–960.

Joachims,T. (1999) Making large-Scale SVM learning practical. In *Advances in Kernel Methods – Support Vector Learning*. MIT-Press, Cambridge, MA.

Kawano,S. *et al*. (2005) Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. *Bioinformatics*, **21**, 3976–3982.

Killian,J.A. and Nyholm,T.K.M. (2006) Peptides in lipid bilayers: the power of simple models. *Curr. Opin. Struct. Biol.*, **16**, 473–479.

Krogh,A. *et al*. (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Lehle,L. *et al*. (2006) Protein glycosylation, conserved from yeast to man: a model organism helps elucidate congenital human diseases. *Angewandte Chemie-Int. Edn*, **45**, 6802–6818.

Lei,Z.D. and Dai,Y. (2005) An SVM-based system for predicting protein subnuclear localizations. *BMC Bioinformatics*, **6**, 291.

Lodhi,H. *et al*. (2002) Text classification using string kernels. *J. Machine Learning Res.*, **2**, 419–444.

Machamer,C.E. *et al*. (1993) Retention of a Cis Golgi protein requires polar residues on one face of a predicted alpha-helix in the transmembrane domain. *Mol. Biol. Cell*, **4**, 695–704.

Mitra,K. *et al*. (2004) Modulation of the bilayer thickness of exocytic pathway membranes by membrane proteins rather than cholesterol. *Proc. Natl Acad. Sci. USA*, **101**, 4083–4088.

Moller,S. *et al*. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**, 646–653.

Murtagh,F. (1983) A Survey of recent advances in hierarchical-clustering algorithms. *Comp. J.*, **26**, 354–359.

Nakai,K. and Kanehisa,M. (1991) Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins-Struct. Funct. Genet.*, **11**, 95–110.

Ng,P.C. *et al*. (2000) PHAT: a transmembrane-specific substitution matrix. *Bioinformatics*, **16**, 760–766.

Puthenveedu,M.A. and Linstedt,A.D. (2005) Subcompartmentalizing the Golgi apparatus. *Curr. Opin. Cell Biol.*, **17**, 369–375.

Remm,M. *et al*. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.

Saeys,Y. *et al*. (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.

Saint-Jore-Dupas,C. *et al*. (2006) Plant N-glycan processing enzymes employ different targeting mechanisms for their spatial arrangement along the secretory pathway. *Plant Cell*, **18**, 3182–3200.

Schneider,G. and Fechner,U. (2004) Advances in the prediction of protein targeting signals. *Proteomics*, **4**, 1571–1580.

Shen,J.W. *et al*. (2007) Predicting protein-protein interactions based only on sequences information. *Proc. Natl Acad. Sci. USA*, **104**, 4337–4341.

Sousa,V.L. *et al*. (2003) Importance of Cys, Gln, and Tyr from the transmembrane domain of human alpha 3/4 fucosyltransferase III for its localization and sorting in the golgi of baby hamster kidney cells. *J. Biol. Chem.*, **278**, 7624–7629.

Sprenger,J. *et al*. (2008) LOCATE: a mammalian protein subcellular localization database. *Nucleic Acids Res*., **36**, D230–D233.

Sprenger,J. *et al*. (2006) Evaluation and comparison of mammalian subcellular localization prediction methods. *BMC Bioinformatics*, **7** (Suppl. 5), S3.

van Meer,G. *et al*. (2008) Membrane lipids: where they are and how they behave. *Nat. Rev. Mol. Cell Biol.*, **9**, 112–124.

Varma,S. and Simon,R. (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, **7**, 91.

Webb,R.J. *et al*. (1998) Hydrophobic mismatch and the incorporation of peptides into lipid bilayers: a possible mechanism for retention in the Golgi. *Biochemistry*, **37**, 673–679.

Yoshizawa,A.C. *et al*. (2006) Extracting sequence motifs and the phylogenetic features of SNARE-dependent membrane traffic. *Traffic*, **7**, 1104–1118.

Yuan,Z. and Teasdale,R.D. (2002) Prediction of Golgi Type II membrane proteins based on their transmembrane domains. *Bioinformatics*, **18**, 1109–1115.

Zerfaoui,M. *et al*. (2002) The cytosolic and transmembrane domains of the beta 1,6 N-acetylglucosaminyltransferase (C2GnT) function as a cis to medial/Golgi-targeting determinant. *Glycobiology*, **12**, 15–24.