Protein Engineering, Design & Selection vol. 17 no. 4 pp. 349–356, 2004
Published online April 28, 2004 doi:10.1093/protein/gzh037

Feature-based prediction of non-classical and leaderless protein secretion

Jannick Dyrløv Bendtsen[1], Lars Juhl Jensen[1,2], Nikolaj Blom[1], Gunnar von Heijne[3] and Søren Brunak[1,4]

[1]Center for Biological Sequence Analysis BioCentrum-DTU, Building 208, Technical University of Denmark, DK-2800 Lyngby, Denmark and [3]Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden

[2]Present address: Computational Biology Unit, EMBL-Heidelberg, D-69117 Heidelberg and Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Strasse 10, D-13092 Berlin, Germany

[4]To whom correspondence should be addressed.
E-mail: brunak@cbs.dtu.dk

We present a sequence-based method, SecretomeP, for the prediction of mammalian secretory proteins targeted to the non-classical secretory pathway, i.e. proteins without an N-terminal signal peptide. So far only a limited number of proteins have been shown experimentally to enter the non-classical secretory pathway. These are mainly fibroblast growth factors, interleukins and galectins found in the extracellular matrix. We have discovered that certain pathway-independent features are shared among secreted proteins. The method presented here is also capable of predicting (signal peptide-containing) secretory proteins where only the mature part of the protein has been annotated or cases where the signal peptide remains uncleaved. By scanning the entire human proteome we identified new proteins potentially undergoing non-classical secretion. Predictions can be made at http://www.cbs.dtu.dk/services/SecretomeP.

Keywords: extracellular/growth factor/hormone/neural network/secretome

Introduction

For targeting a protein to the extracellular space, it has for a long time been believed that an N-terminal signal peptide was strictly required. Recent studies, however, have shown that several extracellular proteins, such as FGF-1, FGF-2, IL-1 and galectins found in the extracellular matrix, can be exported without a classical N-terminal signal peptide (Rubartelli and Sitia, 1997; Hughes, 1999; Cooper, 2002; Nickel, 2003). In addition to the mentioned cases, nuclear HMGB1 and viral proteins, such as HIV-tat or herpes simplex virus VP22, have been shown to enter the non-classical secretory pathway (Goldstein, 1996; Hughes, 1999; Gardella *et al*., 2002). Secretion of proteins without an N-terminal signal peptide is currently known as leaderless secretion or the non-conventional/non-classical secretory pathway (Rubartelli and Sitia, 1997). Eukaryotic protein secretion normally routes through the endoplasmatic reticulum (ER) and Golgi, ending up in a secretory vesicle fusing to the cell membrane. Studies have shown that the non-classical secretory pathway works independently of the ER–Golgi network; the secreted proteins do not enter the ER and have therefore never been observed to be glycosylated even if they carry potential glycosylation motifs. Non-classical secretion of proteins can be verified experimentally as export is not being hampered by inhibitors of the classical secretory pathway, such as monensin and brefeldin A (Tanudji *et al*., 2002).

Accurate protein trafficking and localization are essential for all living organisms. Targeting to the secretory pathway, to mitochondria or to chloroplasts in plants is usually mediated by an N-terminal leader sequence. Targeting signals can also be found internally in the protein sequence, e.g. uncleaved signal peptides, nuclear import and export signals (von Heijne *et al*., 1991). Numerous methods for predicting the subcellular location of proteins have been developed, most of which rely on the presence of these signals (Nielsen *et al*., 1997; Nakai and Horton, 1999; Emanuelsson *et al*., 2000; la Cour *et al*., 2003). Despite the knowledge of alternative secretion pathways, no current machine learning approach directly addresses the problem of predicting proteins entering the non-classical secretory pathway. However, prediction methods based on amino acid composition are in principle capable of predicting proteins entering the non-classical secretory pathway (Reinhardt and Hubbard, 1998). Here we report that essentially all existing methods are unable to predict correctly the majority of the known examples of non-classical secretion.

With the excessive complexity of the secretome, which by definition comprises the secretory proteins and the secretion machinery (Tjalsma *et al*., 2000; Antelmann *et al*., 2001; Greenbaum *et al*., 2001), no current computational prediction method is able to cover all targeting signals.

In this paper, we present a sequence-based prediction method capable of identifying mammalian secretory proteins of the non-classical secretory pathway. The approach we have taken is similar to that of the ProtFun method for predicting protein functional role categories or Gene Ontology classes (Jensen *et al*., 2002a, 2003). It works by mapping protein sequences into protein feature space where they are represented by various sequence-derived features such as predicted post-translational modifications, predicted structure, degradation signals, composition, size and charge. We found that although the method has been trained to identify classical secretory proteins (with their signal peptide removed), it gives similar output scores for proteins secreted via the non-classical secretory pathway. This indicates that extracellular proteins share certain properties and features which can be related to protein function outside the cell and not to the specifics of the secretory process itself. This also enables us to identify secreted proteins where an N-terminal signal peptide is missing owing to gene finding errors. However, the method is not intended to compete with prediction methods identifying N-terminal signal peptides—when such signals are present we recommend using the SignalP and TargetP methods

instead (Nielsen *et al*., 1997; Emanuelsson *et al*., 2000; Bendtsen *et al*., 2004).

Materials and methods

Generation of data sets

Ideally, our positive data set should consists of a large number of proteins secreted via non-classical pathways. Unfortunately, it is not possible to obtain a sufficiently large data set as only a very small number of proteins undergoing non-classical secretion are known. Worse yet, many of these examples exhibit significant sequence similarity to each other.

Since we will be looking for features shared among extracellular proteins, the mechanism by which a protein is secreted should not be important. We therefore use for training the large number of proteins known to be secreted via the classical signal peptide mediated mechanism. A set of 3321 extracellular mammalian proteins were extracted from the Swiss-Prot database based on subcellular localization annotations in the comments block (Bairoch and Apweiler, 2000). Partial sequences and sequences without an annotated signal peptide were not included in the data set. As we wish to train a predictor that works in the absence of signal peptides, the signal peptide part of each sequence was removed.

A set of negative training examples was constructed by extracting 3654 mammalian proteins in Swiss-Prot, which were annotated as residing in the cytoplasm and/or the nucleus. In order to avoid the situation that the method would learn the trivial fact that transmembrane proteins are not extracellular, we did not include transmembrane proteins in the negative set as no such proteins are present in the positive set. Using TMHMM (Krogh *et al*., 2001), we were able to remove all transmembrane proteins with high confidence, as TMHMM can discriminate between soluble and membrane proteins with both specificity and sensitivity better than 99%. Moreover, we did not include proteins from most other subcellular localizations since they go via the endoplasmic reticulum (ER), making it difficult to exclude that they could be secreted.

A test set of 13 non-classically human secretory proteins were collected from Swiss-Prot. Criteria for selection was clear experimental evidence within the literature for the given sequence entry. Only human secretory sequences were used in this test set, discarding viral and parasitic sequences. The Swiss-Prot sequence entries found were CNTF_HUMAN, FGF1_HUMAN, FGF2_HUMAN, HME2_HUMAN, HMG1_HUMAN, IL1A_HUMAN, IL1B_HUMAN, IL18_HUMAN LEG1_HUMAN, LEG3_HUMAN, MIF_HUMAN, THIO_HUMAN, THTR_HUMAN.

Data set partitioning

The data set (positives and negatives) was divided into five cross-validation subsets of roughly the same size with minimal sequence similarity between the subsets. This ensures that when training cross-validation ensembles of neural networks, any two similar sequences will either both be used for training or both used for testing. Compared with a strategy where redundant sequences are discarded, a much larger data set will be available, while we still obtain a correct measure of the predictive performance on the independent test subsets. The most 'significant' match between training and test had a BLAST E-value of only 8×10^{-4} (26% identity for one sequence pair). In general, the similarity is even lower, ruling out completely that the method transfers functional information based on sequence similarity.

Training of a feature-based neural network

Standard feed-forward neural networks with a single layer of hidden neurons were used for predicting which proteins are secreted. Neural networks were trained using as input different combinations of sequence-derived features previously used for other protein function prediction tasks (Jensen *et al*., 2002a,b, 2003). Although our data set was constructed as not containing transmembrane proteins with all signal peptides removed, the SignalP, TargetP and TMHMM predictions were initially included in the feature set, allowing the method to choose features in an entirely data-driven fashion.

The feature selection was performed as follows. First, cross-validation ensembles comprising five neural networks were trained using each single protein feature as input. A robust estimate of the feature performance was calculated as the median test set performance of the five neural networks in the ensemble. Based on the single feature performance estimates, the best features were selected for training of neural networks having pairs of features as input. Again, the most promising features were selected to build up progressively larger feature combinations. As different features and feature combinations will result in very different numbers of input neurons, the number of hidden units was varied to keep the size of the network as constant as possible, aiming for 300 weights.

The six features included in the final neural network were encoded using the following scheme. Calculated sequence properties were presented to the network as single values. Thus, the number of atoms and the number of positively charged amino acid residues were encoded as a single value for each input sequence. In order to preserve positional information, the position specific features were encoded as average values within a number of bins representing different parts of the sequence. For the masking of low-complexity regions by the SEG filter, each sequence was fractioned into five bins and presented to the neural network. Predictions of propeptides with ProP were similarly fractioned into five bins for each sequence. Twenty probabilities were used as input to the neural network from prediction of subcellular localization by PSORT. For transmembrane helix prediction by TMHMM, each sequence was fractioned into five bins and 'inside', 'outside' and 'membrane' predictions were presented to the neural network during training. The other features not retained in the final predictor were represented in a similar fashion.

Screening of the human proteome

For analysis of the human proteome, the IPI database release 2.11 was downloaded (http://www.ebi.ac.uk/IPI/). As the neural network method presented in this paper has been trained to recognize proteins secreted without the use of a signal peptide only, the data set was also filtered to remove all protein sequences predicted by SignalP to contain an N-terminal signal peptide. Furthermore, TMHMM was used to remove all predicted transmembrane proteins from the set as no such proteins were included during training (Krogh *et al*., 2001). This left us with an initial set of 36 585 sequences to be screened. For these sequences, all sequence-derived protein features were calculated, normalized in the same way as for the training examples and finally used as input for all five

neural networks in the cross-validation ensemble. The mean of the five resulting output values was used as the final score for each sequence.

The fairly large number of questionable protein sequences derived from the genes predicted by computational gene finding methods presented a problem in our analysis. All the human protein sequences were therefore compared using gapped BLAST (Altschul *et al*., 1997) to the equivalent IPI database of mouse proteins. Only human proteins for which a mouse homolog could be found with an E-value below 10^{-6} were kept. This requirement reduced our initial set of 36 585 sequences to a more reliable set of 21 771 sequences, which were used for the computational screening and identification of novel putative proteins undergoing non-classical secretion.

Results and discussion

No simple motifs in proteins undergoing non-classical secretion

In the well-described examples of non-classical secretion mentioned above, no sequence motifs have been found that target these proteins to the extracellular surroundings. When submitting a set of 13 known human examples of non-classical secretion to PSORT (Nakai and Horton, 1999), seven are assigned as nuclear, four are predicted to be cytoplasmic and two are predicted to be mitochondrial. None of the 13 sequences are predicted to be extracellular. It is not surprising that PSORT is unable to classify these sequences correctly, as extracellular proteins in PSORT essentially are identified as signal peptide-carrying sequences using McGeoch's method (McGeoch, 1985), later modified by Nakai and Kanehisa (Nakai and Kanehisa, 1991), in combination with the von Heijne weight-matrix for signal peptide prediction (von Heijne, 1986).

The neural network-based method NNPSL, which is based solely on amino acid composition as input, has a reported prediction accuracy of 66% for four eukaryotic compartments (Reinhardt and Hubbard, 1998). Of the 13 protein sequences of non-classical secretion, six are correctly predicted as being secreted, three are predicted to be mitochondrial, two nuclear and two cytoplasmic. Thus, using NNPSL, 46% of this validation set of non-classical secretory proteins could be correctly predicted.

The notion behind the SecretomeP method is firstly that secretory proteins share certain features regardless of the mechanism by which they are secreted and secondly that these combinations of features are not often found in non-secretory proteins. Due to the relatively small number of known non-classically secreted proteins, we use as positive training examples classically secreted proteins (with signal peptide removed) and inspect whether this approach will identify non-classically secreted proteins correctly. We investigated 16 different sequence-derived features and sequence-based prediction methods and tested them for discriminatory value. Subsequently, those contributing most strongly to the predictive performance were combined in a neural network approach (Jensen *et al*., 2002a).

Features of discriminative value for identification of non-classical secretory proteins

Many sequence features will, when assessed independently, be present in proteins undergoing non-classical secretion as well as in non-secretory proteins. We therefore searched for combinations of features with discriminatory value. The full list of features used is given in Table I. Using an iterative scheme (see Materials and methods) where features were tested individually and in combination, we eventually obtained a set of six features together having optimal discriminatory power (Table I).

The number of atoms and the number of positively charged residues turned out to be two of the best performing features. The predictive power of each feature, when used alone, provides a very simple measure of feature importance. As both of these features are strongly correlated to the sequence length, we analyzed the length distributions of the non-classically secreted and non-secreted proteins in addition to that of proteins known to be secreted by conventional mechanisms (see Figure 1).

Several features may well be encoding the same information in different ways—protein size is encoded by both the number

Table I. Features tested for discriminatory value: the table lists features included in the initial training of which some are calculated, wherease others were predicted by various prediction tools from the amino acid sequence

Feature
Number of atoms calculated by ExPASy ProtParam[a]
Number of negatively charged residues calculated by ExPASy ProtParam
Number of positively charged residues calculated by ExPASy ProtParam[a]
Isoelectric point calculated by ExPASy ProtParam
Extinction coefficient calculated by ExPASy ProtParam
Grand average of hydropathicity calculated by ExPASy
ProtParam PEST regions predicted by PESTfind
Low-complexity regions predicted by SEG[a]
Secondary structure predicted by PSI-Pred
Transmembrane helices predicted by TMHMM[a]
N-linked glycosylation predicted by NetNGlyc
O-linked glycosylation predicted by NetOGlyc
Subcellular localization predicted by PSORT[a]
Propeptides predicted by ProP[a]
Tyrosine phosphorylation predicted by NetPhos
Serine and threonine phosphorylation predicted by NetPhos

[a]These features were used in the final method.

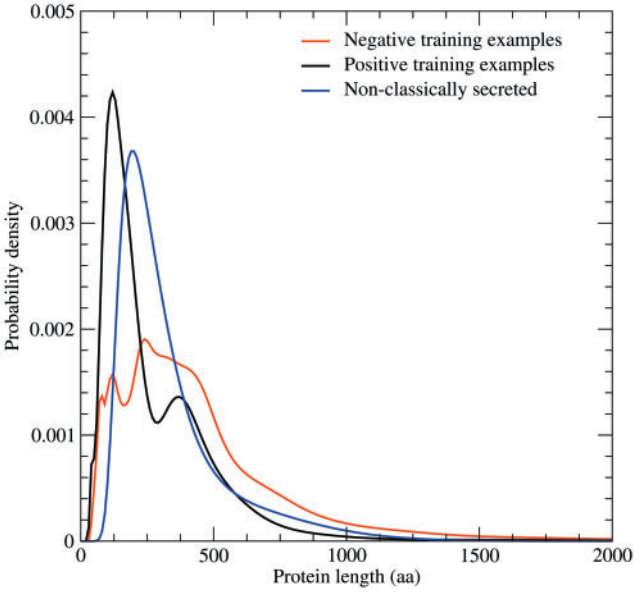

Fig. 1. Protein length distributions. Length distributions for non-classically secreted proteins and cytoplasmic proteins. The length distributions have been normalized and smoothed by a Gaussian kernel density estimation.

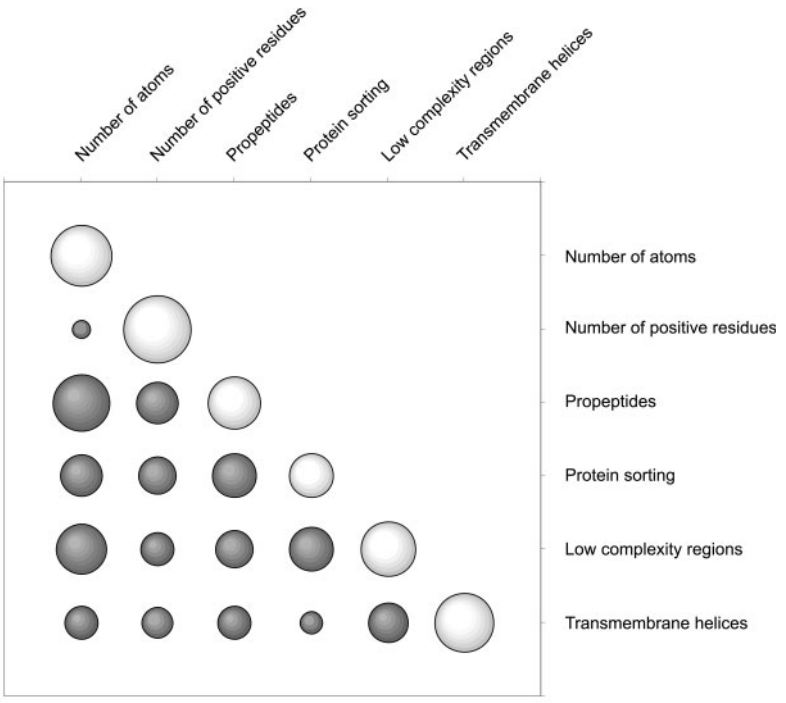

Fig. 2. Correlations in feature information. This plot show the importance of the six features included in the final prediction method. Features included are number of atoms, number of positive residues, propeptide cleavage site (ProP), protein sorting (PSORT), low-complexity regions (SEG) and transmembrane helix predictions (TMHMM). The white diagonal circles shows single feature importance and the gray circles illustrate combined feature importance, that is, e.g. number of atoms and number of positive residues provide much information to the network individually, but in conjunction only limited additional information is obtained. The spot size is proportional to the correlation coefficient.

of atoms and the number of positively charged residues. To discover other less obvious pairs of correlated features, we studied the performance of networks trained for pairs of features. By calculating the difference in the performance of a feature pair and the best of the two individual features, we obtained a measure for the additional information gained by combining features (see Figure 2).

From this feature analysis, it is clear that PSORT and TMHMM also appear to be encoding much the same information in relation to this task. PSORT predicts the probability that a protein resides in each of a number of subcellular localizations. The localization prediction which appears to be of greatest predictive power is 'cytoplasmic', as PSORT generally predicts secreted proteins to be less likely cytoplasmic than non-secreted. PSORT does not predict secreted proteins lacking their signal peptide to be secreted, but predicts them to be cytoplasmic or nuclear. Given the absence of transmembrane proteins in our data set (see Materials and methods), it is more surprising that TMHMM, which predicts transmembrane helices and the topology of transmembrane proteins, can be used to discriminate between secreted and non-secreted proteins. It appears that most of its predictive power stems from the prediction of membrane topology, i.e. 'inside' vs 'outside', rather than the prediction of transmembrane segments.

The two final features included in the prediction show much weaker correlation to other features and therefore each contributes unique information. These features are low-complexity regions as detected by the SEG filter (Wootton and Federhen, 1993) and furin-type propeptide cleavage sites predicted by ProP (Duckert *et al*., 2004). Both features yielded lower scores for secreted than non-secreted proteins, i.e. they contain both fewer regions of low sequence complexity and fewer propeptide cleavage sites. This is puzzling, since propeptide cleavage usually takes place in secreted proteins (Thomas, 2002). The propeptide predictor is not only detecting dibasic sites, but also there is a small bias in the R/K content which is 12.1, 10.4, 10.7 and 10.7% in intracellular, extracellular, non-classically secreted and secreted (mature part), respectively.

In earlier work, differences in single and pair amino acid composition between extracellular and non-secreted proteins have been reported (Nakashima and Nishikawa, 1994). For the over-represented pairs of amino acids we found in general little agreement when comparing the mature part of the classically secreted and the non-secreted proteins. More importantly, when calculating the over-represented pairs in the (small) non-classically secreted data set, these seem to differ strongly from those found in the classically secreted proteins. We concluded that no simple compositional statistics can identify non-classically secreted proteins.

Prediction performance of the neural network

For any prediction method, an important aspect is to assess the prediction performance and to make a trade-off between sensitivity and specificity, i.e. finding as many of the positive examples as possible while still keeping the number of false-positive predictions low. Figure 3 shows this trade-off for the method as a ROC (receiver operating characteristic) curve. The performance corresponds to what we expect for novel proteins as the curve is based on cross-validated test set performances with minimized similarity to the training sets. With the current method we are able to obtain a sensitivity of 40% with a very low level of false-positive predictions of <5%.

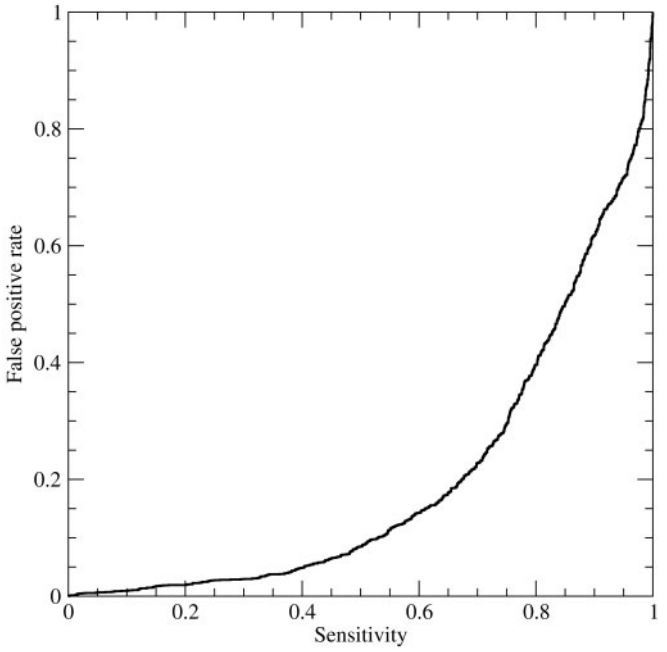

Fig. 3. Sensitivity and rate of false positives for the prediction method. The receiver operating characteristic (ROC) was constructed based on the cross-validation test set performances. Due to the homology partitioning of the cross-validation data set, the performances shown correspond to what can be expected for novel proteins. Random performance would correspond to a diagonal line.

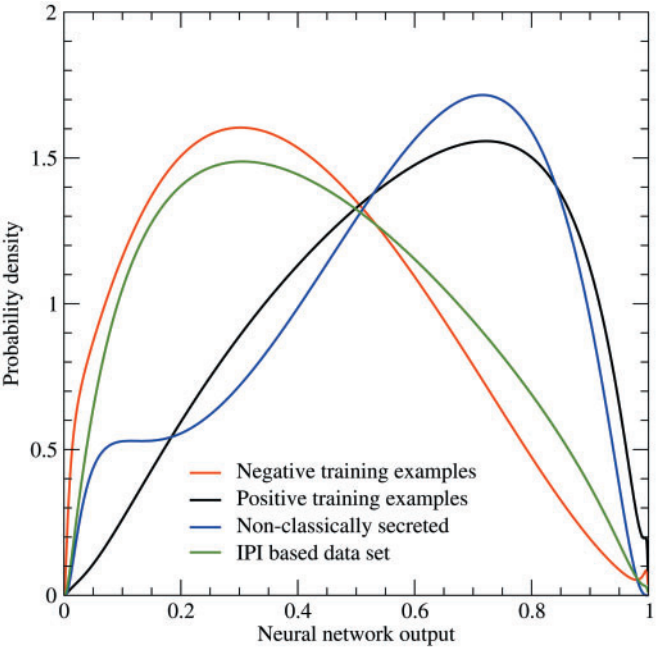

Fig. 4. Neural network score distributions. Known secretory proteins of the non-classical pathway have similar neural network output scores to the positive training examples (truncated secretory proteins of the classical pathway without the signal peptide), easily distinguishable from the negative training examples. The data set extracted from the IPI database display a scoring pattern similar to the negative training set, indicating that the majority of the IPI set is non-secreted. All score distributions were normalized and smoothed by a Gaussian kernel density estimation.

It is commonly believed that most of the sequence information, which is related to classical secretion, resides in the signal peptide itself and that the primary structure of mature protein is not strongly constrained. The score distributions shown in Figure 4 indicate clearly that this view may not be entirely correct. It is in fact possible to predict with reasonable accuracy whether a (classically secreted) sequence is secreted or not by using the mature part of the protein only. From the ROC curve in Figure 3, it is also clear that we cannot expect to find all non-classically secreted proteins without assigning a lot of false positives also. Below we have used this method to rank the entire human proteome and used the top scoring part where there will be no or very few false-positive assignments.

Despite our method being capable of predicting secreted proteins with signal peptide removed, the actual test should be on proteins known to enter the non-classical secretory pathway. To analyze this, the neural network score distribution for the set of 13 such proteins was compared with the score distributions obtained for positive and negative training examples (see Figure 4). Only human non-classically secretory proteins were used as a measure of performance.

It is clear that the neural networks give very similar output scores for the two sets of secreted proteins despite the proteins being secreted by entirely different pathways. This supports the idea that extracellular proteins share certain features, which are related more to the protein functioning outside the cell than to the process by which the protein is secreted.

Prediction results for known and putative non-classically secreted proteins

For the 13 known human examples of non-classical secretion, 10 receive neural network scores exceeding 0.6 using our method (CNTF_HUMAN, FGF1_HUMAN, FGF2_HUMAN, HME2_HUMAN, IL1B_HUMAN, IL18_HUMAN, LEG3_HUMAN, MIF_HUMAN, THIO_HUMAN, THTR_HUMAN) and two (LEG1_HUMAN and HMG1_HUMAN) receive neural network scores less than 0.6. HMG1_HUMAN, a nuclear protein under normal cellular conditions, receives the lowest SecretomeP score. IL1A_HUMAN is only present in the extracellular environment in low concentrations and is predicted as a borderline case. The two FGF examples have significant matches to various FGFs with signal peptides (FGF4, FGF6 and FGF7), which also is the case for IL1B_HUMAN, but the rest have no significant matches to any sequence in the positive part of the training set. Three viral proteins are known to enter the non-classical secretory pathway, HIV-tat, foamy virus Bet and herpes simplex VP22. HIV tat and Bet are clearly predicted to enter the non-classical secretory pathway, whereas the VP22 capsid protein is a borderline case.

Recently, non-classical secretion of Annexin A2 was demonstrated in the small intestinal enterocytes of pigs (Danielsen *et al.*, 2003). Similarly, we expect Annexin A2 of human enterocytes to be secreted by non-classical means. ANX2_HUMAN receives a neural network score of 0.739, well above the threshold for correct classification of a non-classical secreted protein.

FGF-9, -16 and -20 lack the classical secretory N-terminal signal peptide but are indeed efficiently secreted (Miyake *et al.*, 1998; Miyakawa *et al.*, 1999; Ohmachi *et al.*, 2000; Revest *et al.*, 2000). Interestingly, these three proteins are predicted to be secretory using SecretomeP with high confidence. This again shows the power of our prediction method, regardless of whether the secretory protein carries a cleaved N-terminal signal peptide or not.

We also applied the method to the entire human proteome in order to identify novel proteins potentially secreted without N-terminal signal peptides. For this purpose, we used the IPI set of predicted human protein sequences (http://www.ebi.ac.uk/IPI/), removing all sequences predicted to have signal peptides and/or transmembrane helices.

As only a fairly small number of non-classically secreted proteins are expected to exist in the human genome, the score distribution for the predictions on this set of candidate sequences should resemble that of the negative training examples. However, this was clearly not the case as a very large number of proteins received high scores from the neural networks (see Figure 4). When inspecting the predictions, it was clear that most of these high-scoring sequences were short and annotated as hypothetical proteins in IPI. Hence we suspect that a very large fraction of these are in fact conceptual translations of spurious gene predictions. As one of the prominent features of secreted proteins is that they tend to be shorter than other proteins (see Figure 1), the presence of such sequences in the data set presents a serious obstacle for a computational screening as they lead to false-positive predictions. We have addressed this problem by limiting our study to the subset of sequences that show similarity to sequences in the IPI mouse data set, knowing that we will thereby also discard some true protein sequences.

From the search for new putative secretory proteins entering the non-classical secretory pathway, we divided the high-scoring sequences into two groups. Both groups lack SignalP-predicted signal peptides. In the first group we list sequences with homology to entries in Swiss-Prot that carry signal peptides (BLAST E-values $<10^{-10}$ to the SecretomeP training set, Table II). In Table III we list the top-scoring sequences without any functional annotation in IPI or NR. In both lists the sequences have mouse homologs.

The requirements for the first group reduced the IPI set to 237, including immunoglobulins and sequences known from the IPI database to be fragments. By removing known non-classically secreted proteins, immunoglobulins and known fragments, we ended up with a total of 33 sequences, which we predict as being secretory proteins with high confidence. A list of these sequences is presented in Table II. The method predicts, for example, thioredoxin to be secreted. This cytoplasmic protein can also be found in the extracellular surroundings and has previously be found to be secreted via the non-classical pathway (Rubartelli *et al*., 1992).

Another example to highlight is the positive prediction of PAI-2. This protein, lacking an N-terminal signal peptide, but having an internal uncleaved signal peptide, is secreted through the classical secretory pathway (von Heijne *et al*., 1991). Internal uncleaved signal peptides cannot be predicted using SignalP, but SecretomeP correctly classifies the PAI-2 protein as being secreted.

Entries in Table II with initial methionine and match to proteins carrying signal peptides may indeed be proteins secreted via the non-classical secretory pathway, as seen for the fibroblast growth factor family. Only FGF1 and FGF2 of this large family are secreted by non-classical means (Nickel, 2003). The entry 00147465.2 having a disintegrin-like domain and carrying an initial methionine is a good candidate for experimental validation of whether it is actively secreted via non-classical means. Entry 00101605.1 also carries an initial methionine and contains immunoglobulin domains found in different types of cell adhesion molecules (De Juan *et al*., 2002). Entry 00154153.1 shares similarity to plasma kallikrein precursor which usually carries a signal peptide. Entry 00156091.1 displays similarity to a mannose binding C-type lectin. Lectins are indeed a group of proteins known to enter the non-classical secretory pathway (Hughes, 1999; Cooper, 2002).

In eight cases we were able to make a nearly perfect alignment of the predicted secretory IPI sequence to a known protein sequence with an N-terminal signal peptide, except for the N-terminal region. This indicates missing annotation of the signal peptide in the IPI sequences (Reinhardt and Hubbard, 1998), probably due to an automated computer annotation, or it could be novel isoforms of the proteins simply without the signal peptide as in the example of tenascin XB (see below). Features seem to be conserved in extracellular proteins regardless of the secretory pathway they enter.

Table II. Predicted secretory sequences entries from IPI

IPI	NN	Description in IPI/NR
00002785.2	0.928	TNF receptor superfamily member 6 isoform 4
00018926.1	0.904	Granulysin, isoform 519
00002778.2	0.896	TNF receptor superfamily member 6 isoform 7
00010830.1	0.830	Pro-melanin-concentrating hormone-like 1 protein
00003301.1	0.689	Astrocyte-derived trophic factor 2
00009410.1	0.664	Glycosyl hydrolase family 47
00006605.1	0.650	Thioredoxin-related protein
00008412.1	0.621	Disintegrin and metalloproteinase domain 21 preproprotein
00008308.1	0.982	HGC6.3 protein
00154272.1	0.932	TNF (ligand) superfamily, member 6
00147465.2	0.881	A disintegrin-like and metalloprotease
00017554.1	0.871	Similar to sphingomyelin phosphodiesterase precursor
00156091.1	0.868	Probable mannose-binding C-type lectin
00154153.1	0.844	Similar to Plasma kallikrein precursor
00004922.1	0.838	Chymase precursor
00017799.2	0.829	Thioredoxin
00000962.1	0.819	Protein tyrosine kinase
00008869.3	0.804	Pregnancy-specific β_1-glycoprotein 11
00013723.1	0.792	Peptidyl-prolyl *cis–trans* isomerase NIMA-interacting 1
00016958.1	0.772	FGF22
00006498.1	0.750	Deleted in azoospermia 2 protein
00101605.1	0.742	MAM domain containing glycosylphosphatidylinositol anchor 1
00001602.1	0.728	UL16-binding protein 3
00017262.1	0.712	Carbonic anhydrase VA, mitochondrial precursor
00007067.1	0.701	'SCP-like extracellular protein'
00005038.1	0.698	14.5 kDa translational inhibitor protein
00017809.1	0.684	TNF-C
00012668.2	0.669	Serine protease inhibitor
00003011.2	0.654	FGF 12, isoform 2
00016588.1	0.639	Nucleoporin 210
00013371.1	0.635	Sprouty homolog 3
00007117.1	0.611	Plasminogen activator inhibitor-2 precursor
00017590.1	0.605	Superoxide dismutase [Cu–Zn]
00032456.2	0.585	Tenascin XB isoform 2

Thirty-three sequence entries from the IPI database which received a high neural network output score. Only sequences with a neural network score >0.6 and a BLAST E-value $<10^{-10}$ compared with our positive training set are included. Known fragments and immunoglobulins were removed. The first eight sequences align perfectly to their homologous sequences in Swiss-Prot except for the signal peptide, which is absent in the sequences presented to SecretomeP. The major group seems mostly to be extracellular proteins. Tenascin XB isoform 2 is included in the table even if it has a neural network score slightly <0.6, showing alternative translations of two isoforms.

Table III. Novel predicted secretory sequence entries from IPI

IPI	NN	Description in NR
00146298.2	0.950	Similar to KIAA0725 protein
00063270.1	0.927	Hypothetical protein BC008131
00105506.1	0.892	Similar to CG11412-PA
00155049.1	0.883	Similar to hypothetical protein FLJ10948
00099450.1	0.864	Hypothetical protein MGC2560
00146515.2	0.864	Similar to KIAA0454 protein
00054929.2	0.855	Similar to hypothetical protein MGC4276
00157958.1	0.835	Hypothetical protein XP 302443
00099140.1	0.834	Similar to KIAA1285 protein
00065190.1	0.827	Unnamed protein product
00154145.1	0.817	Unknown (protein for MGC57228)
00155209.1	0.805	Unnamed protein product
00102188.2	0.803	Similar to HSPC041 protein
00157201.1	0.782	Unnamed protein product
00156632.1	0.781	Hypothetical gene CG018
00014899.1	0.780	Unnamed protein product
00155040.1	0.778	Similar to KIAA0565 protein
00105917.2	0.776	Similar to hypothetical protein MGC4276
00098105.1	0.767	Hypothetical protein
00042984.3	0.766	Similar to KIAA1596 protein
00106123.1	0.764	Similar to hypothetical protein MGC4827
00097923.1	0.753	Similar to Protein KIAA0685
00097862.2	0.740	Unnamed protein product
00146618.2	0.738	Unnamed protein product
00073380.1	0.735	Hypothetical protein FLJ11811
00145432.2	0.734	Similar to KIAA1641 protein
00036606.1	0.726	Similar to dJ1156J9.1
00032581.1	0.722	Unnamed protein product
00105151.1	0.711	Similar to hypothetical protein MGC10818
00105340.1	0.691	Hypothetical protein FLJ37659

The 30 highest scoring IPI sequences by SecretomeP without functional annotation. In a few cases some of these proteins contain domains which also are found in other protein known to be secreted. All sequence entries carry an initial methionine.

Some of the cases where we were able to perfectly align the IPI sequence to an entry in Swiss-Prot might not always be erroneous annotations in the IPI database, but splice variants of the gene product. This is indeed the case of granulysin isoform 519 as we find this to be a transcript variant with the inclusion of an additional 242 bp segment within intron 1 of the granulysin gene (AC: NP_036615). The inclusion of this additional segment results in the utilization of a different translation start codon. Both isoforms are secreted, but isoform 519 does not contain any signal peptide predicted by SignalP. The other isoform of the granulysin gene (NKG5) has an annotated signal peptide cleavage site at position 15–16; according to SignalP, the cleavage site is located at position 22–23.

The last entry in Table II showing tenascin XB isoform 2 is a truncated version of isoform 1, including the C-terminal 673 residues (AC: NP_115859). This variant (XB-S) is transcribed from a cryptic promoter sequence in intron 32 and thus includes exons 33–45. It encodes isoform 2, which is identical with the C-terminus of the full-length protein, isoform 1. Isoform 1 of tenascin XB does indeed carry an N-terminal signal peptide.

Whether other predicted secretory proteins are splice variants of proteins carrying signal peptides or whether they are erroneous annotations can only be revealed by thorough experimental investigations.

In order to identify completely novel proteins of the non-classical secretory pathway, we investigated human entries in the IPI database with no descriptive annotation and no match to sequences in the NCBI non-redundant sequence database (NR) with well known function. We also discarded sequences lacking the initial methionine as well as sequences shorter than 100 amino acids. A total of 2037 human entries with significant homology to mouse sequences (also with no descriptive annotation) were found. Entries containing N-terminal signal peptides were removed. The resulting 1746 sequence entries were investigated using SecretomeP and the top 30 highest scoring sequences were inspected in detail (all had a SecretomeP score >0.691, indicating a high probability of being a secreted protein). The 30 sequences did not have any significant homology to either the negative or the positive training set (Table III).

Several entries from the list of the 30 highest scoring probable non-classical secretory proteins are worth mentioning. Sequence entry 00063270.1 and 00065190.1 are interesting orphan candidates as they display very little homology to other proteins. Only one BLAST hit with an E-value <2.6 can be found for entry 00063270.1 and two BLAST hits were found with an E-value <0.37 for entry 00065190.1.

Several high-scoring SecretomeP sequences with matches to proteins with known function normally located in the cytoplasmic compartment caught our attention (not included in the table). IPI entry 00154901.1 has very high homology to cytochrome P450. Cytochromes are normally located in mitochondria, thus carrying a mitochondrial transit peptide. As this sequence also is without initial methionine we are not able to determine whether this is due to poor gene finding or whether this protein is actively being translocated without the transit peptide. Another IPI entry 00154495.1 has similarity to the PMS1 human homolog. This protein is involved in mismatch repair of DNA and has an HMG box domain. Even though this is a nuclear protein, proteins carrying similar domains have been shown to enter the non-classical secretory pathway (Gardella *et al*., 2002; Nickel, 2003).

The SecretomeP method presented here complements the highly popular method for detection of classically secreted proteins, SignalP (Nielsen *et al*., 1997). We believe that the SecretomeP method also will have a significant user potential.

Acknowledgements

This work was supported by grants from the Danish National Research Foundation and the Danish Natural Science Research Council and by a grant from Novozymes A/S (to J.D.B.).

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.

Antelmann,H., Tjalsma,H., Voigt,B., Ohlmeier,S., Bron,S., van Dijl,J.M. and Hecker,M. (2001) *Genome Res.*, **11**, 1484–1502.

Bairoch,A. and Apweiler,R. (2000) *Nucleic Acids Res.*, **28**, 45–48.

Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides—SignalP 3.0. *J. Mol. Biol*., in press.

Cooper,D.N.W. (2002) *Biochim. Biophys. Acta*, **1572**, 209–231.

Danielsen,E.M., van Deurs,B. and Hansen,G.H. (2003) *Biochemistry*, **42**, 14670–14676.

De Juan,C., Iniesta,P., Gonzalez-Quevedo,R., Moran,A., Sanchez-Pernaute,A., Torres,A.J., Balibrea, J.L., Diaz-Rubio,E., Cruces,J. and Benito,M. (2002) *Oncogene*, **21**, 3089–3094.

Duckert,P., Brunak,S. and Blom,N. (2004) *Protein Eng. Des. Sel.,* **17**, 107–112.

Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) *J. Mol. Biol.*, **300**, 1005–1016.

Gardella,S., Andrei,C., Ferrera,D., Lotti,L.V., Torrisi,M.R., Bianchi,M.E. and Rubartelli,A. (2002) *EMBO Rep.*, **3**, 995–1001.

Goldstein,G. (1996) *Nature Med.*, **2**, 960–964.

Greenbaum,D., Luscombe,N.M., Jansen,R., Qian,J. and Gerstein,M. (2001) *Genome Res.*, **11**, 1463–1468.

Hughes,R.C. (1999) *Biochim. Biophys. Acta*, **1473**, 172–185.
Jensen,L.J. *et al*. (2002a). *J. Mol. Biol.*, **319**, 1257–1265.
Jensen,L.J., Skovgaard,M. and Brunak,S. (2002b) *Protein Sci.*, **11**, 2894–2898.
Jensen,L.J., Gupta,R., Stærfeldt,H.H. and Brunak,S. (2003) *Bioinformatics*, **19**, 635–642.
Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) *J. Mol. Biol.*, **305**, 567–580.
la Cour,T., Gupta,R., Rapacki,K., Skriver,K., Poulsen, F.M. and Brunak,S. (2003) *Nucleic Acids Res.*, **31**, 393–396.
McGeoch,D.J. (1985) *Virus Res.*, **3**, 271–286.
Miyake,A., Konishi,M., Martin,F.H., Hernday,N.A., Ozaki,K., Yamamoto,S., Mikami,T., Arakawa,T. and Itoh,N. (1998) *Biochem. Biophys. Res. Commun.*, **243**, 148–152.
Miyakawa,K., Hatsuzawa,K., Kurokawa,T., Asada,M., Kuroiwa,T. and Imamura,T. (1999) *J. Biol. Chem.*, **274**, 29352–29357.
Nakai,K. and Horton,P. (1999) *Trends Biochem. Sci.*, **24**, 34–36.
Nakai,K. and Kanehisa,M. (1991) *Proteins*, **11**, 95–110.
Nakashima,H. and Nishikawa,K. (1994) *J. Mol. Biol.*, **238**, 54–61.
Nickel,W. (2003) *Eur. J. Biochem.*, **270**, 2109–2119.
Nielsen,H., Brunak,S., Engelbrecht,J. and von Heijne,G. (1997) *Protein Eng.*, **10**, 1–6.
Ohmachi,S., Watanabe,Y., Mikami,T., Kusu,N., Ibi,T., Akaike,A. and Itoh,N. (2000) *Biochem. Biophys. Res. Commun.*, **277**, 355–360.
Reinhardt,A. and Hubbard,T. (1998) *Nucleic Acids Res.*, **26**, 2230–2236.
Revest,J.M., DeMoerlooze,L. and Dickson,C. (2000) *J. Biol. Chem.*, **275**, 8083–8090.
Rubartelli,A., Bajetto,A., Allavena,G., Wollman,E. and Sitia,R. (1992) *J. Biol. Chem.*, **267**, 24161–24164.
Rubartelli,A. and Sitia,R. (1997) In Kuchler,K., Rubartelli A. and Holland B.I. (eds), *Unusual Secretory Pathways: from Bacteria to Man: Secretion of Mammalian Proteins that Lack a Signal Sequence*. Landes, Austin, TX, pp. 87–114.
Tanudji,M., Hevi,S. and Chuck,S.L. (2002) *J. Cell Sci.*, **115**, 3849–3857.
Thomas,G. (2002) *Nat. Rev. Mol. Cell Biol.*, **3**, 753–766.
Tjalsma,H., Bolhuis,A., Jongbloed,J.D., Bron,S. and van Dijl,J.M. (2000) *Microbiol. Mol. Biol. Rev.*, **64**, 515–547.
von Heijne,G. (1986) *Nucleic Acids Res.*, **14**, 4683–4690.
von Heijne,G., Liljestrom,P., Mikus,P., Andersson,H. and Ny,T. (1991) *J. Biol. Chem.*, **266**, 15240–15243.
Wootton,J.C. and Federhen,S. (1993) *Comput. Chem.*, **17**, 149–163.

Received January 29, 2004; revised April 4, 2004; accepted April 6, 2004

Edited by Andrej Sali