Biochemical and Biophysical Research Communications 391 (2010) 1306–1311

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

SPRED: A machine learning approach for the identification of classical and non-classical secretory proteins in mammalian genomes

Krishna Kumar Kandaswamy [a,b], Ganesan Pugalenthi [c], Enno Hartmann [d], Kai-Uwe Kalies [d], Steffen Möller [a], P.N. Suganthan [c], Thomas Martinetz [a,*]

[a] *Institute for Neuro- and Bioinformatics, University of Lübeck, 23538 Lübeck, Germany*
[b] *Graduate School for Computing in Medicine and Life Sciences, University of Lübeck, 23538 Lübeck, Germany*
[c] *School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore*
[d] *Centre for Structural and Cell Biology in Medicine, Institute of Biology, University of Lübeck, 23538 Lübeck, Germany*

ARTICLE INFO

Article history:
Received 24 November 2009
Available online 6 December 2009

Keywords:
Secretory proteins
Signal peptide
Classical pathway
Non-classical pathway
Random forest
Human proteome

ABSTRACT

Eukaryotic protein secretion generally occurs via the classical secretory pathway that traverses the ER and Golgi apparatus. Secreted proteins usually contain a signal sequence with all the essential information required to target them for secretion. However, some proteins like fibroblast growth factors (FGF-1, FGF-2), interleukins (IL-1 alpha, IL-1 beta), galectins and thioredoxin are exported by an alternative pathway. This is known as leaderless or non-classical secretion and works without a signal sequence. Most computational methods for the identification of secretory proteins use the signal peptide as indicator and are therefore not able to identify substrates of non-classical secretion. In this work, we report a random forest method, SPRED, to identify secretory proteins from protein sequences irrespective of N-terminal signal peptides, thus allowing also correct classification of non-classical secretory proteins. Training was performed on a dataset containing 600 extracellular proteins and 600 cytoplasmic and/or nuclear proteins. The algorithm was tested on 180 extracellular proteins and 1380 cytoplasmic and/or nuclear proteins. We obtained 85.92% accuracy from training and 82.18% accuracy from testing. Since SPRED does not use N-terminal signals, it can detect non-classical secreted proteins by filtering those secreted proteins with an N-terminal signal by using SignalP. SPRED predicted 15 out of 19 experimentally verified non-classical secretory proteins. By scanning the entire human proteome we identified 566 protein sequences potentially undergoing non-classical secretion. The dataset and standalone version of the SPRED software is available at http://www.inb.uni-luebeck.de/tools-demos/spred/spred.

Introduction

After protein synthesis in cytoplasm, newly made polypeptides must be transported to their final destination in the cell. The process of protein transport to a particular cellular location is known as protein sorting [1–3]. Generally, eukaryotic protein secretion occurs via the classical secretory pathway that traverses the endoplasmic reticulum (ER) and Golgi apparatus [4].

Secretory proteins are usually characterized by short N-terminal signal peptides (14–60 amino acids) that have intrinsic signals for their transport and localization in the cell [3,5]. Interestingly, several proteins have been found to be exported directly from the cytoplasm by molecular mechanisms that are independent from a signal peptide or any specific motif known to act as an export signal. The secretion of these proteins is referred to as non-classical or unconventional protein secretion [6–9]. Some of the well studied non-classical secretory proteins are fibroblast growth factors (FGF-1, FGF-2), interleukins (IL-1 alpha, IL-1 beta), galectins, thioredoxin, viral proteins and parasitic surface proteins potentially involved in the regulation of host cell infection [10–14]. Although the phenomenon of non-classical secretion in eukaryotes was discovered more than a decade ago, the molecular mechanisms are still unknown. However, it might be possible that this group contains proteins that leave the cell by cell disruption and not by a well defined pathway.

Several methods have been proposed for the identification of secretory proteins that follow the classical secretory pathway [15,16]. Most prediction methods require the presence of the correct N-terminal end of the preprotein for correct classification. As large scale genome sequencing projects sometimes assign the 5′-end of genes incorrectly, many proteins are annotated without the correct N-terminal end which may lead to an incorrect prediction of subcellular localization [17]. Further, signal peptides are completely absent in secretory proteins that follow non-classical secretion pathways. Therefore, an automated approach is required

[*] Corresponding author. Fax: +49 451 500 5502.
E-mail address: martinetz@inb.uni-luebeck.de (T. Martinetz).

0006-291X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2009.12.019

to predict classical and non-classical secretory proteins, irrespective of the N-terminal signal peptides.

Recently, a webserver SecretomeP has been developed to predict non-classically secreted proteins [18]. It is a neural network based method that uses several features of a protein such as the number of atoms, positively charged residues, propeptide cleavage sites, protein sorting, low complexity regions, and transmembrane helices as an input for a neural network. Despite considering a large number of protein features, the method has achieved a sensitivity of only 40% [18]. SRTPRED is another recently developed method which predicts secretory proteins irrespectively of N-terminal signal peptides. It achieves a sensitivity of 60.4% using hybrid modules [19]. In this work, we report a random forest method, SPRED, to identify classical and non-classical secretory proteins from protein sequence irrespectively of N-terminal signal peptides. We scanned the entire human proteome by SPRED and predicted 566 proteins to be secreted by a non-classical secretory pathway.

Materials and methods

Datasets

Training and test dataset

A set of 9890 extracellular mammalian proteins (positive dataset) were extracted from the Uniprot database based on subcellular localization annotations in the comments block [20]. Partial sequences and sequences without an annotated signal peptide were not included in the data set. Proteins with uncertain annotation labels such as "probable", "potential" and "by similarity" were removed. 3131 extracellular proteins which are annotated with experimental observations were selected from the 9890 proteins. To make the dataset completely non-redundant, we applied the CD-HIT software [21] to remove sequences with greater than 40% sequence similarity to each other. Finally, 780 extracellular proteins were retained for the positive dataset. Similarly, a set of negative examples was constructed by extracting 20,610 mammalian proteins in Uniprot which are annotated as residing in the cytoplasm and/or nucleus. 3891 proteins with experimental support were chosen from the 20,610 proteins after excluding membrane proteins, proteins with uncertain labels, and partial sequences. 1980 sequences remained for the negative dataset after removing redundant sequences which have >40% sequence similarity to each other using CD-HIT [21]. Since non-classical secretory proteins lack N-terminal signal peptides, the method should have the capability to predict secretory proteins irrespective of N-terminal signal peptides. To achieve this, we removed the signal peptides from the positive dataset. Finally, the training dataset consisted of 600 extracellular proteins that form the positive dataset and 600 cytoplasmic and/or nuclear proteins that form the negative dataset. The test dataset consisted of the remaining 180 extracellular proteins and 1380 cytoplasmic and/or nuclear proteins.

Human proteome screening

A human proteome database containing 86845 protein sequences was downloaded from the IPI database release 3.66 (http://www.ebi.ac.uk/IPI/) [22]. Transmembrane proteins were removed using TMHMM [23]. Finally, we obtained 65508 protein sequences for the computational screening and identification of novel putative proteins undergoing either classical or non-classical protein secretion.

Features

In this work, each sequence is encoded by 119 features (provided as a supplement to this paper). We categorized amino acids into 10 functional groups based on the presence of side chain chemical groups such as phenyl (F/W/Y), carboxyl (D/E), imidazole (H), primary amine (K), guanidino (R), thiol (C), sulfur (M), amido (Q/N), hydroxyl (S/T) and non-polar (A/G/I/L/V/P) [24]. Further, we categorized 20 amino acids into three groups, namely hydrophobic (FIWLVMYCA), hydrophilic (RKNDEP) and neutral (THGSQ) amino acid groups.

Frequency of groups

The frequency of the 10 functional groups (number of occurrences of functional group "X" divided by length of the protein) and the frequencies of hydrophobic, hydrophilic, neutral, positively charged, negatively charged, polar and non-polar amino acids were computed for every sequence.

Frequency of tripeptides and short peptides

We utilized tripeptide information for the classification. Generally, 8000 tripeptides can be obtained from all possible combinations of 20 amino acids. To reduce the feature dimension, we derived 27 tripeptides from all possible combinations of the three amino acid groups hydrophobic, hydrophilic and neutral. The frequencies of these 27 tripeptides were calculated for every sequence. Additionally, we incorporated the frequencies of short peptides (10 residue length, in this case) which are rich in hydrophobic, hydrophilic, neutral, polar or non-polar amino acids. For example, a short peptide with more than five hydrophobic residues, we consider as a hydrophobic peptide. Similarly, we calculated hydrophilic, neutral, polar and non-polar short peptides. In addition, we incorporated the frequencies of short peptides which are rich in the 10 functional amino acid groups.

Secondary structure

Secondary structure information for every sequence was assigned using PSIPRED [25]. PSIPRED provides two options for secondary structure prediction. The first option uses homologous sequence information and the second option predicts secondary structure from the query sequence without using homologous sequence information. We employed the second option of the PSIPRED method for all sequences. The overall composition of helix (H), beta sheet (E), coil (C) and the frequencies of 10 functional groups, hydrophobic, hydrophilic and neutral amino acids at helix, sheet, and coil regions were calculated.

Physicochemical properties

Physicochemical properties of amino acids have been successfully employed in many sequence based predictions [24,26,27]. Although there are dozens of physicochemical properties of amino acid, we selected 31 physicochemical properties from the UMBC AAIndex database [28]. For each sequence, a physicochemical property value was calculated as the sum of those values of all amino acids in the given sequence, divided by the number of amino acids in the sequence. Table 1 lists number of feature indices for each feature group.

Random forest classification

The random forest (RF) classification extends the concept of decision trees and has been successfully employed in various biological problems [29–34]. We only give a brief description of the random forest approach. The details can be found in [35–38]. Random forest is a collection of decision trees, where each tree is grown using a subset of the possible attributes in the input feature vector. It has been shown that combining multiple decision trees produced in randomly selected subspaces can improve the generalization accuracy [35]. Random forest constructs an ensemble of decision trees from randomly sampled subspaces of the input space, and the final classification is obtained by combining the re-

Table 1
List of 119 features.

Name of the feature	Number of features
Frequencies of 10 functional groups	10
Frequencies of hydrophobic, neutral, hydrophilic, positive, negative, polar and non-polar amino acids	7
Frequencies of secondary structurally elements (Helix, Strand and Coil)	3
Frequencies of 10 functional groups at Helix, Strand and Coil regions	30
Frequencies of hydrophobic, neutral, hydrophilic, positive, negative, polar and non-polar amino acids at Helix, Strand and Coil regions	21
Frequencies of short peptides rich in 10 functional groups	10
Frequencies of short peptides rich in hydrophobic, neutral, hydrophilic, positive, negative, polar and non-polar amino acids	7
Physicochemical properties	31
Total	119

sults from the trees via voting. The random subspace method is used to avoid overfitting on the training set while preserving the maximum accuracy when training a decision tree classifier [38]. RF performs cross-validation by using out-of-bag (OOB) samples. In training, each tree is constructed using a different bootstrap sample from the original data. Since bootstrapping is sampling with replacement from the training data, some of the sequences will be 'left out' of the sample, while others will be repeated in the sample. The 'left out' sequences constitute the OOB sample. On average, each tree is grown using about $1 - e^{-1} \approx 2/3$ of the training sequences, leaving $e^{-1} \approx 1/3$ as OOB. The RF algorithm was implemented by using the random Forest R package [37].

Feature selection by information gain

To identify the important features that distinguish positive and negative classes, we used the Information Gain algorithm with the ranker method [39], the implementation of Weka 3.5 [40]. The information gain for each feature was calculated and the features were ranked according to this measure. Feature selection was performed by five-fold cross-validation on the training dataset. Different models were built using the 10, 25, 50, 75 and 100 best features.

Prediction assessment

The prediction system is evaluated using accuracy, sensitivity, specificity and Matthew's correlation coefficient (MCC). These measurements are expressed in terms of the fraction of true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). The measurements are defined as follows:

Table 2
Performance of SPRED on the training dataset (600 positive and 600 negative sequences) using different feature subsets.

Feature subset	Sensitivity (%)	Specificity (%)	MCC	Accuracy (%)
10	82.50	85.83	0.6837	84.17
25	85.33	85.83	0.7117	85.58
50	85.67	86.17	0.7183	85.92
75	86.17	85.00	0.7117	85.58
100	86.00	84.17	0.7018	85.08
119	86.50	84.83	0.7134	85.67

MCC, Matthew's correlation coefficient.

Table 3
Performance of SPRED on the test dataset (180 positive and 1380 negative sequences) using different feature subsets.

Feature subset	Sensitivity (%)	Specificity (%)	MCC (%)	Accuracy (%)
10	79.44	80.51	0.4345	80.38
25	83.89	80.94	0.4691	81.28
50	88.33	81.38	0.5036	82.18
75	90.56	81.23	0.5163	82.31
100	89.44	81.16	0.5082	82.12
119	90.56	80.80	0.5109	81.92

MCC, Matthew's correlation coefficient.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{MCC} = \frac{\text{TPTN} - \text{FPFN}}{\sqrt{(\text{TN} + \text{FN})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TP} + \text{FP})}} \tag{4}$$

The Matthew's correlation coefficient ranges from $-1 \leqslant \text{MCC} \leqslant 1$. A value of MCC = 1 indicates the best possible prediction while MCC = −1 indicates the worst possible prediction (or anti-correlation). Finally, MCC = 0 would be expected for a random prediction scheme.

Results and discussion

Classification by SPRED

We trained our random forest model on the training dataset containing 600 extracellular proteins secreted via classical and non-classical pathways and 600 cytoplasmic and/or nuclear proteins. As shown in Table 2, on the training data an overall prediction accuracy of 85.67% with a sensitivity of 86.50% and a specificity of 84.83% was obtained using all features. Then we selected five feature subsets by decreasing the number of features. The prediction rate is improved in each feature selection step. The maximum accuracy of 85.92% with 85.67% sensitivity and 86.17% specificity was obtained using 50 features. This result sug-

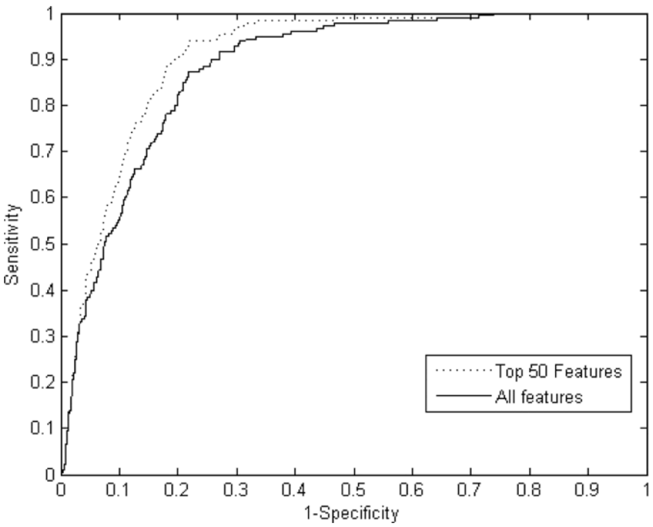

Fig. 1. Receiver operating characteristic (ROC) curves. ROC curves were plotted utilizing sensitivity and specificity values derived from the prediction results of SPRED using the top 50 features and all features.

Table 4
Prediction result for 19 experimentally verified non-classical secretory proteins using SPRED, SecretomeP and SRTPRED. " + " denotes proteins correctly predicted as non-classical secretory proteins and "–" denotes proteins incorrectly predicted as non-classicial secretory proteins.

SwissProt ID	Protein annotation	SPRED	SecretomeP	SRTPRED
P05230	Heparin-binding growth factor 1	+	+	+
P09038	Heparin-binding growth factor 2	+	+	+
P01584	Interleukin 1 beta	+	+	+
P01583	Interleukin 1 alpha	+	+	–
P17931	Galectin-3	+	+	–
P14174	Macrophage migration inhibitory factor	+	+	–
P26447	Protein S100-A4	+	+	–
P09211	Glutathione S-transferase P	+	+	–
Q06830	Peroxiredoxin-1	+	+	–
Q14116	Interleukin 18	+	+	–
P27797	Calreticulin	+	–	+
P62805	Histone H4	+	–	–
P29034	Protein S100-A2	+	–	–
P09382	Galectin-1	+	–	–
P10599	Thioredoxin	+	–	–
P26441	Ciliary neurotrophic factor	–	+	+
P19622	Homeobox protein engrailed-2	–	+	–
Q16762	Thiosulfate sulfurtransferase	–	+	–
P09429	High mobility group protein B1	–	–	–

gests that our feature reduction approach selected useful features by eliminating uncorrelated and noisy features.

In order to examine the performance of the newly developed model, we tested the trained model on a test dataset containing 180 extracellular proteins and 1380 cytoplasmic and/or nuclear proteins. As shown in Table 3, using the top 50 features, we obtained 82.18% accuracy with a sensitivity of 88.33% and a specificity of 81.38%. We also plotted the sensitivity versus specificity chart, i.e. the receiver operator curve (ROC). The area under curve for all features was 0.89 and for the top 50 features was 0.91, respectively (Fig. 1).

Prediction result for known non-classical secretory proteins

For predicting non-classical secretory proteins, we do the following steps. First, SPRED tells us, whether the protein is secretory or non-secretory, and then we look whether the protein has a signal peptide or not. If not, we know that we have a non-classically secreted protein. As a final test we use 19 human proteins that are experimentally verified non-classical secretory proteins from various sources. Criteria for selection were clear experimental evidence within the literature for the given sequence entry. These, secreted but with no signals sequences are not found in any of the above datasets on which SPRED was trained or tested. For comparison, we applied SPRED, SecretomeP [17] and SRTPRED [18] to these 19 proteins. SPRED correctly predicts 15 proteins as non-classical secretory proteins whereas SecretomeP and SRTPRED predict 13 (with low score) and 5 proteins, respectively. The prediction results are given in Table 4.

Screening for classical and non-classical secretory proteins in the human proteome

To identify novel candidates in the human proteome for non-classical secretory proteins, we scanned the human proteome using SPRED (Fig. 2). With SPRED, we classified these 65508 protein sequences into 44611 non-secreted proteins and 20897 proteins located outside of the nucleo-cytoplasm. We removed all the classical secretory proteins (9542 protein sequences) using SignalP, leaving 11355 proteins which do not belong to the classical

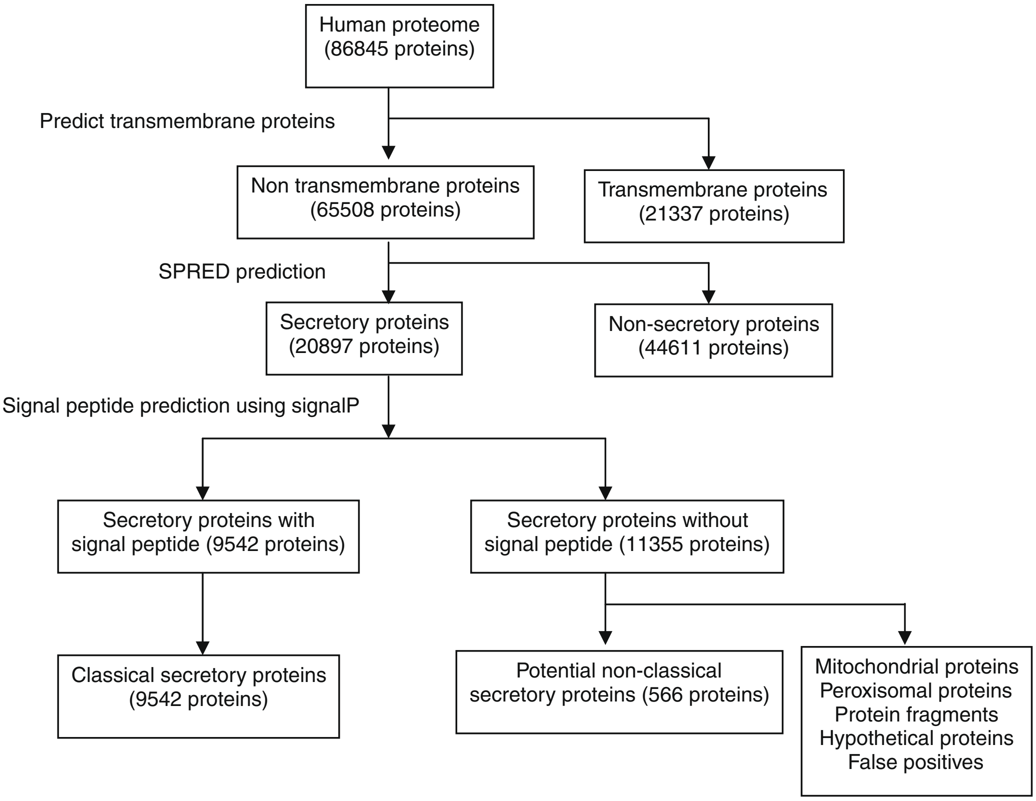
Human proteome (86845 proteins)
Predict transmembrane proteins
Non transmembrane proteins (65508 proteins)
Transmembrane proteins (21337 proteins)
SPRED prediction
Secretory proteins (20897 proteins)
Non-secretory proteins (44611 proteins)
Signal peptide prediction using signalP
Secretory proteins with signal peptide (9542 proteins)
Secretory proteins without signal peptide (11355 proteins)
Classical secretory proteins (9542 proteins)
Potential non-classical secretory proteins (566 proteins)
Mitochondrial proteins
Peroxisomal proteins
Protein fragments
Hypothetical proteins
False positives

Fig. 2. Screening for secretory proteins in human proteome.

Table 5
Comparison of SPRED with other machine learning methods using the top 50 features.

Method	Sensitivity (%)	Specificity (%)	MCC	Accuracy (%)
Na Bayes	70.00	78.28	0.2639	77.79
IBK	57.50	82.34	0.2344	80.88
SVM(Linear)	82.78	82.90	0.4867	82.88
SVM(RBF kernel)	78.89	80.87	0.4351	80.64
SPRED	88.33	81.38	0.5036	82.18

secretory pathway. Subsequently, we removed hypothetical proteins, fragmented proteins, mitochondrial proteins, peroxisomal proteins and false positive proteins. The remaining 566 protein sequences were finally classified as non-classical secretory proteins. Our analysis shows that these 566 proteins include well studied non-classical secretory proteins such as Galectin [8], Interleukin 1 alpha, Interleukin 1 beta [9], thioredoxin [41], S100-A [42], etc. which leave intact cells by defined pathways. However, as the classification of proteins in the training dataset into the positive dataset "extracellular proteins" is often based on the detection of these proteins outside of cells without any knowledge about the export pathway, these predicted proteins may also include proteins that are released during cell disruption and are relatively stable in the extracellular environment. The complete list of predicted non-classical secretory proteins is provided in the supplementary materials.

Comparison of SPRED with other machine learning methods

The proposed SPRED method was compared with several state-of-the-art classifiers such as the Naïve Bayes classifier [43], instance learning based IBK algorithm [44] and the Support Vector Machine (linear and RBF kernel) [45]. The optimal values of the SVM parameters were obtained using a five-fold cross-validation on the training dataset. We compared the performance of SPRED with the other models using the same feature subsets that are mentioned in Table 2. All models were tested on the test dataset containing 180 positive and 1380 negative sequences. With the top 50 features, SPRED and SVM (linear and RBF kernel) achieved comparable accuracy and specificity, however, the sensitivity of SPRED is still higher (Table 5).

Conclusion

Protein secretion is a universal process which occurs in all organisms and has tremendous importance to biological research. Identification of classical and non-classical proteins is an essential and also difficult task.

We implemented a random forest approach to predict protein secretion using sequence derived properties. The validation of SPRED on a test dataset showed 82.18% accuracy with a sensitivity of 88.33% and a specificity of 81.38%. SPRED performed better than SecretomeP and SRTPRED. The next challenge will be to verify the predicted non-classical proteins experimentally.

Acknowledgments

This work was supported by the Graduate School for Computing in Medicine and Life Sciences funded by Germany's Excellence Initiative [DFG GSC 235/1]. GP and PNS acknowledge the financial support offered by the A∗Star (Agency for Science, Technology and Research).

References

[1] G. Palade, Intracellular aspects of the process of protein synthesis, Science 189 (1975) 347–358.

[2] J.E. Rothman, F.T. Wieland, Protein sorting by transport vesicles, Science 272 (1996) 227–234.

[3] P. Walter, R. Gilmore, G. Blobel, Protein translocation across the endoplasmic reticulum, Cell 38 (1984) 5–8.

[4] G. Schatz, B. Dobberstein, Common principles of protein translocation across membranes, Science 271 (1996) 1519–1526.

[5] G. Heijne, The signal peptide, J. Membr. Biol. 115 (1990) 195–201.

[6] A. Müsch, E. Hartmann, K. Rohde, A. Rubartelli, R. Sitia, T.A. Rapoport, A novel pathway for secretory proteins?, Trends Biochem Sci. 15 (1990) 86–88.

[7] A.E. Cleves, Protein transports: the nonclassical ins and outs, Curr. Biol. 7 (1997) R318–R320.

[8] R.C. Hughes, Secretion of the galectin family of mammalian carbohydratebinding proteins, Biochim. Biophys. Acta 1473 (1999) 172–185.

[9] W. Nickel, The mystery of nonclassical protein secretion, Eur. J. Biochem. 270 (2003) 2109–2119.

[10] P. Mignatti, D.B. Rifkin, Release of basic fibroblast growth factor, an angiogenic factor devoid of secretory signal sequence: a trivial phenomenon or a novel secretion mechanism?, J Cell. Biochem. 47 (1991) 201–207.

[11] A. Rubartelli, A. Bajetto, G. Allavena, E. Wollman, R. Sitia, Secretion of thioredoxin by normal and neoplastic cells through a leaderless secretory pathway, J. Biol. Chem. 267 (1992) 24161–24164.

[12] B. Mehul, R.C. Hughes, Plasma membrane targetting, vesicular budding and release of galectin 3 from the cytoplasm of mammalian cells during secretion, J. Cell. Sci. 110 (1997) 1169–1178.

[13] P.W. Denny, S. Gokool, D.G. Russell, M.C. Field, D.F. Smith, Acylation-dependent protein export in Leishmania, J. Biol. Chem. 275 (2000) 11017–11025.

[14] L.C. Trotman, D.P. Achermann, S. Keller, M. Straub, U.F. Greber, Non-classical export of an adenovirus structural protein, Traffic 4 (2003) 390–402.

[15] J.D. Bendtsen, H. Nielsen, A. Krogh, G. von Heijne, S. Brunak, Improved prediction of signal peptides: SignalP 3.0, J. Mol. Biol. 340 (2004) 783–795.

[16] C. Guda, pTARGET: a web server for predicting protein subcellular localization, Nucleic Acids Res. 34 (2006) W210–213.

[17] A. Reinhardt, T. Hubbard, Using neural networks for prediction of the subcellular location of proteins, Nucleic Acids Res. 26 (1998) 2230–2236.

[18] J.D. Bendtsen, L.J. Jensen, N. Blom, G.V. Heijne, S. Brunak, Feature-based prediction of non-classical and leaderless protein secretion, Protein Eng. Des. Sel. 17 (2004) 349–356.

[19] A. Garg, G.P.S. Raghava, A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search, Silico. Biol. 8 (2008) 1–12.

[20] A. Bairoch, R. Apweiler, The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000, Nucleic Acids Res. 28(1) (2000) 45–48.

[21] W. Li, L. Jaroszewski, A. Godzik, Clustering of highly homologous sequences to reduce the size of large protein database, Bioinformatics 17 (2001) 282–283.

[22] P.J. Kersey, J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney, R. Apweiler, The International Protein Index: An integrated database for proteomics experiments, Proteomics 4 (7) (2004) 1985–1988.

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

[24] G. Pugalenthi, K.K. Kumar, P.N. Suganthan, R. Gangal, Identification of catalytic residues from protein structure using support vector machine with sequence and structural features, Biochem. Biophys. Res. Commun. 367 (2008) 630–634.

[25] L.J. McGuffin, K. Bryson, D.T. Jones, The PSIPRED protein structure prediction server, Bioinformatics 16 (2000) 404–405.

[26] M. Bhasin, G.P.S. Raghava, ESLpred: SVM based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST, Nucleic Acids Res. 32 (2004) W414–419.

[27] G. Pugalenthi, K. Tang, P.N. Suganthan, G. Archunan, R. Sowdhamini, A machine learning approach for the identification of odorant binding proteins from sequence-derived properties, BMC Bioinformatics 8 (2007) 351.

[28] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, AAindex: amino acid index database, progress report 2008, Nucleic Acids Res. 36 (2008) D202–205.

[29] S. Dudoit, J. Fridlyan, T.P. Fridlyan, Comparison of discrimination methods for the classification of tumors using gene expression data, J. Am. Stat. Assoc. 97 (2002) 77–87.

[30] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, H. Zhao, Comparison of statistical methods for classification of ovarian cancer using a proteomics dataset, Bioinformatics 19 (2003) 1636–1643.

[31] J.W. Lee, J.B. Lee, M. Park, S.H. Song, An extensive comparison of recent classification tools applied to microarray data, Comput. Stat. Data Anal. 48 (2005) 869–885.

[32] Y. Qi, J.K. Seetharaman, Z.B. Joseph, Random forest similarity for protein–protein interaction prediction from multiple sources, Pac. Symp. Biocomput. (2005) 531–542.

[33] R.D. Uriarte, S.A. Andres, Gene selection and classification of microarray data using. Random forest, BMC Bioinformatics 3 (2006).

[34] A. Statnikov, L. Wang, C.F. Aliferis, A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification, BMC Bioinformatics 9 (2008) 319.

[35] T.K. Ho, Data complexity analysis of comparative advantages of decision forest constructors, Pattern Anal. Appl. 5 (2002) 102–112.

[36] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.

[37] A. Liaw, M. Wiener, Classification and regression by random forest, R. News. 2 (2002) 18–22.

[38] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 832–844.

[39] T.M. Mitchell, Machine Learning, McGraw-Hill, 1997.

[40] E. Frank, M. Hall, L. Trigg, G. Holmes, I.H. Witten, Data mining in bioinformatics using Weka, Bioinformatics 20 (2004) 2479–2481.

[41] A. Ubartelli, A. Bajetto, G. Allavena, E. Wollman, R. Sitia, Secretion of thioredoxin by normal and neoplastic cells through a leaderless secretory pathway, J. Biol. Chem. 267 (34) (1992) 24161–24164.

[42] M. Landriscina, R. Soldi, C. Bagala, I. Micucci, S. Bellum, F. Tarantini, I. Prudovsky, T. Maciag, S100A13 participates in the release of fibroblast growth factor 1 in response to heat shock in vitro, J. Biol. Chem. 276 (2001) 22544–22552.

[43] H. J. George, P. Langley, Estimating continuous distributions in bayesian classifiers, Eleventh Conf. Uncertainty Artif. Intell. San Mateo (1995) 338–345.

[44] D. Aha, D. Kibler, Instance-based learning algorithms, Machine Learning 6 (1991) 37–66.

[45] V. Vapnik, The Nature of Statistical Learning Theory, first ed., Springer, NY, 1995.