BioPublisher©
Computational Molecular Biology 2016, Vol.6, No.4, 1-12
http://cmb.biopublisher.ca

Research Article Open Access

ProtSecKB: The Protist Secretome and Subcellular Proteome Knowledgebase

Brian Powell[1], Vamshi Amerishetty[2], John Meinken[2,3], Geneva Knott[4], Feng Yu[1], Chester Cooper[2,4], Xiang Jia Min[2,4]

1 Department of Computer Science & Information Systems, Youngstown State University, Youngstown, OH 44555, USA

2 Center for Applied Chemical Biology, Youngstown State University, Youngstown, OH 44555, USA

3 Center for Health Informatics, University of Cincinnati, Cincinnati, OH 45267-0840, USA

4 Department of Biological Sciences, Youngstown State University, Youngstown, OH 44555, USA

Corresponding author email: xmin@ysu.edu

Computational Molecular Biology, 2016, Vol.6, No.4 doi: 10.5376/cmb.2016.06.0004

Received: 19 Aug., 2016

Accepted: 01 Nov., 2016

Published: 14 Dec., 2016

Preferred citation for this article:

Powell B., Amerishetty V., Meinken J., Knott G., Feng Y., Cooper C., and Min X.M., 2016, ProtSecKB: the protist secretome and subcellular proteome knowledgebase, Computational Molecular Biology, 6(4): 1-12 (doi: 10.5376/cmb.2016.06.0004)

Abstract Kingdom Protista contains a large group of eukaryotic organisms with diverse lifestyles. We developed the Protist Secretome and Subcellular Proteome Knowledgebase (ProtSecKB) to host information of curated and predicted subcellular locations of all protist proteins. The protist protein sequences were retrieved from UniProtKB, consisting of 1.97 million entries generated from 7,024 species with 101 species including 127 organisms having complete proteomes. The protein subcellular locations were based on curated information and predictions using a set of well evaluated computational tools. The database can be searched using several different types of identifiers, gene names or keyword(s). Secretomes and other subcellular proteomes can be searched or downloaded. BLAST searching against the complete set of protist proteins or secretomes is available. Protein family analysis of secretomes from representing protist species, including *Dictyostelium discoideum*, *Phytophthora infestans*, and *Trypanosoma cruzi*, showed that species with different lifestyles had drastic differences of protein families in their secretomes, which may determine their lifestyles. The database provides an important resource for the protist and biomedical research community. The database is available at http://bioinformatics.ysu.edu/secretomes/protist/index.php.

Keywords Computational Prediction; Protest; Protista; Secreted Protein; Secretome; Signal Peptide; Subcellular Location; Subcellular Proteome; Lifestyle

1 Introduction

Protists consist of a large number of diverse eukaryotic organisms that are not classified into the kingdoms of Fungi, Plantae, or Animalia (Foissner, 1999, 2006; Slapeta et al., 2005). Some protists are parasites of animals and humans, such as *Plasmodium falciparum* causing malaria, and many others cause similar diseases in other vertebrates (D'Acremont et al., 2010). The oomycete *Phytophthora infestans* causes late blight in potato and tomato plants (Nowicki et al., 2012). Understanding the metabolism of these protists and their roles in ecology may allow these diseases to be treated more efficiently.

In eukaryotes, proteins are synthesized within a cell and then transported to different subcellular locations including extracellular space or matrix to perform their biological functions. Identification and analysis of protein subcellular locations in eukaryotes is one of the important subjects for annotating a proteome. The term secretome is often used to describe the set of proteins secreted outside of a cell (Lum and Min, 2011). The parasite *P. falciparum* causes malaria by replicating inside red blood cells of infected individuals. Secreted proteins of *P. falciparum* were identified and experimentally examined (Przyborski and Lanzer, 2004; Hiller et al., 2004; Van Ooij et al., 2008). These secreted proteins are potential targets for drug treatment of the malaria disease (Bhatt, 2012).

Classical eukaryotic secreted proteins contain a secretory signal peptide at the N-terminus (von Heijne, 1990). Classical secreted proteins of eukaryotes can be computationally predicted accurately with our developed computational protocols of combining multiple prediction tools (Min, 2010). Thus we have made efforts to

construct secretome databases for fungi, plants, and animals (Lum and Min, 2011; Lum et al., 2014; Meinken et al., 2014; Meiken et al., 2015). In this work, we describe the Protist Secretome and Subcellular Proteome Knowledgebase (ProtSecKB, http://bioinformatics.ysu.edu/secretomes/protist/index.php). The database will serve a useful resource for the community working with protist organisms for biomedical research.

2 Methods of Database Construction

2.1 Data collection

The protist protein sequences were retrieved from the UniProtKB/Swiss-Prot dataset and the UniProtKB/TrEMBL dataset (release 2016-02) (http://www.uniprot.org/downloads) using our in-house script. As proteins in the Kingdom Protista are actually not labeled as "Protist" or "Protista", we retrieved all entries belonging to "Eukaryota" but not further classified as "Fungi", "Metazoa", or "Viridiplantae". The UniProtKB/Swiss-Prot dataset contains manually annotated and reviewed protein sequences. The UniProtKB/TrEMBL dataset contains computationally analyzed protein sequences. The combined protist dataset consisted of a total of 1,970,022 protein entries with 8,661 and 1,961,361 entries retrieved from the Swiss-Prot dataset and the TrEMBL dataset, respectively. The identifier mapping data including UniProt accession number (AC), UniProt ID, RefSeq accession number, and gi number were retrieved from the UniProt ID mapping data file. All data used in the database construction and analysis can be downloaded from the website at http://proteomics.ysu.edu/publication/data/ProtSecKB/.

2.2 Prediction of protein subcellular locations

As similar approaches to using the same set prediction tools have been employed in construction of FunSecKB (Lum and Min, 2011), FunSecKB2 (Meinken et al., 2014), PlantSecKB (Lum et al., 2014), and MetazSecKB (Meinken et al., 2015) in our group, we only briefly introduce these tools in this work. For detailed information, the relevant references for each tool or the exemplar introduction by Lum and Min (2013) can be consulted. The software tools used in this work include SignalP (version 4.0), TargetP, Phobius, WoLF PSORT, TMHMM, and PS-Scan. In brief, SignalP 4.0 was used for secretory signal peptide prediction (Petersen et al., 2011). However, we also included prediction information from SignalP 3.0 (Bendtsen et al., 2004) as it provides more accurate cleavage site prediction than SignalP 4.0 (Petersen et al., 2011). Phobius is a combined signal peptide and a transmembrane topology predictor (Käll et al., 2007). TargetP predicts the presence of any signal sequences such as signal peptide (SP), chloroplast transit peptide (cTP), or mitochondrial targeting peptide (mTP) in the N-terminus (Emanuelsson et al., 2007). TMHMM predicts the presence and topology of transmembrane helices and their orientation to the membrane (in/out) (Krogh et al., 2001). PS-Scan was used to scan the PROSITE database (http://www.expasy.org/tools/scanprosite/) for identifying ER targeting proteins (Prosite: PS00014) (Sigrist et al., 2010). WoLF PSORT predicts multiple subcellular locations including cytosol, cytoskeleton, ER, extracellular (secreted), Golgi apparatus, lysosome, mitochondria, nuclear, peroxisome, plasma membrane, and vacuolar membrane (Horton et al., 2007). As for all these programs, there were no specific parameters available for protists yet, the default parameters for eukaryotes or fungi, if available, were used, based on our previous evaluation (Min, 2010). We took the following procedure to assign a protein subcellular location. The annotated subcellular location in UniProtKB and our manual curation take precedence over computational prediction. Thus, only proteins not having an annotated subcellular location are subjected to computational assignment. However, the prediction information generated by all the tools is available for all proteins. It should be noted that some of the proteins may have more than one subcellular location.

Membrane proteins: A membrane protein is a protein having one or more transmembrane domains predicted by TMHMM. However, if there is only one transmembrane domain predicted and located within the N-terminus 70 amino acids, and also a signal peptide is predicted by SignalP 4.0, then this protein is not counted as a membrane protein.

Mitochondrial proteins: Assignment of mitochondrial proteins was based on WoLF PSORT prediction. If it is

also classified as a membrane protein, then it is further classified as a mitochondrial membrane protein.

ER proteins: Proteins predicted to contain a signal peptide by SignalP 4.0 and an ER target signal (Prosite: PS00014) by PS-Scan were treated as luminal ER proteins.

Secretomes: A secretome is all secreted proteins from a species. There were four subcategories of secreted proteins. Curated secreted proteins include proteins which are annotated to be "secreted" or "extracellular" or "cell wall" in the subcellular location from the UniProtKB/Swiss-Prot data set which are "reviewed" as well as manually collected secreted proteins from recent literature by our curators. "Highly likely secreted" proteins are predicted to have a secretory signal peptide by at least three of the four predictors including SignalP 4.0, Phobius, TargetP and WoLF PSORT, but are not classified as any of the above categories. "Likely secreted" proteins are predicted to have a secretory signal peptide by two of the four predictors, and "Weakly likely secreted" proteins are predicted to have a secretory signal peptide by one of the four predictors. We recommend combining both curated and highly likely secreted proteins as a secretome for a species (see accuracy evaluation section).

Proteins in other subcellular locations: Other subcellular locations - including cytosol (cytoplasm), cytoskeleton, Golgi apparatus, lysosome, nucleus, peroxisome, plasma membrane and vacuole - were predicted by WoLF PSORT. It should be noted that we did not predict the category of plastid proteins and all entries in this category were from UniProtKB curation.

2.3 Prediction accuracy evaluation of protein subcellular locations

The prediction tools we chose above were based on our previous evaluation (Min, 2010). To further evaluate the prediction accuracy of each subcellular location in this dataset, we retrieved protein entries having an annotated, unique subcellular location from UniProtKB/Swiss-Prot dataset. Proteins having multiple subcellular locations, labeled as "fragment", not starting with "M", or having a length < 70 amino acids were excluded. Proteins with a subcellular location having a term including "By similarity", "Probable", or "Potential" were excluded. The prediction accuracy for each subcellular location was evaluated using prediction sensitivity (Equation 1), specificity (Equation 2) and Matthews Correlation Coefficient (MCC) (Equation 3).

$$\text{Sensitivity (\%)} = TP/(TP + FN) \times 100 \quad (1)$$

$$\text{Specificity (\%)} = TN/(TN + FP) \times 100 \quad (2)$$

$$\text{MCC (\%)} = (TP \times TN - FP \times FN) \times 100 / ((TP + FP)(TP + FN)(TN + FP)(TN + FN))1/2 \quad (3)$$

TP is the number of true positives, FN is the number of false negatives, FP is the number of false positives, and TN is the number of true negatives. The MCC takes into account true and false positives and negatives and is generally regarded as a balanced measure, with +1 representing a perfect prediction and 0 meaning no better than random chance (Matthews, 1975). The dataset contains a total of 2,407 proteins. For each category, the number of actual positives equals TP plus FN and the number of actual negatives equals FP plus TN (Table 1).

3 Results

3.1 Prediction accuracy

3.1.1 Mitochondrial proteins

The prediction accuracy results for each subcellular location are shown in Table 1. As both TargetP and WoLF PSORT can predict mitochondrial proteins, we evaluated the prediction accuracy of these two tools both individually and combined (Table 1a). When an individual tool was used, WoLF PSORT prediction showed a much higher sensitivity but a slightly lower specificity than TargetP prediction. Thus, the MCC value was higher using WoLF PSORT (0.53) than using TargetP (0.32). If only positives predicted by both tools were used, the specificity was slightly increased but the sensitivity decreased. In contrast, including positives predicted by either tool increased the sensitivity but decreased the specificity resulting in a lower MCC value (0.50) than using WoLF PSORT alone. Thus, we based our predictions for mitochondrial proteins on WoLF PSORT alone.

Table 1 Prediction accuracy evaluation of protist protein subcellular locations

	TP	FP	TN	FN	Sn	Sp	MCC
(a) Mitochondrial proteins							
TargetP	136	89	1823	359	27.5	95.3	0.32
WoLF PSORT	278	133	1779	217	56.2	93.0	0.53
TargetP AND WoLF PSORT	118	35	1877	377	23.8	98.2	0.36
TargetP OR WoLF PSORT	296	188	1724	199	59.8	90.2	0.50
(b) Secreted proteins							
Secreted	99	50	2211	47	67.8	97.8	0.65
S + HLS	130	85	2176	16	89.0	96.2	0.71
S + HLS + LS	137	121	2140	9	93.8	94.6	0.68
S+ HLS + LS + WLS	138	280	1981	8	94.5	87.6	0.52
(c) Other subcellular locations							
Cytoplasm	322	167	1714	204	61.2	91.1	0.54
Cytoskeleton	62	13	2180	152	29.0	99.4	0.46
ER	12	26	2254	115	9.4	98.9	0.15
Golgi	0	4	2345	58	0.0	99.8	0.00
Lysosome	0	0	2379	28	0.0	100.0	
Nucleus	466	514	1348	79	85.5	72.4	0.49
Peroxisome	1	2	2381	23	4.2	99.9	0.11
Plasma membrane	18	149	2046	194	8.5	93.2	0.02
Vacuole	0	0	2375	32	0.0	100.0	

Note: TP: true positives; FP: false positives; TN: true negatives; FN: false negatives. Sn: sensitivity; Sp: specificity; MCC: Matthews correlation coefficient. Secreted: predicted by 4 predictors; HLS: highly likely secreted, predicted by 3 out of 4 predictors; LS: likely secreted, predicted by 2 out of 4 predictors; WLS: weakly likely secreted, predicted by 1 out of 4 predictors

3.1.2 Secreted proteins

Our previous evaluation showed that secreted protein prediction accuracy can be improved by removing transmembrane proteins and ER resident proteins (Min, 2010). As we employed four tools - SignalP, TargetP, WoLF PSORT, and Phobius - for predicting secreted proteins or secretory signal peptides, we had to determine which should be included in the secretome set. After removing transmembrane proteins and ER proteins, the protein set predicted to be secreted are divided into four categories: (1) Secreted: predicted by 4 predictors; (b) Highly likely secreted (HLS): predicted by 3 out of 4 predictors; (3) Likely secreted (LS): predicted by 2 out of 4 predictors; and (4) Weakly likely secreted (WLS): predicted by 1 out of 4 predictors. The dataset consisted of 146 curated secreted proteins as positives and 2,261 proteins located in other subcellular locations as negatives. The accuracy results are shown in Table 1b.

As expected, when only entries were predicted by all four tools to be positives as true positives, the prediction specificity was highest. However, the sensitivity was lowest. On the other hand, when including all entries predicted by any of the four tools to be positives as true positives, the prediction specificity was decreased while the sensitivity was increased. Based on the MCC values, the most accurate prediction (0.71) for a secretome includes secreted entries predicted by at least three out of four predictors with a specificity of 96.2% and a sensitivity of 89.0% (Table 1b). Thus, we recommend including only curated secreted proteins and highly likely secreted proteins for estimating the secretome size for a species. It should be noted that both entries predicted by 4 of 4 tools and 3 of 4 tools were assigned to the category of highly like secreted in the database.

3.1.3 Proteins in other subcellular locations

The prediction accuracy results for proteins located in cytoplasm, cytoskeleton, ER, Golgi apparatus, lysosome,

nucleus, peroxisome, plasma membrane, and vacuole are shown in Table 1c. Proteins for the cytoplasm subset also include cytosol as these two terms are used interchangeably in the UniProtKB annotation. The annotated cytoskeleton entries are also annotated as cytoplasm in UniProtKB. However, in our evaluation cytoskeleton proteins were not counted in the subset of cytoplasm. We would also like to point out that plasma membrane proteins were annotated as "cell membrane" in UniProtKB, thus cell membrane proteins were retrieved for evaluating the category of plasma membrane. The prediction accuracies for these subcellular locations vary significantly. Predictions of proteins located in cytoplasm, cytoskeleton, and nucleus were relatively accurate with a MCC value of 0.54, 0.46 and 0.49, respectively. The specificities for cytoskeleton, ER, and peroxisome predictions were high (> 98%), but the sensitivities were low (< 50%). There were no positives predicted for proteins localized in Golgi, lysosome or vacuoles. These results showed there is a need to train the predictors with protist-specific proteins for protist protein subcellular location prediction.

3.2 Overview of subcellular proteome distribution in different species

ProtSecKB contains a total of 1.97 million protein sequences generated from 7,024 protist species including 101 unique species with some of them having multiple strains totaling 127 organisms with complete proteomes. The main categories of subcellular proteomes - including highly likely secreted and likely secreted, cytoplasm, plasma membrane, mitochondrial, and nuclear proteins - for species having complete proteomes are summarized in Table 2. Curated secreted proteins, ER proteins, etc. are not included but can be obtained from the website mentioned above in the Data section (Supplementary Table 1). There are not many proteins with curated subcellular locations in protist species. The curated secreted proteins were mainly from *D. discoideum* with 113 proteins. *D. discoideum* is a soil-living amoeba belonging to the phylum Amoebozoa and commonly referred to as slime mold (Bakthavatsalam and Gomer, 2010). We also curated 29 secreted proteins in *P. falciparum*, a protozoan parasite causing malaria in humans (Singh et al., 2009; Soni et al., 2016).

The species in Protista kingdom have quite variable proteome sizes - from about 5000 proteins in *P. falciparum* to over 50,000 in *Trypanosoma cruzi*, a parasitic euglenoid protozoan causing Chagas' disease in humans (Bern et al. 2011) (Table 2). The distribution of subcellular proteomes varied tremendously in different species, with nucleus, cytoplasm, mitochondria representing the larger subcellular compartments. There were from 14.4% to 77.0% proteins located in the nucleus, from 7.4% to 40.3% in mitochondria, from 4.6% to 35.4% in cytoplasm, and 0.8% to 15.2% secreted. On average for all protist species with complete proteomes, approximately 44% of proteins were located in the nuclear compartment, 22% in mitochondria, 17% in cytoplasm, and 6% secreted outside the plasma membrane of the cell (Table 2).

3.3 Comparative protein family analysis of protist secretomes

Complete comparative evolutionary analyses of protist secretomes or other sub-proteomes were beyond the scope of this study. As complete secretome or other sub-proteome sequences can be downloaded directly from our database, researchers with their specific aims can carry out further detailed comparative study of these sub-proteomes in different species of their interest. However, we performed an rpsBLAST search against the Pfam database for all predicted curated secreted, highly likely secreted and likely secreted proteins (Supplementary Table 2). Here we only included Pfams of the secretomes from the highly likely secreted and curated secreted protein sets of three species to demonstrate the functional diversities of the secreted proteins in protists (Table 3).

The three species were *D. discoideum*, a soil-living amoeba; *P. infestans*, a plant pathogen; and *T. cruzi*, a human parasite. *D. discoideum* had 832 secreted proteins with 388 of them with a Pfam, *P. infestans* had 1,748 secreted proteins with 583 of them with a Pfam, and *T. cruzi* had 4,122 secreted proteins with 1,599 of them with a Pfam (Table 3). The distribution of protein families having at least 6 members in each family was listed in Table 3 and a complete list of data can be downloaded (Supplementary Table 3). In different protist species, not only the total numbers of secreted proteins were different but also the categories of protein families as well as the number of members in each family were vastly different (Table 3).

Table 2 Summary of the protein distribution in some major subcellular locations in different protist species

Specie	Total	HLS	LS	Cyt	Plasm	Mt mem	Mt non-m	Nuc mem	Nuc non-m	Sec(%)
Alveolata; Apicomplexa										
Babesia bigemina	5227	379	116	1452	341	119	997	244	1697	7.3
Babesia bovis	6268	730	178	1302	514	102	1015	71	2611	11.6
Cryptosporidium parvum	6388	426	142	1011	387	168	504	168	3952	6.7
Eimeria acervulina	6909	481	332	928	342	116	1249	125	2895	7.0
Eimeria brunetti	8720	805	515	848	325	155	1365	271	3766	9.2
Eimeria maxima	6132	420	264	883	303	124	1039	91	2686	6.8
Eimeria mitis	10072	986	682	879	317	217	1502	360	4289	9.8
Eimeria necatrix	8623	711	450	918	519	208	1762	221	3149	8.2
Eimeria praecox	7642	683	508	670	246	136	1005	258	3693	8.9
Eimeria tenella	9299	729	432	1185	482	174	2112	133	3543	7.8
Gregarina niphandrodes	6342	316	169	1816	400	67	1161	29	2348	5.0
Hammondia hammondi	8016	588	239	1123	764	179	1948	82	3024	7.3
Neospora caninum	10101	730	362	1505	963	194	2281	116	3697	7.2
Plasmodium berghei	15967	537	379	1173	404	754	2900	1024	8086	3.4
Plasmodium chabaudi	20341	840	440	1833	450	816	3651	1221	10472	4.1
Plasmodium chabaudi chabaudi	5539	507	144	467	176	229	562	594	3222	9.2
Plasmodium cynomolgi strain B	5713	324	92	632	320	129	611	417	3525	5.7
Plasmodium falciparum FCH/4	5762	206	99	440	179	246	590	605	3516	3.6
Plasmodium falciparum IGH-CR14	5035	220	116	448	167	214	516	472	3028	4.4
Plasmodium falciparum MaliPS096_E11	6301	271	130	452	228	281	781	651	3684	4.3
Plasmodium falciparum NF135/5.C10	6324	266	146	458	222	331	832	603	3600	4.2
Plasmodium falciparum Santa Lucia	6175	265	145	435	231	300	772	594	3584	4.3
Plasmodium falciparum Tanzania	6699	259	137	493	235	317	803	649	3990	3.9
Plasmodium falciparum UGT5.1	5904	234	125	461	201	265	694	601	3443	4.0
Plasmodium falciparum Vietnam Oak-Knoll	6218	270	129	447	204	288	751	609	3658	4.3
Plasmodium fragile	5683	357	110	717	383	140	662	328	3217	6.3
Plasmodium inui San Antonio 1	5828	345	95	485	332	124	636	394	3709	5.9
Plasmodium knowlesi	5682	453	124	586	343	144	621	301	3321	8.0
Plasmodium reichenowi	5744	283	148	460	297	347	557	525	3151	4.9
Plasmodium vinckei petteri	5206	429	114	432	156	186	528	508	3183	8.2
Plasmodium vivax Brazil I	6398	377	119	722	429	211	737	676	3495	5.9
Plasmodium vivax India VII	6588	380	115	717	451	193	726	701	3652	5.8
Plasmodium vivax Mauritania I	6297	388	108	680	428	204	732	676	3422	6.2
Plasmodium vivax North Korean	6616	375	121	739	446	205	759	695	3611	5.7
Plasmodium yoelii 17X	6833	397	153	489	201	305	662	1229	3725	5.8
Plasmodium yoelii yoelii	7911	289	174	610	140	296	1043	1297	3822	3.7
Theileria equi strain WA	5313	810	134	991	425	75	702	197	2486	15.2
Toxoplasma gondii FOU	10116	656	342	1395	799	215	2727	107	3765	6.5
Toxoplasma gondii GAB2-2007-GAL-DOM2	9135	620	265	1237	782	192	2392	101	3430	6.8
Toxoplasma gondii MAS	10005	644	342	1372	792	216	2655	105	3755	6.4
Toxoplasma gondii p89	9698	632	326	1307	791	220	2592	103	3611	6.5
Toxoplasma gondii RUB	10027	654	349	1391	804	225	2686	100	3691	6.5
Toxoplasma gondii VAND	9252	632	307	1242	800	212	2454	92	3379	6.8
Toxoplasma gondii VEG	7476	572	212	1003	740	170	1759	84	2782	7.7
Alveolata; Chromerida;										
Chromera velia CCMP2878	29025	2469	713	6915	2399	305	6644	139	9554	8.5
Vitrella brassicaformis CCMP3155	23030	2437	559	4803	2289	316	5645	96	6692	10.6
Alveolata; Ciliophora										

Continuing Table 2

Specie	Total	HLS	LS	Cyt	Plasm	Mt mem	Mt non-m	Nuc mem	Nuc non-m	Sec(%)
Ichthyophthirius multifiliis	8119	165	67	762	336	559	858	820	4771	2.0
Oxytricha trifallax	24577	770	205	1498	1631	452	1373	1742	17191	3.1
Paramecium tetraurelia	39936	1509	367	2569	1961	1304	3552	2207	27495	3.8
Stylonychia lemnae	20784	817	210	1379	1379	471	1289	1389	14206	3.9
Tetrahymena thermophila	27741	1541	454	1275	1336	1365	2481	2688	16630	5.6
Alveolata; Perkinsea										
Perkinsus marinus	23241	1406	415	6525	1823	234	4333	81	7927	6.0
Amoebozoa; Archamoebae										
Entamoeba dispar	8679	386	109	1349	431	362	657	237	5472	4.4
Entamoeba histolytica HM-1:IMSS-A	6315	281	70	845	370	244	416	165	4120	4.4
Entamoeba histolytica HM-1:IMSS-B	6292	269	67	902	363	229	415	160	4049	4.3
Entamoeba histolytica HM-3:IMSS	7331	339	90	1077	419	274	526	186	4656	4.6
Entamoeba histolytica KU27	7398	357	81	1157	424	265	560	185	4653	4.8
Entamoeba invadens IP1	9857	630	189	2323	550	217	901	352	5403	6.4
Entamoeba nuttalli	6223	283	80	891	366	240	434	149	4020	4.5
Amoebozoa; Discosea										
Acanthamoeba castellanii str. Neff	14944	952	291	3844	1160	197	2842	99	5162	6.4
Amoebozoa; Mycetozoa										
Dictyostelium discoideum	13125	719	188	1711	635	609	1246	416	7223	5.5
Dictyostelium fasciculatum	12210	729	249	1359	904	265	1163	363	7110	6.0
Dictyostelium purpureum	12353	783	227	1543	706	402	1091	501	7363	6.3
Polysphondylium pallidum	12451	812	216	1492	831	271	1193	358	7271	6.5
Apusozoa; Apusomonadidae										
Thecamonas trahens ATCC 50062	10606	500	311	3035	1497	119	2056	25	1949	4.7
Cryptophyta; Pyrenomonadales										
Guillardia theta CCMP2712	24987	2268	679	4572	2017	310	4883	183	10082	9.1
Fornicata; Diplomonadida										
Giardia intestinalis	30589	735	539	7798	1798	226	5172	221	12528	2.4
Spironucleus salmonicida	8104	62	103	1175	398	214	1084	287	4853	0.8
Haptophyceae										
Emiliania huxleyi	36436	3111	1733	8517	3003	703	12655	66	5361	8.5
Chrysochromulina sp. CCMP291	16813	936	481	4937	1605	262	4857	36	3036	5.6
Heterolobosea; Schizopyrenida										
Naegleria gruberi	15736	675	204	1909	1362	323	1310	549	9619	4.3
Kinetoplastida; Bodonidae										
Bodo saltans	18292	1187	657	3009	2306	283	4502	87	5263	6.5
Kinetoplastida; Trypanosomatidae										
Angomonas deanei	14609	325	333	3921	1136	309	3169	148	5452	2.2
Leishmania infantum	8311	207	177	1685	800	209	2621	42	2264	2.5
Leishmania major	8426	224	220	1737	838	216	2663	40	2223	2.7
Leishmania mexicana	8239	192	181	1707	789	194	2583	43	2271	2.3
Leishmania panamensis	7754	160	175	1576	680	189	2489	38	2226	2.1
Leptomonas pyrrhocoris	9298	222	225	1981	936	227	2783	54	2647	2.4
Leptomonas seymouri	8500	170	187	1680	778	205	2676	40	2546	2.0
Phytomonas sp. isolate EM1	6361	94	101	1263	488	126	1645	55	2482	1.5
Phytomonas sp. isolate Hart1	6442	112	98	1162	462	127	1636	74	2675	1.7
Strigomonas culicis	10048	295	286	2529	876	258	2814	64	2827	2.9
Trypanosoma brucei brucei	9381	726	312	1910	940	261	2567	41	2542	7.7
Trypanosoma brucei gambiense	9786	408	322	1773	1142	411	2868	76	2598	4.2

Continuing Table 2

Specie	Total	HLS	LS	Cyt	Plasm	Mt mem	Mt non-m	Nuc mem	Nuc non-m	Sec(%)
Trypanosoma congolense	12883	1082	399	2546	1023	310	3876	69	3388	8.4
Trypanosoma cruzi	54882	4122	1765	11430	5395	1152	14339	305	13257	7.5
Trypanosoma cruzi Dm28c	11346	632	396	2016	1164	275	3426	70	2795	5.6
Trypanosoma cruzi marinkellei	10187	554	361	2293	1056	191	2493	62	2443	5.4
Trypanosoma rangeli SC58	7365	284	279	1615	735	167	2308	21	1733	3.9
Trypanosoma vivax	11624	1058	350	2137	835	305	3799	58	2950	9.1
Opisthokonta										
Monosiga brevicollis	9273	980	300	1792	910	174	1813	90	2762	10.6
Salpingoeca rosetta	11703	936	305	2127	938	176	2134	109	4376	8.0
Fonticula alba	6248	254	170	1014	519	130	1717	36	1702	4.1
Capsaspora owczarzaki	9926	525	357	1841	951	185	2134	78	3210	5.3
Sphaeroforma arctica JP610	18649	862	366	4572	1017	168	3335	89	8197	4.6
Parabasalia; Trichomonadida										
Trichomonas vaginalis	50709	399	732	12171	1618	603	4039	1250	33551	0.8
Rhizaria										
Plasmodiophora brassicae	9753	776	275	2048	1241	130	2290	34	2466	8.0
Spongospora subterranea	11127	569	342	1769	861	159	2203	62	4680	5.1
Reticulomyxa filosa	39924	626	720	4680	2900	988	3983	3808	23455	1.6
Rhodophyta										
Cyanidioschyzon merolae	5053	114	115	779	442	150	1884	40	1302	2.3
Galdieria sulphuraria	7333	174	151	1034	764	154	1557	109	3307	2.4
Chondrus crispus	9741	307	248	1869	546	238	3544	31	2678	3.2
Stramenopiles; Bacillariophyta										
Thalassiosira oceanica	34571	2331	921	7179	1402	326	8637	288	13269	6.7
Thalassiosira pseudonana	11874	1335	348	2521	1063	147	1825	183	4404	11.2
Phaeodactylum tricornutum	10730	1214	417	2011	1147	177	2078	93	3305	11.3
Stramenopiles; Blastocystis										
Blastocystis hominis	5832	185	91	1304	411	101	944	81	2677	3.2
Stramenopiles; Eustigmatophyceae										
Nannochloropsis gaditana	15363	1263	619	2809	1317	333	5326	50	3305	8.2
Stramenopiles; Oomycetes										
Albugo candida	13241	608	335	1640	1040	208	2442	172	6902	4.6
Albugo laibachii Nc14	12983	395	273	1745	925	158	2719	180	6482	3.0
Aphanomyces astaci	25026	1434	579	4835	2619	321	5991	81	7843	5.7
Aphanomyces invadans	19594	869	445	4062	2254	271	4657	66	5945	4.4
Hyaloperonospora arabidopsidis	14228	837	387	2620	816	191	4206	62	5082	5.9
Phytophthora infestans	18516	1748	393	3868	1736	282	4015	92	6375	9.4
Phytophthora parasitica P1569	26477	1824	471	4751	2268	322	5802	131	10660	6.9
Phytophthora parasitica P1976	26517	1825	452	4840	2237	305	5772	126	10686	6.9
Phytophthora ramorum	15595	1423	259	3561	1646	195	3132	78	4954	9.1
Phytophthora sojae	26502	2480	526	5571	2318	319	6360	112	8611	9.4
Plasmopara halstedii	15460	811	350	2721	1036	154	3711	102	6664	5.2
Pythium ultimum DAOM BR144	15153	892	278	3146	1745	170	2995	84	5523	5.9
Saprolegnia diclina VS20	18047	1178	466	3825	2586	279	4018	49	4508	6.5
Saprolegnia parasitica	20070	1290	518	4378	2670	317	4585	60	5091	6.4
Stramenopiles; Pelagophyceae										
Aureococcus anophagefferens	11519	1094	417	4073	949	140	2537	19	1645	9.5
Stramenopiles; PX clade										
Ectocarpus siliculosus	16454	1205	454	4387	1310	259	4594	44	3737	7.3
Total	1970022	114788	45807	336720	150373	58519	370334	51070	810058	5.8

Note: Abbreviation: HLS: highly likely secreted; LS: likely secreted; Cyt: cytoplasm (or cytosol); Plasm: plasma membrane; Mt mem: mitochondrial membrane; Mt non-m: mitochondrial non-membrane; Nuc mem: nuclear membrane; Nuc non-m: nuclear non-membrane; Sec: secretome

For example, *D. discoideum* had 52 secreted proteins with DUF3430 domain (unknown function) and 44 secreted proteins with carbohydrate binding domain CBM49, while the other two species had no such protein family at all. As expected, there were a large number of secreted Elicitin, RXLR phytopathogen effector protein, necrosis inducing protein (NPP1), phytotoxin PcF protein, trypsin, etc. in *P. infestans*, which may be related to its lifestyle as a plant pathogen (Meijer et al., 2014). *T. cruzi,* not surprisingly as a human parasite pathogen, had 345 Mucin-like glycoprotein, 198 BNR repeat-like domain, and 102 Peptidase_M8 (Leishmanolysin), etc. in its secretome while the other two species did not have any for those categories. These secreted proteins may play an important role for *T. cruzi* for invading and infecting humans and causing Chagas' disease (Costa et al., 2016).

4 Discussion

We constructed the ProtSecKB to provide a resource of curated and predicted subcellular locations of protist proteins. As all the tools we selected to use were not specifically trained for protists, the prediction accuracies were lower than prediction accuracies in other eukaryotes including fungi, plants and animals (Lum and Min, 2011; Lum et al., 2014; Meiken et al., 2014; Meiken et al., 2015). However, our evaluation using curated protein subcellular locations showed that the prediction specificities for nearly all subcellular locations except nucleus were > 90%, and in particular, prediction of secreted proteins had an MCC value of 0.71 with 89.0% sensitivity and 96.2% specificity (Table 1). Thus we concluded that the prediction of secreted proteins was relatively reliable. Other tools are also available as webservers including the Cell-PLoc servers (Chou and Shen, 2008) and some others (Meinken and Min, 2012). These tools and their related publications can be found at our website (http://bioinformatics.ysu.edu/tools/subcell.html) (Meinken and Min, 2012). As standalone tools are not available for some, such as Cell-PLoc, or too slow to processing large datasets, we were not able to use them for our data processing. However, we suggest users utilize these tools to get a second prediction for proteins of interest as our experience showed that using multiple tools improves prediction specificity.

Recently the efforts had been made by our research group to improve the prediction accuracies of subcellular locations in plant proteins (Neizer-Ashun et al., 2015), fungal proteins (Munyon et al., 2015), and animal/human proteins (Khavari, 2016) using various statistics algorithms. The results were mixed for different subcellular locations using different methods with different eukaryotic proteins. However, some of the algorithms were promising in improving the prediction accuracy. When enough experimental protist protein subcellular location data are available, a specific tool will need to be implemented for protist protein subcellular location prediction.

ProtSecKB contains 101 unique protist species within some of them having multiple strains resulting in a total of 127 organisms having complete proteomes. The database allows that each subcellular proteome in each species can be searched and downloaded for detailed comparative analysis. As an example for the usage of the database, our analysis on protein families using three species having different lifestyles demonstrated that the secretome in each species may play an important role in determining their lifestyles (Table 3). We also have implemented a curation tool accessible through ProtSecKB for the community to manually curate subcellular locations of protist proteins having experimental evidence. We anticipate the database resource will facilitate the protist research community to design further experiments characterizing protist proteins and understanding protist biology, particularly of the plant, human and animal protist pathogens.

Authors' contributions

XM and CC conceived the work; BP, VA and JM implemented the database; GK curated proteins. XM, BP, FY analyzed the data. XM, BP, JM and CC prepared the manuscript. All authors read and approved the final manuscript.

Acknowledgements

BP was supported by the College of Graduate Studies, Youngstown State University (YSU), VA and JM were supported by the Center for Applied Chemical Biology, YSU. The work is also supported by a Research Professorship to XM by YSU.

Table 3 Comparison of protein families in secretomes of three protist species having different lifestyles

PFam ID	*D. discoideum*	*P. infestans*	*T. cruzi*	PFam	Description
pfam11912	52	0	0	DUF3430	Protein of unknown function (DUF3430)
pfam09478	44	0	0	CBM49	Carbohydrate binding domain CBM49
pfam00112	16	8	23	Peptidase_C1	Papain family cysteine protease
pfam03265	9	1	0	DNase_II	Deoxyribonuclease II
pfam02221	8	0	0	E1_DerP2_DerF2	ML domain
pfam07691	8	2	0	PA14	PA14 domain
pfam04562	7	0	0	Dicty_spore_N	Dictyostelium spore coat protein
pfam04916	7	0	0	Phospholip_B	Phospholipase B
pfam10500	7	0	0	SR-25	Nuclear RNA-splicing-associated protein
pfam00144	6	0	0	Beta-lactamase	Beta-lactamase
pfam00383	6	0	2	dCMP_cyt_deam_1	Cytidine and deoxycytidylate deaminase
pfam00759	6	0	0	Glyco_hydro_9	Glycosyl hydrolase family 9
pfam05577	6	1	5	Peptidase_S28	Serine carboxypeptidase S28
pfam09286	6	0	0	Pro-kuma_activ	Pro-kumamolisin
pfam00964	0	39	0	Elicitin	Elicitin
pfam16810	0	36	0	RXLR	RXLR phytopathogen effector protein
pfam05630	0	24	0	NPP1	Necrosis inducing protein (NPP1)
pfam09461	0	22	0	PcF	Phytotoxin PcF protein
pfam00089	0	17	0	Trypsin	Trypsin
pfam00188	3	14	0	CAP	Cysteine-rich secretory protein family
pfam00295	0	14	0	Glyco_hydro_28	Glycosyl hydrolases family 28
pfam03211	0	14	0	Pectate_lyase	Pectate lyase
pfam05642	0	12	1	Sporozoite_P67	Sporozoite P67 surface antigen
pfam13091	1	10	0	PLDc_2	PLD-like domain
pfam00588	0	8	1	SpoU_methylase	SpoU rRNA Methylase family
pfam16656	3	7	0	Pur_ac_phosph_N	Purple acid Phosphatase
pfam00050	0	7	0	Kazal_1	Kazal-type serine protease inhibitor domain
pfam00264	0	7	0	Tyrosinase	Common central domain of tyrosinase
pfam02055	0	7	0	Glyco_hydro_30	O-Glycosyl hydrolase family 30
pfam03639	0	7	0	Glyco_hydro_81	Glycosyl hydrolase family 81
pfam01565	3	6	0	FAD_binding_4	FAD binding domain
pfam04147	1	6	45	Nop14	Nop14-like family
pfam00194	0	6	0	Carb_anhydrase	Eukaryotic-type carbonic anhydrase
pfam01083	0	6	0	Cutinase	Cutinase
pfam01095	0	6	0	Pectinesterase	Pectinesterase
pfam01670	0	6	0	Glyco_hydro_12	Glycosyl hydrolase family 12
pfam01456	0	0	345	Mucin	Mucin-like glycoprotein
pfam13859	0	0	198	BNR_3	BNR repeat-like domain
pfam01457	1	0	102	Peptidase_M8	Leishmanolysin
pfam12517	0	0	90	DUF3720	Protein of unknown function (DUF3720)
pfam05086	0	0	57	Dicty_REP	Dictyostelium (Slime Mold) REP protein
pfam12446	0	0	50	DUF3682	Protein of unknown function (DUF3682)
pfam01762	0	1	41	Galactosyl_T	Galactosyltransferase
pfam01299	4	0	21	Lamp	Lysosome-associated membrane glycoprotein
pfam00183	1	0	18	HSP90	Hsp90 protein

Continuing Table 3

PFam ID	*D. discoideum*	*P. infestans*	*T. cruzi*	PFam	Description
pfam00069	0	1	15	Pkinase	Protein kinase domain
pfam00085	1	2	14	Thioredoxin	Thioredoxin
pfam00012	0	1	14	HSP70	Hsp70 protein
pfam01764	5	0	14	Lipase_3	Lipase (class 3)
pfam08553	0	0	13	VID27	VID27 cytoplasmic protein
pfam01532	2	1	12	Glyco_hydro_47	Glycosyl hydrolase family 47
pfam03388	0	0	11	Lectin_leg-like	Legume-like lectin family
pfam10479	0	0	11	FSA_C	Fragile site-associated protein C-terminus
pfam00175	0	0	10	NAD_binding_1	Oxidoreductase NAD-binding domain
pfam00106	1	0	9	adh_short	short chain dehydrogenase
pfam13458	0	0	9	Peripla_BP_6	Periplasmic binding protein
pfam00450	2	2	8	Peptidase_S10	Serine carboxypeptidase
pfam10446	0	1	8	DUF2457	Protein of unknown function (DUF2457)
pfam11052	0	0	8	Tr-sialidase_C	Trans-sialidase of Trypanosoma
pfam00226	1	4	7	DnaJ	DnaJ domain
pfam02777	0	0	7	Sod_Fe_C	Iron/manganese superoxide dismutases
pfam00160	1	1	6	Pro_isomerase	Cyclophilin type peptidyl-prolyl
pfam00004	0	1	6	AAA	ATPase family associated with various cellular
pfam03985	0	0	6	Paf1	Paf1
pfam07999	0	0	6	RHSP	Retrotransposon hot spot protein
pfam13868	0	0	6	TPH	Trichohyalin-plectin-homology domain

Note: The table only contains protein families having 6 or more members in a species. A complete list can be found as supplementary Table3

References

Bakthavatsalam D., and Gomer R.H., 2012, The secreted proteome profile of developing *Dictyostelium discoideum* cells, Proteomics, 10: 2556-2559
https://doi.org/10.1002/pmic.200900516

Bendtsen J.D., Nielsen H., von Heijne G., and Brunak S., 2004, Improved prediction of signal peptides: SignalP 3.0, Journal of Molecular Biology, 340(4): 783-795
https://doi.org/10.1016/j.jmb.2004.05.028

Bern C., Kjos S., Yabsley M.J., and Montgomery S.P., 2011, Trypanosoma cruzi and Chagas' disease in the United States, Clinical Microbiology Reviews, 24(4): 655-681
https://doi.org/10.1128/CMR.00005-11

Bhatt T.K., 2012, Malaria Parasite 'Secretome': A Potential Drug Target, Research and Reviews: Journal of Computational Biology, 1(2): 1-5

Watanabe C.R., Da S.J., and Bahia D., 2016, Interactions between trypanosoma cruzi secreted proteins and host cell signaling pathways, Frontiers in Microbiology, 7(e102)
https://doi.org/10.3389/fmicb.2016.00388

D'Acremont V., Lengeler C., and Genton B., 2010, Reduction in the proportion of fevers associated with plasmodium falciparum parasitaemia in Africa: a systematic review, Malaria Journal, 9(1): 1
https://doi.org/10.1186/1475-2875-9-240

Emanuelsson O., Brunak S., von Heijne G., and Nielsen H., 2007, Locating proteins in the cell using TargetP, SignalP and related tools, Nature Protocols, 2(4): 953-971
https://doi.org/10.1038/nprot.2007.131

Foissner W., 1999, Protist diversity: estimates of the near-imponderable, Protist, 150: 363-368
https://doi.org/10.1016/S1434-4610(99)70037-4

Foissner W., 2006, Biogeography and dispersal of micro-organisms: a review emphasizing protists, Acta Protozoologica, 45: 111-136

Hiller N.L., Bhattacharjee S., van Ooij C., Liolios K., Harrison T., Lopez-Estrano C., and Haldar K., 2004, A host-targeting signal in virulence proteins reveals a secretome in malarial infection, Science, 306(5703): 1934-1937
https://doi.org/10.1126/science.1102737

Horton P., Park K.J., Obayashi T., Fujita N., Harada H., Adams-Collier C.J., and Nakai K., 2007, WoLF PSORT: protein localization predictor, Nucleic Acids Research, 35(suppl 2): W585-W587
https://doi.org/10.1093/nar/gkm259

Käll L., Krogh A., and Sonnhammer E.L., 2007, Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server,

Nucleic Acids Research, 35(suppl 2): W429-W432
https://doi.org/10.1093/nar/gkm256

Khavari S., 2016, Predicting human and animal protein subcellular location, Thesis for Master of Science in Mathematics, Advisor: Chang G-H, Youngstown State University, pp 1-68

Krogh A., Larsson B., Von Heijne G., and Sonnhammer E.L., 2001, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, Journal of Molecular Biology, 305(3): 567-580
https://doi.org/10.1006/jmbi.2000.4315

Lum G., Meinken J., Orr J., Frazier S., and Min X.J., 2014, PlantSecKB: the plant secretome and subcellular proteome knowledgebase, Computational Molecular Biology, 4(4)

Lum G., and Min X.J., 2011, FunSecKB: the fungal secretome knowledgebase, Database, bar001
https://doi.org/10.1093/database/bar001

Lum G., and Min X.J., 2013, Bioinformatic protocols and the knowledge-base for secretomes in fungi, In Laboratory Protocols in Fungal Biology (pp. 545-557), Springer New York
https://doi.org/10.1007/978-1-4614-2356-0_54

Matthews B.W., 1975, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochimica et Biophysica Acta (BBA)-Protein Structure, 405(2): 442-451
https://doi.org/10.1016/0005-2795(75)90109-9

Meijer H.J., Mancuso F.M., Espadas G., Seidl M.F., Chiva C., Govers F., and Sabidó E., 2014, Profiling the secretome and extracellular proteome of the potato late blight pathogen Phytophthora infestans, Molecular and Cellular Proteomics, 13(8): 2101-2113
https://doi.org/10.1074/mcp.M113.035873

Meinken J., Asch D.K., Neizer-Ashun K.A., Chang G.H., Cooper J.R., C.R., and Min X.J., 2014, FunSecKB2: a fungal protein subcellular location knowledgebase, Computational Molecular Biology, 4(4)

Meinken J., Walker G., Cooper C.R., and Min X.J., 2015, MetazSecKB: the human and animal secretome and subcellular proteome knowledgebase, Database, bav077
https://doi.org/10.1093/database/bav077

Meinken J., and Min J., 2012, Computational prediction of protein subcellular locations in eukaryotes: an experience report, Computational Molecular Biology, 2(1): 1-7
https://doi.org/10.5376/cmb.2012.02.0001

Min X.J., 2012, Evaluation of computational methods for secreted protein prediction in different eukaryotes, Journal of Proteomics and Bioinformatics, 2010

Munyon J.D., Min X., Khavari S., and Chang G.H., 2015, Prediction of subcellular locations for fungal proteins, Proceeding of the Joint Statistics Meeting 2015 (JSM2015), pp. 2497-2508

Neizer-Ashun K., Yu F., Meinken J., Min X., and Chang G.H., 2015, Prediction of plant protein subcellular locations, 7th International Conference on Bioinformatics and Computational Biology (BICoB 2015), Honolulu, Hawaii, USA, pp. 91-96

Nowicki M., Foolad M.R., Nowakowska M., and Kozik E.U., 2012, Potato and tomato late blight caused by Phytophthora infestans: an overview of pathology and resistance breeding, Plant Disease, 96(1): 4-17
https://doi.org/10.1094/PDIS-05-11-0458

Petersen T.N., Brunak S., von Heijne G., and Nielsen H., 2011, SignalP 4.0: discriminating signal peptides from transmembrane regions, Nature methods, 8(10): 785-786
https://doi.org/10.1038/nmeth.1701

Przyborski J., and Lanzer M., 2004, The malarial secretome, Science, 306: 1897-1898
https://doi.org/10.1126/science.1107072

Sigrist C.J.A., Cerutti L., Castro E.D., Langendijk-Genevaux P.S., Bulliard V., and Bairoch A., and Hulo N., 2010, Prosite, a protein domain database for functional characterization and annotation, Nucleic Acids Research, 38(suppl_1): 161-166

Singh M., Mukherjee P., Narayanasamy K., Arora R., Sen S.D., Gupta, S., Natarajan K., and Malhotra P., 2009, Proteome analysis of plasmodium falciparum extracellular secretory antigens at asexual blood stages reveals a cohort of proteins with possible roles in immune modulation and signaling, Molecular and Cellular Proteomics, 8(8): 2102-2118
https://doi.org/10.1074/mcp.M900029-MCP200

Slapeta J., Moreira D., and López-García P., 2005, The extent of protist diversity: insights from molecular ecology of freshwater eukaryotes, Proceedings of the Royal Society B Biological Sciences, 272(1576): 2073-2081
https://doi.org/10.1098/rspb.2005.3195

Soni R., Sharma D., and Bhatt T.K., 2016, Plasmodium falciparum secretome in erythrocyte and beyond, Frontiers in Microbiology, 7(6049)
https://doi.org/10.3389/fmicb.2016.00194

Van O.C., Tamez P., Bhattacharjee S., Hiller N.L., Harrison T., Liolios K., Kooij T., Ramesar J., Balu B., Adams J., Waters A.P., Janse C.J., and Haldar K., 2008, The malaria secretome: from algorithms to essential function in blood stage infection, Plos Pathogens, 4(6): e1000084-e1000084
https://doi.org/10.1371/journal.ppat.1000084

Heijne G.V., 1990, The signal peptide, The Journal of Membrane Biology, 115(3): 195-201
https://doi.org/10.1007/BF01868635