Bioinformatic Protocols and the Knowledge-Base for Secretomes in Fungi

54

Gengkon Lum and Xiang Jia Min

Abstract

Fungal secreted proteins play important roles in cell signaling, metabolism, and regulation of fungal growth and development. The secretome refers to all secreted proteins in a proteome that are identified from completely sequenced genomes. The majority of secreted proteins are classical, signal peptide-dependent proteins that can be predicted using bioinformatics tools. In this chapter, we describe some commonly used tools for secreted protein prediction in fungi and propose a relatively accurate bioinformatic protocol for fungal secretome identification. The protocol combines multiple signal peptide or subcellular location predictors, including SignalP, WoLF PSORT, and Phobius, with TMHMM for removing transmembrane proteins and PROSITE PS-Scan for removing endoplasmic reticulum (ER) proteins. Applying this protocol, we have built the fungal secretome knowledge-base (FunSecKB). The utility of FunSecKB is described in detail. FunSecKB serves the community as a central portal for search and deposition of fungal secretome information.

Keywords

Secreted proteins • Secretome • Signal peptide • Fungi • Prediction • Knowledge-base • Database

Introduction

Secreted proteins are proteins which are synthesized within cells and then secreted to extracellular space and matrix to play their roles. Secreted proteins play important roles in cell signaling, metabolism, and regulation in growth and development of all organisms. As the genomes have been completely sequenced in many organisms, the proteomes could be predicted using the information in

G. Lum
Department of Computer Science and Information Systems, Youngstown State University, Youngstown, OH 44555, USA

X.J. Min (✉)
Department of Biological Sciences, Center for Applied Chemical Biology, Youngstown State University, One University Plaza, Youngstown, OH 44555, USA
e-mail: xmin@ysu.edu

V.K. Gupta et al. (eds.), *Laboratory Protocols in Fungal Biology: Current Methods in Fungal Biology*, Fungal Biology, DOI 10.1007/978-1-4614-2356-0_54,

genomes. The term "secretome" was first used to include all proteins secreted to extracellular space and matrix and proteins involved in the secretion pathway including endoplasmic reticulum, Golgi apparatus, and transportation vesicles [1–3]; however, more recently, the term was used to include secreted proteins only [4,5]. In this work, the secretome only refers the complete set of secreted proteins in an organism.

Secretomes are an important part of the fungal proteome. These secreted proteins include enzymes, growth factors, cell wall proteins, and other bioactive molecules which play important roles in host–pathogen interactions. Fungal secreted enzymes are used to break down potential food sources for transport into the cells. As there are many types of fungi producing a great variety of enzymes that are able to break down lignocelluloses and other biopolymers, fungi have an important function in the biosphere as decomposers. Since secreted proteins are useful in their ability to break down biopolymers, they have found a role in many applications including pharmaceutical and industrial [6]. Therefore, the ability to analyze a protein to determine if it is secreted and what functions it may have is useful as a tool in research to more easily focus on or to eliminate potential targets. Increased understanding of the secretome biology of fungi will further promote exploration of the potential applications of fungal secreted proteins in environmental remediation and industrial processing including bio-fuel production.

Most of secreted proteins in fungi are classical secreted proteins, which have a signal peptide on the N-terminus of protein sequences. A signal peptide is typically 15–30 amino acids long, located at the N-terminus of the protein and is cleaved off during translocation across the membrane. The presence of a signal peptide directs the protein to the rough endoplasmic reticulum (ER) and the Golgi complex in preparation for transport through the secretory vesicles. This is referred to as the classical secretory pathway. Although not all proteins excreted extracellularly contain a signal peptide, it is believed that that the majority of fungal proteins are secreted in this manner [6]. The presence of a transmembrane domain in the protein sequence, however, indicates that although the protein passes through the classical secretory pathway, it is not secreted extracellularly but instead becomes part of the cell membrane. By combining the results of one or more predictions for the presence of a signal peptide along with the absence of a transmembrane domain, the likelihood of the protein being secreted is very high. Our recent evaluation reveals that combining the results of multiple programs increases the accuracy by reducing the number of false positives and negatives [7].

Two fungal-specific secretome databases are currently available. The Fungal Secretome Database[1] developed by Choi et al. used nine bioinformatics tools and protein sequences from completely sequenced genomes including some work in progress draft genomes [8]. The Fungal Secretome Knowledge-Base (FunSecKB)[2] developed by us used all fungal protein sequences available in the NCBI RefSeq database and being linked and supplemented with protein sequences in the UniProt database. The detailed comparison of the two databases was described by Lum and Min [5]. In this work, we focus on how to utilize FunSecKB.

Materials (Data)

For individual secreted protein identification, the input is a fungal protein sequence in FASTA format. For a species-specific secretome prediction from a whole proteome, which often is obtained from a completely sequenced genome, a set of proteins in multiple FASTA format are used as input.[3] We will use a glucoamylase enzyme from *Aspergillus niger* (gi 145235763) and a *Schizosaccharomyces pombe* protein (gi 19115161) as examples to explain the input and output of the tools mentioned in Sect. "Methods."

[1] http://fsd.snu.ac.kr/.

[2] http://proteomics.ysu.edu/secretomes/fungi.php.

[3] A description of the FASTA format may be found at http://www.ncbi.nlm.nih.gov/BLAST/fasta.shtml.

Methods

The programs used for fungal secretome prediction include SignalP 3.0 [9], TargetP 1.1 [10], TMHMM 2.0 [11], Phobius [12], WoLF PSORT [13], PS-Scan for PROSITE [14], and FragAnchor [15]. SignalP and TargetP predict the presence and location of an N signal peptide and a potential cleavage site. TMHMM predicts the presence of a transmembrane domain. Phobius is a combined signal peptide and transmembrane topology predictor. WoLF PSORT (WolfPsort) predicts the subcellular location(s) of a protein. PS-Scan is a PROSITE scanning tool which predicts whether or not a protein contains an endoplasmic reticulum (ER) targeting sequence (Prosite: PS00014). FragAnchor is used to predict if there is a glycosylphosphatidylinositol (GPI) anchor in the protein, which may indicate if the secreted protein is a cell wall protein or attaches to the outside of the plasma membrane.

There are two methods of using these tools. The first one, most often used by a biologist to process an individual protein, uses the online Webserver tool. The second one, often used by bioinformaticians to process proteome-wide secretome identification, use a standalone package which may be downloaded and run on a UNIX (Linux) platform.

SignalP 3.0

SignalP 3.0 uses both neural network (NN) and hidden Markov model (HMM) algorithms in two different predictors to predict whether a protein has a signal peptide and where the most likely cleavage site would be if one is detected [9].[4] For each protein processed by SignalP 3.0, scores are calculated and returned in two sections: SignalP-NN result (Fig. 54.1a) and SignalP-HMM result (Fig. 54.1b).

In the Signal-NN results, two different neural networks are used for each prediction, one for predicting the presence of the signal peptide, the other for predicting the position of the cleavage site. For each position in the protein, a C, S, and Y score is calculated. The C score is the cleavage site score with values being high at potential cleavage sites. The S score is reported for every position submitted with high scores for amino acids which are part of the signal peptide and low score for those which are part of a mature protein. The Y score is a derivative of C and S with a likely cleavage point being when the slope of S is steep and there is a high C score resulting in a high Y score. The mean S score is the average of the S scores from the N-terminus to the highest Y score. The D score is average of the Y score and the mean S score. The D score is used to determine whether or not a protein is predicted to be secreted.

In the Signal-HMM results, the positions are evaluated to determine the likelihood of being a part of the n-region, h-region, or c-region. Signal peptides commonly have a hydrophobic central core (h-region) surrounded by the N- and C-terminal hydrophilic regions. The HMM makes a prediction of a signal peptide, a nonsecretory protein or a signal anchor. A protein with a signal anchor passes through the membrane but the uncleaved signal peptide remains anchored to the membrane resulting in a type II membrane protein [9]. The results also include a probability for both a signal peptide and a signal anchor.

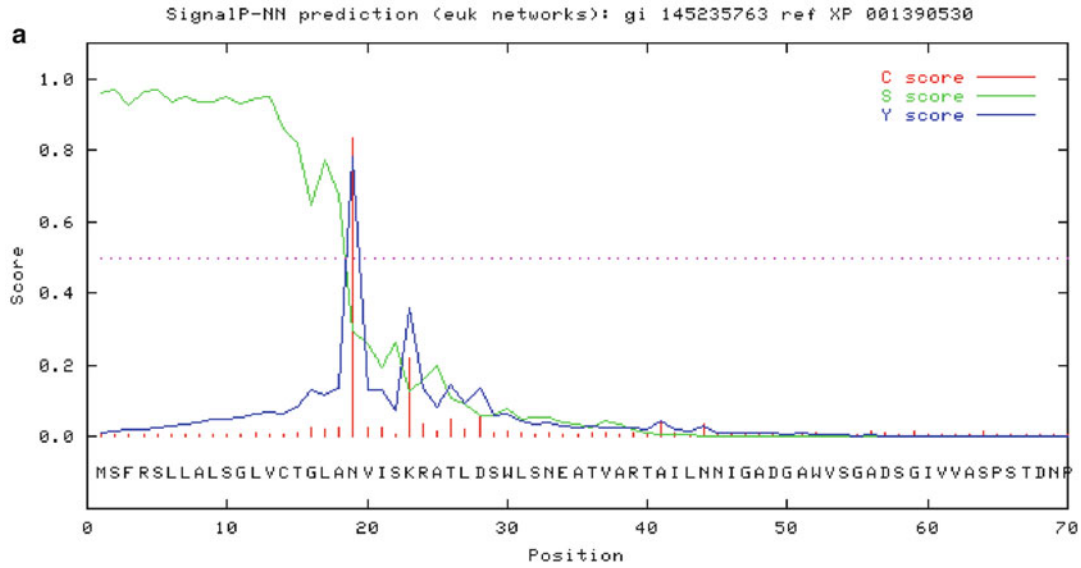
Phobius

Phobius is a combined signal peptide and a transmembrane topology predictor [5] [12]. A known problem with signal peptide and transmembrane topology predictors is the high similarity of the hydrophobic regions of both the signal peptide h-region and transmembrane helix. Due to this similarity, pure signal peptide predictors and transmembrane topology predictors sometimes

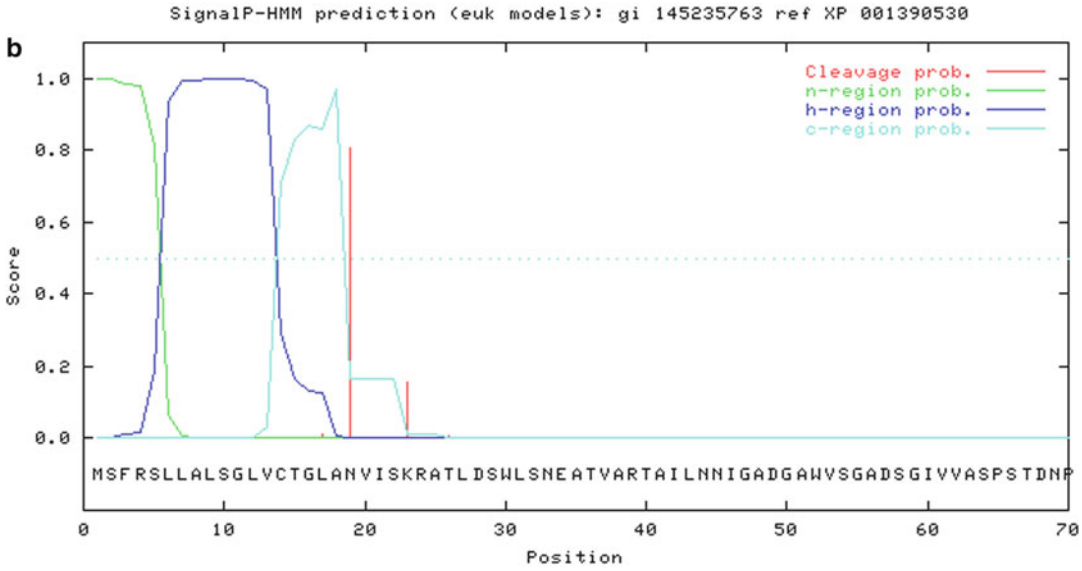
[4] http://www.cbs.dtu.dk/services/SignalP/.

[5] http://phobius.sbc.su.se/index.html.

```
>gi|145235763| glucan 1,4-alpha-glucosidase glaA-Aspergillus niger
SignalP-NN result:
```



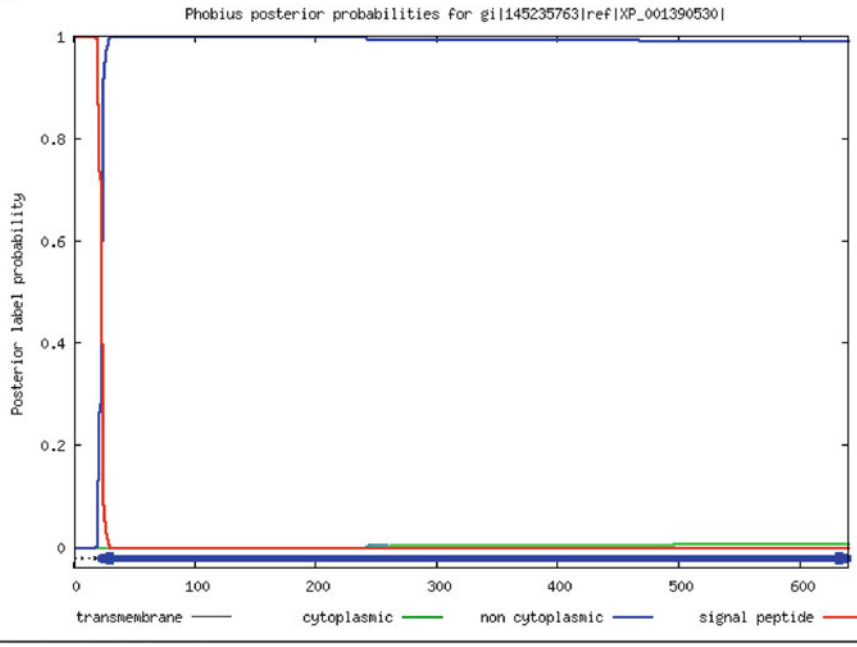

```
# Measure  Position  Value  Cutoff  signal peptide?
  max. C     19       0.833   0.32   YES
  max. Y     19       0.782   0.33   YES
  max. S      2       0.970   0.87   YES
  mean S     1-18     0.894   0.48   YES
       D     1-18     0.838   0.43   YES
# Most likely cleavage site between pos. 18 and 19: GLA-NV
SignalP-HMM result:
```




```
Prediction: Signal peptide
Signal peptide probability: 0.998
Signal anchor probability: 0.001
Max cleavage site probability: 0.806 between pos. 18 and 19
```

Fig. 54.1 (**a**) Neural network results output for SignalP 3.0 Server of *Aspergillus niger* glucoamylase protein. (**b**) Hidden Markov model results output for SignalP 3.0 Server of *Aspergillus niger* glucoamylase protein

```
Long with Graphics output format:
Prediction of gi|145235763|ref|XP_001390530|
ID   gi|145235763|ref|XP_001390530|
FT   SIGNAL        1     22
FT   REGION        1      5       N-REGION.
FT   REGION        6     17       H-REGION.
FT   REGION       18     22       C-REGION.
FT   TOPO_DOM     23    640       NON CYTOPLASMIC.
//
```




```
Short output format:
SEQENCE ID                      TM SP PREDICTION
gi|145235763|ref|XP_001390530|  0  Y n6-17c22/23
```

Fig. 54.2 Results output for Phobius of *Aspergillus niger* glucoamylase protein

results in false classifications. To this end, Phobius was designed to do both signal peptide and transmembrane topology prediction and to distinguish between the two regions.

The output formats available are long with graphics, long without graphics, or short format (Fig. 54.2). The default output (long with graphics) shows the prediction of probable locations for sections of the protein. Some possible predictions are: SIGNAL for signal peptide, REGION for N-, H-, and C-regions, TOPO_DOM for topology (cytoplasmic or non-cytoplasmic) and TRANSMEM for positions predicted to be within the membrane. The range of positions is given for each predicted segment. If the entire sequence is labeled cytoplasmic or non-cytoplasmic though, the prediction is that there are no membrane helices and is not an actual prediction of location, but the most probable location.

The short output format gives **TM** as the number of predicted transmembrane segments, **SP** as the prediction of whether or not there is a signal peptide, and **PREDICTION** as the predicted topology. The format of the predicted topology is given as a series of numbers and letters. If a signal peptide is detected, it is given in the format: n#-#c#/# where # represents a position in the sequence. The numbers between n and c is the range of the hydrophobic h-region and the #/# is the cleavage site. Following this is either an "i" if the loop is cytoplasmic or an "o" if it is on the non-cytoplasmic side and then numbers in the

```
k used for kNN is: 27
gi|145235763|ref|XP_001390530| details extr: 26.0
```

Fig. 54.3 Results output for WoLF PSORT of *Aspergillus niger* glucoamylase protein

```
### targetp v1.1 prediction results ###################################
Number of query sequences:  1
Cleavage site predictions included.
Using NON-PLANT networks.
Name                    Len             mTP     SP   other   Loc   RC   TPlen
--------------------------------------------------------------------------------
gi_145235763_ref_XP_   640           0.183  0.801  0.050    S     2     18
--------------------------------------------------------------------------------
cutoff                               0.000  0.000  0.000
```

Fig. 54.4 Results output for TargetP 1.1 Server of *Aspergillus niger* glucoamylase protein

format #-# indicating the range of the transmembrane helix. This format is repeated until the end of the sequence.

WoLF PSORT

WoLF PSORT[6] is a program for predicting the subcellular location of proteins [13]. It takes the amino acid sequences and converts them into numerical vectors which are then classified using a weighted *k*-nearest neighbor classifier. The predictions are based on known sorting signal motifs and the content of the amino acids. It requires the selection of the organism type: animal, plant, or fungi. For our example protein (Fig. 54.3), the result was based on using $k=27$ nearest neighbors. Of these 27 closest, 26 were extracellular and the result is displayed as extr: 26.0. A list of the localization site definitions is available on the Website and include locations such as golg for the Golgi apparatus, mito for mitochondria, and nucl for nuclear.

TargetP 1.1

TargetP is designed to predict the subcellular locations of eukaryotic proteins[7] [10]. TargetP predicts in the N-terminus the presence of any of the N-terminal presequences such as signal peptide (SP), chloroplast transit peptide (cTP), or mitochondrial targeting peptide (mTP). The output is given in Fig. 54.4. **Name** is the sequence name truncated to 20 characters. **Len** is the length of the sequence. **cTP**, **mTP**, **SP**, **other** are the final neural network (NN) scores. **cTP** is only used if the organism group on the submission page is set to Plant since it is used to detect a cTP.

TMHMM 2.0

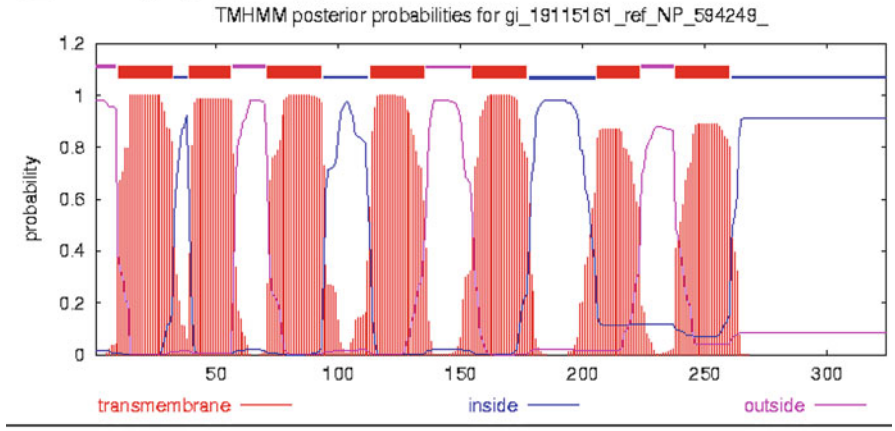
TMHMM 2.0 uses a HMM to predict the presence and topology of transmembrane helices and their orientation to the membrane (in/out)[8][11]. The output shows the results of the prediction (Fig. 54.5).

[6] http://wolfpsort.org/

[7] http://www.cbs.dtu.dk/services/TargetP/.

[8] http://www.cbs.dtu.dk/services/TMHMM/.

```
Long format (default):
# gi_19115161_ref_NP_594249_ Length: 324
# gi_19115161_ref_NP_594249_ Number of predicted TMHs:  7
# gi_19115161_ref_NP_594249_ Exp number of AAs in TMHs: 147.53314
# gi_19115161_ref_NP_594249_ Exp number, first 60 AAs:  40.80425
# gi_19115161_ref_NP_594249_ Total prob of N-in:        0.01841
# gi_19115161_ref_NP_594249_ POSSIBLE N-term signal sequence
gi_19115161_ref_NP_594249_   TMHMM2.0  outside      1     9
gi_19115161_ref_NP_594249_   TMHMM2.0  TMhelix     10    32
gi_19115161_ref_NP_594249_   TMHMM2.0  inside      33    38
gi_19115161_ref_NP_594249_   TMHMM2.0  TMhelix     39    56
gi_19115161_ref_NP_594249_   TMHMM2.0  outside     57    70
gi_19115161_ref_NP_594249_   TMHMM2.0  TMhelix     71    93
gi_19115161_ref_NP_594249_   TMHMM2.0  inside      94   112
gi_19115161_ref_NP_594249_   TMHMM2.0  TMhelix    113   135
gi_19115161_ref_NP_594249_   TMHMM2.0  outside    136   154
gi_19115161_ref_NP_594249_   TMHMM2.0  TMhelix    155   177
gi_19115161_ref_NP_594249_   TMHMM2.0  inside     178   205
gi_19115161_ref_NP_594249_   TMHMM2.0  TMhelix    206   223
gi_19115161_ref_NP_594249_   TMHMM2.0  outside    224   237
gi_19115161_ref_NP_594249_   TMHMM2.0  TMhelix    238   260
gi_19115161_ref_NP_594249_   TMHMM2.0  inside     261   324
```




```
Short format:
gi_19115161  len=324  ExpAA=147.53  First60=40.80  PredHel=7  Topology=o10-
32i39-56o71-93i113-135o155-177i206-223o238-260i
```

Fig. 54.5 Results output for TMHMM Server 2.0 of a *Schizosaccharomyces pombe* protein

Using the default long format: **Length** is the length of the sequence submitted. **Number of predicted TMHs** is the number of predicted transmembrane helices. **Exp number of AAs in TMHs** is the expected number of amino acids in transmembrane helices. **Exp number, first 60 AAs** is the expected number of amino acids in transmembrane helices within the first 60 positions. **Total prob of N-in** is the total probability that the N-terminus is on the cytoplasmic side of the membrane. Following this section is the prediction of where specific parts of the protein are likely to be: inside, outside, or TM helix (part of the transmembrane helix). The structure is the identifier used, followed by the program name (TMHMM2.0), the predicted location, then the starting and ending position of the segment. In this example, the prediction is that the N-terminus is outside the membrane and the protein crosses the membrane seven times and the C-terminus ends on the inside of the cell. Using the short format: **len**, length of sequence, **ExpAA**, expected number of amino acids in transmembrane helices, **First60**, expected number of amino acids in transmembrane helices within the first 60 positions, **PredHel**, number of predicted transmembrane helices by N-best, and **Topology**, the topology predicted by N-best with "o" indicating sections outside and "i" indicating sections inside the cell.

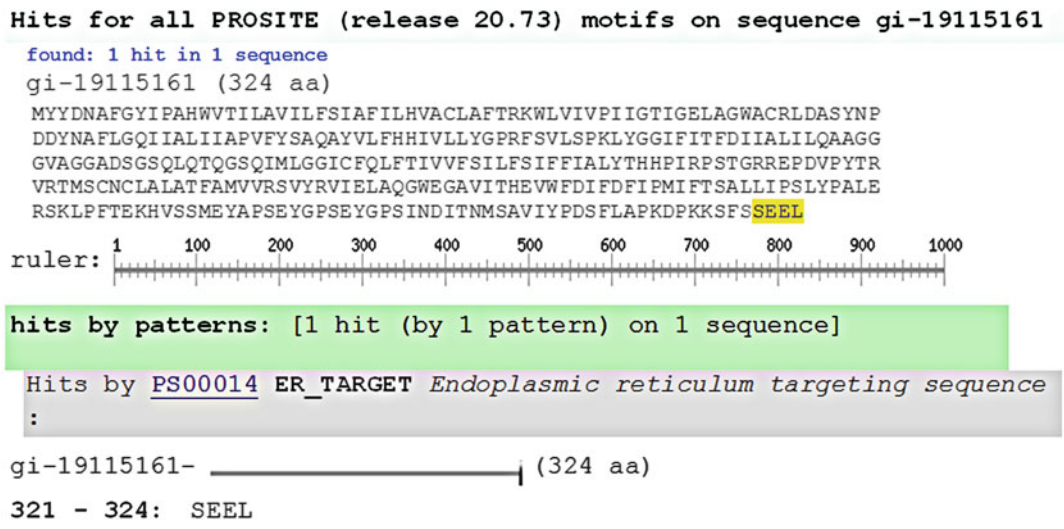


Fig. 54.6 Results output for ScanProsite of a *Schizosaccharomyces pombe* protein

PS-Scan for PROSITE

PROSITE is a database containing protein families, domains, and functional sites. The ScanProsite Website[9] scans the PROSITE database for motifs matching the input sequence [14]. The output from the Website lists any hits found in their database matching sections within our sequence. In our FunSecKB database, we used the standalone program PS-Scan to determine if there was an ER retention signal (Prosite: PS00014), which if found could rule out the possibility that the particular protein would be secreted. The output of a *Schizosaccharomyces pombe* protein (gi 19115161) shows an ER targeting sequence detected (Fig. 54.6) at positions 321–324.

FragAnchor

FragAnchor[10] is a tool to detect the presence of a glycosylphosphatidylinositol (GPI) anchor [15]. It uses a combination of a neural network to select potential GPI-anchored sequences and a HMM to classify those sequences into categories of likelihood. The four categories are highly probable, probable, weakly probable, and potential false positive. Our example did not contain a potential GPI-anchored sequence and thus was rejected with the HMM classification never being run, thus the output is not shown here. However, the detailed information of the GPI-anchored secreted proteins and the correlations with proteome size and genome size can be found in Lum and Min [5]. This Webserver support a batch of sequences, but no standalone tool is available. There are some other tools available for GPI anchor prediction, including Big-PI predictor [11] and PredGPI.[12]

SecretomeP

SecretomeP[13] is a program that uses a sequence-based method for prediction of secreted proteins based on nonclassical secretory pathways. The original program was trained on bacteria and support for mammalian proteins was added

[9] http://expasy.org/tools/scanprosite/.

[10] http://navet.ics.hawaii.edu/~fraganchor/NNHMM/NNHMM.html.

[11] http://mendel.imp.ac.at/sat/gpi/gpi_server.html.

[12] http://gpcr.biocomp.unibo.it/predgpi/.

[13] http://www.cbs.dtu.dk/services/SecretomeP/.

Table 54.1 Linux commandline summary for standalone packages[a]

Tools	Commands
SignalP	signalp –t euk –f summary input_file>output_file
Phobius	phobius input_file –short>output_file
WoLFPsort	runWolfPsortSummaryOnly fungi<input_file>output_file
TargetP	targetp –c –N input_file>output_file
TMHMM	tmhmm input_file –short –noplot>output_file
PS-Scan	ps_scan.pl input_file –p PS00014 –o scan –d prosite.dat>output_file

[a]Input_file is the protein sequences in FASTA format. output_file is the file to save the results of the program

afterward. The Webserver currently has support for gram-negative and gram-positive bacteria along with mammalian proteins but its use in prediction of secreted fungal proteins by nonclassical pathways has not been tested. Choi et al. used this tool to predict nonclassical, signal peptide independent secreted proteins in constructing the Fungal Secretome Database[14] [8]. However, as the accuracy of the tool in fungal secretome prediction was not reported, we did not use this tool in FunSecKB (see discussion in Sect. "TMHMM 2.0").

Commands of Standalone Tools

We described the online Webservers above. The online Webservers normally have a limit for the maximum number of sequences allowed to be submitted at once; therefore, to process a large number (i.e., a proteome of a whole species) the standalone tools are needed. For the standalone tools that need to be installed on a Linux system, the commands of how to run them are summarized in Table 54.1. Detailed explanations of how to run each tool often can be found in the "readme" page in each downloaded package.

Protocol Evaluation

The accuracy of a prediction tool can only be evaluated using a set of sequence data. Min reported the accuracy of some of the tools mentioned above in prediction of fungal secretomes [7]. The tools were evaluated individually and in combination with others. The dataset contained 241 secreted proteins and 5,992 nonsecreted proteins and the results were measured using sensitivity (Sn) (Equation 54.1), specificity (Sp) (Equation 54.2), and Mathews' Correlation Coefficient (MCC) (Equation 54.3) [16–18],

$$\mathrm{Sn}(\%) = \mathrm{TP} / (\mathrm{TP} + \mathrm{FN}) \times 100 \quad (54.1)$$

$$\mathrm{Sp}(\%) = \mathrm{TN} / (\mathrm{TN} + \mathrm{FP}) \times 100 \quad (54.2)$$

$$\mathrm{MCC}(\%) = (\mathrm{TP} \times \mathrm{TN} - -\mathrm{FP} \times \mathrm{FN}) \times 100 / \left((\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})\right)^{1/2} \quad (54.3)$$

where TP represents the number of true positives, FN is the number of false negatives, TN is the number of true negatives, and FP is the number of false positives. When tools were combined, a true positive was counted only when all the tools used predicted the protein as positive. The results were provided in Table 54.2, which was adopted from Ref. [7]. Based on the results, we used the combination of SignalP, Phobius, WoLF PSORT, TMHMM, and PS-Scan, which gave the highest MCC (83.4 %) result, as the prediction protocol for fungal secretome prediction in FunSecKB development [5]. The TargetP 1.1 can be used for

[14]http://fsd.snu.ac.kr/.

Table 54.2 Prediction accuracies of secreted proteins in fungi[a]

Methods	TP	FP	TN	FN	Sn (%)	Sp (%)	MCC (%)
SignalP	232	329	5663	9	96.3	94.5	61.2
Phobius	226	203	5789	15	93.8	96.6	68.8
TargetP	228	583	5409	13	94.6	90.3	48.6
WolfPsort	230	167	5825	11	95.4	97.2	73.1
SignalP/TMHMM	228	168	5824	13	94.6	97.2	72.6
Phobius/TMHMM	224	200	5792	17	92.9	96.7	68.6
TargetP/TMHMM	224	265	5727	17	92.9	95.6	63.5
WolfPsort/TMHMM	227	135	5857	14	94.2	97.7	75.8
SignalP/TMHMM/WolfPsort	226	86	5906	15	93.8	98.6	81.6
SignalP/TMHMM//WolfPsort/Phobius	222	69	5923	19	92.1	98.8	83.1
SignalP/TMHMM/WolfPsort/Phobius/PS-Scan	**222**	**67**	**5925**	**19**	**92.1**	**98.9**	**83.4**
SignalP/TMHMM/WolfPsort/Phobius/TargetP/PS-Scan	218	66	5926	23	90.5	98.9	82.6

TP true positives; *FP*, false positives; TN, true negatives; FN, false negatives; Sn, sensitivity; Sp, specificity; MCC, Mathews' correlation coefficient.

[a]The table is reproduced with permission from Min [7].

individual secreted protein prediction, however, adding it to the pipeline for secretome prediction slightly reduced the accuracy (see Table 54.2).

The Fungal Secretome Knowledge-Base

The Fungal Secretome Knowledge-Base (FunSecKB) is a database of fungal proteins collected from NCBI and UniProt on which we have performed various analyses for prediction of possible extracellular secretion [5].[15] From this site (Fig. 54.7), you can look up specific proteins using either NCBI's gi or RefSeq accession or UniProt's accession numbers. In addition you can enter a keyword to search for such as species, function, or cellular location. You may also search for secreted proteins of a specific species or perform BLAST (Basic Local Alignment Search Tool) search against our fungal database. When a keyword or species secretome search is performed, a list of results will be displayed with an identifier to the left followed by a description. The identifier is a link and clicking on it will display the details page for that protein. Similarly, searching for a specific protein by gi or accession will display that particular protein. This page shows the results of the different tests performed on the protein along with the sequence in FASTA format and any available manually curated data.

The Web page is divided up into five main sections: Search individual proteins by ID or keyword(s), Search secretome information by species, BLAST search, and Community Annotation.

Search by ID or Key Words

This section allows searching for a specific protein by using NCBI's RefSeq accession or gi number or UniprotKB's accession number. A search by keyword(s) will return a list of proteins containing the keyword(s) based on the UniProt Protein name. Details of an individual protein's results may be found by clicking on the identifier. For each protein which has been tested in our database, the results of those tests are displayed. The first area includes the various identifiers from NCBI and UniProt along with a clickable direct link to those sites. Also listed are the species, RefSeq definition, UniProt name, and a UniProt annotation for subcellular location (if any). The second area is a summary of the test results consisting of a yes/no for prediction of a secreted protein for each test. Also listed is a

[15]It is an online resource available at http://proteomics.ysu.edu/secretomes/fungi.php.

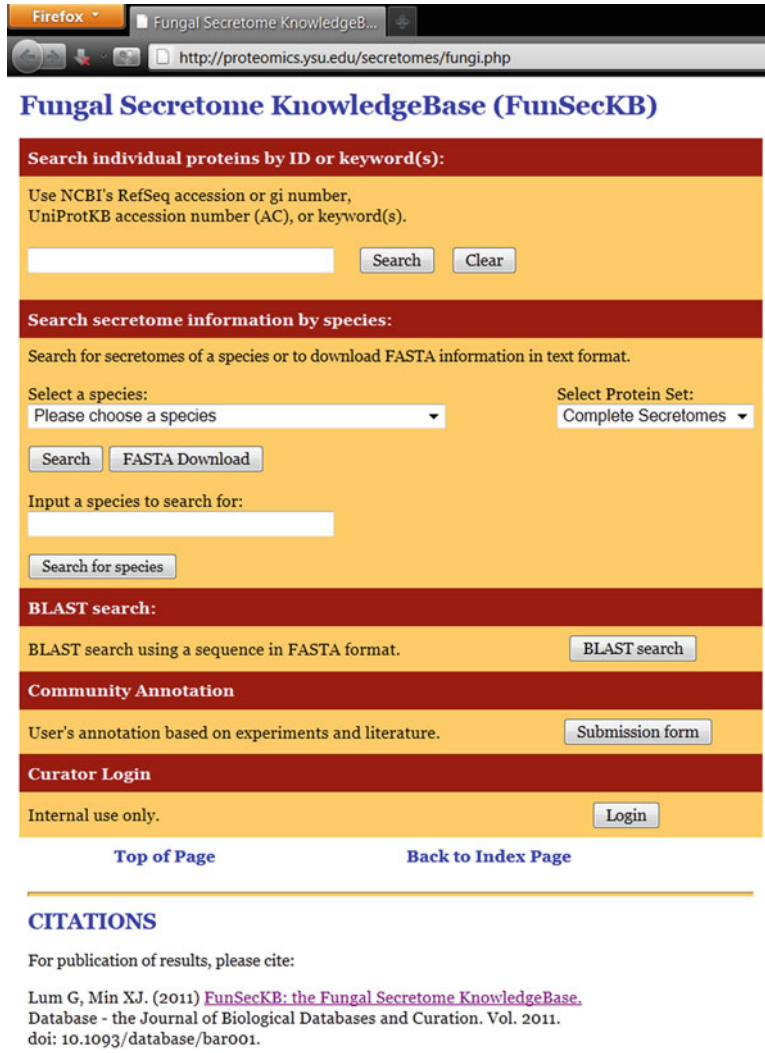


Fig. 54.7 Home page of Fungal Secretome Knowledge-Base

conclusion of whether or not this protein is belonged to a **Secretome**: based on our own combination prediction algorithm as mentioned above, that is, SignalP/Phobius/WoLF PSORT predicted to have a signal peptide, TMHMM predicted not have a transmembrane domain, and PS-Scan did not find an ER retention signal. The third area is the details for each of the tests along with a link to the original site's page on how to interpret the results (if available) or the Web site for the program. After the test results is listed the protein sequence used in FASTA format and if manual curation was done for the particular protein, the experimental evidence and the PubMed reference to the paper is given.

Search or Download Secretomes by Species

This section allows searching by species of secreted proteins, which are either predicted or curated. You can either select from a drop-down menu one of 53 species or manually input a species to search for. When using the drop-down menu, you may also select a protein set, either

Complete Secretome or Curated Proteins. The complete secretome is all proteins in the species predicted or curated by UniProt or our curator to be secretomes. From these options you may either search or download the FASTA. Search gives a listing of proteins similar to searching for keyword as above where an individual protein may be clicked to view details. FASTA download allows you to download the FASTA for a particular species. When "Curated Proteins" is selected, a list of available proteins is given on a Web page which may be copied and pasted. When "Complete Secretomes" is selected a window will appear allowing you to download and save a ".fas" file for that species since usually the entire FASTA file would be too large to display on screen.

BLAST Search

This section allows either a BLASTP or a BLASTX search against either of our two fungal databases. One is the secretome database containing our predicted and confirmed proteins and the other is for all fungal proteins in our database. The input format is a sequence or file in FASTA format. The NCBI BLAST page provides more information about how to use BLAST.[16]

Community Annotation

This section is a Community Annotation submission page allowing the user community to submit a protein for manual curation and addition into our database. The required entries are email, RefSeq gi and accession numbers, subcellular location of the protein, evidence and reference for the submission. Entries will be curated and if confirmed, entered into our database. Currently, we have manually curated secreted proteins from *Aspergillus niger* based on Tsang et al. [19] and *A. oryzae* based on Oda et al. [20] We would like to request the fungal secretome research community to submit experimentally verified secreted fungal proteins to FunSecKB using this utility. Once a protein has been curated, it will be permanently included as part of the database.

[16] http://blast.ncbi.nlm.nih.gov/Blast.cgi.

Acknowledgments We thank Dr. Gary Walker for his mentoring support and Jessica Orr for assistance in manual data curation. The work is supported by Youngstown State University (YSU) Research Council (Grants 2009-10 #04-10 and 2010-2011 #12-11), YSU Research Professorship (2009–2011), and the College of Science, Technology, Engineering, and Mathematics Dean's reassigned time for research to XJM.

References

1. Tjalsma H, Bolhuis A, Jongbloed JD, Bron S, van Dijl JM (2000) Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome. Microbiol Mol Biol Rev 64:515–547
2. Greenbaum D, Luscombe NM, Jansen R, Gerstein M (2001) Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. Genome Res 11:1463–1468
3. Hathout Y (2007) Approaches to the study of the cell secretome. Expert Rev Proteomics 4:239–248
4. Simpson JC, Mateos A, Pepperkok R (2007) Maturation of the mammalian secretome. Genome Biol 8:211
5. Lum, G. and Min, X. J. (2011) FunSecKB: the Fungal Secretome KnowledgeBase. *Database* 2011: doi:10.1093/database/bar001.
6. O'Toole N, Min XJ, Storms R, Butler G, Tsang A (2006) Sequence-based analysis of fungal secretomes. Appl Mycol Biotechnol 6:277–296
7. Min XJ (2010) Evaluation of computational methods for secreted protein prediction in different eukaryotes. J Proteomics Bioinformatics 3:143–147
8. Choi J, Park J, Kim D et al (2010) Fungal secretome database: integrated platform for annotation of fungal secretomes. BMC Genomics 11:105
9. Bendtsen JD, von Nielsen H, Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795
10. Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 30:1005–1016
11. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580
12. Käll L, Krogh A, Sonnhammer ELL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338:1027–1036
13. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ et al (2007) WoLF PSORT: protein localization predictor. Nucleic Acids Res 35:W585–W587 (Web Server issue)

14. Sigrist CJA, Cerutti L, de Casro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A et al (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38:161–166
15. Poisson G, Chauve C, Chen X, Bergeron A (2007) FragAnchor a large scale all Eukaryota predictor of glycosylphosphatidylinositol-anchor in protein sequences by qualitative scoring. Genomics Proteomics Bioinformatics 5:121–130
16. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442–451
17. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16:412–424
18. Menne KM, Hermjakob H, Apweiler R (2000) A comparison of signal sequence prediction methods using a test set of signal peptides. Bioinformatics 16:741–742
19. Tsang A, Butler G, Powlowski J et al (2009) Analytical and computational approaches to define the *Aspergillus niger* secretome. Fungal Genet Biol 46:S153–S160
20. Oda K, Kakizono D, Yamada O, Iefuji H, Akita O, Iwashita K (2006) Proteomic analysis of extracellular proteins from *Aspergillus oryzae* grown under submerged and solid-state culture conditions. Appl Environ Microbiol 72:3448–3457