Yeast
Yeast 2003; **20**: 595–610.
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI:** 10.1002/yea.988

Yeast Functional Analysis Report

An analysis of the *Candida albicans* genome database for soluble secreted proteins using computer-based prediction algorithms

Samuel A. Lee[1,2*†], Steven Wormsley[1†], Sophien Kamoun[3], Austin F. S. Lee[4,5], Keith Joiner[1] and Brian Wong[1,2]

[1] *Infectious Diseases Section, Department of Medicine, Yale University School of Medicine, New Haven, CT, USA*
[2] *Infectious Diseases Section, Department of Medicine, VA Connecticut Healthcare System, West Haven, CT, USA*
[3] *Department of Plant Pathology, The Ohio State University, Ohio Agricultural Research and Development Center, Wooster, OH, USA*
[4] *Department of Mathematics and Statistics, Boston University, Boston, MA, USA*
[5] *Center for Health Quality, Outcomes, and Economic Research, Bedford VA Hospital, Bedford, MA, USA*

**Correspondence to:*
Samuel A. Lee, Infectious Diseases Section, VA Connecticut Healthcare System, 950 Campbell Avenue, Building 8 (111-I), West Haven, CT 06516, USA.
E-mail: Samuel.Lee@yale.edu

† *These authors contributed equally to this work.*

Received: 10 October 2002
Accepted: 10 January 2003

Abstract

We sought to identify all genes in the *Candida albicans* genome database whose deduced proteins would likely be soluble secreted proteins (the secretome). While certain *C. albicans* secretory proteins have been studied in detail, more data on the entire secretome is needed. One approach to rapidly predict the functions of an entire proteome is to utilize genomic database information and prediction algorithms. Thus, we used a set of prediction algorithms to computationally define a potential *C. albicans* secretome. We first assembled a validation set of 47 *C. albicans* proteins that are known to be secreted and 47 that are known not to be secreted. The presence or absence of an N-terminal signal peptide was correctly predicted by SignalP version 2.0 in 47 of 47 known secreted proteins and in 47 of 47 known non-secreted proteins. When all 6165 *C. albicans* ORFs from CandidaDB were analysed with SignalP, 495 ORFs were predicted to encode proteins with N-terminal signal peptides. In the set of 495 deduced proteins with N-terminal signal peptides, 350 were predicted to have no transmembrane domains (or a single transmembrane domain at the extreme N-terminus) and 300 of these were predicted not to be GPI-anchored. TargetP was used to eliminate proteins with mitochondrial targeting signals, and the final computationally-predicted *C. albicans* secretome was estimated to consist of up to 283 ORFs. The *C. albicans* secretome database is available at http://info.med.yale.edu/intmed/infdis/candida/ Copyright © 2003 John Wiley & Sons, Ltd.

Keywords: **secreted proteins; genomics; yeast; fungi**

Introduction

The prevalence of invasive candidiasis has increased dramatically. *Candida* spp. have become the fourth most commonly isolated microorganism from the bloodstream of hospitalized patients in the USA and sixth most common nosocomial pathogen overall (Emori and Gaynes, 1993; Jarvis, 1995). Although *Candida albicans* is an increasingly important opportunistic pathogen, an incomplete understanding of *Candida* pathogenesis and cell biology has limited our ability to diagnose and treat candidiasis.

C. albicans has a diploid genome and has no clearly defined sexual cycle (Hull *et al*., 2000; Magee and Magee, 2000). Consequently, classical genetic approaches have been of limited value for studying this organism. Recent application of molecular genetic techniques in the analysis of medically important fungi has significantly

enhanced fungal pathogenesis research. Important developments in the study of *C. albicans* biology and pathogenesis include the cloning and sequencing of many individual genes, development of integrative and episomal DNA transformation systems (De Backer *et al*., 2000), chromosomal mapping (Tait *et al*., 1997) and the near completion of a genome sequencing project (Magee, 1998; Scherer and Magee, 1990). The *C. albicans* genome sequencing project at Stanford University (**http://www-sequence.stanford.edu/group/candida**) (Tzung *et al*., 2001) has already identified >6000 partial and complete *C. albicans* genes. Based on annotation information from 6165 ORFs in CandidaDB (**http://genolist.pasteur.fr/CandidaDB/**), approximately 3400 of these *C. albicans* genes are structural homologues of known genes from *Saccharomyces cerevisiae*; however, the functions of most of the remaining 2700 genes or gene fragments are unknown. Thus, although our knowledge of *C. albicans* genome structure is growing rapidly, our challenge now is to utilize this information to understand the functional significance of these genes, particularly in relation to *C. albicans* biology and pathogenesis.

Numerous algorithms for prediction of protein structure and function are available either as computer applications or as Internet-based programs, and several have been used for preliminary functional analyses of large sets of predicted proteins. Recent analyses of entire yeast genome databases have included identification of GPI-anchored proteins in *S. cerevisiae* (Caro *et al*., 1997), a comprehensive BLAST analysis of *C. albicans* homologues of *S. cerevisiae* sexual cycle genes (Tzung *et al*., 2001) and a prediction of the subcellular localization of the entire *S. cerevisiae* proteome (Kumar *et al*., 2002). Thus, one approach to rapidly analyse an entire genome is to utilize database information and computer-based algorithms to predict structure and/or function (Tjalsma *et al*., 2000; Kamoun *et al*., 2001).

In *C. albicans*, as in other eukaryotes, proteins are typically targeted for entry into the general secretory pathway by the presence of a N-terminal signal sequence. Signal sequences have a tripartite structure characterized by a central hydrophobic core (h-region) usually consisting of 6–15 amino acid (aa) residues which is flanked by hydrophilic N- and C-terminal regions (Martoglio and Dobberstein, 1998). The h-region is important for correct targeting and membrane insertion of the peptide. The polar C-terminal region often contains helix-breaking proline and glycine residues and small uncharged residues at the −3 and −1 positions which determine the signal peptide cleavage site (von Heijne, 1990). The polar N region is variable in length and frequently is positively charged. Although some proteins lacking N-terminal signal sequences reach the extracellular space, the majority of soluble secreted proteins in *C. albicans* are likely to be transported via the general secretory pathway. Therefore, we took advantage of SignalP version (v)2.0, a program that accurately identified eukaryotic signal peptides (Nielsen *et al*., 1997, 1999; Nielsen and Krogh, 1998) and other predictive algorithms to define a computational secretome of *C. albicans* from the genome sequences.

Methods

We reasoned that soluble secreted proteins should have the following characteristics: (a) an N-terminal signal peptide; (b) no transmembrane domains; (c) no GPI-anchor site; and (d) no localization signal predicted to target the protein to mitochondria or other intracellular organelles. ORFs fulfilling these four criteria gained inclusion in the set of soluble secreted proteins we have defined as the computational secretome.

Data sets

In order to test our SignalP criteria, we assembled a validation set consisting of 47 *C. albicans* proteins that are known to be secreted (or members of known families of secreted proteins) and 47 that are known not to be secreted (see Table 1 and supplementary data). Next, we retrieved the entire set of non-redundant open reading frames (ORFs) from the *C. albicans* genome database from CandidaDB (**http://genolist.pasteur.fr/CandidaDB/**) and divided it into three manageable partial databases. Sequence data from CandidaDB was obtained from the Stanford Genome Technology Center website at **http://www-sequence.stanford.edu/group/candida**. This sequencing of *C. albicans* was accomplished with the support of the NIDR and the Burroughs Wellcome Fund.

Table 1. *Candida albicans* known proteins used as validation set

Gene	Accession No.	Description
A. Secretory proteins		
ALS1.5eoc	CA0909	Agglutinin-like protein, 5′-end
ALS10	CA0448	Agglutinin like protein
ALS11.5f	CA1425	Agglutinin-like protein, 5′-end
ALS2.5f	CA1473	Agglutinin-like protein, 5′-end
ALS3.5eoc	CA0591	Agglutinin-like protein, 5′-end
ALS4.5f	CA1527	Agglutinin-like protein, 5′-end
ALS5	CA2852	Agglutinin-like protein
ALS6	CA5713	Agglutinin-like protein
ALS7	CA5699	Agglutinin-like protein
ALS9.5eoc	CA0315	Agglutinin-like protein, 5′-end
BGL21	CA1541	*endo*-β-1,3-Glucanase
CFL1	CA3460	Ferric reductase
CHT1	CA5859	Endochitinase 1 precursor
CHT2	CA1051	Chitinase 2 precursor
CHT3	CA5987	Chitinase 3 precursor
EXG1	CA0822	Glucan 1,3-β-glucosidase
HEX1	CA4276	β-N-Acetylglucosaminidase
HWP1	CA2825	Hyphal wall protein
HYR1	CA1576	Hyphally regulated protein
KRE9	CA2958	Cell wall synthesis protein
LIP1	CA1079	Secretory lipase
LIP10	CA4757	Secretory lipase
LIP2	CA3068	Secretory lipase
LIP3	CA4731	Secretory lipase
LIP4	CA3182	secretory lipase
LIP5	CA4417	Secretory lipase
LIP6	CA4756	Secretory lipase
LIP7	CA5556	Secretory lipase
LIP8	CA1241	Secretory lipase
LIP9.exon1	CA4423	Secretory lipase 9, exon 1
LIP9.exon2	CA4422	Secretory lipase 9, exon 2
PHR1	CA4857	GPI-anchored pH-responsive glycosyl transferase
PHR2	CA3867	pH-Regulated protein 2
PLB1	CA1975	Phospholipase B
PLB2	CA0825	Phospholipase B
PLB3	CA3834	Phospholipase B (by homology)
PLB4.5f	CA0185	Phospholipase, 5′-end (by homology)
PLB5	CA2223	Putative phospholipase B precursor
SAP1	CA2660	Secreted aspartyl proteinase
SAP2	CA3138	Aspartic protease
SAP3	CA6065	Secreted aspartyl proteinase
SAP4	CA2055	Secreted aspartyl proteinase
SAP5	CA2499	Secreted aspartyl proteinase 5
SAP6	CA0968	Secreted aspartyl protease
SAP7	CA1929	Secreted aspartyl proteinase 7
SAP8	CA1266	Aspartic protease
SAP9	CA4700	Aspartyl proteinase 9 (by homology)
B. Non-secretory proteins		
AAF1	CA5726	Adhesion and aggregation-mediating surface antigen
ACT1	CA5255	Actin
ADE2	CA6139	Phosphoribosylaminoimidazole carboxylase
ARD1	CA6015	Protein N-acetyltransferase subunit
ARG4	CA4292	Argininosuccinate lyase
ARO4	CA1484	3-Dehydro-deoxyphosphoheptonate aldolase, tyrosine-inhibited
CAP1	CA0183	Transcriptional activator
CBF1	CA2473	Putative centromere binding factor 1
CBK1	CA2022	Serine/threonine protein kinase
CDC10	CA4259	Cell division control protein
CDC25	CA4698	Cell division cycle protein
CDC3	CA0844	Cell division control protein
CLA4	CA1710	Protein kinase homologue
CLA4	CA1710	Protein kinase homologue
CPH1	CA0154	Transcription factor
CPP1	CA4721	Probable protein-tyrosine phosphatase
EFG1	CA2787	Enhanced filamentous growth factor
FAB1	CA2179	Phosphatidylinositol 3-phosphate 5-kinase
FAS2.5f	CA6105	Fatty-acyl-CoA synthase, α-chain, 5′-end

Prediction algorithms

We then queried the validation set and the entire *C. albicans* ORF set with SignalP v2.0 (**http://www.cbs.dtu.dk/services/SignalP-2.0/**) to identify N-terminal signal peptides. We defined a positive SignalP hit as the simultaneous presence of three criteria: (a) signal peptide predicted by SignalP-NN; (b) signal peptide predicted by SignalP-HMM; and (c) signal peptide cleavage site located within 10–40 aa from the N-terminus.

Next, we analysed the set of ORFs predicted to encode proteins with N-terminal signal peptides with the following prediction algorithms to determine whether three additional characteristics were present (Table 2). TMHMM (**http://www.cbs.dtu.dk/services/TMHMM/**) was used to predict transmembrane domains (Krogh *et al*., 2001), big-PI Predictor (**http://mendel.imp.univie.ac.at/**

Table 1. Continued

Gene	Accession No.	Description
GAL1	CA4040	Galactokinase
GSP1	CA2675	GTP-binding protein
HEM3	CA0306	Porphobilinogen deaminase
HIS1	CA4792	ATP phosphoribosyltransferase
HK1	CA4676	Histidine kinase
HOG1	CA4677	Ser/thr protein kinase of MAPK family
IMH3.exon1	CA1246	IMP dehydrogenase, exon 1
LEU2	CA5618	Isopropyl malate dehydrogenase
MET3	CA5238	ATP sulphurylase
MIG1	CA1593	Transcriptional regulator
MIG1	CA1593	Transcriptional regulator
MKC1	CA5865	ser/thr Protein kinase of MAP kinase family
NAG1	CA1130	Glucosamine-6-phosphate deaminase
NMT1	CA1063	N-Myristoyltransferase
NRG1	CA5289	Similar to transcriptional repressor Nrg1p/Nrg2p
PFY1	CA3897	Profilin, BINDS TO ACTIN
PMI40	CA0988	Mannose-6-phosphate isomerase (phosphomannose isomerase) (PMI)(phosphohexomutase)
RHO1	CA2866	GTP-binding protein of the rho subfamily of ras-like proteins (by homology)
SEC18.5f	CA5270	Vesicular fusion protein by homology, 5′ end
SEC4	CA2681	GTP-binding protein
SNF1	CA3361	Serine/threonine protein kinase
SSK1	CA5233	Putative response regulator two-component phosphorelay gene
TPS1	CA4084	Trehalose-6-phosphate synthase
TUP1	CA3852	General transcription repressor
URA3	CA2801	Orotidine-5-monophosphate decarboxylase (*Candida albicans*)
VPS34	CA0149	1-Phosphatidylinositol 3-kinase
YPT1	CA5077	GTP-binding protein of the rab family (by homology)
YRB1	CA5822	GTPase-activating protein (by homology)

gpi/gpi_server) was used to identify potential GPI-anchor sites (Eisenhaber *et al*., 1999, 2001), and TargetP v1.01 (**http://www.cbs.dtu.dk/services/TargetP/**) was used to identify mitochondrial localization sequences (Emanuelsson *et al*., 2000). Because some ORFs in CandidaDB are partial, in the case of ORFs containing only the 5′ end of a gene, the corresponding 3′ end of the gene was retrieved from CandidaDB when available and used to query big-PI Predictor for the GPI-anchor analysis. The final dataset comprises all the ORFs whose deduced proteins are potentially soluble secreted proteins in *C. albicans* according to these four major characteristics.

Properties of the computational secretome

As a supplementary analysis, we compared subcellular localization data of *S. cerevisiae* homologues from the Yeast Protein Localization server (**http://bioinfo.mbb.yale.edu/genome/localize/**), which integrates data derived from genome-wide experimental and predicted subcellular localization studies (Drawid and Gerstein, 2000; Kumar *et al*., 2002; Drawid *et al*., 2000; Alexandrov and Gerstein, 2001). Annotation information directly from CandidaDB was used to identify *C. albicans* and *S. cerevisiae* homologues for comparison, and no additional criteria was imposed on these assignments to define homology.

Statistical analysis

We used a discriminant analysis (Kleinbaum *et al*., 1998) based on Mean S and HMM scores from SignalP to analyse the validation set and derive a discriminant function. This discriminant function was applied to the validation set and then to the SignalP predictions of the entire set of *C. albicans* ORFs and used to re-assign classifications to secretory and non-secretory categories.

Results

When the 47 known secretory proteins were analysed with SignalP, the S scores were all >0.6 and the HMM scores were all >0.8. In contrast, the 47 non-secretory *C. albicans* proteins all had S scores <0.25 and HMM scores <0.1 (Figure 1A). The standard criteria provided by SignalP correctly predicted that all 47 secreted proteins had N-terminal signal peptides (SP^+) and that all 47 non-secreted proteins did not (SP^-). In order to generate criteria for predicting the presence or absence of N-terminal signal peptides specifically in *C. albicans*, we used a statistical discriminant analysis based on Mean S and HMM scores from SignalP to derive prediction parameters for the unknowns. The derived discriminant function based on the validation set was: $L = -918.235 - 123.455^*$(Mean S

Table 2. Summary of prediction algorithms used

Algorithm	Prediction	Validation set	Accuracy (%)	Comments	Reference
SignalP v2.0	N-terminal signal peptide	SWISS-PROT version 29	97	Accuracy reported is for eukaryotic data set	Nielsen *et al.*, 1997 **http://www.cbs.dtu.dk/services/SignalP-2.0/**
TMHMM v2.0	Transmembrane domains	Set of 160 experimentally known transmembrane proteins and 645 soluble proteins	97–98	Accuracy reported refers to individual transmembrane helices. Accuracy is 77.5% for correct topology of protein	Krogh *et al.*, 2001 **http://www.cbs.dtu.dk/services/TMHMM/**
		Set of 188 experimentally known transmembrane proteins and 634 soluble proteins	68 or greater	Independent evaluation of 16 different algorithms to predict transmembrane domains. TMHMM was the best performing program in this evaluation	Moller *et al.*, 2001
big-PI Predictor	GPI-anchor site	Set of 177 proteins from SWISS-PROT and SWISS-NEW	>80		Eisenhaber *et al.*, 1999, 2001 **http://mendel.imp.univie.ac.at/gpi/gpi_server**
TargetP v1.01	Mitochondrial or other localization sequence	Set of 2738 mitochondrial and 1652 other proteins from SWISS-PROT	90	Accuracy reported is for non-plant sequences	Emanuelsson *et al.*, 2000 **http://www.cbs.dtu.dk/services/TargetP**

Accuracy is defined as concordance of computational algorithm with experimentally-derived data.

score) + 1983.44*(HMM score), where L values <0 predicted classification to the non-secretory group, and L values >0 predicted classification to the secretory group (an L value of 0 is indeterminate). When the discriminant function was applied to the 94 proteins in the validation set, none required re-classification.

When all 6165 ORFs from CandidaDB were analysed using SignalP v2.0, 83.8% of deduced proteins either had an S score >0.7 and HMM score >0.8 or an S score <0.25 and HMM score <0.4, and the remaining ORFs had intermediate mean S and HMM scores, thus separating most ORFs into a clear bimodal distribution (Figure 1B). Using our three standard SignalP criteria (SP^+ by mean S score, SP^+ by HMM score, signal peptide cleavage site within 10–40 aa of N-terminus), we predicted that 495 of the 6165 ORFs encoded proteins with N-terminal signal peptides. When our *C. albicans*-derived discriminant function was applied to all 6165 ORFs, the classifications were nearly identical except for three of 495 predicted secretory and five of 5607 predicted non-secretory proteins. Because our approach is intended to be inclusive rather than exclusive, we re-assigned only the five ORFs identified as 'non-secretory' by SignalP to the secretory group and analysed these separately (Table 3).

When the 495 deduced proteins predicted to have N-terminal signal peptides were analysed with TMHMM, 103 were predicted to have two or more transmembrane domains, 97 were predicted to have one transmembrane domain, and 295 were predicted to have no transmembrane domains. Of the 97 deduced proteins predicted to have one transmembrane domain, the transmembrane domain was located within the first 40 N-terminal amino acids in 55. Because TMHMM may not distinguish signal peptides from transmembrane domains, the 295 deduced proteins with no transmembrane domains and the 55 deduced proteins with a single transmembrane domain within 40 aa of the N-terminus were considered to be 350 potential soluble secreted proteins (Figure 2).

Next, to identify GPI-anchored proteins which might not be extracellularly secreted, the database of 495 SP^+ ORFs was queried with big-PI Predictor. Because *ALS1*, *ALS3*, *ALS4* and *ALS5* ORFs

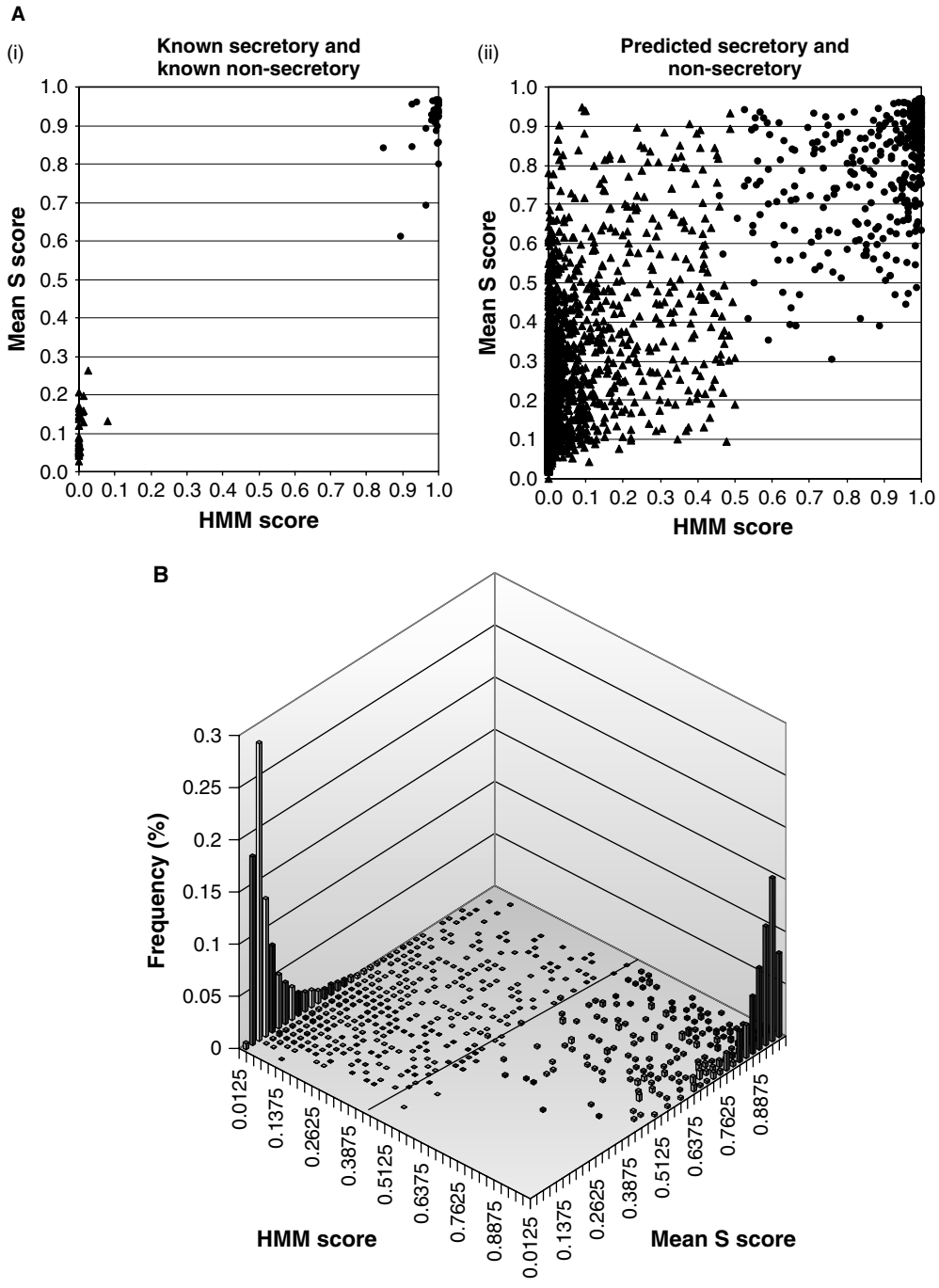

Figure 1. (A) Distribution of SignalP v2.0 scores for (i) 47 known and annotated *C. albicans* secreted and 47 non-secreted proteins and (ii) 6165 ORFs identified from CandidaDB. Raw Mean S and HMM scores were plotted for ORFs encoding proteins in the validation set of known secretory and non-secretory *C. albicans* proteins, and then for the entire set of 6165 ORFs from CandidaDB. Unmodified SignalP predictions are represented as follows: solid circle, presence of a Signal peptide; solid triangle, absence of a signal peptide. (B) Frequency plot of secretory and non-secretory proteins in *C. albicans.* Mean S and HMM scores for the entire set of *C. albicans* ORFs from CandidaDB are shown. The calculated discriminant function generated from the validation set scores is shown as a solid line on the X–Y axis

Table 3. (A) Discriminant analysis of secretory and non-secretory proteins. After generating a discriminant function based on data from the validation set, the SignalP scores for the set of *C. albicans* ORFs from CandidaDB were analysed. The majority of ORFs had concordant predictions using the two methods. The discriminant analysis re-classified five non-secretory predictions to secretory, and three secretory predictions to non-secretory. (B) List of mis-matches between SignalP prediction and discriminant analysis.

(A)

		Discriminant analysis		
		Secretory	**Non-secretory**	**Total**
SignalP analysis	Secretory	492 (8.06%)	3 (0.05%)	495 (8.20%)
	Non-secretory	5 (0.08%)	5602 (91.81%)	5607 (91.80%)
	Total	497 (8.14%)	5605 (91.86%)	6102*

(B)

Gene	Accession No.	Mean S score	HMM score	L score	Trans-membrane domains	GPI	Mito-chondrial SS	Function
Group prior = secretory by SignalP								
IPF11508	CA3023	0.572	0.469	−32.2095	3	N	N	Unknown; similarity to Sc integral membrane proteins Rta1p and Rtm1p
IPF6880	CA2185	0.473	0.443	−55.2707	4	N	N	Unknown; no significant homology to *S. cerevisiae*
IPF8760	CA4221	0.722	0.459	−48.6262	1/SP**	N	N	Unknown; no significant homology to *S. cerevisiae*
Group prior = non-secretory by SignalP								
IPF11449	CA0145	0.093	0.479	20.3525	0	N	Yes	Unknown
IPF1331	CA5115	0.45	0.495	8.0141	4	N	Yes	Unknown
IPF7823	CA3562	0.303	0.483	2.3607	0	N	N	Unknown
URA7	CA1635	0.308	0.499	9.5769	1	N	N	CTP synthase 1 (by homology); Sc homologue is a cytosolic protein
VMA5	CA0711	0.189	0.500	15.3918	0	N	N	H^+-ATPase V1 domain 42 kDa subunit (by homology); Sc homologue is a vacuolar membrane protein

*ORFs predicted to have N-terminal signal peptides by SignalP v2.0 but that did not fulfil our three standard criteria were classified as indeterminate and excluded from this analysis. Thus, percentages shown are based on 6102 analysable ORFs. **Probably represents a signal peptide, not a true transmembrane domain.

consist of 5′ fragments in CandidaDB, the corresponding 3′ fragments were retrieved and used for this analysis. After excluding SP^+ ORFs encoding proteins with greater than one transmembrane-domain, this algorithm identified a total of 58 potential GPI-anchored proteins. In the database of 350 SP^+ ORFs used for further analysis to predict the secretome, there were 50 predicted GPI-anchored proteins (Table 4).

Because in eukaryotic cells secretory proteins may be targeted to intracellular organelles rather than secreted extracellularly, we used TargetP (**http://www.cbs.dtu.dk/services/TargetP/**) to identify mitochondrial targeting sequences in order to eliminate these ORFs from the dataset. In the set of 495 SP^+ ORFs, 21 ORFs were excluded due to the presence of a mitochondrial localization signal in 14 ORFs or other localization signal in seven ORFs (Table 5).

Functional information from CandidaDB was reviewed for the 495 SP^+ ORFs, and 244 of these ORFs encode deduced proteins of unknown

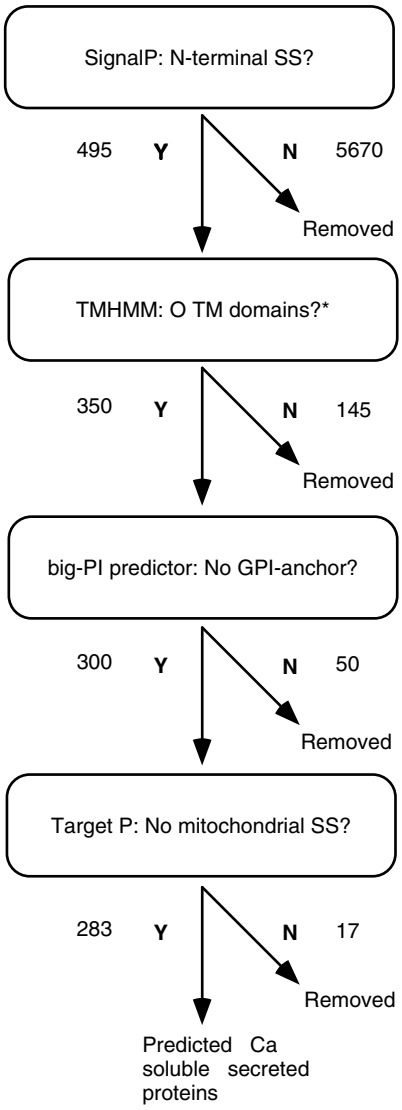

Figure 2. Flowchart of strategy used to identify *C. albicans* soluble secreted proteins using a series of prediction algorithms. A positive SignalP hit was defined as the simultaneous presence of three criteria: (1) Signal peptide predicted by SignalP-NN; (2) Signal peptide predicted by SignalP-HMM; and (3) Signal peptide cleavage site located within 10–40 aa from the N-terminus. *Because TMHMM may not distinguish Signal peptides from transmembrane domains, 295 deduced proteins with no transmembrane domains and 55 deduced proteins with a single transmembrane domain within 10–40 aa of the N-terminus were considered to be 350 potential soluble secreted proteins. Of 58 ORFs predicted to encode GPI-anchored proteins in the set of 495 SP^+ ORFs, 50 remained after the analysis with TMHMM. After eliminating ORFs predicted to encode proteins with mitochondrial signal sequences, 283 ORFs were predicted to be the set of ORFs encoding soluble secreted proteins

function. After the 495 SP^+ ORFs were analysed with TMHMM, big-PI Predictor, and TargetP, 283 remaining ORFs fulfilled our four criteria: (a) presence of an N-terminal signal peptide; (b) lack of a transmembrane domain (unless located at the extreme N-terminus); (c) absence of a GPI-anchor; and (d) no mitochondrial or other localization signal. We propose that these 283 SP^+ ORFs comprise the predicted secretome of *C. albicans*.

Of the 283 SP^+ *C. albicans* ORFs in the predicted secretome, 140 are of unknown function. The remaining 143 have an assigned function by homology to *S. cerevisiae* proteins (105) or are ORFs that encode known *C. albicans* proteins or members of known protein families (38). These 38 known *C. albicans* ORFs encode 25 extracellularly secreted proteins, 10 cell wall-associated proteins, two vacuolar proteins, and one ER-related protein (Table 6).

Comparison of these 283 SP^+ *C. albicans* ORFs to *S. cerevisiae* subcellular localization data identified 73 *S. cerevisiae* homologues that also are secretory pathway proteins, 24 membrane proteins, 22 mitochondrial proteins, seven vacuolar proteins, and 50 homologues with other subcellular localizations. No *S. cerevisiae* homologue was identified by CandidaDB for 124 ORFs (see supplementary data).

Discussion

Soluble secreted *C. albicans* virulence factors, such as the secreted aspartyl proteases (reviewed in Hoegl *et al.*, 1996; Hube *et al.*, 1997; Sanglard *et al.*, 1997) and extracellular phospholipases (reviewed in Ghannoum, 2000; Niewerth and Korting, 2001) have been studied in detail, and many of these are found either on the cell surface or in the extracellular environment. Members of the secreted aspartyl protease (Sap) family of proteins are differentially secreted extracellularly depending on strain and environmental conditions (White and Agabian, 1995). *C. albicans sap1, sap2* and *sap3* mutants, and a triple *sap4, sap5* and *sap6* null mutant are attenuated in virulence in a mouse model of invasive candidiasis (Hube *et al.*, 1997; Sanglard *et al.*, 1997). In addition to the signal peptide, the Sap propeptide is also important for proper secretion (Monod *et al.*, 2000). Extracellular phospholipases have also been implicated as virulence

Table 4. GPI-anchor predictions. A total of 58 ORFs are predicted to encode GPI-anchored proteins from the 495 SP^+ dataset; 35 ORFs are unnamed; 29 ORFs are of unknown function by homology. Analysis of the *ALS* family of genes is preliminary, due to partial and incomplete ORFs in CandidaDB

Gene name	Accession No.	Gene length	Protein length	Prediction	HMM score	Mean S score	Predicted TM	Description	Subcellular localization of Sc homologue
ALS1.5eoc	CA0909	1974	658	Signal peptide	0.997	0.940	0	Agglutinin-like protein, 5′-end	ER
ALS10	CA0448	4761	1586	Signal peptide	1.000	0.956	0	Agglutinin like protein	ER
ALS11.5f	CA1425	2859	952	Signal peptide	1.000	0.960	0	Agglutinin-like protein, 5′-end	ER
ALS3.5eoc	CA0591	2658	886	Signal peptide	0.980	0.912	0	Agglutinin-like protein, 5′-end	ER
ALS4.5f	CA1527	4782	1593	Signal peptide	1.000	0.956	0	Agglutinin-like protein, 5′-end	ER
ALS5	CA2852	4044	1347	Signal peptide	0.995	0.919	0	Agglutinin-like protein	ER
ALS6	CA5713	4101	1366	Signal peptide	0.996	0.899	1/SP	Agglutinin-like protein	ER
CRH11	CA0375	1362	453	Signal peptide	0.999	0.786	0	Probable membrane protein	ER
CRH12	CA1835	1515	504	Signal peptide	0.985	0.901	1	Cell wall protein	ER
CSA1	CA5585	3057	1018	Signal peptide	0.996	0.864	0	Mycelial surface antigen by homology	N/A
DFG5	CA4822	1356	451	Signal peptide	0.994	0.924	1	Required for filamentous growth	PM
EXG2	CA4180	1440	479	Signal peptide	0.999	0.909	0	Glucan 1,3-β-glucosidase-like by homology	ER
HWP1	CA2825	1905	635	Signal peptide	0.896	0.613	0	Hyphal wall protein	ER
HYR1	CA1576	2814	937	Signal peptide	0.995	0.937	0	Hyphally-regulated protein	N/A
IFF2	CA2714	3750	1249	Signal peptide	0.964	0.958	0	Unknown function	ER
IFF4	CA5819	4581	1526	Signal peptide	0.981	0.856	0	Unknown function	ER
IFF7	CA5468	3678	1225	Signal peptide	0.771	0.742	1	Unknown function	ER
IPF10662	CA3827	1179	392	Signal peptide	0.999	0.873	0	Unknown function	N/A
IPF10919	CA2625	660	219	Signal peptide	0.997	0.890	0	Similar to Flo1p (by homology)	ER
IPF11998	CA1898	1554	517	Signal peptide	0.995	0.907	1	Unknown function	N/A
IPF12022	CA3622	3147	1048	Signal peptide	0.836	0.869	0	Extracellular α-1,4-glucan glucosidase (by homology)	N/A
IPF12101	CA2557	660	219	Signal peptide	0.984	0.845	0	Mycelial surface antigen precursor (by homology to *Candida* gene *CSA1*)	N/A
IPF1218	CA4835	699	232	Signal peptide	0.981	0.819	0	Similar to superoxide dismutase (by homology)	CYT
IPF13070	CA3763	891	296	Signal peptide	1.000	0.962	1/SP	Unknown function	N/A
IPF1341	CA5112	1371	456	Signal peptide	0.998	0.813	1	Similarity to mucin proteins (by homology)	N/A
IPF14081	CA1553	924	307	Signal peptide	0.980	0.911	1	Unknown function	N/A
IPF14126	CA1313	999	332	Signal peptide	0.999	0.917	0	Unknown function	N/A
IPF14598	CA1360	2205	734	Signal peptide	0.824	0.608	1	Unknown function	N/A
IPF14706	CA1777	930	309	Signal peptide	1.000	0.961	1	Unknown function	N/A

Table 4. Continued

Gene name	Accession No.	Gene length	Protein length	Prediction	HMM score	Mean S score	Predicted TM	Description	Subcellular localization of Sc homologue
IPF15423	CA2737	951	316	Signal peptide	0.970	0.893	0	Putative superoxide dismutase (by homology)	N/A
IPF15442	CA0188	1155	384	Signal peptide	0.999	0.865	0	Unknown function	ER
IPF15581	CA1720	420	139	Signal peptide	0.995	0.880	0	Unknown function	N/A
IPF1580	CA5418	396	131	Signal peptide	0.988	0.647	0	Unknown function	ER
IPF15911	CA3623	3531	1176	Signal peptide	0.733	0.841	0	Unknown function	N/A
IPF15957	CA0171	255	84	Signal peptide	0.966	0.663	0	Unknown function	N/A
IPF19706	CA0647	723	240	Signal peptide	0.998	0.894	0	Unknown function	N/A
IPF20008	CA4124	342	113	Signal peptide	0.994	0.801	0	Unknown function	N/A
IPF20103	CA2502	2226	741	Signal peptide	0.998	0.929	0	Unknown function	N/A
IPF20148	CA3826	672	223	Signal peptide	0.999	0.845	0	Unknown function	N/A
IPF20161	CA4125	642	213	Signal peptide	0.998	0.755	0	Unknown function	N/A
IPF20169	CA4381	753	250	Signal peptide	1.000	0.915	0	Unknown function	N/A
IPF3233	CA2475	498	165	Signal peptide	0.999	0.939	0	Unknown function	N/A
IPF3844	CA2405	2262	753	Signal peptide	0.952	0.658	0	Unknown function	N/A
IPF4089	CA4863	1362	453	Signal peptide	0.781	0.882	0	Secretory aspartyl proteinase	ER
IPF4123	CA3642	690	229	Signal peptide	0.989	0.952	1/SP	Unknown function	N/A
IPF4299	CA4246	336	111	Signal peptide	1.000	0.887	0	Unknown function	N/A
IPF4722	CA3252	510	169	Signal peptide	0.993	0.807	0	Unknown Function	N/A
IPF4724	CA3253	816	271	Signal peptide	0.989	0.754	0	Unknown Function	N/A
IPF5185	CA1678	1602	533	Signal peptide	1.000	0.879	0	Putative cell wall protein (by homology)	ER
IPF8129	CA3630	681	226	Signal peptide	0.984	0.702	0	Unknown function	N/A
IPF8796	CA4800	1356	451	Signal peptide	0.965	0.934	0	Putative GPI-anchored protein related to Phr1, Phr2 and Phr3 (by homology)	ER
IPF9101	CA2548	594	197	Signal peptide	0.998	0.842	0	Unknown function	N/A
MID1	CA0203	1680	559	Signal peptide	0.999	0.899	0	Involved in Ca^{2+} influx during mating (by homology)	ER/PM
PLB5	CA2223	2265	754	Signal peptide	0.965	0.693	0	Putative phospholipase B precursor	ER
RBT1	CA2830	2145	714	Signal peptide	0.950	0.749	0	Repressed by TUP1 protein 1	N/A
RBT5	CA2558	726	241	Signal peptide	1.000	0.838	0	Repressed by TUP1 protein 5	ER
SAP9	CA4700	1635	544	Signal peptide	0.999	0.935	0	Aspartyl proteinase 9 (by homology)	ER
SSR1	CA5213	705	234	Signal peptide	0.998	0.888	0	Secretory stress response protein 1 (by homology)	ER/CW

factors involved in the pathogenesis of infection with *C. albicans* (Leidich *et al*., 1998; Mukherjee *et al*., 2001). The deduced protein of *C. albicans* *PLB1*, a phospholipase B, is predicted to have a stretch of hydrophobic amino acids at the amino terminus that likely serves as a signal peptide. The

Table 5. List of ORFs predicted by TargetP to contain mitochondrial and other intracellular targeting signals

Gene name	Accession No.	Gene length	Protein length	Prediction	HMM score	Mean S score	Predicted TM	Description	Subcellular localization of Sc homologue	TargetP
Mitochondrial										
ADH1	CA4765	1305	434	Signal peptide	0.983	0.740	0	Alcohol dehydrogenase	MIT	MIT
COQ3	CA2432	984	327	Signal peptide	0.716	0.533	0	3,4-Dihydroxy-5-hexaprenylbenzo-atemethyltransferase	MIT	MIT
CPA1	CA0874	1305	434	Signal peptide	0.987	0.488	0	Arginine-specific carbamoylphosphate synthase, small chain	CYT	MIT
DLD2	CA5942	1602	533	Signal peptide	0.678	0.786	0	D-Lactate ferrycytochrome *c* oxidoreductase	MIT	MIT
FTI1	CA2642	879	292	Signal peptide	0.865	0.620	0	Rad52 inhibitor	MIT	MIT
IPF19578	CA0371	2421	806	Signal peptide	0.992	0.882	0	Unknown function	MIT	MIT
IPF3361	CA4785	756	251	Signal peptide	0.905	0.506	0	Putative mitochondrial ribosomal protein S7 (by homology)	MIT	MIT
IPF7704	CA4114	564	187	Signal peptide	0.762	0.526	0	Unknown function	MIT	MIT
IPF8359	CA3383	456	151	Signal peptide	0.661	0.633	1	Unknown function	MIT	MIT
IPF864	CA5347	366	121	Signal peptide	0.917	0.667	0	Unknown function	NUC	MIT
IPF9370	CA3964	1716	571	Signal peptide	0.649	0.392	12	Unknown function	No homologue	MIT
LAT1	CA4875	1434	477	Signal peptide	0.880	0.598	0	Dihydrolipoamide *S*-acetyltransferase (by homology)	MIT	MIT
MGM1	CA2773	2667	888	Signal peptide	0.872	0.560	0	GTPase	MIT	MIT
MNT1	CA3469	1296	431	Signal peptide	0.559	0.751	1/SP	Mannosyltransferase involved in *n*-linked and *o*-linked glycosylation	ER/Golgi	MIT
Other										
CBP1	CA5559	1470	489	Signal peptide	0.888	0.389	0	Corticosteroid binding protein	NUC	Other
COF1	CA5409	435	144	Signal peptide	0.901	0.732	0	Cofilin	CYT	Other
IPF149	CA6127	1092	363	Signal peptide	0.589	0.355	6	Peroxisomal membrane protein (by homology)	No homologue	Other
IPF19608	CA0674	558	185	Signal peptide	0.761	0.305	2	Unknown function	No homologue	Other
IPF8950	CA2361	690	229	Signal peptide	0.664	0.389	0	Unknown function	MIT	Other
RPN2	CA4988	2859	952	Signal peptide	0.651	0.436	0	Proteasome regulatory subunit (by homology)	?CYT	Other
SOD1.3	CA4120	480	159	Signal peptide	0.852	0.569	0	Cu,Zn-superoxide dismutase, 3′-end	CYT	Other

family of *C. albicans* secretory lipases may also have a role in virulence (Fu *et al*., 1997; Hube *et al*., 2000). In addition, a number of secreted proteins that remain associated with the cell wall or membrane have been identified and shown to have a role in virulence, including the outer mannoprotein Hwp1 (Staab *et al*., 1999), the *ALS* family of genes (reviewed in Hoyer, 2001) and the pH-responsive genes *PHR1-2* (Bernardis *et al*., 1998; Ghannoum *et al*., 1995; Fonzi, 1999; Saporito-Irwin *et al*., 1995). Thus, it is apparent that the ability of *C. albicans* to transport proteins to the cell surface via the secretion pathway and to secrete degradative enzyme out of the cell is required for virulence and pathogenesis (reviewed in Haynes, 2001).

Although it is clear that detailed studies of individual genes and gene products are essential, it is also important to obtain a more global perspective on secreted proteins, including those involved in virulence. The use of computer-based prediction algorithms is a powerful, systematic, and rapid tool to obtain preliminary functional information on gene products of an entire genome. Information

Table 6. List of known genes in the final predicted *Candida albicans* secretome

Gene name	Accession No.	Gene length	Protein length	Prediction	HMM score	Mean S score	Predicted TM	Description	Secretory?
Soluble									
HEX1	CA4276	1689	562	Signal peptide	0.998	0.935	0	N-Acetylglucosaminidase	Y
LIP1	CA1079	1407	468	Signal peptide	0.999	0.968	0	Secretory lipase	Y
LIP10	CA4757	1398	465	Signal peptide	0.999	0.965	0	Secretory lipase	Y
LIP2	CA3068	1401	466	Signal peptide	0.999	0.941	0	Secretory lipase	Y
LIP3	CA4731	1416	471	Signal peptide	1.000	0.952	0	Secretory lipase	Y
LIP4	CA3182	1380	459	Signal peptide	0.995	0.968	0	Secretory lipase	Y
LIP5	CA4417	1392	463	Signal peptide	0.927	0.956	0	Secretory lipase	Y
LIP6	CA4756	1392	463	Signal peptide	0.995	0.930	0	Secretory lipase	Y
LIP7	CA5556	1281	426	Signal peptide	0.993	0.943	0	Secretory lipase	Y
LIP8	CA1241	1383	460	Signal peptide	0.985	0.964	0	Secretory lipase	Y
LIP9.exon1	CA4423	642	213	Signal peptide	0.940	0.961	0	Secretory lipase 9, exon 1	Y
LIP9.exon2	CA4422	792	263	Signal peptide	0.848	0.841	0	Secretory lipase 9, exon 2	Y
PLB1	CA1975	1818	605	Signal peptide	0.987	0.940	0	Phospholipase B	Y
PLB2	CA0825	1830	609	Signal peptide	0.998	0.962	0	Phospholipase B	Y
PLB4.5f	CA0185	1185	394	Signal peptide	1.000	0.958	0	Phospholipase, 5′-end (by homology)	Y
SAP1	CA2660	1176	391	Signal peptide	0.999	0.921	0	Secreted aspartyl proteinase	Y
SAP2	CA3138	1197	399	Signal peptide	0.999	0.935	0	Aspartic protease	Y
SAP3	CA6065	1197	399	Signal peptide	0.999	0.926	0	Secreted aspartyl proteinase	Y
SAP4	CA2055	1254	417	Signal peptide	0.997	0.926	0	Secreted aspartyl proteinase	Y
SAP5	CA2499	1257	418	Signal peptide	0.998	0.925	0	Secreted aspartyl proteinase 5	Y
SAP6	CA0968	1257	418	Signal peptide	0.998	0.925	0	Secreted aspartyl protease	Y
SAP7	CA1929	1767	588	Signal peptide	0.981	0.929	0	Secreted aspartyl proteinase 7	Y
SAP8	CA1266	1218	405	Signal peptide	0.998	0.855	0	Aspartic protease	Y
RBT4	CA0104	1077	358	Signal peptide	0.969	0.624	0	Repressed by TUP1 protein	Y?
RBT7	CA0169	918	305	Signal peptide	0.999	0.914	0	Repressed by TUP1	Y?

can then be analysed in global fashion to organize functional groupings of predicted proteins, or individually, in order to identify genes of particular interest for future experimental study.

Since one of our interests is secreted proteins associated with virulence, we queried the *C. albicans* genome database in an effort to identify all genes whose deduced proteins would likely be soluble secreted proteins in order to: (a) obtain a global perspective on secreted proteins in *C. albicans*; and (b) identify previously uncharacterized genes for further experimental study. We therefore used a series of prediction algorithms available on Internet-based servers to analyse the *C. albicans* genome database. First, we assembled a validation set of known *C. albicans* secretory and non-secretory proteins to train our prediction algorithm. We generated a discriminant function which was applied to the unknown ORFs to derive a new cut-off whereby re-assignments could be made. Then we used our criteria based on the SignalP v2.0 algorithm to identify 495 ORFs with N-terminal signal peptides from a total of 6165 *C. albicans* ORFs. Using the discriminant function we re-classified two ORFs predicted by SignalP to be non-secretory as secretory. Thus, approximately 8% of the entire *C. albicans* genome consists of SP^+ ORFs. In comparison, approximately 11%

Table 6. Continued

Gene name	Accession No.	Gene length	Protein length	Prediction	HMM score	Mean S score	Predicted TM	Description	Secretory?
Cell wall-associated									
ALS2.5f	CA1473	5271	1756	Signal peptide	1.000	0.956	0	Agglutinin-like protein, 5′-end	CW
ALS7	CA5699	6003	2000	Signal peptide	0.993	0.887	0	Agglutinin-like protein	CW
ALS9.5eoc	CA0315	2685	894	Signal peptide	0.998	0.934	0	Agglutinin-like protein, 5′-end	CW
BGL21	CA1541	927	308	Signal peptide	0.986	0.913	0	Endo-β-1,3-glucanase	CW
CHT1	CA5859	1389	462	Signal peptide	0.997	0.957	1/SP	Endochitinase 1 precursor	CW
CHT2	CA1051	1752	583	Signal peptide	1.000	0.857	0	Chitinase 2 precursor	CW
CHT3	CA5987	1704	567	Signal peptide	0.999	0.959	0	Chitinase 3 precursor	CW
KRE9	CA2958	816	271	Signal peptide	0.997	0.927	0	Cell wall synthesis protein	CW
PHR1	CA4857	1647	548	Signal peptide	0.966	0.893	0	GPI-anchored pH responsive glycosyl transferase	CW
PRA1	CA4399	900	299	Signal peptide	1.000	0.965	0	pH-Regulated antigen	CW?
Other									
APR1	CA4476	1260	419	Signal peptide	0.998	0.810	0	Aspartyl protease	VAC
CPY1.5f	CA2123	258	85	Signal peptide	0.998	0.815	0	Carboxypeptidase Y precursor, 5′-end	VAC
CYP51	CA5717	639	212	Signal peptide	0.932	0.891	1/SP	Cyclophilin-peptidylprolyl *cis–trans* isomerase or PPIase	ER

S. cerevisiae ORFs were predicted to encode signal peptides but a different prediction algorithm was used (Caro *et al*., 1997). Next, we used TMHMM to identify ORFs predicted to have no true transmembrane domains. In this subset, we identified 350 ORFs that fulfilled our criteria. Proteins with one or more transmembrane domains were eliminated as they were unlikely to be secreted extracellularly. However, because TMHMM does not necessarily distinguish signal peptides from transmembrane domains, if TMHMM predicted a transmembrane domain at the N-terminus, we did not exclude these ORFs from our dataset. We then identified 50 potential GPI-anchored proteins from this dataset (58 total from the SP^+ TM 0–1 dataset, or 50 total from the SP^+ TM 0 dataset). This is on the same order as the 51 GPI-anchored proteins predicted in *S. cerevisiae* using a similar analysis (Caro *et al*., 1997). Finally, we used TargetP to identify mitochondrial signal sequences to eliminate secretory proteins that are targeted to intracellular organelles, yielding a computationally-defined secretome of 283 ORFs.

Given the inherent limitations of the prediction algorithms, a minority of ORFs are probably assigned incorrectly. Our three SignalP criteria clearly separated the ORFs from the *C. albicans* genome into two distinct categories, although a small number of ORFs fell into an intermediate range. However, by using a discriminant analysis, we generated a function based on the validation sets to generate a new cut-off for assigning ORFs to secretory and non-secretory classifications. Thus, the vast majority of these SP^+ ORFs are most likely proteins that enter the general secretory pathway, and either are secreted extracellularly, GPI-anchored, or in some cases targeted to distinct intracellular organelles. Overall, we predicted that the potential *C. albicans* secretome, according to our set of four prediction algorithms consists of up to an estimated 283 proteins.

In this study, we defined the predicted type II secretome of *C. albicans*. We identified, as expected, genes whose proteins have signal peptides and are known to be cell wall-associated, including *EXG1* (exo-β-1,3-glucanase), *BGL2* (β-1,3-glucan transferase), *CHT1-3* (chitinases),

and *HEX1* (*β-N*-acetylglucosaminidase). We also identified genes whose proteins have signal peptides and are known to be secreted extracellularly, including: *SAP1-9* (secreted aspartyl proteases); *PLB1* (phospholipase B); *LIP1* (secreted lipase); and gene homologues of glucoamylase, carboxypeptidase Y, acid phosphatase and alkaline phosphatase. Interestingly, 160 of these ORFs are unnamed, and 140 of them are ORFs of unknown function.

However, some *C. albicans* proteins are known to reach the extracellular space independently of the Type II secretion pathway. It remains unclear how proteins such as enolase (Mason *et al*., 1989; Franklyn *et al*., 1990; Angiolella *et al*., 1996; Sundstrom and Aliaga, 1994), Hsp70 and Hsp90 (Matthews *et al*., 1988) reach the cell wall and/or extracellular space. At this point it is not possible to predict such extracellular proteins using bioinformatic approaches. Genes encoding cell wall-associated proteins that were correctly predicted to lack signal peptides in our database included: *ENO1* (enolase), *SSA1* (Hsp70), *PGK* (phosphoglycerate kinase) and GAPDH (*TDH1*). Thus, while the majority of secreted proteins in *C. albicans* would be expected to be transported via the general secretory pathway (Lee *et al*., 2001; Mao *et al*., 1999), there may be several potential non-*SEC* dependent pathways in *C. albicans* that permit proteins to reach the extracellular space. In addition to non-specific mechanisms such as cell lysis or leakage, other possibilities include efflux pumps of the *MDR* and *CDR* families (reviewed by White *et al*., 1998), non-classical transport mediated by *NCE1* (Cleves *et al*., 1996) and perhaps other unknown specific transporters.

In order to gain additional insight into the functional properties of these potential *C. albicans* secretory proteins in our dataset, we referred to the extensive subcellular localization data available for the corresponding *S. cerevisiae* homologues. Although no *S. cerevisiae* homologue was identified by CandidaDB for 124 ORFs, the majority of the evaluable *S. cerevisiae* homologues were secretory pathway proteins.

We also compared our database of predicted secretory proteins to an experimentally-derived set of *C. albicans* secreted proteins recently identified in a heterologous, genome-wide genetic screen. In this approach, in-frame fusions of *C. albicans* genomic DNA were fused to episomal vectors bearing mutant *suc2* alleles, encoding invertase lacking the signal peptide region in *S. cerevisiae*, such that growth on sucrose implies the presence of a signal peptide (Monteoliva *et al*., 2002). This screen identified 68 putatively exported *C. albicans* proteins. Of 54 ORFs which could be directly retrieved from CandidaDB, our identification of signal peptides using our three SignalP criteria were concordant in 50 cases (see supplementary data).

Our GPI-anchor predictions should be interpreted with caution, as the big-PI Predictor is not intended to be fungal-specific. A recent report predicts *C. albicans* to encode 54 GPI-anchored proteins (Sundstrom, 2002). Of 44 ORFs available in CandidaDB our predictions correlated in 29 cases.

Important limitations of our approach is that it relies on prediction algorithms with a defined error rate which could potentially be greater in specific organisms. In addition, there are gene fragments in CandidaDB which can potentially confuse the prediction algorithms; thus, results obtained with partial ORFs must be cross-checked to obtain relevant upstream or downstream sequences if available and evaluated cautiously. Finally, these prediction algorithms are useful for rapid preliminary analyses of large amounts of genomic data, but it must be emphasized that these are only predictions, which require experimental validation. Our approach was to be inclusive rather than exclusive, so overall these results probably represent an overestimation of the actual *C. albicans* secretome, especially since many ORFs in the genome database have not been confirmed experimentally and some ORFs may not be expressed. Alternatively, we may have inadvertently excluded secreted proteins, e.g. proteins encoded by ORFs not annotated by CandidaDB, particularly small ORFs that would not fulfil gene prediction criteria.

In future studies, we would like to examine the following questions using proteomics-based approaches to analyse *C. albicans* soluble secreted proteins: (a) can novel secreted proteins be identified, and what is their role in virulence?; (b) are there abundant proteins that are secreted but do not have signal peptides, and if so, how do they reach the extracellular space?; (c) what are the specific targeting signals in *C. albicans* that allow sorting of proteins to their proper intracellular destinations? Fortunately, the extensive

work done in *S. cerevisiae* will provide a roadmap toward answering some of these questions in this pathogenic yeast.

Acknowledgements

We thank Margaret Hostetter, Peter Novick and Craig Roy (all from Yale University) for helpful advice, and Birgit Eisenhaber (Research Institute of Molecular Pathology) for assistance with GPI-anchor predictions. We thank the Galar Fungail Consortium for CandidaDB, and the Stanford Genome Technology Center for the *Candida albicans* genome sequencing project. Sequence data for *Candida albicans* was obtained from the Stanford Genome Technology Center website at **http://www-sequence.stanford.edu/group/candida**. Sequencing of *Candida albicans* was accomplished with the support of the NIDR and the Burroughs Wellcome Fund. This work was supported by grants from the Department of Veterans' Affairs (Career Development Award to S.L. and Merit Review to B.W.) and the National Institute of Allergy and Infectious Diseases (R01 AI-47442 to B.W.).

References

Alexandrov V, Gerstein M. 2001. Calculating populations of subcellular compartments using density matrix formalism. *Int J Quant Chem* **85**: 693–696.

Beggah S, Lechenne B, Reichard U, Foundling S, Monod M. 2000. Intra- and intermolecular events direct the propeptide-mediated maturation of the *Candida albicans* secreted aspartic proteinase Sap1p. *Microbiol* **146**: 2765–2773.

Caro LH, Tettelin H, Vossen JH, *et al*. 1997. *In silicio* identification of glycosyl-phosphatidylinositol-anchored plasma-membrane and cell wall proteins of *Saccharomyces cerevisiae*. *Yeast* **13**: 1477–1489.

De Backer MD, Magee PT, Pla J. 2000. Recent developments in molecular genetics of *Candida albican*s. *Ann Rev Microbiol* **54**: 463–498.

De Bernardis F, Muhlschlegel FA, Cassone A, Fonzi WA. 1998. The pH of the host niche controls gene expression in and virulence of *Candida albicans*. *Infect Immun* **66**: 3317–3325.

Drawid A, Gerstein M. 2000. A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J Mol Biol* **301**: 1059–1075.

Drawid A, Jansen R, Gerstein M. 2000. Genome-wide analysis relating expression level with protein subcellular localization. *Trends Genet* **16**: 426–430.

Cleves AE, Cooper DN, Barondes SH, Kelly RB. 1996. A new pathway for protein export in *Saccharomyces cerevisiae*. *J Cell Biol* **133**: 1017–1026.

Eisenhaber B, Bork P, Eisenhaber F. 1999. Prediction of potential GPI-modification sites in proprotein sequences. *J Mol Biol* **292**: 741–758.

Eisenhaber B, Bork P, Eisenhaber F. 2001. Post-translational GPI lipid anchor modification of proteins in kingdoms of life: analysis of protein sequence data from complete genomes. *Protein Eng* **14**: 17–25.

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* **300**: 1005–1016.

Emori TG, Gaynes RP. 1993. An overview of nosocomial infections, including the role of the microbiology laboratory. *Clin Microbiol* **6**: 428–442.

Fonzi WA. 1999. *PHR1* and *PHR2* of *Candida albicans* encode putative glycosidases required for proper cross-linking of β-1,3- and β-1,6-glucans. *J Bacteriol* **181**: 7070–7079.

Fu Y, Ibrahim AS, Fonzi W, *et al*. 1997. Cloning and characterization of a gene (*LIP1*) which encodes a lipase from the pathogenic yeast *Candida albicans*. *Microbiology* **143**(2): 331–340.

Gale C, Finkel D, Tao N, *et al*. 1996. Cloning and expression of a gene encoding an integrin-like protein in *Candida albicans*. *Proc Natl Acad Sci USA* **93**: 357–361.

Gale CA, Bendel CM, McClellan M, *et al*. 1998. Linkage of adhesion, filamentous growth, and virulence in *Candida albicans* to a single gene, *INT1*. *Science* **279**: 1355–1358.

Ghannoum MA, Spellberg B, Saporito-Irwin SM, Fonzi WA. 1995. Reduced virulence of *Candida albicans PHR1* mutants. *Infect Immun* **63**: 4528–4530.

Ghannoum MA. 2000. Potential role of phospholipases in virulence and fungal pathogenesis. *Clin Microbiol* **13**: 122–143.

Haynes K. 2001. Virulence in *Candida* species. *Trends Microbiol* **9**: 591–596.

Hoegl L, Ollert M, Korting HC. 1996. The role of *Candida albicans* secreted aspartic proteinase in the development of candidoses. *Mol Med* **74**: 135–142.

Hoyer LL. 2001. The *ALS* gene family of *Candida albicans*. *Trends Microbiol* **9**: 176–180.

Hube B, Sanglard D, Odds FC, *et al*. 1997. Disruption of each of the secreted aspartyl proteinase genes *SAP1, SAP2* and *SAP3* of *Candida albicans* attenuates virulence. *Infect Immun* **65**: 3529–3538.

Hube B, Stehr F, Bossenz M, *et al*. 2000. Secreted lipases of *Candida albicans*: cloning, characterisation and expression analysis of a new gene family with at least ten members. *Arch Microbiol* **174**: 362–374.

Hull CM, Raisner RM, Johnson AD. 2000. Evidence for mating of the 'asexual' yeast *Candida albicans* in a mammalian host. *Science* **289**: 307–310.

Jarvis WR. 1995. Epidemiology of nosocomial fungal infections, with emphasis on *Candida* species. *Clin Infect Dis* **20**: 1526–1530.

Kamoun S, Dong S, Hamada W, *et al*. 2002. From sequence to phenotype: functional genomics of *Phytophthora*. *Can J Plant Pathol* **24**: 6–9.

Kleinbaum DG, Kupper LL, Muller KE, Nizam A (eds). 1998. *Applied Regression Analysis and other Multivariable Methods*. Duxbury: Pacific Grove; 656–686.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**: 567–580.

Kumar A, Agarwal S, Heyman JA, *et al*. 2002. Subcellular localization of the yeast proteome. *Genes Dev* **16**: 707–719.

Lee SA, Mao Y, Zhang Z, Wong B. 2001. Overexpression of a dominant-negative allele of *YPT1* inhibits growth and aspartyl

proteinase secretion in *Candida albicans*. *Microbiology* **147**: 1961–1970.

Leidich SD, Ibrahim AS, Fu Y, *et al*. 1998. Cloning and disruption of ca*PLB1*, a phospholipase B gene involved in the pathogenicity of *Candida albicans*. *J Biol Chem* **273**: 26 078–26 086.

Magee BB, Magee PT. 2000. Induction of mating in *Candida albicans* by construction of MTL**a** and MTLα strains. *Science* **289**: 310–313.

Magee PT. 1998. Analysis of the *Candida albicans* genome. In *Methods in Microbiology*, vol 26. Academic Press: New York; 395–415.

Mao Y, Kalb VF, Wong B. 1999. Overexpression of a dominant-negative allele of *SEC4* inhibits growth and protein secretion in *Candida albicans*. *J Bacteriol* **181**: 7235–7242.

Martoglio B, Dobberstein B. 1998. Signal sequences: more than just greasy peptides. *Trends Cell Biol* **10**: 410–415.

Matthews R, Wells C, Burnie JP. 1988. Characterisation and cellular localisation of the immunodominant 47 kDa antigen of *Candida albicans*. *J Med Microbiol* **27**: 227–232.

Moller S, Croning MD, Apweiler R. 2001. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* **17**: 646–653.

Monteoliva L, Matas ML, Gil C, Nombela C, Pla J. 2002. Large-scale identification of putative exported proteins in *Candida albicans* by genetic selection. *Eukaryotic Cell* **1**: 514–525.

Mukherjee PK, Seshan KR, Leidich SD, *et al*. 2001. Reintroduction of the *PLB1* gene into *Candida albicans* restores virulence *in vivo*. *Microbiology* **147**(9): 2585–2597.

Nielsen H, Brunak S, von Heijne G. 1999. Machine learning approaches to the prediction of signal peptides and other protein sorting signals. *Protein Eng* **12**: 3–9.

Nielsen H, Engelbrecht J, Brunak S, von Heijne G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* **10**: 1–6.

Nielsen H, Krogh A. 1998. Prediction of signal peptides and signal anchors by a hidden Markov model. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB 6)*. AAAI Press: Menlo Park, CA; 122–130.

Niewerth M, Korting HC. 2001. Phospholipases of *Candida albicans*. *Mycoses* **44**: 361–367.

Sanglard D, Hube B, Monod M, Odds F, Gow N. 1997. A triple deletion of the secreted aspartyl proteinase genes *SAP4, SAP5* and *SAP6* of *Candida albicans* causes attenuated virulence. *Infect Immun* **65**: 3539–3546.

Saporito-Irwin SM, Birse CE, Sypherd PS, Fonzi WA. 1995. *PHR1*, a pH-regulated gene of *Candida albicans* is required for morphogenesis. *Mol Cell Biol* **15**: 601–613.

Scherer S, Magee PT. 1990. Genetics of *Candida albicans*. *Microbiology* **54**: 226–241.

Staab JF, Bradway SD, Fidel PL, Sundstrom P. 1999. Adhesive and mammalian transglutaminase substrate properties of *Candida albicans* Hwp1. *Science* **283**: 1535–1538.

Sundstrom P. 2002. Adhesion in *Candida* spp. *Cell Microbiol* **4**: 461–469.

Tait E, Simon MC, King S, *et al*. 1997. *Candida albicans* genome project: cosmid contigs, physical mapping, and gene isolation. *Fungal Genet Biol* **21**: 308–314.

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* **64**: 515–547.

Tzung KW, Williams RM, Scherer S, *et al*. 2001. Genomic evidence for a complete sexual cycle in *Candida albicans*. *Proc Natl Acad Sci USA* **98**: 3249–3253.

von Heijne G. 1990. Protein targeting signals. *Curr Opin Cell Biol* **4**: 604–608.

White TC, Agabian N. 1995. *Candida albicans* secreted aspartyl proteinases: isoenzyme pattern is determined by cell type, and levels are determined by environmental factors. *J Bacteriol* **177**: 5215–5221.

White TC, Marr KA, Bowden RA. 1998. Clinical, cellular, and molecular factors that contribute to antifungal drug resistance. *Clin Microbiol* **11**: 382–402.

