Published online February 27, 2004

1414–1421 *Nucleic Acids Research, 2004, Vol. 32, No. 4*
DOI: 10.1093/nar/gkh286

Identifying secretomes in people, pufferfish and pigs

Eric W. Klee[1,3], Daniel F. Carlson[4], Scott C. Fahrenkrug[3,4], Stephen C. Ekker[2,3] and Lynda B. M. Ellis[1,3,*]

[1]Laboratory Medicine and Pathology, [2]Genetics, Cell Biology and Development, [3]Arnold and Mabel Beckman Center for Transposon Research, University of Minnesota, Minneapolis, MN 55455, USA and [4]Animal Science, University of Minnesota, St Paul, MN 55108, USA

Received October 29, 2003; Revised January 7, 2004; Accepted January 26, 2004

ABSTRACT

The proteins processed by the secretory pathway (secretome) are critical players in the development of multi-cellular eukaryotic organisms but have yet to be comprehensively studied at the genomic level. In this study, we use the Target P algorithm to predict human (13–20% of proteins found in individual datasets) and Fugu (14%) secretomes based on analysis of their nearly complete proteomes. We combine internal processing with prediction software to automate secreted protein identification and overcome one of the major challenges associated with EST data: identification of the minority of clones that encode N-terminally-complete proteins. We discuss the use of these methods to predict secreted proteins in EST-based consensus sequence sets, and we validate these predictions using an assay for cell-free cotranslational translocation. Analysis of TIGR Porcine Gene Index 4.0 as a test dataset resulted in the identification of 352 N-terminally-complete, putative secreted proteins. In functional agreement with our predictions, 34 of 40 (85%) of these cDNAs were verified to be cotranslationally translocated in an *in vitro* translation system. The methods developed here are specifically designed to accept partial open reading frames and improve secreted protein predictions in eukaryotic transcriptomes, and are valuable for the analysis and annotation of eukaryotic EST databases.

INTRODUCTION

Secreted proteins, including ligands and receptors, are critical to both short- and long-range intercellular signaling during the development and growth of multi-cellular organisms. Additionally, membrane proteins mediate cellular responses to a myriad of environmental cues. The development of embryos, and the differentiation of tissues important to animal production and vertebrate reproduction, respond to both intrinsic and external signals, a response likely regulated by secreted proteins. Functional understanding of these types of proteins could provide insight into diverse sets of biological processes critical to agricultural animal performance and human disease; these proteins are a high-priority target for functional annotation.

As defined by Tjalsma (1), the term 'secretome' applies to all proteins that are synthesized and processed through the secretory pathway, along with the protein secretion machinery. Many proteins are secreted by targeting the endoplasmic reticulum (ER) membrane by signal peptides, which, if Type-1, are on average 20 amino acids long in eukaryotes and are located at the amino terminus of nascent polypeptides (2). The signal peptide of the nascent polypeptide is recognized by the signal recognition particle (SRP), a cytoplasmic ribonucleoprotein consisting of six different subunits and 7SL RNA. The nascent chain–ribosome–SRP complex associates with the SRP receptor on the ER membrane and SRP is released from the complex. At the ER membrane, the nascent chain–ribosome complex associates with the protein translocation channel. This channel provides a closed and aqueous environment through which the hydrophilic nascent peptide chain can be cotranslationally translocated. Following transport to the ER lumen, the signal peptide is cleaved from the protein at a peptide bond preceded by two small neutral residues (3). The protein then undergoes folding, modification and transport from the ER to locations such as the plasma membrane, extra cellular space or organelles.

A publicly available secreted protein database would provide a source of protein targets for use in research on agricultural animal performance, embryonic development, and human disease. Three projects identifying secreted proteins in *Candida albicans*, mouse and human have recently been published. The *C.albicans* project computationally identified soluble proteins that possessed N-terminal signal sequences and lacked transmembrane domains, GPI anchor sites and mitochondrial targeting sequences, from open reading frames (ORFs) obtained from the yeast genome (4). Unfortunately, many eukaryotes genomes have not been sequenced and higher eukaryote genomes contain significant intron splicing, causing problems in identification of translation initiation sites, and creation of partial ORFs.

Grimmond *et al.* studied a subset of the mouse genome representing the portion of the secretome found in an EST database, which encodes proteins with signal sequences and

*To whom correspondence should be addressed at: Mayo Mail Code 609, 420 SE Delaware Street, Minneapolis, MN 55455, USA. Tel: +1 612 625 9122; Fax: +1 612 624 6404; Email: lynda@umn.edu

lacking transmembrane domains (i.e. candidate ligands and related molecules) (5). This study avoided complications arising from partial ORFs by using the RIKEN RPS, fully sequenced, full-length cDNAs, a type of data not often available for other vertebrates. The human secretome project identified a set of 'novel' transcripts possessing signal peptides or transmembrane domains (6). Neither study represented a comprehensive genome-wide scan for annotating the vertebrate secretome. These projects nevertheless demonstrate a broad interest in secretome databases and illustrate the need for methods that analyze ESTs and identify those encoding full-length proteins.

Prediction of secretory proteins in mammalian genomic and EST sequences has been reviewed (7). It has been shown that secreted protein prediction programs have been designed to effectively identify signal peptides in full-length protein sequences (8–10). However, these programs were not designed to analyze partial ESTs and it is expected the accuracy of predictions would deteriorate. ESTs are difficult targets for signal sequence prediction because they have a high inaccuracy rate (≈2%) and are intrinsically 3′ biased (11). Consensus sequence clusters such as NCBI UniGenes (12) and TIGR Gene Indices (13) provide increased sequence quality and length, but these sequences may still lack the correct 5′ end.

In this study, we describe our computation of human and Fugu secretomes based on public proteomes. We combine internal processing steps with public prediction software to more fully-automate secreted protein identification. We use these methods to predict secreted proteins in TIGR Porcine Gene Index. This large mammal model organism is used since we have access to porcine clones to validate our predictions using an assay for cell-free cotranslational translocation (CTT). The methods described in this study are specifically designed to accept partial ORFs and improve secreted protein predictions in eukaryotic transcriptomes.

MATERIALS AND METHODS

Sequence data

Four secreted protein sequence sets were constructed using *Homo sapiens* and *Takifugu rubripes* (Fugu) proteomes. Human sequences were obtained from the International Protein Index (IPI) database, 03/03/02 download (URL: http://www.ebi.ac.uk/IPI/); NCBI RefSeq database, 01/02/02 download (14); and NCBI GenScan database, 02/04/02 download (15). Fugu sequences were obtained from the Joint Genome Institute, assembly 2.0, 12/06/01 download (16).

15 616 tentative consensus swine sequences were obtained from TIGR Porcine Gene Index, release 4.0, 02/01/02. The ENSEMBL known human proteome dataset (17), version 15.33 (NCBI 33 assembly), downloaded 7/02/03, was used to further annotate porcine–human homologous sequences.

N-terminal subsequences (125 amino acids) of the human and Fugu protein datasets were submitted to the Center for Biological Computation's TargetP v1.01 (18). Protein sequences predicted to be secreted by TargetP, configured for non-plant analysis, with cleavage site prediction and winner-take-all selection, were assigned to their respective secreted protein sequence set. Sequences unique to each human secreted protein sequence set were estimated by comparing a cumulative sequence set (three human secreted protein sequence sets appended together) to itself using BLAST. Sequences matching only with themselves, at a threshold of $E \leqslant 1e^{-10}$ (19), were classified as unique.

Secreted protein identification

Secreted protein identification is carried out in a multi-step process involving sequence comparison to reference secreted protein sequence sets, prediction of signal peptides, identification of putative start codons and N-terminal alignments. First, the target sequence set is compared to each reference secreted protein set using NCBI BLAST v2.1.2 (20), with a selection threshold of $1e^{-10}$. All other parameters have default values. Target sequences possessing at least one homolog meeting the selection threshold are placed in a homolog sequence set; one homolog set is created for each reference set. The nucleotide sequences in the homolog sets are translated to protein sequences using BioPerl's CodonTable module (21). The frame used for this translation corresponds to that used to align with the highest scoring homolog of each reference set.

Each homologous sequence set was independently subjected to signal peptide prediction, translation start codon identification and N-terminal alignment. Signal peptide prediction was performed by TargetP 1.01 Server, using default parameters (18). Sequences predicted to contain a signal peptide in the first 125 N-terminal amino acids of each target protein sequence made up the signal peptide positive sequence set. Target sequences were also analyzed for the presence of at least one 'ATG' in the first 150 5′ base pairs, without reference to ATG context. Those sequences containing a putative start codon made up the ATG positive sequence set. N-terminal alignments of target protein sequences with their homologous reference secreted proteins were carried out using an index residue pair obtained from the BLAST output (Fig. 1). The index residue pair is the first reported sequence positions in the BLAST high-scoring pair alignment. Relative to the index residue pair, an N-terminal offset was calculated for the sequence pair. Target sequences with offsets less than or equal to a designated threshold (50 amino acids) made up the N-terminally aligned sequence set.

For each reference set, target sequences belonging to the signal peptide positive sequence set, the ATG positive sequence set and the N-terminally aligned sequence set, comprise the putative secreted protein sequence set. All putative secreted protein sequence sets were combined, and redundant sequences were removed, to create a non-redundant, putative secreted, protein set.

Test sequence selection

We selected putative secreted protein sequences containing at least one USDA Meat Animal Research Center (MARC) clone in the first 35 nucleotides of the parent TIGR consensus sequence, since these clones were available to us to use in the CTT assay. To increase the utility of the test sequences for further comparative and functional studies, they were enriched in proteins with unknown function, based on homology (BLAST threshold of $E \leqslant 1e^{-10}$) to proteins in the ENSEMBL human proteome dataset.

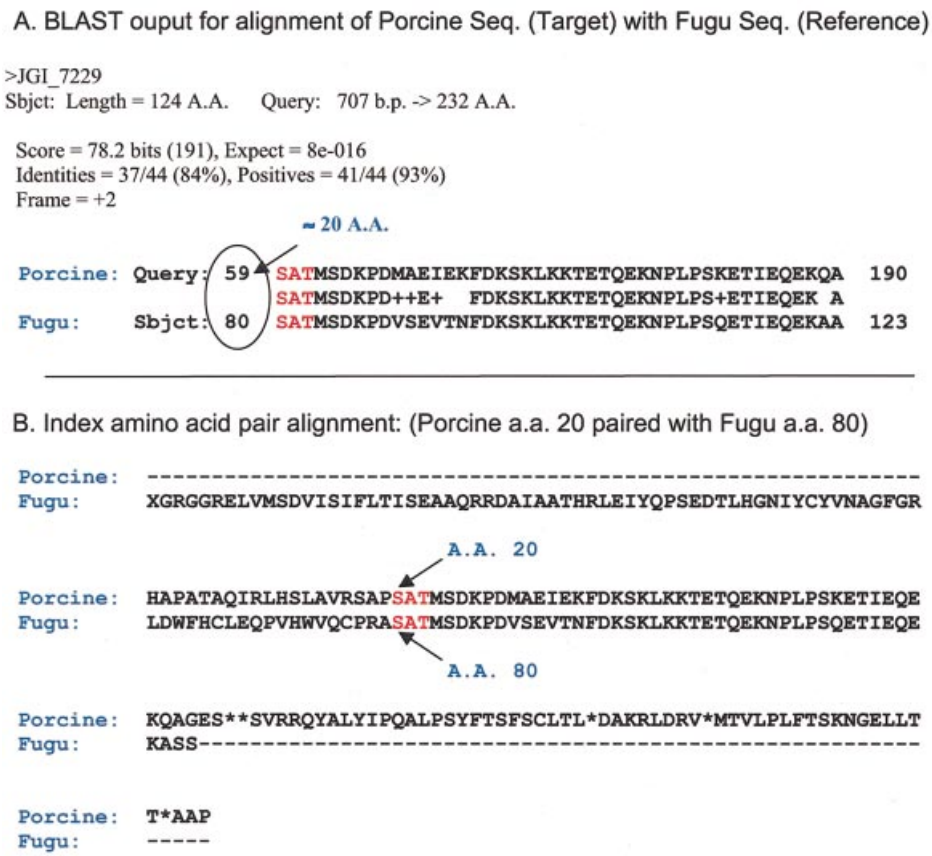

Figure 1. Construction of homologous sequence pair alignments. (**A**) Homologous proteins are aligned using an index residue pair obtained from the BLAST output. (**B**) The alignment is expanded if necessary to include the N-termini of both sequences. Relative to the index residue pair, an N-terminal offset is calculated for the sequences pair and offsets less than or equal to a threshold (50 amino acids) are selected.

ENSEMBL-based annotation

Putative secreted protein sequences were compared to the ENSEMBL known human peptide dataset version 15.33, to further annotate the sequences and estimate the number of novel sequences identified. The two datasets were compared using BLAST ($E \leqslant 1e^{-10}$). Protein annotation for their best human homolog was obtained from the ENSEMBL Description Field. Sequences were considered to have unknown function when the Description Field contained 'No Description'.

Cotranslational translocation (CTT)

Clones from MARC 1PIG and 2PIG cDNA libraries (22) that were predicted to encode secreted proteins were grown overnight at 37°C in 1× LB broth, 50 µg/ml carbenicillin. Plasmid DNA was isolated using standard alkaline lysis. To verify identity, each clone was 5′-end sequenced and compared using pairwise BLAST to its GenBank EST entry. Transcription and translation reactions were carried out using Sp6 TNT® Quick Coupled Transcription/Translation Systems (Promega, Madison, WI). Each reaction contained 20 µl of TNT® Quick Master Mix, 2.0 µl of [^{35}S]methionine (1000 Ci/mmol at 10 Ci/ml) (Amersham, Little Chalfont, UK), 0.5 µg of plasmid DNA, 0.5 µl of Promega Canine Pancreatic Microsomal Membranes (Promega) and nuclease-free water to a final volume of 25 µl. Reactions were incubated for 90 min at 30°C. After the incubation, a 2.5 µl aliquot was removed (pre-protease total) and prepared for SDS–PAGE analysis to assess whether the cDNA produced a protein product.

The remainder of the reaction was incubated for 30 min on ice at 1 mg/ml proteinase K, and then quenched with 1 µl of Complete Proteinase Inhibitor EDTA free (Roche, Indianapolis, IN) at a concentration of one tablet per 300 µl of water. After a 15 min incubation on ice in the presence of the inhibitor, the entire reaction was diluted to 150 µl at a final concentration of 110 mM KOAc, 20 mM K–Hepes, and 2 mM $Mg(OAc)_2$ [KHM]. The entire reaction mixture was then placed on a 0.5 M sucrose/KHM cushion and centrifuged at 40 000 r.p.m. (68 000 RCF) for 5 min at 4°C in a Beckman Optima TLX Ultracentrifuge in a TLA 100.3 rotor.

Three different fractions were processed for SDS–PAGE analysis from each sample; pre-protease total, pellet and supernatant. The presence of protein in the pre-protease sample indicated the cDNA was effectively transcribed and translated. Secreted proteins would be expected in the pellet, since they are protected from Proteinase K in the lumen of microsomes. Proteinase K treatment of the supernatant fraction was expected to degrade any non-secreted proteins or secreted proteins not cotranslationally translocated, possibly synthesized by the surplus of translation components in

Table 1. Number of human and Fugu total protein, and derived secreted protein, reference sequences.

Sequence set	Total proteins	Secreted proteins
H.sapiens		
IPI	54 687	8752
RefSeq	33 524	6716
GenScan	54 113	7302
T.rubripes	38 633	5310

Table 2. Analysis of TIGR porcine index version 4.0

Reference secreted sets	Homologs	ATG	Aligned	Signal peptide	ATG, aligned, signal peptide
H.sapiens					
IPI	2792	2442	1007	522	294
RefSeq	2379	2077	820	450	228
GenScan	2646	2291	888	462	247
T.rubripes	2121	1824	570	333	148
Non-redundant	3934	3422	1487	626	**352**

Number of sequences in the index that have homologs with one or more sequences in the reference secreted protein sequence sets. Of the index sequences with homologs, the number that: have an ATG; are N-terminally aligned with their homolog; have a signal peptide; and meet all three of these criteria. The number of non-redundant index sequences for each column is given, with the predicted porcine secreted protein set in bold.

the reactions. SDS sample buffer was added to all fractions to a concentration of 2% SDS, 66 mM Tris, 4 M urea, 0.01% bromophenol blue and 5% BME. Pre-protease, supernatant and pellet fractions (2, 1 and 11%) along with 6 µl of Kaleidoscope Protein Standards (Bio-Rad), were resolved on a 4–10% gradient Ready Gel (Bio-Rad) at 150 V for ~45 min. Gel running buffer was composed of 27 mM Tris, 190 mM glycine, 5.4 mM NaN_3 and 0.1% SDS. Gels were fixed in 90:5:5 water, isopropyl alcohol and glacial acetic acid for 10 min followed by a 1 h treatment with Autofluor (National Diagnostics, Atlanta, GA), dried, exposed to X-ray film overnight, and developed.

RESULTS

Reference secretome

Secreted protein sequence sets were constructed from full length human and Fugu protein sequence sets. These protein sets composed between 13 and 20% of the respective protein sequence sets (Table 1) and are available as Supplementary Material. Three human protein sequence sets were analyzed to maximize proteome coverage. Each set introduced unique (as defined in Materials and Methods) sequences into the human secretome. The IPI Human secreted protein sequence set contained 14.9% unique sequences, NCBI GenScan, 23.2% unique sequences, and NCBI RefSeq, 3.6% unique sequences. These were combined to create a human secretome containing 10 688 non-redundant protein sequences. 49% of the 5310 sequences in the Fugu secretome had no homolog in the human secretome.

Porcine secretome analysis

TIGR Porcine Gene Index, release 4.0 was analyzed using methods presented in Secreted Protein Identification. 3934 (25.2%) of the porcine nucleotide sequences are homologous to at least one secreted protein in the reference sequence sets, 22.6% homologous to a human protein and 13.6% homologous to a Fugu protein (Table 2). Inspection of the homologous porcine sequences shows that 3422 (87.0%) possess an ATG (putative start site) within the first 150 5′ bases. Following best-hit frame translation, 1487 (37.6%) of the porcine proteins aligned with their homolog at the N-terminus, within the designated threshold. 626 (16.2%) of the homologous protein sequences were predicted to possess N-terminal signal peptides. Collapsing across all three criteria 352 (9%) of the porcine sequences possessed a 5′ ATG, acceptably aligned near the N-terminus, and possessed a predicted signal peptide. These porcine sequences, encoding putative, N-terminally complete, secreted proteins, make up 2.3% of the total porcine gene index and are available in the Supplementary Material. While most of the 352 sequences had homologs in more than one reference secreted protein sequence set, 38 had a homolog in only one set: 21 in IPI, eight in GenScan, two in RefSeq and seven in Fugu.

Annotation using human homologs

The 352 putative secreted porcine protein sequences were further annotated by BLAST comparison to the ENSEMBL known human peptide sequence set. ENSEMBL human homologs were identified for 344 sequences (98%). Forty-six of the 344 were considered to have unknown function since the Description Field of their best ENSEMBL homolog contained 'No Description'. The remaining eight did not have an ENSEMBL human homolog. One had a Fugu reference homolog, and the remaining seven had homologs from our, more inclusive, human reference secreted protein sequence sets.

We examined ENSEMBL homologs for the 352 sequences we identified, specifically looking for annotation of TM domains and signal peptides. Eighty-six homologs contained annotated signal peptides. The ENSEMBL annotation for human homologs of 40 sequences selected for assessment by cell-free CTT is shown in Table 3. Three of the 40 sequences did not possess a human homolog, while 17 were homologous with proteins lacking description and the remaining 20 (50%) were homologous with annotated human proteins. The sequences selected for cell-free testing were enriched for unknown function, to maximize the information gained by further study.

Assessment for cell-free CTT

A cell-free test for CTT was performed on 46 of the putative secreted proteins, selected from 190 of the 352 putative secreted proteins which contained a clone in MARC1PIG and MARC2PIG libraries within 35 bases of the consensus sequence 5′ end. In order to provide insight as to the performance of our algorithm in *de novo* secreted protein identification, our selection of 46 clones for CTT analysis was skewed towards proteins of unknown function. Six sequences failed to yield significant translation product. Even under this rigorous challenge 34 of the 40 translated sequences (85%)

Table 3. Annotation of putative secreted proteins

TC 4.0	CTT results	ENSEMBL annotation of human homologs
TC32232	Secreted	Major epididymis-specific protein E4 precursor (HE4) (epididymal secreted protein e4) (wap four-disulfide core domain protein 2)
TC41141	Secreted	Apolipoprotein c-II precursor (Apo-CII)
TC34159	Secreted	Presenilin-like protein 1 (EC 3.4.99.-) (SPPL2B protein)
TC40379	Secreted	Small inducible cytokine A21 precursor (CCL21) (beta chemokine exodus-2) (6ckine) (secondary lymphoid-tissue chemokine) (SLC)
TC46174	Secreted	IG superfamily protein
TC46246	Secreted	Acrosomal protein SP-10 precursor (acrosomal vesicle protein-1)
TC34642	Secreted	Dolichol phosphate-mannose biosynthesis regulatory protein
TC39921	Secreted	Vacuolar proton-ATPase subunit
TC39343	Secreted	Cathepsin Z precursor (EC 3.4.22.-) (cathepsin X) (cathepsin P)
TC36368	Secreted	Cathepsin W precursor (EC 3.4.22.-) (lymphopain)
TC37131	Secreted	Calreticulin 3 precursor (calreticulin 2)
TC42510	Secreted	Liver-expressed antimicrobial peptide 2 precursor (LEAP-2)
TC39473	Secreted	Cathepsin H precursor (EC 3.4.22.16)
TC46432	Secreted	UDP-galnac:polypeptide N-acetylgalactosaminyltransferase T10; UDP-galnac:polypeptide N-acetylgalactosaminyltransferase T14
TC46084	Secreted	Solute carrier family 22 (organic anion/cation transporter), member 9; organic anion transporter 4
TC42760	Secreted	Glycoprotein VI (platelet); platelet glycoprotein VI
TC45618	Secreted	MCM10 minichromosome maintenance deficient 10; homolog of yeast MCM10
TC39859	Secreted	No description
TC33636	Secreted	No description
TC41349	Secreted	No description
TC43242	Secreted	No description
TC37052	Secreted	No description
TC42487	Secreted	No description
TC44870	Secreted	No description
TC40355	Secreted	No description
TC34663	Secreted	No description
TC41473	Secreted	No description
TC41691	Secreted	No description
TC32518	Secreted	No description
TC38861	Secreted	No description
TC36737	Secreted	No description
TC45770	Secreted	No homolog
TC37624	Secreted	No homolog
TC39439	Secreted	No homolog
TC46416	Not secreted	Cathepsin S precursor (EC 3.4.22.27)
TC36933	Not secreted	Roundabout homolog 4, magic roundabout
TC38793	Not secreted	Tumor necrosis factor-inducible protein TSG-6 precursor (TNF-stimulated gene 6 protein) (hyaluronate-binding protein)
TC35935	Not secreted	No description
TC32823	Not secreted	No description
TC39862	Not secreted	No description

The TC identifier for 40 porcine sequences subjected to CTT validation, CTT results and ENSEMBL description field annotation of their highest scoring human homolog, if present.

were secreted according to cell-free CTT (Table 3). Examples of positive and negative results are shown in Figure 2.

DISCUSSION

Human and Fugu secretomes

We computed human and Fugu secretomes, containing ligands and receptors, through the analysis of publicly available proteomes by a signal peptide prediction program. The resulting human secreted protein sequence sets varied from 13 to 20% of their parent protein set (Table 1). These were combined to create a total human secretome of 10 688 unique protein sequences. The Fugu secretome contained 5310 (14%) proteins. The human secretome had a 51% overlap with the Fugu secretome, less than the 61% overlap reported between mouse and human (5). A greater disparity between human and Fugu is not surprising, since these are more evolutionarily distant vertebrates and indeed bracket both ends of the vertebrate lineage. Mouse sequences were not included in our study as they were not available when we began our experimental analysis of the porcine EST clones. However, our preliminary analysis of mouse data shows very close agreement in size for the computed mouse, human and Fugu secretomes. We conjecture that the secretomes for a majority of other vertebrates are composed of secreted proteins found in the logical union of the human and Fugu.

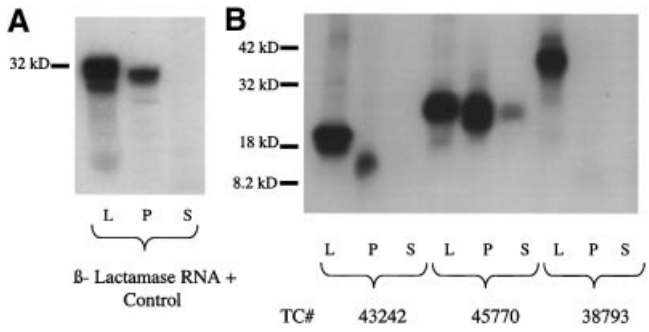

Figure 2. Cell-free assay of CTT. Samples were run on a 4–20% SDS polyacrylamide gradient gel. L, P and S represent pre-proteinase lysate, pellet and supernatant fractions, respectively. (**A**) Gel of signal peptide control RNA (β-lactamase). This demonstrates the ability of the system to differentiate secreted from non-secreted proteins. Bands are present in the pre-proteinase lysate fraction as well as in the pellet, an indication that the product was protected from cleavage by the microsomes. Further evidence that CTT has occurred is a decrease in peptide size going from the pre-protease lysate to the pellet, due to signal peptide cleavage. (**B**) Gel of cDNAs. TC43242 and TC45770 code for secreted proteins; TC38793 does not.

We selected three human protein sets for analysis in this study to maximize the coverage of the proteome since no 'gold standard' human protein set exists. When the secreted protein sequence sets derived from each human protein set were compared, hundreds of unique sequences were found in each. Since sequence comparisons are performed against each reference secreted protein sequence set independently and equivalent value is given for homology with one or more proteins in any of the secretomes, there is negligible bias associated with this redundancy between the reference sets.

Signal peptide prediction on incomplete protein sequences

Nielsen *et al.* (23) state that signal peptide prediction on EST sequences when the start codon may not be present, and on sequences with N-terminal TM domains, can result in false positive signal peptide predictions. We developed a test case for this, truncating several known non-secreted proteins, and known secreted with and without internal TM domains. Our results confirm Nielsen and coworkers' statement (Supplementary Material). *Ab initio* signal peptide prediction programs are designed to analyze full-length protein sequences and suffer reduced performance when used to analyze truncated, incomplete protein sequences, such as those derived from ESTs. Even though most proteins containing one or more TM domains also contain a SP, these misidentifications may lead to the incorrect assumption that the analyzed sequence is N-terminally complete. Since EST data are inherently fragmented, and even consensus sequences built from EST data are not necessarily full-length, SP predictions on these data are often not valid. We demonstrated the methods employed in this study are useful since they correctly predicted only those secreted proteins with N-terminally complete sequences.

Cell-free CTT

Forty-six sequences were tested for translation and translocation into pancreatic microsomes. Detectable protein was translated from 40 of the sequences; 34 of these were translocated into microsomes. This is a conservative validation method. It has a low false-positive rate, since proteins are protected from digestion only when completely translocated into microsomes. However, a receptor or other surface-anchored protein could be translocated and still digested if a majority of the protein remains outside the microsome.

Six cDNAs were predicted to be secreted, but failed CTT. Two are highly homologous to known secreted proteins. The reason these proteins were not protected from proteolysis is not known, but the observation reveals the importance of annotating protein function by multiple techniques, and suggests possible false-negative results from the cell-free CTT assay. Consequently, this assay should be seen as a method for confirmation, not rejection, of our predictions. Alternative methods would be needed to definitively determine whether those cDNAs failing cell-free CTT are in fact secreted (6,24).

Porcine secretome

Our analysis identified 352 putative secreted sequences in the porcine gene index, representing the first attempt to identify the porcine secretome. Extrapolating from the CTT results, 299 ± 6 (95% confidence interval) of the 352 sequences identified are cotranslationally translocated. This extrapolation ignores any bias incurred from selection of sequences for validation, including the requirement for MARC clones in the 5′ end of the consensus sequence assembly and enrichment with proteins of unknown function.

Analysis of the porcine data contains several examples where only one reference set contained a homolog to a putative secreted protein. Since each reference set was created based on different criteria, one may conjecture that putative secreted proteins with homologs in multiple sets are more likely to be true proteins. However, less well-characterized secreted proteins may have fewer homologs, and homologs in fewer reference protein sets.

Computationally derived protein sets such as GenScan, which do not use homology as a criterion for inclusion, may contain a higher percentage of conserved hypothetical or novel proteins. Two of the eight putative secreted porcine sequences only homologous to GenScan proteins were selected for CTT analysis. These sequences both exhibited CTT, and possessed ENSEMBL human protein homologs with unknown function. This supports the above possibility.

Of the 352 putative secreted porcine sequences, 98% possess human ENSEMBL homologs, confirming the close homology of these two species. Only 86 of these ENSEMBL homologs were annotated to contain signal peptides, fewer than expected. Our methods to identify secreted proteins thus add value to protein annotation.

Assumptions

In our analysis, a protein identified as secreted is required to have a homolog in the reference secreted protein sequence sets. Consequently, our methods do not identify putative secreted proteins lacking such homologs. This may occur because the protein sequence sets and derived reference secreted protein sets are incomplete, in part due to lack of quality 5′ annotation of eukaryotic genomes. We miss proteins that have distant homologs or proteins that are unique to the query organism. Additionally, our approach does not identify

proteins secreted by other mechanisms (25,26). Since the functional genomic studies that are being carried out on these putative secreted proteins are expensive and time-consuming, these studies should benefit from N-terminally complete sequences and the minimization of false positives.

Comparison with other secretome projects

Our project was designed to develop a method for the analysis of all proteins targeted to the secretory pathway using type I signal sequences, including ligand and receptor molecules. We used the current, publicly available proteomes from Fugu and humans as they represent two opposite extremes of the vertebrate lineage. Depending on the protein sequence set used for our analysis, we identified 13–20% of the human and ~14% of the Fugu proteins as secreted.

Although no similar comprehensive analysis using as template an entire vertebrate genome has been previously published for analysis of the secretome, two recent studies from mouse and human are noteworthy. Grimmond *et al.* (5) used a proprietary and full-length cDNA library for the identification of candidate ligands encoded by the mouse secretome. Clark *et al.* (6) used human genomic and EST public and private data for the assessment of novel members of the human secretome. Neither example included a genome-wide survey, but instead focused on the identification of unannotated members of the secretome. Indeed, these datasets overlap and nicely complement the study described herein, offering an opportunity for the expansion of the reference protein dataset secretomes using distinct methodological approaches.

Future directions

Further development of our system offers the possibility of a more refined annotation of the vertebrate secretome. Development may include implementation of improved translational start site identification, differentiation between secreted ligands and receptors, and more detailed homology selection criteria. These improvements will allow us to identify a larger percentage of a target secretome and better discriminate between proteins within this complex dataset.

The validated putative secreted proteins identified by us in this study are well suited for analysis by comparative and functional genomic techniques. For example, eight of the 34 validated porcine sequences have likely zebrafish homologs in the current EST dataset for this organism (data not shown), whose sequence information is suitable for morpholino-based 'knockdown' studies using the zebrafish embryo (27). The members of the secretome will continue to receive scientific emphasis in part because of the key roles these genes play in development and disease.

CONCLUSIONS

We have developed data freely available to the greater scientific community, including human, Fugu and porcine secretome databases. We have also developed methods for the analysis of eukaryotic EST sequences that reliably identify N-terminally-complete, secreted proteins, suitable for functional genomic studies. Our methods are useful for the analysis and annotation of ESTs, especially for organisms that do not have fully sequenced genomes.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health (RO1-GM63904), an NLM Predoctoral Training Fellowship (NLM-0704) to E.K., and the Arnold and Mabel Beckman Center for Transposon Research at the University of Minnesota.

REFERENCES

1. Tjalsma,H., Bolhuis,A., Jongbloed,J.D., Bron,S. and van Dijl,J.M. (2000) Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the secretome. *Microbiol. Mol. Biol. Rev.*, **64**, 515–547.
2. Plath,K., Mothes,W., Wilkinson,B.M., Stirling,C.J. and Rapoport,T.A. (1998) Signal sequence recognition in posttranslational protein transport across the yeast ER membrane. *Cell*, **94**, 795–807.
3. von Heijne,G. (1986) A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res.*, **14**, 4683–4690.
4. Lee,S.A., Wormsley,S., Kamoun,S., Lee,A.F., Joiner,K. and Wong,B. (2003) An analysis of the *Candida albicans* genome database for soluble secreted proteins using computer-based prediction algorithms. *Yeast*, **20**, 595–610.
5. Grimmond,S.M., Miranda,K.C., Yuan,Z., Davis,M.J., Hume,D.A., Yagi,K., Tominaga,N., Bono,H., Hayashizaki,Y., Okazaki,Y. *et al.* (2003) The mouse secretome: funcational classification of the proteins secreted into the extracellular environment. *Genome Res.*, **13**, 1350–1359.
6. Clark,H.F., Gurney,A.L., Abaya,E., Baker,K., Baldwin,D., Brush,J., Chen,J., Chow,B., Chui,C., Crowley,C. *et al.* (2003) The secreted protein discovery initiative (SPDI), a large-scale effort to identify novel human secreted and transmembrane proteins: a bioinformatics assessment. *Genome Res.*, **13**, 2265–2270.
7. Ladunga,I. (2000) Large-scale predictions of secretory proteins from mammalian genomic and EST sequences. *Curr. Opin. Biotechnol.*, **11**, 13–18.
8. Menne,K., Hermjakob,H. and Apweiler,R. (2000) A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics*, **16**, 741–742.
9. Chou,K. (2002) Prediction of protein signal sequences. *Curr. Protein Pept. Sci.*, **3**, 615–622.
10. Nair,R. and Rost,B. (2002) Sequence conserved for subcellular localization. *Protein Sci.*, **11**, 2836–2847.
11. Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B., Moreno,R.F. *et al.* (1991) Complementary DNA sequencing: expressed sequence tag and human genome project. *Science*, **252**, 1651–1656.
12. Pontius,J.U., Wagner,L. and Schuler,G.D. (2003) UniGene: a unified view of the transcriptome. *The NCBI Handbook*. National Library of Medicine, Bethesda, MD, USA.
13. Quackenbush,J., Cho,J., Lee,D., Liang,F., Holt,I., Karamycheva,S., Parvizi,B., Pertea,G., Sultana,R. and White,J. (2001) The TIGR Gene Indices: Analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.*, **29**, 159–164.
14. Kim,D. Pruitt,K., Tatusova,T. and Maglott,D. (2003) NCBI Reference Sequence Project: update and current status. *Nucleic Acids Res.*, **31**, 34–37.
15. Yeh,R.-F., Lim,L.P. and Burge,C.B. (2001) Computational inference of homologous gene structures in the human genome. *Genome Res.*, **11**, 803–816.
16. Aparicio,S., Chapman,J., Stupka,E., Putnam,N., Chia,J.M., Dehal,P., Christoffels,A., Rash,S., Hoon,S., Smit,A. *et al.* (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes. Science*, **297**, 1301–1310.
17. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.

18. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
19. Klee,E.W., Ekker,S.C. and Ellis,L.B.M. (2001) Target selection for *Danio rerio* functional genomics. *Genesis*, **30**, 123–125.
20. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
21. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
22. Nielsen,H., Brunak,S. and von Heijne,G. (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.*, **12**, 3–9.
23. Fahrenkrug,S.C., Smith,T.P., Freking,B.A., Cho,J., White,J., Vallet,J., Wise,T., Rohrer,G., Pertea,G., Sultana,R. *et al.* (2002) Porcine gene discovery by normalized cDNA-library sequencing and EST cluster assembly. *Mamm. Genome*, **13**, 475–478.
24. Moffatt,P., Salois,P., Gaumond,M.H., St-Amant,N., Godin,E. and Lanctot,C. (2002) Engineered viruses to select genes encoding secreted and membrane-bound proteins in mammalian cells. *Nucleic Acids Res.*, **30**, 4285–4294.
25. Hughes,R.C. (1999) Secretion of galectin family of mammalian carbohydrate-binding proteins. *Biochim. Biophys. Acta*, **1473**, 172–185.
26. Shurety,W., Merino-Trigo,A., Brown,D., Hume,D.A. and Stow,J.L. (2000) Localization and post-Golgi trafficking of tumor necrosis factor-α in macrophages. *J. Interferon Cytokine Res.*, **20**, 427–438.
27. Nasevicius,A. and Ekker,S. (2000) Effective targeted gene 'knockdown' in zebrafish. *Nature Genet.*, **26**, 216–220.