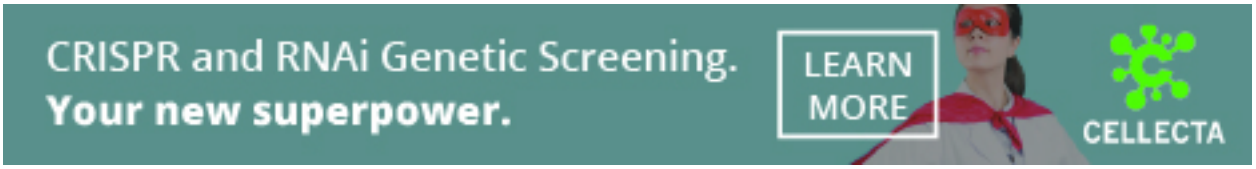
The Secreted Protein Discovery Initiative (SPDI), a Large-Scale Effort to Identify Novel Human Secreted and Transmembrane Proteins: A Bioinformatics Assessment

Hilary F. Clark, Austin L. Gurney, Evangeline Abaya, et al.

Genome Res. 2003 13: 2265-2270
Access the most recent version at doi:10.1101/gr.1293003



Letter

The Secreted Protein Discovery Initiative (SPDI), a Large-Scale Effort to Identify Novel Human Secreted and Transmembrane Proteins: A Bioinformatics Assessment

Hilary F. Clark,[1] Austin L. Gurney, Evangeline Abaya, Kevin Baker, Daryl Baldwin, Jennifer Brush, Jian Chen, Bernard Chow, Clarissa Chui, Craig Crowley, Bridget Currell, Bethanne Deuel, Patrick Dowd, Dan Eaton, Jessica Foster, Christopher Grimaldi, Qimin Gu, Philip E. Hass, Sherry Heldens, Arthur Huang, Hok Seon Kim, Laura Klimowski, Yisheng Jin, Stephanie Johnson, James Lee, Lhney Lewis, Dongzhou Liao, Melanie Mark, Edward Robbie, Celina Sanchez, Jill Schoenfeld, Somasekar Seshagiri, Laura Simmons, Jennifer Singh, Victoria Smith, Jeremy Stinson, Alicia Vagts, Richard Vandlen, Colin Watanabe, David Wieand, Kathryn Woods, Ming-Hong Xie, Daniel Yansura, Sothy Yi, Guoying Yu, Jean Yuan, Min Zhang, Zemin Zhang, Audrey Goddard, William I. Wood, and Paul Godowski

Departments of Bioinformatics, Molecular Biology and Protein Chemistry, Genentech, Inc., South San Francisco, California 94080, USA

A large-scale effort, termed the Secreted Protein Discovery Initiative (SPDI), was undertaken to identify novel secreted and transmembrane proteins. In the first of several approaches, a biological signal sequence trap in yeast cells was utilized to identify cDNA clones encoding putative secreted proteins. A second strategy utilized various algorithms that recognize features such as the hydrophobic properties of signal sequences to identify putative proteins encoded by expressed sequence tags (ESTs) from human cDNA libraries. A third approach surveyed ESTs for protein sequence similarity to a set of known receptors and their ligands with the BLAST algorithm. Finally, both signal-sequence prediction algorithms and BLAST were used to identify single exons of potential genes from within human genomic sequence. The isolation of full-length cDNA clones for each of these candidate genes resulted in the identification of >1000 novel proteins. A total of 256 of these cDNAs are still novel, including variants and novel genes, per the most recent GenBank release version. The success of this large-scale effort was assessed by a bioinformatics analysis of the proteins through predictions of protein domains, subcellular localizations, and possible functional roles. The SPDI collection should facilitate efforts to better understand intercellular communication, may lead to new understandings of human diseases, and provides potential opportunities for the development of therapeutics.

[Supplemental material is available online at www.genome.org and at http://share.gene.com. The cDNA clone sequences from this study have been submitted to GenBank under accession nos. AY358081–AY359127. The following individuals kindly provided reagents, samples, or unpublished information as indicated in the paper: T. Wu.]

Discovery of novel human proteins provides new opportunities for development of drug therapies for treatment of the wide range of diseases for which there is still no cure. In other cases, these proteins play an integral role in a disease state or the biological pathway leading to disease, and their identification and characterization may lead to an understanding of disease paradigms. Secreted and transmembrane proteins, in particular, have properties that lend themselves to be utilized as therapeutic agents or targets. They are accessible to various drug delivery mechanisms, because they are presented on the cell surface or within the extracellular space. A purified secreted protein or a receptor extracellular domain can be utilized directly as a therapeutic (e.g., growth hormone), or may be targeted by specific antibodies or small molecules. Important therapeutics have been created that target proteins present on the cell surface in a specific cell type or disease state. Rituxan is an antibody therapeutic targeting the B lymphocyte-specific CD20 protein and is an effective therapeutic in the treatment of non-Hodgkin's lymphoma. Herceptin is an antibody therapeutic targeting the breast carcinoma-specific HER2 protein and is an effective therapeutic in the treatment of breast cancer.

A number of gene families of secreted and transmembrane

[1]**Corresponding author.**
E-MAIL hclark@gene.com; FAX (650) 225-5389.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1293003. Article published online before print in September 2003.

proteins related by homology have emerged that include members known to have key roles in important biological processes such as morphogenesis, cellular differentiation, angiogenesis, apoptosis, and modulation of the immune response, as well as disease processes such as cancer progression. These gene families include tumor necrosis factors (Flavell 2002), growth factors (Hackel et al. 1999; Cross and Claesson-Welsh 2001; Ornitz and Itoh 2001; Danielsen and Maihle 2002), cytokines (Schooltink and Rose-John 2002), chemokines (Onuffer and Horuk 2002), interferons (Grandvaux et al. 2002), and angiopoietin-related (Yancopoulos et al. 1998) protein families, as well as the protein families of their receptors and other receptors such as the Toll-like receptors (Armant and Fenton 2002), integrins (Dedhar 1999), and disintegrins (Yamamoto et al. 1999; Tang 2001).

In some cases, a protein may have therapeutic potential if it is present in a disease state, even if it does not play a role in the progression or maintenance of the disease. However, there are many factors that influence a protein's potential as an effective and safe therapeutic or therapeutic target; the presence and abundance in normal and diseased tissues, the subcellular localization, the activity, and the biological role of the protein are just some of these factors. Therefore, it is imperative to screen a large number of proteins for a wide variety of such characteristics in order to identify the most promising potential candidates for drug development. Computational methods can be useful in predicting the likelihood of some of a protein's characteristics in order to focus further laboratory investigation on the proteins with the most potential for playing a role in a disease state and leading to a therapeutic. To facilitate the discovery of new therapeutic opportunities, we undertook a large-scale program of biological and computational strategies to identify and classify new secreted and transmembrane proteins.

RESULTS

This effort to identify novel secreted and transmembrane human proteins resulted in 1047 transcripts successfully cloned, representing 1021 genes (Table 1). A complete list of the GenBank accession numbers of the cDNA clone sequences with details of the analysis summarized in this publication is available as a Supplementary Table. The success rate of the SPDI project can be measured by the proportion of these genes that appear to encode secreted or transmembrane proteins, which is 86% (879 genes). A total of 13% (136 genes) appear to encode cytoplasmic and nuclear proteins, and the subcellular localization could not be predicted for 1% (6 genes). Because our identification of these transcripts as representing novel genes, 77% (791) of them have been submitted to GenBank from sources other than this SPDI effort (Table 2). However, 25% of these cDNAs are still unique transcripts. This includes 20% (209) that are variants of genes currently represented in GenBank and 5% (47) that may represent completely novel genes.

A number of SPDI transcripts still represent novel genes at the time of the submission of this work. Evidence of these being bonafide genes includes ESTs, homology with known protein domains, and orthology with a mouse gene. Such evidence is present for most of the novel SPDI transcripts (Table 3). Transcripts with none of this evidence may also represent bonafide genes, but those with small predicted proteins may be more likely to be partial transcripts or other artifacts. Nonetheless, almost all of the SPDI transcripts that initially lacked supporting evidence have been confirmed by cDNAs identified by others over the years.

The first approach to the identification of novel secreted proteins was to exploit biological screens for the ability of cDNA library-encoded fusion proteins to direct the secretion of a reporter protein. Yeast cells provide an easily manipulated system for such screens for secreted proteins (Klein et al. 1996; Baker and

Table 1. Gene Categorization

Gene/domain category	No. of genes
7 Transmembrane domain receptors	9
Acyltransferases	4
Alcohol dehydrogenases	16
Amino oxidases	3
AMP binding proteins	3
C1q domain receptors	10
Cadherins	8
Carbonic anhydrases	3
Carboxylesterases	3
Carboxypeptidases	4
Claudins	7
Collagens	5
CUB/Sushi domain receptors	12
Cystatins	1
Cytochromes	5
Cytokines and chemokines	7
Epidermal growth factor receptors	12
Epidermal growth factors	10
Fibrinogens	11
Fibroblast growth factors	6
Fibronectins	12
Glypicans	2
Hormones	2
Integrins	2
Interferons	2
Laminins	3
Lectins	14
Leucine-rich repeat receptors	51
Lipocalins	3
Low-density lipoprotein receptors	5
Membrane channels	11
Olfactomedins	5
Peptidases	16
Phosphodiesterases	3
Proteases and protease inhibitors	14
Scavenger receptor domain proteins	2
Semaphorins and plexin repeat receptors	9
Thrombospondins and ADAMs	11
Transforming growth factors	2
Trypsins and trypsin inhibitors	17
Tumor necrosis factor receptors	10
Tumor necrosis factors	4
Tyrosine kinase receptors	1
Uteroglobins	1
WNT and WNT induced signaling proteins	5
Other immunoglobulin superfamily	64
Other secreted proteins	217
Other transmembrane proteins	252
Total number of secreted and transmembrane proteins	879
Galactosyltransferases	6
Glycosyltransferases	11
Kinases	1
Mitochondrial carrier proteins	3
Sulfatases	4
Sulfotransferases	3
Thioredoxins	7
Transcription factors	8
Other cytoplasmic and nuclear	93
Total number of cytoplasmic and nuclear proteins	136
Total number of proteins with unpredicted localization	6
Total number of SPDI genes	1021

Gene families and categories as determined by computational assessment of protein domains and sub-cellular localization, as well as curation.

Table 2. Novelty Assessment of Transcripts Identified

No. of cDNAs	Comparison to other GenBank cDNAs
791	Identity with GenBank cDNA currently in GenBank from source other than this SPDI effort
209	Variant of gene with cDNAs in GenBank
47	No other GenBank cDNAs for this gene
1047	Total number of cDNAs included in the SPDI collection

Determined just prior to this publication by BLAST algorithm against GenBank using the ORF sequence (all sequences were absent from GenBank when identified).

Gurney 2000). Large libraries of cDNA fragments inserted before the reporter gene can be screened, and positive yeast colonies secreting the fusion reporter protein can then be identified. PCR amplification of the cDNA insert from the yeast colony allows sequence identification of cDNA clones encoding functional secretion sequences.

One difficulty inherent in biological screens for secreted proteins is that they encounter diminishing yields as the more abundant novel proteins are discovered and the remaining novel proteins become more rare. Computational methods, by comparison, are in principle, well suited to the identification of rare genes, provided there is sequence information to analyze. The overall SPDI strategy was to use both biological and computational methods to identify novel secreted and transmembrane proteins from multiple sources of DNA sequence (Fig. 1). The availability of very large collections of ESTs has greatly facilitated the use of such computational strategies. Two algorithms that detect the properties of signal peptides were developed and utilized, Signal Sensor (C. Watanabe, unpubl.) and Sighmm (Zhang and Wood 2003); both measure the hydropathy of the amino terminus of DNA translations that may encode proteins. Both are effective at identifying signal peptides with robust sensitivities and specificities.

Some proteins known to be secreted or membrane bound cannot currently be identified as such computationally and/or do not possess a signal peptide. Additionally, limitations to the EST collections result in some genes not being represented with EST coverage containing amino-terminal sequence information. However, these proteins may have amino acid sequence homology to known secreted and transmembrane proteins. This homology may suggest a similar role and subcellular localization. Thus, homology-based screening strategies can be a powerful tool to identify putative secreted and transmembrane proteins. We utilized a collection of known ligands and receptors of interest as a homology-based method of identifying new members of these protein families. The protein families used in this search represent key players in cell–cell signaling, such as growth factors, cytokines, chemokines, and their receptors.

The recent availability of large-scale genomic sequence has provided new opportunities to identify rare genes not abundantly present within cDNA libraries and EST collections. The presence of introns in genomic sequence requires that a gene-prediction algorithm such as Genscan be used for gene identification. We have utilized both signal-sequence detection strategies and homology-based approaches to mine both predicted genes and genomic sequence directly for the identification of additional genes.

The SPDI effort utilized multiple gene-identification strategies that were used at different times during the course of the project, and genes already identified by one strategy were bypassed with later strategies. For this reason, it is not possible to evaluate which strategy was most effective at identifying secreted and transmembrane proteins. However, the largest number of genes were identified in this effort by computational signal sequence or homology detection from ESTs (Table 4). The smallest number of genes were detected only from genomic sequence. EST evidence was not sufficient for identification of these genes because of their rarity of expression, as EST coverage did not include a signal sequence, or because they are not highly homologous to the known ligands and receptors used to identify family members. For some genes, multiple methods were required in an iterative strategy in order to attain a full-length cDNA clone. Often, this occurred for particularly long transcripts when a 5′-truncated transcript was identified by EST mining, and then genomic sequence mining revealed the first exon of the gene. The SPDI effort exemplifies the value of utilizing various complementary approaches of gene identification.

Many of the genes identified belong to gene families related by homology, which are known to include important regulators of key physiological processes. These include secreted proteins such as cytokines, chemokines, and growth factors and their receptors. Other genes, such as those that apparently encode cytoplasmic or nuclear proteins, were also identified. In some cases, this was due to the presence of domains such as protease domains that can occur in proteins localized to either intracellular or extracellular spaces. The families of proteins that were found to have the greatest number of new members through this effort were the immunoglobulin (Ig) domain and leucine-rich repeat proteins. Combined, these two structural domains were present in >10% of the proteins identified. Another 10% of the proteins are clearly related to known classes of enzymes. A number of these proteins appear to be localized within subcompartments of secretory pathways and may have roles in regulating protein post-translational modification (e.g., glycoslation). Perhaps surprisingly, new members were identified for most of the major known families of secreted proteins. In some cases, such as in the interferon family, new members were identified despite the considerable previous efforts to identify members of the family.

DISCUSSION

The success of this effort was due to the combined use of multiple strategies for the identification of genes that encode secreted and transmembrane molecules. Each strategy has different strengths and limitations. The strategies were directed at both the source of gene evidence, such as ESTs, and both predicted gene and exon homology from genomic sequence, and at the method of detecting putative proteins with the properties of secreted and transmembrane proteins including biological screens for secretion, algorithms for detecting signal sequences, and homology searches based on a collection of known secreted and transmembrane proteins of interest.

The various methods described have differed in their success at identifying particular types of genes. For instance, novel secreted genes without a recognizable relationship to other known genes can perhaps only be identified with the biological or computational signal-sequence detection methods. Conversely, many secreted and transmembrane proteins of known gene families do not have a detectable signal sequence (e.g., basic FGF), but could be recognized by homology. The success rate of these methods was also influenced by the timing of their introduction. For example, the yeast signal trap screening was gradually discontinued as EST collections became larger and proved to be a more efficient means of gene identification. Similarly, genomic sequence mining was introduced only after EST mining had been fairly exhaustive.

Table 3. Novel Genes

Genentech UNQ ID	Name	Predicted protein (aa)	Predicted domains	Mouse ortholog LocusLink	EST evidence
9220	RRLF9220	250			
9392	VLLR9392	199			
9218	CMRF35A4	194	ig	140497	
6494	QCWQ6494	183			Yes
6490	YPLR6490	168			Yes
5809	AVLL5809	159		66925	Yes
6126	LPEQ6126	157			
9438	TIMM9	152			
6487	LMNE6487	150			Yes
751	HHSL751	143			
6167	NINP6167	142			Yes
3061	GLLV3061	139			Yes
9165	AALS9165	137			Yes
6190	AGVR6190	127			
9374	VCEW9374	127			
3057	VFLL3057	126			
430	RGTR430	125	UPAR_LY6	68311	Yes
1945	VLGN1945	125			
6493	EPWW6493	122			
9368	RTFV9368	120			
3118	GRTR3118	119			Yes
9364	FLFF9364	118			
3104	ACAH3104	115			
9353	micronovel	115			
5815	QIQN5815	114			
6228	MRSS6228	114			Yes
3029	VLCS3029	113			
3112	LVLF3112	113			Yes
6125	ARVP6125	108			
2550	SFVP2550	103			
3028	TSSP3028	102			Yes
6249	KLIA6249	102			Yes
2999	GNNC2999	100			
9419	AHPA9419	99			
5830	AILT5830	95			
2786	PIKR2786	93			Yes
3106	KVVM3106	93	Uteroglobin		Yes
6489	GWSI6489	91			Yes
5840	VGSA5840	90			Yes
5836	HSAL5836	89			Yes
1944	RVLA1944	88			Yes
6484	ISPF6484	88			
6488	AAGG6488	85			Yes
1948	AVPC1948	83			
3035	LCII3035	78			
1829	FRSS1829	73			
466	GVEI466	52			Yes

Characteristics of genes represented in GenBank solely by cDNAs from this SPDI program as of July 16, 2003. EST evidence and domain prediction are described in Materials and Methods, but only PFAM domains are noted here. Mouse orthologs have been identified for some of these genes, but that assessment is not comprehensive.

The novelty of these proteins was a key factor in the criteria for cloning them. Candidates that had identity to cDNA clone sequences in GenBank were not pursued. Therefore, the genes identified do not represent a complete collection of secreted and transmembrane proteins. Large-scale efforts by others have also identified comprehensive collections of cDNA clones for human (Strausberg et al. 1999) and mouse (Kawai et al. 2001) genes.

Many proteins in the SPDI collection have already been shown to have functions in important biological processes through investigations with the cDNA clones identified here. Of particular interest have been newly identified growth factors, cytokines, tumor necrosis factors, and Toll family receptors. Angiogenic mitogens stimulate growth of vascular endothelial cells, which is critical to the development of vascular supply. EG-VEGF induces proliferation, migration, and fenestration in capillary endothelial cells derived from endocrine glands (LeCouter et al. 2001).

Cytokines and their receptors transmit signals that modulate the immune response. The IL-17B and IL-17C cytokines induce the release of the TNF-α and IL-1β cytokines from monocytic immune cells (Li et al. 2000). The IL17E cytokine binds the cytokine receptor IL17Rh1 and induces activation of the NF-kB-signaling pathway and release of the proinflammatory cytokine IL-8 (Lee et al. 2001). The IL-22 cytokine mediates the JAK-STAT signaling pathway on binding its receptors CRF2-4 and IL-22R (Xie et al. 2000). GLM-R is a type-1 cytokine receptor that signals cellular proliferation in the immune system (Ghilardi et al. 2002). The cytokine receptor TCCR is critical for the generation of the adaptive immune response mediated by T-helper cells of the Th1 subset (Chen et al. 2000).

Tumor necrosis factors and their receptors are involved in a number of physiological and pathological responses. DR5 induces apoptosis in tumor cells after binding Apo2L/TRAIL, and DcR1 and DcR2 act as decoy receptors that inhibit this signaling (Marsters et al. 1997; Sheridan et al. 1997). Apo3L induces apoptosis and activatation of the NF-kB signaling pathway after binding Apo3/DR3 (Marsters et al. 1998). DcR3 is a soluble decoy receptor that inhibits apoptosis on binding Fas ligand (Pitti et al. 1998). GITRL activates the NF-kB signaling pathway on binding to its receptor GITR and protects T-lymphocytes against apoptosis-induced cell death (Gurney et al. 1999). XEDAR induces the NF-kB signaling pathway on binding its ligand, an isoform of ectodysplasin (Yan et al. 2000).

The innate immune system uses Toll family receptors to signal for the presence of microbes and initiate host defense. Bacterial lipoproteins are potent activators of Toll-like receptor-2, mediating both apoptosis and NF-kB signaling through myeloid differentation factor 88 (Yang et al. 1998; Aliprantis et al. 2000). Flagellin, the structural component of bacterial flagella, is detected by Toll-like receptor-5, which mediates inflammatory responses to *Salmonella* (Gewirtz et al. 2001).

A number of the SPDI proteins have been implicated in a wide range of other biological processes, and further investigation of others is underway (Pennica et al. 1998; Xie et al. 1999; Holcomb et al. 2000). Currently, microarray expression data from a wide array of normal and diseased cells and tissues is providing valuable information about possible utility for many of these genes in the treatment of cancer and immune diseases. The ongoing challenge for researchers is to continue to develop and explore new genomic scale approaches to best utilize the rich trove of sequence information that has been made possible by this and other efforts to discover and define the genes encoded within the human genome.

METHODS

Biological Screens in Yeast Cells for Detection of Secretion Sequences

Recombinant gene libraries were constructed by replacing the signal peptide encoded by the reporter gene with a library of

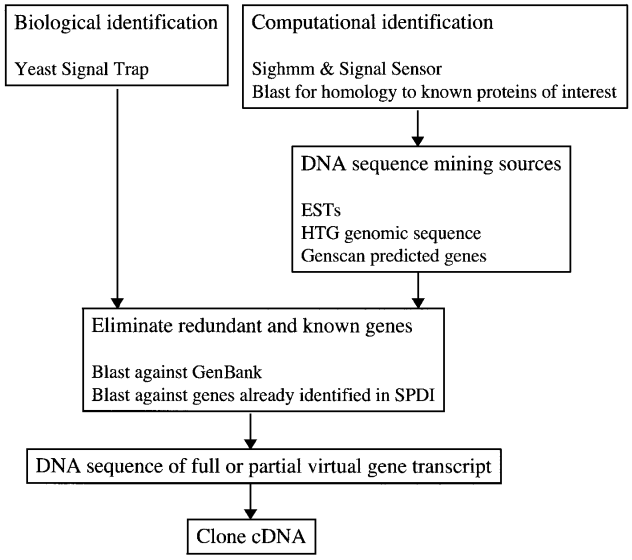


Figure 1 SPDI flow diagram.

cDNA fragments. If a given cDNA fragment encodes a signal peptide, the fusion protein may be secreted by a clonal colony of yeast cells, enabling identification of functional signal sequences. Several reporter genes were utilized in these studies, including invertase, amylase (Klein et al. 1996; Baker and Gurney 2000), and the yeast enzyme BAR1 (V. Smith, unpubl.). In brief, yeast colonies were identified that were positive for secretion of the reporter protein. The cDNA fragment contained within the fusion gene was then isolated by PCR and sequenced. Those cDNA fragments that appeared to encode ORFs containing signal sequence motifs were then further characterized by isolation of corresponding full-length cDNA clones.

Sequence Data Sources for Computational Screens

ESTs from both public (Lennon et al. 1996) and private (Incyte Pharmaceuticals) collections were utilized. ESTs were mined from consensus sequences of EST clusters, as well as individually. Genomic sequence was also mined, both directly from the human high-throughput sequence (HTG) available in the GenBank database, and through an internally compiled database of proteins predicted by the Genscan algorithm (Burge and Karlin 1997) from the HTG sequence. cDNA libraries were screened to isolate corresponding full-length cDNA clones. In some cases, the cDNA clones represented by an EST identified in these screens were purchased from private (Incyte Pharmaceutical) and public (Williamson 1999) sources.

Computational Screen for Signal Peptides

The Signal Sensor (C. Watanabe, unpubl.) and Sighmm (Zhang and Wood 2003) algorithms were used to detect signal peptides at the amino terminus of a putative protein sequence in an EST or Genscan-predicted protein.

Computational Screen for Homology to Proteins of Interest

A collection of known secreted and transmembrane proteins from gene families that include members with demonstrated biological function of particular interest (see Introduction) was used as query sequences with the BLAST algorithm against EST, genomic, and predicted protein databases. The Pfam database was queried using hmmpfam (Eddy 1998; Sonnhammer et al. 1998; Bateman et al. 2002) with each of these known proteins to determine protein domains that are represented in these protein families.

Novelty Assessment of Identified Transcripts

An automated computer algorithm was written to assess the novelty of each sequence, using only the ORF queried by the BLAST algorithm against GenBank. Identity was defined as at least 96% identity over the length of the ORF minus no greater than 36 bp (6 amino acids). A variant is defined as the top BLAST hit to a GenBank entry without identity, but a match of at least 100 bp with at least 96% identity. This identifies both splice variants and truncated versions of the same protein. Sequences that had no identity or variant in GenBank were further queried against GenBank ESTs by BLAST. EST evidence was determined by a match of at least 80 bp with at least 96% identity.

Prediction of Protein Domains and Subcellular Localization

An automated computational strategy was utilized to query each protein translation with the Signal Sensor, Sighmm, Tmdetect (T. Wu, unpubl.), hmmpfam (Eddy 1998), and Protcomp (Softberry, Inc.) algorithms. As described earlier, the Signal Sensor and Sighmm algorithms predict a secretion signal sequence. The Tmdetect algorithm predicts a transmembrane domain. The hmmpfam algorithm queries the Pfam database of protein domains to predict function domains of proteins that are related by sequence homology. The Protcomp algorithm predicts the subcellular localization of a protein, on the basis of homology to well-annotated proteins, a neural net, and various protein motifs.

Assignment of Transcipts to Gene Categories

The Gene Categories in Table 1 were determined by an iterative analysis of protein domain and protein localization predictions from various algorithms to assess the likely subcellular localization and putative functional role of each protein. In general, the genes have been divided into "Secreted and Transmembrane Proteins" and "Cytoplasmic and Nuclear Proteins", for the purpose of evaluating the success of this effort to identify secreted and transmembrane proteins. Further subcategories are delineated, in which categories of particular interest or preponderance became apparent. The categorization of proteins was determined first by domains. Some curation was done to evaluate the hmmpfam scores and determine whether the Pfam domain seemed valid, as well as consideration of the signal sequence and transmembrane domain predictions in that context. A number of genes had no signal sequence, transmembrane, or Pfam domains predicted. In this case, the Protcomp subcellular localization prediction was used to categorize these genes as "Other Secreted", "Other Transmembrane", or "Other Cytoplasmic or Nuclear".

ACKNOWLEDGMENTS

We thank Thomas Wu for the Tmdetect algorithm and David Carpenter for fruitful analysis and discussions.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby

Table 4. Assessment of the Contribution by Multiple Identification Methods

Primary identification method	Genes identified
Yeast signal sequence trap	117
Computational signal sequence detection from ESTs	335
Computational signal sequence detection from genomic sequence	27
Computational homology detection from ESTs	521
Computational homology detection from genomic sequence	25
Other	21

Determined by primary identification method.

marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Aliprantis, A.O., Yang, R.B., Weiss, D.S., Godowski, P., and Zychlinsky, A. 2000. The apoptotic signaling pathway activated by Toll-like receptor-2. *EMBO J.* **19:** 3325–3336.

Armant, M.A. and Fenton, M.J. 2002. Toll-like receptors: A family of pattern-recognition receptors in mammals. *Genome Biol.* **3:** 3011–3016.

Baker, K. and Gurney, A.L. 2000. Method of selection for genes encoding secreted and transmembrane proteins. In *US Patent 6,060,249*.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. *Nucleic Acids Res.* **30:** 276–280.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Chen, Q., Ghilardi, N., Wang, H., Baker, T., Xie, M.H., Gurney, A., Grewal, I.S., and de Sauvage, F.J. 2000. Development of Th1-type immune responses requires the type I cytokine receptor TCCR. *Nature* **407:** 916–920.

Cross, M.J. and Claesson-Welsh, L. 2001. FGF and VEGF function in angiogenesis: Signalling pathways, biological responses and therapeutic inhibition. *Trends Pharmacol. Sci.* **22:** 201–207.

Danielsen, A.J. and Maihle, N.J. 2002. The EGF/ErbB receptor family and apoptosis. *Growth Fact.* **20:** 1–15.

Dedhar, S. 1999. Integrins and signal transduction. *Curr. Opin. Hematol.* **6:** 37–43.

Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* **14:** 755–763.

Flavell, R.A. 2002. The relationship of inflammation and initiation of autoimmune disease: Role of TNF super family members. *Curr. Top. Microbiol. Immunol.* **266:** 1–10.

Gewirtz, A.T., Navas, T.A., Lyons, S., Godowski, P.J., and Madara, J.L. 2001. Cutting edge: Bacterial flagellin activates basolaterally expressed TLR5 to induce epithelial proinflammatory gene expression. *J. Immunol.* **167:** 1882–1885.

Ghilardi, N., Li, J., Hongo, J.A., Yi, S., Gurney, A., and de Sauvage, F.J. 2002. A novel type I cytokine receptor is expressed on monocytes, signals proliferation, and activates STAT-3 and STAT-5. *J. Biol. Chem.* **277:** 16831–16836.

Grandvaux, N., tenOever, B.R., Servant, M.J., and Hiscott, J. 2002. The interferon antiviral response: From viral invasion to evasion. *Curr. Opin. Infect. Diseases* **15:** 259–267.

Gurney, A.L., Marsters, S.A., Huang, A., Pitti, R.M., Mark, M., Baldwin, D.T., Gray, A.M., Dowd, P., Brush, J., Heldens, S., et al. 1999. Identification of a new member of the tumor necrosis factor family and its receptor, a human ortholog of mouse GITR. *Curr. Biol.* **9:** 215–218.

Hackel, P.O., Zwick, E., Prenzel, N., and Ullrich, A. 1999. Epidermal growth factor receptors: Critical mediators of multiple receptor pathways. *Curr. Opin. Cell Biol.* **11:** 184–189.

Holcomb, I.N., Kabakoff, R.C., Chan, B., Baker, T.W., Gurney, A., Henzel, W., Nelson, C., Lowman, H.B., Wright, B.D., Skelton, N.J., et al. 2000. FIZZ1, a novel cysteine-rich secreted protein associated with pulmonary inflammation, defines a new gene family. *EMBO J.* **19:** 4046–4055.

Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., et al. 2001. Functional annotation of a full-length mouse cDNA collection. *Nature* **409:** 685–690.

Klein, R.D., Gu, Q., Goddard, A., and Rosenthal, A. 1996. Selection for genes encoding secreted proteins and receptors. *Proc. Natl. Acad. Sci.* **93:** 7108–7113.

LeCouter, J., Kowalski, J., Foster, J., Hass, P., Zhang, Z., Dillard-Telm, L., Frantz, G., Rangell, L., DeGuzman, L., Keller, G.A., et al. 2001. Identification of an angiogenic mitogen selective for endocrine gland endothelium. *Nature* **412:** 877–884.

Lee, J., Ho, W.-H., Maruoka, M., Corpuz, R.T., Baldwin, D.T., Foster, J.S., Goddard, A.D., Yansura, D.G., Vandlen, R.L., Wood, W.I., et al. 2001. IL-17E, a novel proinflammatory ligand for the IL-17 receptor homolog IL-17Rh1. *J. Biol. Chem.* **276:** 1660–1664.

Lennon, G., Auffray, C., Polymeropoulos, M., and Bento Soares, M. 1996. The I.M.A.G.E. consortium: An integrated molecular analysis of genomes and their expression. *Genomics* **33:** 151–152.

Li, H., Chen, J., Huang, A., Stinson, J., Heldens, S., Foster, J., Dowd, P., Gurney, A.L., and Wood,W.I. 2000. Cloning and characterization of IL-17B and IL-17C, two new members of the IL-17 cytokine family. *Proc. Natl. Acad. Sci.* **97:** 773–778.

Marsters, S.A., Sheridan, J.P., Pitti, R.M., Huang, A., Skubatch, M., Baldwin, D., Yuan, J., Gurney, A., Goddard, A.D., Godowski, P., et al. 1997. A novel receptor for Apo2L/TRAIL contains a truncated death domain. *Curr. Biol.* **7:** 1003–1006.

Marsters, S.A., Sheridan, J.P., Pitti, R.M., Brush, J., Goddard, A., and Ashkenazi, A. 1998. Identification of a ligand for the death-domain-containing receptor apo3. *Curr. Biol.* **8:** 525–528.

Onuffer, J. and Horuk, R. 2002. Chemokines, chemokine receptors and small-molecule antagonists: Recent developments. *Trends Pharmacol. Sci.* **23:** 459–467.

Ornitz, D. and Itoh, N. 2001. Fibroblast growth factors. *Genome Biol.* **2:** 3001–3009.

Pennica, D., Swanson, T.A., Welsh, J.W., Roy, M.A., Lawrence, D.A., Lee, J., Brush, J., Taneyhill, L.A., Deuel, B., Lew, M., et al. 1998. WISP genes are members of the connective tissue growth factor family that are up-regulated in wnt-1-transformed cells and aberrantly expressed in human colon tumors. *Proc. Natl. Acad. Sci.* **95:** 14717–14722.

Pitti, R.M., Marsters, S.A., Lawrence, D.A., Roy, M., Kischkel, F.C., Dowd, P., Huang, A., Donahue, C.J., Sherwood, S.W., Baldwin, D.T., et al. 1998. Genomic amplification of a decoy receptor for Fas ligand in lung and colon cancer. *Nature* **396:** 699–703.

Schooltink, H. and Rose-John, S. 2002. Cytokines as therapeutic drugs. *J. Interfer. Cyto. Res.* **22:** 505–516.

Sheridan, J.P., Marsters, S.A., Pitti, R.M., Gurney, A., Skubatch, M., Baldwin, D., Ramakrishnan, L., Gray, C.L., Baker, K., Wood, W.I., et al. 1997. Control of TRAIL-induced apoptosis by a family of signaling and decoy receptors. *Science* **277:** 818–821.

Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A., and Durbin, R. 1998. Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* **26:** 320–322.

Strausberg, R.L., Feingold, E.A., and Klausner, R.D. 1999. The mammalian gene collection. *Science* **286:** 455–457.

Tang, B.L. 2001. ADAMTS: A novel family of extracellular matrix proteases. *Intl. J. Biochem. Cell Biol.* **33:** 33–44.

Williamson, A.R. 1999. The Merck gene index project. *Drug Discovery Today* **4:** 115–122.

Xie, M.H., Holcomb, I., Deuel, B., Dowd, P., Huang, A., Vagts, A., Foster, J., Liang, J., Brush, J., Gu, Q., et al. 1999. FGF-19, a novel fibroblast growth factor with unique specificity for FGFR4. *Cytokine* **11:** 729–735.

Xie, M.H., Aggarwal, S., Ho, W.H., Foster, J., Zhang, Z., Stinson, J., Wood, W.I., Goddard, A.D., and Gurney, A.L. 2000. Interleukin (IL)-22, a novel human cytokine that signals through the interferon receptor-related proteins CRF2-4 and IL-22R. *J. Biol. Chem.* **275:** 31335–31339.

Yamamoto, S., Higuchi, Y., Yoshiyama, K., Shimizu, E., Kataoka, M., Hijiya, N., and Matsuura, K. 1999. ADAM family proteins in the immune system. *Immunol. Today* **20:** 278–284.

Yan, M., Wang, L.C., Hymowitz, S.G., Schilbach, S., Lee, J., Goddard, A., de Vos, A.M., Gao, W.Q., and Dixit, V.M. 2000. Two-amino acid molecular switch in an epithelial morphogen that regulates binding to two distinct receptors. *Science* **290:** 523–527.

Yancopoulos, G.D., Klagsbrun, M., and Folkman, J. 1998. Vasculogenesis, angiogenesis, and growth factors: Ephrins enter the fray at the border. *Cell* **93:** 661–664.

Yang, R.B., Mark, M.R., Gray, A., Huang, A., Xie, M.H., Zhang, M., Goddard, A., Wood, W.I., Gurney, A.L., and Godowski, P.J. 1998. Toll-like receptor-2 mediates lipopolysaccharide-induced cellular signalling. *Nature* **395:** 284–288.

Zhang, Z. and Wood, W.I. 2003. A profile hidden Markov model for signal peptides generated by HMMER. *Bioinformatics* **19:** 307–308.

Received February 23, 2003; accepted in revised form July 28, 2003.