*AAPS PharmSci* 2003; 5 (1) Article 1 (http://www.aapspharmsci.org).

# A Semiautomated Approach to Gene Discovery Through Expressed Sequence Tag Data Mining: Discovery of New Human Transporter Genes

*Submitted: June 6, 2002; Accepted: November 11, 2002; Published: January 14, 2003*

Shoshana Brown[1], Jean L. Chang[1, 2], Wolfgang Sadee[1, 3] and Patricia C. Babbitt[1,4]

[1]Department of Biopharmaceutical Sciences, School of Pharmacy, 513 Parnassus St., University of California, San Francisco, San Francisco, CA 94143

[2]Whitehead Institute/MIT Center for Genome Research, 320 Charles St., Cambridge, MA 02141

[3]Ohio State University Medical Center, 333 W. 10th Ave., Columbus, OH 43210-1239

[4]Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, San Francisco, CA 94143

## ABSTRACT

Identification and functional characterization of the genes in the human genome remain a major challenge. A principal source of publicly available information used for this purpose is the National Center for Biotechnology Information database of expressed sequence tags (dbEST), which contains over 4 million human ESTs. To extract the information buried in this data more effectively, we have developed a semiautomated method to mine dbEST for uncharacterized human genes. Starting with a single protein input sequence, a family of related proteins from all species is compiled. This entire family is then used to mine the human EST database for new gene candidates. Evaluation of putative new gene candidates in the context of a family of characterized proteins provides a framework for inference of the structure and function of the new genes. When applied to a test data set of 28 families within the major facilitator superfamily (MFS) of membrane transporters, our protocol found 73 previously characterized human MFS genes and 43 new MFS gene candidates. Development of this approach provided insights into the problems and pitfalls of automated data mining using public databases.

**KEYWORDS:** Major facilitator superfamily, transporters, superfamily analysis, expressed sequence tags, data mining.

**Corresponding Author:** Patricia C. Babbitt, Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, San Francisco, CA 94143. Phone: (415) 476-3784;Fax: (415) 476-9819; Email: babbitt@cgl.ucsf.edu.

## INTRODUCTION

Draft sequences of the human genome[1,2] and the genomes of many other organisms have been completed. However, before genomic information can be put to practical use, it must be converted into biologically useful information. A major challenge for the conversion process is to infer the likely molecular and biological functions of each gene in a genome.

In this report we describe a partially automated method for mining expressed sequence tag (EST) data for gene discovery and functional characterization using the major facilitator superfamily (MFS) of transporters as an example. The MFS has been extensively characterized by Saier and colleagues[3] (http://tcdb.ucsd.edu/tcdb/tcfamilybrowse.php?tcname=2.A.1#protein),[FOOTNOTE 1] and its importance has been highlighted by studies analyzing the transport capabilities of various organisms with completely sequenced genomes[4,5] (http://www.biology.ucsd.edu/~ipaulsen/transport/). These studies have shown that the MFS and another large transporter group, the adenosine triphosphate (ATP) binding cassette (ABC) superfamily, together account for about 50% of the total number of transporters in the 18 organisms studied. When our analysis was performed, the MFS had been clustered into 29 transporter families based on both functional and phylogenetic analysis.[FOOTNOTE 2] Although the proportion of transporters belonging to the ABC and MFS superfamilies varies by organism, MFS members have been shown to account for a substantial fraction of transporters. For example, MFS members have been estimated to account for 31% of all transporters in the *Escherichia coli* genome.[6] Because MFS families transport a variety of different substrates into and out of cells, including drugs, essential nutrients, neurotransmitters, peptides, and amino acids, identification of new transporters is important for understanding human response to both natural and xenobiotic molecules.

Because the number of predicted genes in the human and other genomes is large,[1,2,7] functional annotation efforts must rely on automated approaches as much as possible. Unfortunately, the inherent limitations of automated analysis strategies can obviate their usefulness, especially in providing accurate predictions of gene function. Most methods rely on automated database searching, using measures of statistical significance to determine likely relationships useful for functional classification. But this type of search bypasses safeguards provided by experimental verification of function. Even when functional classification is accurate, automated methods may fail to completely extract functional information because statistical significance thresholds for database searches must be set relatively high to avoid recruitment of false positive hits. For example, even in intensively studied organisms such as *E. coli* and *Bacillus subtilis*, determination of functions for approximately 30% to 40% of open reading frames remains incompletely resolved.[8-11]

One way to improve the reliability of functional assignment generated from automated data mining is to use all available knowledge about functionally characterized homologs in the identification and evaluation of new candidate sequences. We achieved this by using a core family of sequences from many species rather than a single sequence or motif in our approach for finding new gene candidates, thus maximizing the amount of relevant sequence information used for mining the human EST database and for inferring the function and specificity of new gene candidates.

Although the original goal for this project was identification of human transporter gene candidates that would be useful for functional genomics studies, now that the draft sequence of the human genome has been completed, we envision that these data provide the most useful dataset for that purpose. However, we note that EST data mining continues to be a useful approach both for gene discovery[12] and for other applications. For example, a recent analysis has shown substantial discrepancies between gene predictions made for the human genome by the public consortium and by Celera genomics.[7] Because ESTs represent a record of expressed genes, their use will continue to be important for the verification and completion of gene predictions.

The recent realization that simply assigning a function to every gene in a genome will not give us the complete "parts list" for that organism, as expected in the early days of the genome projects, points to other useful roles for EST data mining. Identification of splice variants, single nucleotide polymorphisms (SNPs), and posttranslational modifications will be required to infer the full set of functions represented in the proteome.[13] The importance of this additional information is supported by observations that alternative splicing plays a role in tissue-specific and temporal expression of functionally unique gene forms,[14,15] that splice variants and SNPs can be associated with specific disease conditions,[16-20] and that both may be useful in optimizing drug treatments for individual patients based on genetic makeup.[21] The large number of available ESTs (12,148,014 in National Center for Biotechnology Information [NCBI] dbEST as of August 7, 2002) that are annotated based on disease type, developmental stage, and/or tissue of origin represents an enriched information source useful for detecting the most relevant splice variants and SNPs. With slight modification, our protocol could also be used to mine EST databases for splice variants and SNPs.

## MATERIALS AND METHODS

### *Automated Protocol for Identification of New Gene Candidates*

**Figure 1** provides an overview of the automated protocol for the identification of new gene candidates. Default parameters were used for all programs unless otherwise noted. All database searches were performed using local versions of NCBI databases, except for the human genome database, which was searched via the NCBI Web site. Updated versions of the databases used in this study can be accessed at http://www.ncbi.nlm.nih.gov/BLAST/.

*Converged PSI-BLAST*—Iterative Position-Specific Iterative BLAST (PSI-BLAST)[22] analyses are performed using 1 seed sequence each from 28 of the MFS families defined by Saier et al. For each family, the optimal PSI-BLAST expectation value cutoff for the sequence characteristics of that family is determined. A core family of related sequences is collected from the final round of the PSI-BLAST analysis performed using the expectation value cutoff with the least stringent value less than $1*10^{-3}$ (within 2 exponent units) that allows PSI-BLAST to converge within 10 rounds.

*Retrieval of human EST sequences*—Each protein sequence in the core family is used as a query for a TBLASTN[23] search of the human EST database at an expectation value of $1*10^{-5}$. The resulting EST nucleotide sequences are collected.

*Removal of vector sequences*—Nonmammalian vector sequences contained in the ESTs are identified using the NCBI program VecScreen (described at http://ncbi.nlm.nih.gov/VecScreen/VecScreen.html) in conjunction with the UniVec_Core database (Kitts, R.A., Madden, T.L., Sicotte, H. and Ostell, J.A. ftp://ftp.

ncbi.nih.gov/pub/UniVec/README.uv). EST regions identified by VecScreen as strong or moderate matches to vector sequences are removed. Regions in an EST that VecScreen classifies as weak matches to vector or segments of suspect origin are removed only if the EST shows additional evidence of vector contamination (eg, additional regions of the EST that have strong or moderate matches to vector sequence).

*Removal of previously characterized ESTs*—Each EST that has passed the quality and filtering tests described above is used as a query for BLASTX and BLASTN[23] searches performed using expectation value cutoffs of $1*10^{-25}$ against the NCBI nonredundant protein and nucleotide databases, with low complexity filtering turned off. The output from both BLAST runs is then processed to remove any ESTs with greater than or equal to 95% identity to a characterized human gene or protein.

*Contig assembly*—The remaining ESTs are assembled into contigs using the program CAP3.[24] CAP3 output includes a file containing the FAST-All (FASTA) sequences for each contig, and a file containing the FASTA sequences of each singlet (ie, each EST that could not be assembled into a contig). New gene candidates represented by these remaining contigs and singlets are then evaluated for membership within a specific MFS family.

### *Nonautomated Protocol for Validation of New Gene Candidates*

*Removal of gene candidates that correspond to characterized human genes*—Because inconsistencies in annotation methodologies and presentation across the databases accessed by BLAST complicate the application of automated approaches for removal of all ESTs corresponding to characterized human genes, nonautomated analysis of new gene candidates is also required (see Discussion).

For this nonautomated analysis, each sequence identified by the automated protocol as a new gene candidate is used as a query for BLASTX and BLASTN searches against the NCBI nonredundant protein and nucleotide databases. For these searches, low complexity filtering is turned off, and an expectation value cutoff of $1*10^{-25}$ is used. BLAST results for each new gene candidate are examined, and candidates with greater than or equal to 95% identity to a characterized human gene or protein over ~95% or more of their length are eliminated.

*Removal of nonhuman gene candidates*—Because the NCBI human EST database was observed to be contaminated with some nonhuman EST sequences,[25] each new gene candidate is examined to verify that it corresponds to human sequence.

The BLASTX and BLASTN results for each remaining new gene candidate are again examined. All candidates with greater than or equal to 95% identity over ~95% of their length to nonhuman genes or proteins are eliminated.

*Creation of translated protein sequences for new gene candidates*—The nucleotide sequence of each remaining gene candidate is translated into the corresponding protein sequence using either the program FRAMESEARCH (Genetics Computer Group [GCG], Version 10.1, Accelrys Inc., San Diego, CA) or a related method.

*Alignment of gene candidates with core family*—A subset of the core family obtained during the Converged PSI-BLAST stage of the automated protocol is chosen for use in a CLUSTALW multiple alignment[26] with the contigs and singlets found using the entire core family. Generally, a group of core family members is selected for use in the alignment by first creating a histogram of the sizes of the core members and eliminating any outliers that are much larger or much smaller than the main group. The core sequences used in the alignment are then chosen by selecting the set of remaining core members with ≤50% identity to each other. However, for some very large families, a lower percentage identity threshold was used to generate the set of core sequences for use in the multiple alignment. Alternatively, for some very small families, all core family members were used in the alignment. All alterations to the 50% identity threshold were made to minimize degradation of the multiple alignment while maximizing the breadth of the alignment.

*Predictions of transmembrane topology for new gene candidates*—Transmembrane topology predictions for new gene candidates were generated using the program HMMTOP Version number 1.1.[26]

Phylogenetic tree construction—Phylogenetic trees were constructed for some of the core protein families based on their CLUSTALW multiple alignments. Distance matrices were constructed using the program DISTANCES (GCG, Version 10.1), and trees were constructed from

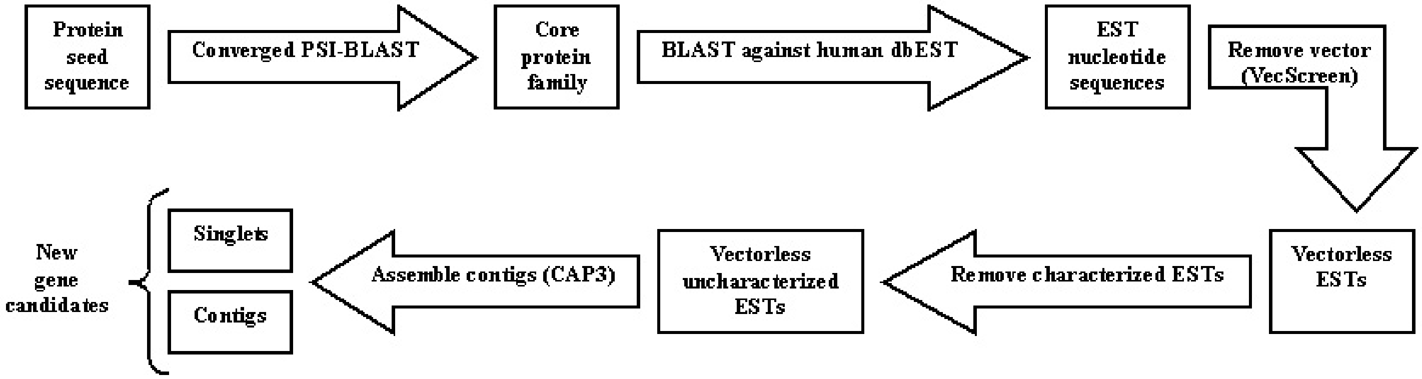

**Figure 1.** Overview of the automated protocol used to find new gene candidates. Rectangles contain descriptions of data; arrows contain descriptions of data manipulations.

these distance matrices using the GCG implementation of the neighbor-joining algorithm.[28]

*Motif analysis*— The twenty-eight core families generated from the protocol shown in **Figure 1** were evaluated using the program FINDPATTERNS, with 1 allowed mismatch (GCG, Version 10.1). The family-specific motifs used as queries for this program were those generated by Pao et al[3] for MFS families 2.A.1.1 to 2.A.1.15, and 2.A.1.17.

*Human genome analysis*—New gene candidates were matched to the most similar region of the human genome using MEGABLAST (http://www.ncbi.nlm.nih.gov/blast/),[29] with an expectation value cutoff of $1*10^{-2}$ and low complexity filtering turned off. New gene candidates that matched the same locus of the same chromosome were assumed to represent the same gene. The percentage of redundancy in new gene candidates was estimated using only those new gene candidates that could be matched to the human genome.

### Run Time

The run time of the automated elements of our protocol increased as the size of the core family increased, ranging from approximately 30 minutes to 8 hours on a 4-processor 500-MHz digital alpha server.

## RESULTS

The protocol shown in **Figure 1** was followed starting with 1 seed sequence each from 28 of the families identified by Saier et al (http://tcdb.ucsd.edu/tcdb/tcfamilybrowse.php?tcname=2.A.1#protein) in the MFS.[FOOTNOTE 3] Fifty-five contigs and 103 singlets were identified as new gene candidates, as shown in **Table 1**.[FOOTNOTE 4]

Using this initial set of new gene candidates, contigs and singlets were validated individually to eliminate any candidates corresponding to previously characterized human genes. Redundant sequences were removed. As an additional check, these candidate sequences were also evaluated for the presence of transmembrane domains by computational methods, which provided additional evidence that their structures are consistent with their identification as transporters. More detailed information about the validated contigs and singlets is given in **Tables 2** and **3**, respectively. Information about the human sequences in each core family used to find these new gene candidates is given in **Table 4**.

It should be noted that because the analysis was performed using the EST database, some singlets and contigs might actually represent different nonoverlapping regions of a single gene. Comparison of the most similar human chromosomal region for each new gene candidate (see **Tables 2** and **3**) suggests that as many as 30% of the new gene candidates could represent different regions of the same gene. When this overlap is taken into account, the number of new gene candidates corresponding to unique genes found by our method is reduced to approximately 18 contigs and 29 singlets. However, for a small number of new gene candidates, an examination of their exons suggests that they may represent different splice variants (see **Table 2**).

**Table 1.** Raw and Corrected Numbers of Contigs and Singlets Found by the Protocol, Starting With 1 Seed Sequence Each From 28 MFS Families*

| Seed Sequence GI | Seed Sequence TC No.† | No. Core Sequences‡ | No. Human Core Sequences | No. Contigs (Raw)§ | No. Contigs (Validated)‖ | No. Singlets (Raw)§ | No. Singlets (Validated)‖ |
|---|---|---|---|---|---|---|---|
| 516515 | 2.A.1.1 | 437 | 14 | 18 | 6 | 27 | 12 |
| 400227 | 2.A.1.2 | 27 | 6 | 0 | 0 | 1 | 0 |
| 1703757 | 2.A.1.3 | 190 | 0 | 1 | 0 | 5 | 4 |
| 3913718 | 2.A.1.4 | 23 | 4 | 0 | 0 | 1 | 1 |
| 125941 | 2.A.1.5 | 9 | 0 | 0 | 0 | 0 | 0 |
| 125382 | 2.A.1.6 | 62 | 0 | 1 | 1 | 6 | 6 |
| 120593 | 2.A.1.7 | 9 | 0 | 0 | 0 | 1 | 1 |
| 127835 | 2.A.1.8 | 55 | 0 | 0 | 0 | 2 | 1 |
| 1346710 | 2.A.1.9 | 41 | 0 | 0 | 0 | 0 | 0 |
| 128909 | 2.A.1.10 | 5 | 0 | 0 | 0 | 0 | 0 |
| 1235993 | 2.A.1.11 | 16 | 0 | 1 | 0 | 0 | 0 |
| 549755 | 2.A.1.12 | 3 | 0 | 0 | 0 | 0 | 0 |
| 1082597 | 2.A.1.13 | 71 | 13 | 8 | 4 | 11 | 4 |
| 2498057 | 2.A.1.14 | 127 | 9 | 6 | 2 | 11 | 4 |
| 6686326 | 2.A.1.15 | 36 | 0 | 0 | 0 | 3 | 2 |
| 731422 | 2.A.1.16 | 182 | 0 | 1 | 0 | 5 | 4 |
| 2506999 | 2.A.1.17 | 11 | 0 | 0 | 0 | 0 | 0 |
| 7673989 | 2.A.1.18 | 5 | 0 | 0 | 0 | 0 | 0 |
| 2143891 | 2.A.1.19 | 107 | 20 | 11 | 9 | 16 | 12 |
| 5777416 | 2.A.1.20 | 5 | 0 | 0 | 0 | 0 | 0 |
| 1669857 | 2.A.1.21 | 5 | 0 | 0 | 0 | 0 | 0 |
| 730883 | 2.A.1.22 | 24 | 3 | 5 | 1 | 9 | 5 |
| 6457645 | 2.A.1.23 | 4 | 0 | 0 | 0 | 0 | 0 |
| 140371 | 2.A.1.24 | 6 | 0 | 0 | 0 | 0 | 0 |
| 2114304 | 2.A.1.25 | 37 | 1 | 1 | 1 | 2 | 1 |
| 2506626 | 2.A.1.26 | 5 | 0 | 0 | 0 | 2 | 1 |
| 2495642 | 2.A.1.27 | 4 | 0 | 0 | 0 | 0 | 0 |
| 5565872 | 2.A.1.28 | 15 | 4 | 2 | 1 | 1 | 1 |
| Sum | | 1521 | 74 | 55 | 25 | 103 | 59 |
| Sum of Nonredundant Sequences | | 1153 | 73 | – | 21 | – | 41 |

*Analyses were performed between July 5, 2000, and August 8, 2000, except for 1 analysis for the seed sequence in family 2.A.1.1, which was performed on May 21 and 22, 2000. MFS indicates major facilitator superfamily; GI, NCBI Geninfo Identifier; TC, Transport Commission #

†MFS families are identified using TC number (30). Alternative nomenclature more relevant to human transporters can be found at http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl but is not used in this article because it is less complete than the compilation represented by TC numbers.

‡Core sequences are those collected during the Converged PSI-BLAST step of the protocol.

§Raw contig and singlet numbers indicate the number of contigs and singlets obtained as output from 1 run of the automated protocol, performed using the single-seed sequence indicated in the table.

‖Validated contig and singlet numbers indicate the number of contigs and singlets remaining after nonautomated validation.

**Table 2.** Contig Sequences Representing Uncharacterized Putative Transporters Found Using 1 Seed Sequence Each From MFS Families 2.A.1.1 to 2.A.1.28, and Their Corresponding Genomic Regions*

| Seed Sequence TC | Contig GIs | Contig Length (bp) | Human Genome Information | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Chrom. No. | Chrom. GI† | Map Element† | Locus ID† | Start to End (bp)‡ | %ID/match length (bp) |
| 2.A.1.1 and 2.A.1.19 | 3279578 1505780 2715430 1505781 3279899 1009986 2694995 2694486 1014306 1025498 3253866 1444660 735857 958734 | 779 | 11 | 20557821 | NT_030106 | 221106 | 63582900-63594000 | 100/623 |
| 2.A.1.1 | 1551869 2900746 3058607 3058608 3237960 3959417 3961426 4018287 4290861 4564782 4851409 4851413 4988771 5232573 5361697 5865017 6639900 6709755 7320001 | 618 | 4 | 20536850 | NT_006350 | 56606 | 18593300-18655300 | 100/584 |
| 2.A.1.1 | 1040396 1081430 1273644 | 499 | 11 | 20557821 | NT_030106 | 221103 | 63377800-63395500 | 96/445 |
| 2.A.1.1 and 2.A.1.19 | 2656651 2669854 | 367 | 11 | 20557821 | NT_030106 | 221105 | 63510600-63522800 | 100/341 |
| 2.A.1.1 | 4440080 830537 | 290 | 6 | 20549474 | NT_025741 | 154091 | 146454500-146480900 | 99/263 |
| 2.A.1.1§ and 2.A.1.19 and 2.A.1.22 | 6439198 2002320 7456480 471428 8155481 2002864 4152490 389299 | 717 | 12 | 20554464 | NT_009660 | 55530 | 107636500-107644000 | 99/722 |
| 2.A.1.1 | 3835081 3214654 7148985 5234142 5110215 | 621 | 22 | 16168698 | NT_011520 | 66035 | 20913500-20921300 | 100/618 |
| 2.A.1.6 | 7904385 7904395 7904393 | 232 | NA‖ | NA | NA | NA | NA | NA |
| 2.A.1.13 | 5110579 5368349 | 485 | 17 | 20560291 | NT_010823 | 162515 | 5541400-5542300 | 99/486 |
| 2.A.1.13 | 5340358 3872877 | 542 | 10 | 20548282 | NT_008769 | 196023 | 93193000-93201400 | 100/509 |
| 2.A.1.13 | 6569405 6989741 | 369 | 6 | 20552094 | NT_033944 | 117247 | 123511000-123511300 | 100/345 |
| 2.A.1.13 | 3231976 4299120 1485147 3085451 3070363 3147430 | 555 | 10 | 20471458 | NT_033984 | 222342 | 61927500-61972800 | 100/555 |
| 2.A.1.14 | 8065870 8065868 8065867 8065869 | 324 | NA | NA | NA | NA | NA | NA |
| 2.A.1.14 | 6443206 6443220 | 264 | NA | NA | NA | NA | NA | NA |
| 2.A.1.19 | 1298735 945608 6498416 946097 | 498 | 14 | 20542417 | NT_010164 | 57100 | 21136700-21137600 | 98/425 |
| 2.A.1.19 | 1301016 8146557 8149658 389504 5744210 991776 889283 990119 943127 946350 2001033 942961 2003991 1977296 769904 | 1251 | 14 | 20542417 | NT_010164 | 57100 | 21136500-21139300 | 97/1250 |
| 2.A.1.19 | 1569428 4373167 1776117 5747433 5744281 5747645 8142502 4622424 4389746 2589728 3648449 2841498 3751794 | 618 | 14 | 20542417 | NT_010164 | 57100 | 21136200-21136800 | 99/567 |
| 2.A.1.19 | 1779879 5430989 | 368 | 14 | 20542417 | NT_010164 | NA | 21141900-21142200 | 96/368 |
| 2.A.1.19 | 2954952 2954901 | 293 | 14 | 20542417 | NT_010164 | 57100¶ | 21136200-21136700 | 100/284 |
| 2.A.1.19 | 891343 889195 | 375 | 14 | 20542417 | NT_010164 | 57100¶ | 21136200-21136700 | 98/375 |
| 2.A.1.25 | 3092410 5767096 | 409 | 3 | 20536551 | NT_005612 | 9197 | 166365800-166366100 | 89/227 |
| 2.A.1.28 | 1957664 697762 | 399 | 14 | 20543860 | NT_026437 | 55640 | 74130500-74143400 | 100/400 |

*Since our initial analysis was performed, some of the proteins corresponding to the contigs in this table have been cloned and characterized&#151;see Table 7 for examples. MFS indicates major facilitator superfamily; TC, Transport Commission #; GI, NCBI Geninfo Identifier; chrom., Chromosome; ID, Identifier; NA, Not Applicable.

†These identifiers may be used to query LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/), the National Center for Biotechnology Information's single-query interface for information about genomic loci.

‡This region may not represent a continuous match because of the presence of introns.

§The contig obtained using a seed sequence from family 2.A.1.1 did not contain the expressed sequence tags with GI number 8155481.

‖NA represents the fact that as of August 5, 2002, no match to the human genome was found using MEGABLAST with an expectation value cutoff of $1*10^{-2}$ at the National Center for Biotechnology Information Web site: http://www.ncbi.nlm.nih.gov/genome/seq/HsBlast.html.

¶Based on examination of the exons matched, the new gene candidate may represent a splice variant.

**Table 3.** Singlet Sequences Representing Uncharacterized Putative Transporters Found Using 1 Seed Sequence Each From MFS Families 2.A.1.1 to 2.A.1.28, and Their Corresponding Genomic Regions*

| Seed Sequence TC | Singlet GI | Singlet Length (bp) | Human Genome Match Information | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Chrom. No. | Chrom. GI[†] | Map Element[†] | Locus ID[†] | Start to End (bp)[‡] | %ID/match length (bp) |
| 2.A.1.1 and 2.A.1.19 | 5339943 | 508 | 11 | 20481208 | NT_033241 | 116085 | 65255400-65256400 | 100/508 |
| 2.A.1.1 and 2.A.1.19 | 5813195 | 629 | 16 | 20562112 | NT_010542 | 146429 | 92321900-92323600 | 96/617 |
| 2.A.1.1 and 2.A.1.19 | 6835893 | 190 | NA[§] | NA | NA | NA | NA | NA |
| 2.A.1.1 and 2.A.1.19 and 2.A.1.22 | 677005 | 258 | NA | NA | NA | NA | NA | NA |
| 2.A.1.1 and 2.A.1.19 | 751568 | 370 | 11 | 20557821 | NT_030106 | 221103 | 63376200-63376900 | 99/239 |
| 2.A.1.1 and 2.A.1.19 | 1101243 | 326 | 11 | 20557821 | NT_030106 | 221106 | 63620400-63621600 | 97/311 |
| 2.A.1.1 and 2.A.1.19 | 2000637 | 307 | 1 | 20532581 | NT_019273 | NA | 111697100-111701700 | 99/292 |
| 2.A.1.1 and 2.A.1.19 | 5361744 | 401 | 11 | 20557821 | NT_030106 | 222417 | 63339500-63355200 | 98/399 |
| 2.A.1.1 and 2.A.1.3 and 2.A.1.15 and 2.A.1.16 | 5130954 | 405 | NA | NA | NA | NA | NA | NA |
| 2.A.1.1 | 5924646 | 622 | 6 | 20549474 | NT_025741 | 154091 | 146481200-146481800 | 98/622 |
| 2.A.1.1 | 1855228 | 470 | 12 | 20554950 | NT_009702 | 6515 | 7218500-7218600 | 92/150 |
| 2.A.1.1 and 2.A.1.3 and 2.A.1.14 and 2.A.1.16 | 6591316 | 512 | 16 | 20562783 | NT_033291 | 83985 | 57781200-57785200 | 100/485 |
| 2.A.1.3 and 2.A.1.16 | 8430516 | 346 | NA | NA | NA | NA | NA | NA |
| 2.A.1.3 and 2.A.1.16 | 6975459 | 610 | NA | NA | NA | NA | NA | NA |
| 2.A.1.4 | 712867 | 380 | 11 | 20560872 | NT_009151 | 2542 | 117556600-117557000 | 95/380 |
| 2.A.1.6 | 2329483 | 250 | NA | NA | NA | NA | NA | NA |
| 2.A.1.6 | 8175633 | 519 | NA | NA | NA | NA | NA | NA |
| 2.A.1.6 | 8007347 | 339 | NA | NA | NA | NA | NA | NA |
| 2.A.1.6 | 6493986 | 281 | NA | NA | NA | NA | NA | NA |
| 2.A.1.6 | 648041 | 225 | NA | NA | NA | NA | NA | NA |
| 2.A.1.6 | 8482636 | 153 | NA | NA | NA | NA | NA | NA |
| 2.A.1.7 | 2619436 | 326 | NA | NA | NA | NA | NA | NA |
| 2.A.1.8 | 7111499 | 459 | NA | NA | NA | NA | NA | NA |
| 2.A.1.13 | 8415074 | 251 | 6 | 20552094 | NT_033944 | 117247 | 123540200-123540300 | 91/147 |
| 2.A.1.13 | 8423571 | 434 | 17 | 20560291 | NT_010823 | 201232 | 544800-546100 | 99/434 |
| 2.A.1.13 | 2008184 | 292 | 1 | 20532581 | NT_019273 | NA | 116299200-116326100 | 100/292 |
| 2.A.1.13 | 1166011 | 598 | 6 | 20552094 | NT_033944 | 117247 | 123552600-123556100 | 98/598 |
| 2.A.1.14 | 2064553 | 472 | NA | NA | NA | NA | NA | NA |
| 2.A.1.14 | 5109512 | 605 | NA | NA | NA | NA | NA | NA |
| 2.A.1.14 | 3851150 | 131 | NA | NA | NA | NA | NA | NA |
| 2.A.1.15 | 6439198 | 487 | 12 | 20554464 | NT_009660 | 55530 | 107636800-107642300 | 98/375 |
| 2.A.1.19 | 1099568 | 456 | 5 | 20547229 | NT_007072 | 6583 | 138809900-138811200 | 99/369 |
| 2.A.1.19 | 1982625 | 336 | NA | NA | NA | NA | NA | NA |

**Table 3.** Continued

| Seed Sequence TC | Singlet GI | Singlet Length (bp) | Human Genome Match Information | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Chrom. No | Chrom. GI† | Map Element† | Locus ID† | Start to End (bp)‡ | %ID/match length (bp) |
| 2.A.1.19 | 3238840 | 193 | 6 | 20549694 | NT_029991 | 6580 | 172576900-172577000 | 92/79 |
| 2.A.1.19 and 2.A.1.22 | 389703 | 441 | 12 | 20554464 | NT_009660 | NA | 107661200-107663100 | 98/441 |
| 2.A.1.22 | 769942 | 181 | NA | NA | NA | NA | NA | NA |
| 2.A.1.22 | 870209 | 208 | 1 | 20534106 | NT_029226 | 9900 | 151654400-151654600 | 94/154 |
| 2.A.1.25 | 8125491 | 398 | NA | NA | NA | NA | NA | NA |
| 2.A.1.26 | 8125464 | 630 | NA | NA | NA | NA | NA | NA |
| 2.A.1.28 | 1505486 | 267 | 14 | 20543860 | NT_026437 | 55640 | 74087500-74087800 | 97/266 |

*Since our initial analysis was performed, some of the proteins corresponding to the singlets in this table have been cloned and characterized–see Table 7 for examples. MFS indicates major facilitator superfamily; TC, Transport Commission #; GI, NCBI Geninfo Identifier; chrom., Chromosome; ID, Identifier.

†These identifiers may be used to query LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/), the National Center for Biotechnology Information's single-query interface for information about genomic loci.

‡This region may not represent a continuous match because of the presence of introns.

§NA represents the fact that as of August 5, 2002, no match to the human genome was found using MEGABLAST with an expectation value cutoff of $1*10^{-2}$ at the National Center for Biotechnology Information Web site: http://www.ncbi.nlm.nih.gov/genome/seq/HsBlast.html.

## DISCUSSION

### *Overall Performance of the Protocol*

The percentage of the human MFS transporter genes our protocol was able to find can be estimated by analogy to other organisms. Paulsen et al[32] estimate that of the 19099 protein coding genes in *Caenorhabolitis elegans*, 153, or 0.8%, are MFS transporters (http://www.biology.ucsd.edu/~ipaulsen/transport/). Using this model and recent estimates of 30000 to 40000 as the total number of human genes,[1,2] 240 to 320 human MFS genes are expected. Our protocol found 73 previously characterized human MFS genes, and approximately 43 new gene candidates representing unique genes (see **Tables 2** and **3**). Extrapolating to the above estimates, our protocol found 35% to 50% of the expected human MFS genes.

However, it may not be possible to extrapolate the number of expected human MFS transporters based on the genomes of simpler eukaryotes, because the transporter composition of different organisms varies considerably.[6] An alternative approximation can be derived from estimates of the proportion of the human genome represented by dbEST at the time our analysis was performed—that is, 40% to 80% of the human genome.[33-35] Thus, if our protocol were working at 100% efficiency, it should have found ~40% to 80% of the genes in a given family, a number roughly in agreement with the estimate provided by the previous method.

It should be noted that because our protocol relies on sequence similarity for identification of related genes, genes that have diverged significantly from the sequence signature of the core family will not be detected. Further, because our protocol requires a seed sequence for each family analyzed, MFS transporter families with no members yet identified will not be represented in our results.

### *Functional Characterization of New Gene Candidates*

Because our protocol uses the information from an entire core family of diverse sequences to find new gene candidates and infer their functions, the quality and completeness of the core families is central to the success of our protocol. Therefore, we present below an evaluation of the clustering of core families and of their usefulness for functional characterization of new gene candidates.

*Validation of core families*—Before using our automatically generated core families as tools for making functional inferences about new gene candidates, we first needed to verify that new gene candidates and core family sequences were properly clustered (ie, that sufficient sequence similarities could be identified to presume that all sequences grouped together are likely to perform similar biological functions). We used 2 common strategies to evaluate sequence similarities: multiple alignments of full-length sequences and identification of conserved motifs.[3,36,37]

**Table 4.** Summary of All Human Sequences From Core Families Found Using 1 Seed Sequence Each From MFS Families 2.A.1.1 to 2.A.1.28*

| Seed Sequence TC | GI | Name |
|---|---|---|
| 2.A.1.1 | 183296 | Glucose transporter |
| 2.A.1.1 | 4375938 | dH28H20.1 (similar to membrane transport protein) |
| 2.A.1.1 | 7688146 | Glucose transporter 8 |
| 2.A.1.1 and 2.A.1.19 | 8923870 | hOAT4 organic anion transporter 4 |
| 2.A.1.1 | 7657681 | Glucose transporter X1 |
| 2.A.1.1 | 8923733 | Sugar transporter (SLC2A6 gene) glucose transporter 9 |
| 2.A.1.1 | 3387905 | Glucose transporter glycoprotein |
| 2.A.1.1 | 7446715 | Fructose transporter |
| 2.A.1.1 | 7512760 | Hypothetical protein DKFZp564K1672.1 |
| 2.A.1.1 | 4557851 | Solute carrier family 2 (facilitated glucose transporter) member 2 |
| 2.A.1.1 | 4507011 | Solute carrier family 2 (facilitated glucose transporter) member 4 |
| 2.A.1.1 | 5730051 | Solute carrier family 2 (facilitated glucose transporter) member 1 |
| 2.A.1.1 | 4507013 | Solute carrier family 2 (facilitated glucose transporter) member 5 |
| 2.A.1.1 | 5902090 | Solute carrier family 2 (facilitated glucose transporter) member 3 |
| 2.A.1.2 | 422997 | Monoamine transport protein |
| 2.A.1.2 | 1722742 | Synaptic vesicle amine transporter |
| 2.A.1.2 | 1770738 | Vesicle monoamine transporter type 2 |
| 2.A.1.2 | 4506989 | Solute carrier family 18 (vesicular monoamine) member 2 |
| 2.A.1.2 | 4506991 | Solute carrier family 18 (vesicular acetylcholine) member 3 |
| 2.A.1.2 | 4506987 | Solute carrier family 18 (vesicular monoamine) member 1 |
| 2.A.1.4 | 4406200 | Glucose-6-phosphate transporter |
| 2.A.1.4 | 6560624 | PRO0685 |
| 2.A.1.4 | 4128045 | Glucose-6-phosphate translocase |
| 2.A.1.4 | 4503847 | Glucose-6-phosphatase transport (glucose-6-phosphate) protein 1 |
| 2.A.1.13 | 4759120 | Solute carrier family 16 (monocarboxylic acid transporters) member 7 |
| 2.A.1.13 | 3834395 | Monocarboxylate transporter 2 |
| 2.A.1.13 | 4759118 | Solute carrier family 16 (monocarboxylic acid transporters) member 6 |
| 2.A.1.13 | 7019529 | Monocarboxylate transporter 3 |
| 2.A.1.13 | 7513431 | X-linked PEST-containing transporter |
| 2.A.1.13 | 8923981 | Hypothetical protein PRO0813 |
| 2.A.1.13 | 4759114 | Solute carrier family 16 (monocarboxylic acid transporters) member 4 |
| 2.A.1.13 | 4759112 | Solute carrier family 16 (monocarboxylic acid transporters) member 3 |
| 2.A.1.13 | 7328162 | Hypothetical protein |
| 2.A.1.13 | 4759116 | Solute carrier family 16 (monocarboxylic acid transporters) member 5 |
| 2.A.1.13 | 2827492 | (al009193) /prediction = (method:""genefinder"" version:""084"") |
| 2.A.1.13 | 5730045 | Solute carrier family 16 (monocarboxylic acid transporters) member 2 (putative transporter) |
| 2.A.1.13 | 4506983 | Solute carrier family 16 (monocarboxylic acid transporters) member 1 |
| 2.A.1.14 | 7513174 | Na+-dependent phosphate cotransporter |
| 2.A.1.14 | 4885441 | Na/PO4 cotransporter |
| 2.A.1.14 | 4827010 | Solute carrier family 17 (sodium phosphate) member 1 |
| 2.A.1.14 | 7328923 | Differentiation-associated Na-dependent inorganic phosphate cotransporter |
| 2.A.1.14 | 5031955 | Sodium phosphate transporter 3 |
| 2.A.1.14 | 6912666 | Solute carrier family 17 (anion/sugar transporter) member 5 |
| 2.A.1.14 | 7328925 | Brain-specific Na-dependent inorganic phosphate cotransporter |

**Table 4.** (Continued)

| Seed Sequence TC | GI | Name |
|---|---|---|
| 2.A.1.14 | 2498056 | Renal sodium-dependent phosphate transport protein 1 |
| 2.A.1.14 | 5730047 | Solute carrier family 17 (sodium phosphate) member 3 |
| 2.A.1.19 | 2337865 | Organic cation transporter; 50% similarity to JC4884 |
| 2.A.1.19 | 3581982 | Extraneuronal monoamine transporter |
| 2.A.1.19 | 4506999 | Solute carrier family 22 (organic cation transporter) member 1 |
| 2.A.1.19 | 7706146 | hBOIT for potent brain-type organic ion transporter |
| 2.A.1.19 | 2511670 | Organic cation transporter |
| 2.A.1.19 | 4193819 | Para-aminohippurate transporter |
| 2.A.1.19 | 6970446 | Multispecific organic anion transporter 4 |
| 2.A.1.19 | 4759042 | Renal organic anion transporter 1 |
| 2.A.1.19 | 3831566 | Putative renal organic anion transporter 1 |
| 2.A.1.19 | 4758846 | Organic anion transporter 3 |
| 2.A.1.19 | 4758852 | Organic cationic transporter-like 3 |
| 2.A.1.19 | 5924012 | dJ261K5.1 (novel organic cation transporter) |
| 2.A.1.19 | 5730049 | Solute carrier family 22 (organic anion transporter) member 7 |
| 2.A.1.19 | 4758854 | Organic cationic transporter-like 4 |
| 2.A.1.19 | 4507003 | Solute carrier family 22 (organic cation transporter) member 4 |
| 2.A.1.19 | 4507001 | Solute carrier family 22 (organic cation transporter) member 2 |
| 2.A.1.19 | 4579723 | hOAT1-1 organic anion transporter |
| 2.A.1.19 | 8648881 | Putative organic anion transporter |
| 2.A.1.19 | 4507005 | Solute carrier family 22 (organic cation transporter) member 5 |
| 2.A.1.22 | 7662272 | KIAA0736 gene product |
| 2.A.1.22 | 5689445 | KIAA1054 protein |
| 2.A.1.22 | 7662270 | KIAA0735 gene product; synaptic vesicle protein 2B homolog |
| 2.A.1.25 | 4757708 | Acetyl-coenzyme A transporter |
| 2.A.1.28 | 7661708 | Feline leukemia virus subgroup C receptor FLVCR |
| 2.A.1.28 | 8923350 | Hypothetical protein FLJ20371 |
| 2.A.1.28 | 5764708 | Unknown |
| 2.A.1.28 | 7022664 | Unnamed protein product |

*Names in third column are derived from the definition line of the National Center for Biotechnology Information entry for a given protein. MFS indicates major facilitator superfamily; TC, Transport Commission #; GI, NCBI Geninfo Identifier.

This analysis shows that the quality of the multiple alignments varies significantly for each core family, presumably reflecting differences in the evolutionary history within the various families, as well as the completeness of coverage represented by available sequences. **Figure 2** provides two examples of core family alignments illustrative of such variations. The multiple alignment of core family 2.A.1.8, shown in **Figure 2A**, reveals a relatively high degree of sequence conservation. Although the characterized members of this core family come from many different organisms, including bacteria, fungi, and plants, all transport either nitrate or nitrite. This high level of sequence conservation suggests that a relatively fixed sequence is required in order to retain both substrate specificity and transport functionality within this family.

**Figure 2B** shows a portion of the multiple alignment of core family 2.A.1.1 along with 3 new gene candidates. Although some elements of this alignment—for example, positions c and d—are well conserved, the overall alignment reveals substantial sequence variation. Because this core family contains a diverse group of transporters, transporting sugars, organic cations, and organic anions, it is not surprising that the multiple alignment shows somewhat greater sequence variation than that of core family 2.A.1.8.

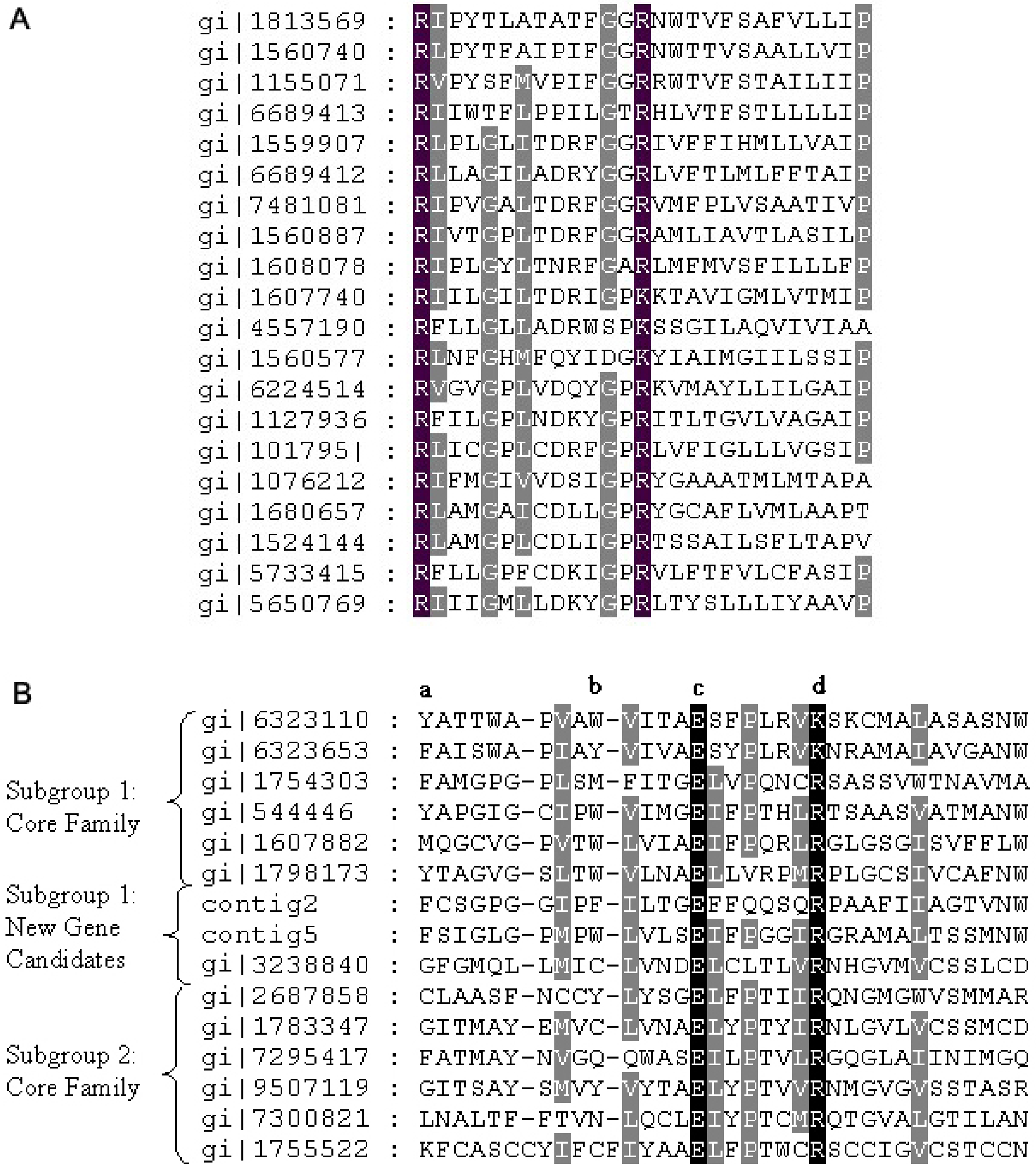
A
gi|1813569 : RIPYTLATATFGGRNWTVFSAFVLLIP
gi|1560740 : RLPYTFAIPIFGGRNWTTVSAALLVIP
gi|1155071 : RVPYSFMVPIFGGRRWTVFSTAILIIP
gi|6689413 : RIIWTFLPPILGTRHLVTFSTLLLLIP
gi|1559907 : RLPLGLITDRFGGRIVFFIHMLLVAIP
gi|6689412 : RLLAGILADRYGGRLVFTLMLFFTAIP
gi|7481081 : RIPVGALTDRFGGRVMFPLVSAATIVP
gi|1560887 : RIVTGPLTDRFGGRAMLIAVTLASILP
gi|1608078 : RIPLGYLTNRFGARLMFMVSFILLLFP
gi|1607740 : RIILGILTDRIGPKKTAVIGMLVTMIP
gi|4557190 : RFLLGLLADRWSPKSSGILAQVIVIAA
gi|1560577 : RLNFGHMFQYIDGKYIAIMGIILSSIP
gi|6224514 : RVGVGPLVDQYGPRKVMAYLLILGAIP
gi|1127936 : RFILGPINDKYGPRITLTGVLVAGAIP
gi|101795| : RLICGPLCDRFGPRLVFIGLLLVGSIP
gi|1076212 : RIFMGIVVDSIGPRYGAAATMLMTAPA
gi|1680657 : RLAMGAICDLLGPRYGCAFLVMLAAPT
gi|1524144 : RLAMGPLCDLIGPRTSSAILSFLTAPV
gi|5733415 : RFLLGPFCDKIGPRVLFTFVLCFASIP
gi|5650769 : RIIIGMLLDKYGPRLTYSLLLIYAAVP
B
a
b
c
d
Subgroup 1: Core Family
gi|6323110 : YATTWA-PVAW-VITAESFPLRVKSKCMALASASNW
gi|6323653 : FAISWA-PIAY-VIVAESYPLRVKNRAMAIAVGANW
gi|1754303 : FAMGPG-PLSM-FITGEIVPQNCRSASSVWTNAVMA
gi|544446 : YAPGIG-CIPW-VIMGEIFPTHLRTSAASVATMANW
gi|1607882 : MQGCVG-PVTW-LVIAEIFPQRLRGLGSGISVFFLW
gi|1798173 : YTAGVG-SLTW-VLNAELLVRPMRPLGCSIVCAFNW
Subgroup 1: New Gene Candidates
contig2 : FCSGPG-GIPF-ILTGEFFQQSQRPAAFIIAGTVNW
contig5 : FSIGLG-PMPW-LVLSEIFPGGIRGRAMALTSSMNW
gi|3238840 : GFGMQL-LMIC-LVNDELCLTLVRNHGVMVCSSLCD
Subgroup 2: Core Family
gi|2687858 : CLAASF-NCCY-LYSGELFPTIIRQNGMGWVSMMAR
gi|1783347 : GITMAY-EMVC-LVNAELYPTYIRNLGVLVCSSMCD
gi|7295417 : FATMAY-NVGQ-QWASEILPTVLRGQGLAIINIMGQ
gi|9507119 : GITSAY-SMVY-VYTAELYPTVVRNMGVGVSSTASR
gi|7300821 : LNALTF-FTVN-LQCLEIYPTCMRQTGVALGTILAN
gi|1755522 : KFCASCCYIFCFIYAAELFPTWCRSCCIGVCSTCCN

**Figure 2.** Alignment columns that show 100% conservation are shaded in black, and those that show ≥80% conservation are shaded in gray. Amino acids with similar physiochemical properties are treated as equivalent for the purposes of determining conservation percentages. (A) Section of the multiple alignment showing several representative core family members found using a seed sequence from family 2.A.1.8. (B) Section of a multiple alignment showing several representative core family members and new gene candidates found using a seed sequence from family 2.A.1.1. Note that positions a and b are better conserved across subgroup 1 than subgroup 2, while positions c and d are conserved throughout both subgroups. GI indicates the Geninfo Identifier, a sequence identifier used by NCBI.

The multiple alignments of several other core families also show a high degree of sequence variation, indicating perhaps that they should be further subdivided to obtain more accurate clustering of functionally distinct proteins. To examine this issue further, we divided 2.A.1.1 into 2 subgroups based on a phylogenetic tree constructed from the multiple alignment of the core family. This division of core family 2.A.1.1 into 2 subgroups reveals additional positions in the multiple alignment that are better conserved across a given subgroup than across the original core family. For example, **Figure 2B** shows that positions a and b are largely conserved across subgroup 1 but not across subgroup 2. Consistent with this tree, available information regarding the function of characterized members of the family suggests that subgroup 1 consists mainly of sugar transporters, while subgroup 2 consists mainly of organic cation and anion transporters.

After evaluating the clustering of core families, we investigated how discretely related core families could be separated. Some degree of overlap between closely related core families is expected, because the boundary of a given family can differ based on the method used to construct it.[36] However, if there is significant overlap between core families, this overlap will have to be considered when the core families are used to make functional inferences about new gene candidates.

**Figure 3** shows the instances of overlap among core family sequences found by our analysis, starting with 1 seed sequence each from MFS families 2.A.1.1 to 2.A.1.28. **Table 5** provides the corresponding fractional overlap. As expected, the majority of the overlap occurs between core families identified using seed sequences known to be closely related. For example, families 2.A.1.1, 2.A.1.19, and 2.A.1.22 are sufficiently similar that they were originally classified as a single family by Pao et al.[3]

To gain supplemental information regarding the overlap of the 28 MFS core families found in this study, we tested all members of each core family for the presence of motifs designed to be specific for individual MFS families (as defined by Pao et al[3]). **Table 6** shows the fraction of members from each core family possessing a family-specific motif, with 1 allowed mismatch. As expected, the motif overlaps shown in **Table 6** and the core sequence overlap shown in **Table 5** follow the same general trends. As shown in **Table 6**, the majority of sequences within a given core family usually possess the motif designed to be specific to the family of the seed sequence. Also shown in the table, a significant number of core family members possess motifs designed to be specific to other, usually related, families. Interestingly, even a single sequence can have more than 1 family-specific motif, showing that not all motifs are entirely specific to the family for which they were initially designed.

*Use of core families for functional classification of new gene candidates*—As mentioned above, one of the most important reasons for using a diverse group of related proteins in each of our core families is that the larger context provided by this approach is especially useful for functional characterization of new gene candidates. The specific pattern of sequence conservation shown in **Figure 2B** illustrates how the multiple alignment of new gene candidates with core family sequences can be used for functional classification of the new gene candidates. For example, a study by Kasahara and Maeda of the Gal2 galactose transporter (gi|6323110 in **Figure 2B**) and the Hxt2 glucose transporter (gi|6323653 in **Figure 2B**) from *Saccharomyces cerevisiae* showed that the presence of specific amino acids in positions a and b in **Figure 2B** strongly influences substrate specificity.[38] Tyr at position a and Trp at position b are important for galactose recognition. Phe at position a is important for glucose transport, and Tyr at this position supports glucose transport, but at highly reduced levels.

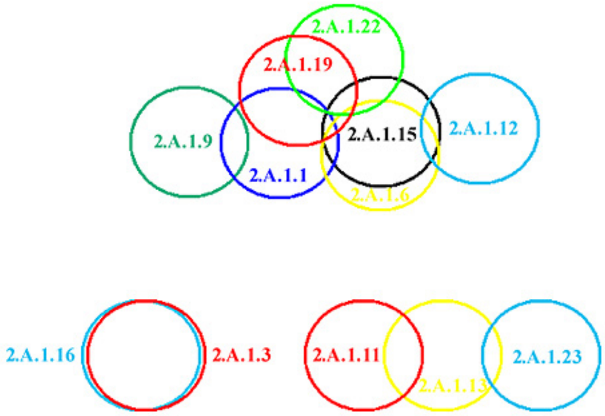

**Figure 3.** Illustration showing the approximate sequence overlap between Major Facilitator Superfamily (MFS) core families. Core families were found using a seed sequence from the MFS family indicated by Transport Commission # (TC) number. The fractional overlap of each core family is listed in Table 5.

Assignment of contigs 2 and 5 to subgroup 1 therefore suggests that they are sugar transporters. Both contigs have aromatic amino acids at positions a and b, like most of subgroup 1. Because neither of the contigs has a Tyr at position a, they probably do not function in galactose

**Table 5.** Fractional Sequence Overlap Between Core Families Obtained Using Seed Sequences From Different MFS Families*

| Core Family 1 | Core Family 2 | Common Sequence/Total Sequence |
|---|---|---|
| 2.A.1.3 | 2.A.1.16 | 182/372 |
| 2.A.1.1 | 2.A.1.19 | 98/544 |
| 2.A.1.6 | 2.A.1.15 | 13/98 |
| 2.A.1.11 | 2.A.1.13 | 11/87 |
| 2.A.1.1 | 2.A.1.9 | 32/478 |
| 2.A.1.13 | 2.A.1.23 | 4/75 |
| 2.A.1.15 | 2.A.1.19 | 6/143 |
| 2.A.1.1 | 2.A.1.15 | 19/473 |
| 2.A.1.19 | 2.A.1.22 | 5/131 |
| 2.A.1.12 | 2.A.1.15 | 1/39 |
| 2.A.1.15 | 2.A.1.22 | 1/60 |
| 2.A.1.6 | 2.A.1.12 | 1/65 |
| 2.A.1.1 | 2.A.1.22 | 6/461 |
| 2.A.1.6 | 2.A.1.22 | 1/86 |
| 2.A.1.1 | 2.A.1.6 | 3/499 |
| 2.A.1.6 | 2.A.1.19 | 1/169 |

*Core families were obtained from analysis using a single-seed sequence from the MFS family indicated by TC number. MFS indicates major facilitator superfamily; TC, Transport Commission #.

transport. Instead, the Phe at position a indicates that the contigs may function in glucose transport, or in the transport of another nongalactose sugar. Subsequent to our analysis, the protein corresponding to contig 2 was cloned and identified as a facilitated glucose transporter,[39] thus validating our functional prediction. **Table 7** provides additional examples of new gene candidates that have now been experimentally characterized by other investigators. In all cases shown in **Table 7**, the experimental data support our functional classification.[FOOTNOTE 5]

Because the singlet shown in the alignment (gi number 3238840 in **Figure 2B**) does not have aromatic amino acids at positions a or b, it is more difficult to make conclusions about transport specificity of this gene candidate. This anomalous pattern may even indicate that this singlet has been misclassified and belongs to a different transporter family or represents a single instance of a new transporter family. These questions can be addressed only after all existing transporter families have

**Table 6.** Distribution of MFS Family-Specific Motifs Within Each Core Family*

| Motif[†] | Core Family[‡] | Fraction of Core Sequence With Motif |
|---|---|---|
| 2.A.1.1 | 2.A.1.1 | 258/437 |
| | 2.A.1.3 | 13/190 |
| | 2.A.1.6 | 15/62 |
| | 2.A.1.8 | 1/55 |
| | 2.A.1.9 | 2/41 |
| | 2.A.1.11 | 1/16 |
| | 2.A.1.13 | 1/71 |
| | 2.A.1.14 | 3/127 |
| | 2.A.1.15 | 7/36 |
| | 2.A.1.16 | 13/182 |
| | 2.A.1.19 | 37/107 |
| | 2.A.1.22 | 16/24 |
| | 2.A.1.24 | 1/6 |
| 2.A.1.2 | None | NA |
| 2.A.1.3 | 2.A.1.3 | 125/190 |
| | 2.A.1.16 | 126/182 |
| 2.A.1.4 | 2.A.1.4 | 16/23 |
| 2.A.1.5 | 2.A.1.5 | 8/9 |
| 2.A.1.6 | 2.A.1.1 | 31/437 |
| | 2.A.1.6 | 46/62 |
| | 2.A.1.12 | 1/3 |
| | 2.A.1.15 | 6/36 |
| | 2.A.1.19 | 25/107 |
| 2.A.1.7 | 2.A.1.7 | 5/9 |
| 2.A.1.8 | 2.A.1.8 | 43/55 |
| 2.A.1.9 | 2.A.1.1 | 31/437 |
| | 2.A.1.9 | 32/41 |
| 2.A.1.10 | 2.A.1.10 | 4/5 |
| 2.A.1.11 | 2.A.1.11 | 5/16 |
| | 2.A.1.13 | 2/71 |
| 2.A.1.12 | 2.A.1.6 | 6/62 |
| | 2.A.1.15 | 6/36 |
| 2.A.1.13 | 2.A.1.11 | 2/16 |
| | 2.A.1.13 | 41/71 |
| 2.A.1.15 | 2.A.1.1 | 9/437 |
| | 2.A.1.6 | 1/62 |
| | 2.A.1.15 | 11/36 |
| | 2.A.1.19 | 1/107 |
| | 2.A.1.22 | 1/24 |
| 2.A.1.17 | 2.A.1.17 | 5/11 |

*MFS indicates major facilitator superfamily; NA, Not Applicable; TC, Transport Commission #.

†Motifs were determined by Pao et al[3] and designed to be specific for the MFS family indicated by TC number. The 2.A.1.1 motif is not a true family-specific motif in that it was constructed using only a portion of the 2.A.1.1 family members known at the time.

‡Core families were obtained from the automated protocol, using a single-seed sequence from the MFS family indicated by TC number.

**Table 7.** Summary of New Gene Candidates Cloned and Experimentally Characterized by Other Groups Subsequent to Identification by Our Protocol*

| Seed Sequence TC | New Gene Candidate Information | | Characterized Protein Match Information | | | |
|---|---|---|---|---|---|---|
| | **GI(s)** | **Length (bp)** | **GI** | **Name** | **%ID/match length (aa)** | **Reference** |
| 2.A.1.1 | 1551869 2900746 3058607 3058608 3237960 3959417 3961426 4018287 4290861 4564782 4851409 4851413 4988771 5232573 5361697 5865017 6639900 6709755 7320001 | 618 | 9910554 | Solute carrier family 2 member 9 (facilitated glucose transporter) | 99/143 | 39 |
| 2.A.1.1 | 3835081 3214654 7148985 5234142 5110215 | 621 | 12802047 | Facilitated glucose transporter glut11 | 98/179 | 40 |
| 2.A.1.1 and 2.A.1.19 | 5339943 | 508 | 21239030 | Renal ureate anion exchanger | 100/83 | 41 |
| 2.A.1.13 | 1166011 | 598 | 18699730 | *t* type amino acid transporter 1 | 100/91 | 42 |
| 2.A.1.13 | 6569405 6989741 | 369 | 18699730 | *t* type amino acid transporter 1 | 98/93 | 42 |

*TC indicates Transport Commission #; GI, NCBI Geneinfo Identifier; ID, Identifier.

been queried by our method and sufficient experimentalevaluation has been performed.

## *Comparison to Other Methods for EST Data Mining*

Besides the approach reported here, several other automated or semiautomated analysis strategies that use ESTs to find and characterize new genes have provided useful results.[43-46]

First, using an approach that is a precursor to that described here, Botka et al identified new members of the proton-oligopeptide transporter gene family.[43] This method uses the iterative neighborhood cluster analysis (INCA) algorithm to assemble gene families. These gene families are then used to search the EST databases in a second screen. Differences between INCA and our current protocol include somewhat different methods for family generation and, of course, the incorporation of family information to aid in the evaluation of new gene candidates.

Using a different approach, Schultz et al developed a semiautomated protocol to mine the EMBL EST database for new human signaling proteins[44] using information gleaned from conserved domains within gene families of interest as a starting point for the data mining process. This strategy is advantageous because a conserved domain may contain useful information, such as motifs required for the specific biological function mediated by a given family. When tested using conserved signaling domains from 100 different families, ESTs representing over 1000 novel human genes were found.

However, despite its advantages, the Schultz protocol requires the user either to limit searches to families for which a conserved domain sequence has already been described, or to define such a sequence prior to implementation of the protocol. In contrast, our protocol requires only a single sequence as input. This input sequence is then used to generate a family of related sequences dynamically, using up-to-date sequence information from available databases, all of which are then used to mine the EST database. Because our protocol performs family generation internally, the loss of information normally resulting from the use of only a single input sequence is minimized (see the following section).

Other strategies have focused on a more generic rules-based system for classifying all known ESTs according to their function. For example, The Institute for Genomic Research (TIGR) has constructed gene indices that classify all ESTs for a given organism into tentative consensus sequences, each of which represents a unique gene.[45] The function of each tentative consensus sequence is

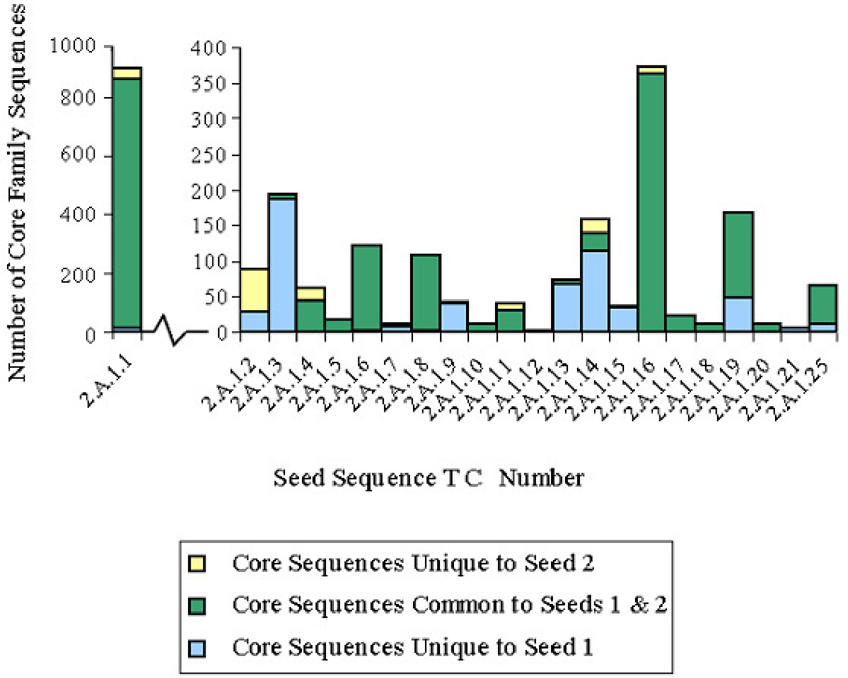

**Figure 4.** Graph showing the number of common core sequences, as well as the number of unique core sequences, found by 2 identical runs of the protocol using different seed sequences within the same Major Facilitator Superfamily (MFS) family. Only 22 of the 29 MFS families designated by Saier et al were tested. TC indicates Transport Commission #.

then annotated with the functional annotation of the expressed transcript sequence (derived from GenBank sequences) that best matches the tentative consensus sequence. This system has the added advantage of being fully automated.

Although the TIGR gene indices contain functional annotations for each tentative consensus sequence, because these annotations are generated from the information available for 1 or a few of the most similar expressed transcript sequences, the indices are likely to be sensitive to annotation errors.[47,48] In contrast, our approach uses family analysis to facilitate functional classification of new gene candidates, an approach that has been shown to be very useful in other classes of proteins.[49,50] The output from the automated phase of our protocol includes both a core family containing characterized protein sequences and the sequences of new gene candidates identified from comparison with the core family. Because new gene candidates are grouped together with a core family containing characterized genes, putative functions for the gene candidates can be inferred from examination of conserved sequence elements important for the function of the characterized family members. Thus, even though our strategy cannot yet be fully automated, it has the potential to provide more accurate and detailed functional information than do the other methods described above.

### *Problems and Dependencies*

Although our results demonstrate the utility of our approach, the current implementation suffers from some specific problems and dependencies that prevent full automation. Among the most important issues we have encountered are the lack of both controlled vocabulary and fixed format in database annotation, which complicates the automated removal of characterized[FOOTNOTE 6] human genes. For example, NCBI definition lines sometimes list multiple organism names and may use more than 1 name for a given organism, making it difficult to identify the organism of a specific gene. It is also difficult to determine whether a given gene had been characterized, because there is no controlled vocabulary term used to designate such genes. The best long-term solution to these problems is the conversion of all database annotation to a common structured format, such as that described by the Gene Ontology Consortium.[51]

Another problem complicating the automation of our analysis is that closely related seed sequences may find different sets of core family members in a PSI-BLAST search. This can be ascribed to the fact that if related seed sequences find a slightly different set of hits in the initial rounds of PSI-BLAST, differences in the final converged results may be amplified because of accumulated differences in the scoring profiles generated in each round of PSI-BLAST. To better understand the importance of this issue, we investigated how much care must be placed on selecting the proper seed sequence. Duplicate runs of the protocol were performed using 2 different seed sequences each from MFS families 2.A.1.1 to 2.A.1.21 and 2.A.1.25. Whenever possible, the second seed sequence was obtained from an organism that was not closely related to the organism from which the first seed sequence was taken. As shown in **Figure 4**, core families found using seeds from the same family overlapped well in only ~50% of the cases. Generally, poor overlap resulted from 1 seed sequence finding many fewer core sequences than the other found. The more disturbing problem, poor overlap resulting from the 2 seed sequences finding different sets of core family sequences, was uncommon but did occur, specifically for family 2.A.1.2, and to a lesser extent for 2.A.1.14.

Because certain seed sequences seem better able than others to generate a complete set of core family members, the automated protocol could be improved by modifications that provide a more intelligent method for choosing a seed sequence. One solution being investigated is a trial run of the initial steps of the protocol using a randomly selected seed, followed by an automated evaluation of the core family members found so that the best seed from among the core family sequences can be chosen as input for a more refined run of the protocol.

## ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health grants GM60595 and U01 HL66600 (P.C.B.), GM61390 (P.C.B and W.S.), and GM43102 (W.S.).

## FOOTNOTES

1. An alternative classification of human solute carrier (SLC) families has also been proposed (http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl). We chose the Saier classification over the human SLC system because it offers more complete coverage of all organisms. Moreover, the Saier classification encompasses many SLC families.

2. As of August 7, 2002, 36 MFS families have been listed at the Saier Web site (http://tcdb.ucsd.edu/tcdb/tcfamilybrowse.php?tcname=2.A.1#protein).

3. Twenty-nine MFS families (2.A.1.1-2.A.1.29) were available at the time this analysis was performed. However, MFS family 2.A.1.29 was not used, because a protein sequence corresponding to the listed genes could not be obtained at the time the analysis was performed.

4. These predicted transporter sequences have been used in the design of an expression microarray created to analyze the majority of human genes encoding transporters and ion exchangers.[31]

5. One of the entries in **Table 7** was found using seed sequences from 2 closely related families: 2.A.1.1 and 2.A.1.19. The experimentally determined function corresponds to that of family 2.A.1.19.

6. Here, the term "characterized" has been used loosely to mean either that the sequence in question has been experimentally determined to perform the function listed in its annotation, or that a human expert has determined that the sequence likely performs the annotated function based on comparison with experimentally characterized proteins.

## REFERENCES

1. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860-921.

2. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science. 2001;291:1304-1351.

3. Pao SS, Paulsen IT, Saier MH, Jr. Major facilitator superfamily. Microbiol Mol Biol Rev. 1998;62:1-34.

4. Paulsen IT, Sliwinski MK, Saier MH, Jr. Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. J Mol Biol. 1998;277:573-592.

5. Paulsen IT, Sliwinski MK, Nelissen B, Goffeau A, Saier MH, Jr. Unified inventory of established and putative transporters encoded within the complete genome of Saccharomyces cerevisiae. FEBS Lett. 1998;430:116-125.

6. Paulsen IT, Nguyen L, Sliwinski MK, Rabus R, Saier MH, Jr. Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. J Mol Biol. 2000;301:75-100.

7. Hogenesch JB, Ching KA, Batalov S, et al. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. Cell. 2001;106:413-415.

8. Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, Pellegrini-Toole A. The EcoCyc and MetaCyc databases. Nucleic Acids Res. 2000;28:56-59.

9. Ogasawara N. Systematic function analysis of Bacillus subtilis genes. Res Microbiol. 2000;151:129-134.

10. Blattner FR, Plunkett G III, Bloch CA, et al. The complete genome sequence of Escherichia coli K-12. Science. 1997;277:1453-1474.

11. Kunst F, Ogasawara N, Moszer I, et al. The complete genome sequence of the gram-positive bacterium Bacillus subtilis. Nature. 1997;390:249-256.

12. Wittenberger T, Schaller HC, Hellebrand S. An expressed sequence tag (EST) data mining strategy succeeding in the discovery of new G-protein coupled receptors. J Mol Biol. 2001;307:799-813.

13. Black DL. Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. Cell. 2000;103:367-370.

14. Graveley BR. Alternative splicing: increasing diversity in the proteomic world. Trends Genet. 2001;17:100-107.

15. Strehler EE, Zacharias DA. Role of alternative splicing in generating isoform diversity among plasma membrane calcium pumps. Physiol Rev. 2001;81:21-50.

16. Ingram VM. Abnormal human haemoglobin, III: the chemical difference between normal and sickle cell haemoglobins. Biochim Biophys Acta. 1959;36:402-411.

17. Qi M, Byers PH. Constitutive skipping of alternatively spliced exon 10 in the ATP7A gene abolishes Golgi localization of the menkes protein and produces the occipital horn syndrome. Hum Mol Genet. 1998;7:465-469.

18. Brett D, Hanke J, Lehmann G, et al. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett. 2000;474:83-86.

19. Cargill M, Altshuler D, Ireland J, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet. 1999;22:231-238.

20. Lai E. Application of SNP technologies in medicine: lessons learned and future challenges. Genome Res. 2001;11:927-929

21. Sadee W. Genomics and drugs: finding the optimal drug for the right patient. Pharm Res. 1998;15:959-963.

22. Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389-3402.

23. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403-410.

24. Huang X, Madan A. CAP3: a DNA sequence assembly program. Genome Res. 1999;9:868-877.

25. Adams MD, Venter JC. Should non-peer-reviewed raw DNA sequence data release be forced on the scientific community? Science. 1996;274:534-536.

26. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673-4680.

27. Tusnady GE, Simon I. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol. 1998;283:489-506.

28. Hillis DM, Moritz C, Mable BK. Molecular Systematics. 2nd ed. Sunderland, MA: Sinauer Associates; 1996.

29. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000;7:203-214.

30. Saier MH, Jr. A functional-phylogenetic system for the classification of transport proteins. J Cell Biochem. 1999; 32/33(suppl):84-94.

31. Anderle P, Rakhmanova V, Woodford K, Zerangue N, Sadee W. Messenger RNA expression of transporter and ion channel genes in undifferentiated and differentiated Caco-2 cells compared to human intestines. Phar.Res. 2003; in press.

32. The C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. 1998;282:2012-2018.

33. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. Gene index analysis of the human genome estimates approximately 120,000 genes. Nat Genet. 2000;25:239-240.

34. Ewing B, Green P. Analysis of expressed sequence tags indicates 35,000 human genes. Nat Genet. 2000;25:232-234.

35. Roest Crollius H, Jaillon O, Bernot A, et al. Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. Nat Genet. 2000;25:235-238.

36. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res. 2000;28:263-266.

37. Sonnhammer EL, Kahn D. Modular arrangement of proteins as inferred from analysis of homology. Protein Sci. 1994;3:482-492.

38. Kasahara M, Maeda M. Contribution to substrate recognition of two aromatic amino acid residues in putative transmembrane segment 10 of the yeast sugar transporters Gal2 and Hxt2. J Biol Chem. 1998;273:29106-29112.

39. Phay JE, Hussain HB, Moley JF. Cloning and expression analysis of a novel member of the facilitative glucose transporter family, SLC2A9 (GLUT9). Genomics. 2000;66:217-220.

40. Doege H, Bocianski A, Scheepers A, et al. Characterization of human glucose transporter (GLUT) 11 (encoded by SLC2A11), a novel sugar-transport facilitator specifically expressed in heart and skeletal muscle. Biochem J. 2001;359(pt 2):443-449.

41. Enomoto A, Kimura H, Chairoungdua A, et al. Molecular identification of a renal urate anion exchanger that regulates blood urate levels. Nature. 2002;417:447-452.

42. Kim DK, Kanai Y, Matsu H, et al. The human T-type amino acid transporter-1: characterization, gene organization, and chromosomal location. Genomics. 2002;79:95-103.

43. Botka C, Wittig T, Graul R, et al. Human proton/oligopeptide transporter (POT) genes: identification of putative human genes using bioinformatics. AAPS PharmSci.

2000; Article 2. Available at: http://www.aapspharmsci.org/scientificjournals/pharmsci/journal/16.html.

44. Schultz J, Doerks T, Ponting CP, Copley RR, Bork P. More than 1,000 putative new human signalling proteins revealed by EST data mining. Nat Genet. 2000;25:201-204.

45. Quackenbush J, Liang F, Holt I, Pertea G, Upton J. The TIGR gene indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res. 2000;28:141-145.

46. Allikmets R, Gerrard B, Hutchinson A, Dean M. Characterization of the human ABC superfamily: isolation and mapping of 21 new genes using the expressed sequence tags database. Hum Mol Genet. 1996;5:1649-1655.

47. Gerlt JA, Babbitt PC. Can sequence determine function? Genome Biol. 2000;1. Available at: http://genomebiology.com/1465-6906/1/REVIEWS0005.

48. Karp PD. What we do not know about sequence analysis and sequence databases. Bioinformatics. 1998;14:753-754.

49. Holm L, Sander C. An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. Proteins. 1997;28:72-82.

50. Babbitt PC, Hasson MS, Wedekind JE, et al. The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. Biochemistry. 1996;35:16489-16501.

51. The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. Genome Res. 2001;11:1425-1433.