Comparison of human solute carriers

Avner Schlessinger,[1,2,3]* Pär Matsson,[1,2,3] James E. Shima,[1,2,3,4] Ursula Pieper,[1,2,3] Sook Wah Yee,[1,2,3] Libusha Kelly,[1,2,3,5] Leonard Apeltsin,[1,2,3,5] Robert M. Stroud,[6] Thomas E. Ferrin,[1,2,3] Kathleen M. Giacomini,[1,2,3] and Andrej Sali[1,2,3]*

[1]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, California
[2]Department of Pharmaceutical Chemistry, University of California, San Francisco, California
[3]California Institute for Quantitative Biosciences, University of California, San Francisco, California
[4]Graduate Program in Pharmaceutical Sciences and Pharmacogenomics, University of California, San Francisco, California
[5]Graduate Program in Biological and Medical Informatics, University of California, San Francisco, California
[6]Department of Biochemistry and Biophysics, University of California, San Francisco, California

Received 14 August 2009; Revised 10 December 2009; Accepted 14 December 2009
DOI: 10.1002/pro.320
Published online 5 January 2010 proteinscience.org

Abstract: Solute carriers are eukaryotic membrane proteins that control the uptake and efflux of solutes, including essential cellular compounds, environmental toxins, and therapeutic drugs. Solute carriers can share similar structural features despite weak sequence similarities. Identification of sequence relationships among solute carriers is needed to enhance our ability to model individual carriers and to elucidate the molecular mechanisms of their substrate specificity and transport. Here, we describe a comprehensive comparison of solute carriers. We link the proteins using sensitive profile–profile alignments and two classification approaches, including similarity networks. The clusters are analyzed in view of substrate type, transport mode, organism conservation, and tissue specificity. Solute carrier families with similar substrates generally cluster together, despite exhibiting relatively weak sequence similarities. In contrast, some families cluster together with no apparent reason, revealing unexplored relationships. We demonstrate computationally and experimentally the functional overlap between representative members of these families. Finally, we identify four putative solute carriers in the human genome. The solute carriers include a biomedically important group of membrane proteins that is diverse in sequence and structure. The proposed classification of solute carriers, combined with experiment, reveals new relationships among the individual families and identifies new solute carriers. The classification scheme will inform future attempts directed at modeling the structures of the solute carriers, a prerequisite for describing the substrate specificities of the individual families.

Keywords: solute carrier transporters; pharmacogenetics; profile–profile alignment; sequence analysis; protein function prediction; family classification

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: National Institutes of Health; Grant numbers: U01 GM61390, U54 GM074929, R01 GM54762, U54 GM074945, P41 RR01081; Grant sponsors: Hewlett Packard, IBM, NetApps, Intel, Ron Conway, Mike Homer.

**Correspondence to:* Avner Schlessinger, Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, University of California, San Francisco, CA. E-mail: schles@salilab.org or Andrej Sali, Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, University of California, San Francisco, CA. E-mail: sali@salilab.org

Introduction

Solute carriers

Solute carriers are eukaryotic membrane proteins that control the uptake and efflux of various solutes, including amino acids, sugars, and drugs.[1,2] Solute carriers and their prokaryotic homologs include facilitative transporters (energy for transport provided by an electrochemical gradient) as well as active transporters, including both cotransporters and exchangers (energy for transport provided by diverse energy coupling mechanisms).[1–4]

Sequence-based classification of solute carriers

The Gene Nomenclature Committee (HGNC) of the Human Genome Organization (HUGO)[2] classifies ~400 human solute carriers into 47 families.[1] According to this classification, members of a family share a similar substrate and at least 20–25% sequence identity to at least one other member of the family.[1] At least initially, atomic structures of transporters were apparently not considered in constructing this classification. Several other classifications have also been proposed.[3,5] The transporter classification (TC) system aims to automatically classify transporters across all organisms (http://www.tcdb.org/).[3] All known transporters, including solute carriers, are divided into a hierarchical system of "classes," "subclasses," and "families." Human solute carriers are represented in several distinct classes of the TC database (TCDB).[4] For example, the neurotransmitter family SLC6 is grouped into the various subclasses of 2.A.22.x.x. Another phylogenetic (tree-based) analysis of human solute carriers proposed that they consists of 15 related families that can be organized into four distinct clusters, as well as 32 additional unlinked families.[5]

Structures of solute carriers and their homologs

Solute carriers are typically composed of one large domain consisting of 10–14 transmembrane α-helices. Currently, the only solute carriers of known structure are the human Rhesus glycoprotein RhCG ammonium transporter (SLC42A3) (Gruswitz et al., submitted for publication), which belongs to the SLC42 family, and the cow mitochondrial ADP/ATP carrier (SLC25A4),[6] which belongs to the SLC25 family. In addition, there are a number of structures for prokaryotic homologs of solute carriers.[7–21] These structures revealed that some solute carrier families are related to each other, despite weak sequence similarities (even at less than 15% sequence identity). For example, the prokaryotic members of the SLC5 (vSGLT),[22] SLC6 (LeuT),[23] SLC7 (Arginine/agmatine antiporter (AdiC),[11,12] and ApcT transporter[13]) families as well as the bacterial transporter NCS1[24] and the Na$^+$/betaine symporter BetP[14] are classified into the same structural family in the orientations of proteins in membranes (OPM) database (Fig. 1)[25] and share similar transport mechanisms.[10,28] In contrast, some solute carrier families have different folds (Fig. 1). For example, the SLC1 member Glp from *Pyrococcus horikoshii* has the proton glutamate symport fold,[15] whereas the SLC6 member LeuT from *Aquifex aeolicus vf5* adopts the SNF-like fold.[8] Therefore, the use of the term superfamily to refer to all solute carriers is misleading, because it generally implies a common ancestor as well as detectably similar sequences and structures.[29]

Function and pharmacology of solute carriers

Solute carriers play a role in a variety of cellular functions. For instance, the sodium- and chloride-dependent neurotransmitter transporter family (SLC6) consists of important neurosignaling proteins, such as the dopamine and serotonin transporters.[30,31] Many antidepressants selectively inhibit the SLC6A4 (SERT) serotonin transporter, which is responsible for serotonin reuptake from synaptic spaces.[32,33] In addition, solute carriers can modulate drug levels within the body by regulating their absorption, distribution, metabolism, and elimination (ADME). For example, the organic ion transporter family (SLC22) mediates the uptake of anticancer and antiviral drugs into the liver and kidney.[34,35] Thus, genetic variation in solute carriers may result in differential response to drugs (pharmacogenetics). For instance, the intracellular concentrations of the antidiabetic drug metformin are affected by genetic variations in the organic cation transporter 1 (SLC22A1 or OCT1).[36–39] Consequently, solute carriers are principal drug targets of utmost clinical relevance.

Toward a description of specificity determinants in solute carriers

As with most proteins, the function of a solute carrier is determined by its structure and dynamics. For example, the shape and physicochemical properties of the binding site on the transporter (i.e., specificity determinants) determine what molecules do and do not bind to the transporter (i.e., binding specificity), which in turn helps determine what molecules do and do not get transported by the transporter (i.e., substrate specificity). The mechanism of transport, which can be different for different familes,[10,20–23,28,40] describes how the specificity determinants of a transporter result in the binding specificity and substrate specificity. Therefore, an important step toward descriptions of their mechanisms of transport includes the characterization of their structures in different conformational states. Biochemical and crystallographic studies[10,20–23,28,40] suggest that solute carriers from several structural families transport solutes using the "alternating

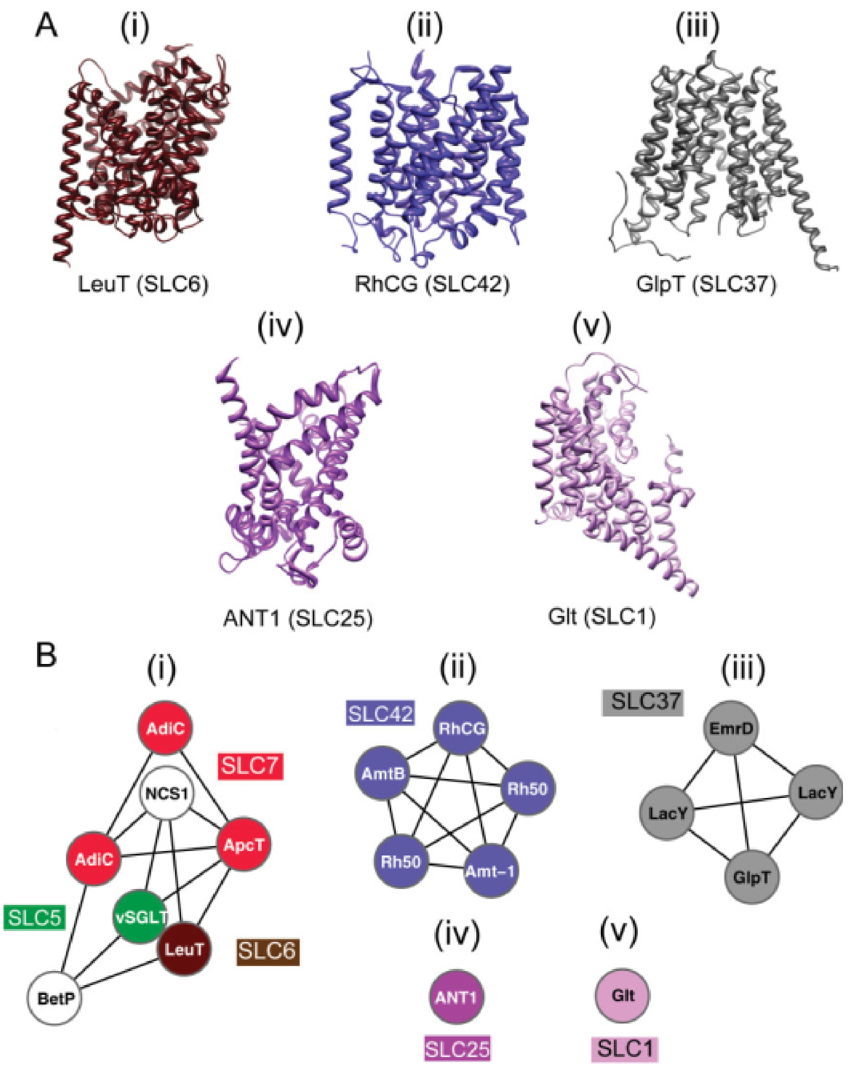

Figure 1. Structure-based classification of solute carriers. A: Structures of solute carriers and their homologs representing the currently known structural classes.[25] The folds include (i) the NSF-like fold, (ii) the ammonium transporter fold, (iii) the MFS general substrate transporter fold, (iv) the mitochondrial carrier fold, and (v) the proton glutamate symport protein fold. B: The relationships between the structures are visualized using cytoscape, based on pairwise structural alignment scores computed by SALIGN.[26] Each link represents a pairwise structural alignment with an SALIGN score of at least 30. The colored nodes represent proteins that are either a human solute carrier or similar in sequence to at least one human solute carrier. Nodes in white correspond to proteins that are not detectably similar in sequence to human solute carriers (i.e., an overlap of at least 150 residues of the target sequence to a sequence of a known structure at the sequence identity cutoff 20%, according to PSI-BLAST[27]).

access" mechanism, in which the transporter alternatively exposes its binding site to either side of the membrane.[41]

As one recent example, X-ray structures of the bacterial leucine transporter LeuT in complex with various amino acid ligands have provided insights into the substrate specificity of several human solute carrier families, including SLC5, SLC6, and SLC7.[42] In particular, it has been suggested that for a substrate to be transported efficiently by LeuT, it needs to bind to the binding site on the transporter surface as well as fit within the binding cavity of the "occluded" transporter state[42]; occluded states are conformations that solute carriers from various structural families adopt during the transport process.[10,21,40,43] Tryptophan is much larger than leucine, but like leucine it binds to the binding site of LeuT. However, it was suggested that tryptophan does not fit into the cavity in the occluded state and consequently traps the LeuT transporter in the "open-to-out" conformation, thereby acting as a competitive inhibitor of Leu transport.[42] While this inhibition model was proposed based on a partially occluded state of LeuT, the recent structure of BetP reveals a fully occluded conformation supporting the transport mechanism for the structural family.[14] Additionally, crystal structures of LeuT with several antidepressants proposed a noncompetitive

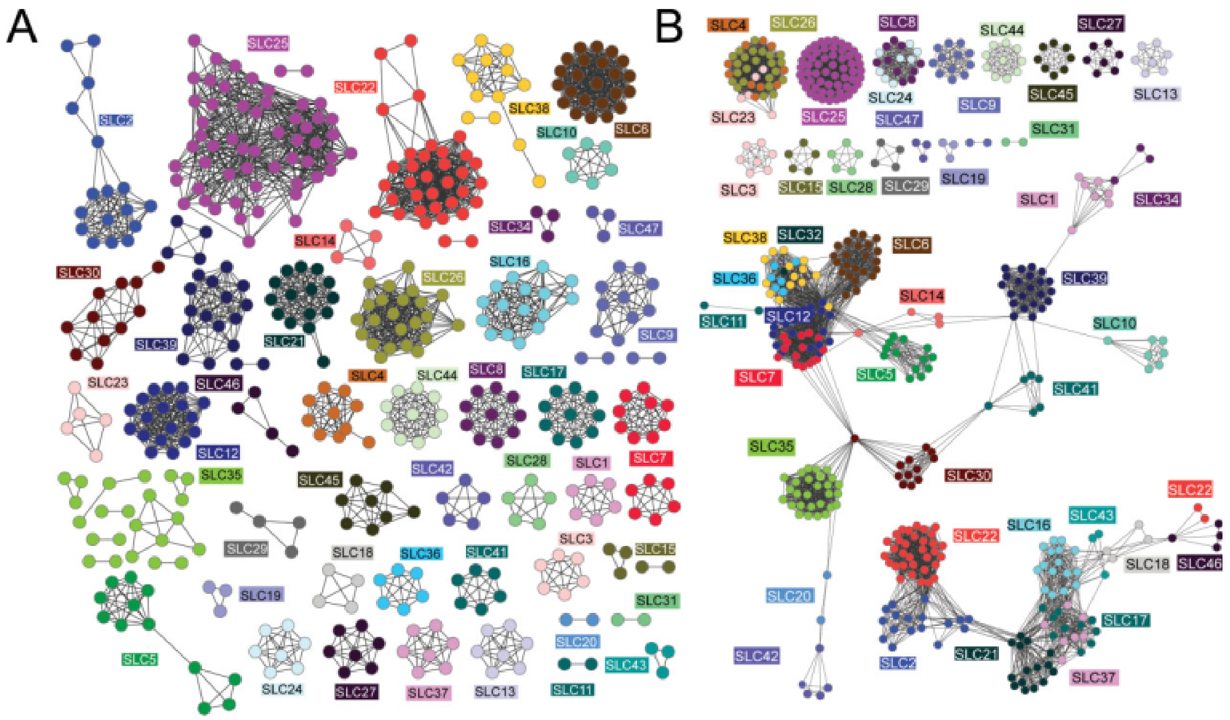

Figure 2. Classification of solute carriers using similarity maps. The relationships between solute carrier sequences are visualized using the modified edge-weighted spring-embedded layout in cytoscape 2.6.1.[55] Briefly, in the spring-embedded algorithm, the connected nodes are being attracted toward each other, while nodes that are unconnected are pushed apart. The nodes are connected by springs with resting lengths proportional to the shortest-path distance between them. The algorithm then iteratively adjusts the positions of each node to minimize the total "energy" of the system. The lengths of the springs are also determined by link weights, which are derived from the alignment scores better than a threshold. A: Each link represents a pairwise alignment with sequence identity of at least 25% and an *E*-value of less than 1. B: Each link represents a pairwise alignment with sequence identity of at least 10% and an *E*-value of less than 1.

inhibition mechanism for transport by LeuT that involves a second inhibitory binding site.[44,45] It was suggested that this site binds another substrate molecule,[46] allosterically triggering the intracellular release of the first substrate molecule. It was also suggested that the second site was occupied by the detergent used for the crystallization of LeuT,[8,47] acting as a noncompetitive inhibitor.[47]

Here, we describe a comprehensive comparison of solute carriers to inform future attempts directed at modeling the structures of the solute carriers, a prerequisite for describing the substrate specificities of the individual families. Proteins with a common evolutionary origin adopt similar structures and tend to have related molecular functions. When their structures are not known, one way to detect relationships between proteins is through comparison of their sequences.[48–51] The resulting sequence-based connections among related proteins can be useful for inferring similarities and differences in structural and functional features (e.g., fold, ligand binding site, and molecular mechanism) of uncharacterized proteins based on their characterized aligned homologs. In this article, similarity networks based on sequence profile alignments are employed to relate all known human solute carrier sequences. We refine the current classification of solute carriers and predict new connections between solute carriers. The connections are interpreted in terms of the structures, substrate type, transport mode, and tissue specificity of the proteins (see Results section). We then discuss our approach, and the utility of our results for improving comparative models of solute carriers and describing substrate specificity within the families (see Discussion section). Finally, we describe our comparison approaches, the databases of sequences and their functional annotations, as well the visualization software used in this study (see Materials and Methods section).

Results

Similarity maps confirm and refine current classification

We illustrate relationships in protein families using sequence similarity networks.[52] The nodes in the graphs represent sequences of the known solute carriers, color coded by their family as annotated by Uniprot[53] (or Genbank,[54] if not in Uniprot); proteins not annotated as solute carriers are not included in this graph. To construct this map, we used a cutoff similar to that used originally by HGNC[1,2] to define the solute carrier families [Fig. 2(A)]. The new similarity network is generally, but not always, in agreement with the current HGNC classification. While many families already appear well defined, the current classification needs to be refined. For instance, the nucleoside-sugar transporter family SLC35 is much more divergent in sequence than other families, and may contain subfamilies with more specific functions [Fig. 2(A)]. As another example, two variant forms of SLC22A18 are not connected to the rest of the SLC22 family. In fact, even when we used a looser similarity cutoff [Fig. 2(B)] to connect protein sequences, these two SLC22A18 isoforms are not connected to the rest of the SLC22 family; instead, they are more similar to members of the SLC46 family. The dissimilarity between the SLC22A18 isoforms and the rest of the SLC22 family is also demonstrated by a phylogenetic tree and co-occurrence of specific motifs in the SLC22 family (Fig. 6).

Sequence-based relationships are consistent with structural data

One potential advantage of a similarity network over a phylogenetic tree analysis is its ability to conveniently represent multiple links between proteins as a function of a similarity cutoff.[52] Thus, to study relationships between solute carriers, we used a variety of cutoffs to connect the nodes (Fig. 2; Supporting Information Figures S1–S2). One of our goals was to use a cutoff that is so permissive to capture most distant relationships, albeit at a cost of requiring subsequent careful inspection to eliminate likely false positives. Therefore, we simultaneously used a sequence identity cutoff of 10% and an E-value cutoff of 1 [Fig. 2(B)] as the threshold to connect the nodes. Although this criterion may not always result in an accurate detection of a conserved structure-function relationship, a similar cutoff was previously shown to work well for the bacterial homologs of the SLC5 and SLC6 families.[7]

The observed structural similarity of SLC5 and SLC6 families was reflected in the close proximity of these two families in the similarity map [Fig. 2(B); brown and green circles]. Furthermore, three additional features of the similarity network at the "permissive" sequence identity cutoff of 10% and E-value cutoff of 1 were consistent with data not used to construct it: (i) the links between SLC6 and SLC7 family members are in agreement with the structural similarities among the prokaryotic members of the SLC7[11–13] and SLC6[8] families (Fig. 1); (ii) the separation between the SLC25, SLC37, SLC42, and SLC1 families is consistent with the structural dissimilarity among the cow SLC25A4[6] mitochondrial ADP/ATP carrier, the bacterial SLC37 homologs,[16–19] the human SLC42A3[6] RhCG ammonium transporter, and the archaean SLC1 homolog[15] (Figs. 1 and 2); and (iii) the separation of the SLC3 family from the rest of the solute carriers is in agreement with its unique fold containing only one predicted transmembrane helix (they are not transporters on their own and only become a part of a transporting complex when bound to SLC7 family members[56]) [Fig. 2(B)].

Unknown relationships between families are proposed

Generally, families that cluster together based on their sequences also have similar substrates [Fig. 3(A)]. For example, the organic anion transporter cluster (red) includes SLC16 (monocarboxylate transporters), SLC21/SLCO (organic anion transporters), SLC37 (sugar-phosphate transporters), and SLC17 (vesicular glutamate transporters). Another example is the amino acid transporter cluster (green), which includes the SLC36 (the proton-coupled amino acid transporters), SLC32 (the vesicular inhibitory amino acid transporters), SLC6 (the sodium- and chloride-dependent neurotransmitter transporters), and SLC7 families (the cationic amino acid transporter/glycoprotein-associated transporters). In contrast, some families with similar functions are not linked in the network, indicating that different folds have evolved to perform similar functions. For example, neurotransmitter transporters are found in three distinct clusters: (i) The SLC1 family (pink; the proton glutamate symport protein fold), (ii) the SLC6 and SLC32 families (brown and dark green, respectively; the NSF-like fold), and (iii) the SLC17 and SLC18 families (teal and light gray, respectively; the MFS general substrate transporter fold).

Interestingly, the organic ion transporter family (SLC22) is highly connected to the facilitative glucose transporter family (SLC2), indicating a possible evolutionary link between the families that has not been previously appreciated and is not obvious from their prototypical substrates.[1] This relationship is also manifested in a high conservation of the type and spatial position of the functionally important residues (Supporting Information Figure S3). Further validation of this relationship is provided by the recent identification of the fructose transporter, GLUT9 (SLC2A9),[57] as a transporter of urate.[58,59] Urate, a prototypical endogenous substrate of SLC22A12 (URAT1),[60] was also identified as a SLC2A9 substrate in our own uptake experiments (Supporting Information Figure S4). The interconnectivity of the SLC2 family with the SLC16, SLC17, SLC18, SLC21, SLC22, SLC37, SLC43, and SLC46 families [Fig. 2(B); bottom right] indicates that all these families share similar functions and the MFS fold.

When the similarity network is constructed based on the E-value cutoff alone (Supporting Information Figures S1–S2), many families are well-

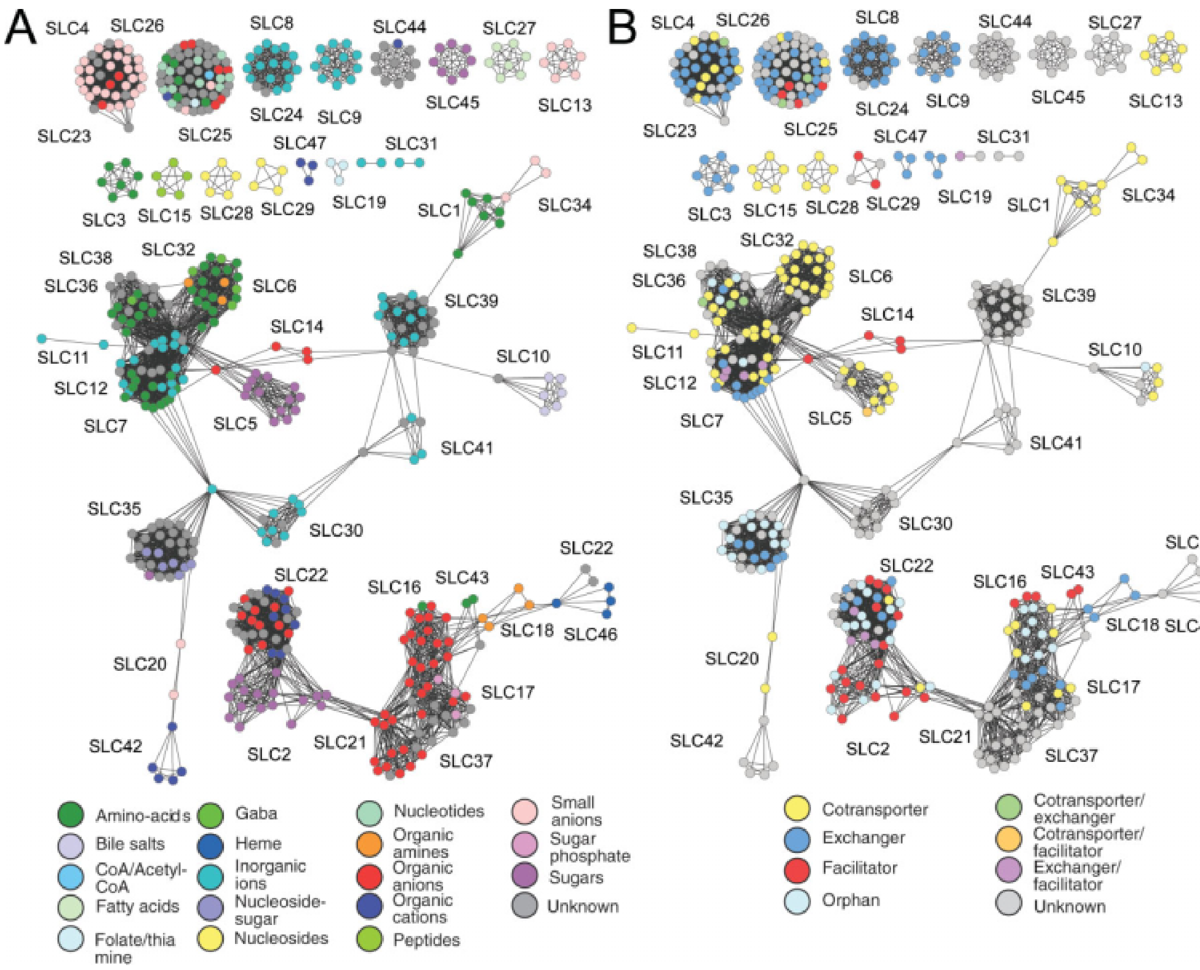

Figure 3. Substrate type and transport mode mapped onto the solute carrier sequences. A: The colors represent the prototypical substrates of the transporters (see Materials and Methods section). In most cases, the sequence-based clustering also correlates with substrate type. For example, amino acid transporters, such as SLC6, SLC36, and SLC38, are clustered together (green). B: The colors represent the transport mode. The three major groups include cotransporters (yellow), exchangers (blue), and facilitators (red). "Orphan" nodes (cyan) represent transporters whose substrates are unknown, and "unknown" nodes (grey) represent transporters without a known transport mode. For some transporters, different modes of transport have been reported. For example, orange nodes mark transporters that are reported to be facilitators or cotransporters.

linked even for stringent cutoff values (E-value $\leq$ 0.001; Supporting Information Figure S2); again, the linked families are generally functionally related (e.g., SLC8 and SLC24). These links are missed with the sequence identity cutoff of 25% [Fig. 2(A)], which is frequently used to infer a sequence-structure relationship. Taken together, these data show that sequence identities alone may not always capture functional and structural similarities in solute carriers. Instead, it is beneficial to construct similarity networks with a variety of cutoff values on both the E-value and sequence identity.

Transport mode is conserved within clusters

We now examine the conservation of the transport mode within and across the HGNC families; the type of transport mode is indicated by color on the solute carrier similarity network [Fig. 3(B)]. Conservation of the transport mode within and across families can assist structural studies (e.g., cocrystallizing the target protein with a ligand to increase the odds of crystallization) as well as shed light onto the evolution of solute carriers (see Discussion section). Unsurprisingly, the mode of transport is conserved within most families; most families consist entirely of facilitative transporters [Fig. 3(B); red, e.g., SLC2], cotransporters (yellow; e.g., SLC6), or exchangers (blue; e.g., SLC8). Furthermore, clusters of sequence-related families also share the mechanism of transport. Most proteins from the cluster containing the SLC5, SLC6, SLC7, SLC11, SLC12, SLC32, SLC36, and SLC38 families are cotransporters. The majority of the proteins from the cluster containing the SLC2 and SLC22 families are facilitative transporters, and the majority of the proteins from the cluster containing the SLC8 and SLC24 families are exchangers. The conservation of functional traits such as substrate type and transport mode provide

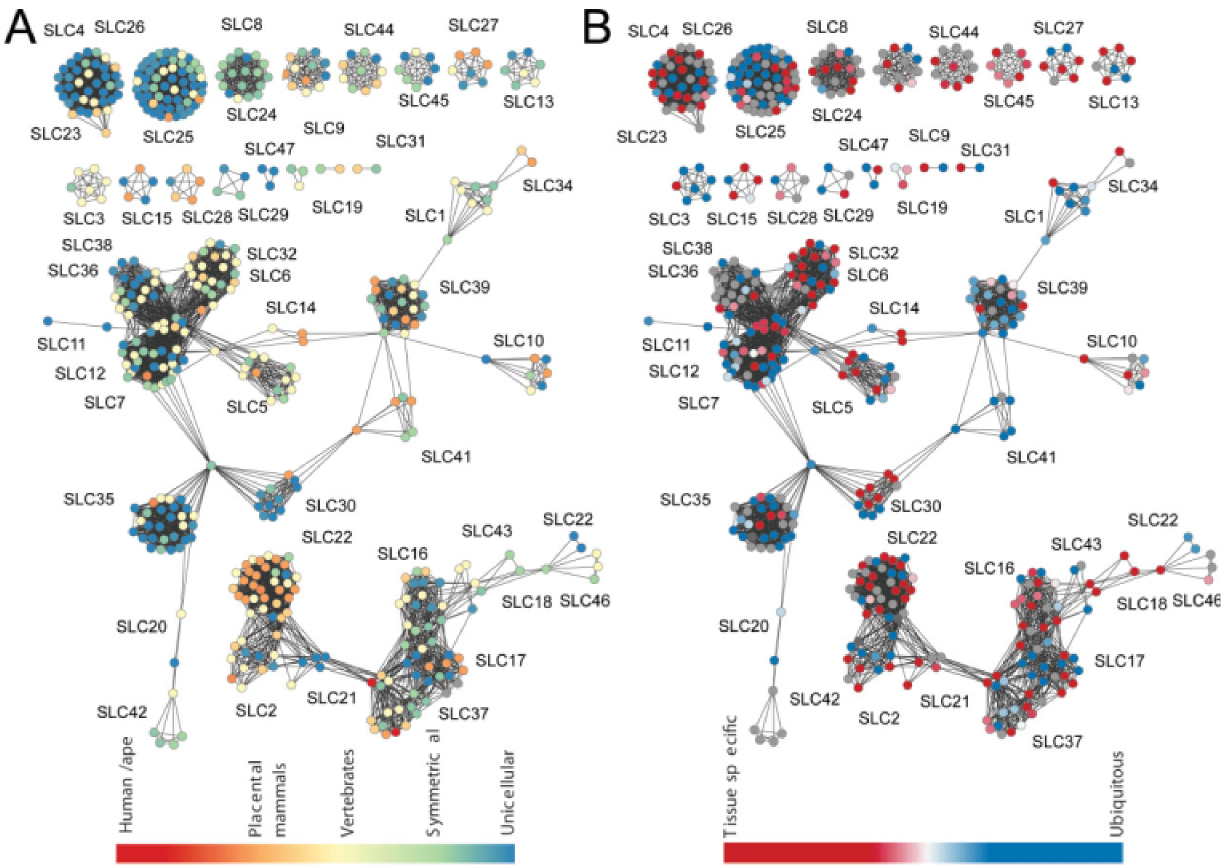

Figure 4. Conservation of solute carriers across eukaryotic organisms and tissues. A: The colors of the nodes indicate the oldest species in which the corresponding transporter is found, ranging from old (blue) to new (red). For instance, most members of the sugar transporter family SLC5 appeared for the first time in symmetrical organisms, whereas most SLC22 members appear only in higher mammals. B: The colors of the nodes indicate tissue specificity of the corresponding transporters, ranging from highly tissue-specific (red) to ubiquitous (blue). For example, SLC32 transporters, which are key proteins for synaptic release of inhibitory amino acids, are highly specific to the nervous system, whereas most members of the mitochondrial transporter family SLC25 are ubiquitously expressed.

some validation for using the permissive cutoff (10% sequence identity and E-value of 1) for constructing the network.

Family function is weakly linked to organism complexity

Next, we looked at conservation of solute carriers across eukaryotes of varying "complexity" level [Fig. 4(A)]. The assignment of the most similar pairs of transporter sequences between any two organisms was obtained from the NCBI HomoloGene database (see Materials and Methods section). For instance, SLC25 proteins are found in all eukaryotes, including unicellular organisms [blue; Fig. 4(A)]. Members of the neurotransmitter transporter family SLC6 are only found in vertebrates and higher organisms [yellow; Fig. 4(A)].

We hypothesized that the solute carriers unique to higher eukaryotes may have evolved to perform specific functions, and will therefore also be tissue specific (e.g., expressed in the brain only). On the basis of mRNA expression levels in different tissues,[38] we calculated a measure of tissue specificity ranging from 0 to 1 [see Materials and Methods section; Eq. (1)].[61] The value of 0 [blue; Fig. 4(B)] indicates that the transporter is expressed at similar levels in all tissues and 1 [red; Fig. 4(B)] means that the protein is expressed in a single tissue. Some families confirm our hypothesis. For example, most SLC22 transporters are expressed in the liver, kidney, intestines, and brain—organs with specialized functions that do not exist in lower eukaryotes. Additional examples include the SLC6, SLC17, SLC18, and SLC32 families that are mostly expressed in specific tissues of the nervous system and also appeared late in evolution. In contrast, SLC25 (mitochondria transporters) and SLC35 (Golgi transporters) are found in all eukaryotes and are expressed ubiquitously. However, on average, there was no correlation between tissue specificity and organism complexity in eukaryotes (the correlation coefficient between tissue specificity and organism complexity rank derived from HomoloGene annotation was −0.03). Most of the families consist of a mix of tissue specific and ubiquitous members.

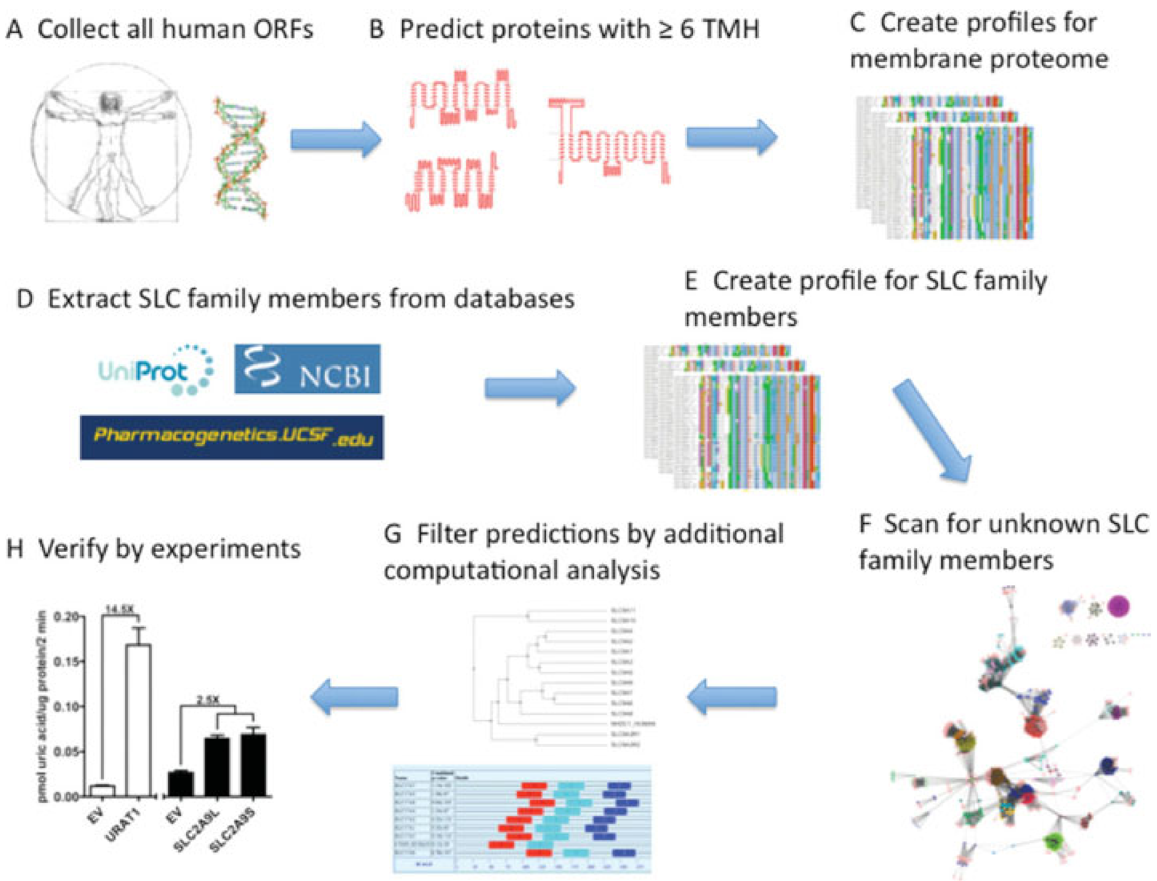
A Collect all human ORFs
B Predict proteins with ≥ 6 TMH
C Create profiles for membrane proteome
D Extract SLC family members from databases
UniProt
NCBI
Pharmacogenetics.UCSF.edu
E Create profile for SLC family members
F Scan for unknown SLC family members
G Filter predictions by additional computational analysis
H Verify by experiments
pmol uric acid/ug protein/2 min
0.20
0.15
0.10
0.05
0.00
14.5X
2.5X
EV
URAT1
EV
SLC2A9L
SLC2A9S

Figure 5. Steps in identifying and classifying unknown solute carriers. (A) We extracted all human ORFs from Ensembl, (B) filtered out proteins with less than six predicted membrane α-helices, and (C) constructed multiple sequence profiles for the remaining sequences. (D) Simultaneously, a list of known solute carriers was extracted from public databases, and (E) again, for each protein sequence we created a multiple sequence profile. (F) We aligned a profile of each known solute carrier sequence with each of the human membrane protein profiles, resulting in a list of human membrane proteins that are similar to at least one known solute carrier. (G) Additional bioinformatics analysis, including construction of phylogenetic trees and detection of family-specific sequence motifs, allows us to identify high confidence predictions. (H) Finally, we verify our computational predictions experimentally by measuring the rate of substrate uptake into cells expressing tested transporters.

Unknown solute carriers are predicted using a sequence-based approach

We used a multistep approach to predict unknown solute carriers in humans (see Materials and Methods section; Fig. 5). To obtain a preliminary list of all human solute carriers, we expanded all currently annotated solute carrier sequences by their matches against all human proteins with at least six predicted transmembrane helices. This preliminary list contained 199 solute carriers representing 34 families. Because of the limitations of sequence-based function annotation,[62–66] further computational analysis was performed. For example, of the 199 putative solute carriers, 21 were potential new members of the SLC22 family. We subsequently excluded entries that had been removed from the updated ENSEMBL[67] database during the preparation of this manuscript, splice variants of known SLC22 transporters, and proteins that were recently annotated manually as SLC22 transporters[54,68] [Fig. 6(A)]. After these sequences were removed, two potential new SLC22 members remained. We evaluated these putative SLC22 members using common sequence analysis approaches, such as multiple sequence alignment, phylogenetic tree construction, and motif detection (Figs. 5 and 6). Interestingly, ENS00000182157 is more similar to SLC22A17 than to any other member of the SLC22 family based on a phylogenetic tree [Fig. 6(B)]. Additionally, this protein contains the majority of the sequence motifs found in known SLC22 transporters by MEME[70] [Fig. 6(C)]. We highlight four high-confidence predictions (Table I).

Discussion

Using profile–profile alignments and sequence-based clustering, we updated and extended the map of the solute carrier relationships. We then annotated different functional characteristics of the transporters onto the sequence-based network and found previously unknown sequence-function relationships within this diverse and important membrane

Table I. *Putative Solute Carriers in the Human Proteome*

Family[a]	SwissProt/ UniProt[b]	Gene name[c]	Protein name[d]	Family function[e]
SLC6	A6NF70_HUMAN	—	Transporter	The sodium- and chloride-dependent neurotransmitter transporter family
SLC9	NHDC1_HUMAN	NHEDC1	Sodium/hydrogen exchanger-like domain-containing protein 1	The Na^+/H^+ exchanger family
SLC22	S22AX_HUMAN	—	Putative solute carrier family 22 member ENSG00000182157	The organic cation/anion/ zwitterion transporter family
	SVOPL_HUMAN	SVOPL	Putative transporter SVOPL	

[a] Family marks the family of the closest known solute carrier homolog based on sequence identity, derived from profile–profile alignment.

[b] SwissProt/UniProt marks the Swissprot/Uniprot[52] identifier that was linked from the ENSEMBL gene found in the profile–profile scan against the human membrane proteome.

[c] Gene name marks the name of the gene according to the "gene names" field in the corresponding SwissProt/UniProt entry.

[d] Protein name marks the name of the protein according to the "protein names" field in the corresponding SwissProt/UniProt entry.

[e] Family function marks the function of the predicted solute carrier family based on the HGNC classification described in Ref. [1].

transporters. Finally, we predicted unknown solute carriers. Our analysis can assist in improving models of solute carriers and highlighting functionally important residues.

Correlating functional attributes with solute carrier similarity networks is informative

Construction and visualization of similarity networks as well as the display of information on these networks were previously applied to study GPCRs, kinases, and crotonase enzymes.[52] Similarity network topology of multidomain proteins, which are highly abundant in the human proteome, can be influenced by variations of domain organization across different families.[52,71] Solute carriers usually consist of only one domain and similar membrane topology; therefore, using similarity networks is especially informative in studying these proteins. One advantage of similarity networks compared with phylogenetic trees is that they preserve all observed connections among proteins.[52] This feature is particularly relevant for analyzing solute carriers because only a subset of distant pairwise relationships may be detectable by sequence comparison. Although some links between the nodes in the different similarity networks do not represent alignment scores traditionally considered for reliably transferring structure and function [Fig. 2(B)], the multiple connections in our network do capture known structural and functional similarities across families; families known to have similar structures, substrate types, and transport modes tend to be clustered together (Fig. 3).

For example, the SLC22 and SLC2 families are highly interconnected [Fig. 2(B)]. Although the links between these families represent weak similarities between their members (typically 8–15% sequence identity), they collectively result into a distinct cluster when visualized via the network (Fig. 2; Supporting Information Figures S1–S2). This clustering is coupled with the localization of functional residues (Supporting Information Figure S3), as well as the overlap in substrate type between particular members of each family (Fig. 3; Supporting Information Figure S4). Interestingly, transport mode differences indicate considerable divergence between the SLC2 and SLC22 families: SLC2 transporters are almost all facilitative transporters, whereas SLC22 transporters may be facilitative transporters, exchangers, or in some cases, cotransporters (e.g., SLC22A5).

The sequence-based link between the SLC2 and SLC22 families is supported by the common substrates among some of their members (Supporting Information Figure S4),[58,59] and suggests that their structures may be similar. However, in the absence of experimentally determined structures from both families, a more definitive confirmation of the structural similarity requires additional experiments, such as site-directed mutagenesis assays, crosslinking experiments, and assessment of residue accessibility to membrane-impermeant reagents.[72–74]

Different information is provided by different classification schemes

One limitation of similarity networks is the absence of an underlying model of sequence duplication, divergence, and selection.[52,75] Therefore, it was suggested that they are not as accurate as phylogenetic (tree-based) approaches when there is substantial variation in the rates of evolution between different solute carriers.[75] To overcome this issue, we (i) constructed maps using a variety of similarity cutoffs (Fig. 2, Supporting Information Figures S1–S2), (ii) used profile–profile alignments that are more sensitive in detecting structural relationships in membrane proteins than other alignment approaches,[76] (iii) verified the maps by known solute carrier structures (Fig. 1), and (iv) clustered the sequences using

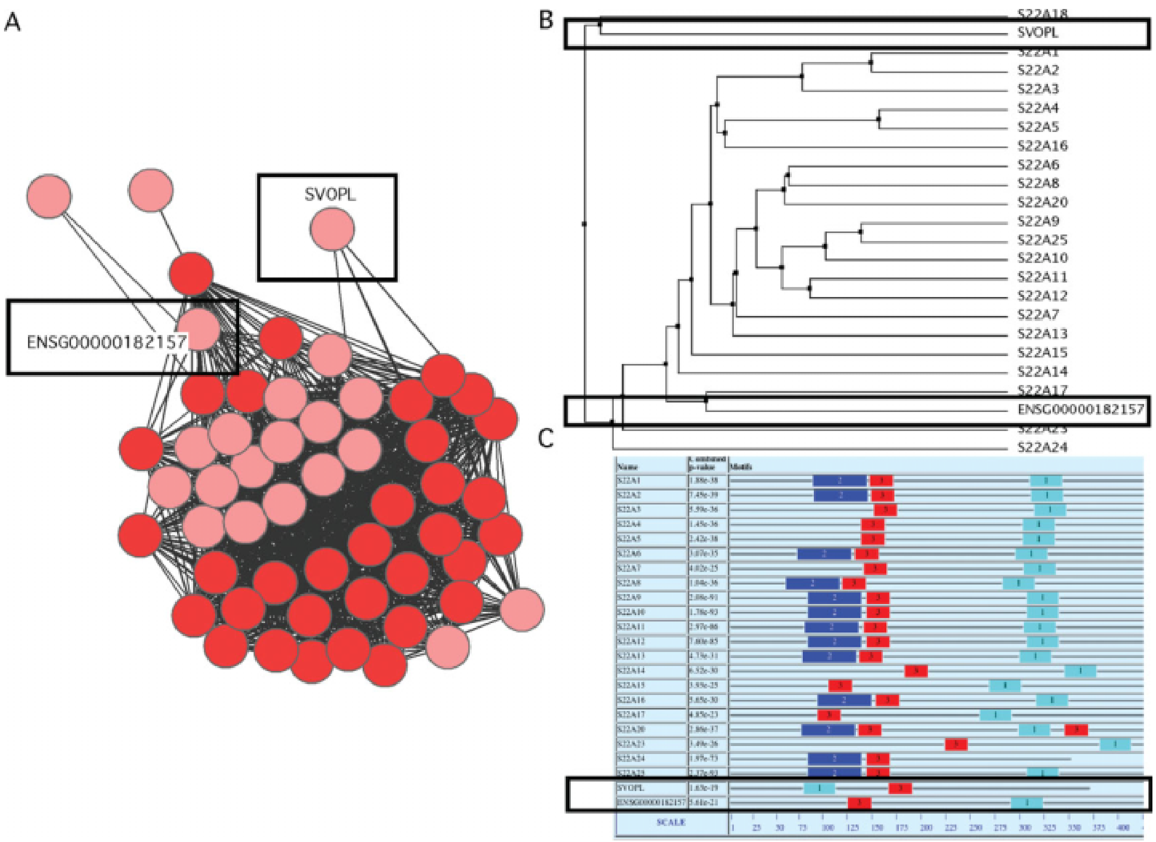

Figure 6. Predicted members of the SLC22 family. (A) Known SLC22 family members are represented by red nodes in the similarity network. An initial set of potential solute carriers is represented by pink nodes. Each link to these unannotated nodes indicates a pairwise alignment with sequence identity of at least 25% and an *E*-value of less than 1. The putative members of the SLC22 family were further evaluated using sequence analysis tools such as (B) phylogenetic trees and (C) analysis of family specific motifs. The SLC22 family is predicted to have two additional members [S22AX_HUMAN (ENS00000182157) and SVOPL_HUMAN] (black rectangles). A previous analysis of the SLC22 family also suggested that SVOPL might be an uncharacterized member.[69]

a complementary method—the Network Filtration Protocol (NFP) that filters dense similarity networks to estimate boundaries between families (see Materials and Methods section) (Supporting Information Figure S5) (Apeltsin et al., submitted for publication). NFP-based clustering indicates that SLC6, SLC16, and SLC22 are "hubs" in sequence space (i.e., they have a relatively large number of neighboring families). Hubs are particularly relevant targets for structural and functional characterization by the structural genomics consortia, which aim for complete coverage of integral membrane proteome,[77] because hub structures can serve as templates for computational modeling of many other proteins with similar sequences (Kelly et al., submitted for publication).[78,79] In addition, similarity to many other families may indicate a broad functional specificity of a family. For example, within the SLC22 family, there are transporters for organic cations [e.g., SLC22A1 (OCT1)] and organic anions [e.g., SLC22A6 (OAT1) and SLC22A12 (URAT1)]. This broad specificity of the SLC22 family is reflected in its similarity to both the SLC16 family, which transports an array of organic anions,[80–82] and the SLC18 family, which transports organic cations (primarily monoamines). Further, as noted previously, URAT1 (SLC22A12) and GLUT9 (SLC2A9) both transport uric acid (Supporting Information Figure S4).

Implication for drug design: improving accuracy of comparative modeling and virtual screening

Computational methodologies can play a key role in several steps of the drug discovery process, significantly reducing its costs.[83] For instance, virtual screening techniques can considerably decrease the time and cost of searching the chemical space for compounds that bind to a protein of known structure.[84,85] When experimentally determined atomic structures are not available, comparative modeling may be used to predict target structures when homologous template structures are known. However, the accuracy of comparative models is correlated with the template-target percentage sequence identity. High-accuracy models are typically based on more than 40% sequence identity to their templates, whereas alignments of 30% sequence identity or less[86–88] usually result in inaccurate models.

Because of the small number of experimentally determined structures of solute carriers and their homologs, they have been previously modeled based on alignments with low-sequence similarity to the templates, sometimes resulting in unreliable models. However, some models of solute carriers that were based on relatively low-sequence identity (as low as 10–15%) accurately captured mechanistic details that are critical for their functions.[7,10,23,89–93] For instance, the outward-facing conformation of vSGLT (bacterial homolog of SLC5) was modeled based on the LeuT structure (a bacterial homolog of SLC6) using an alignment at 11% sequence identity.[7] The model helped identify inverted structural repeats that are important for transport in both proteins[23] and revealed a possible Na^+ binding site in vSGLT based on the corresponding binding site in LeuT. Automatically derived models for all modelable solute carriers are shown in Supporting Information Figure S6 (http://modbase.compbio.ucsf.edu/projects/SLC/).

Identification of sequence conservation patterns across families and within them can potentially result in more accurate structural models of transporter/ligand complexes. First, detection and alignment of family-specific or cluster-specific motifs can improve template-target alignments, a crucial step in comparative modeling. Second, conservation of the transport mode and substrate type across proteins within clusters can help inform when it is appropriate to transfer function of specific residues from functionally characterized to uncharacterized solute carriers (Fig. 3). Third, the selection of ligand libraries for both high-throughput[94] and virtual screening[84] can be facilitated by considering known substrates of similarity network neighbors of the target solute carrier or family.

A glimpse into the evolution of solute carriers in eukaryotes: case study of the SLC25 family

Two proteins may exhibit similar structural and functional features without having similar sequences.[48–51] Such cases may arise because the proteins diverged from a common precursor a long time ago (divergent evolution) or because they evolved independently to have similar structures and functions without sharing a common precursor (convergent evolution).

The mitochondrial transporter family SLC25 is the most functionally diverse solute carrier family as its different members have a wide range of substrates, such as amino acids, nucleotides, coenzyme A, organic ions, and inorganic ions. The majority of the SLC25 family members are found in organisms spanning the whole eukaryotic kingdom [Fig. 4(A)], and they are not typically tissue-specific [Fig. 4(B)], which is in line with the universal need for mitochondrial energy processes. Furthermore, none of the SLC25 members were detectably related in sequence to other solute carriers, despite the use of a sensitive profile–profile alignment method and a variety of similarity cutoffs (Fig. 2; Supporting Information Figures S1–S2).

SLC25 is an ancient family that is found in both prokaryotes and unicellular eukaryotes, so it is not surprising that the SLC25 family has diverged from other solute carrier families sufficiently to prevent us from detecting sequence similarity between them. These changes likely include acquiring short motifs responsible for the various functions of the different family members, rationalizing the wide range of substrate specificity of this family. Interestingly, the SLC25 sequences are relatively highly conserved within the family (Fig. 2; Supporting Information Figures S1–S2; purple circles), suggesting that there are some restraints on their evolution despite their varied substrate specificities.

Some of the SLC25 members share functions with members from other solute carrier families, even though they are not detectably related in sequence. For example, the mitochondrial glutamate carrier 1 (SLC25A22) and the excitatory amino acid transporter 3 (SLC1A1) are glutamate transporters that do not share statistically significant sequence similarity (Fig. 2). Previous studies have shown that it is possible to change a transporter's charge specificity by mutating only two amino acid residues.[78] Therefore, it is plausible that SLC1A1 and SLC25A22 independently gained motifs responsible for their similar substrate specificity through convergent evolution.

Our comparison captures distant relationships between solute carriers

There are three methodological advantages of our approach over previous analyses. First, the profile–profile alignments used here are more sensitive for detecting similarities between proteins than the sequence–sequence alignments used in previous studies.[2,5] Second, evolutionary relationships were previously established relying only on phylogenetic trees.[5] Although phylogenetic tree construction is an extremely powerful tool for inferring evolutionary relationships between proteins, the trees do not show multiple links for a given node.[52] Therefore, distant relationships between proteins, often representing nontrivial functional links between families, may not have been revealed in previous analyses. Third, it tends to be easier to observe trends in functional attributes of a large set of proteins when they are displayed on similarity maps rather than phylogenetic trees.

Previously, four distinct clusters consisting of only 15 solute carrier families have been identified, whereas the remaining 32 solute carrier families were not connected to any of these four clusters.[5] Similarly, our map also links most of the 15 previously clustered families to each other. For example, the SLC2, SLC22, SLC16, SLC17, SLC18, and SLC37 families, which formed one cluster,[5] are also highly connected to each other in the similarity map [Fig. 2(B)]. However, in contrast to the previous finding that the cluster containing SLC7 and SLC12 and the cluster containing SLC32, SLC36, and SLC38 were distinct,[5] we found that both of these clusters are highly linked and also connected to the SLC5, SLC6, and SLC11 families. The links we identified are likely to be biologically meaningful as almost all transporters in this cluster are cotransporters that tend to transport chemically similar substrates (Fig. 3).

In summary, our comparison of the human solute carriers reveals new relationships between the individual families and identifies potential new members of this set. The classification scheme is likely to inform future modeling of the solute carrier structures, a prerequisite for describing their substrate specificities. Analysis of functional trends across the solute carrier similarity network provides guidance about which properties can be transferred across the solute carrier space through sequence similarity.

Materials and Methods

Solute carriers

We extracted sequences of solute carriers from UniProt,[53] NCBI,[54] and from the literature using keyword search. Our final set included 491 unique sequences. Because of different annotation procedures in the different databases, genes with the same name occasionally had slightly different sequences, often representing different isoforms. Sequences differing only in a single amino acid position were manually removed. The clustering results were not expected to depend on whether or not highly similar sequences are included; thus, we used all closely related sequences because only some of them may contain short functional motifs that may be useful in linking more divergent transporters (e.g., sugar binding motifs).

Constructing and visualizing sequence-based similarity maps

A sequence similarity network is made up of links corresponding to pairwise relationships that score better than a defined cutoff.[52] The graphs representing the similarity networks were visualized using Cytoscape 2.6.1.[55] We relied on the yFiles organic layout algorithm, which is partially based on the spring-embedded algorithm (http://www.cytoscape.org/manual/Cytoscape2_6Manual.html), and shows only node connectivity to illustrate relationships. Because of the correlation between sequence similarity within a family and the number of similarity relationships better than a cutoff, the algorithm partly captures relative distances between two nodes in two dimensions, even without including link weights[52]; sets of nodes that are highly interconnected tend to cluster closely in space. Similarity maps were constructed using significant hits corresponding to a variety of cutoffs (Fig. 2, Supporting Information Figures S1–S2).

Determining uric acid uptake in SLC22A12 (URAT1) and SLC2A9 (GLUT9) expressing cells

All cells were grown in a humidified incubator set at 37°C and 5% CO_2. Uptake of uric acid was determined using transiently transfected HEK-293T cells cultured in 24 well poly-D-lysine coated plates. Cells were transfected with pcDNA5 (Invitrogen, Carlsbad, CA) containing either the full-length reference cDNA for human URAT1 or SLC2A9 (either the long or short isoform) or lacking any additional cDNA (EV) using Lipofectamine 2000 (Invitrogen), following the manufacturer's standard protocol. Uptake experiments in URAT1 expressing cells lasted for 2 min with a ^{3}H-uric acid (Moravek Chemicals, Brea, CA) concentration of 25 μ*M* in chloride free uptake buffer (125 m*M* sodium gluconate, 4.8 m*M* potassium gluconate, 1.2 m*M* K_2PO_4, 1.3 m*M* calcium gluconate, 1.2 m*M* $MgSO_4$, and 25 m*M* HEPES, pH 7.0), whereas experiments in GLUT9 expressing cells lasted for 3 min with a radiolabeled uric acid concentration of 50 μ*M* in chloride free uptake buffer. All experiments were performed at 37°C, terminated using ice-cold buffer, and uptake levels were quantified using liquid scintillation counting. All uptake results were normalized to total protein in each well.

Clustering using the NFP

The NFP provides a complementary way to elucidate complex relationships within large superfamilies from sequence data (Apeltsin et al., submitted for publication). The NFP takes as input an all-by-all sequence similarity network based on a BLAST alignment.[27] It then groups the sequences into clusters using TribeMCL.[95] Subsequently, the individual clusters are connected by placing the best-fitting links between the most similar clusters. The final result is a protein cluster topology, which subdivides the plausible clusters into neighbors and non-neighbors. Neighboring clusters are more likely to share greater structural and functional similarities than non-neighbors. This clustering allows us to estimate the structural and functional properties of uncharacterized families from their better-characterized neighbors.

Annotating substrate type

An initial annotation of the substrate type for each solute carrier was made based on data in the Bioparadigms database,[1] followed by a PubMed literature search for individual transporters and transporter families. To enable visualization of functional similarities across different solute carrier families, relatively broad substrate type classes were used. For instance, bicarbonate transporters from the SLC4 family were grouped together with phosphate transporters from the SLC34 family under the common "small anion" label.

Annotating species conservation

Species orthologs of human solute carriers were extracted from the NCBI HomoloGene[96] database, Release 63. The database contains proteins from 20 completely sequenced eukaryotic genomes, clustered to reveal species orthologs, as well as within-species paralogs. Each solute carrier was ranked by the age of the oldest species in which an orthologous protein was found. For example, rank 1 corresponds to transporters found only in humans [red, Fig. 4(A)], and 28 corresponds to transporters already found in the earliest single-cell eukaryotes [blue, Fig. 4(A)].

Calculating tissue specificity

Expression levels for 275 of the solute carriers in 75 different tissues were obtained from GEO dataset GDS596.[97] The tissue specificity of each transporter's expression was determined using the Γ measure[61]:

$$\Gamma = \frac{\sum_{i=1}^{N}(1-\chi_i)}{N-1}. \tag{1}$$

where N is the number of examined tissues and x_i is the expression level of gene i in a certain tissue, normalized to the maximum expression across all tissues. When Γ equals 1 [blue; Fig. 4(B)], the protein is expressed at similar levels in all human tissues; and when it equals 0, it is expressed in only one tissue [red; Fig. 4(B)].

Finding unknown solute carriers

To identify new solute carriers, we extracted sequences of all human open reading frames (ORF) derived from the ENSEMBL[67] database. We then predicted integral membrane proteins containing six or more transmembrane α-helices using TMHMM2,[98] as described previously.[99] For each such sequence, we built a multiple sequence profile by scanning against UniProt,[53] using MODELLER.[100] Specifically, the profile.build command implements the Smith-Waterman local alignment method for scanning a single query sequence against a database of sequences. Scores and significance values were calculated similarly to the procedure described by Pearson[101] and as already applied to membrane proteins (Kelly et al., submitted for publication). We scanned these profiles against profiles of known solute carrier sequences (including different isoforms), which were extracted from NCBI[54] and UniProt.[53] The initial list of predicted solute carriers corresponded to human proteins with at least six transmembrane helices whose profile–profile alignment with at least one known solute carrier has at least 25% sequence identity or an E-value lower than 1.

Overcoming limitations in homology-based annotation

While transferring function based on sequence similarity alone has proven to be beneficial, it is not always accurate, due to several reasons: (1) similar genes within the same organism (paralogs), which have resulted from gene duplication events, are frequently functionally divergent, on average much more so than the most similar homologous genes across multiple organisms (orthologs)[62]; (2) the hit or query may perform multiple functions (moonlighting proteins[63]); and (3) the hit may be a multidomain protein whose functional domain is not aligned to the query.[64] Therefore, to refine our search for homologs, we also evaluated (i) whether or not the candidate homologs contain family signatures such as conserved motifs predicted by MEME[70] and (ii) whether or not the candidate homologs cluster with other family members, based on multiple sequence alignment by MUSCLE[102] and phylogenetic trees by JalView [Fig. 5(G)].[103] Finally, we verified experimentally a high-confidence prediction; we tested the functional overlap between SLC2 and SLC22 families by measuring the rate of uric acid uptake in HEK cells expressing human URAT1 (SLC22A12) or GLUT9 (SLC2A9) [Fig. 5(H); Supporting Information Figure S4].

Visualizing similarity between SLC2 and SLC22 families

To locate the conservation patterns across the members of the SLC2 and SLC22 families, multiple sequence alignments were calculated using the MUSCLE web server.[102] Alignment scores were calculated using JalView 2.4.[104] The membrane topologies of selected members from each family were predicted using a majority vote consensus approach, based on individual predictions of membrane spanning domains by the TMHMM2[98] and TOPCONS (http://topcons.cbr.su.se/) web servers. The topology images were color-coded according to alignment scores, either from multiple alignments including all SLC2 and SLC22 family members, or from separate alignments for each family.

Acknowledgments

The authors thank Ron Kaback, Patricia Babbitt, Amber Dahlin, Holly Atkinson, Jeremy Phillips, and

Keren Lasker for helpful suggestions, as well as Ben Webb, Elina Tjioe, and Narayanan Eswar for technical assistance and maintenance of the computational resources required for this study.

References

1. Hediger MA, Romero MF, Peng JB, Rolfs A, Takanaga H, Bruford EA (2004) The ABCs of solute carriers: physiological, pathological and therapeutic implications of human membrane transport proteins—Introduction. Pflugers Arch 447:465–468.
2. Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H (2001) The HUGO Gene Nomenclature Committee (HGNC). Hum Genet 109:678–680.
3. Saier MH, Jr (2000) A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol Mol Biol Rev 64:354–411.
4. Saier MH, Jr, Yen MR, Noto K, Tamang DG, Elkan C (2009) The Transporter Classification Database: recent advances. Nucl Acids Res 37:D274–D278.
5. Fredriksson R, Nordstrom KJ, Stephansson O, Hagglund MG, Schioth HB (2008) The solute carrier (SLC) complement of the human genome: phylogenetic classification reveals four major families. FEBS Lett 582:3811–3816.
6. Pebay-Peyroula E, Dahout-Gonzalez C, Kahn R, Trezeguet V, Lauquin GJ, Brandolin G (2003) Structure of mitochondrial ADP/ATP carrier in complex with carboxyatractyloside. Nature 426:39–44.
7. Faham S, Watanabe A, Besserer GM, Cascio D, Specht A, Hirayama BA, Wright EM, Abramson J (2008) The crystal structure of a sodium galactose transporter reveals mechanistic insights into Na^+/sugar symport. Science 321:810–814.
8. Yamashita A, Singh SK, Kawate T, Jin Y, Gouaux E (2005) Crystal structure of a bacterial homologue of Na^+/Cl—dependent neurotransmitter transporters. Nature 437:215–223.
9. Gether U, Andersen PH, Larsson OM, Schousboe A (2006) Neurotransmitter transporters: molecular function of important drug targets. Trends Pharmacol Sci 27:375–383.
10. Krishnamurthy H, Piscitelli CL, Gouaux E (2009) Unlocking the molecular secrets of sodium-coupled transporters. Nature 459:347–355.
11. Gao X, Lu F, Zhou L, Dang S, Sun L, Li X, Wang J, Shi Y (2009) Structure and mechanism of an amino acid antiporter. Science 324:1565–1568.
12. Fang Y, Jayaram H, Shane T, Kolmakova-Partensky L, Wu F, Williams C, Xiong Y, Miller C (2009) Structure of a prokaryotic virtual proton pump at 3.2 A resolution. Nature 460:1040–1043.
13. Shaffer PL, Goehring A, Shankaranarayanan A, Gouaux E (2009) Structure and mechanism of a Na^+-independent amino acid transporter. Science 325: 1010–1014.
14. Ressl S, Terwisscha van Scheltinga AC, Vonrhein C, Ott V, Ziegler C (2009) Molecular basis of transport and regulation in the Na(+)/betaine symporter BetP. Nature 458:47–52.
15. Yernool D, Boudker O, Jin Y, Gouaux E (2004) Structure of a glutamate transporter homologue from *Pyrococcus horikoshii*. Nature 431:811–818.
16. Huang Y, Lemieux MJ, Song J, Auer M, Wang DN (2003) Structure and mechanism of the glycerol-3-phosphate transporter from *Escherichia coli*. Science 301:616–620.
17. Mirza O, Guan L, Verner G, Iwata S, Kaback HR (2006) Structural evidence for induced fit and a mechanism for sugar/H^+ symport in LacY. EMBO J 25: 1177–1183.
18. Yin Y, He X, Szewczyk P, Nguyen T, Chang G (2006) Structure of the multidrug transporter EmrD from *Escherichia coli*. Science 312:741–744.
19. Abramson J, Smirnova I, Kasho V, Verner G, Kaback HR, Iwata S (2003) Structure and mechanism of the lactose permease of *Escherichia coli*. Science 301: 610–615.
20. Abramson J, Iwata S, Kaback HR (2004) Lactose permease as a paradigm for membrane transport proteins (Review). Mol Membr Biol 21:227–236.
21. Guan L, Kaback HR (2006) Lessons from lactose permease. Annu Rev Biophys Biomol Struct 35:67–91.
22. Karpowich NK, Wang DN (2008) Structural biology: symmetric transporters for asymmetric transport. Science 321:781–782.
23. Forrest LR, Zhang YW, Jacobs MT, Gesmonde J, Xie L, Honig BH, Rudnick G (2008) Mechanism for alternating access in neurotransmitter transporters. Proc Natl Acad Sci USA 105:10338–10343.
24. Weyand S, Shimamura T, Yajima S, Suzuki S, Mirza O, Krusong K, Carpenter EP, Rutherford NG, Hadden JM, O'Reilly J, Ma P, Saidijam M, Patching SG, Hope RJ, Norbertczak HT, Roach PC, Iwata S, Henderson PJ, Cameron AD (2008) Structure and molecular mechanism of a nucleobase-cation-symport-1 family transporter. Science 322:709–713.
25. Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI (2006) OPM: orientations of proteins in membranes database. Bioinformatics 22:623–625.
26. Madhusudhan MS, Webb BM, Marti-Renom MA, Eswar N, Sali A (2009) Alignment of multiple protein structures based on sequence and structure features. Protein Eng Des Sel 22:569–574.
27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 25:3389–3402.
28. Abramson J, Wright EM (2009) Structure and function of Na(+)-symporters with inverted repeats. Curr Opin Struct Biol 19:425–432.
29. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540.
30. Blakely RD, DeFelice LJ (2007) All aglow about presynaptic receptor regulation of neurotransmitter transporters. Mol Pharmacol 71:1206–1208.
31. Chen NH, Reith ME, Quick MW (2004) Synaptic uptake and beyond: the sodium- and chloride-dependent neurotransmitter transporter family SLC6. Pflugers Arch 447:519–531.
32. Manji HK, Drevets WC, Charney DS (2001) The cellular neurobiology of depression. Nat Med 7:541–547.
33. Nestler EJ, Barrot M, DiLeone RJ, Eisch AJ, Gold SJ, Monteggia LM (2002) Neurobiology of depression. Neuron 34:13–25.
34. Leabman MK, Huang CC, DeYoung J, Carlson EJ, Taylor TR, de la Cruz M, Johns SJ, Stryke D, Kawamoto M, Urban TJ, Kroetz DL, Ferrin TE, Clark AG, Risch N, Herskowitz I, Giacomini KM; Pharmacogenetics of Membrane Transporters Investigators (2003) Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. Proc Natl Acad Sci USA 100:5896–5901.

35. Koepsell H, Schmitt BM, Gorboulev V (2003) Organic cation transporters. Rev Physiol Biochem Pharmacol 150:36–90.
36. Shu Y, Brown C, Castro RA, Shi RJ, Lin ET, Owen RP, Sheardown SA, Yue L, Burchard EG, Brett CM, Giacomini KM (2008) Effect of genetic variation in the organic cation transporter 1, OCT1, on metformin pharmacokinetics. Clin Pharmacol Ther 83:273–280.
37. Shu Y, Sheardown SA, Brown C, Owen RP, Zhang S, Castro RA, Ianculescu AG, Yue L, Lo JC, Burchard EG, Brett CM, Giacomini KM (2007) Effect of genetic variation in the organic cation transporter 1 (OCT1) on metformin action. J Clin Invest 117:1422–1431.
38. Shu Y, Leabman MK, Feng B, Mangravite LM, Huang CC, Stryke D, Kawamoto M, Johns SJ, DeYoung J, Carlson E, Ferrin TE, Herskowitz I, Giacomini KM; Pharmacogenetics of Membrane Transporters Investigators (2003) Evolutionary conservation predicts function of variants of the human organic cation transporter, OCT1. Proc Natl Acad Sci USA 100: 5902–5907.
39. Dresser MJ, Gray AT, Giacomini KM (2000) Kinetic and selectivity differences between rodent, rabbit, and human organic cation transporters (OCT1). J Pharmacol Exp Ther 292:1146–1152.
40. Hilgemann DW, Lu CC (1999) GAT1 (GABA:Na$^+$:Cl$^-$) cotransport function. Database reconstruction with an alternating access model. J Gen Physiol 114:459–475.
41. Jardetzky O (1966) Simple allosteric model for membrane pumps. Nature 211:969–970.
42. Singh SK, Piscitelli CL, Yamashita A, Gouaux E (2008) A competitive inhibitor traps LeuT in an open-to-out conformation. Science 322:1655–1661.
43. Andersen J, Kristensen AS, Bang-Andersen B, Stromgaard K (2009) Recent advances in the understanding of the interaction of antidepressant drugs with serotonin and norepinephrine transporters. Chem Commun (Camb) 25:3677–3692.
44. Singh SK, Yamashita A, Gouaux E (2007) Antidepressant binding site in a bacterial homologue of neurotransmitter transporters. Nature 448:952–956.
45. Zhou Z, Zhen J, Karpowich NK, Goetz RM, Law CJ, Reith ME, Wang DN (2007) LeuT-desipramine structure reveals how antidepressants block neurotransmitter reuptake. Science 317:1390–1393.
46. Shi L, Quick M, Zhao Y, Weinstein H, Javitch JA (2008) The mechanism of a neurotransmitter:sodium symporter—inward release of Na$^+$ and substrate is triggered by substrate in a second binding site. Mol Cell 30:667–677.
47. Quick M, Winther AM, Shi L, Nissen P, Weinstein H, Javitch JA (2009) Binding of an octylglucoside detergent molecule in the second substrate (S2) site of LeuT establishes an inhibitor-bound conformation. Proc Natl Acad Sci USA 106:5563–5568.
48. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823–826.
49. Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68.
50. Brenner SE, Chothia C, Hubbard TJ (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci USA 95:6073–6078.
51. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12:85–94.
52. Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PLoS ONE 4:e4345.
53. The UniProt Consortium (2008) The universal protein resource (UniProt). Nucl Acids Res 36:D190–D195.
54. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2010) GenBank. Nucl Acids Res D46–51.
55. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504.
56. Palacin M, Kanai Y (2004) The ancillary proteins of HATs: SLC3 family of amino acid transporters. Pflugers Arch 447:490–494.
57. Manolescu AR, Augustin R, Moley K, Cheeseman C (2007) A highly conserved hydrophobic motif in the exofacial vestibule of fructose transporting SLC2A proteins acts as a critical determinant of their substrate selectivity. Mol Membr Biol 24:455–463.
58. Vitart V, Rudan I, Hayward C, Gray NK, Floyd J, Palmer CN, Knott SA, Kolcic I, Polasek O, Graessler J, et al. (2008) SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. Nat Genet 40:437–442.
59. Caulfield MJ, Munroe PB, O'Neill D, Witkowska K, Charchar FJ, Doblado M, Evans S, Eyheramendy S, Onipinla A, Howard P, Shaw-Hawkins S, Dobson RJ, Wallace C, Newhouse SJ, Brown M, Connell JM, Dominiczak A, Farrall M, Lathrop GM, Samani NJ, Kumari M, Marmot M, Brunner E, Chambers J, Elliott P, Kooner J, Laan M, Org E, Veldre G, Viigimaa M, Cappuccio FP, Ji C, Iacone R, Strazzullo P, Moley KH, Cheeseman C (2008) SLC2A9 is a high-capacity urate transporter in humans. PLoS Med 5: e197.
60. Uldry M, Thorens B (2004) The SLC2 family of facilitated hexose and polyol transporters. Pflugers Arch 447:480–489.
61. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, Lancet D, Shmueli O (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21:650–659.
62. Theissen G (2002) Secret life of genes. Nature 415: 741.
63. Jeffery CJ (2003) Moonlighting proteins: old proteins learning new tricks. Trends Genet 19:415–417.
64. Hegyi H, Gerstein M (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. Genome Res 11:1632–1640.
65. Ofran Y, Punta M, Schneider R, Rost B (2005) Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. Drug Discov Today 10:1475–1482.
66. Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318:595–608.
67. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G,

Smith J, Searle S, Flicek P (2009) Ensembl 2009. Nucl Acids Res 37:D690–D697.

68. Krogh A (2008) What are artificial neural networks? Nat Biotechnol 26:195–197.
69. Jacobsson JA, Haitina T, Lindblom J, Fredriksson R (2007) Identification of six putative human transporters with structural similarity to the drug transporter SLC22 family. Genomics 90:595–609.
70. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucl Acids Res 34:W369–W373.
71. Bashton M, Chothia C (2007) The generation of new protein functions by the combination of domains. Structure 15:85–99.
72. Kasho VN, Smirnova IN, Kaback HR (2006) Sequence alignment and homology threading reveals prokaryotic and eukaryotic proteins similar to lactose permease. J Mol Biol 358:1060–1070.
73. Vadyvaloo V, Smirnova IN, Kasho VN, Kaback HR (2006) Conservation of residues involved in sugar/H(+) symport by the sucrose permease of *Escherichia coli* relative to lactose permease. J Mol Biol 358: 1051–1059.
74. Sorgen PL, Hu Y, Guan L, Kaback HR, Girvin ME (2002) An approach to membrane protein structure without crystals. Proc Natl Acad Sci USA 99: 14037–14040.
75. Eisen JA (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res 8:163–167.
76. Forrest LR, Tang CL, Honig B (2006) On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. Biophys J 91: 508–517.
77. Li M, Hays FA, Roe-Zurz Z, Vuong L, Kelly L, Ho CM, Robbins RM, Pieper U, O'Connell JD, III, Miercke LJ, Giacomini KM, Sali A, Stroud RM (2009) Selecting optimum eukaryotic integral membrane proteins for structure determination by rapid expression and solubilization screening. J Mol Biol 385:820–830.
78. Punta M, Forrest LR, Bigelow H, Kernytsky A, Liu J, Rost B (2007) Membrane protein prediction methods. Methods 41:460–474.
79. Kelly L, Karchin R, Sali A (2007) Protein interactions and disease phenotypes in the ABC transporter superfamily. Pac Symp Biocomput 51–63.
80. Feng B, Dresser MJ, Shu Y, Johns SJ, Giacomini KM (2001) Arginine 454 and lysine 370 are essential for the anion specificity of the organic anion transporter, rOAT3. Biochemistry 40:5511–5520.
81. Zhang S, Lovejoy KS, Shima JE, Lagpacan LL, Shu Y, Lapuk A, Chen Y, Komori T, Gray JW, Chen X, Lippard SJ, Giacomini KM (2006) Organic cation transporters are determinants of oxaliplatin cytotoxicity. Cancer Res 66:8847–8857.
82. Ahlin G, Karlsson J, Pedersen JM, Gustavsson L, Larsson R, Matsson P, Norinder U, Bergstrom CA, Artursson P (2008) Structural requirements for drug inhibition of the liver specific human organic cation transport protein 1. J Med Chem 51:5932–5942.
83. Blundell TL, Sibanda BL, Montalvao RW, Brewerton S, Chelliah V, Worth CL, Harmer NJ, Davies O, Burke D (2006) Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. Philos Trans R Soc Lond B Biol Sci 361:413–423.
84. Shoichet BK (2004) Virtual screening of chemical libraries. Nature 432:862–865.
85. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov 3:935–949.
86. Sanchez R, Sali A (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. Proc Natl Acad Sci USA 95:13597–13602.
87. Koehl P, Levitt M (1999) A brighter future for protein structure prediction. Nat Struct Biol 6:108–111.
88. Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294:93–96.
89. Landau M, Herz K, Padan E, Ben-Tal N (2007) Model structure of the Na^+/H^+ exchanger 1 (NHE1): functional and clinical implications. J Biol Chem 282: 37854–37863.
90. Forrest LR, Tavoulari S, Zhang YW, Rudnick G, Honig B (2007) Identification of a chloride ion binding site in Na^+/Cl^- dependent transporters. Proc Natl Acad Sci USA 104:12761–12766.
91. Celik L, Sinning S, Severinsen K, Hansen CG, Moller MS, Bols M, Wiborg O, Schiott B (2008) Binding of serotonin to the human serotonin transporter. Molecular modeling and experimental validation. J Am Chem Soc 130:3853–3865.
92. Beuming T, Kniazeff J, Bergmann ML, Shi L, Gracia L, Raniszewska K, Newman AH, Javitch JA, Weinstein H, Gether U, Loland CJ (2008) The binding sites for cocaine and dopamine in the dopamine transporter overlap. Nat Neurosci 11:780–789.
93. Beuming T, Shi L, Javitch JA, Weinstein H (2006) A comprehensive structure-based alignment of prokaryotic and eukaryotic neurotransmitter/Na^+ symporters (NSS) aids in the use of the LeuT structure to probe NSS structure and function. Mol Pharmacol 70: 1630–1642.
94. Macarron R (2006) Critical review of the role of HTS in drug discovery. Drug Discov Today 11:277–279.
95. Enright AJ, van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucl Acids Res 30:1575–1584.
96. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA (2001) Database resources of the National Center for Biotechnology Information. Nucl Acids Res 29:11–16.
97. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA 101: 6062–6067.
98. Sonnhammer EL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6:175–182.
99. Kelly L, Pieper U, Eswar N, Hays FA, Li M, Roe-Zurz Z, Kroetz DL, Giacomini KM, Stroud RM, Sali A (2009) A survey of integral alpha-helical membrane proteins. J Struct Funct Genomics 10:269–280.
100. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815.
101. Pearson WR (1998) Empirical statistical estimates for sequence similarity searches. J Mol Biol 276:71–84.
102. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl Acids Res 32:1792–1797.
103. Retief JD (2000) Phylogenetic analysis using PHYLIP. Methods Mol Biol 132:243–258.

104. Clamp M, Cuff J, Searle SM, Barton GJ (2004) The Jalview Java alignment editor. Bioinformatics 20: 426–427.

105. Eswar N, John B, Mirkovic N, Fiser A, Ilyin VA, Pieper U, Stuart AC, Marti-Renom MA, Madhusudhan MS, Yerkovich B, Sali A (2003) Tools for comparative protein structure modeling and analysis. Nucleic Acids Res 31:3375–3380.

106. Pieper U, Eswar N, Webb BM, Eramian D, Kelly L, Barkan DT, Carter H, Mankoo P, Karchin R, Marti-Renom MA, Davis FP, Sali A (2009) MODBASE, a database of annotated comparative protein structure models and associated resources. Nucl Acids Res 37:D347–D354.

107. Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15:2507–2524.