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
Introduction

Solute carriers
Solute carriers are eukaryotic membrane proteins that control the uptake and efflux of various solutes, including amino acids, sugars, and drugs.\textsuperscript{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).\textsuperscript{1–4}

Sequence-based classification of solute carriers
The Gene Nomenclature Committee (HGNC) of the Human Genome Organization (HUGO)\textsuperscript{2} classifies ~400 human solute carriers into 47 families.\textsuperscript{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.\textsuperscript{1} At least initially, atomic structures of transporters were apparently not considered in constructing this classification. Several other classifications have also been proposed.\textsuperscript{3,5} The transporter classification (TC) system aims to automatically classify transporters across all organisms (http://www.tcdb.org/).\textsuperscript{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).\textsuperscript{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 consist of 15 related families\textsuperscript{1} and 32 additional unlinked families.\textsuperscript{5}

Structures of solute carriers and their homologs
Solute carriers are typically composed of one large domain consisting of 10–14 transmembrane \textalpha-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),\textsuperscript{6} which belongs to the SLC25 family. In addition, there are a number of structures for prokaryotic homologs of solute carriers.\textsuperscript{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),\textsuperscript{22} SLC6 (LeuT),\textsuperscript{23} SLC7 (Arginine/ammonium antiporter (AdiC)),\textsuperscript{11,12} and ApcT transporter\textsuperscript{13} families as well as the bacterial transporter NCS1\textsuperscript{24} and the Na+/betaine symporter BetP\textsuperscript{14} are classified into the same structural family in the orientations of proteins in membranes (OPM) database (Fig. 1)\textsuperscript{25} and share similar transport mechanisms.\textsuperscript{10,28} In contrast, some solute carrier families have different folds (Fig. 1). For example, the SLC1 member Glp from \textit{Pyrococcus horikoshii} has the proton glutamate symport fold,\textsuperscript{15} whereas the SLC6 member LeuT from \textit{Aquifex aeolicus} \textit{v}f5 adopts the SNF-like fold.\textsuperscript{5} 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.\textsuperscript{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.\textsuperscript{20,21} Many antidepressants selectively inhibit the SLC6A4 (SERT) serotonin transporter, which is responsible for serotonin reuptake from synaptic spaces.\textsuperscript{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.\textsuperscript{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).\textsuperscript{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 families,\textsuperscript{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\textsuperscript{10,20–23,28,40} suggest that solute carriers from several structural families transport solutes using the “alternating
access” mechanism, in which the transporter alternatively exposes its binding site to either side of the membrane. 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. 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; occluded states are conformations that solute carriers from various structural families adopt during the transport process.

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. 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. Additionally, crystal structures of LeuT with several antidepressants proposed a noncompetitive
inhibition mechanism for transport by LeuT that involves a second inhibitory binding site.\textsuperscript{44,45} It was suggested that this site binds another substrate molecule,\textsuperscript{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,\textsuperscript{8,47} acting as a noncompetitive inhibitor.\textsuperscript{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.\textsuperscript{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

\textbf{Similarity maps confirm and refine current classification}

We illustrate relationships in protein families using sequence similarity networks.\textsuperscript{52} The nodes in the graphs represent sequences of the known solute carriers, color coded by their family as annotated by Uniprot\textsuperscript{53} (or Genbank\textsuperscript{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\textsuperscript{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. 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.

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 SLC711–13 and SLC68 families (Fig. 1); (ii) the separation between the SLC25, SLC37, SLC42, and SLC1 families is consistent with the structural dissimilarity among the cow SLC25A4 mitochondrial ADP/ATP carrier, the bacterial SLC37 homologs,16–19 the human SLC42A3 RhCG ammonium transporter, and the archaean SLC1 homolog25 (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 members56) [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. 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-
linked even for stringent cutoff values (E-value ≤ 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 (e.g., SLC8), cotransporters (yellow), or exchangers (blue). Furthermore, clusters of sequence-related families also share the mechanism of transport. Most proteins from the cluster containing the SLC5, SLC6, SLC7, SLC11, SLC12, SLC14, SLC2, and SLC22 families are cotransporters. 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
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, we calculated a measure of tissue specificity ranging from 0 to 1 [see Materials and Methods section; Eq. (1)]. 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.
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, 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 database during the preparation of this manuscript, splice variants of known SLC22 transporters, and proteins that were recently annotated manually as SLC22 transporters [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 [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...
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. 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. 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. 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).

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. 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. 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, (iii) verified the maps by known solute carrier structures (Fig. 1), and (iv) clustered the sequences using

<table>
<thead>
<tr>
<th>Family</th>
<th>SwissProt/UniProt</th>
<th>Gene name</th>
<th>Protein name</th>
<th>Family function</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLC6</td>
<td>A6NF70_HUMAN</td>
<td>—</td>
<td>Transporter</td>
<td>The sodium- and chloride-dependent neurotransmitter transporter family</td>
</tr>
<tr>
<td>SLC9</td>
<td>NHDC1_HUMAN</td>
<td>NHEDC1</td>
<td>Sodium/hydrogen exchanger-like domain-containing protein 1</td>
<td>The Na+/H+ exchanger family</td>
</tr>
<tr>
<td>SLC22</td>
<td>S22AX_HUMAN</td>
<td>Putative solute carrier family</td>
<td>Putative transporter SVOPL</td>
<td>The organic cation/anion/zwitterion transporter family</td>
</tr>
<tr>
<td>SVOPL</td>
<td>SVOPL</td>
<td>—</td>
<td>Putative transporter SVOPL</td>
<td></td>
</tr>
</tbody>
</table>

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

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

*Gene name marks the name of the gene according to the “gene names” field in the corresponding SwissProt/UniProt entry.*

*Protein name marks the name of the protein according to the “protein names” field in the corresponding SwissProt/UniProt entry.*

*Family function marks the function of the predicted solute carrier family based on the HGNC classification described in Ref. 1.*
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, because hub structures can serve as templates for computational modeling of many other proteins with similar sequences (Kelly et al., submitted for publication). 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, 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).

**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 (ENSG00000182157) and SVOPL_HUMAN] (black rectangles). A previous analysis of the SLC22 family also suggested that SVOPL might be an uncharacterized member.

**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. 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. 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\textsuperscript{36–38} 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.\textsuperscript{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.\textsuperscript{7} The model helped identify inverted structural repeats that are important for transport in both proteins\textsuperscript{23} and revealed a possible Na\textsuperscript{+} 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\textsuperscript{84} and virtual screening\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{2,5} Second, evolutionary relationships were previously established relying only on phylogenetic trees.\textsuperscript{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.\textsuperscript{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. 2B). 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, NCBI, 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 weights52; 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% CO2. Uptake of uric acid was determined using transiently transfected HEK-293T cells cultured in 24 well poly-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 3H-uric acid (Moravek Chemicals, Brea, CA) concentration of 25 μM in chloride free uptake buffer (125 mM sodium gluconate, 4.8 mM potassium gluconate, 1.2 mM K2PO4, 1.3 mM calcium gluconate, 1.2 mM MgSO4, and 25 mM HEPES, pH 7.0), whereas experiments in GLUT9 expressing cells lasted for 3 min with a radiolabeled uric acid concentration of 50 μM 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.
An initial annotation of the substrate type for each solute carrier was made based on data in the Bioparadigms database, 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.

Species orthologs of human solute carriers were extracted from the NCBI HomoloGene 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)].

Expression levels for 275 of the solute carriers in 75 different tissues were obtained from GEO dataset GDS596. The tissue specificity of each transporter's expression was determined using the \( \Gamma \) measure:

\[
\Gamma = \frac{\sum_{i=1}^{N} (1 - x_i)}{N - 1},
\]

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 \( \Gamma \) 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)].

To identify new solute carriers, we extracted sequences of all human open reading frames (ORF) derived from the ENSEMBL database. We then predicted integral membrane proteins containing six or more transmembrane \( \alpha \)-helices using TMHMM2, as described previously. For each such sequence, we built a multiple sequence profile by scanning against UniProt using MODELLER. 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 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 and UniProt. 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.

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); (2) the hit or query may perform multiple functions (moonlighting proteins); and (3) the hit may be a multidomain protein whose functional domain is not aligned to the query. 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 and (ii) whether or not the candidate homologs cluster with other family members, based on multiple sequence alignment by MUSCLE and phylogenetic trees by JalView.

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].

To locate the conservation patterns across the members of the SLC2 and SLC22 families, multiple sequence alignments were calculated using the MUSCLE web server. Alignment scores were calculated using JalView. 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 TMHMM and TOPCONS 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.

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

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 Tjoe, and Narayanan Eswar for technical assistance and maintenance of the computational resources required for this study.

References


