Special Issue *"Database"*

Vol. 15, pp. 94–103 (2018)
doi: 10.2142/biophysico.15.0_94

Biophysics and Physicobiology

https://www.jstage.jst.go.jp/browse/biophysico/

Regular Article

iMusta4SLC: Database for the structural property and variations of solute carrier transporters

Akiko Higuchi[1], Naoki Nonaka[2] and Kei Yura[3,4,5]

[1] *Graduate School of Frontier Sciences, The University of Tokyo, Minato-ku, Tokyo 108-8639, Japan*
[2] *Graduate School of Engineering, The University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan*
[3] *Graduate School of Humanities and Sciences, Ochanomizu University, Bunkyo-ku, Tokyo 112-8610, Japan*
[4] *Center for Informational Biology, Ochanomizu University, Bunkyo-ku, Tokyo 112-8610, Japan*
[5] *School of Advanced Science and Engineering, Waseda University, Shinjuku-ku, Tokyo 169-0072, Japan*

Received October 25, 2017; accepted January 17, 2018

Membrane transporter proteins play important roles in transport of nutrients into the cell, in transport of waste out of the cell, in maintenance of homeostasis, and in signal transduction. Solute carrier (SLC) transporter is the superfamily, which has the largest number of genes (>400 in humans) in membrane transporter and consists of 52 families. SLC transporters carry a wide variety of substrates such as amino acids, peptides, saccharides, ions, neurotransmitters, lipids, hormones and related materials. Despite the apparent importance for the substrate transport, the information of sequence variation and three-dimensional structures have not been integrated to the level of providing new knowledge on the relationship to, for instance, diseases. We, therefore, built a new database named iMusta4SLC, which is available at http://cib.cf.ocha.ac.jp/slc/, that connected the data of structural properties and of pathogenic mutations on human SLC transporters. iMusta4SLC helps to investigate the structural features of pathogenic mutations on SLC transporters. With this database, we found that the mutations at the conserved arginine were frequently involved in diseases, and were located at a border between the membrane and the cytoplasm. Especially in SLC families 2 and 22, the conserved residues formed a large cluster at the border. In SLC2A1, one third of the reported pathogenic missense mutations were found in this conserved cluster.

Key words: arginine, disease, protein three-dimensional structure, sequence variation, solute carrier transporter

Membrane proteins are roughly divided into three main types, channels, ABC transporters and SLC transporters [1]. Channel proteins passively transport ions and water molecules by diffusion [2], and ABC transporters actively transport the substrates using the energy derived from ATP hydrolysis [3]. Despite numerous studies on these two types of membrane proteins, the study on SLC transporters has been fallen behind, due to its difficulty in three-dimensional structure analyses [4]. SLC transporter carries a small molecule along with conformation changes in the transporter without ATP hydrolysis. So far, about 400 types of SLC

Corresponding author: Kei Yura, School of Advanced Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan.
e-mail: yura.kei@aoni.waseda.jp

◀ ***Significance*** ▶

Solute Carrier (SLC) transporter is the largest superfamily in membrane transporters. SLC transporters have a key role in nutrient intake to the cell. Previous experimental and clinical studies have reported the relationship between variants found in human SLC genes and clinical phenotypes. However the atomic mechanisms of the variation in the amino acid sequence leading to the alteration of the structure, dynamics, function and phenotype are still unknown. We therefore built iMusta4SLC that integrated the known related data to assist the simultaneous analyses of variation data and protein structural properties on SLC transporters.

transporter genes have been identified, and classified into 52 families (SLC1–SLC52) based on the sequence similarity, their function, substrate similarity and the number of transmembrane helices (mostly 10 to 14) [5–7].

The type of ligands for each SLC transporter family extends to a broad range of small molecules such as ions, amino acids, peptides, lipids, neurotransmitters, hormones, drugs and so on. The selectivity of small molecules in each transporter plays an important role in delivering nutrition into the cells, in releasing unnecessary substances outside the cells and in maintaining homeostasis of the cells and tissues. SLC transporters specifically expressed on the blood brain barrier are the crucial suppliers of essential nutrients to the nervous system in the brain [8–10].

A number of studies on the relationship between SLC genes/proteins and diseases have been conducted, because dysfunction of SLC transporter is often related to neurological or metabolic conditions [11–15]. Missense mutations (variations that results in amino acid substitution) in glucose transporter genes (members of SLC2) are highly correlated with glucose transporter deficiency syndrome, dystonia and epilepsy in infants [16–18]. Insufficient supply of glucose to the brain in infancy due to the dysfunction of the SLC proteins is assumed to be the cause of these diseases. Some pathogenic mutations in carnitine transporter genes (members of SLC22) are known to have high correlation with the carnitine deficiency and are seemingly correlated with sudden infant death syndrome [19,20]. Carnitine transporters play a role in carrying the activated fatty acids across the inner mitochondrial membrane, energy production and removal of toxic oxidized fatty acids. Patients with carnitine transporter dysfunction suffer from metabolic cardiomyopathy, central nervous system abnormalities and gastrointestinal symptoms, such as recurring abdominal pain and diarrhea [19,20]. Despite these descriptions on the relationship between the dysfunction and the diseases, there have been a limited number of descriptions of atomic mechanisms that elucidate the cause of the diseases. A crystal structure of lactose permease (a homolog of SLC37) from *Escherichia coli* and that of leucine transporter (a homolog of SLC6) from *Aquifex aeolicus,* were determined [21], and a couple of models that explain the substrate-transport mechanisms have been proposed [22]. None of these models, however, had sufficient details to elucidate the relationship between the sequence variations and the effects of the variations on the function of SLC in humans.

The studies of the relationship between the variation on genes and the diseases in humans have been carried out for the last decades, and the results have been compiled in many databases including, but not limited to, ClinVar [23,24], dbSNP [25] and COSMIC [26,27]. These databases tell that not all variations on amino acid sequences have correlations with diseases and that variations found in SLC are no exception. This fact provided a new question on making a distinction between pathogenic and non-pathogenic variations. The answer to this question seems to be different on different proteins and can only be addressed by integrating data derived from different types of measurements.

ClinVar contains about 2,000 missense mutations in SLC transporters. According to the clinical significance stored with the mutation in ClinVar, over 500 cases of them were reported as ‘pathogenic’ or ‘risk factor,’ which means that the variation is highly correlated with certain diseases or symptoms. Therefore, the questions to be addressed should be the distinction between the pathogenic and non-pathogenic variations and the atomic mechanisms that the pathogenic variation leads to a certain level of dysfunction of the transporter.

To facilitate paths to the elucidation of atomic mechanisms, we built a database tool named iMusta4SLC, which stands for Integrated MUtational and STructural Analysis FOR SLC proteins**,** for integrated analyses of structural properties and variations on SLC proteins. Here, we demonstrate the database tool and some of the findings based on the database.

Materials and Methods

Sources of the original and derivative data

SLC family classification and gene names were taken from SLC Tables (http://slc.bioparadigms.org) built by BioParadigms (http://www.bioparadigms.org) [1] and by Human Gene Nomenclature Committee (HGNC) (https://www.genenames.org). Nucleotide and amino acid sequences of SLC families were obtained from NCBI [28]. The mutations and related disease data of SLC transporters were obtained from ClinVar [23,24] by searching the database with the keywords ‘Solute Carrier.’ Three-dimensional structure data were obtained from PDB [29] by searching the database with BLAST [30] for proteins with a similar sequence to SLC proteins.

Based on these original data, we predicted the membrane protein topology of SLC proteins using TopCons2 [31], which is a tool that combines the prediction results of Philius [32], PolyPhobius [33], SPOCTOPUS [34], OCTOPUS [35] and SCAMPI [36] by taking residue-wise consensus.

Pathogenic and non-pathogenic mutation data in SLC

ClinVar provides the clinical significance for each mutation. The terms ‘pathogenic’ and ‘risk factor’ are used when the variants affect clinical phenotypes, many of which are reported with specific diseases or symptoms. Thus, in this study, we defined the missense mutation annotated with the keyword either ‘pathogenic’ or ‘risk factor’ in the clinical significance column of ClinVar as ‘pathogenic mutation.’ Similarly, ‘non-pathogenic mutation’ was defined as the missense mutation annotated with the keyword ‘benign.’

SLC topology data

Membrane protein topology is originally a term representing the N-terminal and C-terminal positions for the membrane, and the number and position of transmembrane regions [37]. Several topology prediction methods have been developed so far. In this study, TopCons2, which calculates a consensus of several prediction algorithms, was used to predict the location of each residue in SLC transporter at outside, middle and inside of the membrane. Based on the results obtained from TopCons2, we newly defined five topology regions, R1–R5, as follows; R1, a residue in the protein segment located outside of the cell; R2, a residue in the segment within the range of ±two residues from the transmembrane helix terminus facing outside of the cell; R3, a residue in the segment of the transmembrane helix except the ones included in R2 or R4; R4, a residue in the segment within the range of ±two residues from the transmembrane helix terminus facing the cytoplasm; R5, a residue in the segment located in the cytoplasm.

Mutation and topology data integration

In order to investigate the relationship between the site with pathogenic mutation and the topology region, all the data above were integrated in the database. The frequency of the pathogenic mutations in the topology region R*i* ($i = 1, 2, \ldots, 5$) was calculated by,

$$f_{p,Ri} = C_{p,Ri} / \mathrm{C}_p,$$

where $C_{p,Ri}$ is the count of the pathogenic mutations in topology region R*i*, and C_p is the total number of pathogenic mutations. Similarly, the frequency of the non-pathogenic mutations in the topology region R*i* was calculated by,

$$f_{n,Ri} = C_{n,Ri} / \mathrm{C}_n.$$

The expected frequency of the mutations in topology region R*i* was calculated by,

$$f_{exp,Ri} = C_{res,Ri} / \mathrm{C}_{res},$$

where $\mathrm{C}_{res,Ri}$ is the total count of amino acid residues in topology region R*i*, and C_{res} is the total number of amino acid residues in SLC transporter sequences. The log odds ratio of the pathogenic mutations in topology region R*i* was then obtained by $\log_2 (f_{p,Ri}/f_{exp,Ri})$, and that of the non-pathogenic mutations by $\log_2 (f_{n,Ri}/f_{exp,Ri})$.

Sequence similarity visualization

All-against-all similarities among the SLC protein sequences were calculated by FASTA36 (v36.3.6) [38,39]. Similarity threshod was set at E-value = 1.0×10^{-5}. The set of the pairs of similar SLC proteins was visualized in a circular form using the Hierarchical Edge Bundling protocol in D3.js (https://d3js.org) [40].

Results and Discussion

Overview of the database

In iMusta4SLC, data on SLC transporter gene names, accession ID, locus, family classification, sequence similarity, three-dimensional structures, mutations found in SLC genes and associated diseases were stored. The SLC genes have been classified into 52 families according to the previous study [1] and HGNC. Total number of SLC transcripts including isoforms was 573. The SLC transcript is generally named in the form of SLC*n*X*m*, where *n* is for family number, X is for an alphabet representing subfamily and *m* is for a digit assigned to an individual gene, which was the standard nomenclature fixed by HGNC. When the SLC has a splice variant, the SLC transcript is named in the form SLC*n*X*mpl*, where *p* is for an alphabet representing a splice variant, and *l* is for a digit or alphabet assigned to each variant. Out of 52 families of SLC in humans, only four of them (SLC2, 4, 25 and 42) have at least one member of which the three-dimensional structure was registered in PDB, and 26 families have at least one member with sequence identity more than 20% to a protein in PDB.

The number of missense variations of SLC transporters found in ClinVar was 1,216. These variations in amino acid sequences were classified by clinical significance in iMusta4SLC. The counts and proportions in each clinical significance were as follows; Risk factor, 10 (0.8%); Pathogenic, 524 (44.4%); Likely pathogenic, 122 (10.3%); Benign, 64 (5.4%); Likely benign, 55 (4.7%); Uncertain significance, 309 (26.2%); not provided, 78 (6.6%); and Conflicting interpretations, 17 (1.4%). The data seem to have a bias toward pathogenic and missense mutations. In terms of diseases, pathogenic mutations in SLC genes cause at least 100 different types of diseases, including metabolic diseases, neurological diseases, malformations, deafness, developmental disorders, immunodeficiency, renal failure and cancer.

Database architecture

We built iMusta4SLC using a relational database management system and connected the annotations in original databases, such as sequence accession ID, gene locus, three-dimensional structures, mutations and diseases associated with the variations in human SLC transporters (Fig. 1A). The interface of the database was developed on Flask (http://flask.pocoo.org) and the various types of query can be cast through a web browser. The results of the search are displayed in a page-wise manner. In each page of this website, a search option can be set on the left column. The details of the options and items to be displayed are described in 'How to Use' page. In 'Family list' page, the family number, SLC transporter name, isoform, protein name, alias and the brief description of the family are listed. These data are based on SLC Tables [1] and HGNC as mentioned above. 'Function list' page provides the list of ligands, transport-coupling ion and a type of transport mechanism for each SLC transporter. In 'Sequence similarity' page, sequence similarities among all SLC transporter sequences calculated by FASTA36 program (v36.3.6) [38,39] are visualized. A pair of SLC trans-

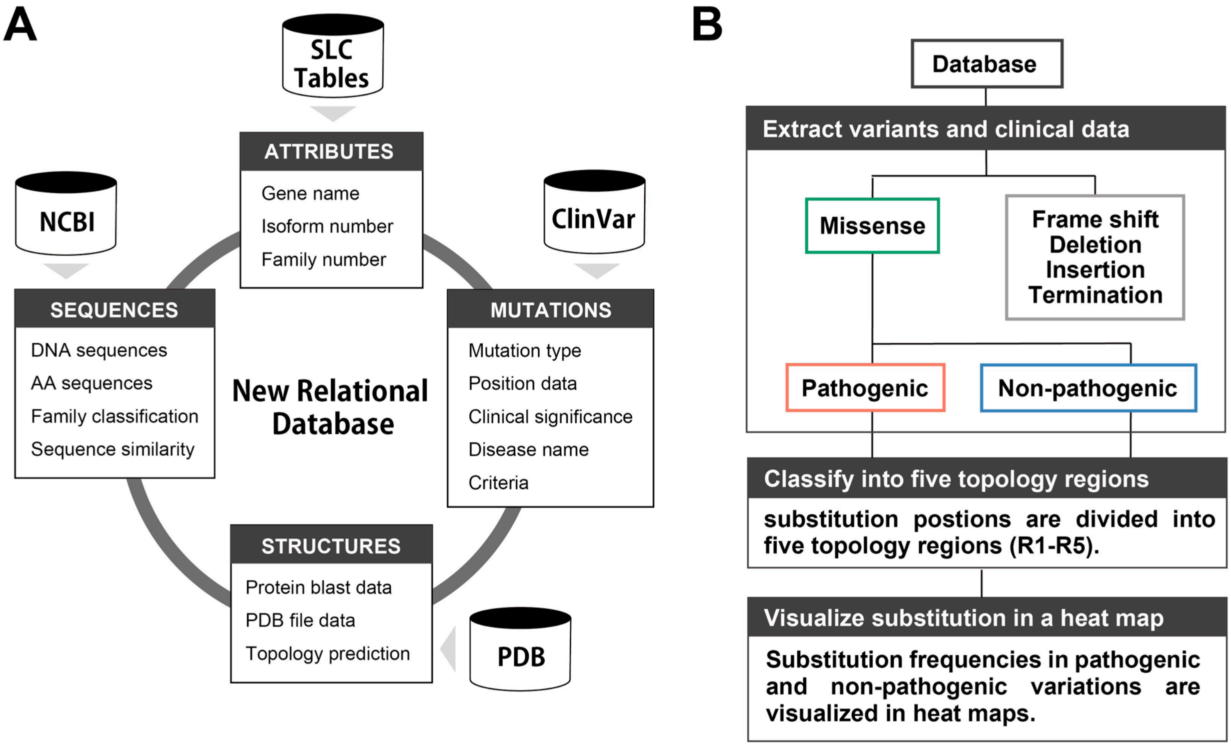


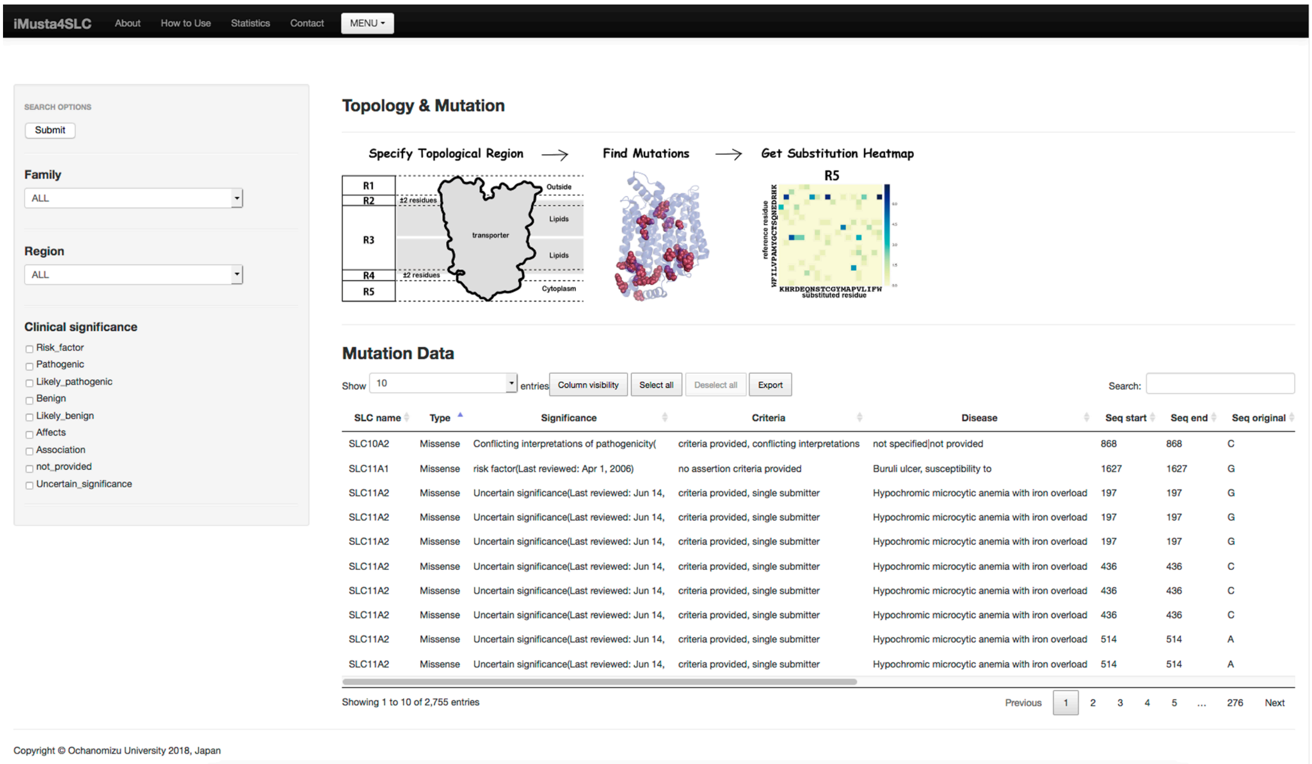
Figure 1 Database overview. A, Schematic view of the database structure. Data were collected from existing databases (SLC Tables [1], NCBI [28], ClinVar [23,24], PDB [29]). These data were linked using SLC gene ID as a key feature. B, The workflow for integrated analyses of mutations and topology region in iMusta4SLC.

porters with E-value less than the given threshold is depicted by a blue line connecting two points representing the two SLC transporters. The threshold can be set by a slide bar between 1.0×10^{-20} and 1.0×10^{-1}. A mouse-over to a name of SLC transporter turns color of the names and of the edges to the other SLC transporters in red, when the E-values to the others are less than the value set by the slide bar. In 'Sequence list' page, sequence accession ID, gene locus, protein sequence length, and the description of the sequence are listed. 'Template list' shows the results of BLAST search of each SLC protein against PDB. In 'Template coverage' page, the detail of 'Template list' is shown, namely the detail results of BLAST search of each SLC protein against PDB are schematically shown. The red box on the top of the figure represents the query sequence, and the boxes in blue stripes represent a sequence in PDB similar to the query sequence. The blue stripes have four different color modes, which represent the degree of matches between query sequence and the protein sequence in PDB. 'Mutation list' shows the list of all variants registered in iMusta4SLC. The displayed data can be narrowed down by conditions, such as family number, mutation type and clinical significance. In 'Topology & Mutation' page, integrated analysis between mutation site and the protein topology is conducted (Fig. 1B). The substitution patterns of amino acid residues in a specific topology region of each family of SLC are displayed in a heat map. The coloring in the heat map tells the frequency of the substitution. The specific substitution is tabulated below the heat map and can be downloaded in CSV or Excel format by pressing 'Export' button (Fig. 2).

Ratios of mutations at different topology regions

We first compared the abundance ratio of pathogenic and non-pathogenic mutation for each topology region, namely, R1-R5 (Fig. 3A). There are 545 pathogenic mutations and 130 non-pathogenic mutations in iMusta4SLC. The imbalance in the numbers of pathogenic and non-pathogenic mutations might have stemmed from the fact that pathogenic mutations have been paid much attention to. The numbers and proportions of all amino acid residues in each region were as follows; R1, 91,115 (28.4%); R2, 22,400 (7.0%); R3, 87,652 (27.3%); R4, 23,424 (7.3%); R5, 96,527 (30.1%). The counts and proportions in each region of pathogenic mutations were as follows; R1, 111 (20.4%); R2, 52 (9.5%); R3, 176 (32.3%); R4, 88 (16.1%); R5, 118 (21.7%). The counts and proportions of non-pathogenic mutations in each region were as follows; R1, 31 (23.8%); R2, 12 (9.2%); R3, 32 (24.6%); R4, 2 (1.5%); R5, 53 (40.8%).

The distribution of the different types of mutations was compared using the log odds ratio of these values (Fig. 3B). The figures evidently show that R4 has peculiar characteristics. Pathogenic mutations are highly localized to R4, and



Family: ALL

Region: ALL

Clinical significance

□ Risk_factor
□ Pathogenic
□ Likely_pathogenic
□ Benign
□ Likely_benign
□ Affects
□ Association
□ not_provided
□ Uncertain_significance

Mutation Data

Show 10 entries

Search:

SLC name	Type	Significance	Criteria	Disease	Seq start	Seq end	Seq original
SLC10A2	Missense	Conflicting interpretations of pathogenicity(	criteria provided, conflicting interpretations	not specified|not provided	868	868	C
SLC11A1	Missense	risk factor(Last reviewed: Apr 1, 2006)	no assertion criteria provided	Buruli ulcer, susceptibility to	1627	1627	G
SLC11A2	Missense	Uncertain significance(Last reviewed: Jun 14,	criteria provided, single submitter	Hypochromic microcytic anemia with iron overload	197	197	G
SLC11A2	Missense	Uncertain significance(Last reviewed: Jun 14,	criteria provided, single submitter	Hypochromic microcytic anemia with iron overload	197	197	G
SLC11A2	Missense	Uncertain significance(Last reviewed: Jun 14,	criteria provided, single submitter	Hypochromic microcytic anemia with iron overload	197	197	G
SLC11A2	Missense	Uncertain significance(Last reviewed: Jun 14,	criteria provided, single submitter	Hypochromic microcytic anemia with iron overload	436	436	C
SLC11A2	Missense	Uncertain significance(Last reviewed: Jun 14,	criteria provided, single submitter	Hypochromic microcytic anemia with iron overload	436	436	C
SLC11A2	Missense	Uncertain significance(Last reviewed: Jun 14,	criteria provided, single submitter	Hypochromic microcytic anemia with iron overload	436	436	C
SLC11A2	Missense	Uncertain significance(Last reviewed: Jun 14,	criteria provided, single submitter	Hypochromic microcytic anemia with iron overload	514	514	A
SLC11A2	Missense	Uncertain significance(Last reviewed: Jun 14,	criteria provided, single submitter	Hypochromic microcytic anemia with iron overload	514	514	A

Showing 1 to 10 of 2,755 entries

Copyright © Ochanomizu University 2018, Japan

Figure 2 Typical interface of iMusta4SLC. Users can select SLC family, topology region and the type of mutation on the left column. The raw data and the heat map of substitution frequency that match the conditions on the left column are displayed. The tabulated data on the bottom can be downloaded in CSV or Excel format.

non-pathogenic mutations are rare in R4, which also means that the variations in R4 are mostly pathogenic. This skewed distribution in R4 was statistically significant, tested by chi-square analysis ($p<2.04\times10^{-19}$ and 1.08×10^{-2}, respectively).

Heat map for amino acid substitution in each topology region

A pattern of amino acid substitution and its log odds value can be analyzed in ‘Topology & Mutation’ in iMusta4SLC as described in the section above. When the heat maps of pathogenic and non-pathogenic mutations in different topology regions were compared, peculiar characteristics can be found as shown in Figure 3C. On the heat map in the pathogenic mutations of R4, the substitution from arginine to tryptophan were highly observed (count: 11, ratio: 11.36%). Arginine to tryptophan substitution around the junction area between the membrane and cytosol may have significant impact on membrane protein folding and protein stability. For membrane protein folding, ‘positive inside rule’ has been proposed, which tells that the positively charged residues in the cytoplasmic region outnumber those in the extracellular region [41–44]. The imbalance in the number of positively charged residues is considered as a remnant of topology formation and a replacement of a charged residue with a neutral one may affect a protein topology formation [45,46]. Hence, the substitutions found in the pathogenic mutation likely affect the topology formation. The arginine in R4 also interacts with the phosphorus group of the lipid and a water molecule, both of which are polar molecules. Substitution of arginine with tryptophan deprives of a stable polar interaction between the side chain of arginine and the oxygen atoms of a phosphorus group or a water molecule, hence may ended in destabilizing SLC protein in the membrane. These perturbations in the protein structure are seemingly the initial step for disturbing the function of SLC transporter which finally leads to the diseases.

Conserved residues on SLC structures

The evolutionary relationship of each transporter was visualized based on the sequence similarity in ‘Sequence similarity’ page (Fig. 4A). We found that the well conserved residues were clustered on the surface of SLC protein. The surface conservation was visualized on the three-dimensional structure of SLC2. The crystal structures of one of the members were reported in two different forms, namely inward-open (a conformation with a cavity on the cytosolic side) (PDB ID: 5EQI) [47] and outward-open (a conformation with a cavity on the extracellular side) (PDB ID: 4ZW9) [48]. The multiple sequence alignment of SLC2 was conducted by MAFFT v7.305b [49, 50], and the result showed that the residues such as Gly, Arg, and Glu are highly conserved in SLC2 (Fig. 5). The topology region of SLC2A1

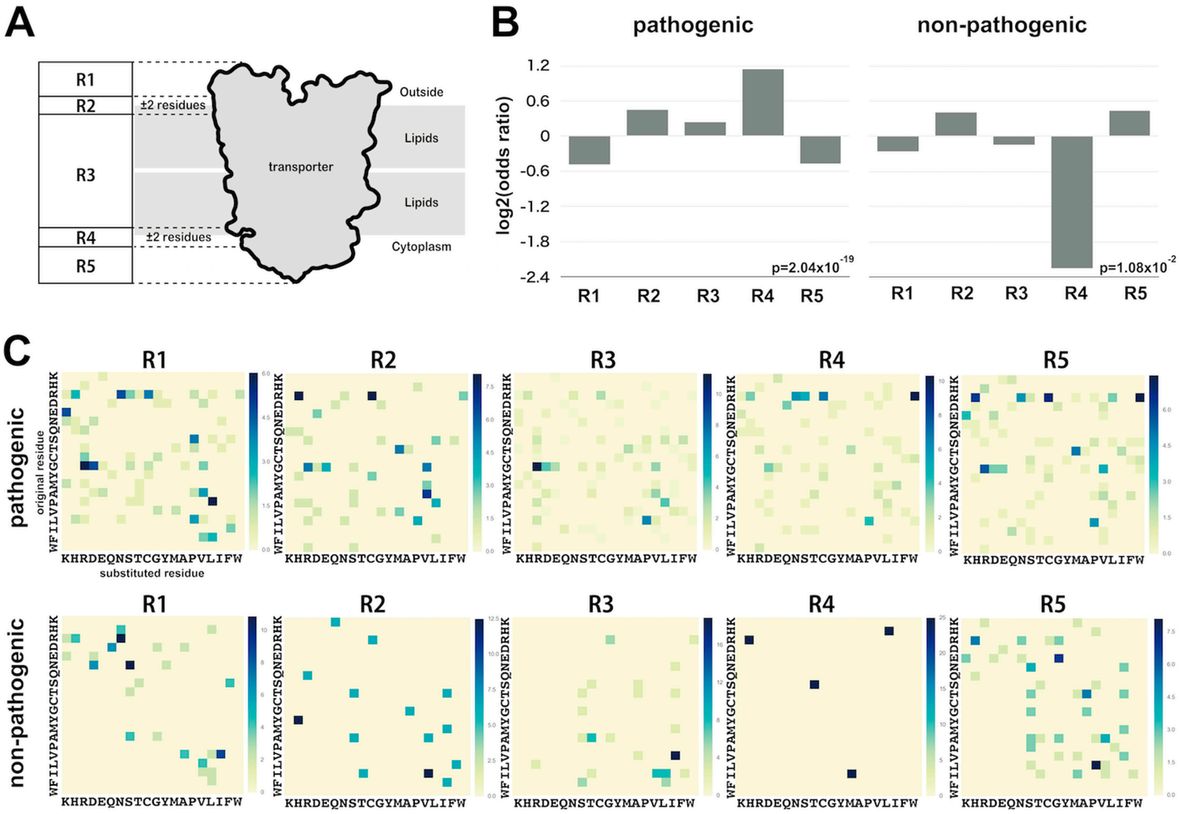


Figure 3 Integrated analysis for mutation and topology regions. A, See Materials and Methods for the definition of the topology region. B, Log odds ratio for mutations in each region. C, Heat map of amino acid substitution in each region. The horizontal axis of each map represents substituted residues and the vertical axis represents original residues.

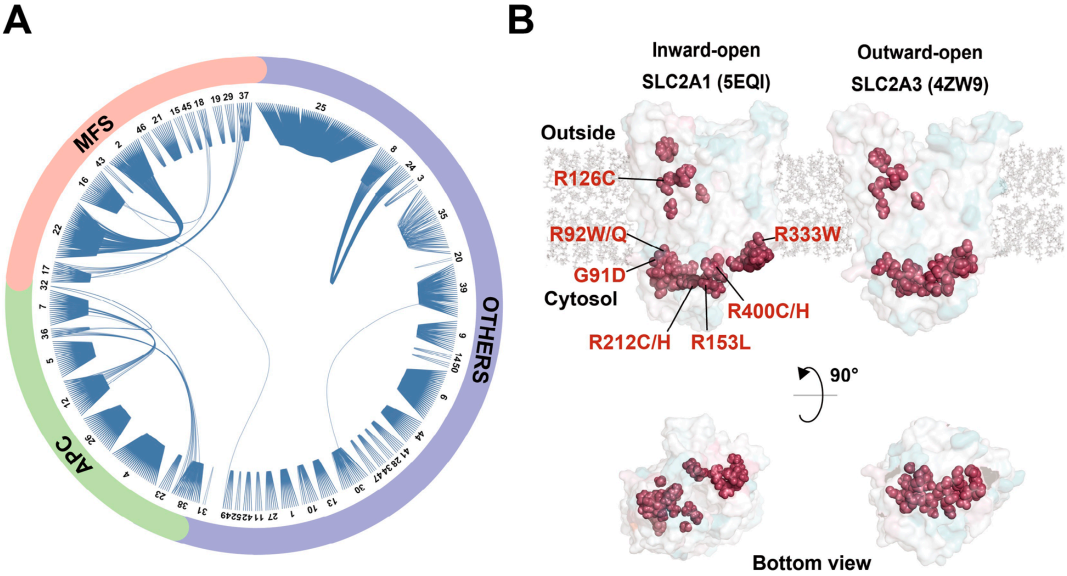


Figure 4 Conserved amino acid residues and their positions in the three-dimentional structure of SLC2 and SLC22. A, A circular map to depict sequence similarities in SLC transporters. SLC transporters were first grouped into three types, namely, SLC with Major Facilitator Superfamily (MFS) domain, SLC with Amino Acid-Polyamine-Organocation (APC) domain and other SLCs. Each SLC was then represented as a node on a circle, initially grouped by the type, then by the family. The SLCs with similar sequences were connected by a blue line. The similarity threshold was E-value less than 1.0×10^{-5} (data-size: 319,831 residues in 572 sequences) obtained by FASTA program. B, Highly conserved residues in SLC2 and SLC22. Red spheres are the conserved residue in SLC2 and 22 mapped on the three-dimensional structures of SLC2A1 (PDB ID: 5EQI, inward-open structure) and SLC2A3 (PDB ID: 4ZW9, outward-open structure).

is schematically illustrated on the top of the alignment in Figure 5. These conserved residues are found in or juxtaposing to R4. Highly conserved residue positions in SLC2 in addition to the ones in SLC22 were mapped onto the three-dimensional structures of SLC2, and most of the highly conserved residues were found at the interface between the transmembrane region and the cytoplasm (Fig. 4B). In addition, the conserved residues formed a single cluster,

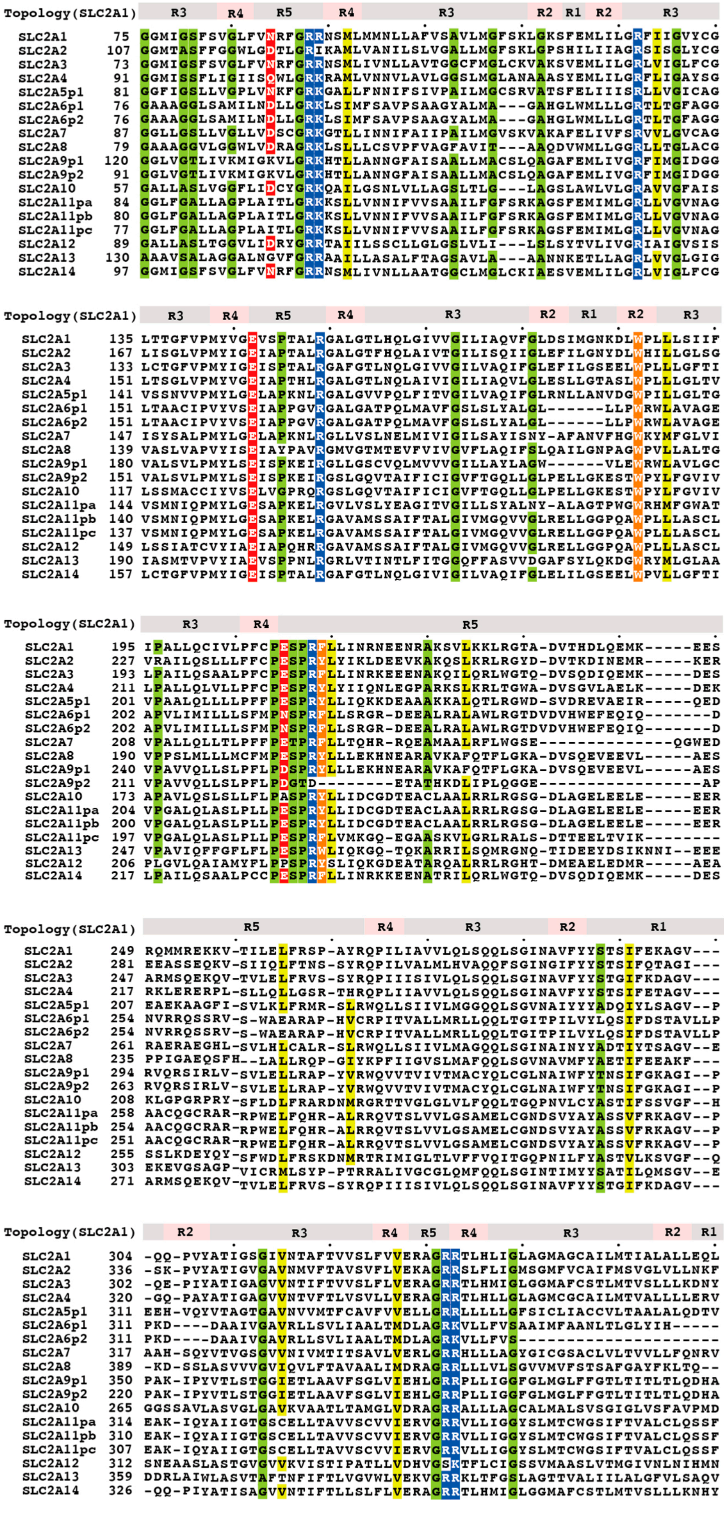


Figure 5 Sequence alignment of SLC2. Multiple sequence alignment of SLC2 was conducted by MAFFT v7.305b [49,50]. Topology region of SLC2A1 is shown on the top of the alignment.

especially on the outward-open structure. The cluster is divided into four clusters on the inward-open structure, forming a space in the center. This space is a putative pathway of a substrate.

The relationship between the positions of pathogenic mutation and conservation

In topology region R4, pathogenic mutations were found in high ratio and there exist a cluster of highly conserved residues at the border between transmembrane and cytoplasm. The specific examples of this correlation can also be found in iMusta4SLC. Table 1 describes all pathogenic mutations in a specific family of SLC, SLC2A1. It is intriguing to find that one third of these mutations, namely Gly91Asp, Arg92Trp/Gln, Arg126Cys, Arg153Leu, Arg212Cys/His, Arg333Trp and Arg400Cys/His are found in the conserved red spheres around topology region R4 shown in Figure 4B. In the members of SLC22, two amino acid residue positions, namely Arg169Trp/Gln and Arg399Trp/Gln, were reported to be correlated with carnitine transport disorder and they correspond to Arg92 and Arg333 in SLC2A1, respectively, which are listed in Table 1.

The function of these conserved residues is still unknown. One may assume that the residues should be involved in substrate binding or transportation, but there has been no supportive evidence for this hypothesis. The substrate-binding residues of human GLUT1 was predicted based on the alignment between human GLUT1 and its bacterial

Table 1 List of reported pathogenic missense mutations on SLC2A1 (Glucose transporter 1, GLUT1)

Amino acid substitution				Clinical phenotype		Genoms position			
No.	position	reference	substituted	significance	disease	chr.	location	variation ID	allele ID
1	34	Asn	Ile	Pathogenic	GLUT1 deficiency syndrome 2	1	42943239	16115	31154
2	91	Gly	Asp	Pathogenic	GLUT1 deficiency syndrome 1	1	42931049	16110	31149
3	92	Arg	Trp	Pathogenic	GLUT1 deficiency syndrome 2	1	42931047	16119	31158
4	93	Arg	Trp	Pathogenic	GLUT1 deficiency syndrome 2	1	42930865	16117	31156
5	95	Ser	Ile	Pathogenic	GLUT1 deficiency syndrome 2	1	42930858	16116	31155
6	126	Arg	Cys	Pathogenic	Dystonia 9, GLUT1 deficiency syndrome 1, GLUT1 deficiency syndrome 2, Epilepsy, idiopathic generalized	1	42930766	16118	31157
7	126	Arg	His	Pathogenic	GLUT1 deficiency syndrome 1, Epilepsy, idiopathic generalized	1	42930765	16111	31150
8	126	Arg	Leu	Pathogenic	GLUT1 deficiency syndrome 1	1	42930765	16109	31148
9	130	Gly	Arg	Pathogenic	not provided	1	42930754	207190	201162
10	140	Val	Met	Pathogenic	not provided	1	42930724	372507	359361
11	153	Arg	Leu	Pathogenic	not provided	1	42930684	207227	201160
12	165	Val	Ile	Pathogenic	not provided	1	42930649	379258	365285
13	212	Arg	Cys	Pathogenic	Dystonia 9	1	42929918	37300	45874
14	212	Arg	His	Pathogenic	not provided	1	42929917	265386	259669
15	223	Arg	Trp	Pathogenic	not provided	1	42929885	207193	201157
16	223	Arg	Pro	risk factor	Epilepsy, idiopathic generalized	1	42929884	39702	48301
17	232	Arg	Cys	Pathogenic, risk factor	Epilepsy, idiopathic generalized	1	42929766	37301	45875
18	275	Ala	Thr	Pathogenic	GLUT1 deficiency syndrome 1, GLUT1 deficiency syndrome 2	1	42929637	16114	31153
19	283	Gln	Arg	Pathogenic	not provided	1	42929612	207197	201152
20	286	Gly	Asp	Pathogenic	Stomatin-deficient cryohydrocytosis with neurologic defects	1	42929603	218333	215041
21	295	Thr	Met	Pathogenic	GLUT1 deficiency syndrome 1	1	42929298	207229	201150
22	313	Ser	Pro	Pathogenic	not provided	1	42929245	280423	264006
23	314	Gly	Ser	Pathogenic	GLUT1 deficiency syndrome 1, GLUT1 deficiency syndrome 2, Epilepsy, idiopathic generalized	1	42929242	16113	31152
24	324	Ser	Leu	Pathogenic	not provided	1	42929211	207201	201147
25	330	Arg	Ter	Pathogenic	not provided	1	42929018	207196	201142
26	333	Arg	Trp	Pathogenic	GLUT1 deficiency syndrome 1, Epilepsy, idiopathic generalized	1	42929009	198842	196002
27	400	Arg	His	Pathogenic	not provided	1	42927684	280046	264004
28	400	Arg	Cys	Pathogenic	not provided	1	42927685	207212	201132
29	411	Asn	Ser	risk factor	Epilepsy, idiopathic generalized	1	42927651	96709	102598
30	458	Arg	Trp	risk factor	Epilepsy, idiopathic generalized	1	42927148	96708	102597
31	468	Arg	Trp	Pathogenic	GLUT1 deficiency syndrome 1	1	42927118	16120	31159

homolog XylE of which the substrate-binding residues had been identified [1]. None of these residues were included in the conserved cluster. In addition, the inhibitor-binding sites of human GLUT1 has been known [47,48], but no residues were included in the cluster.

Conclusion

iMusta4SLC assists a search for the relationship among diseases, mutation sites and structural properties on SLC transporters. With iMusta4SLC, we found that the mutations of several conserved arginines, particularly to tryptophan in R4 region, were frequently involved in diseases. Most of the highly conserved residues in SLC2 form a large cluster on the cytoplasmic side. These residues are apparently involved in conformational change and ligand transport. iMusta4SLC can help to construct a hypothesis on the atomic mechanisms of how amino acid residue variation affects the function of SLC transporter.

Acknowledgement

This research is supported by Platform Project for Supporting Drug Discovery and Life Science Research (Basis for Supporting Innovative Drug Discovery and Life Science Research (BINDS)) from Japan Agency for Medical Research and Development (AMED) under Grant Number JP17am 0101065.

Conflicts of Interest

A. H., N. N. and K. Y. declare that they have no conflict of interest.

Author Contribution

A. H. and K. Y. directed the entire project and A. H., N. N. and K. Y. co-wrote the manuscript. A. H. designed and N. N. constructed the database and the interface. K. Y. prepared the web platform.

References

[1] Hediger, M. A., Clémençon, B., Burrier, R. E. & Bruford, E. A. The ABCs of membrane transporters in health and disease (SLC series): Introduction. *Mol. Aspects Med.* **34**, 95–107 (2013).

[2] Gouaux, E. & Mackinnon, R. Principles of selective ion transport in channels and pumps. *Science* **310**, 1461–1465 (2005).

[3] Tanaka, K. J., Song, S., Mason, K. & Pinkett, H. W. Selective substrate uptake: The role of ATP-binding cassette (ABC) importers in pathogenesis. *Biochim. Biophys. Acta* **1860**, 868–877 (2018).

[4] Schlessinger, A., Khuri, N., Giacomini, K. M. & Sali, A. Molecular modeling and ligand docking for Solute Carrier (SLC) transporters. *Curr. Top. Med. Chem.* **13**, 843–856 (2013).

[5] Diallinas, G. Understanding transporter specificity and the discrete appearance of channel-like gating domains in transporters. *Front. Pharmacol.* **5**, 207 (2014).

[6] Fredriksson, R., Nordström, K. J., Stephansson, O., Hägglund, M. G. & Schiöth, H. B. The solute carrier (SLC) complement of the human genome: Phylogenetic classification reveals four major families. *FEBS Lett.* **582**, 3811–3816 (2008).

[7] Schlessinger, A., Yee, S. W., Sali, A. & Giacomini, K. M. SLC classification: An update. *Clin. Pharmacol. Ther.* **94**, 19–23 (2013).

[8] Nałęcz, K. A. Solute carriers in the blood-brain barrier: Safety in abundance. *Neurochem. Res.* **42**, 795–809 (2017).

[9] Morris, M. E., Rodriguez-Cruz, V. & Felmlee, M. A. SLC and ABC transporters: expression, localization, and species differences at the blood-brain and the blood-cerebrospinal fluid barriers. *AAPS J.* **19**, 1317–1331 (2017).

[10] Suhy, A. M., Webb, A., Papp, A. C., Geier, E. G. & Sadee, W. Expression and splicing of ABC and SLC transporters in the human blood-brain barrier measured with RNAseq. *Eur. J. Pharm. Sci.* **103**, 47–51 (2017).

[11] Benarroch, E. E. Glutamate transporters: diversity, function, and involvement in neurologic disease. *Neurology* **74**, 259–264 (2010).

[12] Palmieri, F. The mitochondrial transporter family SLC25: Identification, properties and physiopathology. *Mol. Aspects Med.* **34**, 465–484 (2013).

[13] Bröer, S. & Palacín, M. The role of amino acid transporters in inherited and acquired diseases. *Biochem. J.* **436**, 193–211 (2011).

[14] Pramod, A. B., Foster, J., Carvelli, L. & Henry, L. K. SLC6 transporters: structure, function, regulation, disease association and therapeutics. *Mol. Aspects Med.* **34**, 197–219 (2013).

[15] Wright, E. M., Hirayama, B. A. & Loo, D. F. Active sugar transport in health and disease. *J. Intern. Med.* **261**, 32–43 (2007).

[16] Wang, D., Kranz-Eble, P. & De Vivo, D. C. Mutational analysis of GLUT1 (SLC2A1) in Glut-1 deficiency syndrome. *Hum. Mutat*. **16**, 224–231 (2000).

[17] Jiang, X., McDermott, J. R., Ajees, A. A., Rosen, B. P. & Liu, Z. Trivalent arsenicals and glucose use different translocation pathways in mammalian GLUT1. *Metallomics* **2**, 211–219 (2010).

[18] Wong, H. Y., Law, P. Y. & Ho, Y. Y. Disease-associated Glut1 single amino acid substitute mutations S66F, R126C, and T295M constitute Glut1-deficiency states in vitro. *Mol. Genet. Metab*. **90**, 193–198 (2007).

[19] Nezu, J., Tamai, I., Oku, A., Ohashi, R., Yabuuchi, H., Hashimoto, N., *et al*. Primary systemic carnitine deficiency is caused by mutations in a gene encoding sodium ion-dependent carnitine transporter. *Nat. Genet.* **21**, 91–94 (1999).

[20] Sharma, S. & Black, S. M. Carnitine homeostasis, mitochondrial function, and cardiovascular disease. *Drug Discov. Today Dis. Mech.* **6**, e31–e39 (2009).

[21] Wang, H., Elferich, J. & Gouaux, E. Structures of LeuT in bicelles define conformation and substrate binding in a membrane-like context. *Nat. Struct. Mol. Biol.* **19**, 212–219 (2012).

[22] Kazmier, K., Claxton, D. P. & Mchaourab, H. S. Alternating access mechanisms of LeuT-fold transporters: trailblazing towards the promised energy landscapes. *Curr. Opin. Struct. Biol.* **45**, 100–108 (2017).

[23] Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., *et al*. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).

[24] Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., *et al*. ClinVar: public archive of rela-

tionships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).

[25] Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., *et al*. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. **29**, 308–311 (2001).

[26] Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate J., *et al*. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).

[27] Forbes, S. A., Beare, D., Bindal, N., Bamford, S., Ward, S., Cole, C. G., *et al*. COSMIC: High-resolution cancer genetics using the catalogue of somatic mutations in cancer. *Curr. Protoc. Hum. Genet.* **91**, 10.11.1–10.11.37 (2016).

[28] Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., *et al*. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res*. **43**, D36–D42 (2015).

[29] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., *et al*. The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542 (1977).

[30] Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S. & Madden, T. L. NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36**, W5–W9 (2008).

[31] Tsirigos, K. D., Peters, C., Shu, N., Käll, L. & Elofsson, A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.* **43**, W401–W407 (2015).

[32] Reynolds, S. M., Kall, L., Riffle, M. E., Bilmes, J. A. & Noble, W. S. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS. Comput. Biol*. **4**, e1000213 (2008).

[33] Kall, L., Krogh, A. & Sonnhammer, E. L. An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* **21**, i251–i257 (2005).

[34] Viklund, H., Bernsel, A., Skwark, M. & Elofsson, A. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics* **24**, 2928–2929 (2008).

[35] Viklund, H. & Elofsson, A. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* **24**, 1662–1668 (2008).

[36] Bernsel, A., Viklund, H., Falk, J., Lindahl, E., von Heijne, G. & Elofsson, A. Prediction of membrane-protein topology from first principles. *Proc. Natl. Acad. Sci. USA* **105**, 7177–7181 (2008).

[37] Melén, K., Krogh, A. & von Heijne, G. Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.* **327**, 735–744 (2003).

[38] Pearson, W. R. Finding Protein and Nucleotide Similarities with FASTA. *Curr. Protoc. Bioinformatics* **53**, 3.9.1–3.9.25 (2016).

[39] Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448 (1988).

[40] Bostock, M., Ogievetsky, V. & Heer, J. D³ Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.* **17**, 2301–2309 (2011).

[41] Baeza-Delgado, C., Marti-Renom, M. A. & Mingarro, I. Structure-based statistical analysis of transmembrane helices. *Eur. Biophys. J.* **42**, 199–207 (2013).

[42] Heijne, G. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the transmembrane topology. *EMBO J.* **5**, 3021–3027 (1986).

[43] Heijne, G. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* **225**, 487–494 (1992).

[44] Baeza-Delgado, C., Marti-Renom, M. A. & Mingarro, I. Structure-based statistical analysis of transmembrane helices. *Eur. Biophys. J.* **42**, 199–207 (2013).

[45] Bogdanov, M., Xie, J. & Dowhan, W. Lipid-protein interactions drive membrane protein topogenesis in accordance with the positive inside rule. *J. Biol. Chem.* **284**, 9637–9641 (2009).

[46] Bogdanov, M., Dowhan, W. & Vitrac, H. Lipids and topological rules governing membrane protein assembly. *Biochim. Biophys. Acta* **1843**, 1475–1488 (2014).

[47] Kapoor, K., Finer-Moore, J. S., Pedersen, B. P., Caboni, L., Waight, A., Hillig, R. C., *et al*. Mechanism of inhibition of human glucose transporter GLUT1 is conserved between cytochalasin B and phenylalanine amides. *Proc. Natl. Acad. Sci. USA* **113**, 4711–4716 (2016).

[48] Deng, D., Sun, P., Yan, C., Ke, M., Jiang, X., Xiong, L., *et al*. Molecular basis of ligand recognition and transport by glucose transporters. *Nature* **526**, 391–396 (2015).

[49] Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. **30**, 3059–3066 (2002).

[50] Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol*. **30**, 772–780 (2013).