User Guide

The Utility of the Database

SeQuery provides two navigation interfaces for exploring the GPCR network, as illustrated in Figure 1. We recommend users to use the Chrome browser for better viewing quality. At its homepage (http://cluster.phy.ntnu.edu.tw), users can select the GPCR2841 dataset and explore the GPCR network from the top down (I). Alternatively, users can also submit a query sequence and explore its role in the GPCR network from the bottom up (II).

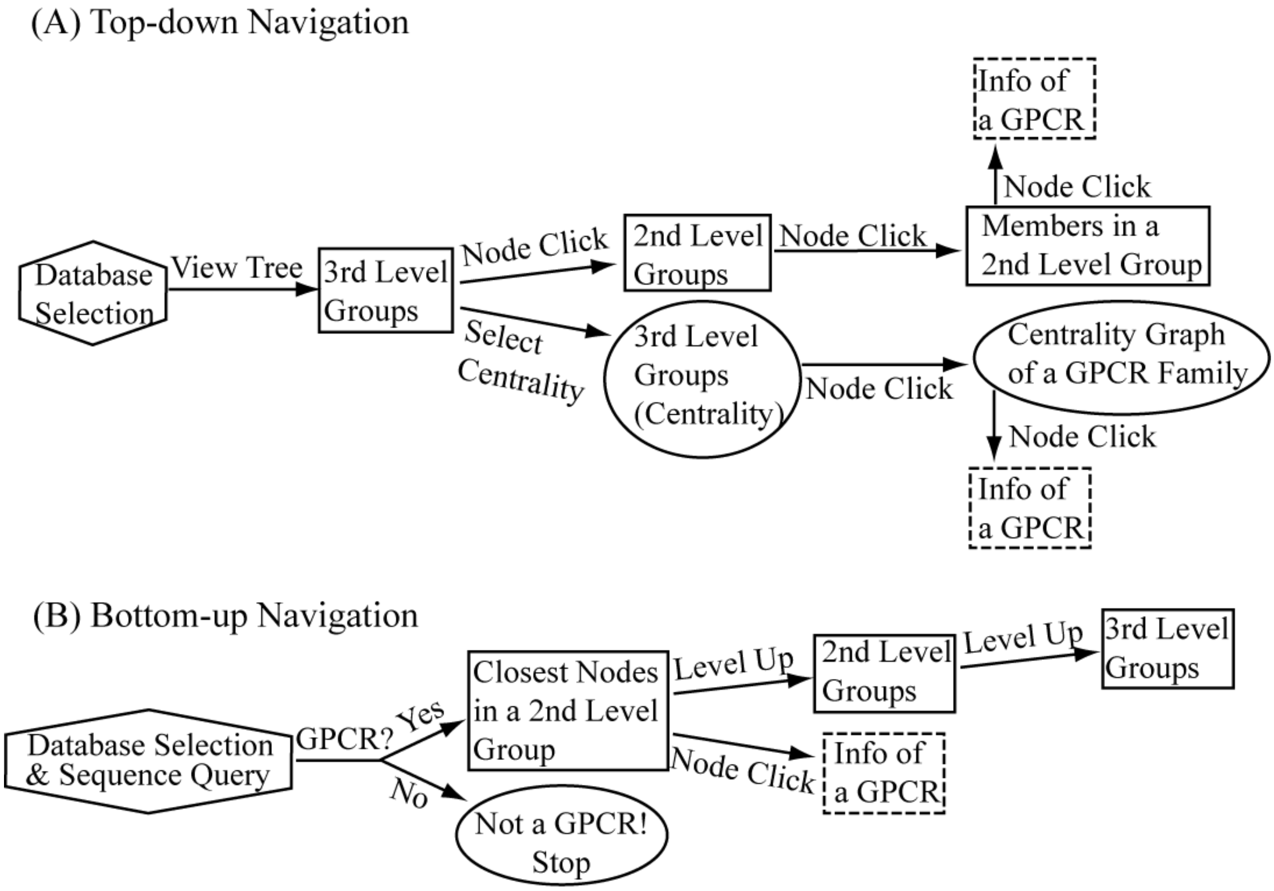
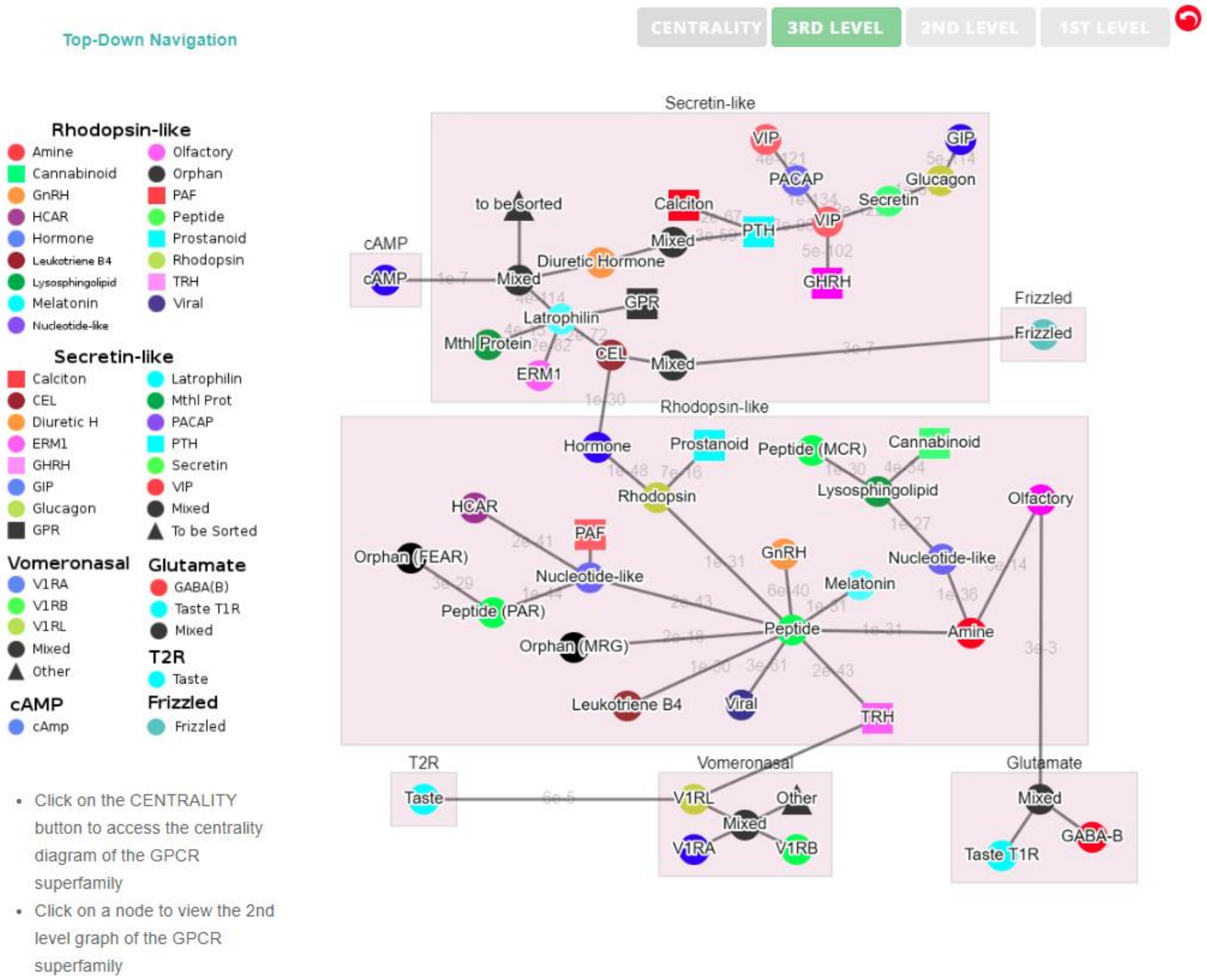


Fig. 1 Flowchart of the two navigation interfaces of the SeQuery database, including the top-down navigation scheme (I) and the bottom-up navigation scheme (II).

(I) In the top-down navigation scheme, SeQuery first displays the minimum spanning tree (MST) graph of the GPCR superfamily at the third level of clustering, which contains seven GPCR classes as shown in Figure 2. Each node in the network graph represents a GPCR family. The same scheme of colors and shapes as shown in the legend is used to denote the functions of GPCR clusters for all levels of clustering. When users click on a node, as shown in Figure 3, SeQuery shows the MST graph of the superfamily at the 2nd level clustering, centering at constituent clusters of the selected GPCR family. At this second level of clustering, the GPCR superfamily is represented by a minimum spanning tree of receptor clusters predicted from the first-level MSC clustering and each node in the graph represents a first-level MSC cluster. Users can explore the 2nd level graph of the network by dragging and zooming the network or locate any GPCR family of interest by clicking on the node in the legend at the left-hand side (SeQuery will center the second level graph at the clusters of the selected family). By clicking on a node in the 2nd level graph, SeQuery shows constituent sequences of the selected cluster in a 1st level graph. Each node in a 1st level graph is a GPCR sequence, whose detailed information can be viewed in an information box by clicking on the node. At the first level of clustering, receptor clusters are represented by a tree graph which may or may not contain a core of zero-distance sequences, as shown in Figure 4. If a sequence has a known protein structure, such as P51681 in Figure 4(A), its structure will be displayed on the node. When the user clicks on a sequence node, SeQuery displays its sequence and function information, as shown in Figure 5.



- Click on the CENTRALITY button to access the centrality diagram of the GPCR superfamily
- Click on a node to view the 2nd level graph of the GPCR superfamily

Fig. 2 The third-level minimum spanning tree diagram of the GPCR network with the base dataset. Each node represents a GPCR family. The legend shows the scheme of nodes' colors and shapes that are used to distinguish GPCR functions annotated in GPCRdb (also labeled on the nodes).

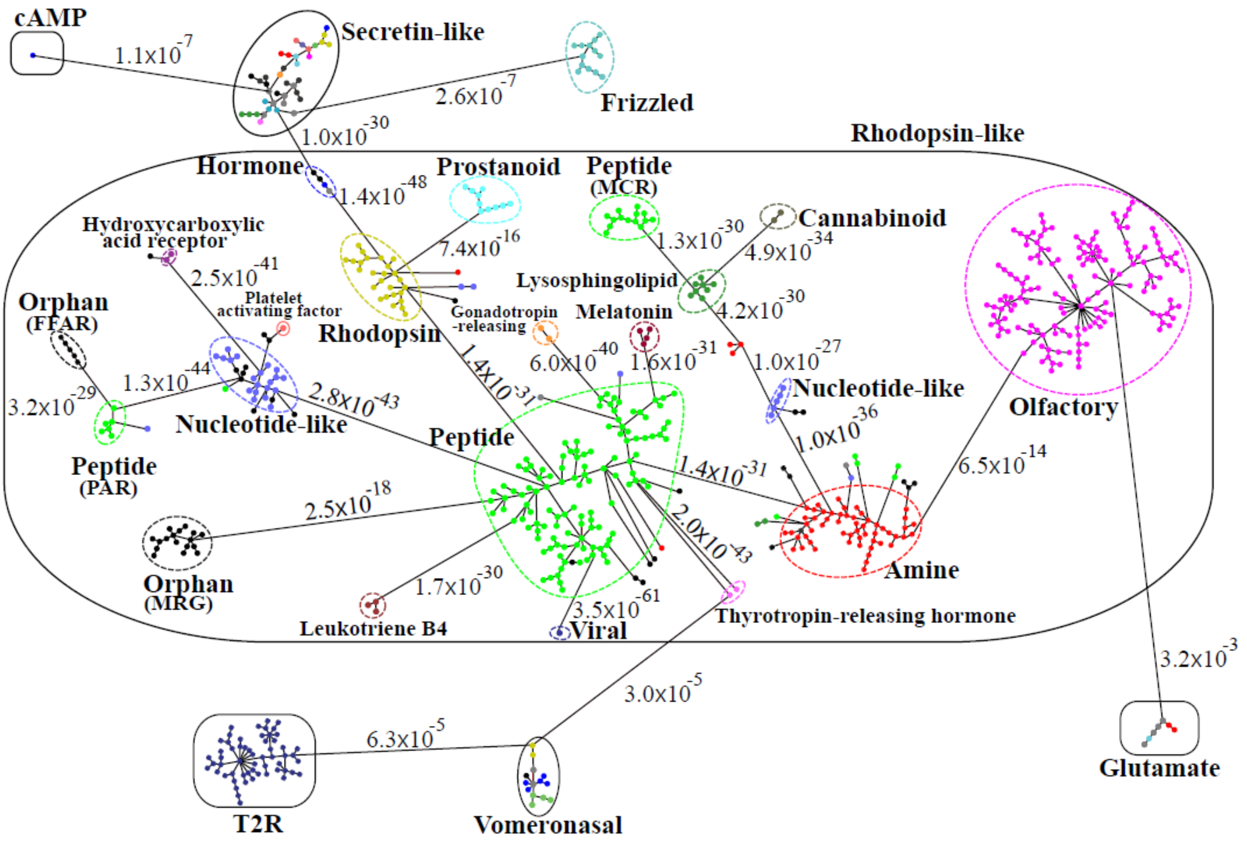


Figure 3. The minimum spanning tree diagram of the GPCR network in the dataset GPCR2841 (outliers not shown). Here each circle represents an MSC cluster the color of which is according to the function of its constituents. The length of the edges is not proportional to their distance, but the distances between subfamilies and classes are labeled to see their sequence similarity.

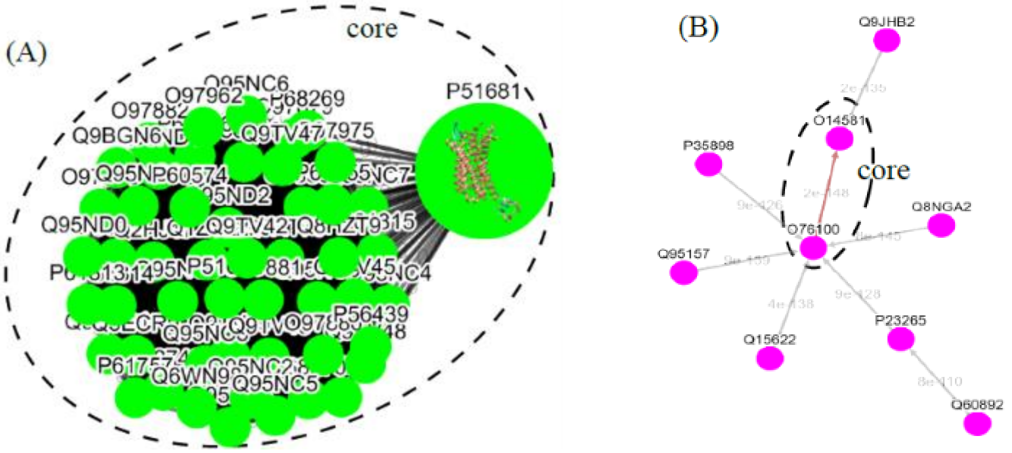
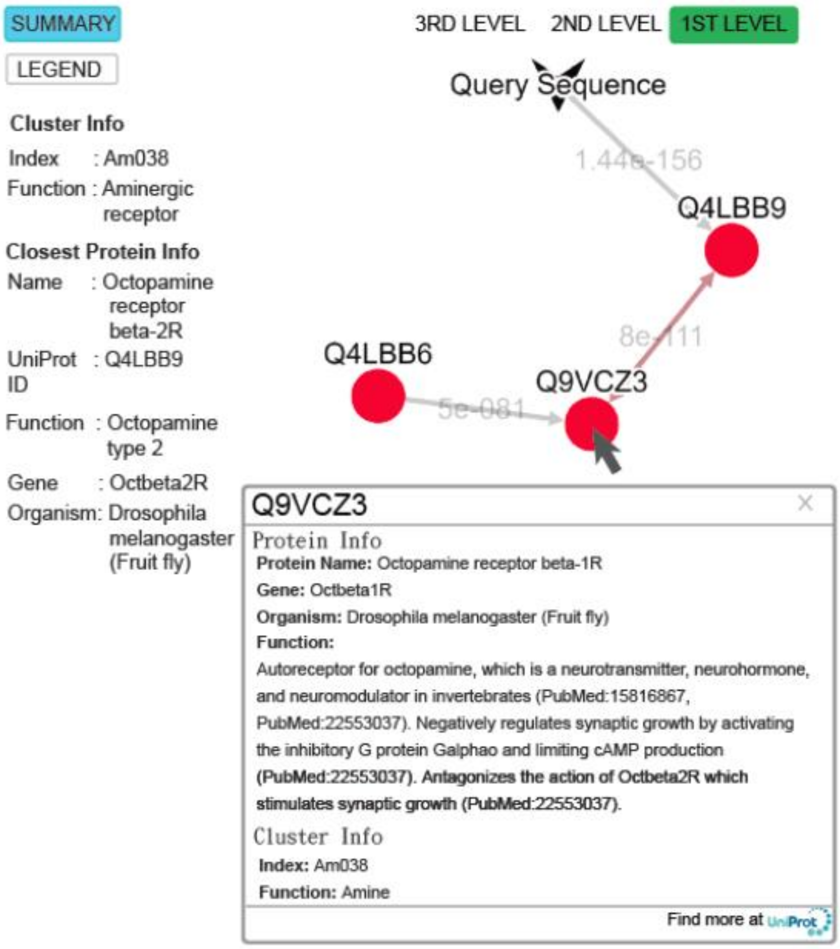


Fig. 4 The tree diagrams of first-level GPCR clusters, showing member sequences of cluster Pe001 which has a conservative core (A), and member sequences of cluster Ol001 which has no conservative core (B). In (A) the three-dimensional protein structure of P51681 is displayed in its node.



Cluster Info

Index : Am038
Function : Aminergic receptor

Closest Protein Info

Name : Octopamine receptor beta-2R
UniProt ID : Q4LBB9
Function : Octopamine type 2
Gene : Octbeta2R
Organism: Drosophila melanogaster (Fruit fly)

Q9VCZ3

Protein Info

Protein Name: Octopamine receptor beta-1R
Gene: Octbeta1R
Organism: Drosophila melanogaster (Fruit fly)
Function:
Autoreceptor for octopamine, which is a neurotransmitter, neurohormone, and neuromodulator in invertebrates (PubMed:15816867, PubMed:22553037). Negatively regulates synaptic growth by activating the inhibitory G protein Galphao and limiting cAMP production (PubMed:22553037). Antagonizes the action of Octbeta2R which stimulates synaptic growth (PubMed:22553037).

Cluster Info

Index: Am038
Function: Amine

Fig. 5 The query result of the submitted sequence of G3M4F8 in SeQuery, showing the cluster graph at the first level. Both cluster information and closest protein information are shown on the left-hand side. Upon clicking on the node Q9VCZ3, a modal box shows both protein information and cluster information of the node.

Alternatively, users can select the centrality tab to see the centrality network on the 3rd level graph. When users click a node on the 3rd level network, SeQuery displays various thresholded network graphs showing statistical properties of the corresponding receptor family, as shown in Figure 6 for the Cannabinoid Receptor family in the Rhodopsin-like class. The available threshold values of sequence distance are 10^{-20} (A), 10^{-50} (B), 10^{-80} (C), and 10^{-100} (D), and edges longer than the threshold are not shown in Figure 6. From Figure 6, it is clear that the edge connecting Q98894 and P47936 has the largest betweenness and thus the largest potential to disconnect graphs if removed. It is also seen that the node Q98894 has the largest weighted degree and betweenness and thus its sequence can be considered as the representative of the Cannabinoid Receptor family. By clicking the MST graph button, MST graphs are also available to show the centrality measures of each GPCR family, as shown in Figure 7 for the Olfactory Receptor family. In the centrality networks, nodes are colored according to the value of their centrality measures (closeness in thresholded graphs and the weighted degree in MST graphs), and edges are colored according to the value of their betweenness centrality. The functional information and centrality data of a GPCR sequence will be displayed in a box after clicking its corresponding node in the network.

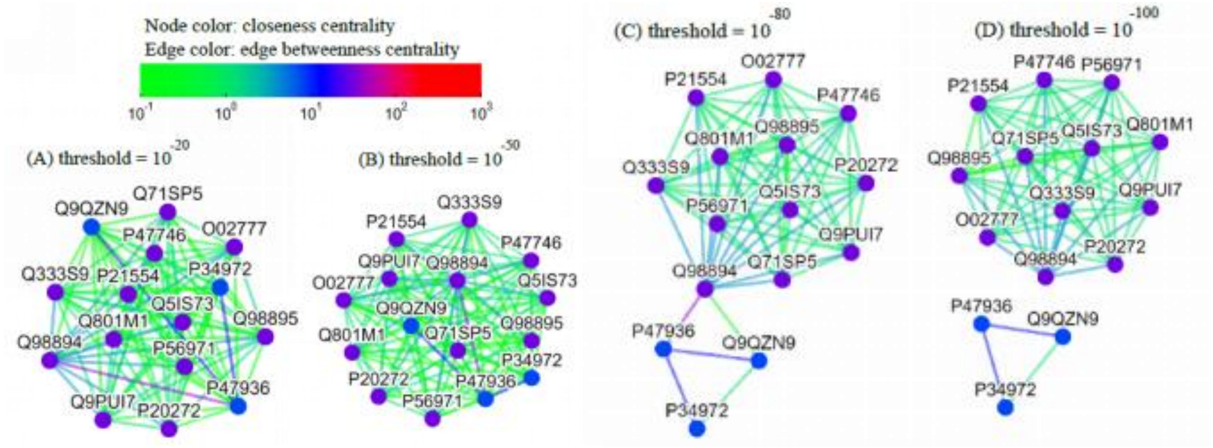


Fig. 6 Thresholded sequence similarity network graphs of the Cannabinoid Receptor family with the threshold distance of 10^{-20}(A), 10^{-50} (B), 10^{-80} (C), and 10^{-100} (D). Nodes and edges are respectively colored by the value of their closeness centrality and betweenness centrality according to the color bar.

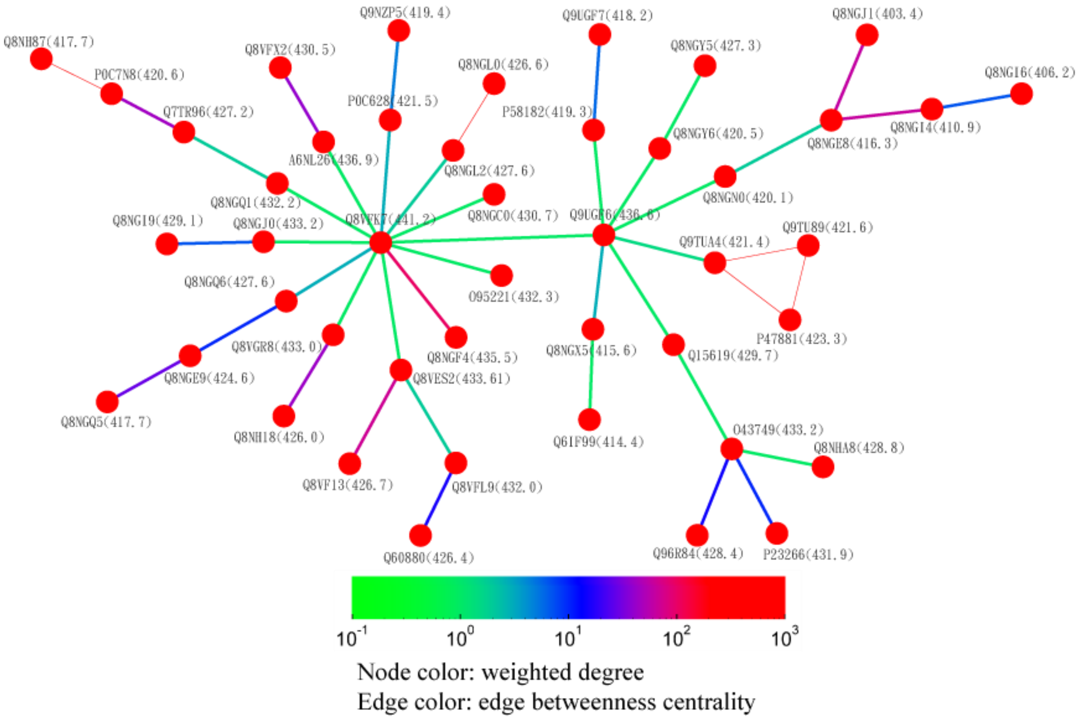


Fig. 7 A partial minimum spanning tree of Olfactory Receptors near the hub sequence Q8VFK7. Nodes are colored based on their weighted degree values which are labeled in the parentheses. Edges are colored based on their betweenness centrality values. Thin edges represent those sequence pairs of zero distance.

(II) In the bottom-up navigation, the database containing the base dataset of 2841 GPCRs is selected, and the query sequence is uploaded or entered at the homepage in the FASTA format. SeQuery determines if it is a GPCR based on its distance to the sequences in the base dataset. If so, as shown in Figure 5, SeQuery will display the graph of network connections between the query sequence and its neighbors at the first level of clustering, as well as information related to the overall cluster and its closest neighbor. The role of this cluster in the GPCR network can be further investigated at the second or third level of clustering.

Centrality measures for the sequence similarity network of GPCRs

In a network, central nodes are those in the thick of things. To evaluate the centrality of important nodes, we define four different centrality measures, namely the weighted degree (C_{WD}), closeness (C_{C}), betweenness (C_{B}), and eigenvector (C_{EV}) centralities, by considering an all-to-all, undirected, and weighted graph $G := (V, E)$ with $|V|$ nodes and $|E|$ edges. The weight matrix **W** of the graph has weights w_{uv} for the edges connecting nodes u and v $\forall u, v \in V$. Equivalently, we can define a distance matrix **D** with elements $\tilde{d}_{uv}$, where $\tilde{d}_{uv} \equiv w_{uv}^{-1} - 1$. To calculate the centrality measures for the GPCR network and avoid numerical errors, we consider the relation $\tilde{d}_{uv} \equiv d_{uv}{}^{0.01} + \delta$, where $\delta = 10^{-200}$ is an arbitrary small distance for $u \neq v$ and d_{uv} ($\forall u, v \in V$) are the sequence distance matrix elements for GPCR pairs. Diagonal elements in both **W** and **D** have a value of 0. For such a graph, the weighted degree centrality of a node u is defined as

$$C_{\mathrm{WD}}(u) = \sum_{v \in V} w_{uv}. \quad (1)$$

The closeness centrality is defined as

$$C_{\mathrm{C}}(u) = (|V| - 1) \cdot \left[\sum_{v \in V} \tilde{d}(u, v)\right]^{-1}, \quad (2)$$

where $\tilde{d}(u, v)$ is the shortest distance between nodes u and v. The betweenness centrality of node u is defined as

$$C_B(u) = \sum_{\substack{j,k \\ j \neq k \neq u}} \frac{g_{jk}(u)}{g_{jk}}, \quad (3)$$

where j and k are other nodes in the network with $j \neq k \neq u$, g_{jk} is the number of shortest paths between node j and node k, and g_{jk} (u) is the number of those paths that go through u. The definition of betweenness centrality in Eq. (3) can be extended to calculate the edge betweenness centrality by calculating g_{jk} (e), the number of shortest paths that go through edge e. The eigenvector centrality is defined as

$$C_{\mathrm{EV}}(u) = \frac{1}{\lambda} \sum_{v \in V} w_{uv} C_{\mathrm{EV}}(v), \quad (4)$$

where λ is the eigenvalue. This equation can be rewritten in vector notation as the eigenvector equation $\mathbf{W}\mathbf{x} = \lambda\mathbf{x}$, where $x_u = C_{\mathrm{EV}}(u)$. As the elements in the weight matrix **W** are non-negative, there is a unique largest eigenvalue, which is real and positive. The eigenvector **x** corresponding to the greatest eigenvalue yields the desired centrality measure.

In studying the sequence similarity network of GPCRs, we use the above centrality indices to characterize important nodes or edges within the network. The weighted degree centrality $C_{\mathrm{WD}}(u)$ of sequence (or node) u is used to characterize its overall connectivity to other sequences in the network. The closeness centrality $C_{\mathrm{C}}(u)$ of sequence u measures the reciprocal of the sum of the length of the shortest paths between u and all other sequences; the more central a sequence is, the closer it is to all other sequences. In graph theory, the betweenness centrality of a node/edge is the number of the shortest paths that pass through the node/edge; a node/edge with higher betweenness centrality would have more control over the network. Lastly, the eigenvector centrality is a measure of the influence of a node in a network; a high eigenvector score means that a node is connected to many nodes of high scores. In general, the closeness/weighted degree/eigenvector centrality measures have similar patterns in a complex network; while the betweenness centrality is characteristically different from the other three measures and has a dynamic role in representing the information flow of the network (31). For the study of GPCRs, the first three centrality measures can be used to find the most representative/influential sequence of a GPCR cluster, sub-family, or family. The betweenness centrality can be used to find particular sequences (nodes) or sequence pairs (edges) that bridge different domains in the sequence space of GPCRs; these sequences or sequence pairs could play a key role in the evolution of GPCRs.