MutHTP: Mutations in Human Transmembrane Proteins

A. Kulandaisamy[1], S. Binny Priya[1], R. Sakthivel[1], Svetlana Tarnovskaya[2], Ilya Bizin[2], Peter Hönigschmid[3], Dmitrij Frishman[2,3] and M. Michael Gromiha[1*]

[1]*Department of Biotechnology, Bhupat and Jyoti Mehta School of BioSciences, Indian Institute of Technology Madras, Chennai 600 036, Tamilnadu, India*

[2]*Department of Bioinformatics, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation*

[3]*Department of Bioinformatics, Technische Universität München, Wissenschaftszentrum Weihenstephan, Freising, Germany*

Abstract

Transmembrane proteins have diverse functional roles, featuring important functions such as cell signaling, cell adhesion, transport of molecules and ions across membranes, energy generation and many others (Traxler *et al.,* 1993; Escriba *et al.,* 2008; Almen *et al.,* 2009; Tuteja *et al.,* 2009; Gromiha and Ou, 2014). Approximately, 20-30% of human genes encode membrane proteins and ~60% of them act as drug targets (Hopkins and Groom, 2002). Mutations and aberrant activity in membrane proteins cause many different developmental disorders and diseases, including several types of cancer, neurodegenerative diseases, cystic fibrosis, diabetes etc. (Overington *et al.,* 2006).

The experimental mutation data are accumulating rapidly in public databases, such as HumSavar (http://www.uniprot.org/docs/humsavar), SwissVar (Mottaz *et al.,* 2010), 1000 Genomes (Auton *et al.,* 2015), ExAC (Lek et al., 2016), COSMIC (Forbes *et al.,* 2015) and ClinVar (Landrum *et al.,* 2014). HumSavar is an index of human polymorphisms and disease mutations. SwissVar portal provides access to a collection of single amino acid polymorphisms and diseases in the UniProtKB/Swiss-Prot database. 1000 Genomes provides a resource of human genetic variation, whereas COSMIC is specific to cancer. ClinVar reports the relationship among human variations and phenotypes, with supporting evidence. The major limitations of the above mentioned databases are (i) information are available only on mutation and disease without any sequence/structure based features (ii) no mapping with structure and type of diseases and (iii) information in these databases are not uniform. Further no specific database on disease causing and neutral mutations is available exclusively for membrane proteins. Hence, constructing a reliable and comprehensive database for membrane proteins would provide a useful resource for understanding the influence of mutations in membrane proteins and developing prediction methods, which provide insights in therapeutic interventions.

In this work, we have developed a comprehensive database for **Mut**ations in **H**uman **T**ransmembrane **P**roteins (**MutHTP**). The membrane proteins were retrieved from UniProtKB (Boutet *et al.,* 2016) using the keyword “transmembrane” in subcellular location and we obtained a set of 5195 human membrane proteins. We collected the mutation data and disease information

from the above mentioned databases such as HumSavar, SwissVar, ExAC, 1000 Genomes, COSMIC and ClinVar. From the overall data, we extracted missense mutations, insertions and deletions for both disease and neutral cases using in-house perl scripts. The combined dataset has 1,73,757 missense mutations, 2455 insertions and 7183 deletions associated with disease and 17516 missense mutations, 39 insertions and 272 deletions considered benign.

In the database we provide the information on each mutation and its chromosomal location, origin, gene name, UniProt ID, PDB code (Rose *et al.,* 2017), location of the mutant with respect to the membrane protein topology (transmembrane domains vs exposed loops), neighbouring residues of the mutant, diseases associated with the mutation, disease classes and the source databases.
For each position in the membrane protein we calculated position-specific conservation scores which reflect the conservation of physico-chemical properties of residues (small, polar, hydrophobic, tiny, charged, negative, positive, aromatic, aliphatic, proline) in the multiple protein sequence alignment (Livingstone and Barton, 1993). We also provide information about the minor allele frequency for each mutation from the ExAC database containing the cohorts from The Cancer Genome Atlas (TCGA) and from non-cancer population (nonTCGA) using its chromosomal location.

The user can search the data by using Gene name, UniProt ID, PDB ID, wild and mutated residues, interface, topology, disease name, disease class and source database. The refinement of the web interface with further search and display options is in progress. This integrated database will be a unique resource to perform wide-range analyses of membrane proteins and to study the effect of mutations using sequence, structure and physiochemical properties of the amino acids.

Acknowledgements: The work was supported by the Department of Science and Technology, Government of India and the Russian Science Foundation grant 16-44-02002.

References

1. Almen, M. S., Nordstrom, K. J., Fredriksson, R., and Schioth, H. B. (2009). Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. BMC Biol, 7,50.
2. Auton *et al.,* (2015) 1000 Genomes Project Consortium: A global reference for human genetic variation. Nature, 526, 68-74.
3. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A. J., and Xenarios, I. (2016). UniProtKB/Swiss-Prot, the manually annotated section of the UniProtKnowledgeBase: how to use the entry view. Methods Mol Biol, 1374, 23-54.
4. Escriba, P. V., Gonzalez- Ros, J. M., Goni, F. M., Kinnunen, P. K., Vigh, L., Sanchez-Magraner, L., and Barcelo- Coblijn, G. (2008). Membranes: a meeting point for lipids, proteins and therapies. J Cell Mol Med, 12, 829-875.

5. Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., and Kok, C. Y. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res, 43, D805-D811.
6. Gromiha, M.M., and Ou, Y.Y. (2014). Bioinformatics approaches for functional annotation of membrane proteins. Brief Bioinform 15, 155-168.
7. Hopkins, A. L., and Groom, C.R. (2002). The druggable genome. Nat Rev Drug Discov 1, 727–730.
8. Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., and Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res, 42, D980-D985.
9. Lek M, Karczewski KJ, Minikel E V, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285–291.
10. Livingstone CD, Barton GJ. 1993. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci 9:745–56.
11. Mottaz, A., David, F. P., Veuthey, A. L., and Yip, Y. L. (2010). Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. Bioinformatics, 26, 851-852.
12. Overington, J. P., Al-Lazikani, B., and Hopkins, A. L. (2006). How many drug targets are there?.Nat Rev Drug Discov, 5, 993-996.
13. Rose, P.W., Prlic, A., Altunkaya, A., Bi, C., Bradley, A.R., Christie, C.H., Di Costanzo, L., Duarte, J.M., Dutta, S., Feng, Z., and Green, R.K. (2017). The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res, 45, D271-D281.
14. Traxler, B., Boyd, D., and Beckwith, J. (1993). The topological analysis of integral cytoplasmic membrane proteins. J Membr Biol, 132, 1-11.
15. Tuteja, N. (2009). Signaling through G protein coupled receptors. Plant Signal Behav, 4, 942-947.