

XXXII. OTDK BIOLÓGIA SZEKCIÓ



A HUMÁN TRANSZMEMBRÁN PROTEOM MEGHATÁROZÁSA ÉS VIZSGÁLATA

Kiegészítő dokumentumok

Szerző:

Dobson László

Info-bionika MSc - I. évfolyam

Pázmány Péter Katolikus Egyetem, MTA Enzimológiai Intézet

Témavezető:

Dr. Tusnády Gábor

MTA Enzimológiai Intézet

Membránfehérje Bioinformatika Kutatócsoport

2014. május 14.

1. Absztrakt

A transzmembrán fehérjék rendkívül fontos szerepet játszanak az élő sejtekben mind egysejtű, mind többsejtű élőlények esetében, hiszen részt vesznek az energiatermelő folyamatoktól kezdve a különböző anyag- és információ "transzport" folyamatokon át a sejt-sejt kapcsolatok kialakításában. Ezek a fehérjék a sejteket, illetve az egyes sejt organelumokat határoló kettős lipidréttegű membránokba ágyazódva helyezkednek el, úgy, hogy a membrán mindkét oldala ("kint" és "bent") felé kinyúlnak a membránból és ezáltal képesek a membrán két oldala közötti anyag- és információáramlásban irányított módon részt venni. Mivel ilyen sok, és ilyen különböző biológiai folyamatban vesznek részt a transzmembrán fehérjék, rendkívül sok betegséggel kapcsolatba hozhatóak, illetve a betegség gyógyításához szükséges gyógyszerek és metabolitok célba juttatásában segédkeznek. A ma forgalomban levő gyógyszerek több mint fele valamilyen transzmembrán fehérjén fejt ki hatását, vagy hat kölcsön a kívánt hatás eléréséhez. Az új, hatásos gyógyszerek tervezéséhez ezért elengedhetetlenül fontos a transzmembrán fehérjék szerkezetének ismerete. Ugyanakkor, ezen fehérjék térszerkezetének meghatározása a ma ismert kísérletes eszközökkel (röntgen-diffrakció, mágneses magrezonancia spektroszkópia) igen nehéz. A humán genom bioinformatikai vizsgálata alapján a genom által kódolt fehérjék mintegy negyede transzmembrán fehérje, azaz kb. 6-8 ezer transzmembrán fehérje található az emberi proteomban, ugyanakkor eddig kevesebb, mint 100 emberi transzmembrán fehérje szerkezete ismert.

A transzmembrán fehérjék szekvencia alapján való pontos azonosítása, illetve a szekvenciák alapján a térszerkezet alacsony felbontású becslése (topológia becslés, "2.5D" becslés), ezen becslések pontosságának és megbízhatóságának a megadása, illetve a topológiai adatok adatbázisba szerkesztése, azok szisztematikus elemzése a jelen TDK munka célkitűzése. Az eredményül kapott humán transzmembrán fehérje proteom és topológiai adatbázis fontos kiinduló pontja lehet ezen fehérjék nagy skálás 3D-s szerkezetének a becsléséhez, illetve új gyógyszer célpontok azonosításához, új gyógyszerek kifejlesztéséhez.

2. Abstract

Transmembrane proteins play important role in living cells, both in unicellular and multicellular organisms, since they take part from energy production processes through material and information transport to establish cell-cell contacts. These proteins are embedded in cell or the cell organell bounding lipid bilayer membrane, extended from one side of the membrane to another ("inner" or "outer"), having the ability to take part in transport and signaling in a controlled way between the two sides of the membrane. The quantity and variety of these functions result a high amount of relation to diseases, thus these proteins can help drug and metabolite delivery. More than half of the currently distributed medicines affects transmembrane proteins, or have interaction with them for the desired result. Knowing the structure of transmembrane proteins is necessary for designing effective medicines. However determining spatial structure with current experimental methods (X-ray, NMR) is really difficult. Bionformatic investigation of the human genome showed, that quarter of the genome coded proteins are transmembrane proteins, which means 6-8000 in the whole human proteome, although less than 100 structures of them are determined.

This paper aims for high reliability identification of transmembrane proteins by sequence and predicting their low resolution structure (topology prediction, "2,5D" prediction) to build databases for systematic analysis. The Human Transmembrane Proteome database can be an important starting-point for high-scale 3D structure prediction of these proteins, identification of new drug targets and designing new medicines.

3. Köszönetnyilvánítás

Köszönettel tartozom konzulensemnek, Dr. Tusnády Gábornak a kapott útmutatásáért, javaslataiért és hogy az elmúlt évben felügyelte a munkámat.

Köszönöm az MTA-TTK Membránfehérje Bioinformatika Kutatócsoportjának a folyamatos támogatást.

Külön köszönettel tartozom Kozma Dánielnek és dr. Mészáros Bálintnak a felmerülő nehézségek esetében nyújtott segítségért.

4. Tételes szerzői hozzájárulás

Alulírott Dobson László a dolgozatban leírt folyamatok közül a:

- Becslő programok kimenetének feldolgozása (DAS,HMMTOP,Membrain,Memsat-SVM,Octopus,Philius,Phobius,Pro,Prodiv,Scampi,SignalP,Sosui,SPOctopus,TMHMM)
- Adatbázisok feldolgozása (PDB,PDBTM,TOPDB,UniProt,TOPDOM)
- Kísérleti eredmények összegyűjtése
- BLAST analízis: UniProt-TOPDB,UniProt-PDBTM
- Adatok szűrése (CD-HIT és "manuálisan")
- Teszthalmaz létrehozása
- Becslő programok tesztelése
- Konszenzus módszerek (transzmembrán filter, szignál peptid filter, topológia becselő) megalkotása és tesztelése
- A megalkotott módszer leprogramozása
- Megbízhatóság tesztelése
- HTP adatbázis létrehozása

végeztem el.

Budapest, 2014. május 14.



Hivatkozások

- [1] Venter, Adams, Myers, Li, Mural, Sutton, Smith, Yandell, Evans, Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.
- [2] Lander, Linton, Birren, Nusbaum, Zody, Baldwin, Devon, Dewar, Doyle, FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.
- [3] Klabunde and Hessler. Drug design strategies for targeting G-protein-coupled receptors. *Chembiochem*, 3(10):928–944, 2002.
- [4] Tusnady, Dosztanyi, and Simon. PDB-TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic acids research*, 33(suppl 1):D275–D278, 2005.
- [5] Kozma, Simon, and Tusnady E. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic acids research*, 41(D1):D524–D529, 2013.
- [6] Tusnady E, Dosztanyi, and Simon. TMDet: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*, 21(7):1276–1277, 2005.
- [7] Fagerberg, Jonasson, von Heijne, Uhlen, and Berglund. Prediction of the human membrane proteome. *Proteomics*, 10(6):1141–1149, 2010.
- [8] Viklund, Granseth, and Elofsson. Structural classification and prediction of reentrant regions in α -helical transmembrane proteins: application to complete genomes. *Journal of molecular biology*, 361(3):591–603, 2006.
- [9] Lasso, Antoniw, and Mullins. A combinatorial pattern discovery approach for the prediction of membrane dipping (re-entrant) loops. *Bioinformatics*, 22(14):e290–e297, 2006.
- [10] Bagos, Liakopoulos, Spyropoulos, and Hamodrakas. PRED-TMBB: a web server for predicting the topology of β -barrel outer membrane proteins. *Nucleic acids research*, 32(suppl 2):W400–W404, 2004.

- [11] Hayat and Elofsson. BOCTOPUS: improved topology prediction of transmembrane β barrel proteins. *Bioinformatics*, 28(4):516–522, 2012.
- [12] Tusnady E and Simon. Topology prediction of helical transmembrane proteins: how far have we reached? *Current Protein and Peptide Science*, 11(7):550–561, 2010.
- [13] von Heijne. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *The EMBO journal*, 5(11):3021, 1986.
- [14] van Geest and Lolkema. Membrane topology and insertion of membrane proteins: search for topogenic signals. *Microbiology and molecular biology reviews*, 64(1):13–33, 2000.
- [15] Uhlen, Oksvold, Fagerberg, Lundberg, Jonasson, Forsberg, Zwahlen, Kampf, Wester, Hober, et al. Towards a knowledge-based human protein atlas. *Nature biotechnology*, 28(12):1248–1250, 2010.
- [16] Tusnady, Kalmar, and Simon. TOPDB: topology data bank of transmembrane proteins. *Nucleic acids research*, 36(suppl 1):D234–D239, 2008.
- [17] Tusnady E, Kalmar, Hegyi, Tompa, and Simon. TOPDOM: database of domains and motifs with conservative location in transmembrane proteins. *Bioinformatics*, 24(12):1469–1470, 2008.
- [18] Gupta, Birch, Rapacki, Brunak, and Hansen. O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic acids research*, 27(1):370–372, 1999.
- [19] Cserzo, Eisenhaber, Eisenhaber, and Simon. TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. *Bioinformatics*, 20(1):136–137, 2004.
- [20] Tusnady E and Simon. Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *Journal of molecular biology*, 283(2):489–506, 1998.

- [21] Tusnady E and Simon. The HMMTOP transmembrane topology prediction server. *Bioinformatics*, 17(9):849–850, 2001.
- [22] Tusnady E. *Sztochasztikus modellek a feherjekutatasban*. PhD thesis, ELTE TTK Szerkezeti Biokemia, 1999.
- [23] Shen and Chou. MemBrain: improving the accuracy of predicting transmembrane helices. *PLoS One*, 3(6):e2399, 2008.
- [24] Nugent and Jones. Transmembrane protein topology prediction using support vector machines. *BMC bioinformatics*, 10(1):159, 2009.
- [25] Jones. Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*, 23(5):538–544, 2007.
- [26] Viklund and Elofsson. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics*, 24(15):1662–1668, 2008.
- [27] Viklund, Bernsel, Skwark, and Elofsson. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics*, 24(24):2928–2929, 2008.
- [28] Reynolds, Kall, Riffle, Bilmes, and Noble. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS computational biology*, 4(11):e1000213, 2008.
- [29] Russell, Norvig, Malik, and Edwards. *Artificial intelligence: a modern approach*, volume 74. Prentice hall Englewood Cliffs, 1995.
- [30] Viklund and Elofsson. Best α -helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Science*, 13(7):1908–1917, 2004.
- [31] Bernsel, Viklund, Falk, Lindahl, von Heijne, and Elofsson. Prediction of membrane-protein topology from first principles. *Proceedings of the National Academy of Sciences*, 105(20):7177–7181, 2008.

- [32] Dyrlov Bendtsen, Nielsen, von Heijne, and Brunak. Improved prediction of signal peptides: SignalP 3.0. *Journal of molecular biology*, 340(4):783–795, 2004.
- [33] Petersen, Brunak, von Heijne, and Nielsen. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8(10):785–786, 2011.
- [34] Hirokawa, Boon-Chieng, and Mitaku. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, 14(4):378–379, 1998.
- [35] Sonnhammer, von Heijne, Krogh, et al. A hidden Markov model for predicting transmembrane helices in protein sequences. In *Ismb*, volume 6, pages 175–182, 1998.
- [36] Kall, Krogh, and Sonnhammer. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic acids research*, 35(suppl 2):W429–W432, 2007.
- [37] Li and Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.
- [38] Kersey, Hermjakob, and Apweiler. VARSPLIC: alternatively-spliced protein sequences derived from SWISS-PROT and TrEMBL. *Bioinformatics*, 16(11):1048–1049, 2000.
- [39] Larkin, Blackshields, Brown, Chenna, McGettigan, McWilliam, Valentin, Wallace Wilm, Lopez, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007.
- [40] Wang, Zhang, Shi, Zhang, and Zhou. Improving transmembrane protein consensus topology prediction using inter-helical interaction. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1818(11):2679–2686, 2012.
- [41] Griep and Hobohm. PDBselect 1992–2009 and PDBfilter-select. *Nucleic acids research*, 38(suppl 1):D318–D319, 2010.