

XXXII. OTDK BIOLÓGIA SZEKCIÓ



A HUMÁN TRANSZMEMBRÁN PROTEOM MEGHATÁROZÁSA ÉS VIZSGÁLATA

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.

Tartalomjegyzék

1. Bevezetés	3
2. Célkitűzések	6
3. Felhasznált adatbázisok, módszerek	7
3.1. Adatbázisok	7
3.1.1. PDBTM	7
3.1.2. TOPDB	7
3.1.3. TOPDOM	8
3.1.4. Egyéb	8
3.2. Szignál peptid, topológia és topográfiabecslő algoritmusok	8
3.2.1. DAS	8
3.2.2. HMMTOP	8
3.2.3. Membrain	9
3.2.4. Memsat-SVM	10
3.2.5. Octopus és SPOctopus	10
3.2.6. Philius	10
3.2.7. Pro és Prodiv	11
3.2.8. Scampi	11
3.2.9. SignalP	12
3.2.10. Sosui	12
3.2.11. TMHMM, Phobius	12
3.3. Egyéb programok	14
3.3.1. BLAST (Basic Local Alignment Searching Tool)	14
3.3.2. CD-hit	14
3.3.3. Swiftnife, Varsplic	15
3.4. Becslés pontosságának mérése	15
3.4.1. A predikció megbízhatóságának becslése	16
4. Módszerfejlesztés	17
4.1. A Constrained Consensus TOPology predikciós eljárás elemei	17
4.2. A CCTOP által használt nyelvtan	18
4.3. Transit szakasz és szignál peptid eltávolítása	18
4.4. TMFilter	19
4.5. Megszorítások keresése	19
4.6. Topológiabecslő eljárások	20
4.7. Megszorítások ellenőrzése	21
4.8. Súlyozás	22
4.9. Végső topológia (HMMTOP)	24
4.10. Az XML fájl felépítése	25

5. Eredmények	26
5.1. Teszthalmaz	26
5.1.1. Szignál peptid predikció és topológia becslés	26
5.1.2. TMFilter	26
5.2. Szignál peptid predikció	27
5.3. Transzmembrán fehérjék szűrése	27
5.4. Topológia becslések eredménye	28
5.5. A módszer megbízhatósága	30
5.6. A HTP adatbázis	31
6. Összefoglalás, kitekintés	32
7. Használt rövidítések	33

1. Bevezetés

A Human Genom Project [1, 2] egy 1990-2003-ig tartó nagyszabású munka volt, mely feltárta az emberi genomot egészen a nukleotidok szintjéig. A genom szekvenciájának meghatározása nem azonos jelentésének megértésével. A bioinformatikai módszerek jelentős része valamilyen funkciót próbál rendelni a megszerzett információhoz, hogy kiváltsa a költséges és időigényes in vivo és in vitro kísérleteket. Az annotáció több szinten történik, amelynek része az ismert proteomok megismerése is, a fehérje térszerkezeti elemek felderítése és ezáltal a funkciók hozzárendelése a fehérjékhez. A jelenleg forgalomban levő gyógyszerek fele a G-protein kapcsolt receptor szupercsaládot (legnagyobb olyan csoportosulás, amely visszavezethető közös őstre) célozza [3], amelyekre 7 transzmembrán szegmens jellemző, így ezek szerkezetének felderítése kiemelt fontosságú. A transzmembrán fehérjék szerkezetének kísérletes meghatározása a tisztítás és a kristályosítás körülményei miatt nehezen megoldhatóak. Jelenleg a PDB-ben található szerkezetek 2%-a tartalmaz transzmembrán szakaszt a PDBTM adatbázis szerint és az összes szerkezet 0.1%-a tartalmaz humán transzmembrán hélixet [4–6]. Különböző predikciós módszerekkel vizsgálva a teljes genomot, jelenlegi ismereteink szerint a teljes proteom körülbelül 25%-a transzmembrán fehérje [7].

A membránfehérjék polipeptid lánc többször is áthatol a membránon. Azok a régiók, melyek az apoláris kettős lipidrétegben helyezkednek el csak meghatározott formákat vehetnek fel. A legkisebb szabad-entalpiájú állapot elérése α -hélix és β -hordó szerkezet kialakításával lehetséges, így a legkevesebb a szabad hidrogénhid akceptor és donor atom. A β -hordó szerkezet szálakból épül fel és csak ezután épülhet a membránba, az ilyen szerkezetek ritkábbak. Az utóbbi 10-15 évben több olyan szerkezeti elemet is azonosítottak, melyek jelentős komplexitást adnak a korábban feltételezett modellhez. A membránba visszahurkoló régió (re-entrant loop) hélix-coil-hélix, hélix-coil vagy coil-hélix részekből áll, a membrán ugyanazon oldalán lép be és halad ki. Az interface hélix a membrán határán, annak síkjával párhuzamosan helyezkedik el [8].

A transzmembrán fehérjékkel kapcsolatos bioinformatikai módszerek nagy része azok topológiájával foglalkozik. Egyes módszerek a topográfiát (elrendezést, jelen esetben csak a membránrégióknak az aminosav szekvenciában való elhelyezkedését) adják meg, míg fejlet-

tebb eljárások képesek a topológia becslésre, azaz meghatározzák, hogy a polipeptidlánc mely része található a membránban és mely annak citoszólikus vagy extra-citoszólikus oldalán. A topológia becslés esetében változó lehet a megbecsült elemek száma (pl.: megkülönböztetik a membránrégió határán elhelyezkedő, a külső és belső részekkel kölcsönhatásban levő részeket), illetve változó lehet a felismert szerkezetek száma és típusa: a legtöbb módszer α -helikális transzmembrán fehérjék topológia becslésére képes (lásd 5.2), de léteznek re-entrant loopot [8, 9], β -hordót [10, 11] felismerő módszerek is. A topológia becselő módszerek hibája, hogy a fehérje N-terminálisán esetlegesen elhelyezkedő szignál peptidet helytelenül transzmembrán hélixnek predikálják. Emiatt a topológia becslés során a szignál peptid rész eltávolítása a szekvencia elejéről több módszer használatának előfeltétele. Bár a fehérjék egyes szegmenseihez rendelt helyzetük önmagukban nem elegendőek a 3 dimenziós felépítésük megalkotásához, fontos lépést jelentenek térszerkezetük pontosabb becsléséhez.

A különböző becselő eljárások a fehérjék kémiai és fizikai tulajdonságain alapulnak, sok esetben szekvenciaillesztéssel javítva a predikciót. Ezek mellett figyelembe veszik a topológiára vonatkozó szabályokat, amik a megjósolt elemek számától függően egyre komplexebbek. A legegyszerűbb szabály, hogy a membránon áthaladó szekvencia darabban az apoláros aminosavak felhalmozódnak. Ezeket a régiókat hidrofóbicitási skála (több is létezik) segítségével, ablakolással lehet megkeresni. Az újabb programok jóval több paramétert vesznek figyelembe és kifinomultabb módszerekkel keresik a membránrégiókat. Összességében a helyesen megválasztott hidrofóbicitási skála és szabadentalpia profil elegendő lehet a pontos topográfia meghatározásához [12]. A módszerek egy része topológiát is becsül. Ehhez a pozitív-bent szabályt alkalmazzák, amely azon a megfigyelésen alapul, hogy a pozitívan töltött aminosavak (arginin és lizin) előfordulása a membránhoz közel eső citoszólikus részben gyakoribb, mint az extra-citoszólikus részben [13]. Ahogy egyre több térszerkezetet sikerül felderíteni, a tanuló gép (machine learning) eljárások, mint a mesterséges neurális hálózat (ANN), a rejtett Markov modell (HMM), a támogató vektor gép (SVM) stb, amelyek ismert topológiájú adatszett alapján képesek tanulni, a különböző térrészekben jellemző aminosav karakterisztikát felhasználva egyre pontosabb eredményt képesek adni. Fontos kiemelni, hogy ezek a módszerek nem feltétlenül visznek közelebb az aminosav eloszlást okozó fizikai vagy kémiai szabályok megértéséhez, illetve folyamatosan

újra kell ezeket tanítani az ismert térszerkezetek bővülésével. A tanítás nem feltétlenül szükséges, ellentétben más HMM alapú módszerekkel, a HMMTOP eljárás a különböző térrészekben fellelhető aminosavak eloszlásának maximális divergenciáját keresve osztja fel részekre a szekvenciát (részletesebben lásd: 5.2). Ahogy egyre több módszer jelent meg, érdekessé vált ezeket együtt is használni a megbízhatóság meghatározása miatt, vagy összevonni ezeket egy még pontosabb konszenzus predikció érdekében.

A predikció pontossága tovább növelhető, ha a program képes figyelembe venni adott részekre vonatkozó, már meglévő topológiai információkat, amelyeket különböző kísérleti eredmények bevonásával nyerhetünk. Ilyen kísérlet lehet fúzió (a riporter enzim fúzióval vagy inzercióval a transzmembrán fehérjéhez való kapcsolódása után a riporter enzim aktivitása meghatározhatja annak lokalizációját), poszt-transzlációs módosítás (N- és O-glikoziláció/foszforiláció rendre az extra- és intracelluláris részben történik), proteáz (fehérjebontó enzim, amely nem tud áthaladni a membránon, így a vágás lokalizációja meghatározható), immunolokalizáció (inzertált epitóp lokalizációjának meghatározása immunokémiai módszerekkel) [14] stb. A kísérleti eredményeket kiegészítheti a konzervatíván, mindig azonos oldalon (citoszólikus vagy extra-citoszólikus) előforduló motívumok vagy domének felhasználása. Néhány módszert ennek megfelelően átalakítottak, hogy ezeket az eredményeket figyelembe tudják venni és egyes részek lokalizációját felhasználva növeljék a predikció pontosságát. A predikciós módszerek pontosságát több dolog befolyásolja. Legtöbb szerkezetet α -helikális transzmembrán fehérjéből áll rendelkezésre, így az ezeket felismerő algoritmusok a legpontosabbak. Ismert re-entrant loop, interface hélix és β hordó szerkezetből sokkal kevesebb érhető el publikus adatbázisokban, így ezeknek felismerése ugyan adhat jó eredményt az ismert halmazon, de a megbízhatósága nem lesz magas. Külön problémát okozhat a re-entrant loop felismerése: egy hosszabb re-entrant loop könnyen összetéveszthető egy transzmembrán α -hélix-szel, egy rövidebb, amely nem hatol olyan mélyen a membránba az interface hélix-szel. Azoknál a módszereknél, amelyek nem tudnak szignál peptidet jósolni problémát okozhat a már korábban említett transzmembrán hélix keresztpredikció, mivel ezek hasonló hosszal és aminosav összetétellel rendelkeznek. A becslést befolyásolja a tanító és teszhalmaz, ezek rendszerint prokarióta szekvenciákat tartalmaznak, amelyek ideálisan nem tartalmazhatnak homológ szekvenciákat.

2. Célkitűzések

Jelenleg nem létezik olyan egységes adatbázis, mely megbízható és naprakész topológiai információt tartalmazna a humán transzmembrán proteomról. A proteinatlas (<http://www.proteinatlas.org>) [7, 15] ugyan tartalmaz topológia becsléseket, illetve konszenzus predikció is elérhető, azonban tesztalmaz hiányában nem állapítható meg a becslések pontossága és az adatok globálisan nem tölthetőek le. A gyógyszerkutatásban és a pontos térszerkezet meghatározásához is hasznos és jól felhasználható lenne egy ilyen adatbázis. Ezért célunk a humán transzmembrán proteom lehető legpontosabb becslése, amit úgy kívánunk elérni, hogy:

- Létrehozunk egy humán transzmembrán fehérje teszt adathalmazt (eddig prokarióta szekvenciákon tesztelték a módszereket).
- Szétválasztjuk a topológia becslés problémáját a globuláris és transzmembrán fehérjék megkülönböztetésének problémájától. Ez utóbbira a továbbiakban, mint transzmembrán fehérjék szűrése, TMFilterként hivatkozunk.
- Létrehozunk egy munkamenetet a becslés menetére, amely a következő lépésekből áll: szekvencia előfeldolgozása, szignál peptid predikció, transzmembrán fehérjék szűrése, topológia becslés.
- Kidolgozunk egy konszenzus becselő eljárást, aminek a legmagasabb a pontossága a csak emberi fehérje szekvenciákat tartalmazó tesztalmazon.
- Kidolgozunk egy módszert a predikció megbízhatóságának a becslésére.
- Olyan eljárást dolgozunk ki, ami képes a kísérleti adatokat figyelembe venni és automatikusan beépíteni a becslésbe.
- A humán transzmembrán fehérjéken végzett nagy áteresztő képességű kísérleti technikákkal meghatározott topológiai adatok összegyűjtése és beépítése a predikcióba.
- A kész programot lefuttatjuk az összes humán fehérjére, a splice variánsok figyelembe vételével, aminek az eredménye a Human Transmembrane Proteom adatbázis.

3. Felhasznált adatbázisok, módszerek

3.1. Adatbázisok

3.1.1. PDBTM

A PDBTM [4, 5] a PDB adatbázis transzmembrán fehérjéket tartalmazó része. A szűrés a TMDET eljárással [6] történik. Az algoritmus először kiválogatja és elveti a vírusokat, píluszt képző fehérjéket és a 15 aminosavnál rövidebb szerkezeteket. Ezután az alacsony-felbontású szerkezeteket válogatja ki, ezeket a továbbiakban máshogy kezeli. A PDB szerkezet alapján felépíti a biomolekulát, eltávolítja azokat a láncokat, melyek nem biológiai kapcsolatot létesítenek. Kiszámítja a víz által hozzáférhető felszínt, majd hidrofóbicitás és szerkezet alapján keresi azokat a részeket, amelyek a membránban lehetnek. Ezekből a részekből megkeresi egy bizonyos célfüggvény segítségével a legnagyobb értékű 15Å-ös szeletet és ha ez egy felső küszöbérték felett van, akkor a szeletet addig tágítja, amíg nem változik a szeleten áthaladó részek száma: ez lesz a membránrégió. A PDBTM adatbázist hetente frissítik, a PDB adatbázis frissítését követően, így a benne levő információ naprakésznek tekinthető.

3.1.2. TOPDB

A TOPDB [16] a legteljesebb és legátfogóbb transzmembrán fehérje topológiai adatbázis. Az adatbázisban található topológiák a fellelhető irodalomban található kísérleti eredmények felhasználásával a HMMTOP algoritmus készítette. A térszerkezeti információk a PDB és a PDBTM adatbázisból vannak átvéve. Az egyes topológiák mellett szerepelnek a felhasznált irodalmi hivatkozások, illetve található keresztreferencia a PDB, a PDBTM és a UniProt adatbázisokkal. Az adatbázis 3 fő lépésben készült:

- Kísérleti adatok feldolgozása (PubMed).
- HMMTOP topológia predikció, felhasználva a megszerzett kísérleti adatokat.
- Megbízhatósági érték számítása a megbecsült topológiákra.

3.1.3. TOPDOM

A TOPDOM [17] azoknak a doméneknek és motívumoknak a gyűjteménye, melyek konzervatív módon a transzmembrán fehérje mindig ugyanazon az oldalán (citoszólus vagy extra-citoszólus) helyezkednek el. A kereső algoritmus a UniProt adatbázis helyesen annotált transzmembrán fehérjéin hajt végre keresést és amennyiben a megtalált domén vagy motívum rendszeresen a membrán ugyanazon oldalán helyezkedik el, bekerül az adatbázisba. A keresést a Hmmpfam (Pfam, Smart adatbázisokon), a Prosite (Prosite adatbázison) és a Fingerprint (Prints) algoritmusok hajtják végre.

3.1.4. Egyéb

A humán transzmembrán proteom a UniProt és az NCBI adatbázisa alapján készült, a teljes proteomon való szűréssel. Az egyes glikozilációs helyek a O-GlycBase-ből [18] illetve az irodalom feldolgozásával keletkeztek.

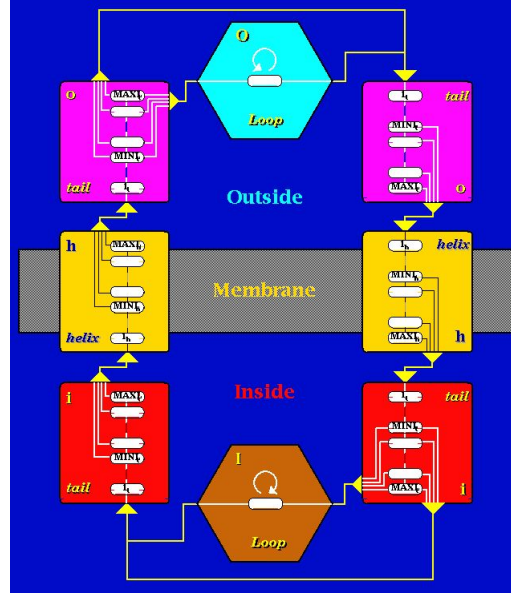
3.2. Szignál peptid, topológia és topográfiabecslő algoritmusok

3.2.1. DAS

Statistikus eljárás segítségével határozza meg, hogy melyik globuláris vagy transzmembrán fehérje. A vizsgált fehérjét ismert topológiájú TM fehérjékkel veti össze, ezt dot-ploton ábrázolja (aminosav hasonlóságot használ, nem csak egyezést). A dot-plot integrálját összegzi a becsléshez összeállított, többnyire prokarióta transzmembrán fehérjéket tartalmazó könyvtár felhasználásával, ez adja a vizsgált fehérje hidrofóbicitás profilját [19].

3.2.2. HMMTOP

A HMMTOP [20–22] azon feltevésen alapszik, hogy a különböző szerkezeti részekben található aminosav eloszlások maximálisan különböznek egymástól, illetve a fehérje egészére jellemző eloszlástól. A maximális különbséget adó szekvencia felosztást egy speciális rejtett Markov modell segítségével lehet megadni. A modell 5 különböző térszerkezeti részből áll: inside loop, inside hélix tail, outside loop, outside hélix tail, transmembrane hélix (1. ábra). Az



1. ábra: A HMMTOP architektúra. Az azonos átmeneti mátrixhoz tartozó állapotok azonos színnel vannak jelölve (hélix: sárga; extracelluláris tail: lila; intracelluláris tail: piros; extracelluláris loop: kék; intracelluláris loop: barna. A négyszög alakú állapotok hossza fix (ebben az esetben láthatóak az alállapotok), a hatszög alakúak hossza nem fix. A nyilak a lehetséges átmeneteket jelölik.

egyes állapotokat az átmeneti valószínűségek alapján származtatja, egymás után. A folyamatot egy pszeudó vektor segít optimalizálni, kísérletileg meghatározott paraméterkészlet alapján. Az egyes térrészek aminosav-összetételének a teljes fehérje aminosav-összetételére vonatkozó divergenciáinak az ötféle térrészben vett összege akkor maximális, ha a megfelelő térrészekből vett relatív gyakoriságok szorzata maximális. Ezt a felosztást nem lehet megtalálni a lehetséges variációk nagyon magas száma miatt, de egy speciális rejtett Markov modell segítségével meghatározható. A HMMTOP képes kísérleti adatokat is figyelembe venni, amelyet a Baum-Welch algoritmusban, mint feltételes valószínűséget épít be.

3.2.3. Membrain

PSSM segítségével kétféle ablakolással keres transzmembrán szegmens "hajlandóságot". Simítás után dinamikus küszöbértéket használva keresi a membránrégiót. Miután az alsó küszöbértékkel talál egy régiót, tovább növeli a küszöböt amíg ennek hatására két részre válik a membránrégiót. Ha a két kapott régió hosszabb 5 aminosavnál, megtartja, ellenkező esetben az eredetileg megtalált szakaszt fogadja el [23].

3.2.4. Memsat-SVM

SVM osztályozáson alapuló módszer. Mivel az SVM-ek bináris klasszifikáció esetén adnak megbízható eredményt, többféle predikciót készít és ezeket vonja össze. Először azt vizsgálja, hogy transzmembrán vagy globuláris-e a fehérje, majd előbbi esetben minden aminosavat négy szempont szerint osztályoz: szignál peptid/nem szignál peptid; TM hélix/nem TM hélix; intracelluláris loop/extracelluláris loop; re-entrant loop/nem re-entrant loop. A Memsat-SVM [24] ezután a korábbi verzióban használt [25] dinamikus programozással keresi azt a modellt, melyre a log-likelihood értékek összege maximális.

3.2.5. Octopus és SPOctopus

Az Octopus [26] a paraméter optimalizáláshoz kétféle címkézést használ. Az első felbontás membrán (membráncentertől vett távolság <13 Ångström), interface (11-18 Ångström távolság), loop (13-23 Ångström távolság), globuláris (>23 Ångström távolság) szétválasztást használ. A második címkézés a loop kategóriába eső, a membrántól legalább 12 pozícióra levő aminosavakat inside/outside címkével látja el. Ezután két ANN-t használva, az első és a második címkézés szerint történik a tulajdonságok becslése. A végső topológiát HMM-el Viterbi algoritmus segítségével számítja ki.

Az SPOctopus [27] a fenti módszer kiegészítése, mely az ANN részben egy "SP preference" pontszámot állapít meg minden aminosavhoz. Amennyiben 13 elem geometriai közepe egy küszöb felett van, egy 3 állapotú HMM állapítja meg a szignál peptid rész határait.

3.2.6. Philius

A Philius [28] a Phobius eljárás kiterjesztése DBN és a HMM általánosításával, melyek a következő előnyöket jelentik:

- Két-állapotú dekódolás (posterior dekódolás és a nyelvtani megszorítást használó Viterbi algoritmus kombinálása).
- A rejtett Markov modell egyszerű állapot átmenet valószínűségeket használva egy-egy állapot hosszúság eloszlása geometriai eloszlást ad. Ez sokszor nem felel meg a

fehérjékben levő szegmensek (pl. rövid hurok) hosszúságeloszlásának. Ezért ebben a modellben a geometriai eloszlás tetszőleges más eloszlásra cserélhető.

- Új megközelítés a bizonytalan elemek címkézésére a tanítás során.
- Minden HMM megfeleltethető DBN-vel. Egy rendszer állapotát annak változói szerint dekomponálva, DBN képes kihasználni az időbeli valószínűségi modell ritkaságát, ami jelentősen kisebb számítási kapacitást igényel [29].

A Philius 4 osztályba sorolja a fehérjéket (globuláris, szignál peptid+globuláris, transzmembrán, szignál peptid+transzmembrán) és a becslés során meghatározza a transzmembrán szegmenseken kívül eső régiók membránhoz viszonyított helyzetét is.

3.2.7. Pro és Prodiv

Profil alapú, HMM alapú módszerek, a profilt a paraméterillesztéshez és a becsléshez is felhasználják. Az illesztéseknél az adott oszlopba tartozó aminosavaknak azonos állapotban kell lenniük. A PRO a TMHMM architektúrájára épül, a PRODIV ennek egy kiterjesztett változata, mely a HMMTOP-hoz hasonlóan az aminosavak egyes szerkezeti részekben mérhető relatív gyakoriságának szorzatának maximumát keresi [30].

3.2.8. Scampi

A Scampi [31] 4-állapotos modellel dolgozik (membrán (M), extra- és intracelluláris (O,I), utóbbinál megkülönböztetve a membránhoz közel eső részeket (i) a pozitív-bent szabály felhasználásával). Az I/O/i részeknél egyenletes eloszlással találja meg a legvalószínűbb aminosavat, i esetén a Lys and Arg kibocsátási valószínűség optimalizálva van. A membrán részeket a hidrofóbicitási momentum és szabadentalpia profil segítségével adja meg. Nem használ tranzíciós valószínűségeket, hanem a kapott nyelvtant tartja meg vagy veti el. A legjobb megoldást Viterbi algoritmus elvével azonos dinamikus programozási eljárás segítségével határozza meg.

3.2.9. SignalP

ANN segítségével keresi a szignál peptid részeket. A tanítás során kétféle negatív adatot használ fel: a korábbi verzió negatív adatszettjén kívül - ami citoplazmatikus vagy nukleáris fehérjéket tartalmazott (SignalP-noTM) - egy olyan halmazon is tanul, ami nem tartalmaz szignál peptidet, de rendelkezik transzmembrán régióval a szekvencia első felében (SignalP-TM). Egy döntési fa segítségével határozza meg, melyik ANN-t használja: amennyiben a SignalP-TM legalább 4 pozíciót transzmembránnak jósol, azt használja fel a végső predikcióra, egyéb esetben a SignalP-noTM-et [32, 33].

3.2.10. Sosui

A SOSUI [34] az aminosavak fizikai és kémiai tulajdonságain alapuló módszer. A program a következő feltevéseket használva határozza meg a transzmembrán fehérjék topográfiáját:

- A transzmembrán fehérjék tartalmaznak legalább egy, részben hidrofób aminosavakat tartalmazó elsődleges hélixet.
- Hidrofil transzmembrán hélixek is előfordulhatnak abban esetben, ha a fehérje polipeptid lánc többször áthalad a membránon.
- Stabilizáló hatások figyelembe vétele: amfifil aminosavak előfordulhatnak a hélixek végén, valamint hidrofób aminosavak a hélix közepén.

A program képes különbséget tenni globuláris és transzmembrán fehérjék között és képes másodlagos, illetve harmadlagos szerkezetet jósolni.

3.2.11. TMHMM, Phobius

A TMHMM [35] a HMMTOP eljárással egyszerre publikált, szintén HMM-en alapuló algoritmus. Az egyes állapotok (összesen 7: hélix, hélix sapka (extra- és intracelluláris oldal külön), intracelluláris loop, rövid/hosszú extracelluláris loop, globuláris domén) több részre vannak osztva, hogy az adott szerkezeti elem hosszát jól modellezzék. Az állapotok hossza szintén eltérő, a hélixek hossza 15 és 35 aminosav között lehet. A "kintről be" és "bentről ki" átmenetek paramétereit azonosnak választották.

A tanítás 160 fehérjén, két lépésben zajlott: először bizonytalannak véve a régiók határait a modell paramétereinek maximalizálását a Baum-Welch algoritmus végzi. A globális maximum megtalálása érdekében az adathoz zajt kevertek, amelyet minden iterációval 5%-kal csökkentettek. A második lépésben az eljárás a topológia megtartásával újraszámítja a régiók határait, 5 aminosavnyi eltérést engedélyezve minden régió határán (azaz összesen 10-et a régiók találkozásánál).

A Phobius [36] ennek a továbbfejlesztett változata, amellyel szignálszekvenciát is lehet becsülni. A szignálszekvencia jellemző részei:

- n régió: n-terminális (1-10)
- h régió: hidrofób szakasz
- c régió: c-terminális (1-12)
- cut: vágás helye

Az becsülő eljárások típusait (topográfia, topológia, szignál peptid, transzmembrán-globuláris szétválasztás) az 1. táblázat foglalja össze.

Módszer	TMFilter	Szignál peptid	Topográfia	Topológia
DAS	+	-	+	-
HMMTOP	-	-	+	+
Membrain	-	-	+	-
Memsat	+	+	+	+
(SP)Octopus	+	+	+	+
Philius	+	+	+	+
Phobius	+	+	+	+
Pro	+	-	+	+
Prodiv	-	-	+	+
Scampi	+	-	+	+
SignalP	-	+	-	-
Sosui	+	+	+	-
TMHMM	+	-	+	+

1. táblázat: A különböző módszerek által adott predikciók típusa.

3.3. Egyéb programok

3.3.1. BLAST (Basic Local Alignment Searching Tool)

Az algoritmus adott hosszúságú, adott értéknél nagyobb hasonlóságú (de nem feltétlenül azonos) szegmenspárokat keres két szekvencia között. Először törli a kereső szekvenciából az alacsony komplexitású részeket. Ezután az összes k elemű szegmensre (aminosavaknál $k=3$) megkeresi, hogy azok milyen más k elemű "szavakkal" adnak az aminosavak hasonlóságát megadó mátrix alapján adott küszöbérték feletti pontszámot. A szavakból készített lista alapján keresést folytat a bemenetként kapott adatbázisban. Találat esetén mindkét irányba növeli a szegmenspárt, hogy High-score Segment Pair (HSP) kapjon eredményül. Az egy szekvencián belül kapott több HSP-t Needleman Wunsch algoritmussal kapcsolja össze és határoz meg egy már betoldásokat (gap) tartalmazó szekvenciaillesztést a kereső és adatbázisbeli szekvencia között.

A kapott találatokat szignifikancia-bebecslés alapján rangsorolja, amelynek legfontosabb mérőszáma az E érték (várható gyakoriság): adatbázison való kereséskor a véletlennek köszönhetően várható, a tényleges találattal azonos vagy nagyobb pontszámmal rendelkező találatok száma.

Az általam írt programok a BLAST eljárás eredményeit abban az esetben fogadják el, ha,

1. E érték: 10^{-10}
2. Az így megmaradt találatok pontozása: a HSP-ken a lekért és a kapott szekvencia hasonlóságát az alábbi képlet határozza meg, ahol az egyes pontok a BLOSUM 62-es mátrixból vannak származtatva. A találatok szűrésénél a küszöbérték: 0,8.

$$\sum_{i=1}^n \frac{\text{score}^2(\text{querysequence}[i], \text{hitsequence}[i])}{\text{score}(\text{querysequence}[i], \text{querysequence}[i]) * \text{score}(\text{hitsequence}[i], \text{hitsequence}[i])} \quad (1)$$

3.3.2. CD-hit

Hatékony és gyors eljárás, ami a fehérje szekvencia adatbázisokat a fehérjék aminosav szekvenciájának hasonlósága alapján szűri meg. A szekvencia rövid szakaszait hasonlítja

össze és ezen darabok statisztikája alapján dönt, hogy két szekvencia bizonyos hasonlósági küszöb alatt van-e. A hasonló találatokból egyet tart meg [37].

3.3.3. Swiffknife, Varsplic

Sok fehérje létezik többféle izoformában, az alternatív splicing-nak köszönhetően. Ezek az izoformák jelentős részben eltérhetnek egymástól. A UniProt-ról letölthető human proteom nem tartalmazza ezeket a fehérjéket, vizsgálatuk azonban hozzátartozik a teljes vizsgálathoz. A Swiffknife és Varsplic programcsomagok a UniProt adatbázisban található annotáció alapján állítják elő ezeket a szekvenciákat [38].

3.4. Becslés pontosságának mérése

Az eredmények kiértékeléséhez a "megtévesztési mátrixban" definiált (2. táblázat) mérőszámokat szokták használni.

	megfigyelt pozitív	megfigyelt negatív
jósolt pozitív	Helyes pozitív	Hibás pozitív
jósolt negatív	Hibás negatív	Helyes negatív

2. táblázat: Megtévesztési mátrix

A becslés pontosságát kétféleképpen is szokás mérni. Az első, a szegmensen értelmezi a valós/becsült pozitív/negatív értékeket (per szegmens predikciós pontosság), a másik, szigorúbb lehetőség, hogy ezeket a fogalmakat a fehérjéken értelmezzük (per fehérje pontosság), úgy, hogy azt a becslést vesszük valós pozitívnak, amelyben az összes szegmens (transzmembrán hélix) helyesen lett meghatározva. Ezeket az értékeket felhasználva definiáljuk a következő fogalmakat:

Szenzitivitás: a ténylegesen a halmazba tartozó elemek helyesen besorolt része.

$$\frac{TP}{TP + FN}$$

Specificitás: a nem a halmazba tartozó elemek helyesen besorolt része.

$$\frac{TN}{TN + FP}$$

Matthews correlation coefficient: bináris klasszifikáció esetén korrelációs koefficiens a megfigyelt és a jósolt osztályok között. Értéke -1 (teljes eltérés a 2 osztály között) és 1 (tökéletes predikció) között változhat, a 0 azt jelöli, ha becslés nem jobb a véletlen összerendezésnél.

$$\frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Per fehérje topográfia pontosság: azon fehérjék aránya a halmazban, amelyekben az összes transzmembrán régió helyesen lett megadva.

Per fehérje topológia pontosság: azon fehérjék aránya a halmazban, melyekben az összes transzmembrán és a köztük levő szegmensek membránhoz viszonyított helyzete is helyesen lett megbecsülve.

3.4.1. A predikció megbízhatóságának becslése

A predikált topológia megbízhatóságának becsléséhez a HMMTOP programot módosítottuk úgy, hogy a program meghatározza minden egyes szekvenciális pozícióban a három lehetséges állapot (kint, bent és hélix) "posterior" valószínűségét és ezek közül a legnagyobbat összegzi a szekvencia mentén. A becsült megbízhatóság alapján csökkenő sorrendbe helyezve az elemeket rendre az első, második (stb) fehérjéig összehasonlítást végezve ellenőrizhető, hogy a megbízhatósági érték mennyire korrelál a topológia pontosságával, illetve az egyes megbízhatósági értékekhez ezen a módon hozzárendelhető a tényleges pontosság.

Szignál peptid predikció esetén figyeltem azt is, hogy a TP találatok esetén az ismert és a megfigyelt szignál peptid hossza mennyire tér el. A 20% alatti eltérést tekintettem megbízhatónak.

4. Módszerfejesztés

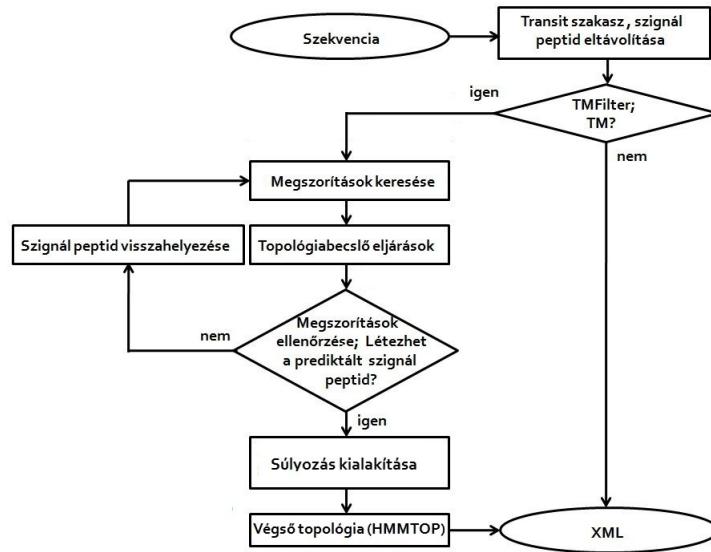
4.1. A Constrained Consensus TOPology predikciós eljárás elemei

A cél egy olyan konszenzus predikció megvalósítása volt, amely jobban teljesít mint más algoritmusok önmagukban és figyelembe tudja venni a becslés során a kísérleti eszközökkel meghatározott biztos topológiai adatokat. Többféle programot teszteltem, melyek eltérő módszerrel és eredményességgel becslik a topológiát (vagy csak topográfiát). Végül 12 módszer választottam ki, melyek könnyen hozzáférhetőek és az általuk adott becslések együttesen jobb eredményt ad, mint külön-külön bármelyiké. A kiválasztott módszerek a következők: DAS, HMMTOP (kísérleti adatok figyelembe vételével), Membrain, Memsat, Octopus, Philius, Phobius, Pro, Prodiv, Scampi-msa, Sosui, TMHMM.

Szignál peptid predikcióra a SignalP programot használtam, melynek eredményét a 6.7 fejezetben leírt módon, a topológia becslést és a kísérleti adatokat is figyelembe véve korrigáltam (SignalP_kor).

A konszenzus becslés során szintén figyelembe vettem az eddig megismert biztos és homológ kísérleti adatokat. A végső kimenet 3 topológiai elemet tartalmaz: citoszólikus, extra-citoszólikus és transzmembrán régió. Ezeken kívül az annotált tranzit és predikált szignál peptid részek szintén bekerültek a végső topológia eredménybe. Az egyes becselő eljárások pontossága növelhető, ha a becslés során figyelembe vesszük a homológ fehérjék aminosav szekvenciáit. Azokban az eljárásokban, ahol lehetőség volt többszörös szekvencia illesztést használni a becsléshez, a rokon szekvenciákat a UniProt adatbázisban található transzmembrán fehérjékből kerestem ki a BLAST eljárás segítségével.

A fenti lépések végrehajtására hoztam létre a Constrained Consensus TOPology predikciós eljárást (továbbiakban CCTOP - 2. ábra), melynek lépéseiről részletes áttekintést adok a továbbiakban.



2. ábra: A CCTOP vázlatos folyamatábrája.

4.2. A CCTOP által használt nyelvtan

A CCTOP több esetben a topológia nyelvtana alapján hoz döntést. Ez azt jelenti, hogy az egyes térszerkezeti elemek csak bizonyos sorrendben követhetik egymást. Jelen esetben a következő megkötéseket tettem:

- Extra- és intracelluláris rész között membránrégióknak kell elhelyezkednie.
- A membránrégióknak legalább 16 aminosav hosszúnak kell lennie (β -hordó és re-entrant loop elemek kizárása).
- Nem lehet azonos oldalra eső rész membránrégió két oldalán (további re-entrant loop elemek kizárása).

Az ismeretlen helyeket nem címkézi fel, ezek tetszőleges értéket vehetnek fel a nyelvtan teljesüléséhez.

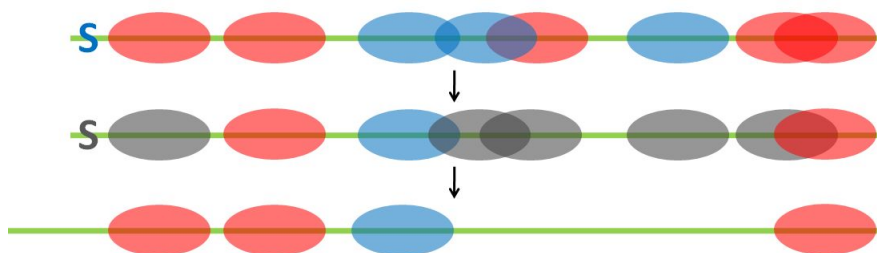
4.3. Transit szakasz és szignál peptid eltávolítása

A CCTOP először a UniProt-on annotált transit peptidet távolítja el. A kapott fehérjéken ezután a SignalP-vel keresi a szignál peptideket. Az előfeldolgozás azért fontos, mert sok-

szor a teljes fehérje szekrécióra kerül, de szignál peptid nem vágódik le. Egy másik fontos szempont, hogy számos módszer csak az érett membránfehérjét tudja hatékonyan feldolgozni. Ezeknél a programoknál gyakori a szignál peptid-transzmembrán keresztpredikció (a hasonló szerkezet és aminosav összetétel miatt), illetve számos eljárás nem lett felkészítve a szignál peptid becslésre (így arra sem, hogy ezután mindig extracelluláris résznek kell következnie).

4.4. TMFilter

A CCTOP a megmaradt fehérjéket két osztályra bontja a Sosui, a TMHMM és a Phobius segítségével: globuláris és transzmembrán fehérjék halmazára. Transzmembránnak minősít egy fehérjét, amennyiben a fenti 3 módszerből legalább kettő legalább egy transzmembrán szegmens jelenlétét valószínűsíti a szekvenciában.



3. ábra: A kísérleti és predikciós eredményekből következő megszorítások ellenőrzésének folyamata I: a vízszintes zöld vonal jelöli a fehérjét, rajta a különböző lokalizációkkal. Kék jelöli az extracelluláris, piros az intracelluláris megszorítást. Az ábra azt a folyamatot mutatja, ahol a CCTOP az egymásnak (lásd 6.7) vagy a nyelvtannak (lásd 6.2) ellentmondó megszorításokat eltávolítja (szürkével jelölve), majd a hibásan becsült szignál peptidek szekvencia darabját visszailleszti. Végül a szignál peptiddel ellentmondó megszorítás visszakerül, mivel a kísérleti eredmény erősebbnek tekinthető mint a predikció.

4.5. Megszorítások keresése

A becslés pontosságát jelentősen javítja, ha a különböző kísérleti eredményeket, mint a rejtett Markov modellben feltételeket, megszorításként vesszük figyelembe. A CCTOP a becslés során a következő adatokat használja fel, mint megszorításokat:

1. Szignál peptid követő rész mindig extracelluláris. Ez a szignál peptid szakaszok jóslására különösen érzékenyvé teszi a CCTOP-t, mivel egy helytelenül pozitív becslés a rossz orientáció miatt a teljes topológiát elronthatja.
2. Glikozilációs helyek (extracelluláris). O-glikoziláció a szerin, threonin, triptofán aminosavak OH csoportján történik, N-glikoziláció az aszparginsav aminocsoportján.

3. TOPDOM keresés eredménye: amennyiben a kérdéses szekvencia tartalmaz olyan szekvencia motívumot, vagy domént, amely megtalálható a TOPDOM adatbázisban, azt is felhasználhatja kényszerfeltételként. A keresést a TOPDOM adatbázis honlapján levő keresőmotorok felhasználásával végzi. A talált domént csak akkor fogadja el, amennyiben a TOPDOM adatbázisban az annotált extracelluláris és intracelluláris helyek számának eltérése 15-nél nagyobb.
4. TOPDB adatbázisban levő fehérjével nagyfokú szekvenciális azonosság. Amennyiben a vizsgált fehérje szekvenciához BLAST keresés segítségével talál hasonló szekvenciát a TOPDB adatbázisban, akkor a TOPDB-ben levő topológiai információkat szintén felhasználja megszorításként a becslés során. A BLAST keresés során használt paraméterek az 5.3.1. fejezetben találhatók. Ezután a CCTOP Clustal-W [39] segítségével párillesztést végez a lekért (UniProt) és a talált (TOPDB) fehérjepárokon. A szekvenciáknak megfelelően a topológiai információkat is illeszti és kizárja a következő fehérjéket:

- Értelmetlen illesztés (ha a topológia által leírt nyelvtan értelmezhetetlen α -helikális transzmembrán fehérjékre).
- Az érett fehérjének nincs transzmembrán régiója.

A megmaradt fehérjéknél a CCTOP nem veszi figyelembe a TOPDB-n tárolt teljes topológiát, csak a különböző kísérleti eredmények által definiált lokalizációkat.

4.6. Topológiabecslő eljárások

A CCTOP ezután meghívja a különböző topográfia és topológiabecslő eljárásokat. Ahol az eredmény 3-nál több kimeneti állapotot tartalmaz (a membrántól való távolság függvényében a topológia tovább van osztva), az eredményt 3 állapotúvá alakítja. A topográfiabecsléseknél ismeretlennek címkézi a nem a membránhoz tartozó részeket.

4.7. Megszorítások ellenőrzése

Egyes esetekben előfordulhat, hogy a kísérleti vagy predikciós megszorítások nem pontosak. Ennek oka lehet emberi hiba (kísérlet során), hibásan becsült szignál peptid vagy egy kivétel a konzervált helyen megjelenő domének esetén.

Ennek elkerülése miatt a CCTOP a különböző kísérleti helyekből kizárja az egymásnak vagy a nyelvtannak ellentmondókat. Kivételt képez a szignál peptid: ez egy predikció, így ellentmondás esetén automatikusan csak ez esik ki, mivel a kísérleti eredmény megbízhatóbb. Amennyiben a szignál peptid predikciót elveti, a szignál peptidnek becsült szakaszt vissza-illeszti (3. ábra).

A topológiabecslő eljárók eredményét felhasználva a CCTOP egy gyors ellenőrzést hajt végre, hogy a hibásan becsült szignál peptidek számát (tovább) csökkentse.

Ehhez egy közelítő topográfiát számol, az alábbi módszerrel. A topológia/topográfia prediktáló programok közül kizárásra kerül egy módszer, ha:

1. egyáltalán nem jósol membránrégiót

Kizárásra kerül a módszer által predikált membránszegmens, ha:

2. teljes részében átlapol egy TOPDOM-ról származó extra- vagy intracelluláris doménnel vagy szekvencia motívummal.
3. a predikált régió és a megszorítás átfedésben van és nincs legalább 3 aminosav átfedetlen rész.
4. bármilyen módon sérti a nyelvtant.

A közelítő topográfia kiszámítása a következő módon történik: egy vektor minden eleme az adott pozícióhoz tartozó aminosav membránrészhez való tartozásának mérőszáma (minél magasabb, annál nagyobb eséllyel tartozik a membránhoz). A különböző módszerek súlya (melyet a topográfiájuk jósága határoz meg) csak akkor kerül a vektorba, ha nincs ellentmondásban a korábbi megszorításokkal.

A következő rész addig fut, amíg lehet membránrégiót predikálni:

A CCTOP megkeresi a vektor legnagyobb elemét és amennyiben ez egy küszöb felett van

(az összes figyelembe vett módszer súlyösszegének fele) a membránrégiót meghosszabbítja mindkét irányba, amíg egy másik küszöb alá nem ér (az előzőleg kapott legnagyobb elem 70%-a). Az összes predikcióban a szakasszal (akár csak) részben átfedő transzmembrán régióknak becsült szakaszokat a továbbiakban nem veszi figyelembe. Ha a vektor minden eleme a küszöb alá kerül, vége a ciklusnak.

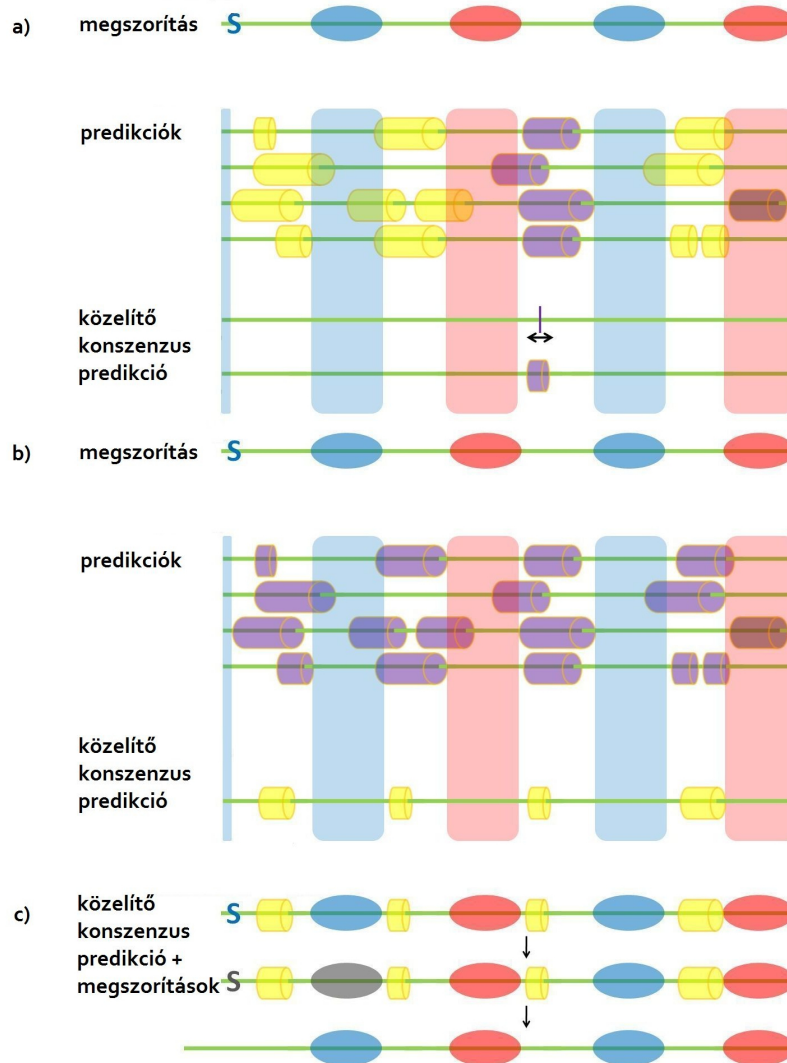
A kapott közelítő topográfiára a megszorításokat ráillesztve ki lehet szűrni a hibásan becsült szignál peptidek egy részét (ha a szignál peptid és az első megadott megszorítás közötti membránrégiók száma nem felel meg a nyelvtani szabályoknak). Amennyiben a szignál peptid predikciót elvetjük, az adott szakaszt visszaillesztjük (4. ábra).

4.8. Súlyozás

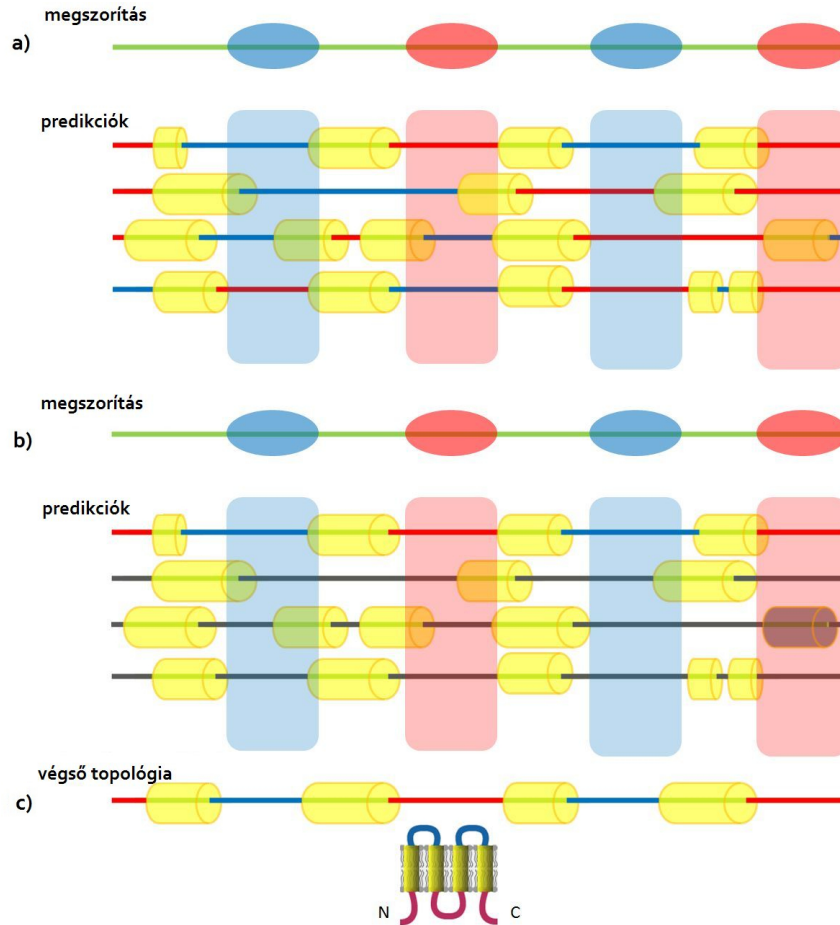
Miután a hamis megszorítások ki lettek zárva, újra futtatja majd validálja a különböző módszereket (lásd kizárások: 6.7/1-4), kiegészítve újabb (már a topológiára) vonatkozó résszel. Kizárásra kerül a módszer által predikált topológia (de a topográfia nem), ha:

5. egy membránrészre való megszorítás nem fed át legalább 3 aminosavval egy becsült membránrégióval.
6. a megszorítás és a predikció között ellentmondás van az extra- vagy intracelluláris lokalizációt illetően.

A megmaradt módszerek különböző régiói adják a súlyozást (membránrégiók a topográfia által adott súllyal, a külső és belső régiók a topológia által meghatározott súllyal), amit a CCTOP a végső topológia meghatározásához fog felhasználni (5. ábra).



4. ábra: A kísérleti és predikciós eredményekből következő megszorítások ellenőrzésének folyamata II: a közelítő topográfia számítása. A vízszintes zöld vonal jelöli a fehérjét, rajta a különböző lokalizációkkal. Kék jelöli az extracelluláris, piros az intracelluláris megszorítást. A szürkével jelölt elemek ellentmondanak egymásnak (lásd 6.7) vagy a nyelvtannak (lásd 6.2). Felül láthatók a megszorítások, alatta a predikciók (a membránrégiók sárgával vannak jelölve), legalul a közelítő konszenzus módszer kimenete. Az a) részen az látható, ahogy a CCTOP a módszerek súlyát összegezve megkeresi a legnagyobb valószínűséggel membránrésznek predikált részt, majd ezt két irányba növeli egy küszöbig. Lilával vannak jelölve azok a régiók, amelyek átfednek a legnagyobb elemmel, ezek később már nem vesznek részt a súlyozásban. A szürkével jelölt membránrégió ellentmondásban van egy megszorítással, ezt nem veszi figyelembe a program. A b) részen a folyamat utolsó állapota látható. A c) részen látható a nyelvtannak (lásd 6.2) ellentmondó megszorítások eltávolítása, majd a hibásan becsült szignál peptid szekvenciájának visszaállítás. A szignál peptiddel ellentmondó megszorítás visszakerül, mivel a kísérleti eredmény erősebbnek tekinthető mint a predikció.



5. ábra: Predikciók és megszorítások összevetése. A vízszintes vonal jelöli a fehérjét, rajta a különböző lokalizációkkal. Kék jelöli az extracelluláris, piros az intracelluláris megszorítást, sárga a membránrészre vonatkozót. A predikcióknál a vonal színe a predikciót, a háttér a megszorítást jelenti. Az a) résznél minden predikált elem megfigyelhető, a b) részben az egymásnak (lásd 6.6 és 6.7) vagy a nyelvtannak (lásd 6.2) ellentmondó részek szürkével vannak jelölve. Ami szürkével van jelölve, az nincs felhasználva a c) részben látható végső HMMTOP predikció súlyozásánál.

4.9. Végső topológia (HMMTOP)

A legtöbb konszenzus predikciót használó módszer (pl.: [7, 40]) nyelvtani szabályok alapján alakítja ki a konszenzus predikciót, anélkül, hogy bármilyen magasabb rendű összefüggést figyelembe venne (ilyen az CCTOP által használt közelítő topográfia is). A CCTOP a végső topológiát a HMMTOP erre a célra átalakított verziójával számolja ki, úgy, hogy az egyes becslő módszerek eredményét - a kísérleti adatokhoz hasonlóan - kényszerfeltételként veszi figyelembe, de ebben az esetben az egyes módszerek globális pontosságával súlyozza a kényszerfeltétel nagyságát.

4.10. Az XML fájl felépítése

A CCTOP eljárás az eredményeket XML formátumba menti el (6. ábra), mely tartalmazza a különböző predikciók kimeneteit, a felhasznált megszorításokat és a mindezeket figyelembe vevő HMMTOP predikció eredményét. A későbbi verziókban az XML file tartalmazni fogja az egyes fehérjékhez található keresztreferenciákat más adatbázisokhoz (UniProt, PDB, PDBTM, TOPDB, NCBI), valamint megtalálhatóak lesznek a fehérje izoformái is. Az XML fájlokból épül fel a Human Transmembrane Proteom (HTP) adatbázis, mely saját 6 jegyű numerikus azonosítókat használ a fehérjék címkézésére. Az XML fájl XSD Schema definíciója a <http://http.enzim.hu/http.xsd> url-en érhető el.

```
<Predictions>  
-<Item AC="A6NHP6">  
-<OriginalSequence>  
MFCCLGYEWLSSGCKTWHSAAWVINTLADHHHCGTGFRGSPWLHFIIAFPRSYKVFIIILWTVTYLVMSLLKTILW:  
</OriginalSequence>  
<Protein type="transmembrane"/>  
-<Sequence>  
MFCCLGYEWLSSGCKTWHSAAWVINTLADHHHCGTGFRGSPWLHFIIAFPRSYKVFIIILWTVTYLVMSLLKTILW:  
</Sequence>  
-<Topology>  
-<Prediction name="HTP" reliability="0.67">  
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOMMMMMMMMMMMMMMMMMMMM  
</Prediction>  
-<Prediction name="DAS">  
#####MMMMMMMMMMMMMM  
</Prediction>  
-<Prediction name="hmmtop">  
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOMMMMMMMMMMMMMMMMMMMM  
</Prediction>  
-<Prediction name="membrain">  
#####MMMMMMMMMMMMMMMMMMMMMMMMMMMMMM  
</Prediction>  
-<Prediction name="memsat">  
#####OOOOOMMMMMMMMMMMMMMMMMMMMMMMMMM  
</Prediction>  
-<Prediction name="octopus">  
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOMMMMMMMMMMMMMMMMMM  
</Prediction>  
-<Prediction name="philius">  
|||||MMMMMMMMMMMMMMMMMMMMMMMMMMMMMOOOOOOOOOOOOOOOOOOOO  
</Prediction>  
-<Prediction name="phobius">  
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOMMMMMMMMMMMMMMMMMMMM  
</Prediction>  
-<Prediction name="pro">  
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOMMMMMMMMMMMMMMMMMMMM  
</Prediction>  
-<Prediction name="prodig">  
IIMMMMMMMMMMMMMMMMMMMMMMOOOOOOOOOOOOOOOOOOOOMMMMMMMMMMMMMMMMMM  
</Prediction>  
-<Prediction name="scampimsa">  
|||||MMMMMMMMMMMMMMMMMMMMMMMMMMMMMOOOOOOOOOOOOOOOOOOO
```

6. ábra: A CCTOP első verziója által generált XML fájl részlete.

5. Eredmények

5.1. Teszthalmaz

5.1.1. Sznál peptid predikció és topológia beclés

A teszthalmazt a UniProt-ról letölthető összes humán fehérjét tartalmazó "fasta" fájl szűrésével állítottam elő. A keresési algoritmus megegyezik az 6.5 4. pontban leírtakkal, a következő kiegészítésekkel:

- A keresést kibővítettem a PDBTM adatbázisra is.
- Az adatbázisban megtalálható teljes topológiát figyelembe vettem.
- A találatokat a CD-HIT programmal szűrtem 40%-os küszöbvel, hogy kizárjam a hasonló szekvenciájú fehérjéket.
- A megmaradt halmazt manuálisan is átnéztem, kiegészítettem a hiányos topológiai információkat és kiszűrtem néhány megmaradt, de nem felhasználható adatot.

Az így kapott adatszettet használtam a topológiabecslő eljárások és a sznál peptid predikciók jóságának mérésére. Összesen 343 szekvencia maradt, ezek hossza és a transzmembrán régiók száma jól lefedi az eddig ismert kategóriákat. Ezen teszthalmaznak használt fehérjek listája és topológiája a HTP adatbázis honlapjáról (<http://htp.enzim.hu>) letölthetők.

5.1.2. TMFilter

A transzmembrán és globuláris fehérjek szétválasztásának pontosságának mérésére az előzőleg generált adatszettet egészítettem ki globuláris fehérje aminosav szekvenciákkal. Ezeket a szekvenciákat a PDBSelect [41] adathalmazból válogattam ki úgy, hogy PDBTM-ben szereplő fehérje ne kerülhessen közéjük. A kapott halmazból véletlenszerűen választottam ki annyit, hogy a teszthalmaz globuláris-transzmembrán fehérje összetétele megegyezzen a humán proteomban becsült 1:3 aránnyal.

5.2. Szignál peptid predikció

Először minden esetben levágtam a UniProt adatbázisban annotált tranzit részeket, a szignál predikciót az így maradt szekvenciákon végeztem. A találatoknál figyeltem azt is, hogy az adott módszer a helyesen becsült találatokon mennyire pontosan jósolja a szakasz hosszát. A SignalP és a Phobius adta a legpontosabb eredményt (MCC: 0.89), azonban a Phobius kevésbé pontosan találta el a TP találatokon a szignál peptid rész hosszát. Az SPOctopus és a SOSUI sok hibás pozitív eredményt adott (3. táblázat).

A CCTOP (a táblázatban: Signalp_kor)a SignalP eredményeit használja fel és korrigálja a 6.7 fejezetben leírtak szerint.

	Memsat	Phobius	SignalP	Sosui	SPOctopus	Signalp_kor
Összes fehérje száma	343	343	343	343	343	343
Helyes pozitív	28	158	157	123	161	157
Hibás pozitív	32	6	5	19	21	3
Helyes negatív	141	167	154	154	152	156
Hibás negatív	142	12	13	47	9	13
Szenzitivitás	0.16	0.93	0.92	0.72	0.94	0.92
Specifitás	0.82	0.97	0.97	0.89	0.88	0.98
MCC	-0.02	0.89	0.89	0.62	0.83	0.90
TP: a hossz eltérés 20% alatt[%]	57	89	93	80	91	93
TP: a hossz eltérés 20% felett [%]	43	11	7	20	9	7

3. táblázat: A vizsgált szignál peptid becselő módszerek eredménye.

5.3. Transzmembrán fehérjék szűrése

A transzmembrán fehérjék szűrését több független módszerrel is elvégeztem. Ezek pontossága általában önmagában is elég jó (MCC 0.93 vagy felette), azonban kombinálva őket még pontosabb becslés is elérhető. A leghatékonyabb a Sosui, a TMHMM és a Phobius által adott predikciók összefésülése volt (4. táblázat), ami alapján transzmembrán fehérjének azt tekintettem, ha a 3 módszerből legalább 2 legalább 1 membrán régiót jósol. Ezt az eljárást TMFilternek neveztem el. További módszerek bevonásával az eredményt nem sikerült javítani.

	Memsat	Octopus	Phobius	Sosui	TMHMM	TMFilter
Összes fehérje száma	1372	1372	1372	1372	1372	1372
Helyes pozitív	339	340	332	327	321	336
Hibás pozitív	36	28	9	14	10	8
Helyes negatív	993	1001	1020	1015	1019	1021
Hibás negatív	4	3	11	16	22	7
Szenzitivitás	0.99	0.92	0.97	0.95	0.93	0.98
Specifititás	0.97	0.99	0.99	0.99	0.99	0.99
<i>MCC</i>	<i>0.93</i>	<i>0.94</i>	<i>0.96</i>	<i>0.94</i>	<i>0.94</i>	<i>0.97</i>

4. táblázat: A transzmembrán fehérjék szűrésének eredménye.

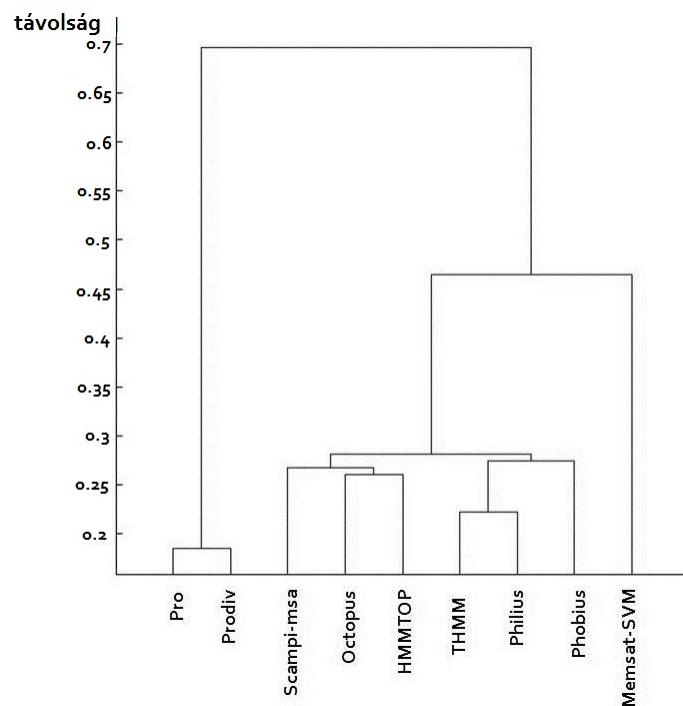
5.4. Topológia becslések eredménye

A topológia becslés eredménye minden esetben a fehérje előfeldolgozása után történt (lásd 6.2-6.8). Mivel ilyenkor a szignál peptid szakaszt egy becsléssel állapítjuk meg, az egyes módszerek hatékonysága pár százalékkal csökken ahhoz képest, mintha a tényleges szignál peptid részt távolítanánk el. Az egyes módszerek között jelentős eltérés található a becsült topológia szintjén, topográfia esetén a különbség nem olyan számottevő (lásd 7. és 8. ábra). Az önálló módszerek közül több is viszonylag pontos (70-80 % között) topográfiát tud becsülni. A legjobb eredményt itt az Octopus érte el (81%). A topológiát is figyelembe véve a kísérleti adatokkal (megszorításokkal, lásd 6.5) megtámogatott HMMTOP és az Octopus volt a legpontosabb - rendre 78% és 77%. A megszorításokat is figyelembe vevő, konszenzus predikción alapuló CCTOP-pal 7%-ot sikerült javítani a legjobb (HMMTOP) topológiabecsítő eljárás eredményén (5. táblázat). Mivel a teszhalmaz 90%-os átfedésben van a TOPDB adatbázissal, amivel az összehasonlítás is történik, a TOPDB kísérleti helyeket nem vettem figyelembe a konszenzus módszer tesztelése során.

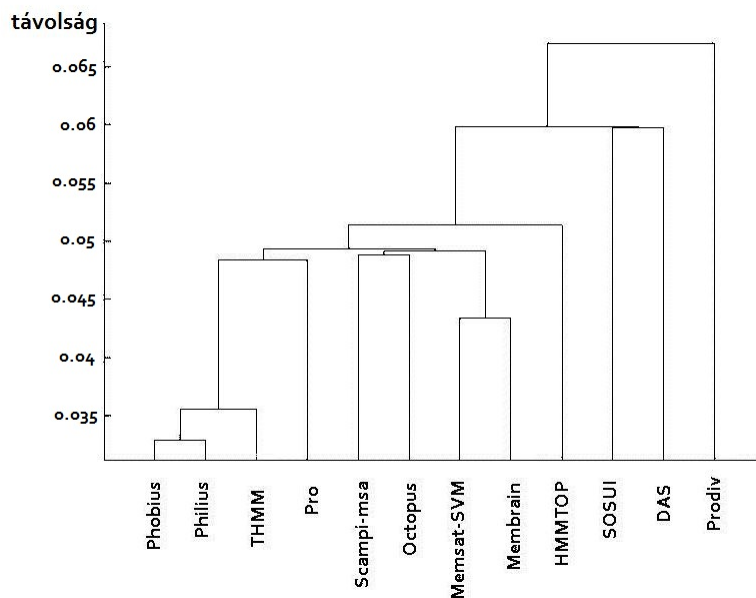
	DAS	HMMTOP	Membrain	Memsat	Octopus	Philius	Phobius
Szenzitivitás (/szegmens)	0.83	0.83	0.95	0.92	0.93	0.9	0.90
Specifititás (/szegmens)	0.85	0.95	0.87	0.92	0.94	0.94	0.96
MCC (/szegmens)	0.84	0.88	0.91	0.92	0.94	0.92	0.93
Topográfia (/fehérje) [%]	62	78	69	78	81	75	79
<i>Topológia (/fehérje) [%]</i>	<i>0</i>	<i>78</i>	<i>0</i>	<i>62</i>	<i>77</i>	<i>66</i>	<i>69</i>

	Pro	Prodiv	Scampi-msa	Sosui	TMHMM	Közelítő topográfia	CCTOP
Szenzitivitás (/szegmens)	0.87	0.94	0.92	0.88	0.87	0.92	0.96
Specifititás (/szegmens)	0.95	0.84	0.9	0.92	0.97	0.97	0.95
MCC (/szegmens)	0.91	0.93	0.91	0.9	0.92	0.94	0.95
Topográfia (/fehérje) [%]	70	67	72	66	74	83	85
<i>Topológia (/fehérje) [%]</i>	<i>54</i>	<i>52</i>	<i>68</i>	<i>0</i>	<i>65</i>	<i>0</i>	<i>85</i>

5. táblázat: A topológia becslések eredménye



7. ábra: A különböző módszerek távolsága topológia szerint. A különböző programok becsléseit minden aminosav pozícióban, páronként összehasonlítva hoztam létre a csoportokat, hierarchikus clusterezéssel. Az egymásra épülő módszerek eredményei közötti korreláció jól látszik (PRO-PRODIV/TMHMM-Phobius-Philius).

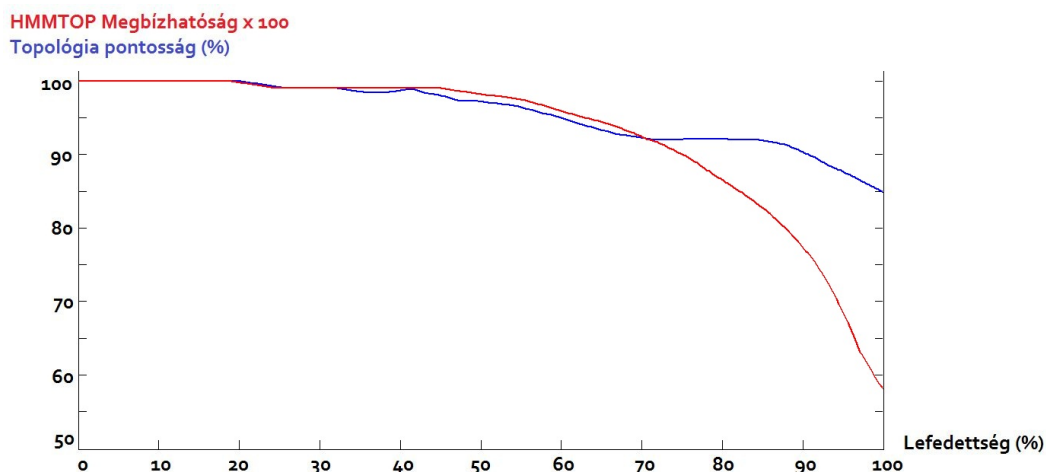


8. ábra: A különböző módszerek távolsága topográfia szerint. A különböző programok becsléseit minden aminosav pozícióban, páronként összehasonlítva hoztam létre a csoportokat, hierarchikus clusterezéssel. Amíg a teljes topológiát nagy eltéréssel jósolják a módszerek, itt az eredmények közötti távolság egy nagyságrenddel kisebb, az eredmények közel egyezők.

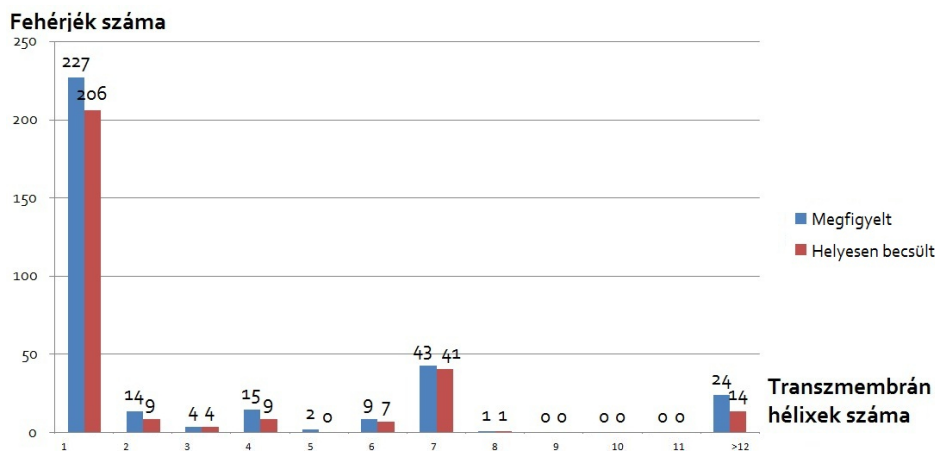
5.5. A módszer megbízhatósága

A HMMTOP kimenete minden predikcióhoz megbízhatósági értéket is tud rendel (lásd 5.4.1 fejezet). A megbízhatóság alapján csökkenő sorrendbe helyezve az elemeket rendre az első, második (stb) fehérjéig összehasonlítást végezve megfigyelhető a teljes adatszett pontosságának alakulása. A CCTOP pontossága az esetek több mint 90%-ban 90% feletti (lásd 8. ábra)

A topológia pontosságát a transzmembrán hélixek eloszlása alapján vizsgálva (9. ábra) az 1 (91%-os pontosság) és 7 (95%-os pontosság) transzmembrán régióval rendelkező fehérjéken a legprecízebb a becslés. Utóbbi azért fontos, mivel a G-protein-kapcsolt receptorok is 7 transzmembrán régióval rendelkeznek és a jelenleg forgalomban levő gyógyszerek fele ezeket célozza [3].



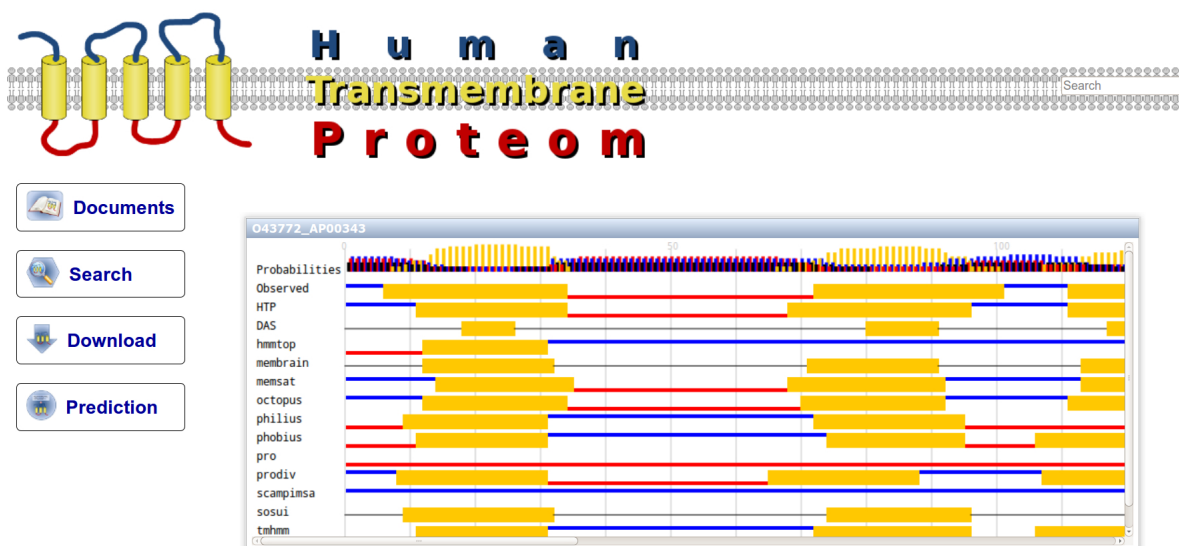
9. ábra: A lefedettség, a predikció pontossága és a HMMTOP megbízhatósági értéke közötti összefüggés. Az adatokat a megbízhatósági érték alapján rendeztem sorba. A vízszintes tengelyen a lefedettség (%) a függőlegesen az addig megszerzett topológia pontossága (kék) és a HMMTOP által számolt megbízhatósági érték százszorosa (piros) látható. Az adatok 50%-hoz több mint 98%-os pontossággal sikerült topológiát predikálni. Az adott helyen becsült megbízhatóság és a lefedett rész pontossága jól korrelál az adatok 70%-án. A simítást mozgó átlag ablakolással végeztem.



10. ábra: A tesztalomban található fehérjék eloszlása a transzmembrán hélixek száma szerint. A kék oszlop jelöli a megfigyelt transzmembrán fehérjék számát, a piros, hogy ezen belül hányhoz sikerült helyes topológiát becsülni. Kiemelkedő a 7 transzmembrán hélixet tartalmazó fehérjéken a becslés pontossága (a gyógyszerek nagy része által célzott G-protein-kapcsolt receptorok is 7 transzmembrán régióval rendelkeznek).

5.6. A HTP adatbázis

A fent részletezett CCTOP predikciót a teljes human proteomon lefuttatva hoztam létre a Human Transmembrane Proteom (HTP) adatbázis, amely a <http://htp.enzim.hu> oldalon érhető el. Az adatbázis 4845 transzmembrán fehérjét tartalmaz (a teljes proteom 24.74%-a). Az adatok XML formátumban vannak (10. ábra), az első (teszt) verzió elérhető a honlapról. Az újabb adatok feltöltése és frissítése jelenleg is folyamatban van, hamarosan a végleges verzió is elérhető lesz. Az oldalon több paraméter szerint lehet keresést folytatni, az egyes predikciók és a konszenzus becslés eredménye grafikusan is megjeleníthető.



11. ábra: A HTP adatbázis első verziója.

6. Összefoglalás, kitekintés

A Human Genome Project óta folyamatosan történik a szekvenálási adatok annotálása. A dolgozatban kidolgozott CCTOP a jelenlegi legmegbízhatóbb transzmembrán topológiabecslő eljárás, amelyet kifejezetten a humán transzmembrán fehérjék topológia becsléséhez dolgoztam ki. A módszert a teljes humán proteomon végigfuttatva jött létre a Human Transmembrane Proteom adatbázis (HTP), amely a teljes humán transzmembrán proteomot tartalmazza, az egyes fehérjékhez tartozó topológia becslésekkel együtt. Az ehhez tartozó weboldal (<http://htp.enzim.hu>) már elérhető az adatbázis első verziójával. Az adatok feltöltése és frissítése (izoformák hozzáadása, az NCBI adatbázis humán proteom részének feldolgozása) folyamatban van.

A CCTOP további optimalizálása után cél egy webszerver létrehozása, ahol tetszőleges szekvencián futtatható a módszer.

A különböző új generációs szekvenálások eredményeképpen meghatározták az emberi populációban természetesen előforduló, illetve a különböző rákos és egyéb megbetegedések során megjelenő szekvencia variációkat is. Ezen nukleotid szekvencia variációk (SNP: single nucleotide polymorphism) egy része a nukleotidok által kódolt aminosavak között is változást okozhat (non-synonym SNP), amelyek az általuk kódolt fehérjék fenotípusát befolyásolhatják. További cél a HTP adatbázisra támaszkova ezeket az SNP-ket a meghatározott topológiára illeszteni.

Több módszer is foglalkozik szekvencia-alapú foldfelismeréssel. Előzetes eredmények alapján a topológiaszűrés közel 20%-al növelheti ilyen módszerek pontosságát. Ez segíthet a pontosabb térszerkezet-meghatározásban és ezáltal a funkciók hozzárendelésében.

A HTP a gyógyszerkutatásban is hasznos lehet: a ma forgalomban levő gyógyszerek több mint fele valamilyen transzmembrán fehérjére fejt ki hatást. Egy megbízható topológiai adatbázis segíthet ezeknek a gyógyszereknek a fejlesztésében.

7. Használt rövidítések

ANN: Artificial Neural Network (Mesterséges Neurális Hálózat)

BLAST: Basic Local Alignment Searching Tool

CCTOP: Constrained Consensus TOPology prediction method

DBN: Dynamic Bayesian Network (Dinamikus Bayes-i Hálózat)

HMM: Hidden Markov Model (rejtett Markov Model)

HTP: Human Transmembrane Proteom

HSP: High-scoring Segment Pair

NCBI: National Center for Biotechnology Information PDB: Protein Data Bank

PDBTM: Protein Data Bank of Transmembrane Proteins

SVM: Support Vector Machine (Támogató vektor gép)

TOPDB: Topology Database of Transmembrane Proteins

TOPDOM: Conservatively Located Domains and Motifs in Transmembrane Proteins

PSSM: Position Specific Scoring matrix (Pozíció Specifikus Pontozó Mátrix)

XML: eXtensible Markup Language (Kiterjeszhető Jelölő Nyelv)