



BUDAPESTI MŰSZAKI ÉS GAZDASÁGTUDOMÁNYI EGYETEM
VEGYÉSZMÉRNÖKI ÉS BIOMÉRNÖKI KAR

TRANSZMEMBRÁN FEHÉRJÉK
SZERKEZETMEGHATÁROZÁSÁT SEGÍTŐ
BIOINFORMATIKAI ESZKÖZÖK FEJLESZTÉSE

DIPLOMAMUNKA

Szerző

Varga Julia Kornélia

Témavezető

Dr. Tusnády Gábor

Belső konzulens

Dr. Vértessy G. Beáta

2017

Tartalomjegyzék

Rövidítésjegyzék.....	4
1. Bevezetés	6
2. Irodalmi áttekintés	7
2.1. Biológiai membránok és alkotórészeik	7
2.2. Transzmembrán fehérjék.....	8
2.2.1. Jelentőség.....	8
2.2.2. Szerkezet.....	8
2.2.3. Funkciók	9
2.3. Transzmembrán fehérjék szerkezetének meghatározása	10
2.3.1. Klónozás, expresszió	11
2.3.2. Szolubilizálás	12
2.3.3. Tisztítás.....	12
2.3.4. Szerkezetmeghatározás röntgenkristallográfiával.....	13
2.3.4.1. Röntgenkristallográfia	13
2.3.4.2. Membránfehérjék kristályosítása	14
2.3.5. Mágneses magrezonancia spektroszkópia	14
2.4. Homológia modellezés.....	15
2.5. Szerkezeti genomikai célpontok kiválasztása	16
2.5.1. Lehetséges megközelítések.....	16
2.5.2. Szerkezeti lefedettség növelése	17
2.6. Tanulógép eljárások	17
2.6.1. Adatok előfeldolgozása.....	18
2.6.2. Teljesítmény mérése	18
2.6.3. A modellek betanítása.....	22
2.6.4. Alul-és túlillesztett modellek	23
2.6.5. Támogatóvektor gépek	24

2.6.6.	Véletlen erdő.....	26
2.7.	Fehérjék kristályosíthatóságát becsülő eljárások áttekintése	28
	Célkitűzések.....	30
3.	Módszerek	32
3.1.	Felhasznált programok	32
3.1.1.	HHblits.....	32
3.1.2.	CCTOP.....	32
3.1.3.	TMFoldrec	33
3.1.4.	BLAST.....	33
3.1.5.	CD-HIT	34
3.1.6.	caret csomag	34
3.1.7.	EVFold.....	34
3.2.	Felhasznált adatbázisok.....	35
3.2.1.	Human Transmembrane Proteome	35
3.2.2.	Protein Data Bank of Transmembrane Proteins	35
3.2.3.	TargetTrack.....	36
3.3.	Az adatbázis felépítéséhez használt munkafolyamat kidolgozása	36
3.4.	A predikciós programhoz szükséges adatok előállítása	38
3.5.	A modell betanítása.....	40
4.	Eredmények.....	42
4.1.	Adatbázis létrehozása, elérése.....	42
4.2.	Az XML fájlok jellemzése	43
4.3.	Az adatbázis jellemzése	44
4.3.1.	Klaszterek	44
4.3.2.	3D kategória.....	45
4.3.3.	Modellezhető kategória.....	45
4.3.4.	Célpont kategória.....	46

4.3.5.	Könnyen és nehezen modellezhető fehérjék.....	47
4.4.	Az adatbázis statisztikái	48
4.4.1.	Számított állapotok	48
4.4.2.	TargetTrack.....	49
4.5.	Összehasonlítás más adatbázisokkal	49
4.6.	Esettanulmány	51
4.7.	A predikciós modellhez összegyűjtött adatok.....	54
4.8.	A betanított modellek kiértékelése	54
4.8.1.	A modellek teljesítménye	54
4.8.2.	Teljesítmény független tesztadatokon	57
4.8.3.	A predikció integrálása az adatbázisba	59
5.	Összefoglalás, kitekintés	60
	Köszönetnyilvánítás.....	61
	Irodalomjegyzék	62
	Ábrajegyzék	70

Rövidítésjegyzék

AUC	area under curve	görbe alatti terület
CSS	cascading style sheets	rangsorolt stíluslapok
HMM	Hidden Markov Model	rejtett Markov modell
MCC	Matthew correlation coefficient	Matthews korrelációs együttható
NMR	nuclear magnetic resonance	mágneses magrezonancia
NPV	negative predictive value	negatív prediktív érték
PPV	positive predictive value	pozitív prediktív érték
PR	precision – recall	precizitás – felidézés
PSI	Protein Structure Initiative	
RMSD	root mean square deviation	négyzetes gyökeltérés
ROC	receiver operating characteristics	vevő működtetési karakterisztika
SVM	support vector machine	támogatóvektor gép
XML	extensible markup language	kiterjeszthető jelölő nyelv

Gyakran használt eljárások és adatbázisok jegyzéke

CCTOP	Consensus Constrained TOPology	http://cctop.enzim.hu/
HTP	Human Transmembrane Proteome	http://htp.enzim.hu/
PDB	Protein Data Bank	http://www.rcsb.org/pdb/
PDBTM	PDB of Transmembrane Proteins	http://pdbtm.enzim.hu

1. Bevezetés

A transzmembrán fehérjék fontos szerepet töltenek be a sejtek működésében. A sejt belsejét a külvilágtól elválasztó membránon keresztüli anyag-és információátvitelben, sejtek adhéziójának biztosításában, különböző jelátviteli folyamatokban játszanak szerepet. A ma forgalomban levő gyógyszerek közel fele transzmembrán fehérjékkel lép kölcsönhatásba. Ennek ellenére az ismert fehérjeszerkezetek igen kis része tartozik transzmembrán fehérjékhez, mivel a szerkezetmeghatározáshoz vezető lépések ezen fehérjék speciális fiziko-kémiai tulajdonságaik miatt gyakran nem járnak sikerrel. Egy humán transzmembrán fehérje szerkezetének meghatározása ezek miatt a nehézségek miatt pár millió dollárba, valamint több évbe is kerülhet¹. Ugyanakkor a fehérjeszerkezetek ismerete nagyban segíti a gyógyszerfejlesztést, a fehérje kölcsönhatások megismerését, funkcióbecslést, egyáltalán a működésük megértését.

A legelső szerkezeti genomikai projektek általában kihagyták a transzmembrán fehérjéket a célpontok közül, mivel már ismert volt, hogy nehezen kezelhetők, és nem álltak rendelkezésre olyan kísérleti protokollok, amelyek nagyobb sikert biztosíthattak volna². Későbbi projektekben, mint az amerikai *Protein Structure Initiative* (PSI), akár teljes kutatócsoportok, centrumok foglalkoztak membránfehérjék kristályosításával, vagy a folyamat egyes lépéseinek optimalizálásával^{3,4}. Az ilyen projekteknek köszönhetően nagyban fejlődtek a rendelkezésre álló kísérleti lehetőségek, protokollok; ugyanakkor a transzmembrán fehérjék szerkezetének megoldása még mindig nagy kihívást jelent

A jelen dolgozatban egy olyan adatbázis, illetve predikciós program elkészítését ismertetem, amelyek humán transzmembrán fehérjék szerkezetmeghatározását segíthetik. A dolgozat az Országos Tudományos Diákköri Konferencián előadott munkám új témával, módszerekkel és eredményekkel kibővített változata.

2. Irodalmi áttekintés

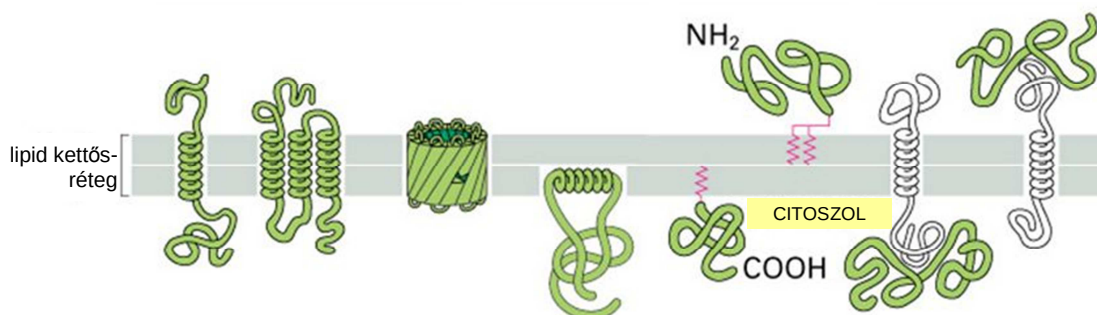
2.1. Biológiai membránok és alkotórészeik

A sejteket, valamint az eukarióta sejtek kompartmentumait kettős lipidrétegből álló membránok határolják. A membránok feladata a különböző térrészek, például a sejt belseje és a külvilág közötti különbségek fenntartása, ugyanakkor kapcsolatot is kell biztosítania ezen terek között. A lipidmolekulák hidrofób alkil láncukkal egymás felé, hidrofil fejükkel kifelé fordulva hoznak létre egy a víz és más hidrofil molekulák számára átjárhatatlan réteget. A membránok nagy részét különböző foszfolipidek (foszfogliceridek, szfingolipidek) alkotják, ezenkívül glikolipidek és a membránok merevségét szabályozó koleszterinek is megtalálható bennük. A membránok pontos összetétele függ az élőlénytől, sejtípustól, környezeti hatásoktól, a membrán elhelyezkedésétől a sejtben, valamint eltérés van a membránokat alkotó két lipidréteg összetételében is^{5,6}.

Az eukarióta sejtek membránjai lipidek mellett szénhidrátokat is tartalmaznak. A szénhidrát részekkel rendelkező glikolipidek eloszlása teljes aszimmetriát mutat: csak a membrán extracitoszolikus oldalán fordulnak elő, nagyrészt a sejt felszínén, ahol fontos feladatuk van a sejt és a környezet közötti kölcsönhatásokban. Többféle funkciójuk lehet: védhetik a sejtek membránját a környezet káros hatásaitól, mint az alacsony pH vagy a degradáló enzimek⁵. Az elektromos töltéssel rendelkező szénhidrátok befolyásolják a membrán környezetében az elektromos teret, valamint az ionösszetételt. Részt vesznek a sejt-sejt felismerésben, sejtek adhéziójában⁵. Néhány glikolipid bakteriális toxinok vagy vírusok receptoraként is szolgálhat, lehetővé téve azok bejutását a sejtbe⁵. A szénhidrát részeket a sejt a Golgi-készülékben köti a lipidekre⁵.

A membránok fontos alkotórészei a membránfehérjék. Ezek mennyisége nagyban függ a sejt típusától, átlagosan a membrán tömegének mintegy a felét teszik ki. Az 1. ábra mutatja a membránfehérjék lehetséges szerkezeteit. A membránfehérjék egy része, amelyeket transzmembrán fehérjéknek nevezünk, teljesen átszeli a membránt, egy (1. **ábra**) vagy több (1. **ábra**2) α -hélixet, vagy β -hordót (1. **ábra**3) alkotva a membránrégióban. A membránfehérjék többi része csak a membrán egyik oldalához csatlakozik. Egyes membránfehérjék a lipidréteg egyik felébe ágyazódnak (1. **ábra**4), mások kovalens kötéssel csatlakoznak a membrán lipidmolekuláihoz, akár közvetlenül (1. **ábra**5), akár oligoszacharid linkeren keresztül (1. **ábra**6). Membránfehérjék

csatlakozhatnak nemkovalens kötéssel transzmembrán fehérjékhez is (1. **ábra**7-1. **ábra**8)⁵. Dolgozatomban a membránrégióban α -hélixet alkotó fehérjékkel foglalkozom.



1. ábra. A membránfehérjék lehetséges szerkezetei⁵.

A membránokban nemcsak a lipidekhez, hanem a fehérjékhez is csatlakozhatnak szénhidrát egységek, leginkább N-vagy O-glikozilációval. Mindkét poszttranszlációs módosítás főként eukariótákban és archeákban történik, de ritkán baktériumokban^{7,8} is lejátszódik. Az eukarióta N-glikoziláció általában NXS/T motívum aszparagin aminosavjának nitrogénjén történik az endoplazmás retikulumban, a transzlálódó fehérjén. A szénhidrát egységek további módosítása a Golgi-készülékben zajlik, a folyamat minden lépését enzimek katalizálják. Az O-glikoziláció során hidroxil-csoporttal rendelkező aminosavak oxigénjéhez kapcsolódik szénhidrát a Golgi-készülékben⁵.

2.2. Transzmembrán fehérjék

2.2.1. Jelentőség

A humán proteom körülbelül 25-30%-át transzmembrán fehérjék teszik ki⁹. A piacon levő gyógyszerhatóanyagok 50%-a transzmembrán fehérjék működését befolyásolja, azokkal lép kölcsönhatásba¹⁰⁻¹².

2.2.2. Szerkezet

A hidrofób membrán átszelése alapvetően befolyásolja a transzmembrán fehérjék szerkezetét, mivel így az energetikailag legkedvezőbb szerkezetet úgy kell felvenniük, hogy az mind a hidrofób membránrégióban, mind a hidrofíl környezetben megfelelő legyen. A membránrégiókban ezek a fehérjék többségében hidrofób aminosavakat tartalmaznak, bár előfordulhatnak töltött aminosavak is, jellemzően csatornát alkotó fehérjék membránrégiójában, hogy a csatornák szelektivitását biztosítsák.

Energetikailag kedvező, ha az összes hidrogénhíd-donor és -akceptor csoport kötést tud kialakítani, ami a peptidlánc gerince esetén csak akkor valósul meg, ha másodlagos szerkezet, α -hélix vagy β -redő alakul ki a membránrégióban. β -redőkből kialakuló β -hordó szerkezetet felvevő membránfehérjék jellemzően Gram-negatív baktériumok külső membránjában, valamint a mitokondriumban és kloroplasztban találhatók. α -hélix szerkezetű membránfehérjék minden sejtben megtalálhatóak.

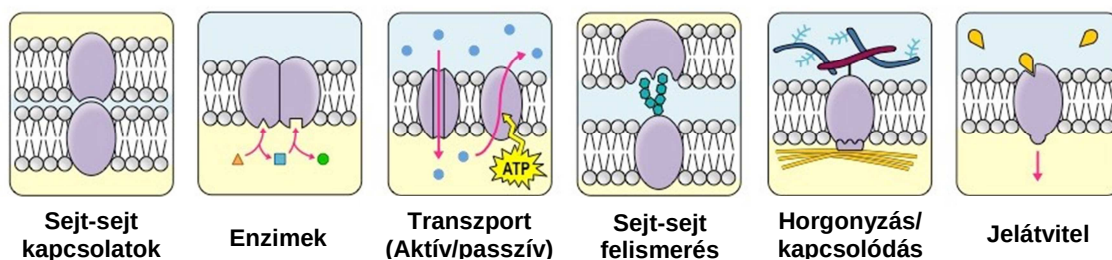
Az aminosav szekvencia ismeretében a transzmembrán fehérjék topográfiája (a membránrégiók elhelyezkedése a peptidláncon belül), valamint topológiája (a membránrégiókat összekötő szakaszok orientációja – extracelluláris vagy citoplazmatikus oldal) is jól becsülhető¹³.

A topológiabecslés folyamán először meghatározzák, hogy a fehérje tartalmaz-e szignálszekvenciát, valamint azt, hogy globuláris vagy transzmembrán fehérjét kódol-e. Majd ezután következik a membránrégiók felismerése, ezt a szekvenciában levő hidrofób aminosavak eloszlása alapján lehet meghatározni, mivel ezeket a részeket nagyrészt hidrofób aminosavak alkotják. Végül a membránrégiók közötti részek elhelyezkedésének meghatározása következik: ehhez legelterjedtebben a „pozitív benn” szabályt alkalmazzák, amely szerint a membrán citoplazma felőli oldalán halmozódnak fel a pozitív töltésű aminosavak^{14,15}. A HMMTOP^{16,17} predikciós program egy másik megközelítést használ: a membránfehérje topológiáját úgy becsli, hogy maximális legyen az egyes térrészekben elhelyezkedő aminosavak eloszlásának különbözősége. A becslés pontossága javítható több predikciós program eredményének egyesítésével¹⁸, ismert kísérletes topológiai adatok felhasználásával^{19,20}, valamint egyéb információk, például a membrán konzisztensen egy oldalán lokalizálódó domének^{21,22} meghatározásával²³.

2.2.3. Funkciók

A transzmembrán fehérjék létfontosságú szerepet töltenek be az élő szervezetek működésében (2. ábra). A membránok elválasztják a különböző térrészeket, ugyanakkor az anyag és az információ áramlását biztosítani kell – ezt a szerepet töltik be a transzmembrán fehérjék. Biztosíthatják sejtek egymás közötti fölismerését és kapcsolódását, receptorként részt vesznek jelátviteli folyamatokban. Egyes transzmembrán fehérjék enzimatisz aktivitással rendelkeznek. A különböző molekulák transzportját a membrán egyik oldaláról a másikra szintén ezek a fehérjék végzik, aktív

(ATP-t használó, koncentrációgradienssel ellentétes) vagy passzív (koncentrációgradienssel megegyező irányú) transzporttal. A sejt extracelluláris mátrixhoz való kapcsolódását, horgonyzását is transzmembrán fehérjék biztosítják. Mivel ilyen fontos szerepet töltenek be a sejtek életében, hiányuk, rendellenes működésük súlyos betegségekhez vezethet, ugyanakkor jó gyógyszercélpontok: hatékonyan lehet rájuk keresztül befolyásolni a celluláris folyamatokat.

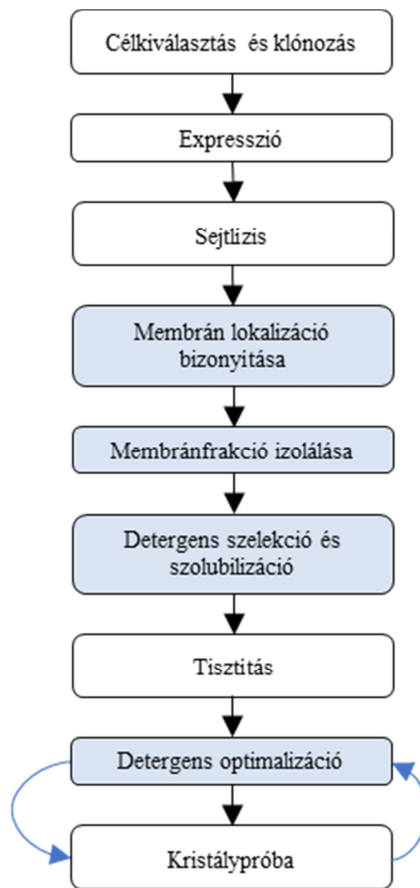


2. ábra. A transzmembrán fehérjék funkciói²⁴.

2.3. Transzmembrán fehérjék szerkezetének meghatározása

A leginkább elterjedt fehérjeszerkezet-meghatározási módszerek a röntgenkristallográfia, a mágneses magrezonancia spektroszkópia, valamint az elektronmikroszkópia. A különböző eljárások kezdeti lépései azonosak, valamint hasonlóak a globuláris fehérjék szerkezetmeghatározásának folyamatához, azonban van néhány lépés, amelyeket csak membránfehérjék esetében alkalmaznak (3. ábra).

Első lépésként expressziós rendszerben klónozni és kellő mennyiségben ki kell fejeztetni a fehérjét, amit ezek után oldatba visznek és tisztítanak. A fehérjének igen nagy tisztaságúnak kell lennie, hogy a különböző módszerekkel a szerkezete meghatározható legyen, például, hogy röntgenkristallográfiai vizsgálatoknál a kristályosítás során a lehető legkevesebb zárvány és egyéb szennyező legyen a kristályban²⁵. E lépések után kerülhet sor a tulajdonképpeni szerkezetmeghatározásra, ez történhet röntgenkristallográfiával, mágneses magrezonancia spektroszkópiával, valamint legújabb technikaként krio-elektronmikroszkópiával is.



3. ábra. A membránfehérjék kristályosításának általános folyamata.

Kék háttérrel a membránfehérjékre specifikus lépések²⁶.

2.3.1. Klónozás, expresszió

A membránfehérjék klónozása megegyezik a többi fehérje klónozási technikájával. A folyamat során figyelembe kell venni, hogy egyes membránfehérjék az N-terminális végükön szignálszekvenciát tartalmaznak, így a tisztításhoz gyakran használt hexahisztidin, vagy egyéb tageket általában a C-terminálisra viszik be.

A klónozás első lépéseként a fehérjét kódoló DNS szakaszt az eredeti stop kodon kivételével PCR-rel felszaporítják, majd emésztik a vektornak megfelelő restriktációs endonukleázokkal. A szekvenciákat expressziós vektorokba építik be, majd szekvenálással alátámasztják, hogy a klónozendó szakasz valóban beépült a vektorba. A fehérje cDNS-ét kódoló vektorokat expressziós rendszerbe transzformálják, majd megfelelő idejű növesztés után kinyerik a sejteket.

A membránfehérjék megfelelő mennyiségben történő kifejeztetése több nehézségbe is ütközik. Egyrészt, a nagymennyiségű membránfehérje gyakran toxikus a sejtre nézve. Másrészt, sokszor nem funkcionális formában történik a fehérje termelése, például prokarióta rendszerekben a poszttranszlációs módosulások nem jelennek meg, vagy a membrán összetétele eltérő attól, ami a fehérje eredeti környezete; ezek a körülmények a fehérje szerkezetét, így stabilitását is befolyásolhatják²⁶. Harmadrészt, a nagy tömegű vagy sok transzmembrán hélixet tartalmazó fehérjéket gyakran egyáltalán nem sikerül kifejeztetni. A siker ebben az esetben növelhető a megfelelő expressziós rendszer kiválasztásával, valamint a tenyésztési körülmények optimalizálásával²⁷.

2.3.2. Szolubilizálás

A fehérjék membránból történő kioldásához detergenseket használunk, amik képesek a membránból kioldódó hidrofób részeket beborítani, utánozva a membránban levő körülményeket, ezzel megakadályozva a fehérjék aggregációját a hidrofób membránrégiókon keresztül. A kemény detergenssek, mint az SDS sokszor nemcsak kioldják a fehérjét, hanem denaturálják is. A lágy detergenssek alkalmasak a fehérjék szerkezetének meghatározásához⁵. A legtöbb fehérje egy detergenst vagy detergens keveréket preferál, ennek megtalálását általában nagy áteresztőképességű szűréssel (*screening*) végzik. Előfordulhat, hogy a fehérjét egy bizonyos detergensben szolubilizálják, majd átoldják egy másik detergensbe, amely megfelelőbb környezetet biztosít a tisztításhoz.

A szolubilizálás első lépéseként homogenizálják a sejteket, majd centrifugálás és a felülúszó eltávolítása után a pellethez hozzáadják a detergenst. A detergenssel való inkubáció és centrifugálás után a kapott felülúszó fogja tartalmazni a membránfehérjét. A fehérje jelenlétét festéssel vagy a hexahisztidin tagre specifikus antitestekkel lehet kimutatni²⁶.

2.3.3. Tisztítás

A fehérjék tisztítása történhet különböző tagok alapján, mint a fent említett hisztidin tag. Ezenkívül alkalmazható streptavidin, biotin, FLAG vagy hemagglutinin tag is. A tisztítást követően a tag-eket általában proteázzal eltávolítják a fehérjéről. Amennyiben elérhető az adott fehérjéhez antitest vagy ligand, a tisztítás ezek segítségével is megtörténhet.

Ezek a módszerek kivül klasszikus biokémiai eljárásokkal is történhet a tisztítás. Használatos ammónium-szulfát precipitáció, szacharóz gradienssel történő centrifugálás, amelyeknek hátránya, hogy a többi módszerhez képest nagyobb mennyiségű fehérje kell az elvégzésükhöz. Lehetséges még méretkizárásos, hidrofób kölcsönhatáson alapuló vagy fordított áramlású kromatográfia alkalmazása is²⁸.

2.3.4. Szerkezetmeghatározás röntgenkrisztallográfiával

2.3.4.1. Röntgenkrisztallográfia

A röntgensugárzás nagyenergiájú elektromágneses sugárzás. Mint minden hullám, ha a hullámhosszával összemérhető (röntgensugárzásnál 0,01 és 10 nm közötti) méretű akadály vagy hasadék kerül az útjába, akkor azon diffrakciót (azaz elhajlást) szenved. Mivel a kristályos anyagok rácssíkjainak a távolsága ebbe a tartományba esik, ezért a röntgensugárzás elhajlik ezen a rácson, lehetővé téve ezzel a kristályos anyagok vizsgálatát. Ehhez megfelelő méretű (pár tizedmilliméteres), rendezett egykristályokra van szükség, amelyek előállítása főleg makromolekulák (például fehérjék) esetében igen nehéz és hosszú folyamat is lehet. Az egykristályokban minden elemi cella azonos orientációjú.

A mérést alacsony hőmérsékleten végzik, megakadályozva ezzel a fehérjék röntgensugár általi roncsolódását, miközben a fehérjekristályt folyamatosan forgatják, hogy a kristályról a lehető legtöbb információt szerezzék. A kapott diffraktogramokat elemezve Fourier-transzformációval kapjuk meg a fehérje elektronsűrűség-térképét, amely a fehérje háromdimenziós kontúrjaként fogható fel. A fehérjeszekvencia ismeretében az elektronsűrűség eloszlásából meghatározható a fehérjeszerkezet^{29,30}.

A módszer hátránya, hogy a fehérjét nem fiziológiás környezetben vizsgálják, a fehérjedinamikájáról nem kapunk információt, valamint bizonyos fehérjék (például a transzmembrán fehérjék) nehezen kristályosíthatók, így a módszer kevésbé alkalmas ilyen fehérjék szerkezetének meghatározására. Előnye ugyanakkor, hogy nagy felbontású szerkezetek felderítésére alkalmas, valamint az, hogy nagyméretű fehérjék és fehérje komplexek, akár egy teljes vírus³¹ szerkezete is meghatározható vele.

2.3.4.2. Membránfehérjék kristályosítása

Az egykristályok létrehozását célzó kristályosítás a szerkezetmeghatározási folyamat legnehezebb lépése, gyakran több száz vagy akár több ezer beállítás tesztelését jelenti, sokszor így is sikertelenül. Többváltozós lépés, ahol rengeteg biokémiai, fizika, kémiai paramétert kell megfelelően beállítani a sikerhez. Automatizálással a próbálgatás sokkal gyorsabb, de még mindig igen nagy kihívást jelent³². Adalékok, például kismolekulás ligandok, kofaktorok, antitestek hozzáadásával javítható a fehérje stabilitása²⁶.

A kristályosítást leggyakrabban *in surfo* technikákkal hatják végre, gőzdiffúziós, mikrodialízis vagy mikrobatch technikákkal. Ezeknél a módszereknél az egykristály növekedéséhez szükséges fehérje kölcsönhatások biztosítása kulcsfontosságú, ezt megfelelő detergensek használatával lehet elérni³². Kis extramembrán régiókkal rendelkező membránfehérjék kristályosítása nehezebb, mivel a fehérjekölcsönhatások a hidrofób részeket burkoló detergensek miatt nehezen tudnak létrejönni³³.

2.3.5. Mágneses magrezonancia spektroszkópia

A mágneses magrezonancia spektroszkópia (*nuclear magnetic resonance spectroscopy*, NMR) az atommagok azon kvantummechanikai tulajdonságán alapul, hogy mágneses térbe helyezve az atommag bizonyos frekvenciákon – a környezetétől függően – egy másik rádiófrekvenciás elektromágneses térrel történő besugárzásakor azt elnyeli, azaz „rezonál” vele. Erre csak NMR aktív atommagok képesek, azaz, olyan atommagok, amelyek rendszáma páratlan és így spinkvantumszáma nem egész szám. Ilyen a hidrogén, vagy egyes izotópok (például a szén 13-as, vagy a nitrogén 15-ös izotópja).

A molekulaserkezet-meghatározást, így a fehérjeszerkezetek megismerését az ún. kémiai eltolódás teszi lehetővé, vagyis az, hogy az atommagok rádiófrekvenciás elnyelése nagyban függ a környezettől, a fehérjén belüli elhelyezkedésüktől, a körülöttük levő elektronoktól és atommagoktól. Mivel egy fehérjén belül még így is sok hasonló atommag van, amelyek a spektrumon hasonló jelet adnak, átlapoló csúcsokat létrehozva, ezért szükséges lehet a fehérjét ¹³C és ¹⁵N izotópokkal jelölni; valamint lehetséges a mérést több dimenzióban is végrehajtani³⁴.

Az NMR spektroszkópiát jellemzően 35 kDa alatti fehérjék esetében használják, azaz nagyjából 350 aminosavnál kisebb fehérjékhez, bár az irodalomban fellelhető néhány, ennél nagyobb fehérje szerkezetének a meghatározása is ezzel a módszerrel³⁵. Az eljárás

hátránya tehát, hogy jobbra kisebb szerkezetek megismerését teszi lehetővé, éppen ezért a transzmembrán fehérjék nagy részére nem alkalmazható, hiszen ezen fehérjék mérete általában nagyobb, akár több száz kDa is lehet. Előnye ugyanakkor, hogy a fehérje oldatban, fiziológias körülmények között is használható, akár a sejten belül, natív körülmények között^{36,37}, valamint a fehérje dinamikája, rendezetlen régiók, fehérje-fehérje kölcsönhatások is vizsgálhatók NMR-rel³⁸.

2.4. Homológia modellezés

A homológia modellezés során a fehérje szekvenciáját egy megfelelő templat szerkezetre illesztve próbálják meghatározni a fehérje szerkezetét. Míg globuláris fehérjék esetében a modellezés teljes mértékben automatizálható, transzmembrán fehérjék esetében eddig nem tudtak automatizált módszert készíteni³⁹.

A modellezés folyamata a következő: a rendelkezésre álló megoldott szerkezetek szekvenciáiból kiválasztják azokat, amelyek a kérdéses fehérjének homológjai lehetnek, a lekereső és találati szekvenciákból illesztést készítenek. A szekvenciaillesztés minősége alapvetően meghatározza az eredményül kapott modell minőségét⁴⁰. A következő lépés a modell elkészítése: az illesztésben nem azonos aminosavakat kicserélik a lekereső szekvencia megfelelő aminosavaira. Amennyiben a teljes szerkezet nem fedi le a szekvenciát, a kimaradó részeket hurokrégióként modellezik. Lehetőség van a modell további finomítására is: iteratíván a legjobb modellek illesztései felhasználhatók újabb illesztések elkészítésére⁴¹.

Általánosan elmondható, hogy minél nagyobb a szekvenciák közötti azonosság, annál pontosabb a kapott modell⁴¹. Nagy pontosságú modellek általában legalább 50%-nál nagyobb azonosságú fehérjék esetében készíthetők. A közepes pontosságú modellek a 30-50% azonosságú illesztések alapján készülnek. Kis pontosságúnak általában a 30% alatti szekvenciaazonossággal rendelkező modelleket nevezzük. Nem minden kutatáshoz van feltétlenül szükség nagy pontosságú modellre, vannak olyan feladatok is, amelyekhez a rosszabb minőségű modellek is felhasználhatók, például új, kismolekulás ligandok keresése esetében⁴². Transzmembrán fehérjék esetében a nagyon kicsi, akár 20% alatti szekvenciaazonosságú szerkezetek is felhasználhatók homológia modellezésre a membránrégiók sajátos aminosav-összetétele miatt, főleg, ha foldfelismerő eljárással a fehérje foldja, azaz a gerincet alkotó peptidváz közelítő szerkezete meghatározható⁴¹.

Ha a kiszámolt modellt egy ismert fehérje szerkezetéhez szeretnék hasonlítani, akkor a modell minőségének meghatározására legegyszerűbb az átlagos négyzetes gyökeltérés (*root mean square deviation*, RMSD) mérőszámot használni:

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta_i^2} \quad (1)$$

Az RMSD kiszámításához a modellezett és referenciaszerkezetet egymásra illesztik, majd az egymásnak megfeleltetett N darab atom közötti távolságokból (δ_i) kiszámolják az RMSD értéket.

2.5. Szerkezeti genomikai célpontok kiválasztása

A szerkezeti genomika egyik fontos kérdése annak kiválasztása, hogy mely fehérjék 3D szerkezetét próbálják megoldani, hiszen egy-egy fehérjeszerkezet megoldása rengeteg időt és pénzt emészt fel. Számos különböző megközelítés lehetséges annak meghatározására, hogy milyen fehérjékkel érdemes foglalkozni.

2.5.1. Lehetséges megközelítések

Orvostudományi szempontok alapján történő választás esetén olyan célpontok szerkezetének megoldása a cél, amelyek betegségek kezelésének vagy kialakulásának megértéséhez járulhatnak hozzá. Ez természetesen nemcsak humán, hanem más olyan élőlények fehérjéi közötti választást is jelent, amelyek befolyásolják az emberi szervezet működését, például vírusok, baktériumok, stb^{43,44}. A megoldott szerkezetek gyógyszertervezés során is felhasználhatóak⁴⁵.

Célpontok azonosítása lehetséges biofizikai, vagy más kísérleti szempontok figyelembevételével is, ekkor azokat a fehérjéket részesítik előnyben, amelyek szerkezete a tapasztalatok alapján nagy valószínűséggel megoldható⁴⁴.

További lehetőség egy kiválasztott modellorganizmushoz tartozó minél több fehérje szerkezetét megoldani, bár nyilvánvaló, hogy ezzel távolról sem fedhető le a lehetséges szerkezetek száma, ugyanakkor mégis sok különféle fehérjéről kaphatunk szerkezeti információt⁴⁴.

Több szerkezeti genomikai centrumban lehetőség van a tudományos közösségnek is kéréseket megfogalmazni. Ezek szerkezetének meghatározását általában a centrumokkal

kapcsolatban levő, együttműködő kutatócsoportok kérték, olyan fehérjékre, amelyeket tanulmányoztak⁴⁶.

2.5.2. Szerkezeti lefedettség növelése

A bioinformatikai analízisek, modellezések szempontjából igen fontos módszer a minél nagyobb szerkezeti lefedettségre való törekvés. Ebben az esetben a fehérjék közül kiválogatják azokat, amelyeknek vagy már létezik megoldott 3D szerkezete, vagy a meglevő szerkezetek alapján modellezhetők. A többi fehérje között meghatározzák a homológ kapcsolatokat, például klaszterező program használatával, és az alapján határozzák meg a célpont fehérjék listáját, hogy megoldott szerkezetük mennyivel járulna hozzá a szerkezeti tér lefedettségének növekedéséhez^{42,46}.

2.6. Tanulógép eljárások

A mesterséges intelligencia egyik ága a tanulógép eljárások, amelyek lehetővé teszik, hogy a számítógépek adatokból szabályszerűségeket ismerjenek fel, ezek alapján modellt építsenek az adott problémára, és a modell segítségével predikciót végezzenek új adatokra. A modell az adatok függő (célváltozó) változója és független (magyarázó) változói között keres összefüggést. A célváltozó értékére szeretnénk becslést adni, a magyarázó változókkal pedig az adathalmaz elemeit jellemezzük.

A tanulógép eljárások két nagy csoportra oszthatók: felügyelt és nem felügyelt tanításra. A felügyelt tanítás címkézett adatokon történik, azaz a tanítóhalmaz minden rekordhoz tartalmazza a célváltozó helyes értékét is. A felügyelt tanítás is kétfelé osztható: klasszifikációra, ami az adatok csoportosítását jelenti (a célváltozó diszkrét értékeket vehet fel), valamint regresszióra, amikor a függő változó folytonos értékeket vesz fel. A leggyakrabban alkalmazott eljárások a támogatóvektor gépek, a döntési fák és a véletlen erdő módszer. A nem felügyelt tanítású eljárások nem címkézett adatokban rejlő csoportosulást, vagy azok belső, rejtett szerkezetét próbálják feltárni. Ide tartozik például a klaszterezés, anomáliaérzékelés, főkomponens-analízis, de rejtett Markov modell (*hidden Markov modell*, HMM) is betanítható nem, felügyelt tanítással¹⁷. Mivel a munkához felügyelt tanítású, klasszifikációt végző algoritmusokat használtam, a továbbiakban ezekkel fogok foglalkozni.

2.6.1. Adatok előfeldolgozása

A megfelelő pozitív és negatív adatok összegyűjtése, valamint megfelelő címkézése kulcsfontosságú egy sikeres modell felállításához. Az előfeldolgozás nagyban függ attól, hogy milyen modellt építünk. Vannak olyan algoritmusok, amelyek használatának előfeltétele az adatok átalakítása például normálással vagy centrálással, különben egyes változók nagyobb súlyt kapnak a végső modellben. A változókat logaritmikus skálára is transzformálhatjuk. Dimenziócsökkentést hajthatunk végre például főkomponens-elemzéssel, amely egymással kölcsönhatásban álló változók lineáris kombinációit állítja elő, úgy, hogy az adatokban levő variancia a lehető legkevesbé csökkenjen, így csökkentve a magyarázó változók számát.

Előfordul, hogy nem minden változóhoz áll rendelkezésre adat az egyedekről. Erre a helyzetre kétféle megoldási módszer létezik: vagy teljesen kihagyjuk az adott egyedet vagy változót, amihez hiányoznak az értékek, vagy becsléssel pótolni tudjuk a hiányzó adatokat. Becslést az adott változó értékei alapján, esetleg más változóval kölcsönhatás alapján adhatunk. Néhány algoritmus esetében erre sincs szükség, ezekkel építhető modell hiányos adatsorok esetében is⁴⁷.

2.6.2. Teljesítmény mérése

Az osztályozó modellek kétféle becslést szolgáltatnak: egy kategorikus besorolást, valamint egy valószínűséget, hogy az adott egyed mennyire tartozik abba a kategóriába. A kategorikus besorolás egyik legnépszerűbb kiértékelése az ún. igazságmátrix (1. táblázat). Ebben a kiértékelés során a becsült és a valós osztályokat vetjük össze. Az igazságmátrix segítségével többféle mérőszámot kiszámolhatunk.

valós osztály	becsült osztály	
	pozitív	negatív
pozitív	VP	ÁN
negatív	ÁP	VN

1. táblázat. Igazságmátrix.

(VP: valós pozitív, ÁN: álnegatív, ÁP: álpozitív, VN: valós negatív becslés.)

$$Pontosság = \frac{VP + VN}{VP + \hat{A}N + \hat{A}P + VN} \quad (2)$$

A pontosság előnye, hogy könnyen átlátható, értelmezhető, összeveti az összes helyesen becsült értéket az összes kimenetellel. Hátránya, hogy nem ad információt arról, hogy milyen típusú hibákat követ el az algoritmus a becslés során. Másik probléma, hogy nem veszi figyelembe az osztályok mintabeli arányát, azaz amennyiben az egyik csoport felülreprezentált a mintában, és pontosságra optimalizáljuk a modellt, a pontosság magas lesz, ha a modell a becslések nagy részét a túlreprezentált osztályba teszi.

$$Kappa = \frac{O-E}{1-E} \quad (3)$$

A Kappa statisztika a mért (O) és várt (E) pontosságot veti össze. A Kappa -1 és 1 közötti értéket vehet fel. A 0 érték azt jelenti, hogy nincs összefüggés a megfigyelt és becsült osztályok között; az 1 érték jelentése, hogy teljes egyezés van. A negatív értékek fordított összefüggést mutatnak a becslés és megfigyelés között.

$$Szenzitivitás = \frac{VP}{VP + \hat{A}N} \quad (4)$$

A szenzitivitás vagy felidézés (valós pozitív arány) azt méri, hogy a pozitív osztályba tartozó egyedek közül hány került a becslés során valóban a pozitív osztályba.

$$Specifititás = \frac{VN}{VN + \hat{A}P} \quad (5)$$

A specifititás (valós negatív arány) a szenzitivitás ellentétét méri, vagyis azt, hogy a negatív osztályba tartozó egyedek közül hányat soroltunk be valóban a negatív osztályba.

A modell betanítása során általában egyensúlyt kell találnunk a szenzitivitás és a specifititás között, mindkettőt egyszerre igen ritkán lehet tökéletesen optimalizálni. Tudnunk kell megállapítani, hogy az adott probléma esetén melyik fontosabb a számunkra: hogy ne legyenek olyan találatok, amik hamis pozitívak, vagy inkább a hamis negatívokat kell elkerülnünk.

$$Precizitás = \frac{VP}{VP + \hat{A}P} = \frac{VP}{\sum \text{pozitív predikciók}} \quad (6)$$

A precizitás vagy pozitív prediktív érték (*positive predictive value*, PPV) azt mutatja meg, hogy mekkora a pozitív eredmény megbízhatósága, vagyis egy pozitív predikció esetén annak a valószínűsége, hogy az adott egyed valóban a pozitív osztályba tartozik. A precizitás 0 és 1 közötti értéket vehet fel.

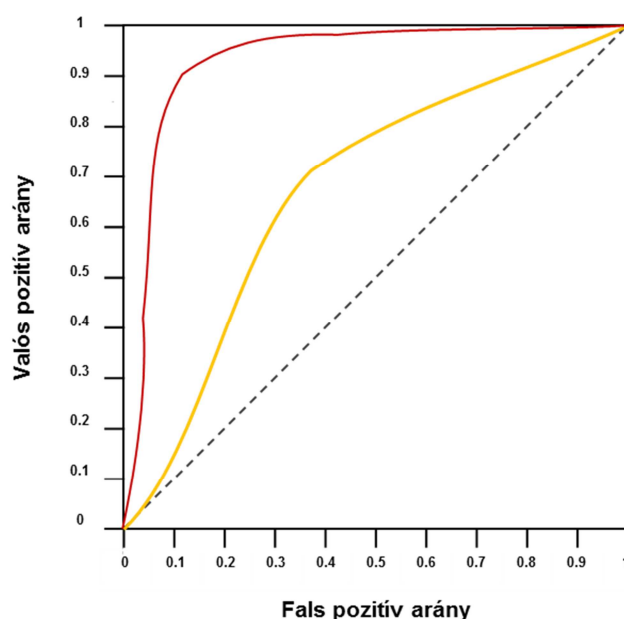
$$\text{Negatív prediktív érték} = \frac{VN}{VN + \hat{A}N} = \frac{VN}{\Sigma \text{negatív predikciók}} \quad (7)$$

A negatív prediktív érték (*negative predictive value*, NPV) a precizitáshoz hasonlóan a negatív predikciók megbízhatóságát méri, szintén 0 és 1 között.

$$F_1 = 2 * \frac{\text{precizitás} * \text{szenzitivitás}}{\text{precizitás} + \text{szenzitivitás}} \quad (8)$$

Az F_1 mérőszám bináris osztályozókra a precizitás és a szenzitivitás harmonikus középértékeként számolható ki. A hátránya, hogy a valós negatív értékekkel nem számol, ezért a teljesítmény mérésére csak korlátozottan alkalmas. Értéke 0 és 1 közötti lehet, ahol 0 a legrosszabb, az 1 pedig a legjobb értéket jelenti.

A vevő működési karakterisztika (*receiver operating characteristics*, ROC) görbén (4. ábra) a szenzitivitást ábrázoljuk az $1 - \text{specificitás}$ (fals pozitív arány) érték függvényében. Különböző küszöbértékek mentén kiszámolva a specificitást és a szenzitivitást, és ezeket az értékeket ábrázolva kapjuk meg a görbét.

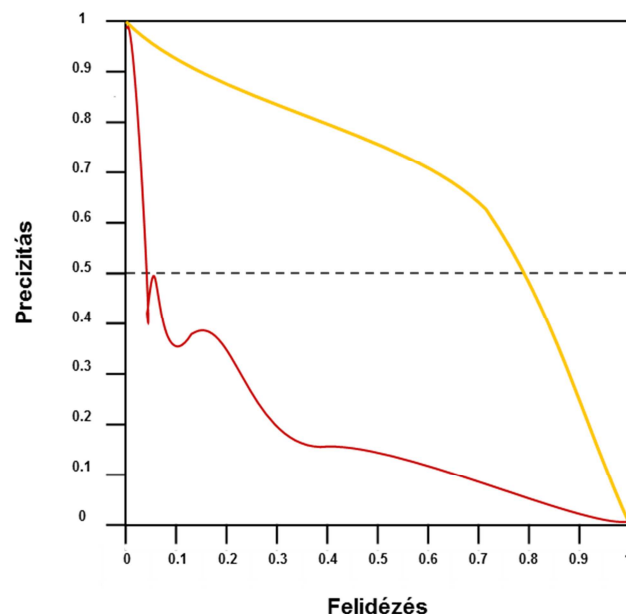


4. ábra. Vevő működési karakterisztika görbék.

A ROC görbe alapján megállapítható a modell jósága. Egy teljesen véletlenszerűen osztályozó modell a (0,0) és (1,1) pontok között egy egyenes vonalként fut (4. ábra, szürke szaggatott egyenes). Ebben az esetben a valós és az álpozitív arány természetesen meg fog egyezni. Az optimális modell ettől a bal felső sarok felé mozdul el (4. ábra, piros és sárga görbe). Különböző modellekből származó görbék összevetésével a modellek (eltérő paraméterek, algoritmusok, magyarázó változók) is összehasonlíthatóvá válnak.

A ROC görbe alatti területet (*area under the curve*, ROC-AUC) is használják a modellek összehasonlításra. Egy nagyobb terület általában jobb modellt is jelent, de ez nagyban függ a görbe lefutásától is, egy tökéletes osztályozóhoz $AUC = 1$ tartozik.

Egy másik gyakran használt ábrázolásmód a precizitás – felidézés görbe (*precision – recall*, PR, 5. ábra). Egy modell optimalizációja esetén ezek között általában fordított összefüggés van, azaz, ahogy nő a precizitás, úgy csökken a felidézés, és viszont; a kettő között egyensúlyt kell találni, annak a figyelembevételével, hogy milyen céllal készítjük a modellt, egy álnegatív vagy egy álpozitív becslést tartunk nagyobb problémának. A ROC görbéhez hasonlóan, erre is definiálhatunk görbe alatti területet (PR-AUC).



5. ábra. Precizitás – felidézés görbe.

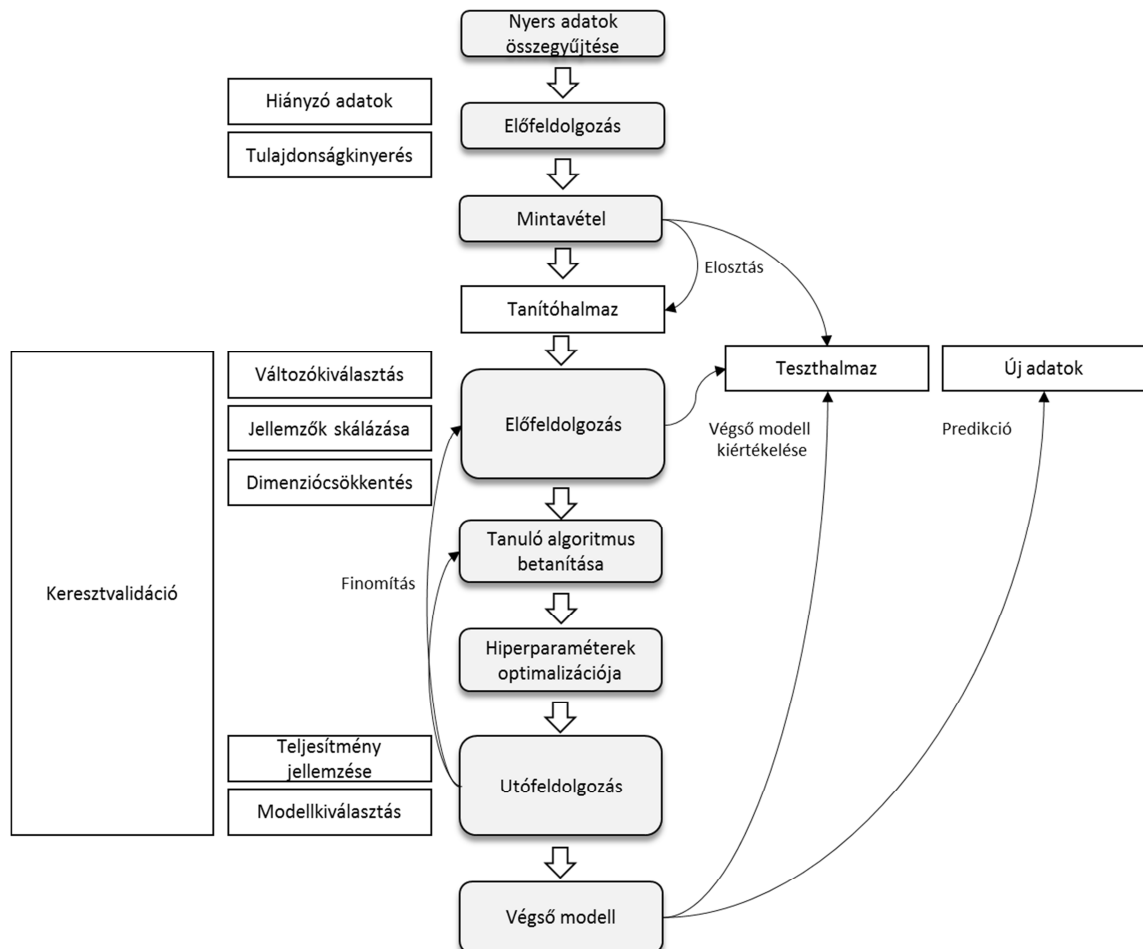
(Piros és sárga: PR görbék, szürke: random osztályozó görbéje).

$$MCC = \frac{VP \cdot VN - \hat{A}P \cdot \hat{A}N}{\sqrt{(VP + \hat{A}P)(VP + \hat{A}N)(VN + \hat{A}P)(VN + \hat{A}N)}} \quad (9)$$

A Matthews korrelációs együttható (*Matthews correlation coefficient*, MCC) bináris osztályozás esetén használható. Mivel mind a valós és álpozitív illetve álnegatív csoportokat is figyelembe veszi, ezért nagyon kiegyensúlyozatlan adatkészletek esetén is elég jól használható a modell teljesítményének értékelésére.

2.6.3. A modellek betanítása

A felügyelt tanítású módszerek általános betanítási és validálási sémáját az 6. ábra mutatja be: az adatok összegyűjtése után meghatározzuk a magyarázó változók értékeit, azaz kvantifikálunk olyan tulajdonságokat, amelyekről úgy gondoljuk, befolyásolhatják az egyedek osztályba sorolását.



6. ábra. Felügyelt tanítású algoritmusok tanítása

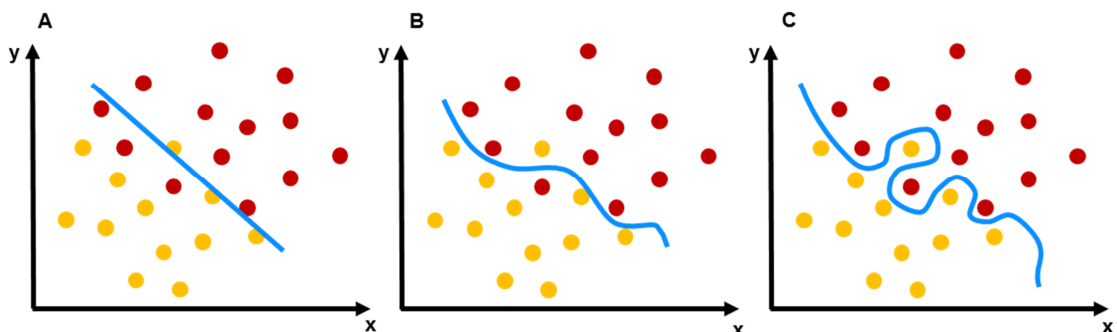
(Sebastian Raschka munkája alapján, <http://sebastianraschka.com/>)

Minden paraméter optimalizációt a tanítóhalmazon végzünk. A paramétereket legáltalánosabban keresztvalidációval határozzuk meg: a tanítóhalmazt k egyenlő részre osztjuk, és k -szor ismételve a folyamatot, $k-1$ halmazt összevonva betanítjuk a modellt, majd a kimaradt részen teszteljük. Így k darab modellt kapunk, a pontosságot pedig a modellek pontosságának az átlaga adja. Amennyiben k megegyezik a tanítóhalmaz elemeinek számával, *leave-one-out* validációról beszélünk, mivel mindig egy kimaradt egyeden tesztelünk. Ha alkalmazunk változókiválasztást – azaz a magyarázó változók számát a modell betanítása előtt csökkentjük, hogy a korreláló, redundáns tulajdonságokat eltávolítsuk, azt is ezekben a lépésekben tesszük meg.

Az előfeldolgozás után betanítjuk az algoritmust, majd elvégezzük a hiperparaméterek optimalizációját. Az utófeldolgozás során kiszámoljuk a teljesítmény, amire több mérőszámot (pontosság, PR-AUC, ROC-AUC) használhatunk. Ide tartozik a modellkiválasztás, azaz a legjobb teljesítményű modell meghatározása is. Az utófeldolgozás eredményétől függően vissza lehet térni az előfeldolgozásra, például további, feleslegesnek vélt változókat kihagyva, majd a modellt újratanítva jobb teljesítményű modellt létrehozni.

2.6.4. Alul-és túlillesztett modellek

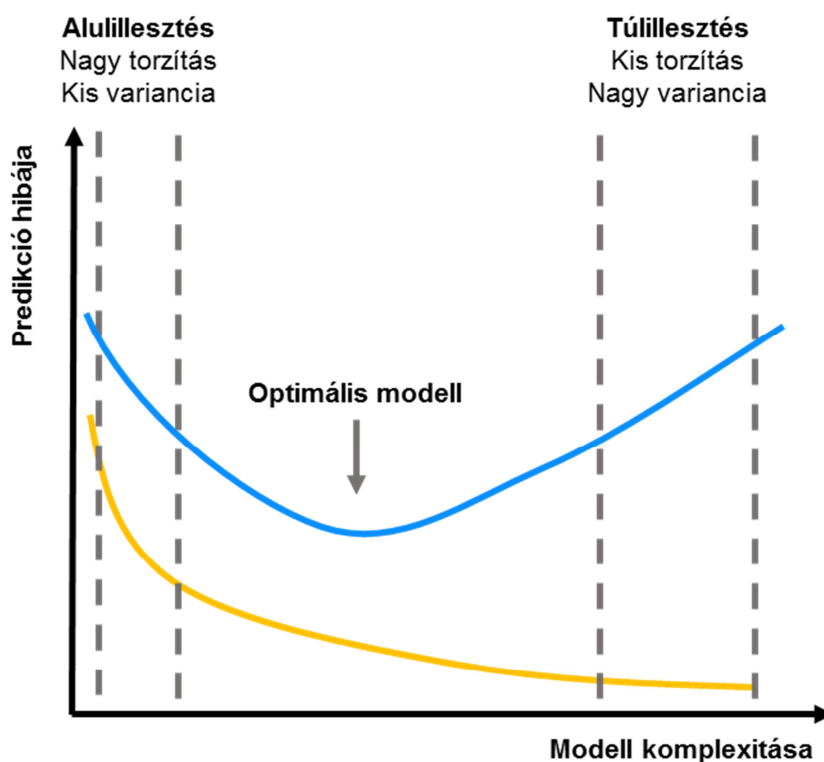
A modellek létrehozásakor a két leggyakoribb hiba, amit elkövethetünk, ha a modellt alul-illetve túlillesztjük a meglevő adatokra. Alulillesztés esetén (7. **ábra**A ábra) a modell nem írja le elég jól a meglevő adatokat, ennek következtében még a tanítóhalmaz elemeire sem ad megfelelő becslést, új adatokra általánosítani pedig még kevésbé tud.



7. **ábra.** Modellillesztések.

A) alulillesztett modell, B) megfelelő modell, C) túlillesztett modell.

Túlillesztés esetén (7. **ábra**C ábra) ennek az ellenkezője történik, a modell nagyon pontosan becsüli a tanítóhalmaz elemeit, ugyanakkor nem általánosít jól, a független teszhalmazon nagy hibát ad. A modell nemcsak az adatokban levő szerkezetet, hanem az adatokban előforduló zajt, random ingadozásokat is megtanulja.



8. ábra. Alul-és túlillesztett modellek hibája.

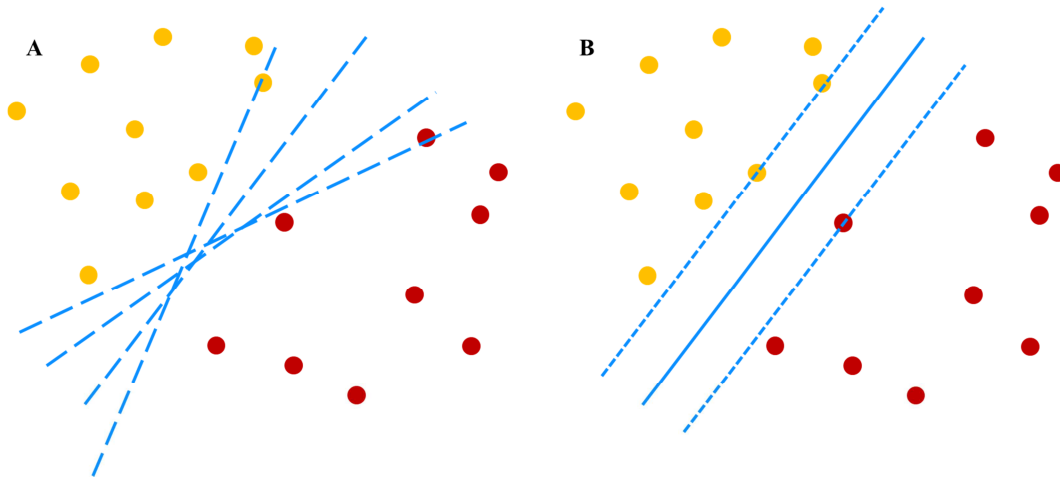
(Sárga: tanítóhalmazon, kék: teszhalmazon mért hiba.)

Elkerülni mindkét esetet vagy keresztvalidációval, vagy teszhalmaz kihagyásával lehet, amelyen a végső modell teljesítményét mérjük (8. ábra). A modell komplexitása függvényében kiszámoljuk az egyes modellek predikciós hibáját, valamint ábrázoljuk a varianciát és a torzítást. Az optimális modellt ott kapjuk, ahol a predikció teljes hibája a legkisebb.

2.6.5. Támogatóvektor gépek

A támogatóvektor gép⁴⁸ (*support vector machine*, SVM) az egyik leggyakrabban használt a tanulógep eljárások közül, mivel gyors tanítást és predikciót tesz lehetővé. Az adatokat egy n dimenziójú vektorként tároljuk, mindegyik dimenzióban egy-egy jellemzőjüket tárolva. A jellemzők numerikusak, a kategorikus jellemzőket mesterséges (*dummy*) változókká konvertáljuk, amelyek 0 vagy 1 értéket vehetnek fel. Az algoritmus lényege, hogy egy olyan hipersíkot keres az n dimenziós térben, amely lehetővé teszi a

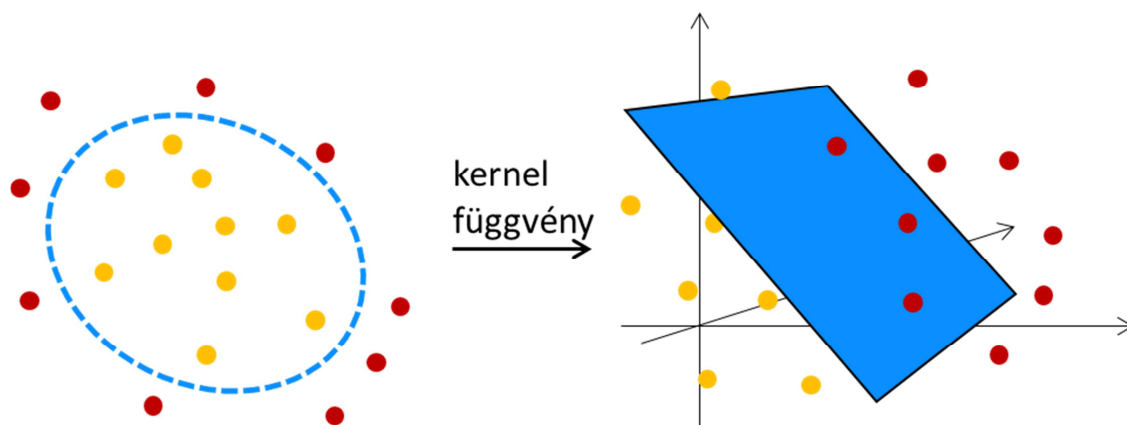
különböző osztályokba tartozó egyedek elválasztását oly módon, hogy a lehető legnagyobb elválasztást (margót) biztosítsa (9. ábra). Ez azt jelenti, hogy az így talált hipersík bármelyik osztály legközelebbi pontjától a lehető legnagyobb távolságra helyezkedik el.



9. ábra. Adatpontok elválasztása támogatóvektor géppel.

A) Különböző megoldások egy lineárisan elválasztható feladatra. B) A feladat legnagyobb margót biztosító megoldása

Az elválasztás nem minden esetben valósulhat meg lineárisan, ekkor különböző ún. kerneleket használhatunk, ami azt jelenti, hogy a magyarázó változók terét transzformáljuk valamilyen függvény szerint, és az így kapott térben keresünk hipersíkot (10. ábra). Ekkor előfordulhat az, hogy a transzformált térben találunk olyan hipersíkot, ami az eredeti térben nem lineáris. Gyakran használt kernelek például polinomiális vagy a Gauss-féle radiális bázisfüggvényt használó kernelek.



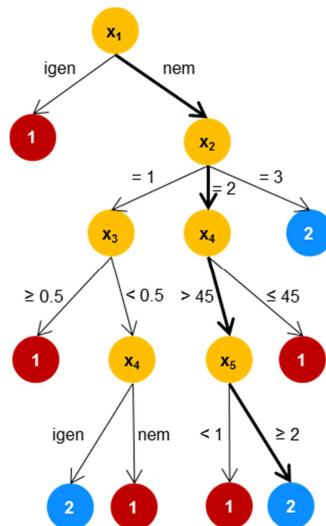
10. ábra. Az adatok transzformálása és elválasztása kernel függvény segítségével.

A módszer előnye, hogy nem probléma számára a jellemzők multikollinearitása, azaz a változók egymás közötti korrelációja, nagyszámú jellemzővel is jól működik, valamint különböző kernelek használatával lineárisan nem szeparálható adatok elválasztása is lehetségessé válik. Ugyanakkor, ha a tanítóhalmaz nagyszámú egyedet tartalmaz, a betanítása hosszú időt vehet igénybe, valamint feladattól függően probléma lehet, hogy a predikció eredményét nem valószínűségekkal, hanem binárisan (egyik vagy másik csoport) adja vissza, bár vannak módszerek arra, hogy az osztályozás megbízhatóságát számszerűsítsük.

2.6.6. Véletlen erdő

A véletlen erdő modell⁴⁹ döntési fák sokaságából áll. A döntési fa olyan modell, ami a tanítóhalmaz alapján kérdéseket fogalmaz meg az egyedek jellemzőivel kapcsolatban, és predikció esetén a kérdésekre adott válaszok alapján végzi el az osztályozást (11. ábra). A döntési fák matematikai értelemben gráfok. Használatuk egyik előnye, hogy mindenféle (ordinális, nominális, intervallum, arány) skálán elhelyezkedő adatokat kezelni tud, és nincs szükség azok további normálására vagy standardizálására sem.

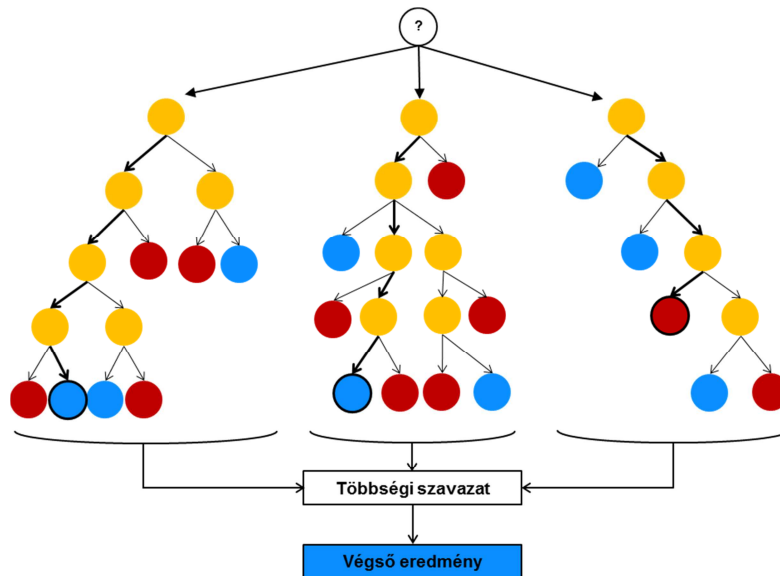
A döntési fákat úgy kell felépíteni, hogy a lehető legtömörebben, legkevesebb lépésben jussanak optimális eredményre. Ehhez a gyökértől kiindulva kell az egyes lépésekben olyan tesztet végeznünk, ami a legnagyobb információnyereséget illetve nyereségarányt biztosítja, azaz az adatokat a legnagyobb mértékben szét tudja választani. Önmagában egy döntési fa hajlamos a túltanulásra, ezt a teszthalmaz alapján a fa mélységének szabályozásával (metszéssel) szokták szabályozni. A fák felépítéséhez illetve metszéséhez többféle algoritmus használható.



11. ábra. Egyszerű döntési fa.

$x_1, x_2, \dots$ a vizsgált változók, 1, 2 a lehetséges kimenetek. A vastag vonalak egy osztályozást mutatnak be.

A véletlen erdők felépítéséhez nagyszámú különböző döntési fát állítunk elő, úgy, hogy a tanítóhalmazból visszatevéses mintavétellel minden fához külön tanítóhalmazt hozunk létre, felépítjük a fákat, majd mindegyikkel becslést adunk a kimenetekre. A modell azt az eredményt adja vissza végső válaszul, amelyikre a legtöbb szavazatot adják a fák (12. ábra). Minden fa elkészítés előtt véletlenszerűen mintavételezik a rendelkezésre álló magyarázó változókat, és ebből a halmazból választják ki azt, ami egy adott elágazásnál a legjobb elválasztást biztosítja. Az algoritmus betanításakor optimalizálni kell, hogy a rendelkezésre álló változókból hányat válasszon ki a fák létrehozásakor.



12. ábra. Véletlen erdő sematikus modellje.

A hatékonyság több paraméteren múlik: a létrehozott fák számán, illetve azon, hogy az egyes fák mennyire korrelálnak egymással – a nagyfokú korreláció rontja a hatékonyságot. A véletlen erdő modell általában védett a túlillesztéssel szemben⁵⁰, a modell magas komplexitása (a generált fák száma illetve mélysége) nem eredményez túlillesztést, a fák nagy száma miatt. A predikció hibája egy értékhez konvergál, így a fákon általában metszést sem alkalmaznak. Mindez természetesen nagyban függ az adatoktól, azok belső struktúrájától, megbízhatóságától is, a túlillesztés sem lehetetlen, ebben az esetben a fák számának a növelésével, mélységük csökkentésével, illetve a változók számának csökkentésével lehet javítani a modellen.

A véletlen erdő előnye, hogy nagy, magas dimenziójú adathalmazon is gyors betanítást tesz lehetővé, valamint nemcsak klasszifikációra, hanem regresszióra is használható. Az egyes paraméterek fontosságát is becsülni tudja a változókhoz tartozó értékek véletlenszerű összekeverésével: egy olyan változó, amelyik megváltozása esetén a predikció hibája nagyobb mértékben nő, nagyobb pontszámot kap. Nagyszámú változó esetén csak a fontosnak ítélt változókat megtartva új, nagyobb hatékonyságú modell építhető.

2.7. Fehérjék kristályosíthatóságát becsülő eljárások áttekintése

A fehérjék kristályosításának elősegítésére több predikciós eljárást kifejlesztettek. Ezek között vannak nagyon egyszerű, csak pár paramétert használó módszerek, de akár több

száz, a kristályosítási folyamat lépéseit külön-külön figyelembe vevők is. A programok egy része interneten érhető el, más részük letölthető, és lokálisan futtatható.

Az első módszert 2006-ban publikálták, ez volt az OB-score⁵¹, ami a GRAVY pontozást (*grand average of hidropathy*) és a fehérje izoelektromos pontját vette alapul. Az izoelektromos pont a fehérje oldhatóságával, tisztíthatóságával van kapcsolatban, a hidrofóbicitás felhasználása pedig elsősorban a membránfehérjék kiszűrését tette lehetővé, mint nehezen kristályosítható fehérjéket.

A CRYSTALP⁵² és annak frissített változata, a CRYSTALP2⁵³, aminosavak, di-és tripeptidek előfordulási gyakoriságát, valamint az izoelektromos pont értékét és a fehérje hidrofóbicitását veszi figyelembe, a predikcióhoz neurális hálót használ. A CRYSTALP2 létrehozásához használt adathalmazt használták egy másik program, az SVMCRYST betanítására is. Ez utóbbi szekvenciából meghatározott adatokat használ bemenetként egy SVM alapú modellhez, mint különböző fiziko-kémiai tulajdonságok az AAIndex⁵⁴ adatbázis alapján, valamint másodlagos szerkezetbecslésből⁵⁵ származó értékeket.

A 2011-ben publikált PPCPred⁵⁶ az első olyan módszer volt, amely már nemcsak magára a kristályosításra, hanem a kristályosítási folyamat különböző lépéseire is becslést adott, ezzel megadva a lehetséges szűk keresztmetszeteket is a kísérlet folyamán. A szekvenciák alapján számolt értékeken kívül felhasználtak másodlagos szerkezet, rendezetlen régió és az oldószer által elérhető relatív felszín becslést is.

Az XtalPred⁵⁷ és az azt megújító XtalPred-RF⁵⁸ a TargetTrack adatbázisból összegyűjtött adatokból épített modellek. A változók között megtalálható a felszíni egyenetlenség, hidrofóbicitás, a felszíni aminosavak oldallánc entrópiája, és a felszín aminosav-összetétele, az XtalPred-RF véletlen erdő módszeren alapul.

Az egyik legújabb eljárás a 2016-ban megjelent Crysalis⁵⁹, amely több feladatot is ellát. Egyrészt becslést ad egy adott fehérje kristályosíthatóságára, másrészt olyan pontmutációkat keres, amelyek valószínűleg növelhetik az egyes lépések sikerességét, valamint annotálja az egyes szekvenciákat másodlagos szerkezet, transzmembrán hélixek, rendezetlenség, oldószer általi feltételezett elérhetőség szerint. A módszer támogatóvektor gépeket használ az egyes lépések, valamint a teljes folyamat becslésére, ilyen módon kétrétegű predikciót építve.

A legújabb eljárás a Cryst⁶⁰, amely a szekvenciából meghatározott változókon túl a UniProt⁶¹ adatbázisból származó tulajdonságokat, mint szignálszekvencia vagy diszulfidhidak megléte, sejten belüli lokalizáció, kölcsönhatások, poszttranszlációs módosulás, alegységek, stb. is felhasznál. Az egyik változata több másik predikciós eljárás eredményeit integrálva ad új becslést.

Az egyetlen, transzmembrán fehérjékre specifikusan kifejlesztett eljárás a MEMEX⁶², amely 2008-ban jelent meg és naiv Bayes modellt használ a predikcióra. A felhasznált változók között globális értékek, mint a GRAVY index, a fehérje töltése, a transzmembrán hélixek valamint a loopok száma és elhelyezkedése, a transzmembrán hélixek közötti interakciót meghatározó GxxG- és GxxxG-szerű motívumok 100 transzmembrán szakaszon elhelyezkedő aminosavra eső gyakorisága, valamint az aminosavak, illetve aminosav csoportok gyakorisága.

Célkitűzések

A human transzmembrán fehérjék igen kis részének oldották meg a szerkezetét. Ahhoz, hogy a lehető leggyorsabban – azaz a legkevesebb fehérje kristályosításával – fedjük le ezen fehérjék szerkezeti terének minél nagyobb részét, azaz minél több fehérjéről készíthető legyen számítógépes modell, elsőként szükség van arra, hogy tudjuk, hogy melyek azok a humán transzmembrán fehérjék, amelyek szerkezetét meghatározták már úgy, hogy a szerkezet minden transzmembrán régiót tartalmazzon, valamint, hogy a meglevő humán és nem humán fehérjék alapján mely fehérjék modellezhetők minden membránrégió lefedésével. A többi fehérjéről el kell döntenünk, hogy a szerkezetük meghatározása hány másik fehérje számára jelenthetne templát szerkezetet. A szerkezeti tér legjobb lefedését a legtöbb homológgal rendelkező fehérjék szerkezetének meghatározása biztosítaná. Humán transzmembrán fehérjékre eddig két, a mostani dolgozatban bemutatotthoz hasonló vizsgálatot végeztek^{42,63}, azonban ezeknek az eredménye a megjelenésük óta (2011 és 2013) nem frissült, annak ellenére, hogy az elmúlt időszakban folyamatosan, ha nem is drasztikus mértékben, de nőtt a meghatározott szerkezetek száma.

A munka során a célunk volt egy olyan adatbázist előállítani, amely szabadon elérhető, tartalmazza a megoldott szerkezetű vagy modellezhető humán transzmembrán fehérjéket, valamint egy olyan listát, amely útmutatásul szolgálhat ahhoz, hogy milyen

módon lehetne minél kevesebb fehérjeszerkezet megoldásával a lehető legnagyobb szerkezeti tér lefedettséget elérni. Célul tűztük ki azt is, hogy a folyamat során mindvégig figyelembe vegyük a fehérjék becsült topológiáját, hogy elkerüljük azt, hogy különböző számú transzmembrán régiókat tartalmazó fehérjéket egymáshoz hasonlóknak és modellezhetőnek állítsunk be. Az adatbázist olyan módon kívántuk előállítani, hogy követve a készítéséhez felhasznált adatbázisok frissülését, a későbbiekben automatikusan frissíthető legyen.

A munka második része egy predikciós program fejlesztése. Kristályosítási predikcióra nagyszámú program áll rendelkezésre, ugyanakkor ezek közül mindössze egy olyan van, amely transzmembrán fehérje specifikus. Ez utóbbi 2008-ban jelent meg, és azóta sem frissítették, tanították újra a modellt, így – a kísérleti protokollok fejlődése miatt – egészen biztosan elavult, hiszen nem használták fel az azóta szerzett új tudást. A programban nemcsak a kristályosítási folyamat utolsó lépésére, hanem az előtte levő szolubilizálási és tisztítási lépésekre is becslést szeretnénk adni. A klónozási és expressziós lépések kihagyását az indokolja, hogy ezek sokkal inkább függenek a fehérjét kódoló DNS szakaszoktól⁶⁴, valamint az expressziós rendszertől, mint az aminosav szekvenciától, azonban sajnos a szerkezeti genomikai kísérletekhez használt DNS szekvenciákat az adatbázisok általában nem tartalmazzák.

3. Módszerek

3.1. Felhasznált programok

3.1.1. HHblits

A HHblits⁶⁵ program egy rejtett Markov modellen (HMM) alapuló szekvenciakereső- és illesztő algoritmus, amely speciális HMM profilokat használ. HMM-en alapuló algoritmusokat gyakran használnak homológia modellezéshez, mivel nagyon szenzitív keresést tesznek lehetővé, távoli homológokat is megtalálva. A program alkalmas párhuzamos futtatásra, ezzel gyors lekeresést és illesztést lehetővé téve, valamint saját lekeresési adatbázis is építhető hozzá. Bemenetként a lekeresendő szekvenciát és az adatbázist kell megadni.

Saját lekeresési adatbázist készítettem két adatszettből: egyrészt az összes humán transzmembrán fehérjéből, valamint a PDBTM adatbázisból. Ezeket az adatbázisokat a program dokumentációjában leírt módon építettem föl.

A program futtatásait mind a két adatbázison való lekeresés esetén ugyanúgy végeztem, és az értékelés is hasonló volt. Futtatáskor profil-profil összehasonlítást használtam, globális keresést, valamint minden lekérdezési-templát szekvenciapárt csak akkor vettem figyelembe, ha ezt visszafelé, azaz akkor is megtalálta, ha a templát szekvenciát használtam lekérdezőként. A két találat *E-value* értékének a geometriai közepét vettem, ahogy a dokumentáció ajánlotta, az így kapott kombinált *E-value* értéke nem haladhatta meg a 10^{-10} értéket.

Mivel az algoritmus topográfiát vagy topológiát nem ellenőriz, ezért ezt szükséges volt külön megtenni, annak érdekében, hogy két fehérjét csak akkor fogadjunk el homológoknak, ha a transzmembrán régiók száma megegyezik. Így minden találatnál leellenőriztem, hogy a transzmembrán régiók átfednek-e, valamint hogy a fehérjék szekvenciaazonossága legalább 25% legyen.

3.1.2. CCTOP

A Constrained Consensus Topology²³ (CCTOP) algoritmussal transzmembrán fehérjék topológiája becsülhető. Már régebben kimutatták, hogy a transzmembrán fehérjék topológiabecslését javítani lehet új kísérletes adatok és egyéb megszorítások bevonásával^{19,17}, valamint több, már meglevő algoritmus eredményének

összevonásával⁶⁶. Az eljárás az eddigieknél pontosabban, mindössze 1,5% hibával dönti el, hogy egy fehérje transzmembrán-e, vagy nem, ehhez három becselő algoritmusból legalább kettőnek kell transzmembránként azonosítani a fehérjét. A topológiabecsléhez 10 algoritmus eredményét kombinálja, és ezekből ad végső eredményt. Ezenkívül használ kísérletes eredményeket a TOPDB^{67,68} adatbázisból, valamint információt a membrán azonos oldalán lokalizálódó fehérjedoménekről a TOPDOM^{21,22} adatbázisból. Amennyiben elérhető a fehérjéhez megoldott homológ szerkezet, a PDBTM⁶⁹ adatbázisból ennek a topológiáját is felhasználja a becslés során. Végeredményben az algoritmus mintegy 84% átlagos pontosságot ér el a topológiabecslésben⁹.

A programot nem a webszerveren keresztül, hanem lokálisan futtattam a TargetTrack fehérjeszekvenciáira. Az alapbeállításokon nem változtattam.

3.1.3. TMFoldrec

A TMFoldrec⁷⁰ topográfiai adatot is felhasználó foldfelismerő algoritmus, amelyet specifikusan transzmembrán fehérjékre fejlesztettek ki. Az algoritmus nem szekvenciahasonlóság alapján határozza meg a fehérje lehetséges foldját, hanem az összes, azonos számú transzmembrán régiót tartalmazó szerkezetre megpróbálja felilleszteni a transzmembrán régiókat, majd minden szerkezet esetében becslést ad a fehérje energiájára, egy speciális statisztikus potenciál függvény felhasználásával, amely figyelembe veszi a fehérje környezetét, például lipid-fehérje kölcsönhatásokat. Ha a legkisebb energiájú szervezetre adott pont meghalad egy küszöbértéket, akkor a program a szerkezetet mint legvalószínűbb foldot adja vissza a fehérjére.

A program futtatásához bemenetként a fehérje szekvenciáját és topológiáját, valamint a lehetséges szerkezeteket tartalmazó szerkezetkönyvtárat kell megadni. A munka során a legalább 0.5 valószínűséggel kapott szerkezeteket vettem figyelembe.

3.1.4. BLAST

A Basic Local Alignment Search Tool⁷¹ (BLAST) algoritmus az egyik leggyakrabban használt szekvenciakereső program. Nagyon gyorsan futtatható, több millió szekvenciából álló adatbázisok esetében is könnyen, gyorsan használható. Az algoritmusnak több változata is elérhető, a munka során a *blastp* programot használtam, amely fehérje szekvenciát keres fehérjékből álló adatbázison.

A lekeresési adatbázist a felhasználó állíthatja össze, a munkához a TargetTrack ún. *trial* szekvenciáiból álló adatbázist használtam. A lekeresés során legalább 95% azonosságú fehérjetalálatokat fogadtam el, amelyek *E-value* értéke maximum 10^{-10} volt, és a fehérjék illesztése lefedte a lekereső fehérje összes transzmembrán régióját.

3.1.5. CD-HIT

A CD-HIT⁷² programcsomag szekvenciákat csoportosít a hasonlóságuk alapján. A beadott szekvenciák közül a leghosszabbal kezdve, a következő szekvenciákat ehhez hasonlítja, és megállapítja, hogy az éppen vizsgált szekvencia redundáns (hasonlít valamelyik reprezentatív szekvenciához) vagy egy új reprezentatív szekvencia. A programcsomag többféle algoritmust tartalmaz, ezek közül a CD-HIT algoritmust, valamint a 2D-CD-HIT-et használtam. Az előbbi egy szekvenciaadatbázist klaszterez adott azonosságra (40%-os azonosságot használtam). Az utóbbi 2 adatbázist összevetve kiválogatja az első adatbázisból azokat, amik a másodikban levő szekvenciákhoz kevesebb, mint adott arányban (40%-ot használtam) hasonlítanak.

3.1.6. caret csomag

A caret⁷³ csomag (*classification and regression training*) az R programnyelv egyik kiegészítő csomagja. Tartalmazza a legtöbb olyan algoritmust, amely tanulógépek sikeres betanításához szükséges, valamint 233 lehetséges modellt. Támogatóvektor gépekhez többféle kernel (lineáris, radiális, polinomiális, stb.), változókiválasztás, mintavételezés, validációk (egyszeres vagy ismételt keresztellenőrzés, leave-one-out validáció, bootstrap), különböző előfeldolgozási lehetőségek érhető el. Az algoritmusok a modellek hangolását a beállítások szerint automatikusan végzik az általunk választott mérőszámok (ROC, AUC, pontosság, Kappa, stb.) alapján. A csomagban levő függvények nagy része több szálon is futtatható, meggyorsítva ezzel a megfelelő modell kiválasztását és betanítását.

3.1.7. EVFold

Az EVFold⁷⁴ (<http://evfold.org>) egy korrelált mutációs analízisen alapuló *ab initio* szerkezetbecslő algoritmus. A program többszörös illesztésekkel dolgozik, ezek minden oszloppárjára kiszámolja az aminosavak együttes előfordulását, majd a kapott eredmények alapján becslést ad, hogy a szekvenciában mely aminosavak állhatnak egymással kölcsönhatásban, és ez alapján szerkezetmodelleket épít. A modelleket

molekuladinamikai szimulációval finomítja, majd az elkészült modelleket rangsorolja a modell valószínűsége szerint. A megfelelő eredmény eléréséhez a többszörös illesztésnek több száz, de inkább több ezer szekvenciát kell tartalmaznia. A módszert 500 aminosavnál rövidebb szekvenciákra, doménekre optimalizálták, a honlapon egyes domének előre kiszámolt szerkezete is elérhető.

Bemenetként megadható a fehérje szekvenciája, amelyhez a program saját maga készít többszörös illesztést, vagy használhatjuk az adott domén Pfam⁷⁵ adatbázisból vett többszörös illesztését is. További lehetőség, hogy saját illesztést készítünk, és azt töltjük fel a honlapra. Transzmembrán fehérjékre lehetséges saját topológiai adatok megadása is, ezzel felülírva a program saját topológia becslését.

3.2. Felhasznált adatbázisok

3.2.1. Human Transmembrane Proteome

A Human Transmembrane Proteome⁹ (HTP, <http://htp.enzim.hu/>) adatbázisban a Consensus Constrained TOPology²³ (CCTOP) algoritmussal meghatározott humán transzmembrán fehérjék és azok kísérletes módszerekkel meghatározott, vagy becsült topológiája van összegyűjtve. Az adatbázis mindegyik topológiához megbízhatósági értéket rendel, az egyes fehérjék topológiája vizuálisan is megjelenítésre kerül. Az adatok az egyes fehérjékről külön-külön, valamint egyesítve is letölthetők a honlapról, a letöltött fájlok tartalmazzák az összes adatot, amit a végső topológia összeállításához a CCTOP algoritmus felhasznált.

Az adatbázis v1.4 (2016.06.24.) verziója 5 609 transzmembrán fehérjét tartalmaz.

3.2.2. Protein Data Bank of Transmembrane Proteins

A Protein Data Bank of Transmembrane Proteins^{69,76} (PDBTM, <http://pdbtm.enzim.hu/>) adatbázis a PDB-ben található transzmembrán fehérjék szerkezeteit tartalmazza, amelyekről a TMDET⁷⁷ algoritmus segítségével határozták meg, hogy az adott fehérjeszerkezet hogyan illeszkedik a membránba. A TMDET algoritmus először kiválogatja a vírus és pílus fehérjéket, valamint a túl rövid láncokat tartalmazó rekordokat. Ezután meghatározza a fehérje biológiailag aktív oligomer szerkezetét, valamint kiszűri azokat a fehérjeláncokat, amelyek biológiailag nem lehetséges kölcsönhatásokkal kapcsolódnak egymáshoz. Majd kiszámolja a lipidek számára

hozzáférhető felületet, és a membránt alkotó síkokat a fehérjeszerkezetre illeszti. Amennyiben ez az illesztés megfelelő, a fehérjét transzmembránként ismeri fel⁷⁸.

Az adatbázis rendszeresen frissül, a 2016.06.03. verzió összesen 2 829 transzmembrán szerkezetből, köztük 2 464 α -helikális és 344 β -hordó szerkezetből áll.

3.2.3. TargetTrack

A TargetTrack⁷⁹ (<http://sbkb.org/tt/>) adatbázis a *Protein Structure Initiative* program alatt jött létre. A cél egy átlátható, követhető adatbázis létrehozása volt, ahol több szerkezeti genomikai centrum munkája egységesen nyomon követhető. Minden egyes célpontról tartalmaz alapvető információkat: melyik centrum, és mikor kezdett hozzá, milyen protokollokkal folyt a munka, milyen státuszban van (klónozott, expresszált, stb.), vagy melyik lépésnél és miért szakadt félbe végleg a munka az adott szekvencián. Az egyes célfehérjékhez több próbálkozást (*trial*) is tartalmazhat, amennyiben az eredeti szekvenciát mutálták, csonkolták, vagy egyéb módon módosították a sikerebb kristályosítás érdekében.

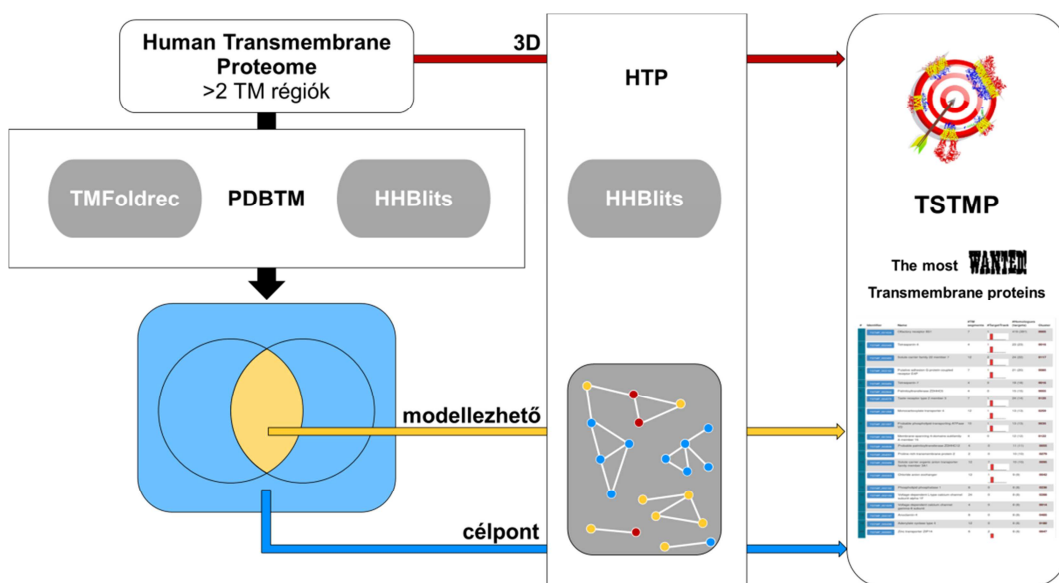
Az adatbázis része a 2004-ben létrehozott TargetDB⁸⁰ és PepcDB (*Protein expression purification and crystallization Database*), amelyeket 2010-ben olvasztottak a TargetTrackbe. A felhasznált 2016.06.18 verziójú adatbázis körülbelül 330 ezer célfehérjét tartalmazott, majdnem 2 millió *trial* szekvenciával.

3.3. Az adatbázis felépítéséhez használt munkafolyamat kidolgozása

Az adatbázis építéséhez olyan munkafolyamatot készítettem, amellyel automatikusan letölthetők a forrásadatbázisok legújabb verziói, futtathatók a különböző programok, valamint feldolgozhatók azok eredményei. A munkafolyamat nemcsak a legújabb adatok esetében, hanem visszamenőleg az előző évek adataira is futtatható, a megfelelő verziójú adatbázisok felhasználásával. Segítségével az adatbázis folyamatosan frissíthető, követve a felhasznált adatbázisok újabb verzióinak megjelenését. A munkafolyamatban levő scripteket Perl programnyelvben, valamint bash shell-ben írtam meg. A folyamat sémája a 13. ábraán látható.

A munka első lépéseként a HTP adatbázisból kiválogattam azokat a transzmembrán fehérjéket, amelyek legalább két transzmembrán régióval rendelkeznek. Az egy

transzmembrán régióval rendelkező fehérjét a vizsgálat során nem vettük figyelembe, többek között azért, mert ezeknél a fehérjéknél a legvalószínűbb, hogy hibás a fehérje transzmembránként való meghatározása (a becselő algoritmusok hajlamosak a szekvencia elején levő hidrofób tulajdonságokat mutató szignálszekvenciát transzmembrán régióként azonosítani⁸¹), másrészt az egy transzmembrán régióval rendelkező fehérjék globuláris része általában megfelelően levágható és külön kristályosítható.



13. ábra. Az adatbázis elkészítésének folyamatábrája.

A többi fehérjét a következőképpen soroltam klaszterekbe: a HHblits programot futtattam a HTP adatbázison, az így kapott találatokat a fentebb leírt módon válogattam ki. Ezzel a módszerrel tulajdonképpen egy gráfot határoztam meg, amely összefüggő komponenseit, mint klasztereket definiáltuk.

A PDBTM adatbázisból kigyűjtöttem mindazokat a humán transzmembrán fehérjeszerkezeteket, amelyeknél a szerkezet az összes transzmembrán régiót lefedi. Ezeket a fehérjét a megoldott 3D szerkezettel rendelkező (3D) kategóriába soroltam.

Lefuttattam a TMFoldRec és HHblits kereséseket a PDBTM adatbázis ellenében. Ha egy fehérjéhez a megfelelő elfogadási kritériumok alapján mindkét algoritmus adott találatot, akkor a fehérjét modellezhetőnek (*modelable*) vettem. Ahhoz, hogy ezekhez legalább 25%-os azonossággal rendelkező modelleket készíthessünk, nem szükséges további fehérjeszerkezetek megoldása.

A transzmembrán fehérjék maradéka a célpont (*target*) fehérjék. Ezekhez a rendelkezésre álló szerkezetekből legfeljebb csak nagyon rossz minőségű modellek készülhetnének, ezért szükséges további fehérjeszerkezetek megoldása.

Minden fehérjéhez, amennyiben elérhető volt, BLAST kereséssel rendeltem megfelelő TargetTrack azonosítót és állapotot. Így az adatbázis böngészése során könnyebb nyomon követni a fehérjék kristályosítási folyamatának állapotát.

3.4. A predikciós programhoz szükséges adatok előállítása

A predikciós programhoz szükséges volt olyan adathalmazok összeállítása, amely minden becsülni kívánt lépéshez tartalmaztak sikeres, valamint sikertelen próbálkozásokat. Ezeket az adatokat a TargetTrack, valamint a PDBTM adatbázisokból nyertem (14. ábra).

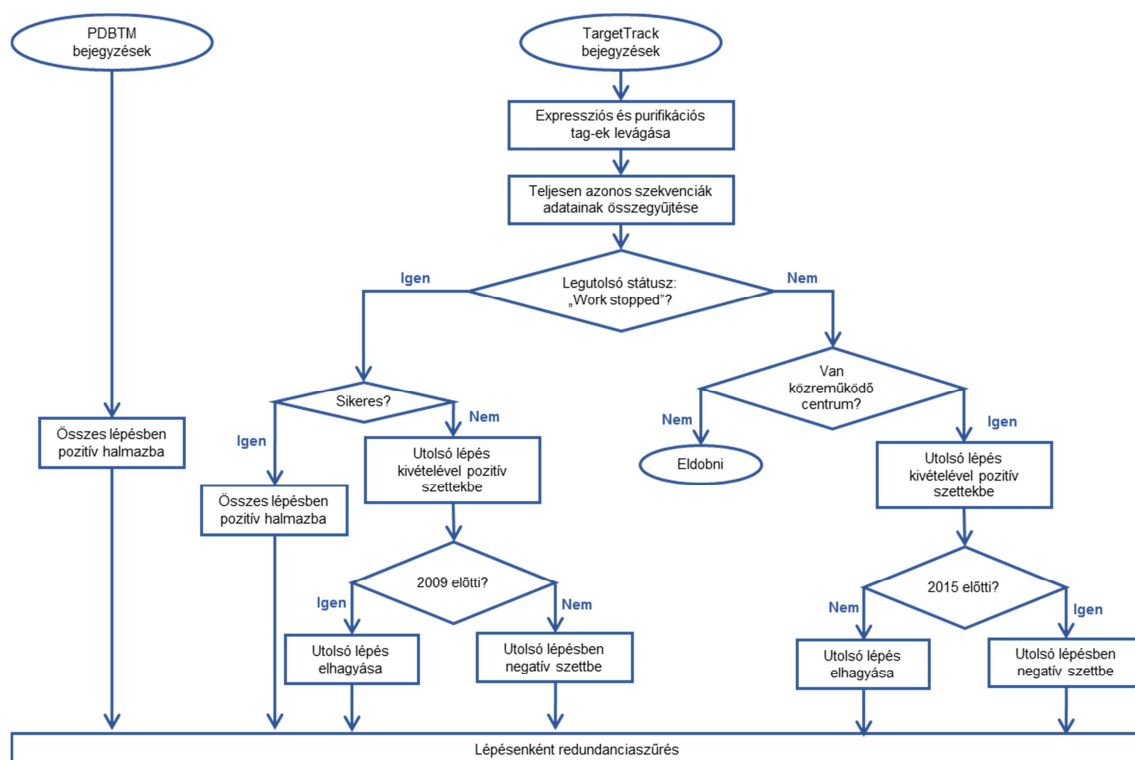
A PDBTM adatbázis feldolgozása egyszerű volt: mivel ebben már teljes szerkezetek vannak, az itt megoldott szekvenciák már minden lépésen sikeresen átmentek. Minden szekvenciát felhasználtam a szolubilizálási, valamint a tisztítási lépések pozitív adathalmazához. A kristályosítási lépéshez értelemszerűen csak a röntgenkristallográfiával meghatározott szerkezeteket használtam.

A TargetTrack adatbázis feldolgozása nagyobb kihívást jelentett. A résztvevő szerkezeti genomika centrumok eltérő minőségű és mennyiségű adatot töltöttek fel, ezért bizonyos feltételezésekkel, vagy elhanyagolásokkal kellett élnem a feldolgozás során. Ezek a feltételezések a szakirodalomban már máshol is használtak, és elfogadottak voltak.

Az egyik feltételezés az volt, hogy mivel az adatbázis hosszú időt ölel fel (az alapjául szolgáló TargetDB 2004-ben indult), ezért a sikertelen próbálkozások meghatározását a módszerek fejlődése nagyban befolyásolhatja. Ennek elkerülésére a sikertelen próbálkozásokból csak a 2009 utániakat vettük figyelembe.

Másik probléma volt, hogy az együttműködő centrumok, kutatócsoportok kiválhattak, megszűnhettek; a PSI program részei lejártak, így nemcsak azért nem folytatódhatott egy kísérlet, mivel technikailag sikertelen volt. Abban az esetben, ha ezt külön jelezték (pl. „*End of PSI-1*”), akkor egyértelmű volt, hogy ezekkel a bejegyzésekkel nem foglalkozom. Ha nem jelölték, akkor csak azokat a sikertelen próbálkozásokat vettem figyelembe, amik még aktív együttműködő partnerekhez tartoztak.

Harmadszor, a centrumok nem mindig jelezték, ha egy kísérlet sikertelen, ezért több ezer bejegyzés évek óta „folyamatban van” az adatbázis szerint. Ebben az esetben azzal a feltételezéssel éltem, hogy a 2015 előtti folyamatban levő kísérletek az utolsó lépésben sikertelenek voltak.



14. ábra. Az adatok összegyűjtésének folyamata.

A TargetTrack bejegyzések esetében először levágtam az expressziós és purifikációs tageket a szekvenciákról, amennyiben ezek jelen voltak, majd összegyűjtöttem azokat a szekvenciákat, amik így teljesen azonossá váltak. Ezek után az azonos szekvenciákhoz tartozó bejegyzéseket összesítve dolgoztam fel. Amennyiben a bejegyzések közötti legutolsó státusz „*Work stopped*” volt, megnéztem, hogy az utolsó lépés sikeres volt-e. Ha igen, az összes vizsgált lépésben, amit az adott fehérjével elértek, a pozitív adathalmazba tettem. Ha nem volt sikeres, csak az utolsó előtti lépésig soroltam a pozitív halmazba. Az utolsó lépésben pedig a 2009 előtti bejegyzéseket elhagytam, és csak a 2009 utániakat soroltam a lépésnek megfelelő negatív halmazba.

Ezek után a kristályosítási folyamat minden vizsgált lépésére a következőket futtattam: a negatív példák közül a 2D-CD-HIT algoritmussal kiválogattam azokat, amelyek 40% azonossággal hasonlítottak egy, a pozitív adathalmazban levő példára. Majd a két

halmazban levő szekvenciákat külön-külön klasztereztem 40%-os azonosságra a CD-HIT algoritmussal.

Végül minden egyes példára kiszámoltam a magyarázó változókat az általam írt programmal, a ProtParam⁸² webszerver segítségével és a CCTOP algoritmus topológiabecslésének felhasználásával, amit lokálisan futtattam. A kiszámított változók a következők voltak (zárójelben a változók száma, összesen 97):

- Hossz, tömeg (2)
- ProtParam: instabilitás index, izoelektromos pont, GRAVY, féléletidő (4)
- Transzmembrán és extramembrán régiók aránya, összhossza, transzmembrán régiók száma (5)
- Leghosszabb transzmembrán és extramembrán régió (2)
- Aminosavak aránya a teljes fehérjében, a transzmembrán illetve az extramembrán régiókban (60)
- Aminosav-csoportok (töltött, nem töltött, pozitív, negatív, alifás, aromás, ként tartalmazó, hidroxilcsoportot tartalmazó) aránya a teljes fehérjében, a transzmembrán illetve az extramembrán régiókban (24)

3.5. A modell betanítása

A modellek az R programnyelv caret csomagjával végeztem. Háromfajta SVM kernelt használtam (lineáris, radiális, polinomiális), valamint véletlen erdő algoritmust, hogy ezek eredményét összehasonlítsam. A folyamat során alulmintavételezést használtam, mivel a pozitív és negatív halmazok számossága közel sem volt egyenlő, a negatív halmazban majdnem kétszer annyi egyed található, mint a pozitívban. A kiegyensúlyozatlan halmazokra az egyik megoldás, hogy véletlenszerűen kihagyunk annyit a nagyobb halmazból, hogy a számossága megegyezzen a másik adatkészlet számosságával. A probléma másik kezelési módjának azt választottam, hogy a modelleket nem pontosságra, hanem ROC-AUC illetve PR-AUC értékekre optimalizáltam.

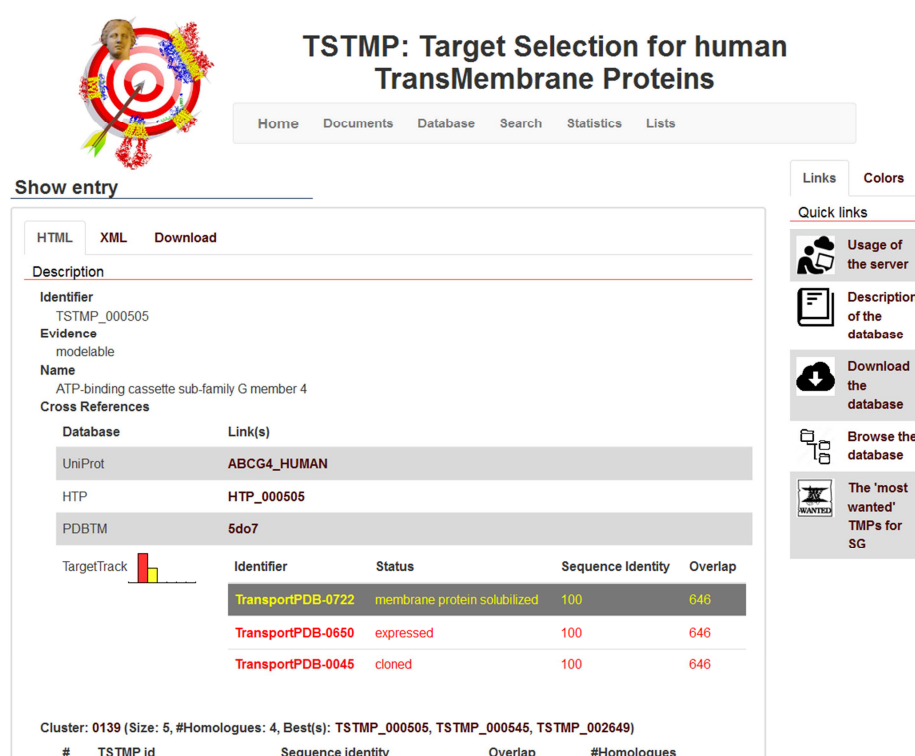
Az SVM-hez és a véletlen erdőhöz különböző előfeldolgozási lépéseket használtam. Az SVM a betanítás során igényli, hogy az adatok azonos skálán legyenek, ezt legegyszerűbben azok standardizálásával lehet elérni, azaz minden egyes változóhoz

tartozó értékekből kivonjuk azok átlagát, valamint osztunk a szórásával. A véletlen erdő nem igényel ilyen lépéseket, ott az adatokat eredeti formájukban meghagytam.

4. Eredmények

4.1. Adatbázis létrehozása, elérése

A munkafolyamat futtatásával kapott adatbázist interneten szabadon elérhetővé tettük (TSTMP, <http://tstmp.enzim.ttk.mta.hu/>). A honlap készítése során az általánosan elterjedt Apache, PHP, MySQL hármast alkalmaztuk, továbbá a Bootstrap CSS keretrendszert. Lehetővé tettük az adatbázis keresését és böngészését különböző azonosítók (UniProt⁶¹, PDB⁸³, HTP), szekvencia, klaszterek, TargetTrack bejegyzések vagy homológok száma alapján.



TSTMP: Target Selection for human TransMembrane Proteins

Home Documents Database Search Statistics Lists

Show entry

HTML XML Download

Description

Identifier
TSTMP_000505

Evidence
modelable

Name
ATP-binding cassette sub-family G member 4

Cross References

Database	Link(s)
UniProt	ABCG4_HUMAN
HTP	HTP_000505
PDBTM	5do7

TargetTrack

Identifier	Status	Sequence Identity	Overlap
TransportPDB-0722	membrane protein solubilized	100	646
TransportPDB-0650	expressed	100	646
TransportPDB-0045	cloned	100	646

Cluster: 0139 (Size: 5, #Homologues: 4, Best(s): TSTMP_000505, TSTMP_000545, TSTMP_002649)

#	TSTMP id	Sequence identity	Overlap	#Homologues
---	----------	-------------------	---------	-------------

Links **Colors**

Quick links

- Usage of the server
- Description of the database
- Download the database
- Browse the database
- The 'most wanted' TMPs for SG

15. ábra. Az elkészült adatbázis egy rekordja (UniProt ID: ABCG4_HUMAN).

Az egyes bejegyzésekben (15. ábra) megjelennek kereszthivatkozások más adatbázisokra, amelyek megkönnyítik a hivatkozott adatbázis megfelelő bejegyzésének az elérését. 3D fehérjék esetében elérhető a fehérje szerkezetének PDBTM azonosítója, modellezhető fehérjék esetében azok a szerkezetek, amelyeket homológiamodellezéshez mint templát szerkezeteket ajánlunk. Ez utóbbiak esetében meghatároztuk a fehérje és az egyes PDBTM szerkezetek illesztéséből a transzmembrán régiók átfedésének a mértéket, valamint a szekvenciaazonosságot. Amennyiben a fehérjéhez tartozik

TargetTrack bejegyzés, ezt is feltüntetjük, az átfedéssel és a szekvenciaazonossággal együtt, azonban ekkor ezeket az értékeket a teljes fehérje alapján számoljuk.

4.2. Az XML fájlok jellemzése

A honlapról letölthetők az egyes rekordok XML fájlformátumban. Lehetőség van az összes rekord, a különböző kategóriákba (3D, modellezhető, célpont) tartozó, valamint a felhasználó által választott rekordok egyidejű letöltésére is. Az adatbázishoz használt XML séma definíciója elérhető a honlapon: <http://tstmp.enzim.ttk.mta.hu/data/tstmp.xsd>).

Az XML fájlok tartalmazzák (16. ábra) az egyes fehérjék UniProt ID és AC azonosítóit, nevét, szekvenciáját, hosszúságát és a fehérje transzmembrán régióinak számát. A 3D kategóriában levő fehérjékre megadtuk az azokhoz tartozó PDBTM azonosítókat, ahogyan a modellezhető csoportba tartozó fehérjékre is. Az utóbbiakhoz továbbá feltüntettük a transzmembrán régiókban levő szekvenciaazonosság és átlapolás mértékét is. Amennyiben a fehérjéhez a fent tárgyalt feltételek mellett találtunk TargetTrack bejegyzést vagy bejegyzéseket, úgy ezeknek az azonosítója, legutolsó státusza, valamint a két szekvencia átlapolása és azonossága is bekerült a fájlokba. A fájl ezeken kívül tartalmazza még az adott fehérje homológjait, valamint annak a klaszternek az azonosítóját, amelybe a fehérje tartozik. A fehérjének a homológokkal való azonossága és átlapolása szintén bekerült a fájlba.

```

<?xml version="1.0" encoding="utf-8"?>
<TSTMPItem id="TSTMP_000505" evidence="modelable">
  <Copyright>Copyrighted by the Institute of Enzymology, CNRS, HAS, see
  http://tstmp.enzim.ttk.mta.hu/data/copyright.txt for details.</Copyright>
  <Name>ATP-binding cassette sub-family G member 4</Name>
  <CrossRef>
    <UniProt id="ABCG4_HUMAN">
      <AC>Q9H172</AC>
      <AC>A8K1B5</AC>
      <AC>Q8WWH0</AC>
      <AC>Q8WWH1</AC>
      <AC>Q8WWH2</AC>
    </UniProt>
    <HTP id="HTP_000505"/>
    <PDBTM id="5do7">
      <Chain id="A" tmsOverlap="124" tmsSequenceIdentity="27"/>
    </PDBTM>
    <TargetTrack id="TransportPDB-0722" sequenceIdentity="100" overlap="646" status="membrane
  protein solubilized"/>
    <TargetTrack id="TransportPDB-0650" sequenceIdentity="100" overlap="646" status="expressed"/>
    <TargetTrack id="TransportPDB-0045" sequenceIdentity="100" overlap="646" status="cloned"/>
  </CrossRef>
  <Cluster id="0139">
    <Homologue id="TSTMP_000546" sequenceIdentity="25" overlap="618"/>
    <Homologue id="TSTMP_000013" sequenceIdentity="72" overlap="646"/>
    <Homologue id="TSTMP_002649" sequenceIdentity="28" overlap="615"/>
    <Homologue id="TSTMP_000545" sequenceIdentity="25" overlap="626"/>
  </Cluster>
  <Sequence length="646" numTM="6" checkSum="d14e3c33f54a5e5de30cf20d38ab4dc4">
    <Seq>MAEKALEAVGCGLGPGAVAMAVTLEDGAEPVLTTHLKKVENHITEAQRFSHLPKRSA
  VDIEFVELSYSVREGPCWRKRGYKTLKCLSGKFCRRELIGIMGPSGAGKSTFMNLAGYRESGM
  KGQILVNGRPRELRTFRKMSCYIMQDDMLLPHLTVLEAMMVSANLKLSEKQEVKKELVTEILTA
  LGLMSCSHTRTALLSGGQRKRLAIALELVNNPPVMFFDEPTSGLDASCFQVVSMLMKSLAQGGR
  TIICTHQPSAKLFEMFDKLYILSQGQCIFKGVVTNLIPYLKGLGLHCPTYHNPADFIEVASGEYG
  DLNPMLFRAVQNGLCAMAEEKSSPEKNEVPAPCPPCPPEVDPIESHTFATSTLTQFCILFKRTFLSI
  LRDTVLTHLRFMSHVIGVLIGLLYLHIGDDASKVFNNNTGCLFFSMLFLMFAALMPTVLTFFLEM
  AVFMREHLNYWYSLKAYYLAKTMADVFPQVVCVVYCSIVYWMGTGQPAETSRFLLFSALATA
  TALVAQSLGLLIGAASNSLQVATFVGVPVTAIPVLLFSGFFVSFKTIPTYLQWSSYLSYVRYGFEGVI
  LTIYGMERGDLTCLERCPCFPREPQSILRALDVEDAKLYMDFLVLGIFFLALRLLAYLVLRYRVKS
  ER</Seq>
  </Sequence>
</TSTMPItem>

```

16. ábra. Az adatbázis egyik rekordját (UniProt ID: ABCG4_HUMAN) leíró XML fájl.

4.3. Az adatbázis jellemzése

4.3.1. Klaszterek

Az adatbázis fehérjéiből 346 klasztert építettünk, amelyek összesen 2 117 fehérjét tartalmaznak. Az egyes fehérjék közötti kapcsolat azt mutatja, hogy az egyik szerkezetének az ismerete segítheti a másik fehérje modellezését. A legnagyobb klaszter 685 G-fehérje kapcsolt fehérjéből áll; amely 24 3D, 239 modellezhető és 422 célpont szekvenciát tartalmaz. A legtöbb célpont homológgal rendelkező fehérje is innen kerül

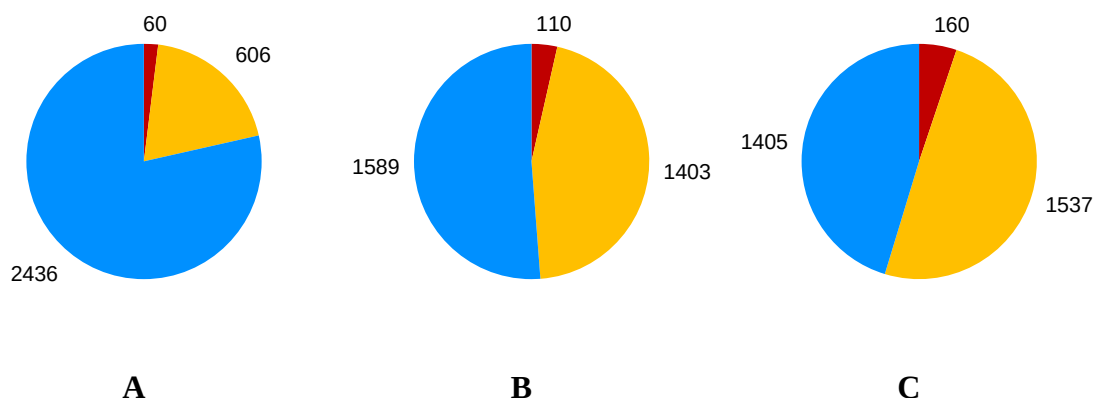
ki: az *Olfactory receptor 8S1* fehérje szerkezetének megoldása majdnem 400 fehérje számára jelenthetne homológia modellezéshez alapul vehető szerkezetet.

4.3.2. 3D kategória

A becslések szerint a transzmembrán fehérjék lehetséges foldjainak mindössze egyötödét sikerült eddig feltérképezni⁴¹, és ez az arány még rosszabb, ha a humán transzmembrán fehérjéket nézzük. A 3 102 fehérjéből összesen 60 rendelkezik az összes transzmembrán régióját lefedő szerkezettel. A 60 fehérjéből 24 ugyanabból a G-fehérje kapcsolt receptorcsaládból származott, 2 fehérjének pedig nincs homológja az adatbázisban. 32 fehérjének van 7 transzmembrán régiója, de ezek együttes figyelembevételével sem lehet minden 7 transzmembrán régióval rendelkező fehérjét megbízhatóan modellezni. További 14 fehérje rendelkezik 4, valamint 7 fehérje 6 transzmembrán régióval. Megvizsgáltuk a megoldott szerkezetek szekvenciáinak hosszát is. A szerkezetek majdnem kétharmada 300 és 500 aminosav közötti hosszúságú volt.

4.3.3. Modellezhető kategória

A modellezhető csoportban 606 olyan fehérje található, amelyekhez legalább egy olyan templát szerkezet van, amelynek szekvencia-azonossága 25%-nál nagyobb (17A ábra). Egy fehérje modellezhetőségének megállapításához mind szerkezeti, mind szekvenciális, mind topológiai információt felhasználtunk. A „Legjobb célpontok” lista (lásd 4.3.4 fejezet) első 50 szerkezetének megoldásával a modellezhetők száma 797-tel emelkedne (606-ról 1 403-ra, 17B ábra), míg az utána következő 50 fehérje további 134 fehérjét tenne modellezhetővé (17C ábra).



17. ábra. A csoportok arányának változása az adatbázisban.

A) Jelenlegi helyzet. B) a „Legjobb célpontok” lista első 50 szerkezetének megoldása után. C) Az első 100 szerkezet megoldása után. (Piros: 3D, sárga: modellezhető, kék: célpont).

4.3.4. Célpont kategória

A célpont fehérjék csoportja tartalmazza azokat a fehérjéket, amelyekhez sem 3D szerkezet, sem megfelelő minőségű homológ szerkezet nem áll rendelkezésre, ezeket a fehérjéket javasoljuk kristályosításra.

Ezeket a fehérjéket sorba rendeztük aszerint, hogy hány célpont homológjuk van, és a listából kiválasztottuk a legtöbb homológgal rendelkező fehérjét, ez lett a „Legjobb célpontok” lista (18. ábra) első eleme. Az eredeti listából eltávolítottuk ezt a fehérjét és az összes homológját, és újból sorba rendezés után addig ismételtük az eljárást, amíg csak olyan fehérjék maradtak, amelyeknek nincs homológjuk.

The most **WANTED** TMPs for structural genomics

The most **WANTED** list shows, how the human transmembrane proteome could be modeled with the minimum number of steps (i.e. with the smallest number of crystallized proteins). We have sorted the proteins based on how many target homologues they have, then selected the one with the most relative. Then we removed all entries from the list that could be modeled based on the selected protein(s) and iterated this process until every human transmembrane protein were covered.

Number of entries: 335 Download selected item(s)

☐	#	Identifier	Name	#TM segments	#TargetTrack	#Homologues (targets)	Cluster
☐	1	TSTMP_001634	Olfactory receptor 8S1	7	1	418 (396)	0005
☐	2	TSTMP_002349	Tetraspanin-4	4	1	23 (23)	0016
☐	3	TSTMP_002483	Solute carrier family 22 member 7	12	2	23 (21)	0119
☐	4	TSTMP_003152	Putative adhesion G protein-coupled receptor E4P	7	1	21 (20)	0084
☐	5	TSTMP_003483	Tetraspanin-7	4	0	18 (18)	0016
☐	6	TSTMP_002844	Palmitoyltransferase ZDHHC5	4	0	14 (14)	0055
☐	7	TSTMP_004575	Taste receptor type 2 member 3	7	1	24 (14)	0125
☐	8	TSTMP_001268	Monocarboxylate transporter 4	12	1	13 (13)	0263

18. ábra. A „Legjobb célpontok” lista.

4.3.5. Könnyen és nehezen modellezhető fehérjék

A munka során elkészítettünk további két listát is a modellezhető fehérjékből. Ezekhez a listákhoz szükség volt az egyes modellezhető fehérjékhez javasolt PDBTM szerkezetek transzmembrán régióinak átlapolására és szekvenciaazonosságára. Az egyik lista tartalmazza azokat a modellezhető fehérjéket, amelyekhez van legalább egy modellszerkezet, ahol a transzmembrán régióban levő azonosság legalább 50%. Ezt a listát nevezzük a könnyen modellezhető fehérjék listájának, ami 237 rekordot tartalmaz. Ezekből a fehérjékhez a rendelkezésre álló szerkezetek alapján legalább közepes minőségű modellek építhetők viszonylag egyszerűen.

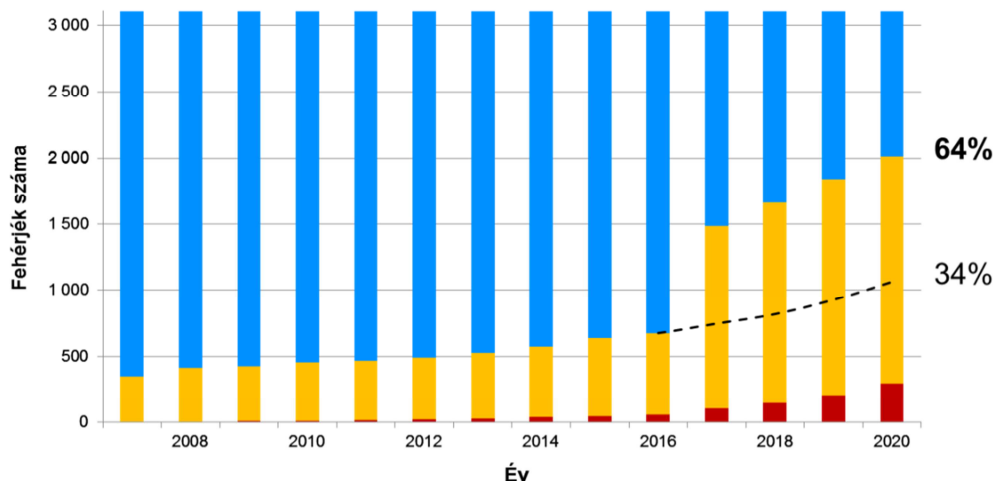
A másik listában ehhez hasonlóan a nehezen modellezhető fehérjéket gyűjtöttük össze, amelyekhez jó modellt nagyobb kihívás építeni. Ez a lista azon fehérjéket tartalmazza, amelyekhez nincs olyan PDBTM bejegyzés, ahol a transzmembrán régióban való azonosság 25%-nál nagyobb lenne. A listán 50 fehérje szerepel.

4.4. Az adatbázis statisztikái

4.4.1. Számított állapotok

A létrehozott munkafolyamatot nemcsak a forrásadatbázisok jelenlegi állapotára, hanem az elmúlt évek adataira is lefuttattam. Az adatok alapján elmondható, hogy a korábbi évek feltételezett „Legjobb célpontok” listájáról egyik fehérje szerkezetét sem oldották meg, ugyanakkor sok közülük modellezhetővé vált. Az is nyilvánvaló az adatok alapján, hogy inkább olyan szerkezeteket oldanak meg, amelyekhez már van valamilyen homológ fehérje. Ezek megoldása nyilvánvalóan egyszerűbb a már meglévő protokollok, kísérletek alapján. A 2016-ban 3D csoportba sorolt fehérjék közül mindössze 26,7% volt azoknak az aránya, amelyek az előző évek során célpont állapotból egyből 3D kategóriába kerültek (azaz nem volt közben olyan év, amikor modellezhetőek voltak).

Az adatokat extrapoláltam 2016 utáni évekre is, figyelembe véve a 3D fehérjék növekedésének eddigi ütemét, valamint, hogy hogyan változna a modellezhető fehérjék száma, ha a „Legjobb célpontok” lista sorrendje szerint történne a fehérjék kristályosítása. Ezt összehasonlítottam azzal az esettel, amikor a kristályosítandó fehérjéket véletlenszerűen választanánk ki (19. ábra)



19. ábra. A fehérjék eloszlásának változása az évek függvényében.

2016 után az adatokat extrapoláltuk a korábbi növekedés ütemét figyelembe véve (piros: 3D, sárga: modellezhető, kék: célpont, fekete: véletlen kiválasztás). Feketével jelöltem azt a vonalat, ahol a modellezhető fehérjék oszlopai futnának abban az esetben, ha a jövőben véletlen módon választanánk ki a fehérjéket kristályosításra.

4.4.2. TargetTrack

A TargetTrack adatbázis információt tartalmaz az egyes fehérjék kristályosításának előrehaladásáról. A honlapon elérhető egy statisztika (2. táblázat), ami azt mutatja, hogy az egyes célpont fehérjék a homológjaik száma szerint csoportosítva milyen státuszban vannak. A táblázatból látszik, hogy nagyon kevés azoknak a fehérjéknek a száma, amelyek kristályosítása előrehaladott állapotban van, és sok homológgal rendelkeznek. Az egyik ilyen fehérje a 22 homológgal rendelkező *Solute carrier family 22 member 1*, amelyet már sikerült megfelelően tisztítani. Ezzel szemben *selected* státuszban – azaz kristályosításra kiválasztva – sok olyan fehérje van, ami csak 1-2 homológgal rendelkezik, azaz kristályosításuk a szerkezeti tér lefedésének szempontjából nem járna sok haszonnal.

TargetTrack státusz	Homológok száma																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
selected	107	54	32	26	9	23	7	7	12	10	6	5	17	4	4			3	1	2	5	8	3	1
cloned	15	19	7	3	6	4	3	3		1			2			1							2	
expressed	24	10	9	7	4	7	5	2	3	1	1	2	5	1				1	2	1	1			
work stopped	9	6	2	3	2	5	2		3															
other	4	1				1																		
soluble	3			1																				
membrane protein solubilized	4	5			2	1	2		2			1	1											
purified	11	4	3	5	1	2	2	1	2													1		
crystallized	2																							
diffraction	1			1																				

2. táblázat. A célpont fehérjék megoszlása homológjaik száma és kristályosítási folyamatban levő állapotuk szerint.

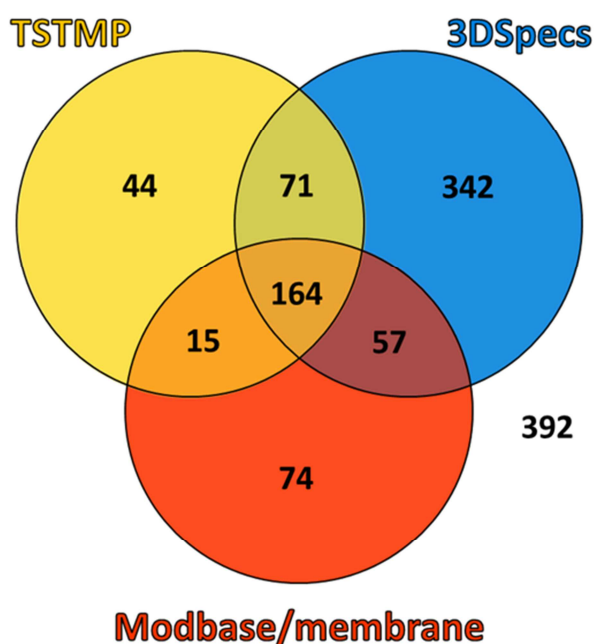
4.5. Összehasonlítás más adatbázisokkal

Elvégeztük az adatbázis összehasonlítását, más, hasonló célt szolgáló adatbázisokkal, a 2011-ben készült 3DSpecs és a 2013-ban készült Modbase/membrane adatbázisokkal (3. táblázat). Ehhez az előbbieken bemutatott, a korábbi évekre futtatott adatokat használtuk, mivel ezek az adatbázisok elkészülésük óta nem lettek frissítve.

	Modbase /membrane	3DSpecs	TSTMP
Szekvenciális információ	Azonosság	Azonosság	Azonosság Topológia
Szerkezeti információ	Homológia-modellezés	Foldfelismerés	TMP specifikus foldfelismerés
TMP szűrés	TMHMM	TMHMM	CCTOP
Frissítés	2013	2011	2016

3. táblázat. A TSTMP adatbázis összehasonlítása más adatbázisokkal.

Az egyes módszerek más-más definíciót használtak a transzmembrán fehérjékre, az egyik a 2, a másik a 3 TMH-val rendelkező fehérjékkel dolgozott, valamint mindkettő egy másik topológiabecslő algoritmussal, a TMHMM⁸⁴-el dolgozott. Az általunk használt CCTOP algoritmus topológiabecslés pontossága több mint 20%-kal nagyobb, így pontosabb kiindulóponttal dolgoztunk. A másik 2 módszer jelentősen felülbecsüli a modellezhető és megoldott szerkezettel rendelkező fehérjék számát (20. ábra) – köszönhetőn annak, hogy az egyik csak szekvenciális információt használt, a másik pedig nem transzmembrán specifikus foldfelismerést, mi pedig mind szekvenciális, mind szerkezeti, mind topológiai információt felhasználtunk.



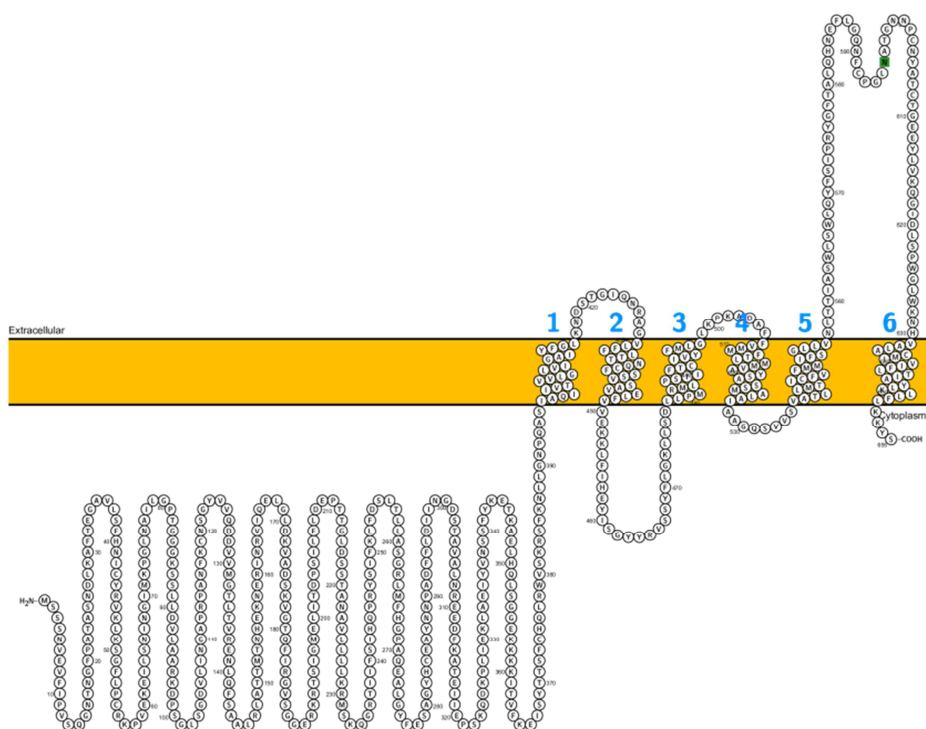
20. ábra. A TSTMP adatbázis összehasonlítása más adatbázisokkal.

A számok a modellezhető és meghatározott 3D szerkezettel rendelkező fehérjék együttes számát jelentik.

4.6. Esettanulmány

Az adatbázisban levő információk, adatok nemcsak a szerkezetmeghatározáshoz, hanem a transzmembrán fehérjék modellezéséhez is új támpontot szolgáltatnak. Ennek illusztrálására a következő esettanulmányt végeztem.

Az EVFold harmadlagos szerkezet predikciós algoritmus honlapján elérhető egyik fehérje, a humán ABCG2 transzporter (UniProt ID: ABCG2_HUMAN). Konszenzus topológia predikció alapján ennek a fehérjének hat transzmembrán szegmense van (21. ábra), és ezt az információt epitóp jelöléssel kísérletesen is megerősítették⁸⁵.

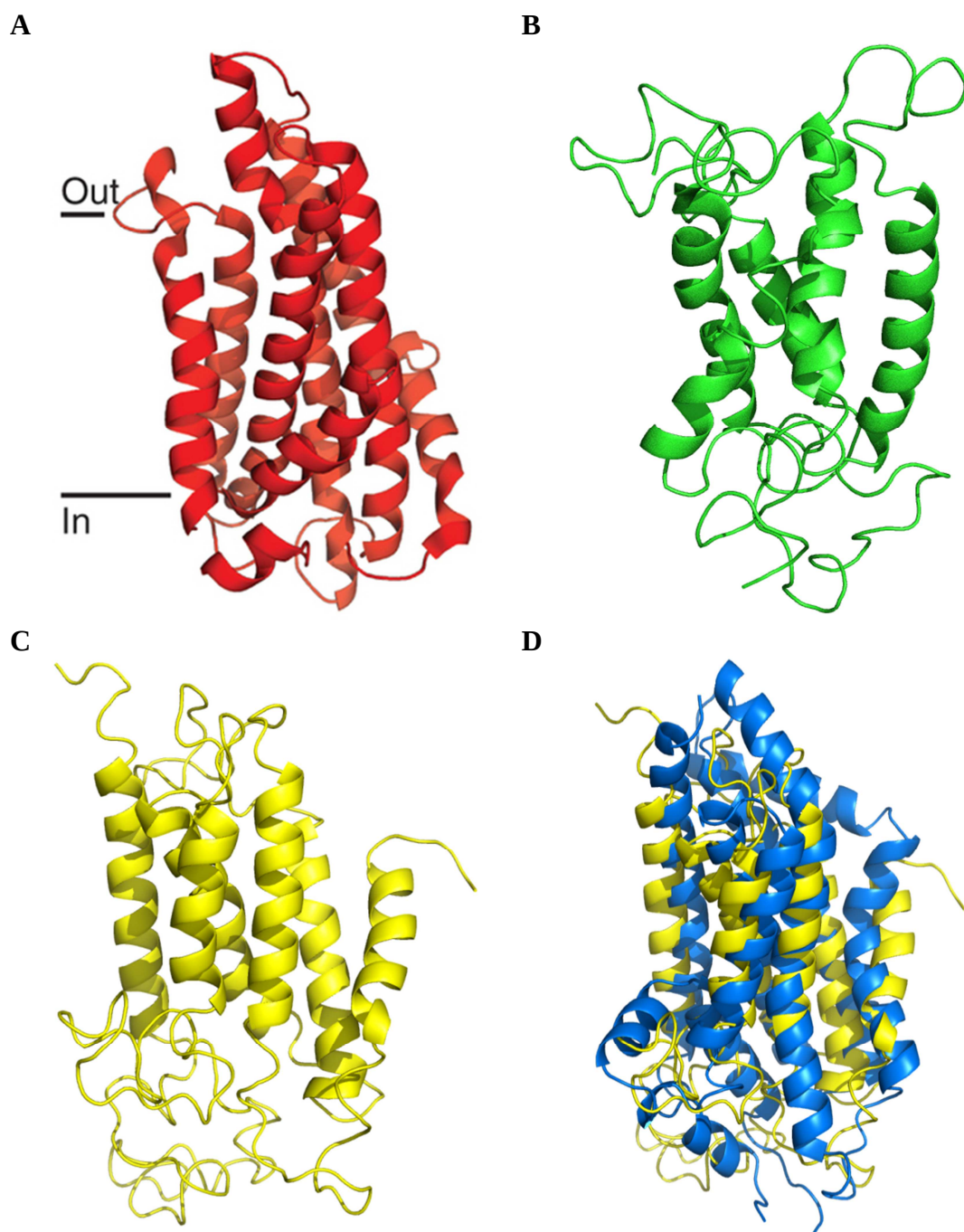


21. ábra. Az ABCG2 transzmembrán fehérje topológiája⁸⁶.

Az EVFold honlapon az előre kiszámolt szerkezet 7 transzmembrán régiót tartalmaz (22. ábra), valószínűleg a topológiabecsléshez használt algoritmus hibás predikciója miatt, vagyis ez a szerkezet biztosan tartalmaz hibát. Az ABCG2_membrane Pfam domén (PF01061) többszörös illesztésének használatával lefutott foldpredikció is hibás eredményt, 5 transzmembrán régióval rendelkező szerkezetet ad (22. ábra), mivel a Pfam illesztés nem tartalmazza az utolsó transzmembrán hélixet, így hiába adjuk meg a bemenetben a helyes topológiát, mégis rossz eredményt kapunk. Amennyiben az

alapértelmezett beállításokkal futtatjuk az algoritmust, szintén nem kapunk jó eredményt, a többszörös illesztésben a 6. membránrégiót jelentő oszlopok kiesnek, mivel ezek az oszlopok sok szekvenciában hézagokat tartalmaznak, és ezért az algoritmus kihagyja azokat.

A jó megoldást a helyes topológia megadása mellett a helyes szekvenciaillesztés megadásával lehet elérni (22. **ábra**): vagy a Pfam illesztését kell kiegészíteni a kihagyott szekvenci részekkel, vagy azon paramétert beállításait kell változtatni, hogy mennyi hézag engedhető meg az illesztés egy oszlopában úgy, hogy az algoritmus ne hagyja ki ezeket az oszlopokat. Mind a két módszerrel elérhető, hogy a becsült szerkezet végül hat transzmembrán hélixet tartalmazzon. Referencia szerkezetként a humán ABCG5 megoldott szerkezetét⁸⁷ használtam (PDB id: 5do7). Az ABCG5 kísérleti és ABCG2 becsült térszerkezetének illesztésekor kapott RMSD érték 7,38 Å, ami nagyon jónak mondható. A kihagyandó hézagok számának csökkentésével becsült szerkezet ABCG5-re való illesztése 5,55 Å RMSD-t eredményezett (22. **ábra**).



22. ábra. A humán ABCG2 modellje.

A) 7, B) 5, C) 6 transzmembrán hélix topológia alapján készült modellek. D) a 6 transzmembrán hélixes modell (sárga) illeszkedése az ABCG5 szerkezetére (kék).

Az eredmények alapján elmondható, hogy a transzmembrán fehérjék szerkezetének predikciójához elengedhetetlen a minél pontosabb topológiabecslés, hogy a transzmembrán hélixek megfelelően modellezhetők legyenek; valamint a többszörös illesztés létrehozásakor is figyelni kell arra, hogy a transzmembrán régiókat lefedjük a korrelált mutációs számításokhoz igényelt megfelelő mennyiségű szekvenciával.

4.7. A predikciós modellhez összegyűjtött adatok

A kristályosítási folyamat egyes lépéseire összegyűjtött adatok számát a 4. táblázat tartalmazza. Ahogy a folyamatban haladunk előre, természetesen egyre kevesebb adat áll rendelkezésre, hiszen az egyes lépések negatív adathalmazában levő egyedeiről semmit sem tudunk a következő lépésekben. Az adatkészlet minden egyes eleméhez 97 magyarázó változót számoltam ki.

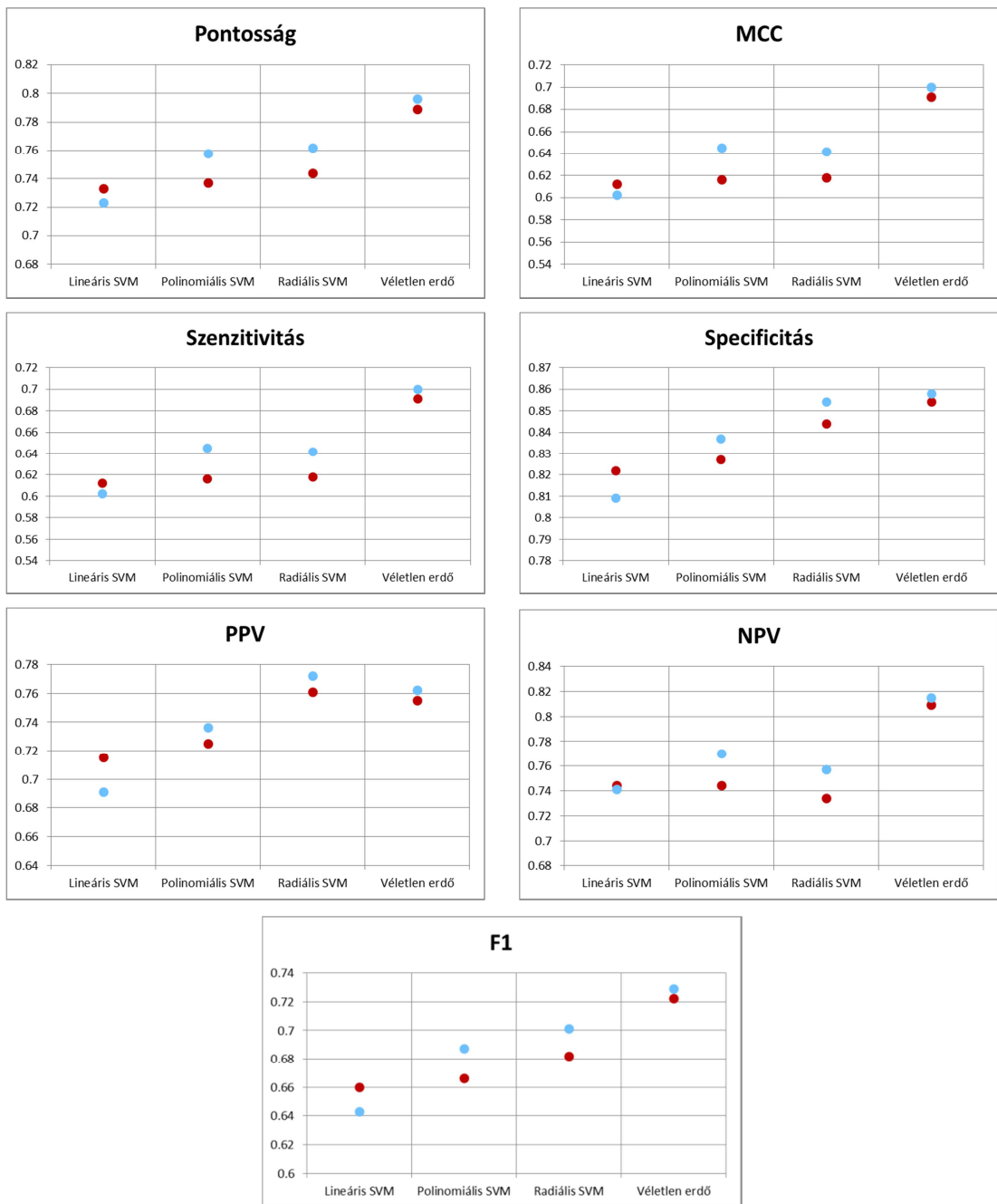
Lépés	Pozitív	Negatív
Szolubilizálás	1965	1000
Tisztítás	1469	518
Kristályosítás	443	800

4. táblázat. A predikciós programhoz összegyűjtött adatok száma a kristályosítási folyamat egyes lépéseiben.

4.8. A betanított modellek kiértékelése

4.8.1. A modellek teljesítménye

A különböző lépésekre készülő predikciós programok közül dolgozatomban a kristályosítási lépésre készült modelleket ismertetem. A betanított modellekhez kiszámítottam az irodalmi bevezetőben ismertetett mérőszámokat (5. táblázat, 23. ábra). Megállapítható, hogy az ROC-AUC-re optimalizált modellek a lineáris SVM kivételével mind jobb eredményt adtak, mint a PR-AUC-re optimalizáltak. Az eddigi vizsgálatokból az is valószínűsíthető, hogy ezekre az adatokra a véletlen erdő jobb predikciót ad, mint a különböző SVM modellek, a ROC-AUC-re tréningezett véletlen erdő is a legtöbb mérés szerint összehasonlítható, vagy jobb eredményt ad, mint a radiális SVM, ami a legjobb az SVM modellek közül.

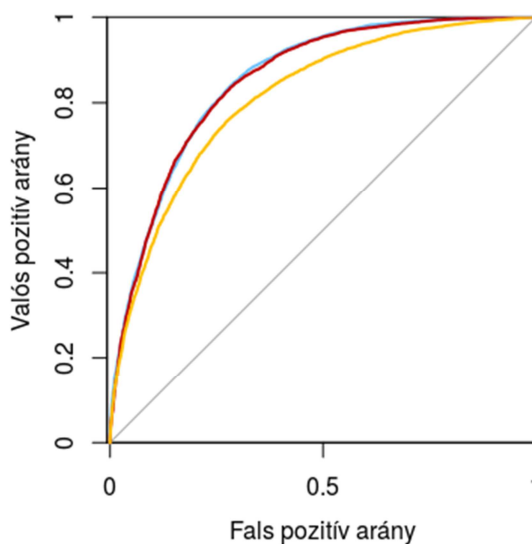


23. ábra. A betanított modellek összehasonlítása.
Kék: PR-AUC-re, piros: ROC-AUC-re optimalizált modellek.

Modell	Optimalizáció (AUC)	Szenzitivitás	Specifititás	PPV	NPV	F1	MCC	Pontosság
Radiális SVM	ROC	0.642	0.854	0.772	0.757	0.701	0.512	0.762
Radiális SVM	PR	0.618	0.844	0.761	0.734	0.682	0.479	0.744
Lineáris SVM	ROC	0.602	0.809	0.691	0.741	0.643	0.421	0.723
Lineáris SVM	PR	0.612	0.822	0.715	0.744	0.660	0.446	0.733
Polinomiális SVM	ROC	0.645	0.837	0.736	0.770	0.687	0.494	0.758
Polinomiális SVM	PR	0.616	0.827	0.725	0.744	0.666	0.456	0.737
Véletlen erdő	ROC	0.700	0.858	0.762	0.815	0.729	0.567	0.796
Véletlen erdő	PR	0.691	0.854	0.755	0.809	0.722	0.554	0.789

5. táblázat. A modellek teljesítménye. A félkövérrel szedett értékek minden oszlop legmagasabb értékét mutatják.

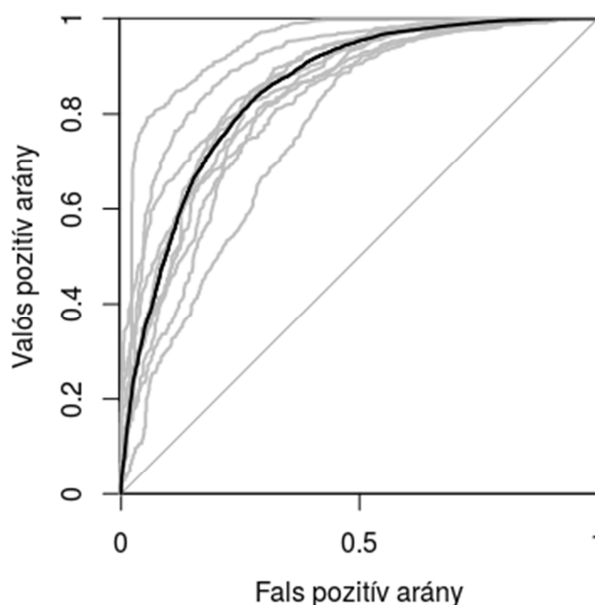
A vizsgált modellek közül a három legjobban teljesítő ROC görbét is ábrázoltam (24. ábra). A ROC görbe alapján is megállapítható a véletlen erdő modellek SVM-nél jobb teljesítménye, ugyanakkor az előbbieket a görbe alapján nem tehető különbség.



24. ábra. Modellek teljesítményének összehasonlítása.

(Kék: véletlen erdő – ROC-AUC, piros: véletlen erdő – PR-AUC, sárga: SVM – ROC-AUC)

A továbbiakban a legjobbnak ítélt, ROC-AUC-re optimalizált modellt vizsgálom. Ábrázoltam a modell minden keresztellenőrzési lépéshez tartozó ROC görbét, valamint az átlagos görbét (25. ábra). Ezzel a fajta ábrázolással képet kaphatunk arról, hogy a tanítóhalmaz változása mennyiben befolyásolja a modell teljesítményét. Az egyes lépésekre nagy a görbék szórása, jelentősen eltérnek egymástól, a modell teljesítménye tehát nagyban függ attól, hogy az adatok melyik részén tanítjuk be.



25. ábra. A véletlen erdő modell keresztellenőrzés alapján készült ROC görbéi.
(Fekete: átlagos görbe, szürke: a keresztellenőrzés egyes lépéseiben készült görbék.)

A modell teljesítményének növelésére több lehetőség áll rendelkezésre. Elsőként, növelhetjük a paraméter optimalizálásnál azon beállítások számát a jelenlegi 15 helyett, amelyeket az algoritmus véletlenszerűen kipróbál. Másodszor, az egyes szekvenciákhoz több változót kiszámolva (például AAIndex alapján, másodlagos szerkezetbecslést, csúszóablakot, stb. alkalmazva) olyan változókat találhatunk, amelyek segítik az adatpontok minél optimálisabb elválasztását. Harmadszor, egyes változókat más formában (például az értékek logaritmusát véve) is kipróbálhatjuk, bizonyos esetekben ezzel is jobb modell hozható létre.

4.8.2. Teljesítmény független tesztadatokon

A modell teljesítményét a tanítóhalmaztól független tesztadatokon is vizsgáltam. Ehhez a PDBTM adatbázisból a 2016.06.03 verzió utáni bejegyzéseket gyűjtöttem ki, ezek alkották a pozitív tesztalmazt. A negatív tesztalmazhoz a TargetTrack adatbázisból

kiválogattam azokat a bejegyzéseket, amelyeket 2015.01.01 és 2015.07.01 között frissítettek, és „*Purified*” vagy „*Mass Spec verified*” volt az utolsó bejegyzett státuszuk. Az utóbbi időpontok meghatározásában a tanítóhalmaz kialakításakor alkalmazott logikát követtem, abban a lépésben ugyanis negatív halmazba csak a „*Work stopped*”, illetve a 2015 előtti, be nem fejezett státuszú bejegyzések kerülhettek. Mivel a tanító- és a teszhalmaz létrehozása között fél év telt el, ezért fél évvel elcsúsztattam ezt a kritériumot, és az így kapott bejegyzések alkották a negatív adatokat. A halmazokból kihagytam azokat az egyedeket, amelyek legalább 40% azonosságot mutattak a tanítóhalmaz bármelyik elemével, valamint a teszhalmazt is CD-HIT-tel klasztereztem 40 %-ra. Ezekre a halmazokra ugyanúgy kiszámoltam az összes változó értéket, majd a betanított modellt használva, becslést adtam az összes egyed csoportjára. A pozitív teszhalmazban 88, a negatívban 29 egyed volt. A negatív egyedek alacsony száma valószínűleg annak köszönhető, hogy 2015-ben ért véget a PSI program, és ezután nem minden közreműködő centrum frissítette az adatbázist rendszeresen.

valós osztály	becsült osztály	
	pozitív	negatív
pozitív	106	54
negatív	2	27

6. táblázat. A független teszhalmaz becslésével kapott igazságmátrix.

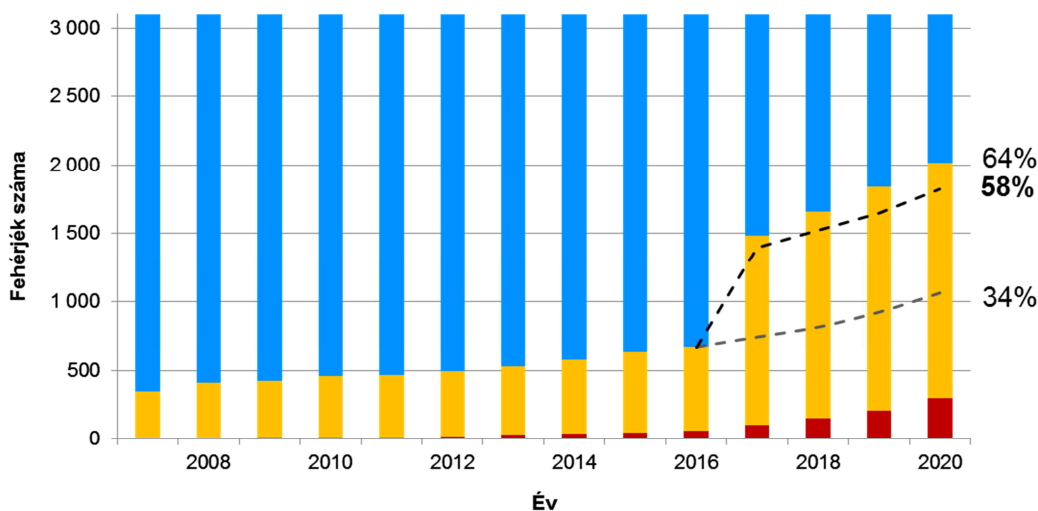
A teszhalmazon kapott eredményeket a várakozásnak megfelelően rosszabbak, mint a keresztvalidációval kapott értékek. Az igazságmátrixból (6. táblázat), valamint a specificitás értékéből (7. táblázat) látszik, hogy a modell jelentősen felülbecsüli a negatív osztályba tartozó elemeket, ugyanakkor alacsony az álpozitív becslések száma. Ez valószínűleg annak köszönhető, hogy a tanítóhalmaz kiegyensúlyozatlan volt a tanítás során, és még az alkalmazott technikákkal (alulmintavételezés, ROC-AUC-re tanítás), sem sikerült ennek a hatását teljes mértékben kiküszöbölni. A modell további optimalizálása során meg kell keresni azokat a további módszereket, paramétereket, amelyek megfelelő beállításával ezek az értékek javíthatók.

Mérőszám	Érték
Szenzitivitás	0.981
Specifitás	0.333
PPV	0.663
NPV	0.931
F1	0.791
MCC	0.432
Pontosság	0.704

7. táblázat. A független tesztalmazon kapott eredmények.

4.8.3. A predikció integrálása az adatbázisba

A modellt felhasználva, újra megvizsgáltuk, hogyan alakulna a modellezhető fehérjék számának alakulása, ha csak azokat a fehérjéket vehetjük fel a Legjobb célpontok listára, amelyeket az előzetes modell kristályosíthatónak becsül (26. ábra). Megállapítható, hogy az eredeti listához képest várhatóan nem csökken nagymértékben azon fehérjék száma, amelyek modellezhetők.



26. ábra. A fehérjék eloszlásának változása az évek függvényében.

(Piros: 3D, sárga: modellezhető, kék: célpont, szürke: véletlen kiválasztás, fekete: kiválasztás predikciós eredmények bevonásával).

5. Összefoglalás, kitekintés

A transzmembrán fehérjéknek minden sejt működésében fontos szerepe van és rendellenes működésük súlyos betegségek kialakulásához vezethet, ennek ellenére, a mai napig 60, a fehérjék transzmembrán régióit teljesen lefedő humán transzmembrán fehérje szerkezetét sikerült megoldani, valamint további körülbelül 600 fehérje szerkezete modellezhető, ugyanakkor ezek gyakran egymással rokon fehérjék.

A munka során elkészítettünk a TSTMP adatbázist, amely tartalmazza mindezeket a fehérjéket, a hozzájuk tartozó, illetve modellezhető fehérjék esetében a sablonként szolgáló szerkezetek. A célpont fehérjékből készítettük a „Legjobb célpontok” listát, amely azokat a humán transzmembrán fehérjéket tartalmazza, amelyek szerkezetének meghatározásával a lehető legtöbb másik humán transzmembrán fehérje lenne modellezhető. Az adatbázist más, kiegészítő adatokkal és esettanulmányokkal együtt elérhetővé tettük. A munkából készült publikáció a Nucleic Acids Research folyóirat 2017. január 1-jén megjelent Database számában tettük közzé⁸⁸.

A membránfehérjék kristályosítását más módon is tudjuk számításokkal segíteni. Transzmembrán fehérjékre jelenleg nem elérhető olyan predikciós program, amely becslést adna a kristályosíthatóságra. Egy ilyen modell kidolgozásába kezdtem bele, a megfelelő adatbázisokból összeválogatva, saját programokkal kiszámolva a szükséges adatokat. A létrejött adatokon a folyamat egyik lépésére betanítottam és kiértékeltem az adatokat, valamint előzetes vizsgálatot végeztem arra, hogyan befolyásolná az adatbázisban már meglevő adatokat. A kristályosítási folyamat többi lépésére a modellek betanítása jelenleg is folyik, valamint minden lépésre el fogom végezni a modellek optimalizálását is.

A végső modellek alapján olyan szervert szeretnénk létrehozni, amely szabadon elérhető, és használható a transzmembrán fehérjék kristályosításának becslésére. Ezenkívül futtatni szeretnénk a teljes TSTMP adatbázisra is, és létrehozni egy olyan „Legjobb célpontok” listát, amely ezen adatokat is figyelembe veszi. A programból készült publikációt végül majd tudományos folyóiratban szeretnénk közzé tenni.

Köszönetnyilvánítás

Szeretnék köszönetet mondani Dr. Tusnády Gábornak, hogy lehetővé tette a dolgozat elkészülését és irányította, felügyelte a munkát.

Köszönetet mondok konzulensemnek, Dr. Vértessy Beáta egyetemi tanárnak.

Szeretném megköszönni Dobson László és Reményi István hozzájárulását, a munka elkészülésében nyújtott segítséget és tanácsokat.

Köszönöm az MTA TTK Membránfehérje Bioinformatika Kutatócsoportnak a folyamatos támogatást.

A dolgozat az Emberi Erőforrások Minisztériuma Új Nemzeti Kiválóság Programjának támogatásával készült.

Irodalomjegyzék

1. Stevens, R. C. The cost and value of three-dimensional protein structure. *Drug Discov.* **4**, 35–48 (2003).
2. Büssow, K. *et al.* Structural genomics of human proteins--target selection and generation of a public catalogue of expression clones. *Microb. Cell Fact.* **4**, 21 (2005).
3. Love, J. *et al.* The New York Consortium on Membrane Protein Structure (NYCOMPS): a high-throughput platform for structural genomics of integral membrane proteins. *J. Struct. Funct. Genomics* **11**, 191–9 (2010).
4. Stevens, R. C. *et al.* The GPCR Network: a large-scale collaboration to determine human GPCR structure and function. *Nat. Rev. Drug Discov.* **12**, 25–34 (2013).
5. Alberts, B. *et al.* *Molecular Biology of the Cell.* (Garland Science, 2014).
6. Ubarretxena-Belandia, I. & Stokes, D. L. Present and future of membrane protein structure determination by electron crystallography. *Adv. Protein Chem. Struct. Biol.* **81**, 33–60 (2010).
7. Nothaft, H. & Szymanski, C. M. Protein glycosylation in bacteria: sweeter than ever. *Nat. Rev. Microbiol.* **8**, 765–778 (2010).
8. Nothaft, H. & Szymanski, C. M. Bacterial Protein N-Glycosylation: New Perspectives and Applications. *J. Biol. Chem.* **288**, 6912–6920 (2013).
9. Dobson, L., Reményi, I. & Tusnády, G. E. The human transmembrane proteome. *Biol. Direct* **10**, 31 (2015).
10. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nat. Rev. Drug Discov.* **1**, 727–30 (2002).
11. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nat.Rev.Drug Discov.* **5**, 993–996 (2006).
12. Klabunde, T. & Hessler, G. Drug Design Strategies for Targeting G-Protein-Coupled Receptors. *ChemBioChem* **3**, 928–944 (2002).
13. E. Tusnády, G. & Simon, I. Topology Prediction of Helical Transmembrane

- Proteins: How Far Have We Reached? *Curr. Protein Pept. Sci.* **11**, 550–561 (2010).
14. Sipos, L. & von Heijne, G. Predicting the topology of eukaryotic membrane proteins. *Eur. J. Biochem.* **213**, 1333–40 (1993).
 15. Heijne, G. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J.* **5**, 3021–7 (1986).
 16. Tusnády, G. E. & Simon, I. Principles Governing Amino Acid Composition of Integral Membrane Proteins: Application to Topology Prediction. (1998).
 17. Tusnády, G. E. & Simon, I. The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17**, 849–50 (2001).
 18. Nilsson, J., Persson, B. & von Heijne, G. Consensus predictions of membrane protein topology. *FEBS Lett.* **486**, 267–269 (2000).
 19. Bernsel, A. & Von Heijne, G. Improved membrane protein topology prediction by domain assignments. *Protein Sci.* **14**, 1723–8 (2005).
 20. Langó, T. *et al.* Identification of Extracellular Segments by Mass Spectrometry Improves Topology Prediction of Transmembrane Proteins. *Sci. Rep.* **7**, 42610 (2017).
 21. Varga, J., Dobson, L. & Tusnády, G. E. TOPDOM : database of conservatively located domains and motifs in proteins. *Bioinformatics* 1–2 (2016). doi:10.1093/bioinformatics/btw193
 22. Tusnády, G. E., Kalmár, L., Hegyi, H., Tompa, P. & Simon, I. TOPDOM: Database of domains and motifs with conservative location in transmembrane proteins. *Bioinformatics* **24**, 1469–1470 (2008).
 23. Dobson, L., Reményi, I. & Tusnády, G. E. CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic Acids Res.* **43**, W408–W412 (2015).
 24. Campbell, N. A. & Recce, B. *Biology*. (Pearson Education Inc., 2005).
 25. Kim, Y. *et al.* Chapter 3. High-throughput protein purification for x-ray crystallography and NMR. *Adv. Protein Chem. Struct. Biol.* **75**, 85–105 (2008).

26. *Structural Proteomics*. (Humana Press, 2008). at <http://link.springer.com/10.1007/978-1-60327-058-8>
27. Bernaudat, F. *et al.* Heterologous expression of membrane proteins: Choosing the appropriate host. *PLoS One* **6**, e29191 (2011).
28. Lundstrom, K. Structural genomics for membrane proteins. *Cell. Mol. Life Sci.* **63**, 2597–2607 (2006).
29. Smyth, M. S. & Martin, J. H. X Ray Crystallography. *J. Clin. Pathol. Mol. Pathol.* **53**, 8–14 (2000).
30. Ladd, M. & Palmer, R. *Structure Determination by X-ray Crystallography*. (Springer US, 2003). doi:10.1007/978-1-4615-0101-5
31. Verdaguer, N., Garriga, D. & Fita, I. X-ray crystallography of viruses. *Subcell. Biochem.* **68**, 117–44 (2013).
32. Moraes, I., Evans, G., Sanchez-Weatherby, J., Newstead, S. & Stewart, P. D. *Membrane protein structure determination - the next generation*. *Biochim Biophys Acta* **1838**, (2014).
33. Lacapère, J. J., Pebay-Peyroula, E., Neumann, J. M. & Etchebest, C. Determining membrane protein structures: still a challenge! *Trends Biochem. Sci.* **32**, 259–270 (2007).
34. Lian, L.-Y. & Roberts, G. C. K. (Gordon C. K. *Protein NMR spectroscopy practical techniques and applications*. (Wiley, 2011).
35. Velyvis, A., Ruschak, A. M. & Kay, L. E. An economical method for production of (2)H, (13)CH₃-threonine for solution NMR studies of large protein complexes: application to the 670 kDa proteasome. *PLoS One* **7**, e43725 (2012).
36. Serber, Z., Corsini, L., Durst, F. & Dötsch, V. In-Cell NMR Spectroscopy. *Methods Enzymol.* **394**, 17–41 (2005).
37. Barbieri, L., Luchinat, E. & Banci, L. Characterization of proteins by in-cell NMR spectroscopy in cultured mammalian cells. *Nat. Protoc.* **11**, 1101–1111 (2016).
38. O'Connell, M. R., Gamsjaeger, R. & Mackay, J. P. The structural analysis of

- protein-protein interactions by NMR spectroscopy. *Proteomics* **9**, 5224–32 (2009).
39. de Brevern, A. G. 3D structural models of transmembrane proteins. *Methods Mol. Biol.* **654**, 387–401 (2010).
 40. Forrest, L. R., Tang, C. L. & Honig, B. On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophys. J.* **91**, 508–17 (2006).
 41. Tusnády, G. E. & Kozma, D. in *Protein Modelling* 199–221 (Springer International Publishing, 2014). doi:10.1007/978-3-319-09976-7_9
 42. Pieper, U. *et al.* Coordinating the impact of structural genomics on the human α -helical transmembrane proteome. *Nat. Struct. Mol. Biol.* **20**, 135–8 (2013).
 43. Lundstrom, K. *Structural Genomics on Membrane Proteins*. (CRC Press, 2006). doi:10.1201/9781420016086
 44. Linial, M. & Yona, G. Methodologies for target selection in structural genomics. *Prog. Biophys. Mol. Biol.* **73**, 297–320 (2000).
 45. Grabowski, M., Chruszcz, M., Zimmerman, M. D., Kirillova, O. & Minor, W. Benefits of structural genomics for drug discovery research. *Infect. Disord. Drug Targets* **9**, 459–74 (2009).
 46. Punta, M. *et al.* Structural genomics target selection for the New York consortium on membrane protein structure. *J. Struct. Funct. Genomics* **10**, 255–68 (2009).
 47. Kuhn, M. & Johnson, K. *Applied Predictive Modeling*. (Springer New York, 2013). doi:10.1007/978-1-4614-6849-3
 48. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
 49. Tin Kam Ho. Random decision forests. *Proc. 3rd Int. Conf. Doc. Anal. Recognit.* **1**, 278–282 (1995).
 50. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
 51. Overton, I. M. & Barton, G. J. A normalised scale for structural genomics target

- ranking: The OB-Score. *FEBS Lett.* **580**, 4005–4009 (2006).
52. Chen, K., Kurgan, L. & Rahbari, M. Prediction of protein crystallization using collocation of amino acid pairs. *Biochem. Biophys. Res. Commun.* **355**, 764–769 (2007).
 53. Kurgan, L. *et al.* CRYSTALP2: sequence-based protein crystallization propensity prediction. *BMC Struct. Biol.* **9**, 50 (2009).
 54. Kawashima, S., Ogata, H. & Kanehisa, M. AAindex: Amino Acid Index Database. *Nucleic Acids Res.* **27**, 368–369 (1999).
 55. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
 56. Charoenkwan, P. *et al.* SCMCRYST: Predicting Protein Crystallization Using an Ensemble Scoring Card Method with Estimating Propensity Scores of P-Collocated Amino Acid Pairs. *PLoS One* **8**, (2013).
 57. Slabinski, L. *et al.* XtalPred: a web server for prediction of protein crystallizability. *Bioinformatics* **23**, 3403–3405 (2007).
 58. Jahandideh, S., Jaroszewski, L. & Godzik, A. Improving the chances of successful protein structure determination with a random forest classifier. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **70**, 627–635 (2014).
 59. Wang, H. *et al.* CrysAlis: an integrated server for computational analysis and design of protein crystallization. *Sci. Rep.* **6**, 21383 (2016).
 60. Huilin Wang, Liubin Feng, Geoffrey I. Webb, Lukasz Kurgan, J. S. and D. L. Critical evaluation of bioinformatics tools for the prediction of protein crystallization propensity. *Brief. Bioinform.* **7**, 1–15 (2017).
 61. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
 62. Martin-Galiano, A. J., Smialowski, P. & Frishman, D. Predicting experimental properties of integral membrane proteins by a naive Bayes approach. *Proteins Struct. Funct. Genet.* **70**, 1243–1256 (2008).
 63. Bray, J. E. Target selection for structural genomics based on combining fold

- recognition and crystallisation prediction methods: application to the human proteome. *J. Struct. Funct. Genomics* **13**, 37–46 (2012).
64. Saladi, S. M., Javed, N., Müller, A. & Clemons, W. M. Decoding sequence-level information to predict membrane protein expression. 1–53 (2017).
 65. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011).
 66. Bernsel, A., Viklund, H., Hennerdal, A. & Elofsson, A. TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res.* **37**, W465–8 (2009).
 67. Tusnády, G. E., Kalmár, L. & Simon, I. TOPDB: Topology data bank of transmembrane proteins. *Nucleic Acids Res.* **36**, 234–239 (2008).
 68. Dobson, L., Lango, T., Remenyi, I. & Tusnady, G. E. Expediting topology data gathering for the TOPDB database. *Nucleic Acids Res.* **43**, D283–D289 (2014).
 69. Kozma, D., Simon, I. & Tusnády, G. E. PDBTM: Protein Data Bank of transmembrane proteins after 8 years. *Nucleic Acids Res.* **41**, D524–D529 (2013).
 70. Kozma, D. & Tusnády, G. E. TMFoldRec: a statistical potential-based transmembrane protein fold recognition tool. *BMC Bioinformatics* **16**, 201 (2015).
 71. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 1–9 (2009).
 72. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–2 (2012).
 73. Kuhn, M. Building Predictive Models in R Using the **caret** Package. *J. Stat. Softw.* **28**, 1–26 (2008).
 74. Hopf, T. A. *et al.* Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607–1621 (2012).
 75. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
 76. Tusnady, G. E. PDB_TM: selection and membrane localization of

- transmembrane proteins in the protein data bank. *Nucleic Acids Res.* **33**, D275–D278 (2004).
77. Tusnády, G. E., Dosztányi, Z. & Simon, I. TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics* **21**, 1276–1277 (2005).
 78. Tusnády, G. E., Dosztányi, Z. & Simon, I. Transmembrane proteins in the Protein Data Bank: Identification and classification. *Bioinformatics* **20**, 2964–2972 (2004).
 79. Gabanyi, M. J. *et al.* The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. *J. Struct. Funct. Genomics* **12**, 45–54 (2011).
 80. Chen, L., Oughtred, R., Berman, H. M. & Westbrook, J. TargetDB: A target registration database for structural genomics projects. *Bioinformatics* **20**, 2860–2862 (2004).
 81. Lao, D. M., Arai, M., Ikeda, M. & Shimizu, T. The presence of signal peptide significantly affects transmembrane topology prediction. *Bioinformatics* **18**, 1562–6 (2002).
 82. *The Proteomics Protocols Handbook.* (Humana Press, 2005). doi:10.1385/1592598900
 83. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
 84. Sonnhammer, E. L., von Heijne, G. & Krogh, A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**, 175–82 (1998).
 85. Horsey, A. J., Cox, M. H., Sarwat, S. & Kerr, I. D. The multidrug transporter ABCG2: still more questions than answers. *Biochem. Soc. Trans.* **44**, 824–30 (2016).
 86. Omasits, U., Ahrens, C. H., Müller, S. & Wollscheid, B. Protter: interactive protein feature visualization and integration with experimental proteomic data. *Bioinformatics* **30**, 884–6 (2014).

87. Lee, J.-Y. *et al.* Crystal structure of the human sterol transporter ABCG5/ABCG8. *Nature* **533**, 561–564 (2016).
88. Varga, J., Dobson, L., Reményi, I. & Tusnády, G. E. TSTMP: target selection for structural genomics of human transmembrane proteins. *Nucleic Acids Res.* **45**, D325–D330 (2017).

Ábrajegyzék

1. ábra. A membránfehérjék lehetséges szerkezetei ⁵	8
2. ábra. A transzmembrán fehérjék funkciói ²⁴	10
3. ábra. A membránfehérjék kristályosításának általános folyamata.	11
4. ábra. Vevő működési karakterisztika görbék.....	20
5. ábra. Precizitás – felidézés görbe.....	21
6. ábra. Felügyelt tanítású algoritmusok tanítása.....	22
7. ábra. Modellillesztések.	23
8. ábra. Alul-és túlillesztett modellek hibája.	24
9. ábra. Adatpontok elválasztása támogatóvektor géppel.....	25
10. ábra. Az adatok transzformálása és elválasztása kernel függvény segítségével.....	26
11. ábra. Egyszerű döntési fa.	27
12. ábra. Véletlen erdő sematikus modellje.....	28
13. ábra. Az adatbázis elkészítésének folyamatábrája.	37
14. ábra. Az adatok összegyűjtésének folyamata.	39
15. ábra. Az elkészült adatbázis egy rekordja.....	42
16. ábra. Az adatbázis egyik rekordját leíró XML fájl.	44
17. ábra. A csoportok arányának változása az adatbázisban.	46
18. ábra. A „Legjobb célpontok” lista.	47
19. ábra. A fehérjék eloszlásának változása az évek függvényében.....	48
20. ábra. A TSTMP adatbázis összehasonlítása más adatbázisokkal.	50
21. ábra. Az ABCG2 transzmembrán fehérje topológiája ⁸⁶	51
22. ábra. A humán ABCG2 modellje.....	53
23. ábra. A betanított modellek összehasonlítása.	55
24. ábra. Modellek teljesítményének összehasonlítása.....	56
25. ábra. A véletlen erdő modell keresztellenőrzés alapján készült ROC görbéi.....	57
26. ábra. A fehérjék eloszlásának változása az évek függvényében.....	59