GENOM ANNOTÁCIÓ
Részletes szakmai zárójelentés

A projekt célja

Az orvostudomány, a gyógyszerfejlesztés, a mezőgazdaság és a biotechnológia nagymértékben támaszkodik a genom projektekből származó adatokra. A genom szekvencia értelmezésének első lépése a gének bioinformatikai úton történő azonosítása és szerkezetük meghatározása: minden további biológiai kutatás ezekre az adatokra épül.

A génazonosítás nehézségeit illusztrálhatjuk azzal, hogy – jóllehet a humán genom szekvenciájának „draft" verzióját már több mint egy évtizede publikálták – az emberi genomban található fehérjekódoló gének pontos száma még mindig nem ismert. Ennél is súlyosabb problémát jelez, hogy - a génpredikciós módszerek fejlődése ellenére - a fehérjekódoló gének szerkezetének predikciója nem kellően megbízható: a megjósolt gének kevesebb, mint 50%-ának helyes a szerkezete. Ennek következményeként a biológiai kutatások alapjául szolgáló nyilvános fehérje adatbázisok erősen szennyezettek hibás és félrevezető adatokkal, és mindez súlyos problémákat okozhat a genom információkat hasznosító kutatások szempontjából. Mindezek miatt jelentős az igény megbízhatóbb génpredikciós eljárások kifejlesztésére, illetve a meglévő adatok minőségellenőrzésére.

A korábbi munkánk során kifejlesztettük az első öt MisPred eszközt, amelyek lehetővé teszik a tévesen megjósolt gének/fehérjék automatizált azonosítását. A megközelítés alapja, hogy egy fehérjekódoló gén valószínűleg tévesen megjósolt, ha a gén (vagy az általa kódolt fehérje) jellemzői nincsenek összhangban a fehérjekódoló génekről és a fehérjékről alkotott jelenlegi tudásunkkal. A kifejlesztett eszközökkel elemeztük különböző Metazoa fajok fehérje adatbázisait, azonosítottuk a hibás fehérje-szekvenciákat és a vizsgálatok eredményeinek publikálására létrehoztuk a MisPred adatbázist.

A 2011.09.01-én elkezdődött PD 101201 „Genom-annotáció" projekt fő célkitűzései:
1. Új MisPred hibaazonosítási módszerek kidolgozása további hibatípusok azonosítására.
2. További eukarióta genomok bevonása a vizsgálatokba. Nyilvános fehérje adatbázisok legutóbbi verzióinak elemzése és a MisPred adatbázis folyamatos frissítése.
3. A MisPred által hibásként azonosított gének/fehérjék szerkezetének kijavítására alkalmas FixPred módszer kidolgozása és a kijavított szekvenciákat tartalmazó FixPred adatbázis létrehozása.
4. A kijavított gének/fehérjék funkciójának és biológiai szerepének predikciója, új predikciós módszerek alkalmazásával.
5. Automatizált eljárás kidolgozása az orvosbiológiai, agrárbiológiai vagy biotechnológiai szempontból hasznosítható gének kiválasztására.

A projekt folyamán elvégzett szakmai feladatok

1. MISPRED

1. Továbbfejlesztettük a MisPred módszert annak érdekében, hogy tovább növeljük a minőségellenőrző vizsgálatok hatékonyságát, növeljük az automatikusan azonosítható hibatípusok és a vizsgált eukarióta genomok számát.

a) Új MisPred eszközök kidolgozása további biológiai dogmák alapján

i) Az eredeti célkitűzéseknek megfelelően kidolgoztuk a 7-10. típusú hibák azonosítási módszerét és kifejlesztettük a 7-10. MisPred eszközöket:

- 7. konfliktus: GPI-horgonyt jelenléte és szignál peptid hiánya.
- 8. konfliktus: GPI-horgony és intracelluláris domén együttes előfordulása.
- 9. konfliktus: GPI-horgony és nukleáris domén együttes előfordulása.
- 10. konfliktus: GPI-horgony és transzmembrán szegmens együttes előfordulása.

A tervezett négy új eszközzel (7-10. konfliktus) a fehérjéknek azt a csoportját vizsgáljuk, amelyek C-terminális végéhez egy GPI (glikozil-foszfatidil-inozitol)-horgony kapcsolódik. Ezek a fehérjék egy szignál peptidet is kell, hogy tartalmazzanak, amely a sejtből az extracelluláris térbe irányítja őket, ahol a sejtmembrán külső feléhez kapcsoltan maradnak. A GPI-kötött fehérjék esetén a teljes fehérje az extracelluláris térben helyezkedik el, ezért csak extracelluláris doménnel rendelkezhet (de nem szükségszerűen tartalmaz extracelluláris domént), az intracelluláris és nukleáris domén előfordulása, illetve a transzmembrán szegmens jelenléte tiltott. Mindezek alapján hibásnak tekintjük azokat a GPI-kötött fehérjéket, amelyek nem tartalmaznak szignál peptidet, illetve obligát intracelluláris vagy obligát nukleáris domént vagy transzmembrán szegmenst tartalmaznak.

ii) Az eredeti célkitűzéseken kívül kidolgoztunk és kifejlesztettünk két új MisPred eszközt (6. és 11. konfliktus). Ennek oka, hogy a két új eszköz jelentősen megnöveli a MisPred teljesítményét:

- 6. konfliktus: A szignál peptid és intracelluláris domén együttes jelenléte, valamint a transzmembrán domén hiánya.
- 11. konfliktus: Domén architektúra deviáció.

A 6-os számú eszközzel a fehérjéknek azt a csoportját vizsgáljuk, amelyek tartalmaznak szignál peptidet és citoplazmikus Pfam-A domént. Ezen fehérjéknek az N-terminális része az extracelluláris részben helyezkedik el, van azonban olyan része, amely a sejten belül, ezért szükségszerűen tartalmaznia kell legalább egy transzmembrán hélixet. Mindezek alapján hibásnak tekintjük azokat a szignál peptidet és citoplazmikus Pfam-A domént tartalmazó fehérjéket, amelyek nem tartalmaznak transzmembrán szegmenst.

A 11-es számú eszköz a fehérjék domén architektúráját, azaz a fehérjékben a domének sorrendjét vizsgálja. A hibaazonosítási módszer alapja, hogy az evolúció során a fehérjék domén architektúrája ritkán változik meg, ezért ha találunk egy olyan domén architektúrát, amely eltér a közeli rokon fajok ortológ (azonos eredetű és funkciójú) fehérjéinek domén architektúrájától, akkor ez valószínűleg nem új domén architektúrát, hanem hibás fehérjét jelez.

A fenti biológiai dogmán alapuló új eszközök elérhetőek és a kutatók számára szabadon használhatók a www.mispred.com webhely „Analyze your sequence” menüpontja alatt.

b) További eukarióta genomok bevonása a vizsgálatokba

i) Új, evolúciós szempontból kulcsfontosságú Metazoa fajok bevonása

A projekt során négy új faj vizsgálatára tettük alkalmassá a MisPred eszközöket: az *Amphimedon queenslandica (a Poriferák, szivacsok képviselője), Mnemiopsis leidyi (a Ctenophorák képviselője), Petromyzon marinus* és *Aplysia californica* teljes proteómája a projekt időtartama alatt vált hozzáférhetővé.

Így jelenleg az alábbi Metazoa fajok proteómáinak vizsgálatára alkalmasak a MisPred eszközök: *Homo sapiens, Mus musculus, Rattus norvegicus, Gallus gallus, Monodelphis domestica, Xenopus tropicalis, Danio rerio, Fugu rubripes, Ciona intestinalis, Branchiostoma floridae, Strongylocentrotus purpuratus, Drosophila melanogaster, Drosophila simulans, Drosophila pseudoobscura, Caenorhabditis elegans, Caenorhabditis briggsae, Hydra magnipapillata, Nematostella vectensis, Trichoplax adhaerens + Amphimedon queenslandica, Mnemiopsis leidyi, Petromyzon marinus, Aplysia californica.*

ii) Növények bevonása a MisPred vizsgálatokba

A pályázatban nem terveztük, de mivel több külföldi kutatócsoport azzal a javaslattal keresett meg minket, hogy a MisPred eszközöket növényi proteómák vizsgálatára szeretnék használni, elkezdtük optimalizálni a növényekre is a MisPred eszközöket.
Ennek során megvizsgáltuk, hogy az 1-3. és 6. konfliktus vizsgálata során használt obligát extracelluláris, citoplazmikus és nukleáris doménlistákban milyen változtatások szükségesek, illetve megvizsgáltuk a MisPred pipeline által használt szignál peptid és GPI-horgony jósló programok teljesítőképességét növények esetén. A tapasztalatok azt mutatják, hogy a MisPred eszközök optimalizálása a növényi proteómák vizsgálatára jelentősebb bioinformatikai és informatikai munkát igényel és ez egy új projekt feladata lehet.

2. Frissítettük a MisPred vizsgálatok hátterében működő programokat és adatbázisokat.

A pályázat feladatai között nem szerepelt, de szükségessé vált a MisPred pipeline által használt programok, adatbázisok és az 1-3. és 6. konfliktusok azonosításának alapjául szolgáló obligát extracelluláris, intracelluláris szignalling és nukleáris doménccsaládokat tartalmazó doménlisták frissítése.

3. Elemeztük a UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, EnsEMBL és NCBI/RefSeq adatbázisok legújabb verzióit és folyamatosan frissítettük a hibás fehérjéket tartalmazó MisPred adatbázist.

A MisPred pipeline segítségével megvizsgáltuk 23 Metazoa faj fehérje szekvenciáit a UniProtKB/Swiss-Prot, a UniProtKB/TrEMBL, az NCBI/RefSeq és az EnsEMBL adatbázisokban, kiszűrtük az 1-10. konfliktusban szenvedő szekvenciákat, és ezeket elhelyeztük a MisPred adatbázisban. Ennek eredményeként a MisPred adatbázis (www.mispred.com) 6.0 verziója 80.890 db, a 7.0 verziója 92.016 db, a 8.0 verziója pedig 110.186 db hibás fehérjét tartalmaz.
A legfrisebb adatbázis verzióhoz tartozó hibás fehérjék annotációi a www.mispred.com webhelyen a „Search MisPred” menüpont alatt érhetők el.
A hibás fehérjék vizsgálatára vonatkozó statisztikai adatok a http://www.mispred.com/statistics oldalon láthatóak.

2. FIXPRED

1. Kidolgoztuk és kifejlesztettük a MisPred 1-11. eszközök által hibásként azonosított szekvenciák kijavítására alkalmas FixPred eljárást.

a) Az eredeti terveknek megfelelően kidolgoztuk a MisPred eszközök által hibásként azonosított, az 1-5. és 7-10. konfliktusokban szenvedő fehérjeszekvenciák hibáinak kijavítására szolgáló eljárásokat, és kifejlesztettük az 1-5. és 7-10. konfliktusok hibáinak kijavítására alkalmas FixPred pipeline-t:

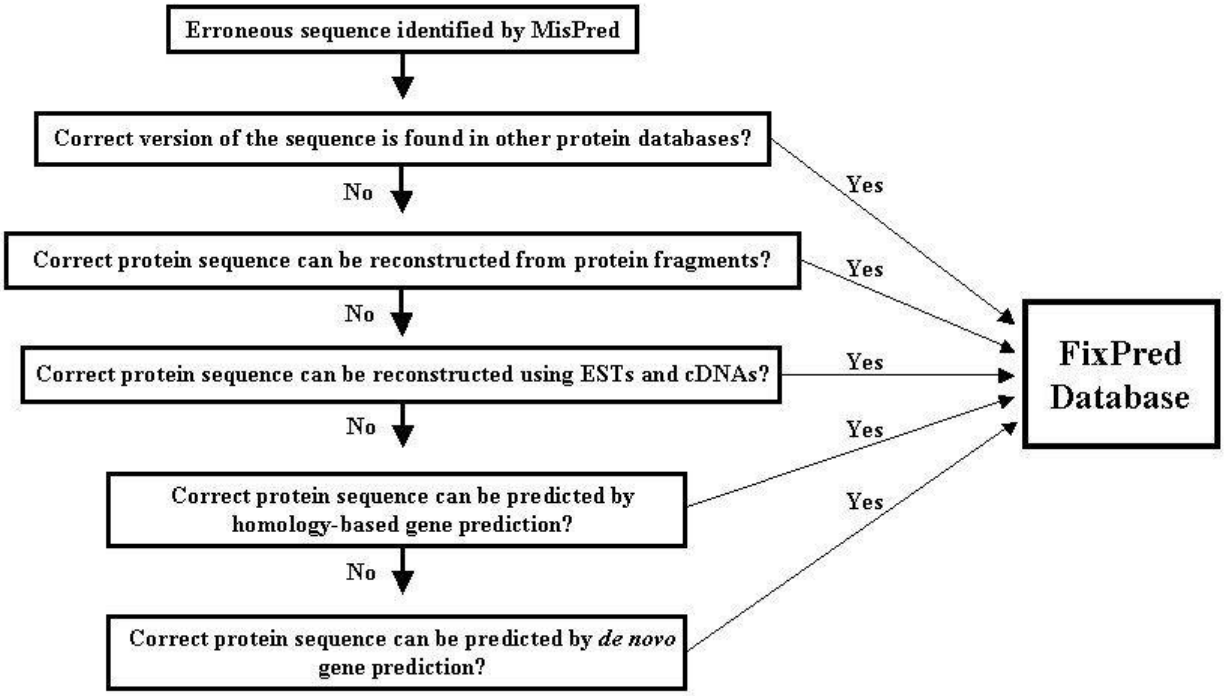
- 1. konfliktus: A fehérje sejten belüli lokalizációja és a megfelelő szekvencia-szignál hiánya.
- 2. konfliktus: Az extracelluláris és intracelluláris domének jelenléte és a transzmembrán domének hiánya.
- 3. konfliktus: Az extracelluláris és nukleáris domének együttes előfordulása.
- 4. konfliktus: Domén méret deviáció.
- 5. konfliktus: Különböző kromoszómák által kódolt kiméra fehérjék.
- 7. konfliktus: GPI-horgony jelenléte és szignál peptid hiánya.
- 8. konfliktus: GPI-horgony és intracelluláris domén együttes előfordulása.
- 9. konfliktus: GPI-horgony és nukleáris domén együttes előfordulása.
- 10. konfliktus: GPI-horgony és transzmembrán szegmens együttes előfordulása.

b) Az eredetileg tervezett 9 hibajavító módszerhez képest további 2 módszert dolgoztunk és fejlesztettünk ki (6. és 11. konfliktus), hogy az új eszközökkel megnöveljük a FixPred teljesítményét:

- 6. konfliktus: A szignál peptid és intracelluláris domén együttes jelenléte, valamint a transzmembrán domén hiánya.
- 11. konfliktus: Domén architektúra deviáció.

A FixPred pipeline többféle megközelítést alkalmaz a hibák kijavítására és a hibás szekvenciák korrekt verzióit a FixPred adatbázisban helyezi el. A pipeline a legegyszerűbb megoldásokkal indul (a hibás szekvencia korrekt verzióját azonosítja, vagy rekonstruálja az adatbázisokban található kísérletes adatok alapján) és ha ezek a megoldások nem vezetnek eredményre akkor alkalmazza az időigényesebb génpredikciós módszereket. A FixPred eljárás folyamatát az 1. ábrán látható dichotomikus „döntési fa” illusztrálja:

1. lépés: A MisPred hibaazonosítása.
2. lépés: A fehérje helyes verziójának megkeresése más fehérje adatbázisokban.
3. lépés: A javított fehérje kijavítása átfedő fehérje fragmentumok felhasználásával.
4. lépés: A hibás fehérje szekvencia kijavítása átfedő cDNS vagy EST szekvenciák segítségével.
5. lépés: A hibás fehérje szekvencia kijavítása genomikus szekvencia felhasználásával, nem hibás szekvenciával szekvencia-hasonlóság alapján.
6. lépés: A hibás fehérje szekvencia kijavítása genomikus szekvencia felhasználásával, *de novo* predikciók révén.



A FixPred pipeline jelenlegi verziója az alábbi Metazoa fajok hibás fehérje szekvenciáinak kijavítására alkalmas: *Homo sapiens, Mus musculus, Rattus norvegicus, Gallus gallus, Monodelphis domestica, Xenopus tropicalis, Danio rerio, Fugu rubripes, Ciona intestinalis, Branchiostoma floridae, Drosophila melanogaster, Caenorhabditis elegans, Amphimedon queenslandica, Mnemiopsis leidyi, Petromyzon marinus, Aplysia californica.*

A FixPred hibajavító eszközök elérhetőek és a kutatók számára szabadon használhatók a www.fixpred.com webhely „Correct your sequence" menüpontja alatt.

2. Létrehoztuk a FixPred adatbázist, amely a MisPred eszközök által hibásként azonosított és a FixPred által kijavított fehérje szekvenciákat tartalmazza.

Megterveztük és létrehoztuk a kijavított szekvenciákat tartalmazó FixPred adatbázist és a hozzá kapcsolódó FixPred portált (www.fixpred.com). A javítások során az 1-10. számú MisPred eszközökkel hibásként azonosított UniProtKB/Swiss-Prot , illetve NCBI/RefSeq fehérjéket javítottuk ki. A javítások helyességét ellenőriztük, és a helyes fehérje szekvenciákat és azok annotációit elhelyeztük a FixPred adatbázisban. Az adatbázis a www.fixpred.com webhelyen a „Search FixPred" menüpont alatt érhető el.

3. TARGETPRED

1. Kidolgoztuk az orvosbiológiai szempontból hasznosítható humán fehérjék kiválasztására szolgáló TargetPred eljárást.

Korábbi kutatásaink alapján megállapítottuk, hogy a humán gyógyszercélpont adatbázisok fehérjekészlete jelentősen eltér a fehérjék általános készletétől a biológiai funkció, a szubcelluláris lokalizáció és az evolúciós eloszlás szempontjából. Mindezek alapján a TargetPred eljárás a fehérje három paramétere, 1. az evolúciós eloszlása, 2. a szubcelluláris lokalizációja, 3. és a biológiai

funkciója alapján dönt arról, hogy az adott fehérje milyen valószínűséggel hasznosítható gyógyszercélpontként.

2. Kifejlesztettük a humán fehérjék ortológjai „Evolúciós eloszlásának" meghatározására szolgáló eljárást, a TargetPred módszer első komponensét.

A kifejlesztett eszköz meghatározza azt az élőlénycsoportot (evolúciós vonalak, taxonómiai csoportok), amelyben előfordulnak az adott humán fehérje funkcionálisan ekvivalens ortológjai (utalva arra, hogy az adott biológiai funkció az ortológokat tartalmazó vonalak közös ősében jelent meg).
A kifejlesztett eszköz a humán fehérjéket négy fő kategóriába sorolja:

1. Univerzális (ortológjai a prokariótákban és eukariótákban egyaránt előfordulnak)
2. Eukarióta-specifikus (ortológjai csak eukariótákban fordulnak elő)
3. Metazoa-specifikus (ortológjai csak Metazoákban fordulnak elő)
4. Gerinces-specifikus (ortológjai csak gerincesekben fordulnak elő)

3. Kifejlesztettük a humán fehérjék „Szubcelluláris lokalizációjának" meghatározására szolgáló eljárást, a TargetPred módszer második komponensét.

A „szubcelluláris lokalizáció" paraméter meghatározza a fehérje lokalizációját a sejten belül vagy azon kívül.
A kifejlesztett TargetPred eszköz a fehérjéket hat nagyobb kategóriába sorolja:

1. Szekretált/extracelluláris
2. Mitokondriális
3. Nukleáris
4. Plazmamembrán
5. Citoplazmikus
6. Citoplazmikus-nukleáris

A fehérjék besorolásához a TargetPred öt megközelítést alkalmaz:

1. Szubcelluláris lokalizáció meghatározása a fehérje Swiss-Prot annotációja alapján
2. Szubcelluláris lokalizáció meghatározása a fehérje GO annotációja alapján
3. A fehérje szubcelluláris lokalizációjának jóslása MisPred eszközökkel
4. A fehérje szubcelluláris lokalizációjának jósolása a WoLF PSORT programmal
5. A fehérjére vonatkozó Swiss-Prot annotáció következetességének ellenőrzése

Az ötféle vizsgálattal kapott besorolás alapján a többségi szabályt használva dönt az eszköz a fehérje szubcelluláris lokalizációjáról.

4. Kidolgoztuk a fehérjék „Biológiai funkciójának" meghatározására szolgáló eljárást, a TargetPred módszer harmadik komponensét.

Ez a paraméter meghatározza a fehérjék biológiai funkcióját, biológiai szerepét.
Ebben az elemzésben a TargetPred a fehérjéket három funkcionális szuperosztályba sorolja:

1. Energia, metabolizmus és a sejtes élet alapvető szerkezetei
2. Információ
3. Kommunikáció a környezettel

Az eredeti tervek szerint a TargetPred-nek ez a komponense az általánosan elfogadott, széleskörűen használt Swiss-Prot és GO adatbázisok funkcionális annotációt kívánta hasznosítani. Munkánk során azt tapasztaltuk, hogy a biológiai funkció meghatározásához használni kívánt funkcionális annotációs adatok sem a Swiss-Prot, sem a GO adatbázisokban nem következetesek, nem

egyértelműek, nem megbízhatóak, nem ellentmondásmentesek és nem teljeskörűek. Megállapítottuk, hogy a szakirodalomban és az adatbázisokban található, funkcióra vonatkozó információk kigyűjtéséhez és rendszerezéséhez új, lényegesen több kifejezésből álló „kulcsszó gyűjteményt" kell összeállítani, és a fehérjék funkcionális kategorizálása is lényeges fejlesztést igényel. Mindez a vártnál lényegesen bonyolultabb feladat, ezért a TargetPred módszer harmadik komponensének tökéletesítése, megbízhatóbbá tétele további munkát igényel.