# Átfedő modulok molekuláris biológiai kölcsönhatási hálózatokban

MTA doktori értekezés tézisfüzete

Farkas Illés


Magyar Tudományos Akadémia
Támogatott Kutatóhelyek Irodája
MTA-ELTE Statisztikus és Biológiai Fizika Kutatócsoport

Budapest, 2015 április

# A kutatások előzménye

A doktori dolgozat a biológia és a statisztikus fizika határterületén található kutatási témákat és eredményeket ismertet. A biológia hosszú ideig döntően leíró tudományág volt, az élettel kapcsolatos jelenségek felismerését, rendszerezését és részletes leírását végezte. A 20. század elejétől kezdődően a fizikai és kémiai kutatások eljutottak az anyag korábban ismeretlen alkotóelemeinek leírásáig. Az ezt követően megjelenő új, molekuláris szintű mérési lehetőségek által a 20. században a biológia jelentős részben kvantitatív tudománnyá vált. A biológia kvantitatívabbá válására egy kíváló példa a hálózatos leírási módszer. Ha egy biológiai jelenségben (i) sikerül a pár kölcsönhatások nagy részét azonosítani, és ha (ii) a kölcsönhatások nagy része valóban két elem között lép fel, akkor a hálózatos leírási módszer igen intuitív és hatékony lehet. A megfigyelt valós jelenség modelljeként használt hálózat könnyen lerajzolható és értelmezhető, továbbá számos matematikai (például lineáris algebrai és kombinatorikai) eszközhöz kíválóan kapcsolható.

A technológiai és társadalmi rendszerekhez képest a biológiai (élő) rendszerekben jóval gyakoribb, hogy két résztvevő (például fehérje) kölcsönhatását sok másik résztvevő (fehérje) állapota módosíthatja. Például ha két fehérje térbeli szerkezete (felszíne) olyan, hogy ez képessé teszi kettőjüket a kapcsolódásra, akkor is még nagyon sok körülmény (például fizikai és kémiai paraméter) befolyásolhatja, hogy a kölcsönhatásuk valóban bekövetkezik-e. A biológiai feladatokat ugyanis gyakran kettőnél több, egymással kapcsolatban álló molekula (például fehérje) végzi [1]. Ha a fehérjék egy csoportja képes egy biológiai feladat elvégzésére, akkor az adott csoportot szokás modulnak nevezni[1]. További érdekesség, hogy egy fehérje modul tagjai gyakran nincsenek azonos időpontban mindannyian szorosan összekapcsolódva, mégis az adott feladat elvégzéséhez gyakori kölcsönhatásaikra szükség van. (A „fehérje modul”-tól kissé eltér

[1] A „modul” szó és a „fehérjék által alkotott funkcionális modul” kifejezés a biológiai hálózatos irodalomban igen elterjedt és elfogadott. A „modul” szónak a molekuláris biológiában van más – jóval régebben használatos – jelentése is. A fehérjék szerkezetét és funkcióit kutató szakirodalomban az aminosav szekvencia, a szerkezet és a funkció alapján együttesen megállapítható jellegzetes fehérje szakaszokat gyakran doménnek hívják, és a több élőlényben előforduló doméneket szokás modulnak nevezni.

a „fehérje komplex”, mert ez utóbbi általában azt jelöli, hogy a résztvevő fehérjék mindannyian folyamatosan fizikailag kapcsolatban vannak, tehát azonos időpontban jelen vannak ugyanazon a helyen.) A fehérje-fehérje kölcsönhatási hálózatok átfedő moduljainak azonosítása érdekében először fel kell sorolni az összes megfigyelhető pár kölcsönhatást, és utána kereshetőek a pár kölcsönhatásokból kapott hálózatban olyan csoportok, amelyeken belül sokkal nagyobb a kapcsolatok sűrűsége, mint a hálózatban máshol átlagosan.

# Célkitűzések

A dolgozatban bemutatott kutatás célja, hogy az élő rendszerek molekuláris biológiai szintű jelenségeiben a modern hálózatkutatás és a hálózati modul keresés segítségével általános érvényű összefüggéseket azonosítson, valamint ilyen összefüggések felismerésére alkalmas eszközöket biztosítson. A kutatási eredményeket a dolgozat három fejezetben mutatja be. A dolgozat első és második fejezetének célkitűzése döntően alapkutatási eredmények ismertetése, míg a harmadik fejezetben az alkalmazásokon van a hangsúly.

# Vizsgálati módszerek és matematikai eszközök

A dolgozat főként hálózatokkal kapcsolatos és DNS szekvencia alapú módszereket használ.

Egy hálózat (a legegyszerűbb esetben) $N$ darab csúcspontból áll, amelyek közül $E$ darab csúcs pár össze van kapcsolva egy egyszerű éllel. A hálózat egy csúcspontjával szomszédos (hozzá éllel kapcsolt) további csúcspontok száma az adott csúcspont fokszáma. A hálózat (gráf) $k$ csúcspontja egy $k$-klikket alkot pontosan akkor, ha a $k$ csúcs közül tetszőleges pár között van él. Két $k$-klikk egymással szomszédos, ha pontosan $(k-1)$ csúcspontjuk közös. Egy gráfban egy $k$-klikk perkolációs klaszter a csúcsoknak egy olyan maximális részhalmaza, amely bejárható $k$-klikk szomszédsági kapcsolatokon keresztül. A $k$-klikk perkolációs módszer (Clique Percolation Method, CPM) által definiált átfedő hálózati modulok a $k$-klikk perkolációs klaszterek. Ezek a modulok egymással átfedhetnek $(k-1)$-nél kevesebb csúcsot tartalmazó klikkek által. A $k$-klikk fogalma

kiterjeszthető irányított vagy súlyozott élekből álló hálózatokra is. Ennek megfelelően a $k$-klikk perkolációs módszerhez hasonlóan definiálható két további módszer: a CPMw és a CPMd (w: weighted, d: directed).

A dolgozatban felhasznált DNS szekvencia alapú kapcsolatok (géneket, mint csúcsokat összekötő hálózati élek) alapja a heurisztikus BLAST szekvencia-illesztő algoritmus. A BLAST (vagy más szekvencia-illesztő algoritmus) segítségével szokás azt definiálni, hogy ha egy $F_1$ faj (élőlény) $g_1$ génjéhez az $F_2$ élőlény génjei közül leginkább hasonló a $g_2$, és a $g_2$ génhez az $F_1$ élőlény génjei közül leginkább hasonló a $g_1$, akkor ez a két gén egy „Bidirectional Best Hit”-et (BBH) alkot. A gének hasonlósága kapcsán szintén központi fogalom a homológia. Két gén egymás homológja, ha egy közös ős leszármazottai. Ez a közös ős általában egy olyan gén, ami napjainkban már nem létezik. Ha ez a két gén azonos fajban található, akkor egymás paralógjai. Ha két különböző faj genomjában találhatóak, akkor egymás ortológjai. Megjegyzés: a BBH nem mindig jelent ortológ kapcsolatot [2, 3].

# Megjegyzések

A doktori dolgozat (a pályázati kiírásban: „doktori mű”) és a tézisfüzet áttekinthetősége érdekében a dolgozat mindhárom fejezetének „Eredmények” részében a számozott alfejezetek számozása azonos a tézisfüzetben található tézispontok számozásával: **1.1**, **1.2**, **1.3**, **1.4**, **1.5**, **2.1**, **2.2**, **3.1** és **3.2**. A két dokumentumban a tézispontokhoz kapcsolódó saját publikációk számozása szintén azonos: [T1, T2, T3, T4, T5, T6, T7, T8, T9, T11, T10, T12]. A dolgozat minden alfejezetének címe mellett megtalálható azoknak a saját publikációknak a sorszáma, amelyeknek az eredményeit az adott alfejezet használja. Ugyanígy a tézisfüzetben minden tézispont címe mellett megtalálható az adott tézisponthoz felhasznált saját publikációk sorszáma.

A Ph.D. fokozatom megszerzése (2004 március) óta részt vettem több olyan kutatásban, amelyeknek a témája eltér a D.Sc. dolgozat témájától. Ezeknek a kutatásoknak az eredményeként nemzetközi referált folyóiratokban megjelent 7 cikk, amelyeknek a számozása a doktori dologzatban és a tézisfüzetben szintén

azonos: [M1, M2, M3, M4, M5, M6, M7]. A doktori dolgozathoz hasonlóan a tézisfüzet PDF fájlja is tartalmaz hiperlinkeket. A tézisfüzet PDF fájljából a hivatkozott publikációk elérhetőek a cikk alatt található „teljes cikk PDF” szövegű linkre vagy DOI (Digital Object Identifier) linkre történő klikkeléssel.

# Új tudományos eredmények

**Fehérje-fehérje kölcsönhatási hálózatok moduljai**

A doktori dolgozat első fejezete fehérje-fehérje kölcsönhatási (Protein-Protein Interaction, PPI) hálózatokban történő modul keresés eredményeit mutatja be. A módszerünk segítségével azonosított átfedő fehérje-fehérje kölcsönhatási modulok [T1, T2, T3] azért jelentősek, mert (i) hozzájárulnak ismeretlen funkciójú fehérjék biológiai funkcióinak előrejelzéséhez és (ii) segítik már ismert funkciójú nagyobb csoportokon belüli kisebb csoportok azonosítását. Az (i) pontban említett egyedi fehérje funkciók előrejelzésével kapcsolatban egy irodalmi összefoglaló [4] beszámol munkánkról, és nagy számú további független hivatkozás is érkezett cikkeinkre. A (ii) pontban említett fehérje-fehérje kölcsönhatási részcsoportoknak a gyógyászati szerepe lehet jelentős, mert egy modul fehérjéinek részleges gátlásával elérhető, hogy pontosan az adott modul biológiai funkciója jóval erősebben legyen gátolva, mint a vele átfedő más modulok funkciói.

A következő tézispontok a doktori dolgozat 1. fejezetében található számozott alfejezetekkel azonos számozásúak, és az ott leírt eredményeket összegzik:

1.1. **Átfedő modulok azonosítása fehérje-fehérje kölcsönhatási (PPI) hálózatokban [T1, T2, T3]**

*Csapatmunkával elért eredményeink:* Részt vettem a CPM (Clique Percolation Method) hálózati modulkereső algoritmus kidolgozásában, amely egymással átfedő és a (hálózati) környezethez képest nagy belső élsűrűséggel rendelkező hálózati modulokat azonosít. Javaslatot tettünk a módszerben felhasznált $k$-klikk méret paraméter lehetséges értékei közül az

optimális érték kiválasztására. A modulkereső CFinder alapkutatási program a [T1] publikációink 2005-ös megjelenése óta non-profit célra ingyenesen elérhető, és a felhasználók a weboldalon (Manual, FAQ) és e-mailben támogatást kapnak. A CFinder tartalmazza az általunk kidolgozott súlyozott és irányított klikk perkolációs módszert.

*Egyéni munkával elért eredményeim:* A klikk perkolációs módszer teszteléséhez hálózatokat gyűjtöttem: PPI, szó asszociációs és tudományos társszerzőségi hálózatokat. Az adatokat az eredeti formájukból hálózatokká alakítottam és az adat hibákat szűrtem. Beprogramoztam nagy hálózatokra alkalmazhatóan a(z irányítatlan) csúcs köztiség (node betweenness) számítását és az agglomeratív Girvan-Newman klaszterezést. A [T3] cikkünkben a súlyozott klikk perkolációs módszer vizsgálata érdekében analitikusan levezettem azt a függvényt, amely a perkolációs pontban a módszer két paramétere közötti kapcsolatot mutatja. Ezt az eredményt ugyanebben a cikkben a társszerzők numerikus eredményei igazolták.

### 1.2. **PPI hálózatok átfedő moduljainak biológiai funkciói [T1, T2]**

*Csapatmunkával elért eredményeink:* A $k$-klikk-perkolációs módszerrel azonosítottuk az élesztőgomba (*S. cerevisiae*) fehérje-fehérje kölcsönhatási (PPI) hálózatának átfedő moduljait.

*Egyéni munkával elért eredményeim:* A CFinder szoftver segítségével a modulok azonosítását elvégeztem több különböző paraméterrel, több esetben automatizálva az eredmények feldolgozását. Az adatok feldolgozásához és az eredmények elemezhetőségéhez átalakítottam a PPI hálózatokban szereplő gén és fehérje neveket olyan név típusra, amelyet a funkciókat felsoroló Gene Ontology [5] adatbázis használ. A PPI hálózati modulok kiszámítása után részben automatizáltam a csoportok legszignifikánsabb közös funkcióinak szisztematikus keresését. Azonosítottam a modulok statisztikailag legszignifikánsabb biológiai funkcióit.

1.3. **Fehérjék modul száma PPI hálózatok átfedő moduljaiban [T1]**

*Csapatmunkával elért eredményeink:* A tézispontban tárgyalt modulokat a csapatmunkában kidolgozott klikk perkolációs módszerrel határoztam meg.

*Egyéni munkával elért eredményeim:* A doktori dolgozathoz végzett számításokkal megállapítottam, hogy a három vizsgált élőlény friss PPI hálózataiban a modulokban található csúcsok átlagos modul száma $1.1$ és $1.2$ közötti, és a három élőlény legtöbb modullal rendelkező fehérjéi az átlagosnál nagyobb számú biológiai funkcióval rendelkeznek.

1.4. **Átfedő PPI hálózati modulok mérete [T1]**

*Csapatmunkával elért eredményeink:* A tézispontban tárgyalt modulokat a csapatmunkában kidolgozott klikk perkolációs módszerrel határoztam meg.

*Egyéni munkával elért eredményeim:* Szintén a doktori dolgozathoz kiszámítottam ebben a három friss hálózatban a modulok méretének eloszlását. Az azonosított modulok méretének eloszlása hatványfüggvény-szerű, széles eloszlás. Ez az eloszlás típus azonos a PPI hálózat csúcsainak fokszámánál ismert eloszlás típusával, és arra utal, hogy a vizsgált méret tartományon ($3$ és $30$ közötti csúcs) a moduloknak nincsen karakterisztikus mérete.

1.5. **PPI hálózatok átfedő moduljainak kapcsolat számai [T1]**

*Csapatmunkával elért eredményeink:* Az azonosított átfedő PPI hálózati modulok segítségével definiáltunk egy új mennyiséget, a modulok fokszámát: egy modul fokszáma a vele átfedő modulok száma. A modulok közötti átfedések segítségével definiáltuk a modulok hálózatát. Ebben a hálózatban egy csúcspont az eredeti hálózat egy modulja, és egy él az eredeti hálózat két modulja közötti átfedést jelöl. Ez a hálózat az eredeti fehérje-fehérje kölcsönhatási hálózatot mutatja be egy jóval tömörebb formában. Ebben a tömörített (átskálázott) hálózatban azt találtuk, hogy az eredeti

PPI hálózattól eltérően a fokszám eloszlás kezdeti szakasza gyorsan lecsengő, exponenciális típusú, majd az eloszlás vége hatványfüggvényszerű. Ennek az exponenciális lecsengésnek az oka a klikk méret paraméter, amely egy karakterisztikus csúcs számot okoz, és ezen keresztül egy karakterisztikus modul kapcsolat számot (modul fokszámot) jelent.

*Egyéni munkával elért eredményeim:* A doktori dolgozathoz végzett számításokkal megállapítottam, hogy a vizsgált három élőlényben (*S. cerevisiae*, *M. musculus*, *H. sapiens*) a PPI hálózati modulok kapcsolat számának eloszlása teljes egészében exponenciális típusú, nincsen hatványfüggvényszerű szakasza.

**Transzkripció és transzláció szabályozási hálózatok moduljai**

A molekuláris biológia fejlődése során a DNS szekvenciában tárolt információ legkorábban ismert formája a fehérjéket kódoló gén volt. A génekből az élő sejtekben először mRNS (messenger RNA, azaz: „hírvivő” RNS), majd aminosav lánc (fehérje) készül. Erre és további ismeretekre építette Francis Crick a molekuláris biológia centrális dogmájának nevezett elméletet [6]. A centrális dogma szerint a három, információt hordozó biológiai molekula típus (a DNS, az RNS és a fehérje) közötti összesen 9 irány közül nem mindegyik irányban lehetséges információ áramlás. A jelenlegi ismeretek alapján az információ áramlás lehetséges mindegyik irányban.

A centrális dogma által megengedett egyik információ áramlási irány a DNS$\rightarrow$RNS irány, amelyre példa a transzkripció. A transzkripció során a DNS-ben található nukleotidok sorrendje (a DNS szekvencia) által meghatározott információ alapján RNS molekulák készülnek. A legtöbb biológiai folyamathoz hasonlóan a transzkripció is szabályozható: a transzkripció szabályozását a transzkripciós faktornak (TF) nevezett fehérjék végzik. Ezek a fehérjék képesek a target (célpont) gének átírási gyakoriságát csökkenteni vagy növelni. Hálózatos formában ezt a kölcsönhatást lehet úgy ábrázolni, hogy a TF fehérjétől a target génhez egy negatív (gátló) vagy pozitív (serkentő) kölcsönhatást jelölő irányított hálózati él fut. A transzkripció szabályozási hálózatok elemzésekor egy gyakori

egyszerűsítés (közelítés) az a feltételezés, hogy a gének és a fehérjék közötti megfeleltetés kölcsönösen egyértelmű. Ha ezt a közelítést elfogadjuk, akkor egy gén és a hozzá tartozó fehérje egyetlen hálózati csúcspontként ábrázolható, tehát a TF hálózat egy gének (vagy fehérjék) közötti szabályozási hálózat.

Az élő sejtek képesek a hírvivő (messenger) RNS-ből aminosav láncot készíteni. Ennek a folyamatnak a neve transzláció, magyarul (le)fordítás (mRNS-ről aminosav láncra). A sikeres transzlációhoz szükséges, hogy a fordítást végző fehérje-RNS komplex (a riboszóma) le tudja olvasni az egyszálú mRNS nukleotidjait. A rövid gátló RNS-ek (amelyeket összefoglaló névvel gyakran mikroRNS-nek neveznek) a DNS kettős spiráljánál megismert módon képesek templát képződéssel messenger DNS-ekhez (RNS-ek) kapcsolódni, és ezzel a transzlációt gátolni. Ennek a gátlási módszernek a gyakori neve „RNA silencing” (RNS csendesítés).

A következő tézispontok a doktori dolgozat 2. fejezetében található számozott alfejezetekkel azonos számozásúak, és az ott leírt eredményeket összegzik:

### 2.1 Irányított hálózati modulok az élesztőgomba transzkripció szabályozási hálózatában [T4, T5, T6]

*Csapatmunkával elért eredményeink:* Irányított élekből álló hálózatokban definiáltuk az irányított $k$-klikket, majd az irányított $k$-klikkekből képezhető irányított $k$-klikk perkolációs klaszterek segítségével definiáltuk az irányított $k$-klikk perkolációs hálózati modulkereső algoritmust (CPMd) [T4]. Az átfedő irányított modulokban található csúcspontok összehasonlítására a csúcsok relatív kimenő fokszámát javasoltuk és használtuk. A transzkripció szabályozással kapcsolatos friss eredményeket és egy friss kísérleti módszert mutat be egy, a Cell folyóiratban megjelent kitekintő cikk, amelynek első szerzője vagyok [T6].

*Egyéni munkával elért eredményeim:* A PPI hálózatokhoz hasonlóan (az adatok feldolgozásához és az eredmények elemezhetőségéhez) átalakítottam a TR hálózatokban szereplő gén és fehérje neveket olyan név típusra, amelyet a funkciókat felsoroló Gene Ontology [5] adatbázis használ. A

CPMd módszerrel meghatároztam több, az irodalomban használt transzkripció szabályozási hálózat irányított moduljait és a GO TermFinder funkció keresővel a modulok legjelentősebb biológiai funkcióit. A CMPd algoritmus tesztelésére a transzkripció szabályozási hálózatokon túl a következő adatokat gyűjtöttem és alakítottam hálózatos formába [T4]: irányított szó asszociációk és a Google statikus saját oldalai közötti hiperlinkes kapcsolatok. Az utóbbi hálózat elérhető a http://CFinder.org/data helyen. Szintén a CPMd algoritmus tesztelése érdekében numerikusan meghatároztam azt az él sűrűséget (a kritikus él sűrűséget a rendszer méret függvényében), amelynél az irányított Erdős-Rényi hálózatban bekövetkezik az irányított $k$-klikk perkolációs átalakulás (a [T4] publikáció 3. ábrája).

Az irányított TR (transcription regulation) hálózati modulokban a csúcspontok relatív kimenő fokszáma és modul száma alapján a csúcspontok két nagy csoportját azonosítottam. Ez a két csoport biológiailag valószínűleg a „fő" irányító csúcspontokat (fehérjéket) és az „alközpontokat" jelenti. A doktori dolgozathoz a számításokat megismételtem az eredeti cikkben [T4] szereplő egy TR hálózat helyett két TR hálózatra.

Egy, a CPMd-től eltérő TR hálózati modul keresés [T5] során beprogramoztam kis irányított részgráfok keresését és az előfordulási számuk szignifikanciájának meghatározását olyan élkeverési tesztekkel, amelyek változatlanul hagyják a csúcsok ki- és be-fokszámait, és a csúcsok rögzített csoportjainak (a TR hálózat „szintjeinek") elem számát.

## 2.2 Emberi mikroRNS-ek szerepeinek összehasonlítása a transzláció csendesítési hálózat moduljai alapján [T7]

*A vezetésemmel végzett csapatmunka során elért eredményeink:* Megmértük a korábban (számítógépes szekvencia-elemzéssel) előrejelzett mikroRNS→mRNS kölcsönhatás listák hasonlóságát. Eredményeink szerint bármely két adatbázis szerinti $1000$ legerősebb kölcsönhatás relatív átfedése legfeljebb néhány százaléknyi. Mivel az egyes mikroRNS-mRNS kölcsönhatások száma nagy és önmagában egy-egy kölcsönhatás

erőssége kevéssé megbízható, ezért javasoltuk, hogy a kölcsönhatások nagyobb csoportokba történő összesítése informatívabb lehet, mint az egyenként történő elemzésük. Definiáltuk a mikroRNS-ek hálózatát az általuk végzett szabályozási feladatok hasonlósága alapján és a CFinder modulkereső programmal meghatároztuk a mikroRNS-ek moduljait. Miután a mikroRNS-ek moduljait összehasonlítottuk a mikroRNS-ek expressziójával, definiáltuk a modulokban lévő mikroRNS-ek fontosságát („essentiality of a microRNA"). Ha egy mikroRNS (vagy általánosabban: szabályozó egység) feladata több más mikroRNS feladatához hasonló, de ezt a feladatot speciális körülmények között végzi, akkor a definíciónk alapján ez a mikroRNS fontosabb („essential") a többinél. Biológiailag mindez azt jelenti, hogy az adott mikroRNS eltávolítása nagyobb eséllyel okoz mérhető fenotípus változást.

*Egyéni munkával elért eredményeim:* A fehérje-fehérje kölcsönhatásokhoz és a TR hálózatokhoz hasonlóan a mikroRNS-ek esetén is elvégeztem számos név átalakítást a különböző forrásokból származó adatok összehasonlíthatósága érdekében. Kiszámítottam az egyes mikroRNS-ek fontosságát. Ezeknek a lépéseknek egy részét a társszerzők is elvégezték. Teszteltem az eredményeket úgy, hogy a Pearson (kovariancia) korreláció helyett a Spearman (sorrend) korrelációt használtam. Teszteltem több eredményt élkeveréssel (az irányított mikroRNS→mRNS hálózat csúcsainak be- és ki-fokszámait változatlanul tartva). A tesztek megtalálhatóak a [T7] publikáció kiegészítő anyagában („supplement"-jében).

**Jelátviteli hálózatok moduljai (útvonalai)**

A jelátvitel (signal transduction) a biokémiai jelek továbbítása a sejt külső felszínétől a sejt belsejéig, általában a sejten kívül található jelző fehérjéktől (ligand fehérjéktől) a sejtmagban található transzkripciós faktor fehérjékig [7, 8]. Ha a jel terjedése során két fehérje egymással kölcsönhatásba lép, akkor – jelentős egyszerűsítéssel – ezt a két fehérjét a jelátviteli hálózat két csúcspontjaként

ábrázolhatjuk és a két fehérje közötti kölcsönhatást a jelátviteli hálózat egy irányított éle jelölheti.

Az előző két fejezettől eltérően ebben a fejezetben az alkalmazásokon van a hangsúly. Az alkalmazások során a hálózatok segítségével történő elemzésekhez szükség van jó minőségű hálózatokra (kölcsönhatás listákra) és további adatokra (annotációkra). A jelátviteli kölcsönhatások esetében több élőlény jelátviteli útvonalait átfogóan és egységesen kezelő adatbázis a 2010-es évek előtt nem volt elérhető, de azóta a fejlődés jelentős. Véleményem szerint az itt bemutatott SignaLink adatbázis ebben a folyamatban vesz részt eredményesen. A SignaLink adatai jelenleg elérhetőek adatbázisokat integráló központi adatbázisokban. A jelátviteli adatbázisok fejlődési folyamatának részeként a SignaLink-et is bemutatja egy 2015-ben megjelent összefoglaló cikk 1-es ábrája [9].

A következő tézispontok a doktori dolgozat 3. fejezetében található számozott alfejezetekkel azonos számozásúak, és az ott leírt eredményeket összegzik:

3.1 **Átfedő jelátviteli útvonalak összeállítása egységesített gyűjtési kritériumokkal és adatszerkezettel [T8, T9]**

*Csapatmunkával elért eredményeink:* Két fős csapatként együttműködve pontosan leírtuk a SignaLink adatbázis „manual curation” típusú („kézi”, tehát nem automatizált) adat gyűjtésének kritériumait. A részletes leírás a 2010-ben megjelent [T8] publikációnk kiegészítő anyagában található. A SignaLink adatai ma (2015-ben) már megtalálhatóak központi „aggregált” adatbázisokban, például a következőkben: FlyBase, WormBase, UniProt (a dolgozat 3. fejezetének első oldala mutat példákat). Ez jelentős részben a SignaLink adatbázis minőségének és időszerűségének elismerése. A SignaLink adatbázis friss verziója elérhető a http://SignaLink.org web címen, a kapcsolódó publikáció 2013-ban jelent meg [T9].

*Egyéni munkával elért eredményeim:* A társszerzők folyamatos javaslatai és visszajelzései alapján létrehoztam a SignaLink weboldal és az adat tárolás első verzióját, amely jelenleg a http://signalink1.netbiol.org web címen érhető el. A weboldalon elérhetőek – többek között – fehérje név keresési és interaktív hálózat elemzési funkciók. A SignaLink jelátviteli útvonalaiban lehetséges gyógyszer célpontok kijelöléséhez kereséseket végeztem, és azok összesített eredményeit ábrázoltam [T10].

3.2 **Fehérje csoportok jelátviteli elemzése online, valamint részvétel jelátviteli szerepek előrejelzésében és lehetséges gyógyszer célpontok kijelölésében [T10, T11, T12]**

*Csapatmunkával elért eredményeink:* A SignaLink adatai segítségével előrejeleztük jelátviteli útvonalak (modulok) új fehérjéit. Hat *C. elegans* fehérje esetén az adatokból *in silico* megjósolt eredményt (a Notch jelátviteli útvonalban való részvételt) kísérletek igazolták [T11]. Létrehoztuk a PathwayLinker online kereső szolgáltatást, amelynek célja, hogy egy vagy több fehérje hálózati szomszédai és a talált összes fehérje részcsoportjainak vagy egészének jelátviteli funkciói gyorsan kereshetőek legyenek [T12].

*Egyéni munkával elért eredményeim:* Beprogramoztam a PathwayLinker számára a fehérje neveket (szinonimákat) és a kölcsönhatásokat tároló adatbázist, valamint a kereső és a weboldal közelítőleg 80%-át. Ezek során megterveztem és megvalósítottam többek közt egy index-elést, amely a Linux fájlrendszerben egymásba ágyazott könyvtár nevekkel tárolja a kereshető karakterláncokat az egymás utáni karaktereik szerint. A PathwayLinker Help oldalain megírtam az adatbázis és a szolgáltatás részletes megvalósítási és felhasználási dokumentációját (utóbbit közös megbeszélések alapján), és nagyjából 80%-ban a PathwayLinker API, PDF és interaktív hálózat megjelenítő kimenetét. A társszerzők javaslatai és a saját javaslataim alapján a felhasználók számára a megjelenített hálózat és az elemző ablakok („dialogs”) között kommunikációt építettem be a CytoscapeWeb és a jQuery szoftver csomagok felhasználásával. A szolgáltatás a http://PathwayLinker.org web címen érhető el [T12].

# Irodalmi hivatkozások listája

[1] L. H. Hartwell, J. J. Hopfield, S. Leibler, A. W. Murray: From molecular to modular cell biology. *Nature* **402**, C47–52 (1999).
Link a cikk ingyenes PDF verziójára

[2] Y. I. Wolf, E. V. Koonin: A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol. Evol.* **4**, 1286–1294 (2012). DOI: 10.1093/gbe/evs100

[3] D. A. Dalquen, C. Dessimoz: Bidirectional Best Hits miss many orthologs in duplication-rich clades such as plants and animals. *Genome Biol. Evol.* **5**, 1800–1806 (2013). DOI: 10.1093/gbe/evt132

[4] R. Sharan, I. Ulitsky, R. Shamir: Network-based prediction of protein function. *Mol. Syst. Biol.* **3**, 88 (2007). DOI: 10.1038/msb4100129

[5] M. Ashburner, *et. al.*: Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000). DOI: 10.1038/75556

[6] F. Crick: Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).
DOI: 10.1038/227561a0

[7] A. Pires-daSilva, R. J. Sommer: The evolution of signalling pathways in animal development. *Nat. Rev. Genet.* **4**, 39–49 (2003). DOI: 10.1038/nrg977

[8] A. Bauer-Mehren, L. I. Furlong, F. Sanz: Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol. Syst. Biol.* **5**, 290 (2009). DOI: 10.1038/msb.2009.47

[9] S. Chowdhury, R. R. Sarkar: Comparison of human cell signaling pathway databases – evolution, drawbacks and challenges. *Database*, cikk szám: bau126 (2015). DOI: 10.1093/database/bau126

# A tézispontokhoz kapcsolódó publikációk társszerzőségemmel a Ph.D. fokozat megszerzése után

[T1] G. Palla, I. Derényi, I. Farkas, T. Vicsek: Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814–818 (2005).
Kivonat, Teljes cikk PDF, Kiegészítő anyagok PDF, Weboldal

[T2] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi, T. Vicsek: CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* **22**, 1021–1023 (2006).
Kivonat, Teljes cikk PDF, Weboldal

[T3] I. J. Farkas, D. Ábel, G. Palla, T. Vicsek: Weighted network modules. *New Journal of Physics* **9**, 180:1–18 (2007).
Kivonat, Teljes cikk PDF, Weboldal

[T4] G. Palla, I. J. Farkas, P. Pollner, I. Derényi, T. Vicsek: Directed network modules. *New Journal of Physics* **9**, 186:1–21 (2007).
Kivonat, Teljes cikk PDF, Weboldal

[T5] I. J. Farkas, C. Wu, C. Chennubhotla, I. Bahar, Z. N. Oltvai: Topological basis of signal integration in the transcriptional-regulatory network of the yeast, *Saccharomyces cerevisiae*. *BMC Bioinformatics* **7**, 478:1–12 (2006).
Kivonat, Teljes cikk PDF

[T6] I. J. Farkas, Q. K. Beg, Z. Oltvai: Exploring transcriptional regulatory networks in the worm. *Cell* **125**, 1032–1034 (2006).
Kivonat, Teljes cikk PDF

[T7] G. Boross, K. Orosz, I. Farkas: Human microRNAs co-silence in well-separated groups and have different predicted essentialities. *Bioinformatics* **25**, 1063–1069 (2009).
Kivonat, Teljes cikk PDF, Kiegészítő anyagok PDF

[T8] T. Korcsmáros *, I. J. Farkas *, M. S. Szalay, P. Rovó, D. Fazekas, Z. Spiró, C. Böde, K. Lenti, T. Vellai, P. Csermely: Uniformly curated signaling pathways reveal tissue-specific cross-talks and support drug target discovery. *Bioinformatics* **26**, 2042–2050 (2010).
*: egyenlő hozzájárulás (megosztott első szerzők)
Kivonat, Teljes cikk PDF, Kiegészítő anyagok PDF, Weboldal

[T9] D. Fazekas *, M. Koltai *, D. Türei *, D. Módos, M. Pálfy, Z. Dúl, L. Zsákai, M. Szalay-Bekő, K. Lenti, I. J. Farkas, T. Vellai, P. Csermely, T. Korcsmáros: SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks. *BMC Systems Biology* **7**, 7:1–15 (2013).
*: egyenlő hozzájárulás (megosztott első szerzők)
Kivonat, Teljes cikk PDF, Weboldal

[T10] I. J. Farkas, T. Korcsmáros, I. A. Kovács, Á. Mihalik, R. Palotai, G. I. Simkó, K. Z. Szalay, M. Szalay-Bekő, T. Vellai, S. Wang, P. Csermely: Network-based tools for the identification of novel drug targets. *Science Signaling* **4**, pt3:1–12 (2011).
Kivonat, Teljes cikk PDF

[T11] T. Korcsmáros, M. S. Szalay, P. Rovó, R. Palotai, D. Fazekas, K. Lenti, I. J. Farkas, P. Csermely, T. Vellai: Signalogs: orthology-based identification of novel signaling pathway components in three metazoans. *PLoS ONE* **6**, e19240:1–13 (2011).
Kivonat, Teljes cikk PDF

[T12] I. J. Farkas, Á. Szántó-Várnagy, T. Korcsmáros: Linking proteins to signaling pathways for experiment design and evaluation. *PLoS ONE* **7**, e36202:1–5 (2012).
Kivonat, Teljes cikk PDF

# További publikációk társszerzőségemmel a Ph.D. fokozat megszerzése után

[M1] G. Palla, I. Farkas, I. Derenyi, A.-L. Barabasi, T. Vicsek: Reverse engineering of linking preferences from network restructuring. *Phys. Rev. E* **70**, 046115 (2004). DOI: 10.1103/PhysRevE.70.046115

[M2] I. J. Farkas, T. Vicsek: Initiating a mexican wave: an instantaneous collective decision with both short and long range interactions. *Physica A* **369**, 830-840 (2006). DOI: 10.1016/j.physa.2006.01.075

[M3] P. Pollner, G. Palla, D. Ábel, A. Vicsek, I. J. Farkas, I. Derényi, T. Vicsek. Centrality properties of directed module members in social networks. *Physica A* **387**, 4959 (2008). DOI: 10.1016/j.physa.2008.04.025

[M4] G. Palla, I. J. Farkas, P. Pollner, I. Derényi, T. Vicsek: Fundamental statistical features and self-similar properties of tagged networks. *New J. Phys.* **10**, 123026 (2008). DOI: 10.1088/1367-2630/10/12/123026

[M5] A. Szanto-Varnagy, P. Pollner, T. Vicsek, I. J. Farkas: Scientometrics: untangling the topics. *National Science Review*, cikk szám: nwu027 (2014). DOI: 10.1093/nsr/nwu027

[M6] M. Xu, Y. Wu, Y. Ye, I. Farkas, H. Jiang, Z. Deng: Collective crowd formation transform with mutual information-based runtime feedback. *Comp. Graph. Forum* (2014). DOI: 10.1111/cgf.12459

[M7] I. J. Farkas, J. Kun, Y. Jin, G. He, M. Xu: Keeping speed and distance for aligned motion. *Phys. Rev. E* **91**, 012807 (2015). DOI: 10.1103/PhysRevE.91.012807