



ComPPI: kompartmentalizált fehérje-fehérje kölcsönhatási adatbázis

Biológia Országos Tudományos Diákköri Dolgozat - 2013

Szerző

Veres Dániel

Semmelweis Egyetem, Általános Orvostudományi Kar, VI. évfolyam

Témavezetők

Prof. Csermely Péter

Semmelweis Egyetem

Orvosi Vegytani, Molekuláris Biológiai és Patobiokémiai Intézet

Dr. Korcsmáros Tamás

Eötvös Loránd Tudomány Egyetem

Genetikai Tanszék

Tartalomjegyzék

I.	Összefoglaló	<i>(3. old.)</i>
II.	Bevezető	<i>(3. old.)</i>
III.	Célkitűzések	<i>(4. old.)</i>
IV.	Források	
	<i>Organizmusok</i>	<i>(5. old.)</i>
	<i>Lokalizációs forrás adatbázisok</i>	<i>(5. old.)</i>
	<i>Fehérje-fehérje kölcsönhatási forrás adatbázisok</i>	<i>(6. old.)</i>
V.	Módszerek	
	<i>Az adatbázisok összesítése</i>	<i>(8. old.)</i>
	<i>Az adatbázis szerkezete</i>	<i>(8. old.)</i>
	<i>A ComPPI megbízhatósági érték leírása</i>	<i>(9. old.)</i>
VI.	Eredmények és diszkusszió	
	<i>Összesített adatok</i>	<i>(12. old.)</i>
	<i>A kiszolgáló felület – fehérje keresés</i>	<i>(13. old.)</i>
	<i>Fehérje keresés – alkalmazási példa</i>	<i>(13. old.)</i>
	<i>A kiszolgáló felület – adatbázis letöltés</i>	<i>(15. old.)</i>
	<i>Adatbázis letöltés – alkalmazási példa</i>	<i>(15. old.)</i>
	<i>A ComPPI megbízhatósági érték alkalmazása</i>	<i>(17. old.)</i>
VII.	Konklúzió	<i>(18. old.)</i>
VIII.	Szerzői hozzájárulás	<i>(19. old.)</i>
IX.	Köszönetnyilvánítás	<i>(19. old.)</i>
X.	Irodalom	<i>(20. old.)</i>

I. Összefoglaló

ComPPI: kompartmentalizált fehérje-fehérje kölcsönhatási adatbázis

A fehérje-fehérje kölcsönhatási hálózatok (interaktómok) nagyban segítik az ép és a beteg sejt szerveződésének és funkcióinak rendszerszintű megértését. Az interaktómokban a sejten belüli lokalizációt eddig csak ritkán vették figyelembe. Emiatt a legtöbb esetben az interaktómok sok hamis-pozitív kölcsönhatást tartalmaztak, hiszen a kölcsönható fehérjék a sejten belül a valóságban nem tartózkodtak azonos helyen.

Az interaktómok hamis kapcsolatainak szűrésére az egyik módszer a szubcelluláris lokalizáció hozzárendelése a fehérjékhez. A munka során négy élőlény (*S. cerevisiae*, *C. elegans*, *D. melanogaster* és *H. sapiens*) adatbázisát készítettük el. Ennek alapját a fehérje-fehérje kölcsönhatási adatok adták, amelyek kapcsolatait több adatbázis összegzéséből származó lokalizációs információk alapján korrigáltuk.

Az elkészült kompartmentalizált fehérje-fehérje kölcsönhatási adatbázist ComPPI-nak neveztük el (www.linkgroup.hu/comppi). Az adatbázist kiszolgáló webes felületen a letöltendő adatbázisok struktúrája a későbbi vizsgálatokhoz optimalizálható.

A kompartmentalizált interaktómok vizsgálata rávilágított arra, hogy a fehérjék lokalizációs térképe az eddig használt interaktómok tulajdonságait alapvetően megváltoztatja. Az adatbázis segítségével így nagyobb biztonsággal kereshetőek olyan hálózati pontok vagy kapcsolatok, amelyek gyógyszeres befolyásolásával lehetőség nyílna az olyan komplex patomechanizmussal rendelkező betegségek kezelésére is, mint amilyen a diabétesz vagy a rák.

II. Bevezetés

A biológiai rendszerek szépsége komplexitásukban rejlik. Az egyes molekuláktól egészen a Föld ökoszisztémájáig a látszólag önálló egységek folyamatosan változó, dinamikus egyensúlya biztosítja a stabilitást, és a fenntartható fejlődést. Ezeket a komplex rendszereket

tekinthetjük hálózatoknak, ami segíti a bonyolult működések megértését [1]. A 21. század egyik fő kutatási iránya a hálózatos biológia, mely a hálózatos viselkedés egységes törvényei alapján egy új keretrendszert nyújt a sejten belüli, illetve sejtek közötti kapcsolatok vizsgálatára [2].

A proteomika térnyerésével hatalmas mennyiségű adathoz jutunk, melyek értelmezése azonban nehéz feladat elé állítja a kutatókat. A fehérje-fehérje kölcsönhatási hálózatok (interaktómok) kiválóan alkalmasak a fehérjék közti kapcsolatok elemzésére, új funkciók és interakciók jóslására [3]. Az így felépített hálózatok kapcsolatszáma a genom méreténél jobban korrelál a biológiai komplexitással [4].

A nagy áteresztőképességű (*high-throughput*) technikák számos hamis kapcsolatot eredményezhetnek, melyek szűréséhez szükség van a fehérje-fehérje interakciós hálózatok kapcsolatainak megbízhatóságát vizsgálni. Ennek egyik módja a fehérjék szubcelluláris lokalizációja alapján történő szűrés.

Az eukarióta sejtben zajló biokémiai folyamatok térbeli elkülönítése biztosítja, hogy azok egymástól időben függetlenül mehessenek végbe. Ez az elhatárolás segíti a sejtet a környezethez való alkalmazkodásban, és biztosítja a hatékony működést. A plazmamembrán által határolt sejtben belüli térben membránnal körülvett sejtszervecskék találhatók, melyek funkcionális egységeket alkotva garantálják a folyamatok elkülönülését, és segítik a differenciált jelátvitelt. A membránnal határolt organellek mellett azonban számos nem membránnal elválasztott sejtkompartment is kialakult, melyek dinamikusan változó működési egységeket alkotnak. Ezen sejt-alatti szerveződések figyelembevétele elengedhetetlen a fehérje-fehérje interakciós hálózatok megértéséhez [5][6].

III. Célkitűzések

Célkitűzésem volt egy olyan új, szubcelluláris lokalizációs információval kiegészített fehérje-fehérje interakciós adatbázis készítése, melynek segítségével a hamis kapcsolatok egy része kiszűrhető. Ezt az adatbázist és hozzá kapcsolódó web alkalmazást neveztem el ComPPI-nak (**Comp**artmentalized **Protein-Protein** Interaction Database). (<http://www.linkgroup.hu/comppi>)

Az elérhető lokalizációs, illetve fehérje-fehérje interakciós adatbázisok vizsgálata során kiderült, hogy a már meglévő adatok között csekély az átfedés, így szükségesnek tartottam a források összesítését az új adatbázis minőségi szerkezetének megalkotásához [7].

A kapcsolatokat a szubcelluláris lokalizáció, a hálózatos topológia és az adatok minősége alapján értékeljük, és egyúttal kialakítunk egy olyan megbízhatósági értéket, melynek segítségével kiszűrhetők a nem valószínű kapcsolatok, újak jósolhatóak, valamint még fel nem ismert, illetve hamis lokalizációkat is prediktálhatunk az interaktóm tulajdonságai alapján [8].

IV. Források

Organizmusok

Az interaktómok felépítéséhez olyan organizmusok kiválasztására volt szükség, melyekről nagy mennyiségű és jó minőségű adat érhető el. Emellett fontosnak tartottam az evolúciós aspektus vizsgálatának lehetőségét, melynek segítségével a későbbiekben mód nyílhat lokalizációs (ún. localog) [9] vagy interakciós (ún. interolog) [10] ortológok keresésére. A ComPPI adatbázis jelenleg négy fajra vonatkozóan tartalmaz adatokat: *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*. A későbbiekben szeretnénk kiegészíteni az adatbázist további, a bioinformatikai elemzésekben gyakran használt fajok adataival.

Lokalizációs forrás adatbázisok

A források keresése során olyan adatbázisokat kerestem, melyek viszonylag nagy mennyiségű, de megbízható adatokat tartalmaznak. Az irodalomból hiányzik a lokalizációs adatbázisok átfogó összehasonlító elemzése, pedig az adatok eltérő mennyisége és minősége ezt indokolná. A keresés és válogatás után az **1. táblázat**ban feltüntetett adatbázisokat használtuk fel.

Forrás adatbázis neve	Elérhető fajok	Fehérjék száma	Lokalizációk száma
BaCelLo [11]	<i>S. cerevisiae</i>	6 544	6 544
	<i>C. elegans</i>	23 626	25 714
	<i>H. sapiens</i>	29 795	33 551
eSLDB [12]	<i>S. cerevisiae</i>	6 561	8 462
	<i>C. elegans</i>	22 278	31 470
	<i>D. melanogaster</i>	20 815	20 815
	<i>H. sapiens</i>	39 805	81 806
Human Proteinpedia [13]	<i>H. sapiens</i>	2 133	2 900
LOCATE [14]	<i>H. sapiens</i>	13 450	18 723
MatrixDB [15]	<i>H. sapiens</i>	8 999	8 999
OrganelleDB [16]	<i>S. cerevisiae</i>	5 024	8 237
	<i>C. elegans</i>	493	551
	<i>D. melanogaster</i>	2 714	3 816
	<i>H. sapiens</i>	3 954	4 955
PA-GOSUB [17]	<i>S. cerevisiae</i>	3 178	3 412
	<i>C. elegans</i>	11 230	12 227
	<i>D. melanogaster</i>	10 511	11 354
	<i>H. sapiens</i>	23 507	25 741
GO [18]	<i>S. cerevisiae</i>	4 420	12 625
	<i>C. elegans</i>	23 113	29 215
	<i>D. melanogaster</i>	4 972	15 025
	<i>H. sapiens</i>	35 825	154 949
The Human Protein Atlas [19]	<i>H. sapiens</i>	7 815	10 264

1. táblázat: A szubcelluláris lokalizációs adatbázisok fehérje- és lokalizációs száma

A források prediktált és kísérletes adatokat is tartalmaznak. Az egyes fehérjékhez tartozó lokalizációk forrásainak számát, illetve az adatok minőségét (prediktált vagy kísérletes) felhasználtuk a megbízhatósági érték kialakításához.

A szubcelluláris lokalizációs adatok eltérő nevezéktannal, és felbontással rendelkeztek. Ennek megfelelően előfordult, hogy egy fehérje a sejtmag mellett a sejtmagi pórus komplexben is megtalálható, ami redundanciát okozott az adatokban. Ezt elkerülendő a lokalizációkat az AmiGO [20] szerver segítségével GO Cellular Component Term [18] alapján egy nem redundáns, hierarchikus lokalizációs fába rendezve standardizáltam (1512 GO azonosító), melynek segítségével a lokalizáció alapján különböző felbontású lekérések alakíthatóak ki. A legkisebb felbontásban hat nagy lokalizációt különítettem el: citoplazma, sejtmag, mitokondrium, szekréciós út, plazmamembrán, extracelluláris tér.

Fehérje-fehérje kölcsönhatási forrás adatbázisok

A fehérje-fehérje kölcsönhatási adatbázisok *high-throughput* és *low-throughput* technikák segítségével detektált fehérjék közötti interakciókat tartalmazzak. Adatbázisunkban nem használunk genetikai, csak fizikai kölcsönhatásokat. A források ebben az esetben is eltérő adatokat tartalmazzak [7], így szükségesnek tartottuk több adatbázis összesítését. Az adatok nagyobb része kísérletes módszerrel született, a jóslott kapcsolatok [21] hányada alacsony. A ComPPI a **2. táblázat**ban szereplő forrásokat tartalmazza.

Forrás adatbázis neve	Elérhető fajok	Fehérjék száma	Kapcsolatok száma
BioGRID [22]	<i>S. cerevisiae</i>	5 822	103 168
	<i>C. elegans</i>	2 882	4 877
	<i>D. melanogaster</i>	7 407	24 884
	<i>H. sapiens</i>	10 505	66 311
CCSB [23]	<i>S. cerevisiae</i>	2 018	2 930
	<i>C. elegans</i>	2 525	3 945
	<i>H. sapiens</i>	2 187	3 881
DiP [25]	<i>S. cerevisiae</i>	5 034	22 367
	<i>C. elegans</i>	2 595	3 945
	<i>D. melanogaster</i>	7 105	22 080
	<i>H. sapiens</i>	2 875	3 872
DroID [26]	<i>D. melanogaster</i>	16 267	89 541
HomoMINT [27]	<i>H. sapiens</i>	7 876	33 501
HPRD [28]	<i>H. sapiens</i>	9 644	39 240
IntAct [29]	<i>S. cerevisiae</i>	5 930	97 786
	<i>C. elegans</i>	4 352	11 352
	<i>D. melanogaster</i>	8 469	27 834
	<i>H. sapiens</i>	10 552	58 051
MatrixDB [15]	<i>H. sapiens</i>	121	281
MINT [30]	<i>S. cerevisiae</i>	5 467	48 529
	<i>C. elegans</i>	3 316	7 394
	<i>D. melanogaster</i>	7 428	23 515
	<i>H. sapiens</i>	9 662	31 008
MIPS [31]	<i>H. sapiens</i>	121	281
PIP's [21]	<i>H. sapiens</i>	4 255	25 598

2. táblázat: A fehérje-fehérje kölcsönhatási adatbázisok fehérje- és interakciós száma

V. *Módszerek*

Az adatbázisok összesítése

A forrás adatbázisok eltérő fehérje nevezéktant alkalmaztak, így szükség volt ezek egységesítésére. Ennek érdekében több forrásból (UniProt [33], HPRD [28], BioGRID [22]) származó fehérje azonosító táblát építettünk fel, melynek segítségével a legmegbízhatóbb, manuálisan ellenőrzött UniProtKB/Swiss-Prot nevezéktanra [34] térképeztük a fehérjéket. Amennyiben a UniProtKB/Swiss-Prot nem volt elérhető, abban az esetben az UniProtKB/TrEmbl nevekre keresünk, majd a további elérhető szinonimákra. Az egyes fehérjék teljes, eredeti neve (pl. MHC) is elérhető.

A beépített nevezéktan-térképek mellett fennmaradó neveket manuálisan fordítottam le a Synergizer (<http://llama.mshri.on.ca/synergizer/translate/>) illetve a Protein Identifier Cross-Reference Service (PICR) (<http://www.ebi.ac.uk/Tools/picr/>) fordító szerverek segítségével.

Az egységesített fehérje nevezéktan-térkép segítségével a felhasználó által keresett fehérje kapcsolatait és lokalizációját az elérhető legmegbízhatóbb azonosító alapján tudjuk visszaadni.

Az adatbázis szerkezete

Az adatbázis felépítése hierarchikus (**3. ábra**). A forrás adatbázisok egy univerzális bemeneten keresztül, Symfony® (<http://www.symfony-project.org/>) keretrendszerben megírt illesztőfelületek segítségével csatlakoznak a MySQL szerveren futó adatbázishoz, mely saját nevezéktant használ. Ennek segítségével könnyen kapcsolhatóak be új források, illetve frissíthetők a már betöltött adatok. A PHP 5-alapú, szintén Symfony® keretrendszerben implementált webes felhasználói felület és az adatréteg között a kapcsolat kétirányú. A kért adatok testre szabhatóan megjeleníthetők, vagy egy export modul segítségével többféle formátumban és különféle szűréssel letölthetők. A ComPPI nyílt forráskódú, így a hozzáértő felhasználók számára a működése átlátható, a fejlesztés folyamata követhető (<http://bificom2.sote.hu/redmine/projects/compqi>).



3. ábra: Az adatbázis vázlatos felépítése

A ComPPI megbízhatósági érték

A fehérje-fehérje kölcsönhatásokat számos paraméter alapján lehet rangsorolni, melyek segítségével megbízhatósági értéket (ún. *confidence score*) rendelhetünk a kapcsolatokhoz. Ez az érték azt írja le, mekkora a valószínűsége annak, hogy az adott kapcsolat ténylegesen létrejön-e az élő sejtben.

A megbízhatósági érték számolására számos lehetőség létezik. Az egyik elterjedt értékelési rendszer a PRINCESS [35], mely hat paraméter alapján rendel megbízhatósági értéket a humán fehérje-fehérjeinterakciós adatokhoz. E hat paraméter közül kettő a ComPPI megbízhatósági értékhez hasonló: a GO *Biological Process Term* [18] alapján osztályozza az azonos funkciójú fehérjéket, illetve funkcionális csoportok keresésével a hálózatos topológiát is figyelembe veszi.

A ComPPI megbízhatósági érték a fehérjék szubcelluláris lokalizációja (GO *Cellular Component Term* [18]), a hálózatos topológia és az adatok kísérletes vagy prediktált minősége alapján rangsorolja a kapcsolatokat. (4. ábra)

$$p(\text{comp}pi) = \left\{ \bigvee_{Loc} \left[\left(\varphi_{v1} \vee \sum_{n(v1)} \frac{\varphi_n}{\deg(v1)} \right) \cdot \left(\varphi_{v2} \vee \sum_{n(v2)} \frac{\varphi_n}{\deg(v2)} \right) \right] \right\} \cdot \varphi_i$$

$$a \vee b = 1 - (1 - a)(1 - b)$$

4. **ábra: A ComPPI megbízhatósági érték számolásához használt képlet.** A $v1$ és $v2$ az interakció két oldalán álló fehérjéket jelöli. Az $n(v1)$ a $v1$ szomszédjainak számát adja meg. A φ_{v1} , φ_{v2} és φ_n a fehérje adott lokalizációbeli valószínűségét határozza meg, míg a φ_i az interakcióhoz rendel valószínűséget az alapján, hogy a forrás adatok között mennyi kísérletes és prediktált adat szerepel. A $\deg(v1)$ és $\deg(v2)$ az interaktáló fehérjék fokszámát, azaz kapcsolatainak számát jelöli.

A képlet alapján minden nagy lokalizációra (citoplazma, sejtmag, mitokondrium, szekréciós út, plazmamembrán, extracelluláris tér) kiszámoljuk, hogy az adott fehérje és első szomszédjainak lokalizációja alapján mennyire tartozik bele az adott kompartmentbe. Ennek feltétele, hogy $v1$ és $v2$ legalább egy-egy szomszédja tartalmazza az adott lokalizációt. Amennyiben egyik sem tartalmazza, úgy arra a nagy lokalizációra a valószínűség 0.

A következő lépésben a kölcsönhatás két végén lévő fehérjékre kapott kompartmentbeli valószínűségeket összeszorozzuk, ami annál nagyobb érték lesz, minél inkább közös lokalizációban helyezkednek el.

A végső lépésben a nagy lokalizációkra kapott megbízhatósági értéket egy VAGY művelettel összevonjuk, és az interakció valószínűségével szorozva megkapjuk a kapcsolat megbízhatósági értékét. Ez az érték akkor a legmagasabb, ha a kölcsönható fehérjék stabilan egy, vagy több közös lokalizációval rendelkeznek, és a forrás adatok kísérletes eredményből származnak.

A potenciálisan új lokalizációk jóslására a fehérje szomszédjainak elhelyezkedése alapján nyílik lehetőség. Így ha egy fehérje első szomszédjai az átlagosnál szorosabban tartoznak bele egy adott lokalizációba, de maga a fehérje nem rendelkezik azzal az adott lokalizációval, akkor egy állítható erősség alapján jósolhatunk számára kompartmentbe tartozási valószínűséget. Ehhez első lépésben minden nagy lokalizációra kiszámoljuk az egyes fehérjék

átlagos beletartozási értékét, majd a predikció erősségének beállításával az átlag feletti szomszédság növeli az adott fehérje lokalizációs valószínűségét, míg az átlag alatti csökkenti azt. Abban az esetben, ha ezt a funkciót nem alkalmazzuk, akkor az erősséget 0-ra állítjuk.

A ComPPI megbízhatósági érték számítása során beállítható paraméterek:

- *kísérletes / prediktált / ismeretlen eredetű* kölcsönhatási adathoz rendelt valószínűség (ϕ_i)
- *kísérletes / prediktált / ismeretlen eredetű* lokalizációs adathoz rendelt valószínűség (ϕ_{v1} , ϕ_{v2} és ϕ_n)
- szomszédság alapú lokalizációs predikció erőssége

VI. *Eredmények és diszkusszió*

Az elkészült kompartmentalizált fehérje-fehérje kölcsönhatási adatbázis a következő fő tulajdonságokkal rendelkezik:

- A. több forrásból származó fehérje-fehérje interakciós és lokalizációs adatokat összegez;
- B. a ComPPI megbízhatósági érték segítségével alkalmas azon kapcsolatok szűrésére, melyek eltérő kompartmentek között jönnének létre;
- C. az adatbázist kiszolgáló webes felületen egyszerűen kiválaszthatóak a kívánt tulajdonságok, így a letöltendő adatbázisok struktúrája a későbbi vizsgálatokhoz optimalizálható;
- D. az adatbázis letöltés mellett lehetőség van kompartment-specifikus részhálózatok készítésére, illetve az egyes fehérjék kapcsolatainak böngészésére is.

A fenti tulajdonságok alkalmassá teszik az adatbázist kísérletes eredmények vizsgálatára a biokémiában, molekuláris biológiában, illetve bioinformatikai és hálózatbiológiai elemzésekre proteóm- és interaktóm-szinten.

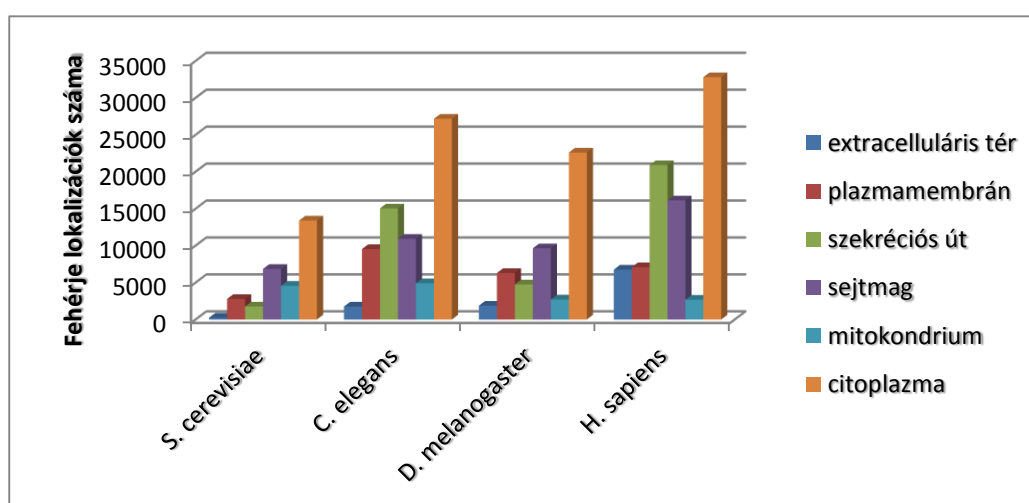
Összesített adatok

Az adatok összesítése után a ComPPI a **5. táblázatban** feltüntetett adatkészlettel rendelkezik.

Faj neve	Fehérjék száma	Kapcsolatok száma	Lokalizációk száma
<i>S. cerevisiae</i>	6 790	274 958	39 280
<i>C. elegans</i>	20 477	31 436	99 177
<i>D. melanogaster</i>	19 742	187 937	51 010
<i>H. sapiens</i>	60 250	234 055	341 888

5. táblázat: A fehérje-fehérje interakciók és az összes lokalizáció száma fajonkénti bontásban. A fehérjék száma UniProt nevezéktan alapján, míg a lokalizációk száma GO azonosító [18] alapján került feltüntetésre, de a fa struktúra miatt egy fehérje több GO azonosító [18] alá is tartozhat.

A lokalizációs fa 1512 GO azonosító [18] alapján készült el, melyből a ComPPI adatkészlet 1160-t tartalmaz. A fennmaradó 352 a fa olyan ágainak neveit tartalmazza, melyek a létező lokalizációk beágyazáshoz szükségesek. A hat nagy kompartment adatait az **6. ábra** mutatja.



6. ábra: A fehérjék eloszlása a hat nagy lokalizáció szerint fajonként. Egy fehérje több nagy lokalizációba is tartozhat, így az oszlopok összege több, mint az összes fehérje száma.

A kiszolgáló felület – fehérje keresés

A *Protein Search* oldalon lehetőség van fehérjék kompartmentalizált fehérje-fehérje interakciós adatainak keresésére és megjelenítésére. A kereső mező használatával egy fehérje kapcsolatait kaphatjuk meg, míg az *Open Advanced Search* gombra kattintva fajonként, vagy egyszerre több fehérjére kereshetünk. A **7. ábra** az NF- κ B p100 alegységének kapcsolatait tartalmazza.

ComPPI Compartmentalized Protein-Protein Interaction Database		Home Protein Search Download DB Help	
Protein Name:	Q00653	Submit Search	Open Advanced Search
3 proteins with 0 synonyms and 186 interactions were found.			
Protein A	Localization A	Protein B	Localization B
P41279	cytoplasm src	Q00653	nucleus src
Q00653	nucleus src	Q969H0	nucleus src
Q00653	nucleus src	Q04864	nucleus src
Q00653	nucleus src	Q04206	nucleus src
Q00653	nucleus src	O75604	nucleus src
Q00653	nucleus src	O75604	plasma membrane src
Q00653	nucleus src	O75604	cytoplasm src
Q00653	nucleus src	O00571	nucleus src
Q00653	nucleus src	Q00653	nucleus src
Q00653	nucleus src	Q01201	nucleus src
Q00653	nucleus src	P62888	cytoplasm src
Q01201	nucleus src	Q00653	nucleus src
Q00653	nucleus src	P20749	nucleus src
Q00653	nucleus src	Q9Y6K9	nucleus src
Q00653	nucleus src	Q9Y6K9	cytoplasm src
Q00653	nucleus src	P41279	cytoplasm src
Q00653	nucleus src	O00255	nucleus src
P08047	nucleus src	Q00653	nucleus src
P04150	nucleus src	Q00653	nucleus src
Q99558	cytoplasm src	Q00653	nucleus src
Q99558	nucleus src	Q00653	nucleus src
Q99558	cytoplasm src	Q00653	nucleus src
Q99558	vesicle src	Q00653	nucleus src
Q00653	nucleus src	Q9Y297	cytoplasm src
O15226	nucleus src	Q00653	nucleus src
O15226	nucleus src	Q00653	nucleus src
Q04206	nucleus src	Q00653	nucleus src
Q00653	nucleus src	O15111	cytoplasm src
Q00653	nucleus src	Q9BYH8	nucleus src
Q00653	nucleus src	O00231	cytoplasm src

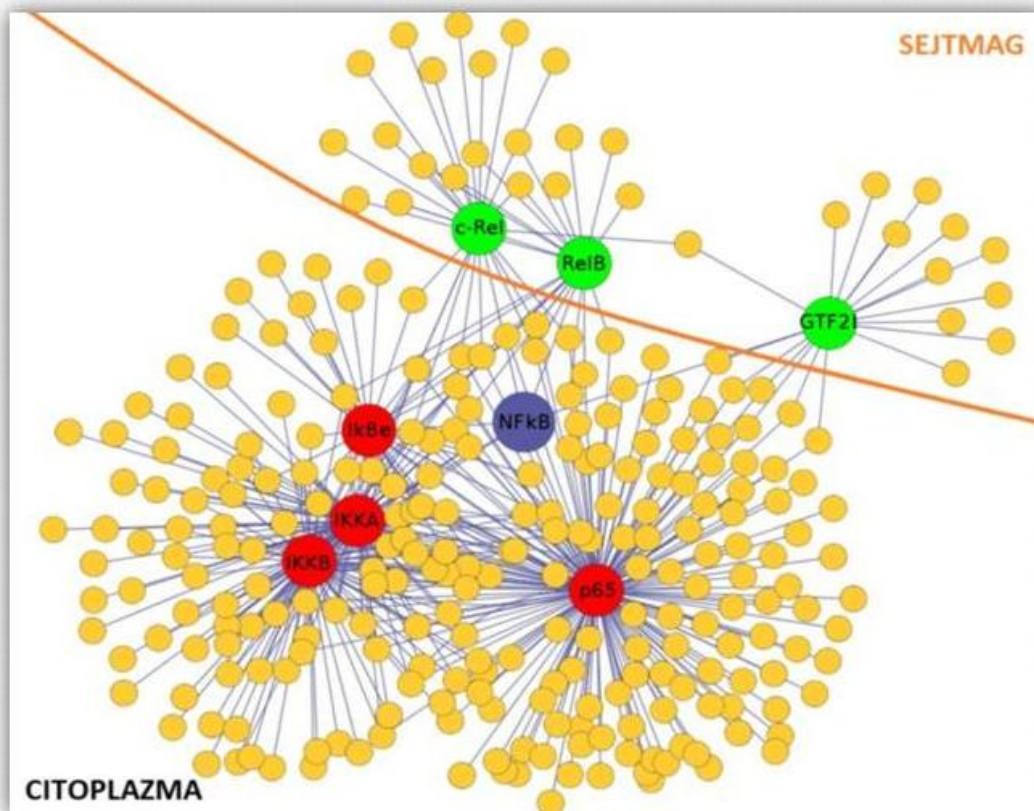
7. ábra: Az NF- κ B p100 alegység (Q00653) kapcsolatainak egy része lokalizációval kiegészítve

Fehérje keresés – alkalmazási példa

Sok fehérje kompartmentek közti transzlokációja alapvető szerepet játszik betegségek kialakulásában, illetve terápiájában egyaránt. Az ilyen tulajdonságú fehérjék vizsgálatánál kulcsfontosságú a szubcelluláris lokalizáció figyelembevétele. Példaként mutatnám be a nukleáris faktor NF- κ B-t (**7. ábra**), melynek citoplazmából a sejtmagba történő transzlokációjának gátlása antiapoptotikus hatást fejt ki a dopaminerg neuronokra Parkinson kórban [36].

A **8. ábra** az NF- κ B p100 alegységének kapcsolatait, illetve lokalizációs viszonyait ábrázolja. A zölddel jelölt fehérjék az NF- κ B által mediált transzkripciók faktorok. Ezek a

sejtmagi részhálózat elemei, míg a pirossal jelölt NF-kB inhibitor complex tagjai a citoplazmában találhatók. Ezen fehérjék kapcsolatai a ComPPI segítségével kompartment szinten vizsgálhatók, ezzel új betekintést engedve a patomechanizmusban betöltött szerepükbe.



8. **ábra:** az NF-kB p100 alegységének kapcsolatait, illetve lokalizációs viszonyait ábrázoló részhálózat (NF-kB: nuclear factor kappa-light-chain-enhancer of activated B cells, C-Rel: proto-oncogene c-Rel, RelB: transcription factor RelB, GTF2I: general transcription factor II-I, IkBe: NF-kappa-B inhibitor epsilon, IKKA: inhibitor of NF-kappa-B subunit A, IKKB: inhibitor of NF-kappa-B subunit B, p65: transcription factor p65)

A kiszolgáló felület – adatbázis letöltés

A *Download DB* oldalon (9. ábra) a faj kiválasztása után vagy egy előre definiált adatbázis, vagy a lokalizációk megjelölésének segítségével tetszőlegesen kialakított sejtsejtszervecske specifikus részhálózatok letöltésére van lehetőség. Az adatbázis különböző formátumokban exportálható, így a leggyakrabban használt bioinformatikai kiterjesztésekben is (CSV, PSI-MI, BioPAX és SBML). A hálózatok a Cytoscape [37] vizualizációs program által támogatott CYS formátumban is elérhetőek, így azok letöltés után azonnal megjeleníthetőek.

A letöltött adatbázis tartalmazza a kölcsönható fehérjék neveit, a lokalizációs viszonyokat, a kapcsolat megbízhatósági értékét, és amennyiben elérhető, úgy az interakciós és lokalizációs adatok forrását, hivatkozását, illetve az adat minőségét (kísérletes vagy prediktált).

9. ábra: Az adatbázis letöltő felület képe

Adatbázis letöltés – alkalmazási példa

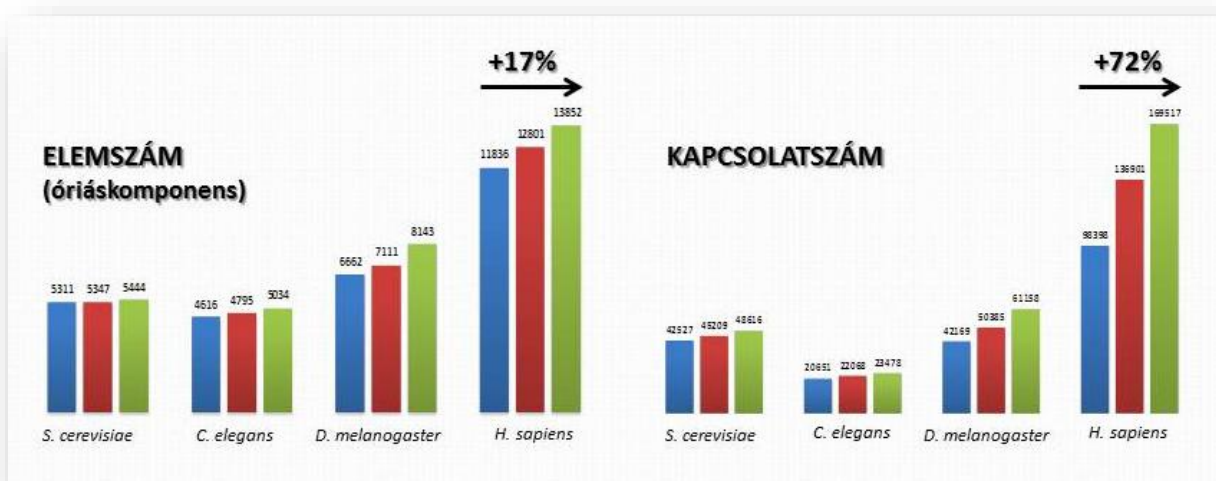
Az öregedés komplex folyamatának megértéséhez szükség van a bioinformatikai módszerek alkalmazására. Az általam megalkotott öregedési modellben azt a megfigyelést vettük alapul, miszerint a sejtmagmembrán permeabilitása az öregedés során nő [38].

Modellemben a kompartmentalizáltság teljessége alapján három teoretikus öregedési fázist különítettem el.

- A „**fiatal**” modellben azon fehérjék között alakul ki kapcsolat, melyek legalább egy esetben azonos lokalizációval rendelkeznek. Ez áll legközelebb biológiai relevanciájában az élő sejtben megfigyelhető állapotokhoz.
- Az „**öregedő**” modellben a sejtmagi és citoplazmai részhálózatokat egybeolvasztottam. Ez szemlélteti a sejtmagmembrán átjárhatósága esetén kialakuló képet.

- Az „**öreg**” fázisban az összes részhálózatot egyesítettem. Ez a forma áll legtávolabb a valóságtól, ennek ellenére az eddigi elemzések ilyen hálózatokon zajlottak.

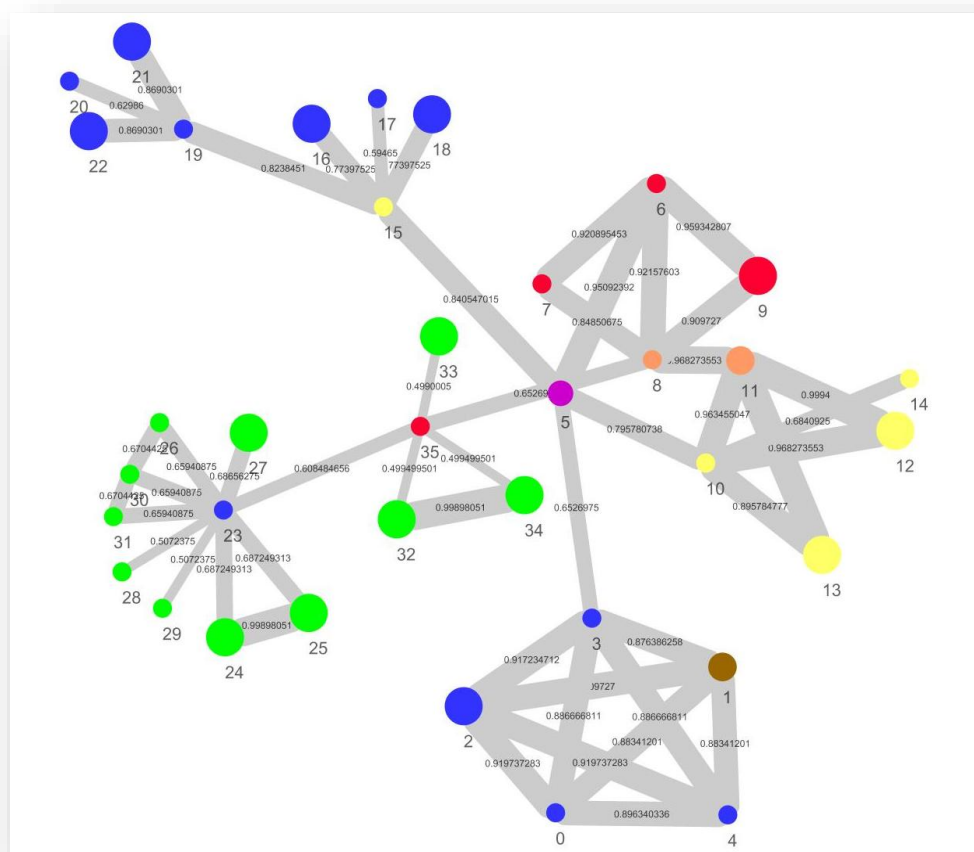
Az öregedés során megfigyelhető volt, hogy a modell tulajdonságaiból adódóan nőtt a kapcsolatok száma (**10. ábra**). Ezen kapcsolatok egy része azonban az élő sejtben nem létezhetne, hiszen nem rendelkeznek azonos lokalizációval. Az eredmények felhívják a figyelmet az öreg és fiatal hálók közti jelentős különbségekre, és az embernél megfigyelhető 72%-os kapcsolatszám növekedés megalapozza a ComPPI szükségességét.



10. ábra: Az öregedési modell elem- és kapcsolatszámának változása „fiatal” - „öregedő” - „öreg” sorrendben. (Az öregedési modell vizsgálata előzetes adatokkal történt a ComPPI létjogosultságának bizonyítása érdekében. A végleges adatbázison alapuló elemzések folyamatban vannak.)

A ComPPI megbízhatósági érték alkalmazása

A megbízhatósági értéket egy teszt hálózaton próbáltam ki (**11. ábra**), melybe beépítettem olyan eseteket is, melyek nem feltétlenül kapnának magas értéket, de a megbízhatósági érték kialakításának ismeretében mégis azt várnánk. Ilyen a 23-as és 35-ös fehérje, melyek egymástól és szomszédjaiktól is eltérő helyen találhatóak, azonban a szomszédok egy kivételével azonos lokalizációval rendelkeznek. A kapcsolatra kapott megbízhatósági érték így 0,61, ami magasabb, mint például a 35-ös és 33-as pont között (0,5), ahol a 33-as pontnak nincsenek azonos lokalizációval rendelkező szomszédjai.



11. ábra: A megbízhatósági érték vizsgálatára használt teszt hálózat (35 pont, 50 kapcsolat). A hálózat elrendezésekor a megbízhatósági érték az élek súlyát adta. Az élek vastagsága a kapcsolat valószínűségével arányos, míg a pontok mérete kísérletes adatok használata esetén a legnagyobb.

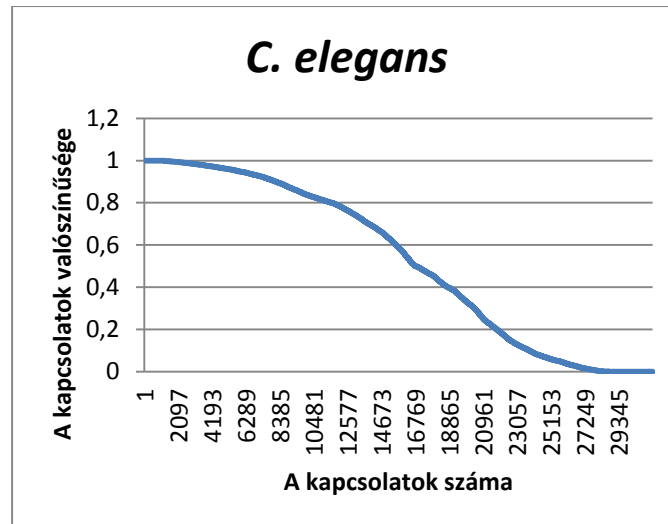
Színkódolás: *sárga*: citoplazma; *narancs*: citoplazma és sejtmag; *vörös*: sejtmag; *zöld*: szekréciós út; *kék*: mitokondrium; *lila*: citoplazma, mitokondrium és sejtmag; *barna*: mitokondrium és extracelluláris tér.

Az 1-es pont két lokalizációval rendelkezik, de az egyik hamis információ. A megbízhatósági érték azonban nem csökken jelentősen (0,89 a 0,92-vel szemben), így az előforduló hamis lokalizációk nem rontják érdemben az interakciók megbízhatóságát.

A fenti két példa is mutatja, hogy a ComPPI megbízhatósági érték segítségével lehetőség nyílik eddig még fel nem ismert lokalizációk jóslására, illetve hamis lokalizációk szűrésére az interakciós adatokra támaszkodva. Hasonlóképpen tervezzük lokalizáció alapján kapcsolatok prediktálását is.

A ComPPI alapvető célja azonban a fehérje-fehérje interakciók szűrése a fehérjék lokalizációja alapján. Ehhez a ComPPI megbízhatósági érték állítható paramétereinek finom

hangolása szükséges. Az előzetes eredményeket a *C. elegans* példáján a **12. ábra** tartalmazza. Ez alapján a kutatás következő fázisában statisztikai módszerekkel megalapozott megbízhatósági szinteket állíthatunk fel, így lehetőség nyílt magas, közepes és alacsony valószínűségű kapcsolatok differenciálására.



12. ábra: A kapcsolatok megbízhatósági értékének eloszlása a *C. elegans* példáján

A fejlesztés jelen állapotában a következő beállításokat használtuk:

- kölcsönhatási adathoz rendelt valószínűség (ϕ_i)
 - *kísérletes* = 1.0
 - *prediktált* = 0.8
 - *ismeretlen eredetű* = 0.9
- lokalizációs adathoz rendelt valószínűség (ϕ_{v1} , ϕ_{v2} és ϕ_n)
 - *kísérletes* = 1.0
 - *prediktált* = 0.6
 - *ismeretlen eredetű* = 0.8
- szomszédság alapú lokalizációs predikció erőssége

VII. Konklúzió

Az ép mellett a beteg sejt szerveződésének és funkciójának megértésben is kiemelkedő szerepet kap a sejt-szintű hálózatok vizsgálata, melynek segítségével jobban megérthetünk olyan komplex patomechanizmussal rendelkező állapotokat, mint az öregedés [39], vagy az inzulin rezisztencia. A hálózatbiológiai gondolkodás lehetőséget teremt a mainál kevesebb mellékhatással rendelkező, de hatásosabb gyógyszerek kifejlesztésére [40]. A ComPPI segítségével olyan minőségi hálózatokat építhetünk, melyek szerkezeti és dinamikai elemzésével a potenciális gyógyszer célpontok száma emelkedhet, és így új módszerekhez juthatunk például a daganatos megbetegedések terápiájában [41].

(A ComPPI még fejlesztés alatt áll. A tesztelés és javítás után publikáljuk a szakirodalomban.)

VIII. Szerzői hozzájárulás

A ComPPI saját ötlet eredményeként jött létre, melynek megvalósulását egy interdiszciplináris projekt keretében koordinálom.

A forrás adatbázisok kijelölését, az illesztőfelületek kialakítását, a manuális kereséseket (lokalizációs fa, fehérje térképezés), a megbízhatósági érték kidolgozását, az adatbázis szerkezetének kialakítását és a webes felület tervezését végeztem.

A ComPPI tesztelését, az eredmények értékelését, a további felhasználás kereteinek kijelölését a biológiai és orvosi tudás és haszon függvényében hajtottam végre.

IX. Köszönetnyilvánítás

A ComPPI egy multidiszciplináris együttműködés eredménye. Munkatársaim:

- Dr. Gyurkó M. Dávid
- Szalay Kristóf

- Thaler Benedek

Köszönettel tartozom témavezetőimnek, Prof. Csermely Péternek és Korcsmáros Tamásnak, illetve az általuk vezetett LINK-group (<http://www.linkgroup.hu/>) és NetBiol (<http://netbiol.elte.hu>) kutatócsoportok tagjainak.

X. Irodalom

- [1] Veres Dániel, Csermely Péter: Hálózatok az élővilágban (<http://www.dgci.sote.hu/file.download.php?id=4529>)
- [2] Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5**, 101–13 (2004).
- [3] Koh, G. C. K. W. *et al.* Analyzing protein-protein interaction networks. *Journal of Proteome Research* **11**, 2014–31 (2012).
- [4] Kelly, W. & Stumpf, M. Protein-protein interactions: from global to local analyses. *Current Opinion in Biotechnology* **19**, 396–403 (2008).
- [5] Gandhi, T. K. B. *et al.* Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics* **38**, 285–293 (2006).
- [6] Kumar, G. & Ranganathan, S. Network analysis of human protein location. *BMC Bioinformatics* **11 Suppl 7**, S9 (2010).
- [7] Mathivanan, S. *et al.* An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics* **7 Suppl 5**, S19 (2006).
- [8] Shin, C. J. *et al.* a Protein-protein interaction as a predictor of subcellular location. *BMC Systems Biology* **3**, 28 (2009).
- [9] Chi, S.-M. Prediction of protein subcellular localization by weighted gene ontology terms. *Biochemical and Biophysical Research Communications* **399**, 402–5 (2010).
- [10] Huang, T.-W. *et al.* Reconstruction of human protein interolog network using evolutionary conserved network. *BMC Bioinformatics* **8**, 152 (2007).
- [11] Pierleoni, A. *et al.* BaCellLo: a balanced subcellular localization predictor. *Bioinformatics (Oxford, England)* **22**, e408–16 (2006). (<http://gpcr.biocomp.unibo.it/bacello/>)
- [12] Pierleoni, A. *et al.* eSLDB: eukaryotic subcellular localization database. *Nucleic Acids Research* **35**, D208–12 (2007). (<http://gpcr.biocomp.unibo.it/esldb/>)
- [13] Kandasamy, K. *et al.* Human Proteinpedia: a unified discovery resource for proteomics research. *Nucleic Acids Research* **37**, D773–81 (2009). (<http://www.humanproteinpedia.org/>)
- [14] Sprenger, J. *et al.* LOCATE: a mammalian protein subcellular localization database. *Nucleic Acids Research* **36**, D230–3 (2008). (<http://locate.imb.uq.edu.au/>)

- [15] Chautard, E. *et al.* MatrixDB, a database focused on extracellular protein-protein and protein-carbohydrate interactions. *Bioinformatics (Oxford, England)* **25**, 690–1 (2009). (<http://matrixdb.ibcp.fr/>)
- [16] Wiwatwattana, N. *et al.* Organelle DB: an updated resource of eukaryotic protein localization and function. *Nucleic Acids Research* **35**, D810–4 (2007). (<http://organelledb.lsi.umich.edu/>)
- [17] Lu, P. *et al.* PA-GOSUB: a searchable database of model organism protein sequences with their predicted Gene Ontology molecular function and subcellular localization. *Nucleic Acids Research* **33**, D147–53 (2005). (<http://webdocs.cs.ualberta.ca/~bioinfo/PA/GOSUB/>)
- [18] Gene, T. *et al.* Gene Ontology : tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2011). (<http://www.geneontology.org/>)
- [19] Uhlen, M. *et al.* Towards a knowledge-based Human Protein Atlas. *Nature Biotechnology* **28**, 1248–50 (2010). (<http://www.proteinatlas.org/>)
- [20] Carbon, S. *et al.* AmiGO: online access to ontology and annotation data. *Bioinformatics (Oxford, England)* **25**, 288–9 (2009).
- [21] McDowall, M. D. *et al.* PIPs: human protein-protein interaction prediction database. *Nucleic Acids Research* **37**, D651–6 (2009). (<http://www.compbio.dundee.ac.uk/www-pips/>)
- [22] Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Research* **34**, D535–9 (2006). (<http://thebiogrid.org/>)
- [23] Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–8 (2005). (<http://ccsb.dfci.harvard.edu/web/www/ccsb/>)
- [24] Kamburov, A. *et al.* ConsensusPathDB--a database for integrating human functional interaction networks. *Nucleic Acids Research* **37**, D623–8 (2009). (<http://cpdb.molgen.mpg.de/>)
- [25] Xenarios, I. *et al.* DIP: the database of interacting proteins. *Nucleic Acids Research* **28**, 289–91 (2000). (<http://dip.doe-mbi.ucla.edu/dip/>)
- [26] Yu, J. *et al.* Droid: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. *BMC Genomics* **9**, 461 (2008). (<http://www.droidb.org/>)
- [27] Persico, M. *et al.* HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics* **6 Suppl 4**, S21 (2005). (<http://mint.bio.uniroma2.it/HomoMINT/Welcome.do>)
- [28] Keshava Prasad, T. S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Research* **37**, D767–72 (2009). (<http://www.hprd.org/>)
- [29] Kerrien, S. *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Research* **40**, D841–6 (2012). (<http://www.ebi.ac.uk/intact/>)
- [30] Licata, L. *et al.* MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research* **40**, D857–61 (2012). (<http://mint.bio.uniroma2.it/mint/Welcome.do>)
- [31] Pagel, P. *et al.* The MIPS mammalian protein-protein interaction database. *Bioinformatics (Oxford, England)* **21**, 832–4 (2005). (<http://mips.helmholtz-muenchen.de/proj/ppi/>)

- [32]Apweiler, R. *et al.* UniProt : the Universal Protein knowledgebase. *Nucleic Acids Research* **32**, D115–D119 (2004). (<http://www.uniprot.org/>)
- [33]Huang, H. *et al.* A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics (Oxford, England)* **27**, 1190–1 (2011).
- [34]Boutet, E. *et al.* UniprotKB/Swiss-Prot: The manually annotated section of the Uniprot Knowledgebase. *Methods in Molecular Biology* **406**, 89–112 (2007).
- [35]Li, D. *et al.* PRINCESS, a protein interaction confidence evaluation system with multiple data sources. *Molecular & Cellular Proteomics* **7**, 1043–52 (2008).
- [36]Hunot, S. *et al.* Nuclear translocation of NF-kappaB is increased in dopaminergic neurons of patients with parkinson disease. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 7531–6 (1997).
- [37]Smoot, M. E. *et al.* Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)* **27(3)**, 431–2 (2011).
- [38]D’Angelo, M. A. *et al.* Age-dependent deterioration of nuclear pore complexes causes a loss of nuclear integrity in postmitotic cells. *Cell* **136**, 284–95 (2009).
- [39]Simkó, G. I. *et al.* Network strategies to understand the aging process and help age-related drug design. *Genome Medicine* **1**, 90 (2009).
- [40]Nussinov, R. *et al.* Allo-network drugs: harnessing allostery in cellular networks. *Trends in Pharmacological Sciences* **32**, 686–93 (2011).
- [41]Csermely, P. *et al.* Structure and dynamics of molecular networks : A novel paradigm of drug discovery. *Pharmacology & Therapeutics*, in press