PH.D. ÉRTEKEZÉS

Sipos Botond

SZEGEDI TUDOMÁNYEGYETEM
BIOLÓGIA DOKTORI ISKOLA
TERMÉSZETTUDOMÁNYI
ÉS INFORMATIKAI KAR

MTA SZEGEDI BIOLÓGIAI KÖZPONT
GENETIKAI INTÉZET

2010
SZEGED

SZEGEDI TUDOMÁNYEGYETEM
BIOLÓGIA DOKTORI ISKOLA
TERMÉSZETTUDOMÁNYI
ÉS INFORMATIKAI KAR

MTA SZEGEDI BIOLÓGIAI KÖZPONT
GENETIKAI INTÉZET

A NIMROD SZUPERGÉNCSALÁD EVOLÚCIÓJA

Ph.D. értekezés

Sipos Botond

TÉMAVEZETŐ:
Dr. Pénzes Zsolt,
egyetemi docens,
SZTE TTIK, Ökológiai Tanszék,
tudományos munkatárs,
MTA SZBK, Genetikai Intézet

KONZULENS:
Dr. Somogyi Kálmán,
tudományos munkatárs,
MTA SZBK, Genetikai Intézet

2010
SZEGED

Tartalomjegyzék

1. Bevezetés **1**

1.1. Génduplikációk, géncsaládok és homológia 5
1.2. Filogenetikai fák és hálózatok 8
1.3. A filogenetikai rekonstrukciós módszerek áttekintése 10
1.4. A géncsaládok evolúciójának modelljei 13
1.4.1. Az „összehangolt” és „születés és halál” modellek 14
1.4.2. Az ismétlődő szekvenciaelemek összehangolt evolúciója 17
1.4.3. Az összehangolt evolúció vizsgálata 17
1.4.4. Az összehangolt evolúció jelentősége 19

2. Célkitűzések **21**

3. Anyag és módszer **23**

3.1. A szekvenciák összegyűjtése 24
3.2. A NIM domének azonosítása profil Rejtett Markov Modellek segítségével . . . 25
3.3. Profil–profil összehasonlítások 26
3.4. A szinonim és nemszinonim távolságok becslése 26
3.5. A szaturációs ábrák 27
3.6. Filogenetikai rekonstrukció 28
3.6.1. Az aminosav-szekvenciák többszörös illesztése 29
3.6.2. Modellszelekció 32
3.6.3. A filogenetikai szignál tesztelése: likelihood térképezés 32
3.6.4. Maximum Likelihood filogenetikai rekonstrukció 33
3.6.5. Filogenetikai rekonstrukció Bayes statisztikai módszerekkel 33
3.6.6. Neighbor-Joining filogenetikai rekonstrukció 35
3.6.7. A génfák gyökereztetése 35
3.7. A klaszterizáció statisztikai szignifikanciájának vizsgálata 36

4. Eredmények **38**

4.1. A NIM és az EGF domének közötti hasonlóság 39

4.2. A ismétlődő NIM domének evolúciója 41
4.2.1. Egy vizuális módszer a domén-evolúciós trendek vizsgálatára 41
4.2.2. A NIM domének evolúciójának vizsgálata PRHD-k segítségével 48
4.3. A filogenetikai analízisek eredményei 59
4.4. A Nimrod klaszter konzerváltságának statisztikai szignifikanciája 67

5. Az eredmények tárgyalása **69**
5.1. A Nimrod génklaszter integritását fenntartó tényezők 70
5.2. Mintázatok az ismétlődő NIM domének evolúciójában 71
5.3. A Nimrod szupercsalád evolúciós története 73

Köszönetnyilvánítás **77**

Irodalomjegyzék **88**

Összefoglaló **89**

Summary **92**

Függelék **96**
1. A vizsgált Nimrod szekvenciák adatai 96
2. A NIM domének és a CCxGY/W motívumok többszörös illesztésekben való vizualizációját szemléltető példák. 97
3. A Bayes MCMC filogenetikai rekonstrukció során használt MrBayes nexus blokkok 100
4. NimB, NimC és Eater szekvenciákban azonosított domének NJ algoritmussal számolt filogenetikája. 103
5. Különböző Drosophila Nim ortológok szekvenciáinak párbaállításával készült PRHD-k. 109
6. A *D. pseudoobsucra* és *D. yakuba* Eater szekvenciákban azonosított domének közötti szinonim és nemszinonim távolságok ábrázolása 113
7. A NimA szekvenciák NJ és Bayes MCMC módszerekkel rekonstruált filogenetikái 114
8. A NimB és NimC szekvenciák NJ algoritmussal rekonstruált filogenetikái 115
9. A NimB és NimC szekvenciák filogenetikájának Bayes MCMC rekonstrukciója nyomán kapott poszterior mintákból számolt konszenzushálózatok 118
10. A *D. virilis* NimB szekvenciák párbaállításával készült PRHD-k. 120
11. A Nimrod klaszter konzerváltságának vizsgálata során felhasznált adatok a források megjelölésével 121

Rövidítések jegyzéke

d	két szekvencia közötti divergencia
I	színintenzitás
T	egy filogenetikai fa ághosszainak összege
x	a konszenzushálózatok felépítése során a splitgyakorságra vonatkozó küszöbérték
AIC	Akaike Information Criterion
BD	a géncsaládevolúció „születés és halál” („Birth-and-Death”) modellje
BIC	Bayesian Information Criterion
CD	„complete deletion”
DDC	a szubfunkcionalizáció „Duplication-Degeneration-Complementation” modellje
EGF	Epidermal Growth Factor
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood
MP	Maximális Parszimónia
NJ	Neighbor-Joining
PD	„pairwise deletion”
pHMM	profile Hidden Markov Model (profil Rejtett Markov Modell)
PRHD	Pairwise Repeat Homology Diagram
PSRF	Potential Scale Reduction Factor
SVG	Scalable Vector Graphics

Ábrák jegyzéke

1.1. A *Drosophila melanogaster nimrod* génjei által kódolt fehérjék prediktált doménstruktúrája. Kurucz és mtsai (2007) nyomán, módosítva. 3
1.2. A *Drosophila melanogaster* genom 34E régiójának sematikus ábrázolása. Kurucz és mtsai (2007) nyomán, módosítva. 5
1.3. A génduplikációk termékeként létrejött génpárok néhány lehetséges sorsa. A fekete szín a funkcióvesztést, míg a piros szín új funkciót jelöl. Innan (2009) alapján, módosítva. 6
1.4. Egy géncsalád evolúciós történetét szemléltető filogenetikai fa. A kék színű belső csomópontok speciációs, míg a piros színűek génduplikációs eseményeknek felelnek meg. A, B és C különböző fajokat jelölnek. A piros nyíl horizontális géntranszfert jelöl. Fitch (2000) alapján, módosítva. 7
1.5. **A** A monofiletikus, parafiletikus és polifiletikus csoportok jelentését szemléltető filogenetikai fa. A kékkel jelölt csomópont a fa gyökere. Baldauf (2003) alapján, módosítva. A **D** ábrarészleten látható konszenzushálózat a **C** és **D** ábrarészleteken látható négylevelű gyökértelen topológiák közötti konfliktust szemlélteti. 9
1.6. A filogenetikai rekonstrukció során követett leggyakoribb munkamenetek molekuláris adatok esetén. A Holder és Lewis (2003) felosztása szerint „tradicionálisnak" tekinthető módszereket zöld színnel, míg a Bayes statisztikai alapokon nyugvó módszereket piros színnel jelöltük. A kék szín az adatokat jelöli a munkamenet különböző fázisaiban. 12
1.7. A géncsaládok evolúciója a „Birth-and-Death" (**A**) és a „concerted" (**B**) modellek szerint. A különböző színű téglalapok különböző tulajdonságokkal bíró géneket jelképeznek. 15
1.8. Két gén közötti divergencia jellemző alakulása abban az esetben ha a génkonverzió gyakorisága jelentős. d a divergenciát, d_e az egyensúlyi divergenciát, míg d_t azt a küszöbértéket jelöli, aminél nagyobb divergencia esetén a homogenizáció már nem lehetséges. Innan (2009) nyomán, módosítva. 16
1.9. A szekvencia-evolúció illusztrációja egy kéttagú géncsalád esetében a homogenizáció hiánya (**A**) és jelenléte (**B**) esetén. A **C** és **D** ábrarészleteken a szekvenciaadatokból rekonstruált filogenetikai fák várható topológiái láthatóak a két említett esetben. A különböző színű téglalapok a szekvenciák közötti különbségeket jelképezik. Innan (2009) alapján, módosítva. 18

3.1. A vizsgált fajok gyökeres filogenetikája. A zöld háttér a Drosphila génuszban tartozó fajokat jelöli. Az ábra az iTOL webalkalmazás (Letunic és Bork, 2007) segítségével készült. A fa topológiája a Zdobnov és Bork (2007), valamint a Russo és mtsai (1995) által közölt fákon alapszik. 24
3.2. A becsült távolság (d) és a különbséget mutató pozíciók aránya (p) közötti összefüggés a Jukes-Cantor szubsztitúciós modell esetében. Az ábra az R 2.8.1 szoftver segítségével készült. 27
3.3. A Nimrod szekvenciák filogenetikai analízise során követett munkament. A Holder és Lewis (2003) felosztása szerint „tradicionálisnak" tekinthető módszereket zöld színnel, míg a Bayes statisztikai alapokon nyugvó módszereket piros színnel jelöltük. 28
3.4. A többszörös szekvencia-illesztés progresszív stratégiájának illusztrációja. Az **A** ábrarészleten a vezérfa, míg a **B** ábrarészleten a páronkénti profil-profil illesztések általa meghatározott sorrendje látható. Baldauf (2003) alapján, módosítva. 29
3.5. Egy kvartett esetében lehetséges három alternatív gyökértelen topológia poszterior valószínűségeit tartalmazó vektor ábrázolása egy háromszögű koordinátarendszer segítségével. T_1, T_2 és T_3 az alternatív topológiák, míg p_1,p_2 és p_3 a nekik megfelelő poszterior valószínűségek. Strimmer és von Haeseler (1997) alapján, módosítva 33

4.1. A NIM domént modellező pHMM logója. A bal felső sarokban látható vonal egy aminosavnyi hozzájárulásnak felel meg. Az oszlopok alatti számok a pHMM-ben elfoglalt pozíciót jelölik. 39
4.2. Az EGF-szerű (Pfam:PF00008) és NIM domének (**A**), valamint az EGF-szerű és EMI (Pfam:PF07546) domének (**B**) lokális-lokális illesztéseit ábrázoló páros HMM logók. Az összekötött oszlopok a megfelelő pozíciók illeszkedését, míg a szürkített hátterű oszlopok a nem illeszkedő pozíciókat jelképezik. A kék színű számozás a konzervált ciszteineket, míg fekete színű a pHMM-ben elfoglalt pozíciót jelöli. 40
4.3. A NimB szekvenciákban azonosított NIM domének gyökértelen filogenetikájának ML becslése (illesztés: `hmmalign`, szoftver: ProtTest 1.3, a BIC szerint legjobban illeszkedő szubsztitúciós modellel számolva). A leveleket a fajok szerint színeztük, az azonosítók a szekvenciákban elfoglalt pozíciót tükrözik.Az ághosszak egysége egy szubsztitúció pozíciónként. Az ábra az iTOL 1.7 (Letunic és Bork, 2007) webalkalmazás segítségével készült. 42
4.4. A NimC szekvenciákban (a *T. castaneum* NimCII és az *A. gambiae* NimCII kivételével) azonosított NIM domének gyökértelen filogenetikájának ML becslése (illesztés: `hmmalign`, szoftver: ProtTest 1.3, a BIC szerint legjobban illeszkedő szubsztitúciós modellel számolva). A leveleket a fajok szerint színeztük, az azonosítók a szekvenciákban elfoglalt pozíciót tükrözik.Az ághosszak egysége egy szubsztitúció pozíciónként. Az ábra az iTOL 1.7 (Letunic és Bork, 2007) webalkalmazás segítségével készült. 43

4.5. Az Drosophila Eater szekvenciákban, valamint a *T. castaneum* NimCII és az *A. gambiae* NimCIII szekvenciákban azonosított NIM domének gyökértelen filogenetikájának ML becslése (illesztés: `hmmalign`, szoftver: ProtTest 1.3, a BIC szerint legjobban illeszkedő szubsztitúciós modellel számolva). A leveleket a fajok szerint színeztük, az azonosítók a szekvenciákban elfoglalt pozíciót tükrözik.Az águhosszak egysége egy szubsztitúció pozíciónként. Az ábra az iTOL 1.7 (Letunic és Bork, 2007) webalkalmazás segítségével készült. 44
4.6. A *D. melanogaster* és *D. yakuba* NimB2 szekvenciákból származó NIM domének NJ algoritmussal rekonstruált (Clustal W 1.83) filogenetikája (**A**) és az ennek megfelelő PRHD (**B**). A **B** ábra jobb oldalán látható skála egysége egy pozíció. A fa águhosszainak egysége egy szubsztitúció pozíciónként. A levelek azonosítói a fajneveket (M - *D. melanogaster*, Y - *D. yakuba*) és a domének szekvenciákban elfoglalt pozícióit tükrözik. A piros téglalapok a két szekvenciában azonosított NIM doméneket jelképezik. A PRHD-n a kék vonalak az ortológia kapcsolatoknak felelnek meg, a fán nekik megfelelő kládokat kék színnel kiemeltük. 45
4.7. A *D. melanogaster* és *D. yakuba* Eater szekvenciákból származó NIM domének NJ algoritmussal rekonstruált (Clustal W 1.83) filogenetikája (**A**) és a neki megfelelő PRHD (**B**). A fa águhosszainak egysége egy szubsztitúció pozíciónként. A levelek azonosítói a fajneveket (M - *D. melanogaster*, Y - *D. yakuba*) és a domének szekvenciákban elfoglalt pozícióit tükrözik. A piros téglalapok a két szekvenciában azonosított NIM doméneket jelképezik. A PRHD-n a kék vonalak az ortológia kapcsolatoknak, míg a vörös vonalak a paralógia kapcsolatoknak felelnek meg. A fán nekik megfelelő kládokat kék, illetve vörös színnel kiemeltük. 46
4.8. A Drosophila NimB2 ortológ szekvenciák párbaállításával számolt PRHD-k. Az ábrák a `t2prhd` 1.7 segítségével ($w = 2$) készültek a Clustal W 1.83 által számolt NJ domén-filogenetikák alapján. A piros téglalapok a két szekvenciában azonosított NIM doméneket jelképezik. A kék vonalak az ortológia kapcsolatoknak felelnek meg. 50
4.9. A Drosophila NimB4 ortológ szekvenciák párbaállításával számolt PRHD-k. Az ábrák a `t2prhd` 1.7 segítségével ($w = 2$) készültek a Clustal W 1.83 által számolt NJ domén-filogenetikák alapján. A piros téglalapok a két szekvenciában azonosított NIM doméneket jelképezik. A kék vonalak az ortológia kapcsolatoknak felelnek meg. 51
4.10. A Drosophila NimC2 ortológ szekvenciák párbaállításával számolt PRHD-k. Az ábrák a `t2prhd` 1.7 segítségével ($w = 2$) készültek a Clustal W 1.83 által számolt NJ domén-filogenetikák alapján. A piros téglalapok a két szekvenciában azonosított NIM doméneket jelképezik. A kék vonalak az ortológia, míg a vörös vonalak a paralógia kapcsolatoknak felelnek meg. 52
4.11. A Drosophila NimC1 ortológ szekvenciák párbaállításával számolt PRHD-k. Az ábrák a `t2prhd` 1.7 segítségével ($w = 2$) készültek a Clustal W 1.83 által számolt NJ domén-filogenetikák alapján. A piros téglalapok a két szekvenciában azonosított NIM doméneket jelképezik. A kék vonalak az ortológia, míg a vörös vonalak a paralógia kapcsolatoknak felelnek meg. 53

4.12. A Drosophila Eater ortológ szekvenciák párbaállításával számolt PRHD-k. Az ábrák a `t2prhd` 1.7 segítségével ($w = 2$) készültek a Clustal W 1.83 által számolt NJ domén-filogenetikák alapján. A piros téglalapok a két szekvenciában azonosított NIM doméneket jelképezik. A kék vonalak az ortológia, míg a vörös vonalak a paralógia kapcsolatoknak felelnek meg. 54
4.13. A *D. melanogaster*, *D. sechellia* és a *D. virilis* Eater szekvenciákban azonosított, és a pozíciójuk alapján sorbaállított NIM domének közötti szinonim ($\mathbf{d_S}$) és nemszinonim ($\mathbf{d_N}$) távolságok ábrázolása. A távolsági mátrixokat a módosított Nei-Gojobori módszerrel, a MEGA 3.1 szoftver segítségével számoltuk ki. A gömbök méretei a megfelelő távolságokkal arányosak. A téglalapokkal jelzett összehasonlítások esetében a távolság nem számolható ki a módosított Nei-Gojobori módszerrel. 55
4.14. A *D. melanogaster* NimC1 és NimC2 szekvenciákban azonosított, és a pozíciójuk alapján sorbaállított NIM domének közötti szinonim ($\mathbf{d_S}$) és nemszinonim ($\mathbf{d_N}$) távolságok ábrázolása. A távolsági mátrixokat a módosított Nei-Gojobori módszerrel, a MEGA 3.1 szoftver segítségével számoltuk ki. A gömbök méretei a megfelelő távolságokkal arányosak. A téglalapokkal jelzett összehasonlítások esetében a távolság nem számolható ki a módosított Nei-Gojobori módszerrel. . 56
4.15. A Drosophila NimC1, NimC2 és Eater szekvenciákban azonosított NIM doméneket kódoló nukleotid-szekvenciák többszörös illesztéseiből számolt szaturációs ábrák. A pontok az egyes szekvenciapárok között megfigyelhető különbségek számát ábrázolják, az F84 szubsztitúciós modell által számolt távolságok függvényében. A vonalak a megfelelő ponthalmazokra illesztett lineáris regressziós egyenesek. A zöld háromszögek és egyenesek a transzverziós, míg a kék keresztek és egyenesek a tranzíciós különbségekhez tartoznak. Az ábrák a DAMBE 4.2.13 szoftverrel készültek. 58
4.16. A NimA , NimB és NimC szekvenciák ProbCons illesztéseinek likelihood térképezése. A jobb oldalon látható háromszögű koordinátarendszerek értelmezéséhez lásd a 3.5. ábrát. A bal oldalon látható koordinátarendszerek a megfelelő régiókba eső pontok arányát szemléltetik. 59
4.17. A NimA szekvenciák gyökértelen filogenetikájának ML becslése (illesztés: ProbCons 1.1, szoftver: PhyML 3.0, szubsztitúciós modell: JTT+Γ + F, gamma paraméter: 1,215, gamma kategóriák száma: 4, bootstrap ismétlések száma: 1000). A kládok bootstrap támogatottsága a megfelelő csomópontban található kördiagramok piros szeletével arányos. Az águhosszak egysége egy szubsztitúció pozíciónként. Az ábra az APE R csomag (Paradis és mtsai, 2004) segítségével készült. 60
4.18. A NimB szekvenciák gyökértelen filogenetikájának ML becslése (illesztés: ProbCons 1.1, szoftver: PhyML 3.0, szubsztitúciós modell: VT+Γ + F, gamma paraméter: 2,291, gamma kategóriák száma: 4, bootstrap ismétlések száma: 1000). A kládok bootstrap támogatottsága a megfelelő csomópontban található kördiagramok piros szeletével arányos. A Drosophila ortológokat azonos színnel jelöltük. Az águhosszak egysége egy szubsztitúció pozíciónként. Az ábra az APE R csomag (Paradis és mtsai, 2004) segítségével készült. 62

4.19. A NimB szekvenciák filogenetikájának Bayes MCMC rekonstrukciója nyomán kapott poszterior mintákból az $x = 0,2$-es szinten számolt konszenzushálózat. A konszenzushálózat csak a topológiák közötti konfliktus szemléltetésére szolgál, így az ághosszaknak nincs jelentése. A Drosophila ortológokat azonos színnel jelöltük. 63
4.20. A NimC szekvenciák gyökértelen filogenetikájának ML becslése (illesztés: ProbCons 1.1, szoftver: PhyML 3.0, szubsztitúciós modell: VT+Γ + F, gamma paraméter: 2,526, gamma kategóriák száma: 4, bootstrap ismétlések száma: 1000). A kládok bootstrap támogatottsága a megfelelő csomópontban található kördiagramok piros szeletével arányos. A Drosophila ortológokat azonos színnel jelöltük. Az ághosszak egysége egy szubsztitúció pozíciónként. Az ábra az APE R csomag (Paradis és mtsai, 2004) segítségével készült. 64
4.21. A NimC szekvenciák filogenetikájának Bayes MCMC rekonstrukciója nyomán kapott poszterior mintákból az $x = 0,2$-es szinten számolt konszenzushálózat. A konszenzushálózat csak a topológiák közötti konfliktus szemléltetésére szolgál, így az ághosszaknak nincs jelentése. A Drosophila ortológokat azonos színnel jelöltük. 65
4.22. A *D. melanogaster* Nimrod paralógok illesztésének és filogenetikájának együttes Bayes MCMC becslése nyomán kapott poszterior mintákból az $x = 0,2$-es szinten számolt konszenzushálózat. A konszenzushálózat csak a topológiák közötti konfliktus szemléltetésére szolgál, így az ághosszaknak nincs jelentése. Az ágakon a poszterior valószínűségek láthatók. A **Dmel.Drp** a *D. melanogaster* Draper szekvenciát jelöli. A színezés a NimA/Draper, NimB és NimC kládokat emeli ki. 66
4.23. A Nimrod klaszter szerveződése a vizsgált genomokban (sematikus ábrázolás). A fajok filogenetikáján csak a génuszneveket tüntettük fel. A színezés a különböző géncsaládokba tartozó gének megkülönböztetésére szolgál, a *D. melanogaster* esetében ezek nevét is feltüntettük. Az *ance3* gén esetében a *Vajk* géneket tartalmazó nagyméretű intront is feltüntettük. A divergencia idők forrása: Honeybee Genome Sequencing Consortium (2006). 68

5.1. A Nimrod klaszter közvetlen környezetének sematikus ábrázolása a *D. melanogaster*-ben, néhány idevonatkozó irodalmi adat feltüntetésével. A legfelső sor téglalapjai a körülbelül két megabázisnyi 2L:13490000-14490000 régióban található 102 gént jelképezik. A géneket a 4.23. ábrának megfelelően színkódoltuk. A további sorok a következő, a régió génjeire vonatkozó genomikai adatokat jelképezik: együttesen expresszálódó gének klaszterei (Spellman és Rubin, 2002); a Drosphila fajokban a génexpressziós szint összehangolt evolúcióját mutató klaszterek (Mezey és mtsai, 2008), a különböző méretű mozgó ablakokkal kapott átfedő klasztereket összeolvasztottuk; a CTCF és BEAF-32 fehérjéket kötő inzulátor-elemek hozzávetőleges pozíciója (modENCODE, http://modencode.org), a Nimrod gének többségét tartalmazó inzulátoroktól mentes intervallumok kiemelésével; a Polycomb fehérje kötődési doménjei (Tolhuis és mtsai, 2006); az Eve („Even-skipped”) kötődési helyeit jellemző legnagyobb azonosított „Eve BRICK” régió (de Wit és mtsai, 2008). . . 70

5.2. Az EMI és a NIM domének eredetére vonatkozó hipotézis szemléltetése pHMM logók segítségével. A kék színű számozás a konzervált ciszteineket, míg fekete színű a pHMM-ben elfoglalt pozíciót jelöli. 73
5.3. A Nimrod szupergéncsalád evolúciójával kapcsolatos hipotéziseink összefoglalása. 74

1. fejezet

Bevezetés

A veleszületett immunitás mechanizmusai a patogénekkel való kapcsolatba lépés után azonnal működésbe lépnek, így ezen folyamatok a kórokozók elleni védekezés első védelmi vonalát képezik (Kimbrell és Beutler, 2001).

A veleszületett immunitás a patogének elleni védekezés ősi és univerzálisan elterjedt mechanizmusa, amely korlátozott számú, a csíravonalban öröklődő receptororra hagyatkozik. Ezek képesek felismerni a mikroorganizmusok által általában termelt, de a gazdaszervezetben nem előforduló molekulákat (Janeway és Medzhitov, 2002).

A veleszületett immunitás mechanizmusainak rendelkezésére álló receptorkészlet bővítésében nagy szerepe van a génduplikációknak, amelyek különböző specificitással rendelkező receptorokat kódoló géncsaládokat hozhatnak létre. Ennek egyik legismertebb példája a Toll-szerű receptorokat kódoló géncsalád (Takeda és Akira, 2007).

Nagyszámú vizsgálat utal arra, hogy a gerincesek és a gerinctelenek védekezési mechanizmusai molekuláris szinten erősen konzerváltak, ami ígéretes távlatokat nyit a veleszületett immunitás megértésében (Kimbrell és Beutler, 2001; Govind és Nehm, 2004)

A filogenetikai rekonstrukció a géncsaládok tanulmányozásának alapvető és széles körben alkalmazott eszköze, hasznosnak bizonyult például a Toll-szerű receptorok vizsgálata során is (Kimbrell és Beutler, 2001; Roach és mtsai, 2005). Ezen vizsgálatok a homológia-viszonyok felderítésének legpontosabb módját képezik (Futuyma, 2005), ennek köszönhetően jelentős mértékben hozzájárulnak a modellorganizmusokban végzett kísérletes munkák eredményeinek általánosításához. A filogenetikai fák ugyancsak segítenek kijelölni a további vizsgálatok irányát is: a felderítő célú kutatások során érdemes például olyan génekre fókuszálni, amelyek olyan kládokból származnak, ahol a génekre nézve még nem állnak rendelkezésre funkcionális adatok.

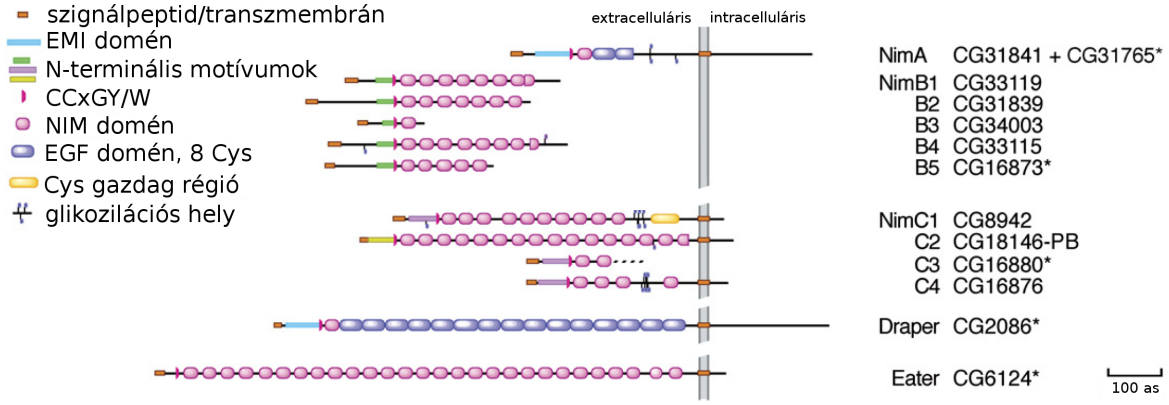
A filogenetikai módszerek nagyobb léptékben is alkalmazhatóak. Így például egy adott genetikai rendszer részét képező géncsaládok filogenetikájának rekonstrukciója, majd ezen filogenetikáknak a fajok evolúció történetével való összehasonlítása révén következtetéseket vonhatunk ezek kialakulására nézve (Litman és mtsai, 2005). Egyes szerzők hasonló vizsgálatok nyomán vonták le azt a következtetést, hogy a gerincesek és az ízeltlábúak veleszületett immunitásának egyes fontos elemei valószínűleg egymástól függetlenül alakultak ki (Hughes, 1998).

A nemrégiben leírt Nimrod szupergéncsalád (Kurucz és mtsai, 2007) tagjai által kódolt proteinek sajátossága egy vagy több speciális EGF-szerű domén, a CxPxCxxxCxNGxCxxPxxCxCxxGY konszenzus-szekvenciával leírható úgynevezett NIM domén jelenléte. Az EGF-szerű domének (Pfam clan: CL0001) az extracelluláris fehérjék gyakori strukturális elemei, amelyek gyakran több példányban fordulnak elő, domén-ismétlődéseket képezve (Bork, 1991; Bork és mtsai, 1996).

A Nimrod szupercsalád által kódolt fehérjék másik jellegzetessége egy, a CCxGY/W

konszenzussal leírható motívum jelenléte az aminosav-szekvenciák N-terminális régiójában, amit egy rövid összekötő szekvencia után mindig egy NIM domén követ (Kurucz és mtsai, 2007). A kódolt fehérjék ugyancsak minden esetben rendelkeznek szignálpeptiddel, ami arra utal, hogy funkciójukat az extracelluláris térben látják el.

A szupercsalád tagjai által kódolt fehérjék minden esetben rendelkeznek a fenti közös elemekkel, de emellett a doménstruktúrájuk igen nagy változatosságot mutat (1.1. ábra).



1.1. ábra. A *Drosophila melanogaster nimrod* génjei által kódolt fehérjék prediktált doménstruktúrája. Kurucz és mtsai (2007) nyomán, módosítva.

A doménstruktúrák alapján a Nimrod proteinek több fő csoportba sorolhatók. A Draper típusú doménstruktúrával rendelkező fehérjék N-terminális régiójában a szignálpeptid után egy EMI domén (Callebaut és mtsai, 2003), és a NIM domén egy példánya található, amit változó számú „klasszikus" EGF domén követ. Ez a típus széles körben elterjedt, az ilyen típusú fehérjéket kódoló gének megtalálhatóak például a *Caenorhabditis elegans* (Mangahas és Zhou, 2005), a *Drosophila melanogaster* (Manaka és mtsai, 2004) és az ember genomjában (Hamon és mtsai, 2006). Ezzel szemben ismétlődő NIM doméneket tartalmazó architektúrával („poli-NIM") rendelkező fehérjéket kódoló géneket eddig csak rovargenomokban sikerült azonosítani (Kurucz és mtsai, 2007). Ezen architektúrával rendelkező proteinek nem tartalmaznak „klasszikus" EGF doméneket és a CCxGY/W motívumot változó számú rövid összekötő szekvenciák által határolt NIM domének követik (1.1. ábra).

A poli-NIM architektúrával rendelkező fehérjéket a Nimrod C és Nimrod B típusokra oszthatjuk. A Nimrod C típusú proteinek, mint például a Drosophila *nimrodC1-4* és *eater* gének termékei, általában rendelkeznek transzmembrán doménnel. Ezzel szemben a Nimrod B típusú proteinek (például a Drosophila *nimrodB1-B5* gének termékei) rövidebb szekvenciával rendelkeznek és nem tartalmaznak transzmembrán domént.

A *D. melanogaster* Nimrod C típusú fehérjéi közül az Eater fehérje rendelkezik a legtöbb NIM doménből álló doménismétlődéssel (1.1. ábra).

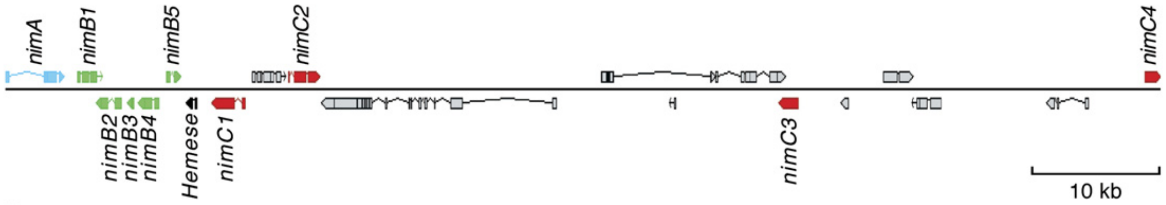
A Nimrod szupercsalád tagjairól jelentős mennyiségű olyan irodalmi adat áll rendelkezésre,

ami ezen géneknek a veleszületett immunitásában játszott szerepére utal. Számos organizmusban azonosítottak a fagocitózisban szerepet játszó Draper típusú doménstruktúrával rendelkező fehérjéket. Ilyen például a *C. elegans* Ced-1 (Mangahas és Zhou, 2005), a Drosophila Draper (Manaka és mtsai, 2004; Cuttell és mtsai, 2008), valamint a humán MEGF-10 (Hamon és mtsai, 2006) fehérje.

Számos Nimrod C típusú fehérjét kódoló génről is bebizonyosodott, hogy szerepet játszik a fagocitózisban. Ilyenek például a *D. melanogaster eater* (Kocks és mtsai, 2005) és *nimrodC1* (Kurucz és mtsai, 2007). Mindkét gén expresszálódik a hemocitákban és az Eater proteinek esetében a baktérimkötés közvetlenül is bizonyított. A *nimrodC4* (*SIMU*) génről is bebizonyosodott, hogy szerepet játszik az apoptotikus törmelékek fagocitózisában (Kurant és mtsai, 2008). Ugyancsak megtalálható egy, a Nimrod C típusú és a fagocitózisban szerepet játszó proteint kódoló gén a *Sarcophaga peregrina* húslégy genomjában (Nishikawa és Natori, 2001) és egy, a *Holotrichia diomphalia* bogárban található Nimrod B típusú fehérjéről is kimutatták, hogy bakteriális lipopoliszacharid-receptorként működik és *in vivo* elősegíti az *Escherichia coli* fagocitózisát (Ju és mtsai, 2006). Mindezen adatok arra utalnak, hogy a szupercsalád tagjai a rovarok veleszületett immunitásásának fontos elemeit képezik.

A *D. melanogaster hemese* gén által kódolt fehérje szekvenciájának egyes elemei arra utalnak, hogy ez a szupercsalád részét képezheti (Dan Hultmark, személyes közlés), de kellő evidencia hiányában ezt nem vontuk be a vizsgált gének körébe. Ezen kívül Kurucz és mtsai (2007) egy, a fentebb ismertetettektől eltérő doménstrukturával rendelkező, de a szupercsalád jellegzetességeit (CCxGY/W motívum, NIM domén) magán viselő proteint kódoló Drosophila gént is megemlít (CG7447). Ezen gén esetében a szupercsalád többi tagjával feltételezhetően homológiát mutató szekvencia-régiók meglehetősen rövidek és a doménstruktúra nagymértékben eltérő, ezért vizsgálataink során a fentebb ismertetett típusokra szorítkoztunk.

Kurucz és mtsai (2007) a Nimrod szupergéncsalád leírása során észrevették, hogy a *nimrod* gének nagy része a *Drosophila melanogaster* genomjában egymás közvetlen közelében helyezkedik el a második kromoszóma 34E régiójában (1.2. ábra) és ez a klaszterizáció egyéb fajok genomjaiban (*Anopheles gambiae* (Diptera), *Apis mellifera* (Hymenoptera)) is megfigyelhető.



1.2. ábra. A *Drosophila melanogaster* genom 34E régiójának sematikus ábrázolása. Kurucz és mtsai (2007) nyomán, módosítva.

További ízeltlábú genomok, vagyis a *Tribolium castaneum* (Coleoptera), *Pediculus humanus* (Phthiraptera) és a *Daphnia pulex* (Cladocera) fajok genomjainak vizsgálata során azt találtuk, hogy a szupercsalád tagjai egyéb géncsaládokkal együtt egy nagyobb időléptékben konzervált génklasztert alkotnak, amit a továbbiakban „Nimrod klaszter”-nek fogunk nevezni.

A Nimrod klaszter vizsgálatát a konzerváltság időléptéke (300-350 millió év) önmagában is jól indokolja, de az is vonzóvá teszi, hogy irodalmi adatok szerint a *D. melanogaster* genomban az immungének hajlamosak klaszterekbe tömörülni (Wegner, 2008).

1.1. Génduplikációk, géncsaládok és homológia

Az Ohno (1970) által végzett úttörő munka nyomán a génduplikációk, a nagyobb genomi régiókat érintő „en bloc” duplikációk, illetve a genomduplikációk evolúcióban betöltött szerepe széles körben elismertté vált és ezeket a folyamatokat mind a mai napig a molekuláris és fenotipikus evolúció legfontosabb hajtóerejének tartják (Koonin, 2009a).

A gradualista szemléletbe jól illeszkedő pontmutációkkal ellentétben a duplikációs események hatása ugrásszerűbb. A már meglévő gének új példányai a funkcionális kényszerek alól felszabadulva mintegy „nyersanyagként” szolgálnak, amiből az evolúció a közismert metafora szerint „barkácsol”. A duplikációk szerepét illetően még sok a nyitott kérdés, de fontosságukat jól szemlélteti, hogy az eukarióták esetében a génduplikációk fixálódásának egy génre becsült rátája nagyságrendileg megközelíti a pontmutációk egy pozícióra becsült rögzülésének rátáját (Lynch és Conery, 2003).

A kisebb genomi régiókat érintő génduplikációk mechanizmusaként leggyakrabban az egyenlőtlen rekombinációt és a retrotranszpozíciót szokták megnevezni (Zhang, 2003a; Hahn, 2009). Bármi is legyen az a mechanizmus, ami révén a kópiaszám-változat[1] a populációban felbukkan, rögzülését a populációgenetikai tényezők, vagyis a mutációk természete és rátája, a genetikai sodródás (drift) és a természetes szelekció határozzák meg (Korbel és mtsai, 2008).

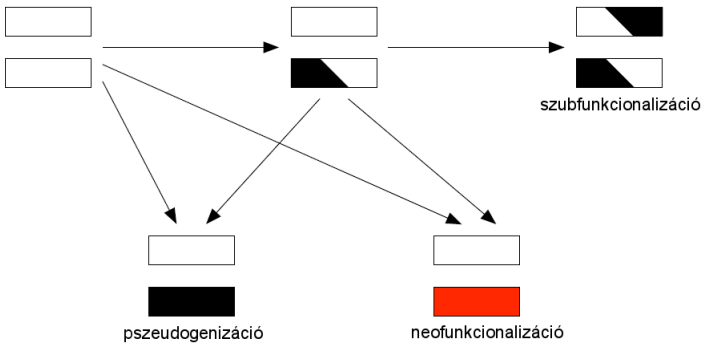
Ohno (1970) munkája nyomán a génduplikációk fixálódás utáni lehetséges sorsával

[1]Copy Number Variant

kapcsolatban a következő főbb útvonalakat tárgyalják (Walsh, 2003; Zhang, 2003a; Hahn, 2009, 1.3. ábra):

- *Deléció* - az egyik példányt a genomból eltávolító deléció rögzülése.
- *Pszeudogenizáció* - inaktiváló mutációk hatására az egyik kópia elveszti funkcióját.
- *A génfunkció konzerválódása* - a duplikáció után mindkét gén megtartja eredeti funkcióját, ami a géntermék dózisának megduplázódásához vezethet.
- *Szubfunkcionalizáció* - mindkét példányban részleges funkcióvesztést okozó mutációk fixálódnak, vagy megváltozik a gének expressziós mintázata, aminek eredményeként mind a két példány szükségessé válik az eredeti funkció ellátásához. Ezt a mechanizmust a szubfunkcionalizáció DDC vagyis „Duplication-Degeneration-Complementation" modelljének nevezik. Ide sorolják továbbá azt az esetet is, amikor a duplikáció révén elkerülhetővé válik a szegregáció és így a heterozigóta előny állandósul[2].
- *Neofunkcionalizáció* - az egyik példányban rögzült mutációk egy új funkció ellátását teszik lehetővé.

Ezek az események még a mutáció rögzülése előtt megtörténhetnek, de az egyszerűség kedvéért sokszor a duplikációk rögzülését és további sorsát külön tárgyalják (Walsh, 2003).



1.3. ábra. A génduplikációk termékeként létrejött génpárok néhány lehetséges sorsa. A fekete szín a funkcióvesztést, míg a piros szín új funkciót jelöl. Innan (2009) alapján, módosítva.

A gének funkciójában bekövetkezett változásokban nem csak a kódoló szekvenciákban, hanem a szabályozó régiókban fixálódott mutációknak is jelentős szerepe lehet, ezek ugyanis változásokat okozhatnak a gének expressziós mintázatában. Ezek a változások, úgy tűnik, gyakran fellépnek a duplikációs események után (Zhang, 2003a).

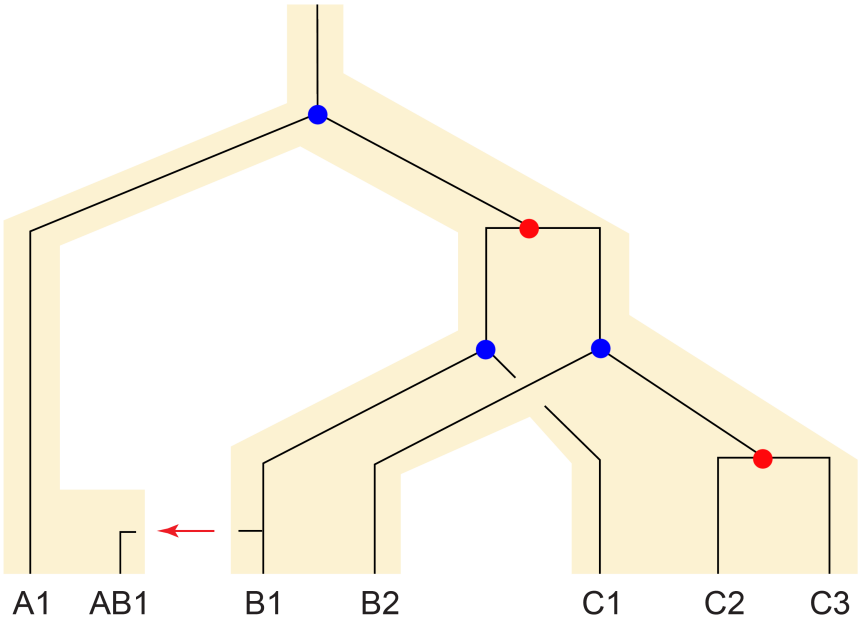
[2] „segregation avoidance"

A rögzült duplikációk sorsát továbbra is a már említett populációgenetikai tényezők (mutáció, szelekció, genetikai sodródás) határozzák meg. A természetes szelekció nagy szerepet játszhat a megváltozott funkcióval rendelkező mutáns gének rögzülésében és a populációban való megtartásában, de ehhez nem minden esetben van szükség pozitív szelekciós hatásra, mint ahogy azt a szubfunkcionalizáció DDC modellje példázza.

A speciációs események és a duplikációk révén megjelenő új példányok fixálódása révén homológ gének csoportjai jönnek létre, amelyeket géncsaládoknak nevezünk.

Homológia alatt a biológiában két olyan struktúra (pl. szerv, szekvencia) közötti relációt értünk, amelyek közös evolúciós eredetüknél fogva számottevően hasonlítanak egymásra (Fitch, 2000). A homológ struktúrák gyakran ugyanazt a funkciót töltik be, de ez egyáltalán nem szükségszerű.

A homológia egy hipotézis, amivel a struktúrák közötti hasonlóságot magyarázzuk. Legtöbbször ez a feltevés magán a hasonlóságon alapszik (különösen a szekvenciák esetében), de a homológia-hipotézisek felállításában nem csak a struktúrák közötti hasonlóság, hanem azok pozíciója is segítségünkre lehet, így például egy gén genomban elfoglalt viszonylagos pozíciója.



1.4. ábra. Egy géncsalád evolúciós történetét szemléltető filogenetikai fa. A kék színű belső csomópontok speciációs, míg a piros színűek génduplikációs eseményeknek felelnek meg. A, B és C különböző fajokat jelölnek. A piros nyíl horizontális géntranszfert jelöl. Fitch (2000) alapján, módosítva.

Attól függően, hogy milyen események révén léptek a gének a független evolúció útjára (1.4. ábra), a homológiának több vállfaját különböztetjük meg. Ezekre nézve a terminológia

nem teljesen egységes, a jelen dolgozatban a Fitch (2000) által javasolt definíciókat fogjuk használni, amelyek a következők:

- *Ortológia* - ortológnak nevezzük azokat a génpárokat, amelyek egy speciációs esemény következtében indultak el az önálló evolúció útján. Ilyenek például az A1-B2 és B2-C2 génpárok az 1.4. ábrán.
- *Paralógia* - paralógoknak nevezzük azokat a génpárokat, amelyek egy génduplikációs esemény révén indultak el a független evolúció útján. Paralógok például a C2-C3 és B1-C2 génpárok az 1.4. ábrán.
- *Xenológia* - xenológoknak nevezzük azokat a génpárokat, amelyeknek kapcsolata horizontális génátvitellel értlemezhető (AB1 és B1 az 1.4. ábrán).

A géncsaládok fentebb említett definíciója szerint ortológ gének csoportja (fajonként egy-egy génnel) is egy géncsaládnak tekintendő. Az egymással valószínűsíthetően leszármazási kapcsolatban lévő géncsaládokat nagyobb csoportokba szokták sorolni, amelyeket Nei és Rooney (2005) nyomán „szupergéncsaládoknak”[3] nevezünk.

1.2. Filogenetikai fák és hálózatok

A fastruktúrákat már Darwin óta használják a különböző biológiai objektumok (pl. fajok, szekvenciák) közötti leszármazási kapcsolatok szemléltetésére. A filogenetikai fák és módszerek központi szerepet játszanak a szisztematikai és molekuláris evolúciós vizsgálatokban (így a géncsaládok tanulmányozásában is), ezek célja ugyanis bizonyos objektumok evolúciós történetének megértése.

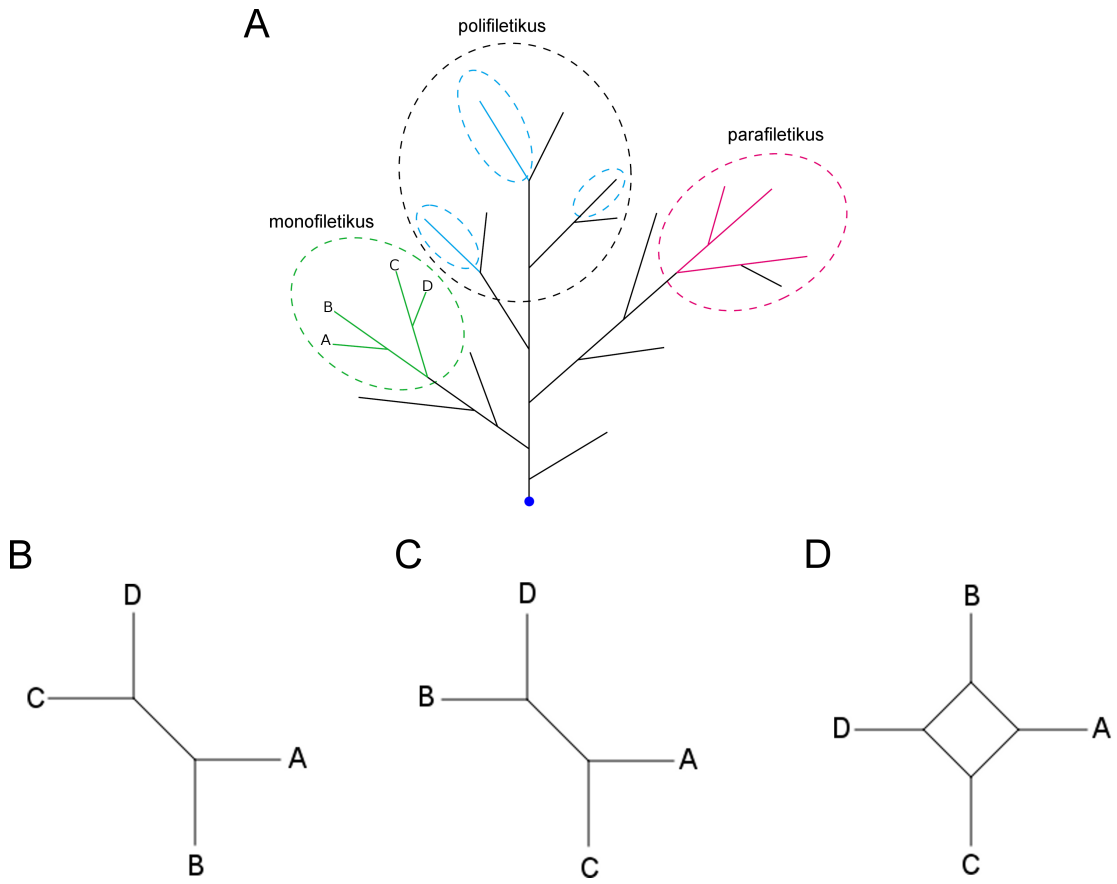
A filogenetikai fák fő elemei a csomópontok és az őket összekötő ágak (1.5. **A** ábra). A fa végpontjai (a „levelek”) legtöbbször, de nem minden esetben, a recens objektumokat jelölik, míg a belső csomópontok az illető részfában található objektumok közös, potenciálisan „kihalt” hipotetikus őseinek felelnek meg.

A filogenetikai módszerek célja, hogy különböző természetű adatokra alapozva (morfológia, szekvenciák) következtessenek a vizsgált objektumok evolúciós történetére, aminek egyik módja az ezt leírni hivatott fastruktúra „felépítése”. A csomópontokat összekötő ágak hosszának értelmezése függ a felhasznált filogenetikai módszertől, vagy akár a felhasznált karakter-evolúciós modelltől is (Liò és Goldman, 1998; Felsenstein, 2002). Így jelképezheti az időt vagy a két csomópont között lezajlott változások becsült értékét (a divergenciát).

A filogenetikai fákon látható objektumokat különbözőképpen csoportosíthatjuk, erre vonatkozóan használni szokták a következő terminusokat (1.5. **A** ábra):

[3] „supergene family”

- *„monofiletikus csoportnak”*, vagy *„kládnak”* nevezzük a belső csomópontokból eredő részfákban található objektumokat, az illető belső csomópontot is beleértve. Az eredmények tárgyalása során a kládokat, valamint a kládokon belüli elágazások mintázatát az elterjedt zárójeles jelöléssel írjuk le (Newick formátum). Az 1.5. **A** ábrán zöld színnel jelölt klád például az „((A, B), (C, D))” zárójeles jelölésnek felel meg.
- *„parafiletikusnak”* nevezzük a fentihez hasonló csoportokat, amelyek az adott belső csomópontnak nem tartalmazzák minden egyes leszármazottját.
- *„polifiletikusnak”* nevezünk egy csoportot, ha az az adott belső csomópont leszármazottain kívül más objektumokat is tartalmaz.



1.5. ábra. **A** A monofiletikus, parafiletikus és polifiletikus csoportok jelentését szemléltető filogenetikai fa. A kékkel jelölt csomópont a fa gyökere. Baldauf (2003) alapján, módosítva. A **D** ábrarészleten látható konszenzushálózat a **C** és **D** ábrarészleteken látható négylevelű gyökértelen topológiák közötti konfliktust szemlélteti.

A filogenetikai fa elágazási mintázatát (vagyis az általa meghatározott kládok összességét) topológiának nevezzük. A filogenetikai fa minden ága a leveleket két diszjunkt halmazra osztja, amit „split”-nek nevezünk. Egy filogenetikai fa által meghatározott összes splitek

„kompatibilis” halmaza az összes információt hordozza az evolúciós történetről és ezek a fa topológiáját teljes mértékben meghatározzák (Felsenstein, 2002).

Az 1.4. ábra alapján nyilvánvaló, hogy a géncsaládok esetében a filogenetikai fa elágazásai génduplikációs, illetve speciációs eseményeknek felelhetnek meg, valamint hogy a géncsaládok és az őket „hordozó” fajok filogenetikája nem feltétlenül egyezik meg. A géncsaládok és a fajok filogenetikája egymással „összeegyeztethető” („reconciliation”), ha a géncsaládok filogenetikájának belső csomópontjaihoz hozzárendeljük a megfelelő speciációs és duplikációs eseményeket, mint ahogy az a 1.4. ábrán látható.

A filogenetikai fák rendelkezhetnek gyökérrel (1.5. **A** ábra), ami meghatározza az evolúció irányát. Csak a gyökeres fák írják le pontosan a vizsgált objektumok evolúciós történetét. A gyökérhez közelebb leágazó kládokat bazálisnak szokás nevezni.

Mivel a filogenetikai módszerek többsége eredményként gyökértelen fákat ad, szükség van olyan eljárásokra, amelyeknek segítségével következtethetünk a gyökér pozíciójára. Ez megállapítható egy, a vizsgált objektumokkal homológ, de rájuk nézve feltehetőleg bazális objektumok (külcsoport) bevonásával, vagy „molekuláris óra” felhasználásával. Géncsaládok esetében a gyökér pozíciójára következtethetünk a fajok gyökeres filogenetikájával való összeegyeztetés révén (Felsenstein, 2002).

A filogenetikai fákra jellemző, hogy minden csomópont maximum egy „bemeneti ággal” rendelkezik. A filogenetikai fa ezen megkötés miatt nem mindig alkalmas az evolúciós történet leírására. Ilyen esetekben úgynevezett „filogenetikai hálózatokat” használnak, amelyeken az „extra ágak” a módszertől és vizsgált objektumoktól függően jelképezhetnek például hibridizációt, rekombinációt, horizontális génátvitelt, illetve kifejezhetik az evolúciós történet becslésében rejlő bizonytalanságokat (Huson és Bryant, 2006). Például az úgynevezett „konszenzushálózatok” (Holland és Moulton, 2003; Huson és Bryant, 2006) a konszenzufákhoz hasonlóan alkalmasak több topológia összegzésére, azonban az előbbiekkel ellentétben a konfliktusokat nem többszörös elágazásokkal (politómiákkal), hanem „extra ágakkal” jelképezik (1.5. **D** ábra).

1.3. A filogenetikai rekonstrukciós módszerek áttekintése

A különböző biológiai objektumok megfigyelhető struktúrái (karakterek) információt hordozhatnak azok evolúciós történetéről (Wiley és mtsai, 1991; Felsenstein, 2002). A homológ karakterek különböző objektumokban felvett állapotai képviselik azt az információt, amiből a filogenetikára következtethetünk. Ezen okból kifolyólag a homológia-hipotézisek felállítása és a karakterállapotok megállapítása nagy jelentőséggel bír.

A biológiai szekvenciák esetében a karakterállapotok megállapítása nem okoz különösebb gondot, míg morfológiai karakterek esetében ez nehezebb lehet. Ezzel ellentétben a

szekvenciák esetében a homológia-hipotézisek felállítása - ami a többszörös illesztés feladata - nagyobb gondot okozhat, mint a morfológiai karakterek esetében, ahol általában erre nézve több információ áll rendelkezésre (Wiley és mtsai, 1991; Kumar és Filipski, 2007). A továbbiakban a „karakter" kifejezés alatt egy homológ karakterkészletet értünk, hacsak nem jelezzük az ellenkezőjét.

Elegendő számú, megfelelő változatosságot mutató, és a tulajdonképpeni evolúciós történettel teljes összhangban lévő (azzal kompatibilis) karakter birtokában a filogenetika viszonylag könnyen és egyértelműen megállapítható lenne (Felsenstein, 2002). A valóságban azonban a felvehető karakterállapotok számának véges volta miatt fellépő reverzió és konvergens evolúció miatt gyakran előfordul, hogy egy karakter két különböző objektumban nem a leszármazási kapcsolatok miatt mutatja ugyanazt az állapotot. Így a karakterek már nem tükrözik tökéletesen a tulajdonképpeni filogenetikát és egymással sem szükségszerűen kompatibilisek.

A filogenetikai rekonstrukciós módszerek célja, hogy ezekből a „zajos" karakterekből következtessen a leszármazási kapcsolatokat megjelenítő filogenetikai fára (Holder és Lewis, 2003). A biológiai szekvenciákból való filogenetikai rekonstrukció során alkalmazott gyakori munkameneteket a 1.6. ábra foglalja össze.

Egy adott filogenetika keretein belül értelmezve a konvergens evolúció miatt fennálló hasonlóságok mint a karakterállapotok változásaira nézve fenntartott extra hipotézisek jelentkeznek (homoplázia). A parszimónia (MP), az egyik legrégebbi múltra visszatekintő optimalitási kritériumon alapuló rekonstrukciós módszer, a homoplázia mértékét minimalizáló fa megkeresésével próbál következtetni a filogenetikára.

A parszimónia azonban a karakterevolúció számos fontos elemét elhanyagolja, ami jelentős torzításokat okozhat, emellett az optimális fa kiválasztása is erőforrásigényes lehet. Ezen problémák orvoslására tett erőfeszítések számos új, optimalitási kritériumokon alapuló, valamint a fát bizonyos rögzített eljárások alapján felépítő algoritmikus módszert hívtak életre (Holder és Lewis, 2003), melyek a szisztematika alapvető eszköztárának részeként, de különböző filozófiai/módszertani hátterű iskolákban teljesedtek ki (Felsenstein, 2001).

Napjainkra a legelterjedtebb nézőpont az összetett parametrikus modelleket felvonultató statisztikus filogenetika, amely a rekonstrukciót statisztikai becslésnek tekinti és az eredményeket is ebben a keretben értelmezi (Felsenstein, 2001; Whelan és mtsai, 2001b). Ezen értelmezés szerint tehát a filogenetikai rekonstrukció során a rendelkezésre álló adatokból, legtöbbször egy, a karakterek evolúcióját leíró modell segítségével megbecsüljük a fa topológiáját, az ághosszakat és az egyéb paramétereket.

A különböző optimalitási kritériumokon alapuló becslések (például a Maximum Likelihood - ML) és a Bayes statisztikai módszerek mellett még számos algoritmikus módszer is jól illeszkedik ebbe a keretbe, amennyiben valamelyik optimalitási kritérium szerinti becslés

1.6. ábra. A filogenetikai rekonstrukció során követett leggyakoribb munkamenetek molekuláris adatok esetén. A Holder és Lewis (2003) felosztása szerint „tradicionálisnak” tekinthető módszereket zöld színnel, míg a Bayes statisztikai alapokon nyugvó módszereket piros színnel jelöltük. A kék szín az adatokat jelöli a munkamenet különböző fázisaiban.

közelítéseként fogható fel. Ilyen például a népszerű Neighbor-Joining módszer, ami a minimális evolúció kritérium egyik változatának mohó optimalizációjaként fogható fel (Gascuel és Steel, 2006).

A Holder és Lewis (2003) felosztása alapján „tradicionális” módszerek - amelyeknek körébe a legtöbb alkalmazott eljárás (ML, NJ, stb.) tartozik - eredményként egy fát adnak, vagyis a filogenetika egy pontbecslését, míg a Bayes statisztika elvein alapuló módszerek ezekkel ellentétben egy, az összes lehetséges topológián értelmezett poszterior eloszlást becsülnek meg. A tradicionális módszerek esetében tehát szükség van egy olyan eljárásra, aminek segítségével a becslés bizonytalansága felmérhető. Ezen bizonytalanság felmérésére számos módszer létezik (Felsenstein, 2002), ezek közül a leggyakrabban alkalmazott módszer az úgynevezett „nemparaméteres bootstrap” (Felsenstein, 1985).

A nemparaméteres bootstrap során az illesztés pozícióit függetlennek tekintjük és ezekből visszatevéses mintázással az eredeti illesztéssel megegyező hosszúságú „újramintázott”

illesztéseket generálunk (1.6. ábra). Nagyszámú (általában több mint ezer) újramintázott illesztésből az eredeti illesztés esetében alkalmazott rekonstrukciós módszer segítségével filogenetikai fák egy halmazát kapjuk. A becsült filogenetikai fa kládjainak bootstrap támogatottsága nem más, mint az így kapott fahalmazban azon topológiák aránya, amelyek az adott kládot tartalmazzák. Az irodalomban a bootstrap értékeket általában százalékban adják meg és a 70%-nál nagyobb értékekkel rendelkező kládokat „szignifikánsnak" tekintik (Holder és Lewis, 2003; Alfaro és Holder, 2006).

A többszörös szekvencia-illesztés és a filogenetika rekonstrukciója tulajdonképpen nem két jól elhatárolható független probléma, ugyanis az illesztés során az inszerciók és deléciók helyes súlyozásához szükség van a filogenetika ismeretére (Lunter és mtsai, 2005). Ezen kívül, ahogy az már említésre került, a szekvenciák esetében a homológia-hipotézisek felállítása (illesztés) jelentős bizonytalanságot foglalhat magába. Azok a módszerek, amelyek egy rögzített többszörös illesztésből indulnak ki, ezt a bizonytalanságot elhanyagolják, ami jelentős torzításokhoz vezethet (Redelings és Suchard, 2005).

Ezen torzítások kiküszöbölhetők olyan módszerek alkalmazásával, amelyek a többszörös illesztést és a filogenetikát egyidejűleg becsülik meg. Ezek közül a legnépszerűbbek a Markov Lánc Monte Carlo[4]-t (MCMC) használó Bayes statisztikai módszerek (Lunter és mtsai, 2005; Redelings és Suchard, 2005). Ezen módszerek erőforrásigénye jelentős lehet, de alkalmazásukkal a homológia-hipotézisek felállításában rejlő bizonytalanság megfelelően propagálható.

1.4. A géncsaládok evolúciójának modelljei

A géncsaládok evolúciója komplex folyamat, így nehéz lenne egy olyan modellt konstruálni, ami ennek minden aspektusát figyelembe veszi. Ennek következményeként a géncsaládok evolúciójának leírására és vizsgálatára számos modell létezik, amelyek gyakran különböző folyamatokra és paraméterekre koncentrálnak.

Így például vannak modellek, amelyek a populációkban felbukkanó génduplikációk rögzülésének valószínűségére és a duplikációval keletkezett gének sorsára nézve próbálnak meg predikciókat tenni különböző tényezők függvényében, mint amilyenek például az effektív populációméret vagy a szelekció iránya és erőssége (áttekintve: Walsh, 2003; Innan, 2009). Ezen modellek közül egyesek a génkonverzió hatását is megpróbálják figyelembe venni. A gének funkciójában bekövetkezett változásokat magyarázni próbáló populációgenetikai modellek predikciói között azonban jelentős átfedések vannak, ami ezek magyarázó erejét csökkenti (Hahn, 2009).

Más modellek a géncsaládok méretének nagyobb időléptékben bekövetkezett változásaira

[4]Markov Chain Monte Carlo

koncentrálnak. Ezek felhasználásával megbecsülhetjük a méret változásának globális rátáját (Hahn és mtsai, 2005), vagy akár a duplikációk és deléciók rátáját az egyes leszármazási vonalakban (Csűrös és Miklós, 2009). A becsült rátákból következtethetünk a genomevolúció általános trendjeire, illetve az egyes géncsaládok méretének adaptív változásaira.

A géncsaládok evolúciójának vizsgálata és leírása során alkalmazott modellek közül nem mindegyik rendelkezik egységes és explicit matematikai megfogalmazással. A félreértések elkerülése végett az ilyen modelleket Hahn és mtsai (2005) nyomán „konceptuális modellek"-nek fogjuk nevezni. Ezek közé tartoznak például a vizsgálataink során központi szerepet játszó „összehangolt" („concerted") és „születés és halál"[5] (rövidítve „BD") modellek.

1.4.1. Az „összehangolt" és „születés és halál" modellek

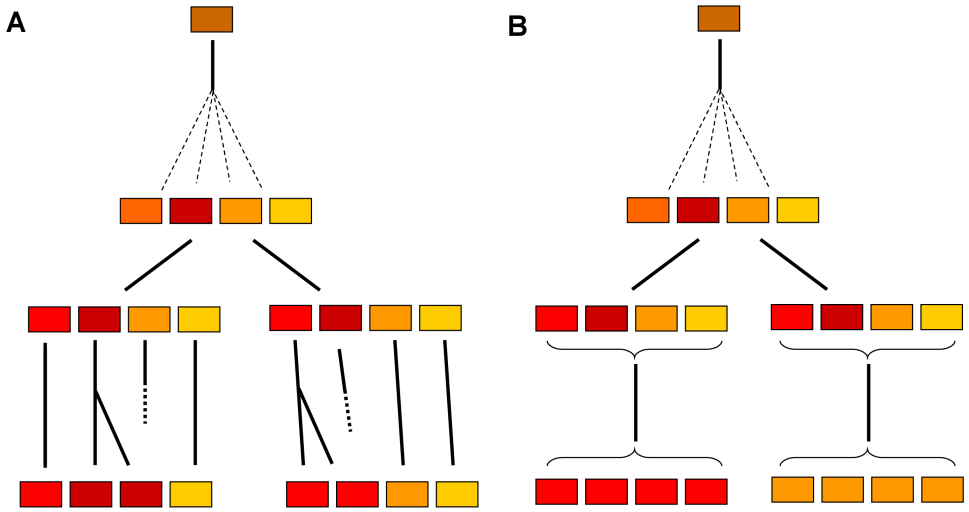
Az „összehangolt", „születés és halál" és divergens modellek a géncsaládok evolúciójának leírására használt legrégebbi konceptuális modellek (áttekintve: Nei és Rooney, 2005), amelyek még a géncsaládok tanulmányozásának „hőskorában" kerültek megfogalmazásra. Ezen modellek az egységek (a géncsaládok esetében gének) evolúciójának függetlenségére, illetve az ezt megsértő folyamatokra koncentrálnak.

A BD modell alatt alapvetően a gének független evolúcióját értjük, ami magába foglalja az új gének keletkezését duplikációs folyamatok révén, gének delécióját és a mutációk független halmozódását a különböző génekben (divergencia), ahogy azt az 1.7. **A** ábra szemlélteti. A mutációk halmozódásának következményeként a gének funkciójában változások következhetnek be (neo- és szubfunkcionalizáció), és a nonszensz, valamint a frameshift mutációk nyomán pszeudogének jöhetnek létre.

Az irodalomban számtalan példát találunk a BD modell szerint viselkedő géncsaládokra (például Ota és Nei, 1994; Nei és mtsai, 2000; Quesada és mtsai, 2005). Jelenleg úgy tartják, hogy a géncsaládok többségének viselkedését ez a konceptuális modell jól leírja és gyakran hangsúlyozzák ennek jelentőségét a genetikai rendszerek (pl. gerincesek adaptív immunrendszere, gerinces szaglóreceptorok) kialakulásában (Nei, 2005; Nei és Rooney, 2005).

A „divergens" modell a BD modellhez hasonló, de statikusabb viselkedést ír le, ami feltételezi a gének szekvenciája közötti divergenciát, de nem számol a géncsaládok méretének változásaival és a pszeudogenizációval, ezért a továbbiakban ezt a konceptuális modellt a BD modell egy speciális esetének tekintjük.

[5]Az angol nyelvű terminológiában „Birth-and-Death". Nem összetévesztendő az azonos nevet viselő sztochasztikus folyamattal.



1.7. ábra. A géncsaládok evolúciója a „Birth-and-Death” (**A**) és a „concerted” (**B**) modellek szerint. A különböző színű téglalapok különböző tulajdonságokkal bíró géneket jelképeznek.

A BD modellel ellentétben az összehangolt evolúciós modell szerint a géncsalád tagjainak evolúciója egymástól nem független, hanem a különböző génekben felbukkanó mutációk egy homogenizáló mechanizmus révén az egész géncsaládban elterjednek, és ilyen módon a géncsalád tagjai egy egységként viselkednek (1.7. **B** ábra). Fontos leszögezni, hogy az összehangolt evolúció egy jelentős ideig ható és intenzív homogenizációs folyamatot feltételez (Nei és mtsai, 2000).

Összehangolt evolúciót számos géncsalád esetében figyeltek meg a prokariótáktól az emlősökig (Liao, 1999). Az egyik legismertebb példa a tandem riboszomális RNS lókuszok evolúciója (Eickbush és Eickbush, 2007), aminek jellemzésére a modellt eredetileg megfogalmazták, és aminek nyomán egészen az 1990-es évekig úgy tartották, hogy a géncsaládok többségére érvényes (Nei és Rooney, 2005).

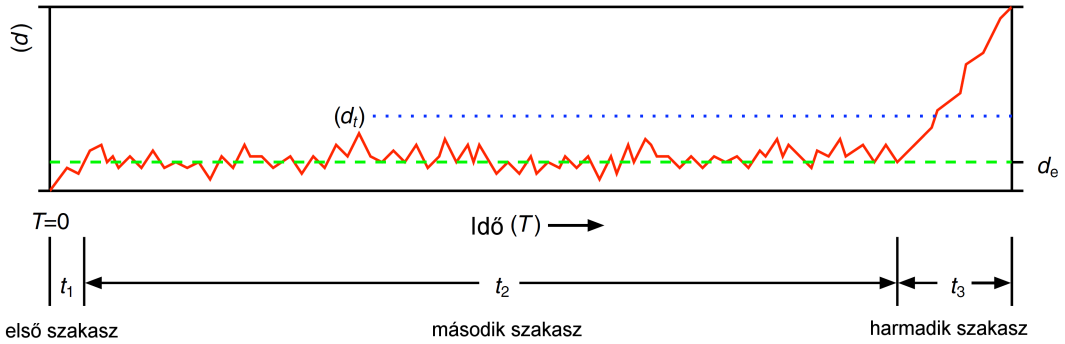
Az irodalomban a homogenizáció mechanizmusaként a leggyakrabban a génkonverziót (Chen és mtsai, 2007) és az egyenlőtlen rekombinációt (Hastings és mtsai, 2009) nevezik meg. Általában nem ismert, hogy a két mechanizmus közül melyik járul hozzá jelentősebb mértékben a homogenizációhoz, de úgy tartják, hogy a kisméretű géncsaládok (két gén) esetében a fő mechanizmus a génkonverzió, míg a közepes és nagyméretű géncsaládok esetében mindkettő szerepet játszik (Innan, 2009).

Az összehangolt és BD modellek túl egyszerűek ahhoz, hogy a géncsaládok evolúcióját külön-külön teljesen és minden időléptékben jellemezzék, így számos olyan példát találunk az irodalomban, amikor csak egyes gének vesznek részt a homogenizációban, illetve a homogenizáció epizodikus és a hatása csak bizonyos taxonokban figyelhető meg (Zhang és mtsai, 2003b; Das és mtsai, 2008). Ezek a megfigyelések jól összhangban vannak Teshima és Innan (2004) eredményeivel, amelyeket két gén génkonverzió által összehangolt evolúciójának

szimulációja során kaptak.

Az 1.8. ábrán a két szekvencia közötti divergencia (d) tipikus alakulása látható abban az esetben ha a génkonverzió gyakorisága jelentős. Teshima és Innan (2004) a divergencia alakulásában három szakaszt különböztet meg:

- Az első szakaszban a génduplikációval létrejött szekvenciapár közötti divergencia növekszik és elér egy egyensúlyi értéket (d_e), amit alapvetően a szubsztitúciós és a génkonverziós ráta viszonya határoz meg.
- A második szakaszban a divergencia az egyensúlyi érték körül ingadozik, ebben a fázisban zajlik a tulajdonképpeni összehangolt evolúció.
- Ha a divergencia átlépi azt a küszöbértéket, ami fölött a homogenizáció már nem lehetséges, a gének elindulnak a független evolúció útján és a divergencia többé-kevésbé lineárisan növekszik.



1.8. ábra. Két gén közötti divergencia jellemző alakulása abban az esetben ha a génkonverzió gyakorisága jelentős. d a divergenciát, d_e az egyensúlyi divergenciát, míg d_t azt a küszöbértéket jelöli, aminél nagyobb divergencia esetén a homogenizáció már nem lehetséges. Innan (2009) nyomán, módosítva.

Az összehangolt evolúciós epizódok (a második szakasz) hosszát a génkonverzió, szubsztitúció és a természetes szelekció összjátéka határozza meg (Innan, 2009).

Az első két tényező esetében a hatás iránya nyilvánvaló, nem úgy a szelekció esetében. Az új szubsztitúciók fixálódását elősegítő szelekciós hatások (pl. irányított szelekció) a divergencia növelése által az összehangolt epizódok időtartamát csökkentik, ezzel szemben az erős kényszerek alatt álló szekvenciákban a szubsztitúciók rögzülése ellen ható „negatív” szelekciós hatások általában növelik.

Ugyancsak elképzelhető, hogy a géntermékek nagyobb dózisát favorizáló szelekciós hatások az összehangolt epizódok időtartamát növelik azáltal, hogy előnyben részesítik az egyes kópiákban található előnytelen mutációkat „kijavító” konverziós események termékeit.

Ezt a tényezőt gyakran hangsúlyozzák az irodalomban a riboszomális RNS lókuszok esetében megfigyelt összehangolt evolúció tárgyalásakor (Liao, 1999). Sugino és Innan (2006) szignifikáns pozitív korrelációt talált a sörélesztő esetében az összehangolt epizódok időtartama és a génexpresszió intenzitása között, ami arra enged következtetni, hogy ez a tényező általában is fontos lehet.

1.4.2. Az ismétlődő szekvenciaelemek összehangolt evolúciója

A fehérjék jelentős része tartalmazza ugyanazon domén több példányát („domén ismétlődéseket", „domén repeateket"), illetve kisebb ismétlődő szekvenciaelemeket (Andrade és mtsai, 2001; Börklund és mtsai, 2006). A fentebb ismertetett konceptuális modellek ezen géneknél kisebb egységek evolúciójának jellemzésére is alkalmazhatók. Így az ezen ismétlődő elemek közötti hasonlóság mutációk felhalmozódása révén csökkenhet, mint például a HEAT típusú ismétlődéseket tartalmazó proteinek esetében (Andrade és mtsai, 2001), de ugyanúgy számos példát találunk az összehangolt evolúcióra (Swanson és Vacquier, 1998; Carmon és mtsai, 2007; Johannesson és mtsai, 2005).

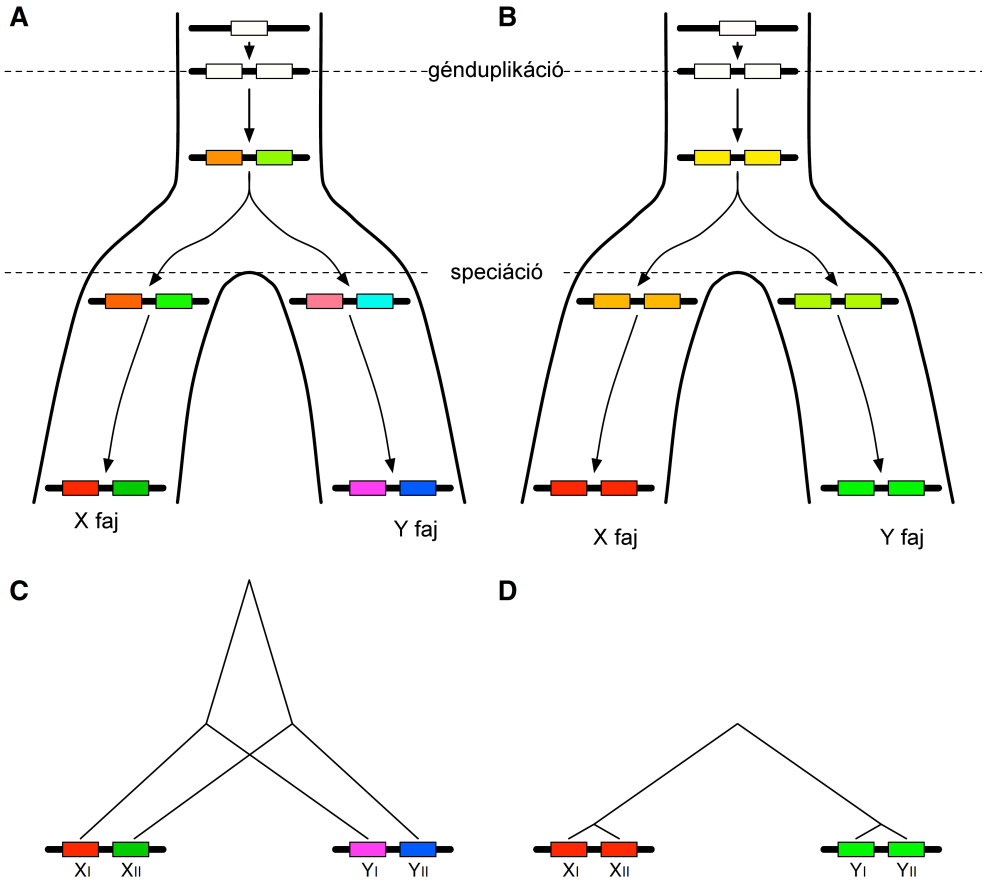
A géncsaládokhoz hasonlóan a domének összehangolt evolúciója is lehet epizodikus, és nem minden domén vesz részt szükségszerűen a homogenizációs folyamatban, ahogy azt jól példázzák Hughes (1999) az emlős Tenascin fehérjékben található EGF és fibronektin-III domének evolúciójával kapcsolatos eredményei.

Fontos megjegyezni, hogy a jelen dolgozatban a „domén" kifejezés alatt nem a harmadlagos fehérjeszerkezet (Orengo és Thornton, 2005), hanem a protein-evolúció egy egységét értjük. Ezen értelmezés szerint a domén nem más, mint potenciálisan több fehérjében is előforduló homológ szekvenciarészletek családja, ami megegyezik a Pfam adatbázis (Finn és mtsai, 2006) által használt definícióval. A fehérjék aminosav-szekvenciája és funkciója közötti összefüggés nehezen általánosítható, ennek ellenére a fenti módon definiált domének jelenlétére és az irodalmi adatokra támaszkodva használható hipotézisek állíthatók fel a funkcióval kapcsolatban, ami a szekvenciák automatikus annotációjának alapját képezi. Az ilyen következtetések összefüggést feltételeznek a szekvencia, és a funkciót végsősoron meghatározó struktúra között, ami világossá teszi a „domén" evolúciós és strukturális értelemben vett meghatározása közötti kapcsolatot.

1.4.3. Az összehangolt evolúció vizsgálata

Az összehangolt evolúció a paralógok közötti hasonlóság növelésének irányába hat és így drámai hatása lehet a szekvenciaadatokból rekonstruált filogenetikai fák topológiájára (Nei és Rooney, 2005).

Innan (2009) nyomán tekintsünk egy paralóg génpárt, ami egy, a speciációs esemény előtti duplikáció terméke. Ha a gének evolúciója egymástól független, a két szekvenciában a mutációk függetlenül halmozódnak, és így információt hordoznak a fajok evolúciós történetéről (1.9. **A** ábra). Ebben az esetben azt várjuk, hogy az egyazon lókuszokon található gének egy kládba csoportosuljanak és ez a klád összhangban legyen a fajok filogenetikájával (1.9. **C** ábra).



1.9. ábra. A szekvencia-evolúció illusztrációja egy kéttagú géncsalád esetében a homogenizáció hiánya (**A**) és jelenléte (**B**) esetén. A **C** és **D** ábrarészleteken a szekvenciaadatokból rekonstruált filogenetikai fák várható topológiái láthatóak a két említett esetben. A különböző színű téglalapok a szekvenciák közötti különbségeket jelképezik. Innan (2009) alapján, módosítva.

A homogenizációs folyamatok (1.9. **B** ábra) következményeként a fajok filogenetikájával inkongruens topológiákat kaphatunk, amelyeken a paralóg gének csoportosulnak egy kládba (1.9. **B** ábra). Ez az inkongruencia abban az esetben lép fel, ha a speciáció időpontjában a divergencia alakulása a második szakaszban van, így ennek valószínűsége a duplikációs és speciációs események között eltelt időtől és az összehangolt epizód időtartamától függ (Innan, 2009).

Az összehangolt és BD modellek tehát különböző predikciókat tesznek a szekvenciaadatokból rekonstruált fa és a fajok filogenetikájának inkongruenciájára nézve

(Nei és Rooney, 2005). Ezt úgy is megfogalmazhatjuk, hogy az összehangolt modellel ellentétben a BD modell esetében azt várjuk, hogy a pozicionális- és szekvencia-információk alapján felállított homológia-hipotézisek egymással összhangban vannak.

Az inkongruencia alapján azonban nem minden esetben lehetséges annak eldöntése, hogy melyik modell írja le jobban a vizsgált géncsalád evolúcióját, mivel nagyobb időléptékekben, a duplikációk és deléciók nyomán, a BD modell szerint viselkedő géncsaládok esetében is kaphatunk hasonló mintázatokat. Az összehangolt evolúció azonban definíció szerint egy intenzív homogenizációs folyamat (Nei és mtsai, 2000), aminek hatása még közeli rokon fajok esetében is érvényesül, ahol egyébként nem várunk nagyobb mértékű kópiaszám-változásokat. Ilyen esetekben tehát a szekvenciaadatokból rekonstruált filogenetikai fa és a fajok filogenetikájának inkongruenciáját a homogenizációra mutató evidenciaként kezelhetjük (Nei és Rooney, 2005; Innan, 2009), ami az összehangolt evolúció detektálásának leggyakoribb módját képezi.

Az összehangolt evolúció során a homogenizáció a nukleotid-szekvenciák szintjén történik, de sok esetben praktikusabb a kódolt aminosav-szekvenciák vizsgálata. Ezekben az esetekben előfordulhat, hogy nagyfokú hasonlóságot találunk bizonyos paralóg szekvenciák között, de ezt nem homogenizációs folyamatok, hanem egy nagyon erős negatív szelekciós kényszer tartja fenn. Ezekben az esetekben a kódoló nukleotid-szekvenciák közötti szinonim távolságok vizsgálatával a homogenizáció ténye megerősíthető (Nei és mtsai, 2000). A szinonim mutációkra a szelekció kevésbé hat, ezért ha a szinonim távolságok is homogenizációra utalnak, akkor a szelekció hatása kizárható.

Lehetséges továbbá egy fajon belül megfigyelhető polimorfizmusból következtetni az összehangolt evolúció jelenlétére. Ilyen célra használhatóak például a Sawyer (1989) által kifejlesztett statisztikai módszerek, amelyek a polimorf pozíciók eloszlására alapozva próbálnak következtetni a génkonverziós és rekombinációs eseményekre. Ezek azonban rosszul teljesítenek abban az esetben, ha a homogenizáció intenzív és a szekvenciák szinte azonosak (Innan, 2009).

1.4.4. Az összehangolt evolúció jelentősége

Az összehangolt evolúció a molekuláris evolúciós vizsgálatok egyik legrégebbi és máig is népszerű témája, amiről rendszerint minden átfogóbb tankönyv is említést tesz (Ridley, 1996; Futuyma, 2005). Az ilyen viselkedést mutató géncsaládok, valamint a jelenség elterjedtségének vizsgálata feltehetőleg azért maradt napjainkig népszerű, mert a gének közötti információcsere a géncsaládok evolúciójának szinte minden aspektusára jelentős, és nem minden esetben könnyen belátható hatással van. A teljesség igénye nélkül felhozunk néhány példát, ami az összehangolt evolúció jelentőségét szemlélteti:

- A duplikáció után ható génkonverzió jelentős hatással lehet a duplikáció rögzülésének valószínűségére és a gének további sorsára. (Innan, 2009).

- Mano és Innan (2008) egy kétlókuszos modell vizsgálata során arra a következtetésre jutott, hogy mivel a két lókusz között a mutációk cseréje zajlik, ami két azonos méretű dém közötti migrációhoz hasonlítható. Ebben az esetben az összehangolt evolúció az effektív populációméretet mintegy megduplázza, és ennek következményeként a természetes szelekció jobban érvényesül, vagyis az előnyös mutációk nagyobb eséllyel rögzülnek és a hátrányos mutációk hatékonyabban szelektálódnak ki. Ugyanezen oknál fogva a „majdnem semleges" mutációk rögzülésének gyakorisága csökken.

- Számos, a szekvenciaadatokra alapozó statisztikai módszer nem számol az összehangolt evolúció hatásával, így amennyiben az számottevő, az általuk szolgáltatott eredményekből hibás következtetéseket vonhatunk le, ahogy ez például a pozitív szelekciót detektáló módszerek esetében bebizonyosodott (Casola és Hahn, 2009).

- A génkonverzió nagyobb időléptékben érvényesülő homogenizáló hatása ellenére a populációk szintjén a haplotípus-diverzitást növeli (Innan, 2009). Egyes esetekben feltételezik, hogy ennek szerepe lehet a fajképződésben (Swanson és Vacquier, 1998), vagy a gazda-patogén kölcsönhatásokban (Johannesson és mtsai, 2005).

2. fejezet

Célkitűzések

A Nimrod szupergéncsalád tagjait a Kurucz és mtsai (2007) által prediktált doméntstruktúrákra alapozva besorolhatjuk ugyan főbb csoportokba (NimA/Draper, NimC, NimB), de a doménstruktúrák nem informatívak a szupercsalád részletes evolúciós történetére nézve. Vizsgálataink egyik fő célja tehát a szupercsalád evolúciójának, vagyis az egyes családok filogenetikájának, valamint családok közötti leszármazási kapcsolatoknak a felderítése volt, a molekuláris filogenetika és általában a bioinformatika eszköztárának bevetésével. A Nimrod szekvenciák filogenetikai analízise a doménismétlődések jelenléte és egyes paralógok közötti nagyfokú divergencia miatt fokozott körültekintést igényel. Mivel a szupercsalád egy kísérletesen aktívan kutatott géncsoport, remélhető, hogy a ráfordított figyelem a jövőben megtérül.

Ugyanezen módszerekre alapozva megpróbáltunk kidolgozni egy, a szupercsalád karakterisztikus doménstruktúrájának eredetét magyarázó hipotézist.

Vizsgálataink második fő célja a NimB és NimC szekvenciákban található ismétlődő NIM domének evolúciójának jellemzése volt, az összehangolt és a BD evolúciós modellek fényében.

Céljaink közé tartozott még a Nimrod klaszter konzerváltságának vizsgálata, az irodalomra és a nyilvánosan elérhető adatbázisokban található adatokra alapozva.

3. fejezet

Anyag és módszer

3.1. A szekvenciák összegyűjtése

A szupercsalád tagjait a következő fajok genomjaiban azonosítottuk: *Drosophila melanogaster*, *D. pseudoobscura*, *D. sechellia*, *D. yakuba*, *D. virilis*, *Anopheles gambiae*, *Tribolium castaneum* és *Apis mellifera* (3.1. ábra).

3.1. ábra. A vizsgált fajok gyökeres filogenetikája. A zöld háttér a Drosphila génuszban tartozó fajokat jelöli. Az ábra az iTOL webalkalmazás (Letunic és Bork, 2007) segítségével készült. A fa topológiája a Zdobnov és Bork (2007), valamint a Russo és mtsai (1995) által közölt fákon alapszik.

A géncsalád tagjai már előzőleg leírásra kerültek a *D. melanogaster* (Kurucz és mtsai, 2007), *T. castaneum* (Zou és mtsai, 2007), *A. mellifera* (Evans és mtsai, 2006) és az *A. gambiae* genomjaiban (Holt és mtsai, 2002).

A fennmaradó genomokban a szupergéncsalád tagjainak azonosítását a következőképpen végeztük: TblastN (Altschul és mtsai, 1990) keresést végeztünk a CCxGY motívummal és a rákövetkező NIM doménnel. Az így azonosított prediktált génmodelleket szükség esetén manuálisan javítottuk a kimaradt konzervált szekvenciarészek hozzáadásával és a hibásan összeolvasztott génmodellek szétválasztásával. A génmodelleket tovább finomítottuk a rendelkezésre álló EST és cDNS szekvenciákkal és az egyéb fajokban azonosított génmodellekkel való összehasonlítás által. Néhány esetben a szekvenciák az olvasási keretet eltoló (frameshift) mutációkat tartalmaztak, ezek a kromatogram-archívumokkal való összehasonlítás révén javításra kerültek. A *nimrod* gének szekvenciáinak összegyűjtését és a génmodellek manuális annotálását Dan Hultmark és Somogyi Kálmán végezte.

A génmodellek exonjait összeolvasztva és az így kapott kódoló szekvenciákat a standard genetikai kód segítségével lefordítva kaptuk a továbbiakban használt aminosav-szekvenciákat (azonosítóikat lásd az 1. függelékben). A dolgozatban a *nimrodA*, *nimrodB* és *nimrodC*

gének által kódolt aminosav-szekvenciákra mint NimA, NimB és NimC szekvenciákra fogunk hivatkozni. A doménstruktúrákra alapozva az *eater* gének által kódolt szekvenciákat a NimC szekvenciák közé soroljuk.

3.2. A NIM domének azonosítása profil Rejtett Markov Modellek segítségével

Mivel a NIM domén konszenzusa (Kurucz és mtsai, 2007) meglehetősen laza, ezért érzékenysége és elméleti megalapozottsága miatt a profil Rejtett Markov Modell (pHMM, Durbin és mtsai, 1998) módszert alkalmaztuk a *nimrod* gének által kódolt aminosav-szekvenciákban található NIM domének azonosítására.

A pHMM felépítését és a NIM domének azonosítását a HMMER 2.3.2 programcsomag (`http://hmmer.janelia.org`) segítségével végeztük. A modell betanítását a *D. melanogaster nimrod* és *eater* szekvenciákban található NIM domének szekvenciáival végeztük. Ezen szekvenciákban egy Perl szabályos kifejezés segítségével[1] azonosítottuk a CCxG motívum után található hat ciszteint tartalmazó szekvenciaegységeket.

A NIM konszenzushoz viszonyítva túl hosszú vagy túl rövid szekvenciákat kiszűrtük, majd a fennmaradó szekvenciákat a ProbCons 1.1 szoftver (Do és mtsai, 2005) segítségével illesztettük. A kapott illesztésből eltávolítottuk a GY motívum utáni pozíciókat. Ezen illesztés alapján, a `hmmbuild` program (alapértelmezett paraméterek) segítségével, betanítottunk egy pHMM-et, majd ezt a `hmmcalibrate` program segítségével kalibráltuk. A pHMM-et teljes NIM domének azonosítására kívántuk felhasználni, ezért választottuk az alapértelmezett `ls` modellarchitektúrát.

Az így kapott „előzetes” pHMM és a `hmmsearch` program segítségével (alapértelmezett paraméterekkel) azonosítottuk a NIM doméneket a teljes adathalmazban. A keresés által azonosított szekvenciákat egy BioPerl (Stajich és mtsai, 2002) modulokat felhasználó Perl szkript segítségével egy Fasta formátumú állományban összegyűjtöttük, majd az „előzetes” pHMM és a `hmmalign` program segítségével (alapértelmezett paraméterekkel) illesztettük őket. Az így kapott illesztéssel a fentebb ismertetett eljárás alkalmazásával egy új, „finomított” pHMM-et tanítottunk be és kalibráltunk.

Az így kapott pHMM-eket a Logomat-M szerver[2] által készített HMM logók (Schuster-Böckler és mtsai, 2004) segítségével ábrázoltuk. A „finomított” és az „előzetes” pHMM-ek logói nem különböztek számottevően, de az előbbi néhány szekvenciában több domént azonosított, ami nagyobb érzékenységre utalt. A továbbiakban minden esetben a „finomított” pHMM-et használtuk, kivéve a potenciálisan doménszintű összehangolt evolúciót mutató

[1] m/(?:C.*?){5}C(?:(.*?)(?=C)|)/gsx

[2] `http://www.sanger.ac.uk/cgi-bin/software/analysis/logomat-m.cgi`

szekvenciák többszörös illesztés előtti azonosítása során, amikor is az „előzetes” modellt használtuk.

A NIM domének azonosítását a `hmmsearch` program segítségével végeztük. A legtöbb esetben az alapértelmezett beállításokat és „cutoff” értékeket használtuk, amennyiben nem, azt jelezzük.

3.3. Profil–profil összehasonlítások

A fentebb említett pHMM módszereket használhatjuk két szekvenciacsalád közötti homológia vizsgálatára azáltal, hogy az egyik család modelljét egy, a másik családba tartozó szekvenciával illesztjük. Ilyen esetekben azonban a profil–szekvencia összehasonlításra alapozó eljárásnál sokkal jobban teljesítenek a profil–profil összehasonlítások (Söding, 2005; Madera, 2008), melyek esetében az illesztés a szekvenciacsaládokat valamilyen módon modellező „profilok” szintjén történik.

A PRC szoftver (Madera, 2005) többféle algoritmust is implementál a családokat modellező pHMM-ek különböző illesztéseinek (lokális–lokális, lokális–globális, stb.) kiszámolására, de mivel a PRC „nyers” kimenete nehezen értelmezhető, ezért a pHMM-ek összehasonlítására a PRC által számolt lokális-lokális illesztéseket páros HMM logók[3] által ábrázoló Logomat-P szervert[4] használtuk (Schuster-Böckler és Bateman, 2005).

A Logomat-P szerver által használt PRC paraméterek a következők:

```
-align prc -hits 1 -algo viterbi -mode local-local
-MMfn dot2 model1.hmm model2.hmm
```

3.4. A szinonim és nemszinonim távolságok becslése

A vizsgált *nimrod* gének által kódolt fehérjék aminosav-szekvenciáiban található NIM doméneket az előzőekben betanított profil HMM segítségével azonosítottuk és illesztettük. Az így kapott illesztéseket a találatok pozíciója és a kódoló szekvencia alapján kodonszekvencia-illesztéssé alakítottuk át egy Perl szkript segítségével, ilyen módon elkerülvén a gapek felbukkanását a kodonok belsejében. Az átalakítást úgy ellenőriztük, hogy a kódoló szekvenciákban elfoglalt pozíciójuk alapján sorbaállított NIM doméneket kódoló szekvenciák összefűzésével képzett szekvenciákat a megfelelő kódoló szekvenciákkal páronként illesztettük. Ez az összekötő szekvenciák DNS-szinten való azonosítását is lehetővé tette.

[3] pairwise HMM logos

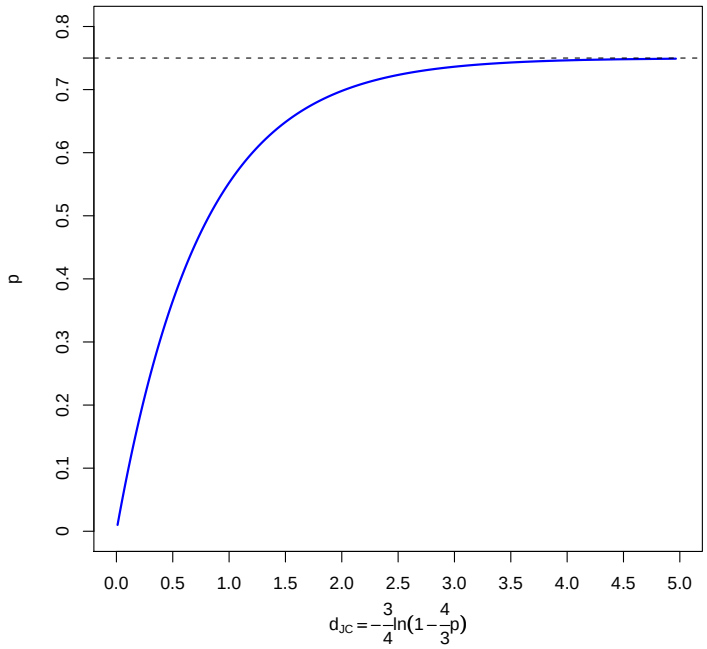
[4] http://www.sanger.ac.uk/Software/analysis/logomat-p

Az egy adott szekvenciában található NIM doméneket kódoló szekvenciák közötti szinonim és nemszinonim távolságok módosított Nei-Gojobori módszerrel (Nei és Gojobori, 1986; Zhang és mtsai, 1998) való kiszámolására a MEGA 3.1 (Kumar és mtsai, 2004) szoftvert használtuk. A számolás során az illesztésekből a PAUP* 4.0b10 (Swofford, 2003) és a Modeltest 3.7 (Posada és Crandall, 1998) által számolt, az Akaike súlyok alapján modellátlagolt tranziciós-transzverziós arányt használtuk. A kapott távolsági mátrixokat egy Perl szkript és a PGF/TikZ LaTeX csomag segítségével ábrázoltuk.

3.5. A szaturációs ábrák

Az, hogy egy illesztés szekvenciái közötti távolság milyen mértékben közelíti meg a szaturáció szintjét, vizuálisan felmérhető a Philippe és mtsai (1994) által javasolt szaturációs ábrák által.

A szaturációs ábrákon az illesztés szekvenciapárjai között megfigyelt különbségeket ábrázoljuk egy szubsztitúciós modell által számolt távolság függvényében. Kis távolságoknál a különbséget mutató pozíciók aránya (p) megközelítően lineárisan növekszik, majd a távolság növekedésével ez egy adott értékhez tart és tovább nem növekszik (telítődik, szaturálódik).



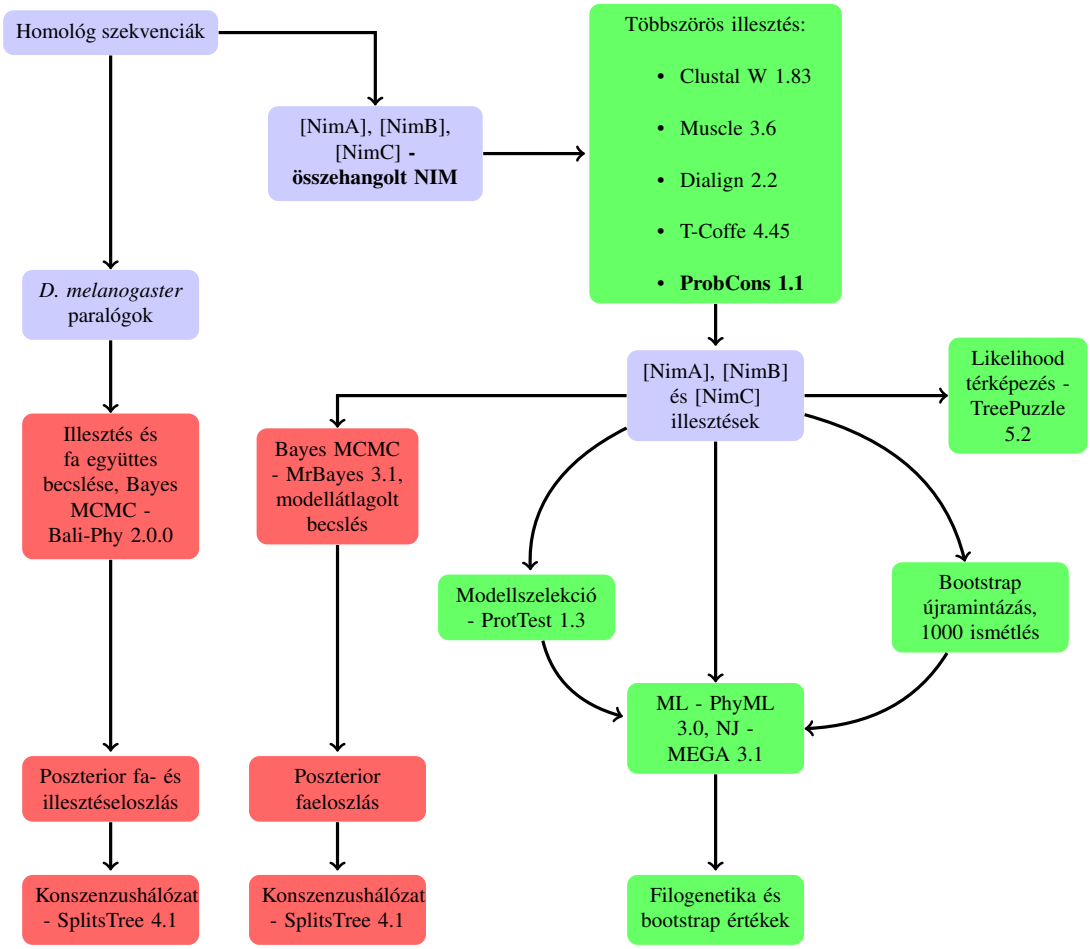
3.2. ábra. A becsült távolság (d) és a különbséget mutató pozíciók aránya (p) közötti összefüggés a Jukes-Cantor szubsztitúciós modell esetében. Az ábra az R 2.8.1 szoftver segítségével készült.

A tranziciós és transzverziós mutációk rátája és így a szaturáció foka is jelentősen különbözhet. Az ezen mutációs típusokra felrajzolt szaturációs ábrák összehasonlításával ez a különbség felmérhető.

A Drosophila NimC és Eater szekvenciákban található NIM doméneket kódoló szekvenciarészek génenkénti illesztéseiből a szaturációs ábrákat a DAMBE 4.2.13 szoftver (Xia és Xie, 2001) segítségével készítettük el. A görbék elkészítése során a tranziciókat és transzverziókat külön kezelő F84 szubsztitúciós modellt (Felsenstein, 2002) használtuk.

3.6. Filogenetikai rekonstrukció

A *nimrod* gének által kódolt aminosav-szekvenciák filogenetikai analízise során követett munkamenetet a 3.3. ábrán látható folyamatábra szemlélteti.

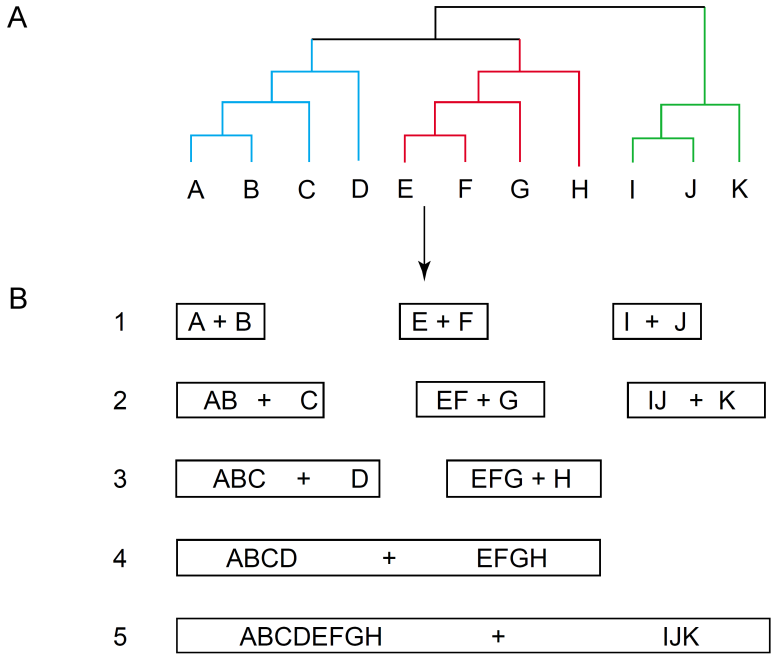


3.3. ábra. A Nimrod szekvenciák filogenetikai analízise során követett munkament. A Holder és Lewis (2003) felosztása szerint „tradicionálisnak” tekinthető módszereket zöld színnel, míg a Bayes statisztikai alapokon nyugvó módszereket piros színnel jelöltük.

3.6.1. Az aminosav-szekvenciák többszörös illesztése

A gapek egy adott súlyozási sémájára és egy szubsztitúciós modellre nézve optimális illesztés multidimenzionális dinamikus programozás segítségével kiszámolható (Durbin és mtsai, 1998), azonban ez nagy erőforrásigénye miatt a gyakorlatban csak kevés számú szekvencia illesztésére alkalmazható. Ebből kifolyólag számos olyan „heurisztikus" algoritmus létezik, ami megad ugyan egy illesztést, de nem garantálja annak optimális voltát (Thompson és mtsai, 1999).

A legtöbb heurisztikus módszer az úgynevezett „progresszív profil" stratégiát használja az illesztés során (3.4. ábra). Ennek első lépéseként az illesztetlen szekvenciákból valamilyen eljárás segítségével, például a páronkénti illesztésekből kiindulva, felépítünk egy „vezérfát"[5] (3.4. **A** ábra). A második lépésben a vezérfa mentén, a levelektől a gyökér felé haladva, minden belső csomópontban páronként illesztjük a két leszármazott kládban található szekvenciák már illesztett csoportjait („profilokat"), ahogy ez a 3.4. ábrán látható.



3.4. ábra. A többszörös szekvencia-illesztés progresszív stratégiájának illusztrációja. Az **A** ábrarészleten a vezérfa, míg a **B** ábrarészleten a páronkénti profil-profil illesztések általa meghatározott sorrendje látható. Baldauf (2003) alapján, módosítva.

A progresszív illesztés „mohó" stratégiát követ, ugyanis a már egyszer illesztett szekvenciák csoportjait egy egységként kezeli és egy megfelelően módosított súlyozási séma segítségével páronként illeszti őket, ezeken belül a pozíciók összerendezése tehát az illesztés

[5] „guide tree"

végéig nem változik. A progresszív algoritmusok esetében nincs tehát lehetőség az illesztés elején vétett „hibák” kijavítására.

Ha az illesztés során használt vezérfa topológiája nagyban eltér a szekvenciák „valódi” filogenetikájától, az jelentős torzításokat okozhat („guide tree bias”), amelyek a számolt illesztésre alapozott további vizsgálatok eredményére is kihat (Redelings és Suchard, 2005; Kumar és Filipski, 2007). Ennek a veszélye általában fennáll, ugyanis a szekvencia-analízis ezen fázisában még nyilvánvalóan nem áll rendelkezésünkre egy többszörös illesztés, amiből a filogenetikát megbecsüljük, a vezérfát kénytelenek vagyunk durva módszerekkel megbecsülni. Ezt a „tyúk-tojás” problémát a filogenetika és az illesztés együttes becslése révén kiküszöbölhetjük (lásd az 1.3. részben).

Az aminosav-szekvenciák heurisztikus illesztésére használható számos módszer által számolt illesztések, és így az azokból a további analízisek során levont következtetések jelentősen különbözhetnek (Kumar és Filipski, 2007; Wong és mtsai, 2008). Ezen okból kifolyólag a Nimrod szekvenciák vizsgálata során a nagyfokú divergenciára, és az illesztést megnehezítő repetitív elemek jelenlétére való tekintettel (Higgins, 2003) több népszerű, és többé-kevésbé különböző eljárásokat implementáló szoftvert is felhasználtunk (alapértelmezett paraméterekkel):

- *Clustal W 1.83* (Thompson és mtsai, 1994) - progresszív illesztő, a gapek súlyozási sémájának számos ötletes módosítása által próbál javítani az illesztés minőségén.
- *Muscle* 3.6 (Edgar, 2004) - progresszív iteratív illesztő, a kapott illesztésből egy új vezérfát számol majd újra elvégzi az illesztést. Ezt többször is megismétli.
- *T-Coffee* 4.45 (Notredame és mtsai, 2000) - progresszív illesztő, amely a „konzisztencia” elvére alapozva megpróbálja felépíteni a páronkénti lokális és globális illesztések által támogatott legtöbb pozíciót tartalmazó többszörös illesztést.
- *ProbCons* 1.1 (Do és mtsai, 2005) - progresszív, a konzisztencia elvét és páros HMM-eket használó módszer.
- *Dialign 2.2* (Morgenstern, 1999) - jól illeszkedő szegmenseket a konzisztencia elvére alapozva többszörös illesztéssé rendező módszer. Ebben az esetben az illesztéseket egy, a NIM konszenzus lazább változatával[6] leírható szegmensek illeszkedésére alkalmazott „bónusz” jelenlétében is elvégeztük.

A fentebb tárgyalt okok miatt nyilvánvalóan nem mindegy, hogy melyik heurisztikus módszerrel illesztjük a szekvenciákat, nem egyszerű azonban eldönteni, hogy az adott esetben melyik által szolgáltatott illesztés tekinthető biológiai szempontból a legrelevánsabbnak.

[6] „CxPxCxxxCxxGxCxxPxxCxCxxGx”

Az illesztések minőségének felmérése során a vizuális kiértékelésen kívül a következő kritériumokat is használtuk:

- A CCxGY/W motívumok és a NIM domének illeszkedése, amit egy, az illesztésekben az említett elemeket vizualizáló Perl szkript segítségével vizsgáltunk. A szkript a NIM doméneket az előzőekben betanított pHMM segítségével azonosítja. A szkript csak saját használatra készült, ezért implementációját nem részletezzük.

  Az így felfedezhető különbségeket a NimC szekvenciák Clustal W és ProbCons illesztéseit vizualizáló ábrákkal szemléltetjük, amelyek a 2. függelékben láthatók. Ezeket összehasonlítva például szembetűnik, hogy a Clustal W illesztésben a nyilvánvalóan homológ CCxGY/W motívumok egymással nem mindig, és a NIM domének néhol az N-terminális régióval illeszkednek, és így az illesztés biológiai relevanciája kétségbe vonható.

- A T-Coffee által számolt konzisztencia-értékek. Ezek használata bizonyos mértékben körbenforgó, ugyanis a konzisztencia alapú módszerek pontosan ennek az optimalizációjára törekednek, de mivel jól kifejezik, hogy a többszörös illesztés mennyire felel meg a lokális és globális illesztési kritériumoknak, jobbnak láttuk ezeket is figyelembe venni.

- A Neighbor-Joining algoritmussal (lásd később a 3.6.6. részben) felépített „előzetes” fák segítségével felmértük az illesztések topológiára gyakorolt hatását.

Az összehangolt evolúciós epizódokat megélt doméneket tartalmazó szekvenciák többszörös illesztésének lehet ugyan strukturális relevanciája, de a homogenizáció miatt nem várhatjuk, hogy ezek az evolúciós történetre nézve informatívak legyenek. Ezért a többszörös illesztésekből, és így a további filogenetikai analízisekből is, kizártunk minden olyan szekvenciát, amelyek a becsült doménfilogenetikák alapján összehangoltan evolválódó doméneket tartalmazhatnak, így a *T. castaneum* NimCI és NimCII, az *A. mellifera* NimCI és a Drosophila fajok Eater szekvenciáit.

A Nimrod szekvenciák analízise során megkíséreltük a fenti módszerekkel a szekvenciákat a szupercsalád szintjén illeszteni, az összes szekvencia bevonásával. Mivel a kritériumaink szerint rossznak minősülő illesztéseket kaptunk, és bizonyos esetekben a vezérfa használata miatt fellépő torzításokra utaló jeleket figyeltünk meg, ezeket az illesztéseket a további analízisek során nem használtuk.

Mind a NimA, mind a NimB és NimC szekvenciák esetében a ProbCons által számolt illesztéseket ítéltük legjobbnak és ezeket használtuk a további analízisek során.

3.6.2. Modellszelekció

A modellekre alapozó filogenetikai rekonstrukciós módszerek (pl. Maximum Likelihood) esetében csak akkor várható, hogy ezek jó tulajdonságai érvényesüljenek, ha a használt modell nem sérül lényegesen, ezért ajánlott az adatokra jól illeszkedő modellek használata (Yang és mtsai, 1995).

A modell kiválasztásában segítségünkre vannak az úgynevezett „modellszelekciós módszerek" (áttekintve: Posada és Buckley, 2004), melyek közül a legnépszerűbbek az úgynevezett „információs kritériumok", mint az Akaike kritérium (AIC[7]) és a Bayes kritérium (BIC[8], Reiczigel és mtsai, 2007). Az információs kritériumok segítségével lehetőség van egyszerre több, nem szükségszerűen egymásba ágyazott modell illeszkedésének összehasonlítására.

A nukleotid-szekvenciákból történő filogenetikai rekonstrukció során a modellszelekciós eljárásokat rutinszerűen alkalmazzák. Egyre több eredmény utal arra, hogy az aminosav-szubsztitúciós modellek esetében sem indokolható ennek elhanyagolása (Keane és mtsai, 2006), ezért vizsgálataink során ezek alkalmazása mellett döntöttünk.

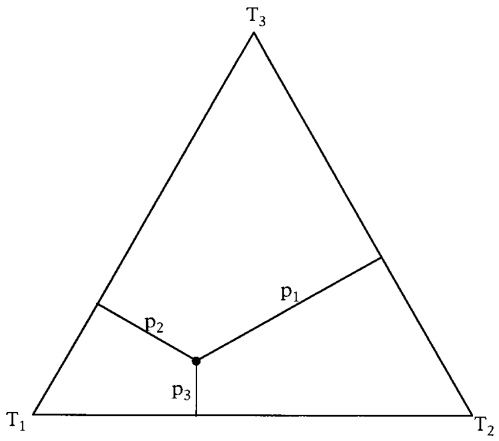
A modellszelekciót a legtöbb empirikus aminosav-szubsztitúciós modellt kezelő ProtTest 1.3 szoftverrel (Abascal és mtsai, 2005) végeztük, illetve ezen szoftver online futtatható változatával[9]. Az illesztésekre legjobban illeszkedő modellek kiválasztására a BIC kritériumot használtuk. Mintanagyságként az illesztések hosszát adtuk meg. A filogenetikai analízisek során az adott szoftverben implementált modellek közül mindig a legnagyobb BIC értékkel rendelkezőt használtuk.

3.6.3. A filogenetikai szignál tesztelése: likelihood térképezés

A „filogenetikai szignál", vagyis az, hogy egy többszörös illesztés mennyire informatív a topológiára nézve, felmérhető többek között a „likelihood térképezés"[10]-nek nevezett módszerrel (Strimmer és von Haeseler, 1997).

A likelihood térképezés során az illesztésben található szekvenciákból négyes csoportokat („kvartetteket") mintázunk. A kvartettek esetében mindhárom lehetséges gyökértelen topológia poszterior valószínűsége pontosan kiszámolható. A poszterior valószínűségeket tartalmazó vektorokat egy háromszögű koordinátarendszerben ábrázoljuk (3.5. ábra).

[7] „Akaike Information Criterion"
[8] „Bayesian Information Criterion"
[9] http://darwin.uvigo.es/software/prottest_server.html
[10] „likelihood mapping"



3.5. ábra. Egy kvartett esetében lehetséges három alternatív gyökértelen topológia poszterior valószínűségeit tartalmazó vektor ábrázolása egy háromszögű koordinátarendszer segítségével. T_1, T_2 és T_3 az alternatív topológiák, míg p_1,p_2 és p_3 a nekik megfelelő poszterior valószínűségek. Strimmer és von Haeseler (1997) alapján, módosítva

A pontok eloszlásából következtethetünk a filogenetikai szignál mértékére: ha ezek többsége a koordinátarendszer szögleteiben helyezkedik el, akkor kvartettek többségének esetében az adatok egy bizonyos topológiát támogatnak, ami megfelelő filogenetikai szignálra utal.

A ProbCons illesztések likelihood-térképezését a TreePuzzle 5.2 szoftverrel végeztük (Strimmer és von Haeseler, 1997) a legjobban illeszkedő szubsztitúciós modell felhasználásával. A NimB és NimC illesztések esetében az analízis 100 000 random kvartett mintázásával, míg a NimA illesztés estében az összes lehetséges kvartett (szám szerint 70) figyelembevételével történt.

3.6.4. Maximum Likelihood filogenetikai rekonstrukció

A Maximum Likelihood rekonstrukciókat a PhyML 3.0 (Guindon és Gascuel, 2003) szoftver segítségével végeztük az előzetesen kiválasztott legjobban illeszkedő modellek felhasználásával.

A pozíciók közötti rátaváltozatosság modellezésére egy négy rátakategóriával rendelkező diszkretizált gamma eloszlást használtunk a ProtTest által becsült megfelelő paraméterekkel. A BioNJ algoritmus által generált kiindulási topológiákat és az NNI keresési stratégiát használtuk. A kládok támogatottságát nemparaméteres bootstrap segítségével becsültük meg (1000 replikáció).

3.6.5. Filogenetikai rekonstrukció Bayes statisztikai módszerekkel

A rögzített ProbCons illesztésekből a Bayes MCMC filogenetikai rekonstrukciókat a

MrBayes 3.1 szoftver (Huelsenbeck és Ronquist, 2001) segítségével végeztük.

A szekvencia-illesztésekben a gapek inszerciós és deléciós eseményeknek felelnek meg, amelyek informatívak lehetnek a topológiára nézve. Hogy képet alkothassunk arról, hogy a gapek mennyire informatívak, valamint arról, hogy felhasználásuk milyen hatással van a becsült topológiára, ezeket Simmons és Ochoterena (2000) egyszerű kódolási eljárásával bináris karakterekké alakítottuk egy Perl szkript segítségével. Ezeket a bináris karaktereket restrikciós karakterekként kezeltük a MrBayes wiki[11] ajánlásának megfelelő módon.

Minden illesztés esetében három analízist végeztünk. Ezek közül az elsőben csak szekvenciaadatokat, a másodikban csak a gapek alapján bekódolt bináris karaktereket, míg a harmadikban mindkét információt felhasználtuk.

A futtatások során alapértelmezett paramétereket és *a priori* eloszlásokat használtunk, egy fontos kivétellel. A MrBayes ugyanis megengedi, hogy a rögzített rátájú aminosav-szubsztitúciós mátrixokra nézve egy egyenletes *a priori* eloszlást értelmezzünk (`prset aamodelpr=mixed`) és a modellszelekciót a filogenetika becslésével összekapcsoljuk („model jumping”), ami révén modellátlagolt becslésekhez jutunk.

A rátaváltozatosság modellezésére használt gamma eloszlás paraméterének ProtTest által szolgáltatott modellátlagolt értéke minden esetben megközelítette vagy meghaladta az 1-et (NimA: 1.22, NimB: 2.29, NimC: 0.98). Mivel az ilyen értékeknél a rátaváltozatosság nem tekinthető számottevőnek (Yang és mtsai, 1995), úgy gondoltuk, hogy a rátaváltozatosság elhanyagolása nem jár jelentősebb kockázattal és így éltünk a fentebb részletezett lehetőséggel.

A MrBayes futtatások egy részét egy Linux klaszteren, párhuzamosítva végeztük (Altekar és mtsai, 2004). A futtatások során használt paramétereket tartalmazó nexus blokkok a 3. függelékben találhatók.

A Nimrod szupergéncsalád *D. melanogaster*-ben található tagjai által kódolt aminosav-szekvenciákból az illesztés és a filogenetika együttes becslését a Bali-Phy 2.0.0 (Suchard és Redelings, 2006) segítségével végeztük. A felhasznált szekvenciák viszonylag kis mennyiségét a módszer erőforrásigénye indokolta. A Bali-Phy futtatásokat a WAG+Γ szubsztitúciós modell (4 gamma kategória) és az alapértelmezett „fragment-based indels +T” modellel végeztük. Három független láncot futtatunk 30000 generáción keresztül illesztetlen szekvenciákból kiindulva. Az MCMC által szolgáltatott minták közül az első tízezret figyelmen kívül hagytuk.

Mindkét szoftverrel végzett analízisek során a konvergencia diagnosztikáját a likelihood és a splitgyakoriságok alakulása, valamint Gelman és Rubin (1992) által javasolt PSRF (Potential Scale Reduction Factor) kritériumot használtuk, ami a független láncokon belüli és a láncok közötti variancia összehasonlításán alapszik. A Bali-Phy futtatások esetében a PSRF értékeket a 0.13-4 R csomag (Plummer és mtsai, 2009) segítségével számoltuk ki és a splitgyakoriságok

[11] http://mrbayes.csit.fsu.edu/wiki

kiértékeléséhez igénybe vettük az AWTY szoftvert (Nylander és mtsai, 2008).

A poszterior minták kiértékelése során a többségi konszenzusfákat[12] a megfelelő szoftverekkel, míg a konszenzushálózatokat a SplitsTree 4.10 (Huson és Bryant, 2006) segítségével építettük fel. A konszenzushálózatok a poszterior mintában egy bizonyos küszöbértéknél (x) gyakoribb spliteket jelenítik meg, így ezek segítségével sokkal könnyebben képet alkothatunk a poszterior mintában található alternatív topológiákról.

3.6.6. Neighbor-Joining filogenetikai rekonstrukció

A Neighbor-Joining (a továbbiakban „NJ", Saitou és Nei, 1987; Studier és Keppler, 1988) filogenetikai rekonstrukciót a MEGA 3.1 (Kumar és mtsai, 2004) szoftver segítségével végeztük. Az ML rekonstrukcióhoz hasonlóan a szoftverben implementált legnagyobb BIC értékkel rendelkező szubsztitúciós modellt használtuk. A pozíciók közötti rátaváltozatosságot gamma eloszlással modellező ($+\Gamma$) modellek estében a modellszelekció során a ProtTest 1.3 által becsült modellátlagolt gamma paramétert használtuk.

A páronkénti távolságok számolása során a szoftver a gapeket két stratégia szerint képes kezelni:

- az akár egy gapet is tartalmazó pozícióknak a teljes illesztésből való eltávolításával („complete deletion" - CD)
- az éppen aktuális szekvenciapár esetében a gapeket tartalmazó pozíciók eltávolításával („pairwise deletion" - PD).

A rekonstrukciót minden esetben mindkét stratégiával elvégeztük. A kládok támogatottságát nemparaméteres bootstrap segítségével becsültük meg, 1000 ismétlést végezve.

3.6.7. A génfák gyökereztetése

A géncsaládok filogenetikájának rekonstrukciója során az illesztési problémák miatt nem álltak a rendelkezésünkre megfelelő külcsoportok, ezért megkíséreltünk a gyökér pozíciójára következtetni a fajok gyökeres filogenetikájával (3.1. ábra) való összeegyeztetés által, a Notung 2.1 (Chen és mtsai, 2000) szoftver segítségével (alapértelmezett paraméterekkel).

Ezen analízisek azonban a gyökeret mindig az *A. mellifera*-ból származó szekvenciákhoz vezető ágra helyezték. Mivel ez egyenértékű ezen szekvenciák külcsoportként való felhasználásával, az eredmények tárgyalása során a fákat gyökértelennek ábrázoljuk és feltételezzük, hogy a gyökér az említett ágakon helyezkedik el.

[12] majority rule consensus

3.7. A klaszterizáció statisztikai szignifikanciájának vizsgálata

A genomok struktúrájának alakításában ugyanúgy részt vesznek a semleges („neutrális”), mint a szelekciós tényezők. Nagyobb időléptékekben a genomokat érintő, és túlnyomórészt neutrálisnak tekinthető átrendeződések következményeként a génsorrend gyakorlatilag randomizálódik, és csak kevés régió őrzi meg a génösszetételét. Ezek esetében ésszerű feltételezni, hogy a konzerváltságot a természetes szelekció tartja fenn (Hurst és mtsai, 2004; Koonin, 2009b).

Tegyük fel, hogy két olyan genomot hasonlítunk össze, amelyek elég régen „szétváltak” ahhoz, hogy bennük a génsorrend egymáshoz viszonyított randomizálódásához elegendő idő álljon rendelkezésre. Hogy az összehasonlítás során talált, egy bizonyos konzervált génösszetétellel rendelkező régió nagysága (pl. a gének számában kifejezve) mennyiben tekinthető rendkívülinek, az nyilvánvalóan függ a genomok méretétől, ugyanis a kisebb genomokban a randomizálódás során a gének nagyobb eséllyel kerülnek egymás mellé.

Az egyes géncsaládokon belül a pontos homológia-viszonyok nem mindig állapíthatók meg könnyen, melynek következményeként az összehasonlító genomikai vizsgálatok során sokszor a géncsaládokat tekintik egységnek, ami bonyolítja a randomizációt feltételező „null modellt”. Ugyanúgy hatással lehet a genom struktúrája egy random semleges modell alapján várt klasztermeretek eloszlására. Szükség van tehát olyan statisztikai eljárásokra, melyeknek segítségével eldönthető, hogy egy klaszter konzerváltságának tárgyalása során a null modellt, vagy egy másik, mondjuk szelekciós hatást feltételező alternatív magyarázatot tekintsünk relevánsnak.

A Nimrod klaszter statisztikai szignifikanciájának felmérésére Raghupathy és Durand (2009) közelítő kombinatorikai módszerét használtuk, ami a genomokban található gének számát és a géncsaládok jelenlétét is figyelembe veszi. A módszer a szerzők szimuláción alapuló vizsgálatai során konzervatívnak bizonyult.

A klaszter vizsgálata során az adatok begyűjtéséhez publikus adatbázisokat (FlyBase[13], Vectorbase[14], Wfleabase[15], NCBI Blast[16] keresések), valamint a rendelkezésre álló irodalmat vettük igénybe. Néhány esetben szükség volt a génmodellek manuális javítására, illetve ezek *de novo* predikciójára. A klaszter vizsgálata során felhasznált adatok és azok forrásai a 11. függelékben találhatók.

A különböző fajokban található klaszterek konzerváltságát a *D. melanogaster*-ben található

[13] http://flybase.org
[14] http://vectorbase.org
[15] http://wfleabase.org
[16] http://blast.ncbi.nlm.nih.gov/Blast.cgi

klaszterrel párba állítva vizsgáltuk, a számolt p-értékek tehát az ecetmuslica klaszterral közös géncsaládok számára vonatkoznak. A p-értékeket a módszert leíró cikk mellékleteként közzétett Mathematica függvények segítségével számoltuk ki. A géncsaládok méretét kifejező paraméter (f) értékeként 2-t adtunk meg, ahogy azt Raghupathy és Durand (2009) a 25000-nél kevesebb gént tartalmazó genomok esetében ajánlja. A többszörös összehasonlításokból adódó problémák elkerülése véget a kapott p-értékeket Bonferroni-korrekciónak (Reiczigel és mtsai, 2007) vetettük alá a `p.adjust` R (2.8.1) függvény segítségével.

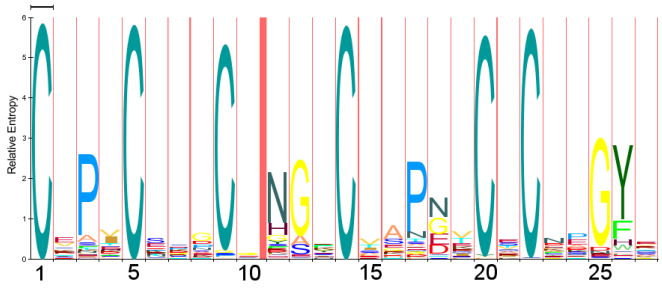
4. fejezet

Eredmények

Az eredmények bemutatása során a fajok neveinek következő rövidítéseit használjuk: Dmel - *Drosophila melanogaster*, Dpse - *D. pseudoobscura*, Dsec - *D. sechellia*, Dyak - *D. yakuba*, Dvir - *D. virilis*, Agam - *Anopheles gambiae*, Amel - *Apis mellifera*.

4.1. A NIM és az EGF domének közötti hasonlóság

Az általunk betanított, a NIM domént modellező pHMM logója (4.1. ábra) összhangban van a domén korábban leírt konszenzus-szekvenciájával, de az egyes pozíciók aminosav-eloszlását pontosabban jellemzi.



4.1. ábra. A NIM domént modellező pHMM logója. A bal felső sarokban látható vonal egy aminosavnyi hozzájárulásnak felel meg. Az oszlopok alatti számok a pHMM-ben elfoglalt pozíciót jelölik.

A Pfam domén-adatbázis több különböző, de egymáshoz hasonló EGF-szerű domént is tartalmaz, melyeket a CL0001 jelzésű „klánba" csoportosítanak. Ezen domének modellezése körülményes, mert a „klán" tagjai egymáshoz számottevően hasonlítanak. Az „EGF-szerű"[1] domén (Pfam:PF00008), ami hat konzervált pozíciójában tartalmaz ciszteint, meglehetősen hasonlít a „Laminin EGF-szerű" doménhez (Pfam:PF00053), ami nyolc konzervált ciszteint tartalmaz.

A korábbi elemzések (Kurucz és mtsai, 2007) a Draper-típusú doménarchitektúrával rendelkező proteinekben a NIM domén után nyolc konzervált ciszteint tartalmazó EGF-szerű doméneket azonosítottak (1.1. ábra). Ha azonban a `hmmpfam` segítségével az EGF klánba tartozó (`ls` architektúrájú) modelleket illesztjük a Draper típusú szekvenciákkal, akkor a legtöbb esetben a hat ciszteint tartalmazó domének (EGF2 - Pfam:PF07974 és EGF-szerű) rendelkeznek a teljes szekvenciára nézve legnagyobb „bit score"-ral.

Már Kurucz és mtsai (2007) is észrevették, hogy a NIM domén konszenzusa a „tipikus" EGF konszenzushoz képest „el van tolódva", de ezt részletesen nem vizsgálták. Mivel a fenti eredmények hat konzervált ciszteint tartalmazó EGF domének jelenlétét valószínűsítik a Draper-típusú fehérjékben, ezért a „tipikusabb" EGF konszenzussal jobban összhangban lévő

[1] „EGF-like"

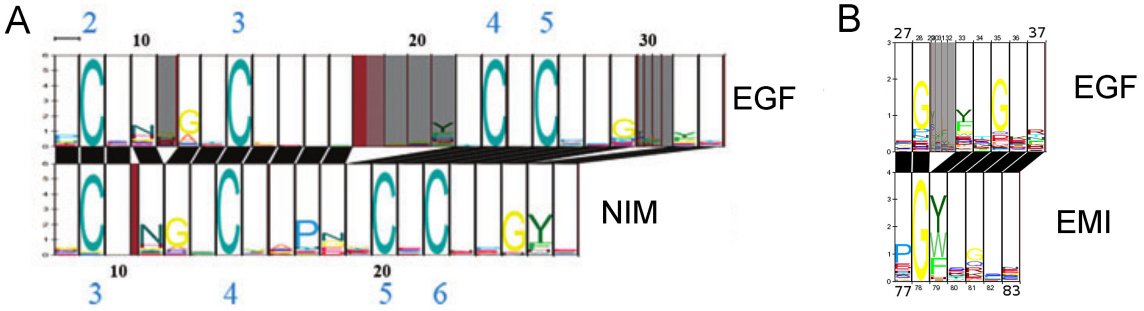
EGF-szerű domén és a NIM domén közötti hasonlóságot vizsgáltuk meg a páros HMM logó módszer segítségével.

A NIM domén körülbelül harmadik konzervált ciszteinje, és a konzervált, tirozint tartalmazó pozíciója közötti régió illeszthető az EGF-szerű doménnel, nem úgy a domén első része (4.2. **A**).

Arról, hogy a modellillesztések mennyire támasztják alá a kétséges régiók homológiáját, képet alkothatunk a PRC által számolt „score”-ok alapján, amelyek kifejezik, hogy az adott illesztés a hasonlóságokat mennyire jól magyarázza bizonyos null modellekhez viszonyítva.

A NIM és EGF-szerű domének az ábrán látható illesztésének „dot2 score”-ja (Madera, 2005) 11,5, amit úgy értelmezhetünk, hogy az illesztés a hasonlóságokat egy egyszerű null modellhez[2] képest több mint 98000-szer jobban magyarázza.

Az illesztés „reverz score”[3]-ja 5,8, ami azt jelenti, hogy az illesztés dot2 score-ja 330,29-szer nagyobb, mint a „reverz null” modell score-ja. A reverz null nem más, mint az első modell és a második modell megfordításával kapott modell illesztése, így ez megfelelő módon figyelembe veszi a modellezett szekvenciacsaládok jellegzetes aminosav-eloszlását (Madera, 2005, 2008).



4.2. ábra. Az EGF-szerű (Pfam:PF00008) és NIM domének (**A**), valamint az EGF-szerű és EMI (Pfam:PF07546) domének (**B**) lokális-lokális illesztéseit ábrázoló páros HMM logók. Az összekötött oszlopok a megfelelő pozíciók illeszkedését, míg a szürkített hátterű oszlopok a nem illeszkedő pozíciókat jelképezik. A kék színű számozás a konzervált ciszteineket, míg fekete színű a pHMM-ben elfoglalt pozíciót jelöli.

A Draper-típusú fehérjék N-terminális régiójában található EMI domén (Pfam:PF07546) Pfam adatbázisban fellelhető modellje magába foglalja a CCxGY/W motívumot is, és a modell két konzervált ciszteint követő hét konszenzus-pozíciója illeszthető az EGF-szerű domén egy régiójával (4.2. **B** ábra).

Az illesztés egyszerű score-ja ebben az esetben 5,9, vagyis az illesztés az egyszerű null modellnél 365-ször jobban magyarázza a hasonlóságokat. A reverz score ebben az esetben

[2] „joint null”
[3] „reverse null”

1,8, amit úgy értelmezhetünk, hogy az illesztés az adatokat a reverz null modellnél körülbelül hatszor jobban írja le.

4.2. A ismétlődő NIM domének evolúciója

A NimB szekvenciákban azonosított NIM domének ML módszerrel rekonstruált filogenetikáján (4.3. ábra), amelyen a leveleket a fajok szerint színeztük, szembetűnő, hogy a terminális kládok nagy többsége „vegyes színezésű", és ez nincs másként az NJ algoritmussal rekonstruált fák esetében sem (4.1. és 4.2. függelék).

Ez annak következménye, hogy a domének általában az ortológ szekvenciákban hasonló pozíciót elfoglaló doménekkel alkotnak kládokat. Ezen kládok közül egyesek, a szekvenciák rövid volta ellenére, pontosan tükrözik a Drosphila fajok aktuálisan elfogadott filogenetikáját (3.1. ábra, Russo és mtsai, 1995). A NimB szekvenciákból származó domének filogenetikája tehát ezek független evolúciójára utal, ami jól leírható a BD konceptuális modellel.

Ezzel ellentétben a NimC típusú szekvenciák ML filogenetikáit (4.4. és 4.5. ábrák) szemügyre véve számos olyan kládot találunk, amelyeket egyazon szekvenciából származó domének alkotnak, és amelyeken belül az ágak hossza is meglehetősen rövid, és ezzel az NJ algoritmussal számolt fák is összhangban vannak (4.3-6. függelék).

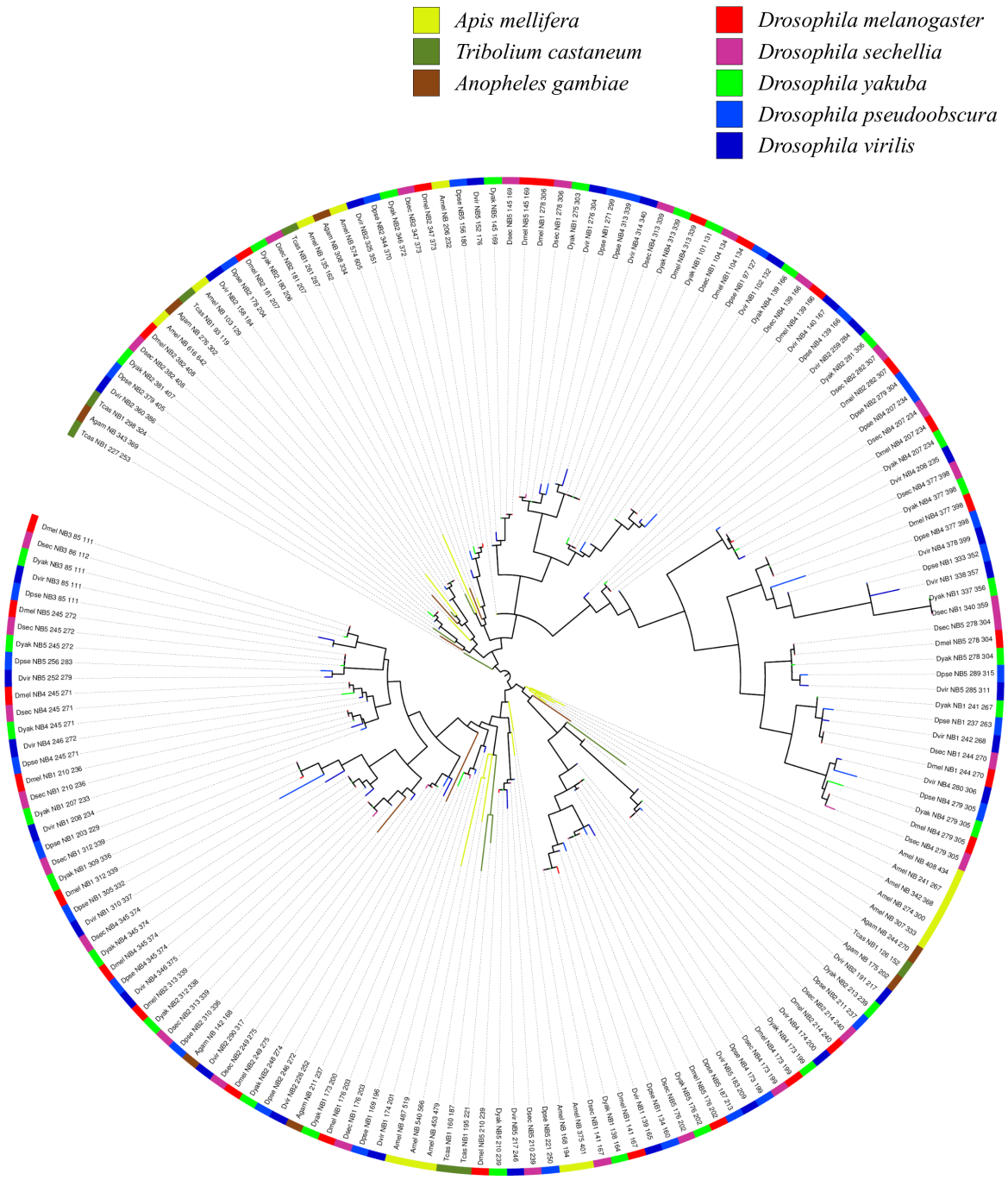
Ilyen kládokat alkotnak például egyes, az *A. mellifera* NimCI, a *T. castaneum* NimCI és NimCII szekvenciákból származó domének, de különösen szembetűnő ez az trend a Drosophila Eater ortológ szekvenciákból származó NIM domének esetében. A fentebb említett szekvenciákban található domének esetében a rekonstruált filogenetikai fák topológiája összehangolt evolúciót valószínűsít.

4.2.1. Egy vizuális módszer a domén-evolúciós trendek vizsgálatára

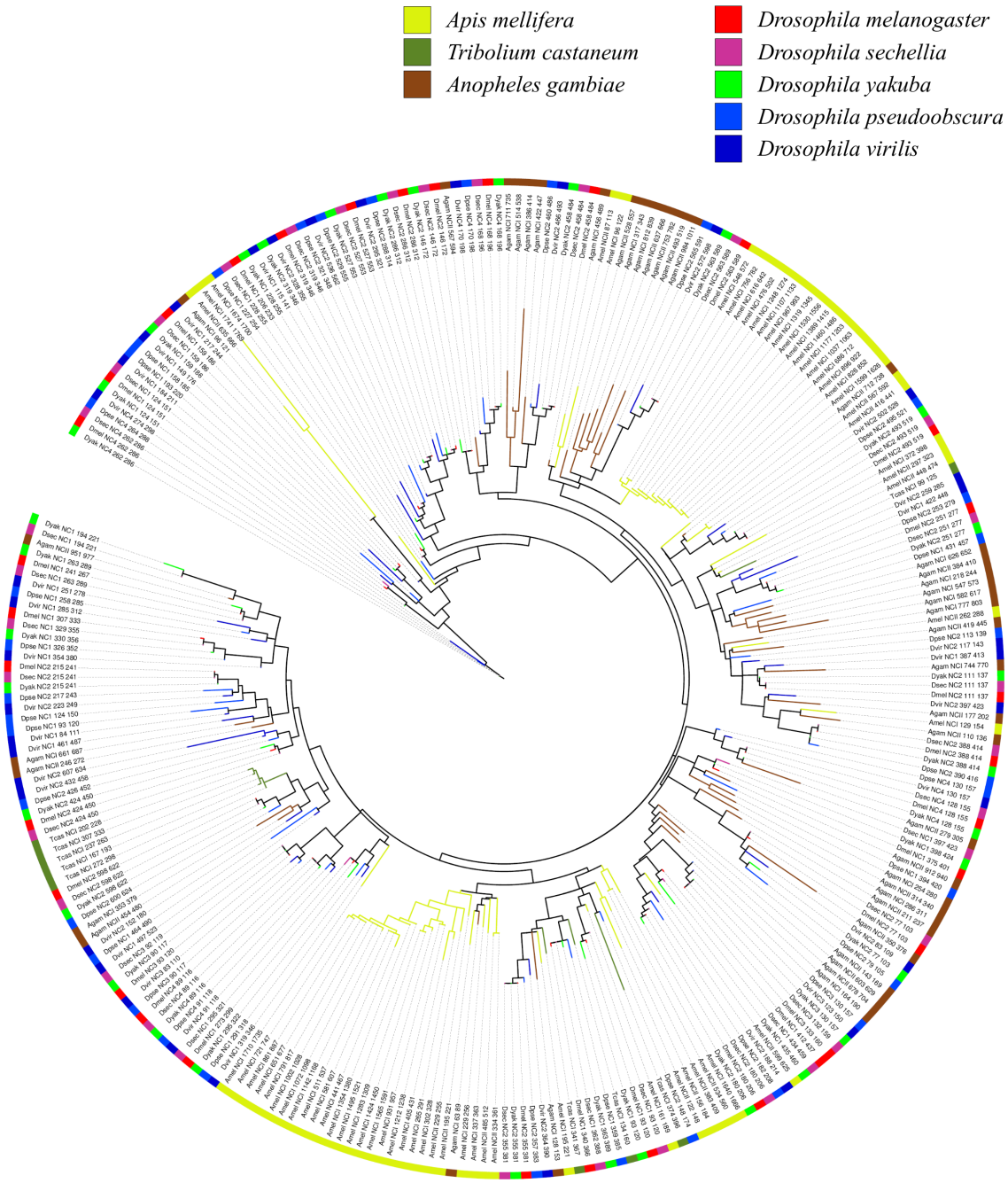
Az összehangolt evolúció vizsgálata során tulajdonképpen a pozicionális információk és a szekvencia-információk alapján felállított homológia-hipotéziseket „ütköztetjük" egymással.

A gének esetében a pozicionális információt a „lókusz", vagyis a genomi pozíció testesíti meg. Ennek megfelelően a domének esetében a szekvenciákban elfoglalt poziciót ésszerű figyelembe venni. A domének egyszerű, a teljes szekvenciában elfoglalt sorrendet tükröző „megszámozása" ellen az szól, hogy ha akár csak egy domén is deléciót szenved, vagy a domének azonosítására hivatott módszer nem ismeri azt fel, akkor a számozás „eltolódhat" és már nem várhatjuk, hogy az azonos sorszámot viselő domének egymás ortológjai legyenek.

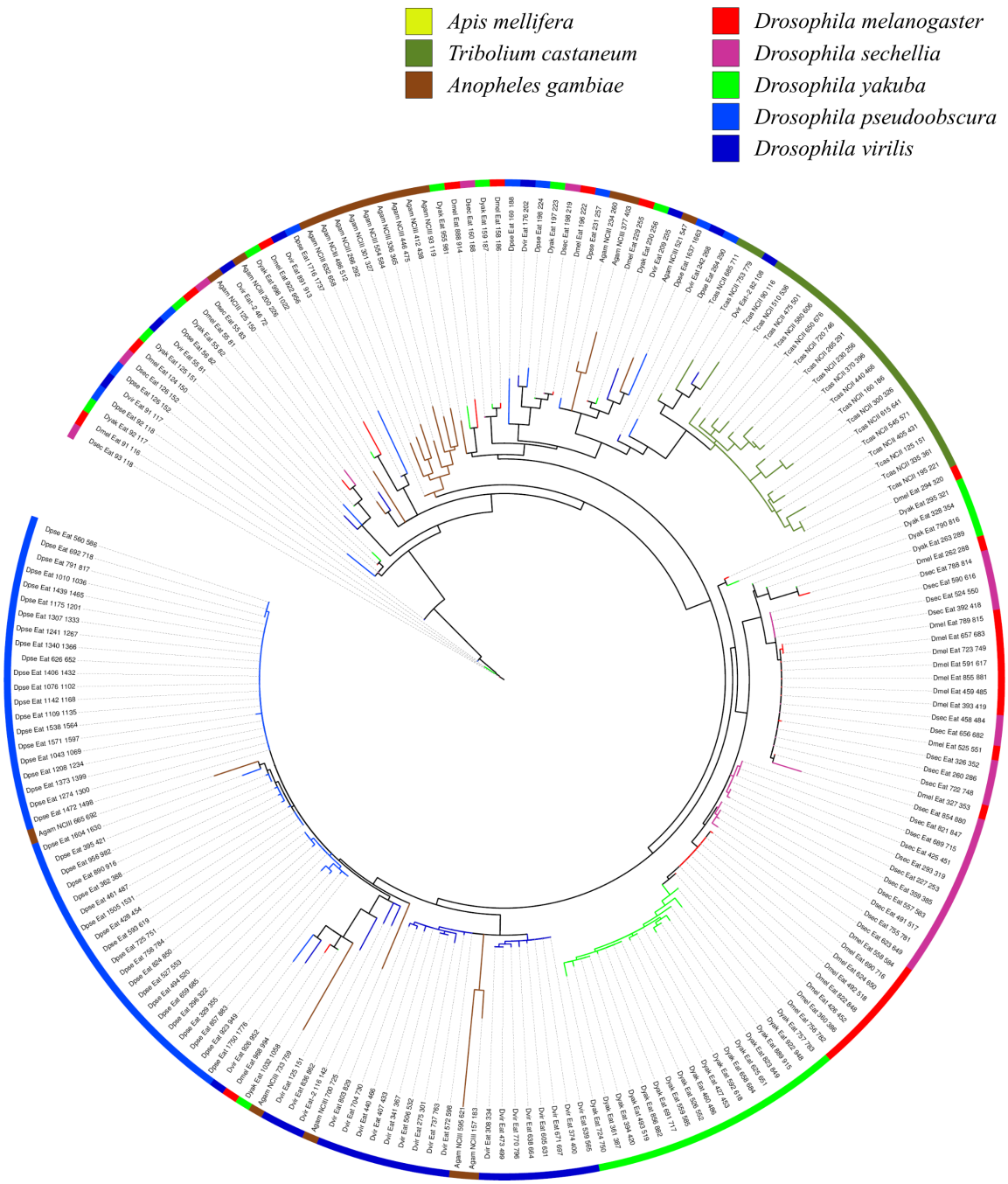
A domének pontos pozíció szerinti azonosítása (mint ahogy az a bemutatott doménfilogenetikák esetében látható) tehát korrektebb, de így a szekvenciák és a domének hosszának változatossága miatt válik nehézkessé a fák értelmezése. Nagyszámú domén



4.3. ábra. A NimB szekvenciákban azonosított NIM domének gyökértelen filogenetikájának ML becslése (illesztés: `hmmalign`, szoftver: ProtTest 1.3, a BIC szerint legjobban illeszkedő szubsztitúciós modellel számolva). A leveleket a fajok szerint színeztük, az azonosítók a szekvenciákban elfoglalt pozíciót tükrözik.Az ághosszak egysége egy szubsztitúció pozíciónként. Az ábra az iTOL 1.7 (Letunic és Bork, 2007) webalkalmazás segítségével készült.



4.4. ábra. A NimC szekvenciákban (a *T. castaneum* NimCII és az *A. gambiae* NimCII kivételével) azonosított NIM domének gyökértelen filogenetikájának ML becslése (illesztés: `hmmalign`, szoftver: ProtTest 1.3, a BIC szerint legjobban illeszkedő szubsztitúciós modellel számolva). A leveleket a fajok szerint színeztük, az azonosítók a szekvenciákban elfoglalt pozíciót tükrözik.Az águhosszak egysége egy szubsztitúció pozíciónként. Az ábra az iTOL 1.7 (Letunic és Bork, 2007) webalkalmazás segítségével készült.



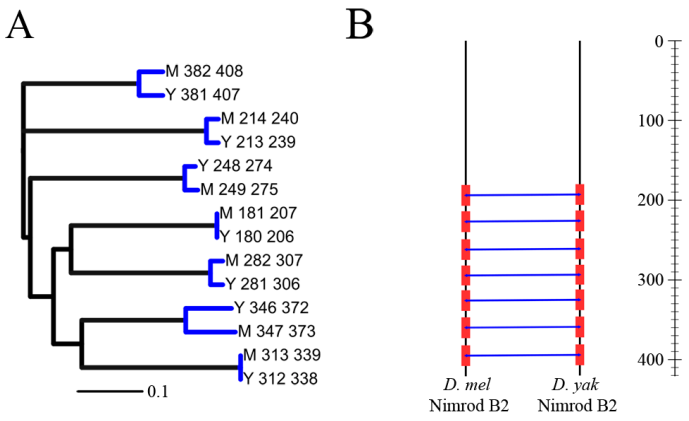
4.5. ábra. Az Drosophila Eater szekvenciákban, valamint a *T. castaneum* NimCII és az *A. gambiae* NimCIII szekvenciákban azonosított NIM domének gyökértelen filogenetikájának ML becslése (illesztés: `hmmalign`, szoftver: ProtTest 1.3, a BIC szerint legjobban illeszkedő szubsztitúciós modellel számolva). A leveleket a fajok szerint színeztük, az azonosítók a szekvenciákban elfoglalt pozíciót tükrözik.Az águhosszak egysége egy szubsztitúció pozíciónként. Az ábra az iTOL 1.7 (Letunic és Bork, 2007) webalkalmazás segítségével készült.

vizsgálata esetén a fák mérete is megnehezítheti azok értelmezését, és ha ezeket kisebb csoportokra osztva vizsgáljuk, az is csak részben oldja meg a problémát.

Mindezekből láthatjuk, hogy a doménfilogenetikák alkalmasak arra, hogy jellemezzük ezek evolúciójának általános trendjeit, de kevésbé praktikusak, ha meg szeretnénk vizsgálni, hogy a szekvenciák különböző régióiban a domének evolúcióját melyik konceptuális modellel írhatjuk le legjobban. A fenti problémák orvoslására, valamint azért, hogy megkönnyítsük az ismétlődő domének evolúciójának vizsgálatát, kifejlesztettünk egy, a doménfilogenetikán alapuló egyszerű vizuális módszert.

A módszer lényege, hogy egy szekvenciapárban azonosított domének filogenetikáján azonosítja azokat a kládokat, amelyeket csupán két levél, és az őket összekapcsoló egyetlen belső csomópont alkot (4.6. **A** és 4.7. **A** ábrák). Az azonosított levél-pároknak[4] megfelelő domének nyilvánvalóan egy, a teljes vizsgált domén-halmazra nézve legközelebbi közös őssel rendelkeznek. Így ezeket egymás „legközelebbi" homológjának tekintjük, vagyis „ortológoknak", ha különböző, illetve „paralógoknak", ha egyazon szekvenciából származnak.

A szekvencia-információk alapján felállított homológia-hipotéziseket a pozicionális információval a következő módon „ütköztethetjük" praktikusan: a két szekvenciát egymás mellett ábrázoljuk az azonosított doméneknek megfelelő régiók kiemelésével, majd az azonosított levélpároknak megfelelő doméneket összekötjük, egy kék egyenessel, ha különböző szekvenciákból származnak („ortológia kapcsolatok"), illetve egy vörös vonallal, ha azonos szekvenciákból származnak („paralógia kapcsolatok").



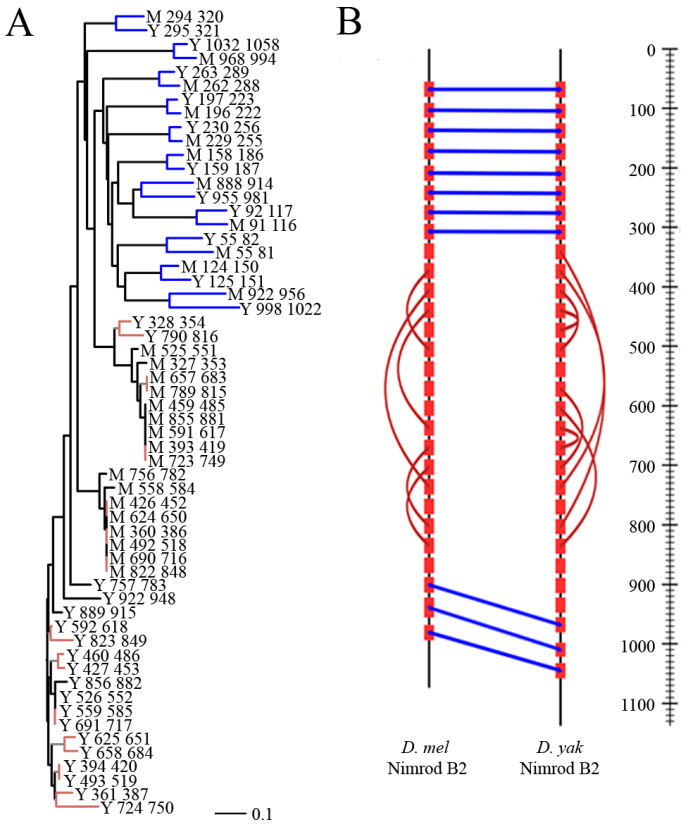
4.6. ábra. A *D. melanogaster* és *D. yakuba* NimB2 szekvenciákból származó NIM domének NJ algoritmussal rekonstruált (Clustal W 1.83) filogenetikája (**A**) és az ennek megfelelő PRHD (**B**). A **B** ábra jobb oldalán látható skála egysége egy pozíció. A fa águhosszainak egysége egy szubsztitúció pozíciónként. A levelek azonosítói a fajneveket (M - *D. melanogaster*, Y - *D. yakuba*) és a domének szekvenciákban elfoglalt pozícióit tükrözik. A piros téglalapok a két szekvenciában azonosított NIM doméneket jelképezik. A PRHD-n a kék vonalak az ortológia kapcsolatoknak felelnek meg, a fán nekik megfelelő kládokat kék színnel kiemeltük.

[4] „sister leaf nodes"

Az így kapott ábra segítségével, amit „Pairwise Repeat Homology Diagram”-nak (a továbbiakban PRHD) nevezünk, könnyen felmérhető, hogy a vizsgált szekvenciák különböző régióiban található domének evolúciója melyik konceptuális modellel írható le a legjobban.

Ha az adott régióban, vagy a teljes szekvenciában az ortológia kapcsolatok egy többé-kevésbé „létraszerű” mintázatot alkotnak, az azt jelenti, hogy a domének a filogenetikai fán nagyrészt az ortológ szekvenciákban hasonló pozícióban lévő doménekkel asszociálódnak, ami független, a BD modellel jól leírható evolúcióra utal (4.6. **B** ábra).

Ha egy régióban az ortológia kapcsolatok nem mutatják az említett mintázatot, hanem a paralógia kapcsolatok dominálnak, az az adott régióban található domének összehangolt evolúciójára utal (4.7. **B** ábra). A domének vizualizálása révén a módszer megkönnyíti az összekötő régiók és egyéb nem ismétlődő szakaszok hosszának felmérését is.



4.7. ábra. A *D. melanogaster* és *D. yakuba* Eater szekvenciákból származó NIM domének NJ algoritmussal rekonstruált (Clustal W 1.83) filogenetikája (**A**) és a neki megfelelő PRHD (**B**). A fa ághosszainak egysége egy szubsztitúció pozíciónként. A levelek azonosítói a fajneveket (M - *D. melanogaster*, Y - *D. yakuba*) és a domének szekvenciákban elfoglalt pozícióit tükrözik. A piros téglalapok a két szekvenciában azonosított NIM doméneket jelképezik. A PRHD-n a kék vonalak az ortológia kapcsolatoknak, míg a vörös vonalak a paralógia kapcsolatoknak felelnek meg. A fán nekik megfelelő kládokat kék, illetve vörös színnel kiemeltük.

Mivel a PRHD-k alapvetően a domén-filogenetikákon alapulnak, esetükben is közeli

fajokból származó szekvenciákra van szükség a doménevolúció jellemzéséhez.

Az összehangolt evolúció epizodikus is lehet. A filogenetikában különböző pozíciókat elfoglaló fajokból származó szekvenciák párbaállításával kapott PRHD-k segítségével következtethetünk arra, hogy ezen epizódok melyik kládokban, és melyik szekvencia-régiókon belül zajlottak le.

A módszer implementációja

A módszert implementáló Perl szkript (`t2prhd`[5]) elérhető a `http://t2prhd.sf.net` weblapon.

A szkript az első lépésben egy, a vizsgált domént modellező pHMM és a `hmmsearch` program segítségével azonosítja a két szekvenciában található doméneket, és `BioPerl` modulok segítségével a nyers keresési eredményekből azok szekvenciáit kinyeri és a `hmmalign` program segítségével ezeket illeszti. Az így kapott domén-illesztéseket a szkript szekvenciális PHYLIP és Fasta formátumokba konvertálja, majd ezekből a Clustal W (Thompson és mtsai, 1994) vagy a PhyML (Guindon és Gascuel, 2003) programot felhasználva megbecsüli a domének filogenetikáját, az NJ vagy az ML módszer segítségével.

Az alkalmazás a becsült filogenetikai fát a `Bio::Tree::TreeI` objektumként beolvassa, majd azonosítja az ortológia és paralógia kapcsolatoknak megfelelő kládokat. Ezen kládok, valamint a két szekvenciában azonosított domének pozíciói alapján felépített PRHD-t a szkript az `XML::Writer` modul segítségével SVG[6] formátumban elmenti.

A domének közötti kapcsolatok színének intenzitása a két domén között az ághosszak összegeként számolt patrisztikus távolság (d_{ij}) és a fa ághosszainak összegétől (T) függ:

$$I(d_{ij}, T, w) = \left(1 - \frac{d_{ij}}{T}\right)^{w} \tag{4.1}$$

A kapcsolatok alapjául szolgáló kládokat alkotó levelek közötti patrisztikus távolság szükségszerűen kicsi. A w paraméter növelésével a színezés intenzitása úgy módosítható, hogy az jobban elősegíti a kis különbségek észrevételét. A w paraméter alapértelmezett értéke 1, ami a lineáris színskálának felel meg. A színek intenzitása és a patrisztikus távolságok közötti összefüggést egy opcionális színskála szemlélteti.

A szkript opcionálisan LaTeX formátumban is elmenti a PRHD-t, és a kimeneti fájlok között megőrzi a felhasznált segédprogramok kimeneteit és naplófájljait. Az eredményként kapott SVG formátumú állományok megjeleníthetők például a `Firefox`[7] böngészővel, vagy raszterizálhatóak a `Batik`[8] szoftvercsomag segítségével. A LaTeX formátumú állományok a

[5] „Tree To PRHD”
[6] „Scalable Vector Graphics”
[7] `http://www.mozilla.com/en-US/firefox`
[8] `http://xmlgraphics.apache.org/batik`

`pdflatex` program, valamint a `pgf/TikZ`, `fancyhdr`, `xcolor` és `fullpage` csomagok felhasználásával PDF állományokká alakíthatóak. A `t2prhd` szkript tartalmazza a parancssori paramétereket ismertető POD formátumú kézikönyvet is.

A szkript elsősorban a doménevolúció vizsgálatának megkönnyítésére szolgál, de nukleotid-szekvenciákat és nukleotid-szekvenciákkal betanított pHMM-eket is képes kezelni. Nem veszi azonban figyelembe a szekvenciák reverz komplemensét, ezért genomikai vizsgálatokban csak korlátozott mértékben használható.

A `t2prhd` szkript 1.7-es verziója minden azonosított kapcsolatot feltüntet a PRHD-n, ez azonban nem minden esetben célszerű. Ha például a domének intenzív homogenizációs epizódokon esnek át, akkor gyakran előfordul, hogy szekvenciáik teljes mértékben megegyeznek. A teljesen azonos szekvenciák által alkotott kládokon belül azonban az elágazások a felhasznált programtól függően vagy véletlenszerűek, vagy a szekvenciák illesztésben elfoglalt pozíciójától függenek, és így a nekik megfelelő paralógia kapcsolatok semmiképpen nem tekinthetőek informatívnak a domének evolúciójára nézve.

Ez a jelenség nem érinti a különböző régiók domén-evolúciójának jellemzését, de problémát jelent, ha a belső kapcsolatok mintázatából következtetni próbálunk például a homogenizáció egységeire. Ez a probléma, sokkal kisebb eséllyel, de az ortológia kapcsolatok esetében is jelentkezhet.

A szkript 2.0-ás verziójában lehetőség van a teljesen azonos szekvenciák által alkotott kládokból származó kapcsolatok kiszűrésére. A NIM domének vizsgálata során a PRHD-k még az előző verzióval (1.7) készültek, de ez nem befolyásolja az evolúciójukra vonatkozó következtetéseinket.

4.2.2. A NIM domének evolúciójának vizsgálata PRHD-k segítségével

A Drosophila NimB ortológok párbaállításával készült PRHD-k (4.8. és 4.9. ábrák, 5.1-2. függelék) megerősítik a doménfilogenetikák alapján levont azon következtetést, miszerint az ezen szekvenciákban található NIM domének evolúciója jól leírható a BD modellel.

Hasonlóképpen a Drosophila NimC2 (4.10. ábra), valamint a NimC3 és NimC4 ortológok párbaállításával készített diagramok (5.3-4. függelék) is egyértelműen a domének független evolúciójára utalnak. A NimC1 ortológok esetében (4.11. ábra) az ortológia kapcsolatok már nem minden esetben mutatnak egy jól kivehető létraszerű mintázatot, de összefüggés mutatkozik ezen jelenség és a két szekvencia közötti evolúciós távolság között, amit az illető fajoknak a Drosophila filogenetikában elfoglalt pozíciója alapján becsülhetünk meg (3.1. ábra).

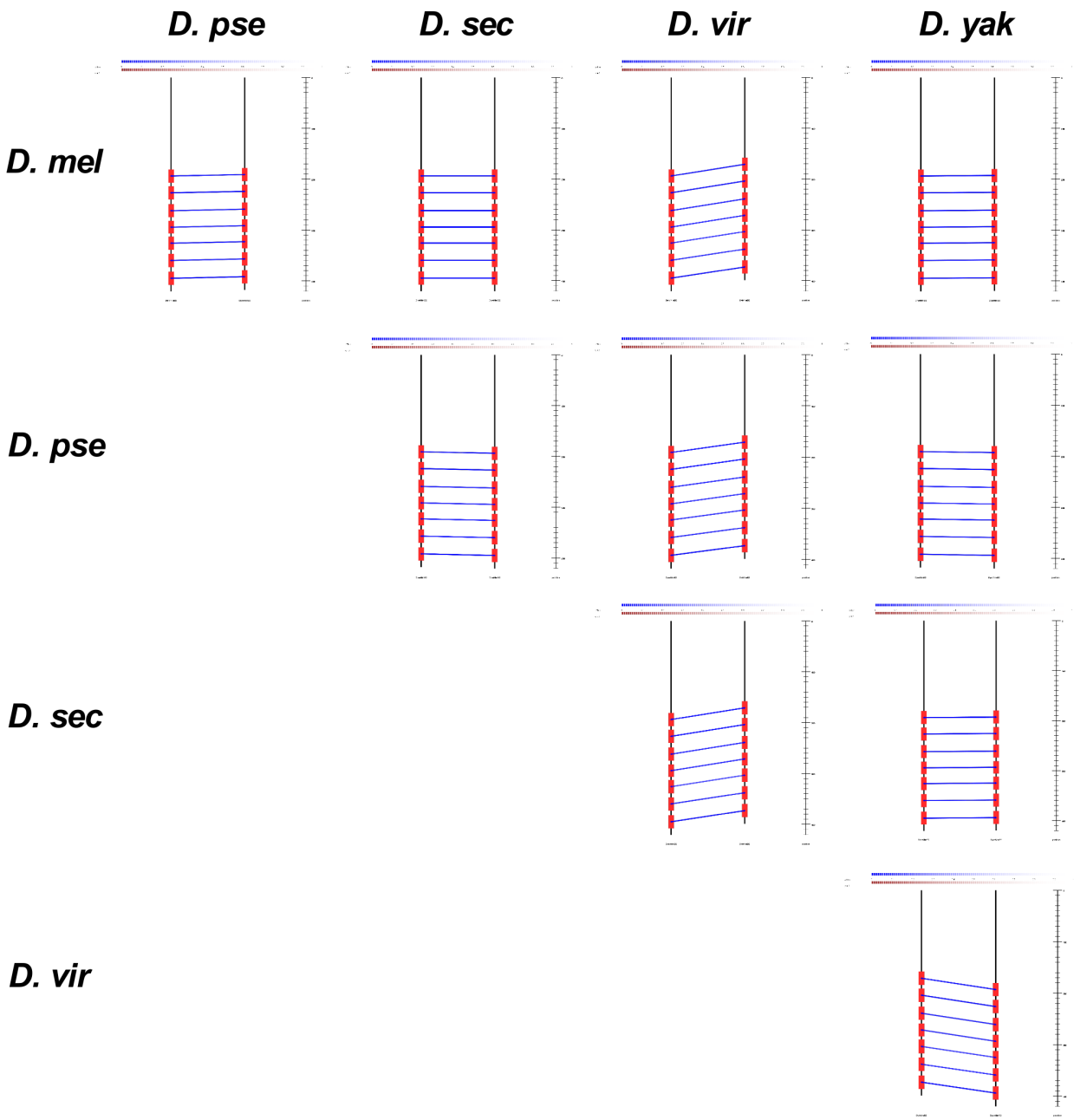
A NimC1 szekvenciákkal ellentétben a Drosophila Eater ortológok esetében mindegyik szekvenciapár esetében az összehangolt evolúcióra utaló mintázatot kapunk a szekvencia középső régiójában (4.12. ábra). Ugyancsak megfigyelhető, hogy a domén-ismétlődés N-terminális régiójában a C-terminális régióhoz képest több domén marad ki a homogenizációs

folyamatból. Ezekben az inkább a BD modellel leírható régiókban - a NimC1 szekvenciákhoz hasonlóan - jelentkezik a létraszerű mintázat hangsúlyos volta és a szekvenciák közötti evolúciós távolság közötti összefüggés.

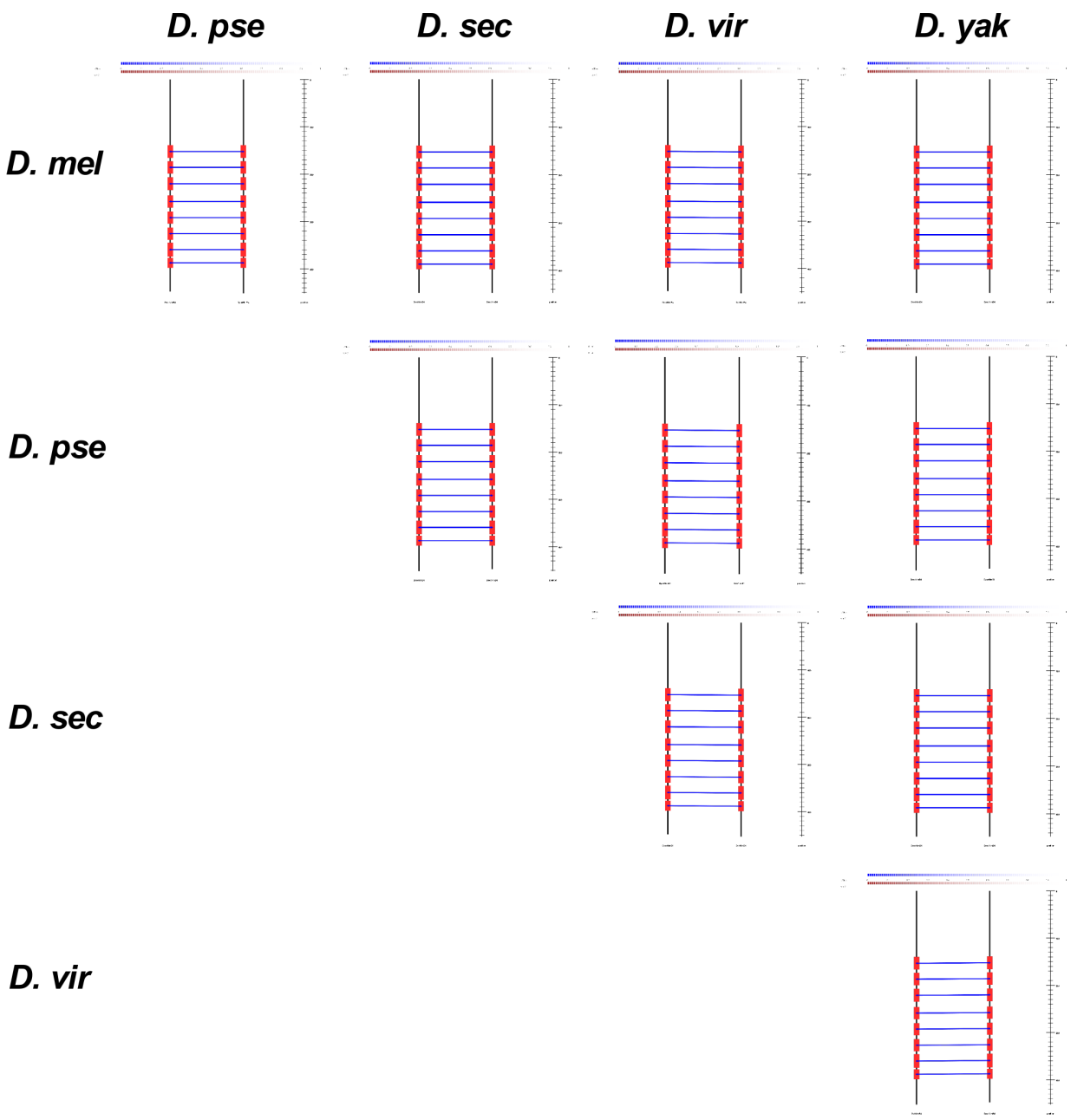
A NIM domének evolúciójának vizsgálata során a kapcsolatok színezésének intenzitása nem bizonyult informatívnak, így ezeket nem tárgyaljuk.

A 4.13. ábrán megfigyelhető, hogy az Eaterekben található doménismétlődések középső régiójában a domének közötti szinonim távolságok a többi összehasonlításokhoz képest kisebbek, ami a PRHD-k alapján feltételezett homogenizációt a nukleotid-szekvenciák szintjén is megerősíti. A becsült távolságokra alapozva tehát kizárhatjuk, hogy az említett domének a BD modell szerint viselkednek és közöttük a hasonlóságot egy erős, az aminosav-szekvenciák szintjén ható negatív szelekció tartja fenn.

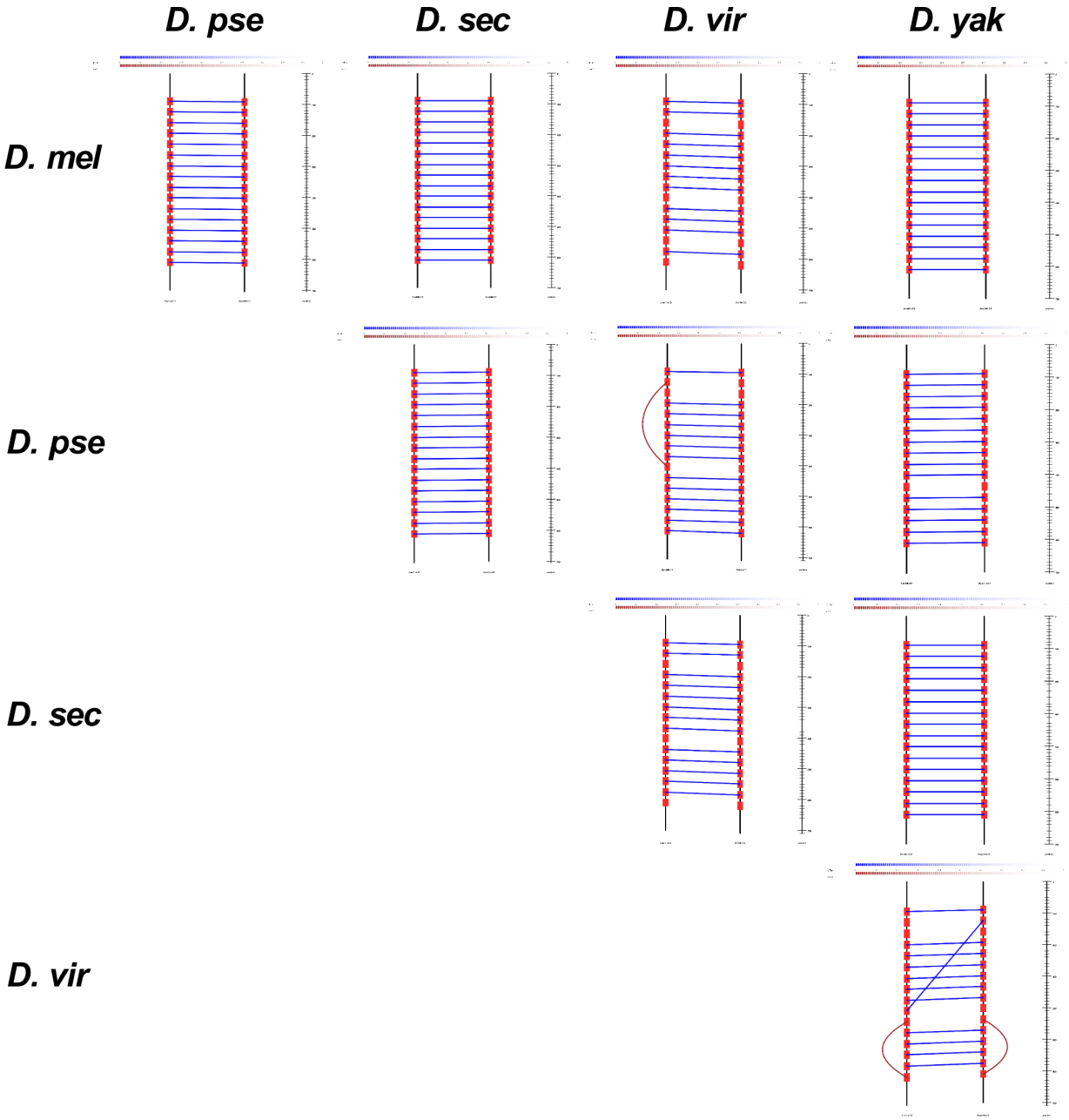
Az Eater-ekkel ellentétben, a NimC1 és NimC2 szekvenciákban azonosított NIM domének távolsági mátrixait elemezve nem találtunk az összehangolt evolúcióra utaló mintázatokat (lásd például a 4.14. ábrát).



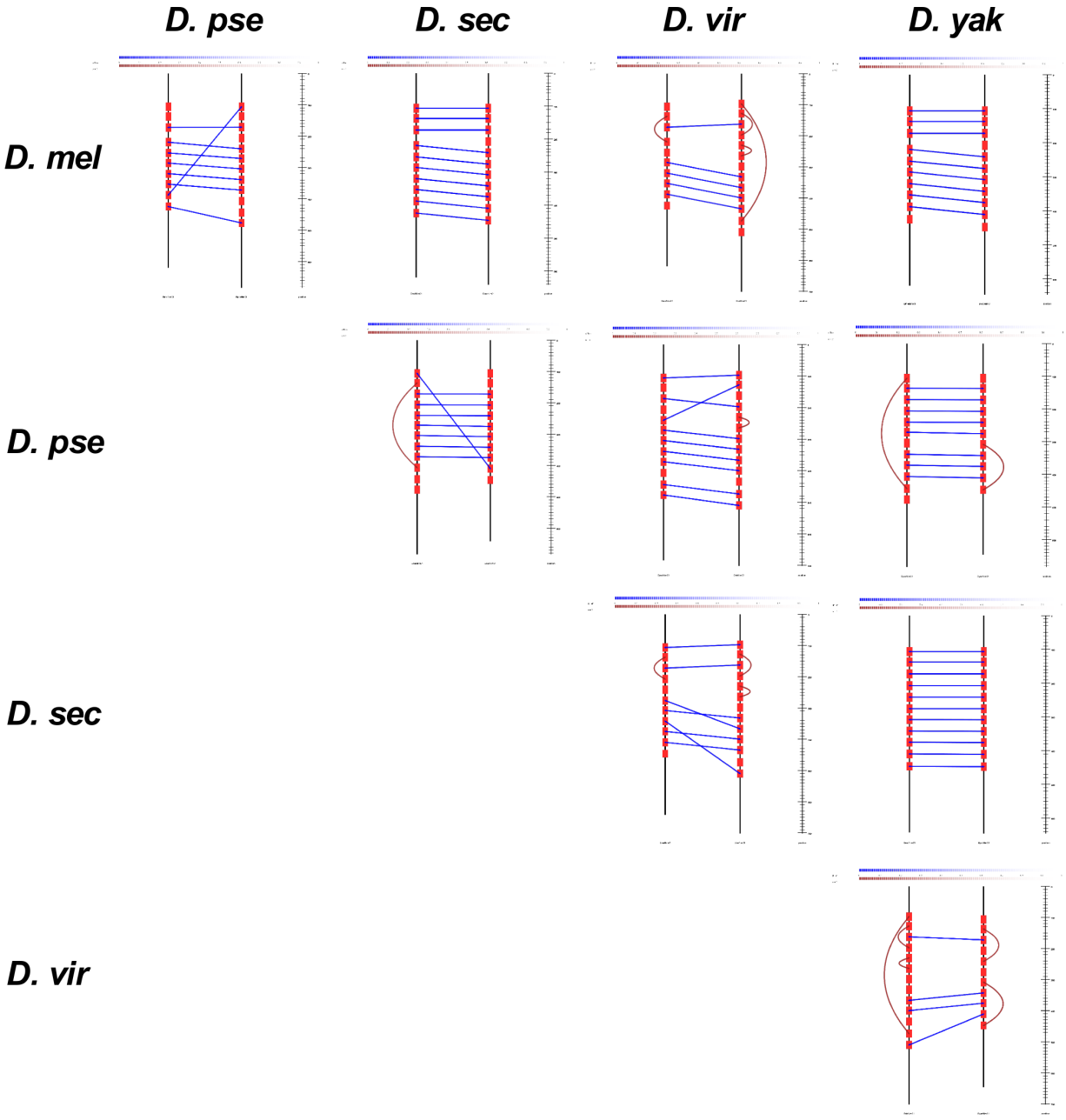
4.8. ábra. A Drosophila NimB2 ortológ szekvenciák párbaállításával számolt PRHD-k. Az ábrák a `t2prhd` 1.7 segítségével ($w = 2$) készültek a Clustal W 1.83 által számolt NJ domén-filogenetikák alapján. A piros téglalapok a két szekvenciában azonosított NIM doméneket jelképezik. A kék vonalak az ortológia kapcsolatoknak felelnek meg.



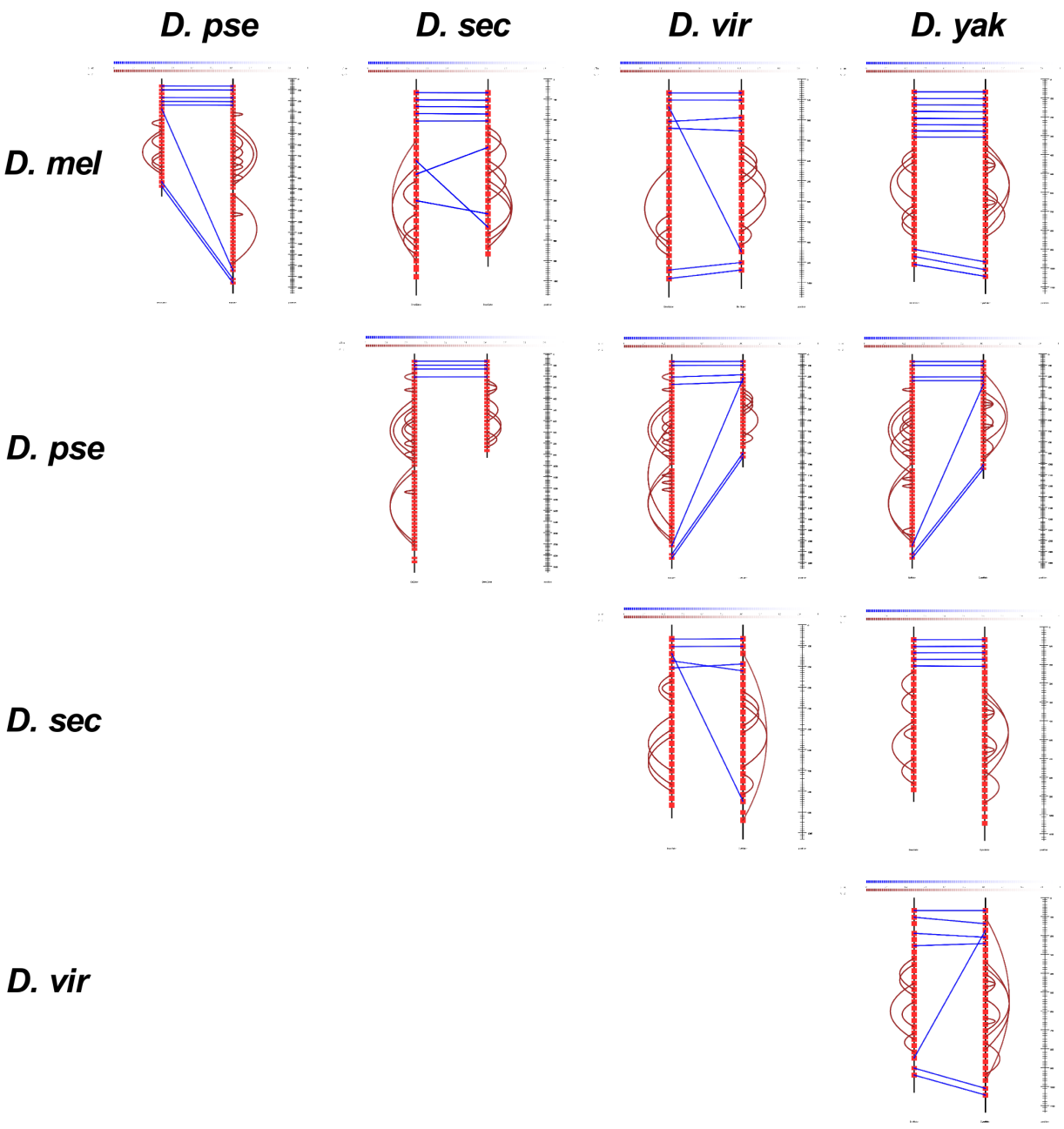
4.9. ábra. A Drosophila NimB4 ortológ szekvenciák párbaállításával számolt PRHD-k. Az ábrák a `t2prhd` 1.7 segítségével ($w = 2$) készültek a Clustal W 1.83 által számolt NJ doménfilogenetikák alapján. A piros téglalapok a két szekvenciában azonosított NIM doméneket jelképezik. A kék vonalak az ortológia kapcsolatoknak felelnek meg.



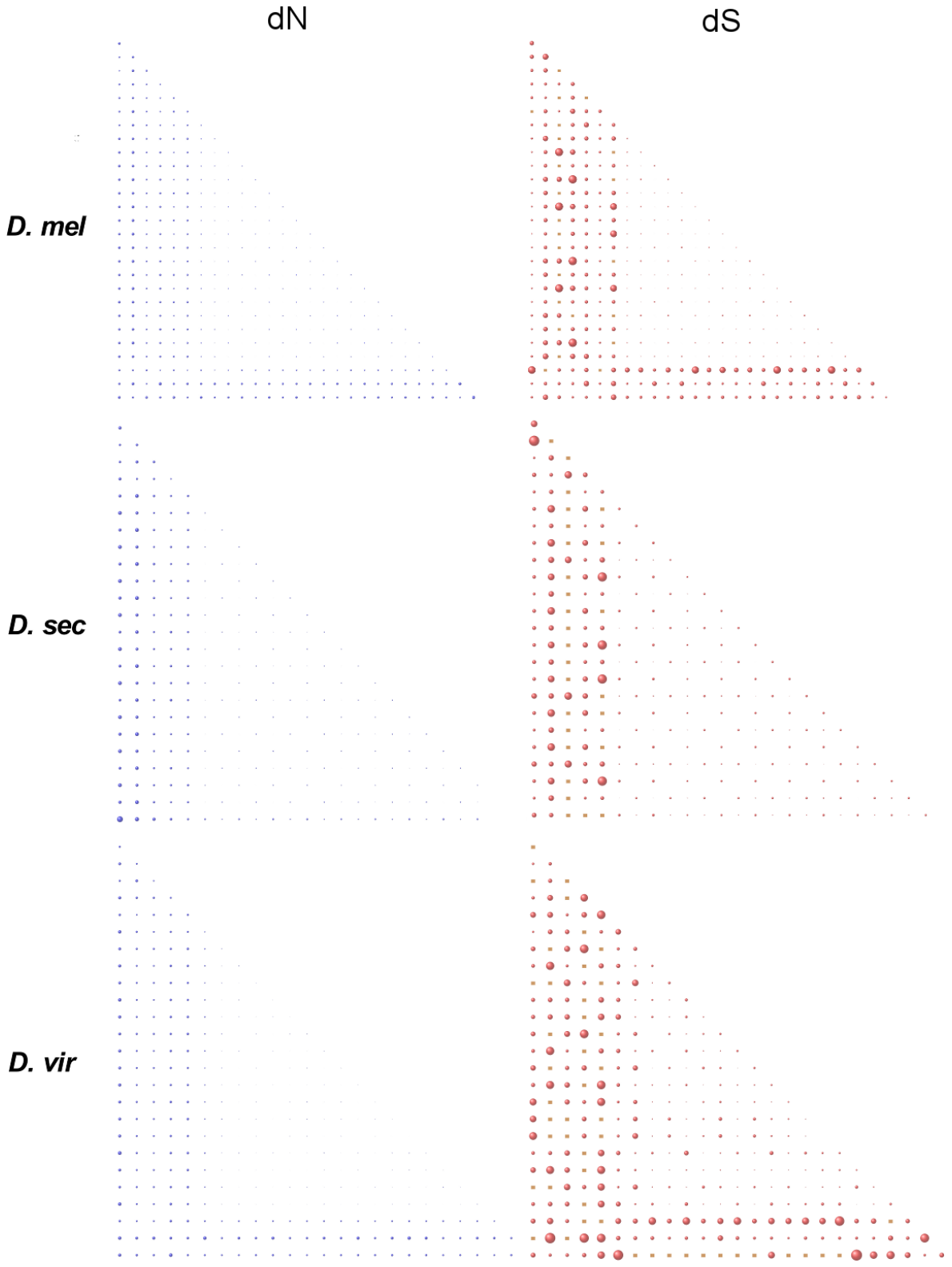
4.10. ábra. A Drosophila NimC2 ortológ szekvenciák párbaállításával számolt PRHD-k. Az ábrák a `t2prhd` 1.7 segítségével ($w = 2$) készültek a Clustal W 1.83 által számolt NJ doménfilogenetikák alapján. A piros téglalapok a két szekvenciában azonosított NIM doméneket jelképezik. A kék vonalak az ortológia, míg a vörös vonalak a paralógia kapcsolatoknak felelnek meg.



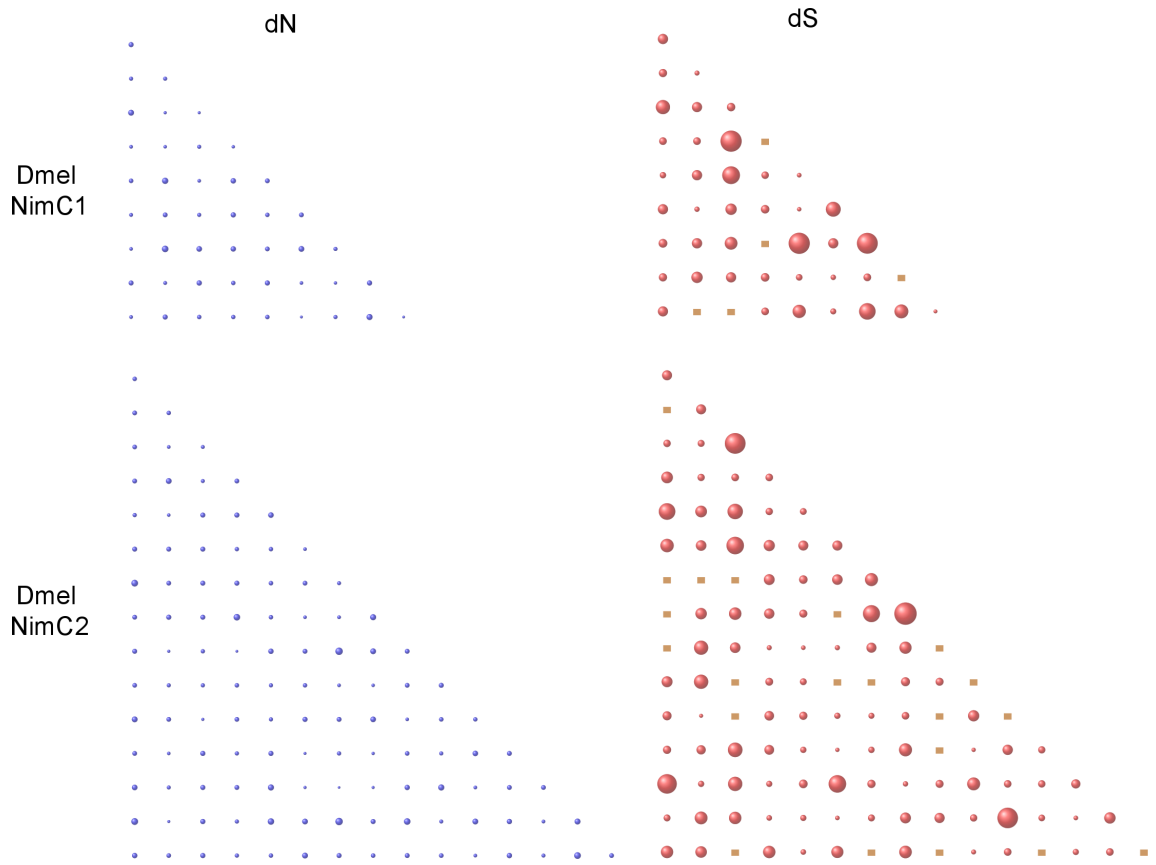
4.11. ábra. A Drosophila NimC1 ortológ szekvenciák párbaállításával számolt PRHD-k. Az ábrák a `t2prhd` 1.7 segítségével ($w = 2$) készültek a Clustal W 1.83 által számolt NJ doménfilogenetikák alapján. A piros téglalapok a két szekvenciában azonosított NIM doméneket jelképezik. A kék vonalak az ortológia, míg a vörös vonalak a paralógia kapcsolatoknak felelnek meg.



4.12. ábra. A Drosophila Eater ortológ szekvenciák párbaállításával számolt PRHD-k. Az ábrák a `t2prhd` 1.7 segítségével ($w = 2$) készültek a Clustal W 1.83 által számolt NJ domén-filogenetikák alapján. A piros téglalapok a két szekvenciában azonosított NIM doméneket jelképezik. A kék vonalak az ortológia, míg a vörös vonalak a paralógia kapcsolatoknak felelnek meg.



4.13. ábra. A *D. melanogaster*, *D. sechellia* és a *D. virilis* Eater szekvenciákban azonosított, és a pozíciójuk alapján sorbaállított NIM domének közötti szinonim ($\mathbf{d_S}$) és nemszinonim ($\mathbf{d_N}$) távolságok ábrázolása. A távolsági mátrixokat a módosított Nei-Gojobori módszerrel, a MEGA 3.1 szoftver segítségével számoltuk ki. A gömbök méretei a megfelelő távolságokkal arányosak. A téglalapokkal jelzett összehasonlítások esetében a távolság nem számolható ki a módosított Nei-Gojobori módszerrel.



4.14. ábra. A *D. melanogaster* NimC1 és NimC2 szekvenciákban azonosított, és a pozíciójuk alapján sorbaállított NIM domének közötti szinonim ($\mathbf{d_S}$) és nemszinonim ($\mathbf{d_N}$) távolságok ábrázolása. A távolsági mátrixokat a módosított Nei-Gojobori módszerrel, a MEGA 3.1 szoftver segítségével számoltuk ki. A gömbök méretei a megfelelő távolságokkal arányosak. A téglalapokkal jelzett összehasonlítások esetében a távolság nem számolható ki a módosított Nei-Gojobori módszerrel.

A szinonim és nemszinonim távolságok kiszámolásához szükséges volt megbecsülni a tranzíciós/transzverziós mutációk arányát[9] a megfelelő illesztésekből, és ennek során tettük a következő megfigyelést: egy kivétellel (*D. yakuba* NimC1) az Eater szekvenciákból származó doméneket kódoló szekvenciák illesztéseiből becsült tranzíciós/transzverziós arányok mindig lényegesen nagyobbak voltak a NimC szekvenciák doménjeit kódoló szekvenciák illesztéseiből becsülteknél.

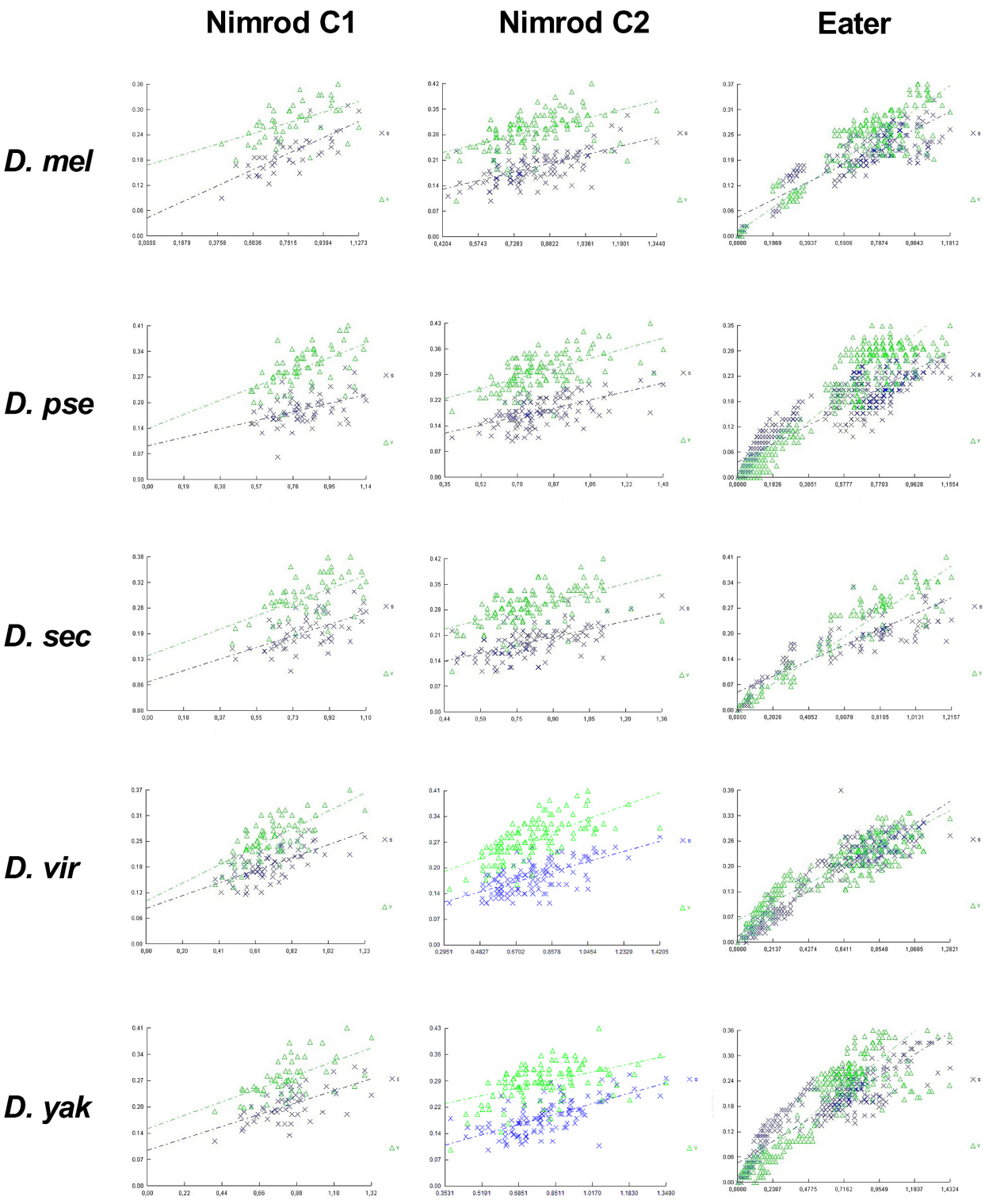
Ez magyarázható azzal, hogy a feltehetően alacsonyabb rátával bekövetkező transzverziós mutációkat a homogenizációs események elfedik. Az egymástól függetlenül evolválódó domének esetében az idő múlásával mindkét típusú mutáció megközelíti a szaturáció szintjét, ami kisebb tranzíciós/transzverziós arányok becsléséhez vezet.

Ezt a magyarázatot megerősíteni látszanak a NimC1, NimC2 és Eater szekvenciákban található NIM doméneket kódoló nukleotid-szekvenciák illesztései alapján számolt szaturációs

[9]Nem összetévesztendő a tranzíciós/transzverziós ráta-aránnyal.

ábrák (4.15. ábra).

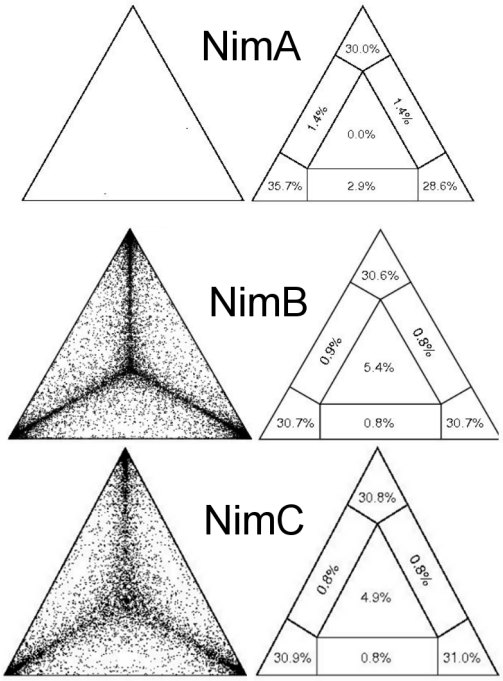
Az Eater szekvenciák doménjeit kódoló illesztések esetében a korrekció nyomán a távolságok lényegesen nem változnak, a NimC illesztéseknél azonban mindkét típusú mutáció esetében a pontokra illesztett egyenesek a vízszinteshez közel állnak, ami arra utal hogy az ezen szekvenciák között felhalmozódott különbségek száma megközelíti a szaturáció szintjét.



4.15. ábra. A Drosophila NimC1, NimC2 és Eater szekvenciákban azonosított NIM doméneket kódoló nukleotid-szekvenciák többszörös illesztéseiből számolt szaturációs ábrák. A pontok az egyes szekvenciapárok között megfigyelhető különbségek számát ábrázolják, az F84 szubsztitúciós modell által számolt távolságok függvényében. A vonalak a megfelelő ponthalmazokra illesztett lineáris regressziós egyenesek. A zöld háromszögek és egyenesek a transzverziós, míg a kék keresztek és egyenesek a tranzíciós különbségekhez tartoznak. Az ábrák a DAMBE 4.2.13 szoftverrel készültek.

4.3. A filogenetikai analízisek eredményei

A likelihood térképezések értelmezése (4.16. ábra) csak a szekvenciák hosszának függvényében lehetséges (Strimmer és von Haeseler, 1997). A Nimrod szekvenciák illesztései azonban sok olyan pozíciót tartalmaznak, melyek nagyrészt gapekből állnak, és így nem informatívak a topológiára nézve. Ezért nem lenne ésszerű az illesztések hosszára alapozni a likelihood térképezések értelmezésénél, úgyhogy erre a célra az illesztésekben található parszimónia-informatív[10] pozíciók számát használtuk, amit a MEGA 3.1 szoftverrel számoltunk ki.



4.16. ábra. A NimA , NimB és NimC szekvenciák ProbCons illesztéseinek likelihood térképezése. A jobb oldalon látható háromszögű koordinátarendszerek értelmezéséhez lásd a 3.5. ábrát. A bal oldalon látható koordinátarendszerek a megfelelő régiókba eső pontok arányát szemléltetik.

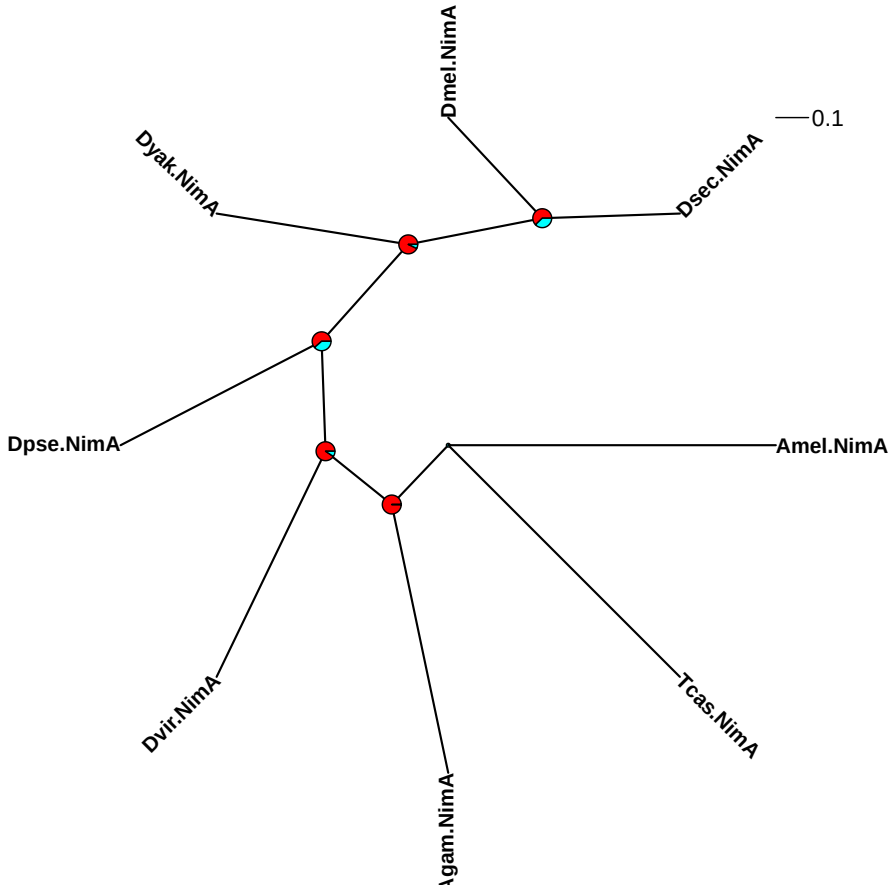
A NimA illesztés likelihood térképezésének esetében a középső régiókban található pontok aránya 0%, ami az adat mennyiségétől függetlenül erős filogenetikai szignálra utal.

A NimB és NimC illesztések esetében, melyekben a parszimónia-informatív pozíciók száma 352 és 587, a középső régióban a pontok aránya 5,4% és 4,9%. Strimmer és von Haeseler (1997) eredményeire támaszkodva ezt úgy értelmezhetjük, hogy ezekben az illesztésekben a filogenetikai szignál ugyan megfelelő erősségű, de bizonyos mértékű bizonytalanság várható a becsült topológiákban.

[10] „parsimony informative”

A likelihood és a splitgyakoriságok alakulása, valamint a PSRF értékek mindegyik Bayes MCMC futtatás esetében a konvergenciára utaltak, beleértve a Bali-Phy futtatásokat is (PSRF = 1,09).

A NimA ortológok ML módszerrel becsült filogenetikája (4.17. ábra) teljes mértékben összhangban van a fajok jelenleg elfogadott filogenetikájával (3.1. ábra) és nem tartalmaz jelentősebb bizonytalanságot.



4.17. ábra. A NimA szekvenciák gyökértelen filogenetikájának ML becslése (illesztés: ProbCons 1.1, szoftver: PhyML 3.0, szubsztitúciós modell: JTT+Γ + F, gamma paraméter: 1,215, gamma kategóriák száma: 4, bootstrap ismétlések száma: 1000). A kládok bootstrap támogatottsága a megfelelő csomópontban található kördiagramok piros szeletével arányos. Az ághosszak egysége egy szubsztitúció pozíciónként. Az ábra az APE R csomag (Paradis és mtsai, 2004) segítségével készült.

A szekvencia-információk alapján Bayes MCMC-vel becsült többségi konszenzusfa topológiája (7.3. függelék), valamint az NJ PD módszerrel felépített topológia (7.1. függelék) ezzel teljes mértékben megegyezik, és ettől az NJ CD fa topológiája is csak két taxon tekintetében tér el (7.2. függelék).

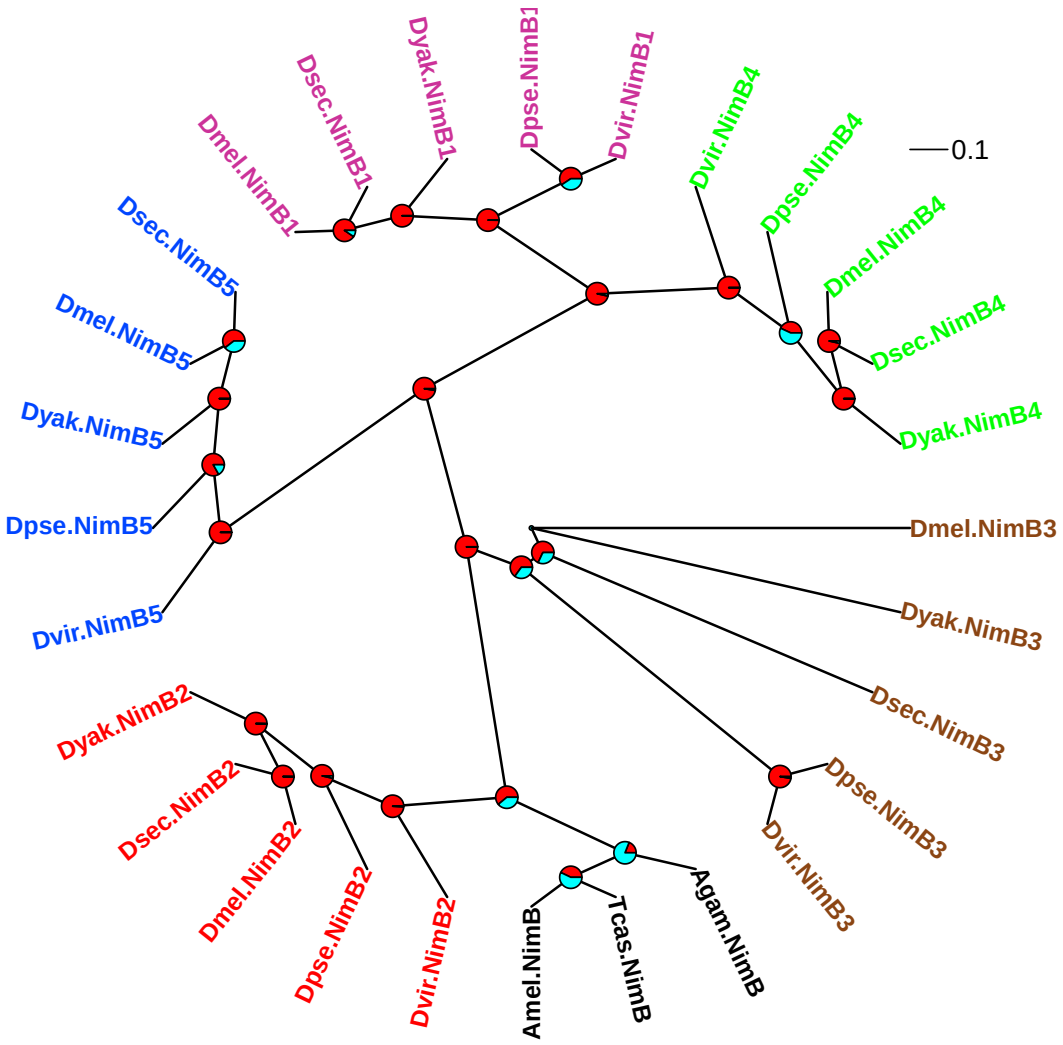
A NimA szekvenciák illesztésében kevés gap volt, így ezek nem voltak számottevően informatívak a topológiára nézve, amit a csak gapinformációk felhasználásával becsült poszterior mintából felépített többségi konszenzusfa is tükrözött. Erre való tekintettel a szekvencia- és gapinformációk felhasználásával becsült NimA filogenetikát nem tárgyaljuk.

A NimB szekvenciák ML módszerrel becsült filogenetikai fáján (4.18. ábra) - ami jól összhangban van az NJ PD módszerrel becsült fával (8.1. függelék) - megfigyelhetjük, hogy a Drosophila NimB ortológok többé-kevésbé jól tükrözik a fajok filogenetikáját, valamint azt, hogy a ((NimB1, NimB4), NimB5) kládok bootstrap támogatottsága magas. Ugyancsak magas a NimB2 és NimB3 paralógokat, valamint a többi fajokból származó NimB szekvenciákat tartalmazó klád támogatottsága, de ezen belül a bootstrap értékek egyes csomópontok esetében jelentős bizonytalanságot sugallnak, így nehéz eldönteni, hogy melyik kládot tekintsük bazálisabbnak.

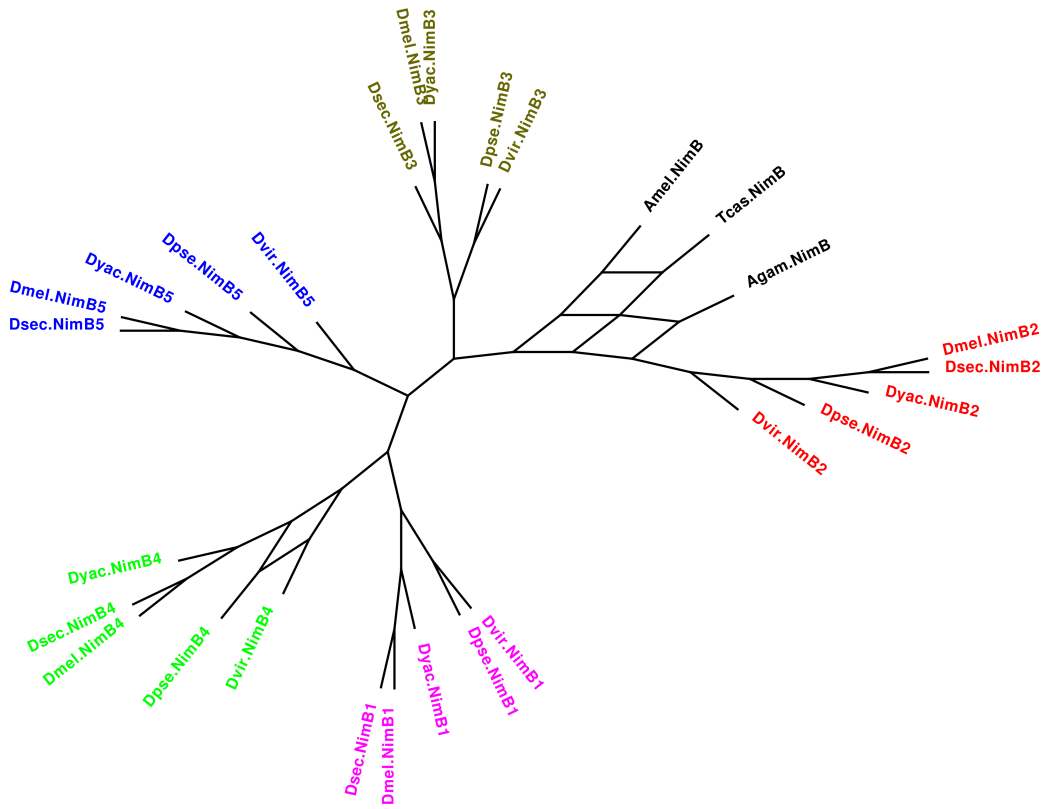
Az NJ CD módszerrel kapott fa topológiája (8.2. függelék) az előzőekben tárgyaltaktól jelentősen eltér. Ezt a gapeket tartalmazó pozíciók figyelmen kívül hagyása miatt fellépő információ-vesztésnek tudjuk be.

A NimB szekvenciák esetében a csak gapinformációk alapján becsült poszterior mintából felépített konszenzushálózat (9.1. függelék) arra utalt, hogy ebben az esetben a gapek a topológiára nézve bizonyos mértékben informatívak. Azonban a csak szekvencia-információk (4.19. ábra) és a szekvencia- és gapinformációk alapján kapott (9.2. függelék) konszenzushálózatokat összehasonlítva arra jutottunk, hogy a gapek bekódolásával kapott karakterek felhasználása a hálózat kritikus részeiben nem hogy csökkentené, de inkább növeli a bizonytalanságot. Ezen okból csak a szekvencia-információk alapján becsült konszenzushálózatot tárgyaljuk.

A szekvencia-információkból Bayes MCMC-vel becsült konszenzushálózat (4.19. ábra) összhangban van az ML módszerrel becsült fa topológiájával, de ennek segítségével jobban felmérhető a NimB2 és NimB3 paralógokat tartalmazó kládon belüli bizonytalanság.



4.18. ábra. A NimB szekvenciák gyökértelen filogenetikájának ML becslése (illesztés: ProbCons 1.1, szoftver: PhyML 3.0, szubsztitúciós modell: VT+Γ + F, gamma paraméter: 2,291, gamma kategóriák száma: 4, bootstrap ismétlések száma: 1000). A kládok bootstrap támogatottsága a megfelelő csomópontban található kördiagramok piros szeletével arányos. A Drosophila ortológokat azonos színnel jelöltük. Az águhosszak egysége egy szubsztitúció pozíciónként. Az ábra az APE R csomag (Paradis és mtsai, 2004) segítségével készült.



4.19. ábra. A NimB szekvenciák filogenetikájának Bayes MCMC rekonstrukciója nyomán kapott poszterior mintákból az $x = 0,2$-es szinten számolt konszenzushálózat. A konszenzushálózat csak a topológiák közötti konfliktus szemléltetésére szolgál, így az ághosszaknak nincs jelentése. A Drosophila ortológokat azonos színnel jelöltük.

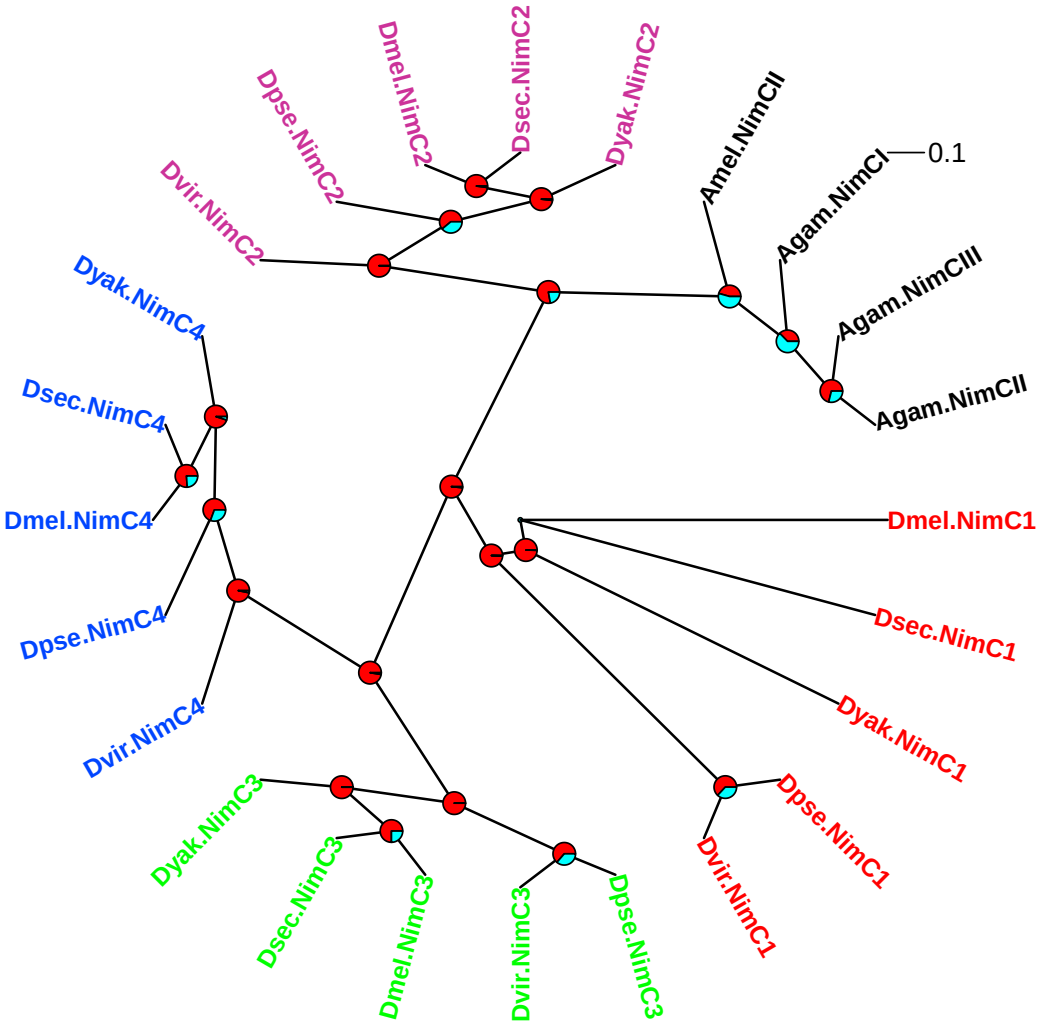
A hálózat alapján arra a következtetésre juthatunk, hogy a Drosophila paralógok közül a NimB2 a legbazálisabb, és a NimB3 szekvenciák a Drosophila fajok leszármazási vonalában, de az Anopheles leszármazási vonaltól való szétválást követően megtörtént duplikációs esemény termékei.

A *D. melanogaster nimrodB* paralógok esetében még nem állnak rendelkezésre funkcionális adatok. A filogenetika ismerete segítséget nyújthat a kísérletes munka fókuszálásához. Eredményeink nyomán elsőként a két bazálisabb klád egy-egy génjének (*nimrodB2* és például *nimrodB4*) géncsendesítéses vizsgálata javasolható.

A NimC szekvenciák ML módszerrel becsült filogenetikájának topológiája (4.20. ábra), valamint az NJ PD algoritmussal felépített topológiák (8.3. függelék) között egyetértés van a ((NimC3, NimC4), NimC1) kládot illetően.

Az ML filogenetikán az említett klád bootstrap támogatottsága nagy, és jelentősebb bizonytalanságokat a NimC2 Drosophila ortológokat, valamint a többi fajokból származó szekvenciákat tartalmazó kládban találunk.

Az NJ PD módszerrel rekonstruált topológia (8.3. függelék) ebben az esetben is a fentiektől eltér, feltehetőleg az információ-veszteség következményeként.

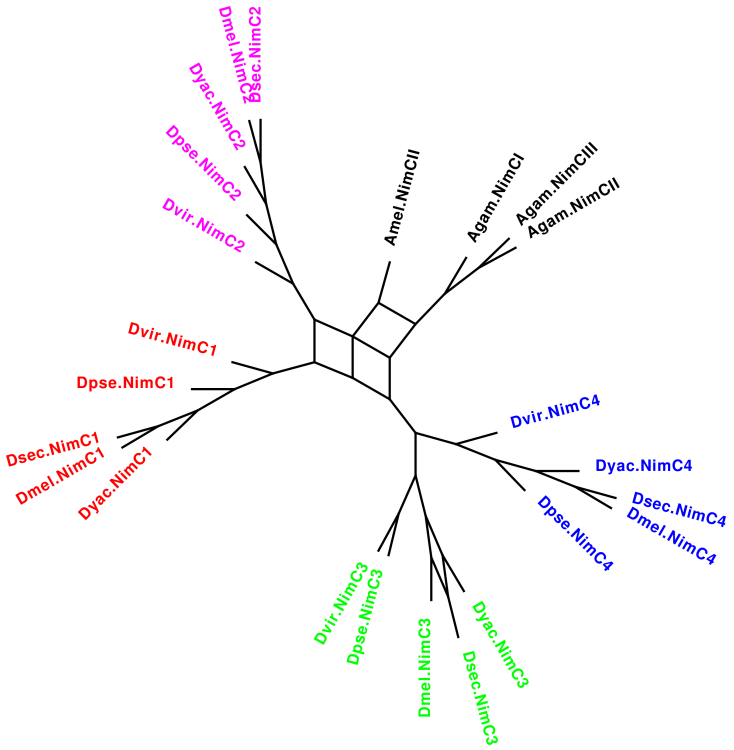


4.20. ábra. A NimC szekvenciák gyökértelen filogenetikájának ML becslése (illesztés: ProbCons 1.1, szoftver: PhyML 3.0, szubsztitúciós modell: VT+Γ + F, gamma paraméter: 2,526, gamma kategóriák száma: 4, bootstrap ismétlések száma: 1000). A kládok bootstrap támogatottsága a megfelelő csomópontban található kördiagramok piros szeletével arányos. A Drosophila ortológokat azonos színnel jelöltük. Az ághosszak egysége egy szubsztitúció pozíciónként. Az ábra az APE R csomag (Paradis és mtsai, 2004) segítségével készült.

A poszterior mintákból számolt konszenzushálózatok (9.3. függelék, 4.21. ábra, 9.4. függelék) összehasonlítása nyomán, a NimC szekvenciák esetében is azt találtuk, hogy a

gapinformációk felhasználása nem csökkentette a topológia becslésében rejlő bizonytalanságot, sőt, okunk van azt feltételezni, hogy ebben az esetben az illesztés alacsonyabb minősége miatt a gapek bekódolásával kapott karakterek felhasználása torzításokat okozhat. Ezen okból a NimC szekvenciák esetében is csak a szekvencia-információk alapján számolt konszenzushálózatot tárgyaljuk.

Ezen konszenzus hálózat (4.21. ábra) teljes mértékben támogatja a (NimC3, NimC4) kládot, de bizonytalanságot sugall a többi paralógokat tartalmazó kládok pozícióját illetően.

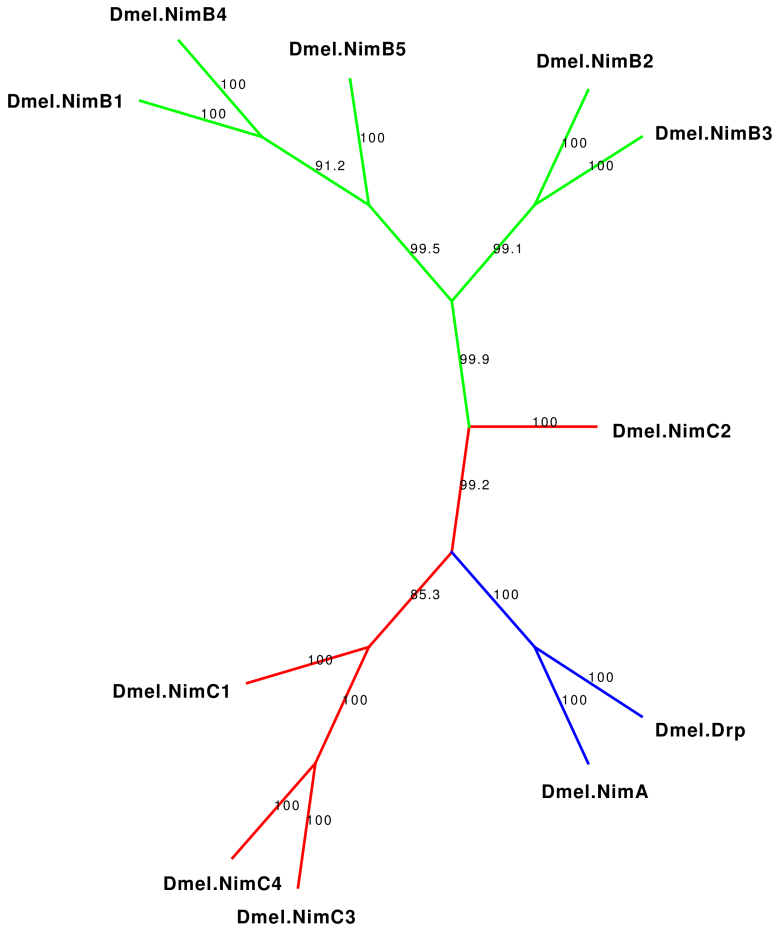


4.21. ábra. A NimC szekvenciák filogenetikájának Bayes MCMC rekonstrukciója nyomán kapott poszterior mintákból az $x = 0,2$-es szinten számolt konszenzushálózat. A konszenzushálózat csak a topológiák közötti konfliktus szemléltetésére szolgál, így az ághosszaknak nincs jelentése. A Drosophila ortológokat azonos színnel jelöltük.

Az *A. gambiae* NimC paralógokra nézve az ML fa és a Bayes MCMC konszenzushálózat topológiája alapján ugyanazt a következtetést vonhatjuk le, mégpedig azt, hogy ezek az Anopheles és Drosophila leszármazási vonalak szétválása után rögzült duplikációk termékei, ezen gének tehát nem rendelkeznek a *D. melanogaster*-ben egy-az-egyhez ortológgokkal. Ennek nyomán érdekes lenne megvizsgálni, hogy ezen gének milyen szerepet játszanak a fagocitózisban az *A. gambiae*-ben.

A *D. melanogaster* Nimrod szekvenciák illesztésének és filogenetikájának együttes Bayes MCMC becslése során kapott poszterior mintából felépített konszenzushálózat (4.22. ábra) nem tartalmaz nagyobb bizonytalanságokat. Ezt jól tükrözik az ágakon látható poszterior

valószínűségek, valamint az, hogy ebben az esetben a konszenzushálózat tulajdonképpen egy fa. A konszenzushálózaton a Draper típusú doménstruktúra széles taxonómiai eloszlása alapján a (NimA, Draper) kládot bazálisnak tekintjük.



4.22. ábra. A *D. melanogaster* Nimrod paralógok illesztésének és filogenetikájának együttes Bayes MCMC becslése nyomán kapott poszterior mintákból az $x = 0,2$-es szinten számolt konszenzushálózat. A konszenzushálózat csak a topológiák közötti konfliktus szemléltetésére szolgál, így az ághosszaknak nincs jelentése. Az ágakon a poszterior valószínűségek láthatók. A **Dmel.Drp** a *D. melanogaster* Draper szekvenciát jelöli. A színezés a NimA/Draper, NimB és NimC kládokat emeli ki.

A konszenzushálózat megerősíti ML módszerrel a családok szintjén becsült filogenetikák esetében kapott (NimC1, (NimC3, NimC4)) és ((NimB1, NimB4), NimB5) kládok létét. A NimB3 paralógok pozícióját illetően azonban a konszenzushálózat konfliktusban van a B családok szintjén rekonstruált ML filogenetikával. A NimB3 szekvenciák rövid voltára, valamint az ML filogenetikán a rájuk vonatkozó bootstrap értékek által sugallt bizonytalanságra való tekintettel ezek topológiában elfoglalt pozícióját bizonytalannak tekintjük.

A konszenzushálózat alapján arra is következtethetünk, hogy a NimB szekvenciák valószínűleg a NimC2 szekvenciákkal egy, a NimC1, NimC3 és NimC4 szekvenciákat

tartalmazó kládtól elkülönülő leszármazási vonal részét képezik.

A Nimrod paralóg szekvenciák közötti divergenciára az ezek párbaállításával készített PRHD-k alapján is megpróbáltunk következtetni a következő elgondolás szerint: abban az esetben, ha a paralógok recens duplikációs események termékei, valamint ha kizárhatjuk az összehangolt evolúció jelenlétét, akkor azt várjuk, hogy a két szekvenciában azonosított pozicionálisan homológ domének közötti ortológia kapcsolatok egy többé-kevésbé létraszerű mintázatot alkossanak.

A paralógok közötti divergencia növekedésével a domének szekvenciáiban felhalmozódott különbségek mértéke egyre jobban megközelíti a szaturáció szintjét, ami a filogenetikai rekonstrukció pontosságának csökkenéséhez vezet, és ennek következményeként az ortológia kapcsolatok által alkotott létraszerű mintázat egyre kevésbé lesz kihangsúlyozott.

Az egyes Drosophila fajokból származó szekvenciák párbaállításával kapott PRHD-k (lásd például a 10. függeléket) elemzése nyomán azt találtuk, hogy az ortológia kapcsolatok által alkotott létraszerű mintázat hangsúlyos volta összhangban van a NimB szekvenciák ML módszerrel rekonstruált filogenetikájával. Ezzel ellentétben a NimC paralógok esetében nem figyeltünk meg semmilyen trendet az ortológia kapcsolatok mintázatára vonatkozóan. Ezt az eredményt az ezen szekvenciák közötti nagyfokú divergenciával magyarázzuk, ami vélhetően ezek korábbi radiációjának tudható be, ahogyan azt a 4.22. ábra sugallja.

A paralógok párbaállításával kapott PRHD-k arra utalnak tehát, hogy viszonylag alacsony divergencia esetén a domén-ismétlődések filogenetikája informatív lehet az őket tartalmazó szekvenciák közötti divergenciára nézve.

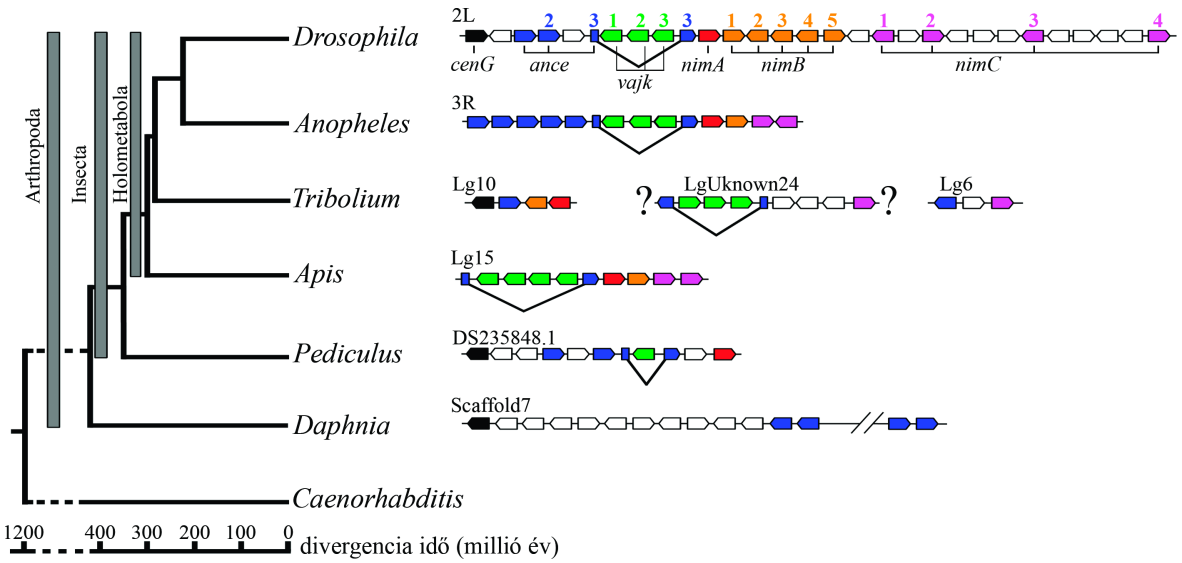
4.4. A Nimrod klaszter konzerváltságának statisztikai szignifikanciája

A Nimrod klasztert a *D. melanogaster*-ben a Nimrod géncsaládokba tartozó géneken kívül az *ance*[11] géncsaládba (Macours és Hens, 2004) tartozó gének, három, a CPF családba (Togawa és mtsai, 2007) tartozó, általunk *Vajk*1, 2, 3-nak nevezett gén és a *centaurin gamma* (*cenG*) gén alkotja (4.23. ábra).

A *Vajk* gének, amennyiben jelen vannak a klaszterben, mindig az egyik *Ance* gén nagyméretű intronjában foglalnak helyet. A *P. humanus* az a *D. melanogaster*-től legtávolabbi faj, melynek genomjában a klaszter hasonló összetételben megtalálható.

A megszekvenált Drosophila genomok túlnyomó részében a klaszter egysége nem bomlik meg (Bhutkar és mtsai, 2008), de a *D. yakuba* és a *D. mojavensis* klaszterben nincs jelen a *cenG*

[11] angiotenzin konvertáló enzimek



4.23. ábra. A Nimrod klaszter szerveződése a vizsgált genomokban (sematikus ábrázolás). A fajok filogenetikáján csak a génuszneveket tüntettük fel. A színezés a különböző géncsaládokba tartozó gének megkülönböztetésére szolgál, a *D. melanogaster* esetében ezek nevét is feltüntettük. Az *ance3* gén esetében a *Vajk* géneket tartalmazó nagyméretű intront is feltüntettük. A divergencia idők forrása: Honeybee Genome Sequencing Consortium (2006).

és két *ance* gén. Ennek alapján úgy tűnik, hogy a *centaurin gamma* a klaszternek egy többé-kevésbé „fakultatív" tagja, különösen, mivel nincs jelen az *A. gambiae* klaszterben, valamint nem azonosítható az ortológja az *A. mellifera* genomban. Mindezek ellenére a *cenG* feltehetően fontos szerepet töltött be a klaszter létrejöttében, mert az ábrán külcsoportként feltüntetett *Daphnia pulex* genomjában is találunk egymás közelében fekvő *ance* és *centaurin gamma* géneket.

A *T. castaneum* esetében a klaszter egy része egy olyan kontigban található, amelynek a genomban elfoglalt pozíciója ismeretlen. Mivel ebben az esetben a klaszter összetételét bizonytalannak tartjuk, a *T. castaneum*-ot a konzerváltság tesztelése során figyelmen kívül hagytuk.

Az *A. gambiae*, *A. mellifera* és *P. humanus* klaszereknek a *D. melanogaster*-ben található klaszterhez viszonyított konzerváltságára vonatkozó, a Raghupathy és Durand (2009) által kidolgozott módszerrel számolt p-értékek a Bonferroni korrekció után is mindhárom esetben kisebbek voltak mint 10^{-6}. Ezen eredményeknek megfelelően elvethető a neutrális, randomizációt feltételező null modell és ésszerű a klaszter konzerváltságát szelekciós tényezőkkel magyarázni.

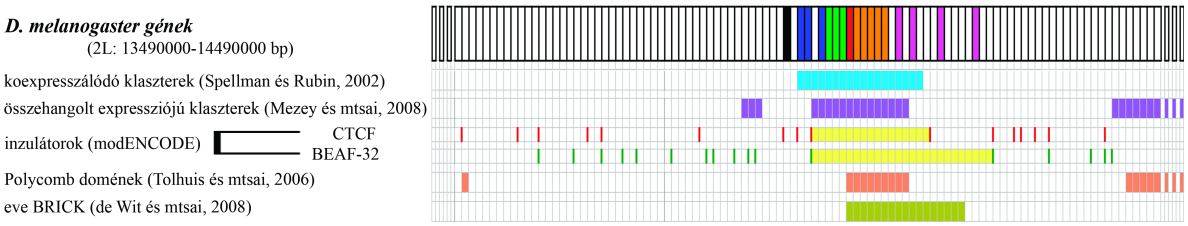
5. fejezet

Az eredmények tárgyalása

5.1. A Nimrod génklaszter integritását fenntartó tényezők

A Nimrod klaszter nagy időléptékekben is jól megőrzi génösszetételét annak ellenére, hogy a gének orientációja nagy változatosságot mutat, ami ezen evolúciós egység fontosságára utal, és egyúttal felveti azt a kérdést, hogy melyek lehetnek a klaszter integritását fenntartó tényezők.

A konzerváltságot fenntartó negatív szelekció motorjai lehetnek például a klaszter génjeinek expressziójában szerepet játszó olyan közös elemek, amelyek a gének közelségét igénylik (Hurst és mtsai, 2004; de Wit és van Steensel, 2009; Koonin, 2009b). Az irodalmi adatok arra utalnak, hogy ez a Nimrod klaszter esetében is egy fontos tényező lehet (5.1. ábra): Spellman és Rubin (2002) átfogó elemzése alapján megállapítható, hogy a Nimrod klaszter „magját” képező gének egy, a *D. melanogaster*-ben koexpressziót mutató klasztert alkotnak. Ezen koexpressziós klasztert alkotó gének egy csoportjának expressziós szintje a Drosophila fajok evolúciója során is korreláltan változott (Mezey és mtsai, 2008), ami ugyancsak egy közös génexpressziót szabályozó elem jelenlétére utal.



5.1. ábra. A Nimrod klaszter közvetlen környezetének sematikus ábrázolása a *D. melanogaster*-ben, néhány idevonatkozó irodalmi adat feltüntetésével. A legfelső sor téglalapjai a körülbelül két megabázisnyi 2L:13490000-14490000 régióban található 102 gént jelképezik. A géneket a 4.23. ábrának megfelelően színkódoltuk. A további sorok a következő, a régió génjeire vonatkozó genomikai adatokat jelképezik: együttesen expresszálódó gének klaszterei (Spellman és Rubin, 2002); a Drosphila fajokban a génexpressziós szint összehangolt evolúcióját mutató klaszterek (Mezey és mtsai, 2008), a különböző méretű mozgó ablakokkal kapott átfedő klasztereket összeolvasztottuk; a CTCF és BEAF-32 fehérjéket kötő inzulátor-elemek hozzávetőleges pozíciója (modENCODE, http://modencode.org), a Nimrod gének többségét tartalmazó inzulátoroktól mentes intervallumok kiemelésével; a Polycomb fehérje kötődési doménjei (Tolhuis és mtsai, 2006); az Eve („Even-skipped”) kötődési helyeit jellemző legnagyobb azonosított „Eve BRICK” régió (de Wit és mtsai, 2008).

A génklaszterek konzerváltságát fenntartó tényezők között jelentős szerepet játszhatnak a kromatinstruktúra szintjén megvalósuló génszabályozási mechanizmusok (Spellman és Rubin, 2002; Hurst és mtsai, 2004; Sproul és mtsai, 2005). A Nimrod klaszter esetében ezen mechanizmusok potenciális szerepére utal az, hogy sejtvonalakon (Tolhuis és mtsai, 2006) és embriókon (de Wit és mtsai, 2008) végzett kísérletek során a klaszterrel jól átfedő, kromatin-kötő fehérjék által meghatározott „doméneket” azonosítottak (5.1. ábra).

Ugyancsak a kromatin-szintű szabályozási mechanizmusok lehetséges szerepére utal egyes,

a kromatin-struktúra szerveződésében részt vevő, úgynevezett „inzulátor”-elemek eloszlása a klaszter környezetében. A Spellman és Rubin (2002) által a régióban azonosított koexpressziós klaszter például jól átfed két CTCF fehérjét kötő inzulátor (Wallace és Felsenfeld, 2007) által közrefogott régióval.

A közös szabályozás alatt álló klaszterek génjei gyakran ugyanazon biológiai folyamatban játszanak szerepet (Lee és Sonnhammer, 2003; Hurst és mtsai, 2004). A klaszterben található *nimrod* gének mellett az *ance* géncsalád (Macours és Hens, 2004) és a *centaurin gamma* gén (Stuart és mtsai, 2007), valamint a *Vajk* gének is kapcsolatba hozhatóak az immunitással. Mindezeket, valamint a klaszter konzerváltságát figyelembe véve elképzelhető, hogy a Nimrod klaszter a rovarok veleszületett immunitásában fontos szerepet játszó „funkcionális modul”, ami kétségtelenül ennek további vizsgálatát indokolja. A legfontosabb irányvonal annak a pontos biológiai folyamatnak a felderítése lenne, amelyben a statisztikai módszerekkel kimutatott koexpresszió szerepet játszik.

5.2. Mintázatok az ismétlődő NIM domének evolúciójában

Az előző fejezetben részletezett eredmények elegendő evidenciát szolgáltatnak a Drosophila Eater szekvenciák középső régiójában található NIM domének összehangolt evolúciójára nézve. A homogenizáció hatása mindegyik szekvenciapár esetében megmutatkozik, ami ennek folyamatosságát sugallja a Drosophila kládon belül. Ha nem feltételezzük a homogenizáció folyamatosságát, akkor az eredmények magyarázatához legalább azt kell feltennünk, hogy a domének mindegyik speciációs esemény után megéltek egy összehangolt epizódot.

Magyarázatra szorul azonban az Eater szekvenciákban található domén-ismétlődésen belül az evolúciós trendekben megmutatkozó „aszimmetria”, vagyis az, hogy az N-terminális régióban a C-terminális régióhoz képest több domén „marad ki” a homogenizációs folyamatokból.

A mechanizmustól (génkonverzió, illetve egyenlőtlen rekombináció) függetlenül az ismétlődések végein található egységek, a velük szomszédos nem homológ szekvenciák jelenléte miatt, kevésbé hajlamosak részt venni a homogenizációs folyamatokban, ahogy ezt jól példázzák a Drosophila *dumpy* gének által kódolt fehérjékben található ismétlődések (Carmon és mtsai, 2007), illetve egyéb repetitív elemek (McAllister és Werren, 1999). A határoló nem homológ szekvenciák jelenléte azonban nem ad megfelelő magyarázatot a fentebb említett aszimmetriára.

A több exon által kódolt szekvenciák esetében az intronok, mint az ismétlődő egységekkel nem homológ szekvenciák hatását is számba kell vennünk. Az Eaterek esetében azonban az ismétlődő egységeket 1-3 domén kivételével egy nagy közös exon kódolja, vagyis az intron-

exon struktúra sem ad kielégítő magyarázatot az aszimmetriára, ezért úgy gondoljuk, hogy ezt szelekciós tényezők okozzák. Ez arra is utal, hogy az Eater szekvenciák esetében az exon-átrendeződések[1], melyeknek általában nagy szerepet tulajdonítanak (Börklund és mtsai, 2006), nem játszanak jelentős szerepet a domén-ismétlődések evolúciójában.

Kocks és mtsai (2005) is észrevették a Drosophila Eater fehérje vizsgálata során, hogy a középső régió repetitív elemei kevésbé változatosak, és kimutatták, hogy az első néhány N-terminális egységnek megfelelő régió elégséges a baktériumkötéshez. Ugyancsak nagyobb számban prediktáltak az N-terminális régióban N-glikozilációs helyeket. Mindezekre alapozva feltételezték, hogy a középső régiónak nincsen szerepe a baktériumkötésben, és mintegy „pálcaként" szolgálva csupán strukturális szerepet tölt be.

A fentiek alapján úgy gondoljuk, hogy az asszimetriát az okozza, hogy az N-terminális régióban található domének a rájuk fokozottabban ható kényszerek miatt nem vesznek részt az összehangolt evolúciós folyamatokban. Ha ugyanis az N-terminális domének közül valamelyiknek, például egy génkonverziós esemény hatására, jelentősen megváltozik a szekvenciája, az nagy eséllyel negatív hatással lesz a fehérje baktériumkötő képességére, ami csökkenti egy ilyen konverziós mutáció rögzülési esélyeit. Ezen magyarázat szerint tehát a funkcionálisan már divergálódott N-terminális domének közötti különbségeket egy erős negatív szelekció tartja fenn.

A távolsági mátrixok kiértékelése nyomán arra a következtetésre jutottunk, hogy az összehangolt evolúció egysége nem minden esetben egy NIM domén. A *D. melanogaster* esetében például minden második domén között figyelhetjük meg a legnagyobb hasonlóságot. Ez arra utal, hogy ebben az esetben a homogenizáció egysége két domén, amit az összekötő szekvenciák közötti hasonlóságok is alátámasztanak.

Az *A. mellifera* NimCI, a *T. castaneum* NimCI és kisebb mértékben a *T. castaneum* NimCII szekvencia esetében a domén-filogenetikák alapján felmerül ugyan annak a gyanúja, hogy a domének az összehangolt modellnek megfelelően viselkednek, azonban közeli ortológok hiányában, csak a távolsági mátrixokra alapozva, ezt nem tudjuk megerősíteni.

A Drosophila NimC1 szekvenciáknál a domén-filogenetika és a távolsági mátrixok nem utalnak összehangolt evolúcióra. A PRHD-k esetében az ortológia kapcsolatok által alkotott létraszerű mintázat hangsúlyos volta összefüggést mutat a divergenciával. Ezért itt a létraszerű mintázat megbomlását nem az összehangolt evolúció hatásával, hanem a NimC1 ortológok nagyobb szubsztitúciós rátájával magyarázzuk. Ezt megerősíti az, hogy a NimC szekvenciák ML módszerrel becsült filogenetikáján (4.20. ábra) a NimC1 kládban az águhosszak a többi NimC ortológot tömörítő kládban megfigyelt águhosszakhoz képest jelentősen nagyobbak.

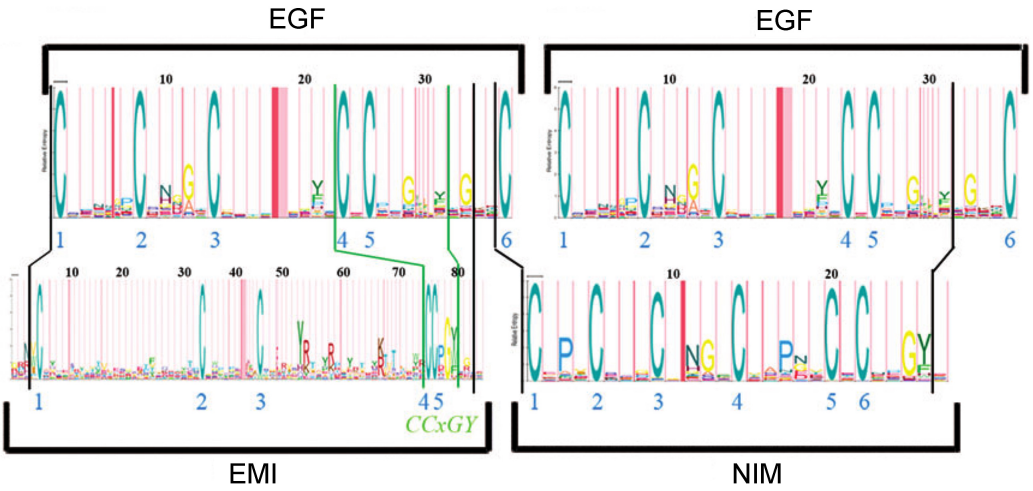
Azt, hogy az Eater szekvenciákhoz hasonló méretű domén-ismétlődéseket tartalmazó egyéb NimC típusú Drosophila szekvenciák esetében miért nem találunk összehangolt evolúcióra

[1] „exon shuffling"

utaló jeleket, több tényezővel is magyarázhatjuk, mint amilyen például a genomban elfoglalt pozíció vagy az eltérő intron-exon struktúra. Ezek közül a *nimrod* gének klaszterizációja nyomán a genomi pozíciót tekintjük a legfontosabbnak, különösen mivel a lokális kromatin-struktúráról bebizonyították, hogy hatással van a génkonverzióra (Cummings és mtsai, 2007). Az irodalomban elterjedtek az összehangolt evolúció esetleges adaptív voltára vonatkozó spekulációk (lásd például Swanson és Vacquier, 1998; Johannesson és mtsai, 2005; Paoletti és mtsai, 2007). Ennek lehetőségét a jövőben az Eaterek esetében is érdemes lehet megvizsgálni, de megfelelő evidencia hiányában egyelőre az előző, neutrális magyarázatot részesítjük előnyben.

5.3. A Nimrod szupercsalád evolúciós története

Az EMI és EGF-szerű doméneket modellező pHMM-ek illesztése (4.2. **B** ábra) megfelelő mértékben alátámasztja az EMI domén részét képező CCxGY/W motívum és az EGF-szerű domén egy régiójának homológiáját. Az EMI és EGF-szerű domének pHMM logóit összehasonlítva egyéb hasonlóságokat is felfedezhetünk a két domén fennmaradó részében (5.2. ábra): az illesztés alapján homológnak tekinthető régiók előtti részben mindkét domén tartalmaz három konzervált ciszteint, valamint egy többé-kevésbé konzervált, tirozint tartalmazó pozíciót.



5.2. ábra. Az EMI és a NIM domének eredetére vonatkozó hipotézis szemléltetése pHMM logók segítségével. A kék színű számozás a konzervált ciszteineket, míg fekete színű a pHMM-ben elfoglalt pozíciót jelöli.

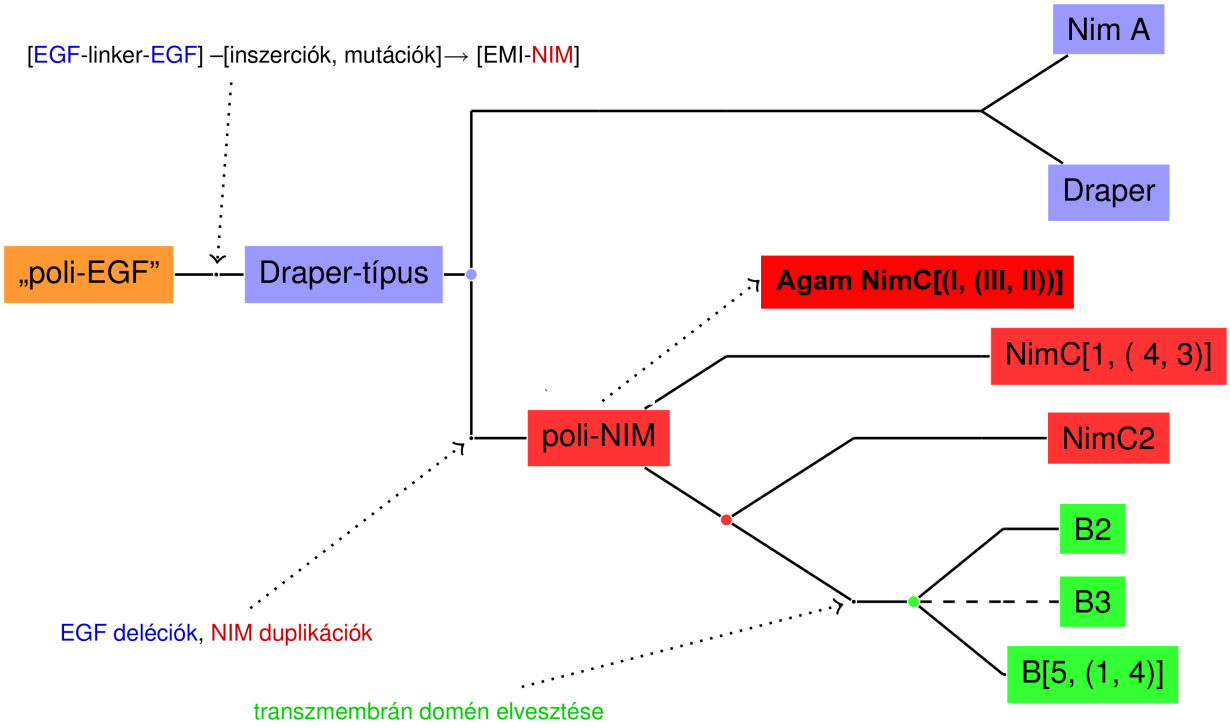
Az EGF-szerű és NIM domének illesztéséből arra következtethetünk, hogy a NIM domén első két konzervált ciszteinje közötti régió nem homológ az EGF-szerű doménnel, ami megmagyarázza a Kurucz és mtsai (2007) által megfigyelt „eltolódást”.

Mindezen megfigyelések azt sugallják, hogy az EMI domén és a NIM domén két EGF-szerű doménből és az őket összekötő rövid összekötő régióból jött létre egy strukturális átrendeződés révén, ahogy azt az 5.2. ábra szemlélteti.

Ez a strukturális átrendeződés nyilvánvalóan mutációs, inszerciós és deléciós eseményeket feltételez. Ahogy az az ábrán is látható, az EMI domén az EGF-szerű doménnél jelentősen hosszabb, így feltételezhetjük, hogy az inszerciók ebben a folyamatban kiemelkedő szerepet töltöttek be. Ez összhangban van Jiang és Blouin (2007) következtetéseivel, ami szerint a változékony régiókban bekövetkezett egymásba ágyazott inszerciós események a domén-evolúció fontos tényezői.

A fenti hipotézist az is támogatja, hogy a taxonómiai eloszlása alapján ősibbnek tekinthető Draper-típusú domén-architektúra a CCxGY/W motívumot követően csak egy NIM domént tartalmaz, melyet változó számú EGF ismétlődések követnek. Ez arra utal, hogy ez a domén-architektúra egy EGF-szerű doménismétlődést és egy transzmembrán régiót is tartalmazó „poli-EGF" architektúra leszármazottja.

A NIM domén duplikációjával és a „klasszikus" EGF-szerű domének elvesztésével létrejöhetett a NimC típusú, majd a NimC2 leszármazási vonalon a transzmembrán domén elvesztése révén a NimB típusú architektúra (5.3. ábra). Ezt a hipotézist a különböző családokat modellező pHMM-ek páros logók általi, az eredmények ismertetése során nem említett illesztései is támogatják.



5.3. ábra. A Nimrod szupergéncsalád evolúciójával kapcsolatos hipotéziseink összefoglalása.

A szupercsalád evolúciója kétségtelenül a BD konceptuális modellel írható le legjobban. A szupercsalád tagjainak funkciójára nézve rendelkezésre álló adatok arra utalnak, hogy a duplikációk révén megjelenő új gének funkciója mind köthető ugyan a veleszületett immunitáshoz és a fagocitózishoz, de a *nimrod* gének az említett folyamatok különböző lépéseiben vesznek részt. Ez azt valószínűsíti, hogy a szupercsalád evolúciója során a neo- és szubfunkcionalizáció volt a duplikációval rögzült gének leggyakoribb sorsa. Ezek az események feltehetően fontos szerepet játszott az ízeltábúak veleszületett immunitásának evolúciójában azáltal, hogy létrehoztak egy változatos ligand-specificitással rendelkező receptor-készletet.

A NIM domének eredetére vonatkozó forgatókönyv egy általánosabb hipotézist is felvet az aminosav-szekvenciákban található ismétlődő elemek evolúciójára nézve.

Az új típusú ismétlődő elemek aminosav-szekvenciákban való megjelenésének általános mechanizmusa egy még nem teljesen tisztázott folyamat. Andrade és mtsai (2001) például tárgyalnak egy erre vonatkozó problémát: ha feltételezzük, hogy az ismétlődések esetében a „kooperativ folding”-nak nagy jelentősége van, akkor valószínűtlennek tűnik egy eltérő típusú egység megjelenése az ismétlődéseket tartalmazó régióban.

A szerzők ezt az ellentmondást úgy próbálják feloldani, hogy azt feltételezik, hogy az egységek még különálló oligomerekként vannak jelen, amikor az új típusú egységek megjelennek, de azt is megjegyzik, hogy erre nézve egy példa sem ismeretes. A NIM domének példájára alapozva azonban egyszerűbben is feloldhatjuk a fentebb ismertetett ellentmondást: ha feltételezzük is a kooperatív folding hatását, annak jelentősége kisebb lehet az ismétlődések szélső egységei esetében, így ezekben a jelentősebb szerkezeti módosulásokat okozó mutációk nagyobb eséllyel rögzülhetnek.

Ugyancsak érdekesnek bizonyulhat a NIM domének eredetére vonatkozó hipotézist Lavorgna és mtsai (2001) eredményeinek fényében vizsgálni, a szerzők ugyanis pozitív összefüggést találtak egy organizmus „komplexitása” és a genomjában található, ismétlődéseket tartalmazó proteinek aránya között. Ennek tárgyalása során a szerzők az ismétlődéseknek a komplexitás evolúciójában betöltött szerepét hangsúlyozzák, a népszerű „barkácsoló evolúció” metaforáját felhasználva. Ennek relevanciáját nem nehéz belátni, ugyanis az új domének kialakulásának esetében egy már valamilyen harmadlagos struktúrával rendelkező ismétlődés jobb „nyersanyagnak” bizonyulhat, mint egy többé-kevésbé „véletlenszerű” szekvencia (lásd például Söding, 2005).

A NIM és EMI domének eredete jól példázza a domén-ismétlődések „rekrutálását”[2] az új domének kialakulása során. Ennek nyomán érdekes lehet az egyes proteómok szintjén megvizsgálni a doménismétlődéseket tartalmazó doménszerkezetekben az ismétlődő egység és az ugyanazon doménszerkezetben előforduló egyéb domének közötti hasonlóságot.

[2] „recruiting”, „exaptation”

Mindezek összefoglalásaként elmondhatjuk, hogy eredményeink nyomán összetett kép rajzolódik ki a Nimrod szupercsalád evolúciójáról, ami jól példázza a géncsaládok szerepét a bonyolult genetikai rendszerek kialakulásában. Reméljük, hogy eredményeink hasznosnak bizonyulnak majd a szupercsalád génjeire, és így a rovarok veleszületett immunitására irányuló további vizsgálatok során.

Köszönetnyilvánítás

Mindenek előtt köszönettel tartozom témavezetőmnek, Dr. Pénzes Zsolt egyetemi docensnek és munkatársamnak, Dr. Somogyi Kálmánnak. Munkájuk és támogatásuk nélkül ez a dolgozat nem jöhetett volna létre.

Ugyancsak köszönettel tartozom a Dr. Andó Istvánnak és csoportjának, valamint Dr. Dan Hultmarknak a gyümölcsöző együttműködésért, és a Molekuláris Biodiverzitás Csoport jelenlegi és volt tagjainak, különösen Bihari Péternek, a dolgozat megírása során nyújtott segítségért.

Szeretném még megköszönni Dr. Rákhely Gábornak, Dr. Leitgeb Balázsnak és Kerényi Ádámnak a linux klaszter használatához nyújtott segítséget, Páll Szilárdnak az algoritmikai kérdésekben adott tanácsokat, valamint Dr. Kós Péternek, Dr. Papp Balázsnak és Dr. Gyurkovics Henriknek a hasznos észrevételeket.

Szeretném megköszönni a Magyar Ösztöndíj Bizottságnak és a romániai CNBSS-nek (Centrul Naţional pentru Burse de Studii în Străinătate), hogy magyarországi PhD-tanulmányaimat finanszírozta, valamint az SZBK Genetikai Intézetének a támogatást.

Végül, de nem utolsósorban köszönettel tartozom szüleimnek, akik tanulmányaim során folyamatosan támogattak és ösztönöztek.

Irodalomjegyzék

Abascal, F., Zardoya, R., és Posada, D. (2005). ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, 21(9):2104–5.

Alfaro, M. E. és Holder, M. (2006). The Posterior and the Prior in Bayesian Phylogenetics. *Annu Rev Ecol Sys*, 37:19–42.

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., és Ronquist, F. (2004). Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, 20(3):407–15.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., és Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–10.

Andrade, M. A., Perez-Iratxeta, C., és Ponting, C. P. (2001). Protein repeats: structures, functions, and evolution. *J Struct Biol*, 134(2-3):117–31.

Baldauf, S. L. (2003). Phylogeny for the faint of heart: a tutorial. *Trends Genet*, 19(6):345–51.

Bhutkar, A., Schaeffer, S. W., Russo, S. M., Xu, M., Smith, T. F., és Gelbart, W. M. (2008). Chromosomal rearrangement inferred from comparisons of 12 Drosophila genomes. *Genetics*, 179(3):1657–80.

Bork, P. (1991). Shuffled domains in extracellular proteins. *FEBS Lett*, 286(1-2):47–54.

Bork, P., Downing, A. K., Kieffer, B., és Campbell, I. D. (1996). Structure and distribution of modules in extracellular proteins. *Q Rev Biophys*, 29(2):119–67.

Börklund, A. K., Ekman, D., és Elofsson, A. (2006). Expansion of protein domain repeats. *PLoS Comput Biol*, 2(8):e114.

Callebaut, I., Mignotte, V., Souchet, M., és Mornon, J. P. (2003). EMI domains are widespread and reveal the probable orthologs of the Caenorhabditis elegans CED-1 protein. *Biochem Biophys Res Commun*, 300(3):619–23.

Carmon, A., Wilkin, M., Hassan, J., Baron, M., és MacIntyre, R. (2007). Concerted evolution within the Drosophila dumpy gene. *Genetics*, 176(1):309–25.

Casola, C. és Hahn, M. W. (2009). Gene conversion among paralogs results in moderate false detection of positive selection using likelihood methods. *J Mol Evol*, 68(6):679–87.

Chen, J.-M., Cooper, D. N., Chuzhanova, N., Férec, C., és Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet*, 8(10):762–75.

Chen, K., Durand, D., és Farach-Colton, M. (2000). NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J Comput Biol*, 7(3-4):429–47.

Csűrös, M. és Miklós, I. (2009). Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. *Mol Biol Evol*, 26(9):2087–95.

Cummings, W. J., Yabuki, M., Ordinario, E. C., Bednarski, D. W., Quay, S., és Maizels, N. (2007). Chromatin structure regulates gene conversion. *PLoS Biol*, 5(10):e246.

Cuttell, L., Vaughan, A., Silva, E., Escaron, C. J., Lavine, M., Van Goethem, E., Eid, J.-P., Quirin, M., és Franc, N. C. (2008). Undertaker, a Drosophila Junctophilin, links Draper-mediated phagocytosis and calcium homeostasis. *Cell*, 135(3):524–34.

Das, S., Nikolaidis, N., Klein, J., és Nei, M. (2008). Evolutionary redefinition of immunoglobulin light chain isotypes in tetrapods using molecular markers. *Proc Natl Acad Sci U S A*, 105(43):16647–52.

de Wit, E., Braunschweig, U., Greil, F., Bussemaker, H. J., és van Steensel, B. (2008). Global chromatin domain organization of the Drosophila genome. *PLoS Genet*, 4(3):e1000045.

de Wit, E. és van Steensel, B. (2009). Chromatin domains in higher eukaryotes: insights from genome-wide mapping studies. *Chromosoma*, 118(1):25–36.

Do, C. B., Mahabhashyam, M. S. P., Brudno, M., és Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*, 15(2):330–40.

Drosophila 12 Genomes Consortium (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167):203–18.

Durbin, R., Eddy, S. R., Krogh, A., és Mitchison, G. (1998). *Biological sequence analysis*. Cambridge University Press, Cambridge, UK.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–7.

Eickbush, T. H. és Eickbush, D. G. (2007). Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics*, 175(2):477–85.

Evans, J. D., Aronstein, K., Chen, Y. P., Hetru, C., Imler, J.-L., Jiang, H., Kanost, M., Thompson, G. J., Zou, Z., és Hultmark, D. (2006). Immune pathways and defence mechanisms in honey bees Apis mellifera. *Insect Mol Biol*, 15(5):645–56.

Felsenstein, J. (1985). Confidence Limits on Phylogenies An Approach Using the Bootstrap. *Evolution*, 39(4):783–791.

Felsenstein, J. (2001). The Troubled Growth of Statistical Phylogenetics. *Syst Biol*, 50(4):465–467.

Felsenstein, J. (2002). *Inferring Phylogenies*. Sinauer Associates Inc., Sunderland, MA, USA.

Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. L., és Bateman, A. (2006). Pfam: clans, web tools and services. *Nucleic Acids Res*, 34(Database issue):D247–51.

Fitch, W. M. (2000). Homology a personal view on some of the problems. *Trends Genet*, 16(5):227–31.

Futuyma, D. J. (2005). *Evolution*. Sinauer Associates, Sunderland, MA, USA.

Gascuel, O. és Steel, M. (2006). Neighbor-joining revealed. *Mol Biol Evol*, 23(11):1997–2000.

Gelman, A. és Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511.

Govind, S. és Nehm, R. H. (2004). Innate immunity in fruit flies: a textbook example of genomic recycling. *PLoS Biol*, 2(8):E276.

Guindon, S. és Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5):696–704.

Hahn, M. W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered*, 100(5):605–17.

Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C., és Cristianini, N. (2005). Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res*, 15(8):1153–60.

Hamon, Y., Trompier, D., Ma, Z., Venegas, V., Pophillat, M., Mignotte, V., Zhou, Z., és Chimini, G. (2006). Cooperation between engulfment receptors: the case of ABCA1 and MEGF10. *PLoS One*, 1:e120.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., és Ira, G. (2009). Mechanisms of change in gene copy number. *Nat Rev Genet*, 10(8):551–64.

Higgins, D. G. (2003). *Multiple alignment*, pages 45–71. Cambridge University Press.

Holder, M. és Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nat Rev Genet*, 4(4):275–84.

Holland, B. és Moulton, V. (2003). *Consensus Networks: A Method for Visualising Incompatibilities in Collections of Trees*, volume 2812 of *Lecture Notes in Computer Science*, pages 165–176. Springer, Berlin/Heidelberg.

Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., Wincker, P., Clark, A. G., Ribeiro, J. M. C., Wides, R., Salzberg, S. L., Loftus, B., Yandell, M., Majoros, W. H., Rusch, D. B., Lai, Z., Kraft, C. L., Abril, J. F., Anthouard, V., Arensburger, P., és további 208 szerző (2002). The genome sequence of the malaria mosquito Anopheles gambiae. *Science*, 298(5591):129–49.

Honeybee Genome Sequencing Consortium (2006). Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, 443(7114):931–49.

Huelsenbeck, J. P. és Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–5.

Hughes, A. L. (1998). Protein phylogenies provide evidence of a radical discontinuity between arthropod and vertebrate immune systems. *Immunogenetics*, 47(4):283–96.

Hughes, A. L. (1999). Concerted evolution of exons and introns in the MHC-linked tenascin-X gene of mammals. *Mol Biol Evol*, 16(11):1558–67.

Hurst, L. D., Pál, C., és Lercher, M. J. (2004). The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet*, 5(4):299–310.

Huson, D. H. és Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*, 23(2):254–67.

Innan, H. (2009). Population genetic models of duplicated genes. *Genetica*, 137(1):19–37.

Janeway, C. A. és Medzhitov, R. (2002). Innate immune recognition. *Annu Rev Immunol*, 20:197–216.

Jiang, H. és Blouin, C. (2007). Insertions and the emergence of novel protein structure: a structure-based phylogenetic study of insertions. *BMC Bioinformatics*, 8:444.

Johannesson, H., Townsend, J. P., Hung, C.-Y., Cole, G. T., és Taylor, J. W. (2005). Concerted evolution in the repeats of an immunomodulating cell surface protein, SOWgp, of the human pathogenic fungi Coccidioides immitis and C. posadasii. *Genetics*, 171(1):109–17.

Ju, J. S., Cho, M. H., Brade, L., Kim, J. H., Park, J. W., Ha, N.-C., Söderhäll, I., Söderhäll, K., Brade, H., és Lee, B. L. (2006). A novel 40-kDa protein containing six repeats of an epidermal growth factor-like domain functions as a pattern recognition protein for lipopolysaccharide. *J Immunol*, 177(3):1838–45.

Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J., és Mclnerney, J. O. (2006). Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol*, 6:29.

Kimbrell, D. A. és Beutler, B. (2001). The evolution and genetics of innate immunity. *Nat Rev Genet*, 2(4):256–67.

Kocks, C., Cho, J. H., Nehme, N., Ulvila, J., Pearson, A. M., Meister, M., Strom, C., Conto, S. L., Hetru, C., Stuart, L. M., Stehle, T., Hoffmann, J. A., Reichhart, J.-M., Ferrandon, D., Rämet, M., és Ezekowitz, R. A. B. (2005). Eater, a transmembrane protein mediating phagocytosis of bacterial pathogens in Drosophila. *Cell*, 123(2):335–46.

Koonin, E. V. (2009a). Darwinian evolution in the light of genomics. *Nucleic Acids Res*, 37(4):1011–34.

Koonin, E. V. (2009b). Evolution of genome architecture. *Int J Biochem Cell Biol*, 41(2):298–306.

Korbel, J. O., Kim, P. M., Chen, X., Urban, A. E., Weissman, S., Snyder, M., és Gerstein, M. B. (2008). The current excitement about copy-number variation: how it relates to gene duplications and protein families. *Curr Opin Struct Biol*, 18(3):366–74.

Kumar, S. és Filipski, A. (2007). Multiple sequence alignment: in pursuit of homologous DNA positions. *Genome Res*, 17(2):127–35.

Kumar, S., Tamura, K., és Nei, M. (2004). MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform*, 5(2):150–63.

Kurant, E., Axelrod, S., Leaman, D., és Gaul, U. (2008). Six-microns-under acts upstream of Draper in the glial phagocytosis of apoptotic neurons. *Cell*, 133(3):498–509.

Kurucz, E., Márkus, R., Zsámboki, J., Folkl-Medzihradszky, K., Darula, Z., Vilmos, P., Udvardy, A., Krausz, I., Lukacsovich, T., Gateff, E., Zettervall, C.-J., Hultmark, D., és Andó, I. (2007). Nimrod, a putative phagocytosis receptor with EGF repeats in Drosophila plasmatocytes. *Curr Biol*, 17(7):649–54.

Lavorgna, G., Patthy, L., és Boncinelli, E. (2001). Were protein internal repeats formed by "bricolage"? *Trends Genet*, 17(3):120–3.

Lee, J. M. és Sonnhammer, E. L. L. (2003). Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res*, 13(5):875–82.

Letunic, I. és Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–8.

Liao, D. (1999). Concerted evolution: molecular mechanism and biological implications. *Am J Hum Genet*, 64(1):24–30.

Liò, P. és Goldman, N. (1998). Models of molecular evolution and phylogeny. *Genome Res*, 8(12):1233–44.

Litman, G. W., Cannon, J. P., és Dishaw, L. J. (2005). Reconstructing immune phylogeny: new perspectives. *Nat Rev Immunol*, 5(11):866–79.

Lunter, G., Miklós, I., Drummond, A., Jensen, J. L., és Hein, J. (2005). Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics*, 6:83.

Lynch, M. és Conery, J. S. (2003). The evolutionary demography of duplicate genes. *J Struct Funct Genomics*, 3(1-4):35–44.

Macours, N. és Hens, K. (2004). Zinc-metalloproteases in insects: ACE and ECE. *Insect Biochem Mol Biol*, 34(6):501–10.

Madera, M. (2005). PhD értekezés, 4. fejezet.

Madera, M. (2008). Profile Comparer: a program for scoring and aligning profile hidden Markov models. *Bioinformatics*, 24(22):2630–1.

Manaka, J., Kuraishi, T., Shiratsuchi, A., Nakai, Y., Higashida, H., Henson, P., és Nakanishi, Y. (2004). Draper-mediated and phosphatidylserine-independent phagocytosis of apoptotic cells by Drosophila hemocytes/macrophages. *J Biol Chem*, 279(46):48466–76.

Mangahas, P. M. és Zhou, Z. (2005). Clearance of apoptotic cells in Caenorhabditis elegans. *Semin Cell Dev Biol*, 16(2):295–306.

Mano, S. és Innan, H. (2008). The evolutionary rate of duplicated genes under concerted evolution. *Genetics*, 180(1):493–505.

McAllister, B. F. és Werren, J. H. (1999). Evolution of tandemly repeated sequences: What happens at the end of an array? *J Mol Evol*, 48(4):469–81.

Mezey, J. G., Nuzhdin, S. V., Ye, F., és Jones, C. D. (2008). Coordinated evolution of co-expressed gene clusters in the Drosophila transcriptome. *BMC Evol Biol*, 8:2.

Morgenstern, B. (1999). DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15(3):211–8.

Nei, M. (2005). Selectionism and neutralism in molecular evolution. *Mol Biol Evol*, 22(12):2318–42.

Nei, M. és Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, 3(5):418–26.

Nei, M., Rogozin, I. B., és Piontkivska, H. (2000). Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proc Natl Acad Sci U S A*, 97(20):10866–71.

Nei, M. és Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. *Annu Rev Genet*, 39:121–52.

Nishikawa, T. és Natori, S. (2001). Targeted disruption of a pupal hemocyte protein of Sarcophaga by RNA interference. *Eur J Biochem*, 268(20):5295–9.

Notredame, C., Higgins, D. G., és Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–17.

Nylander, J. A. A., Wilgenbusch, J. C., Warren, D. L., és Swofford, D. L. (2008). AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics*, 24(4):581–3.

Ohno, S. (1970). *Evolution by gene duplication*. Springer-Verlag, Berlin.

Orengo, C. A. és Thornton, J. M. (2005). Protein families and their evolution-a structural perspective. *Annu Rev Biochem*, 74:867–900.

Ota, T. és Nei, M. (1994). Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. *Mol Biol Evol*, 11(3):469–82.

Paoletti, M., Saupe, S. J., és Clavé, C. (2007). Genesis of a fungal non-self recognition repertoire. *PLoS One*, 2(3):e283.

Paradis, E., Claude, J., és Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290.

Philippe, H., Sörnhaus, U., Baroin, A., Perasso, R., Gasse, F., és Adouette, A. (1994). Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. *J Evol Biol*, 7:247–265.

Plummer, M., Best, N., Cowles, K., és Vines, K. (2009). *coda: Output analysis and diagnostics for MCMC*. R package version 0.13-4.

Posada, D. és Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst Biol*, 53(5):793–808.

Posada, D. és Crandall, K. A. (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics*, 14(9):817–8.

Quesada, H., Ramos-Onsins, S., és Aguadé, M. (2005). Birth-and-death evolution of the Cecropin multigene family in Drosophila. *J Mol Evol*, 60(1):1–11.

Raghupathy, N. és Durand, D. (2009). Gene cluster statistics with gene families. *Mol Biol Evol*, 26(5):957–68.

Redelings, B. D. és Suchard, M. A. (2005). Joint Bayesian estimation of alignment and phylogeny. *Syst Biol*, 54(3):401–18.

Reiczigel, J., Harnos, A., és Solymosi, N. (2007). *Biostatisztika nem statisztikusoknak*. Pars könyvek, Budapest, HU.

Ridley, M. (1996). *Evolution*. Blackwell Science Inc.

Roach, J. C., Glusman, G., Rowen, L., Kaur, A., Purcell, M. K., Smith, K. D., Hood, L. E., és Aderem, A. (2005). The evolution of vertebrate Toll-like receptors. *Proc Natl Acad Sci U S A*, 102(27):9577–82.

Russo, C. A., Takezaki, N., és Nei, M. (1995). Molecular phylogeny and divergence times of drosophilid species. *Mol Biol Evol*, 12(3):391–404.

Saitou, N. és Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–25.

Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Mol Biol Evol*, 6(5):526–38.

Schuster-Böckler, B. és Bateman, A. (2005). Visualizing profile-profile alignment: pairwise HMM logos. *Bioinformatics*, 21(12):2912–3.

Schuster-Böckler, B., Schultz, J., és Rahmann, S. (2004). HMM Logos for visualization of protein families. *BMC Bioinformatics*, 5:7.

Simmons, M. P. és Ochoterena, H. (2000). Gaps as characters in sequence-based phylogenetic analyses. *Syst Biol*, 49(2):369–81.

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7):951–60.

Spellman, P. T. és Rubin, G. M. (2002). Evidence for large domains of similarly expressed genes in the Drosophila genome. *J Biol*, 1(1):5.

Sproul, D., Gilbert, N., és Bickmore, W. A. (2005). The role of chromatin structure in regulating the expression of clustered genes. *Nat Rev Genet*, 6(10):775–81.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., és Birney, E. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–8.

Strimmer, K. és von Haeseler, A. (1997). Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc Natl Acad Sci U S A*, 94(13):6815–9.

Stuart, L. M., Boulais, J., Charriere, G. M., Hennessy, E. J., Brunet, S., Jutras, I., Goyette, G., Rondeau, C., Letarte, S., Huang, H., Ye, P., Morales, F., Kocks, C., Bader, J. S., Desjardins, M., és Ezekowitz, R. A. B. (2007). A systems biology analysis of the Drosophila phagosome. *Nature*, 445(7123):95–101.

Studier, J. A. és Keppler, K. J. (1988). A note on the neighbor-joining algorithm of Saitou and Nei. *Mol Biol Evol*, 5(6):729–31.

Suchard, M. A. és Redelings, B. D. (2006). BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*, 22(16):2047–8.

Sugino, R. P. és Innan, H. (2006). Selection for more of the same product as a force to enhance concerted evolution of duplicated genes. *Trends Genet*, 22(12):642–4.

Swanson, W. J. és Vacquier, V. D. (1998). Concerted evolution in an egg receptor for a rapidly evolving abalone sperm protein. *Science*, 281(5377):710–2.

Swofford, D. L. (2003). *PAUP*: phylogenetic analysis using parsimony (* and other methods) Version 4*. Sinauer Associates Inc., Sunderland, MA, USA.

Takeda, K. és Akira, S. (2007). Toll-like receptors. *Curr Protoc Immunol*, Chapter 14:Unit 14.12.

Teshima, K. M. és Innan, H. (2004). The effect of gene conversion on the divergence between duplicated genes. *Genetics*, 166(3):1553–60.

Thompson, J. D., Higgins, D. G., és Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80.

Thompson, J. D., Plewniak, F., és Poch, O. (1999). A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res*, 27(13):2682–90.

Togawa, T., Augustine Dunn, W., Emmons, A. C., és Willis, J. H. (2007). CPF and CPFL, two related gene families encoding cuticular proteins of Anopheles gambiae and other insects. *Insect Biochem Mol Biol*, 37(7):675–88.

Tolhuis, B., de Wit, E., Muijrers, I., Teunissen, H., Talhout, W., van Steensel, B., és van Lohuizen, M. (2006). Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster. *Nat Genet*, 38(6):694–9.

Wallace, J. A. és Felsenfeld, G. (2007). We gather together: insulators and genome organization. *Curr Opin Genet Dev*, 17(5):400–7.

Walsh, B. (2003). Population-genetic models of the fates of duplicate genes. *Genetica*, 118(2-3):279–94.

Wegner, K. M. (2008). Clustering of Drosophila melanogaster immune genes in interplay with recombination rate. *PLoS One*, 3(7):e2835.

Whelan, S., Liò, P., és Goldman, N. (2001b). Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet*, 17(5):262–72.

Wiley, E. O., Siegel-Causey, D., Brooks, D. R., és Funk, V. A. (1991). *The compleat cladist - A Primer of Phylogenetic Procedures*. The University of Kansas, Museum of Natural History, Lawrence, USA.

Wong, K. M., Suchard, M. A., és Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science*, 319(5862):473–6.

Xia, X. és Xie, Z. (2001). DAMBE: software package for data analysis in molecular biology and evolution. *J Hered*, 92(4):371–3.

Yang, Z., Goldman, N., és Friday, A. (1995). Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst Biol*, 44:384–399.

Zdobnov, E. M. és Bork, P. (2007). Quantification of insect genome divergence. *Trends Genet*, 23(1):16–20.

Zhang, J. (2003a). Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18(6):292–298.

Zhang, J., Rosenberg, H. F., és Nei, M. (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci U S A*, 95(7):3708–13.

Zhang, Z., Inomata, N., Yamazaki, T., és Kishino, H. (2003b). Evolutionary history and mode of the amylase multigene family in Drosophila. *J Mol Evol*, 57(6):702–9.

Zou, Z., Evans, J. D., Lu, Z., Zhao, P., Williams, M., Sumathipala, N., Hetru, C., Hultmark, D., és Jiang, H. (2007). Comparative genomic analysis of the Tribolium immune system. *Genome Biol*, 8(8):R177.

Összefoglaló

A nemrégiben leírt Nimrod szupergéncsalád számos tagja rendelkezik a baktérium-kötéshez és a fagocitózishoz társítható funkcióval, ami arra utal, hogy ez a géncsoport a veleszületett immunitás fontos elemét képezi. A szupercsalád által kódolt fehérjéket egy, az N-terminális régióban található szekvencia-motívum (CCxGY/W), és egy vagy több speciális EGF-szerű domén, az úgynevezett „NIM domén" jelenléte jellemzi. A fehérjék minden esetben rendelkeznek szignálpeptiddel, ami arra utal, hogy funkciójukat az extracelluláris térben látják el.

A szupercsaládon belül a doménstruktúra nagy változatosságot mutat, ami alapján három fő csoport körvonalazódik. A Draper típusú fehérjék egy N-terminális EMI domént, egy NIM domént és számos „klasszikus" EGF-szerű domént tartalmaznak. A NimC és NimB típusú („poli-NIM") fehérjék egy változatos felépítésű N-terminális domént, valamint egy NIM doménekből álló dómenismétlődést tartalmaznak. A Draper és NimC típusú fehérjék, a NimB típusúakkal ellentétben általában rendelkeznek egy transzmembrán doménnel is.

Egyes, a Nimrod szupercsaládba tartozó gének, a *centaurin gamma* génnel, valamint az *ance* és a CPF géncsaládokba tartozó génekkel egyetemben egy nagy időléptékben (300-350 millió év) konzervált klasztert alkotnak („Nimrod klaszter").

A Nimrod szupercsaládra irányuló vizsgálataink egyik fő célja a „poli-NIM" architektúrával rendelkező fehérjékben található NIM domének evolúciójának jellemzése volt, az „összehangolt" („concerted") és a „születés és halál" („Birth-and-Death") konceptuális modellek fényében. Vizsgálataink másik fő célja a Nimrod szupercsalád részletes evolúciós történetének rekonstrukciója volt, a molekuláris filogenetika, és általában a bioinformatika módszereinek segítségével. Céljaink közé tartozott még a Nimrod klaszter konzerváltságának vizsgálata, az irodalomra és a nyilvánosan elérhető adatbázisokban található adatokra alapozva.

A klaszter konzerváltságának statisztikai szignifikanciáját egy, a géncsaládok jelenlétét is figyelembe vevő közelítő analitikus módszerrel vizsgáltuk. A *D. melanogaster*-hez viszonyított konzerváltság szignifikanciáját felmérő tesztek eredményei arra utalnak, hogy a klaszter konzerváltsága nem magyarázható egy semleges, a génsorrend teljes átrendeződését feltételező null modellel. Az irodalmi adatokra alapozva azt feltételezzük, hogy a Nimrod klaszter egységét a gének közelségét igénylő, a kromatin-struktúra szintjén ható közös szabályozási mechanizmusok jelenléte tartja fenn. A konzerváltság és a gének expressziójára, valamint funkciójára

vonatkozó adatok alapján ugyancsak valószínűsíthető, hogy ez klaszter a rovarok veleszületett immunitásában fontos szerepet játszó „funkcionális modul”.

A szupercsalád által kódolt fehérjékben található NIM domének aminosav-szekvenciáiból becsült filogenetikai fák elemzése nyomán azt találtuk, hogy a szupercsalád legtöbb tagjával ellentétben, az *eater* gének által kódolt egyes domének evolúciója az összehangolt modellel írható le legjobban. Ezen következtetésünket a doméneket kódoló nukleotid-szekvenciák közötti becsült szinonim távolságok is megerősítették. A NIM domének összehangolt evolúciójára utaló jeleket találtunk még az *A. mellifera* NimCI, a *T. castaneum* NimCI és kisebb mértékben a *T. castaneum* NimCII szekvencia esetében is, azonban közeli ortológok hiányában a homogenizáció tényét nem tudjuk megerősíteni.

Mivel a gének összehangolt evolúciójának vizsgálata során a leggyakrabban alkalmazott filogenetikai módszerek a domének esetében kevésbé praktikusnak bizonyultak, kifejlesztettünk egy, a domén-filogenetikákon alapuló vizuális módszert, amelynek segítségével egyszerűbben felmérhető, hogy a szekvenciák egyes régióiban található domének evolúcióját melyik konceptuális modell írja le a legjobban. A módszer lényege, hogy egy diagramon ábrázolja az egy adott szekvenciapárban található doméneket, valamint a domén-filogenetikából kinyert, a konceptuális modellekre nézve releváns információkat. A módszert egy Perl nyelven írt alkalmazás formájában implementáltuk, ami elérhető a `http://t2prhd.sf.net` weblapon.

A vizuális módszer segítségével megállapítottuk, hogy az Eater szekvenciák középső régiójában található NIM domének vesznek részt az összehangolt evolúcióban. Ugyancsak megfigyeltük, hogy a domén-ismétlődés N-terminális régiójában a C-terminális régióhoz képest több domén viselkedik a „születés és halál” konceptuális modellnek megfelelően, amit szelekciós tényezőkkel magyarázunk.

A domén NIM eredetének felderítése érdekében a domént modellező profil Rejtett Markov Modellt (pHMM-et) az EGF-szerű domén modelljével összehasonlítottuk, az ezen modellek lokális-lokális illesztésén alapuló páros HMM logó módszer segítségével. Ugyanezen módszer segítségével megvizsgáltuk a taxonómiai eloszlása alapján ősibbnek tekinthető Draper típusú domén-architektúra N-terminális részében található, a CCxGY/W motívumot is magába foglaló EMI domén és az EGF-szerű domén közötti hasonlóságot.

A páros HMM logók, valamint egyéb megfigyelhető hasonlóságok alapján azt feltételezzük, hogy az EMI és NIM domének két EGF-szerű doménből és egy linker régióból jöttek létre egy olyan strukturális átrendeződés-sorozat révén, amiben az inszerciók kiemelkedő szerepet játszottak. Ennek fényében valószínűsíthető, hogy a Draper típusú domén-architektúra egy transzmembrán régiót és egy EGF-szerű doménekből álló ismétlődést tartalmazó „poli-EGF” architektúra leszármazottja. A NIM és EMI domének eredete jó példája lehet a doménismétlődések rekrutálásának az új funkciót ellátó domének kialakulása során.

Klasszikus (Maximum Likelihood, Neighbor-Joining) és Bayes-statisztikai alapokon nyug-

vó filogenetikai rekonstrukciós módszerek segítségével rekonstruáltuk a NimA, NimB és NimC szekvenciák filogenetikáit, a domén-szintű összehangolt evolúció jeleit mutató szekvenciák kizárásával. A Bayes MCMC analízisek során egy egyszerű módszerrel bináris karakterekké alakítottuk az illesztésben található gapeket, és megvizsgáltuk, hogy ezek bevonásával mennyire csökkenthető a topológia becslésének bizonytalansága.

A *D. melanogaster* Nimrod paralógok illesztésének és filogenetikájának egyidejű becslése által következtetni próbáltunk a NimA/Draper, NimC és NimB szekvenciák közötti leszármazási kapcsolatokra. A becslés során kapott poszterior mintából felépített konszenzus-hálózat alapján feltételezhető, hogy a NimB típusú szekvenciákat tömörítő klád a NimC2 leszármazási vonal részét képezi.

Eredményeinket összegezve kidolgoztunk egy hipotézist, ami a szupercsalád evolúciós történetét magyarázza az első jellegzetes doménstruktúra kialakulásától az egyes családok radiációjáig, és ami reményeink szerint hasznosnak bizonyul majd a szupercsalád génjeire irányuló további vizsgálatok során.

Summary

Many members of the recently described Nimrod gene superfamily have been demonstrated to play a role in phagocytosis and binding of bacteria, which suggests that this group of genes may be an important constituent of innate immunity. The proteins encoded by the members of the superfamily are characterized by the presence of a sequence motif (CCxGY/W) in their N-terminal region, and of at least one "NIM domain", a special type of EGF domain. They also possess a signal peptide suggesting that they act extracellularly.

The domain structure within the superfamily is highly variable. Based on the domain architecture, we can define three major groups. The proteins with the Draper-type architecture have an N-terminal EMI domain, and only one copy of the NIM domain, followed by a variable number of "classic" EGF-like repeats. The NimC and NimB type ("poly-NIM") proteins posses a variety of N-terminal domains, and an array of NIM domains separated by short linker sequences. The Draper and NimC type proteins, in contrary to the NimB type, generally have a transmembrane domain.

Some genes belonging to the Nimrod superfamily, along with the *centaurin gamma* gene, and with genes from the *ance* and CPF gene families, are a part of a gene cluster (the "Nimrod cluster") which is conserved on large evolutionary timescales (300-350 million years).

One of the major objectives of our study on the Nimrod was the characterization of the evolution of the domain repeats in the "poly-NIM" type protein in the light of the "concerted" and "Birth-and-Death" conceptual models. The other main objective of our study was to reconstruct the detailed evolutionary history of the Nimrod superfamily by using phylogenetic reconstruction and other sequence analysis methods. We also intended to study the conserved Nimrod cluster, based on the literature and other publicly available data.

We used an approximate analytical method, which also deals with the presence of gene families, to asses the statistical significance of the conservation of the Nimrod cluster. Pairwise tests assessing the conservation against the cluster from the *D. melanogaster* genome indicate that this cannot be explained by a neutral null model assuming randomization of the gene order. Based on previously published data, we propose, that the conservation of the Nimrod cluster is maintained by some common regulatory elements requiring the vicinity of the genes, probably acting at the level of chromatin structure. Considering the timescale of the conservation and

the available functional and expression data concerning the member gene families, it seems possible that the cluster is an important "functional module" of insect innate immunity.

During the analysis of phylogenetic trees based on the amino acid sequences of the NIM domains, we found that, in contrary to the majority of the members of the superfamily, the evolution of a group of domains from the proteins encoded by the Drosophila *eater* genes could be described best by the concerted model. This finding was also corroborated by the estimated synonymous distances between the respective domains. Based on the domain phylogenies, we also suspect that domains from the *A. mellifera* NimCI, *T. castaneum* NimCI and also *T. castaneum* NimCII may have been involved in episodes of concerted evolution, but in the lack of closely related ortologues, we were not able to confirm that.

Since the analysis of phylogenetic trees, which is routinely used to study gene family evolution, proved to be less practical in the case of domain repeats, we have developed a new visual method. The method is based on domain phylogenies, and makes possible the rapid assessment of which conceptual model describes the best the evolution of the domains from the different regions. The basic idea behind the method is to visualize the domains from a pair of sequences and the information relevant to the conceptual models from the phylogeny of the domains. The Perl script implementing the method can be downloaded from `http://t2prhd.sf.net`.

By using the newly developed visual method we found that the domains from the middle of the Eater sequences are the ones involved in the concerted evolution. We also found, that more domains avoid homogenization at the N-terminal end of the domain array as compared to the C-terminal region, which we explain as caused by natural selection.

In order to study the evolutionary origins of the NIM domain, we have used the pairwise HMM logo method based on the local-local alignment of profile Hidden Markov Models (pHMMs) to asses the similarity between the EGF-like and NIM domains. The Draper-type architecture can be considered more ancient based on its taxonomic distribution. This architecture contains an N-terminal EMI domain which also includes the characteristic CCxGY/W sequence motif. By using pairwise HMM logos we also assessed the similarity between the EMI and the EGF-like domains.

Based on the pairwise logos and other similarities, we proposed that the EMI and NIM domains originated from two EGF-like domains and a short linker sequence through structural reorganizations in which insertions played an important role. This means that the Draper-type architecture is the descendant of a "poly-EGF" architecture containing a transmembrane domain and also an array of EGF-like domains. Our hypothesis for the origin of the EMI and NIM domains could be a good example for the recruitment of domain repeats during the emergence of domains with new functions.

By using classical (Maximum Likelihood, Neighbor-Joining) and Bayesian methods, we

reconstructed the phylogenies of the NimA, NimB and NimC sequences, with the exclusion of those with domains supposedly engaged in concerted evolution. We used a simple method to encode the gaps from the multiple sequence alignments into binary characters, and during the Bayesian analyses we have studied the effect of the inclusion of these characters on the uncertainty of the topology estimation.

We also tried to infer the evolutionary relationship between the NimA/Draper, NimB and NimC sequences by the joint estimation of the alignment and phylogeny of the sequences of the *D. melanogaster* Nimrod paralogs. The consensus network built from the posterior sample suggests that the NimB sequences are part of the NimC2 lineage.

In conclusion, we have developed a hypothesis discussing the evolutionary history of the Nimrod superfamily from the emergence of the first characteristic domain architecture to the diversification of the individual families, and which we hope will prove to be useful for further studies aiming its member genes.

Függelék

A függelékben a következő fajokat fogjuk említeni, a génusznév rövidítésével: *Drosophila melanogaster*, *Drosphila pseudoobscura*, *Drosophila sechellia*, *Drosphila yakuba*, *Drosphila virilis*, *Anopheles gambiae*, *Apis mellifera*, *Pediculus humanus*, *Tribolium castaneum*.

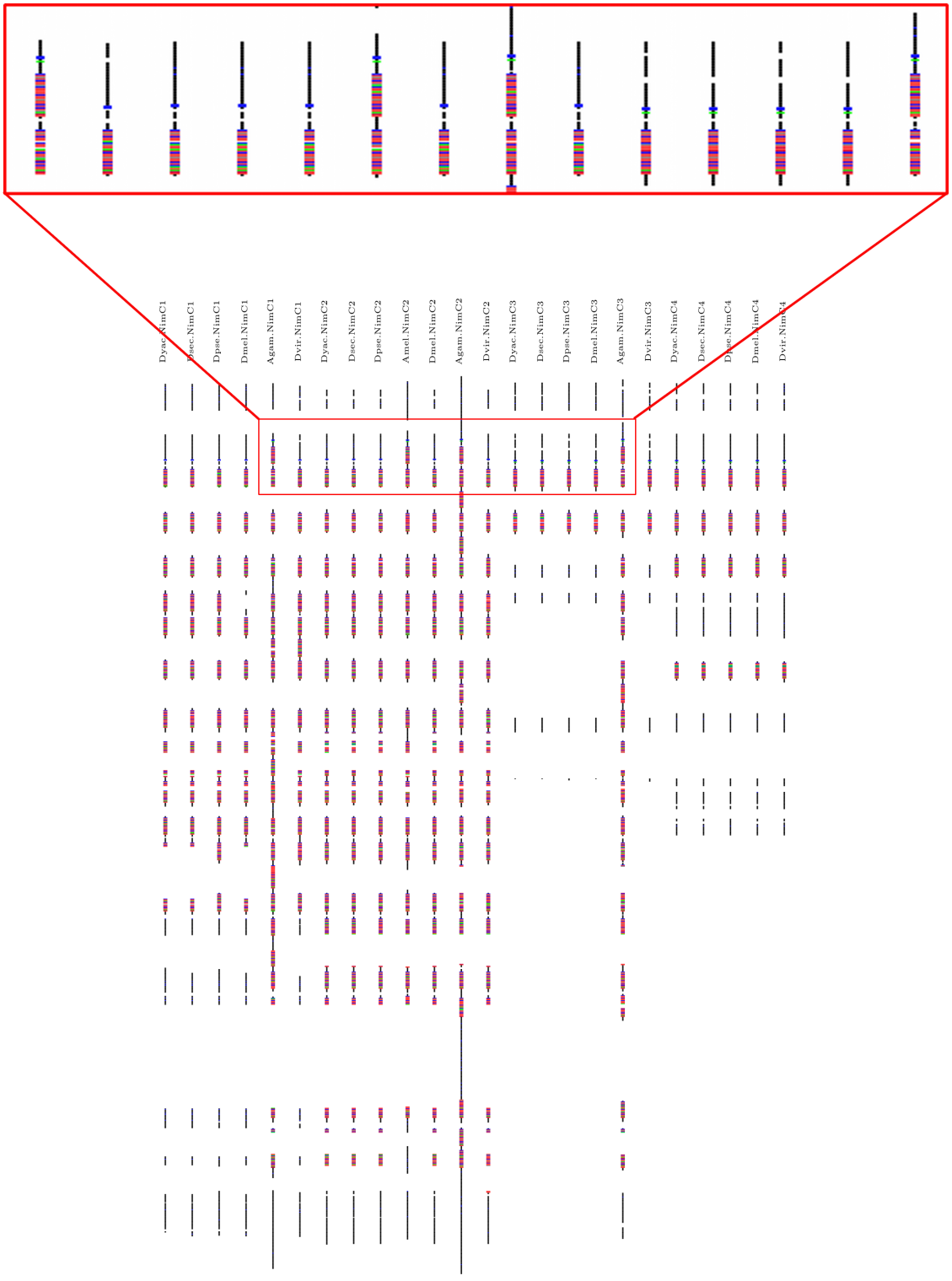
1. A vizsgált Nimrod szekvenciák adatai

Faj	**Név** (saját jelölés)	**Azonosító**	**Adatbázis**	**Referencia**
D. melanogaster	NimA	CG31841+CG31765	FlyBase version FB2008_02, released February 20, 2008	Kurucz és mtsai, 2007
	NimB1	CG33119		
	NimB2	CG31839		
	NimB3	CG34003		
	NimB4	CG33115		
	NimB5	CG16873		
	NimC1	CG8942		
	NimC2	CG18146		
	NimC3	CG16880		
	NimC4	CG16876		
	Eater	CG6124		
	Draper	CG2086		
D. pseudoobscura	NimA	GA16517 + GA16460		Evans és mtsai, 2006
	NimB1	GA17296		
	NimB2	GA16515		
	NimB3	GA29057		
	NimB4	GA17294		
	NimB5	GA14188		
	NimC1	GA21423		
	NimC2	GA14817		
	NimC3	GA14193		
	NimC4	GA14190		
	Eater	GA19372		
	Draper	GA15231		
D. sechellia	NimA	GM15396 + GM1540		
	NimB1	GM15419		
	NimB2	GM15024		
	NimB3	GM15018		
	NimB4	GM15007		
	NimB5	GM15430		
	NimC1	GM14996		
	NimC2	GM15452		
	NimC3	GM14974		
	NimC4	GM15468		
	Eater	GM10134		
	Draper	GM14179		
D. virilis	NimA	GJ18159		
	NimB1	GJ18160		
	NimB2	GJ18075		
	NimB3	GJ18074		
	NimB4	GJ18073		
	NimB5	GJ18161		
	NimC1	GJ18072		
	NimC2	GJ18162		
	NimC3	GJ18068		
	NimC4	GJ18170		
	Eater	GJ24134		
	Draper	GJ12529		
D. yakuba	NimA	GE19013 + GE19014		
	NimB1	GE19015		
	NimB2	GE25257		
	NimB3	GE25246		
	NimB4	GE25236		
	NimB5	GE19016		
	NimC1	GE25203		
	NimC2	GE19018		
	NimC3	GE25181		
	NimC4	GE19022		
	Eater	GE10589		
	Draper	GE20925		
A. gambiae	NimA	AGAP009760	Ensembl release 48	Kurucz és mtsai, 2007
	NimB	AGAP009761		
	NimCI	AGAP009762		
	NimCII	AGAP009763		
	NimCIII	AGAP012386		
	Draper	AGAP007256		
A. mellifera	NimA	GB12883	GenBank	Kurucz és mtsai, 2007; Evans és mtsai, 2006
	NimB	GB12454		
	NimCI	GB14645		
	NimCII	GB13979		
	Draper	GB14962		
T. castaneum	NimA	GLEAN_11427	Tribolium Genome Assembly 2.0	Zou és mtsai, 2007
	NimB	GLEAN_11428		
	NimCI	GLEAN_15258		
	NimCII	GLEAN_02053		
	Draper	GLEAN_00689		

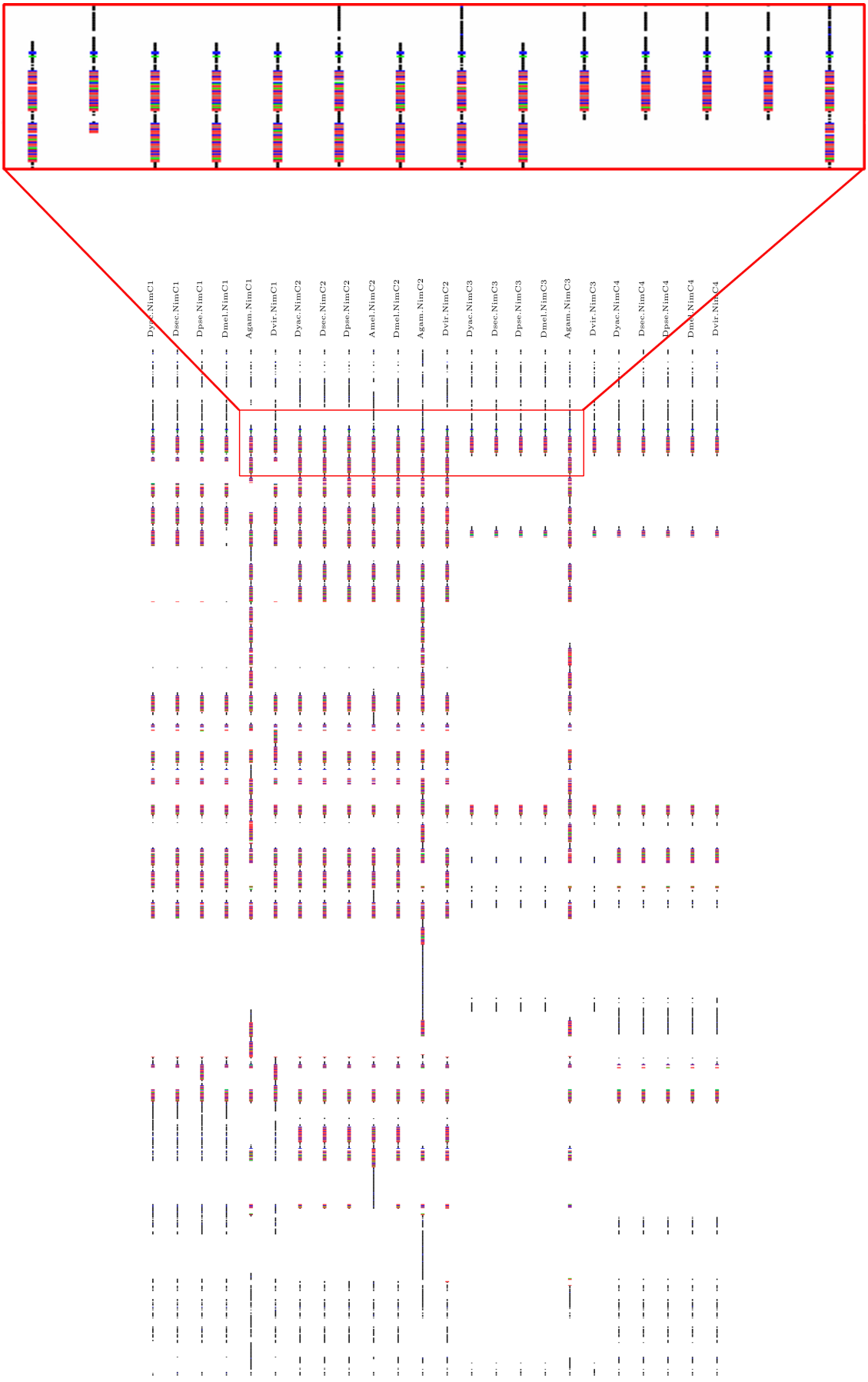
2. A NIM domének és a CCxGY/W motívumok többszörös illesztésekben való vizualizációját szemléltető példák.

Az ábrákon a téglalapok a NIM doméneknek megfelelő pozíciókat jelképezik. A színek a NIM konszenzus konzervált aminosavainak felelnek meg, így az N-terminális régióban a kék-zöld színű vastagítások a CCxG motívumokat jelzik.

2.1 A NimC szekvenciák Clustal W 1.83 által számolt illesztésének ábrázolása



2.2 A NimC szekvenciák ProbCons 1.1 által számolt illesztésének ábrázolása



3. A Bayes MCMC filogenetikai rekonstrukció során használt MrBayes nexus blokkok

Nimrod B, szekvenciaillesztés

```
begin mrbayes;
prset aamodelpr=mixed;
charstat;
mcmcp ngen=2000000 nruns=4 samplefreq=100
printfreq=100 savebrlens=yes;
showmodel;
mcmc;
sump burnin=5000;
sumt burnin=5000;
end;
```

Nimrod B, gapek, párhuzamosított futtatás

```
begin mrbayes;
lset coding=variable;
charstat;
showmodel;
mcmcp ngen=2000000 nchains=5 nruns=3 samplefreq=100
printfreq=100 savebrlens=yes;
mcmc;
sump burnin=5000;
sumt burnin=5000;
end;
```

Nimrod B, szekvenciaillesztés és gapek, párhuzamosított futtatás

```
begin mrbayes;
lset applyto=(1) coding=variable;
prset applyto=(2) aamodelpr=mixed;
unlink statefreq=(all);
unlink revmat=(all);
unlink pinvar=(all);
charstat;
showmodel;
mcmcp ngen=2000000 nchains=5 nruns=3 samplefreq=100
printfreq=100 savebrlens=yes;
mcmc;
sump burnin=5000;
sumt burnin=5000;
end;
```

Nimrod C, szekvenciaillesztés

```
begin mrbayes;
prset aamodelpr=mixed;
charstat;
mcmcp ngen=2000000 nruns=4 samplefreq=100
printfreq=100 savebrlens=yes;
```

```
showmodel;
mcmc;
sump burnin=5000;
sumt burnin=5000;
end;
```

Nimrod B, gapek

```
begin mrbayes;
lset coding=variable;
charstat;
showmodel;
mcmcp ngen=2000000 nchains=4 nruns=4 samplefreq=100
printfreq=100 savebrlens=yes;
mcmc;
sump burnin=5000;
sumt burnin=5000;
end;
```

Nimrod C, szekvenciaillesztés és gapek

```
begin mrbayes;
lset applyto=(1) coding=variable;
prset applyto=(2) aamodelpr=mixed;
charstat;
mcmcp ngen=2000000 nruns=4 samplefreq=100
printfreq=100 savebrlens=yes;
showmodel;
mcmc;
sump burnin=5000;
sumt burnin=5000;
end;
```

Nimrod A, szekvenciaillesztés

```
begin mrbayes;
prset  aamodelpr=mixed;
charstat;
showmodel;
mcmcp ngen=2000000 nruns=4 samplefreq=100
printfreq=100 savebrlens=yes;
mcmc;
sump burnin=5000;
sumt burnin=5000;
end;
```

Nimrod B, gapek, párhuzamosított futtatás

```
begin mrbayes;
lset coding=variable;
charstat;
showmodel;
mcmcp ngen=2000000 nchains=5 nruns=3 samplefreq=100
printfreq=100 savebrlens=yes;
```

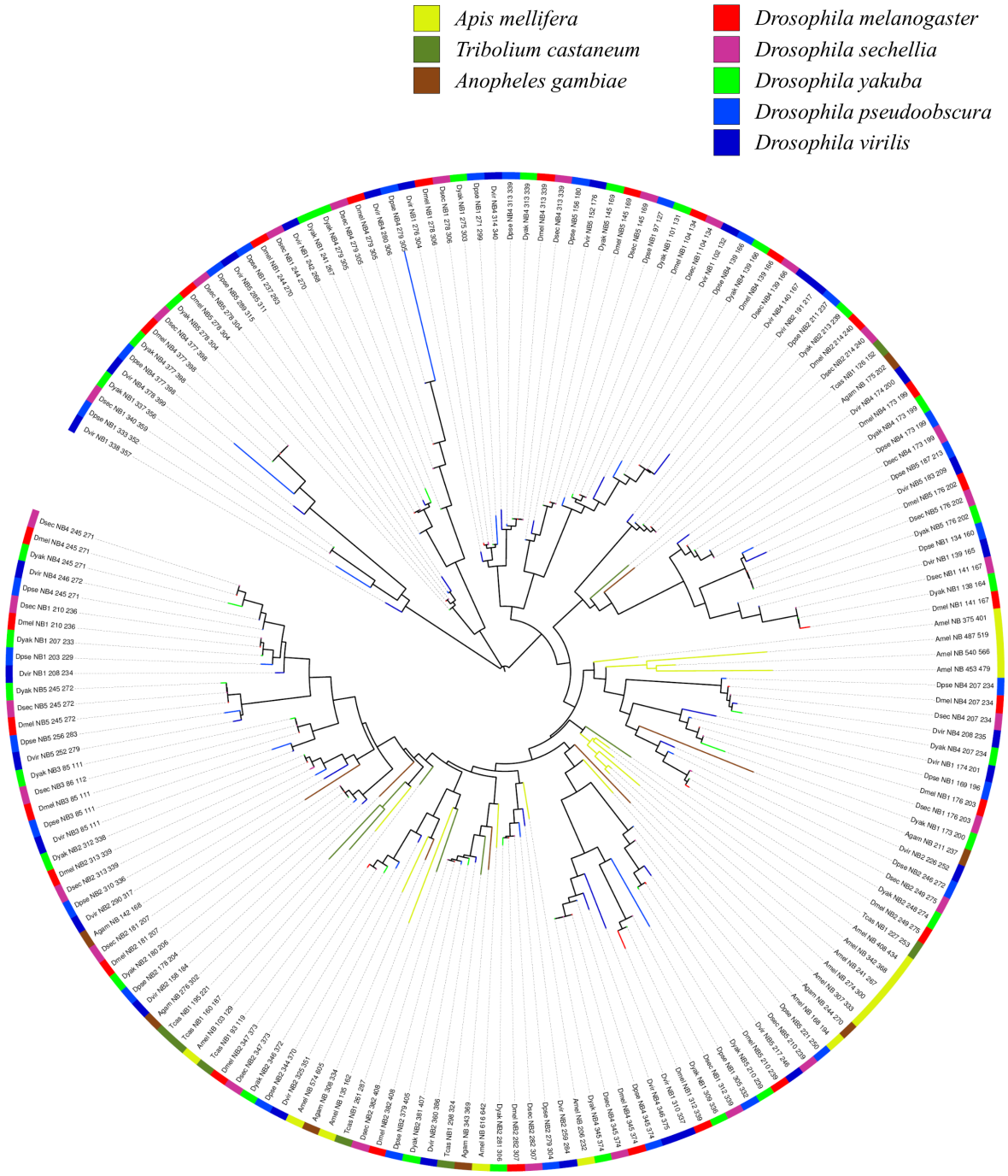
```
mcmc;
sump burnin=5000;
sumt burnin=5000;
end;
```

Nimrod B, szekvenciaillesztés és gapek, párhuzamosított futtatás

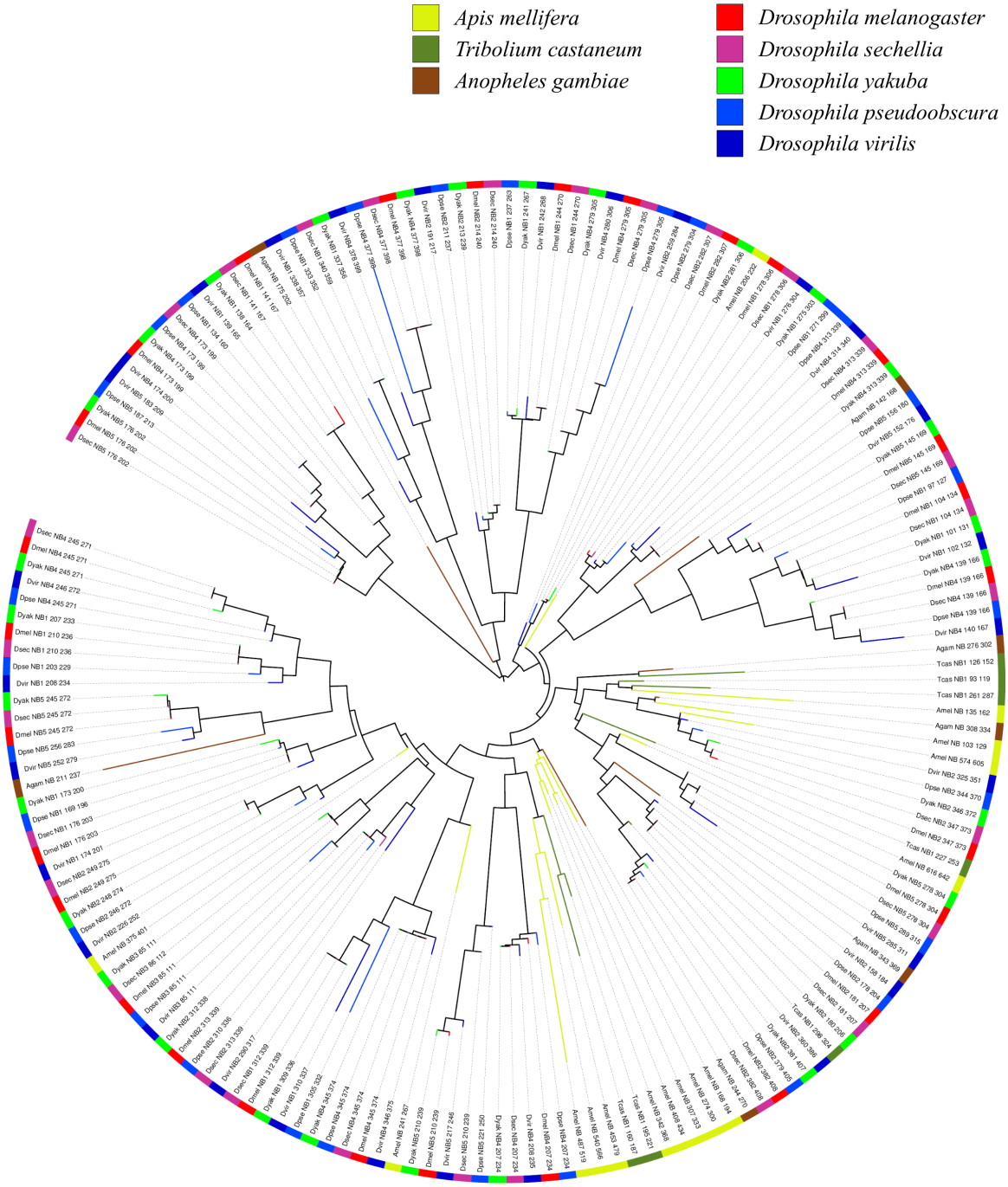
```
begin mrbayes;
lset applyto=(1) coding=variable;
prset applyto=(2) aamodelpr=mixed;
unlink statefreq=(all);
unlink revmat=(all);
unlink pinvar=(all);
charstat;
mcmcp ngen=2000000 nchains=5 nruns=3 samplefreq=100
printfreq=100 savebrlens=yes;
showmodel;
mcmc;
sump burnin=5000;
sumt burnin=5000;
end;
```

4. NimB, NimC és Eater szekvenciákban azonosított domének NJ algoritmussal számolt filogenetikája.

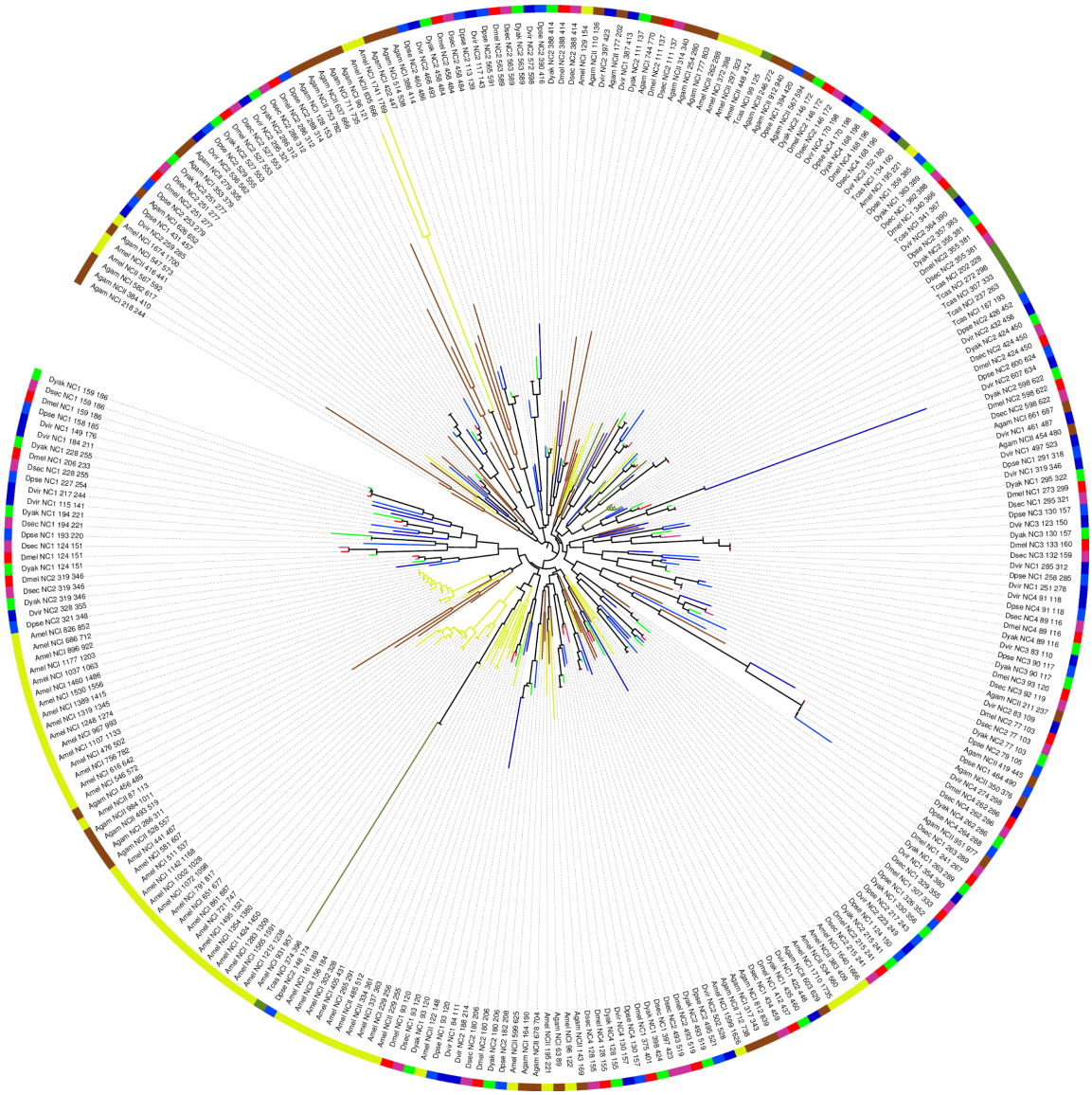
4.1 NimB, NJ, PD



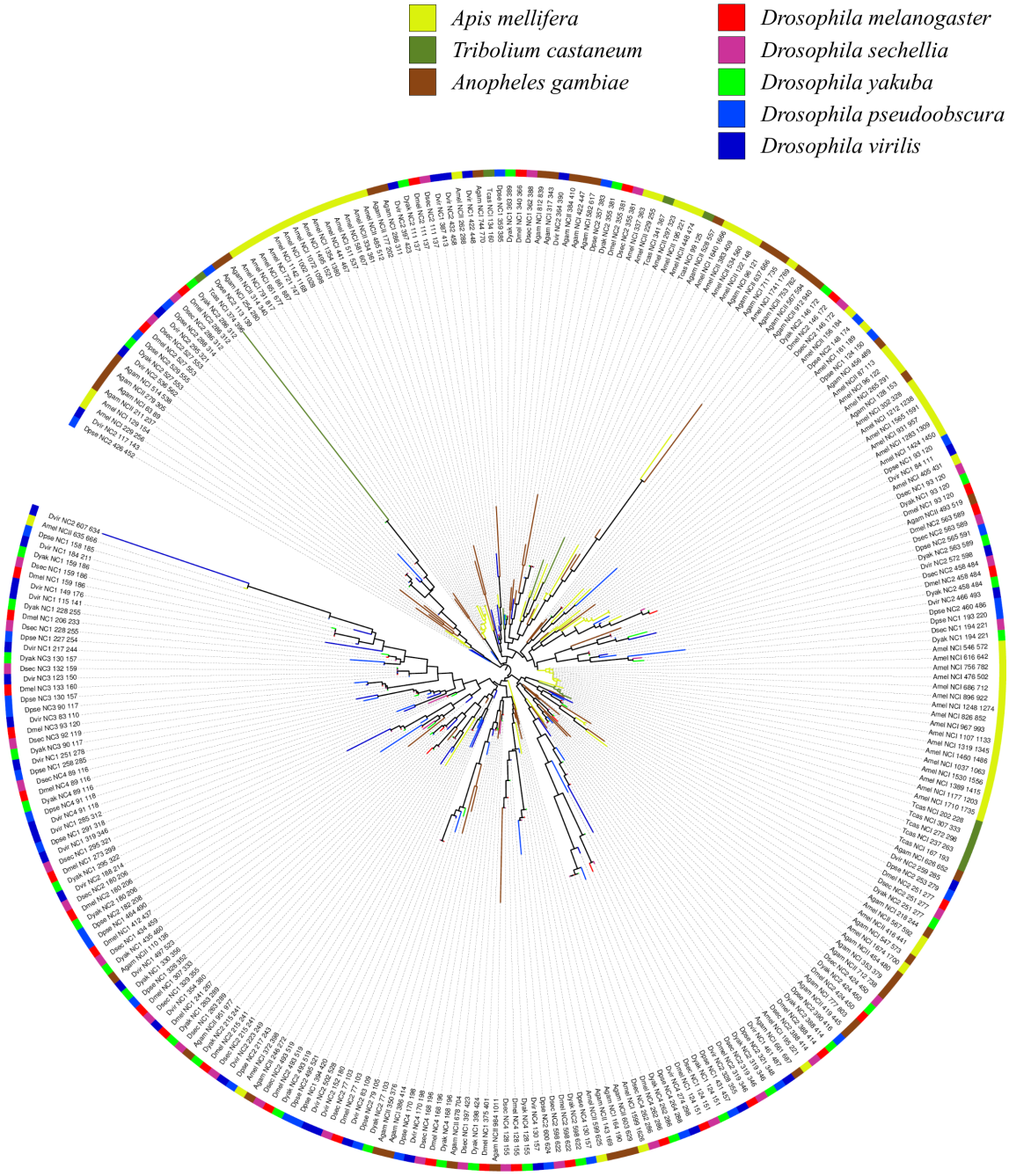
4.2 NimB, NJ, CD


Apis mellifera
Tribolium castaneum
Anopheles gambiae
Drosophila melanogaster
Drosophila sechellia
Drosophila yakuba
Drosophila pseudoobscura
Drosophila virilis

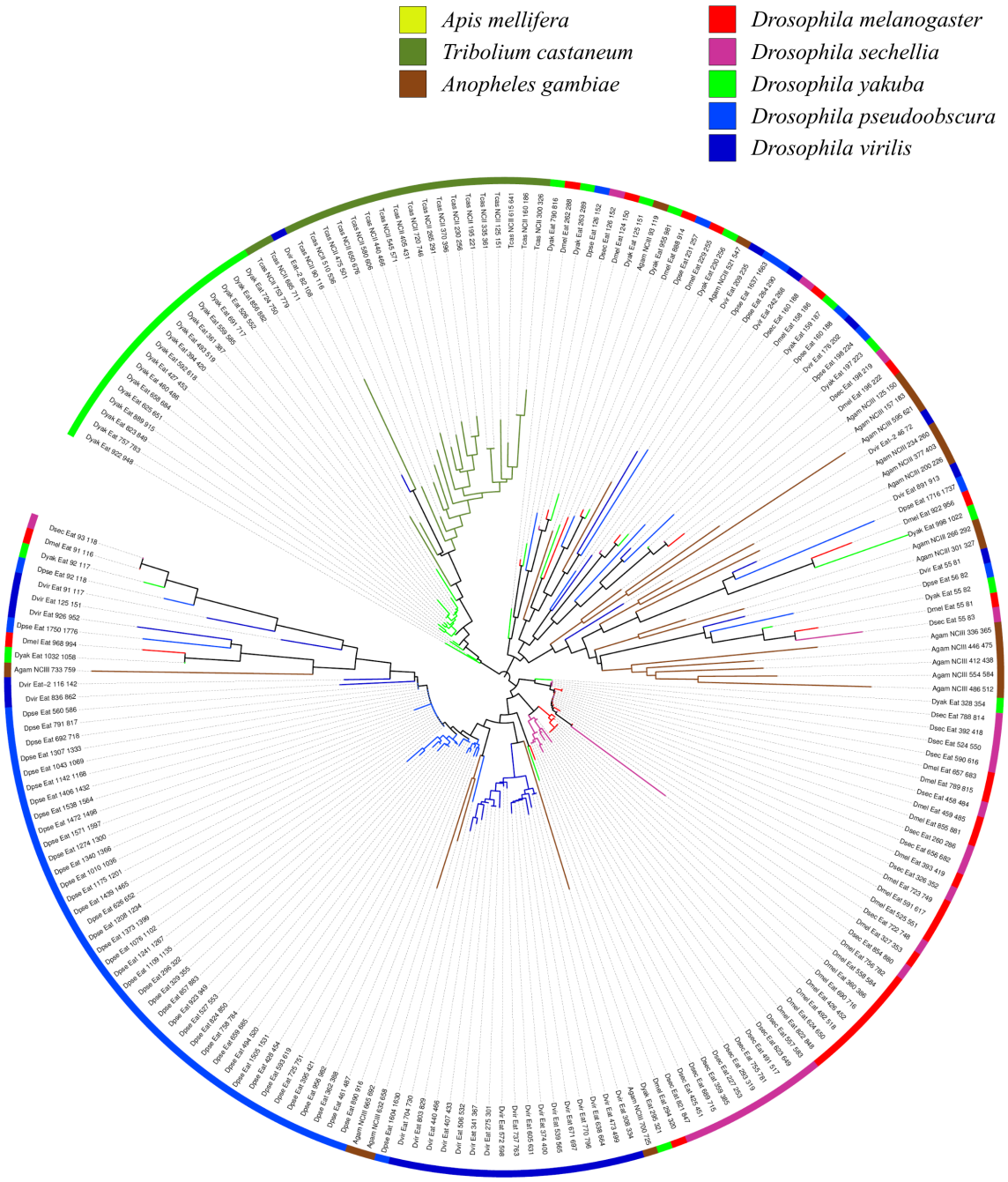
4.3 NimC, NJ, PD


Apis mellifera
Tribolium castaneum
Anopheles gambiae
Drosophila melanogaster
Drosophila sechellia
Drosophila yakuba
Drosophila pseudoobscura
Drosophila virilis

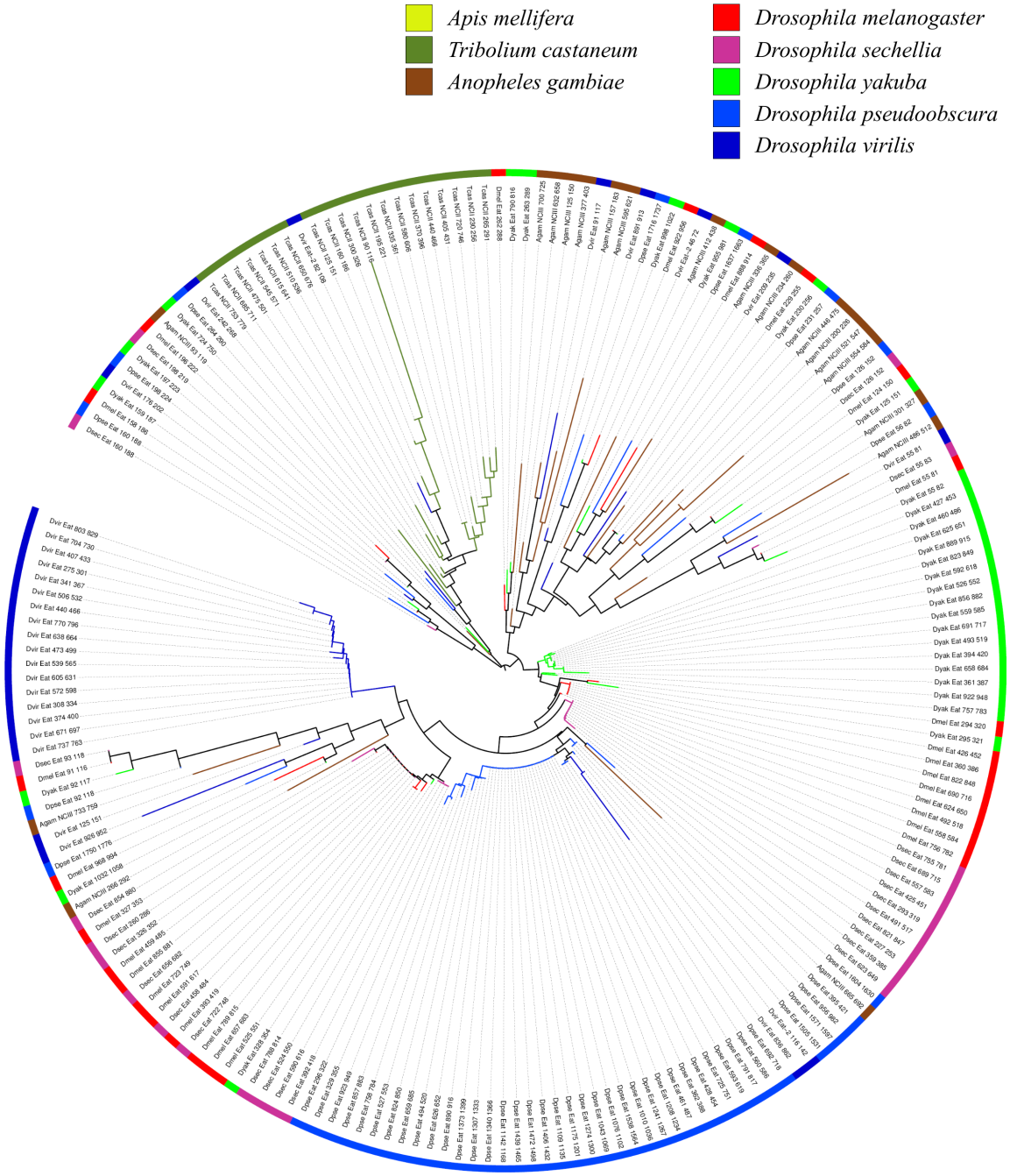
4.4 NimC, NJ, CD


Apis mellifera
Tribolium castaneum
Anopheles gambiae
Drosophila melanogaster
Drosophila sechellia
Drosophila yakuba
Drosophila pseudoobscura
Drosophila virilis

4.5 Eater, NJ, PD

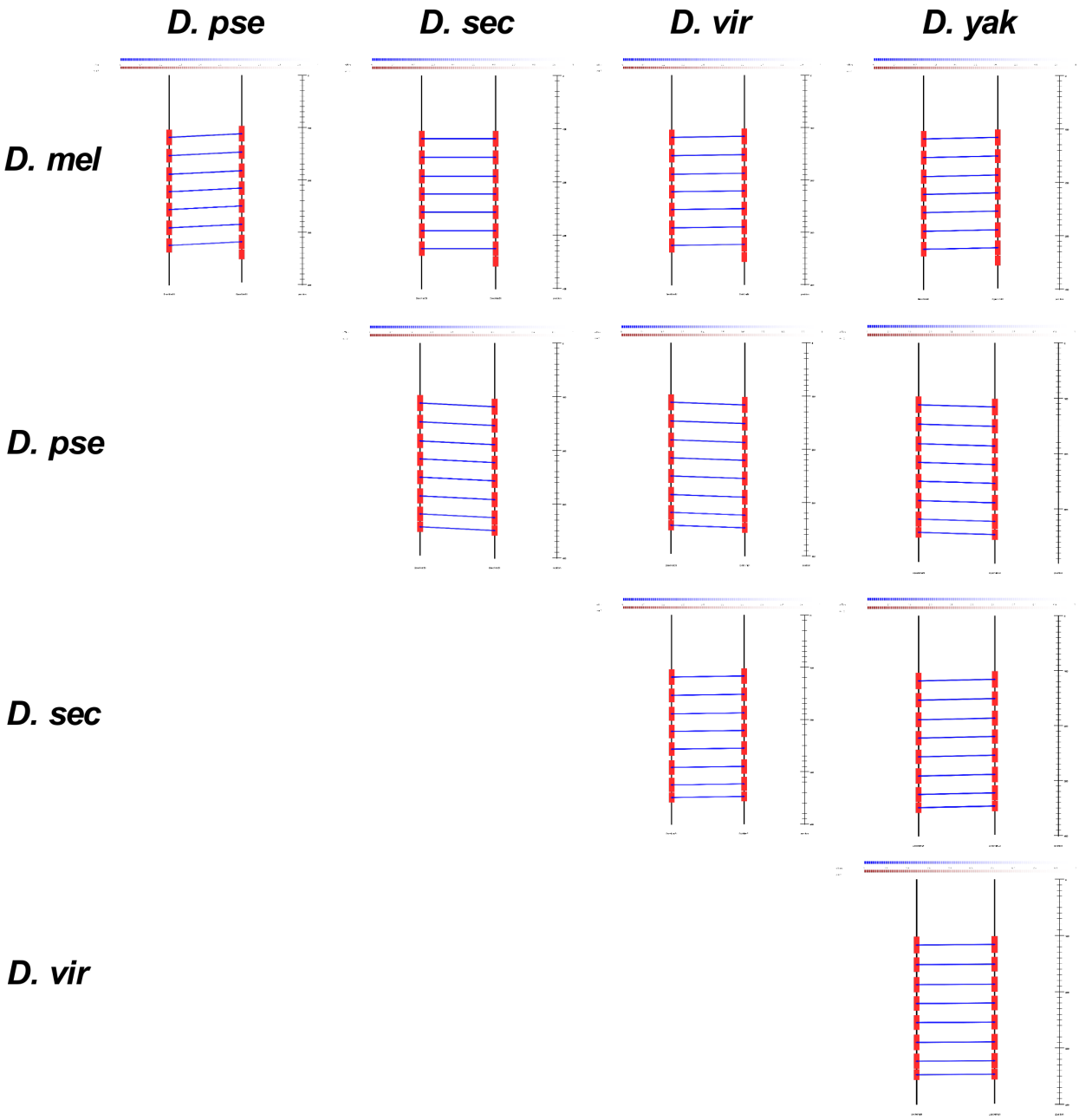


4.6 Eater, NJ, CD

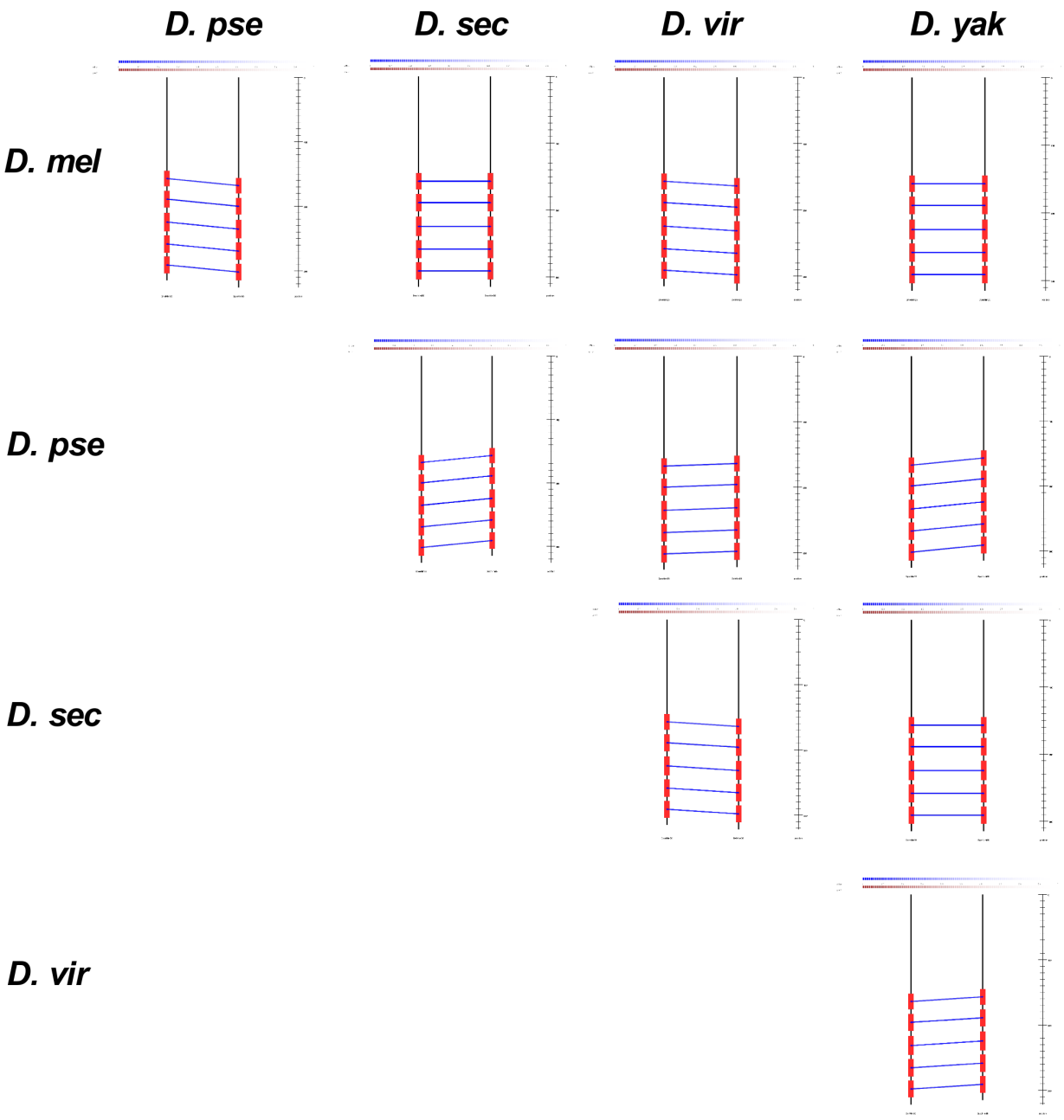


5. Különböző Drosophila Nim ortológok szekvenciáinak párbaállításával készült PRHD-k.

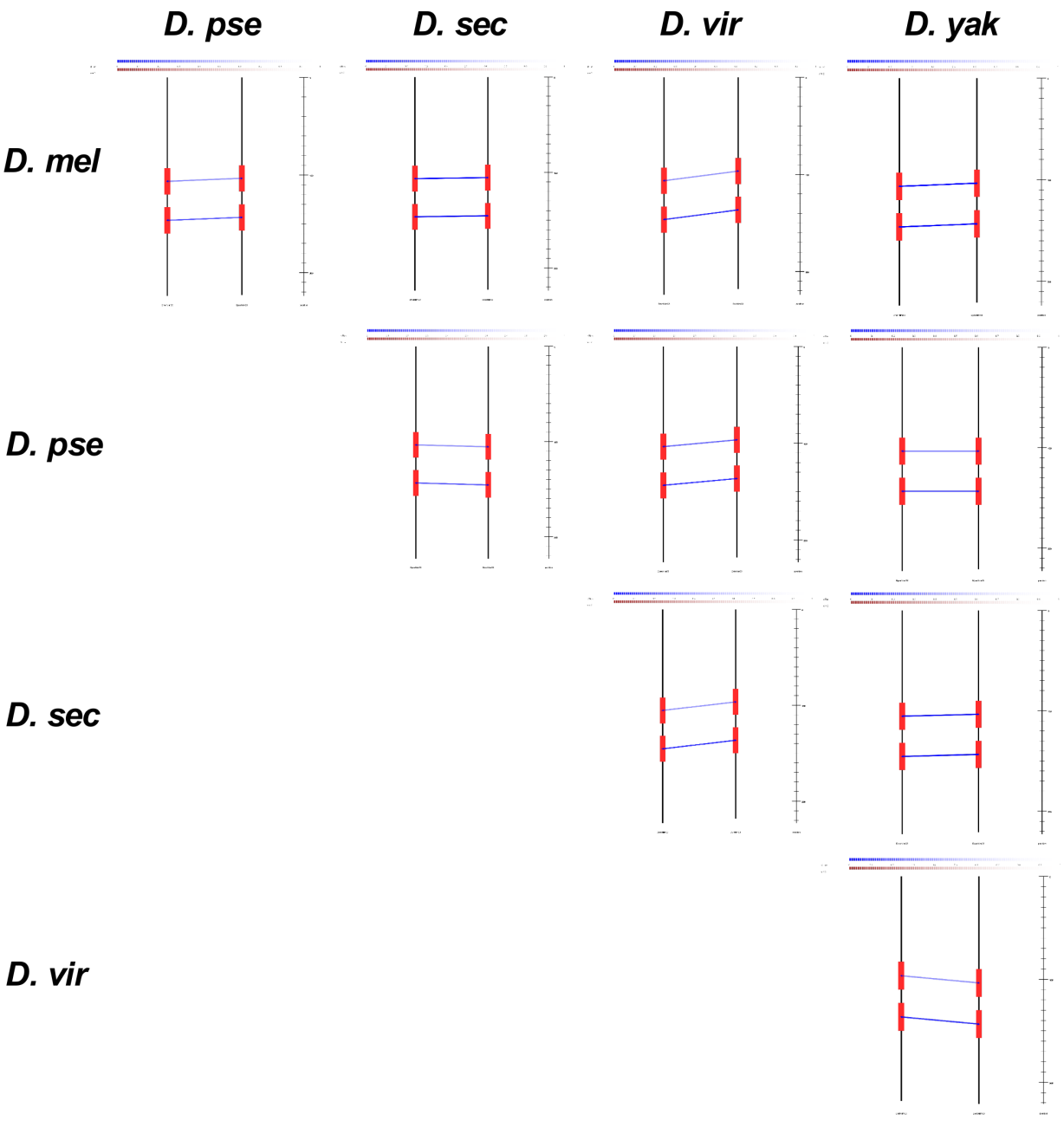
5.1 Drosophila NimB1 ortológok



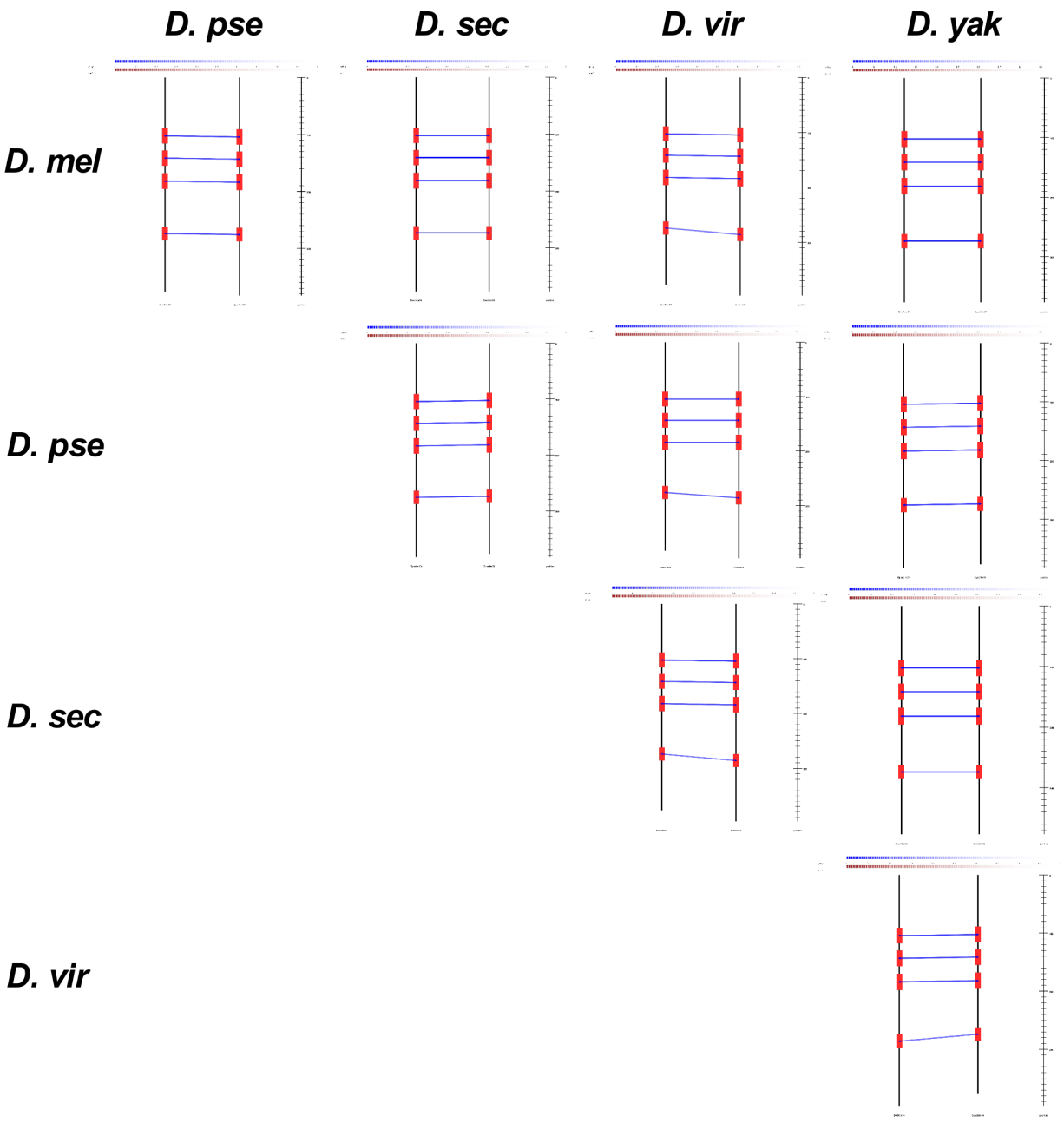
5.2 Drosophila NimB5 ortológok



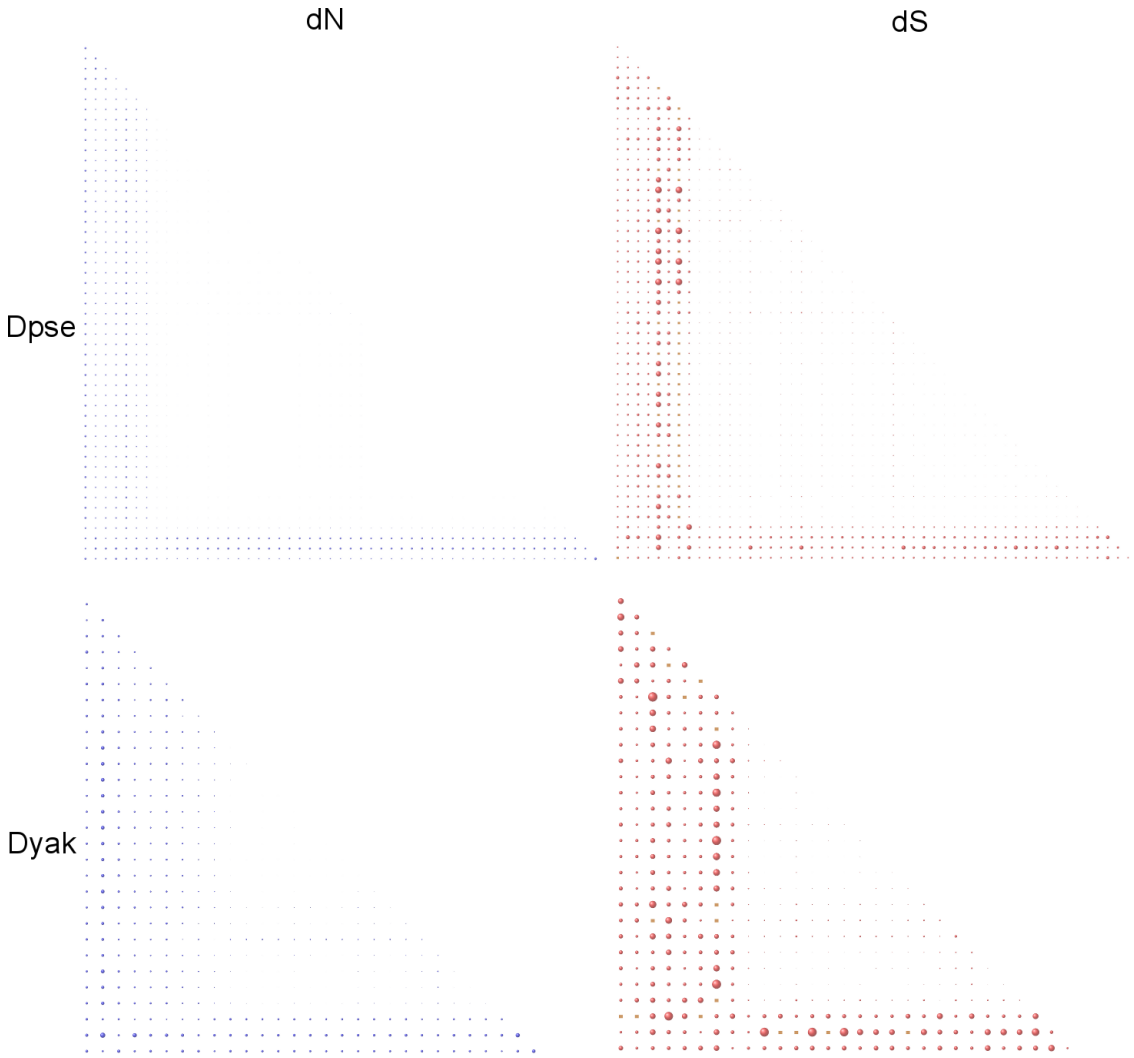
5.3 Drosophila NimC3 ortológok



5.4 Drosophila NimC4 ortológok



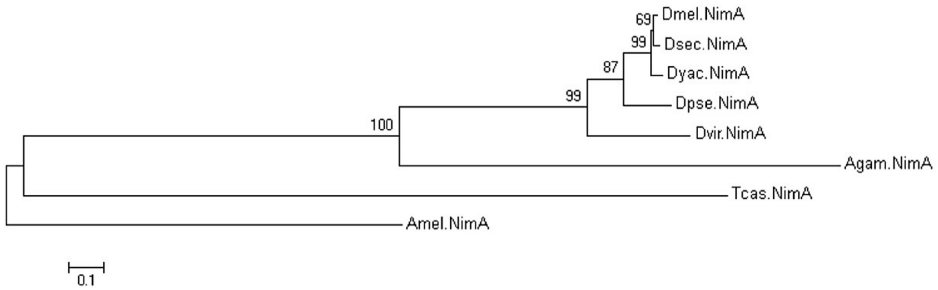
6. A *D. pseudoobsucra* és *D. yakuba* Eater szekvenciákban azonosított domének közötti szinonim és nemszinonim távolságok ábrázolása



Az ábra értelmezéséhez lásd a 4.13. ábra magyarázatát.

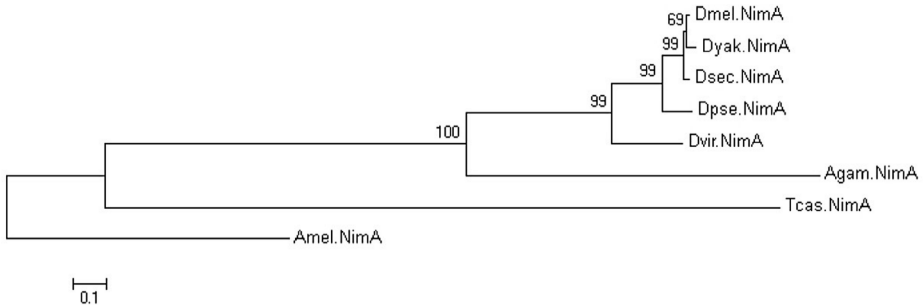
7. A NimA szekvenciák NJ és Bayes MCMC módszerekkel rekonstruált filogenetikái

7.1 NJ, PD



Illesztés: ProbCons 1.1, Szoftver: MEGA 3.1, Szubsztitúciós modell: JTT+Γ, Gamma paraméter: 1,22, gamma kategóriák száma: 4, Bootstrap ismétlések száma: 1000 Az ábra az MEGA 3.1 (Kumar és mtsai, 2004) szoftver segítségével készült.

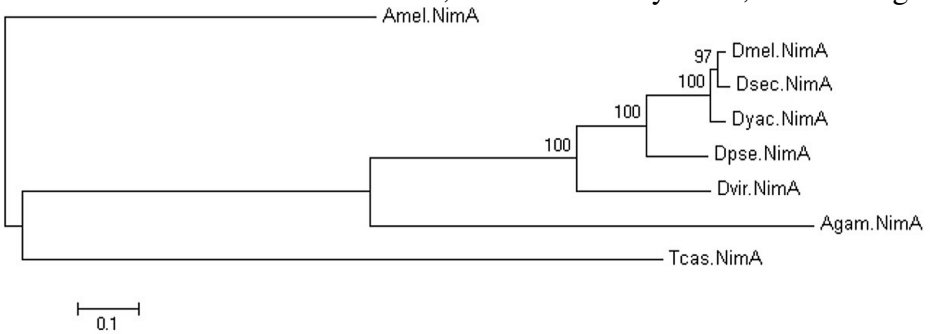
7.2 NJ, CD



Illesztés: ProbCons 1.1, Szoftver: MEGA 3.1, Szubsztitúciós modell: JTT+Γ, Gamma paraméter: 1,22, gamma kategóriák száma: 4, Bootstrap ismétlések száma: 1000 Az ábra az MEGA 3.1 (Kumar és mtsai, 2004) szoftver segítségével készült.

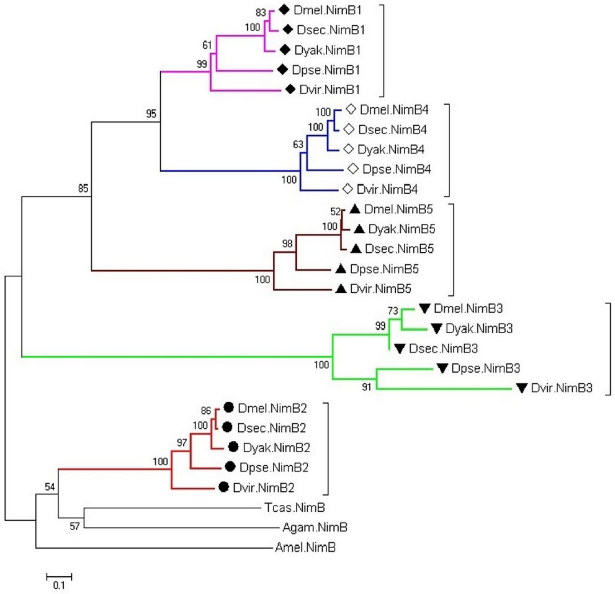
7.3 Bayes MCMC

Többségi konszenzusfa. Illesztés: ProbCons, szoftver: MrBayes 3.1, modellátlagolt becslés.



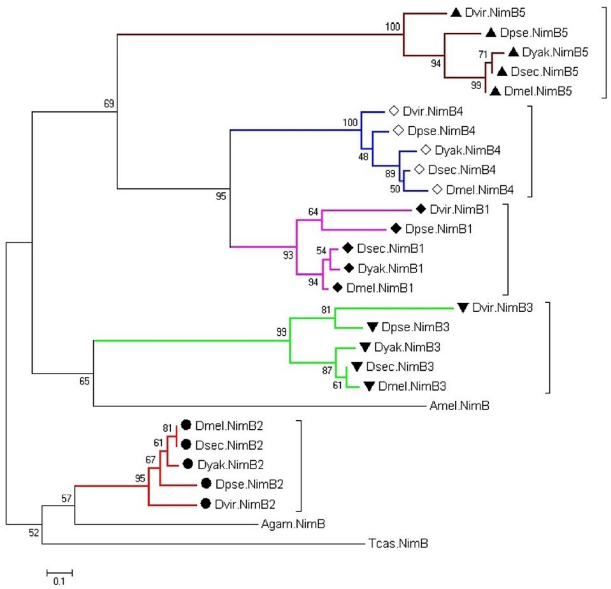
8. A NimB és NimC szekvenciák NJ algoritmussal rekonstruált filogenetikái

8.1 NimB, NJ, PD



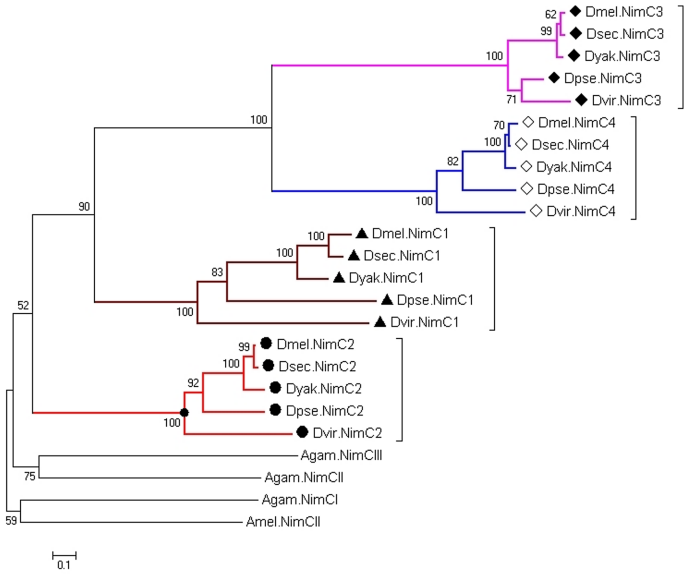
Illesztés: ProbCons 1.1, Szoftver: MEGA 3.1, Szubsztitúciós modell: JTT+Γ, Gamma paraméter: 2,29, gamma kategóriák száma: 4, Bootstrap ismétlések száma: 1000 Az ábra az MEGA 3.1 (Kumar és mtsai, 2004) szoftver segítségével készült.

8.2 NimB, NJ, CD



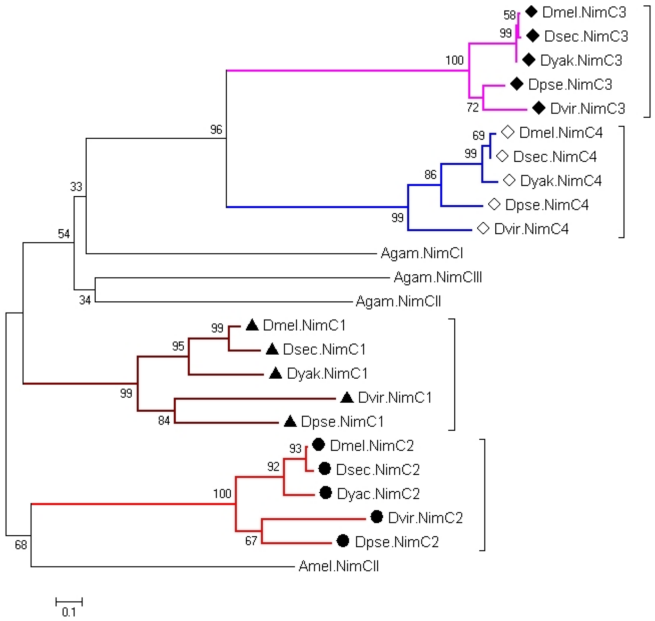
Illesztés: ProbCons 1.1, Szoftver: MEGA 3.1, Szubsztitúciós modell: JTT+Γ, Gamma paraméter: 2,29, gamma kategóriák száma: 4, Bootstrap ismétlések száma: 1000 Az ábra az MEGA 3.1 (Kumar és mtsai, 2004) szoftver segítségével készült.

8.3 NimC, NJ, PD



Illesztés: ProbCons 1.1, Szoftver: MEGA 3.1, Szubsztitúciós modell: PAM+Γ, Gamma paraméter: 2,78, gamma kategóriák száma: 4, Bootstrap ismétlések száma: 1000 Az ábra az MEGA 3.1 (Kumar és mtsai, 2004) szoftver segítségével készült.

8.4 NimC, NJ, CD



Illesztés: ProbCons 1.1, Szoftver: MEGA 3.1, Szubsztitúciós modell: PAM+Γ, Gamma

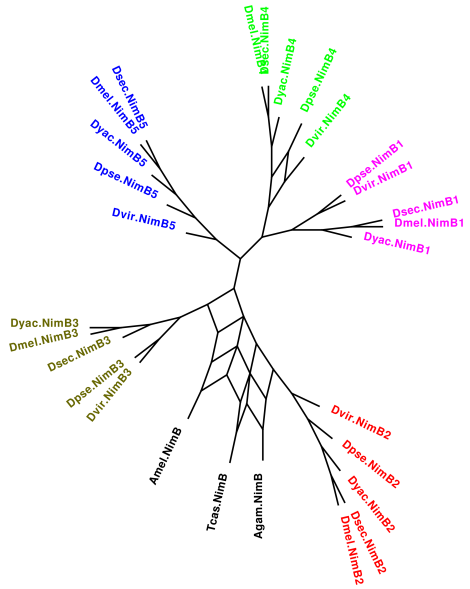
paraméter: 2,78, gamma kategóriák száma: 4, Bootstrap ismétlések száma: 1000 Az ábra az MEGA 3.1 (Kumar és mtsai, 2004) szoftver segítségével készült.

9. A NimB és NimC szekvenciák filogenetikájának Bayes MCMC rekonstrukciója nyomán kapott poszterior mintákból számolt konszenzushálózatok

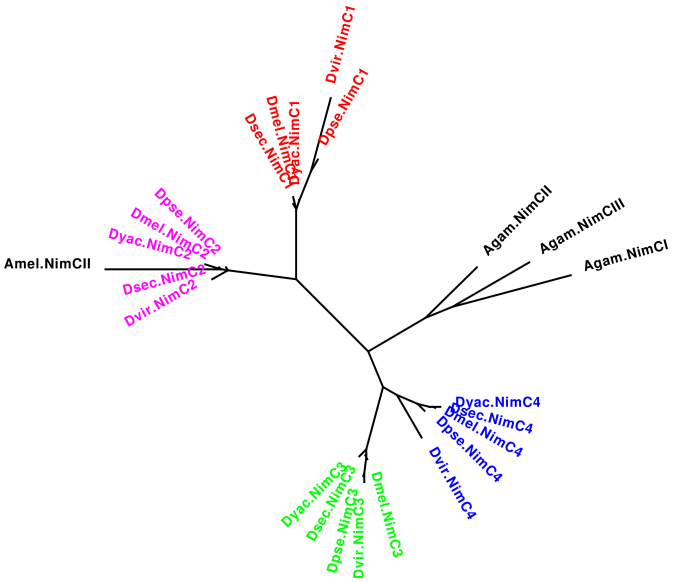
A konszenzushálózatokat az $x = 0,2$-es szinten számoltuk. A Drosophila ortológokat azonos színekkel kiemeltük.

9.1 NimB, csak gap információk

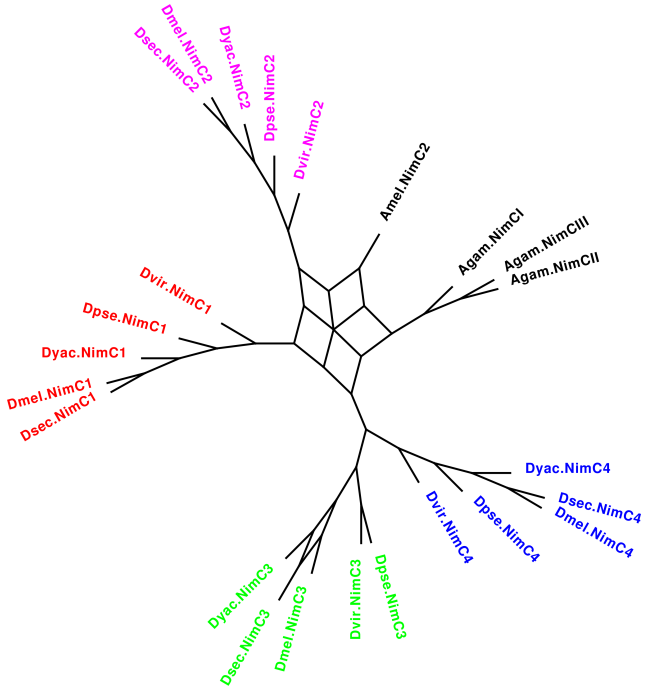
9.2 NimB, gap információk és szekvenciák



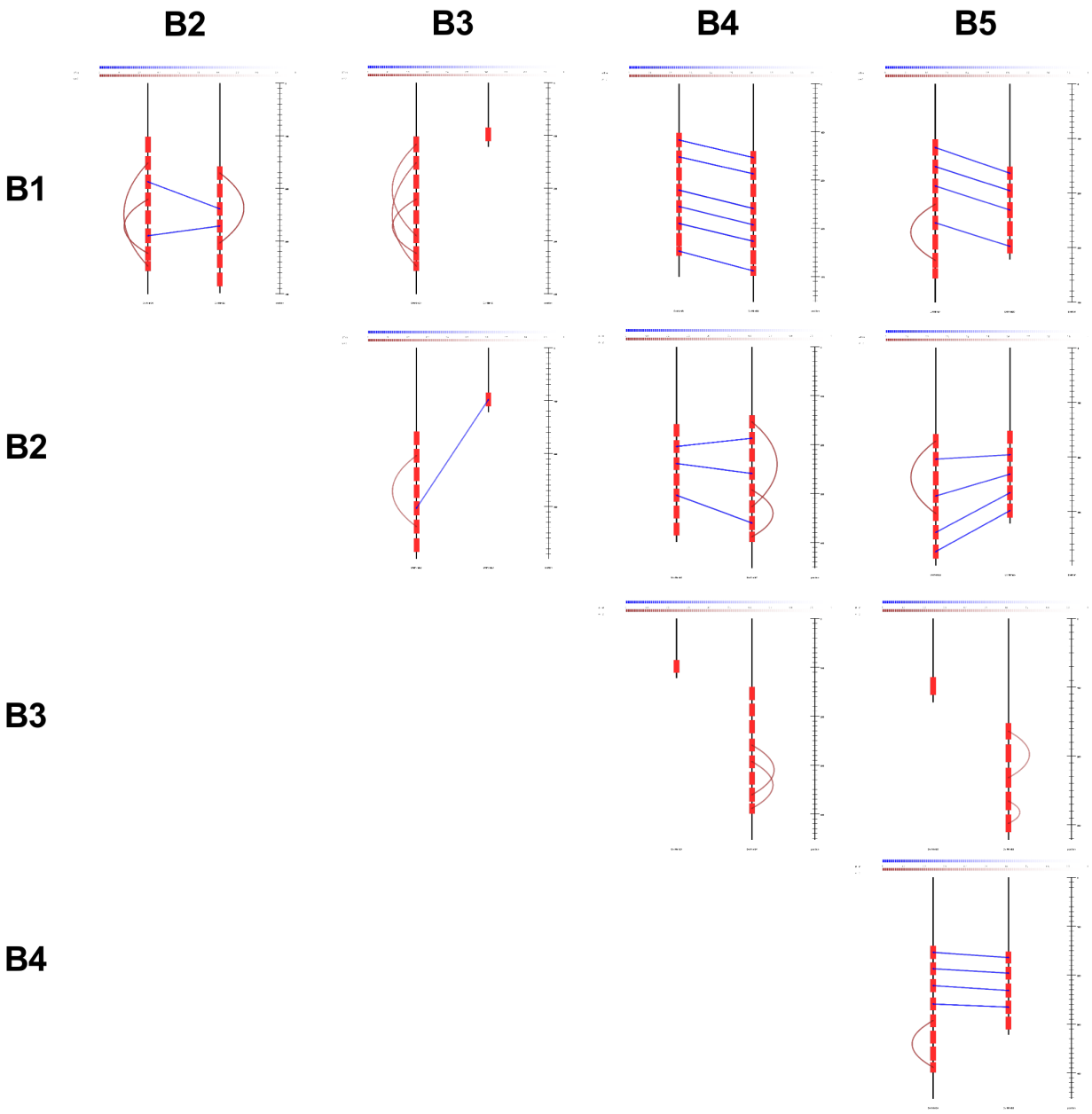
9.3 NimC, csak gap információk



9.4 NimB, gap információk és szekvenciák



10. A *D. virilis* NimB szekvenciák párbaállításával készült PRHD-k.



11. A Nimrod klaszter konzerváltságának vizsgálata során felhasznált adatok a források megjelölésével

Genom	**A gének száma**	**A gének számának forrása**	**A géncsaládok száma a klaszterben**	**A *D. mel.*-el közös géncsaládok száma**
D. melanogaster	13733	Drosophila 12 Genomes Consortium, 2007	28	-
A. gambiae	13683	Holt és mtsai, 2002	13	5
A. mellifera	10157	Honeybee Genome Sequencing Consortium, 2006	9	5
P. humanus	11194	Vectorbase	7	4