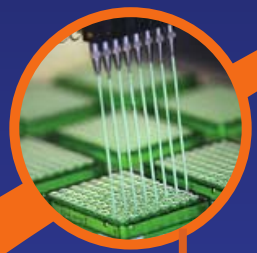
Genom annotáció: a genomszekvencia értelmezése

Patthy László
MTA TTK Enzimológiai Intézet, Budapest

Budapesti Műszaki és Gazdaságtudományi Egyetem
Matematikai Modellalkotás szeminárium
2012. november 20

Az első sejtes organizmus, a Haemophilus influenzae genomjának szekvenciáját 1995-ben határozták meg, a humán genom szekvenciáját 2001-ben közölték.

A remények szerint a genom-méretekben gyűjtött adatok bioinformatikai elemzésének köszönhetően ***hatalmas lendületet kap a gyógyszeripar*** *(új gyógyszercélpontok azonosításának köszönhetően)...*



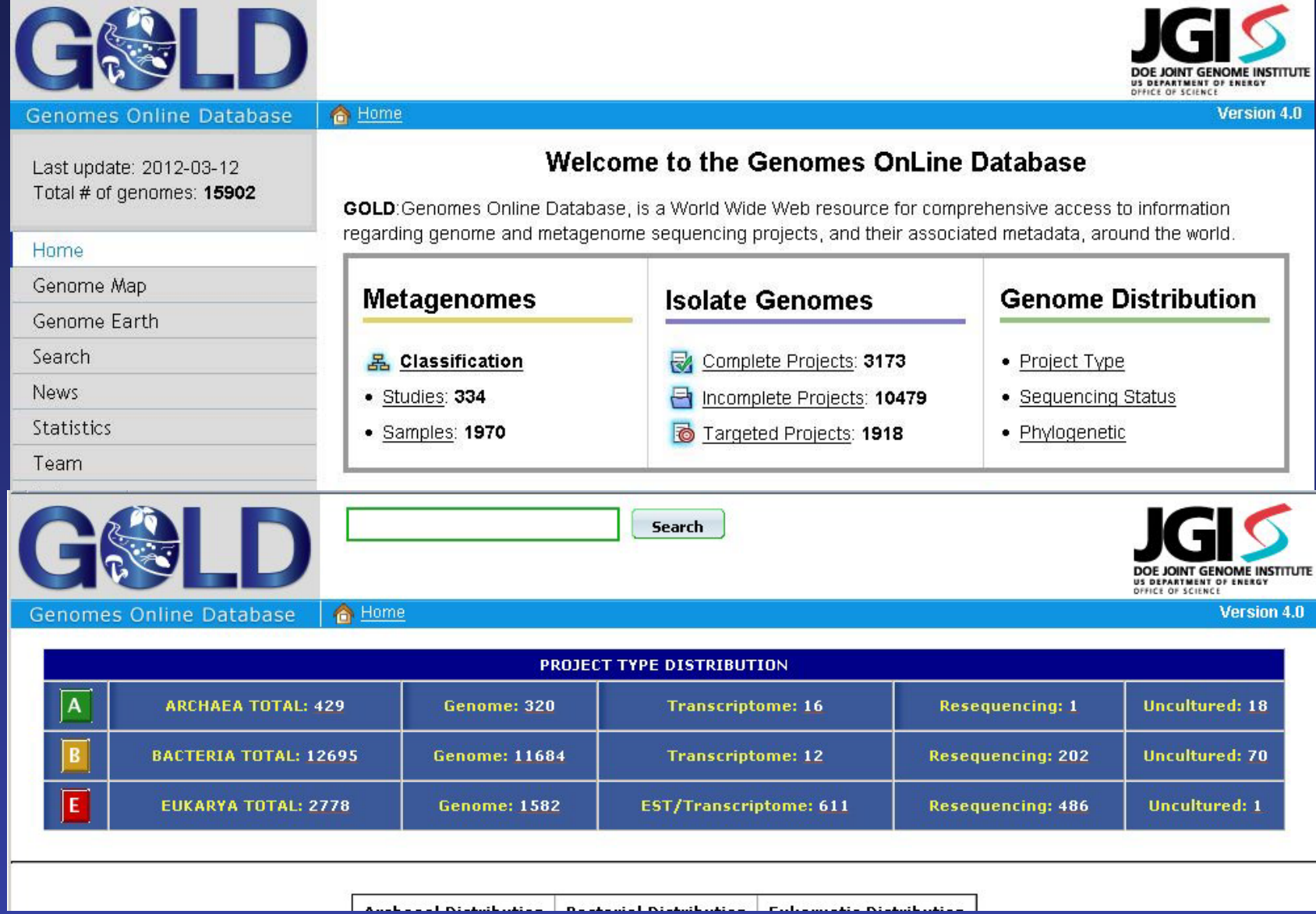
orvostudomány

mezőgazdaság

biotechnológia

környezetvédelem

....és általánossá válik a ***személyre szabott orvoslás,*** *amikor az egyén genomszekvenciájának ismeretében határozhatjuk meg az orvoslás legcélszerűbb módjait….*



Ma már három ezernél több teljes genomszekvencia ismert és tíz ezernél több genomprojekt van folyamatban.

http://www.genomesonline.org/

Fontos hangsúlyozni, hogy a genom szekvenciájának meghatározása nem azonos jelentésének megértésével.

„Last June, we announced that researchers had collected 90 percent of the DNA letters that make up the text of the human genome sequence. Now we have achieved another major advance - by ***reading, from cover to cover, the first draft of this "Book of Life"*** *and reporting on the stunning surprises we encountered along the way.*

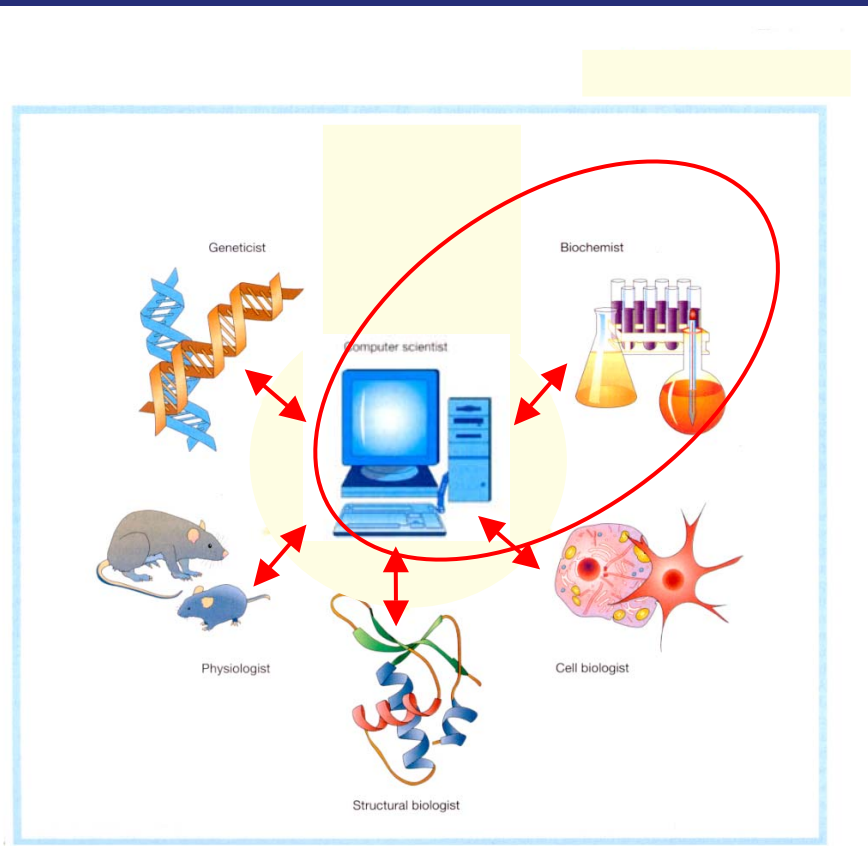
As you will hear today, this Book of Life is actually at least three books. ***It's a history book****: a narrative of the journey of our species through time. It's a shop manual: an incredibly detailed* ***blueprint for building every human cell****. And it's a transformative* ***textbook of medicine****: with insights that will give health care providers immense new powers to treat, prevent and cure disease. We are delighted by what we've already seen in these books. But we are also profoundly humbled by the privilege of turning the pages that describe the miracle of human life,* ***written in the mysterious language of all the ages, the language of God.”***

Remarks at the Press Conference Announcing Sequencing and Analysis of the Human Genome
Dr. Francis S. Collins, Director, National Human Genome Research Institute, February 12, 2001
http://www.genome.gov/10001379

A funkcionális genomika feladata a genom annotáció, a genomszekvencia értelmezése.

A funkcionális genomika egyik fontos eszlöze a bioinfiormatika. A bioinformatika interdiszciplináris tudományterület, az élettudományok által felvetett kérdésekre az informatika és matematika eszköztárának felhasználásával keres választ.

A bioinformatika a biológiai adatok számítógépes tárolásával, rendszerezésével, elemzésével és értelmezésével foglalkozó tudomány.



DNS Annotáció

Gén definíció (alternativ splicing)

- *Regulátorok & Promóterek*
- *Variációk (CNVk és SNPk)*

RNS Annotáció

- *Expresszió*

Proteom annotáció

- ***Fehérje családok***
- ***Fehérje szerkezet***
- *Poszttranszlációs módosítások,*
- ***szubcelluláris lokalizáció***

Funkcionális annotáció

- *Molekuláris funkciók*
- *Kölcsönhatások*
- *Útvonalak és hálózatok*
- ***Biológiai szerep***

Megfogalmazódott az a remény, hogy megbízható bioinformatikai módszerekkel végzett predikciók segíthetik (vagy sok esetben feleslegessé is tehetik) a költséges és időigényes in vitro vagy in vivo kisérleti munkát

A genom annotáció a bioinformatikával szemben azt az igényt támasztja, hogy olyan számítógépes eljárásokat fejlesszen ki, melyek segítségével a genom szekvencia elemzése alapján in silico módszerek alkalmazásával juthatunk el a genom részletes és megbízható funkcionális annotációjáig.

A genom-szekvencia bioinformatikai úton történő értelmezésének egyik alapvető lépése a ***fehérjekódoló gének azonosítása.***

A génazonosítás még mindig meglévő bizonytalanságait illusztrálhatjuk azzal, hogy közel egy évtizeddel az emberi genom szekvenciájának meghatározása után ***még mindig bizonytalan az emberi genomban található fehérjekódoló gének száma.***

Distinguishing protein-coding and noncoding genes in the human genome

Michele Clamp*†, Ben Fry*, Mike Kamal*, Xiaohui Xie*, James Cuff*, Michael F. Lin‡, Manolis Kellis*‡, Kerstin Lindblad-Toh*, and Eric S. Lander*†§¶‖

*Broad Institute of Massachusetts Institute of Technology and Harvard, 7 Cambridge Center, Cambridge, MA 02142; ¶Department of Biology and *Co
Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139; §Whitehead Institute for Biomedical Res
9 Cambridge Center, Cambridge, MA 02142; and ‖Department of Systems Biology, Harvard Medical School, Boston, MA 02115

Contributed by Eric S. Lander, October 3, 2007 (sent for review August 1, 2007)

Although the Human Genome Project was completed 4 years ago, the catalog of human protein-coding genes remains a matter of controversy. Current catalogs list a total of ≈24,500 putative protein-coding genes. It is broadly suspected that a large fraction of these entries are functionally meaningless ORFs present by chance in RNA transcripts, because they show no evidence of evolutionary conservation with mouse or dog. However, there is currently no scientific justification for excluding ORFs simply because they fail to show evolutionary conservation: the alternative hypothesis is that most of these ORFs are actually valid human genes that reflect gene innovation in the primate lineage or gene loss in the other lineages. Here, we reject this hypothesis by carefully analyzing the nonconserved ORFs—specifically, their properties in other primates. We show that the vast majority of these ORFs are random occurrences. The analysis yields, as a by-product, a major revision of the current human catalogs, cutting the number of protein-coding genes to ≈20,500. Specifically, it suggests that nonconserved ORFs should be added to the human gene catalog only if there is clear evidence of an encoded protein. It also provides a principled methodology for evaluating future proposed additions to the human gene catalog. Finally, the results indicate that there has been relatively little true innovation in mammalian protein-

large-scale cDNA sequencing projects yield ever-larger numb
transcripts (2). The three most widely used human gene ca
[Ensembl (4), RefSeq (5), and Vega (6)] together contain a t
≈24,500 protein-coding genes. It is broadly suspected that
fraction of these entries is simply spurious ORFs, because the
no evidence of evolutionary conservation. [Recent studies in
that only ≈20,000 show evolutionary conservation with do
However, there is currently no scientific justification for exc
ORFs simply because they fail to show evolutionary conser
the alternative hypothesis is that these ORFs are valid human
that reflect gene innovation in the primate lineage or gene
other lineages. As a result, the human gene catalog has rem
in considerable doubt. The resulting uncertainty hampers bi
ical projects, such as systematic sequencing of all human ge
discover those involved in disease.

The situation also complicates studies of comparative ge
and evolution. Current catalogs of protein-coding genes vary
among mammals, with a recent analysis of the dog genom
reporting ≈19,000 genes and a recent article on the mouse g
(2) reporting at least 33,000 genes. The difference is attribut
nonconserved ORFs identified in cDNA sequencing projec

Proc Natl Acad Sci U S A. 2007 Dec 4;104(49):19428-33.

Ennél is súlyosabb problémát jelent, hogy ***az azonosított gének jelentős hányadáról bizonyosodik be, hogy a bioinformatikai módszerekkel megjósolt szerkezetük téves.***

A jelenlegi génpredikciós módszerek bizonytalanságai így komoly problémákat okoznak a (tévesen) megjósolt gének/fehérjék funkciójának további vizsgálatában, az expressziójukat szabályozó genomikai elemek meghatározásában és ***megnehezítik a genominformáció gyakorlati hasznosítását.***

A génazonosításra használt számítógépes programmok fő típusai

i) Expressziós adatokat (cDNS-, EST-, fehérje-szekvenciák) használó, extrinsic megközelítések

ii) Intrinsic (ab initio vagy de novo) megközelítések, melyek csak az adott genom szekvenciájának információtartalmára támaszkodnak.

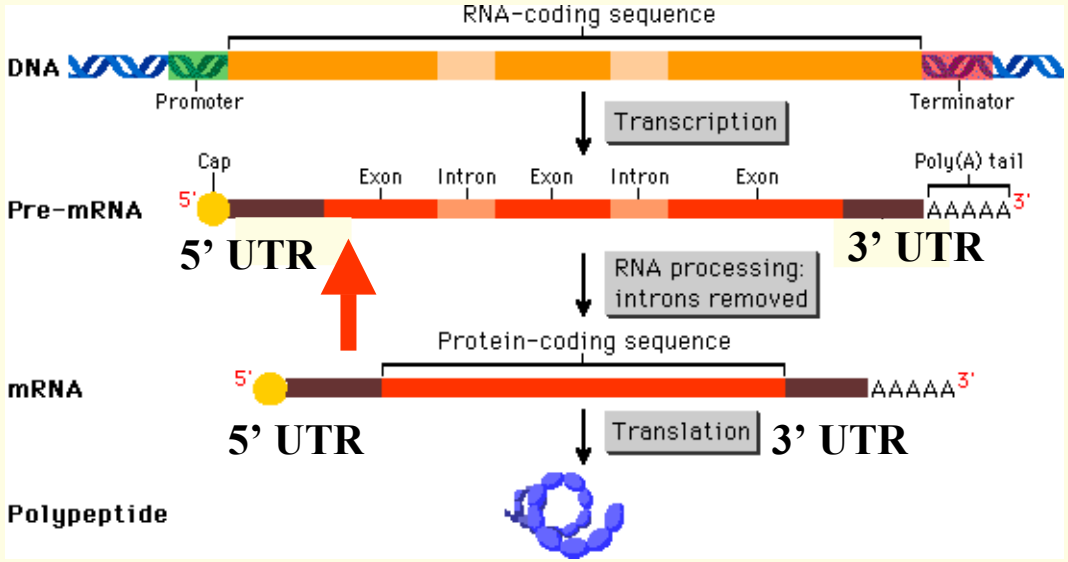
iii) Összehasonlító genomikai megközelítések, melyek két vagy több rokon faj genom-szekvenciájának összehasonlításával szerzett evolúciós információt hasznosítják.

iv) Komplex megközelítések, melyek valamennyi információ-típust felhasználják a génazonosításra.

Extrinsic megközelítések

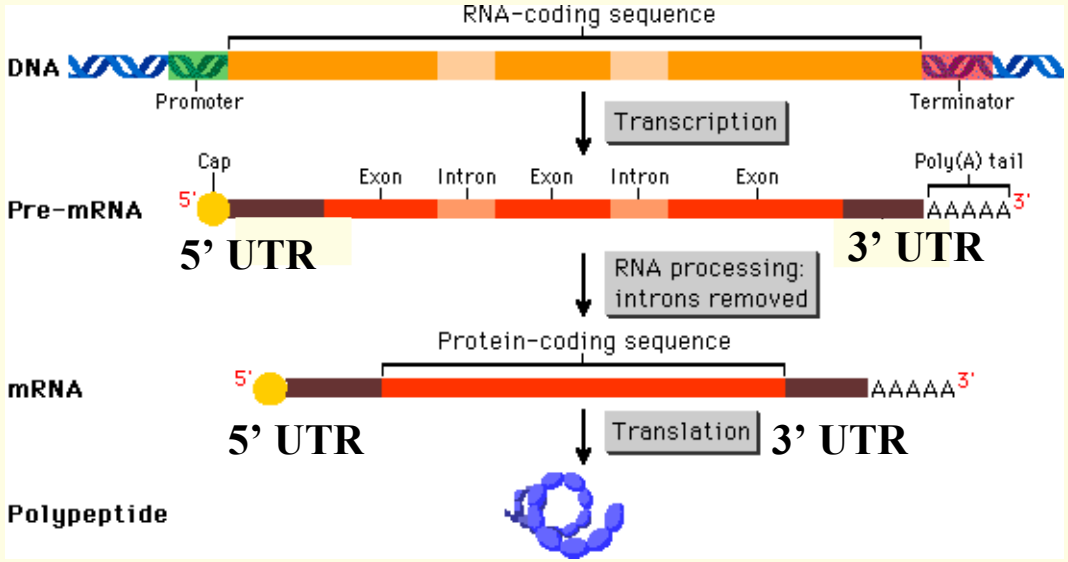
Ezek a módszerek azt vizsgálják, hogy az adott genomban hol találhatók olyan szekvenciák, melyek ***azonosak*** *vagy* ***szignifikáns hasonlóságot mutatnak ismert*** *(ugyanabból a fajból, vagy rokon fajokból származó)* ***mRNS-, cDNS-, EST-*** *vagy* ***fehérje-szekvenciákkal****: a génazonosítás alapja extrinsic információ.*

Például, ha egy adott génről származó ***teljes hosszúságú mRNS, cDNS*** *szekvenciáját ismerjük akkor egyértelműen azonosíthatjuk azt a genomrégiót (a* ***transzkripció iniciációs helytől a poliadenilációs helyig****) amelyről az átíródott; a teljes hosszúságú mRNS, cDNS definiálja azt az* ***exon-intron szerkezetet*** *is amely az adott mRNS szempontjából releváns.*



Extrinsic megközelítések

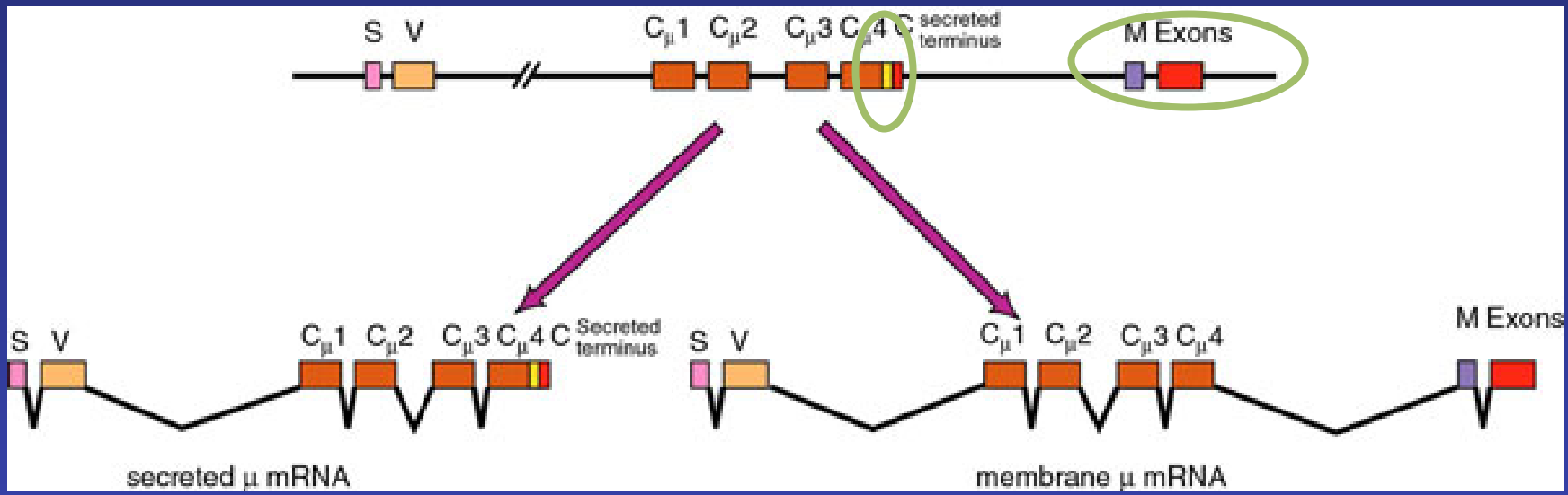
A módszer elvi korlátja, hogy a mRNS szekvencia ismerete nem alkalmas a transzkripció iniciációs hely előtt található regulációs elemek azonosítására ....



Funded by | BIOSAPIENS NETWORK | Coordinated by EBI

Extrinsic megközelítések

... és nem nyerünk információt a gén esetleges alternatív exon-intron szerkezetére vonatkozóan sem.

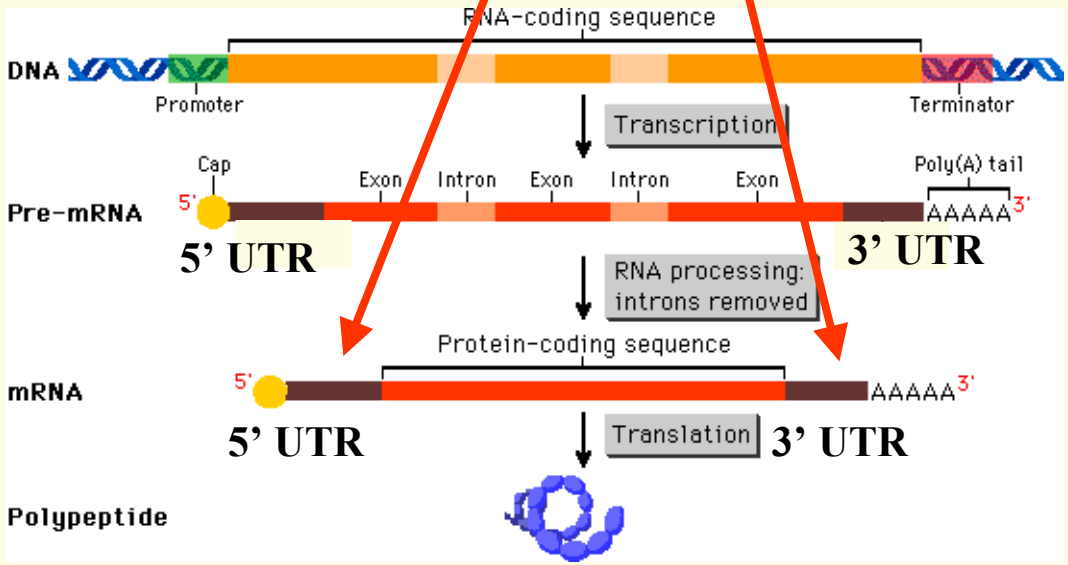


Alternative splicing of the gene for the μ heavy chain of the mouse IgM immunoglobulin

Funded by | BIOSAPIENS NETWORK | EBI Coordinated by

Extrinsic megközelítések

Ha csak fehérje szekvencia áll rendelkezésre, az csak a fehérje-kódoló régióra vonatkozóan ad információt (a transzláció iniciációs helytől a stop kodonig), de nem ad felvilágosítást a gén 5' és 3' nemtranszlált régióira vonatkozóan.



Extrinsic megközelítések

Nem elvi, hanem gyakorlati korlátja ezeknek a megközelítéseknek, hogy az adatbázisokból gyakran hiányoznak az alacsony szinten expresszált gének transzkriptumai, valamint, hogy az adatbázisokban található mRNS, cDNS szekvenciák jelentős hányada nem teljes hosszúságú.

Az új szekvenálási technológiáknak köszönhetően azonban ezek a gyakorlati problémák a közeljövőben eliminálhatók lesznek.

Intrinsic megközelítések

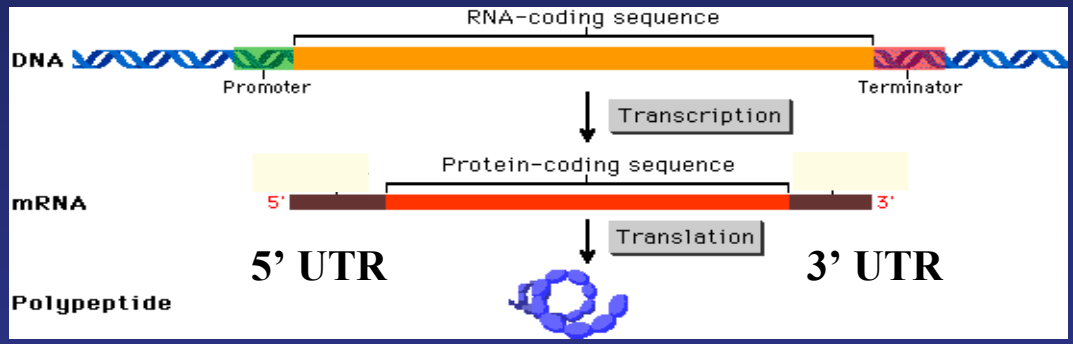
Az intrinsic (ab initio, de novo) gén predikciós megközelítések a géneket a ***fehérjekódoló génekre jellemző tulajdonságok felismerése*** *alapján azonosítják a genom szekvenciában.*

Az azonosítás alapjául két fő karakter-típust használnak.

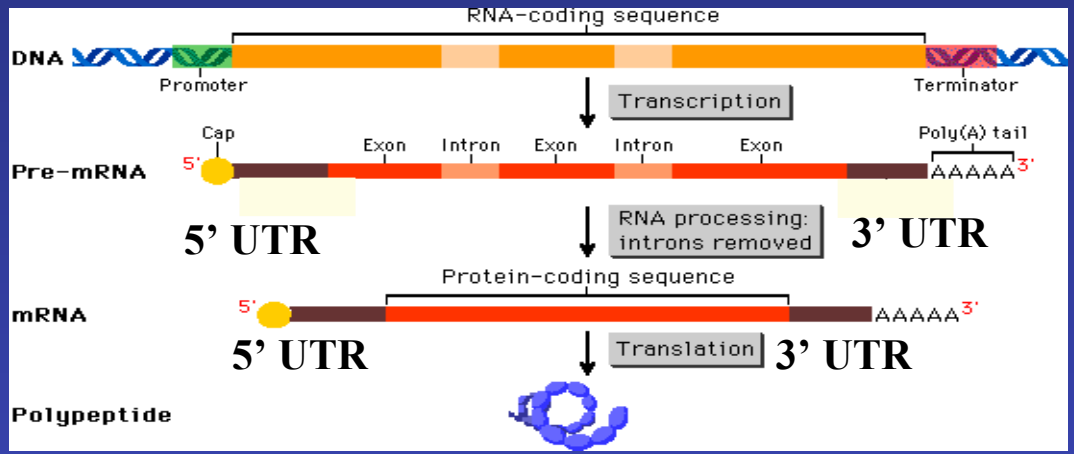
Az első csoportba ***specifikus szekvencia jelek*** *tartoznak, melyek pl. promoter régióra, transzkripció start helyre, exon/intron határra, poliadenilációs helyre stb. utalnak.*

A második csoportba ***tartalom-típusú információk*** *tartoznak:* ***a fehérjekódoló gének olyan jellegzetes statisztikus tulajdonságai, melyek megkülönböztetik őket a nem-kódoló régióktól.***

A prokarióta és eukarióta genomok valamint a prokarióta és eukarióta fehérjekódoló gének közötti jelentős eltérések hatással vannak az intrinsic génpredikciós módszerek teljesítőképességére.



A prokarióta fehérje-kódoló gén ***nem tartalmaz intronokat***



Az eukarióta fehérje-kódoló gén ***intronokat tartalmaz***

A National Human Genome Research Institute, National Institute of Health ***2003-ban*** *indította el az* ***ENCODE*** *(the ENCyclopedia Of DNA Elements)* ***projektet*** *azzal a céllal, hogy* ***a humán genom valamennyi funkcionális elemét azonosítsa.***

http://www.genome.gov/10005107

Funded by | BIOSAPIENS NETWORK | Coordinated by EBI

Intrinsic megközelítések

*Minthogy a **prokarióta fehérjekódoló génekben** nincsenek intronok, **a génekre hosszú, folytonos Open Reading Frame jellemző** (vagyis hosszú szakaszokban nem fordulnak elő stop kodonok).*

*Minthogy nem-kódoló genomikus régiókban stop kodonok (három a 64 triplet közül) véletlenszerűen nagy gyakorisággal fordulnak elő, a **stop kodonok statisztikailag szignifikáns hiánya megbízhatóan és érzékenyen jelzi fehérjekódoló gén jelenlétét prokarióta genomok esetén.***

Funded by | BIOSAPIENS NETWORK | Coordinated by EBI

Intrinsic megközelítések

A prokarióta gének kodon-használata és bázis összetétele ugyancsak szignifikánsan eltér a nem-kódoló (intergénikus) régiókétól.

A felsorolt tulajdonságoknak (és a prokarióta genomok nagy géndenzitásának) köszönhetően az ***intrinsic génpredikciós módszerek prokarióta genomok esetén viszonylag egyszerűek és megbízható eredményt adnak.***

Funded by | BIOSAPIENS NETWORK | Coordinated by EBI

Intrinsic megközelítések

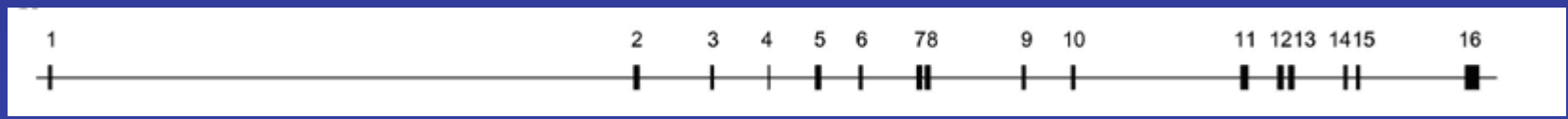
Az ab initio gene predikciós módszerek az eukarióta genomok esetén sokkal kevésbé megbízhatóak: megbízhatóságuk rohamosan csökken a genom méret, az intron-exon arány növekedésével: eukarióta genomok esetén a fehérjekódoló régiók sokszor csak a genom szekvencia néhány százalékát teszik ki.

	Gene density (genes/Mb)	Intergenic (% of total)	Intron (% of total)	Exon (% of total)	Intron/exon ratio
Eukaryotes					
Saccharomyces cerevisiae	446	31	1	68	0.01
Arabidopsis thaliana	215	45	21	31	0.77
Caenorhabditis elegans	196	47	26	27	0.96
Drosophila melanogaster	113	63	17	20	0.85
Homo sapiens	11	75	24	1	24.00

Funded by | BIOSAPIENS NETWORK | Coordinated by EBI

Intrinsic megközelítések

Az eukarióta fehérje-kódoló génekben számos, nagyméretű intron választja el a, gyakran kisméretű, fehérje-kódoló exonokat. Minthogy a fehérjekódoló szekvenciákat sokszor kis fragmentumokra darabolják az intronok, azok a statisztikus eszközök melyek sikerrel használhatóak a prokarióta genomok esetén (stop kodonok elkerülése, kodon használat és bázis összetétel különbözősége a kódoló és nem-kódoló régiókban stb.) kevésbé megbízható eredményt adnak a magasabb rendű eukarióták intron-gazdag genomjai esetén.



A peptidilarginin deiminász 4-t kódoló, 16 exonból álló PADI4 gén exon-intron szerkezete

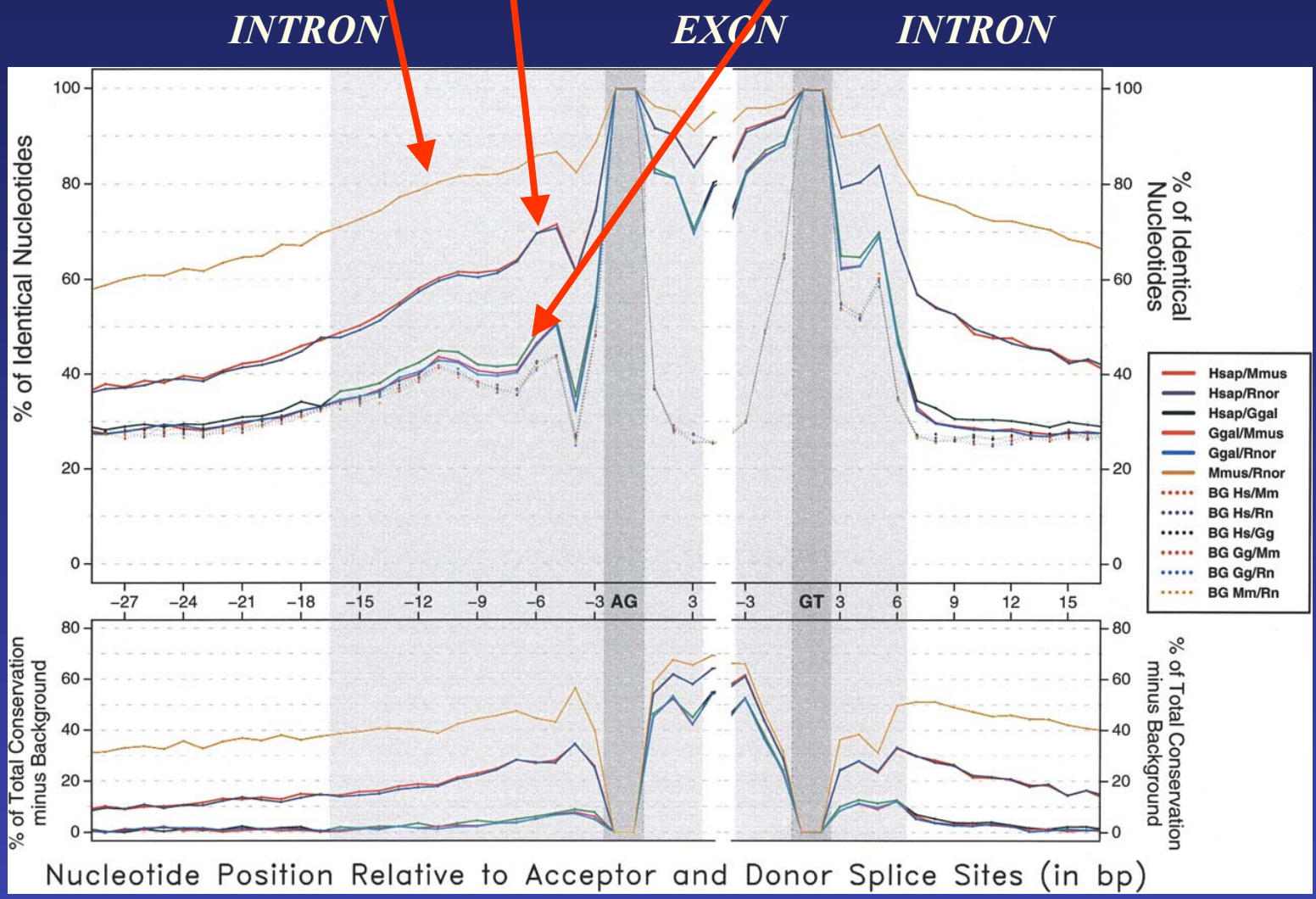
Összehasonlító genomikai megközelítések

A genomprojekteknek köszönhetően egyre több teljes genomszekvencia áll rendelkezésünkre, ennek köszönhetően egyre nagyobb teret nyernek az összehasonlító genomikai megközelítésen alapuló génpredikciós módszerek.

Ezeknek a megközelítéseknek az az elvi alapja, hogy a funkciót hordozó régiók általában konzervatívabbak, lassabban változnak az evolúció során, mint az esszenciális funkciót nem hordozó régiók.

A fehérjekódoló gének esetén ezért várható, hogy ha rokon fajok genom szekvenciáit összehasonlítjuk, akkor a gének azonosíthatók esszenciális régióik (exonjaik, szabályozó elemeik stb.) kiugró konzervativizmusa alapján.

Sequence conservation level in the vicinity of orthologous GT-AG splice sites at different evolutionary distances (mouse-rat, rodent-human, mammalian-chicken comparisons).



Abril JF, Castelo R, Guigo R. Comparison of splice sites in mammals and chicken.Genome Res. 2005 Jan;15(1):111-9.

Összehasonlító genomikai megközelítések

Hangsúlyozni kell, hogy egy genomikus régió konzervativizmusa csak a régió fontosságát bizonyítja, de nem szükségszerűen jelenti, hogy a régió fehérjét kódol. ***A kodon használat, a mutációs mintázat*** *(nem-szinoním/szinoním mutációk aránya)* ***elemzése alapján lehet eldönteni, hogy a konzervált régió fehérjét kódol-e vagy sem.***

Komplex megközelítések

A génpredikciók megbízhatósága jelentősen növelhető ha ***az extrinsic, intrinsic és összehasonlító genomikai megközelítésekből nyert valamennyi információkat egyesítjük.***

A különböző számítógépes megközelítések teljesítőképességét, megbízhatóságát pontosan, kvantitatíve jellemezte egy szisztematikus vizsgálat *(Guigo et al., 2006, EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol. 2006;7 Suppl 1:S2.1-31.)*

Az összehasonlított számítógépes módszereket a következő kategóriákba sorolták:

1) ***EST-, mRNS-, és fehérje-alapú módszerek*** *(AUGUSTUS-EST, PARAGON+NSCAN_EST, ACEVIEW,* ***ENSEMBL****, EXOGEAN, EXONHUNTER, ACEMBLY, ECGene, MGCGene)*
2) ***Egy-genom vizsgálatán alapuló ab initio módszerek*** *(AUGUSTUSabinit, GENEMARKhmm, GENEZILLA, GENEID,* ***GENESCAN****)*
3) ***Összehasonlító genomikai módszerek*** *(AUGUST-dual, ACESCAN, DOGFISH-C, NSCAN, SAGA, MARS, SGP2,* ***TWINSCAN****)*
4) ***Komplex módszerek*** *(AUGUSTUSany,* ***FGENESH++****, JIGSAW, PARAGONany, CCDSGene, KNOWNGene,* ***REFSEQ****)*

Megbízhatóan annotált, ismert humán genomikus szekvenciákat (vakon) különböző génpredikciós programokkal elemezték és a predikciókat összehasonlították az annotációkkal.

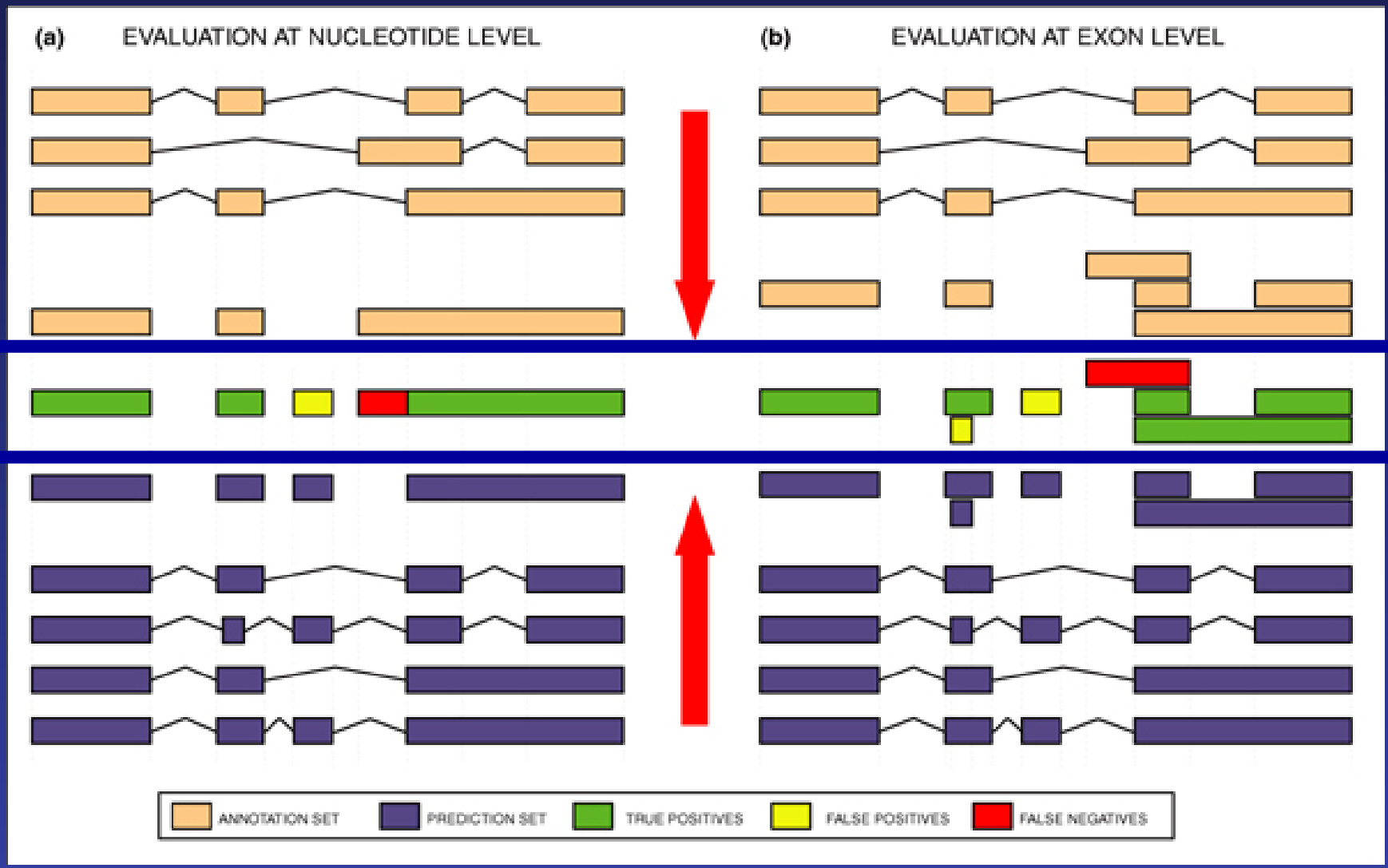
A predikciókat nukleotid-, exon-, transzkriptum- és gén szinten elemezték, hogy megállapítsák, hogy a különböző predikciók, mennyire egyeznek az annotációkkal.

Minden szinten két paramétert határoztak meg:

- ***Szenzitivitás, Sn:*** *az annotált (valós) „tulajdonság” (nukleotid, transzkriptum, exon, gén) mekkora hányadát jósolta meg helyesen az adott módszer.*

- ***Specificitás, Sp:*** *a prediktált tulajdonság (nukleotid, transzkriptum, exon, gén) mekkora hányada korrekt (megegyezik az anotációval).*

Minden egyes program esetén meghatározták a ***szenzitivtás és specificitás átlagát ((Sn + Sp)/2), minthogy ez tükrözi legjobban az adott módszer megbízhatóságát.***

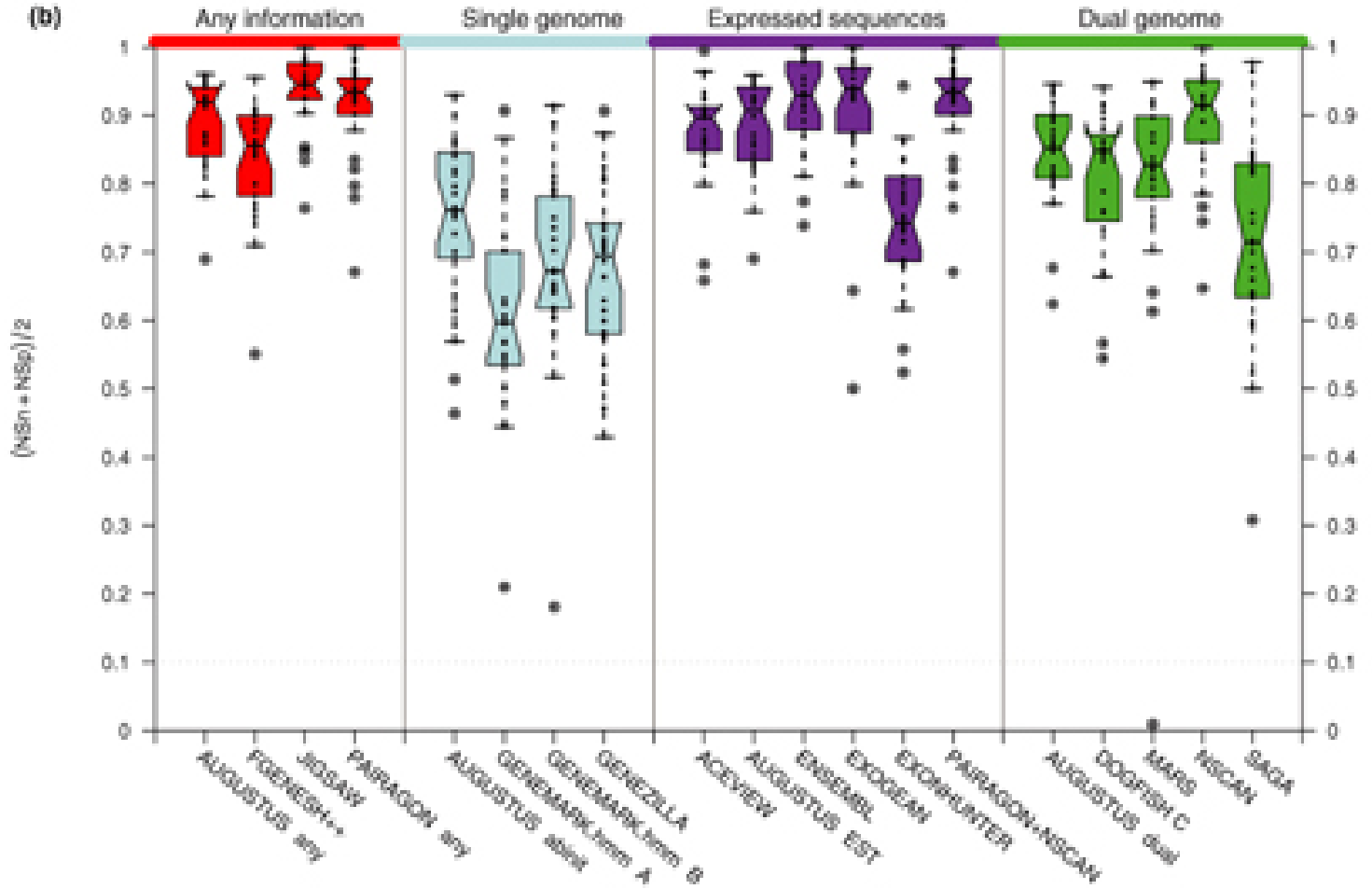


Gene feature projection for evaluation of the accuracy of predictions

missing exons

wrong exons

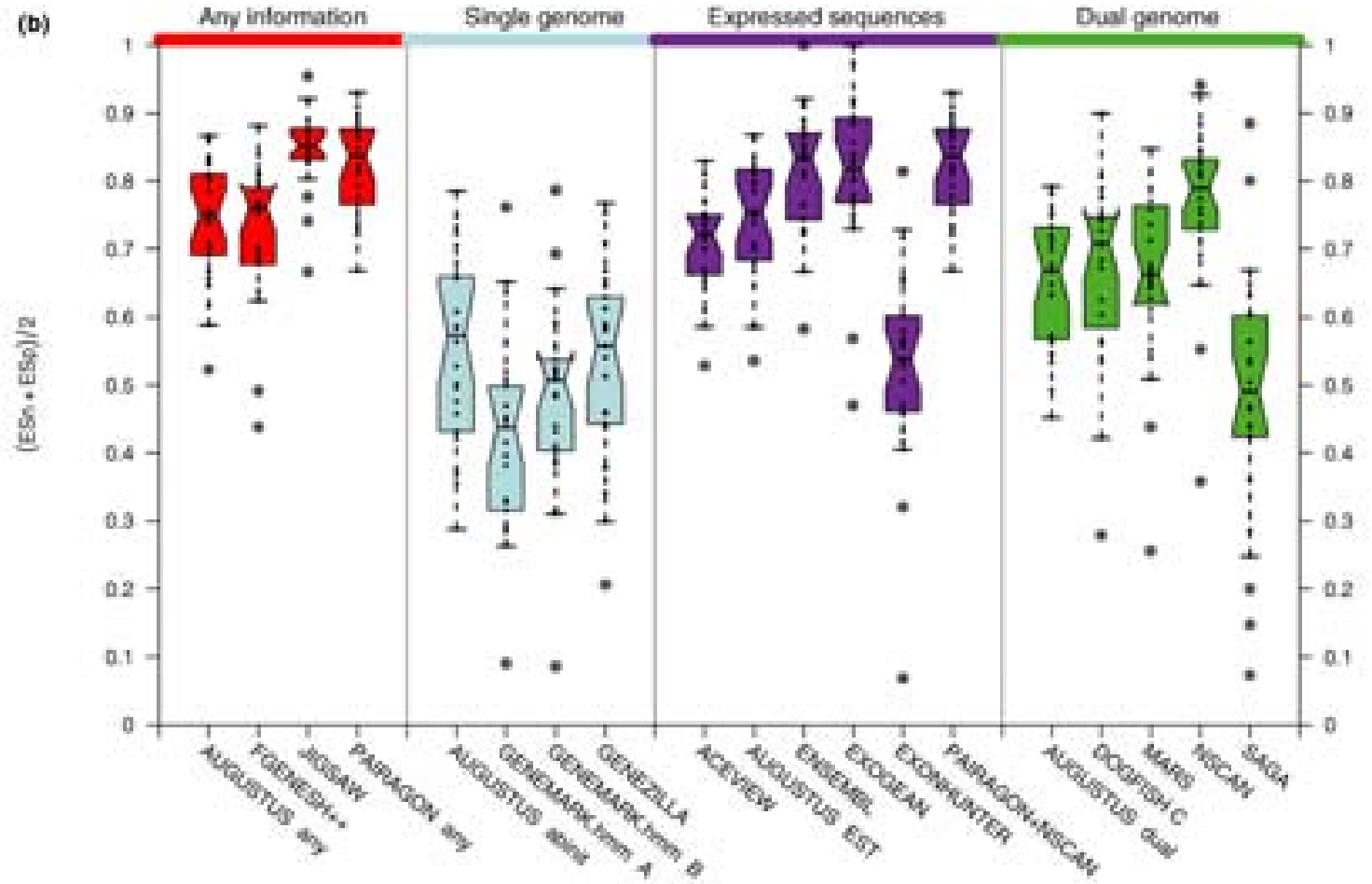
Guigo et al., Genome Biol. 2006;7 Suppl 1:S2.1-31.



Gene prediction accuracy at the nucleotide level. Boxplots of the average sensitivity and specificity ((Sn + Sp)/2) for each program.

At the nucleotide level, sensitivity (Sn) is the proportion of annotated nucleotides that is correctly predicted, and specificity (Sp) is the proportion of predicted nucleotides that is correct.

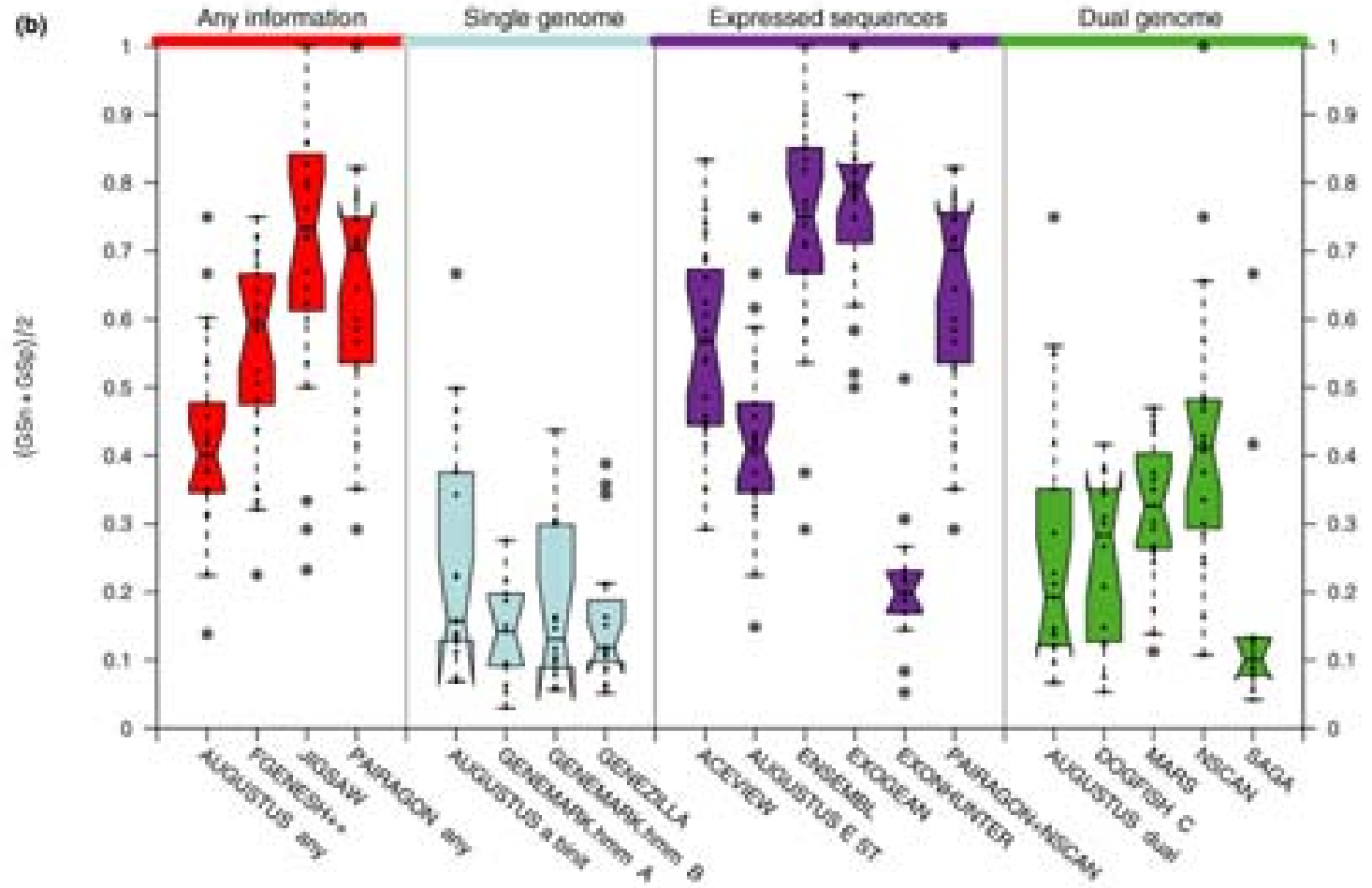
Guigo et al., Genome Biol. 2006;7 Suppl 1:S2.1-31.



Gene prediction accuracy at the <u>exon level</u>. Boxplots of the average sensitivity and specificity ((Sn + Sp)/2) for each program.

The exon level accuracy is calculated with the requirement that an exon in the prediction must have identical start and end coordinates as an exon in the annotation to be counted correct.

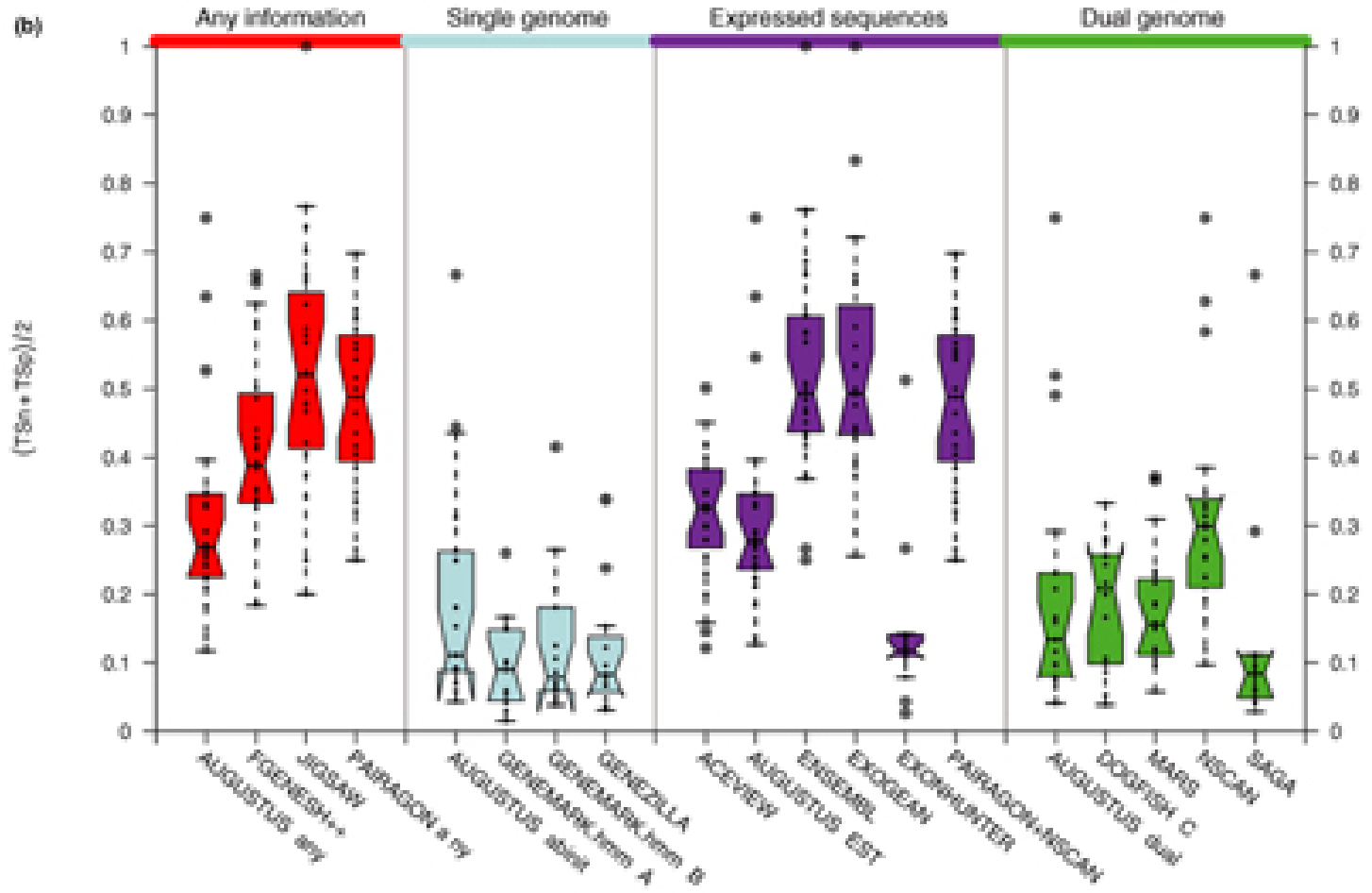
Guigo et al., Genome Biol. 2006;7 Suppl 1:S2.1-31.



Gene Prediction Accuracy at the gene level. Boxplots of the average sensitivity and specificity ((Sn + Sp)/2) for each program.

A gene is counted correct if at least one transcript in the locus is correct .

Guigo et al., Genome Biol. 2006;7 Suppl 1:S2.1-31.



Gene prediction accuracy at the transcript level. Boxplots of the average sensitivity and specificity ((Sn + Sp)/2) for each program.

A transcript is accurately predicted if the beginning and end of translation are correctly annotated and each of the 5' and 3' splice sites for the coding exons are correct.

Guigo et al., Genome Biol. 2006;7 Suppl 1:S2.1-31.

Újabb vizsgálatok is megerősítették, hogy

i) *Egyetlen módszer sem ad tökéletes predikciót*

ii) *Általában azok a módszerek a legmegbízhatóbbak, melyek mRNS- és fehérje-szekvencia információra, illetve valamennyi információ típusra támaszkodnak*

iii) *A több genom összehasonlító elemzését alkalmazó módszerek megbízhatóbbak, mint az egyetlen genomot használó, ab initio predikciós módszerek*

iv) *Nukleotid szinten (és exon szinten) – (a legkevésbé szigorú megbízhatósági kritériumok) – egyetlen predikciós módszer sem azonosította helyesen a nukleotidok több mint ~90%-t (az exonok több, mint ~85%-t).*

v) *Transzkriptum szinten – (a legszigorúbb megbízhatósági kritérium) – a legjobb predikciós módszer is csak a kódoló transzkriptumok ~50%-t képes helyesen megjósolni.*

(Guigo et al., 2006, EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol. 2006;7 Suppl 1:S2.1-31.)
Harrow J, Nagy A, Reymond A, Alioto T, Patthy L, Antonarakis SE, Guigó R. Identifying protein-coding genes in genomic sequences. Genome Biol.2009;10(1):201.

A rendelkezésre álló ***eukarióta génpredikciós módszerek viszonylagos megbízhatatlansága*** *következtében, várható, hogy a prediktált adatokat (is) tartalmazó* ***adatbázisok szennyezettek, tévesen prediktált*** *(mispredicted)* ***nukleotid és fehérje szekvenciákkal.***

Csoportunk munkájának fő célkitűzése az volt, hogy a jelenleg alkalmazott módszerek hibáinak kiderítésével elősegítsük az eddigieknél megbízhatóbb génpredikciós eljárások kidolgozását.

A nyilvánvaló kérdések:

- Hogyan tudjuk eldönteni, hogy egy prediktált szekvencia ***korrekt vagy téves?***

- Milyen jelek utalhatnak arra, hogy egy prediktált szekvencia téves?

A MisPred projekt

A MisPred projekt elvi alapja az a megfontolás volt, hogy ***egy prediktált fehérjekódoló gén valószínűleg téves (mispredicted)*** *ha annak valamely szerkezeti/funkcionális „tulajdonsága” ellentmond a fehérjekódoló génekre és fehérjékre vonatkozó tudásunk valamely törvényszerűségének.*

Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Bányai L, Patthy L. Identification and correction of abnormal, incomplete and mispredicted proteins in public databases. BMC Bioinformatics. 2008 Aug 27;9:353.

Fehérje szintjén megfogalmazva:

Ha egy prediktált gén által kódolt hipotetikus fehérje sérti azoknak a törvényszerűségeknek valamelyikét, melyek korrekt térszerkezetű, korrekt szubcelluláris lokalizációjú funkcióképes fehérjék különböző csoportjaira érvényesek, akkor az a fehérje életképtelennek, funkcióképtelennek minősül és az azt kódoló gén annotációja valószínűleg téves.

Az ilyen „tudásalapú" megközelítések számát csak a fehérjékre és fehérje-kódoló génekre vonatkozó törvényszerűségek száma korlátozza....

A MisPred eljárás néhány ***minőségellenőrző eszköze*** *azt a kérdést vizsgálja, hogy egy prediktált fehérje* ***eljuthat-e abba szubcelluláris kompartmentbe, ahol elnyeri korrekt, stabil és funkcióképes szerkezetét.*** *Ezeknek az eszközöknek az alapja az a megfontolás, hogy a* ***nem megfelelő kompartmentbe kerülő*** *(mislocalized)* ***fehérje térszerkezete hibás, a fehérje instabil és/vagy funkcióképtelen.***

Például, azok a prediktált fehérjék, melyek ***extracelluláris doméneket*** *tartalmaznak,* ***de hiányzanak belőlük*** *azok* ***a szekvencia jelek,*** *melyek az extracelluláris doméneket a számukra megfelelő extracelluláris térbe irányítják* ***nem nyerik el jellegzetes, stabil térszerkezetüket és így nem tölthetnek be biológiai funkciót.***

Funded by | BIOSAPIENS NETWORK | Coordinated by EBI

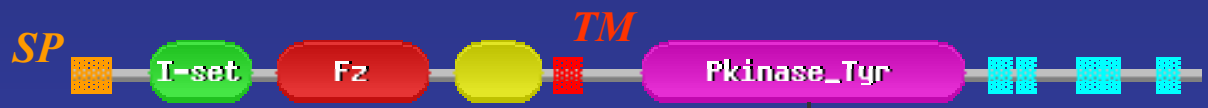
1.eszköz. Konfliktus a fehérje prediktált szubcelluláris lokalizációja és a megfelelő szekvencia jelek hiánya között.

Elméleti háttér: ***azok a fehérjék, melyek olyan doméneket tartalmaznak, melyek kizárólag az extracelluláris térben*** ***(pl. szekretált fehérjékben vagy különböző transzmembrán fehérjék extracitoplazmatikus régióiban)*** ***fordulnak elő hasítható szignál peptidet és vagy transzmembrán régiót tartalmaznak.***

killer cell lectin-like receptor

complement factor masp-3

leukocyte activation antigen m6



receptor tyrosine kinase-like orphan receptor 2

latrophilin-2

Ennek megfelelően, ***azok a fehérjék, melyek obligát extracelluláris domént tartalmaznak, de nincs szignál peptidjük és transzmembrán szegmentjük, abnormálisnak tekinthetők – valószínűleg téves predikciót jeleznek.***

Az 1. Eszköz bioinformatikai komponensei:

- Domén azonosítás (Pfam)

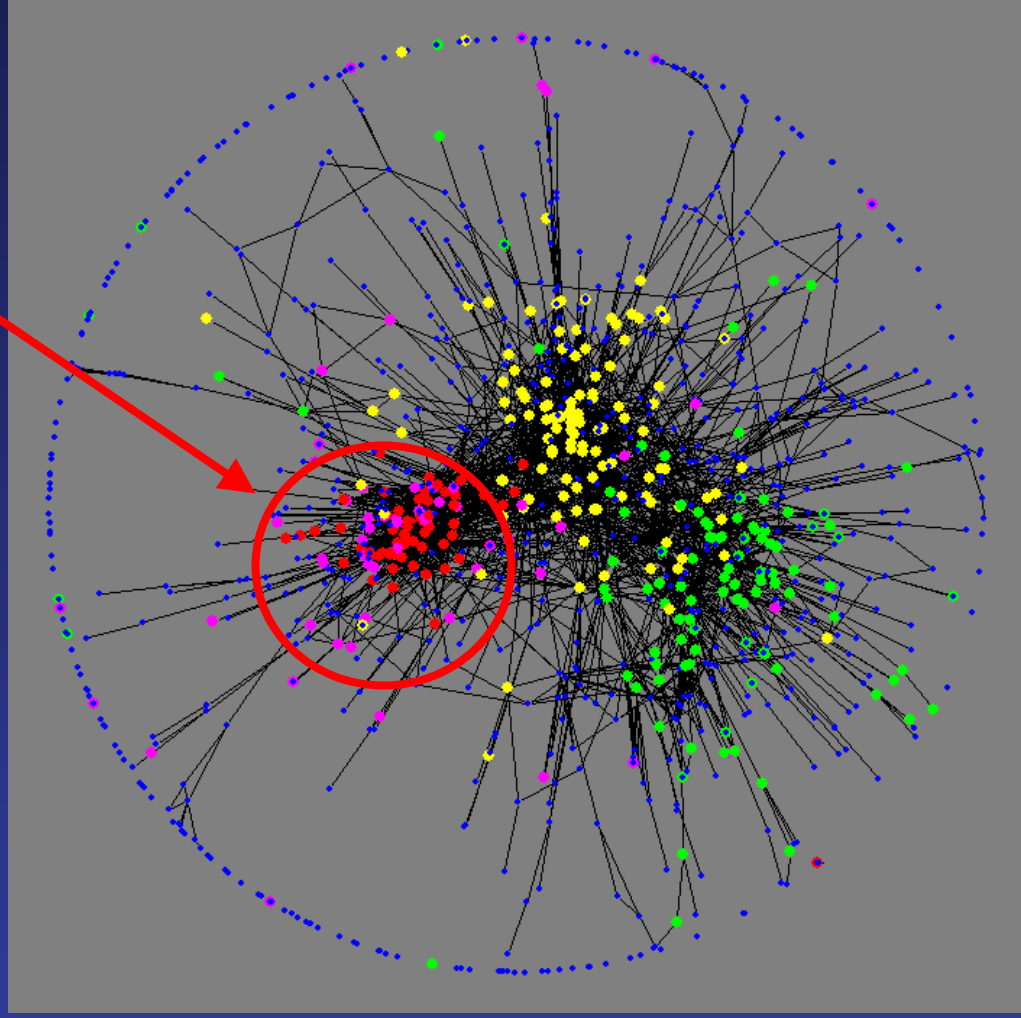
- Extracelluláris doméneket tartalmazó fehérjék azonosítása (extracelluláris Pfam A domének listája)

- Szignál peptid, szignál anchor és transzmembrán szegmentek azonosítása (SignalP, Phobius, TMHMM)

Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Bányai L, Patthy L. Identification and correction of abnormal, incomplete and mispredicted proteins in public databases. BMC Bioinformatics. 2008 Aug 27;9:353.

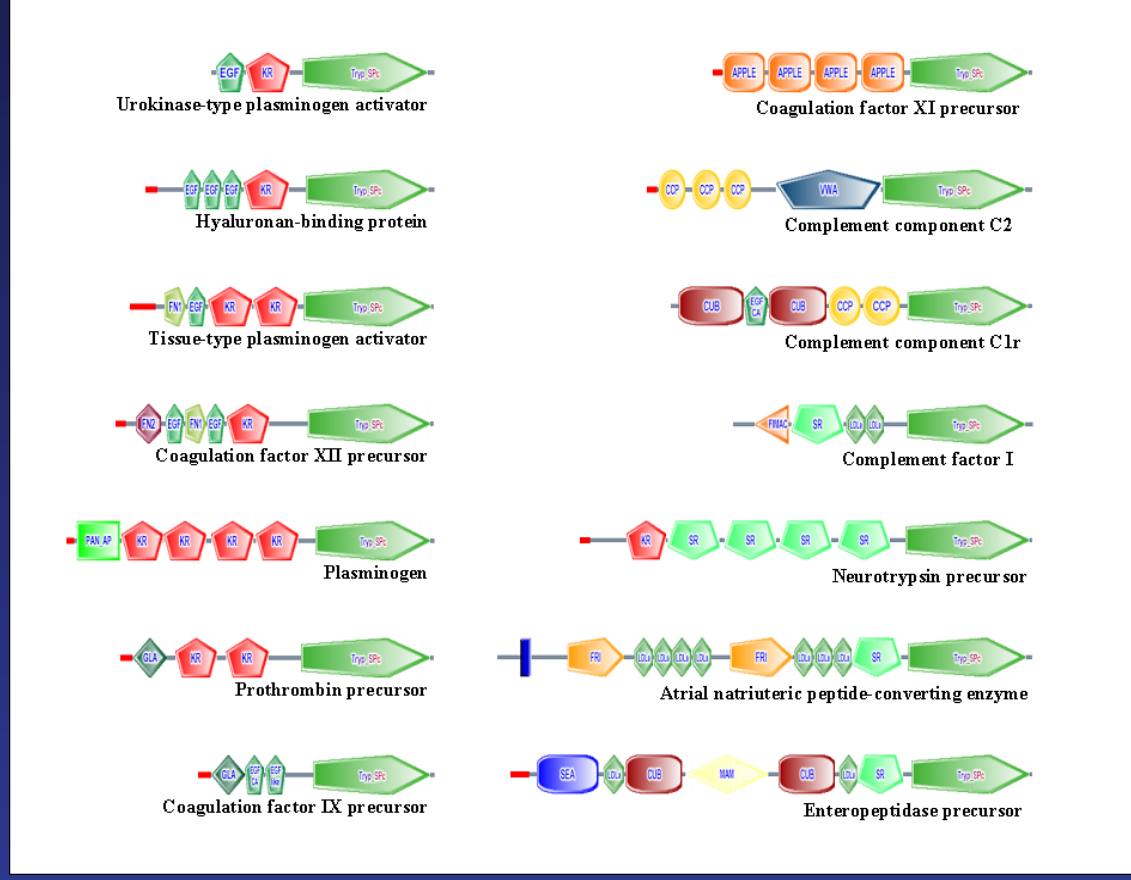
Az extracelluláris domének listáját „domain co-occurrence” alapján hálózat elemzés segítségével határoztuk meg.

Extracelluláris domének szekretált fehérjékben és különböző transzmembrán fehérjék extracitoplazmatikus régióiban találhatók



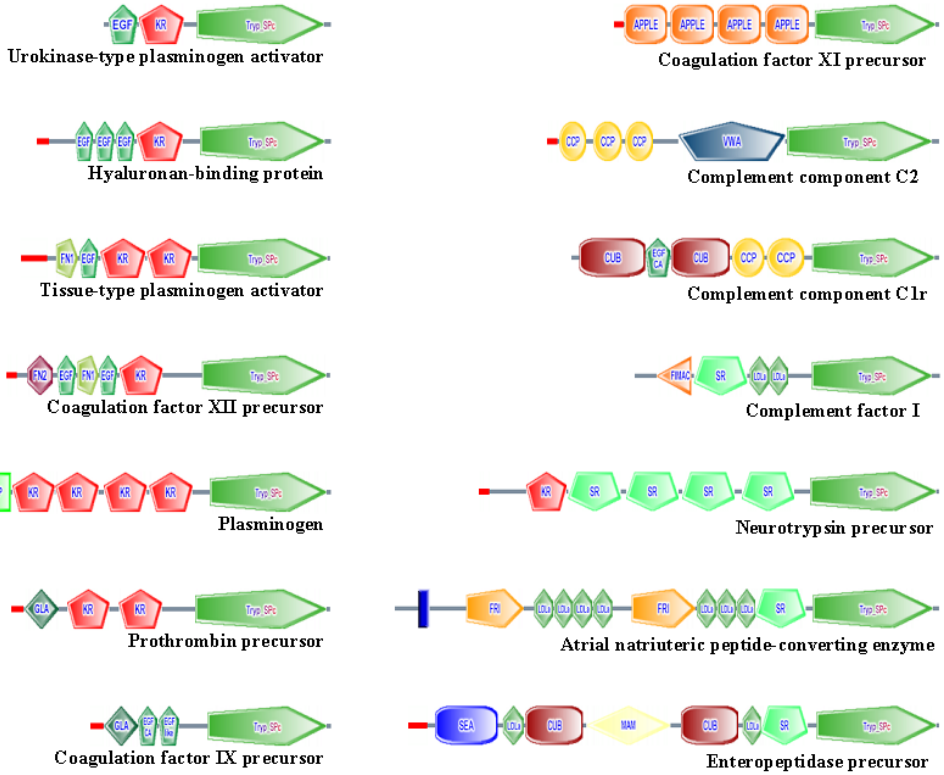
Tordai H, Nagy A, Farkas K, Bányai L, Patthy L. Modules, multidomain proteins and organismic complexity. FEBS J. 2005 Oct;272(19):5064-78.

Multidomén szerin-proteázok domén-kombinációi

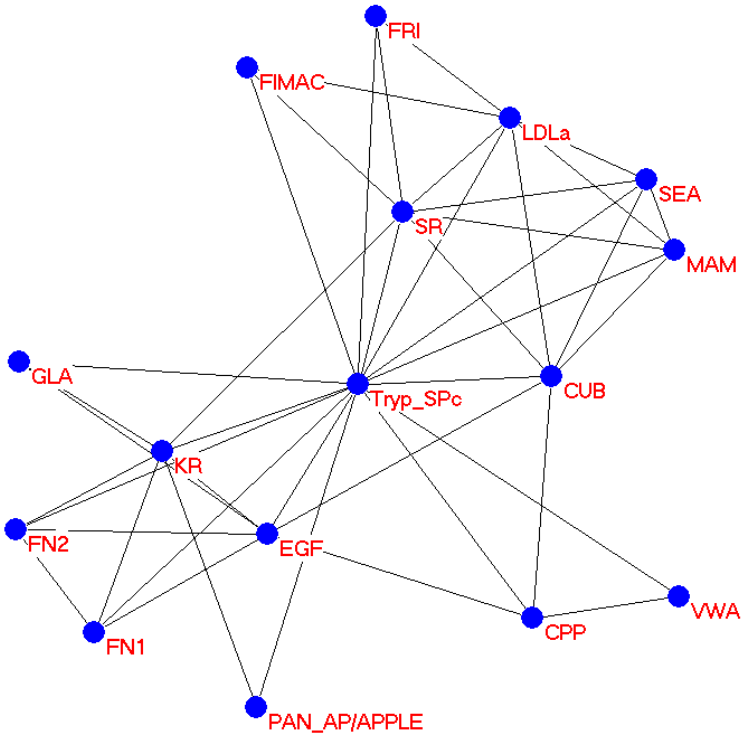


Multidomén fehérjékben a fehérjedomén-típusok véletlenszerű kombinációit többek között az korlátozza, hogy az evolúció során a különböző doméntípusok különböző szubcelluláris kompartmentekre specializálódtak.

Domén-kombinációs hálózatok vizsgálata alapján következtethetünk a domének kompartment-preferenciájára.

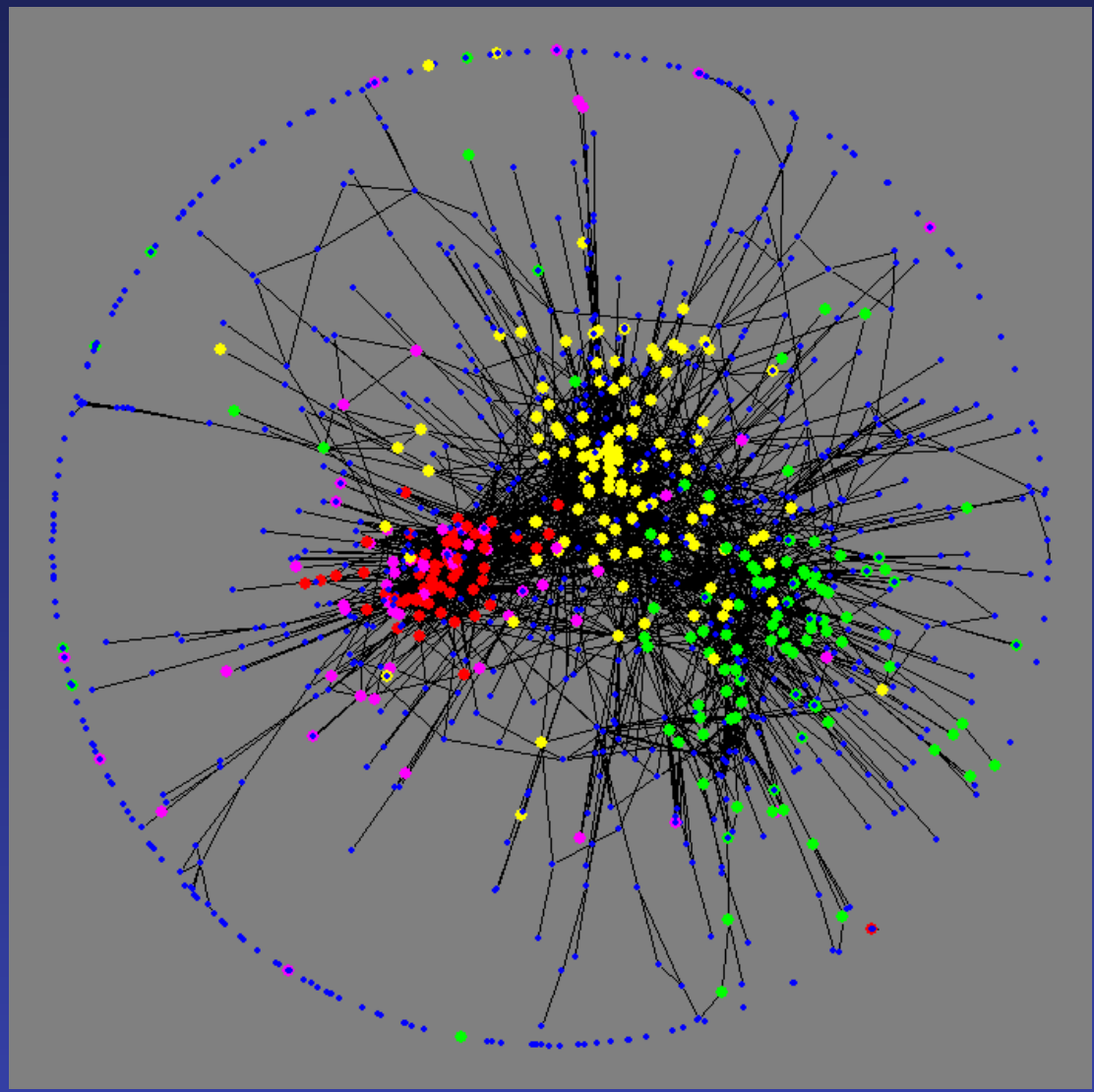


Multidomén szerin-proteázok domén-kombinációinak hálózat elemzése



Metazoa multidomén fehérjék domén összetételének hálózat elemzése

- *Extracelluláris domén*
- *Citoplazmatikus domén*
- *Nukleáris domén*



Tordai H, Nagy A, Farkas K, Bányai L, Patthy L. Modules, multidomain proteins and organismic complexity. FEBS J. 2005 Oct;272(19):5064-78.

Information

- About MisPred
- Statistics
- Gencode info
- Publications
- Useful links
- Contacts

MisPred Database

- Search MisPred
- Analyze Your Sequence

Statistics

Table 1. List of extracellular Pfam-A domain families

Description	Smart name	Pfam name	Pfam
Amyloid A4	A4_EXTRA	A4_EXTRA	PF02
Plant lipid transfer protein / seed storage protein / trypsin-alpha amylase inhibitor domain family	AAI	Tryp_alpha_amyl	PF00
Serum albumin	ALBUMIN	Serum_albumin	PF00
Ami_2	Ami_2	Amidase_2	PF01
Ami_3	Ami_3	Amidase_3	PF01
Adhesion-associated domain present in MUC4 and other proteins	AMOP	AMOP	PF01
APPLE domain	APPLE	PAN	PF00
Bulb-type mannose-specific lectin	B_lectin	B_lectin	PF01
Bacterial Ig-like domain (group 1)	BID_1	Big_1	PF02
Bacterial Ig-like domain 2	BID_2	Big_2	PF02
Bacterial OsmY and nodulation domain	BON	BON	PF04
Bowman-Birk type proteinase inhibitor	BowB	Bowman-Birk_leg	PF00
BPI/LBP/CETP N-terminal domain	BPI1	LBP_BPI_CETP	PF01
Complement component C1q domain	C1Q	C1q	PF00
Netrin C-terminal Domain	C345C	NTR	PF01
C-terminal tandem repeated domain in type 4 procollagens	C4	C4	PF01
Cadherin repeats	CA	Cadherin	PF00
Calcitonin	CALCITONIN	Calc_CGRP_IAPP	PF00
Cellulose Binding Domain Type IV	CBD_IV	CBM_6	PF03
Domain abundant in complement control proteins; SUSHI repeat; short complement-like repeat (SCR)	CCP	Sushi	PF00
CFEM	CFEM	CFEM	PF05
A domain in the BMP inhibitor chordin and in microbial proteins	CHRD	CHRD	PF07
Chitin binding domain	ChtBD1	Chitin_bind_1	PF00
Chitin-binding domain type 2	ChtBD2	CBM_14	PF01
Chitin-binding domain type 3	ChtBD3	CBM_5_12	PF02

LPLC4_HUMAN, P59827, Long palate, lung and nasal epithelium carcinoma-associated protein 4 [575 residues]

Domain	Start	End
LBP_BPI_CETP	156	303
LBP_BPI_CETP_C	416	575

SP

```
                1                                                   50
lplc4_mouse     MWTAWCVAAL SVAAVCGIRQ DTTTVLRVTK DVLGNAISGT IQKSDAFRSA
lplc4_human     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~M LQQSDALHSA

               51                                                  100
lplc4_mouse     LREVPVGVGG VPYNDFHVRE PPPKYTNGRQ LGGNYKYGHI KANDNRAQLG
lplc4_human     LREVPLGVGD IPYNDFHVRG PPPVYTNGKK LDGIYQYGHI ETNDNTAQLG

               101                                                 150
lplc4_mouse     GKYRYGEILD SDGSLRDLRH EDYRPPDSAY ..HRGSGRYR SAADSSSVGR
lplc4_human     GKYRYGEILE SEGSIRDLRN SGYRSAENAY GGHRGLGRYR AA....PVGR
```

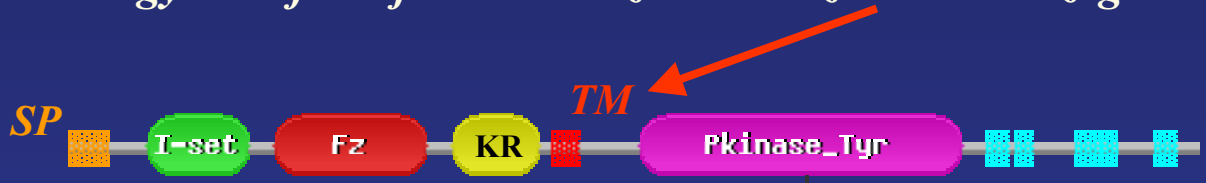
A hibás szekvenciát korrigálni lehet: a humán genom célzott vizsgálatával, a korrekt egér ortológ segítségével azonosítani lehetett a hiányzó szignál peptidet tartalmazó exont is

```
                          1                                                  50
lplc4_human_corrected     MWMAWCVAAL SVVAVCGTSH ETNTVLRVTK DVLSNAISGM LQQSDALHSA
          lplc4_human     ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~M LQQSDALHSA
          lplc4_mouse     MWTAWCVAAL SVAAVCGIRQ DTTTVLRVTK DVLGNAISGT IQKSDAFRSA

                         51                                                 100
lplc4_human_corrected     LREVPLGVGD IPYNDFHVRG PPPVYTNGKK LDGIYQYGHI ETNDNTAQLG
          lplc4_human     LREVPLGVGD IPYNDFHVRG PPPVYTNGKK LDGIYQYGHI ETNDNTAQLG
          lplc4_mouse     LREVPVGVGG VPYNDFHVRE PPPKYTNGRQ LGGNYKYGHI KANDNRAQLG

                        101                                                 150
lplc4_human_corrected     GKYRYGEILE SEGSIRDLRN SGYRSAENAY GGHRGLGRYR AA....PVGR
          lplc4_human     GKYRYGEILE SEGSIRDLRN SGYRSAENAY GGHRGLGRYR AA....PVGR
          lplc4_mouse     GKYRYGEILD SDGSLRDLRH EDYRPPDSAY ..HRGSGRYR SAADSSSVGR

                        151                                                 200
lplc4_human_corrected     LHRRELQPGE IPPGVATGAV GPGGLLGTGG MLAADGILAG QGGLLGGGGL
          lplc4_human     LHRRELQPGE IPPGVATGAV GPGGLLGTGG MLAADGILAG QGGLLGGGGL
          lplc4_mouse     LYRRELRPGE IPAGVATGAL GPGGLLGTGG MLANEGILAG QGGLLGGGGL

                        201                                                 250
lplc4_human_corrected     LGDGGLLGGG GVLGVLGEGG ILSTVQGITG LRIVELTLPR VSVRLLPGVG
          lplc4_human     LGDGGLLGGG GVLGVLGEGG ILSTVQGITG LRIVELTLPR VSVRLLPGVG
          lplc4_mouse     LGDGGLLGGG GVLGVLGEGG ILSTVQGITG LRIVELTLPR VSVRLLPGVG
```

Funded by | BIOSAPIENS NETWORK | Coordinated by EBI

2. Eszköz. Konfliktus extracelluláris ÉS citoplazmatikus domének jelenléte és transzmembrán szegment hiánya között.

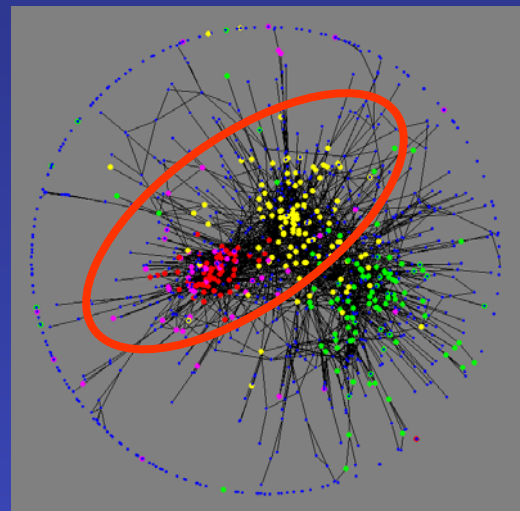
Elméleti háttér: extracelluláris és citoplazmatikus domének csak akkor fordulhatnak elő egyetlen fehérjében ha köztük transzmembrán szegment található.



receptor tyrosine kinase-like orphan receptor 2

Ennek megfelelően, azok a fehérjék, melyek mind extracelluláris, mind citoplazmikus doménT tartalmaznak, de nincs köztük transzmembrán szegment rendellenesnek és/vagy tévesen prediktáltnak minősíttetnek.

Domain co-occurrence network of metazoan multidomain proteins



Tordai et al., FEBS J. 2005; 272(19):5064-78.

A 2. Eszköz bioinformatikai komponensei:

- Domén azonosítás (Pfam)

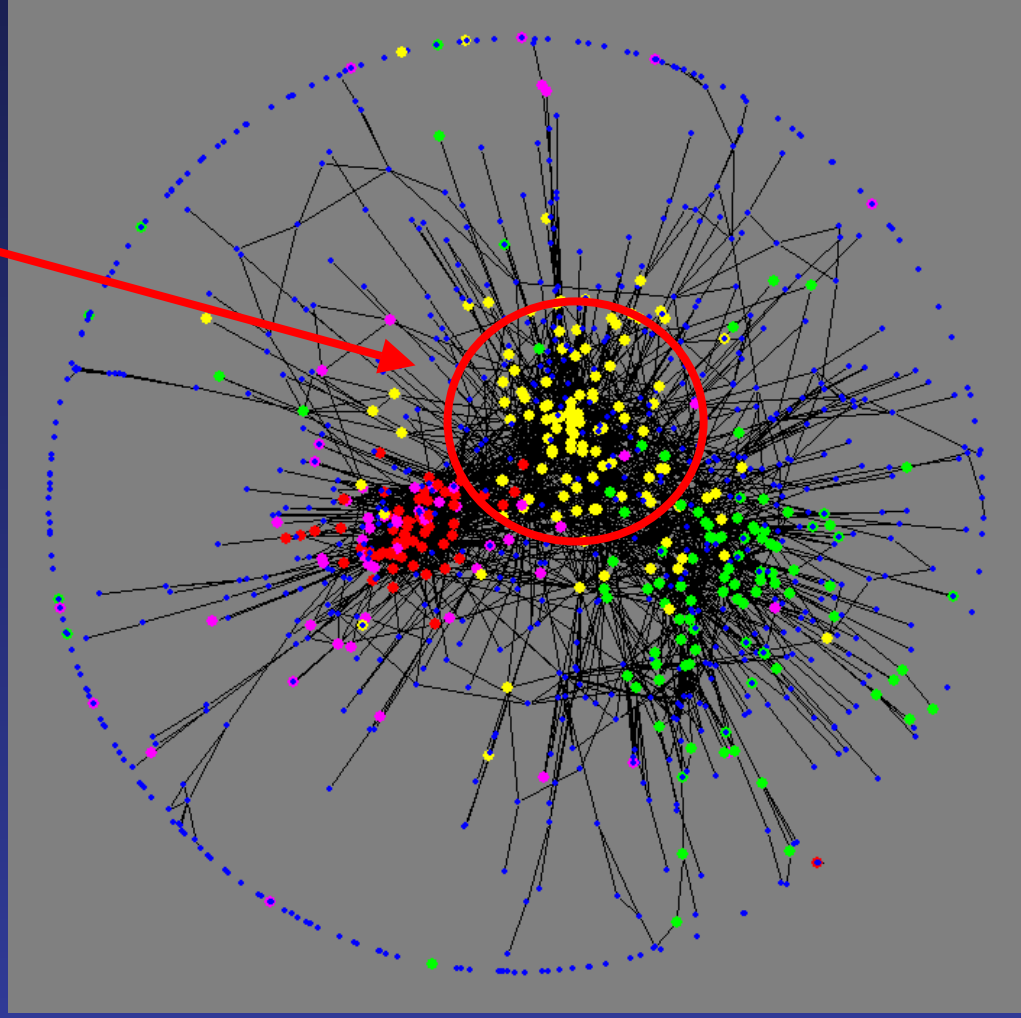
- Extracelluláris és citoplazmatikus doméneket egyaránt tartalmazó fehérjék azonosítása (extracelluláris és citoplazmatikus domének listája)

- Transzmembrán szegmentek azonosítása (TMHMM)

Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Bányai L, Patthy L. Identification and correction of abnormal, incomplete and mispredicted proteins in public databases. BMC Bioinformatics. 2008 Aug 27;9:353.

Az citoplazmatikus domének lidstáját„domain co-occurrence” alapján hálózat elemzés segítségével határoztuk meg.

Citoplazmatikus domének intracelluláris fehérjékben vagy transzmembrán fehérjék citoplazmatikus oldalán találhatók



Tordai H, Nagy A, Farkas K, Bányai L, Patthy L. Modules, multidomain proteins and organismic complexity. FEBS J. 2005 Oct;272(19):5064-78.

Table 2. List of intracellular Pfam-A signaling domain families

Description	Smart name	Pfam name	Pfam ID
14-3-3 homologues	14_3_3	14-3-3	PF00244
Actin depolymerisation factor/cofilin -like domains	ADF	Cofilin_ADF	PF00241
Ankyrin repeats	ANK	Ank	PF00023
ARF-like small GTPases; ARF, ADP-ribosylation factor	ARF	Arf	PF00025
Putative GTP-ase activating proteins for the small GTPase, ARF	ArfGap	ArfGap	PF01412
Band 4.1 homologues	B41	Band_41	PF00373
Cytochrome b-561 / ferric reductase transmembrane domain	B561	Cytochrom_B561	PF03188
Bulb-type mannose-specific lectin	B_lectin	B_lectin	PF01453
BAG domains, present in regulator of Hsp70 proteins	BAG	BAG	PF02179
BH4 Bcl-2 homology region 4	BH4	BH4	PF02180
Bruton's tyrosine kinase Cys-rich motif	BTK	BTK	PF00779
Protein kinase C conserved region 1 (C1) domains (Cysteine-rich domains)	DAG_PE-bind	C1_1	PF00130
Calponin homology domain	CH	CH	PF00307
Two component signalling adaptor domain	CheW	CheW	PF01584
Domain found in NIK1-like kinases, mouse citron and yeast ROM1, ROM2	CNH	CNH	PF00780
Cyclic nucleotide-monophosphate binding domain	cNMP	cNMP_binding	PF00027
Cullin	CULLIN	Cullin	PF00888
Adenylyl- / guanylyl cyclase, catalytic domain	CYCc	Guanylate_cyc	PF00211
Diacylglycerol kinase accessory domain (presumed)	DAGKa	DAGK_acc	PF00609
Diacylglycerol kinase catalytic domain (presumed)	DAGKc	DAGK_cat	PF00781
Domain present in Dishevelled and axin	DAX	DIX	PF00778
Death effector domain	DED	DED	PF01335
Domain found in Dishevelled, Egl-10, and Pleckstrin	DEP	DEP	PF00610
Dual specificity phosphatase, catalytic domain	DSPc	DSPc	PF00782
Domain of Unknown Function with GGDEF motif	DUF1	GGDEF	PF00990
Domain of Unknown Function 2	DUF2	EAL	PF00563
Dynamin, GTPase	DYNc	Dynamin_N	PF00350
Epsin N-terminal homology (ENTH) domain	ENTH	ENTH	PF01417
A Receptor for Ubiquitination Targets	FBOX	F-box	PF00646
Fes/CIP4 homology domain	FCH	FCH	PF00611
Contains two conserved F residues	FF	FF	PF01846
Formin Homology 2 Domain	FH2	FH2	PF02181
Forkhead associated domain	FHA	FHA	PF00498

ENSXETP00000040601 (Xenopus tropicalis) ***hibás mivel nincs benne transzmembrán szegment, jóllehet mind extracelluláris mind citoplazmatikus domént tartalmaz.***

Trusted matches - domains scoring higher than the gathering threshold (A)

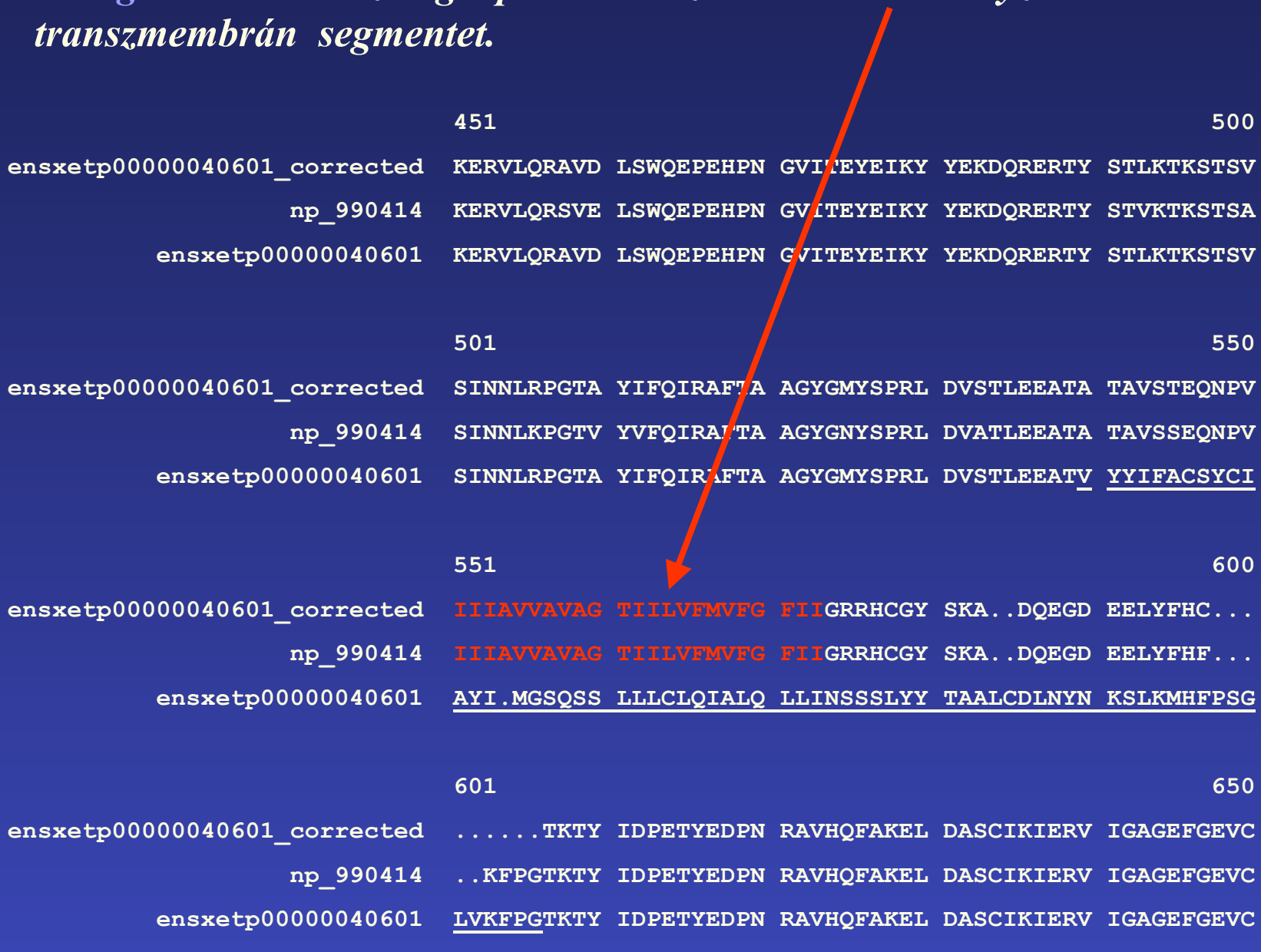
Domain	Start	End	Bits	Evalue	Alignment	Mode
fn3	152	247	43.00	8.8e-10	Align	ls
fn3	263	347	73.50	5.7e-19	Align	ls
Pkinase	454	712	151.80	1.6e-42	Align	ls
Pkinase_Tyr	454	712	503.60	1.9e-148	Align	ls
SAM_2	742	809	95.30	1.6e-25	Align	ls
SAM_1	743	807	107.30	3.7e-29	Align	ls
Ephrin_lbd	1	25	63.20	2.4e-18	Align	fs

Az ENSXETP00000040601 (Xenopus tropicalis) csirke ortológja, az EPHA7_CHICK Ephrin type-A receptor 7 (np_990414), tartalmaz a megfelelő helyen transzmembrán szegmentet.

Az ENSXETP00000040601 (Xenopus tropicalis) éppen ebben a régióban tér el legjelentősebben a EPHA7_CHICK-től

```
Query  181  SDVTYRVVCKRCSWEQGECIPCANTIGYVPQQSGLVDTYISIVDLVAHANYTFEVEAVNG  240
            +DVTYR++CKRCSWEQGEC+PC + IGY+PQQ+GLVD Y++++DL+AHANYTFEVEAVNG
Sbjct  361  NDVTYRILCKRCSWEQGECVPCGSNIGYMPQQTGLVDNYVTVMDLLAHANYTFEVEAVNG  420

Query  241  VSDLSRSQRLFAAVSVTTGQAAPSQVSGVMKERVLQRAVDLSWQEPEHPNGVITEYEIKY  300
            VSDLSRSQRLFAAVS+TTGQAAPSQVSGVMKERVLQR+V+LSWQEPEHPNGVITEYEIKY
Sbjct  421  VSDLSRSQRLFAAVSITTGQAAPSQVSGVMKERVLQRSVELSWQEPEHPNGVITEYEIKY  480

Query  301  YEKDQRERTYSTLKTKSTSVSINNLRPGTAYIFQIRAFTAAGYGMYSPRLDVSTLEEATV  360
            YEKDQRERTYST+KTKSTS SINNL+PGT Y+FQIRAFTAAGYG YSPRLDV+TLEEAT
Sbjct  481  YEKDQRERTYSTVKTKSTSASINNLKPGTVYVFQIRAFTAAGYGNYSPRLDVATLEEATA  540

Query  361  YYIFA-CSYCIAYIMGSCSSLLLCLQIALQLLINSSSLYYTAALCDLCYNKSLKMHFPSG  419
              + +  +  I   + + +  ++ + +    +I      Y+ A  D   ++ L  HF
Sbjct  541  TAVSSEQNPVIIAVVAVAGTIILVFMVFGFIIGRRHCGYSKA--DQEGDEELYFHF---  595

Query  420  LVKFPGTKTYIDPETYEDPNRAVHQFAKELDASCIKIERVIGAGEFGEVCSGRLKLPGKR  479
              KFPGTKTYIDPETYEDPNRAVHQFAKELDASCIKIERVIGAGEFGEVCSGRLKLPGKR
Sbjct  596  --KFPGTKTYIDPETYEDPNRAVHQFAKELDASCIKIERVIGAGEFGEVCSGRLKLPGKR  653

Query  480  DVPVAIKTLKVGYTEKQRRDFLCEASIMGQFDHPNVVHLEGVVTRGKPVMIVIEFMENGA  539
            DV VAIKTLKVGYTEKQRRDFLCEASIMGQFDHPNVVHLEGVVTRGKPVMIVIE+MENGA
Sbjct  654  DVAVAIKTLKVGYTEKQRRDFLCEASIMGQFDHPNVVHLEGVVTRGKPVMIVIEYMENGA  713

Query  540  LDAFLRKLDGQFTVIQLVGMLRGIAAGMRYLADMGYVHRDLAARNILVNSNLVCKVSDFG  599
            LDAFLRK DGQFTVIQLVGMLRGIAAGMRYLADMGYVHRDLAARNILVNSNLVCKVSDFG
Sbjct  714  LDAFLRKHDGQFTVIQLVGMLRGIAAGMRYLADMGYVHRDLAARNILVNSNLVCKVSDFG  773

Query  600  LSRIIEDDPDAVYTTTQGGKIPVRWTAPEAIQYRKFTSASDVWSYGIVMWEVMSYGERPY  659
```

Az ENSXETP00000040601 (Xenopus tropicalis) hibás részét korrigálni lehet: ***célzott génpredikció azonosította a 'hiányzó' transzmembrán segmentet.***



4. Eszköz. Domén méret eltérés

Elméleti háttér: *egy adott globuláris domén-család esetén a domén méret (a doménhatárokon belüli aninosavak száma) szűk határokon belül mozog, a domén-család különböző tagjainak mérete kevéssé tér el a családra jellemző átlagmérettől.*

A jelenség magyarázata az, hogy hosszabb szakaszok inszerciója vagy delációja nagy valószínűséggel olyan fehérjét eredményez, mely nem képes hatékonyan felvenni, a natív, életképes és stabil térszerkezetet.

Ennek megfelelően*, azok a fehérjék, melyek olyan domént tartalmaznak, melynek mérete jelentősen eltér a család többi tagjának méretétől, rendellenesnek és/vagy tévesen prediktáltnak minősíttetnek.*

Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Bányai L, Patthy L. Identification and correction of abnormal, incomplete and mispredicted proteins in public databases. BMC Bioinformatics. 2008 Aug 27;9:353.

A 4. Eszköz komponensei:

- Domén azonosítás és domén méret meghatározás (Pfam)

- A normális domén mérettől jelentősen eltérő méretű doméneket tartalmazó fehérjék azonosítása

Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Bányai L, Patthy L. Identification and correction of abnormal, incomplete and mispredicted proteins in public databases. BMC Bioinformatics. 2008 Aug 27;9:353.

RP11-247A12.5-001 [544 aa]

ABNORMAL

CACP_HUMAN, Carnitine O-acetyltransferase [626 residues]

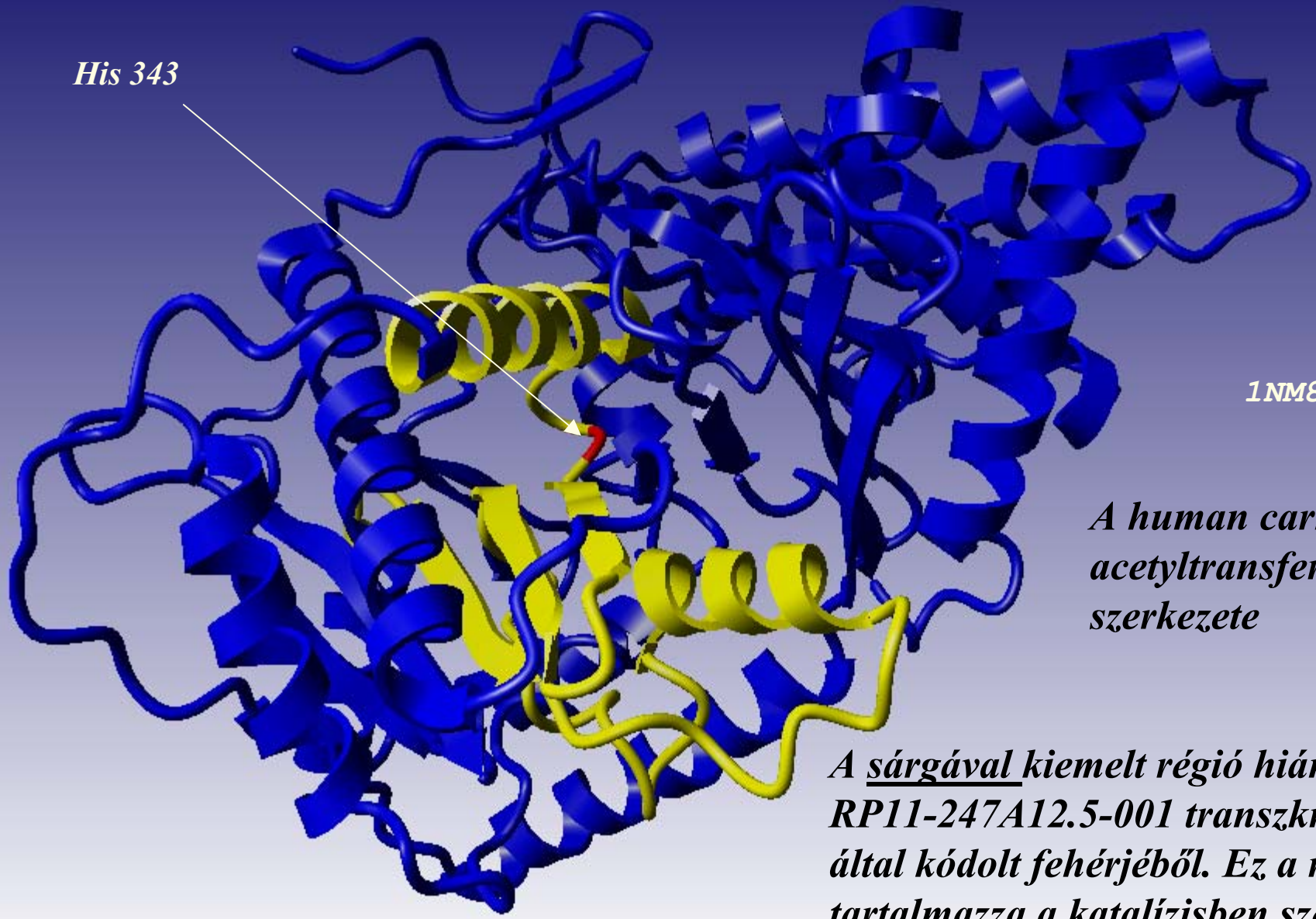
CORRECT

RP11-247A12.5-001 doménen belüli deléciót tartalmaz

Carn_acyltransf domain

Az RP11-247A12.5-001 transzkriptumból hiányzó régió

```
           1                                                                                                   100
 cacp_human  MLAFAARTVV KPLGFLKPFS LMKASSRFKA HQDALPRLPV PPLQQSLDHY LKALQPIVSE EEWAHTKQLV DEFQASGGVG ERLQKGLERR ARKTENWLSE
rp11-247a12  MLAFAARTVV KPLGFLKPFS LMKASSRFKA HQDALPRLPV PPLQQSLDHY LKALQPIVSE EEWAHTKQLV DEFQASGGVG ERLQKGLERR ARKTENWLSE

           101                                                                                                 200
 cacp_human  WWLKTAYLQY RQPVVIYSSP GVMLPKQDFV DLQGQLRFAA KLIEGVLDFK VMIDNETLPV EYLGGKPLCM NQYYQILSSC RVPGPKQDTV SNFSKTKKPP
rp11-247a12  WWLKTAYLQY RQPVVIYSSP GVMLPKQDFV DLQGQLRFAA KLIEGVLDFK VMIDNETLPV EYLGGKPLCM NQYYQILSSC RVPGPKQDTV SNFSKTKKPP

           201                                                                                                 300
 cacp_human  THITVVHNYQ FFELDVYHSD GTPLTADQIF VQLEKIWNSS LQTNKEPVGI LTSNHRNSWA KAYNTLIKDK VNRDSVRSIQ KSIFTVCLDA TMPRVSEDVY
rp11-247a12  THITVVHNYQ FFELDVYHSD GTPLTADQIF VQLEKIWNSS LQTNKEPVGI LTSNHRNSWA KAYNTLIKDK VNRDSVRSIQ .......... ..........

           301                                                                                                 400
 cacp_human  RSHVAGQMLH GGGSRLNSGN RWFDKTLQFI VAEDGSCGLV YEHAAAEGFP IVTLLDYVIE YTKKPELVRS PMVPLPMPKK LRFNITPEIK SDIEKAKQNL
rp11-247a12  .......... .......... .......... .......... .......... .......... ..KKPELVRS PLVPLPMPKK LRFNITPEIK SDIEKAKQNL

           401                                                                                                 500
 cacp_human  SIMIQDLDIT VMVFHHFGKD FPKSEKLSPD AFIQMALQLA YYRIYGQACA TYESASLRMF HLGRTDTIRS ASMDSLTFVK AMDDSSVTEH QKVELLRKAV
rp11-247a12  SIMIQDLDIT VMVFHHFGKD FPKSEKLSPD AFIQMALQLA YYRIYGQACA TYESASLRMF HLGRTDTIRS ASMDSLTFVK AMDDSSVTEH QKVELLRKAV

           501                                                                                                 600
 cacp_human  QAHRGYTDRA IRGEAFDRHL LGLKLQAIED LVSMPDIFMD TSYAIAMHFH LSTSQVPAKT DCVMFFGPVV PDGYGVCYNP MEAHINFSLS AYNSCAETNA
rp11-247a12  QAHRGYTDRA IRGEAFDRHL LGLKLQAIED LVSMPDIFMD TSYAIAMHFH LSTSQVPAKT DCVMFFGPVV PDGYGVCYNP MEAHINFSLS AYNSCAETNA

           601                    626
 cacp_human  ARLAHYLEKA LLDMRALLQS HPRAKL
rp11-247a12  ARLAHYLEKA LLDMRALLQS HPRAKL
```



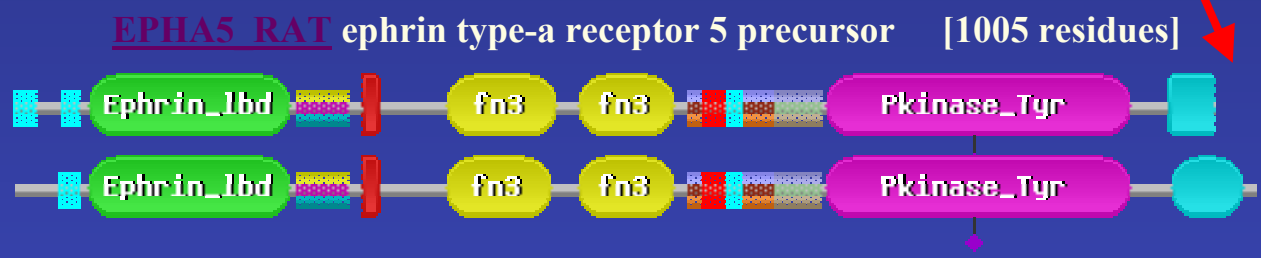
1NM8.pdb

A human carnitine O-acetyltransferase 3D szerkezete

A sárgával kiemelt régió hiányzik az RP11-247A12.5-001 transzkriptum által kódolt fehérjéből. Ez a régió tartalmazza a katalízisben szerepet játszó His-343-t

```
                901                                                 950
epha5_human     PSPMDCPAAL YQLMLDCWQK ERNSRPKFDE IVNMLDKLIR NPSSLKTLVN
  epha5_rat     PSPMDCPAAL YQLMLDCWQK DRNSRPKFDD IVNMLDKLIR NPSSLKTLVN
epha5_chick     PSPMDCPAAL YQLMLDCWQK DRNSRPKFDE IVSMLDKLIR NPSSLKTLVN
epha5_mouse     PSPMDCPAAL YQLMLDCWQK DRNSRPKFDE IVNMLDKLIR NPSSLKTLVN

                951                                                1000
epha5_human     ASCRVSNLLA EHSPLGSGAY RSVGEWLEAI KMGRYTEIFM ENGYSSMDAV
  epha5_rat     ASSRVSTLLA EHGSLGSGAY RSVGEWLEAT KMGRYTEIFM ENGYSSMDAV
epha5_chick     ASSRVSNLLV EHSPVGSGAY RSVGEWLEAI KMGRYTEIFM ENGYSSMDSV
epha5_mouse     ASSRVSTLLA EHGSLGSGAY RSVGEWLEAI KMGRYTEIFM ENGYSSMDAV

                1001                                    1041
epha5_human     AQVTLEDLRR LGVTLVGHQ. KKIMNSLQEM KVQLVNGMVP L
  epha5_rat     AQVTLE.... .......... .......... .......... .
epha5_chick     AQVTLEDLRR LGVTLVGHQ. KKIMNSLQEM KVQLVNGMVP L
epha5_mouse     AQVTLEDLRR LGVTLVGHQK KKIMSSLQEM KVQMVNGMVP V
```

EPHA5_RAT egy C-terminálisán csonkolt SAM_1 domént tartalmaz. A fehérje egér, csirke és humán ortológjai intakt SAM_1 domént tartalmaznak.



```
                        901                                                950
epha5_rat_corrected   PSPMDCPAAL YQLMLDCWQK DRNSRPKFDD IVNMLDKLIR NPSSLKTLVN
          epha5_rat   PSPMDCPAAL YQLMLDCWQK DRNSRPKFDD IVNMLDKLIR NPSSLKTLVN
        epha5_human   PSPMDCPAAL YQLMLDCWQK ERNSRPKFDE IVNMLDKLIR NPSSLKTLVN
        epha5_chick   PSPMDCPAAL YQLMLDCWQK DRNSRPKFDE IVSMLDKLIR NPSSLKTLVN
        epha5_mouse   PSPMDCPAAL YQLMLDCWQK DRNSRPKFDE IVNMLDKLIR NPSSLKTLVN

                        951                                               1000
epha5_rat_corrected   ASSRVSTLLA EHGSLGSGAY RSVGEWLEAT KMGRYTEIFM ENGYSSMDAV
          epha5_rat   ASSRVSTLLA EHGSLGSGAY RSVGEWLEAT KMGRYTEIFM ENGYSSMDAV
        epha5_human   ASCRVSNLLA EHSPLGSGAY RSVGEWLEAI KMGRYTEIFM ENGYSSMDAV
        epha5_chick   ASSRVSNLLV EHSPVGSGAY RSVGEWLEAI KMGRYTEIFM ENGYSSMDSV
        epha5_mouse   ASSRVSTLLA EHGSLGSGAY RSVGEWLEAI KMGRYTEIFM ENGYSSMDAV

                       1001                                          1042
epha5_rat_corrected   AQVTLEDLRR LGVTLVGHQ. KKIMNSLQEM KVQLVNGMVP V
          epha5_rat   AQVTLE~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~
        epha5_human   AQVTLEDLRR LGVTLVGHQ. KKIMNSLQEM KVQLVNGMVP L
        epha5_chick   AQVTLEDLRR LGVTLVGHQ. KKIMNSLQEM KVQLVNGMVP L
        epha5_mouse   AQVTLEDLRR LGVTLVGHQK KKIMSSLQEM KVQMVNGMVP V
```

Korrigált szekvencia

Zhang et al. BMC Genomics 2012, 13:206
http://www.biomedcentral.com/1471-2164/13/206

BMC Genomics

CORRESPONDENCE Open Access

Limitations of the rhesus macaque draft genome assembly and annotation

Xiongfei Zhang, Joel Goodsell and Robert B Norgren, Jr.*

Abstract

Finished genome sequences and assemblies are available for only a few vertebrates. Thus, investigators studying many species must rely on draft genomes. Using the rhesus macaque as an example, we document the effects of sequencing errors, gaps in sequence and misassemblies on one automated gene model pipeline, Gnomon. The combination of draft genome with automated gene finding software can result in spurious sequences. We estimate that approximately 50% of the rhesus gene models are missing, incomplete or incorrect. The problems identified in this work likely apply to all draft vertebrate genomes annotated with any automated gene model pipeline and thus represent a pervasive challenge to the analysis of draft genomes.

Bányai László
Farkas Krisztina
Hegyi Hédi
Kozma Evelin
Nagy Alinda
Szarka Eszter
Szláma György
Tordai Hedvig
Trexler Mária

*A **MisPred** projekt a **„BioSapiens”** program keretében indult el. A BioSapiens programot a European Commission FP6 "Life sciences, genomics and biotechnology for health„ alprogramja támogatta (szerződés szám: LHSG-CT-2003-503265). A MisPred projektet támogatta az NKTH eScience RET14/2005 programmja is.*

*A **FixPred9** projektet a Nemzeti Innovációs Hivatal a TECH_09-A1-2009-0116 számú, **„Genom-információk hasznosítása”** című pályázat keretében támogatja.*