

Биоинформатика
(вычислительная биология)
и биология развития

Методы биоинформатики основаны на

- Последовательности (сиквенс)
- Структуре
- Контексте
- Времени/месте/уровне проявления эффекта (например, экспрессии гена)
- Взаимодействии по принципу «информационной сети»



Алгоритмы

- Определения степени сходства и поиска сходных последовательностей
- Предсказания структуры
- Предсказания функции

bioinformatics in the 90s

primary databases containing
large-scale experimental data
in genomics and proteomics



understanding functions and
utilities of individual genes
or proteins



bioinformatics now

secondary databases (knowledge
bases) containing accumulated
biological knowledge



understanding functions and
utilities at the molecular,
cellular and organism levels



bioinformatics in the future

a complete computer
representation of the cell
and the organism

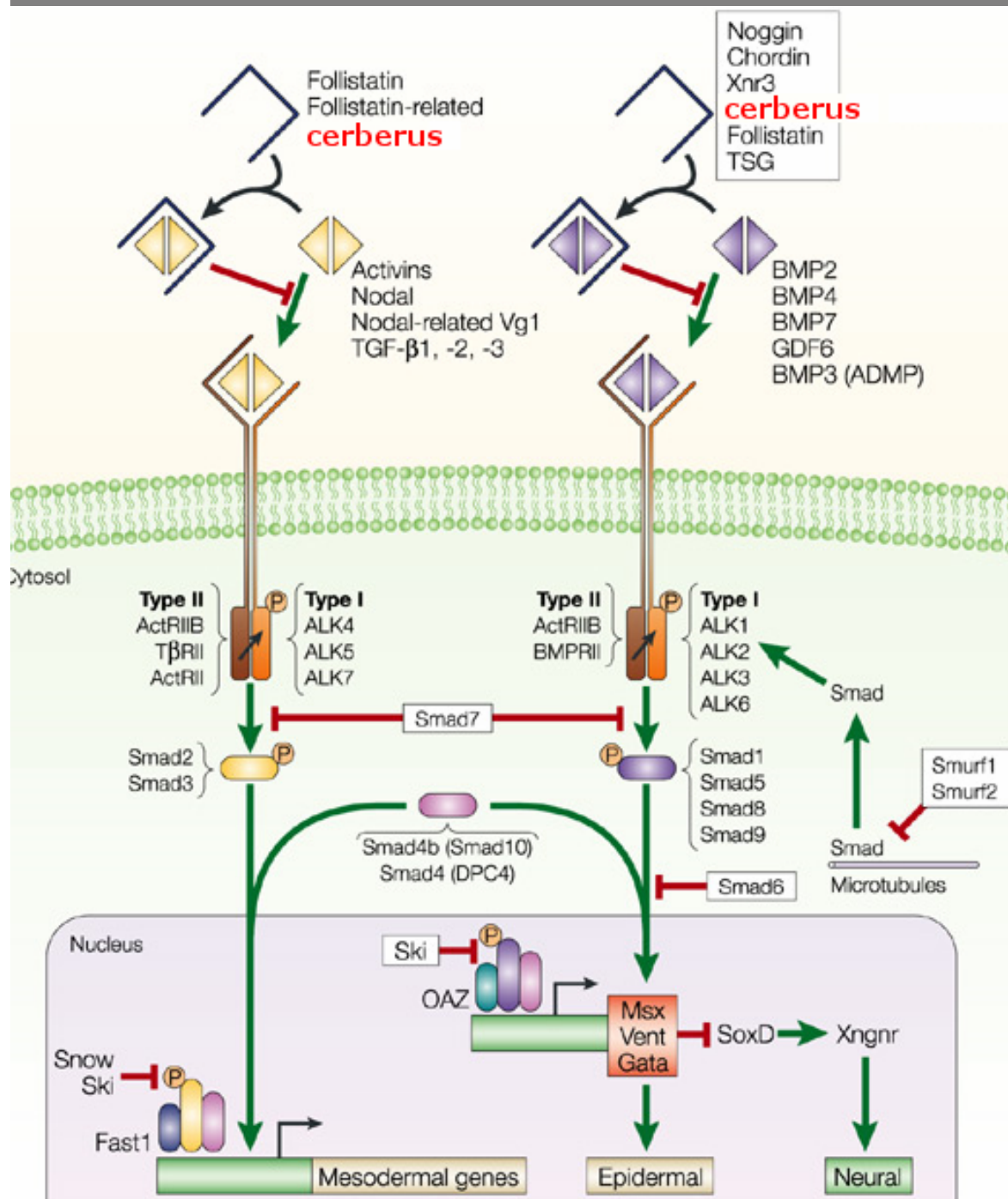


understanding basic principles
of the higher complexity of
biological systems

Быстрый поиск по базам последовательностей

- 1980-ые: FASTA — поиск оптимального выравнивания последовательностей, так, чтобы сумма совпадений или гомологичных соответствий знаков в последовательности была максимальна с точки зрения матрицы аминокислотных замен
- 1990-ые: BLAST — статистическое вычисление high-scoring segment pairs (HSPs) — локальных последовательностей, не содержащих «гэпов» (пробелов), таких, что сумма совпадений или гомологичных соответствий знаков не увеличивается при удлинении или укорачивании размера последовательности.

Cerberus – блокатор активности BMP



NCBI/BLAST/blastp suite/ Formatting Results - VJW5Y55V01R

ⓘ Your search is limited to records matching entrez query: txid9606 [ORGN].

[Edit and Resubmit](#) [Save Search Strategies](#) [▶Formatting options](#) [▶Download](#)

gi|4885135|ref|NP_005445.1| cerberus 1 [Homo...

Query ID lc|92405
Description gi|4885135|ref|NP_005445.1| cerberus 1 [Homo sapiens]
Molecule type amino acid
Query Length 267

Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.2.20+ [▶Citation](#)

Other reports: [▶Search Summary](#) [\[Taxonomy reports\]](#) [\[Distance tree of results\]](#)

▼ Graphic Summary

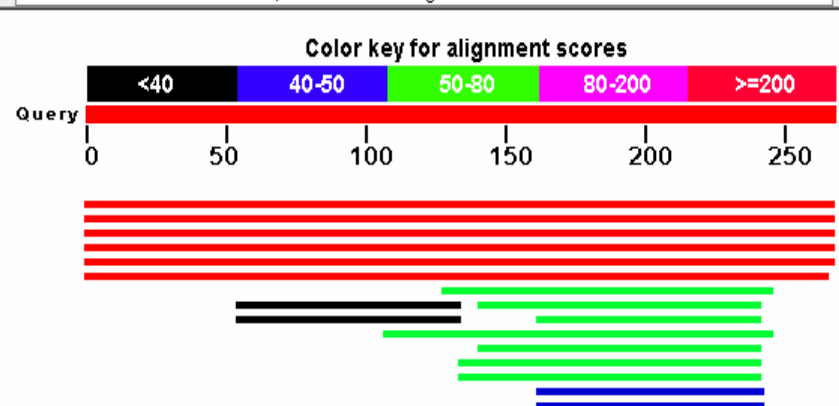
▼ [Show Conserved Domains](#)

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of 45 Blast Hits on the Query Sequence ⓘ

Mouse over to see the define, click to show alignments



Представление результатов поиска в BLAST

NCBI Blast:gi|4885135|ref|... x

http://blast.ncbi.nlm.nih.gov/Blast.cgi

▼ Descriptions

Sequences producing significant alignments:	Score (Bits)	E Value	
ref NP_005445.1 cerberus 1 [Homo sapiens] >sp 095813.1 CER1...	550	2e-156	UG
gb AAH69491.1 Cerberus 1 [Homo sapiens]	548	8e-156	GG
gb AAH69405.1 Cerberus 1 [Homo sapiens]	547	2e-155	GG
gb AAI03978.1 Cerberus 1, cysteine knot superfamily, homolog...	547	2e-155	GG
gb AAH69371.1 CER1 protein [Homo sapiens]	544	1e-154	GG
gb AAH69503.1 Cerberus 1, cysteine knot superfamily, homolog...	538	5e-153	GG
gb EAW84336.1 DAN domain family, member 5, isoform CRA_a [Ho...	70.1	8e-12	GG
dbj BAC04643.1 unnamed protein product [Homo sapiens] >gb EA...	69.7	1e-11	GG
ref NP_037504.1 gremlin-1 precursor [Homo sapiens] >sp 06056...	69.3	1e-11	UG
ref NP_689867.1 dante precursor [Homo sapiens] >sp Q8N907.1 ...	68.9	1e-11	UG
gb EAW92269.1 gremlin 1, cysteine knot superfamily, homolog ...	67.4	4e-11	GG
ref NP_071914.3 gremlin 2 precursor [Homo sapiens] >sp Q9H77...	62.4	1e-09	UG
gb AAH46632.1 Gremlin 2, cysteine knot superfamily, homolog ...	62.0	2e-09	GG
ref NP_002448.2 mucin 2 precursor [Homo sapiens]	46.6	9e-05	GG
sp Q02817.2 MUC2_HUMAN RecName: Full=Mucin-2; Short=MUC-2; Al...	46.6	9e-05	GG
ref NP_877421.2 neuroblastoma, suppression of tumorigenicity...	45.8	2e-04	UG
gb AAA59164.1 MUC2	45.4	2e-04	GG
gb EAX02421.1 mucin 2, oligomeric mucus/gel-forming, isoform...	45.1	2e-04	GG
ref NP_005371.1 neuroblastoma, suppression of tumorigenicity...	45.1	2e-04	UG
gb AAL15440.1 MBL1 [Homo sapiens] >emb CAI19019.1 neuroblas...	45.1	2e-04	GG
gb EAX02422.1 mucin 2, oligomeric mucus/gel-forming, isoform...	45.1	3e-04	GG
emb CAM45837.1 neuroblastoma, suppression of tumorigenicity ...	43.9	6e-04	GG
emb CAM45834.1 neuroblastoma, suppression of tumorigenicity ...	42.4	0.002	GG
emb CAM45836.1 neuroblastoma, suppression of tumorigenicity ...	42.4	0.002	GG
emb CAM45835.1 neuroblastoma, suppression of tumorigenicity ...	42.4	0.002	GG
emb CAM45833.1 neuroblastoma, suppression of tumorigenicity ...	40.0	0.007	GG
emb CAM45832.1 neuroblastoma, suppression of tumorigenicity ...	39.7	0.009	GG
emb CAA26503.1 unnamed protein product [Homo sapiens] >gb AA...	33.1	0.96	GG
sp Q9HC84.2 MUC5B_HUMAN RecName: Full=Mucin-5B; Short=MUC-5B;...	32.7	1.3	UG
ref XP_001719401.1 PREDICTED: similar to mucin 5, subtype B,...	32.7	1.3	UG
ref NP_002449.1 mucin 5, subtype B, tracheobronchial [Homo s...	32.7	1.3	UG
gb EAW82841.1 carboxylesterase 7, isoform CRA_a [Homo sapiens]	31.6	2.8	GG
ref NP_659461.1 carboxylesterase 7 isoform 2 [Homo sapiens] ...	31.6	2.8	UG
gb AAB59512.1 von Willebrand factor prepropeptide	31.6	3.0	GG
gb EAW88814.1 von Willebrand factor, isoform CRA_a [Homo sap...	31.2	3.3	GG

Выравнивание последовательностей

	10	20	30	40	50	60	70	80	90	
Drm	CKTQPLKQT	IHEEGCNS	RTIINRF	CYQCNS	FYIP	PR. HIRKEEGSFQS	CSFCKP	KKFTTMMVT	LNCPE. LQPPTKKKRV. . . TRVKQCR	
PRDC	CKTQPLRQT	VSEEGCRS	RTILNRF	CYQCNS	FYIP	PR. HVKKEEDSFQS	CAFCKP	QRTSVIV	LECPG. LDPPFRIKKI. . . QKVKHCR	
CeCan1	CDGQKFKQR	IRVDGL	TKVVVNRL	CHGACAS	IFIP	PRMHSKKLKAAFRS	CAACAPAEY	DYVDITL	DCPG. RTPPTATKT. . . VKVKSCK	
Cer1	CRTVPFNQT	IAHEDCQ	KVVVQNNL	CFGKCSS	IRFP	. . . GEGADAHSF	CSHCSPT	KFTT	VHMLNCT. . . SPTPVVKMV. . . MQVEECQ	
cerberus	CKTLPFTQN	IVHENC	DRMVQNNL	CFGKCI	SLHVP	 NQQDRRNT	CSHCLPS	KFTLNH	LTNCTG. . . SKNVVKVV. . . MMVEECT	
Dan	CEAKNITQI	VGHSGCE	AKSIQNR	ACLGQC	FSYSVP	P. NTFPQSTESL	VHCDSCMP	AQSMWEI	VTLECPDHEEVPRVDKL. . . EKIVHCS	
Dte	AALVRTAQV	ISRPGCT	SARVLNHL	CFGRCSS	FYIP	. . . SSDPTPVVF	CNSCV	PAKRWTSVT	LWCG. . AGQLASPRRVRI	STVLVQKCQ

Сходство двух последовательностей не обязательно свидетельствует о гомологии белков или их генов. Играет роль также сходство их положения в геноме (например, в составе кластера), сходство интрон-экзонной структуры, сходство регулирования их экспрессии и/или промоторных участков, сходство паттернов экспрессии, доменной и трехмерной структур белков, а также их функциональных характеристик.

Поиск консервативных доменов Cerberus

NCBI

HOME SEARCH SITE MAP NewSearch CDD Home PubMed Protein Structure Taxonomy Help

Conserved Domains

Conserved domains on [gi|4885135|ref|NP_005445.1|]

Local query sequence

SHOW FULL DISPLAY ?

Graphical summary [show options](#) ?

Query seq. 1 50 100 150 200 250 267

Superfamilies

Multi-domains CpaC SEC6 GHB superfamily

[Search for similar domain architectures](#) ?

List of domain hits ?

Description	Pssmid	Multi-dom	E-value
[+] cl00070, GHB, Glycoprotein hormone beta chain homologues. Gonadotropins; reproductive hormones...	119503	N/A	3e-56
[+] COG5173, SEC6, Exocyst complex subunit SEC6 [Intracellular trafficking and secretion]	34772	yes	0.63
[+] COG4964, CpaC, Flp pilus assembly protein, secretin CpaC [Intracellular trafficking and secretion]	34571	yes	2.8

List of domain hits ?

Description	Pssmid	Multi-dom	E-value
[+] cl00070, GHB, Glycoprotein hormone beta chain homologues. Gonadotropins; reproductive hormones...	119503	N/A	3e-56

Glycoprotein hormone beta chain homologues. Gonadotropins; reproductive hormones consisting of two glycosylated chains (alpha and beta) of similar topology with Cysteine-knot motifs. The best-scoring hit on this query sequence is by member pfam03045:

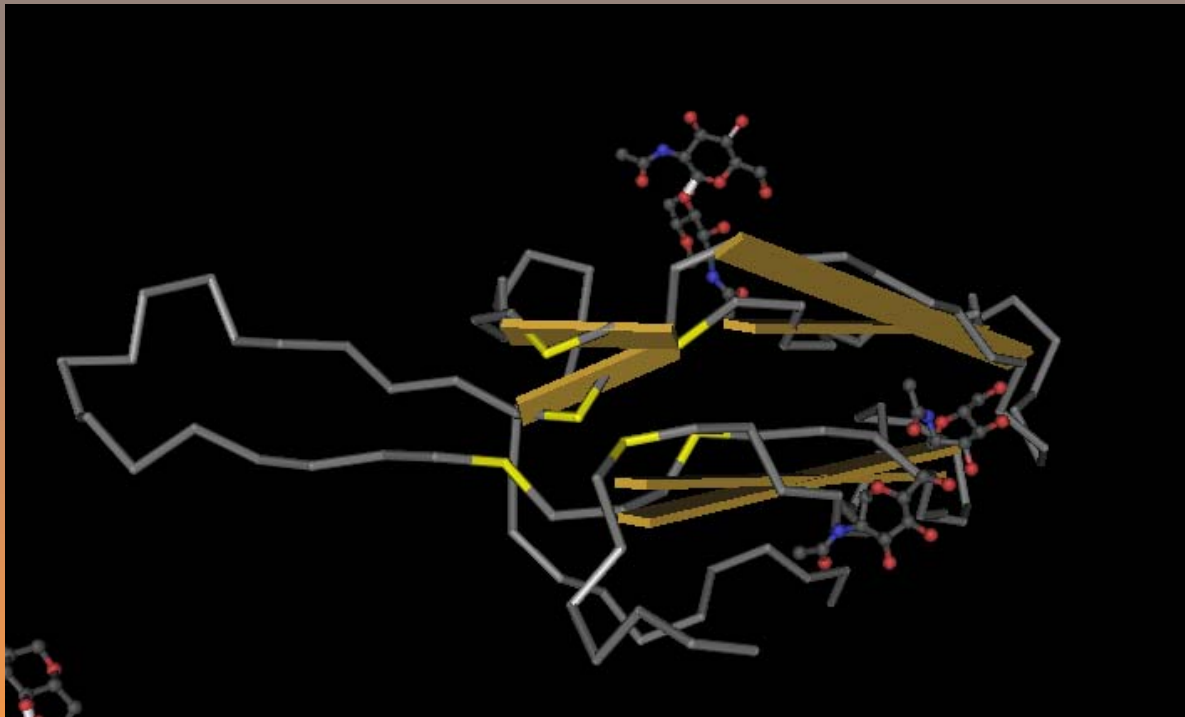
CD Length: 127 Bit Score: 213.56 E-value: 3e-56

gi 4885135	121	MEKSP	PLREEAK	FWHHF	FRKTP	ASQGV	LPIKSH	EVHWET	CRTP	PFSQT	ITHEG	CEKV	VVQNN	LCFGK	CGSV	HFPGAA	Q	200
pfam03045	1	MEEV	PLREEAK	FWHHF	VLKNS	APQLV	VLPIK	YLESH	WETCK	TLPT	QTIA	HEGCE	KVVIQ	NNLCF	GKCN	SFHV	PGKEE	80

gi 4885135	201	HSHT	SCSHCL	PAKFT	TMHL	PLNCTE	LSSVI	KVVM	LVEEC	CQCKV	KTEH	247
pfam03045	81	DSFQ	SCSHCL <td>PAKFT <th>TVHVT</th> <th>LNCTG</th> <th>LEP</th> <th>VVKVM</th> <th>KVEEC</th> <th>RC</th> <th>EA</th> <th>KKLE 127</th> </td>	PAKFT <th>TVHVT</th> <th>LNCTG</th> <th>LEP</th> <th>VVKVM</th> <th>KVEEC</th> <th>RC</th> <th>EA</th> <th>KKLE 127</th>	TVHVT	LNCTG	LEP	VVKVM	KVEEC	RC	EA	KKLE 127

Description	Pssmid	Multi-dom	E-value
[+] COG5173, SEC6, Exocyst complex subunit SEC6 [Intracellular trafficking and secretion]	34772	yes	0.63
[+] COG4964, CpaC, Flp pilus assembly protein, secretin CpaC [Intracellular trafficking and secretion]	34571	yes	2.8

Поиск 3D структур



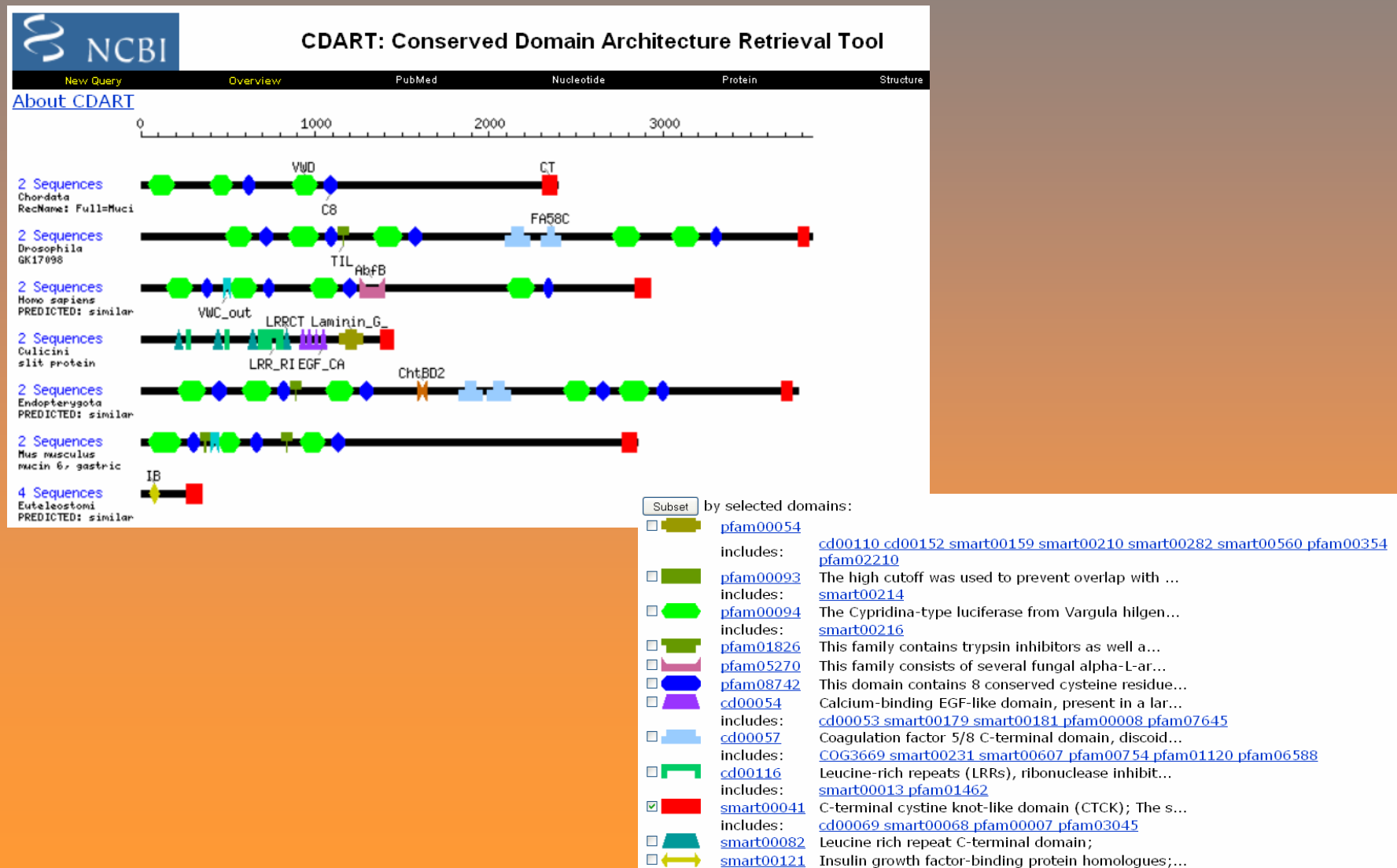
Glycoprotein hormone
beta chain
homologues.
Gonadotropins;
reproductive hormones
consisting of two
glycosylated chains
(alpha and beta) of
similar topology with
Cysteine-knot motifs

GHB - Sequence/Alignment Viewer	
View	Edit Mouse Mode Unaligned Justification Imports
<i>IHRP_B</i>	p l r p r C R P I N A T L A V E k e g C p v C I T V N T T I C A G Y C P T M T R V L Q g v l ~ ~ p a L P Q V V C N Y r d V R F E S I R L P G C p r g v n P V V S Y
<i>IFL7_B</i>	~ ~ ~ n s C E L T N I T I A I E k e e C r f C I S I N T A W C A G Y C Y T R D L V Y K d p a ~ ~ r p K I Q K T C T F k e L V Y E T V R V P G C a h h a d S L Y T Y
<i>gi 121732</i>	s l l l p C E P I N E T I S V E k d g C p k C L V F Q T S I C S G H C I T K D P S Y K s p l ~ ~ s t V Y Q R V C T Y r d V R Y E T V R L P D C r p g v d P H V T F
<i>gi 126481</i>	~ ~ r h v C H L A N A T I S A E k d h C p v C I T F T T S I C T G Y C Q T M D P V Y K t a l ~ ~ s s F K Q N I C T Y k e I R Y D T I K L P D C l p g t d P F F T Y
<i>gi 136442</i>	q a m s f C I P T E Y M M H V E r k e C a y C L T I N T T V C A G Y C M T R D V N G K l f l p k y a L S Q D V C T Y r d F M Y K T A E I P G C p r h v t P Y F S Y
<i>gi 399830</i>	v g c f g C H L K N V S I P M E r ~ ~ C g q R V C I H T T I C E G L C F S E D A V F E s p d ~ ~ e a P E H R V C N G d ~ W S Y E V K H I Q G C p ~ ~ ~ E S I T Y
<i>gi 399831</i>	f q l p r C Q L L N Q T I S L E k r g C s g C H R V E T T I C S G Y C A T K D P N Y K t s y ~ ~ n k A I Q H V C T Y g d L Y Y K T F E F P E C v p g v d P V V T Y
<i>gi 417262</i>	n l r n l C R P I N A T L A A Z h z a C n v C I T F T T S I C A G Y C P S M V R V L P a a l ~ ~ n n V P Z P V C T Y r z L R F A S I R L P G C n n g v h P M V S F
<i>gi 585226</i>	

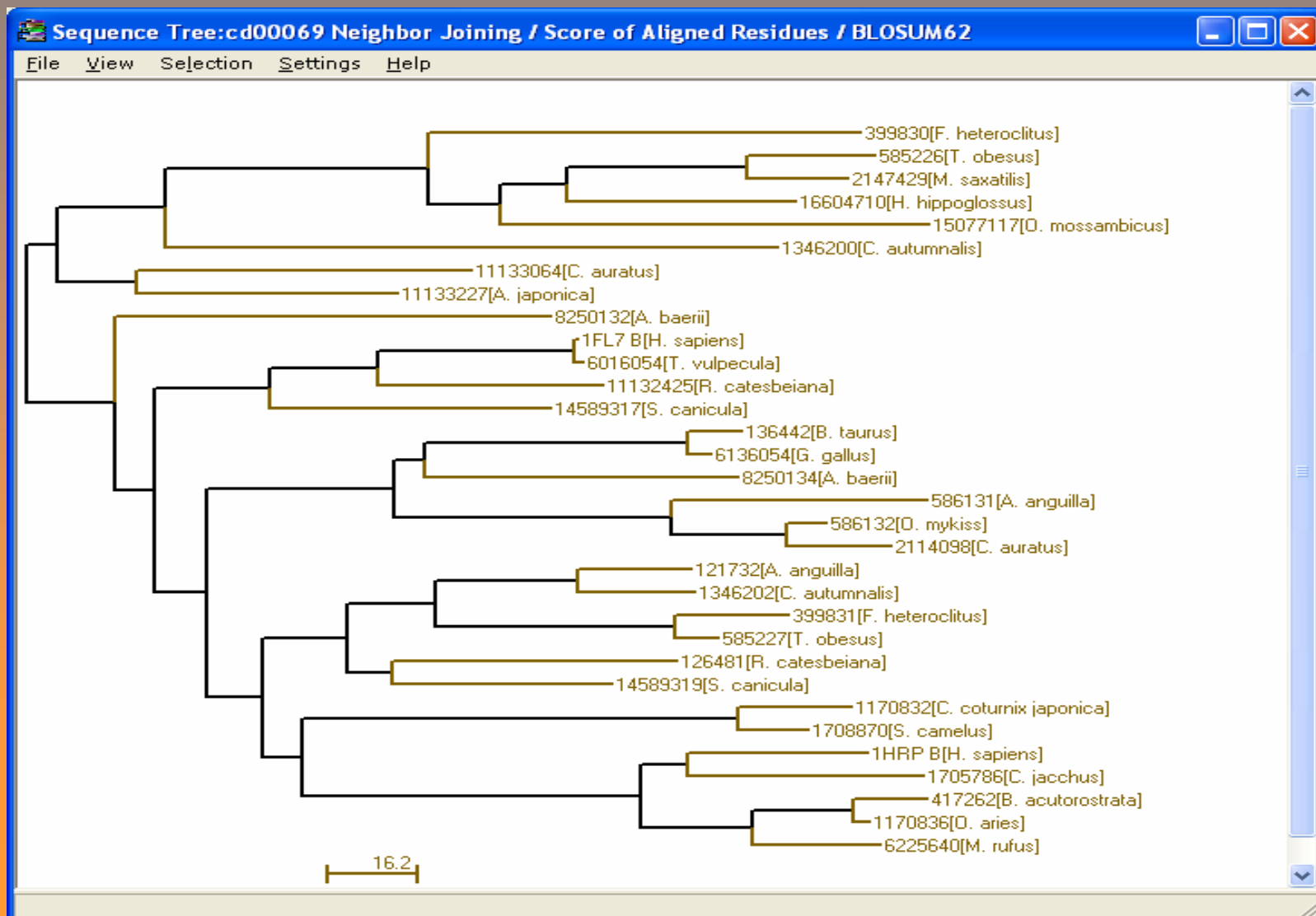
The cystine knot superfamily of ligands comprises many homodimeric and heterodimeric proteins that are involved in organogenesis, as well as tissue remodeling and repair.

- TGF-beta
- Growth differentiation factors (GDFs)
- Bone morphogenetic proteins (BMPs),
- BMP antagonists (CAN family of proteins, Twisted gastrulation protein, Chordin, Ventroptin, Noggin)
- Gonadotropins
- Platelet-derived growth factors

Архитектура консервативных доменов в различных белках

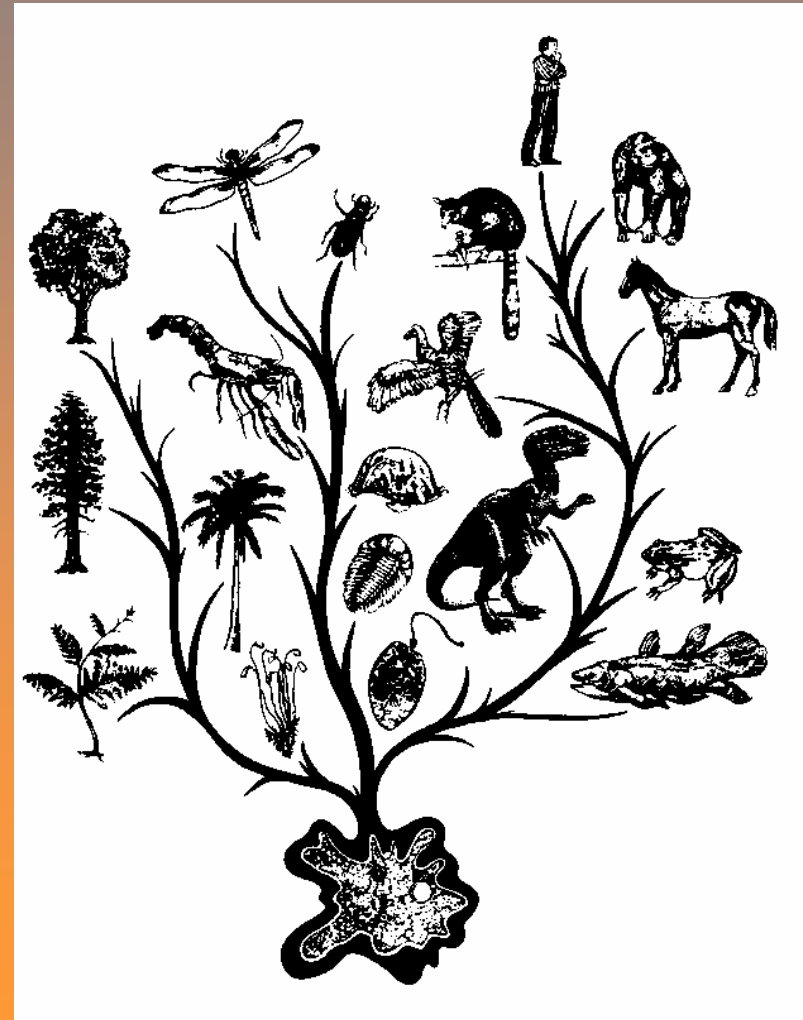


Представленность белка, гена или домена у различных групп организмов



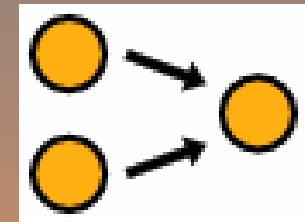
Построение филогенетических деревьев

- Параметрические и непараметрические методы построения дерева
- Бутстреп-анализ и анализ «устойчивости» ветвей
- Выявление суперсемейств генов и белков
- «Геномный» филогенетический анализ

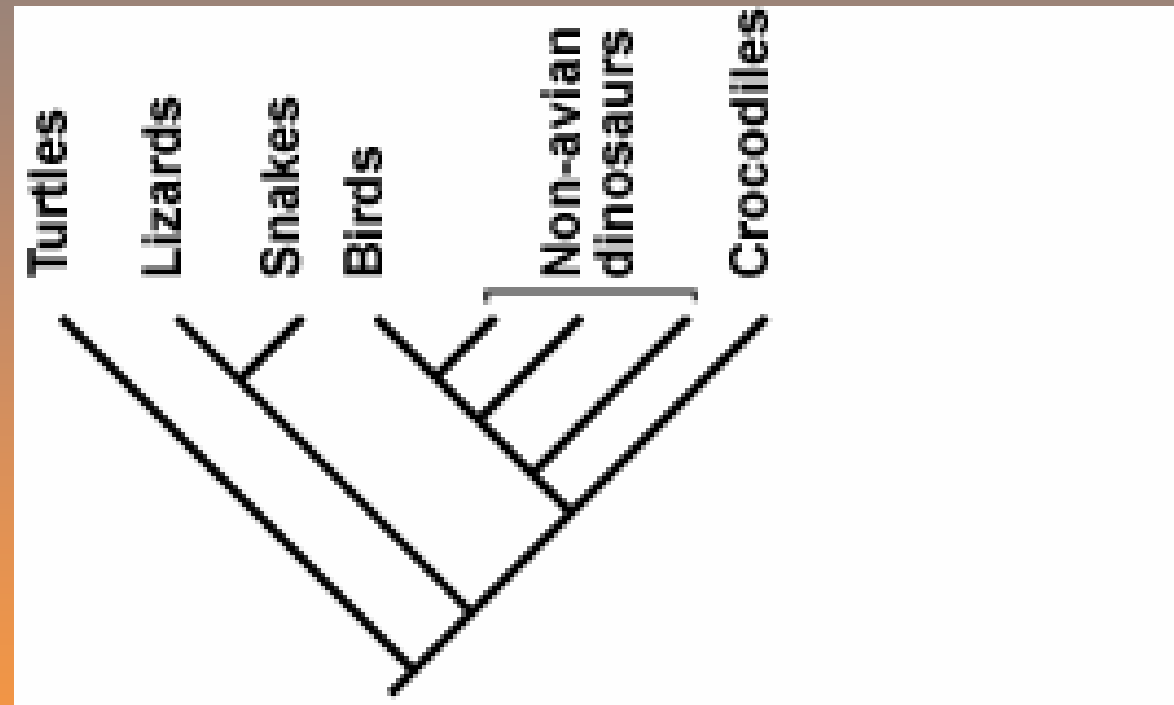


Метод связывания ближайших соседей (Neighbour-joining)

- Выбирается одна из последовательностей и сравнивается поочередно со всеми остальными.
- По результатам сравнений определяется её «ближайший сосед». После этого определяется сходство между данной парой последовательностей и всей остальной выборкой.
- Далее такая же процедура производится для всех оставшихся последовательностей, но уже без определенной ранее пары.
- Описанный выше цикл повторяется до момента исчерпания выборки.



Кладистическая систематика. Монофилия таксонов.



Самостоятельными таксонами считаются клады – группы, произошедшие от общего предка. Как пример – таксономическое положение птиц и рептилий.

Филогенетические отношения BMP антагонистов человека

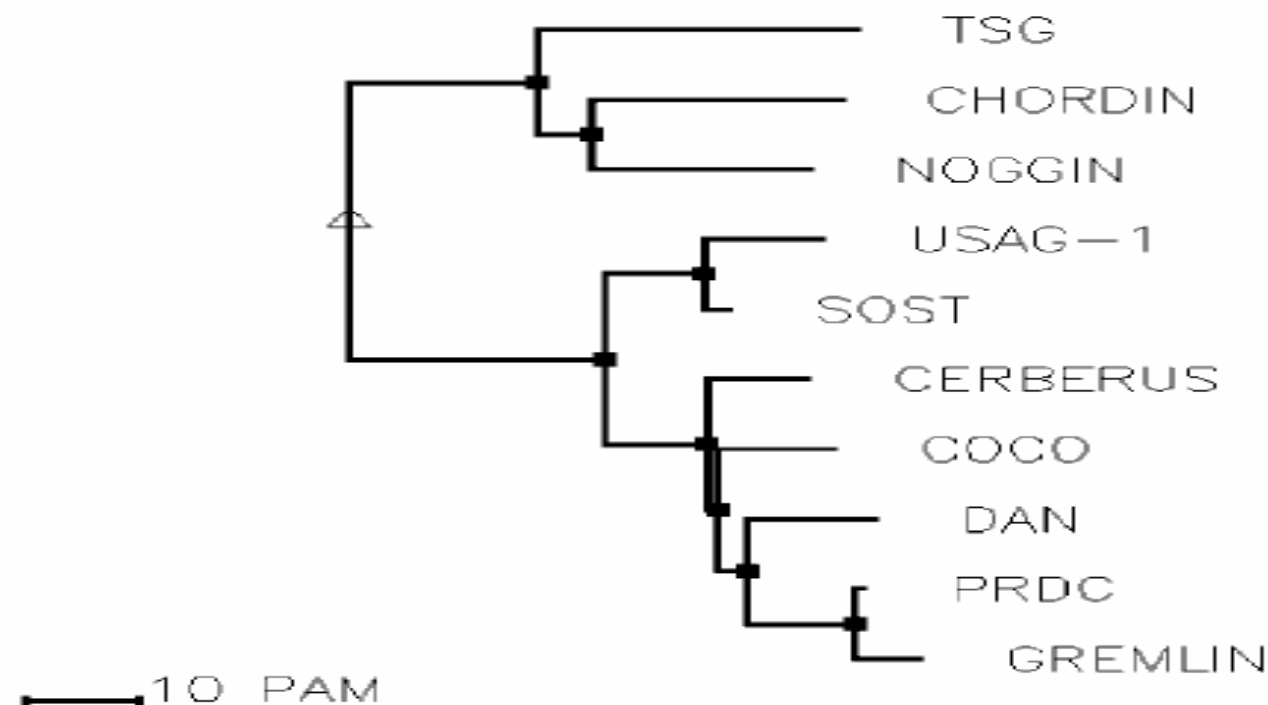


Fig. 3. Phylogenetic Relationship of BMP Antagonists

Phylogenetic tree of human BMP antagonists based on the alignment of cystine knot sequences of representative members from each subfamily. Chordin has four cystine knot domains, all of which gave the same result—the fourth cystine knot domain for chordin is used here. The MultAlin server at <http://prodes.toulouse.inra.fr/multalin/multalin.html> was used for phylogenetic tree construction (66).

СТАТИСТИЧЕСКОЕ ПОДТВЕРЖДЕНИЕ КОРРЕКТНОСТИ ДЕНДРОГРАММ

- Методы построения дендрограмм могут дать несколько вариантов в зависимости от исходных параметров. Чтобы получить корректную дендрограмму, необходимо проверять каждую из полученных дендрограмм и сравнивать их между собой.
- Конечным результатом проверки является **консенсусная дендрограмма**, оптимально сочетающая в себе преимущества всех полученных изначально.

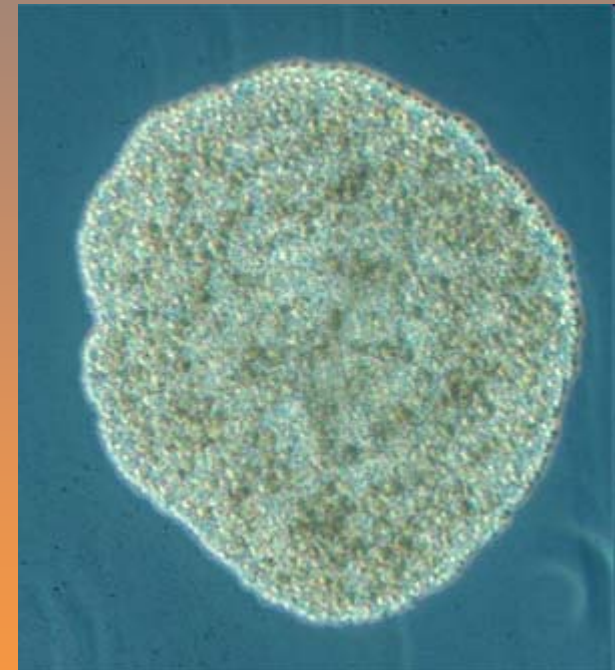
Bootstrap

- Общий принцип состоит в создании нескольких сотен новых последовательностей из случайным образом перемешанных участков исходных, и построении на основании сравнения новых последовательностей множества дендрограмм с использованием исходного метода.
- Если положение ветви в одной из новых дендрограмм соответствует её положению в исходной, **индекс BCL (bootstrap confidence level) данной ветви принимает значение 1**, в обратном случае **BCL = 0**.
- Такая процедура выполняется для каждой вновь построенной дендрограммы (то есть, несколько сотен раз), и каждый раз индексы BCL определяются для всех ветвей.
- Далее **для каждой ветви вычисляется среднее значение BCL** (результат деления суммы значений всех индексов на количество дендрограмм).
- После этого строится согласованная с тестом bootstrap дендрограмма (консенсусная), в которой средние значения BCL для каждой ветви должны быть больше порогового значения (например, 0,95)



Уточнение данных и построение филогенетических отношений генов и белков

- Например: гомолог Cerberus можно найти в базе данных предсказанных белков первичноротых и вторичноротых беспозвоночных и *Nematostella vectensis*. При этом они содержат основной функциональный домен: Cysteine-knot motif
- Также в базах данных можно найти ген, экспрессирующий белок, относящийся к **cystine knot superfamily** и близкородственный DAN/Cerberus у *Trichoplax adhaerens*
- Какую функцию могли бы исходно выполнять белки **cystine knot superfamily** в этом случае?



Trichoplax adhaerens

Фрагмент выравнивания Cerberus-подобных белков в районе cysteine-knot motif

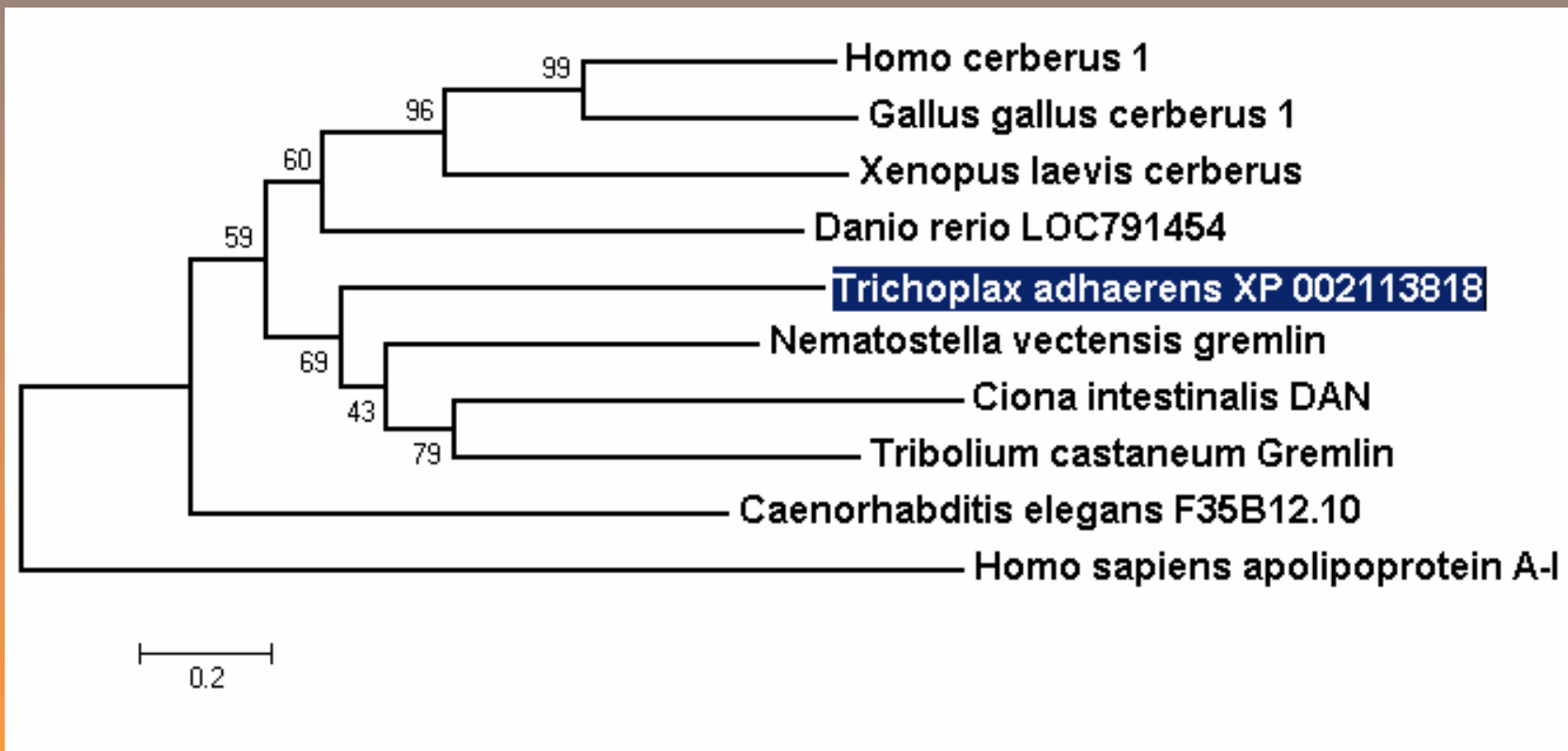
```

Homo_cerberus_l_gi|5566 HEMERKTPASQGVILPIKSHEVHWETCRTVPFSQTITHEGCEKVVVQNNLCFGKCGSVHFP-----
Xenopus_laevis_cerberus NFLVKMNGAPQNTSHGSKAQEIMKEACKTLPTQNIWHENCDRMVIQNNLCFGKCSLHVP-----
Gallus_gallus_cerberus HEMLRKNSASEEVVLPIKTNEMHQETCRTLPFSQSVAME SCEKVIVQNNLCFGKCSSFHVP-----
Danio_rerio_L0C791454_g KVVHKSERKKEAVALRINPKDMNQSCAAVPFTQRITEGCEVTVTVMNNLCYGCSSMFPVSSG----
Ciona_intestinalis_DAN_ ---HNILKTSAKAKTITERKYLKRDWCKSQPVRYQYIRFADGCKGELINQFCYGCNSSEFYIPKDIEIDE
Caenorhabditis_elegans_ GDTVEEGRKHALQLADPDALIMNQRCDGQKEFKQRIKRVGCLTKVVVNRLCHGACASIFIPRMHS---
Tribolium_castaneum_sim FNGRKLLKSSKNALFVTKKEYLQRDWCKTEPLIQKVKEEGCLTRTVINRECYGCNSSEFYIPKNPKKRH
Nematostella_vectensis_ GKYLPLKGGSTIMSLKLRNDLRGDWCKLRPVLQKLHHPGCNSSFIMNMICYGQCMSEFFIP-----
Trichoplax_adhaerens_XP ---INSHHSYGISQLPIAPN--QKAWCKLAKIEQVLSHPGCI SKTISNHICVGGCYSYRIPKSYPP--

-----GAAQHSHTSCSHCLPAKFETTMHLPLNCTELSSVI-----KVVMLVEECQCKVKTE:
-----NQQDRRNT-CSHCLPSKFETLNHLTLNCTGSKNVV-----KVVMMVEECTCEANKS:
-----GPDDRLYTFCSKCLPTKFSMKHFDLNCTSSVPVV-----KKVMIVEECNCETQKI:
-----GSHGQQAQCTRCGP SRARSVLLHLRCG-SEVRE-----RRVLIVEECKCETSSE:
-----RDGEIETDYFRFCAECKPKTEEWSVRLRCRRKGRKKRGRYVIKRVKRVKGCTCISVPE:
-----KKLKAAEFERSCAACAPAEYDYVDITLDCPGRTPTTAT---KTIVKVKSCCKKEVRI:
RHLPVSEETEDDQNGPAFKACAFCRPSKFTWISVTLKCPSLMPPFRK---KRIQRIRKQCKCIAANV:
-----RHETSCAECTPVSKNVVSVHLKCAGLDKVV-----KKVSIQSCSCRPCGN:
-----EAGQENLQHCECHVVDHTWNTVELKCP TLKNNVD-----KLVQYIRSCDCRRCHH:

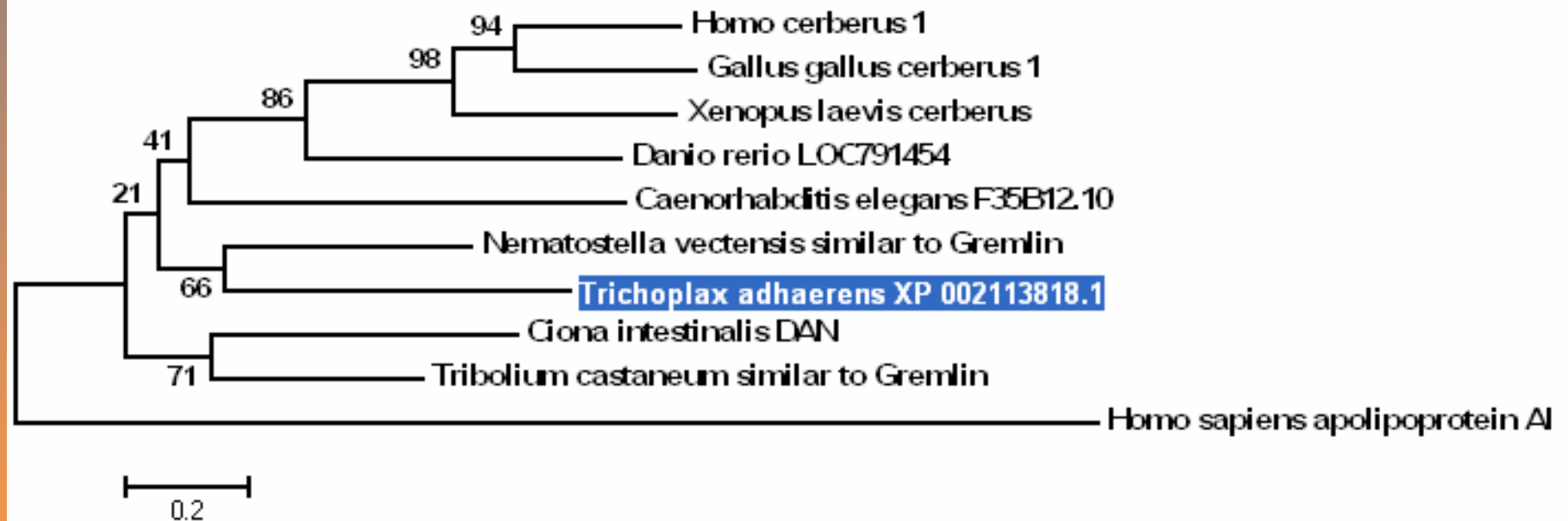
```

Филогенетические отношения гомологов Cerberus



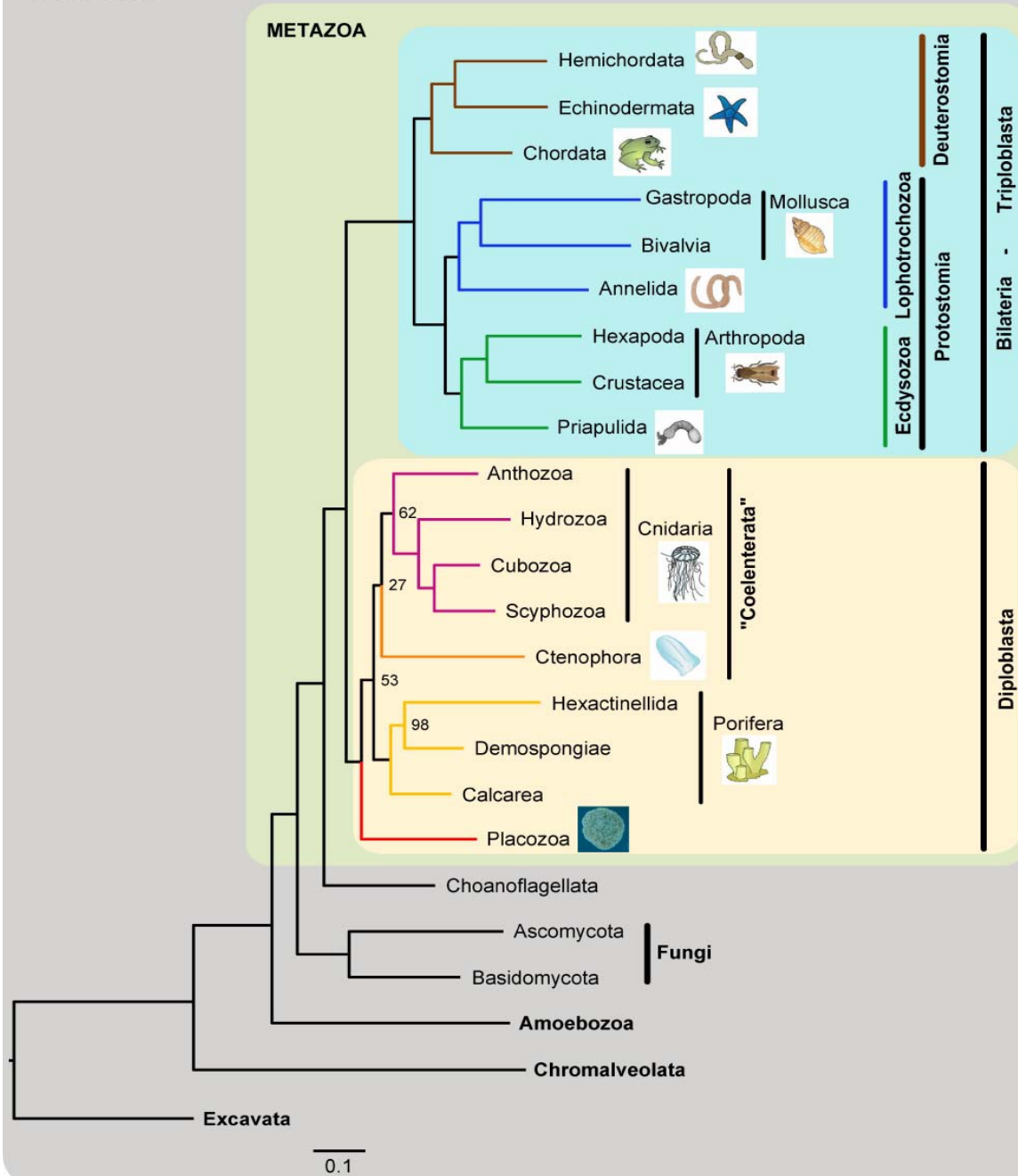
Neighbour-joining, bootstrap 1000, матрица аминокислотных замен с коррекцией Пуассона, **полные аминокислотные последовательности**

Филогенетические отношения гомологов Cerberus

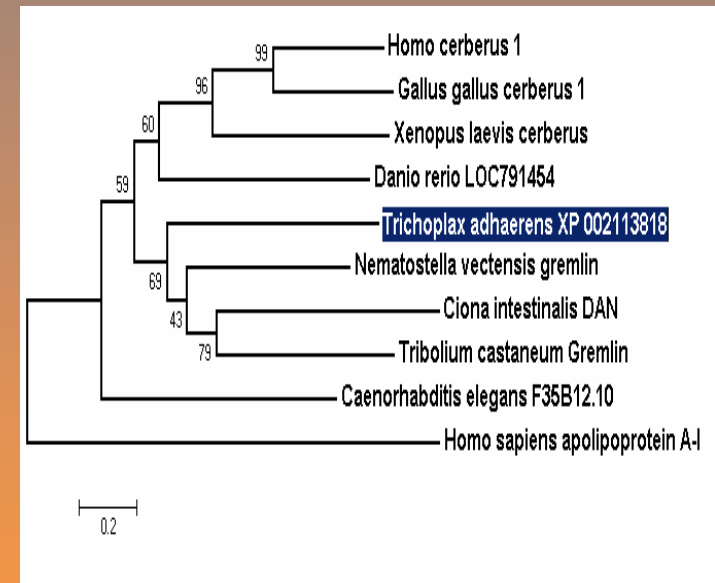


Neighbour-joining, bootstrap 1000, матрица аминокислотных замен с коррекцией Пуассона, аминокислотные последовательности cysteine-knot motif

EUKARYOTA



Дендрограмма, построенная с использованием последовательностей тех или иных консервативных белков, не обязательно совпадает с филогенетическим древом, построенным с использованием общепринятых методов

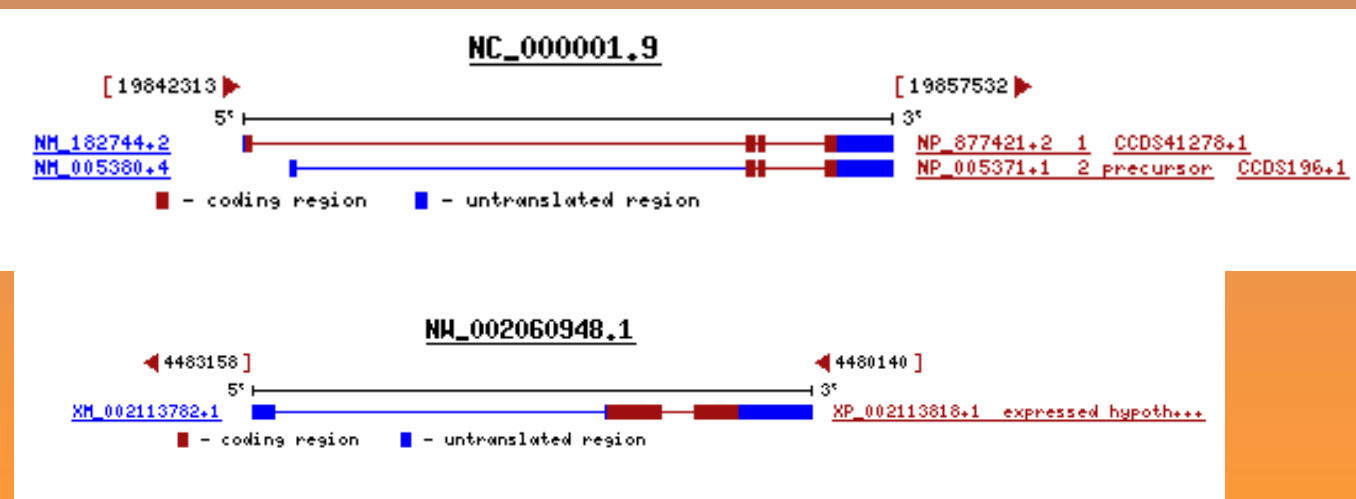


Скорость эволюционной дивергенции последовательности белков может быть различной в разных таксонах. Кроме того, возможны утраты генов, кодирующих данный белок, в различных группах организмов

NBL1 (DAN)

- This gene product is the founding member of the evolutionarily conserved CAN (Cerberus and DAN) family of proteins, which contain a domain resembling the CTCK (**C-terminal cystine knot-like**) motif found in a number of signaling molecules. These proteins are secreted, and act as BMP (bone morphogenetic protein) antagonists by binding to BMPs and preventing them from interacting with their receptors. They may thus play an important role during growth and development. Alternatively spliced transcript variants encoding distinct isoforms have been identified for this gene.

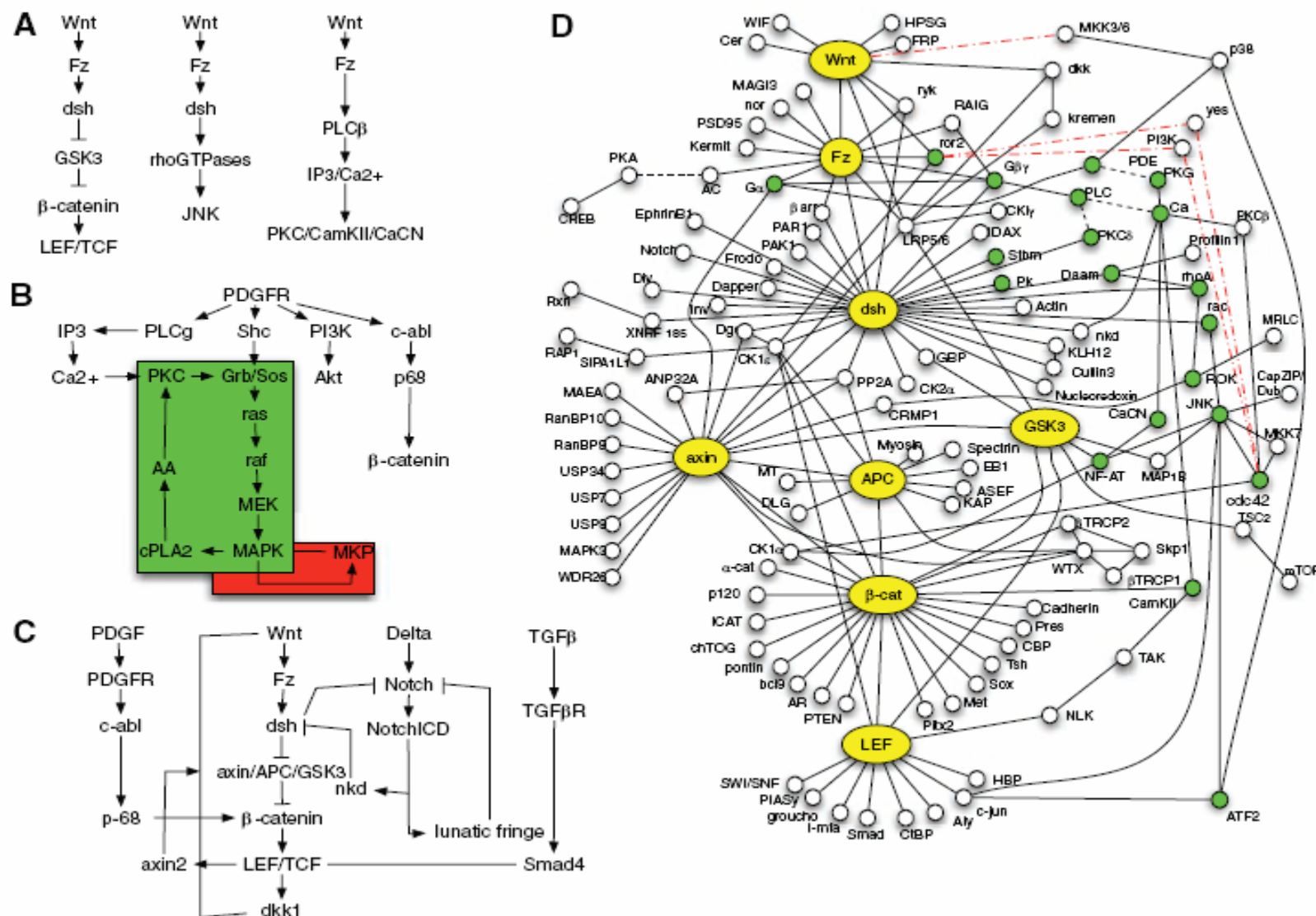
DAN
(HOMO)



TRIADDRAFT_64027
Trichoplax adhaerens

	large-scale data	computational tools	databases of biological knowledge
1990	ESTs	BLAST	
1995	complete genomes	CLUSTAL W HMM	SCOP KEGG
	gene expressions SNPs	PSI-BLAST	COG Pfam SMART
2000	protein interactions		GO

Моделирование внутриклеточного сигналинга на основе белок-белковых взаимодействий



Databases of organized biological knowledge

Knowledge	Database	URL
protein functional sites	PROSITE	http://www.expasy.ch/prosite/
	BLOCKS	http://www.blocks.fhcrc.org/
	PRINTS	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
	ProDom	http://prodes.toulouse.inra.fr/prodom/doc/prodom.html
	Pfam	http://pfam.wustl.edu/
	SMART	http://smart.embl-heidelberg.de/
	TIGRFAMs	http://www.tigr.org/TIGRFAMs/
protein 3D folds	SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/
	CATH	http://www.biochem.ucl.ac.uk/bsm/cath_new/
transcription factors	TRANSFAC	http://transfac.gbf.de/TRANSFAC/
protein interactions	BIND	http://www.bind.ca/
	DIP	http://dip.doe-mbi.ucla.edu/
protein pathways	KEGG	http://www.genome.ad.jp/kegg/
	EcoCyc	http://www.ecocyc.org/
ortholog groups	COG	http://www.ncbi.nlm.nih.gov/COG/
controlled vocabulary	GO	http://www.geneontology.org/

bioinformatics now

mathematics

informatics

physics

biology

chemistry

medicine

bioinformatics in the future

mathematics

informatics

physics

biology

chemistry

medicine