Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 19

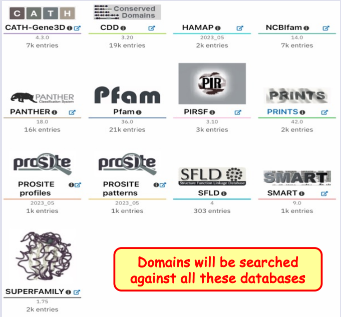
TODAY'S TOPICS

- ➢ Protein family/domain databases (InterPro, SMART, etc.)
- ➢ Distance estimation
  - • Nucleotide substitutions

InterPro Classification of protein families

https://www.ebi.ac.uk/interpro/

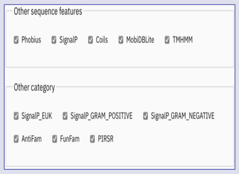
[Member databases]



CATH-Gene3D 4.3.0 7k entries
CDD 3.20 19k entries
HAMAP 2023_05 2k entries
NCBIfam 14.0 7k entries
PANTHER 18.0 16k entries
Pfam 36.0 21k entries
PIRSF 3.10 3k entries
PRINTS 42.0 2k entries
PROSITE profiles 2023_05 1k entries
PROSITE patterns 2023_05 1k entries
SFLD 4 303 entries
SMART 9.0 1k entries
SUPERFAMILY 1.75 2k entries

Domains will be searched against all these databases

[Other information]



Other sequence features

☑ Phobius ☑ SignalP ☑ Coils ☑ MobiDBLite ☑ TMHMM

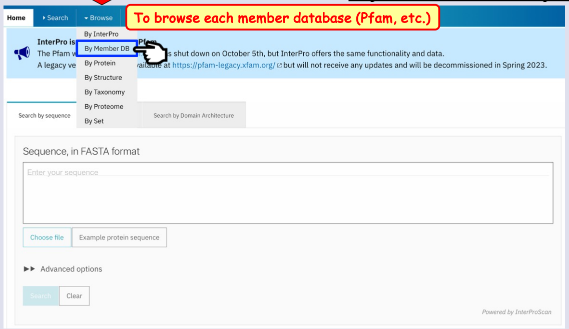
Other category

☑ SignalP_EUK ☑ SignalP_GRAM_POSITIVE ☑ SignalP_GRAM_NEGATIVE
☑ AntiFam ☑ FunFam ☑ PIRSR

Transmembrane, signal peptide, coils, disordered regions will be predicted

InterPro Classification of protein families

https://www.ebi.ac.uk/interpro/



To browse each member database (Pfam, etc.)

InterPro is ... Pfam
The Pfam w... s shut down on October 5th, but InterPro offers the same functionality and data.
A legacy ve... vailable at https://pfam-legacy.xfam.org/ but will not receive any updates and will be decommissioned in Spring 2023.

Sequence, in FASTA format

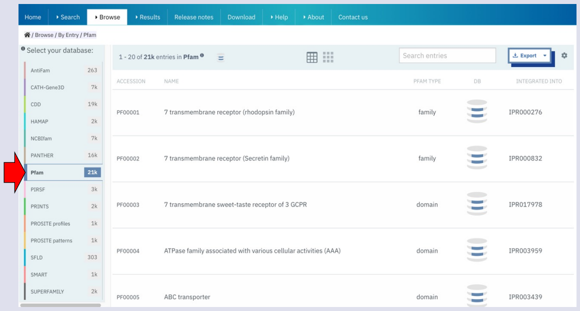
Enter your sequence

Powered by InterProScan

InterPro Classification of protein families

[Pfam database entries]

https://www.ebi.ac.uk/interpro/



/ Browse / By Entry / Pfam

Select your database:

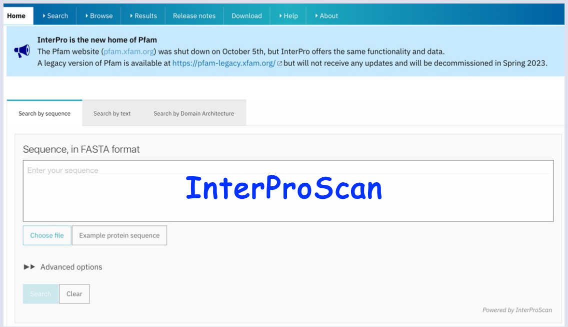
AntiFam	263
CATH-Gene3D	7k
CDD	19k
HAMAP	2k
NCBIfam	7k
PANTHER	16k
Pfam	21k
PIRSF	3k
PRINTS	2k
PROSITE profiles	1k
PROSITE patterns	1k
SFLD	303
SMART	1k
SUPERFAMILY	2k

1 - 20 of 21k entries in Pfam

ACCESSION	NAME	PFAM TYPE	DB	INTEGRATED INTO
PF00001	7 transmembrane receptor (rhodopsin family)	family		IPR000276
PF00002	7 transmembrane receptor (Secretin family)	family		IPR000832
PF00003	7 transmembrane sweet-taste receptor of 3 GCPR	domain		IPR017978
PF00004	ATPase family associated with various cellular activities (AAA)	domain		IPR003959
PF00005	ABC transporter	domain		IPR003439

InterPro Classification of protein families

https://www.ebi.ac.uk/interpro/



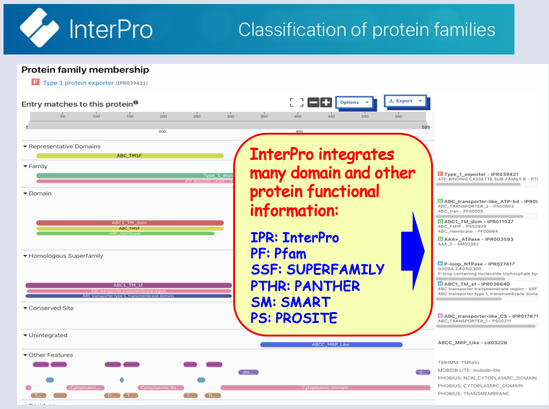
InterPro is the new home of Pfam
The Pfam website (pfam.xfam.org) was shut down on October 5th, but InterPro offers the same functionality and data.
A legacy version of Pfam is available at https://pfam-legacy.xfam.org/ but will not receive any updates and will be decommissioned in Spring 2023.

Sequence, in FASTA format

Enter your sequence

InterProScan

Powered by InterProScan



Protein family membership

F Type 1 protein exporter (IPR039421)

Entry matches to this protein

F Type_1_exporter - IPR039421
ATP-BINDING CASSETTE SUB-FAMILY B - PTI

D ABC_transporter-like_ATP-bd - IPR0
ABC_TRANSPORTER_2 - PS50893
ABC_tran - PF00005

D ABC1_TM_dom - IPR011527
ABC_TM1F - PS50929
ABC_membrane - PF00664

D AAA+_ATPase - IPR003593
AAA_5 - SM00382

H P-loop_NTPase - IPR027417
G3DSA:3.40.50.300
P-loop containing nucleoside triphosphate hy

H ABC1_TM_sf - IPR036640
ABC transporter transmembrane region - SSF
ABC transporter type 1, transmembrane doma

S ABC_transporter-like_CS - IPR017871
ABC_TRANSPORTER_1 - PS00211

ABCC_MRP_Like - cd03228

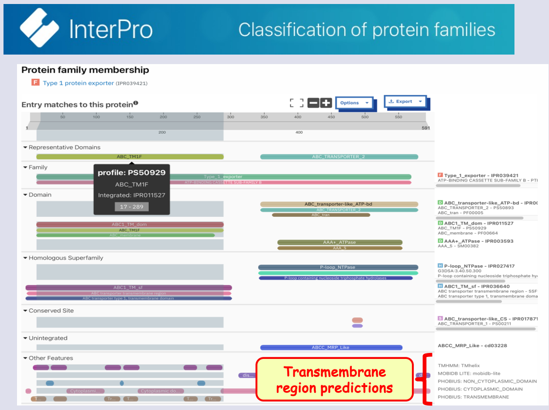
TMHMM: TMhelix
MOBIDB LITE: mobidb-lite
PHOBIUS: NON_CYTOPLASMIC_DOMAIN
PHOBIUS: CYTOPLASMIC_DOMAIN
PHOBIUS: TRANSMEMBRANE

InterPro integrates many domain and other protein functional information:

IPR: InterPro
PF: Pfam
SSF: SUPERFAMILY
PTHR: PANTHER
SM: SMART
PS: PROSITE

https://www.ebi.ac.uk/interpro/

7



Protein family membership

F Type 1 protein exporter (IPR039421)

Entry matches to this protein

F Type_1_exporter - IPR039421
ATP-BINDING CASSETTE SUB-FAMILY B - PTI

D ABC_transporter-like_ATP-bd - IPR0
ABC_TRANSPORTER_2 - PS50893
ABC_tran - PF00005

D ABC1_TM_dom - IPR011527
ABC_TM1F - PS50929
ABC_membrane - PF00664

D AAA+_ATPase - IPR003593
AAA_5 - SM00382

H P-loop_NTPase - IPR027417
G3DSA:3.40.50.300
P-loop containing nucleoside triphosphate hy

H ABC1_TM_sf - IPR036640
ABC transporter transmembrane region - SSF
ABC transporter type 1, transmembrane doma

S ABC_transporter-like_CS - IPR017871
ABC_TRANSPORTER_1 - PS00211

ABCC_MRP_Like - cd03228

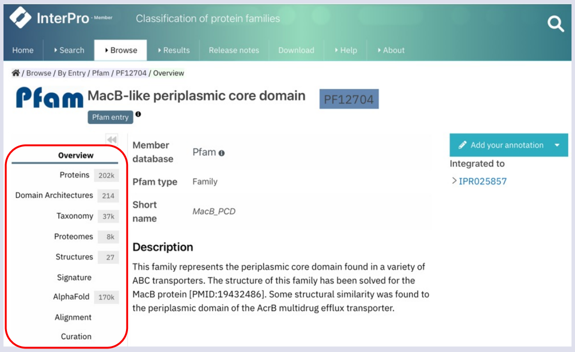
TMHMM: TMhelix
MOBIDB LITE: mobidb-lite
PHOBIUS: NON_CYTOPLASMIC_DOMAIN
PHOBIUS: CYTOPLASMIC_DOMAIN
PHOBIUS: TRANSMEMBRANE

Transmembrane region predictions

https://www.ebi.ac.uk/interpro/

8

InterPro/Pfam family/domain information



/ Browse / By Entry / Pfam / PF12704 / Overview

Pfam MacB-like periplasmic core domain PF12704

Overview	
Proteins	202k
Domain Architectures	214
Taxonomy	37k
Proteomes	8k
Structures	27
Signature	
AlphaFold	170k
Alignment	
Curation	

Member database: Pfam

Pfam type: Family

Short name: *MacB_PCD*

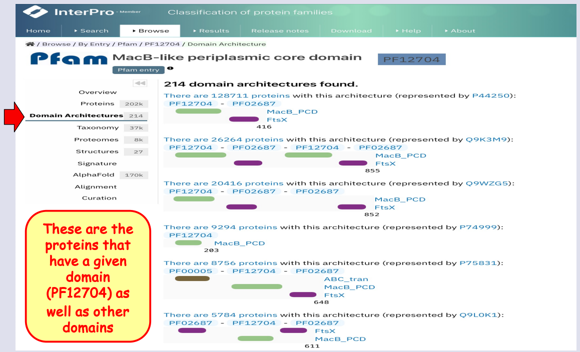
Integrated to
> IPR025857

Description

This family represents the periplasmic core domain found in a variety of ABC transporters. The structure of this family has been solved for the MacB protein [PMID:19432486]. Some structural similarity was found to the periplasmic domain of the AcrB multidrug efflux transporter.

9

InterPro/Pfam family/domain information



/ Browse / By Entry / Pfam / PF12704 / Domain Architecture

Pfam MacB-like periplasmic core domain PF12704

Overview	
Proteins	202k
Domain Architectures	214
Taxonomy	37k
Proteomes	8k
Structures	27
Signature	
AlphaFold	170k
Alignment	
Curation	

214 domain architectures found.

There are 128711 proteins with this architecture (represented by P44250):
PF12704 - PF02687
MacB_PCD
FtsX
416

There are 26264 proteins with this architecture (represented by Q9K3M9):
PF12704 - PF02687 - PF12704 - PF02687
MacB_PCD
FtsX
855

There are 20416 proteins with this architecture (represented by Q9WZG5):
PF12704 - PF02687 - PF02687
MacB_PCD
FtsX
852

There are 9294 proteins with this architecture (represented by P74999):
PF12704
MacB_PCD
203

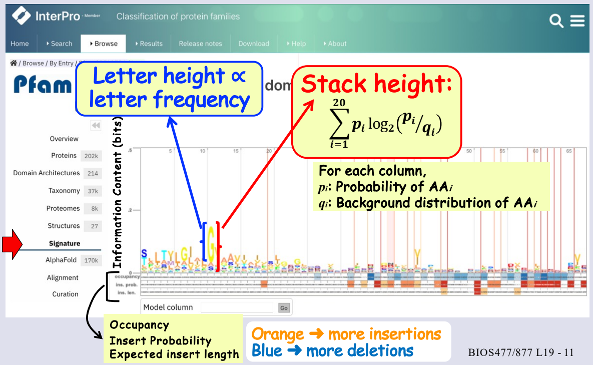
There are 8756 proteins with this architecture (represented by P75831):
PF00005 - PF12704 - PF02687
ABC_tran
MacB_PCD
FtsX
648

There are 5784 proteins with this architecture (represented by Q9L0K1):
PF02687 - PF12704 - PF02687
FtsX
MacB_PCD
611

These are the proteins that have a given domain (PF12704) as well as other domains

10

Profile HMM logo (signature)



/ Browse / By Entry /

Overview	
Proteins	202k
Domain Architectures	214
Taxonomy	37k
Proteomes	8k
Structures	27
Signature	
AlphaFold	170k
Alignment	
Curation	

Letter height ∝ letter frequency

Stack height:

$$\sum_{i=1}^{20} p_i \log_2(p_i/q_i)$$

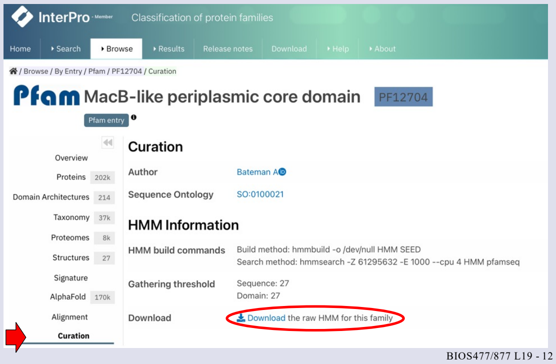
For each column,
p_i: Probability of AA_i
q_i: Background distribution of AA_i

Occupancy
Insert Probability
Expected insert length

Orange ➔ more insertions
Blue ➔ more deletions

11

IntperPro/Pfam family/domain information



/ Browse / By Entry / Pfam / PF12704 / Curation

Pfam MacB-like periplasmic core domain PF12704

Overview	
Proteins	202k
Domain Architectures	214
Taxonomy	37k
Proteomes	8k
Structures	27
Signature	
AlphaFold	170k
Alignment	
Curation	

Curation

Author: Bateman A

Sequence Ontology: SO:0100021

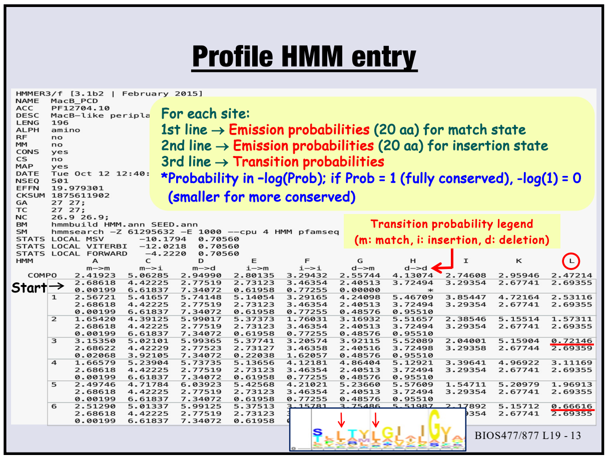
HMM Information

HMM build commands:
Build method: hmmbuild -o /dev/null HMM SEED
Search method: hmmsearch -Z 61295632 -E 1000 --cpu 4 HMM pfamseq

Gathering threshold:
Sequence: 27
Domain: 27

Download: Download the raw HMM for this family

12



Profile HMM entry

```
HMMER3/f [3.1b2 | February 2015]
NAME  MacB_PCD
ACC   PF12704.10
DESC  MacB-like peripla
LENG  196
ALPH  amino
RF    no
MM    no
CONS  yes
CS    no
MAP   yes
DATE  Tue Oct 12 12:40:
NSEQ  501
EFFN  19.979301
CKSUM 1875611902
GA    27 27;
TC    27 27;
NC    26.9 26.9;
BM    hmmbuild HMM.ann SEED.ann
SM    hmmsearch -Z 61295632 -E 1000 --cpu 4 HMM pfamseq
STATS LOCAL MSV       -10.1794  0.70560
STATS LOCAL VITERBI   -12.0218  0.70560
STATS LOCAL FORWARD    -4.2220  0.70560
HMM          A        C        D        E        F        G        H        I        K        L
            m->m     m->i     m->d     i->m     i->i     d->m     d->d
  COMPO   2.41923  5.06285  2.94990  2.80135  3.29432  2.55744  4.13074  2.74608  2.95946  2.47214
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355
          0.00199  6.61837  7.34072  0.61958  0.77255  0.00000        *
      1   2.56721  5.41657  5.74148  5.14054  3.29165  4.24098  5.46709  3.85447  4.72164  2.53116
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355
          0.00199  6.61837  7.34072  0.61958  0.77255  0.48576  0.95510
      2   1.65420  4.39125  5.99017  5.37373  1.76031  3.16932  5.51657  2.38546  5.15514  1.57311
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355
          0.00199  6.61837  7.34072  0.61958  0.77255  0.48576  0.95510
      3   3.15350  5.02101  5.99365  5.37741  3.20574  3.92115  5.52089  2.04001  5.15904  0.72146
          2.68622  4.42229  2.77523  2.73127  3.46358  2.40516  3.72498  3.29358  2.67744  2.69359
          0.02068  3.92105  7.34072  0.22038  1.62057  0.48576  0.95510
      4   1.66579  5.23904  5.73735  5.13656  4.12181  4.86404  5.12921  3.39641  4.96922  3.11169
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355
          0.00199  6.61837  7.34072  0.61958  0.77255  0.48576  0.95510
      5   2.49746  4.71784  6.03923  5.42568  4.21021  5.23660  5.57609  1.54711  5.20979  1.96913
          2.68618  4.42225  2.77519  2.73123  3.46354  2.40513  3.72494  3.29354  2.67741  2.69355
          0.00199  6.61837  7.34072  0.61958  0.77255  0.48576  0.95510
      6   2.51290  5.01337  5.99125  5.37513  3.15781  3.75486  5.51987  2.17892  5.15712  0.66616
          2.68618  4.42225  2.77519  2.73123                             354  2.67741  2.69355
          0.00199  6.61837  7.34072  0.61958  0
```

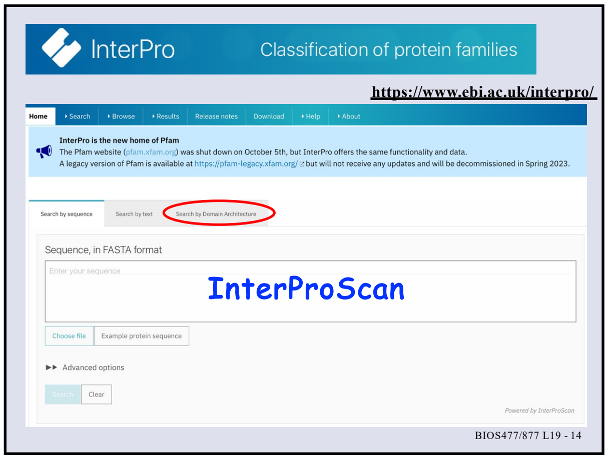
Start →

For each site:
1st line → Emission probabilities (20 aa) for match state
2nd line → Emission probabilities (20 aa) for insertion state
3rd line → Transition probabilities
*Probability in -log(Prob); if Prob = 1 (fully conserved), -log(1) = 0
(smaller for more conserved)

Transition probability legend
(m: match, i: insertion, d: deletion)

BIOS477/877 L19 - 13

13



InterPro Classification of protein families

https://www.ebi.ac.uk/interpro/

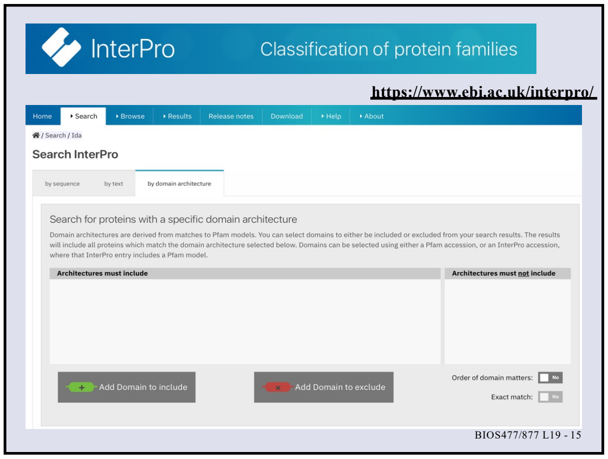
InterPro is the new home of Pfam
The Pfam website (pfam.xfam.org) was shut down on October 5th, but InterPro offers the same functionality and data.
A legacy version of Pfam is available at https://pfam-legacy.xfam.org/ but will not receive any updates and will be decommissioned in Spring 2023.

Sequence, in FASTA format

InterProScan

BIOS477/877 L19 - 14

14



InterPro Classification of protein families

https://www.ebi.ac.uk/interpro/

Search InterPro

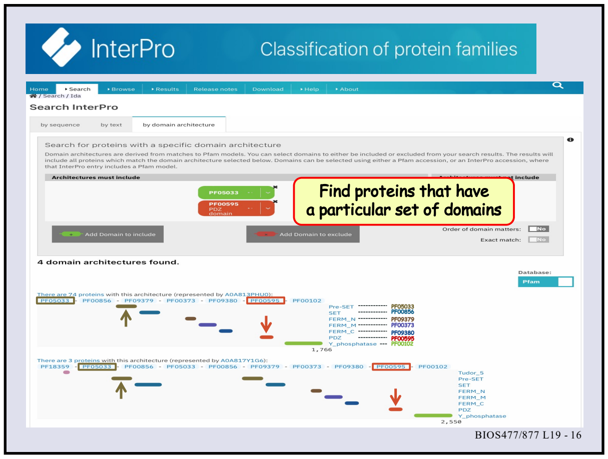
Search for proteins with a specific domain architecture

Domain architectures are derived from matches to Pfam models. You can select domains to either be included or excluded from your search results. The results will include all proteins which match the domain architecture selected below. Domains can be selected using either a Pfam accession, or an InterPro accession, where that InterPro entry includes a Pfam model.

Order of domain matters: No
Exact match: No

BIOS477/877 L19 - 15

15



InterPro Classification of protein families

Search InterPro

Search for proteins with a specific domain architecture

Domain architectures are derived from matches to Pfam models. You can select domains to either be included or excluded from your search results. The results will include all proteins which match the domain architecture selected below. Domains can be selected using either a Pfam accession, or an InterPro accession, where that InterPro entry includes a Pfam model.

Find proteins that have
a particular set of domains

Order of domain matters: No
Exact match: No

4 domain architectures found.

Database: Pfam

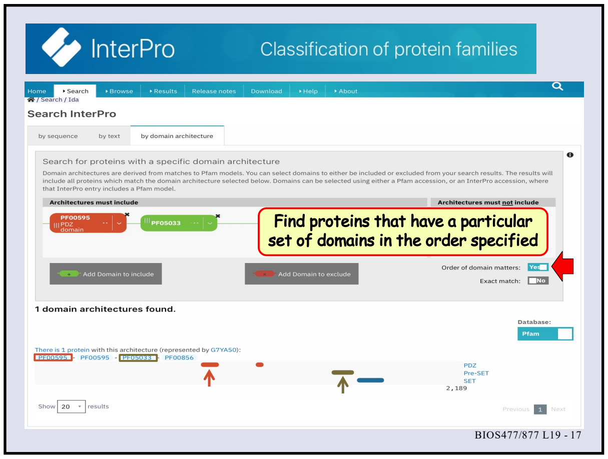
There are 74 proteins with this architecture (represented by A0A813PHU0):
PF05033 - PF00856 - PF09379 - PF00373 - PF09380 - PF00595 - PF00102

Pre-SET ------------ PF05033
SET ------------ PF00856
FERM_N ------------ PF09379
FERM_M ------------ PF00373
FERM_C ------------ PF09380
PDZ ------------ PF00595
Y_phosphatase --- PF00102
1,766

There are 3 proteins with this architecture (represented by A0A817Y1G6):
PF18359 - PF05033 - PF00856 - PF05033 - PF00856 - PF09379 - PF00373 - PF09380 - PF00595 - PF00102

BIOS477/877 L19 - 16

16



InterPro Classification of protein families

Search InterPro

Search for proteins with a specific domain architecture

Domain architectures are derived from matches to Pfam models. You can select domains to either be included or excluded from your search results. The results will include all proteins which match the domain architecture selected below. Domains can be selected using either a Pfam accession, or an InterPro accession, where that InterPro entry includes a Pfam model.

Find proteins that have a particular
set of domains in the order specified

Order of domain matters: Yes
Exact match: No

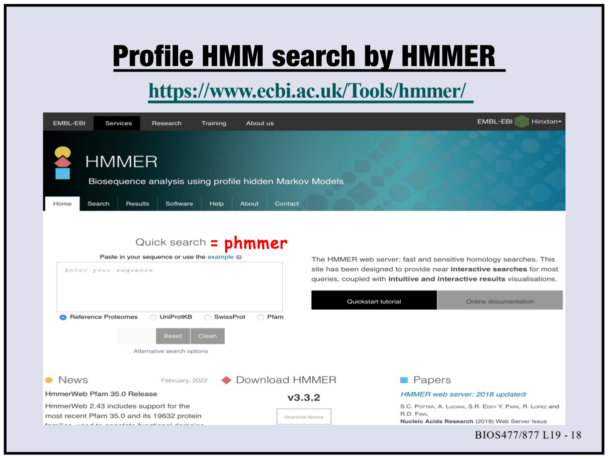
1 domain architectures found.

Database: Pfam

There is 1 protein with this architecture (represented by G7YA50):
PF00595 - PF00595 - PF05033 - PF00856

BIOS477/877 L19 - 17

17



Profile HMM search by HMMER

https://www.ecbi.ac.uk/Tools/hmmer/

Quick search = phmmer

Paste in your sequence or use the example

● Reference Proteomes ○ UniProtKB ○ SwissProt ○ Pfam

The HMMER web server: fast and sensitive homology searches. This site has been designed to provide near **interactive searches** for most queries, coupled with **intuitive and interactive results** visualisations.

News February, 2022
HmmerWeb Pfam 35.0 Release
HmmerWeb 2.43 includes support for the most recent Pfam 35.0 and its 19632 protein

Download HMMER
v3.3.2

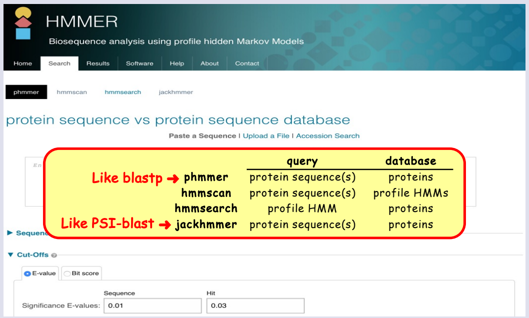
Papers
HMMER web server: 2018 update
S.C. Potter, A. Luciani, S.R. Eddy Y. Park, R. Lopez and R.D. Finn,
Nucleic Acids Research (2018) Web Server Issue

BIOS477/877 L19 - 18

18

Profile HMM search by HMMER

https://www.ecbi.ac.uk/Tools/hmmer/



protein sequence vs protein sequence database

Paste a Sequence | Upload a File | Accession Search

		query	database
Like blastp →	phmmer	protein sequence(s)	proteins
	hmmscan	protein sequence(s)	profile HMMs
	hmmsearch	profile HMM	proteins
Like PSI-blast →	jackhmmer	protein sequence(s)	proteins

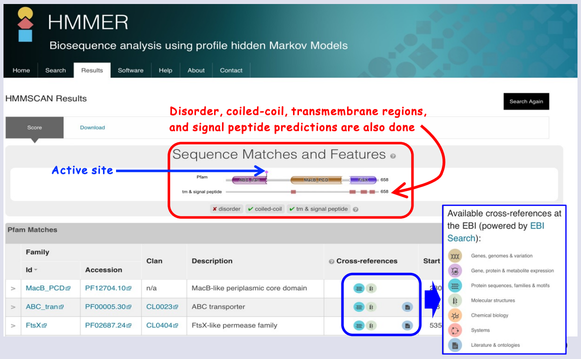
▶ Sequen

▼ Cut-Offs

● E-value ○ Bit score

Significance E-values: Sequence 0.01 Hit 0.03

HMMER hmmscan result



HMMSCAN Results

Disorder, coiled-coil, transmembrane regions, and signal peptide predictions are also done

Sequence Matches and Features

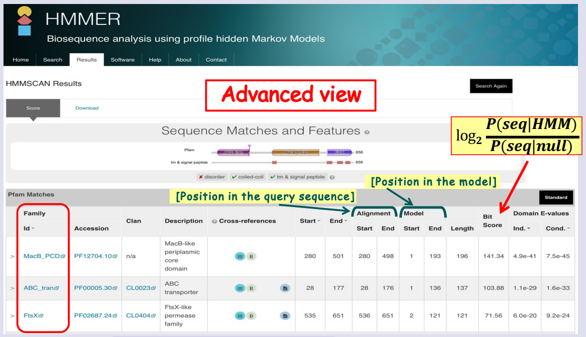
Active site

Available cross-references at the EBI (powered by EBI Search):

Pfam Matches

Family Id	Accession	Clan	Description	Cross-references	Start
MacB_PCD	PF12704.10	n/a	MacB-like periplasmic core domain		
ABC_tran	PF00005.30	CL0023	ABC transporter		
FtsX	PF02687.24	CL0404	FtsX-like permease family		535

HMMER hmmscan resultc



HMMSCAN Results

Advanced view

Sequence Matches and Features

$$\log_2 \frac{P(seq|HMM)}{P(seq|null)}$$

[Position in the model]

[Position in the query sequence]

Pfam Matches

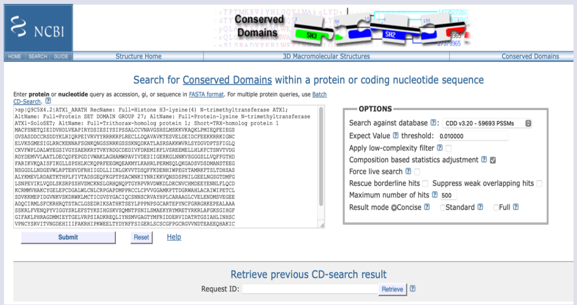
Family Id	Accession	Clan	Description	Cross-references	Start	End	Alignment Start	Alignment End	Model Start	Model End	Length	Bit Score	Domain E-values Ind.	Domain E-values Cond.
MacB_PCD	PF12704.10	n/a	MacB-like periplasmic core domain		280	501	280	498	1	193	196	141.34	4.9e-41	7.5e-45
ABC_tran	PF00005.30	CL0023	ABC transporter		28	177	28	176	1	136	137	103.88	1.1e-29	1.6e-33
FtsX	PF02687.24	CL0404	FtsX-like permease family		535	651	536	651	2	121	121	71.56	6.0e-20	9.2e-24

Domain family

Clan: a collection of families

NCBI Conserved Domains

https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi



Search for Conserved Domains within a protein or coding nucleotide sequence

Enter **protein** or **nucleotide** query as accession, gi, or sequence in FASTA format. For multiple protein queries, use Batch CD-Search.

```
>sp|Q9C5X4.2|ATX1_ARATH RecName: Full=Histone H3-lysine(4) N-trimethyltransferase ATX1;
AltName: Full=Protein SET DOMAIN GROUP 27; AltName: Full=Protein-lysine N-trimethyltransferase
ATX1-SoloSET; AltName: Full=Trithorax-homolog protein 1; Short=TRX-homolog protein 1
MACFSNETQIEIDVHDLVEAPIRYDSIESIYSIPSSALCCVNAVGSHSLMSKKVKAQKLPMIEQFEIEGS
GVSASDDCCRSDDYKLRIQRPEIVRVYYRRRKRPLRECLLDQAVAVKTESVELDEIDCFEEKKRRKIGNC
ELVKSGMESIGLRRCKENNAFSGNKQNGSSRRKGSSSKNQDKATLASRSAKKWVRLSYDGVDPTSFIGLQ
CKVFWPLDALWYEGSIVGYSAERKRYTVKYRDGCDEDIVFDREMIKFLVSREEMELLHLKFCTSNVTVDG
RDYDEMVVLAATLDECQDFEPGDIVWAKLAGHAMWPAVIVDESIIGERKGLNNKVSGGGSLLVQFFGTHD
FARIKVKQAISFIKGLLSPSHLKCKQPRFEEGMQEAKMYLKAHRLPERMSQLQKGADSVDSDMANSTEEG
NSGGDLLNDGEVWLRPTEHVDFRHIIGDLLIINLGKVVTDSQFFKDENHIWPEGYTAMRKFTSLTDHSAS
ALYKMEVLRDAETKTHPLFIVTADSGEQFKGPTPSACWNKIYNRIKKVQNSDSPNILGEELNGSGTDMFG
LSNPEVIKLVQDLSKSRPSSHVSMCKNSLGRHQNQPTGYRPVRVDWKDLDKCNVCHMDEEYENNLFLQCD
KCRMMVHAKCYGELEPCDGALWLCNLCRPGAPDMPPRCCLCPVVGGAMKPTTDGRWAHLACAIWIPETCL
SDVKKMEPIDGVNKVSKDRWKLMCTICGVSYGACIQCSNNSCRVAYHPLCARAAGLCVELENDMSVEGEE
ADQCIRMLSFCKRHRQTSTACLGSEDRIKSATHKTSEYLPPPNPSGCARTEPYNCFGRRGRKEPEALAAA
SSKRLFVENQPYVIGGYSRLEFSTYKSIHGSKVSQMNTPSNILSMAEKYRYMRETYRKRLAFGKSGIHGF
GIFAKLPHRAGDMMIEYTGELVRPSIADKREQLIYNSMVGAGTYMFRIDDERVIDATRTGSIAHLINHSC
VPNCYSRVITVNGDEHIIIFAKRHIPKWEELTYDYRFFSIGERLSCSCGFPGCRGVVNDTEAEEQHAKIC
```

OPTIONS

Search against database: CDD v3.20 - 59693 PSSMs

Expect Value threshold: 0.010000

Apply low-complexity filter □

Composition based statistics adjustment ☑

Force live search □

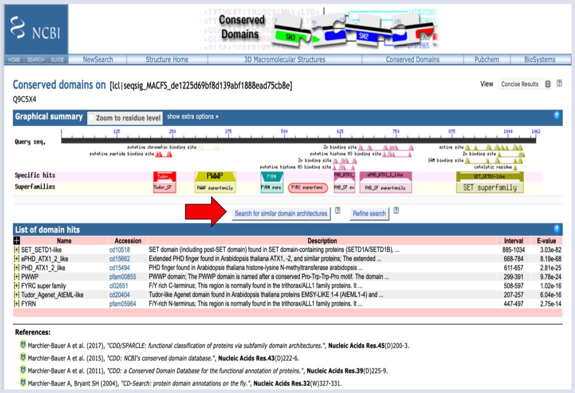
Rescue borderline hits □ Suppress weak overlapping hits □

Maximum number of hits 500

Result mode ● Concise ○ Standard ○ Full

Retrieve previous CD-search result

Request ID:



Conserved domains on [lcl|seqsig_MACFS_de1225d69bf8d139abf1888ead75cb8e]

Q9C5X4

Graphical summary

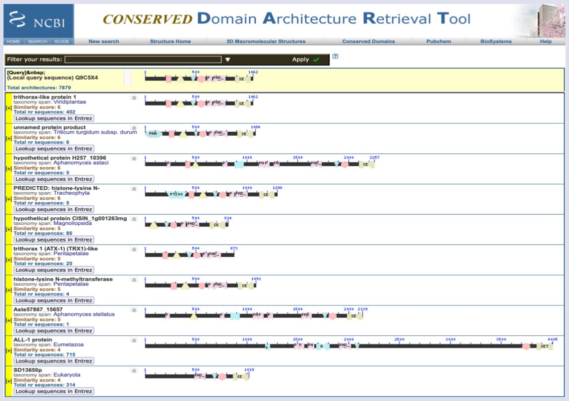
List of domain hits

Name	Accession	Description	Interval	E-value
[+] SET_SETD1-like	cd10518	SET domain (including post-SET domain) found in SET domain-containing proteins (SETD1A/SETD1B), ...	885-1034	3.03e-82
[+] ePHD_ATX1_2_like	cd15662	Extended PHD finger found in Arabidopsis thaliana ATX1, -2, and similar proteins; The extended ...	668-784	8.19e-68
[+] PHD_ATX1_2_like	cd15494	PHD finger found in Arabidopsis thaliana histone-lysine N-methyltransferase arabidopsis ...	611-657	2.81e-25
[+] PWWP	pfam00855	PWWP domain; The PWWP domain is named after a conserved Pro-Trp-Trp-Pro motif. The domain ...	299-391	9.78e-24
[+] FYRC super family	cl02651	F/Y rich C-terminus; This region is normally found in the trithorax/ALL1 family proteins. It ...	508-597	1.02e-16
[+] Tudor_Agenet_AtEML-like	cd20404	Tudor-like Agenet domain found in Arabidopsis thaliana proteins EMSY-LIKE 1-4 (AtEML1-4) and ...	207-257	6.04e-16
[+] FYRN	pfam05964	F/Y-rich N-terminus; This region is normally found in the trithorax/ALL1 family proteins. It ...	447-497	2.75e-14

References:

- Marchler-Bauer A et al. (2017), *"CDD/SPARCLE: functional classification of proteins via subfamily domain architectures."*, **Nucleic Acids Res.45**(D)200-3.
- Marchler-Bauer A et al. (2015), *"CDD: NCBI's conserved domain database."*, **Nucleic Acids Res.43**(D)222-6.
- Marchler-Bauer A et al. (2011), *"CDD: a Conserved Domain Database for the functional annotation of proteins."*, **Nucleic Acids Res.39**(D)225-9.
- Marchler-Bauer A, Bryant SH (2004), *"CD-Search: protein domain annotations on the fly."*, **Nucleic Acids Res.32**(W)327-331.

https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi



CONSERVED Domain Architecture Retrieval Tool

Filter your results:

[Query]
(Local query sequence) Q9C5X4
Total architectures: 7879

[+] trithorax-like protein 1
taxonomy span: Viridiplantae
Similarity score: 6
Total nr sequences: 402
Lookup sequences in Entrez

[+] unnamed protein product
taxonomy span: Triticum turgidum subsp. durum
Similarity score: 6
Total nr sequences: 6
Lookup sequences in Entrez

[+] hypothetical protein H257_10396
taxonomy span: Aphanomyces astaci
Similarity score: 6
Total nr sequences: 5
Lookup sequences in Entrez

[+] PREDICTED: histone-lysine N-
taxonomy span: Tracheophyta
Similarity score: 6
Total nr sequences: 5
Lookup sequences in Entrez

[+] hypothetical protein CISIN_1g001263mg
taxonomy span: Magnoliopsida
Similarity score: 5
Total nr sequences: 86
Lookup sequences in Entrez

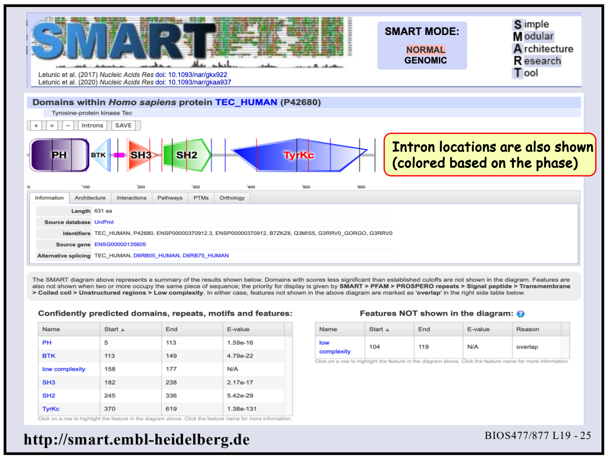
[+] trithorax 1 (ATX-1) (TRX1)-like
taxonomy span: Pentapetalae
Similarity score: 5
Total nr sequences: 20
Lookup sequences in Entrez

[+] histone-lysine N-methyltransferase
taxonomy span: Pentapetalae
Similarity score: 5
Total nr sequences: 4
Lookup sequences in Entrez

[+] Aste57867_15657
taxonomy span: Aphanomyces stellatus
Similarity score: 5
Total nr sequences: 1
Lookup sequences in Entrez

[+] ALL-1 protein
taxonomy span: Eumetazoa
Similarity score: 4
Total nr sequences: 715
Lookup sequences in Entrez

[+] SD13650p
taxonomy span: Eukaryota
Similarity score: 4
Total nr sequences: 314
Lookup sequences in Entrez



Letunic et al. (2017) *Nucleic Acids Res* doi: 10.1093/nar/gkx922
Letunic et al. (2020) *Nucleic Acids Res* doi: 10.1093/nar/gkaa937

Domains within *Homo sapiens* protein TEC_HUMAN (P42680)

Tyrosine-protein kinase Tec

Intron locations are also shown (colored based on the phase)

Length	631 aa
Source database	UniProt
Identifiers	TEC_HUMAN, P42680, ENSP00000370912.3, ENSP00000370912, B7ZKZ6, Q3MIS5, G3RRV0_GORGO, G3RRV0
Source gene	ENSG00000135605
Alternative splicing	TEC_HUMAN, D6RB05_HUMAN, D6RB75_HUMAN

The SMART diagram above represents a summary of the results shown below. Domains with scores less significant than established cutoffs are not shown in the diagram. Features are also not shown when two or more occupy the same piece of sequence; the priority for display is given by **SMART > PFAM > PROSPERO repeats > Signal peptide > Transmembrane > Coiled coil > Unstructured regions > Low complexity**. In either case, features not shown in the above diagram are marked as **'overlap'** in the right side table below.

Confidently predicted domains, repeats, motifs and features:

Name	Start ▲	End	E-value
PH	5	113	1.59e-16
BTK	113	149	4.79e-22
low complexity	158	177	N/A
SH3	182	238	2.17e-17
SH2	245	336	5.42e-29
TyrKc	370	619	1.38e-131

Click on a row to highlight the feature in the diagram above. Click the feature name for more information.

Features NOT shown in the diagram: ?

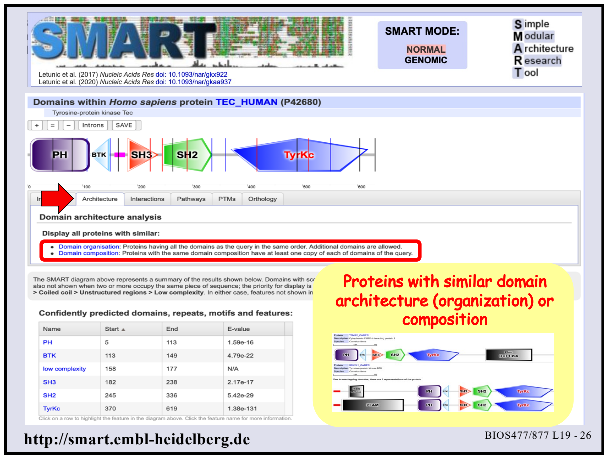
Name	Start ▲	End	E-value	Reason
low complexity	104	119	N/A	overlap

Click on a row to highlight the feature in the diagram above. Click the feature name for more information.

http://smart.embl-heidelberg.de

BIOS477/877 L19 - 25

25



Letunic et al. (2017) *Nucleic Acids Res* doi: 10.1093/nar/gkx922
Letunic et al. (2020) *Nucleic Acids Res* doi: 10.1093/nar/gkaa937

Domains within *Homo sapiens* protein TEC_HUMAN (P42680)

Tyrosine-protein kinase Tec

Domain architecture analysis

Display all proteins with similar:

- Domain organisation: Proteins having all the domains as the query in the same order. Additional domains are allowed.
- Domain composition: Proteins with the same domain composition have at least one copy of each of domains of the query.

The SMART diagram above represents a summary of the results shown below. Domains with sco also not shown when two or more occupy the same piece of sequence; the priority for display is > **Coiled coil > Unstructured regions > Low complexity**. In either case, features not shown in

Proteins with similar domain architecture (organization) or composition

Confidently predicted domains, repeats, motifs and features:

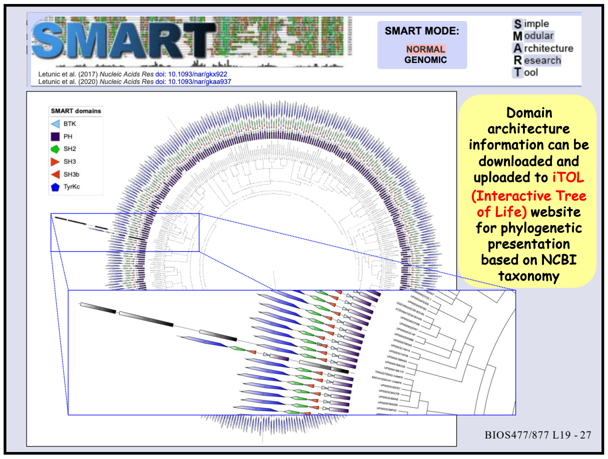
Name	Start ▲	End	E-value
PH	5	113	1.59e-16
BTK	113	149	4.79e-22
low complexity	158	177	N/A
SH3	182	238	2.17e-17
SH2	245	336	5.42e-29
TyrKc	370	619	1.38e-131

Click on a row to highlight the feature in the diagram above. Click the feature name for more information.

http://smart.embl-heidelberg.de

BIOS477/877 L19 - 26

26

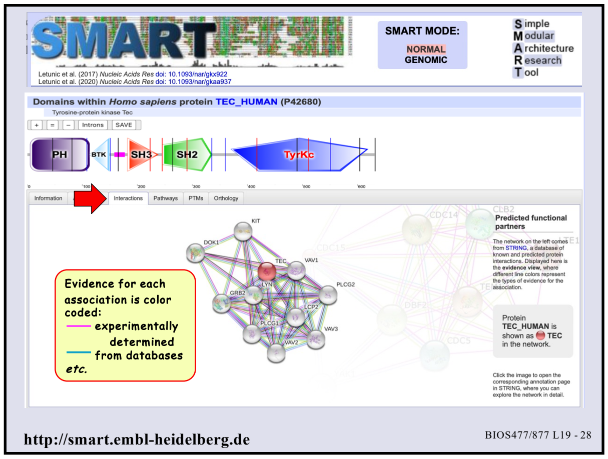


Letunic et al. (2017) *Nucleic Acids Res* doi: 10.1093/nar/gkx922
Letunic et al. (2020) *Nucleic Acids Res* doi: 10.1093/nar/gkaa937

Domain architecture information can be downloaded and uploaded to iTOL (Interactive Tree of Life) website for phylogenetic presentation based on NCBI taxonomy

BIOS477/877 L19 - 27

27



Letunic et al. (2017) *Nucleic Acids Res* doi: 10.1093/nar/gkx922
Letunic et al. (2020) *Nucleic Acids Res* doi: 10.1093/nar/gkaa937

Domains within *Homo sapiens* protein TEC_HUMAN (P42680)

Tyrosine-protein kinase Tec

Predicted functional partners

The network on the left comes from STRING, a database of known and predicted protein interactions. Displayed here is the **evidence view**, where different line colors represent the types of evidence for the association.

Protein **TEC_HUMAN** is shown as ● **TEC** in the network.

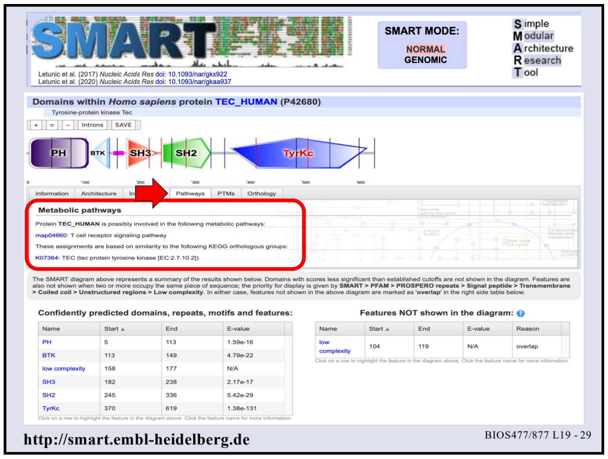
Click the image to open the corresponding annotation page in STRING, where you can explore the network in detail.

Evidence for each association is color coded:
— experimentally determined
— from databases
etc.

http://smart.embl-heidelberg.de

BIOS477/877 L19 - 28

28



Letunic et al. (2017) *Nucleic Acids Res* doi: 10.1093/nar/gkx922
Letunic et al. (2020) *Nucleic Acids Res* doi: 10.1093/nar/gkaa937

Domains within *Homo sapiens* protein TEC_HUMAN (P42680)

Tyrosine-protein kinase Tec

Metabolic pathways

Protein **TEC_HUMAN** is possibly involved in the following metabolic pathways:

map04660: T cell receptor signaling pathway

These assignments are based on similarity to the following KEGG orthologous groups:

K07364: TEC (tec protein tyrosine kinase [EC:2.7.10.2])

The SMART diagram above represents a summary of the results shown below. Domains with scores less significant than established cutoffs are not shown in the diagram. Features are also not shown when two or more occupy the same piece of sequence; the priority for display is given by **SMART > PFAM > PROSPERO repeats > Signal peptide > Transmembrane > Coiled coil > Unstructured regions > Low complexity**. In either case, features not shown in the above diagram are marked as **'overlap'** in the right side table below.

Confidently predicted domains, repeats, motifs and features:

Name	Start ▲	End	E-value
PH	5	113	1.59e-16
BTK	113	149	4.79e-22
low complexity	158	177	N/A
SH3	182	238	2.17e-17
SH2	245	336	5.42e-29
TyrKc	370	619	1.38e-131

Click on a row to highlight the feature in the diagram above. Click the feature name for more information.

Features NOT shown in the diagram: ?

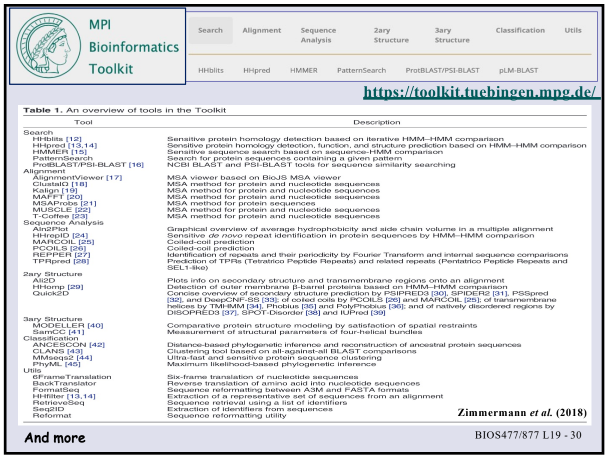
Name	Start ▲	End	E-value	Reason
low complexity	104	119	N/A	overlap

Click on a row to highlight the feature in the diagram above. Click the feature name for more information.

http://smart.embl-heidelberg.de

BIOS477/877 L19 - 29

29



https://toolkit.tuebingen.mpg.de/

Table 1. An overview of tools in the Toolkit

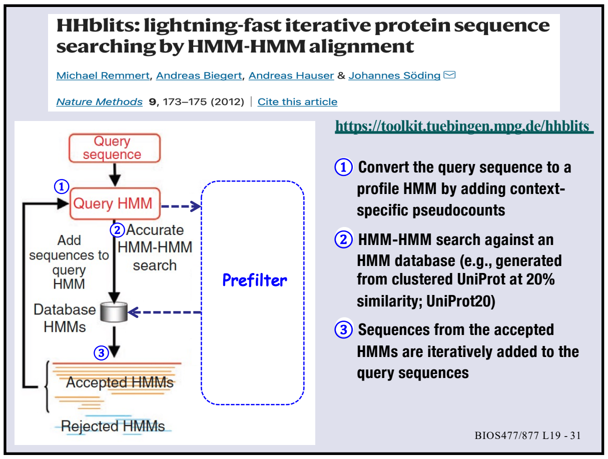
Tool	Description
Search	
HHblits [12]	Sensitive protein homology detection based on iterative HMM–HMM comparison
HHpred [13,14]	Sensitive protein homology detection, function, and structure prediction based on HMM–HMM comparison
HMMER [15]	Sensitive sequence search based on sequence-HMM comparison
PatternSearch	Search for protein sequences containing a given pattern
ProtBLAST/PSI-BLAST [16]	NCBI BLAST and PSI-BLAST tools for sequence similarity searching
Alignment	
AlignmentViewer [17]	MSA viewer based on BioJS MSA viewer
ClustalΩ [18]	MSA method for protein and nucleotide sequences
Kalign [19]	MSA method for protein and nucleotide sequences
MAFFT [20]	MSA method for protein and nucleotide sequences
MSAProbs [21]	MSA method for protein sequences
MUSCLE [22]	MSA method for protein and nucleotide sequences
T-Coffee [23]	MSA method for protein and nucleotide sequences
Sequence Analysis	
Aln2Plot	Graphical overview of average hydrophobicity and side chain volume in a multiple alignment
HHrepID [24]	Sensitive *de novo* repeat identification in protein sequences by HMM–HMM comparison
MARCOIL [25]	Coiled-coil prediction
PCOILS [26]	Coiled-coil prediction
REPPER [27]	Identification of repeats and their periodicity by Fourier Transform and internal sequence comparisons
TPRpred [28]	Prediction of TPRs (Tetratrico Peptide Repeats) and related repeats (Pentatrico Peptide Repeats and SEL1-like)
2ary Structure	
Ali2D	Plots info on secondary structure and transmembrane regions onto an alignment
HHomp [29]	Detection of outer membrane β-barrel proteins based on HMM–HMM comparison
Quick2D	Concise overview of secondary structure prediction by PSIPRED3 [30], SPIDER2 [31], PSSpred [32], and DeepCNF-SS [33]; of coiled coils by PCOILS [26] and MARCOIL [25]; of transmembrane helices by TMHMM [34], Phobius [35] and PolyPhobius [36]; and of natively disordered regions by DISOPRED3 [37], SPOT-Disorder [38] and IUPred [39]
3ary Structure	
MODELLER [40]	Comparative protein structure modeling by satisfaction of spatial restraints
SamCC [41]	Measurement of structural parameters of four-helical bundles
Classification	
ANCESCON [42]	Distance-based phylogenetic inference and reconstruction of ancestral protein sequences
CLANS [43]	Clustering tool based on all-against-all BLAST comparisons
MMseqs2 [44]	Ultra-fast and sensitive protein sequence clustering
PhyML [45]	Maximum likelihood-based phylogenetic inference
Utils	
6FrameTranslation	Six-frame translation of nucleotide sequences
BackTranslator	Reverse translation of amino acid into nucleotide sequences
FormatSeq	Sequence reformatting between A3M and FASTA formats
HHfilter [13,14]	Extraction of a representative set of sequences from an alignment
RetrieveSeq	Sequence retrieval using a list of identifiers
Seq2ID	Extraction of identifiers from sequences
Reformat	Sequence reformatting utility

Zimmermann *et al.* (2018)

And more

BIOS477/877 L19 - 30

30



HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment

Michael Remmert, Andreas Biegert, Andreas Hauser & Johannes Söding

Nature Methods **9**, 173–175 (2012) | Cite this article

https://toolkit.tuebingen.mpg.de/hhblits

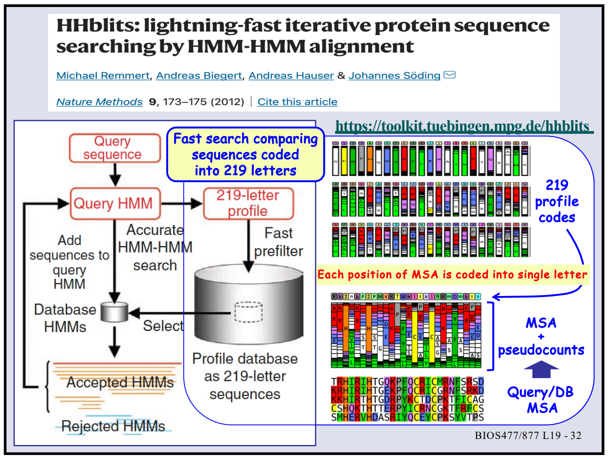
① Convert the query sequence to a profile HMM by adding context-specific pseudocounts

② HMM-HMM search against an HMM database (e.g., generated from clustered UniProt at 20% similarity; UniProt20)

③ Sequences from the accepted HMMs are iteratively added to the query sequences

BIOS477/877 L19 - 31

31



HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment

Michael Remmert, Andreas Biegert, Andreas Hauser & Johannes Söding

Nature Methods **9**, 173–175 (2012) | Cite this article

https://toolkit.tuebingen.mpg.de/hhblits

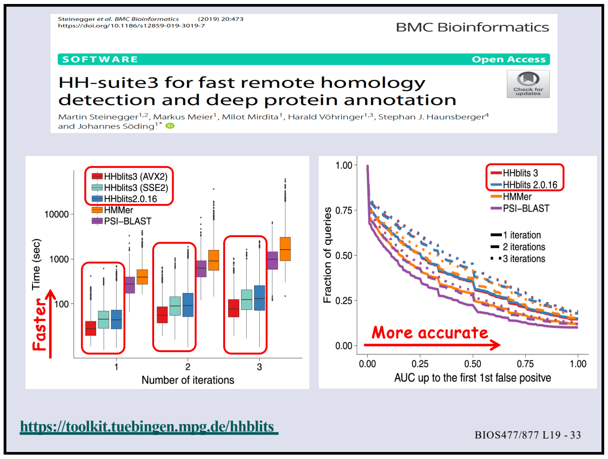
Query sequence
Query HMM
Add sequences to query HMM
Accurate HMM-HMM search
Database HMMs
Select
Accepted HMMs
Rejected HMMs
219-letter profile
Fast prefilter
Profile database as 219-letter sequences
219 profile codes
MSA + pseudocounts
Query/DB MSA
TRHIRIHTGQKPFQCRICMRNFSRSD
KRHIRIHTGEKPFQCRICGRNFSRKD
KKHIRTHTGDRPYKCTDCPKTFICAG
CSHQKTHTTERPYICRNCGKTFRFCS
SMHERVHDASRIYQCEYCPKSYVTPS

Fast search comparing sequences coded into 219 letters

Each position of MSA is coded into single letter

BIOS477/877 L19 - 32

32



Steinegger *et al. BMC Bioinformatics* (2019) 20:473
https://doi.org/10.1186/s12859-019-3019-7

BMC Bioinformatics

SOFTWARE — Open Access

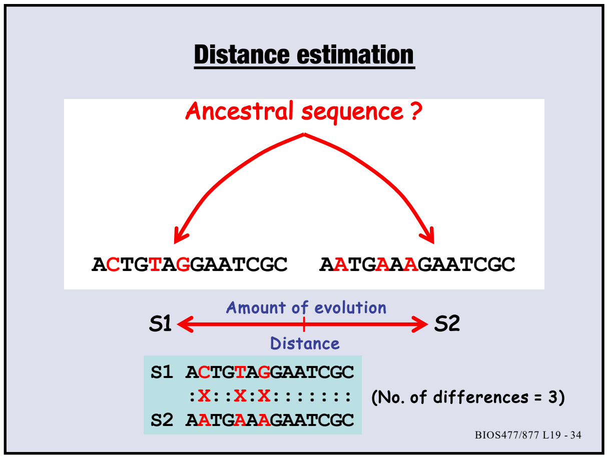
HH-suite3 for fast remote homology detection and deep protein annotation

Martin Steinegger[1,2], Markus Meier[1], Milot Mirdita[1], Harald Vöhringer[1,3], Stephan J. Haunsberger[4] and Johannes Söding[1*]

https://toolkit.tuebingen.mpg.de/hhblits

BIOS477/877 L19 - 33

33

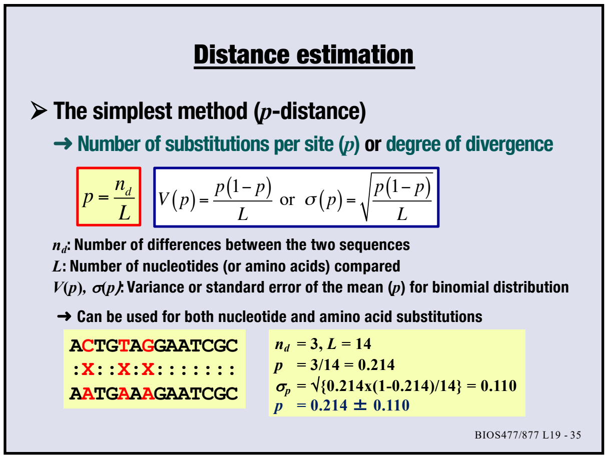


Distance estimation

```
S1  ACTGTAGGAATCGC
     :X::X:X:::::::   (No. of differences = 3)
S2  AATGAAAGAATCGC
```

BIOS477/877 L19 - 34

34



Distance estimation

➢ The simplest method (p-distance)

→ Number of substitutions per site (p) or degree of divergence

$$\boxed{p = \frac{n_d}{L}} \quad \boxed{V(p) = \frac{p(1-p)}{L} \text{ or } \sigma(p) = \sqrt{\frac{p(1-p)}{L}}}$$

n_d: Number of differences between the two sequences

L: Number of nucleotides (or amino acids) compared

$V(p)$, $\sigma(p)$: Variance or standard error of the mean (p) for binomial distribution

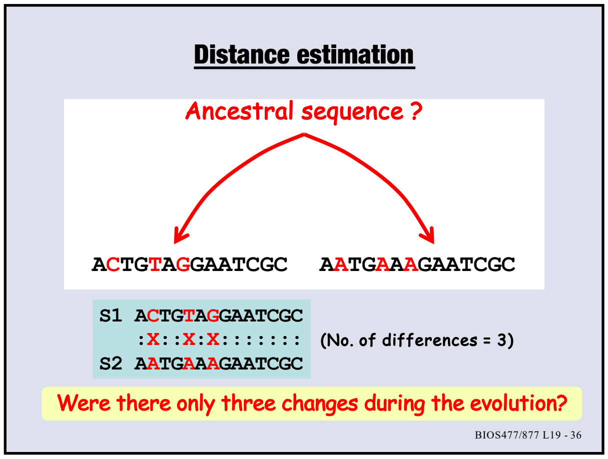
→ Can be used for both nucleotide and amino acid substitutions

```
ACTGTAGGAATCGC
:X::X:X:::::::
AATGAAAGAATCGC
```

n_d = 3, L = 14
p = 3/14 = 0.214
σ_p = √{0.214x(1-0.214)/14} = 0.110
p = 0.214 ± 0.110

BIOS477/877 L19 - 35

35



Distance estimation

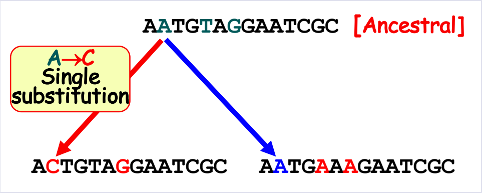
```
S1  ACTGTAGGAATCGC
     :X::X:X:::::::   (No. of differences = 3)
S2  AATGAAAGAATCGC
```

Were there only three changes during the evolution?

BIOS477/877 L19 - 36

36

Distance estimation



AATGTAGGAATCGC [Ancestral]

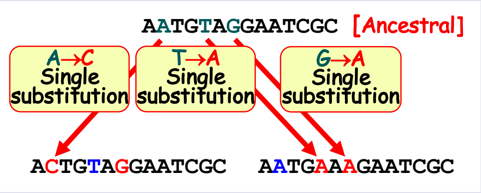
A→C Single substitution

ACTGTAGGAATCGC AATGAAAGAATCGC

```
ACTGTAGGAATCGC
:X::X:X:::::::
AATGAAAGAATCGC
```

37

Distance estimation



AATGTAGGAATCGC [Ancestral]

A→C Single substitution

T→A Single substitution

G→A Single substitution

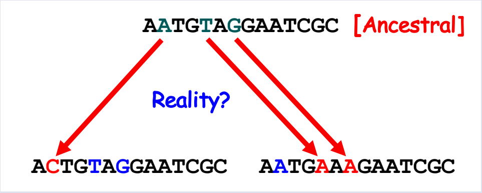
ACTGTAGGAATCGC AATGAAAGAATCGC

```
ACTGTAGGAATCGC
:X::X:X:::::::
AATGAAAGAATCGC
```

3 substitutions: All substitutions found

38

Distance estimation



AATGTAGGAATCGC [Ancestral]

Reality?

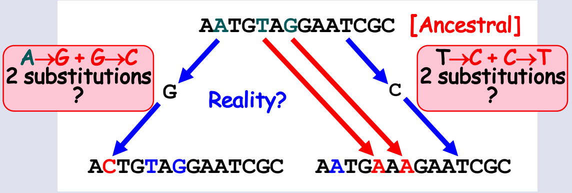
ACTGTAGGAATCGC AATGAAAGAATCGC

```
ACTGTAGGAATCGC
:X::X:X:::::::
AATGAAAGAATCGC
```

Observed number of differences = 3

39

Distance estimation



AATGTAGGAATCGC [Ancestral]

A→G + G→C 2 substitutions ?

G

Reality?

C

T→C + C→T 2 substitutions ?

ACTGTAGGAATCGC AATGAAAGAATCGC

```
ACTGTAGGAATCGC
:X::X:X:::::::
AATGAAAGAATCGC
```

Observed number of differences = 3

^

Actual number of substitutions = 6

[3 hidden substitutions]

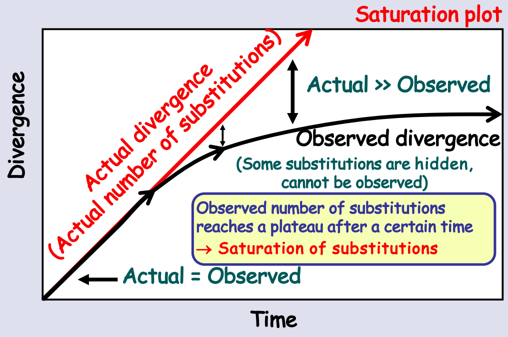
Each site has only a limited number of possibilities (4 nucleotides or 20 amino acids)

More hidden substitutions can happen with longer evolution

40

Effect of multiple substitutions

Saturation plot



Observed number of substitutions reaches a plateau after a certain time

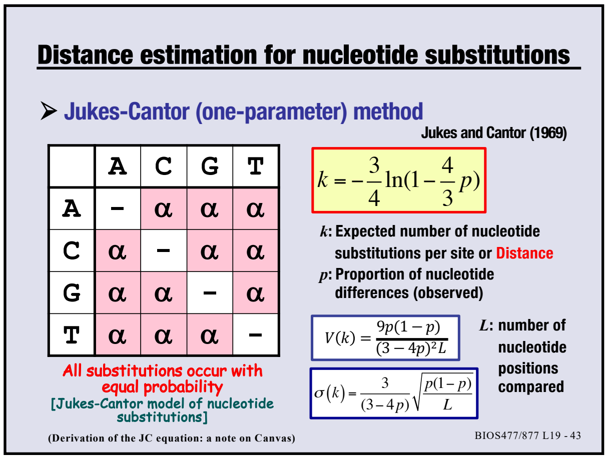
→ Saturation of substitutions

41

Effect of multiple substitutions

- When the degree of divergence between two sequences is small,
  - ➜ the chance of having more than one substitution at any site is negligible
  - ➜ Observed divergence ≈ actual divergence
- When the degree of divergence becomes larger,
  - ➜ more than one substitution could happen at any site [multiple substitutions or multiple hits]
  - ➜ Observed divergence << actual divergence [Saturation effect]
- Effect of multiple hits is larger for nucleotide substitutions
- Methods to uncover the number of hidden substitutions need to be used [Multiple hit correction]
  - ➜ Actual divergence level is estimated based on the observed degree of divergence

42



Distance estimation for nucleotide substitutions

➢ Jukes-Cantor (one-parameter) method

Jukes and Cantor (1969)

	A	C	G	T
A	–	α	α	α
C	α	–	α	α
G	α	α	–	α
T	α	α	α	–

All substitutions occur with equal probability

[Jukes-Cantor model of nucleotide substitutions]

$$k = -\frac{3}{4}\ln(1-\frac{4}{3}p)$$

k: Expected number of nucleotide substitutions per site or Distance

p: Proportion of nucleotide differences (observed)

$$V(k) = \frac{9p(1-p)}{(3-4p)^2 L}$$

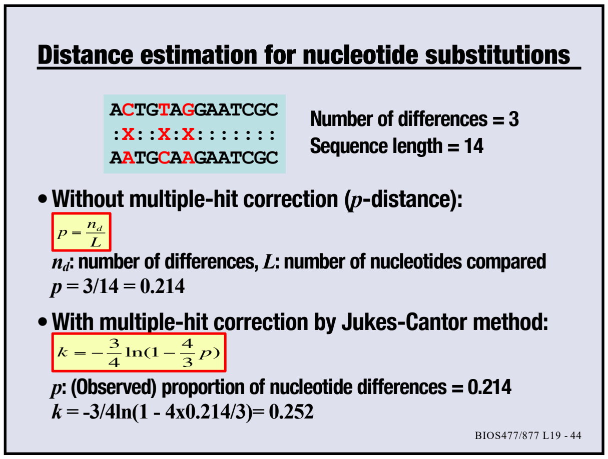
$$\sigma(k) = \frac{3}{(3-4p)}\sqrt{\frac{p(1-p)}{L}}$$

L: number of nucleotide positions compared

(Derivation of the JC equation: a note on Canvas)

BIOS477/877 L19 - 43

43



Distance estimation for nucleotide substitutions

```
ACTGTAGGAATCGC
:X::X:X:::::::
AATGCAAGAATCGC
```

Number of differences = 3

Sequence length = 14

- Without multiple-hit correction (p-distance):

$$p = \frac{n_d}{L}$$

n_d: number of differences, L: number of nucleotides compared

$p = 3/14 = 0.214$

- With multiple-hit correction by Jukes-Cantor method:

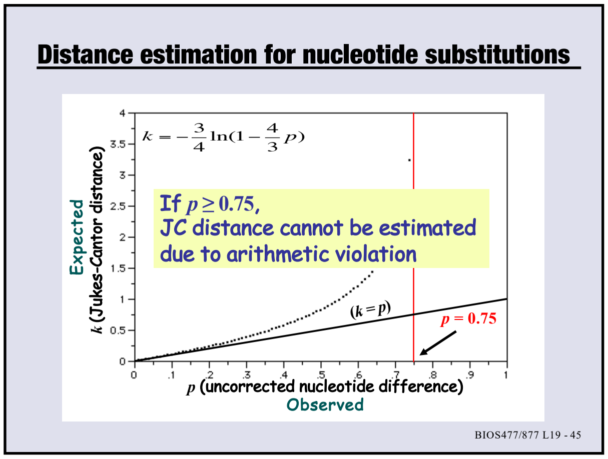
$$k = -\frac{3}{4}\ln(1-\frac{4}{3}p)$$

p: (Observed) proportion of nucleotide differences = 0.214

k = -3/4ln(1 - 4x0.214/3)= 0.252

BIOS477/877 L19 - 44

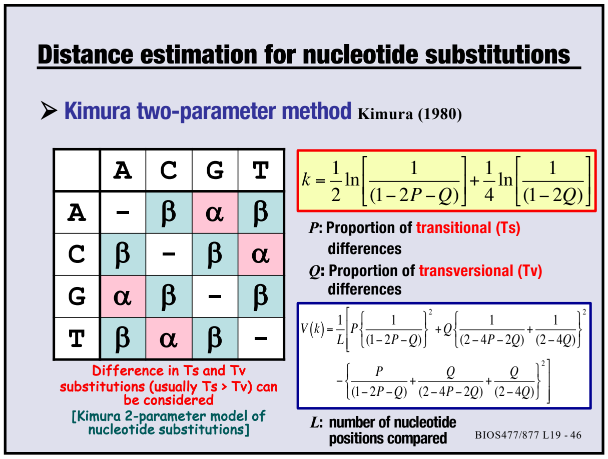
44



Distance estimation for nucleotide substitutions

BIOS477/877 L19 - 45

45



Distance estimation for nucleotide substitutions

➢ Kimura two-parameter method Kimura (1980)

	A	C	G	T
A	–	β	α	β
C	β	–	β	α
G	α	β	–	β
T	β	α	β	–

Difference in Ts and Tv substitutions (usually Ts > Tv) can be considered

[Kimura 2-parameter model of nucleotide substitutions]

$$k = \frac{1}{2}\ln\left[\frac{1}{(1-2P-Q)}\right] + \frac{1}{4}\ln\left[\frac{1}{(1-2Q)}\right]$$

P: Proportion of transitional (Ts) differences

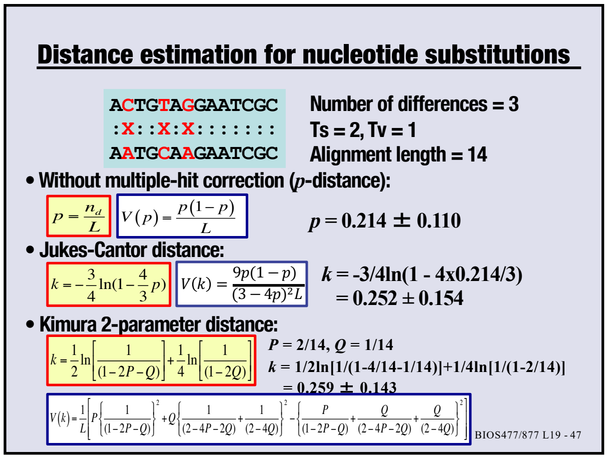
Q: Proportion of transversional (Tv) differences

$$V(k) = \frac{1}{L}\left[P\left\{\frac{1}{(1-2P-Q)}\right\}^2 + Q\left\{\frac{1}{(2-4P-2Q)} + \frac{1}{(2-4Q)}\right\}^2 - \left\{\frac{P}{(1-2P-Q)} + \frac{Q}{(2-4P-2Q)} + \frac{Q}{(2-4Q)}\right\}^2\right]$$

L: number of nucleotide positions compared

BIOS477/877 L19 - 46

46



Distance estimation for nucleotide substitutions

```
ACTGTAGGAATCGC
:X::X:X:::::::
AATGCAAGAATCGC
```

Number of differences = 3

Ts = 2, Tv = 1

Alignment length = 14

- Without multiple-hit correction (p-distance):

$$p = \frac{n_d}{L} \qquad V(p) = \frac{p(1-p)}{L}$$

$p = 0.214 \pm 0.110$

- Jukes-Cantor distance:

$$k = -\frac{3}{4}\ln(1-\frac{4}{3}p) \qquad V(k) = \frac{9p(1-p)}{(3-4p)^2 L}$$

k = -3/4ln(1 - 4x0.214/3)

= 0.252 ± 0.154

- Kimura 2-parameter distance:

$$k = \frac{1}{2}\ln\left[\frac{1}{(1-2P-Q)}\right] + \frac{1}{4}\ln\left[\frac{1}{(1-2Q)}\right]$$

P = 2/14, Q = 1/14

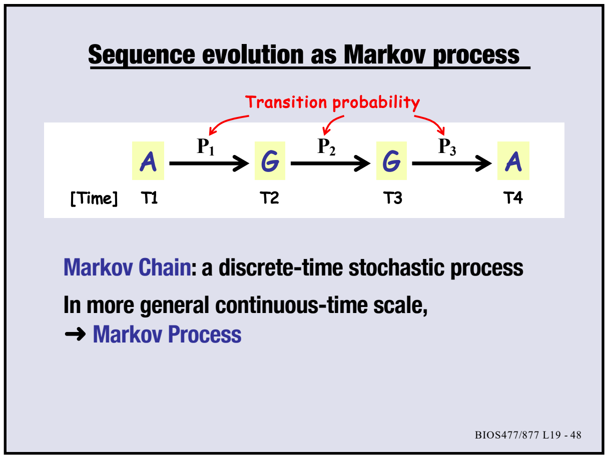
k = 1/2ln[1/(1-4/14-1/14)]+1/4ln[1/(1-2/14)]

= 0.259 ± 0.143

$$V(k) = \frac{1}{L}\left[P\left\{\frac{1}{(1-2P-Q)}\right\}^2 + Q\left\{\frac{1}{(2-4P-2Q)} + \frac{1}{(2-4Q)}\right\}^2 - \left\{\frac{P}{(1-2P-Q)} + \frac{Q}{(2-4P-2Q)} + \frac{Q}{(2-4Q)}\right\}^2\right]$$

BIOS477/877 L19 - 47

47



Sequence evolution as Markov process

$$\text{[Time]}\quad \underset{T1}{A} \xrightarrow{P_1} \underset{T2}{G} \xrightarrow{P_2} \underset{T3}{G} \xrightarrow{P_3} \underset{T4}{A}$$

Transition probability

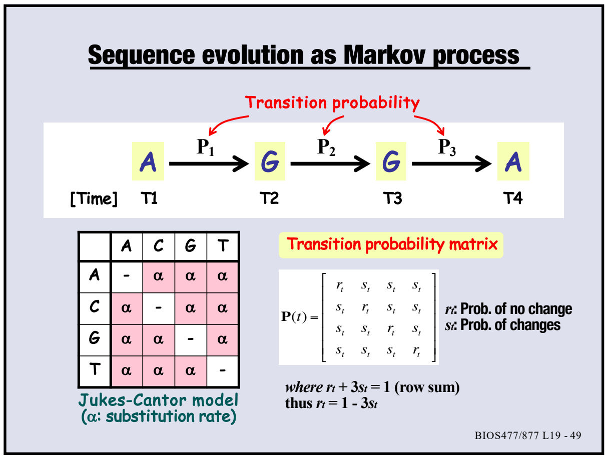
Markov Chain: a discrete-time stochastic process

In more general continuous-time scale,

➔ Markov Process

BIOS477/877 L19 - 48

48



Sequence evolution as Markov process

Transition probability

$$A \xrightarrow{P_1} G \xrightarrow{P_2} G \xrightarrow{P_3} A$$

[Time] T1 T2 T3 T4

	A	C	G	T
A	-	α	α	α
C	α	-	α	α
G	α	α	-	α
T	α	α	α	-

Jukes-Cantor model
(α: substitution rate)

Transition probability matrix

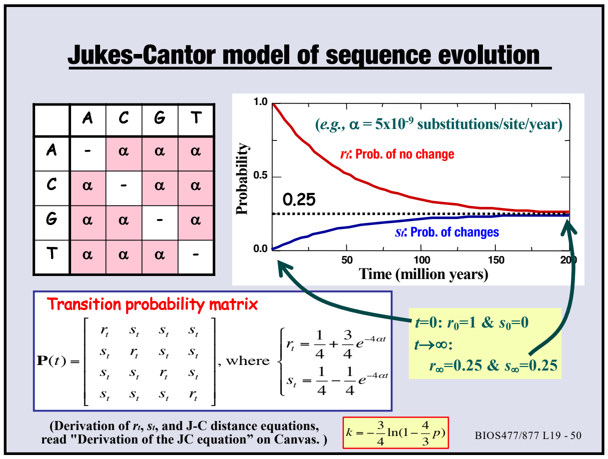
$$\mathbf{P}(t) = \begin{bmatrix} r_t & s_t & s_t & s_t \\ s_t & r_t & s_t & s_t \\ s_t & s_t & r_t & s_t \\ s_t & s_t & s_t & r_t \end{bmatrix}$$

r_t: Prob. of no change
s_t: Prob. of changes

where $r_t + 3s_t = 1$ (row sum)
thus $r_t = 1 - 3s_t$

BIOS477/877 L19 - 49

49



Jukes-Cantor model of sequence evolution

	A	C	G	T
A	-	α	α	α
C	α	-	α	α
G	α	α	-	α
T	α	α	α	-

(*e.g.*, α = 5x10^{-9} substitutions/site/year)

r_t: Prob. of no change

s_t: Prob. of changes

Transition probability matrix

$$\mathbf{P}(t) = \begin{bmatrix} r_t & s_t & s_t & s_t \\ s_t & r_t & s_t & s_t \\ s_t & s_t & r_t & s_t \\ s_t & s_t & s_t & r_t \end{bmatrix}, \text{ where } \begin{cases} r_t = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t} \\ s_t = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t} \end{cases}$$

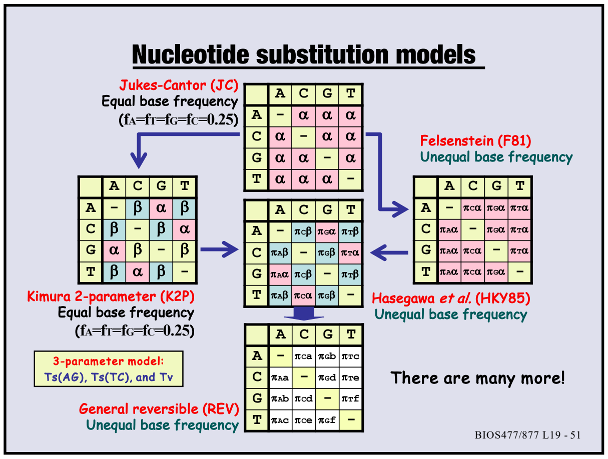
t=0: r_0=1 & s_0=0
$t \rightarrow \infty$:
r_∞=0.25 & s_∞=0.25

(Derivation of r_t, s_t, and J-C distance equations,
read "Derivation of the JC equation" on Canvas.)

$$k = -\frac{3}{4}\ln(1 - \frac{4}{3}p)$$

BIOS477/877 L19 - 50

50



Nucleotide substitution models

Jukes-Cantor (JC)
Equal base frequency
(f_A=f_T=f_G=f_C=0.25)

	A	C	G	T
A	–	α	α	α
C	α	–	α	α
G	α	α	–	α
T	α	α	α	–

Felsenstein (F81)
Unequal base frequency

	A	C	G	T
A	–	πCα	πGα	πTα
C	πAα	–	πGα	πTα
G	πAα	πCα	–	πTα
T	πAα	πCα	πGα	–

Kimura 2-parameter (K2P)
Equal base frequency
(f_A=f_T=f_G=f_C=0.25)

	A	C	G	T
A	–	β	α	β
C	β	–	β	α
G	α	β	–	β
T	β	α	β	–

3-parameter model:
Ts(AG), Ts(TC), and Tv

	A	C	G	T
A	–	πCβ	πGα	πTβ
C	πAβ	–	πGβ	πTα
G	πAα	πCβ	–	πTβ
T	πAβ	πCα	πGβ	–

Hasegawa *et al.* (HKY85)
Unequal base frequency

	A	C	G	T
A	–	πCa	πGb	πTC
C	πAa	–	πGd	πTe
G	πAb	πCd	–	πTf
T	πAC	πCe	πGf	–

General reversible (REV)
Unequal base frequency

There are many more!

BIOS477/877 L19 - 51

51