NIBB GITC「BLAST 自由自在」
September 6 -7, 2018@Okazaki

Beyond BLAST

Shuji Shigenobu /重信秀治

Aim
- BLAST以外の配列解析の手法を概観する。
- Motif search をマスターする。
- 遺伝子アノテーションとGene Ontologyについて理解する。

Beyond BLAST

- Other sequence analysis tools
- Motif search
- Gene Ontology

Advanced BLAST search tools

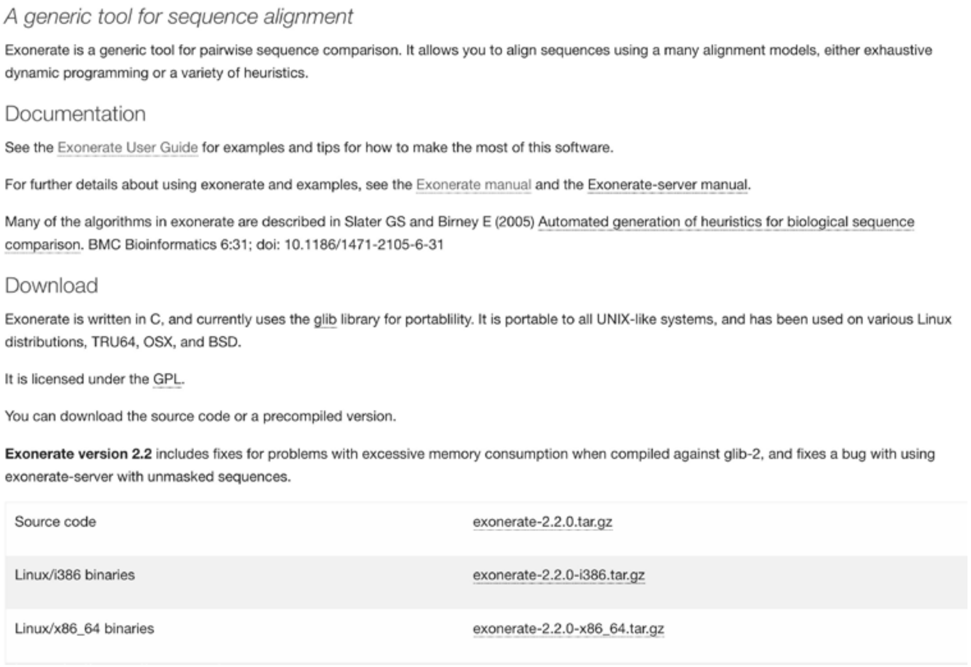
- PSI-BLAST: Position Specific Iterative BLAST
  - Automatically generates a position specific score matrix (PSSM)
  - PSI-BLAST finds sequences significantly similar to the query in a database search and uses the resulting alignments to build a PSSM for the query. With this PSSM the database is scanned again to eventually pull in more significant hits, and further refine the scoring model.
  - More sensitive than standard BLAST
- RPS-BLAST: Reverse Position-Specific BLAST
  - RPS-BLAST uses the query sequence to search a database of pre-calculated PSSMs, and report significant hits in a single pass.
  - Used in CD-search (Conserved Domain search) at NCBI website.
- DELTA-BLAST
  - DELTA-BLAST searches a protein sequence database using a PSSM constructed from conserved domains matching a query. It first searches the NCBI CDD database to construct the PSSM.

Sequence analysis tools for specific purposes

Splicing-aware alignment	exonerate
NGS	
short read	bowtie2, bwa, hisat2
long read	blasr, minimap2
large genome	lastz, last
Multiple alignment	clustal omega, muscle, mafft, PRANK

Exonerate

- Slater GS and Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31
  - https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate



A generic tool for sequence alignment

Exonerate is a generic tool for pairwise sequence comparison. It allows you to align sequences using a many alignment models, either exhaustive dynamic programming or a variety of heuristics.

Documentation

See the Exonerate User Guide for examples and tips for how to make the most of this software.

For further details about using exonerate and examples, see the Exonerate manual and the Exonerate-server manual.

Many of the algorithms in exonerate are described in Slater GS and Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31; doi: 10.1186/1471-2105-6-31

Download

Exonerate is written in C, and currently uses the glib library for portablility. It is portable to all UNIX-like systems, and has been used on various Linux distributions, TRU64, OSX, and BSD.

It is licensed under the GPL.

You can download the source code or a precompiled version.

Exonerate version 2.2 includes fixes for problems with excessive memory consumption when compiled against glib-2, and fixes a bug with using exonerate-server with unmasked sequences.

Source code	exonerate-2.2.0.tar.gz
Linux/i386 binaries	exonerate-2.2.0-i386.tar.gz
Linux/x86_64 binaries	exonerate-2.2.0-x86_64.tar.gz

ex8-1

Exonerate: map cDNA onto genome

Intron/exon構造を考慮してtranscriptをゲノムにマッピングする。
（BLASTでは不可能なdonor/acceptor siteのGU/AGルールを考慮するマッピングソフトウェアが必要）
Exonerate を使う

キイロショウジョウバエのnos遺伝子のORFの配列が手元にある。ゲノムにマッピングせよ。(ex1-1と同じ問題)

- Transcript: Dmel_nos-PA.nuc.fasta
- Genome: dmel-all-chromosome-r6.13.fasta

```
exonerate --model est2genome --bestn 1 \
     Dmel_nos-PA.nuc.fasta Dmel_genome.3R.fasta
```

ex8-3

Exonerate: map protein onto genome

Intron/exon構造を考慮してprotein をゲノムにマッピングする。

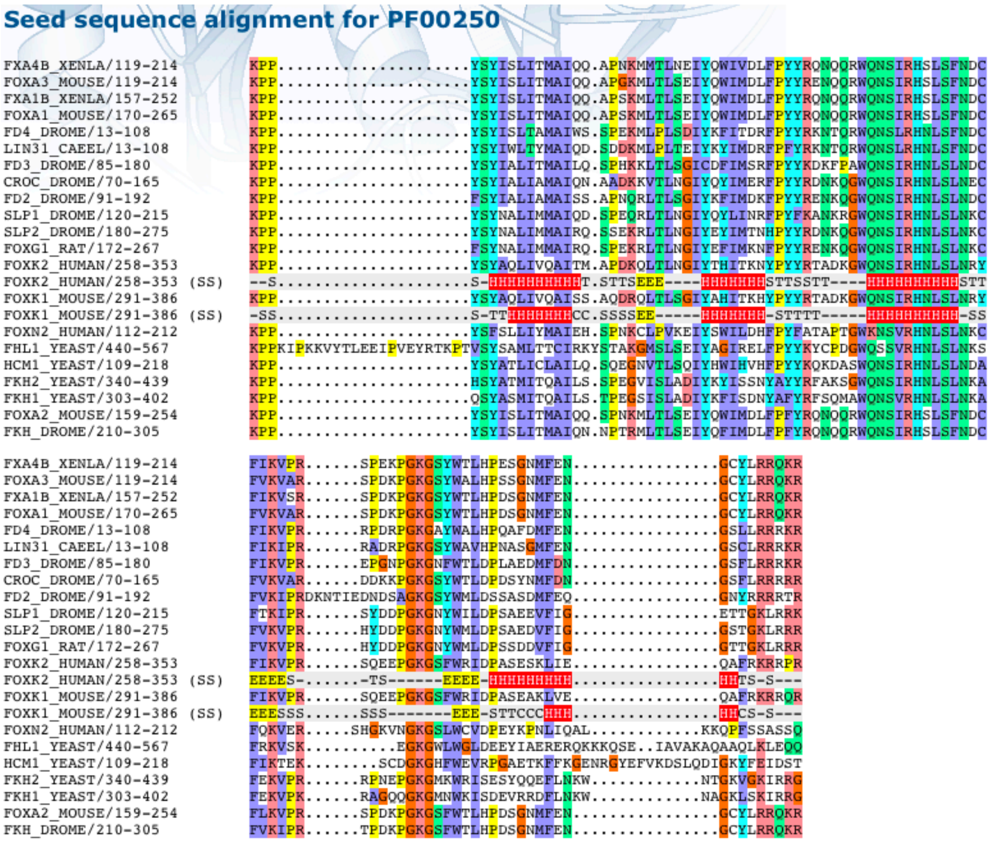
Exonerate を使う

キイロショウジョウバエのnos遺伝子のタンパク質の配列が手元にある。ゲノムにマッピングせよ。

- Transcript: Dmel_nos-PA.pep.fasta
- Genome: dmel-all-chromosome-r6.13.fasta

```
exonerate --model protein2genome --bestn 1 \
     Dmel_nos-PA.nuc.fasta Dmel_genome.3R.fasta
```

Multiple Alignment



Seed sequence alignment for PF00250

```
FXA4B_XENLA/119-214       KPP.....................YSYISLITMAIQQ.APNKMMTLNEIYQWIVDLFPYYRQNQQRWQNSIRHSLSFNDC
FOXA3_MOUSE/119-214       KPP.....................YSYISLITMAIQQ.APGKMLTLSEIYQWIMDLFPYYRENQQRWQNSIRHSLSFNDC
FXA1B_XENLA/157-252       KPP.....................YSYISLITMAIQQ.APSKMLTLSEIYQWIMDLFPYYRQNQQRWQNSIRHSLSFNDC
FOXA1_MOUSE/170-265       KPP.....................YSYISLITMAIQQ.APSKMLTLSEIYQWIMDLFPYYRQNQQRWQNSIRHSLSFNDC
FD4_DROME/13-108          KPP.....................YSYISLTAMAIWS.SPEKMLPLSDIYKFITDRFPYYRKNTQRWQNSLRHNLSFNDC
LIN31_CAEEL/13-108        KPP.....................YSYIWLTYMAIQD.SDDKMLPLTEIYKYIMDRFPFYRKNTQRWQNSLRHNLSFNDC
FD3_DROME/85-180          KPP.....................YSYIALITMAILQ.SPHKKLTLSGICDFIMSRFPYYKDKFPAWQNSIRHNLSLNDC
CROC_DROME/70-165         KPP.....................YSYIALIAMAIQN.AADKKVTLNGIYQYIMERFPYYRDNKQGWQNSIRHNLSLNEC
FD2_DROME/91-192          KPP.....................FSYIALIAMAISS.APNQRLTLSGIYKFIMDKFPYYRENKQGWQNSIRHNLSLNDC
SLP1_DROME/120-215        KPP.....................YSYNALIMMAIQD.SPEQRLTLNGIYQYLINRFPYFKANKRGWQNSIRHNLSLNKC
SLP2_DROME/180-275        KPP.....................YSYNALIMMAIRQ.SSEKRLTLNGIYEYIMTNHPYYRDNKQGWQNSIRHNLSLNKC
FOXG1_RAT/172-267         KPP.....................FSYNALIMMAIRQ.SPEKRLTLNGIYEFIMKNFPYYRENKQGWQNSIRHNLSLNKC
FOXK2_HUMAN/258-353       KPP.....................YSYAQLIVQAITM.APDKQLTLNGIYTHITKNYPYYRTADKGWQNSIRHNLSLNRY
FOXK2_HUMAN/258-353 (SS)  --S.....................S-HHHHHHHHHHT.STTSEEE----HHHHHHHSTTSSTT----HHHHHHHHHHSTT
FOXK1_MOUSE/291-386       KPP.....................YSYAQLIVQAISS.AQDRQLTLSGIYAHITKHYPYYRTADKGWQNSIRHNLSLNRY
FOXK1_MOUSE/291-386 (SS)  -SS.....................S-TTHHHHHHHCC.SSSSEE-----HHHHHHH-STTTT-----HHHHHHHHHH-SS
FOXN2_HUMAN/112-212       KPP.....................YSFSLLIYMAIEH.SPNKCLPVKEIYSWILDHFPYFATAPTGWKNSVRHNLSLNKC
FHL1_YEAST/440-567        KPPKIPKKVYTLEEIPVEYRTKPTVSYSAMLTTCIRKYSTAKGMSLSEIYAGIRELFPYYKYCPDGWQSSVRHNLSLNKS
HCM1_YEAST/109-218        KPP.....................YSYATLICLAILQ.SQEGNVTLSQIYHWIHVHFPYYKQKDASWQNSIRHNLSLNDA
FKH2_YEAST/340-439        KPP.....................HSYATMITQAILS.SPEGVISLADIYKYISSNYAYYRFAKSGWQNSIRHNLSLNKA
FKH1_YEAST/303-402        KPP.....................QSYASMITQAILS.TPEGSISLADIYKFISDNYAFYRFSQMAWQNSVRHNLSLNKA
FOXA2_MOUSE/159-254       KPP.....................YSYISLITMAIQQ.SPNKMLTLSEIYQWIMDLFPFYRQNQQRWQNSIRHSLSFNDC
FKH_DROME/210-305         KPP.....................YSYISLITMAIQN.NPTRMLTLSEIYQFIMDLFPFYRQNQQRWQNSIRHSLSFNDC

FXA4B_XENLA/119-214       FIKVPR......SPEKPGKGSYWTLHPESGNMFEN................GCYLRRQKR
FOXA3_MOUSE/119-214       FVKVAR......SPDKPGKGSYWALHPSSGNMFEN................GCYLRRQKR
FXA1B_XENLA/157-252       FIKVSR......SPDKPGKGSYWTLHPDSGNMFEN................GCYLRRQKR
FOXA1_MOUSE/170-265       FVKVAR......SPDKPGKGSYWTLHPDSGNMFEN................GCYLRRQKR
FD4_DROME/13-108          FIKVPR......RPDRPGKGAYWALHPQAFDMFEN................GSLLRRRKR
LIN31_CAEEL/13-108        FIKIPR......RADRPGKGSYWAVHPNASGMFEN................GSCLRRRKR
FD3_DROME/85-180          FIKVPR......EPGNPGKGNFWTLDPLAEDMFDN................GSFLRRRKR
CROC_DROME/70-165         FVKVAR......DDKKPGKGSYWTLDPDSYNMFDN................GSFLRRRRR
FD2_DROME/91-192          FVKIPRDKNTIEDNDSAGKGSYWMLDSSASDMFEQ................GNYRRRRTR
SLP1_DROME/120-215        FTKIPR......SYDDPGKGNYWILDPSAEEVFIG................ETTGKLRRK
SLP2_DROME/180-275        FVKVPR......HYDDPGKGNYWMLDPSAEDVFIG................GSTGKLRRR
FOXG1_RAT/172-267         FVKVPR......HYDDPGKGNYWMLDPSSDDVFIG................GTTGKLRRR
FOXK2_HUMAN/258-353       FIKVPR......SQEEPGKGSFWRIDPASESKLIE................QAFRKRRPR
FOXK2_HUMAN/258-353 (SS)  EEEES-......-TS------EEEE-HHHHHHHHH................HHTS-S---
FOXK1_MOUSE/291-386       FIKVPR......SQEEPGKGSFWRIDPASEAKLVE................QAFRKRRQR
FOXK1_MOUSE/291-386 (SS)  EEESSS......SSS-------EEE-STTCCCHHH................HHCS-S---
FOXN2_HUMAN/112-212       FQKVER.....SHGKVNGKGSLWCVDPEYKPNLIQAL............KKQPFSSASSQ
FHL1_YEAST/440-567        FRKVSK..........EGKGWLWGLDEEYIAERERQKKKQSE..IAVAKAQAAQLKLEQQ
HCM1_YEAST/109-218        FIKTEK........SCDGKGHFWEVRPGAETKFFKGENRGYEFVKDSLQDIGKYFEIDST
FKH2_YEAST/340-439        FEKVPR......RPNEPGKGMKWRISESYQQEFLNKW............NTGKVGKIRRG
FKH1_YEAST/303-402        FEKVPK......RAGQQGKGMNWKISDEVRRDFLNKW............NAGKLSKIRRG
FOXA2_MOUSE/159-254       FLKVPR......SPDKPGKGSFWTLHPDSGNMFEN................GCYLRRQKR
FKH_DROME/210-305         FVKIPR......TPDKPGKGSFWTLHPDSGNMFEN................GCYLRRQKR
```

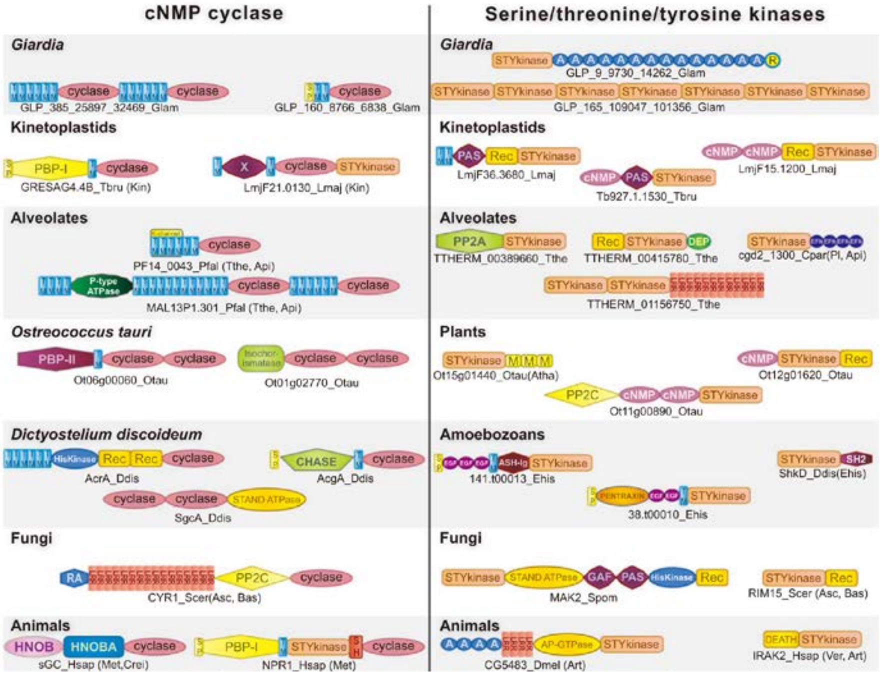
Why Multiple Sequence Alignments?

- Compare multiple sequences
- Identify conserved regions, patterns, and domains
  - Predicting function
  - Predicting structure
  - Identifying new members of protein families
- Perform phylogenetic analysis
- Generate position-specific scoring matrices for profile search

Software for Multiple Alignment

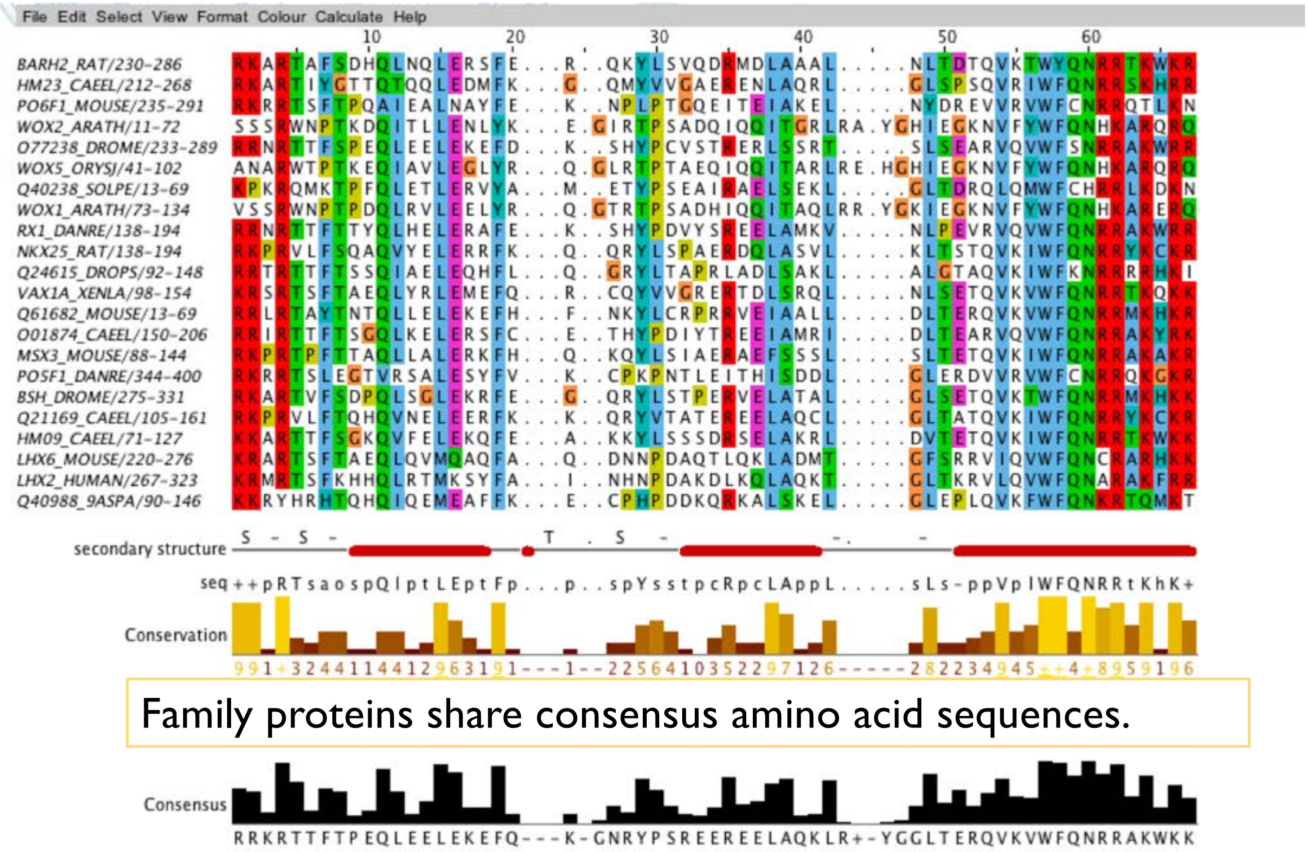
- Clustal Omega
  - http://www.clustal.org/omega/
- MUSCLE
  - http://www.drive5.com/muscle/index.htm
- MAFFT
  - http://mafft.cbrc.jp/alignment/server/
- PRANK
  - http://wasabiapp.org/software/prank/

Protein Families and Motif Search



Anantharaman et al. (2007)

- Proteins are composed of functional modules.
- The modules are conserved among species.

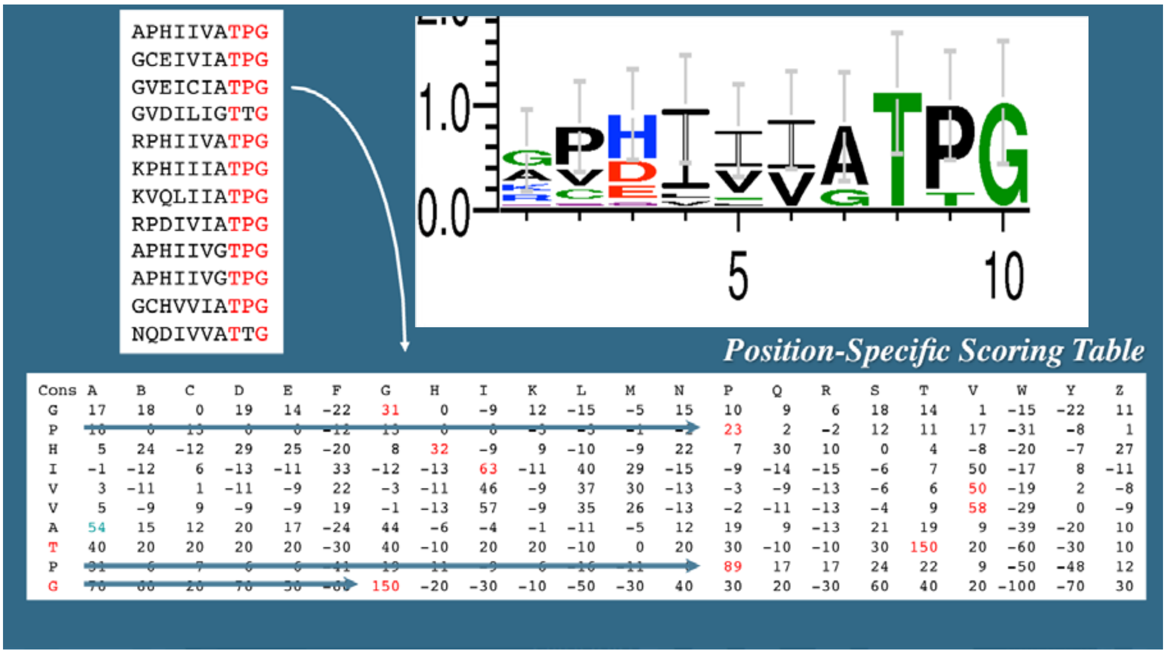


```
                       10        20        30        40        50        60
BARH2_RAT/230-286      RKARTAFSDHQLNQLERSFE...R..QKYLSVQDRMDLAAAL.....NLTDTQVKTWYQNRRTKWKR
HM23_CAEEL/212-268     RKARTIYGTTQTQQLEDMFK...G..QMYVVGAERENLAQRL.....GLSPSQVRIWFQNRRSKHRR
PO6F1_MOUSE/235-291    RKRRTSFTPQAIEALNAYFE...K..NPLPTGQEITEIAKEL.....NYDREVVRVWFCNRRQTLKN
WOX2_ARATH/11-72       SSSRWNPTKDQITLLENLYK...E.GIRTPSADQIQQITGRLRA.YGHIEGKNVFYWFQNHKARQRQ
O77238_DROME/233-289   RRNRTTFSPEQLEELEKEFD...K..SHYPCVSTRERLSSRT.....SLSEARVQVWFSNRRAKWRR
WOX5_ORYSJ/41-102      ANARWTPTKEQIAVLEGLYR...Q.GLRTPTAEQIQQITARLRE.HGHIEGKNVFYWFQNHKARQRQ
Q40238_SOLPE/13-69     KPKRQMKTPFQLETLERVYA...M..ETYPSEAIRAELSEKL.....GLTDRQLQMWFCHRRLKDKN
WOX1_ARATH/73-134      VSSRWNPTPDQLRVLEELYR...Q.GTRTPSADHIQQITAQLRR.YGKIEGKNVFYWFQNHKARERQ
RX1_DANRE/138-194      RRNRTTFTTYQLHELERAFE...K..SHYPDVYSREELAMKV.....NLPEVRVQVWFQNRRAKWRR
NKX25_RAT/138-194      RKPRVLFSQAQVYELERRFK...Q..QRYLSPAERDQLASVL.....KLTSTQVKIWFQNRRYKCKR
Q24615_DROPS/92-148    RRTRTTFTSSQIAELEQHFL...Q..GRYLTAPRLADLSAKL.....ALGTAQVKIWFKNRRRRHKI
VAX1A_XENLA/98-154     KRSRTSFTAEQLYRLEMEFQ...R..CQYVVGRERTDLSRQL.....NLSETQVKVWFQNRRTKQKK
Q61682_MOUSE/13-69     RRLRTAYTNTQLLELEKEFH...F..NKYLCRPRRVEIAALL.....DLTERQVKVWFQNRRMKHKR
O01874_CAEEL/150-206   RRIRTTFTSGQLKELERSFC...E..THYPDIYTREEIAMRI.....DLTEARVQVWFQNRRAKYRK
MSX3_MOUSE/88-144      RKPRTPFTTAQLLALERKFH...Q..KQYLSIAERAEFSSSL.....SLTETQVKIWFQNRRAKAKR
PO5F1_DANRE/344-400    RKRRTSLEGTVRSALESYFV...K..CPKPNTLEITHISDDL.....GLERDVVRVWFCNRRQKGKR
BSH_DROME/275-331      RKARTVFSDPQLSGLEKRFE...G..QRYLSTPERVELATAL.....GLSETQVKTWFQNRRMKHKK
Q21169_CAEEL/105-161   RKPRVLFTQHQVNELEERFK...K..QRYVTATEREELAQCL.....GLTATQVKIWFQNRRYKCKR
HM09_CAEEL/71-127      KKARTTFSGKQVFELEKQFE...A..KKYLSSSDRSELAKRL.....DVTETQVKIWFQNRRTKWKK
LHX6_MOUSE/220-276     KRARTSFTAEQLQVMQAQFA...Q..DNNPDAQTLQKLADMT.....GFSRRVIQVWFQNCRARHKK
LHX2_HUMAN/267-323     KRMRTSFKHHQLRTMKSYFA...I..NHNPDAKDLKQLAQKT.....GLTKRVLQVWFQNARAKFRR
Q40988_9ASPA/90-146    KKRYHRHTQHQIQEMEAFFK...E..CPHPDDKQRKALSKEL.....GLEPLQVKFWFQNKRTQMKT

secondary structure    S - S -             T . S -             -.    -

                   seq ++pRTsaospQIptLEptFp...p..spYsstpcRpcLAppL.....sLs-ppVpIWFQNRRtKhK+

Conservation           991+32441144129631 91---1--2256410352297126-----282234945++4+8959196

Consensus              RRKRTTFTPEQLEELEKEFQ---K-GNRYPSREEREELAQKLR+-YGGLTERQVKVWFQNRRAKWKK
```

Family proteins share consensus amino acid sequences.

Profiles

- Numerical representations of multiple sequence alignments
- Represent the common characteristics of a protein family



```
APHIIVATPG
GCEIVIATPG
GVEICIATPG
GVDILIGTTG
RPHIIVATPG
KPHIIIATPG
KVQLIIATPG
RPDIVIATPG
APHIIVGTPG
APHIIVGTPG
GCHVVIATPG
NQDIVVATTG
```

Position-Specific Scoring Table

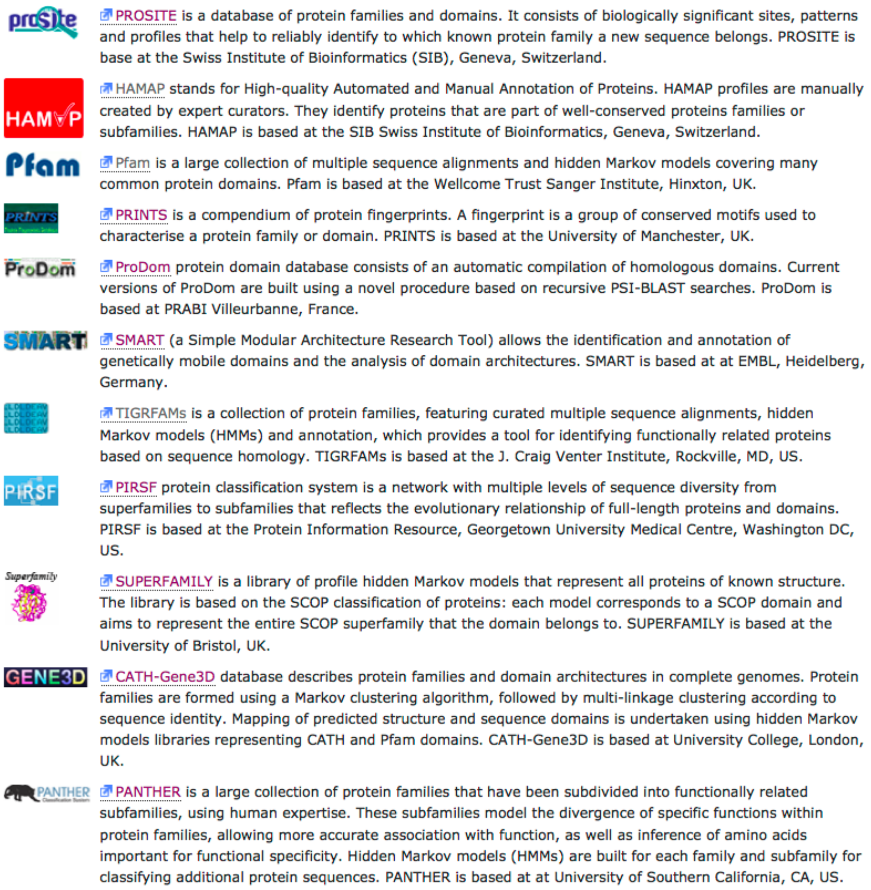
Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
G	17	18	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22	11
P	10	0	13	0	0	-12	13	0	8	-5	-5	-1	-2	23	2	-2	12	11	17	-31	-8	1
H	5	24	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7	27
I	-1	-12	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-6	7	50	-17	8	-11
V	3	-11	1	-11	-9	22	-3	-11	46	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2	-8
V	5	-9	9	-9	-9	19	-1	-13	57	-9	35	26	-13	-2	-11	-13	-4	9	58	-29	0	-9
A	54	15	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20	10
T	40	20	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30	10
P	31	6	7	6	6	-11	19	11	-9	6	-16	-11	[illegible]	89	17	17	24	22	9	-50	-48	12
G	70	60	20	70	50	-60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30

Profile Search

- Uses "collective characteristics" of a family of proteins, rather than individual sequences.
- The "collective characteristics" can be represented as **sequence profile**, or **weight matrices**.

- Tools:
  - HMMER
  - PSI-BLAST
- Profile search is more sensitive than sequence-query searches.
  - => Distantly related genes/proteins can be detected.

Profile/Domain/Motif Databases



PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family a new sequence belongs. PROSITE is base at the Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland.

HAMAP stands for High-quality Automated and Manual Annotation of Proteins. HAMAP profiles are manually created by expert curators. They identify proteins that are part of well-conserved proteins families or subfamilies. HAMAP is based at the SIB Swiss Institute of Bioinformatics, Geneva, Switzerland.

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains. Pfam is based at the Wellcome Trust Sanger Institute, Hinxton, UK.

PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family or domain. PRINTS is based at the University of Manchester, UK.

ProDom protein domain database consists of an automatic compilation of homologous domains. Current versions of ProDom are built using a novel procedure based on recursive PSI-BLAST searches. ProDom is based at PRABI Villeurbanne, France.

SMART (a Simple Modular Architecture Research Tool) allows the identification and annotation of genetically mobile domains and the analysis of domain architectures. SMART is based at at EMBL, Heidelberg, Germany.

TIGRFAMs is a collection of protein families, featuring curated multiple sequence alignments, hidden Markov models (HMMs) and annotation, which provides a tool for identifying functionally related proteins based on sequence homology. TIGRFAMs is based at the J. Craig Venter Institute, Rockville, MD, US.

PIRSF protein classification system is a network with multiple levels of sequence diversity from superfamilies to subfamilies that reflects the evolutionary relationship of full-length proteins and domains. PIRSF is based at the Protein Information Resource, Georgetown University Medical Centre, Washington DC, US.

SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure. The library is based on the SCOP classification of proteins: each model corresponds to a SCOP domain and aims to represent the entire SCOP superfamily that the domain belongs to. SUPERFAMILY is based at the University of Bristol, UK.

CATH-Gene3D database describes protein families and domain architectures in complete genomes. Protein families are formed using a Markov clustering algorithm, followed by multi-linkage clustering according to sequence identity. Mapping of predicted structure and sequence domains is undertaken using hidden Markov models libraries representing CATH and Pfam domains. CATH-Gene3D is based at University College, London, UK.

PANTHER is a large collection of protein families that have been subdivided into functionally related subfamilies, using human expertise. These subfamilies model the divergence of specific functions within protein families, allowing more accurate association with function, as well as inference of amino acids important for functional specificity. Hidden Markov models (HMMs) are built for each family and subfamily for classifying additional protein sequences. PANTHER is based at at University of Southern California, CA, US.

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

InterPro and InterProScan

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

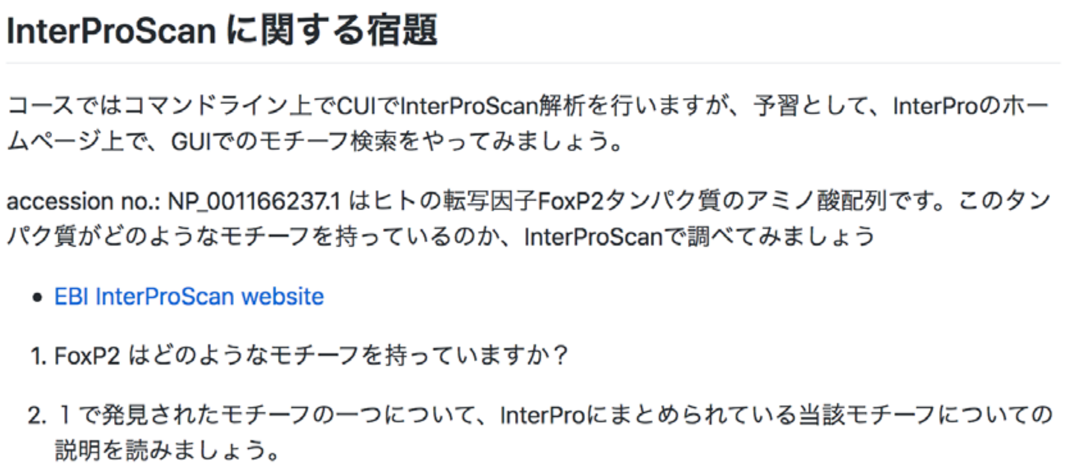
- What is InterPro?
  - InterPro is a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites. To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the InterPro consortium.
- What is InterProScan?
  - InterProScan is the software package that allows sequences to be scanned against InterPro's signatures.

InterPro and InterProScan

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

- Why is InterPro useful?
  - InterPro combines signatures from multiple, diverse databases into a single searchable resource, reducing redundancy and helping users interpret their sequence analysis results.
- Who uses InterPro?
  - InterPro is used by research scientists interested in the large-scale analysis of whole proteomes, genomes and metagenomes, as well as researchers seeking to characterise individual protein sequences. Within the EBI, InterPro is used to help annotate protein sequences in UniProtKB. It is also used by the Gene Ontology Annotation group to automatically assign Gene Ontology terms to protein sequences

InterProScan in GUI



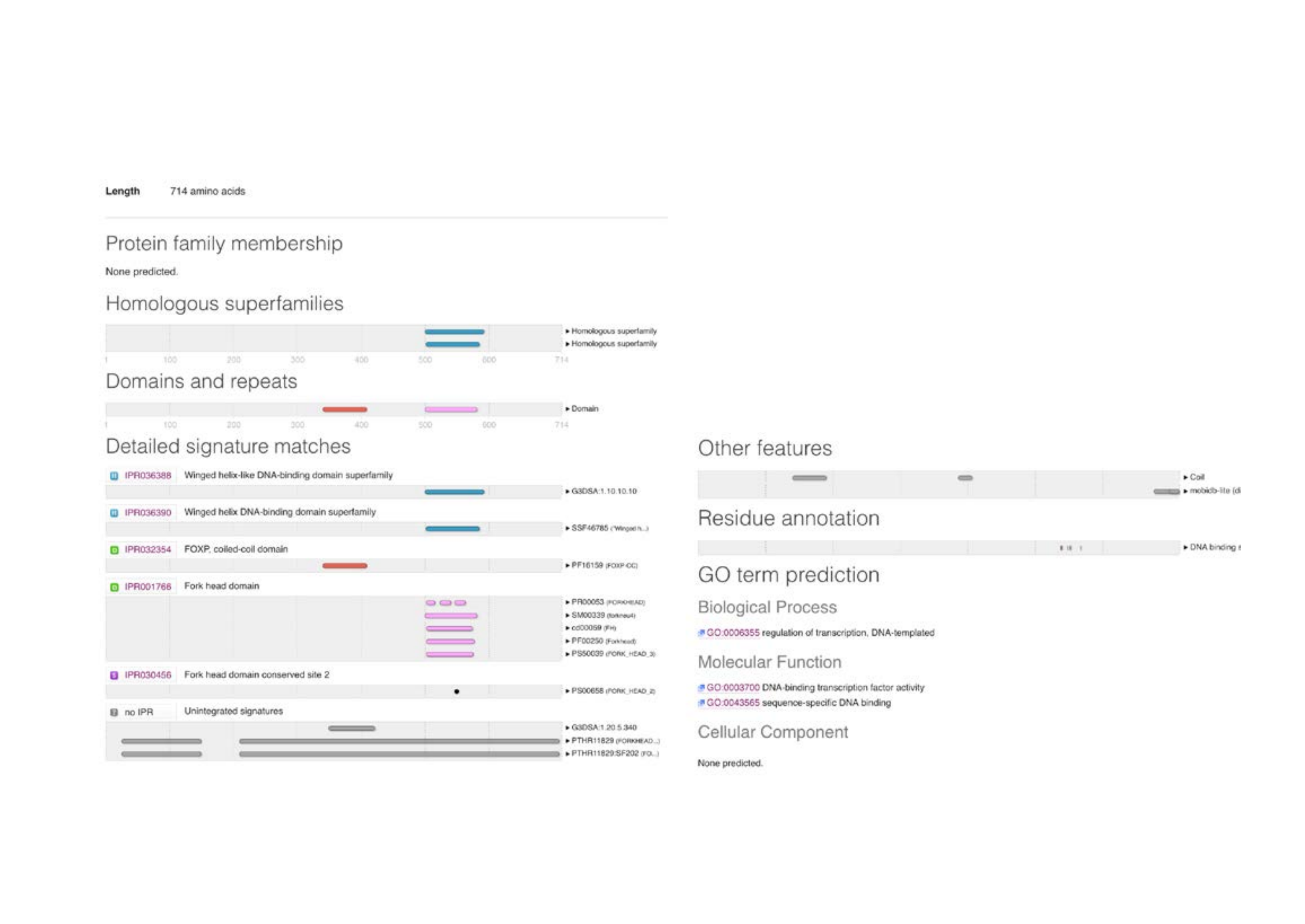
InterProScan に関する宿題

コースではコマンドライン上でCUIでInterProScan解析を行いますが、予習として、InterProのホームページ上で、GUIでのモチーフ検索をやってみましょう。

accession no.: NP_001166237.1 はヒトの転写因子FoxP2タンパク質のアミノ酸配列です。このタンパク質がどのようなモチーフを持っているのか、InterProScanで調べてみましょう

- EBI InterProScan website

1. FoxP2 はどのようなモチーフを持っていますか？
2. １で発見されたモチーフの一つについて、InterProにまとめられている当該モチーフについての説明を読みましょう。



Length 714 amino acids

Protein family membership

None predicted.

Homologous superfamilies

Domains and repeats

Detailed signature matches

H IPR036388 Winged helix-like DNA-binding domain superfamily

H IPR036390 Winged helix DNA-binding domain superfamily

D IPR032354 FOXP, coiled-coil domain

D IPR001766 Fork head domain

S IPR030456 Fork head domain conserved site 2

no IPR Unintegrated signatures

Other features

Residue annotation

GO term prediction

Biological Process

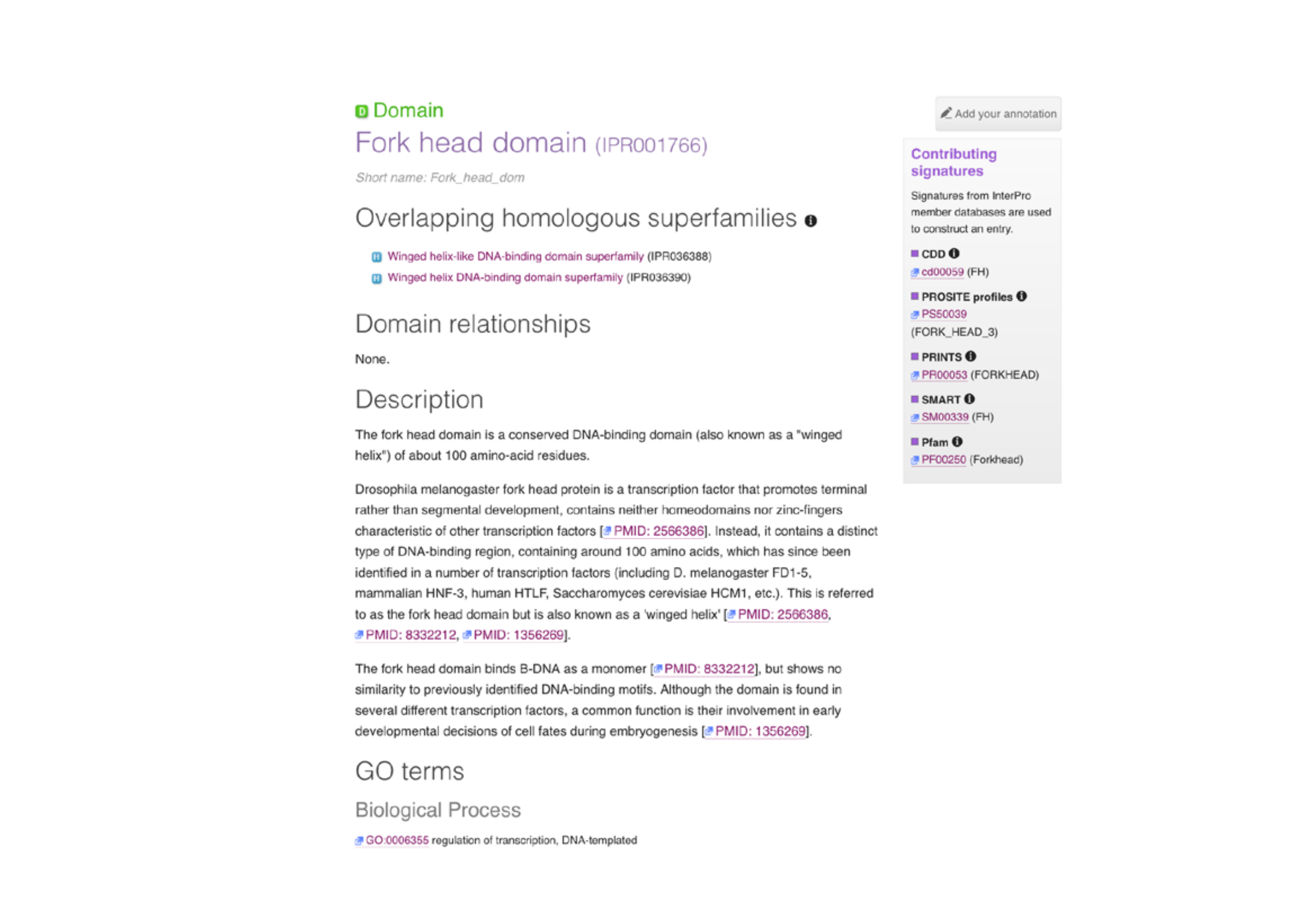
GO:0006355 regulation of transcription, DNA-templated

Molecular Function

GO:0003700 DNA-binding transcription factor activity
GO:0043565 sequence-specific DNA binding

Cellular Component

None predicted.



D Domain

Fork head domain (IPR001766)

Short name: Fork_head_dom

Overlapping homologous superfamilies

- H Winged helix-like DNA-binding domain superfamily (IPR036388)
- H Winged helix DNA-binding domain superfamily (IPR036390)

Domain relationships

None.

Description

The fork head domain is a conserved DNA-binding domain (also known as a "winged helix") of about 100 amino-acid residues.

Drosophila melanogaster fork head protein is a transcription factor that promotes terminal rather than segmental development, contains neither homeodomains nor zinc-fingers characteristic of other transcription factors [PMID: 2566386]. Instead, it contains a distinct type of DNA-binding region, containing around 100 amino acids, which has since been identified in a number of transcription factors (including D. melanogaster FD1-5, mammalian HNF-3, human HTLF, Saccharomyces cerevisiae HCM1, etc.). This is referred to as the fork head domain but is also known as a 'winged helix' [PMID: 2566386, PMID: 8332212, PMID: 1356269].

The fork head domain binds B-DNA as a monomer [PMID: 8332212], but shows no similarity to previously identified DNA-binding motifs. Although the domain is found in several different transcription factors, a common function is their involvement in early developmental decisions of cell fates during embryogenesis [PMID: 1356269].

GO terms

Biological Process

GO:0006355 regulation of transcription, DNA-templated

Contributing signatures

Signatures from InterPro member databases are used to construct an entry.

CDD
cd00059 (FH)

PROSITE profiles
PS50039 (FORK_HEAD_3)

PRINTS
PR00053 (FORKHEAD)

SMART
SM00339 (FH)

Pfam
PF00250 (Forkhead)

Protein motif search using InterProScan

- Query: protein sequence(s)
- Software: InterProScan
- DB: 21 databases are available. Pfam etc.

Search example

```
$ interproscan.sh  -i protein.aa.fas -f TSV --goterms --
pathways --appl Pfam
```

Protein motif search using InterProScan

ヒトFoxP2タンパク質がどのようなモチーフを持っているのか、InterProScanを使って調べる。ここではデータベースはPfamを使う。

- Query: human FoxP2 (HsFoxP2.NP_01166237.aa.fas)
- Software: InterProScan
- DB: Pfam

Search

```
$ interproscan.sh  -I proteins.fasta -f XML,TSV --goterms
--pathways
```

Result (TSV)

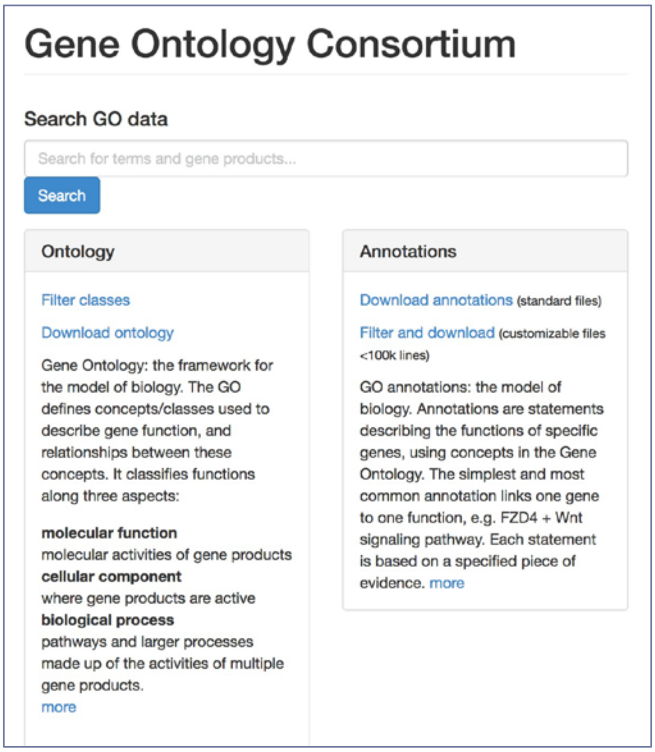
```
NP_001166237.1  94f57  714  Pfam  PF00250 Forkhead domain         503   578  3.2E-27 T  03-09-2018  IPR001766  Fork head domain
GO:0003700|GO:0006355|GO:0043565
NP_001166237.1  94f57  714  Pfam  PF16159 FOXP coiled-coil domain 341   409  3.4E-35 T  03-09-2018  IPR032354  FOXP, coiled-coil
domain
```

Gene annotation and Gene Ontology

- Gene annotation
- GO
  - What is GO?
  - Why GO is required?
  - GO and BLAST?

Two components of GO

- Ontology
- Gene associations



Gene Ontology Consortium

Search GO data

Ontology

Filter classes

Download ontology

Gene Ontology: the framework for the model of biology. The GO defines concepts/classes used to describe gene function, and relationships between these concepts. It classifies functions along three aspects:

molecular function
molecular activities of gene products
cellular component
where gene products are active
biological process
pathways and larger processes made up of the activities of multiple gene products.
more

Annotations

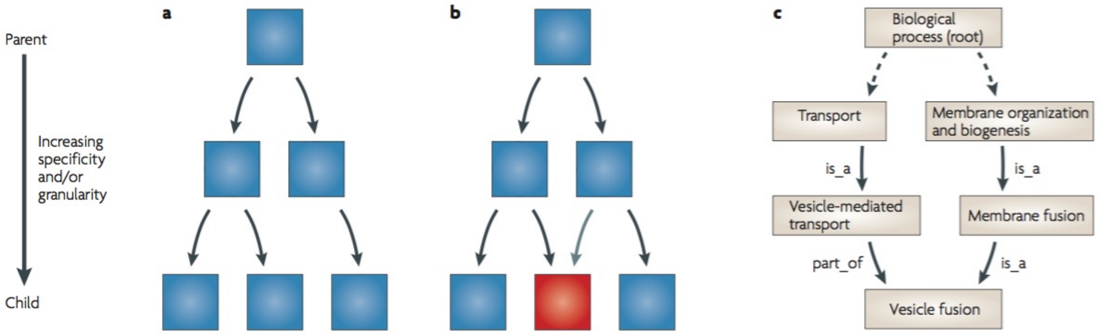
Download annotations (standard files)

Filter and download (customizable files <100k lines)

GO annotations: the model of biology. Annotations are statements describing the functions of specific genes, using concepts in the Gene Ontology. The simplest and most common annotation links one gene to one function, e.g. FZD4 + Wnt signaling pathway. Each statement is based on a specified piece of evidence. more

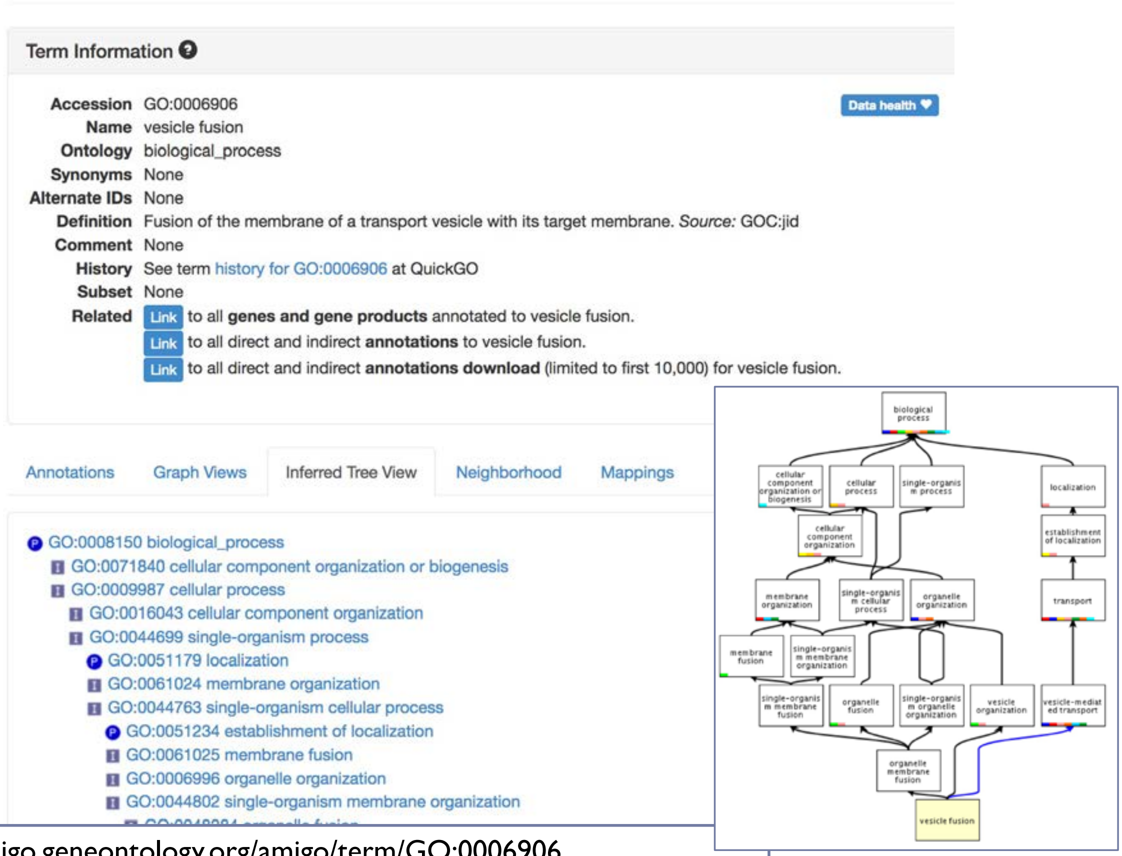
Ontology structure

- Ontologies are represented as a directed acyclic graph (DAG).
- Parent-child relationship
  - is_a
  - part_of
- Ontology can be changed / updated


Parent
Increasing specificity and/or granularity
Child
a
b
c
Biological process (root)
Transport
Membrane organization and biogenesis
is_a
is_a
Vesicle-mediated transport
Membrane fusion
part_of
is_a
Vesicle fusion

Rhee et al., 2008

vesicle fusion



Term Information

Accession	GO:0006906
Name	vesicle fusion
Ontology	biological_process
Synonyms	None
Alternate IDs	None
Definition	Fusion of the membrane of a transport vesicle with its target membrane. *Source:* GOC:jid
Comment	None
History	See term history for GO:0006906 at QuickGO
Subset	None
Related	Link to all **genes and gene products** annotated to vesicle fusion. Link to all direct and indirect **annotations** to vesicle fusion. Link to all direct and indirect **annotations download** (limited to first 10,000) for vesicle fusion.

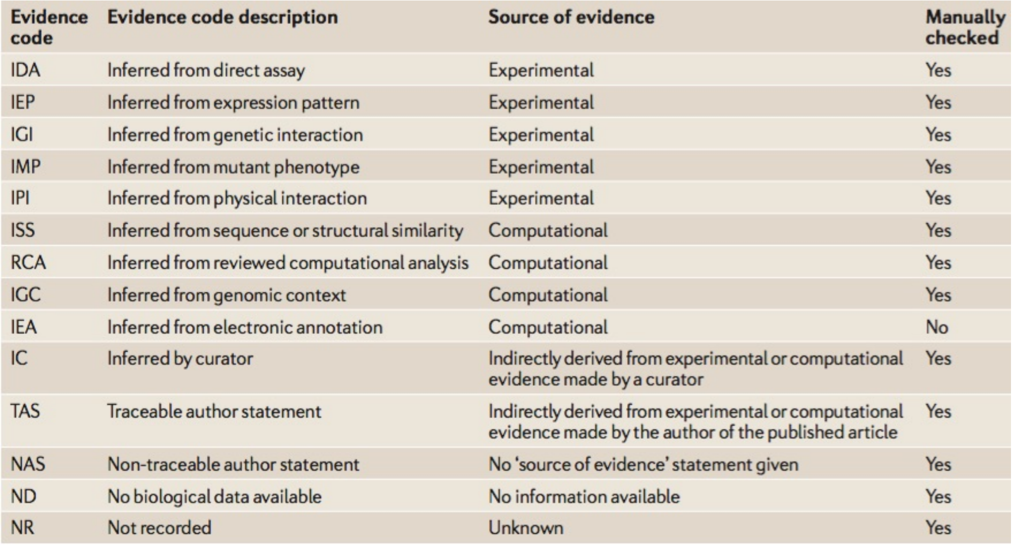
- GO:0008150 biological_process
  - GO:0071840 cellular component organization or biogenesis
  - GO:0009987 cellular process
    - GO:0016043 cellular component organization
    - GO:0044699 single-organism process
      - GO:0051179 localization
      - GO:0061024 membrane organization
      - GO:0044763 single-organism cellular process
        - GO:0051234 establishment of localization
        - GO:0061025 membrane fusion
        - GO:0006996 organelle organization
        - GO:0044802 single-organism membrane organization

biological process
cellular component organization or biogenesis
cellular process
single-organism process
localization
cellular component organization
establishment of localization
membrane organization
single-organism cellular process
organelle organization
transport
membrane fusion
single-organism membrane organization
single-organism membrane fusion
organelle fusion
single-organism organelle organization
vesicle organization
vesicle-mediated transport
organelle membrane fusion
vesicle fusion

http://amigo.geneontology.org/amigo/term/GO:0006906

Gene association

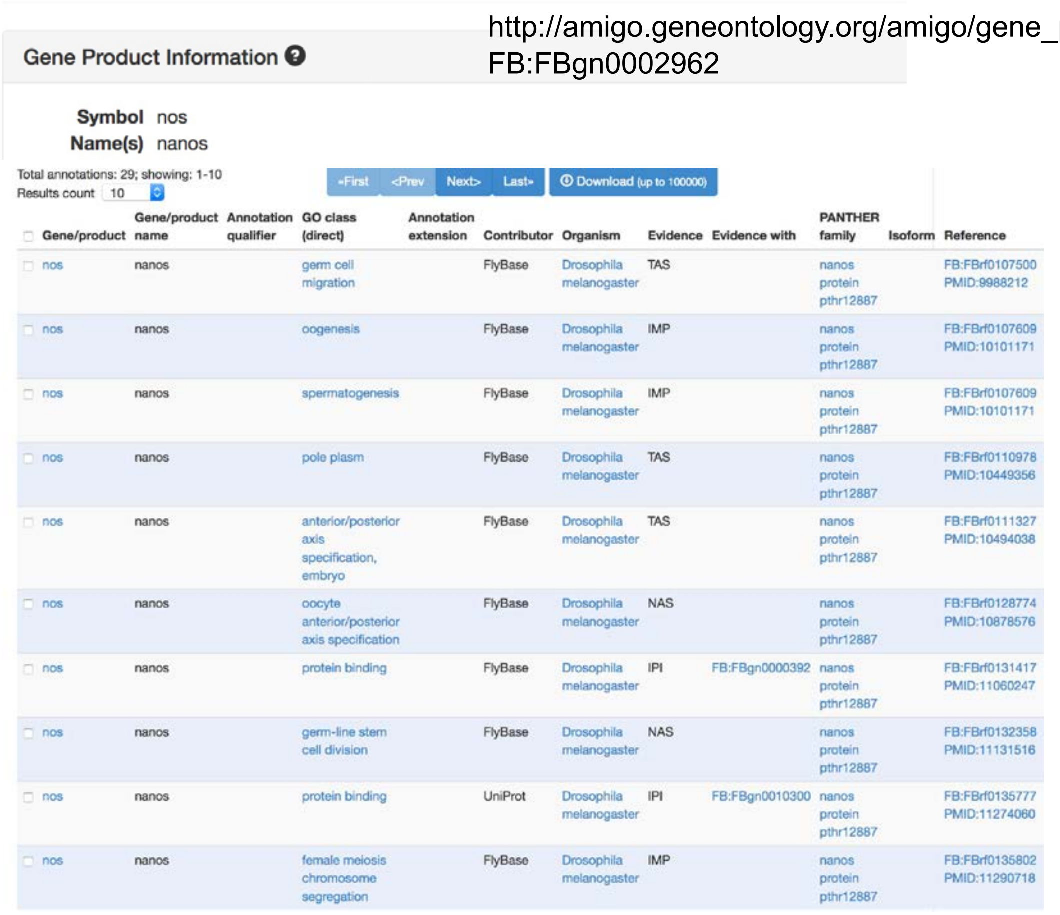
- Gene <=> GO
- A gene may associate with multiple GO terms.
- Evidence codes.



Evidence code	Evidence code description	Source of evidence	Manually checked
IDA	Inferred from direct assay	Experimental	Yes
IEP	Inferred from expression pattern	Experimental	Yes
IGI	Inferred from genetic interaction	Experimental	Yes
IMP	Inferred from mutant phenotype	Experimental	Yes
IPI	Inferred from physical interaction	Experimental	Yes
ISS	Inferred from sequence or structural similarity	Computational	Yes
RCA	Inferred from reviewed computational analysis	Computational	Yes
IGC	Inferred from genomic context	Computational	Yes
IEA	Inferred from electronic annotation	Computational	No
IC	Inferred by curator	Indirectly derived from experimental or computational evidence made by a curator	Yes
TAS	Traceable author statement	Indirectly derived from experimental or computational evidence made by the author of the published article	Yes
NAS	Non-traceable author statement	No 'source of evidence' statement given	Yes
ND	No biological data available	No information available	Yes
NR	Not recorded	Unknown	Yes

nanos

http://amigo.geneontology.org/amigo/gene_product/FB:FBgn0002962



Gene Product Information

Symbol nos
Name(s) nanos

Total annotations: 29; showing: 1-10
Results count 10

□ Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family	Isoform	Reference
□ nos	nanos		germ cell migration		FlyBase	Drosophila melanogaster	TAS		nanos protein pthr12887		FB:FBrf0107500 PMID:9988212
□ nos	nanos		oogenesis		FlyBase	Drosophila melanogaster	IMP		nanos protein pthr12887		FB:FBrf0107609 PMID:10101171
□ nos	nanos		spermatogenesis		FlyBase	Drosophila melanogaster	IMP		nanos protein pthr12887		FB:FBrf0107609 PMID:10101171
□ nos	nanos		pole plasm		FlyBase	Drosophila melanogaster	TAS		nanos protein pthr12887		FB:FBrf0110978 PMID:10449356
□ nos	nanos		anterior/posterior axis specification, embryo		FlyBase	Drosophila melanogaster	TAS		nanos protein pthr12887		FB:FBrf0111327 PMID:10494038
□ nos	nanos		oocyte anterior/posterior axis specification		FlyBase	Drosophila melanogaster	NAS		nanos protein pthr12887		FB:FBrf0128774 PMID:10878576
□ nos	nanos		protein binding		FlyBase	Drosophila melanogaster	IPI	FB:FBgn0000392	nanos protein pthr12887		FB:FBrf0131417 PMID:11060247
□ nos	nanos		germ-line stem cell division		FlyBase	Drosophila melanogaster	NAS		nanos protein pthr12887		FB:FBrf0132358 PMID:11131516
□ nos	nanos		protein binding		UniProt	Drosophila melanogaster	IPI	FB:FBgn0010300	nanos protein pthr12887		FB:FBrf0135777 PMID:11274060
□ nos	nanos		female meiosis chromosome segregation		FlyBase	Drosophila melanogaster	IMP		nanos protein pthr12887		FB:FBrf0135802 PMID:11290718

How to annotate GO for non-model organisms?

- Ortholog grouping with a model organism and then transfer the GO terms from the reference organism to your target organism.
- BLAST2GO
- InterProScan