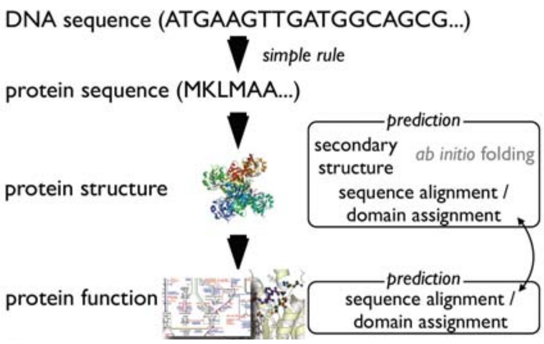
Functional Annotation / Protein Annotation

- Protein structure
- Structure/Function prediction
- Functional Classification

Functional prediction

- In absence of a proven function or mutant phenotype, a gene is no more than a transcribed piece of DNA
- If we can predict gene function in a genome, we can create a “parts list” of molecular functions that allow to make assumptions about the organism
- Actual experiments to elucidate or even just validate gene function usually take many years. Per gene!

From sequence to function



DNA sequence (ATGAAGTTGATGGCAGCG...)

simple rule

protein sequence (MKLMAA...)

protein structure

prediction

secondary structure

ab initio folding

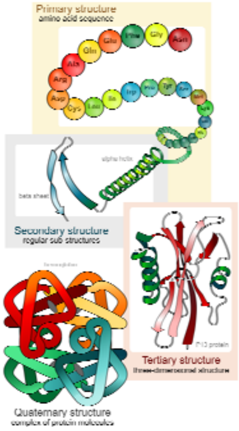
sequence alignment / domain assignment

protein function

prediction

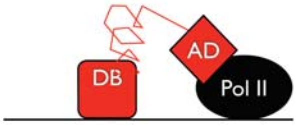
sequence alignment / domain assignment

- Primary
  - AA sequence
  - Post-translational modifications
- Secondary
  - α-helix
  - β-sheet
- Tertiary
  - 3d folding driven by non-specific hydrophobic interactions
  - Stabilised by specific tertiary interactions
- Quaternary
  - Like tertiary but with multiple protein chains



Protein domains

- Many proteins consist of several structural and functional entities called “domain”

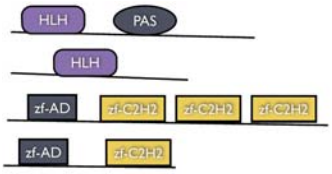


E.g. Transcription factor

DNA-binding and trans-activating domain structurally and functionally different.

- Often, proteins of the same family are ‘mixtures’ of a set of standard domains

Domain architectures



Descriptions of domains:

http://pfam.sanger.ac.uk/

Ab initio prediction of secondary structure from primary structure

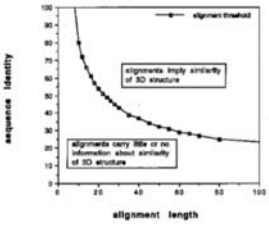
- learning directly from X-ray structures
- consideration of environment
- neural network-based training

e.g. Dor *et al.* - Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training – Proteins 2007

Prediction by alignment

- Primary sequence similarity >30% can be assumed to have the same 3D structure (but not necessarily function - beware of details!)
- Any available structural or functional data on orthologues (the “same” protein in a different organism) can be of great relevance.

The 30% sequence identity rule



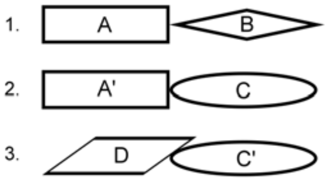
Sander & Schneider, Proteins 9(1):56-68, 1991

BLAST
bi-directional best hit

is often the most simple tool to establish orthology

BLAST of $A_{species\ 1}$ identifies $B_{species\ 2}$
BLAST of $B_{species\ 2}$ identifies $A_{species\ 1}$
as best hits
and sequence similarity > 30%.
BINGO.

Intermediate sequences increase the detection of homology between sequences

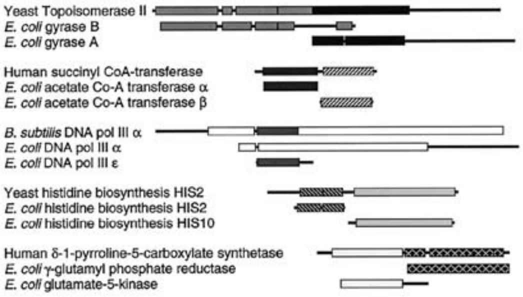


1 and 2 are homologous
2 and 3 are homologous
1 and 3 are **not** homologous

- Multi-domain proteins introduce errors

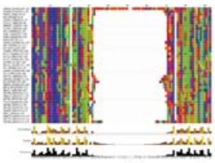
Park *et al.* - Intermediate sequences increase the detection of homology between sequences – J Mol Bio. 1997

However: Function from partial matches


Yeast Topoisomerase II
E. coli gyrase B
E. coli gyrase A
Human succinyl CoA-transferase
E. coli acetate Co-A transferase α
E. coli acetate Co-A transferase β
B. subtilis DNA pol III α
E. coli DNA pol III α
E. coli DNA pol III ε
Yeast histidine biosynthesis HIS2
E. coli histidine biosynthesis HIS2
E. coli histidine biosynthesis HIS10
Human δ-1-pyrroline-5-carboxylate synthetase
E. coli γ-glutamyl phosphate reductase
E. coli glutamate-5-kinase

Identifying domains

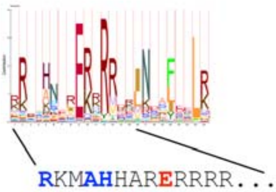
- Structural domains show a high degree of conservation.
- Multiple sequence alignments [of the amino acids found in the domain] from multiple proteins (from different or the same species) can be used for a prediction.



sequence alignment

hidden Markov model

Learning a profile



RKM**AH**HAR**E**RRRR...
RGL**MH****NELE**KRRR...
KRARA**NELE**KQMV...
RTA**AH**KQN**E**RKMR...

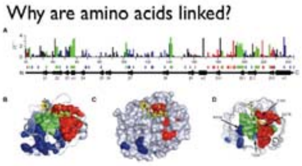
HMM profiles have a higher sensitivity than alignment-based approaches AND effectively compare the candidate sequence against all sequences in the alignment profile at the same time.

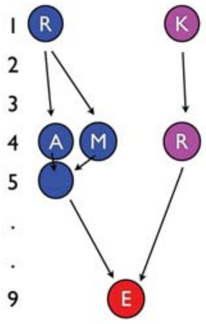
Limitations of profile HMMs

Higher-order relationships are not preserved (but this is also true for sequence alignments).

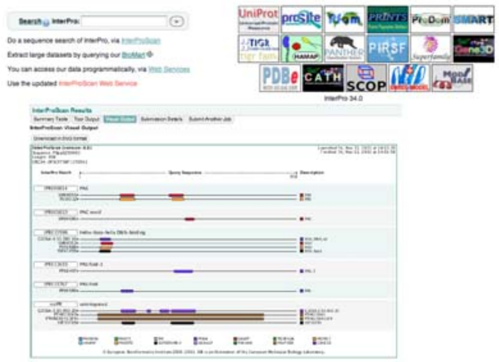
RKM**AH**HAR**E**RRRR...
RGL**MH****NELE**KRRR...
KRARA**NELE**KQMV...
RTA**AH**KQN**E**RKMR...
...

Why are amino acids linked?





Annotating domains - Interpro



Search InterPro:

Do a sequence search of InterPro, via InterProScan

Extract large datasets by querying our BioMart

You can access our data programmatically, via Web Services

Use the updated InterProScan Web Service

InterPro 34.0

InterProScan Results

Summary Table | Text Output | Visual Output | Submission Details | Submit Another Job

InterProScan Visual Output

Download in SVG format

Functional annotation

- For enzymes: EC number
- 6 groupings – Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases, Ligases

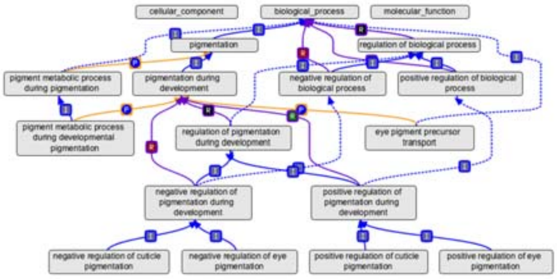
EC 2.2.1.6

Functional annotation

For other proteins: Gene Ontology, Reactome, KEGG

- GO is the *de facto* standard used by all major model organism databases
- Founded in 2000 by the GO Consortium lead by Michael Ashburner
- Database curators read the scientific literature and assign functional classifications along with an evidence code to proteins
- These annotations follow a controlled vocabulary that is organized into an ontology.

GO Structure



cellular_component
biological_process
molecular_function
pigmentation
regulation of biological process
pigment metabolic process during pigmentation
pigmentation during development
negative regulation of biological process
positive regulation of biological process
pigment metabolic process during developmental pigmentation
regulation of pigmentation during development
eye pigment precursor transport
negative regulation of pigmentation during development
positive regulation of pigmentation during development
negative regulation of cuticle pigmentation
negative regulation of eye pigmentation
positive regulation of cuticle pigmentation
positive regulation of eye pigmentation

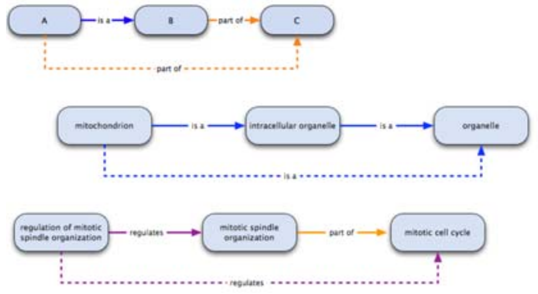
I: is a

P: part of

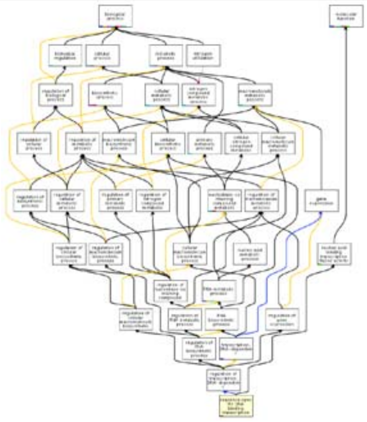
R: regulates

GO can be navigated as a directed acyclic graph

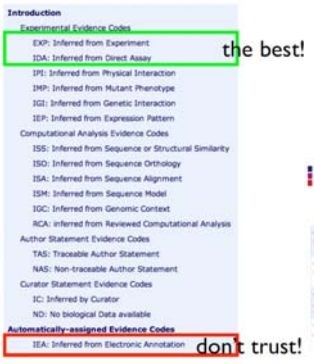
Inferences and logic



Annotation granularity and inconsistencies are a big issue:
1: Index finger *part of* Extremity
2: Ring finger *is a* Finger *part of* Hand *part of* Extremity



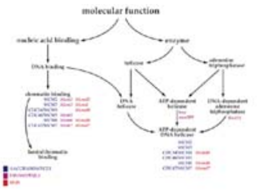
Evidence codes



Introduction
Experimental Evidence Codes
EXP: Inferred from Experiment
IDA: Inferred from Direct Assay
IPI: Inferred from Physical Interaction
IMP: Inferred from Mutant Phenotype
IGI: Inferred from Genetic Interaction
IEP: Inferred from Expression Pattern
Computational Analysis Evidence Codes
ISS: Inferred from Sequence or Structural Similarity
ISO: Inferred from Sequence Orthology
ISA: Inferred from Sequence Alignment
ISM: Inferred from Sequence Model
IGC: Inferred from Genomic Context
RCA: inferred from Reviewed Computational Analysis
Author Statement Evidence Codes
TAS: Traceable Author Statement
NAS: Non-traceable Author Statement
Curator Statement Evidence Codes
IC: Inferred by Curator
ND: No biological Data available
Automatically-assigned Evidence Codes
IEA: Inferred from Electronic Annotation

the best!

don't trust!



Species	Genes	Annot.
Drosophila melanogaster FlyBase	12517	72277 (56665 non-IEA)
Escherichia coli EcoCyc & EcoliHub	3543	38310 (1898 non-IEA)
Ehrlichia chaffeensis Arkansas JCVI	1091	2861 (2861 non-IEA)
Gallus gallus GO Annotations @ EBI	16306	70675 (2035 non-IEA)

References

- Dor *et al.* - Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training – Proteins 2007
- Sander & Schneider - Database of homology-derived protein structures and the structural meaning of sequence alignment – Proteins 1991
- Park *et al.* - Intermediate sequences increase the detection of homology between sequences – J Mol Bio. 1997

References

- Must read: The documentation at
  - www.geneontology.org

Assignment 2

- 20 min group presentations
- Annotate *Drosophila* species, compare annotations across species
- Groups on Moodle – either Gene or Protein

- ALL submit a copy of the presentation on Moodle; PER GROUP print out a copy of slides and give to me on day of presentations

Online resources to explore in your own time - I

Resource centres:

The EBI:	www.ebi.ac.uk
The NCBI:	www.ncbi.nlm.nih.gov
UCSC:	genome.ucsc.edu

Model organism databases:

Budding yeast:	www.yeastgenome.org
Worm:	www.wormbase.org
Fly:	www.flybase.org
Mouse:	www.informatics.jax.org
Rat:	rgd.mcw.edu
Zebrafish:	zfin.org

Database building tool: www.intermine.org

InterMine databases:

www.flymine.org
intermine.modencode.org
ratmine.mcw.edu
yeastmine.yeastgenome.org
www.metabolicmine.org
www.flytf.org
mitominer.mrc-mbu.cam.ac.uk
targetmine.nibio.go.jp

Ontologies:
Gene ontology:
www.geneontology.org
Sequence ontology:
www.sequenceontology.org

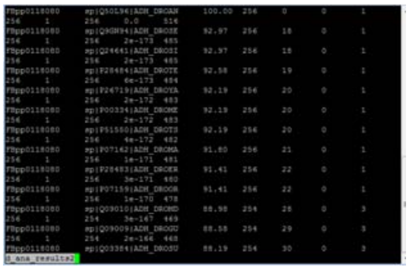
DNA sequence/ genomes:
Ensembl: www.ensembl.org
Short read archive:
www.ncbi.nlm.nih.gov/sra

Proteins:
www.uniprot.org
www.ebi.ac.uk/interpro

RNA:
rfam.sanger.ac.uk

Pathways:
www.reactome.org
www.genome.jp/kegg/

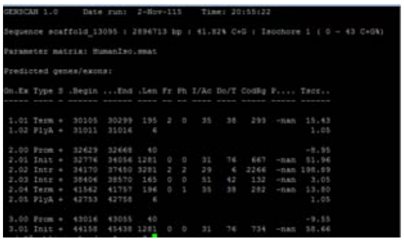
/local/data/public/genome_informatics/programs/ncbi-blast-2.5.0+/bin/blastp -query dana-all-translation-r1.04.fasta -db /local/data/public/genome_informatics/assignment_2/swissprot_database/uniprot_sprot.fasta -out d_ana_results -outfmt 6 -num_threads 60



```
FBpp0118080     sp|Q50L96|ADH_DROAN     100.00  256     0       0       1
256     1       256     0.0     516
FBpp0118080     sp|Q9GN94|ADH_DROSE     92.97   256     18      0       1
256     1       256     2e-173  485
FBpp0118080     sp|Q24641|ADH_DROSI     92.97   256     18      0       1
256     1       256     2e-173  485
FBpp0118080     sp|P28484|ADH_DROTE     92.58   256     19      0       1
256     1       256     6e-173  484
FBpp0118080     sp|P26719|ADH_DROYA     92.19   256     20      0       1
256     1       256     2e-172  483
FBpp0118080     sp|P00334|ADH_DROME     92.19   256     20      0       1
256     1       256     2e-172  483
FBpp0118080     sp|P51550|ADH_DROTS     92.19   256     20      0       1
256     1       256     4e-172  482
FBpp0118080     sp|P07162|ADH_DROMA     91.80   256     21      0       1
256     1       256     1e-171  481
FBpp0118080     sp|P28483|ADH_DROER     91.41   256     22      0       1
256     1       256     3e-171  480
FBpp0118080     sp|P07159|ADH_DROOR     91.41   256     22      0       1
256     1       256     1e-170  478
FBpp0118080     sp|Q09010|ADH_DROMD     88.98   254     28      0       3
256     1       254     3e-167  469
FBpp0118080     sp|Q09009|ADH_DROGU     88.58   254     29      0       3
256     1       254     2e-166  468
FBpp0118080     sp|Q03384|ADH_DROSU     88.19   254     30      0       3
d_ana_results2
```

- Query ID
- Subject ID
- Percentage Identity of alignment
- Alignment length
- Mismatch count
- Gap open count
- Query start
- Query end
- Subject start
- Subject end
- E-value
- bitScore

/local/data/public/genome_informatics/programs/genscan/genscan /local/data/public/genome_informatics/programs/genscan/lib/HumanIso.smat test.fasta 1> genscan.txt

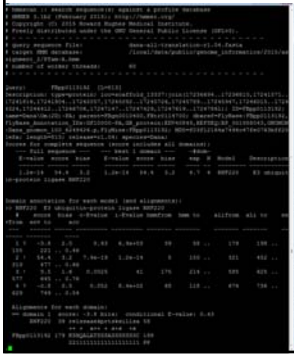


```
GENSCAN 1.0     Date run:  2-Nov-115    Time: 20:55:22

Sequence scaffold_13095 : 2896713 bp : 41.82% C+G : Isochore 1 ( 0 - 43 C+G%)

Parameter matrix: HumanIso.smat

Predicted genes/exons:

Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------

 1.01 Term +  30105  30299  195  2  0   35   38   293 -nan  15.43
 1.02 PlyA +  31011  31016    6                                1.05

 2.00 Prom +  32629  32668   40                               -8.95
 2.01 Init +  32776  34056 1281  0  0   31   76   667 -nan  51.96
 2.02 Intr +  34170  37450 3281  2  2   29    6  2266 -nan 198.89
 2.03 Intr +  38406  38570  165  0  0   51   42   132 -nan   3.05
 2.04 Term +  41562  41757  196  0  1   35   38   282 -nan  13.80
 2.05 PlyA +  42753  42758    6                                1.05

 3.00 Prom +  43016  43055   40                               -9.55
 3.01 Init +  44158  45438 1281  0  0   31   76   734 -nan  58.66
```

- Term Terminal exon
- PlyA PolyA tail
- Prom Promoter
- Init Initial exon
- Intr Internal exon
- Sngl Single exon

```
/local/data/public/genome_informatics/programs/hmmer-3.1b2-linux-intel-x86_64/binaries/hmmscan --cpu 60 /local/data/public/genome_informatics/assignment_2/Pfam-A.hmm dana-all-translation-r1.04.fasta 1> dana_test.txt
```

--tblout <f> : table of per-sequence hits



--domtblout <f> : table of per-domain hits