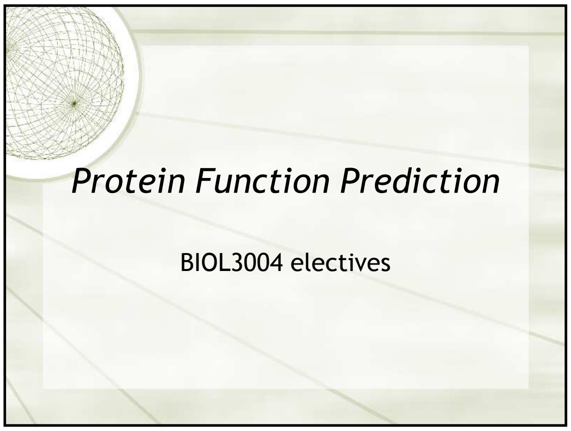
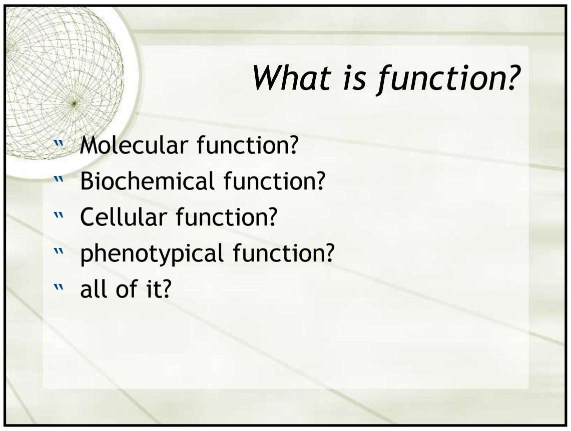


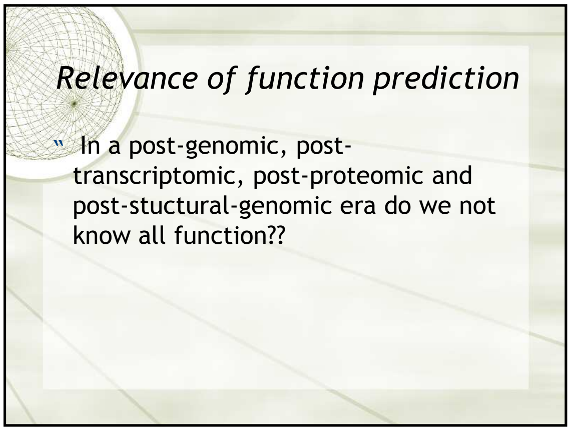
Protein Function Prediction

BIOL3004 electives



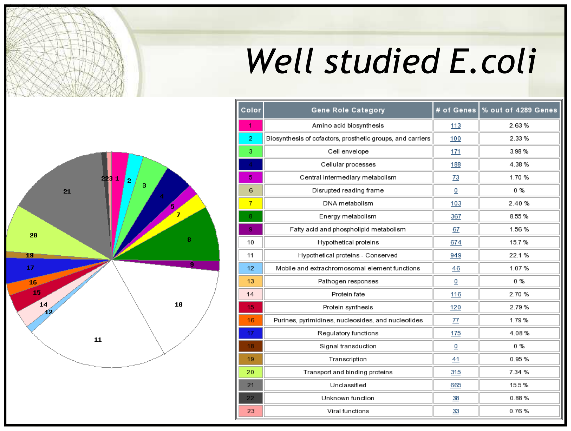
What is function?

- Molecular function?
- Biochemical function?
- Cellular function?
- phenotypical function?
- all of it?



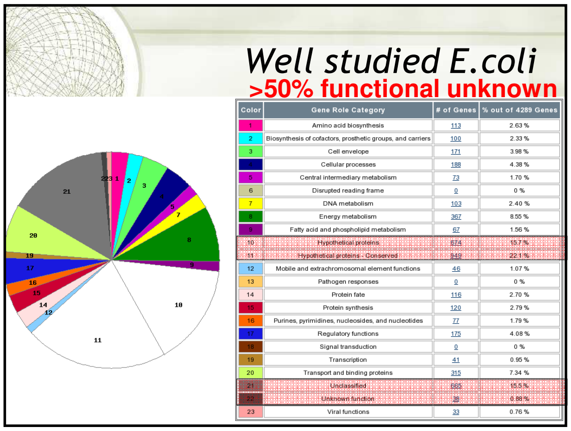
Relevance of function prediction

- In a post-genomic, post-transcriptomic, post-proteomic and post-stuctural-genomic era do we not know all function??



Well studied E.coli

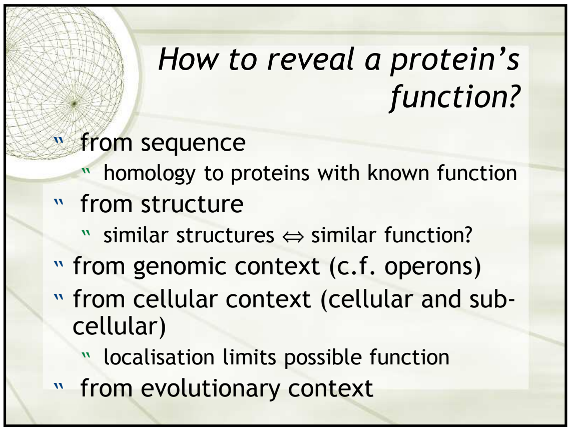
Color	Gene Role Category	# of Genes	% out of 4289 Genes
1	Amino acid biosynthesis	113	2.63 %
2	Biosynthesis of cofactors, prosthetic groups, and carriers	100	2.33 %
3	Cell envelope	171	3.98 %
4	Cellular processes	188	4.38 %
5	Central intermediary metabolism	73	1.70 %
6	Disrupted reading frame	0	0 %
7	DNA metabolism	103	2.40 %
8	Energy metabolism	367	8.55 %
9	Fatty acid and phospholipid metabolism	67	1.56 %
10	Hypothetical proteins	674	15.7 %
11	Hypothetical proteins - Conserved	949	22.1 %
12	Mobile and extrachromosomal element functions	46	1.07 %
13	Pathogen responses	0	0 %
14	Protein fate	116	2.70 %
15	Protein synthesis	120	2.79 %
16	Purines, pyrimidines, nucleosides, and nucleotides	77	1.79 %
17	Regulatory functions	175	4.08 %
18	Signal transduction	0	0 %
19	Transcription	41	0.95 %
20	Transport and binding proteins	315	7.34 %
21	Unclassified	665	15.5 %
22	Unknown function	38	0.88 %
23	Viral functions	33	0.76 %



Well studied E.coli

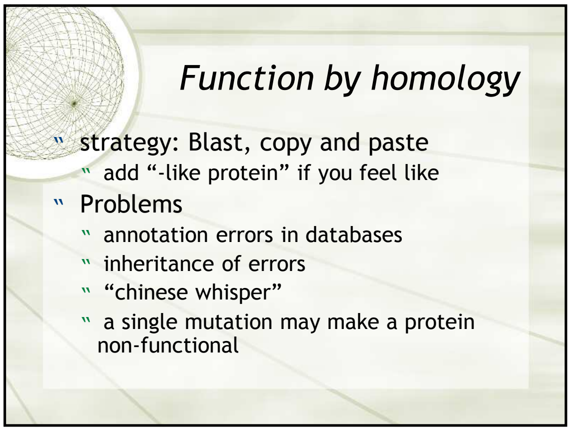
>50% functional unknown

Color	Gene Role Category	# of Genes	% out of 4289 Genes
1	Amino acid biosynthesis	113	2.63 %
2	Biosynthesis of cofactors, prosthetic groups, and carriers	100	2.33 %
3	Cell envelope	171	3.98 %
4	Cellular processes	188	4.38 %
5	Central intermediary metabolism	73	1.70 %
6	Disrupted reading frame	0	0 %
7	DNA metabolism	103	2.40 %
8	Energy metabolism	367	8.55 %
9	Fatty acid and phospholipid metabolism	67	1.56 %
10	Hypothetical proteins	674	15.7 %
11	Hypothetical proteins - Conserved	949	22.1 %
12	Mobile and extrachromosomal element functions	46	1.07 %
13	Pathogen responses	0	0 %
14	Protein fate	116	2.70 %
15	Protein synthesis	120	2.79 %
16	Purines, pyrimidines, nucleosides, and nucleotides	77	1.79 %
17	Regulatory functions	175	4.08 %
18	Signal transduction	0	0 %
19	Transcription	41	0.95 %
20	Transport and binding proteins	315	7.34 %
21	Unclassified	665	15.5 %
22	Unknown function	38	0.88 %
23	Viral functions	33	0.76 %



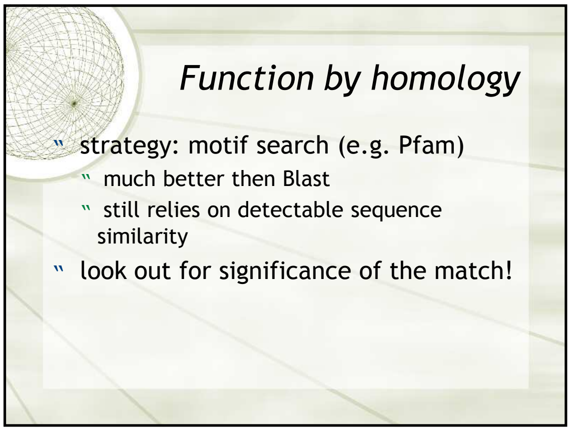
How to reveal a protein's function?

- from sequence
  - homology to proteins with known function
- from structure
  - similar structures ⇔ similar function?
- from genomic context (c.f. operons)
- from cellular context (cellular and sub-cellular)
  - localisation limits possible function
- from evolutionary context



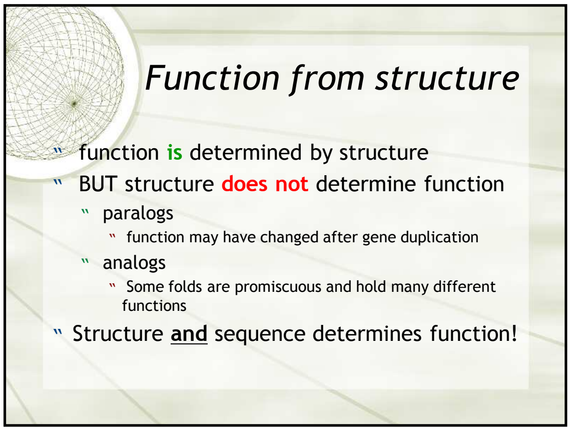
Function by homology

- strategy: Blast, copy and paste
  - add “-like protein” if you feel like
- Problems
  - annotation errors in databases
  - inheritance of errors
  - “chinese whisper”
  - a single mutation may make a protein non-functional



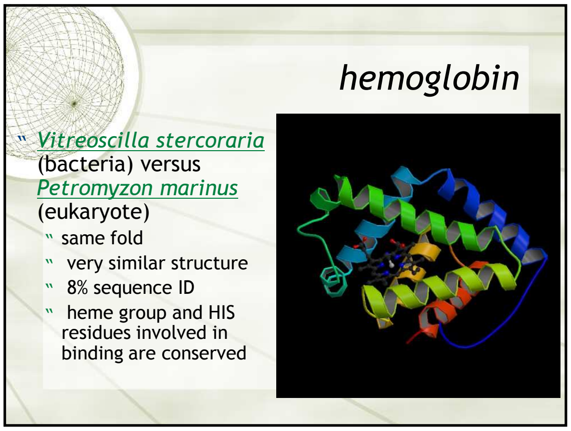
Function by homology

- strategy: motif search (e.g. Pfam)
  - much better then Blast
  - still relies on detectable sequence similarity
- look out for significance of the match!



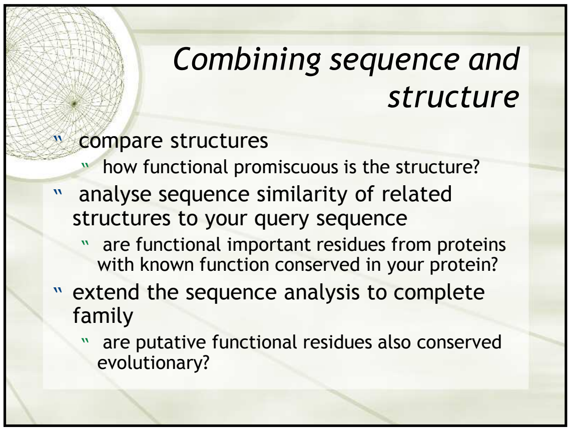
Function from structure

- function **is** determined by structure
- BUT structure **does not** determine function
  - paralogs
    - function may have changed after gene duplication
  - analogs
    - Some folds are promiscuous and hold many different functions
- Structure **and** sequence determines function!



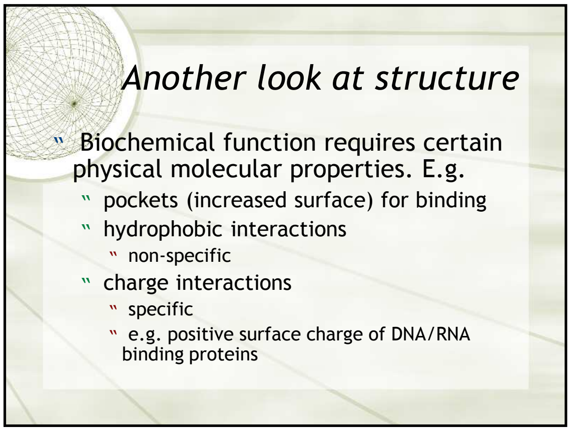
hemoglobin

- *Vitreoscilla stercoraria* (bacteria) versus *Petromyzon marinus* (eukaryote)
  - same fold
  - very similar structure
  - 8% sequence ID
  - heme group and HIS residues involved in binding are conserved



Combining sequence and structure

- compare structures
  - how functional promiscuous is the structure?
- analyse sequence similarity of related structures to your query sequence
  - are functional important residues from proteins with known function conserved in your protein?
- extend the sequence analysis to complete family
  - are putative functional residues also conserved evolutionary?



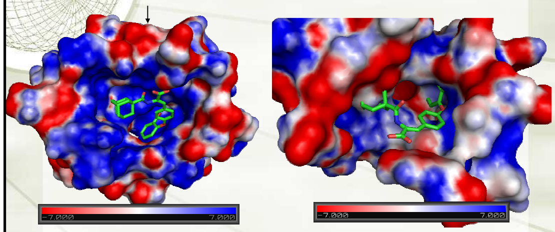
Another look at structure

- Biochemical function requires certain physical molecular properties. E.g.
  - pockets (increased surface) for binding
  - hydrophobic interactions
    - non-specific
  - charge interactions
    - specific
    - e.g. positive surface charge of DNA/RNA binding proteins

HCV inhibitors on site I

HCV pdb1Z4U

Dengue model



G410 ⇒ L476 mutation

Protein surfaces

- To highlight surface features
  - high quality visualistion for nice figures in your paper
- You can calculate them within PyMOL
  - different surface properties (e.g. electrostatic surface)
    - both PyMOL and APBS is on the DVD

Other data supporting function

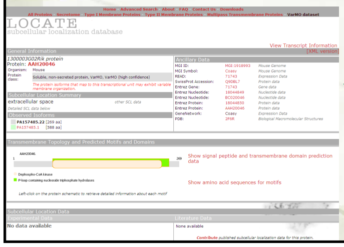
- genomic context
  - bacterial protein
    - functional units (operons) are conserved
    - analyse functional commonalities of co-locating genes
  - eukaryotic proteins
    - functionally related proteins get often physically joint during evolution
    - look for fusion proteins of your target with other proteins

Other data supporting function

- Protein-protein interactions
  - physical interaction suggest functional interaction
  - interaction networks of proteins (interactomes) are available for several model organisms
  - Data quality varies significantly
    - yeast two hybrid
    - bait tag purification
    - Interaction reports from literature

Other data supporting function

- sub-cellular context
  - Sub-cellular location of proteins can either be predicted or experimentally determined
  - both are available for mouse proteins through the LOCATE database


Home - Advanced Search - About - FAQ - Contact Us - Downloads
All Proteins - Secretome - Type I Membrane Proteins - Type II Membrane Proteins - Multipass Transmembrane Proteins - VarMO dataset

LOCATE
subcellular localization database

View Transcript Information
[XML version]

General Information

1300003G02Rik protein
Protein: **AAH20046**
Organism: Mouse
Protein class: Soluble, non-secreted protein, VarMO, VarMO (high confidence)
The protein isoforms that map to this transcriptional unit may exhibit variable membrane organization.

Subcellular Location Summary
extracellular space *other SCL data*
Detailed SCL data below

Observed Isoforms
PA157485.22 [269 aa]
PA157485.1 [588 aa]

Ancillary Data

MGI ID:	MGI:1918993	*Mouse Genome*
MGI Symbol:	Coasy	*Mouse Genome*
READ:	71743	*Expression Data*
SwissProt Accession:	Q9DBL7	*Protein data*
Entrez Gene:	71743	*Gene data*
Entrez Nucleotide:	18044849	*Nucleotide data*
Entrez Nucleotide:	BC020046	*Nucleotide data*
Entrez Protein:	18044850	*Protein data*
Entrez Protein:	AAH20046	*Protein data*
GeneNetwork:	Coasy	*Expression Data*
PDB:	2F6R	*Biological Macromolecular Structures*

Transmembrane Topology and Predicted Motifs and Domains

Show signal peptide and transmembrane domain prediction data

Show amino acid sequences for motifs

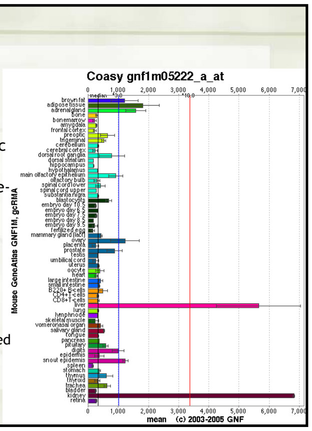
Left-click on the protein schematic to retrieve detailed information about each motif

Subcellular Location Data

Experimental Data	Literature Data
No data available	None available

Contribute *published subcellular localization data for this protein.*

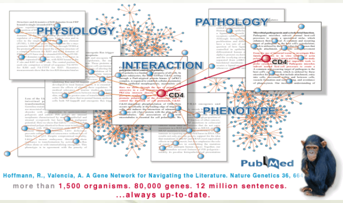
- **cellular context**
  - cellular function (and to some extent molecular function) are tissue specific
  - for the mouse ortholog of your target there are tissue-specific transcriptional regulation data available through BioInfoWeb
  - microarray data is intrinsically noisy
    - potentially compare regulation data of other genes known to be involved in the putative function



Literature context

- Chances are high that someone has worked on your target
  - but publication may be hard to find because another name was used
- iHOP

  Information hyperlinked over proteins



Hoffmann, R., Valencia, A. A Gene Network for Navigating the Literature. Nature Genetics 36, 66

more than 1,500 organisms. 80,000 genes. 12 million sentences. ...always up-to-date.

Summary

- Function prediction most accurate when evidence is cumulated
- Use holistic, hypothesis-driven approach and try to support (disproof) putative function (alternative functions)