I1: Introduction to protein function

pp2_intro_func1

Protein Prediction 2 (for Computational Biology) - Protein function
TUM winter semester
Announcements

Videos: YouTube / www.rostlab.org
THANKS:

Special lectures:
- (TBC)

No lecture:
- 10/30 SVV (student rep)
- 11/01 All Saints
- 11/22 Thanksgiving
- 12/06 Dies Academicus (TUM)
- 12/20-01/06 - no lecture Xmas+

LAST lecture: Jan 22 (followed by 2 wrap-up sessions)

Examen: Feb 07 10:00-13:00, LMU physics

Makeup: TBC: Apr 23 & Apr 25, 2019 - lecture time
THANKS for slides!

© Michael Leunig:
The Traveling Leunig
now
Bar Ilan Univ
Israel
& Boston Univ

Yanay Ofran
Marco Punta

Marco Punta (now ICR London) & Avner Schlessinger (now: Mount Sinai, New York)
Kazimierz O. Wrzeszczynski

Kazimierz O. Wrzeszczynski (now NYGC)
I1
Introduction
protein function
I1a: Basics of life
Life begins how?
RNA at the beginning

RNA at the beginning


Andrei Lupas
MPI Tuebingen
Central dogma of molecular biology

DNA → RNA → Protein

information, code, library, manual
intermediate step
machinery of life
Any dent in that picture?
Central dogma of molecular biology

DNA → RNA → Protein

- DNA: information, code, library, manual
- RNA: intermediate step
- Protein: machinery of life

1) Ribosomes without proteins
2) sRNA/RNAi short RNA important regulator
FANTOM 3: Functional annotation of mouse

Science 2005 309:1559-63
The Transcriptional Landscape of the Mammalian Genome

Numbers:
- Total transcripts: 181,047
- New protein-coding transcripts: 16,247
- New proteins: 5,154
- Multiple splice variants: 65%
- 1.35 5’ start sites for each 3’ end
- 1.83 3’ ends for each 5’ end
Central dogma of molecular biology

DNA → RNA → Protein

- DNA: information, code, library, manual
- RNA: intermediate step
- Protein: machinery of life

1) Ribosomes without proteins
2) sRNA/RNAi short RNA important regulator
3) Universe of non-coding RNA

May be time to bury dogma
Intro: protein function
Proteins are the machinery of life
... but what is a protein?
Define protein
Protein sequence

>gi|16128674|ref|NP_415226.1| potassium translocating ATPase, subunit A [Escherichia coli K12]
MAAQGFLILIATFLVVLMVLARPLGSGLARLINDIPLPGTTGVERVLFRALGVSDREMNWK
QYLCAILGLNMLGLAVLFFMLLGQHYLPLNPQQLPGLSVDLALNTAVSFVTNTNWQSYSG
ETTLSYFSQMAGLTVQNFLSAASGIAVIFALIRAFTRQSMSTLGNAWVDLLRIRILWVLVP
VALLIALFFIQQQGALQNLPYQAVNTVEGAGQQLLPMGPVASQEAIKMLGTNGGGFFNANS
SHPFENPTALTNTFQMLAIFLIPTALCFAFGEVMGDRRQGRMLLWAMSVIFVICVGVVMW
AEVQGNPHLLLALGTDSSINMSEGKESRGFGVLVSSLFAVTTAAASCAGAVIAMHDSFTALGGM
VPMWLMQIGEVVFGGVSGLYGMMLFVLLAVFIAGLMIGRTPEYLGGKIDVREMKLTLA
ILVTPTLVLMAALAMTDAGRSMNPGPHGFSEVLYAINESSANNSAFAGLSANSPF
WNCLLAFCMFVGRFGVIIKPVMAIAGSLVSRSQASSGTLPTHGPLLFGVLLIGTVLLVGA
LTFIPALALGPPVAEYLS
Goal of protein prediction

Epstein & Anfinsen, 1961: sequence uniquely determines structure

**INPUT:** protein sequence  
**OUTPUT:** protein structure & function
Colorful universe of protein structure

Colorful universe of protein structure 2 - assemblies

SLAC1 anion channel

HIV gp120

OCTN1

Crohn's disease, rheumatoid arthritis


Protein sequence

Protein ? - Mentha piperita - peppermint

MELLQLWSALIIILVVTTYTISLLIQWRKPKPQGFPPGPPKLPGLIGHLHLHLLWGLKLPLQHAL
ASVAKEYGPVAHVQLGEVFVSVVLSSREATKEAMKLVDPACANRFESIGTRIMWYDNEDII
FSPYSEHWRQMRKIVVSSELSSRNVRSGFIRQDEVSRLLRLHRSSAGAAVDMTERIELT
TCSIICRAAFGSVIRDNAELVGLVCLKDALSASGFELADMFPSSKLLNLCCWNKSKLWRMR
RRVDTILAEIVDEHKFKSGEFGGEDIIIDVLMRQMKTIQIKNPITTNSIKAFIFDTFSAG
TETSSTTTLWVLAMLMRNPAMAKAQAEVRAALEKETNWDVVDDVQELKYMKSVKETRMRA
HPPIPLIPRSCREECVVNGYTIPNKARIMINVWSSMRNPAMEKPDTFWPERFDQVSKDF
MGNDDEFVFPFGAGRRICPGLNFGLANVEVPLAQLLYHFDWKLAEGLKPSDMEDMSEAEGLT
GILKNLLLVPTPYDPSS
ATPase synthase

converts electrochemical potential into mechanical energy (stalk rotation)

ADP→ATP+rotation

1 cycle 3 Na+ out & 2 K+ in
1-10k/min; 80k-30m pumps per cell

© Proteopedia
Protein sequence

Protein ?

- Mentha piperita - peppermint
MELLQLWSALIIILVVTYISLLINQWRKPKPQGKFPPGPPKLPPLIQLIGHLHLLWGKLPLQPML
ASVAVEKEYGPVAHVQLGEVFSVVLSSREATKEAMKLVDPCANRFESIGTRIMWYDNEDII
FSPYSEHWRQMRKICVSELSSRNVRSGFIRQDEVSRLLRLHRRSSAGAAVDMTERIETL
TCSIICRAAFGSVIRDNAELVGLVDKALSASGFLADMFPPSSKLLNLLCNNKSKLWRMR
RRVDTILEAIVDEHKFKKSGEFGGEDIIIDVLFRMQKDTQIKVPIITNSIKAFIFDTFSAAG
TETSSTTTTLLWVLAELMRNPAMAKAQAEVRAALKEKTNWDVDVQELKYMKSVVKETMRRM
HPPIPLIPRSCREECVVNGYTIPKNKARIMINVINWSMGRNPLYEKPDFTWPERFDQVSKDF
MGNDFEFVFPGAGRRICPGLNFGLANVEVPLAQLLYHFDWKLAEGMKPSDMDEMSEAEGLT
GILKNNNLLLVTPTYPDPSS

Cytochrome P450

- Mentha spicata - spearmint
MELDLLSAIIIILVVTYISLLINQWRKSKSSQQLPLPSPPKLPVIGHLHLWGLPQHVFVSIAQKYGPVAHVQLGEVSVVLSSAEAAKQAMKLVDPNADRFDFGSDRTMDYDKDDIIF
SPYNDHWRQMRRICVTELELLSPKVRSFGYIRQEIEIERLIRLLGSGGAPVDVTEEVSKMES
CVVVCRAAFGSVLKDQGSLAEVVKSLALASGFLADLYPSSWLNLNSLNKRYLQRMRR
RLDHILDGFLFEHREKSGEFGGEIDVLFMRQKGSĐIKIPITSNCIKGFIFDTFSAAGA
ETSSTTISWALSELMRPAMAKVQAEVREALKGKTVDDLSEVQELKYLRSVLKETLRLH
PPFPLIPRQSREECVEVNGYTIPAKTRIFINVWAIGRDPQYWEDPTFRPERDFEVSRSDFM
GNDFFEFPFGAGRRICPGLHFGLANVEVPLAQLLYHFDWKLQPQGMDADLDMTETPGLS
GPKKKNVCLVPTLYKSP

© Burkhard Rost
ROSTLAB. TUM
Protein sequence

Cytochrome P450 - Mentha piperita - peppermint
MELLQLWSALIIILVVTYTISSLINOQWRKPKPQGKFPPGPGLPLGHLHLLWGBKLPQHAL
ASVAKEYGPVAVQVQLGEVFSVVLSSREATKEAMLKVDVPCANRFESIGTRIMWYDENII
FSPYESEHRQMRKICVSELLSSRNRSFGFIRQDEVSRLLRHLRSSAGAADMTERIETL
TCSIICRAAFGSVIRDNAELVGLVLDASMAGFELADMFPSSSLNLCCWNKSKLWRMR
RRVDITLEAVDEHKFKKSFGEDIDIVLFRMQKDTQIKVPITTSIKAFIFDTFSAG
TETSSTTLWLWLAELMRNPAVMAKAQAEVRAALKETKNDVDDVQELKYMKSVKETMRM
HPPIPLIPRSCEECVVNGYTIPNKARIMINVWSMRPLYWEKPDTFWPERFDQVSKDF
MGNDCEFVPFGARGRICPGGLNFLGLANVEVPLAQLLLYHFDWKLAEMKPSDMSDMSEAEGLT
GILKNLLLVLPTYPDSS

Cytochrome P450 - Mentha spicata - spearmint
MELDLSAIILILVATYIVSLINOQWRKSQSNLPPSPPKGVLPHGLHFLWGLPLQHVFR
SIAQKYGPVAVQVQLGEVYSVVLSSAEEAKVQAMKVLDPNFAQDFDGGSRTRMWYDKDIIIF
SPYNDHWRQMRICVTELLSPKNRSFGYIRQEEIERLIRLLGSSGGAPVPDYTEEVSKMS
CVVVCRAAAFSZVLDQGSLAESLALASFGELADLYPSSWLNLWLSNLNYRQLQMRMR
RLDHLIDGFLEEHHREEKKSFGEDIVDVLFRMQKGSIDKIPITSNCKGFIIFDTFSAGA
ETSSTISWASLSELMPPKMAKVQAEVREALKGKTVDLSEVQELKYLRSVLKETLRLH
PPFPIPRSQREECEVNFGYITIPAKTRIFINVWAIGRDQYPWEDPDTRPERFDKSFVRDFM
GNDFEFIPFGAGRIRCPGLHFLGLANVEIPLAQLLLYHFDWKLPQGMCDLDLMTEXTPGSLG
PKKKNVLVPTLYKSP
Can you define protein function?
Define protein function
Function of ADH

Alcohol dehydrogenase (ADH, EC number 1.1.1.1) is an 80kDa enzyme that catalyzes the 4th step in the metabolism of fructose before glycolysis. In the 4th step, glyceraldehyde is converted to the glycolytic intermediate DHAP by the NADH-dependent, ADH catalyzed reduction to glycerol.[1] ADH catalyzes the oxidation of primary and secondary alcohols to their corresponding aldehydes and ketones through a mechanism that involves the removal of a hydrogen. For detailed discussion of horse liver alcohol dehydrogenase see Horse Liver Alcohol Dehydrogenase.

http://www.proteopedia.org/wiki/index.php/Alcohol_dehydrogenase
Function of ADH

- Enzyme
- EC number 1.1.1.1
  - EC 1 Oxireductase
  - EC 1.1 CH-OH group of donors (alcohol oxidoreductases)
  - EC 1.1.1 with NAD or NADP+ as acceptor
  - EC 1.1.1.1 Aldehyde dehydrogenase
- 4th step in metabolism of fructose
- ADH catalyzes reduction to glycerol

$$\text{CH}_3\text{CH}_2\text{OH} + \text{NAD}^+ \rightarrow \text{CH}_3\text{COH} \text{ (acetaldehyde)} + \text{NADH} + \text{H}^+$$

(oxidation of alcohol to aldehyde in concert with transfer of a hydride to NAD)

http://www.proteopedia.org/wiki/index.php/Alcohol_dehydrogenase
Enzyme EC 1.1.1.1: Aldehyde dehydrogenase

Many pathways, e.g. glycolysis

KEGG

http://www.genome.jp/kegg-bin/show_pathway?ec00010+1.2.1.3
Protein function

Intuitive but not well-defined:
- chemical
- biochemical
- cellular (kinase)
- developmental
- physiological
- genetic

how atom bound?
transferase
cell cycle
time, regulatory
related to disease
dominant/recessive

Protein function as action:

\[ \text{Function} = \text{anything that happens to or through a protein} \]
Protein function: everything that happens to or through a protein
Diversity of function LEPTIN

↓ Food intake  
↑ Energy expenditure

By activating appetite-diminishing (anorexigenic) and stimulating (oxigenic) neuropeptides

Ana Luiza Arruda, TUM
What is protein function?
MOLECULAR CONSEQUENCES (ON P53)

- R175H: Metal-binding
- V143A: Stability
- K120R: Acetylation
- R273H: DNA-binding
- G245S: Protein-binding

p53 – tumor suppressor protein

PDB structures: 2ybg, 2j1w, 1ycs and 1tup
Sickle Cell Disease: Gain of Function

Sickle Cell Disease:
- Autosomal recessive disorder
- E6V in HBB causes interaction with F85 and L88
- Formation of fibrils
- Abnormally shaped red blood cells, leads to sickle cell anemia
- Manifestation of disease vastly different over patients


http://gingi.uchicago.edu/hbs2.html
Ilc:

sequence -> structure -> motion
Goal of protein prediction

Epstein & Anfinsen, 1961: sequence uniquely determines structure

INPUT: protein sequence

OUTPUT: protein structure & function

Christian B Anfinsen
Structure provides a scaffold for function

SLAC1 anion channel

HIV gp120

OCTN1

Crohn's disease, rheumatoid arthritis


Molecular dynamics:
function is motion
Protein movie

Sergio Decherchi, Anna Berteotti, Giovanni Bottegoni, Walter Rocchia & Andrea Cavalli (2014)
Nature Comm 6 doi:10.1038/ncomms7155
Molecular dynamics

simulations: courtesy of
Marco Punta (Pfam Sanger Inst Hinxton) &
Marco de Vivo (ISS Geneva)
High speed protein simulations with ANTON

David E Shaw of DE Shaw Research
Keynote @ U of Washington, Seattle
© UWTV/youtube (search: high speed protein simulations with anton)
Order
Disordered regions

Wu & Shi et al Science 2000 287(5450):92-7
Disorder in “ordered” form

A Schlessinger 2007 Thesis
D Petrey & B Honig 2003 Methods Enzymol 374:492-509 (GRASP2)
Eukaryotes reign by disorder?
Intro:
protein function
evolution-
terminology
Metrics for protein function
Predictions and analyses use metrics

- **Sequence:**
  - E-value, pairwise sequence identity, etc.

- **Structure:**
  - RMSD

- **Function?**
Function metric - Enzyme Nomenclature

EC1 oxidoreductases
EC2 transferases
EC3 hydrolases
EC4 lyases
EC5 isomerase
EC6 ligases

EC4.1 carbon-carbon
EC4.2 carbon-oxygen
EC4.3 carbon-nitrogen
EC4.4 carbon-sulfur
EC4.5 phosphorus-oxygen
EC4.99 others

EC4.1.1 pyruvate decarboxylase
EC4.1.2 oxolate decarboxylase
4.1.1.1 pyruvate decarboxylase
4.1.1.2 oxolate decarboxylase
54/156
GO: Gene Ontology

Three classes:

- Biological process
- Molecular function
- Cellular component/localization
Gene Ontology (GO): Biological process

© Marco Punta & Yanay Ofran
GO: Molecular function


© Burkhard Rost
GO: Cellular component/localization


© Marco Punta & Yanay Ofran
GO: Gene Ontology

Three classes:

• Biological process
  e.g. cell cycle control or signal transduction

• Molecular function
  e.g. RNA-binding or “is enzyme”

• Cellular component/localization
  e.g. extra-cellular space or nuclear matrix

Sequence determines structure determines function
Sequence determines structure

similar sequence \( \Rightarrow \) similar structure
Annotation transfer from structure

Similar Structure

Dissimilar Structure


© Marco Punta & Yanay Ofran
Power of comparative modeling

Structure prediction from sequence

Sequence-function asymmetry

similar sequence $\Rightarrow$ similar function
Same sequence, different tissues
-> different function
## Moonlighting proteins

<table>
<thead>
<tr>
<th>One function</th>
<th>Additional functions</th>
<th>Rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>PutA prolne dehydrogenase</td>
<td>Transcriptional repressor</td>
<td>1, 2</td>
</tr>
<tr>
<td>Phosphoglucone isomerase</td>
<td>Neuroleukin, autocrine motility factor, differentiation and maturation mediator</td>
<td>3, 8</td>
</tr>
<tr>
<td>Thymidine phosphorylase</td>
<td>Platelet-derived endothelial cell growth factor</td>
<td>9</td>
</tr>
<tr>
<td>Neuropilin (VEGF receptor)</td>
<td>Receptor for semaphorin III (nerve axons)</td>
<td>10</td>
</tr>
<tr>
<td>Uracil-DNA glycosylase</td>
<td>Glyceraldyde-3-phosphate dehydrogenase</td>
<td>11</td>
</tr>
<tr>
<td>Aconitase</td>
<td>Iron-responsive-element binding protein (IRE-BP)</td>
<td>12, 15</td>
</tr>
<tr>
<td>Carbinolamine dehydratase</td>
<td>Dimerization cofactor (DcoH)</td>
<td>16</td>
</tr>
<tr>
<td><em>Escherichia coli</em> thioredoxin</td>
<td>Subunit of T7 DNA polymerase</td>
<td>17</td>
</tr>
<tr>
<td><em>E. coli</em> aspartate receptor</td>
<td>Maltose-binding-protein receptor</td>
<td>13, 14</td>
</tr>
<tr>
<td>PMS2 mismatch-repair enzyme</td>
<td>Hypermution of antibody variable chains</td>
<td>24</td>
</tr>
<tr>
<td>Ribosomal proteins</td>
<td>DNA repair, translational regulators, development regulators, etc.</td>
<td>25</td>
</tr>
<tr>
<td>Lens crystallins</td>
<td>Heat-shock proteins, lactate dehydrogenase, argininosuccinate, retinaldehyde dehydrogenase, lyase, enolase, quinone oxidoreductase, glyceraldyde-3-phosphate dehydrogenase, etc.</td>
<td>26</td>
</tr>
<tr>
<td>CFTR chloride channel</td>
<td>Regulator of other epithelial anion channels</td>
<td>18</td>
</tr>
<tr>
<td>P-glycoprotein (transporter)</td>
<td>Regulator of cell-swelling ion channel</td>
<td>19, 20</td>
</tr>
<tr>
<td>Thrombin protease</td>
<td>Ligand for cell surface receptors</td>
<td>21</td>
</tr>
<tr>
<td>Thymidylate synthase</td>
<td>Translation inhibitor</td>
<td>22</td>
</tr>
<tr>
<td><em>E. coli</em> biotA biotin synthetase</td>
<td>Bio operon repressor</td>
<td>23</td>
</tr>
<tr>
<td>Mitochondrial ION protease</td>
<td>Chaperone</td>
<td>27</td>
</tr>
<tr>
<td>Bacterial FtsH chaperone</td>
<td>Metalloprotease</td>
<td></td>
</tr>
<tr>
<td>Band-3 anion exchanger</td>
<td>Regulator of glycolysis</td>
<td></td>
</tr>
</tbody>
</table>

Moonlighting proteins

Sequence-function asymmetry
Different structures -> same function

Chymotrypsin (5cha)  
Subtilin (5sic)
“The domain problem”

PSEUDOURIDINE SYNTHASE

HYPOTHETICAL PROTEIN

ATP SULFURYLASE
I2: Homology-based inference
General challenges for homology-based inference of function
ADH: Alcohol dehydrogenase

© Wikipedia  ADH5 PDBid: 1m6h

Human glutathione-dependent formaldehyde dehydrogenase

ADH: Alcohol dehydrogenase

© Wikipedia

alcohol dehydrogenase

Crystallographic structure of the homodimer of human ADH5.[1]

Identifiers

| EC number  | 1.1.1.1 |  
| CAS number | 9031-72-5 |  

Databases

| IntEnz     | IntEnz view |  
| BRENDA     | BRENDA entry |  
| ExPASy     | NiceZyme view |  
| KEGG       | KEGG entry |  
| MetaCyc    | metabolic pathway |  
| PRIAM      | profile |  
| PDB structures | RCSB PDB PDBe |  
| Gene Ontology | AmiGO EGO |  

Search

© GO

View this term in QuickGO.

QuickGO - http://www.ebi.ac.uk/QuickGO

operational definition: homologs have common ancestor
Homolog

common ancestor tree might be

ADH1-yeast
ADH1-bacteria
ADH1-human
ADH1-plants

operational definition:
homologs have common ancestor
Homolog

common ancestor

? ADH1-bacteria

? ADH1-yeast

? ADH1-human

? ADH1-plants

operational definition:

homologs have common ancestor
Homolog, Ortholog

**operational definition:**
orthologs = homologs with corresponding function
Homolog, Ortholog, Paralogs

**common ancestor**

- ADH1-bacteria
- ADH1-yeast
- ADH1-plants
- ADH2-plants

**operational definition:**

- orthologs: similar function
- paralogs: diverged in function
Homolog, Ortholog, Paralogs, convergent

NO common ancestor

Tree hypothetical

ADH1-human
ADH1-plants
ADH1-yeast
ADH1-bacteria
ADH2b-plants
Translation of terms to proteins

- homologous proteins: are related
- orthologs have similar function
- paralogs may evolve a different function
Tree uncertain, but story more complicated for proteins:
problem 1: genes/proteins do not “reproduce”
problem 2: domains decoupled
“The domain problem”
problem 3: moonlighting
Homology-based inference: concept
Homology transfer accurate for very similar proteins

TRUE

methyltransferase

identity protein
100% guanidinoacetate N-methyltransferase
Homology transfer accurate for very similar proteins

<table>
<thead>
<tr>
<th>Protein</th>
<th>Identity</th>
<th>Protein</th>
<th>Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>methyltransferase</td>
<td>100%</td>
<td>guanidinoacetate N-methyltransferase</td>
<td>99%</td>
</tr>
<tr>
<td>100% guanidinoacetate N-methyltransferase</td>
<td></td>
<td>magnesium protoporphyrin IX methyltransferase</td>
<td></td>
</tr>
</tbody>
</table>
| 99% magnesium protoporphyrin IX methyltransferase

TRUE

**methyltransferase**

identity    protein
100% guanidinoacetate N-methyltransferase
99% magnesium protoporphyrin IX methyltransferase

FALSE
Homology transfer accurate for very similar proteins

<table>
<thead>
<tr>
<th>methyltransferase</th>
<th>identity</th>
<th>protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% guanidinoacetate N-methyltransferase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>99% magnesium protoporphyrin IX methyltransferase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>70% phosphoribosylglycinamidase formyltransferase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>66% inositol 3-methyltransferase</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2/3 accuracy ; 2/4 coverage
Homology transfer accurate for very similar proteins

<table>
<thead>
<tr>
<th>Identity</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>guanidinoacetate N-methyltransferase</td>
</tr>
<tr>
<td>99%</td>
<td>magnesium protoporphyrin IX methyltransferase</td>
</tr>
<tr>
<td>70%</td>
<td>phosphoribosylglycinamide formyltransferase</td>
</tr>
<tr>
<td>66%</td>
<td>inositol 3-methyltransferase</td>
</tr>
<tr>
<td>65%</td>
<td>phosphoribosylglycinamide formyltransferase</td>
</tr>
<tr>
<td>63%</td>
<td>aspartate carbamoyltransferase</td>
</tr>
<tr>
<td>62%</td>
<td>glycine amidinotransferase</td>
</tr>
<tr>
<td>61%</td>
<td>inositol 3-methyltransferase</td>
</tr>
</tbody>
</table>

**TRUE**

methyltransferase

2/3 accuracy ; 2/4 coverage

**FALSE**

2/3 accuracy ; 2/4 coverage
what’s better?

66% acc @ 50% cov

or

38% acc @ 100% cov
Specific challenges
(homology-based inference of function):
sub-cellular localization
Goal: predict sub-cellular localization

Predict sub-cellular localization
Zones

- Midnight Zone
- Twilight Zone
- Save Zone

Sequences similar
Structures similar
Performance of homology-based inference
Known-localization all-against-all ok?

proteins of known localization (SWISS-PROT)
Databases biased: MUST remove bias!

all proteins of known localization

sequence-unique subset
Homology-based inference structure

Structure

Homology-based inference of localisation

dots pairs of proteins with
● different
× same localization
Homology-based inference depends on structural and subcellular localization.

R Nair and B Rost (2002) *Protein Science* 11: 2836-47
Homology inference localization

R Nair and B Rost (2002) *Protein Science* 11: 2836-47
Homology inference localization

R Nair and B Rost (2002) *Protein Science* 11: 2836-47
Homology inference localization
Homology-based inference: Conservation of enzymatic activity
Detour: conservation of enzymatic activity

☐ How well is enzymatic activity conserved?
☐ Can we predict enzymatic activity by homology?
☐ Can we predict that a protein is an enzyme?
quick guess: homology inference
enzyme – localization
which works better?
**Enzyme classification (EC)**

(http://www.chem.qmw.ac.uk/iubmb/enzyme/)

<table>
<thead>
<tr>
<th>Table 1. Description of the different levels in the EC classification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First figure</strong></td>
</tr>
<tr>
<td><strong>A. OXIDOREDUCTASES</strong></td>
</tr>
<tr>
<td>Substrate is oxidised-regarded as the hydrogen or electron donor</td>
</tr>
<tr>
<td><strong>B. TRANSFERASES</strong></td>
</tr>
<tr>
<td>Transfer of a group from one substrate to another</td>
</tr>
<tr>
<td><strong>C. HYDROLASES</strong></td>
</tr>
<tr>
<td>Hydrolytic cleavage of bond</td>
</tr>
<tr>
<td><strong>D. LYASES</strong></td>
</tr>
<tr>
<td>Cleavage of bonds by elimination</td>
</tr>
<tr>
<td><strong>E. ISOMERASES</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>F. LIGASES</strong></td>
</tr>
<tr>
<td>Enzyme catalysing the joining of two molecules in concert with hydrolysis of ATP</td>
</tr>
</tbody>
</table>

An enzyme reaction is assigned a four-digit EC number, where the first digit denotes the class of reaction. Note that the meaning of subsequent levels depends upon the primary number, e.g. the substrate acted upon by the enzyme is described at the second level for oxidoreductases, whereas it is described at the third level for hydrolases. Different enzymes clustered together at the third level are given a unique fourth number, and these enzymes may differ in substrate/product specificity or cofactor-dependency, for example. Peptidases (EC 3.4.1.1) have a different classification scheme (Barrett, 1994). Note also that it is a classification of overall enzyme reactions, and not enzymes, and takes no account of the details of the reaction chemistry involved (see caveats below).
EC: Enzyme Commission number

oxido-reductases

transferases

hydrolases

lyases

isomerase

ligases

carbon-carbon

carbon-oxygen

carbon-nitrogen

carbon-sulfur

phosphorus-oxygen

others

carboxy lyases

aldehyde lyases

oxo-acid lyases

other carbon-carbon lyases

Similar reaction/different structure


*Figure 7. MOLSCRIPT (Kraulis, 1991) diagrams of the homologous enzymes (a) chloramphenicol acetyltransferase (PaXAT), and (b) UDP-N-acetylglucosamine acyltransferase (LpxA). The catalytic histidine residues putatively involved in deprotonation of the substrate hydroxyl are shown in ball-and-stick and circled in blue.*
Measuring conservation of enzymatic activity

All proteins

all enzymes

enzymes known function
Conservation of function

Devos & Valencia 2000 *Proteins* 41, 98-107
Measuring conservation of enzymatic activity

All proteins

all enzymes

enzymes known function
Measuring conservation of enzymatic activity

Enzymes of known function

unique

All proteins
George Kingsley Zipf

- American, born Freeport IL
- Linguist and philologist
- Studied Univ Bonn & Berlin

Zipf law: word frequencies:

\[ P_n \sim 1 / n^a \]

\( P_n \) frequency of word ranked \( n \)th
\( a \sim 1 \)
Zipf’s law

\[ y = \frac{1}{x} \]
If you plotted the histogram of settlement sizes, how would that look?
Sizes of metropolitan areas in the USA


Figure I
Log Size versus Log Rank of the 135 largest U. S. Metropolitan Areas in 1991
Source: Statistical Abstract of the United States [1993].
Pick points at random: then what?
Half a Zipf is a Zipf
Expected distribution of protein families?

Number of proteins per family

Histogram (number of families with that number of proteins)

Number of proteins per family
Expected distribution of protein families?

Histogram (number of families with that number of proteins)

Number of proteins per family
Family size

Number of proteins in family

Percentage of families (biased)

Percentage of families (unbiased)

B Rost 2002 J Mol Biol 318, 595-608
REAL conservation of EC number

Accuracy 1st: bias
Accuracy all: bias

Bias 1st EC
Bias 4 EC

Percentage of protein pairs

Percentage pairwise sequence identity

B Rost 2002 J Mol Biol 318, 595-608
REAL conservation of EC number

B Rost 2002 J Mol Biol 318, 595-608
REAL conservation of EC number

bias:
50% found at >90% right

real:
50% found at <15% right!

B Rost 2002 J Mol Biol 318, 595-608
sequence identity bad!
Conservation in detail

B Rost *J Mol Biol* **318**, 595-608
Conservation of EC: PSI- vs. pair-BLAST

- **First EC digit:**
  - Accuracy
  - Coverage

- **All EC digits:**
  - Accuracy
  - Coverage

Number of proteins
- **PAIR-BLAST**
- **PAIR**
- **PSI-BLAST**
- **PSI**

Distance from threshold (identity/length)

Corresponding percentage sequence identity

Log(BLAST E)

© Burkhard Rost
Statistical scores better when statistics kick in.

R Nair & B Rost 2002 Protein Science 11, 2836-47
B Rost 2002 J Mol Biol 318, 595-608
B Rost 1999 Prot Engng 12, 85-94
Homology-based inference: Cell cycle control
Kazimierz O. Wrzeszczynski

Kazimierz O. Wrzeszczynski (now NYGC)
Cell Cycle Control and Data Set

![Cell cycle diagram](image)

### Numbers of Cell Cycle Control Proteins Found in SWISS-PROT

<table>
<thead>
<tr>
<th>Species</th>
<th>Cell Cycle Control</th>
<th>G1/S</th>
<th>G2/M</th>
<th>M Phase</th>
<th>S Phase</th>
<th>Other</th>
<th>Multiple</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eukaryotes</td>
<td>582</td>
<td>135</td>
<td>86</td>
<td>66</td>
<td>156</td>
<td>229</td>
<td>90</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>99</td>
<td>28</td>
<td>11</td>
<td>23</td>
<td>41</td>
<td>24</td>
<td>28</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>68</td>
<td>25</td>
<td>8</td>
<td>10</td>
<td>30</td>
<td>18</td>
<td>23</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>15</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>10</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>87</td>
<td>20</td>
<td>11</td>
<td>5</td>
<td>19</td>
<td>46</td>
<td>14</td>
</tr>
</tbody>
</table>
## Thresholds for cell cycle annotation

### Sequence Identity

<table>
<thead>
<tr>
<th>Seq id.</th>
<th>Accu.</th>
<th>Cov.</th>
</tr>
</thead>
<tbody>
<tr>
<td>53</td>
<td>27%</td>
<td>38%</td>
</tr>
</tbody>
</table>

### HSSP Distance

<table>
<thead>
<tr>
<th>HSSP D</th>
<th>Accu.</th>
<th>Cov.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>70%</td>
<td>58%</td>
</tr>
</tbody>
</table>

### PSI-BLAST E-value

<table>
<thead>
<tr>
<th>E-value</th>
<th>Accu.</th>
<th>Cov.</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^-8</td>
<td>64%</td>
<td>60%</td>
</tr>
</tbody>
</table>

---

**True Positive:** 582 Cell Cycle Control Proteins

**True Negative:** 15,192 w/ No Cell Cycle Annotation

| Seq id. = 53 | Accu. 27% - Cov. 38% |
| HSSP D = 5   | Accu. 70% - Cov. 58% |
| E-value = 10^-8 | Accu. 64% - Cov. 60% |
# Cell cycle vs. EC class inference

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Cell Cycle</th>
<th>1st EC level</th>
<th>4th EC level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E-value</td>
<td>Accuracy</td>
<td>1.x.x.x</td>
</tr>
<tr>
<td>$E \times 10^{-8}$</td>
<td>64%</td>
<td>98%</td>
<td>50%</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>60%</td>
<td>85%</td>
</tr>
<tr>
<td>$E \times 10^{-3}$</td>
<td>29%</td>
<td>95%</td>
<td>48%</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>81%</td>
<td>80%</td>
</tr>
<tr>
<td>HSSP</td>
<td>Accuracy</td>
<td>70%</td>
<td>99%</td>
</tr>
<tr>
<td>$D = 5$</td>
<td>Coverage</td>
<td>58%</td>
<td>82%</td>
</tr>
<tr>
<td>HSSP</td>
<td>Accuracy</td>
<td>25%</td>
<td>90%</td>
</tr>
<tr>
<td>$D = -5$</td>
<td>Coverage</td>
<td>81%</td>
<td>55%</td>
</tr>
</tbody>
</table>

EC - Enzyme Classification:

EC 2.x.x.x Transferase
EC 2.7.x.x Transferring phosphorus-containing groups
EC 2.7.11.x Protein-serine/threonine kinases
EC 2.7.11.22 Cyclin-dependent kinase

Discover new cell cycle control proteins

<table>
<thead>
<tr>
<th>Proteome</th>
<th>Known cell cycle control proteins</th>
<th>Predicted cell cycle control proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D=0 (55%)</td>
<td>D=15 (65%)</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>99</td>
<td>3073</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>68</td>
<td>3162</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>15</td>
<td>970</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>10</td>
<td>1005</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>5</td>
<td>1888</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>87</td>
<td>513</td>
</tr>
<tr>
<td>Sum</td>
<td>284</td>
<td>10611</td>
</tr>
</tbody>
</table>

1 Distance from HSSP-Threshold chosen as seen in Fig. 2 for various levels of percent accuracy using the PSI-BLAST curve. Levels of accuracy are estimated according to Fig. 2, e.g. at a threshold of D=40 more than 95% of the proteins for which we infer the involvement in cell cycle control by homology are supposedly correctly inferred.

2 The number of previously known annotated cell cycle control proteins represented in each specific proteome as used in our trusted data set is given for comparison.
# CellCycleDB

**CellCycleDB** (Database of Cell Cyle Control Proteins in Eukaryotes)

<table>
<thead>
<tr>
<th>It is</th>
<th>CellCycleDB Catalogues proteins involved in the Cell Cycle Control Process through homology transfer from experimental annotations.</th>
</tr>
</thead>
<tbody>
<tr>
<td>It does</td>
<td>CellCycleDB allows the user to submit a protein sequence to determine estimates for involvement in the cell cycle process or search CellCycleDB for predicted cell cycle proteins among six eukaryotic proteomes. CellCycleDB is currently a first detailed analysis through homology assignment for identifying proteins functioning in the cell cycle process focusing on cell cycle control. Single sequence queries are evaluated against a trusted annotated data set of experimentally identified cell cycle control proteins. An overall accuracy estimate for involvement in the cell cycle process based on HSSP-distance threshold values is presented for any specific query. CellCycleDB provides various accuracy levels for cell cycle function assignment of all proteins among six eukaryotic proteomes.</td>
</tr>
<tr>
<td>You can</td>
<td></td>
</tr>
<tr>
<td>• Use CellCycleDB <strong>online</strong> <em>(currently: single protein sequence submissions only)</em></td>
<td></td>
</tr>
<tr>
<td>• Search CellCycleDB using SRS: CellCycleDB</td>
<td></td>
</tr>
<tr>
<td>• <strong>download</strong> the CellCycleDB</td>
<td></td>
</tr>
<tr>
<td>• CellCycleDB Content Summary: Content Summary Tables</td>
<td></td>
</tr>
<tr>
<td>From Here</td>
<td><strong>Who are we?</strong></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Contact</td>
<td>Email: <a href="mailto:kaz@cubic.bioc.columbia.edu">kaz@cubic.bioc.columbia.edu</a></td>
</tr>
<tr>
<td></td>
<td>Tel: +1-212-305-4018 (Kazimierz O. Wrzeszczynski)</td>
</tr>
<tr>
<td></td>
<td>Fax: +1-212-305-7932</td>
</tr>
</tbody>
</table>
Homology-based inference: how much of human?
Homology transfer accurate for very similar proteins

![Graph showing the accuracy of homology transfer with respect to percentage sequence identity. The graph illustrates the percentage conservation/accuracy and coverage for enzymatic activity and localization.]
Some problems of homology transfer

- not all annotations as informative as “methyltransferase”

ID  1433_TRIHA   STANDARD;  PRT;  262 AA.
DE  14-3-3 PROTEIN HOMOLOG (TH1433).
CC  -!- DEVELOPMENTAL STAGE: HIGHEST EXPRESSION DURING THE ACTIVE GROWTH PERIOD 10-12 HOURS AFTER GERMINATION.
CC  -!- SIMILARITY: BELONGS TO THE 14-3-3 FAMILY.

- 70% multi-domain proteins

Schlessinger unpublished

Less than 25% have *some* annotation coverage of homology transfer

< 10-25%

we clearly need something more!

B Rost, Nair, Liu, Wrzeszczynski & Ofran (2003) *CMLS* 60: 2637-50
### Four problems for function annotation

<table>
<thead>
<tr>
<th>Paralogy problem</th>
<th>Moonlighting problem</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>query</strong></td>
<td><strong>query</strong></td>
</tr>
<tr>
<td>template</td>
<td>template</td>
</tr>
</tbody>
</table>

*Template is a paralog, likely diverged functionally*

*Template may have more than one function*

<table>
<thead>
<tr>
<th>Multi-domain problem</th>
<th>Database annotation problem</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>query</strong></td>
<td><strong>query</strong></td>
</tr>
<tr>
<td>template</td>
<td>template</td>
</tr>
</tbody>
</table>

*Template annotation based on other domain*

*Template mis-annotated, possibly due to multi-domain problem*
Evolutionary profile capture information
Today: Intro into function

**LAST WEEKs**
- **Tue:**
- **Thu:** Introduction to protein function I

**THIS WEEK**
- **Tue:** Introduction to protein function II
- **Thu:** CAFA prediction of protein function

**NEXT WEEK**
- **Tue:** No lecture - SVV (student rep meeting)
- **Thu:** Introduction to protein function III - motifs
preliminary Lecture plan (PP2 function)

01: 2018/10/16: No lecture (makeup examen; PP last year)
02: 2018/10/18: No lecture (makeup)
03: 2018/10/23: Welcome: who we are
04: 2018/10/25: Intro function 1: concept of protein function
05: 2018/10/30: No lecture (SVV)
06: 2018/11/01: No lecture (holiday, All Saints)
07: 2018/11/06: Localization 1 (chalk talk)
08: 2018/11/08: Localization 2
09: 2018/11/13: Localization 3
11: 2018/11/20: PPI 1 - interaction sites (chalk talk)
12: 2018/11/22: No lecture (Thanksgiving)
14: 2018/11/29: PPI 3 - sites/ protein-DNA/RNA
15: 2018/12/04: PPI 4 - PPI pairing(chalk)
16: 2018/12/06: No lecture (Dies Academicus)
17: 2018/12/11: PPI 5 - PPI pairing
18: 2018/12/13: PPI 6 - PPI pairing
19: 2018/12/18: Motifs
20: 2018/12/20: No lecture
21-24: no lectures - winter break (2018/12/24 - 2019/01/06)
25: 2019/01/08: SNP effect 1 (chalk talk)
26: 2019/01/10: SNP effect 2
27: 2019/01/15: SNP effect 3
28: 2019/01/17: SNP effect 4
29: 2019/01/22: WRAP up 1
30: 2019/01/24: WRAP up 2
31: 2019/01/29: WRAP up 3
32: 2019/01/31: ?
33: 2019/02/05: ?
34: 2019/02/07: Examen (10:00-13:00, lecture room LMU physics)