title: I1: Introduction to protein function

short title: cb2_intro_func1

lecture: Protein Prediction 2 (for Computer Science) - Protein function
TUM winter semester
Announcements

☐ Videos:  YouTube / www.rostlab.org

THANKS:

Dmitrij Nechaev

☐ 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:  no retake

CONTACT: teaching@rostlab.org
THANKS for slides!

© Michael Leunig: The Traveling Leunig
Yanay Ofran

now
Bar Ilan Univ
Israel
Marco Punta (now Paris) & Avner Schlessinger (now: Mount Sinai)
Kazimierz O. Wrzeszczynski
(now NY Genome Center)
I1
Intro function
I1a
Intro biology: genes and numbers
(repetition)
Life begins how?
RNA at the beginning

Trouble: missing some enzymatic activity

RNA at the beginning

enslave proteins to get the job done

RNA at the beginning

How to proceed to the next level?

ocean + RNA + proteins +?
life is about self
Simple cell

Illustration by David S. Goodsell, the Scripps Research Institute, UCSD, USA
how densely packed are cells?
Mycoplasma genitalium

Protein synthesis (labels in black)
1. DNA
2. DNA polymerase
3. single-stranded-DNA binding protein (protects single-stranded portions during replication)
4. RNA polymerase
5. messenger RNA
6. ribosome
7. transfer RNA (in pink) and elongation factor Tu and Ts
8. elongation factor G
9. aminoacyl-tRNA synthetases
10. topoisomerases
11. Rec system for DNA repair: a) RecA, b) RecBC
12. chaperonin GroEL (helps folding of new proteins)
13. proteasome ClpA (destroys old proteins)

Enzymes for energy production (labels in red)
14. glycolytic enzymes
15. pyruvate dehydrogenase complex

Membrane proteins (labels in blue)
17. ATP synthase
18. secretory proteins
19. sodium pump
20. zinc transporter
21. magnesium transporter
22. ABC transporter (different ABC transporters transport different types of molecules-ABC is short for "ATP-binding cassette")
23. magnesium transporter
24. lypoglycan (long carbohydrate chains)

Illustration by David S. Goodsell, the Scripps Research Institute, UCSD, USA
Central dogma of molecular biology

DNA → RNA → Protein

information, code, library, manual

intermediate step

machinery of life
<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mycoplasma genitalium</td>
<td>470</td>
</tr>
<tr>
<td>Haemophilus influenzae</td>
<td>1,740</td>
</tr>
<tr>
<td>Methanococcus jannaschi</td>
<td>1,738</td>
</tr>
<tr>
<td>Escherichia coli</td>
<td>4,288</td>
</tr>
<tr>
<td>Sacharomyces cerevisiae - yeast</td>
<td>5,800</td>
</tr>
<tr>
<td>Drosophila melanogaster - fruit fly</td>
<td>15,500</td>
</tr>
<tr>
<td>Caenorhabditis elegans - worm</td>
<td>20,500</td>
</tr>
<tr>
<td>Arabidopsis thaliana - mustard</td>
<td>27,000</td>
</tr>
<tr>
<td>Oryza sativa - rice</td>
<td>50,000</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>20,000</td>
</tr>
</tbody>
</table>
Central dogma of molecular biology

DNA → RNA → Protein

- information, code, library, manual
- intermediate step
- machinery of life

Ribosomes (translation = peptide synthesis) can function without proteins
I1b

Intro:

protein function
Predict protein function?
what is a protein?
Protein sequence

>gi|16128674|ref|NP_415226.1| potassium translocating ATPase, subunit A [Escherichia coli K12]
MAAQGFLLIATFLLVLMVLARPLGSGLARLINDIPLPGTTGVERVLFRALGVDREMNWK
QYLCAILGLNMLGLAVLFFMLLGQHYLPLNPQQLPGLSVDLALNTAVSFVNTTNWQSYSG
ETTLSYFSQMAGLTVQNFLSAASGIAVIFALIRAFTSRQSMTLGNAWVDLLRITLWVLVP
VALLIALFFIQQGALQNFLPYQAVNTVEGAQQLLPMGPVASQEAIKMLGTNGGFFNANS
SHPFENPTALTNFVQMMLAIFLIPTALCFAFGEVMGDRRQGRMLLWAMSVIFVICVGVVMW
AEVQGPNHLLALGTDSSINMEGKESRFGVLVSSLFAVVTAAASCGAVIAMHDSFTALGGM
VPMWLMQIGEVVFVGGVGSGLYGMMLFVLLAVFIAGLMIGRTPEYLGKIDVREMKTALAL
ILVTPTLVLGMGAALAMMTDAGRSAMLNPGBFHGFSEVLYAYSSAANNGSAGFLSANSFP
WNCLLAFCMFVGRFGVIIIPVMAIAGSLVSKKSQAASSGTLPHTHGPLFVGLLIGTVLLVGA
LTIFIPALALGPVAEYLS
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

>gi|16128674|ref|NP_415226.1| potassium translocating ATPase, subunit A [Escherichia coli K12]
MAAQGFLLIATFLLVLMVLARPLGSGLARLINDIPLPGTTGVERVLFRA"GVDREM"NWKQYLCAILGLNMLGLAVLFFMMLGQHYLPLNPQLPGLSDLALNNTAVSFVNTNTNWQSYSGETTLSYFSQMAGLTVQNFLSAASGIAVIFALIRAFTRQSMSTLGNAWVDLLRRITLWVLPVALLIALFFIQQGALQNFLPYQAVNTVEGAQQLPMPGVPASQEAIKMLGTNGGGFFNANSHPFENPTALTNFVQMLAIFLIPTALCFAFGEVMGDRRQGRMLLWAMSVIFVICVGVMWAEVQGNPHLLALGTDSSINMEGKESRFVGLVSSLFAVVTTAASCGAVIAMHDSFTALGGMVPMWLMQIGEVVFGGGVGSGLYGMMLFVLAVFIAGLMIGRTPEYLGKIDVREMKLTAAILVTPLLVMGAALAMMTDAGRSMNPQPHGFSEVLYAYSSAANNNGSAFAGLSANSPFWNCLLAFCMFVGRFVIIIPVMAIAAGSLVSKKSQAASSGTLPTHGPLFVGLLIGTVLTVGALTFlPALALGVPVAEYLS
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
MELLQLWSALIILVVTVYTISSLINQWRKPQPKPQKFPPGPPKLPLIGHLHLWWGKLPLQHAL
ASVAKEYGPVAVHQLGEVSLVSSTREATKEAMKLVDPCAANRFESIGTRIMWYDNEDII
FSFSEHWRQMRKICVSELLSRNVRSFGRIFQDEVSRLLRHLRSSAGAAVMTERIETL
TCSIIICRAAFGVIRDNAELVLGKLDSMAGFELADMFPSKLLNLLCWNKSLWHRMR
RRVDTILEAIVDEHKFKKSGEGGEDIIDVLFMRQKDTQIKVPTTTNSIKAFIFDTFSAAG
TETSSTTLWVLALMRNPAVMAKADQAEMRALEKTNWDVDDVQELKYMKSVVKETMRM
HPPIPLIPRSCREECVVNGYTVIPNKARIMINWMSGHRPNLYWEPDPERFDQVSKDF
MGNDEFVFPGAGGGICPGAGNLGFGLANVEVPLAQLLYHDFWKLAEOMKPSDMDDMSEAEGLT
GILKNNLLLVPTPYDSS

Cytochrome P450
- Mentha spicata - spearmint
MELDLSAIIILVTYIVSLINQWRKSQSQQNLPPSPPKLPLVIGHLHLWGGLPLQHVF
SIAQKYGPAVAVQIGEVSSVSSLASAAAKQAMKVLDPNFADRDGISSRTMWYDKDDIIIF
SPYNHZHRQMRRICVTELLSSPKNSQFQYIRQEEEIERLIRLLGSSGAPVDFVTEVSISMS
CVVVCRAAFGVSLKDGQSLALVKESSLALSGFELADLYPSSWLLNLSLKRYRLQRMRR
RLDHILOFGLEYHEKKGSGEGGEDIVDVLFMRQKQGSDIKPITSNCSIKGFIFDTFSAAG
ETSSTTSAWSLREMNPANKAVQAEVREALKGTVDVSLSEVQELKYLRSVLFKELRLHM
PPFPLIPRQSREECEVNGYTVIPAKTRIFINVWAIARPQYWEDPDTPFRPERFDEVSRDFM
GNDGFIFPGAGGGICPGAGNLGFGLANVEVPLAQLLYHDFWKLQPGMDADLDMTETPGSLG
PKKKNVCLVPTLYKSP
Protein sequence

Cytochrome P450 - Mentha piperita - peppermint
MELLQLWSALIIILVVTYITISLLINQWRKPKPQGKFPPGGPPKPLPLIGHLHLLWGKLPQHAL
ASVAKEYGPVAHVQLGEVFSVVLSSREATKEAMLKVDPACANRFESIGTRIMWYDNEDII
FSPYSEHWRQMRKICVSELLSSRNRSFGIFQRDEVSRLLRHLRSSAGAADVDMTERIETL
TCSIICRAAFTSGVIRDNAELVGLVKDALSASGFEADMFPSKLLNLCCWKNSTKLWMMR
RRVDTILEAIVDEHKFKKGFEFEDIIIDVLFRMQKDTQIKVPICTWNISIKAFIFDTFSAG
TETSSTTLWVLALMRNPAVMAKAQAETVAALKEKTNWDVDDVQELKYMKSVVKETMRR
HPPIPIPLIPRSCRCEECCVVNGYTIIPNKAIRIMINVWSMGPNKYWEKPDFTFWPERFDQVSKDF
MGNDFEFVPFGAGRRICPGNLNGFLAVELNVEPLQLLLHFDWKLAEKGKPSDMDMSEAEGLT
GILKNLLLVPTYPDSS

Cytochrome P450 - Mentha spicata - spearmint
MELDLSAIIILVATYIVSSLINQWRKSKSQNLPSPPPKLPVIGHLHFLWGLLPQHVFR
SIAQKYGPAHVQLGEVYSVVLSAEAAKQAMKVLDPNFADRFDFGDSRTMWYDKDDIIIF
SPYNDHWRQMRVICVTELLSPKNVRSGYIRQEEIERLIRLLGSGGAPVDTEEVSKMS
CVVVCRAAFGSVKDQGSLAEDELKESLALASGFEADLYPSSWLNNLSSLNKRILQRMRR
RLDHIKDFLEEHREKKSCGFEFGEDIVDVLFRMQKGSIDIKPITSNCKKGFIIFDTFSAGA
ETSSTISWALSELMRPMAKVQAEVGREALKGTVDFLSEVQELKYLRSVFLKETLRLH
PPFLIPRQSREECECVNVYTIIPAKTRIFINWAIGRDQPQYWEDPDTFRPERFDDEVSRDFM
GNDQEFIPFGAGRRICPGNLHFLANVEIPLAQLLYHFDWKLPQGMTDADLDMTETPGSLG
PKKKNVCLVPTLYKSP
Proteins - genetic code

From the book: “DNA: The Secret of Life” by James Watson and Andrew Berry

http://img.sparknotes.com/figures/1/132572f07836571a1eb4f16b3/53/15a3/15.png

© Burkhard Rost
Some facts about proteins

- How many in human?
  - 20-25K in human

- How long are they?
  - ~35-30k, median around 400
Protein sequences: dime a dozen

UniProt
length distribution
(from 40-40,000)

2013/06: 35 Million
## Big data - big CPU

<table>
<thead>
<tr>
<th></th>
<th>Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009/09</td>
<td>$8 \cdot 10^6$</td>
</tr>
<tr>
<td>2013/06</td>
<td>$35 \cdot 10^6$</td>
</tr>
</tbody>
</table>

- double every 2 years
### Big data - big CPU

<table>
<thead>
<tr>
<th></th>
<th>Proteins</th>
<th>comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009/09</td>
<td>$8 \cdot 10^6$</td>
<td></td>
</tr>
<tr>
<td>2013/06</td>
<td>$35 \cdot 10^6$</td>
<td></td>
</tr>
</tbody>
</table>

- **Cytochrome P450 - Mentha piperita - peppermint**
  - MELLQLWSALIIILVTITISLLINQWRKPQKFPGPPPPLGLPLHLGWGLPQHAL
  - ASVAKEYGPVAHVQLGEVFVLSREATKEAMKLVDACANRFESIGTRIMWYDNEDII

- **Cytochrome P450 - Mentha spicata - spearmint**
  - MELDLLSAIIILVATYIVSLLINQWRKSKQQNLPPPPPLPVIGHLHLWGGLPQHVFR
  - SIAQKYGPVAHVQLGEVSSAEAAKQAMKVLDPNFADRFDGIGSRTMWDKDDII

**double every 2 years**
# Big data - big CPU

<table>
<thead>
<tr>
<th></th>
<th>Proteins</th>
<th>comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2009/09</strong></td>
<td>$8 \cdot 10^6$</td>
<td>$6.4 \cdot 10^{12}$</td>
</tr>
<tr>
<td><strong>2013/06</strong></td>
<td>$35 \cdot 10^6$</td>
<td>$1.2 \cdot 10^{15}$</td>
</tr>
</tbody>
</table>

- *Proteins* doubled every 2 years.
- *Comparisons* increased 20-fold every 2 years.
### Big data - big CPU

<table>
<thead>
<tr>
<th>Year</th>
<th>Proteins</th>
<th>Comparisons</th>
<th>Storage Needed</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009/09</td>
<td>$8 \cdot 10^6$</td>
<td>$6.4 \cdot 10^{12}$</td>
<td>5 TB</td>
</tr>
<tr>
<td>2013/06</td>
<td>$35 \cdot 10^6$</td>
<td>$1.2 \cdot 10^{15}$</td>
<td>150 TB</td>
</tr>
</tbody>
</table>

- Proteins: double every 2 years
- Comparisons: 20-fold every 2 years
- Storage: 5-fold every 2 years
Some facts about proteins

- **how many in human?**
  - 20-25K in human

- **how long are they?**
  - ~35-30k, median around 400

- **do they consist of units?**
  - most proteins have more than 2 domains
Domains
Some facts about proteins

- how many in human?
  - 20-25K in human

- how long are they?
  - ~35-30k, median around 400

- do they consist of units?
  - most proteins have more than 2 domains

- what mediates protein function?
  - the same sequence always folds into the same 3D structure

- does the same human protein always have the same sequence?
  - mostly
Side chain properties

- Alanine
- Asparagine
- Aspartic Acid
- Arginine
- Cysteine
- Glutamine
- Glutamic Acid
- Glycine
- Histidine
- Isoleucine
- Leucine
- Lysine
- Methionine
- Phenylalanine
- Proline
- Serine
- Threonine
- Tryptophan
- Tyrosine
- Valine
Positively charged amino acids

- Alanine
- Arginine
- Cysteine
- Glutamine
- Glutamic Acid
- Histidine
- Isoleucine
- Leucine
- Lysine
- Methionine
- Phenylalanine
- Proline
- Serine
- Threonine
- Trytophan
- Tyrosine
- Valine
Negatively charged amino acids

- Alanine
- Aspartic Acid
- Arginine
- Asparagine
- Glutamic Acid
- Glutamine
- Cysteine
- Glycine
- Histidine
- Isoleucine
- Leucine
- Lysine
- Methionine
- Phenylalanine
- Proline
- Serine
- Threonine
- Tryptophan
- Tyrosine
- Valine
Some sort of Polar amino acids

- Alanine
- Arginine
- Aspartic Acid
- Asparagine
- Glutamic Acid
- Glutamine
- Glycine
- Histidine
- Isoleucine
- Leucine
- Lysine
- Methionine
- Phenylalanine
- Proline
- Serine
- Threonine
- Tyrosine
- Valine
Some sort of Hydrophobic amino acids

A: Alanine
B: Arginine
C: Asparagine
D: Aspartic Acid
E: Cysteine
F: Glycine
G: Glutamine
H: Glutamic Acid
I: Histidine
J: Isoleucine
K: Leucine
L: Lysine
M: Methionine
N: Phenylalanine
O: Proline
P: Serine
Q: Threonine
R: Tyrosine
S: Valine
What is protein structure?
Doyle et al. (1998) Science 280:69-77 - The structure of the potassium channel: molecular basis of K+ conduction and selectivity
Alcohol dehydrogenase (ADH)

http://www.proteopedia.org/wiki/images/7/7b/1htb2.png

homodimer ADH5


http://upload.wikimedia.org/wikipedia/commons/thumb/a/a5/Protein_ADH5_PDB_1m6h.png/800px-Protein_ADH5_PDB_1m6h.png
What is protein function?
Alcohol dehydrogenase (ADH5)

ADH5
D6R9G2_HUMAN
Alcohol dehydrogenase class-3
PDBid: 1m6h (human glutathione-dependent formaldehyde dehydrogenase)

UniProtKB:D6R9G2

sequence (95 residues):
MANEVIKCAAVAWEAGKPLSIEEIEVAPP
KAHEVRIKIIATAVCHTDAYTLSGADPEGC
FPVILGHEGAGIVESVGEVTKLKAVERMK
ILSKS

homodimer ADH5


http://upload.wikimedia.org/wikipedia/commons/thumb/a/a5/
Protein_ADH5_PDB_1m6h.png/800px-Protein_ADH5_PDB_1m6h.png
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.
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)
Enzyme EC 1.1.1.1: Aldehyde dehydrogenase

many pathways, e.g. glycolysis

Function of ADH

KEGG
GO: Gene-Ontology

GO:0004029:

Name:

aldehyde dehydrogenase (NAD) activity

Definition:

Catalysis of the reaction:
an aldehyde + NAD+ + H2O
an acid + NADH + H+.

http://www.ebi.ac.uk/QuickGO/
Protein function

Intuitive but not well-defined:

- chemical
- how atom bound?
- biochemical
- transferase
- cellular (kinase)
- cell cycle
- developmental
- time, regulatory
- physiological
- related to disease
- genetic
- dominant/recessive

Protein function as action:

*Function* = anything that happens to or through a protein
Protein function: everything that happens to or through a protein
Quite a challenge!

our special angle:
evolutionary information + machine learning
Gap sequence/annotation grows!

>35 million protein sequences known by summer 2013

![Graph showing the growth in the number of protein sequences in databases over time. The graph indicates a significant increase in the number of sequences from 1996 onwards, with a peak date of 3-2004.]
function known for 10–50% of human
function known for 10-50% in human
function known for 10-50% in human

annotation precision
function known for 10–50% in human

uncertainty in level of detail
function known for 10-50% in human

uncertainty in level of detail

http://i42.tinypic.com
Sequence determines structure determines function

- Epstein & Anfinsen, 1961: sequence uniquely determines structure

**INPUT:** protein sequence

**OUTPUT:** 3D structure and function
Task in terms of energy landscapes
Task in terms of energy landscapes

Point mutation
Binding (Substrate/Protein)
Environmental change (DNA close/pH)
**Need to know history to predict!**
Evolution is history!

Sequence identity implies structural similarity!

Don't know region

Distance from curve = +10

Distance from curve = -10

Chris Sander & Reinhard Schneider 1991 Proteins 9:56-68
B Rost 1999 Prot Engin 12:85-94
**SH3**
Src-homology 3 domain
one domain of proteins such as Src tyrosine kinase (STK)
Evolution improves prediction

Evolutionary profile implicitly captures history of and individual protein!
Using evolution to predict structure

Sequence → PSI-BLAST → Filter

PHDsec

1993

MaxHom

60% -> 72% / 77%

B Rost 1996 Meth Enzymol 266:525-539

© Burkhard Rost

SWISS-PROT
Exciting projects

- LOCtree etc: predict localization
- Predict enzymatic activity & flexibility
- Protein disorder
- Predict membrane regions, epitomes,
- Improve alignment methods
- SNP-pipeline: predict nsSNP effects
- PredictProtein: web service since 1992
- NESG & NYCOMPS: structural genomics

© Burkhard Rost

Q9P2H0
If we succeeded

- medical applications
  - better diagnosis
  - drug discovery, development, specificity

- molecular biology
  - reducing costs
  - improving advances

- other
  - dump disposal; bioremediation of toxic materials
  - biocompatible detergents
  - more system-oriented fertilizers
  - adhesives
  - ....
Diversity of function LEPTIN

↓ Food intake
↑ Energy expenditure

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

Ana Luiza Arruda, TUM
I1c

sequence - structure - motion
Molecular dynamics: function is motion
Motions in proteins

- Allosteric regulation
- Enzyme catalysis
- Ligand binding
- Folding

Protein movie

© Youtube - user “zla smo” - 3BC group, MedILS Croatia
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)
Disordered regions

Wu & Shi et al Science 2000 287(5450):92-7
Eukaryotes reign by disorder?
Intro: protein function evolution-terminology
Protein function: intuitive but not well-defined
Metrics for protein function
Predictions and analyses use metrics

☐ Sequence:

☐ Structure:

☐ Function?
Predictions and analyses use metrics

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

- Structure:

- 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
- EC4.1.1: carboxylases
- EC4.1.2: aldehyde carboxylases
- EC4.1.3: oxo-acid carboxylases
- EC4.1.99: other carbon-carbon carboxylases

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

© Marco Punta & Yanay Ofman
GO: Gene Ontology

Three classes:

- Biological process
- Molecular function
- Cellular component/localization

Gene Ontology (GO): Biological process

GO: Molecular function


© Marco Punta & Yanay Ofran

© 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

GO: No strict hierarchy

Transcription, DNA-dependent
GO:0006351

Transcription initiation
GO:0006352

Transcription initiation from Pol I promoter
GO:0006261

Transcription from Pol I promoter
GO:0006360

Transcription termination from Pol I promoter
GO:0006361

© Jennifer Clark GO Consortium, RSC
Structure conserved in sequence?
Annotation transfer

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

Assymetry and exceptions
Sequence-function asymmetry
Same sequence, different tissues
-> different function

Structure-function asymmetry

similar structure ⇄ similar function

similar function ⇄ similar structure
Different structures -> same function

Chymotrypsin (5cha)

Subtilin (5sic)
RECAP
First step: find related proteins

protein unk

MELLQLWSALIIILVTISLLINQWRKPKPQGKFPPGPPKLPILGLHLHLLWGKLPLQHAL
ASVAKEGVPAHVLQEGVFSDVVLSSREATKEAMKLVDPALCANRFESIGTRIMWYDNEDII
FSFYSEHWRQMRKICVSELSSRNRSFGFIRQDEVSRLLRHLRRGAGAADVDMTERIETL
TCSIICRAAFGSVIRDNAELENGLVKGALSMASGFELADMFPSSKLNNLLCWNKSKLWRMR
RRVDTILEAIVDEHKFKSKGESFEGGEDIIDVLFMRQKDTIQIKVPITNISIKAFIFDTSAG
TETSSTTLWVLABLEMNPAVMAKAQAEVRAALKEKTNWDVDVQELKYMKSVVKTMRM
HPPIPLIPSCREECVNGYTIIPNARIMINVWSMGRNPLYWEKPDTFWPERFDQVSKDF
MGNDFEFVPFGAGRRICPGLNFGLANVEVPLAQLLLHFDWKLAEKMPSDMDMSEAEGLT
GILKNNLLLVPTYPDSS
First step: find related proteins

PSI-BLAST (NCBI, NLM, NIH, Bethesda)

BLAST @ UniProt (EBI, Hinxton England)
First step: find related proteins

**Cytochrome P450 - Mentha piperita - peppermint**
MELLQLWSALIIILVVTVYTSLLINQWRKPKPQGKFPPGPPLPLGLGLHLHLLWGLKLPQHAL
ASVAKEYGPVAHVQLGEVFSVVLSSREATKEAMLVDPCAANFESIGTRIMWYDNEII
FSYSEHWRQMRKICVSELLRSSNVRSGFIRQDEVSRLLRLHRSSAGAAVDMTERIETL
TCSIICRAAGFVIRDNAELVGLV KDALSMSGFELADMFPPSKLLNLCLCNKSKLWRMR
RRVDTILEAIVDEHKFFKGSFGGEDIIIDVLFRMQKDTQIKVPITTNSIKAFIFDTFSAG
TETSSTTLWVLAMELNPAVMKAAQAEVRAALKETNWDVDVQELKYMKSVVKETMRM
HPPPIPLIPRSCREECVVNGYTIPNKARIMINVWSMGRNPLOYEEKPDTPFPERFDQVSKDF
MGNDFEFVFPFGAGRRCICPGLNFGLANVPLAQLLLYHFDWKLAEGMKPSDMDMSEAEGLT
GILKNLLLVLTPYPDSS

**Cytochrome P450 - Mentha spicata - spearmint**
MELDLSAIIILVATYIVSLLINQWRKSKSQQNLPPSPPKLVPVIGHLHFLWGLLPQHVFR
SIAQKYGPVAHVQLGEVYSVVLSSAAKQAAMKVLDPNDRFDGIGSRTMWYDKDIIIF
SPYNDHWRRQMRICVTELLSPKNSRFGYIRQEEIERLIRLLGSSGGAPVDVTEEVSKMS
CVVVCRAAAGSVLKDQGSLAESLAKELASFGELADLYPSSWLLNLLSRLNKRYLQRMRR
RLDHILDGFLEEHREKKSGEFGGEGDIDVDLFRMQKGSDIKIPITSNCKIGFIFDTFSAGA
ETSSTTISWALSELMHNPAMKMAKVQAERVALKGKTVDLESEVQELKYLRSVLLKETLRLH
PPFPLIPRQSREECEVNGYTIPAKTRIFINVWAIGRDPQYWEDEPDTFPRPERFDEVSRTDFM
GNDCEFIPFGAGRRCICPGLHFGLANVPLAQLLLYHFDWKLPQGMTDADLDMETPGLSG
PKKNVCLVPTLYKSP
First step: find related proteins

Cytochrome P450 - Mentha piperita - peppermint

MELLQLWSALIILVV VTY TISLLIN QWRKP KPK QGFPPGPPKLPLIGHLHLWLWGKLPQHAL
ASV AKEY GPVAH VQLGEVFSVVLSSREAT KEMKLVDPA CANRFESIGTRIMWYDNE DI
FSYSEHWRQMRKICVSsell SSRNVRSFGFIRQDEVSRLLRHLR SAGAAVDMTERIETL
TCSIIICRAAFGSVIRDNAELVGLVKDALS MASGFELADMFPSSKLLNLLCWNKSKLWRMR
RRVDTILEAIVDEHKFKKSgef GEDIDI DVLFRMQ KDTQIK VPTTNSIKAFIFDTSAG
TETSS TTTTLWLVEL MINPAVMAKAQA E VRAALKEKTNW DDVDV QELKYM KSVVKETMRM
HPPIP LI PR S CREECVV NGYTIPN KARIMINV WSMGR NLY WEKPD TFWPERFDQVSKDF
MG N D FEVF PFGAG RRICPQLNFGL ANVEVPLAQLLYHFDWK LAEG MKPSD M DSEA EGLT
GILKNNLLLTV PTYPDSS
function known for
3,000–12,500 human
proteins

unknown for:
22,000–12,500
function known for 10–50% of human
Goal of protein prediction

Epstein & Anfinsen, 1961:
sequence uniquely determines structure

INPUT: protein sequence
OUTPUT: protein structure & function
Goals of predicting protein function
Alcohol dehydrogenase (ADH5)

ADH5
D6R9G2_HUMAN
Alcohol dehydrogenase class-3
PDBid: 1m6h (human glutathione-dependent formaldehyde dehydrogenase)

UniProtKB:D6R9G2

sequence (95 residues):
MANEVIKCKAAVAWEAGKPLSIEEIEVAPP
KAHEVRIKIIATAVCHTDAYTLSGADPEGC
FPVILGHEGAGIVESVGEVTKLKAVWRMQ
ILSKS

homodimer ADH5


http://upload.wikimedia.org/wikipedia/commons/thumb/a/a5/Protein_ADH5_PDB_1m6h.png/800px-Protein_ADH5_PDB_1m6h.png
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
How could we accomplish this end?
preliminary Lecture plan (CB2 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
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)
28: 2019/01/10: SNP effect 2
29: 2019/01/15: SNP effect 3
30: 2019/01/17: SNP effect 4
31: 2019/01/22: WRAP up 1
32: 2019/01/24: WRAP up 2
33: 2019/01/29: WRAP up 3
34: 2019/01/31: ?
35: 2019/02/05: ?
36: 2019/02/07: Exam (10:00-13:00, lecture room LMU physics)

© Burkhard Rost 2019