I1: Introduction to protein function

short title: cb2_intro_func1

lecture: Protein Prediction 2 (for Computer Science) - Protein function
TUM winter semester
Videos: YouTube / www.rostlab.org

THANKS:

Special lectures:
- none planned

No lecture:
- 10/31 THU All Saints
- 11/12 THU SVV (student rep)
- 11/28 THU Thanksgiving
- 12/05 THU TUM Dies Academicus
- 12/19-01/07 - no lecture Xmas+

LAST lecture: Jan 21 (followed by 2 wrap-up sessions)

Examen: Jan 30 11:00-13:00 Room LMU Physics HS019
- Makeup: NOT (emergency: Apr 21 & Apr 23, 2020 lecture time)
THANKS for slides!

© Michael Leunig: The Traveling Leunig
Yanay Ofran

now
Bar Ilan Univ
Israel
Marco Punta

Marco Punta (now ICR London) & 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

RNA at the beginning

Trouble: missing some enzymatic activity

RF Gesteland, TR Cech, JF Atkins:
the RNA World. Cold Spring Harbor Press. (Cover of Book)
RNA at the beginning

enslave proteins to get the job done

RNA at the beginning


A Lupas MPI Tuebingen
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 factors 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)
15. glycolytic enzymes
16. 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. lypoglycan (long carbohydrate chains connected to lipid in the membrane)

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
## Genome sizes

<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
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]
MAAQGFLLIATFLLVLMVLARPLGSGLARLINDIPLPGTTGVERVLFRALGVSDREMNWK
QYLCAILGLNMLGLAVLFFMLLGQHYLPLNQPQLPGGLSDLALNATAVSFVTNTNWQSYSG
ETTLSYFSQMAGLTVQNFLSAASGIAVIFALIRAFTQSMSTLGNAWVLRRITLWVLVP
VALLIALFFIQQGALQNFLPYQAVNTVEGAQQLPMPGVASQEAIKMLGTNGGGFFNANS
SHPFENPTALTNFLVQMLAIFLIPALTCLFACFGEVMGDRQRGRMLLWAMSVIIFVICGVMW
AEVQGNPHLLALGTDSSINMGEKESRFGVGLVSLSLFAVVTTAASCGAVIAMHDSFTALGGM
VPMWLMQIGEVVVFGVGSGLYGMLFLVLLAVFIAGLMIGRTPEYLGKKIDVREMKLTLA
ILVTPTLVLMGAALAMMTDAGRSAMLNPGPHGFSEVLYAVSSAANNNGSAFAGLGSANSPF
WNCLLAFCMVFGRFGVIIIPVMAIAAGSLVSKKSQAASSGTLPHTHPGLFVGLLIGTVLLLVGA
LTFIPALALGPVAEYLS
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


doi: 10.1016/j.cell.2012.04.012
Protein sequence

>gi|16128674|ref|NP_415226.1| potassium translocating ATPase, subunit A [Escherichia coli K12]
MAAQGFLLIATFLLVLMLARPLGSLARLINDIPLPGTTGVERVLFRALGVSDREMNWK
QYLCAILGLNMLGLAVLFFMLLQGHYLPNLNPQQLPGLSWDLALNTAVSFVTNTNWQSYSG
ETTLSYSQSAMGLTVQNFLSAASGIAVIFALIRAIFTRQSMSTLGNANWDLLRITLWVLVP
VALLIALFFIQQGALQNFLPYQAVNTVEGAQQQLPMGPVSAQEAIKMLGTNGGGFFNANS
SHPFENPALTNFVQMLAIFIPTALCFAFGEVMGDRRQGRMLLWAMSVIFVICVGVCVMW
AEVQGNPHLLALGDSSINMEGKRSFGRGVLVSSLFAVVTAAASCAGAVIAMHDSFTALGGM
VPMWLMQIGEVVGGMGGMMLFVLLAVFIAGLMIGRTPEYLGGKIDVREMKLTLA
ILVTPTLVLNGAALAMMTDAGRSMNLNPQPHGFSEVLYAVSSAANNSAFAGLSANSPF
WNCLLAFCFVGFRFGVIIYPVMAIGSLVSKKSQAASSGLTPTHGPLLFLVGLLIGTVLLVGA
LTFIPALALGPVAEYLS
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
MELLQLWSALIIIVVTYTISSLINQWRKPQKPGFPGLPLIGHLHLWLGKLPQHAL
ASVAKEYGVPVAHVQLGEVFSVVSREATKEAMLKDPCAANRFESIGTRIMWYDNEDIEI
FSPYSEHWRQMRKICVSELSSRSNFSGFIRQDEVSRALLHRSSAGAAVDMTERIELTL
TCSIIICRAAFGVSIRDNALVEVGLKDALSMASGEFLADMPFSSKLNLCCWRNKSCLWRMR
RRVDTILEAIVDEHKFKKSGEGGEDIDIVLFRMQKDTQIKVPITTSIKAFIFDTFSAG
TETSSITTLWVLAELMRNPAMAKAQAEVRAALKETNWDVDVVQELKYMKSVVKETMRM
HPPIPLIPRSCREECVVNGYTPNKRIMINTWNSMRPLLWEKPDTFWPERFDQVSKDF
MGNDFEFVFPGAGRRICPGLNFGLANVEVPLAQFLLYHFDWKLAEMKPSDMDMSEAEGLT
GILKNNLLLVLPTYPDPSS

Cytochrome P450 - Mentha spicata - spearmint
MELDLSSAIIIILVATIYISLILINQWRKSQQLPPSPPKLPVIGHLHFLWGLPQHPHR
SIAQKYGPVAHVQLGEVSVVSLSAEAAKQAMKVLDNPADFRFDGSGRTMWYDKDDIIIF
SPYNDHWRQMRICVTELLSPKNFRSGFYIRQQEIEIERLIRLLGSGGAPVDVTEESKMS
CVVVCRAAFGSSLKDQGSLAELVKESSLALASGEFLADLYPSSWLLNLLSNNKYRLQRMRR
RLDHILDGFLEEHREKSGEGGEDIVDVLFRMQKGDIDKIPITSNCIKGFIFDTFSAGA
ETSSITISWALSELMRPAMAKVQAERVALKGKTVDLSVEVQELKYLRSVLKETLRHL
PPFLIPRQSREECEVNGYTPAKTRIFINVWAIQGRDPQFWEPDTPFRPERFDEVSRDFM
GNDFFIFPGAGRRICPGLHFLGLANVEIPLAQFLLYHFDWKLPQGMTDADLMETYTPGSLG
PKKNNCLVLPTYKSP
Protein sequence

Cytochrome P450 - Mentha piperita - peppermint
MELLQLWSALIIILVVTYTISSLLINQWRKPKPQGKFPPGPPKLPILGLHHLWGLKLPQHAL
ASVAKEYGPVAHVQLGGEYFSVVLSSREATKEAMKLVDPAACANRFESIGTRIMWYDNEDII
FSYPSEHWRQMRKICVSELSSRNVSFGFIRQDENVSSLRLHRSSAGAAVDMTERIETL
TCSIICRAAAGSVIRDNAELVGLVKKDALSMSGFELADMFPSKLLNLLLCSWNKSKLWRMR
RRVDTILEAIVDEHKFFKKGEGQEDIDVLFMRQKDQIKNPVITTNSIKAFIFDFTSAAG
TETSSTTLGVLAMELMRNPAMAKAQAAMQRAALEKEKTNDVDDVQELKYMSVVKETMRM
HPPPIPLPRSCREECVVNGYTIPNKAIRIMINVWSMGRNPEVEKPDFTWPERFDQVSKDF
MGNDSEPFPFGAGRRICPGLNFGLANVPLAQLLYHFDFKLRAGMKSMDMSEAEGLT
GILKNNLLLVPTYPDSPS

Cytochrome P450 - Mentha spicata - spearmint
MELDLSAILIVATYIVSLLINQWRKSQSNLPPSPPKLPVIGHLHFLWGLPQHVFR
SIAQKYGPVAHVQLGGEYFSVVLSSAEAKQAMKVLDPNFLDFDGGRTMWDKIDDIF
SPYNKDHRQMRIRCTEELSPKNVRFQGIRQKIEIRLIRILGLSSGGAPVDVTEEVSKMS
CVVVCRAAAGSVLKDOQGSLAEVLKESLALASGFELADLYPSSWLLNNLSSLKRYLQRMMR
RLDHIILDGFLEEHREKSGEFGGEDIVDVFMRQKGSIDIKIPSNCIKGFIFDFTSAAG
ETSSTISWALSELMRNPAMAKVQAEMAKLKGKTVDLSEVQELKYLRSVLKETLRLH
PPFPLIPRQSREECEVNGYTIPAKTRIFINVWAIAGRDPOQYWEDPDTFRPERFDENVSRDFM
GNDSEPFPFGAGRRICPGLNFGLANVPLAQLLYHFDFKLRAGMKSMDMSEAEGLT
PKKKNVCLVPTLYKSP
Proteins - genetic code

Table showing the genetic code:

<table>
<thead>
<tr>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUU</td>
<td>UUC</td>
<td>UUA</td>
<td>UUG</td>
</tr>
<tr>
<td>CUU</td>
<td>CUC</td>
<td>CUA</td>
<td>CUG</td>
</tr>
<tr>
<td>AUA</td>
<td>AUG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GUU</td>
<td>GUC</td>
<td>GUA</td>
<td>GUG</td>
</tr>
</tbody>
</table>

Amino acid names:

- Ala = alanine
- Arg = arginine
- Asp = aspartate
- Cys = cysteine
- Gln = glutamine
- Glu = glutamate
- His = histidine
- Ile = isoleucine
- Leu = leucine
- Lys = lysine
- Met = methionine
- Phe = phenylalanine
- Ser = serine
- Thr = threonine
- Trp = tryptophan
- Tyr = tyrosine
- Val = valine

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**
  - MELLQLWSALIIIVTISLLLINQWRKPKPGKIFPQPPPKLPLIGHLHLLWGLPQHAL
  - ASVAKEYGPVAHVQLGEVFVLSREATKEAMKLVDACANRFESIGTRIMWYDNEDII

- **Cytochrome P450 - Mentha spicata - spearmint**
  - MELDLLSAIIIIVTISLLLINQWRKSQQLPPPKLPLVIGHLHFLWGGLPOHVFR
  - SIAQKYGPVAHVQLGEVYSVLSAAEAKQAMKVLDPNFAFRFDGIGSRTMWDYDKDIIIF

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

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

- Proteins double every 2 years.
- Comparisons increase 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>$8 \times 10^6$</td>
<td>$6.4 \times 10^{12}$</td>
<td>5 TB</td>
</tr>
<tr>
<td>2013/06</td>
<td>$35 \times 10^6$</td>
<td>$1.2 \times 10^{15}$</td>
<td>150 TB</td>
</tr>
</tbody>
</table>

- Proteins double every 2 years.
- Comparisons increase 20-fold every 2 years.
- Storage needed increases 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
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
- Asparagine
- Aspartic Acid
- Cysteine
- Glycine
- Glutamine
- Glutamic Acid
- Histidine
- Isoleucine
- Leucine
- Lysine
- Methionine
- Phenylalanine
- Proline
- Serine
- Threonine
- Tryptophan
- Tyrosine
- Valine
Positively charged amino acids

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

- Alanine
- Cysteine
- Histidine
- Methionine
- Threonine
- Tyrosine
- Valine
- Arginine
- Asparagine
- Glutamine
- Aspartic Acid
- Glutamic Acid

© Burkhard Rost ROSTLAB TUM
Some sort of Polar amino acids

- Alanine
- Arginine
- Asparagine
- Aspartic Acid
- Cysteine
- Glycine
- Glutamine
- Glutamic Acid
- Histidine
- Isoleucine
- Leucine
- Lysine
- Methionine
- Phenylalanine
- Proline
- Serine
- Threonine
- Trytophan
- 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: PhenyIalanine
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):
MANEVIKCKAAVWEAGKPLSIEEIEVAPP
KAHEVRIKIIATAVCHTDAYTLSGADPEGC
FPVILGHEGAGIVESVGEVTKLKAVWRMQ
ILSKS


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

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

(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: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
function known for 10–50% of human
function known for 10–50% in human

annotation precision
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
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!

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

MaxHom

PHDsec

1993

<table>
<thead>
<tr>
<th>Protein</th>
<th>Alignments</th>
<th>profile table</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>GGGG</td>
<td>GSAPD NTEEQ CVHRI LMYFW</td>
</tr>
<tr>
<td>Y</td>
<td>YYYY</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>IEE</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>YYYY</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>DDDDD</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>EEE</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>GGGG</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>DDDD</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>PPPP</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>DDDD</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>DDDD</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>NQNN</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>GGGG</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>VVVV</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>EPKK</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>PPPP</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>GGGG</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>TTHT</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>EKSA</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>FFFF</td>
<td></td>
</tr>
</tbody>
</table>

60% -> 72% / 77%

corresponds to the the 21*3 bits coding for the profile of one residue

B Rost 1996 Meth Enzymol 266:525-539
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
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 lyases
- **EC4.2**: carbon-oxygen lyases
- **EC4.3**: carbon-nitrogen lyases
- **EC4.4**: carbon-sulfur lyases
- **EC4.5**: phosphorus-oxygen lyases
- **EC4.99**: other lyases

- **EC4.1.1**: carboxylates
- **EC4.1.2**: aldehyde lyases
- **EC4.1.3**: oxo-acid lyases
- **EC4.1.99**: other carbon-carbon lyases

4.1.1.1 pyruvate decarboxylase
4.1.1.2 oxolate decarboxylase

© Marco Punta & Yanay Ofra
Three classes:

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

GO: Molecular function

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

© 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

similar sequence  $\Rightarrow$  similar function

similar function  $\Rightarrow$  similar sequence
Same sequence, different tissues
-> different function
Structure-function asymmetry

similar structure $\rightarrow$ similar function

similar function $\rightarrow$ similar structure
Different structures -> same function

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

protein unk

MELLQLWSALIILVVVTYTISLLINQWRKPKPQGKFPPGPPKLPLIGHLHLLWGWKLPQHAL
ASVAKEYGPVAHVQGLGEVFSVVLSSLREATKEAMKLVDPAACANRFESIGTRIMWYDNEII
FSPYSEHWRQMVKICVSELSSRNVRSGFIRQDEVSRLLRHLRSSAGAAVDMLRERIETL
TCSIICRAAFGSVIRDNAELVGLVLKDALSMASGFELADMFPSSKLNLCLCNKSKLWRRMR
RRVDTILEAIVDEHKFKKSGEGFGEDIIIIDVLFRMQKDITIQIIPITTSIKAFIFDTFSA
GTTETSTSSTTLLWVLALMRNPAMAKAQAQAEVRAALKETKNWDVDDVQLKYMKSVKETRM
MPPPIPLIPRSCREECVNQYIPKARNKIRMINVWSMGRNPLOYWEKPDTFWPERFDQVSKDF
MGNDFEFVPGAGRCLCPGLNFLALVEVPLAQLLYHFDWKLAEMKPSDMDMSEAEGLT
GILKNNLLLVPDYPSS
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
MELLQLWSALIIILVVTYTISSLINQWRKPKPQGKFPPPGPPLPLIGHLHLLWGKLIPQHAL
ASVAKEYGPVAVHQLGEVFSVLSSREATKEAMKLVDPAACANRFESIGTRIMWYDNEDII
FSPYSEHWRQMRKICVSELLSSRNRSFGFIRQDEVSRLLRLHRSSAGAAVDMTERIEYT
TCSIIICRAAFGSVIRDNAELVGLVKDLASMASGFELADMFPSSKLNLCLWNKSKLWRMR
RRVDTILEAIVDEHKFKKSCEFGGEDIDVLFMRQKDTQIKVPITTSIKAFIFDTSAG
TETSSTTTLWVLAELMRNPAMAKAQAERVRAALKEKTNWDDVQELKMKSVVKETMRM
HPPIPLIPRSCEECVNVGYTIPNKAIRIMINVWSMGRNPLOYWEKFQTFWPERFDQVSKDF
MGNDIEFVPFAGRRICPGNLFGVNLVMEVEQLYHFDWKLAEAGMPSDMDMSEAEST
GILKNLLLVPTYPDPSS

Cytochrome P450 - Mentha spicata - spearmint
MELDLSAIILILVATYIVSSLINQWRKSQKSQNLPPSPPKLPVIGHLHFLWGLPLQPQHVF
SIAQKYGPVAVHQLGEVSYSVLSSAAEAKQAMKVLDPNFADRFDSRGRTMYDKDDIIIF
SPYNDHWRQMRICVTELLSPKNVRSFGYIRQIEIERLLRIILGGSSGAPVDEPVEEVSKMS
CVVVCRAAFGSVLKDQGSLAELVKESLALASGFELADLYPSSWLLNLSSLNLYRLQMR-R
RLDHILDGFLEEHHREKSGEFGGEDIVDLFRMKQGSDIKIPITSNCIKGFIFDTSAGA
ETSSTTISWALSELMNPAKMAKVQAEVREALKGTKTVDLSEVQELKYLRSVLKETLRLH
PPFPLIPRQSREECEVNVGYTIPAKTRIFINVWAIGRPQYWEDPDTRPERFDREVSRDFM
GNDFEFIPFGAGRRICPGNLFGVNLVMEVEQLYHFDWKLPQGMTADLMETPGLSG
PKKKNVCLVPHTYKSP
First step: find related proteins

Cytochrome P450 - Mentha piperita - peppermint

MELLQLWSALIIILVVTYTISLLINQWRKPKPQKFGFPPGPPKLPPLIGHLHLHLLWGKLPQHAL
ASVAKEYGPVAHVQLGGEVFSVVLSSREATKEAMKLVDPACANRFESIGTRIMWYDNEDII
FSPYSEHWRQMRKICVSELLSSRNVRSGFIRQDEVSRLLRHLRSSAGAAVDMTERIEL
TCSIIICRAAFGSVIRDNAELVGLVLDLSMASGFELADMFPSKLLNLLCWNKSKLWRMR
RRVDITLEAIVDEHKFKKSGFEGGEDIIDVLFMRMQKDTQIKVPITTSIKAFIFDTSAG
TETSSTTLWLVAELMRNPAMAKAQAQEVARAILKEKTNWDDVDDQELKYMKSVVKETMRM
HPPILIIPRSCREECVVNGYTIINPKARIINIVWSMGRNPLYWEKPDTFWPERFDQVSKDF
MGNDVFVFPGAGRRCPLGLNFGLANVPLAQLLYHFDWLKLAEMKPSDMDMSEAEGLT
GILKNNLLLVPTPYDPSS
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):
MANEVIKCKAANWAEEAGKPLSIEEIEVAPP
KAHEVRIKIATAVCHTDAYTLSGADPEGC
FPVLGHEGAGIVESVGEVTKLKAVWRMQ
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: 2019/10/15: No lecture (makeup examen: PP last year)
02: 2019/10/17: No lecture (makeup)
03: 2019/10/22: No lecture
04: 2019/10/24: Welcome: who we are
05: 2019/10/29: Intro function 1: concept of protein function
06: 2019/10/31: No lecture (holiday, All Saints)
07: 2019/11/05: Intro function 2: homology-based inference
08: 2019/11/07: Localization 1 (chalk talk)
09: 2019/11/12: No lecture (SVV)
10: 2019/11/14: Localization 2 (homology, motifs)
12: 2019/11/21: Localization 4
13: 2019/11/26: Localization 5
14: 2019/11/28: No lecture (Thanksgiving)
15: 2019/12/03: Localization 6
16: 2019/12/05: No lecture (Dies Academicus)
17: 2019/12/10: PPI 1 - sites (chalk)
18: 2019/12/12: PPI 2 - sites / PPI pairing
19: 2019/12/17: PPI 3 - sites / DNA / RNA (Jia Jun Qiu)
20: 2019/12/19: No lecture
22-24: no lectures - winter break (2019/12/24 - 2020/01/06)
25: 2020/01/07: No lecture
28: 2020/01/09: PPI 4 - sites: DNA / RNA (Jia Jun Qiu) + PPI pairing 1
29: 2020/01/14: SAV effect 1 (chalk talk)
30: 2020/01/16: SAV effect 2
31: 2020/01/21: SAV effect 3
32: 2020/01/23: WRAP up 1
33: 2020/01/28: WRAP up 2
34: 2020/01/30: Examen (10:00-13:00, lecture room TBA - LMU physics?)
35: 2020/02/04: TBA
36: 2020/02/06: TBA