title: Protein Prediction 2 (for Bioinformaticians) - Protein function:

Intro function - homology inference

short title: pp2_introfunc1

lecture: Protein Prediction 2 - Protein function
TUM winter 2014/2015
Announcements

Videos: YouTube / www.rostlab.org

THANKS:
Tim Karl + Jonas Reeb

Special lectures:
- xxx - Tobias Hamp

No lecture:
- Nov 12 Tue (Student assembly)
- Dec 12 Thu (TUM Dies Academicus)

LAST lecture: January 20

Examen: January 22
- Makeup: Apr 14, 2015 - morning/noon

CONTACT: Tanya Goldberg goldberg@rostlab.org

© Burkhard Rost (TUM Munich)
Today: Intro into function

LAST WEEKs
- Intro into lecture

THIS WEEK
- Tuesday: From molecular predictions to individualized medicine
- Introduction to protein function I

NEXT WEEK
- Tobias Hamp: function prediction
Function Intro:
Homology-based inference
Homology vs. machine learning
How to start the prediction of function?

Search Sequence (of unknown protein):

EVERY NE HAS T KN W IN TEST
why 2 predict function?
Protein kinases in human


© Kaz Wrzeszczynski: Thesis
Protein function classification

Protein Space:

X=Positive  Y=Negative

- Close Homology (Sequence Id. > 60% Psi-Blast Eval < 10^{-20})
- Distant Homology (Domain, Motif)
- Machine Learning (NN, SVM)
General challenges for homology-based inference of function
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/

species=?
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/

species=mating
Evolution: speciation

© [http://evolution.berkeley.edu(evosite/evo101)](http://evolution.berkeley.edu/evosite/evo101/)

happily munching
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/

disaster strikes
Evolution: speciation

https://evolution.berkeley.edu/evosite/evo101/

populations diverge
Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/

rejoined - yet separated
Evolution: speciation


Evolution: speciation

© http://evolution.berkeley.edu/evosite/evo101/

Happy Face Spider *Theridion grallator*  
(same species-interbreed)

Carrion/Hooded crow  
(same species?)
ADH: Alcohol dehydrogenase

© Wikipedia  ADH5 PDBid: 1m6h

Human glutathione-dependent formaldehyde dehydrogenase

ADH: Alcohol dehydrogenase

© Wikipedia

© GO

Homologs have the same ancestor

- ADH1-human
- ADH1-yeast
- ADH1-bacteria
- ADH1-plants
Homologs have the same ancestor.

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

A common ancestor tree might be
Homologs have the same ancestor?

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

Tree hypothetical
Homolog, Ortholog

- Homologs have the same ancestor.
- Orthologs are genes separated by speciation.

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

common ancestor

Tree hypothetical
Homolog, Ortholog

- Homologs have the same ancestor
- Orthologs are genes separated by speciation
- Paralogs are separated AFTER speciation

ADH1-human

ADH1-yeast

ADH1-plants

ADH1-bacteria

ADH2-plants

No common ancestor
Translation of terms to proteins

- homologous proteins: are related
- orthologs have similar function
- paralogs may evolve a different function
Tree uncertain, but story gets even more complicated for proteins:
why?
problem 1: genes/proteins do not “reproduce”
problem 2:
domains decoupled
“The domain problem”

A

Function 1

B

Function 1

C

Function 1
problem 3: moonlighting
## Moonlighting proteins

<table>
<thead>
<tr>
<th>One function</th>
<th>Additional functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>PutA proline dehydrogenase</td>
<td>Transcriptional repressor</td>
</tr>
<tr>
<td>Phosphoglucone isomerase</td>
<td>Neuroleukin, autocrine motility factor, differentiation and maturation mediator</td>
</tr>
<tr>
<td>Thymidine phosphorylase</td>
<td>Platelet-derived endothelial cell growth factor</td>
</tr>
<tr>
<td>Neuropilin (VEGF receptor)</td>
<td>Receptor for semaphorin III (nerve axons)</td>
</tr>
<tr>
<td>Uracil-DNA glycosylase</td>
<td>Glyceraldehyde-3-phosphate dehydrogenase</td>
</tr>
<tr>
<td>Aconitase</td>
<td>Iron-responsive-element binding protein (IRE-BP)</td>
</tr>
<tr>
<td>Carbinolamine dehydratase</td>
<td>Dimerization cofactor (DcoH)</td>
</tr>
<tr>
<td><em>Escherichia coli</em> thioredoxin</td>
<td>Subunit of T7 DNA polymerase</td>
</tr>
<tr>
<td><em>E. coli</em> aspartate receptor</td>
<td>Maltose-binding-protein receptor</td>
</tr>
<tr>
<td>PMS2 mismatch-repair enzyme</td>
<td>Hypermutation of antibody variable chains</td>
</tr>
<tr>
<td>Ribosomal proteins</td>
<td>DNA repair, translational regulators, development regulators, etc.</td>
</tr>
<tr>
<td>Lens crystallins</td>
<td>Heat-shock proteins, lactate dehydrogenase, argininosuccinate, retinaldehyde dehydrogenase, lyase, enolase, quinone oxidoreductase, glyceraldehyde-3-phosphate dehydrogenase, etc.</td>
</tr>
<tr>
<td>CFTR chloride channel</td>
<td>Regulator of other epithelial anion channels</td>
</tr>
<tr>
<td>P-glycoprotein (transporter)</td>
<td>Regulator of cell-swelling ion channel</td>
</tr>
<tr>
<td>Thrombin protease</td>
<td>Ligand for cell surface receptors</td>
</tr>
<tr>
<td>Thymidylate synthase</td>
<td>Translation inhibitor</td>
</tr>
<tr>
<td><em>E. coli</em> birA biotin synthetase</td>
<td>Bio operon repressor</td>
</tr>
<tr>
<td>Mitochondrial LON protease</td>
<td>Chaperone</td>
</tr>
<tr>
<td>Bacterial FtsH chaperone</td>
<td>Metalloprotease</td>
</tr>
<tr>
<td>Band-3 anion exchanger</td>
<td>Regulator of glycolysis</td>
</tr>
</tbody>
</table>

How well does it work?
Homology transfer accurate for very similar proteins

<table>
<thead>
<tr>
<th>TRUE</th>
<th>FALSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>methyltransferase</td>
<td>methyltransferase</td>
</tr>
<tr>
<td>identity</td>
<td>protein</td>
</tr>
<tr>
<td>100%</td>
<td>guanidinoacetate N-methyltransferase</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>100%</th>
<th>guanidinoacetate N-methyltransferase</th>
</tr>
</thead>
</table>

1 | 50

- syn_human VTLFALDY EARTDOMTS KGERQIVIN SSEGQWEAR SLTGQETYI
- yrk_chick VTLFALDY EARTDOMTS KGERQIVIN SSEGQWEAR SLTGQETYI
- fgr_human VTLFALDY EARTDOMTS KGERQIVIN SSEGQWEAR SLTGQETYI
- yes_chick VTVFALDY EARTDOMTS KGERQIVIN SSEGQWEAR SLTGQETYI
- src_avis2 VTTFALDY EARTDOMTS KGERQIVIN SSEGQWEAR SLTGQETYI
- src_avisv VTTFALDY EARTDOMTS KGERQIVIN SSEGQWEAR SLTGQETYI
- src_avisr VTTFALDY EARTDOMTS KGERQIVIN SSEGQWEAR SLTGQETYI
- stk_hydat VTVFLDY EARTDOMTS KGERQIVIN SSEGQWEAR SLTGQETYI
- hck_chick VTVFLDY EARTDOMTS KGERQIVIN SSEGQWEAR SLTGQETYI
- src_rsvpa VTTFALDY EARTDOMTS KGERQIVIN SSEGQWEAR SLTGQETYI
- blk_mouse VTVFALDY EARTDOMTS KGERQIVIN SSEGQWEAR SLTGQETYI
- hck_chick VTVFALDY EARTDOMTS KGERQIVIN SSEGQWEAR SLTGQETYI
Homology transfer accurate for very similar proteins

**methyltransferase**

<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>
</tbody>
</table>
Homology transfer accurate for very similar proteins

<table>
<thead>
<tr>
<th>methyltransferase</th>
<th>TRUE</th>
<th>FALSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>identity protein</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>guanidinoacetate N-methyltransferase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>magnesium protoporphyrin IX methyltransferase</td>
<td>99%</td>
<td></td>
</tr>
<tr>
<td>phosphoribosylglycinamide formyltransferase</td>
<td>70%</td>
<td></td>
</tr>
<tr>
<td>inositol 3-methyltransferase</td>
<td>65%</td>
<td></td>
</tr>
</tbody>
</table>

| 2/3 accuracy ; 2/4 coverage                  |      |       |

© Burkhard Rost (TUM Munich)
Homology transfer accurate for very similar proteins

TRUE FALSE
methyltransferase

<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>96%</td>
<td>phosphoribosylglycinamide formyltransferase</td>
</tr>
<tr>
<td>95%</td>
<td>inositol 3-methyltransferase</td>
</tr>
<tr>
<td>75%</td>
<td>phosphoribosylglycinamide formyltransferase</td>
</tr>
<tr>
<td>65%</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>

2/3 accuracy ; 2/4 coverage
3/8 accuracy ; 4/4 coverage
what’s better?

66% acc @ 50% cov

or

38% acc @ 100% cov
Specific challenges

(homology-based inference of function):

sub-cellular
localization
Rajesh Nair

Rajesh Nair
now: FDA, Washington
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

- Pairwise sequence similarity
- Percentage of pairs
- Midnight zone, Twilight zone, Safe zone
- Accuracy, Coverage
Different sequences adopt similar 3D

Sequence identity implies structural similarity!

Don't know region

Distance from curve = +10

Distance from curve = -10

C Sander & R Schneider 1991 Proteins 9:56-68
B Rost 1999 Prot Engin 12:85-94
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
Annotation transfer: Localization
Annotation transfer: 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

R Nair and B Rost (2002) *Protein Science* 11: 2836-47
Same sequence, different tissues

-> different function
Quick fix? - NO!

Quick fix:
- Establish family-specific similarity cutoff
- Establish a function-specific similarity cutoff

Difficult to realize and not enough!
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?
EC: Enzyme Commission number

- **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**: others

**4.1.1.1** pyruvate decarboxylase
**4.1.1.2** oxolate decarboxylase

Enzyme classification (EC)

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

### Table 1. Description of the different levels in the EC classification

<table>
<thead>
<tr>
<th>First figure</th>
<th>Second figure</th>
<th>Third figure</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. OXIDOREDUCTASES</td>
<td>Describes substrate acted on by enzyme</td>
<td>Type of acceptor</td>
</tr>
<tr>
<td>Substrate is oxidised-regarded as the hydrogen or electron donor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B. TRANSFERASES</td>
<td>Describes group transferred</td>
<td>Further information on the group transferred</td>
</tr>
<tr>
<td>Transfer of a group from one substrate to another</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C. HYDROLASES</td>
<td>Describes type of bond</td>
<td>Nature of substrate</td>
</tr>
<tr>
<td>Hydrolytic cleavage of bond</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D. LYASES</td>
<td>Type of bond</td>
<td>Further information on the group eliminated</td>
</tr>
<tr>
<td>Cleavage of bonds by elimination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E. ISOMERASES</td>
<td>Type of reorganisation</td>
<td>Type of substrate</td>
</tr>
<tr>
<td>F. LIGASES</td>
<td>Describes type of bond formed</td>
<td>Describes type of compound formed</td>
</tr>
<tr>
<td>Enzyme catalysing the joining of two molecules in concert with hydrolysis of ATP</td>
<td></td>
<td></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.-.-) 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).

Enzyme classification nomenclature

http://www.chem.qmul.ac.uk/iubmb/enzyme/

G.P. Moss, Queen Mary University of London, g.p.moss@qmul.ac.uk

EC 1  Oxidoreductases
EC 2  Transferases
EC 2.1 Transferring one-carbon groups
   EC 2.1.1 Methyltransferases
   EC 2.1.2 Hydroxymethyl-, Formyl- and Related Transferases
   EC 2.1.3 Carboxyl- and Carbamoyltransferases
   EC 2.1.4 Amidinotransferases
      EC 2.1.1.1 nicotinamide N-methyltransferase
      EC 2.1.1.2 guanidinoacetate N-methyltransferase
      EC 2.1.1.3 thetin-homocysteine S-methyltransferase
EC 2.2 Transferring aldehyde or ketonic groups
EC 2.3 Acyltransferases
EC 2.4 Glycosyltransferases
EC 2.5 Transferring alkyl or aryl groups, other than methyl groups
EC 2.6 Transferring nitrogenous groups
EC 2.7 Transferring phosphorus-containing groups
EC 2.8 Transferring sulfur-containing groups
EC 2.9 Transferring selenium-containing groups
EC 3  Hydrolases
EC 4  Lyases
EC 5  Isomerases
EC 6  Ligases

2.x: group transferred
2.x.x: details of group
Similar reaction/different structure


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

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

Enzymes of known function

All proteins

unique
Family size

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
Conservation of EC: sequence identity

B Rost 2002 J Mol Biol 318, 595-608
Conservation of EC number: BLAST

E-value better
but no 100% accuracy

B Rost 2002 J Mol Biol 318, 595-608
Conservation in detail

B Rost 2002 J Mol Biol 318, 595-608
Conservation of EC: PSI- vs. pair-BLAST

**First EC digit: accuracy**

**First EC digit: coverage**

**All EC digits: accuracy**

**All EC digits: coverage**

Number of proteins

Distance from threshold (identity/length)

Corresponding percentage sequence identity

**PAIR-BLAST**

**PAIR**

**PSI-BLAST**

**PSI**

+ Number of proteins

© Burkhard Rost (TUM Munich)
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: ER/Golgi
Inferring ER/Golgi localization

Trusted dataset of annotated proteins

ER and Golgi (true positives) 676-312

Other (true negatives) 8417

PSI-BLAST

%Seq.Id. HSSP E-value


© Kaz Wrzeszczynski: Thesis
# Homology-based inference

<table>
<thead>
<tr>
<th>HSSP</th>
<th>Predicted</th>
<th>Annotated as Golgi in Swiss-Prot</th>
<th>Other Swiss-Prot annotation</th>
<th>Hypothetical protein</th>
<th>estimated # of errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>23 (98%)</td>
<td>800</td>
<td>565</td>
<td>235</td>
<td>40</td>
<td>16</td>
</tr>
<tr>
<td>16 (95%)</td>
<td>1110</td>
<td>675</td>
<td>435</td>
<td>66</td>
<td>55</td>
</tr>
<tr>
<td>12 (90%)</td>
<td>1358</td>
<td>728</td>
<td>630</td>
<td>99</td>
<td>136</td>
</tr>
<tr>
<td>8 (85%)</td>
<td>1726</td>
<td>812</td>
<td>914</td>
<td>125</td>
<td>259</td>
</tr>
<tr>
<td>7 (78%)</td>
<td>1853</td>
<td>826</td>
<td>1027</td>
<td>134</td>
<td>407</td>
</tr>
</tbody>
</table>

Homology-based inference:

Cell cycle control
Kazimierz O. Wrzeszczynski

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

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

HSSP Distance

PSI-BLAST E-value

True Positive: 582 Cell Cycle Control Proteins
119 Sequence Unique

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>Accuracy</td>
<td>l.x.x.x.</td>
<td>x.x.x.x.IV</td>
</tr>
<tr>
<td>E-value 1x10^-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-value 1x10^-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 D= 5</td>
<td>70%</td>
<td>99%</td>
<td>82%</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>58%</td>
<td>60%</td>
</tr>
<tr>
<td>HSSP D= -5</td>
<td>25%</td>
<td>90%</td>
<td>55%</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>81%</td>
<td>85%</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></td>
<td>D=0 (55%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D=15 (65%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D= 25 (90%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D= 40 (95%)</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>99</td>
<td>3073</td>
</tr>
<tr>
<td></td>
<td></td>
<td>782</td>
</tr>
<tr>
<td></td>
<td></td>
<td>476</td>
</tr>
<tr>
<td></td>
<td></td>
<td>299</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>68</td>
<td>3162</td>
</tr>
<tr>
<td></td>
<td></td>
<td>574</td>
</tr>
<tr>
<td></td>
<td></td>
<td>310</td>
</tr>
<tr>
<td></td>
<td></td>
<td>203</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>15</td>
<td>970</td>
</tr>
<tr>
<td></td>
<td></td>
<td>181</td>
</tr>
<tr>
<td></td>
<td></td>
<td>96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>10</td>
<td>1005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>185</td>
</tr>
<tr>
<td></td>
<td></td>
<td>87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>5</td>
<td>1888</td>
</tr>
<tr>
<td></td>
<td></td>
<td>303</td>
</tr>
<tr>
<td></td>
<td></td>
<td>148</td>
</tr>
<tr>
<td></td>
<td></td>
<td>63</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>87</td>
<td>513</td>
</tr>
<tr>
<td></td>
<td></td>
<td>148</td>
</tr>
<tr>
<td></td>
<td></td>
<td>119</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>Sum</td>
<td>284</td>
<td>10611</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2173</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1236</td>
</tr>
<tr>
<td></td>
<td></td>
<td>747</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.

© Kaz Wrzeszczynski: Thesis

CellCycleDB

(Database of Cell Cycle Control Proteins in Eukaryotes)

It is

CellCycleDB Catalogues proteins involved in the Cell Cycle Control Process through homology transfer from experimental annotations.

It does

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.

You can

- Use CellCycleDB online (currently: single protein sequence submissions only)
- Search CellCycleDB using SRS: CellCycleDB
- download the CellCycleDB
- CellCycleDB Content Summary: Content Summary Tables

From Here

Who are we?

Contact
Email: kaz@cubic.bioc.columbia.edu
Tel: +1-212-305-4018 (Kazimierz O. Wrzeszczynski)
Fax: +1-212-305-7932
Homology-based inference: how much of human?
**Homology transfer accurate for very similar proteins**

<table>
<thead>
<tr>
<th>TRUE</th>
<th>FALSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>methyltransferase</td>
<td></td>
</tr>
</tbody>
</table>

<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>65%</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>

| 2/3 accuracy ; 2/4 coverage | 3/8 accuracy ; 4/4 coverage |

© Burkhard Rost (TUM Munich)
Homology transfer accurate for very similar proteins
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
A. **Paralogy problem**

```
query

template
```

*Template is a paralog, more likely to have diverged functionally*

---

B. **Moonlighting problem**

```
query

template
```

*Template may have more than one function*

---

C. **Multi-domain proteins problem**

```
query

template
```

*Template annotation may be based on a non-matching domain*

---

D. **Database mis-annotations problem**

```
query

template
```

*Template is mis-annotated

e.g. by homology with a multi-domain protein (see C)*
Evolutionary profile capture information
Conclusions today

- Function introduction
  - Molecular biology knows it all?
  - Can we compute life?
  - Protein function: terminology
  - Homology-based inference (examples)
    - Machine learning vs. homology
    - Challenges for homology-based inference
    - Inferring enzymatic activity
    - Inferring ER/Golgi
    - Inferring Cell-cycle control

NEXT

- Motif-based inference
How to answer the question?

© Wikipedia
comp sci: continue here
comp bio: continue here
this needs work!!
PUT in slide: idea of how to align