title: Introduction PP2
Winter 2012

short title: intro_2

lecture: Protein Prediction II - Protein Function / Burkhard Rost, TUM, 2012 winter
Announcements

 Videos: SciVe / www.rostlab.org

THANKS:
Tim Karl + Julia Gerke

Special lectures:
• Nov 15: Tobias Hamp
• Dec 16: Tatyana Goldberg
• Dec 18: Andrea Schafferhans
• Jan 29: Marco Punta (Pfam)
• Jan 31: Marco De Vivo (ISS Geneva)

No lecture:
• Dec 6 (Dies Academicus)
• Dec 20-Jan 6 (winter break+)

LAST lecture: Feb 5

Examen: Feb 7, 11:00 (likely this room)
• Makeup: may be Apr 18 - morning

CONTACT: Marlena Drabik assistant@rostlab.org

Let it go. Let it out. Let it all unravel. Let it free and it can be: A path on which to travel.
Today: Intro into function

- LAST WEEKs
  - Introduction to protein function I - localization

- THIS WEEK
  - Introduction to protein function 2-3

- NEXT WEEK
  - Tobias Hamp: Homology-based inference of function
Function introduction
  - Homology-based inference of function (concept)
  - Homology-based inference: methods
    - localization generic
    - ER/Golgi
    - cell cycle control

TODAY
  - Homology-based inference (examples)
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

Tuesday November 6, 2012
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)

© Kaz Wrzeszczynski: Thesis
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/](http://evolution.berkeley.edu/evosite/evo101/)

**species=mating**
Evolution: speciation

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

happily munching
Evolution: speciation

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

disaster strikes
Evolution: speciation

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

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

Carrion/Hooded crow  
(same species?)

Tuesday November 6, 2012
ADH: Alcohol dehydrogenase

© Wikipedia  ADH5 PDBid: 1m6h
Human glutathione-dependent formaldehyde dehydrogenase

ADH: Alcohol dehydrogenase

© Wikipedia

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

<table>
<thead>
<tr>
<th>Identifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC number</td>
</tr>
<tr>
<td>CAS number</td>
</tr>
<tr>
<td>IntEnz</td>
</tr>
<tr>
<td>KEGG</td>
</tr>
<tr>
<td>MetaCyc</td>
</tr>
<tr>
<td>PDBstructures</td>
</tr>
<tr>
<td>Gene</td>
</tr>
</tbody>
</table>

© GO

View this term in QuickGO.

Homolog

common ancestor

ADH1-bacteria

ADH1-yeast

ADH1-human

ADH1-plants
Homolog

common ancestor tree might be

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

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

Tree hypothetical
Homologs have the same ancestor
Orthologs are genes separated by speciation
Homolog, Ortholog

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

- ADH1-human
- ADH1-plants
- ADH1-bacteria
- ADH2-plants

Tree hypothetical

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”
problem 3: moonlighting
# Moonlighting proteins

<table>
<thead>
<tr>
<th>One function</th>
<th>Additional functions</th>
<th>Rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>PutA proline dehydrogenase</td>
<td>Transcriptional repressor</td>
<td>1, 2</td>
</tr>
<tr>
<td>Phosphogluco 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>Glyceraldehyde-3-phosphate dehydrogenase</td>
<td>11</td>
</tr>
<tr>
<td>Aconitase</td>
<td>Iron-responsive-element binding protein (IRE-BP)</td>
<td>12,</td>
</tr>
<tr>
<td>Carbinolamine dehydratase</td>
<td>Dimerization cofactor (DcoH)</td>
<td>15</td>
</tr>
<tr>
<td><em>Escherichia coli</em> thioredoxin</td>
<td>Subunit of T7 DNA polymerase</td>
<td>16</td>
</tr>
<tr>
<td><em>E. coli</em> aspartate receptor</td>
<td>Maltose-binding-protein receptor</td>
<td>17</td>
</tr>
<tr>
<td>PMS2 mismatch-repair enzyme</td>
<td>Hypermutation of antibody variable chains</td>
<td>18</td>
</tr>
<tr>
<td>Ribosomal proteins</td>
<td>DNA repair, translational regulators, development regulators, etc.</td>
<td>19</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>
<td>20</td>
</tr>
<tr>
<td>CFTR chloride channel</td>
<td>Regulator of other epithelial anion channels</td>
<td>21</td>
</tr>
<tr>
<td>P-glycoprotein (transporter)</td>
<td>Regulator of cell-swelling ion channel</td>
<td>22</td>
</tr>
<tr>
<td>Thrombin protease</td>
<td>Ligand for cell surface receptors</td>
<td>23</td>
</tr>
<tr>
<td>Thymidylate synthase</td>
<td>Translation inhibitor</td>
<td>24</td>
</tr>
<tr>
<td><em>E. coli birA</em> biotin synthetase</td>
<td>Bio operon repressor</td>
<td>25</td>
</tr>
<tr>
<td>Mitochondrial LON protease</td>
<td>Chaperone</td>
<td>26</td>
</tr>
<tr>
<td>Bacterial FtsH chaperone</td>
<td>Metalloprotease</td>
<td>27</td>
</tr>
<tr>
<td>Band-3 anion exchanger</td>
<td>Regulator of glycolysis</td>
<td></td>
</tr>
</tbody>
</table>
How well does it work?
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>
</tbody>
</table>
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>
<tr>
<td>99%</td>
<td>magnesium protoporphyrin IX methyltransferase</td>
</tr>
</tbody>
</table>
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>65%</td>
<td>inositol 3-methyltransferase</td>
</tr>
</tbody>
</table>

2/3 accuracy ; 2/4 coverage
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**

<table>
<thead>
<tr>
<th>src_avis2</th>
<th>src_aviss</th>
<th>src_avisr</th>
<th>src_avis</th>
<th>src_avis1</th>
<th>src_aviss1</th>
<th>src_avisr1</th>
<th>true_avis1</th>
<th>true_avis2</th>
<th>true_avis3</th>
</tr>
</thead>
<tbody>
<tr>
<td>src_avis2</td>
<td>src_aviss</td>
<td>src_avisr</td>
<td>src_avis</td>
<td>src_avis1</td>
<td>src_aviss1</td>
<td>src_avisr1</td>
<td>true_avis1</td>
<td>true_avis2</td>
<td>true_avis3</td>
</tr>
</tbody>
</table>

**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
Performance of homology-based inference

![Graph showing pairwise sequence similarity with accuracy and coverage metrics.]

- Midnight zone
- Twilight zone
- Safe zone
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

Tuesday November 6, 2012
How to assess alignment accuracy?
How to assess alignment accuracy?
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

Tuesday November 6, 2012
Annotation transfer: Localization

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

Tuesday November 6, 2012
Annotation transfer: Localization

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

Tuesday November 6, 2012
Homology-based inference depends

Structure

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

Tuesday November 6, 2012
Homology inference localization

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

Tuesday November 6, 2012
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

Subclasses:
- EC4.1: carbon-carbon lyases
  - EC4.1.1: carboxy lyases
  - EC4.1.2: aldehyde lyases
  - EC4.1.3: oxo-acid lyases
  - EC4.1.99: other 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

References:

# Enzyme classification (EC)

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

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

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

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

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

Conservation of function

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

![Graph showing the percentage of families against the number of proteins in the family, with two lines indicating biased and unbiased percentages.](image)

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

B Rost 2002 J Mol Biol 318, 595-608

Tuesday November 6, 2012
REAL conservation of EC number

Bias:
50% found at >90% right

Real:
50% found at <15% right!
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 bad!

B Rost *J Mol Biol* **318**, 595-608

Tuesday November 6, 2012
Conservation of EC number: BLAST

E-value better
but no 100% accuracy


Tuesday November 6, 2012
Conservation in detail

Conservation of EC: PSI- vs. pair-BLAST

- Pair-BLAST
- PSI-BLAST

- Number of proteins
- Distance from threshold (identity/length)
- Corresponding percentage sequence identity

- First EC digit: accuracy
- All EC digits: accuracy
- First EC digit: coverage
- All EC digits: coverage
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


© Kaz Wrzeszczynski: Thesis

Tuesday November 6, 2012
Inferring ER/Golgi localization

Trusted dataset of annotated proteins

ER and Golgi (true positives) 676-312

Other (true negatives) 8417


© Kaz Wrzeszczynski: Thesis

Tuesday November 6, 2012
Inferring ER/Golgi localization

Trusted dataset of annotated proteins

ER and Golgi (true positives) 676-312

PSI-BLAST

Other (true negatives) 8417


© Kaz Wrzeszczynski: Thesis

Tuesday November 6, 2012
Homology-based inference: Golgi


Tuesday November 6, 2012
Homology-based inference: Golgi

Homology-based inference: Golgi


Tuesday November 6, 2012
Homology-based inference: Golgi


Tuesday November 6, 2012
Homology-based inference: Golgi

Homology-based inference: Golgi

Homology-based inference: ER


Tuesday November 6, 2012
Homology-based inference: ER & Golgi

Homology-based inference: ER & Golgi


Tuesday November 6, 2012
Homology-based inference: ER & Golgi

Homology-based inference: ER & Golgi

Homology-based inference: ER & Golgi


Tuesday November 6, 2012
Homology-based inference: ER & Golgi

- PSI-BLAST accuracy
- PSI-BLAST coverage

Homology-based inference: ER & Golgi


Tuesday November 6, 2012
### Homology: ER & Golgi 4 proteomes

#### Annotation at HSSP>23 -> 98% accuracy

<table>
<thead>
<tr>
<th>Proteome</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><em>S. cerevisiae</em></td>
<td>70</td>
<td>53</td>
<td>17</td>
<td>8</td>
<td>1-2</td>
</tr>
<tr>
<td><em>A. thaliana</em></td>
<td>70</td>
<td>31</td>
<td>39</td>
<td>5</td>
<td>1-2</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>57</td>
<td>23</td>
<td>34</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td><em>D. melanogaster</em></td>
<td>61</td>
<td>40</td>
<td>32</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>M. musculus</em></td>
<td>195</td>
<td>145</td>
<td>50</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td>347</td>
<td>273</td>
<td>74</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td><strong>All 6</strong></td>
<td>800</td>
<td>565</td>
<td>235</td>
<td>40</td>
<td>16</td>
</tr>
</tbody>
</table>
### 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>


Tuesday November 6, 2012
Homology-based inference: Cell cycle control
Kazimierz O. Wrzeszczynski

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

![Diagram of the cell cycle]

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

Accu./Cov.

Homology

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>E-value = $1 \times 10^{-8}$</td>
<td>E-value = $1 \times 10^{-3}$</td>
</tr>
<tr>
<td></td>
<td>Coverage</td>
<td>64% 60%</td>
<td>29% 81%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>98% 58%</td>
<td>95% 62%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>50% 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


© Kaz Wrzeszczynski: Thesis


Tuesday November 6, 2012
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><strong>Homo sapiens</strong></td>
<td>99</td>
<td>3073</td>
</tr>
<tr>
<td><strong>Mus musculus</strong></td>
<td>68</td>
<td>3162</td>
</tr>
<tr>
<td><strong>Drosophila melanogaster</strong></td>
<td>15</td>
<td>970</td>
</tr>
<tr>
<td><strong>Caenorhabditis elegans</strong></td>
<td>10</td>
<td>1005</td>
</tr>
<tr>
<td><strong>Arabidopsis thaliana</strong></td>
<td>5</td>
<td>1888</td>
</tr>
<tr>
<td><strong>Saccharomyces cerevisiae</strong></td>
<td>87</td>
<td>513</td>
</tr>
<tr>
<td><strong>Sum</strong></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 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 HSPS-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

© Kaz Wrzeszczynski: Thesis


Tuesday November 6, 2012
Homology-based inference: how much of human?
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>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
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

Less than 25% have *some* annotation
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**

Template is a paralog, more likely to have diverged functionally

B. **Moonlighting problem**

Template may have more than one function

C. **Multi-domain proteins problem**

Template annotation may be based on a non-matching domain

D. **Database mis-annotations problem**

Template is mis-annotated

* e.g. by homology with a multi-domain protein (see C)
Evolutionary profile capture information
Complex computation for prediction

Sequence → PSI-BLAST → Filter

PROFsec → PROFacc

1999
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
Lecture plan (PP2 function)

01: 2012/10/16: no lecture
02: 2012/10/18: welcome: who we are
03: 2012/10/23: individualized medicine
04: 2012/10/25: Intro - function 1: concepts
05: 2012/10/30: Tatyana Goldberg: localization
06: 2012/11/01: no lecture: All Saints
07: 2012/11/06: Intro - function 2: homology
08: 2012/11/08: Intro - function 3: motifs
09: 2012/11/13: Localization 1
11: 2012/11/20: Localization 2
13: 2012/11/27: Localization 4
14: 2012/11/29: SNP effect 1
15: 2012/12/04: SNP effect 2
16: 2012/12/06: no lecture: Dies Academicus
17: 2012/12/11: SNP effect 3
18: 2012/12/13: Protein-protein interaction 1
19: 2012/12/18: Andrea Schafferhans: 3D function prediction
20: 2012/12/20: no lecture
21-24: no lectures - winter break (2012/12/23 - 2013/01/06)
25: 2013/01/08: Protein-protein interaction 2
26: 2013/01/10: Protein-protein interaction 3
27: 2013/01/15: Protein-DNA interaction 1
28: 2013/01/17: Protein-DNA/RNA interaction 2
29: 2013/01/22: Andrea Schafferhans: Docking
30: 2013/01/24: networks
31: 2013/01/29: Marco Punta (Pfam)
32: 2013/01/31: Marco De Vivo (ISS Geneva)
33: 2013/02/05: wrap-up
34: 2013/02/07: examen