Homology-based prediction of protein function

Tobias Hamp
Winter term 2013/2014
Overview

- Methods
  - (PSI-)BLAST
  - Motifs
  - Machine learning
  - Text-mining
  - ...
- Target functions
  - EC numbers
  - Localization
  - Transmembrane yes/no
Overview

- **Methods**
  - (PSI-)BLAST
  - Motifs
  - Machine learning
  - Text-mining
  - ...

- **Target functions**
  - EC numbers
  - Localization
  - Transmembrane yes/no
Lecture "Protein Prediction II", 2010/11

• Lecture held by Prof. Burkhard Rost
Lecture "Protein Prediction II", 2010/11

• Lecture held by Prof. Burkhard Rost

• Accompanying exercises: Esmeralda Vicedo, Christian Schaefer, Tobias Hamp
Lecture "Protein Prediction II", 2010/11

- Lecture held by Prof. Burkhard Rost
- Accompanying exercises: Esmeralda Vicedo, Christian Schaefer, Tobias Hamp
- Exercise task: Implement simple function predictor
- 16 students split into 3 groups, each group supposed to implement own predictor
Target Functions

• “… function is everything that happens to or through a protein.” Rost et al., *CMLS*, 2003
Target Functions

• “… function is everything that happens to or through a protein.” Rost et al., *CMLS*, 2003

• It's complex
Target Functions

- “… function is everything that happens to or through a protein.” Rost et al., *CMLS*, 2003
- It's complex
- Best we can do today: The Gene Ontology
  - Quasi-standard for function annotation
  - 3 Categories:
    - Molecular Function, Biological Process, Cellular Component
    - 1 protein => 3 types of annotations
  - Hierarchical structure
Human hemoglobin subunit alpha
Nearest Neighbor Principles

1. PSI-BLAST target against GO annotated part of SwissProt

2. Transfer GO terms of the best hit(s) to the target
   - 1-Nearest-Neighbor: only consider first hit (best e-Value)
   - k-Nearest-Neighbor: consider first k hits (alternatively: hits below a certain e-Value)
   - Weighted k-Nearest-Neighbor: GO terms of more significant hits are more important
Methods in Detail - Group A

- Members
  - Ariane Boehm
  - Tatjana Braun
  - Rebecca Kassner
  - Cedric Landerer
  - Yannick Mahlich

- Method (key features)
  - Weighted 6-Nearest-Neighbor
  - Considers hit count and e-Value of each term
  - Outputs 3 terms per ontology
  - Outputs 3 scores (0.0, 0.5, 1.0)
Methods in Detail - Group A

Top 6 BLAST Hits

protein 1 → GO7, GO4, GO9
...
protein 6 → GO7, GO11

GO Term Scoring; GO Tree Assembly

Redundancy Reduction of Branches; Output

Homology-based inference sets the bar high for protein function prediction. Hamp et al. 2012
Methods in Detail - Group B

• Members
  – Mark Heron
  – Thomas Hopf
  – Stefanie Kaufmann
  – Denis Krompass
  – Stefan Seemayer

• Method (key features)
  – Weighted k-Nearest-Neighbor
  – Sophisticated continuous scoring scheme based on hit count, e-Value, ontology structure
  – Score normalization
  – Parameter optimization
Methods in Detail: Group B

Homology-based inference sets the bar high for protein function prediction. Hamp et al. 2012

Top BLAST Hits
- protein 1  e-val 1e-20
  - GO7, GO4, GO9
- ...
- protein N  e-val 1e-03
  - GO4, GO37, GO13

Raw Template Score
- evals = log(1e-20),...,log(1e-03)
- Raw Template Score = Ø(evals)+stddev(evals)

Template Quality Score

GO Tree Assembly: E-Value Assignment
- GO4
- log(1e-20), log(1e-03)

Term Support Calculations
- 0.74
- 1.0
- 0.66
- 0.43
- 0.22
- 0.30

Combined Leaf Score
- 0.72
- 0.63
- 0.65

Output: Template Quality Score * Combined Leaf Score
- 0.91*0.72 = 0.66
- 0.91*0.63 = 0.57
- 0.91*0.65 = 0.59
Methods in Detail - Group C

- Dominik Achten
- Florian Auer
- Maximilian Hecht
- Peter Hoenigschmid
- Michael Kiening
- Manfred Roos

Method (key features)

- Weighted e-Value based Nearest-Neighbor
- Custom scoring scheme based on Blast score and hit count
- Continuous term scores
Methods in Detail: Group C

Homology-based inference sets the bar high for protein function prediction. Hamp et al. 2012
Homology-based inference sets the bar high for protein function prediction. Hamp et al. 2012
Evaluation Measure

Precision vs. Recall graph with a point labeled P1.
Evaluation Measure

Precision vs. Recall
Evaluation Measure

Precision

Recall
Evaluation Measure

\[ F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \]
Evaluation

• Targets: 10,000 random SwissProt proteins

• Templates: the rest of SwissProt (~400,000 proteins)

• Evaluation based on:
  
  – All 10,000 targets (⇒ recall of an unpredicted target: 0.0)
  
  – Commonly predicted targets, i.e. targets for which all three methods have made a prediction
Evaluation Results

All Targets

Common Targets

Precision

Recall

Method A
Method B
Method C
Meta
Random

Method A
Method B
Method C
Meta
Random
Summary

• Hasty implementation, but ...
• Nearest-Neighbor approaches work surprisingly well (?)
• Technical details matter
• Maybe too many "random giveaways" in measures based on entire DAG
CAFA - CRITICAL ASSESSMENT OF FUNCTION ANNOTATIONS
Going Big

- Great goal: Use all data to predict all functions
- Last 10 years: many function predictors, but
  - Different data sets
  - Different input features
  - Different target classes
  - Different evaluation measures
  - Different strengths/weaknesses
Critical Assessment

- Year 2011: 1st Critical Assessment of Function Annotations (CAFA)
  - Community effort to measure the current state of the art in function prediction
  - Independent assessors, supported by renowned principal investigators
CAFA 2011 – Basic Steps

1. Find organizers

2. Agree on target functions

3. Agree on target proteins

4. Develop evaluation measures

5. Evaluate predictions
CAFA - Step 2: Target Functions

- Best we can do today: The Gene Ontology

Easy, right?
CAFA - Step 2: Target Functions

- Gene Ontology difficulties
  - Different types of relations
  - Different parent ⇔ child inference rules
  - Different evidence codes
  - Different qualifiers
  - Different GOs/GO versions
  - Different annotation resources
CAFA - Step 2: Target Functions

- Gene Ontology difficulties
  - Different types of relations
  - Different parent↔child inference rules
  - Different evidence codes
  - Different qualifiers
  - Different GOs/GO versions
  - Different annotation resources
CAFA - Step 2: Target Functions

- Gene Ontology difficulties
  - Different types of relations
  - Different parent ⇔ child inference rules
  - Different evidence codes
  - Different qualifiers
  - Different GOs/GO versions
  - Different annotation resources
CAFA - Step 2: Target Functions

• Gene Ontology difficulties
  – Different types of relations
  – Different parent ⇔ child inference rules
  – Different evidence codes
  – Different qualifiers
  – Different GOs/GO versions
  – Different annotation resources

The **Qualifier** column is used for flags that modify the interpretation of an annotation. Allowable values are **NOT**, **contributes_to**, and **colocalizes_with**.
CAFA - Step 2: Target Functions

• Gene Ontology difficulties
  – Different types of relations
  – Different parent↔child inference rules
  – Different evidence codes
  – Different qualifiers
  – Different GOs/GO versions
  – Different annotation resources
CAFA - Step 2: Target Functions

- Gene Ontology difficulties
  - Different types of relations
  - Different parent $\leftrightarrow$ child inference rules
  - Different evidence codes
  - Different qualifiers
  - Different GOs/GO versions
  - Different annotation resources
CAFA - Step 2: Target Functions

• Gene Ontology: Simplifications in CAFA
  - Only one type of relation/Precomputed graphs
  - Experimental evidence codes only
  - Agreement on GO version
  - Swissprot

  IDA: Inferred from direct assay
  IPI: Inferred from phys. interaction
  IEP: Inferred from expr. pattern
  IC: Inferred by curator
  IMP: Inferred from mutant phenotype
  IGI: Inferred from genetic interaction
  TAS: Traceable author statement
  EXP: Inferred from experiment

• Molecular Function Ontology
  - 8728 terms; 7003 leaves; (max_depth=11; branching_factor=5.9)

• Biological Process Ontology
  - 18982 terms; 8125 leaves; (max_depth=14; branching_factor=6.4)
Human Hemoglobin subunit alpha

GO:0015701 bicarbonate transport

GO:0008150 biological_process
GO:0065007 biological regulation

GO:0005575 cellular_component
GO:0005623 cell

GO:0003674 molecular_function
GO:0016209 antioxidant activity
GO:0003824 catalytic activity
CAFA 2011 – Basic Steps

1. Find organizers ✓
2. Agree on target functions ✓
3. Agree on target proteins
4. Develop evaluation measures
5. Evaluate predictions
CAFA - Step 3: Target Proteins

- From each CAFA competitor, we want predictions of proteins which ...
  - do not have experimental annotations right now
  - will have experimental annotations in 4 months
  - suffice to establish statistical significance (in terms of number of predictions)
CAFA - Step 3: Target Proteins

• From each CAFA competitor, we want predictions of proteins which ...
  – do not have experimental annotations right now
  – will have experimental annotations in 4 months
  – suffice to establish statistical significance (in terms of number of predictions)

• How would you do that?
CAFA - Step 3: Target Proteins

1a. Reduce databases $t_0$ to a few organisms (10)
1b. Remove all proteins with exp. annotations from DB's $t_0$ => Target proteins $t_0$ (= CAFA targets; 50k)

2. Target annotations $t_0 =$
Swiss-Prot($t_2$) – (Swiss-Prot($t_1$) ∪ GOA($t_1$) ∪ GO($t_1$))
CAFA - Step 3: Target Proteins

Enzymes of known function

Proteins of known function

All proteins

unique
rand new

unique

All proteins

© Burkhard Rost
CAFA - Step 3: Target Proteins

Annotations per organism in Swissprot (Jan 2011)

CAFA targets

<table>
<thead>
<tr>
<th>Organism</th>
<th>Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>32</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>3</td>
</tr>
<tr>
<td>Xenopus laevis</td>
<td>7</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>123</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>136</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>23</td>
</tr>
<tr>
<td>Dictyostelium discoideum</td>
<td>16</td>
</tr>
<tr>
<td>Pseudomonas aeruginosa</td>
<td>1</td>
</tr>
<tr>
<td>Bacillus subtilis</td>
<td>1</td>
</tr>
<tr>
<td>Escherichia coli K-12</td>
<td>253</td>
</tr>
</tbody>
</table>

© Predrag Radivojac, Ph.D.
CAFA - Step 3: Target Proteins

Evidence codes in Swissprot (Jan 2011)

CAFA targets

© Predrag Radivojac, Ph.D.
CAFA - Step 3: Target Proteins

Leaves per protein in Swissprot (Jan 2011)

- MFO:
  - 1 leaf
  - 2 leaves
  - 3 leaves

- BPO:
  - 2 leaves
  - 3 leaves
  - 4 leaves

CAFA targets:

- 1 leaf
- 2 leaves
- 3 leaves
- 4 leaves
- 5 leaves
- 6 leaves
- 7 leaves
- 8+ leaves

© Predrag Radivojac, Ph.D.
CAFA - Step 3: Target Proteins

MFO Leaves
- protein binding
- protein homodimerization activity
- zinc ion binding

BPO Leaves
- cell adhesion
- response to DNA damage stimulus
- nuclear mRNA splicing, via spliceosome
CAFA - Step 3: Target Proteins

Sequence identity to already annotated proteins

- **MFO**
- **BPO**

Sequence identity vs Count

© Predrag Radivojac, Ph.D.
CAFA 2011 – Basic Steps

1. Find organizers
2. Agree on target functions
3. Agree on target proteins
4. Develop evaluation measures
5. Evaluate predictions
CAFA - Step 4: Evaluation Measures

- We now have ...
  - the targets
  - their annotations
  - predictions by ~50 different methods

- Q: How can we compare their accuracies?
- A: Threshold measure and maximum F1 score (see previous slides)
Evaluation Measure

\[ F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \]
CAFA 2011 – Basic Steps

1. Find organizers ✔
2. Agree on target functions ✔
3. Agree on target proteins ✔
4. Develop evaluation measures ✔
5. Evaluate predictions
CAFA – Step 6: Evaluation

- Only model 1 is evaluated for every group
- Evaluations are measured separately for Molecular-Function and Biological-process Ontologies
- No measure is perfect
- Implementation of evaluation code a joint effort
- Identities of the groups hidden
- Baseline classifiers
  - Priors: score is simply based on prevalence of a specific annotation
  - BLAST: transfer annotations based on E-value threshold
<CONTENT REMOVED>
Threshold Measure
MFO

<CONTENT REMOVED>
Threshold Measure
BPO

<CONTENT REMOVED>
BPO difficult targets
Max F1 Scores

<CONTENT REMOVED>
CAFA – Step 6: Evaluation

<CONTENT REMOVED>
CAFA – Step 6: Evaluation

Sequence-based prediction works well, but a lot of room for improvement
METHOD OPTIMIZATIONS AFTER CAFA
Post CAFA Optimizations

• Our classifiers for CAFA
  – StudentA-C
  – PSI-BLAST
Post CAFA Optimizations

• Our classifiers for CAFA
  – **StudentA-C**
  – PSI-BLAST

• Issues
  – Only one of three classifiers evaluated
Post CAFA Optimizations

• Our classifiers for CAFA
  – StudentA-C
  – PSI-BLAST

• Issues
  – Only one of three classifiers evaluated
  – No/False optimizations
Post CAFA Optimizations

• Our classifiers for CAFA
  – StudentA-C
  – PSI-BLAST

• Issues
  – Only one of three classifiers evaluated
  – No/False optimizations
  – No time to submit meta predictions
Post CAFA Optimizations

• Our classifiers for CAFA
  – **StudentA-C**
  – PSI-BLAST

• Issues
  – Only one of three classifiers evaluated
  – No/False optimizations
  – No time to submit meta predictions
  – Bug(s)
Post CAFA Optimizations

• Our classifiers for CAFA
  – StudentA-C
  – PSI-BLAST

• Issues
  – Only one of three classifiers evaluated
  – No/False optimizations
  – No time to submit meta predictions
  – Bug(s)
  – No original measure implementations available
Post CAFA Optimizations

• Our classifiers for CAFA
  – StudentA-C
  – PSI-BLAST

• Issues
  – Only one of three classifiers evaluated
  – No/False optimizations
  – No time to submit meta predictions
  – Bug(s)
  – No original measure implementations available
  – Details (evidence codes, ...) unavailable
Post CAFA Optimizations

- Let's improve!

Optimization idea

2x

Proteins of known function

All proteins

brand new

unique
# Post CAFA Optimizations

Time:
- $t_0$: Jan 1, 2010
- $t_1$: Jan 18, 2011
- $t_2$: May 31, 2011

<table>
<thead>
<tr>
<th>Templates</th>
<th>Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;2010</td>
<td>2010</td>
</tr>
<tr>
<td>Swiss-Prot($t_0$)</td>
<td>Swiss-Prot($t_1$) – Swiss-Prot($t_0$)</td>
</tr>
<tr>
<td>&lt;2011</td>
<td>2011</td>
</tr>
<tr>
<td>Swiss-Prot($t_1$)</td>
<td>Swiss-Prot($t_2$) – Swiss-Prot($t_1$)</td>
</tr>
</tbody>
</table>
Post CAFA Optimizations

- **Data sets**
  - Templates for parameter optimization:
    Anything before Jan 2010 (<2010)
  - Targets for parameter optimization:
    Annotations added between Jan 2010 and Dec 2010 (2010)
  - Templates for CAFA predictions:
    Anything before Jan 2011 (<2011)
  - Targets for CAFA predictions:
    Original CAFA targets (Jan 2011 – May 2011; 2011)
Post CAFA Optimizations

- Optimization of single classifiers
  - Free parameters: $k$'s, $e$-Value's, other algorithmic alternatives (e.g. score cutoff switches, %seq. id. instead of %pos. id.)

$\Rightarrow$ 36, 54, 72 parameter combinations for StudentA-C
Post CAFA Optimizations

• Optimization of single classifiers
  – Free parameters: $k$'s, $e$-Value's, other algorithmic alternatives (e.g. score cutoff switches, %seq. id. instead of %pos. id.)
  
  => 36, 54, 72 parameter combinations for StudentA-C

• Data set
  – Templates: <2010
  – Targets: random subset of 2010
Post CAFA Optimizations

- Optimization of single classifiers
  - Free parameters: k's, e-Value's, other algorithmic alternatives (e.g. score cutoff switches, %seq. id. instead of %pos. id.)
  => 36, 54, 72 parameter combinations for StudentA-C

- Data set
  - Templates: <2010
  - Targets: random subset of 2010
  => Try all parameter combinations and pick best
  => StudentA'-C'
Post CAFA Optimizations

- Training of meta classifier **MetaStudent**
  - **Principle**: Use predictions of single classifiers as input to another classifier
  - Here: meta classifier = linear regression:
    \[ xA' + yB' + zC' + i = p \]
    - \( A', B', C' \) are probability estimates for the same protein-GO term association by the three different student methods
    - \( x, y, z, i \) are the weights to be optimized
    - \( p \) is the new output probability for this protein-GO term association
Post CAFA Optimizations

• Training of meta classifier
  – Problem:
    We used the whole set 2010 for single method parameter optimization. Hence, it is no longer an ideal set for meta classifier training.
    ⇒ Random split of set 2010: 2010a and 2010b
    ⇒ First re-train methods StudentA-C with 2010a to predict set 2010b
    ⇒ Then change roles of 2010a and 2010b and repeat
    ⇒ This creates independent, yet optimized predictions for the entire set 2010
    ⇒ Use these predictions as input for the linear regression
Post CAFA Optimizations

• Training of meta classifier

Having trained StudentA-C and MetaStudent without using any data after Jan 2011, we can now predict the original CAFA targets
Homology-based inference sets the bar high for protein function prediction. Hamp et al. 2012
Homology-based inference sets the bar high for protein function prediction. Hamp et al. 2012
### Post CAFA Optimizations – Result Ranks

<table>
<thead>
<tr>
<th></th>
<th>BPO</th>
<th></th>
<th></th>
<th>MFO</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-20</td>
<td>Threshold</td>
<td>Leaf</td>
<td>Top-20</td>
<td>Threshold</td>
<td>Leaf</td>
</tr>
<tr>
<td>Priors</td>
<td>8</td>
<td>8</td>
<td>11</td>
<td>7</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>Priors'</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>BLAST</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>6</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>GOtcha</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>2</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>StudentA</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>StudentA'</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>StudentB</td>
<td>11</td>
<td>11</td>
<td>7</td>
<td>11</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>StudentB'</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>StudentC</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>9</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>StudentC'</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>MetaStudent</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

Homology-based inference sets the bar high for protein function prediction. Hamp et al. 2012
Post CAFA Optimiizations - Summary

• Loose coupling of simple nearest neighbors method works well
• No single measure is enough
• No single method excells in all categories
• Still a long road ahead, especially for leaf terms
• Problems of homology inference persist