Homology-based prediction of protein function

Protein Prediction 2 - Protein function
TUM Winter 2014/15

Tobias Hamp
Recap

- Previous lectures
  - Methods
    - (PSI-)BLAST
  - Target functions
    - EC numbers
    - Localization
    - Cell cycle control yes/no
Target functions

- We want to predict protein function from protein sequence

DNA → Protein

E.g.:
- Binding ADP (MFO)
- For ATP Synthesis (BPO)
- In the mitochondrion (CCO)
Human hemoglobin subunit alpha

- GO:0015701 bicarbonate transport
- GO:0008150 biological_process
- GO:0005575 cellular_component
- GO:00065007 biological regulation
- GO:0005623 cell
- GO:003674 molecular_function
- GO:0016209 antioxidant activity
- GO:0003824 catalytic activity
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

$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

Precision

Recall

max $F_1$
Evaluation Measure

CRITICISM?

\[ F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \]
CAFA - Step 3: Target Proteins

Sequence identity to already annotated proteins

MFO

BPO

Sequence identity

Count
BPO difficult targets

Graph showing precision vs. recall for various models with different F and C scores:
- Jones–UCL F=0.43
- Argot2 F=0.42 C=0.91
- PANNZER F=0.39 C=0.76
- PDCN F=0.39
- ESG F=0.38 C=0.95
- Rost Lab F=0.38 C=0.80
- BAR+ F=0.35 C=0.58
- Team Orengo F=0.32 C=0.67
- MS-kNN F=0.32
- Tian Lab F=0.28 C=0.69
- Naive F=0.26
- BLAST F=0.22 C=0.97
CAFA – Step 6: Evaluation

- Top 3 method
  - **Jones-UCL**: sequence, gene expression, protein-protein interaction data, text mining
  - **Argot2**: BLAST + HMMer
  - **PANNZER**: BLAST + HMMer + text mining

- Another popular method
  - **GOstruct** (MouseFunc winner): BLAST

- Money quote:
  "...we found that the high throughput data integration based predictions [...] contributed least to the final predictions" (D. Jones)

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

- 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
Let's improve!

Optimization idea

Proteins of known function

All proteins

2x

brand new

unique

Post CAFA Optimizations
**Post CAFA Optimizations**

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

<table>
<thead>
<tr>
<th>Templates</th>
<th>Targets</th>
<th>Date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&lt;2010$</td>
<td>$2010$</td>
<td></td>
<td>Swiss-Prot(t₀)</td>
</tr>
<tr>
<td>$2011$</td>
<td>$2011$</td>
<td></td>
<td>Swiss-Prot(t₁) – Swiss-Prot(t₀)</td>
</tr>
<tr>
<td>$&lt;2011$</td>
<td></td>
<td></td>
<td>Swiss-Prot(t₁)</td>
</tr>
<tr>
<td></td>
<td>$2011$</td>
<td></td>
<td>Swiss-Prot(t₂) – Swiss-Prot(t₁)</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. scorings, %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:
  - \[ xA' + yB' + zC' + i = p \]
  - Where
    - \( A', B', C' \) 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.
  - $\Rightarrow$ Random split of set $2010$: $2010a$ and $2010b$
  - $\Rightarrow$ First re-train methods StudentA-C with $2010a$ to predict set $2010b$
  - $\Rightarrow$ Then change roles of $2010a$ and $2010b$ and repeat
  - $\Rightarrow$ This creates independent, yet optimized predictions for the entire set $2010$
  - $\Rightarrow$ 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

Here are the results
Post CAFA Optimizations - Results

Homology-based inference sets the bar high for protein function prediction. Hamp et al. 2012 © Tobias Hamp (TUM Munich)
Homology-based inference sets the bar high for protein function prediction. Hamp et al. 2012 © Tobias Hamp (TUM Munich)
### Post CAFA Optimizations – Result Ranks

<table>
<thead>
<tr>
<th></th>
<th>BPO</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>
</tr>
<tr>
<td>Priors</td>
<td>8</td>
<td>8</td>
<td>11</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Priors'</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>BLAST</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>G0tcha</td>
<td>6</td>
<td>6</td>
<td>8</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>StudentA</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>StudentA'</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>StudentB</td>
<td>11</td>
<td>11</td>
<td>7</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>StudentB'</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>StudentC</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>StudentC'</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>MetaStudent'</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Homology-based inference sets the bar high for protein function prediction. Hamp et al. 2012 © Tobias Hamp (TUM Munich)
Post CAFA Optimizations - 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
CAFA II
CAFA II

- CAFA I was a big success (many new projects, methods and publications)

- CAFA II: ISMB/ECCB 2013/14 (3 years later)

- Same principle, but ...
  - New rules
  - New prediction disciplines
  - New targets
CAFA II - Rules

- The Gene Ontology got wind...
  - New 'Annotation Blacklist': protein<>GO term annotations that should not exist

<table>
<thead>
<tr>
<th>Gene Product ID</th>
<th>GO Identifier</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1ZBR5</td>
<td>GO:0004842</td>
<td>Lacks the conserved Cys residue necessary for ubiquitin-conjugating enzyme E2 activity</td>
</tr>
<tr>
<td>B0YPL7</td>
<td>GO:0015979</td>
<td>This organism being non-photosynthetic, the role of this protein is uncertain</td>
</tr>
</tbody>
</table>
CAFA II - Rules

- The Gene Ontology got wind...
  - New 'Annotation Blacklist': protein<=>GO term annotations that should not exist
  - Similarly: 'Taxon restrictions' – GO terms that should not be used for certain taxa
CAFA II - Rules

<table>
<thead>
<tr>
<th>Taxon Rule</th>
<th>Ancestor GO ID</th>
<th>Ancestor GO Term Name</th>
<th>Relationship</th>
<th>Taxon ID</th>
<th>Taxon</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOTAX:0000182</td>
<td>GO:0009908</td>
<td>flower development</td>
<td>only_in_taxon</td>
<td>3398</td>
<td>Magnoliophyta</td>
</tr>
<tr>
<td>GOTAX:0000299</td>
<td>GO:0032501</td>
<td>multicellular organismal process</td>
<td>only_in_taxon</td>
<td>2759</td>
<td>Eukaryota</td>
</tr>
<tr>
<td>GOTAX:0000537</td>
<td>GO:0032501</td>
<td>multicellular organismal process</td>
<td>never_in_taxon</td>
<td>4932</td>
<td>Saccharomyces cerevisiae</td>
</tr>
<tr>
<td>GOTAX:0000538</td>
<td>GO:0032501</td>
<td>multicellular organismal process</td>
<td>never_in_taxon</td>
<td>4896</td>
<td>Schizosaccharomyces pombe</td>
</tr>
</tbody>
</table>

More information on taxon constraints in GO is available [here](#).
CAFA II - Rules

The Gene Ontology got wind...

- New 'Annotation Blacklist': protein$\Leftrightarrow$GO term annotations that should not exist

- Similarly: 'Taxon restrictions' – GO terms that should not be used for certain taxa

- Some terms now unsuitable for automatic annotation. E.g.:
  - 'S phase' (GO:0051320): describes only time period
  - 'response to stress' (GO:0006950): too general
CAFA II - Rules

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

© Gene Ontology
CAFA II – New disciplines

- Targets so far: every protein not experimentally annotated in any database at submission deadline
- Now 2 additional categories
  - Targets that are not annotated in all ontologies (→ predict annotations in missing ontologies)
  - Targets that are already annotated in the same ontology (→ predict additional annotations)
- New evaluation measure: information theory & semantic distances
- Two evaluation modes: all possible targets vs. all predicted targets
CAFA II – Metastudent 2.0

- Only minimal changes to Metastudent 1.0
- All methods must use the same BLAST parameters
- Database for template annotations as another free parameter

Exp. annotations in SwissProt

vs.

All annotations in SwissProt

vs.

Exp. annotations in UniProt (many exp. annotations in TrEMBL!)

vs.

(All annotations in UniProt - unfeasible)
We measured how much new template annotations in SwissProt improved predictions since CAFA1.

→ Exchange template SwissProt database “Jan 2013” with “Jan 2011”

→ Re-train + re-evaluation on targets added in 2013

Result

Max F1 relatively stable for BPO & CCO predictions

MFO: old database better by about 15% points

Same effect even for exactly the same free parameters
CAFA II – Old databases better?

Changes in the number of exp. GO annotated proteins from Jan 2011 to Jan 2013

<table>
<thead>
<tr>
<th></th>
<th>MFO</th>
<th>MFO'</th>
<th>BPO</th>
<th>MFO'+BPO</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Proteins</td>
<td>29540 → 22243 (25% red.)</td>
<td>21766 → 22243 (2% inc.)</td>
<td>22% inc.</td>
<td>17% inc.</td>
</tr>
<tr>
<td>Proteins lost entirely</td>
<td>39%</td>
<td>20%</td>
<td>5%</td>
<td>5%</td>
</tr>
<tr>
<td>Proteins reduced</td>
<td>29%</td>
<td>21%</td>
<td>18%</td>
<td>32%</td>
</tr>
<tr>
<td>68% have lost annotations!</td>
<td>41%</td>
<td>23%</td>
<td>37%</td>
<td></td>
</tr>
<tr>
<td>Proteins expanded</td>
<td>11%</td>
<td>12%</td>
<td>33%</td>
<td>37%</td>
</tr>
<tr>
<td>Proteins w/o GO:0005515</td>
<td>100%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© Tobias Hamp (TUM Munich)
## CAFA II – Data sets

<table>
<thead>
<tr>
<th>Category 1</th>
<th>Category 2</th>
</tr>
</thead>
</table>
| **SWISS-PROT*** | 232 benchmarks for MFO  
410 benchmarks for BPO  
608 benchmarks for CCO |
| **EBI** | 238 benchmarks for MFO  
281 benchmarks for BPO  
560 benchmarks for CCO |
| **SWISS-PROT + EBI + UNIPROT-GOA + GO*** | 656 benchmarks for MFO  
773 benchmarks for BPO  
991 benchmarks for CCO |
| **EBI +** | 239 benchmarks for MFO  
157 benchmarks for BPO  
306 benchmarks for CCO |
| **UNIPROT-GOA** | 667 benchmarks for MFO  
751 benchmarks for BPO  
637 benchmarks for CCO |
CAFA II – Data set statistics

EBI SEQUENCE IDENTITIES (MFO)
CAFA II – Data set statistics
CAFA II - Results

- The following slides show 6 results: \{EBI, Swiss-Prot\} x \{MFO, BPO, CCO\}
- 121 methods in total, top 10 shown
- All results under construction and subject to change
CAFA II - Results

SWISS-PROT: TOP 10 MODEL 1s

MFO

Swiss-Prot
CAFA II - Results

EBI: Top 10 Model 1s

MFO

EBI
CAFA II - Results

**SWISS-PROT: TOP 10 MODEL 1s**

- BPO

<table>
<thead>
<tr>
<th>Method</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tian_Lab (0.457)</td>
<td></td>
</tr>
<tr>
<td>ORENGO_FUNF_HMMER (0.449)</td>
<td></td>
</tr>
<tr>
<td>dcGO (0.444)</td>
<td></td>
</tr>
<tr>
<td>jfpred-PB (0.443)</td>
<td></td>
</tr>
<tr>
<td>jfpred-RF (0.443)</td>
<td></td>
</tr>
<tr>
<td>jfpred-FP (0.443)</td>
<td></td>
</tr>
<tr>
<td>dcGO_predictor (0.441)</td>
<td></td>
</tr>
<tr>
<td>Paccanaro_Lab (0.441)</td>
<td></td>
</tr>
<tr>
<td>PFPDB (0.434)</td>
<td></td>
</tr>
<tr>
<td>CONS (0.432)</td>
<td></td>
</tr>
<tr>
<td>BLAST2011 (0.272)</td>
<td></td>
</tr>
<tr>
<td>BLAST2014 (0.298)</td>
<td></td>
</tr>
<tr>
<td>Naive (0.305)</td>
<td></td>
</tr>
</tbody>
</table>

**Recall vs. Precision**
CAFA II - Results

EBI: Top 10 Model 1s

BPO model

- Tian_Lab (0.377)
- BonneauLab (0.349)
- TU (0.347)
- PaccanaroLab (0.342)
- profun (0.335)
- jfpred–RF (0.334)
- IASL_Academia_Sinica (0.333)
- jfpred–FP (0.332)
- orengo–mda (0.330)
- jfpred–PB (0.329)
- BLAST2011 (0.261)
- BLAST2014 (0.292)
- Naive (0.287)

Precision vs. Recall
CAFA II - Results

**SWISS-PROT: TOP 10 MODEL 1s**

![Graph showing precision for different models](image)

- metastudent2 (0.507)
- GoughGroup (0.493)
- EVEX (0.491)
- IASL_Academia_Sinica (0.487)
- Tian_Lab (0.486)
- PFPDB (0.486)
- CONS (0.485)
- Jpred-FP (0.482)
- BonneauLab (0.473)
- orengo-md (0.472)
- BLAST2011 (0.372)
- BLAST2014 (0.381)
- Naive (0.482)

**CCO**

Swiss-Prot
CAFA II - Results

EBI: Top 10 Model 1s

![Graph showing precision-recall curves for various models with CCO and EBI labels.](image)
## CAFA II - Results

### Ranks of Metastudent 2.0 among all methods

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Max. F1</th>
<th>SemSim</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MFO</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sprot</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>EBI</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>ALL</td>
<td>28</td>
<td>2</td>
</tr>
<tr>
<td><strong>BPO</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sprot</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>EBI</td>
<td>24</td>
<td>15</td>
</tr>
<tr>
<td>ALL</td>
<td>23</td>
<td>8</td>
</tr>
<tr>
<td><strong>CCO</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sprot</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>EBI</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>ALL</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>
## CAFA II - Results

### Ranks of Metastudent 2.0 among all methods

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Max. F1</th>
<th>SemSim</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MFO</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sprot</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>EBI</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td>ALL</td>
<td>28</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Max. F1</th>
<th>SemSim</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CCO</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sprot</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>EBI</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>ALL</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

**ALL targets & all ontologies & SemSim & Ranksum → 4th place (out of 55)**
Bias in computational biology
Bias in computational biology

CAFA

Own evaluation

Recall

All Targets

Precision

© Tobias Hamp (TUM Munich)
Bias in computational biology

- Bias in computational biology is...
  - Every non-formal mistake you can make
  - The lack of data power
  - Decreasing accuracy over time
  - What comes after publication

- Various types
  - Sequence bias
  - Experimental bias
  - Statistical bias
  - Training bias
  - Class bias
  - Circular information flow
  - PPI cross-validation bias
  - …
Methods work better if query sequences similar to training sequences

However: most un-annotated sequences dissimilar to annotated sequences

Databases often redundant (e.g. PDB)

$\rightarrow$ Tests (cross-validation, new data) mainly measure performance on easy queries

$\rightarrow$ Overestimation of actual prediction performance

Solution: test on new data + consider sequence similarity
Assume you wanted to know how many eukaryotic proteins have more than 1 domain.

How would you answer this question?
Experiments may distort reality; E.g.:
- Crystallization only possible for part of protein
- Interaction detection dependent on protein localization, protein expression
- Error rates in high-throughput experiments
- …

Fixing requires in-depth knowledge of experimental methods and/or selecting high-quality data subsets

Probably greatest factor in long-term changes of prediction accuracy
Bias in computational biology – Statistical bias

- If data is limited, performance estimates must be tested for statistical significance
- Statistical tests for virtually all types of predictions (Std. err., McNemar, Fisher, rank sums, …)
- However:
  - Choice of the test not always clear ("Choose std. dev. if you want errors to be large, std. err. if you want them to be small")
  - Data sets may be smaller than they appear (e.g. random samples from the PDB)
  - Smooth ROC curves despite lack of data
- Solution: knowledge of statistics, scientific honesty, rigorous peer-review
Bias in computational biology – Training bias

- Cheating in machine learning: using your test data to optimize the prediction model (≠ overfitting)
- → training, cross-training, test, holdout sets
- However: free parameters not only in machine learning: E.g.:
  - Blast parameters
  - Method design
  - …
- Easy to fool yourself
- Solution: tests on new data
Q: When is 99.9% prediction accuracy not a good result?
Bias in computational biology – Class bias

Q: When is 99.9% prediction accuracy not a good result?
A: If occurrence of one class >=99.9%
Distribution of class labels in existing data may not be the same as in natural data
→ Cross-validation and even tests on new data may fail
Solution: Adapt to actual class distribution as well as possible
Bias in computational biology –
Circular information flow

1. Use data A to predict or annotate data B
2. Use data B to predict or annotate data A
→ Good accuracy for A

Now, new data comes...

Example:
- Use mass-spec data to annotate protein complexes
- Use protein complexes to annotate proteins with CCO terms
- Use GO terms to predict protein complexes
- (Encore: use mass-spec data to validate complexes)
The accuracy for prediction a new protein-protein interaction (PPI) A-B depends on whether

- Both A and B (C1)
- Either A or B (C2)
- Neither A nor B (C3)

... were used for training the PPI predictor

Randomly sampling test PPIs from known interactions: most test PPIs in class C1 (most PPIs >1 interaction partner)

However: most PPIs in nature in class C3

Biased estimates even for non-redundant protein sets
Now PPI networks are even more complicated

- Protein
- Sequence similar
- Positive training IA
- Query IA
- Negative training IA
Avoiding every bias may be more difficult than developing the method.

... or simply impossible.

Computational biology is not alone.

Keep your eyes open when mining the literature.
THANKS!