Phylogenomic inference

Hauptseminar Frishman WS2013/2014

Uli Köhler

February 3rd 2014

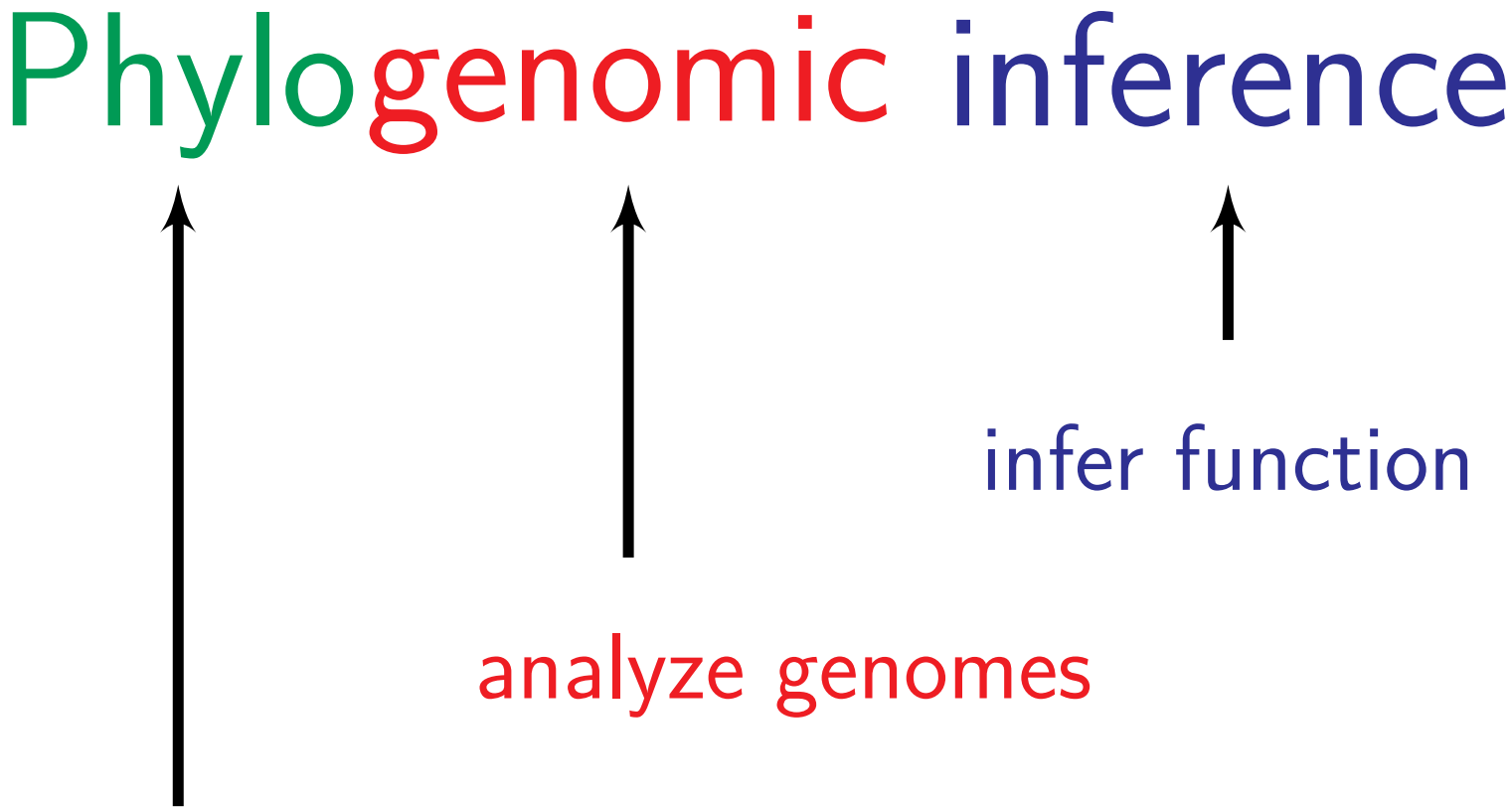
Structure of this talk

- Issues of non-phylogenic functional prediction
- What is phylogenomic inference?
- Phylogenetic tree reconciliation
- Phylogenomic inference methodology
- Phylogenomic databases and algorithms:
  - SIFTER
  - PhyloFacts
- Common problems of phylogenomic predictions
- Future of phylogenomics
- Seminar conclusion

Non-phylogenomic function prediction

- *High-throughput sequencing*
  $\rightarrow$ Many proteins, few information available:
  $\sim$90000 PDB structures vs 5.1×10^6 UniProt/TrEMBL sequences
- Alignment score does not distinguish between matching domains
- Difficult to separate *orthologs* and *paralogs*

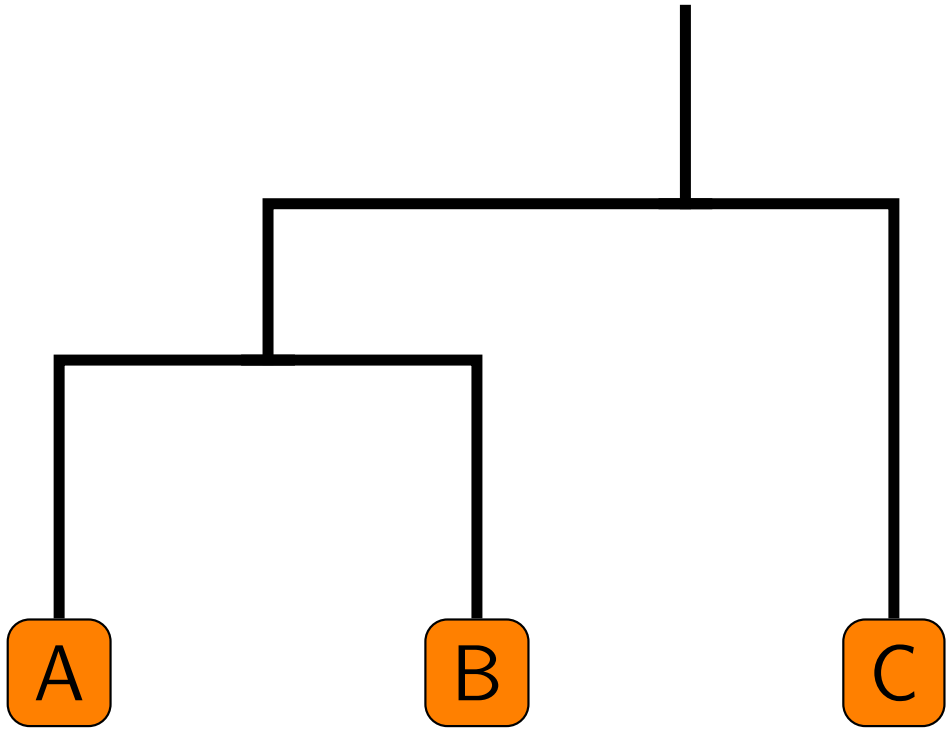
What is phylogenomic inference? I



What is phylogenomic inference? II

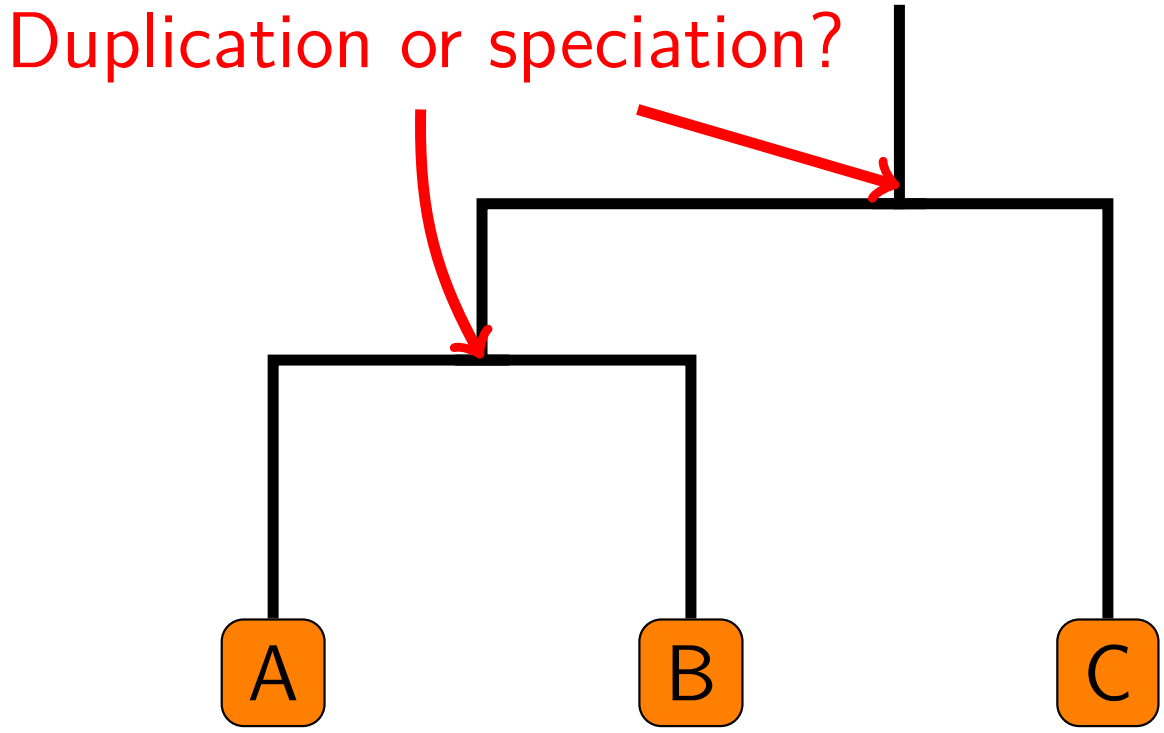
- Concept to enhance homology-based function predictions
- Can be applied to both genes and proteins
- Attempt to **separate *orthologs* and *paralogs*** $\rightarrow$ *ortholog* $=$ high probability of similar or identical function
- *Phylogenetic tree reconciliation*:
  Identify *speciation* and *duplication* events in phylogenetic trees

Tree reconciliation



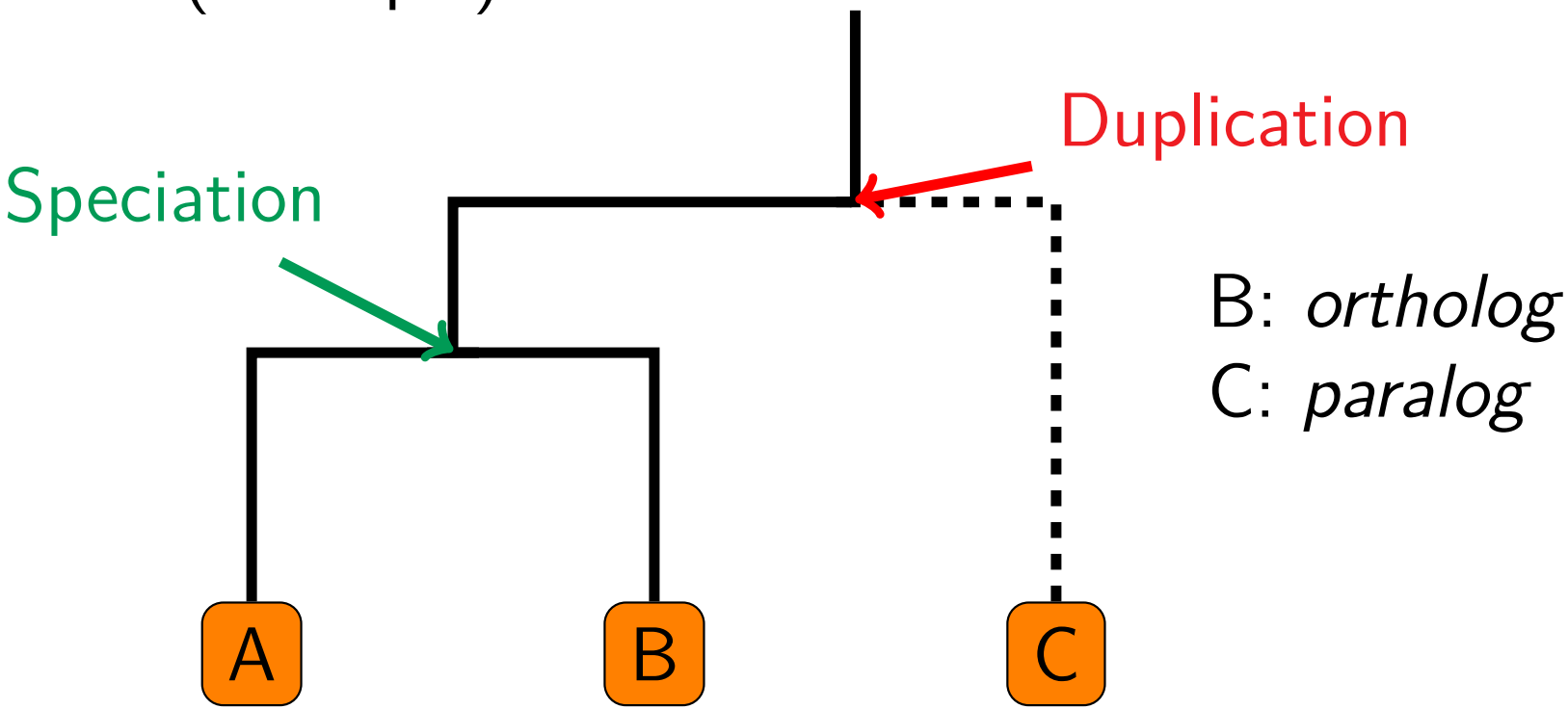
Are B and C *ortholog* or *paralog* in respect to A?

Tree reconciliation



Tree reconciliation

(Example)



B: *ortholog*
C: *paralog*

Phylogenomic inference methodology I

1. Cluster homolog proteins
2. Compute multiple alignment
3. Edit alignment (remove potential non-homologs)
4. Mask less-conserved regions in alignment
5. Construct phylogenetic tree
6. Identify closely related subtrees
7. Overlay with experimental data
8. Differentiate *orthologs* and *paralogs* (*Tree reconciliation*)
9. Infer function from *orthologs*

Phylogenomic inference methodology II

1. Cluster homolog proteins
2. Compute multiple alignment
3. Edit alignment
4. Mask less-conserved regions in alignment

- Raw alignments would introduce noise
- Retain only high-scoring homology & highly-conserved domains

Phylogenomic inference methodology III

5. Construct phylogenetic tree

- Core problems:
  - No information about actual ancestors is available
  - High computational complexity (optimal solution: NP-Hard!)
- Use algorithms like *maximum parsimony* or *maximum likelihood*

Phylogenomic inference methodology IV

6. Identify closely related subtrees
7. Overlay with experimental data

- More filtering to reduce noise
- Given the tree topology, use only closely related subgroups (in addition to filtering distant homologs in step 1)

Phylogenomic inference methodology V

8. Differentiate *orthologs* and *paralogs*

- Computational tree reconciliation – examples:
  - NCBI COG DB: Bidirectional top BLAST hits
  - Complex statistical algorithms like RIO (*Resampled inference of orthologs*), *orthostrapper* or *BETE*
- Computationally intensive, requires highly-filtered input data

SIFTER

9. Infer function from *orthologs*

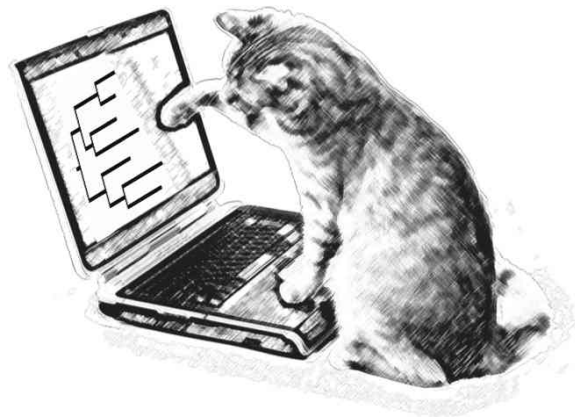
- *Statistical Inference of Function Through Evolutionary Relationships*
- Predicts protein function (homology-based) given a reconciled tree
  $\rightarrow$ Tree construction & reconciliation remains a problem
- Based on bayesian statistics
- Complex mathematics (not shown here)

PhyloFacts I

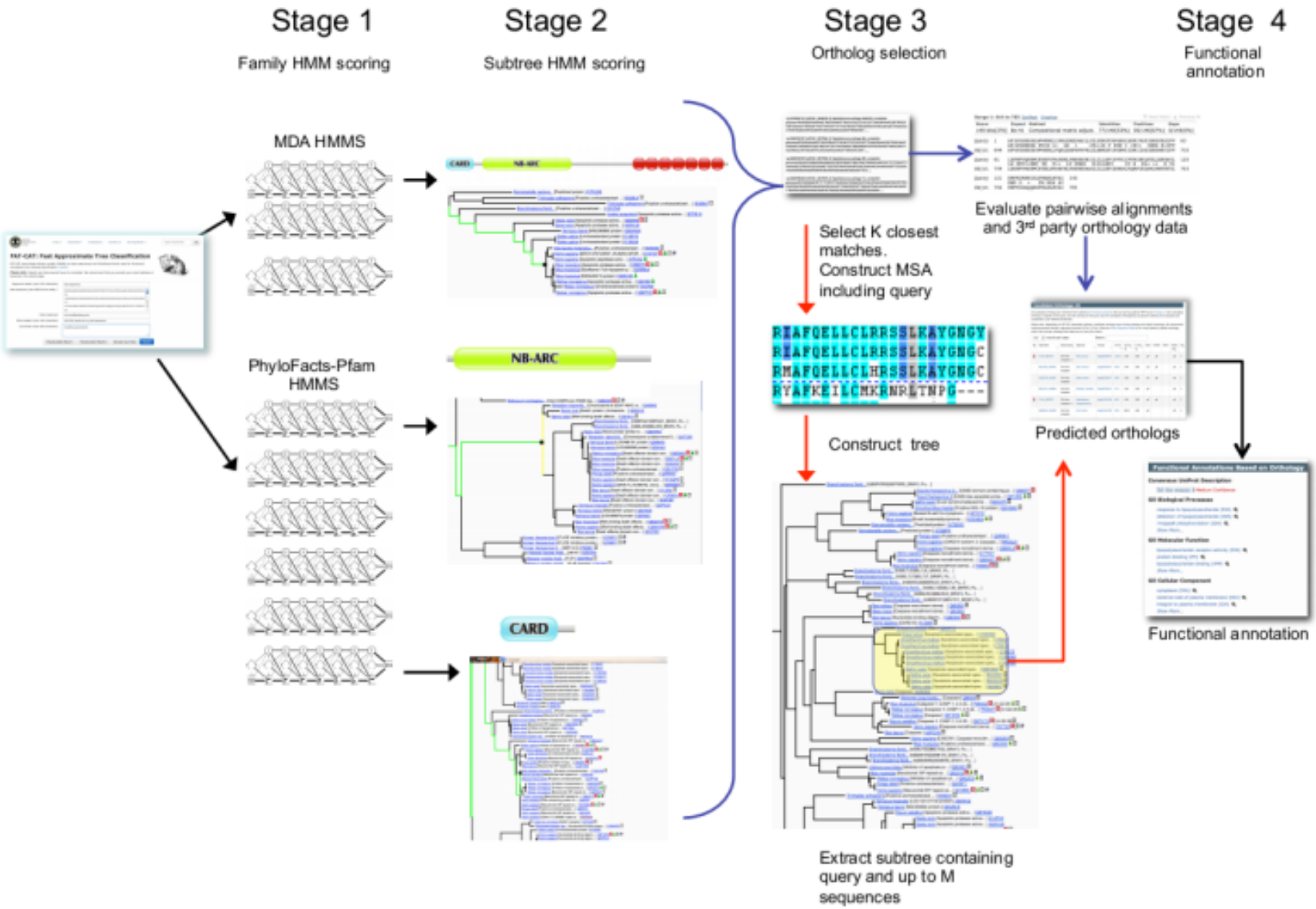
- „Encyclopedia“of „books“for known protein (super)families and structura domains
- 92800 families (as of 2013-02-03)
- Precomputed phylogenetic trees & phylogenomic family HMMs
  $\rightarrow$ Reasonably fast, but
  „*Some results can take hours to complete*“
- Provides structured access to annotated phylogenomic information about protein (super)families

PhyloFacts II

- **FAT-CAT**: PhyloFacts Webservice to predict protein function using phylogenomic methods
- Integrates with *Pfam* and uses HMMs to find the sequence position in the precomputed tree



PhyloFacts III



Issues of phylogenomic methods I

in-silico – Involves manual steps

1. Cluster homolog proteins
2. Compute multiple alignment
3. Edit alignment
4. Mask less-conserved regions in alignment
5. Construct phylogenetic tree
6. Identify closely related subtrees
7. Overlay with experimental data
8. Differentiate *orthologs* and *paralogs*
9. Infer function from *orthologs*

Issues of phylogenomic methods II

1. Cluster homolog proteins
2. Compute multiple alignment
3. Edit alignment
4. Mask less-conserved regions in alignment

- Manual annotation & selection
  $\rightarrow$ Subjective, error-prone, time/cost-intensive
- Information will be lost, does the annotator just select what he wants to see?
- Algorithms too sensitive, are results always reliable?

Issues of phylogenomic methods III

5. Construct phylogenetic tree

- *Distance-based* vs. *character-based* construction algorithms
- Small, highly-conserved protein families perform better than large (super)families
- Lack of consistency across methods
- Algorithms scale poorly $\rightarrow$ Can't be used for large (super)families
- Some methods produce millions of equivalently scored topologies

Issues of phylogenomic methods IV

7. Overlay with experimental data

- Database = Experimental data + inferred data
- Experimental datasets available $\leftrightarrow$ Protein function already know
- Protein function unknown $\leftrightarrow$ few experimental datasets available

Issues of phylogenomic methods V

- Multiple subsequent filter passes
- Huge sets of parameters, impossible to select optimal values
- Requires manual annotation & experimental data
- Sometimes even *orthology* is not sufficient for annotation transfer
- Doesn't work well with distant homologs, requires highly-conserved domains

Future of phylogenomic inference

- Phylogenomics alone has too many problems and open questions, but...

Future of phylogenomic inference

- Phylogenomics alone has too many problems and open questions, but...
- ...**together with other concepts** functional prediction accuracy can be enhanced
- Computational complexity: Moore's law and alternative computational hardware
  $\rightarrow$ Large-scale application feasible in the future?
- Phylogenomic inference for DB verification
- Can also be applied to other attributes (besides protein function)
- PhyloFacts & SIFTER: Usable tools, but apparently not widely adopted or actively developed

Conclusion (Phylogenomic inference)

- Powerful concept for enhancing function prediction accuracy by identifying *orthologs*

Conclusion (Phylogenomic inference)

- Powerful concept for enhancing function prediction accuracy by identifying *orthologs*
- ... if it would actually work in practice
- Too complex, too manual, too many parameters
- Pure *in-silico* phylogenomics
  $\rightarrow$ Low quality results
- Manual annotation can't keep up with *HTS*
- PhyloFacts provides a useful database for function prediction using phylogenomic approaches

Conclusion (Seminar)

- *in-silico* protein function inference is a yet unsolved problem in computational biology
- Combine any information that is available, including:
  - Context-based prediction
  - Alternative splicing
  - SNPs
  - Phylogenomics
  - Experimental results
- **Only with all this information combined sufficient accurracy for *in-silico* function prediction is achievable**

References

Kimmen Sjölander

Phylogenomic inference of protein molecular function: advances and challenges
Bioinformatics, 2004

Barbara E. Engelhardt et al.

Protein Molecular Function Prediction by Bayesian Phylogenomics
PLoS Computational Biology, 2005

Jonathan A. Eisen & Claire M. Frasier

Phylogenomics:Intersection of Evolution and Genomics
Science, 2003

Duncan Brown, Kimmen Sjölander

Functional Classification using Phylogenomic Inference
PLoS Computational Biology, 2006

Nandini Krishnamurthy et al.

PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification
Genome Biology, 2006

Barbara E. Engelhardt et al.

A graphical model for predicting protein molecular function
Proceedings of the International Conference on Machine Learning (ICML), 2006

Web & image sources

http://phylogenomics.berkeley.edu/

Thank you for your attention!

References and sources available at

`https://github.com/ulikoehler/Hauptseminar`

Questions?