Seminar report

# Phylogenomic inference

Uli Köhler – 2014-02-03

## Contents

# 1 Introduction

Since high-throughput sequencing is commercially available, computational biologist are facing an ever-increasing amount of genomic and proteomic data.

In order to leverage all this information for applications like in-silico drug target searching, advanced methods have been developed that use databases of known protein properties in order to infer properties of new proteins.

The most interesting property of any protein is its function – if the function would be known for any protein in the proteome, it would be relatively easy to derive potential drug targets from this information.

One of the most important concepts in this area of bioinformatics is homology-based function prediction: Proteins of high similarity have a high probability of having a similar or identical function that has possibly been conserved in the evolutional history of the organism.

However, in order to assess protein function with high significance, the analysis has to use information from all available sources in order to eliminate as many sources of error as possible.

This report will summarize Phylogenomics, a methodology to augment prediction quality by using phylogenetics in order to differentiate orthologs and paralogs, with the former having a higher probability of conserving their function during evolution.

# 2 Issues of non-phylogenomic function prediction

Classical function prediction methods are often built on the comparison of protein features like tertiary structure.

This concept is based on the assumption that if structural elements like domains are conserved in a protein family, there is a significant probability that the element is critical for the family-specific function.

For most proteins that are known today however, the tertiary structure is not known – the PDB database currently contains ~90000[1] structures, whereas ~$5.1 \times 10^6$ UniProt/TrEMBL[2] sequences.

This huge discrepancy is mainly based on the expensiveness of structural assessment methods like X-ray crystallography that not only require a specialized laboratory, but also – for each individual experiment – expert knowledge to crystallize the protein and assess the X-ray crystallogram. To date, only few advances have been made in automation of those methods (a summary is provided at [Gro07]) and most of them require a high-quality X-ray beam only available in particle accelerators like Synchrotrons which are not available to most laboratories.

In the future, this discrepancy is expected to increase even more, because commercially available sequencers drop in price rapidly and allow semi-automatic sequencing with a speed of several Gigabases per day. Besides these devices yielding a large amount of predicted proteins, recent advances have been made in an area called *shotgun proteomics* that use shotgun sequencing methods well-known from genomic sequencing methodologies in order to sequence proteoms directly. As described in [WM+02], these methods are available since more than ten years and will.

Therefore, in order to cope with the massive amount of sequence-only proteins, computational biology tries to infer protein functions from sequence and relations to known proteins.

[1] 2014-02-03

[2] 2014-02-03, combined

**Systematic errors of non-phylogenomic prediction** While those methods yield high-quality predictions for a large number of proteins, there are several systematic errors outlined in [BS06]:

- *Gene duplication* is not taken into account, yielding a predicted orthology, while the proteins would have to be considered paralog
- *Domain shuffling* events that cause highly similar sequences to have different protein functions due to a largely different structure
- *Evolutionary distance* is not used to augment the prediction, therefore the chance of homolog proteins having the same function in distantly-related species is overrated.

# 3 Phylogenomic methodology

Phylogenomics attempts to minimize the influence of aforementioned errors by improving homology metrics using phylogenetic information that provides knowledge about the evolutionary history of the protein in question.

The core concept of phylogenomics is the differentiation of homologs into orthologs and paralogs.

Orthologs have a higher probability of conserving function over evolution, whereas two proteins identified as paralog usually have different functions.

If sufficient data is available to build a phylogenetic tree[3], phylogenomic algorithms classify branching points in the tree as either duplication or speciation, with the former yielding paralog proteins whereas the latter yields orthologs (see [EF03]). This process is called *tree reconciliation*

A conceptual example of what is classified is visualized in figure 1.

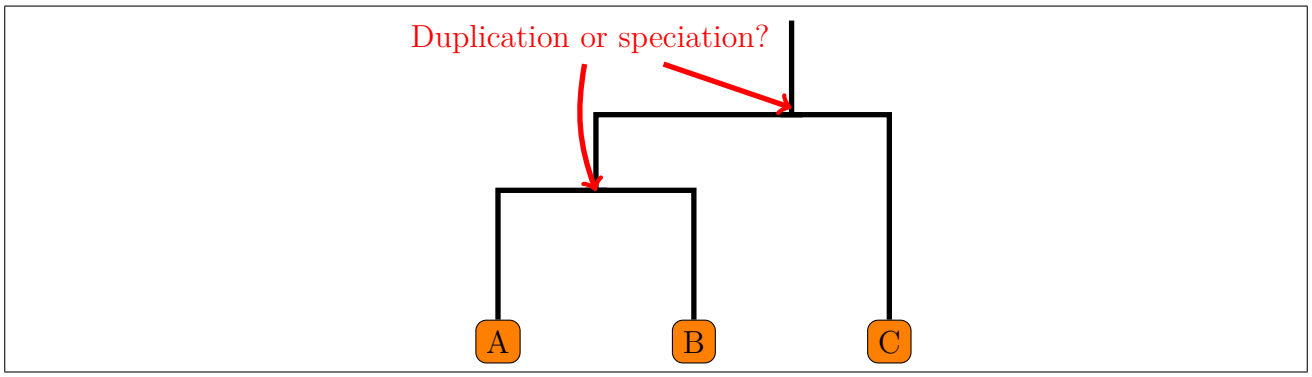


**Figure 1:** Concept of Phylogenomics: Differentiating orthologs and paralogs

**Phylogenomic workflow** A typical workflow for phylogenomic analysis, as described in [Sjö04], consists of ten steps, some of which are shared with classical methods.

1. Cluster homolog proteins
2. Compute multiple alignment

[3]Generally, the tree can be deduced from a sequence set alone, however this approach w ould introduce too much noise

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*

The first two steps are, to some extent, shared with most classical methods, however, contrary to said classical methods, the described workflow involves manual steps, most notably steps 3), 4) and 7).

**Phylogenetic tree construction** Algorithms to construct a phylogenetic tree can be classified into two main subgroups: *distance-based* and *character-based* methods.

Whereas distance-based methods only compute pairwise distances between any two sequences and construct this tree solely based on the distance matrix – therefore gaining a huge advantage in computational complexity – character-based methods have advantages regarding accuracy, but are too slow for practical use on medium to large trees (see [Sjö04]).

A detailed comparison of distance-based and character-based methods can be found in [Fel78].

Besides the high computational complexity of both classes of methods, there are other severe issues with tree construction in general (as outlined in [Sjö04, p. 4f]):

- Trees generated with different methods are inconsistent with each other[4]
- For some algorithms, several millions of trees have the same score even for medium-sized datasets. Often, those trees have major differences, making it impossible to select a consistent topology)
- Small, highly-conserved protein families perform better than large (super)families

**Filtering overview** The steps in the workflow that have not been described so far have the purpose to remove noise from the input data. However, this concept yields a number of issues that are described in section 5 on page 7

Because of the limited extent of this report, these filter steps are not explained in detail here. See [Sjö04, p. 3ff] for a detailed description of the purpose and algorithms for each individual step.

**Manual steps** According to [Sjö04] his specific workflow proved to yield better results than alternatives (especially those that don't involve manual steps). Issues in relation to manual aspects of the analysis will be discussed in detail in section 5 on page 7.

[4]It therefore can be assumed, that for real, large-scale data, the trees are unalike the true, unknown biological tree

# 4 Phylogenomic tools & databases

## 4.1 SIFTER

*SIFTER* (see [EJMB05]) is a tool to infer protein function from a reconciled phylogenetic tree based on bayesian statistics.

As the mathematical methodology of *SIFTER* is highly complex and would require many pages of introductory mathematics, this report will not describe in detail how *SIFTER* works internally.

However, it should be noted that *SIFTER* circumvents most of the general phylogeny issues by only solving the last step of the workflow outlined in section 3 on page 3. While SIFTER itself may be mathematically correct and work well on test data, the rest of the phylogenomic workflow will often provide a low-quality reconciled tree as input for *SIFTER*, effectively rendering the output data useless.

## 4.2 PhyloFacts

*PhyloFacts*, first described in [KBKS06] is an "Encyclopedia" of "books" for known protein (super)families and structural domains. Since its initial publication, it changed significantly and now provides the *FAT-CAT* webserver (see [ASD+13]) that allows interactive protein classification.

Although it is published by the same group at Berkeley University as the workflow (see [Sjö04]) and *SIFTER* (see [EJMB05]), it circumvents most of the complexity by using a pre-computed phylogenetic tree and tries to place the query protein inside an appropriate position in said tree.

As the tree is already reconciled (even if assessing the quality of the reconciliation seems to be error-prone considering the results of this report, yet outside the scope of this report), *FAT-CAT* can use methods like *SIFTER*[5].

The full workflow used by fatcat is visualized in figure 2 on the following page[6]. A highly-detailed description of their methodology, is not within the scope of this report, however it's noteworthy that they first perform a family search using hidden markov models in order to place the query sequence in an appropriate position in the reconciled tree.

The red arrows in figure 2 on the next page denote the workflow of *FAST-CAT*, a new, experimental variant of *FAT-CAT* that avoids some of the computational complexity of *FAT-CAT*.

**How useful is PhyloFacts?** Within the scope of this report, there are very few aspects that allow to make assumptions about the actual usefulness of SIFTER.

Because these results are likely not significant, this report will list some aspects that seem related to the usefulness without drawing any conclusions.

- PhyloFacts is not open-sourced – in the opinion of the author of this report, it therefore clearly violates the scientific principle of reproducibility. It will be almost impossible to find bugs in both the implementation and the data basis in a complex application like PhyloFacts if the code is not provided to the general public.

[5] In [KBKS06] the authors state that they intend to use *SIFTER* itself in the future, yet it is unknown whether it is actually used in the newest version

[6] Image source: http://makana.berkeley.edu/phylofacts/fatcat/about/

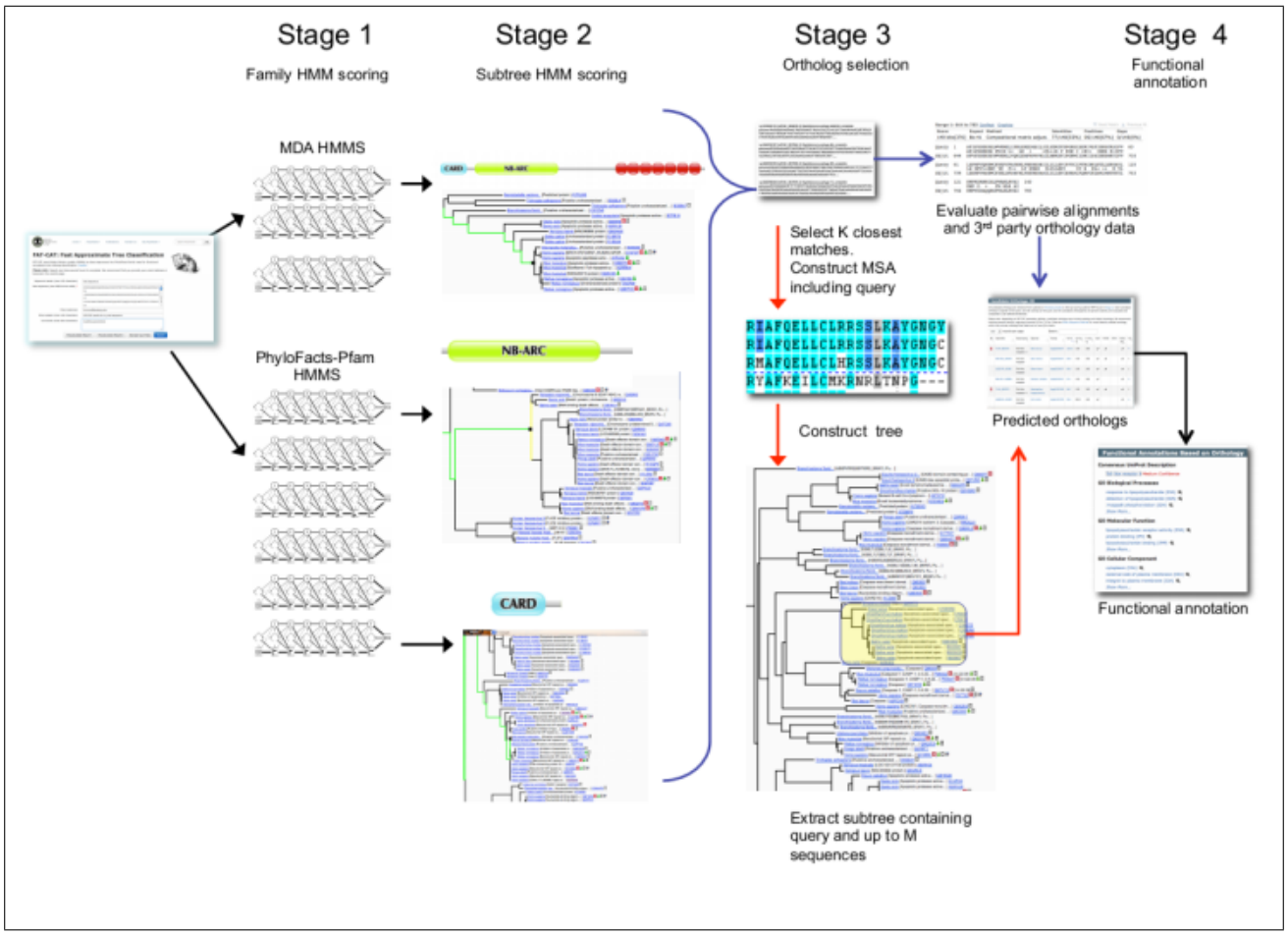


**Figure 2:** The *FAT-CAT* workflow pipeline

- Within the 3-day long *FAT-CAT* research for the talk corresponding to this report, the job ID numbers of FAT-CAT ascended monotonically, and increased (on average) by only about 10 per day[7]. It can therefore be assumed that PhyloFacts is not very widely used
- [KBKS06] has been cited only 27 times since 2006. While this is more than e.g. *SIMAP* (see [ART+05]) which has been cited 18 times since 2005 and since 2011 been used as STRING data source (see [SFK+11]), it can be assumed that a widely and publicly adopted tool would be cited more often
- The runtime of the tool takes dozends of minutes even for $< 300$ AA query sequences
- For multiple randomly selected predicted *TrEMBL* sequences[8], PhyloFacts did not return any results beside PFam families (said families equivalent to those listed in *UniProt*). Therefore, it did not predict any function for these queries. This aspect might require further research, but raises doubts whether PhyloFacts is useful also for de-novo analyses that don't use known proteins with close relatives in the phylogenetic tree
- The SIFTER homepage (`http://sifter.berkeley.edu/`) says "Check back here for a SIFTER server sometime in the next few months!" since 2006. Although PhyloFacts seems

[7] *STRING* had more than 80000 queries per day, averaged over a 7-day period
[8] Example: A1ULI6_MYCSK

to be updated in several-year intervals, it is neither clear when the underlying database is updated (and, for example, if the parameters change) nor can anyone predict the future of the tool. The "Encyclopedia" has completely disappeared since the initial PhyloFacts release.

- Conceptually, new sequences require the reconciled tree to incorporate close relatives. For many biologically relevant applications it is possible the tree doesn't contain any such sequence, and it's impossible to reproduce based on a specialized dataset because the application source code is not publically accessible

What seems to be useful about PhyloFacts is the attempt of the database to integrate information (especially phylogenetic information) from different sources (e.g. PFAM). If a user has a phylogeny-related question about a query sequence (e.g. *distant clades*), PhyloFacts seems to be able to provide a user-friendly interface (e.g. for biologists that don't want to deal with complex bioinformatics toolsets), as depicted in figure 3.

Additionally, the *FAT-CAT* query supports selectable preset parameters, for example *High recall* and *High precision*.

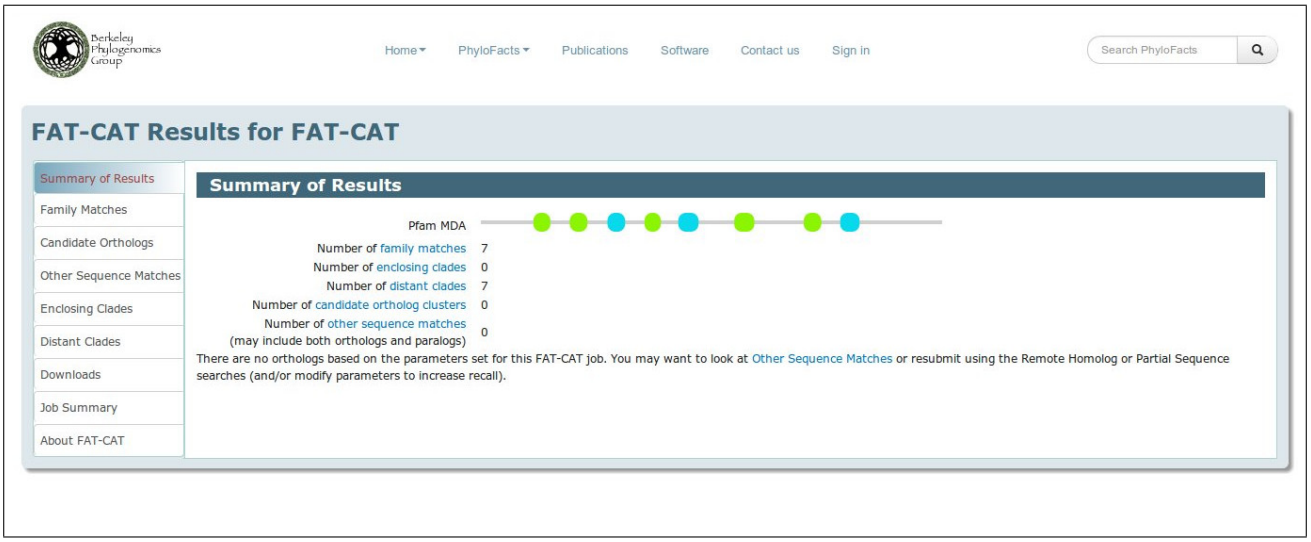


**Figure 3:** The *FAT-CAT* result page

# 5 Common problems of phylogenomics

**Manual annotation requirement** As described in section 3 on page 3 the workflow described in [Sjö04]. While this might improve the quality of the result, essentially it renders the method useless in the context of large-scale analysis as described in section 2 on page 2.

Because of limited availability of expert human resources, phylogenomic analyses with sufficient quality can't be performed on a large scale, even if the computational complexity of the automatable steps would not be an issue.

**Huge parameter sets** The workflow described in [Sjö04] incorporates several layers of filtering that depend on a large parameterset and even on human factors. It can be assumed that for any real analysis being performed, a large subset of the parameters is chosen suboptimally, because

no inherently correct method is known that allows one to choose an optimal parameterset of that size.

In order to achieve optimal results, the parameters would need to be selected for each prediction individually – however, that is not possible in lack of an absolute reference (representing the biological reality). Almost all algorithms in bioinformatics suffer from similar problems, however in case of phylogenomics the parameterset is – caused by the large number of steps – extremely large and contains a large number of known and unknown interdependencies..

Because of the human factor involved in the manual steps, a result may be irreproducible by another researcher, rendering the results subjective.

**Parameter sensitivity** Although a detailed analysis is outside the scope of this report, in [Sjö04] it is observed that the quality of the results depends on all the filtering passes being present in the workflow. This leads to the conclusion that the non-filter steps are so sensitive to noise that any minimal change in the input data could yield a totally different result.

Because interdependencies in the parameters are largely unknown and the high dimensionality of the parameter space, it is essentially impossible to conclude that any result is not the case of an overyl

This issue gets even more significant in cotext of manual processing: Even if assuming two researchers will select similar options for the same input data, a minor change in the parameters they use could cause the result to change significantly.

**Loss of information** Effectively, the huge parameter sets create a paradox: The phylogenomic method which is intended to augment a prediction by adding new information actually loses a significant amount of information in the filtering passes – up to an extent where the prediction quality is detrimente. For real predictions (i.e. those lacking an absolute reference), this effect is impossible to detect

**Deducing information from itself** Non-phylogenomic methods usually work well for closely-related protein families that will only infrequently express behaviour that yields systematic errors as outlined in section 2 on page 3. It can be assumed that if a large amount of information is available about a protein, classical methods provide a sufficient prediction quality, rendering phylogenomics useful especially for cases where few information is available.

However, few information being available could hypothetically be detrimental to the phylogenetic tree – in many cases the information contained in said tree will be derived only from the protein family sequence set itself. However, the same exact sequence set will be used for the classical prediction that will be augmented by phylogenetic information.

Therefore, in an extreme the prediction would actually be augmented with information it already contains, rendering the augmentation useless and (under some circumstances) even decreasing the prediction quality because some aspects of the original prediction are overrated in the final score

It is outside of this report's scope to discuss or assess the effect of this behaviour in general, however it shall be noted that most augmentation methods express a similarly problematic behaviour. The author of this report assumes it is highly problematic to filter out these causality loops, because detailed information about the data sources is not available for any database.

**Resolving database issues using phylogenomics** While [BS06] mentions – besides the systematic errors outlined in section 2 on page 3 – the "propagation of errors in databases" as a systematic error that can be detrimental to data quality, there seems to be no apparent reason why phylogenomics should be better in resolving those issues than any other method leveraging additional data sources.

To the author of this report, the problems outlined before seem to be too severe to conclude without any doubt that phylogenomics will resolve database issues instead of creating even more ones and making manual result assessment difficult by adding yet another, potentially inaccurate layer of complexity to the prediction.

# 6 Conclusion & Outlook

Phylogenomics is a method that attempts to augment function prediction by leveraging evolutional information to differentiate orthology and paralogs. Even if a full analysis based on carefully selected test data is outside the scope of this report, the issues described in this report arise doubts whether phylogenomics provides a tool that can be used for real data.

However, issues like the high computational complexity might be partially resolved by advancements in computational hardware or algorithms in the future. *PhyloFacts* provides an example of how a computationally-feasible phylogenomic tool can be predicted, although the quality of the results in contrast to other databases

In practical function prediction, however, it's neccessary to combine any available information and method. Although care has to be taken to ensure the issues of phylogenomics are not detrimental to the result, in many cases other information sources like *context-based prediction* are not sufficient for a significant result.

Therefore, disregarding any practical issues, phylogenomics provide one of many conceptual sources of information, all of which have to be combined to improve in-silico predictions.

# Bibliography

[ART+05] Arnold, Roland ; Rattei, Thomas ; Tischler, Patrick ; Truong, Minh-Duc ; Stümpflen, Volker ; Mewes, Werner: SIMAP—the similarity matrix of proteins. In: *Bioinformatics* 21 (2005), Nr. suppl 2, S. ii42–ii46

[ASD+13] Afrasiabi, Cyrus ; Samad, Bushra ; Dineen, David ; Meacham, Christopher ; Sjölander, Kimmen: The PhyloFacts FAT-CAT web server: ortholog identification and function prediction using fast approximate tree classification. In: *Nucleic acids research* 41 (2013), Nr. W1, S. W242–W248

[BS06] Brown, Duncan ; Sjölander, Kimmen: Functional classification using phylogenomic inference. In: *PLoS computational biology* 2 (2006), Nr. 6, S. e77

[EF03] Eisen, Jonathan A. ; Fraser, Claire M.: Phylogenomics: intersection of evolution and genomics. In: *Science* 300 (2003), Nr. 5626, S. 1706–1707

[EJMB05] Engelhardt, Barbara E. ; Jordan, Michael I. ; Muratore, Kathryn E. ; Brenner, Steven E.: Protein molecular function prediction by Bayesian phylogenomics. In: *PLoS computational biology* 1 (2005), Nr. 5, S. e45

[Fel78] Felsenstein, Joseph: Cases in which parsimony or compatibility methods will be positively misleading. In: *Systematic Biology* 27 (1978), Nr. 4, S. 401–410

[Gro07] Groves, Matthew R.: Recent Advances in Automation of X-Ray Crystallographic Beamlines at the Embl Hamburg Outstation. In: *Brilliant Light in Life and Material Sciences*. Springer, 2007, S. 133–139

[KBKS06] Krishnamurthy, Nandini ; Brown, Duncan P. ; Kirshner, Dan ; Sjölander, Kimmen: PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. In: *Genome biology* 7 (2006), Nr. 9, S. R83

[SFK+11] Szklarczyk, Damian ; Franceschini, Andrea ; Kuhn, Michael ; Simonovic, Milan ; Roth, Alexander ; Minguez, Pablo ; Doerks, Tobias ; Stark, Manuel ; Muller, Jean ; Bork, Peer u. a.: The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. In: *Nucleic acids research* 39 (2011), Nr. suppl 1, S. D561–D568

[Sjö04] Sjölander, Kimmen: Phylogenomic inference of protein molecular function: advances and challenges. In: *Bioinformatics* 20 (2004), Nr. 2, S. 170–179

[WM+02] Wu, Christine C. ; MacCoss, Michael J. u. a.: Shotgun proteomics: tools for the analysis of complex biological systems. In: *Curr Opin Mol Ther* 4 (2002), Nr. 3, S. 242–250

## Open Data

All data used in the creation of this report can be accessed at
`https://github.com/ulikoehler/Hauptseminar`